

XOPSIS: An Explainable AI Method based on the Order of Preference by Similarity to Ideal Solution

by

Anika Rahman
21266028

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
M.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
July 2023

© 2023. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

A handwritten signature in black ink, appearing to read 'Anika Rahman', is written over a horizontal line.

Anika Rahman
21266028

Approval

The thesis titled “XOPSIS: An Explainable AI Method based on the Order of Preference by Similarity to Ideal Solution”

Submitted by:

Anika Rahman (21266028)
E-mail: anika.rahman@g.bracu.ac.bd

Of Summer, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science on July 26, 2023.

Examining Committee:

External Examiner:
(Member)

Prof. Mohammad Mehedi Hassan, Ph.D.
Professor
Information Systems Department
King Saud University, Riyadh, Kingdom of Saudi Arabia
E-mail: mmhassan@ksu.edu.sa

Internal Examiner:
(Member)

Dr. Amitabha Chakrabarty
Professor
Department of Computer Science and Engineering
BRAC University
E-mail: amitabha@bracu.ac.bd

Internal Examiner:
(Member)

Dr. Md. Ashraful Alam
Associate Professor
Department of Computer Science and Engineering
BRAC University
E-mail: ashraful.alam@bracu.ac.bd

Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
BRAC University
E-mail: rabiul.alam@bracu.ac.bd

Chairperson:
(Member)

Sadia Hamid Kazi, Ph.D.
Associate Professor
Department of Computer Science and Engineering
BRAC University
E-mail: skazi@bracu.ac.bd

Abstract

Explainable AI (XAI) techniques are essential for comprehending machine learning model predictions in a variety of fields, including healthcare. In this study, we highlight XOPSIS, a newly developed XAI method built on the TOPSIS score and intended to offer thorough justifications for gradient boosting models. We compare the performance of XOPSIS with two established XAI techniques, LIME and SHAP, using a comprehensive dataset encompassing different domains, including maternal health records, the benchmark Iris dataset, and the benchmark breast cancer dataset. Our findings demonstrate significant similarities between XOPSIS and LIME in generating explanations and consistently identifying the most influential features contributing to the model's predictions. In addition, by utilizing SHAP values, we acquire a comprehensive comprehension of the model's behavior and the unique contributions of each feature to the predictions. The significance of our proposed approach lies in its ability to enhance interpretability in machine learning models, enabling stakeholders to make informed decisions across various domains. While we showcase the effectiveness of XOPSIS in maternal health risk prediction, the benchmark Iris dataset, and breast cancer diagnosis, its applicability extends to diverse domains such as finance, cybersecurity, and customer behavior analysis. The flexibility and generalizability of XOPSIS make it a valuable tool for understanding the underlying factors driving model predictions. In addition to the maternal health records, the benchmark Iris dataset, and the breast cancer dataset, we also apply XOPSIS to the Car Acceptability dataset, further expanding its applicability across diverse domains. By including this dataset, we demonstrate the versatility of XOPSIS in providing comprehensive explanations for machine learning models in various contexts. Furthermore, future studies should focus on the practical implementation of XOPSIS in different domains, evaluating its effectiveness in real-world scenarios, and assessing its impact on decision-making processes. Furthermore, exploring the integration of XOPSIS in various industries and applications can provide valuable insights into the interpretability and transparency of machine learning models. By advancing XAI techniques like XOPSIS, we can foster trust, accountability, and widespread adoption of AI technologies in diverse fields, ultimately benefiting both industry practitioners and end-users. The continued development and refinement of XOPSIS and similar XAI methods will contribute to the responsible and ethical use of AI, promoting transparency and understanding in complex machine learning models.

Keywords: Explainable AI; XOPSIS; LIME; SHAP; Maternal Health; Interpretability; Transparency

Dedication

This thesis is dedicated to all individuals whose lives are impacted by the application domains explored in this research: maternal health, and breast cancer diagnosis. It is a heartfelt dedication to those who face the challenges and uncertainties of these domains, whether as patients, healthcare professionals, or individuals striving to make informed decisions. May the knowledge gained from this work contribute to improving the well-being and outcomes of those affected, and may it serve as a reminder of the importance of research and innovation in addressing critical issues in various fields.

Acknowledgement

My heartfelt gratitude to my supervisor, Dr. Md. Golam Rabiul Alam, Professor, Department of Computer Science and Engineering, BRAC University, for his unwavering support throughout this research endeavor.

I'd like to thank my parents and the rest of my family members for their support. I would not be in this position now if it weren't for them.

I'd want to express my gratitude to all of the faculty members, staff, students, and other BRAC University stakeholders for their direct and indirect aid during various events of my education and research.

Table of Contents

Declaration	i
Approval	ii
Ethics Statement	iv
Abstract	iv
Dedication	v
Acknowledgment	vi
Table of Contents	vii
List of Figures	x
List of Tables	xiv
Nomenclature	xv
1 Introduction	1
1.1 Contributions	2
2 Background Study	3
2.1 LIME	3
2.2 SHAP	4
2.3 TOPSIS	5
3 Related Work	8
3.1 Explainable AI	8
3.2 Maternal Health	10
3.3 IRIS Species	12
3.4 Breast Cancer	12
3.5 Car Acceptability	13
4 Methodology	14
4.1 Proposed Explainable AI Method	14
4.1.1 Algorithm Application on a Small Dataset	17
4.2 Maternal Health Risk Prediction and Explanation	24
4.2.1 Data Collection	24
4.2.2 Dataset Preprocessing	24

4.2.3	Model Training	26
4.2.4	Random Forest	28
4.2.5	XGBoost	28
4.3	IRIS Species Classification and Explanation	29
4.3.1	Dataset Description	29
4.3.2	Dataset Preprocessing	30
4.3.3	Model Training	31
4.4	Breast Cancer Prediction and Explanation	32
4.4.1	Dataset Description	32
4.4.2	Dataset Preprocessing	33
4.4.3	Model Training	35
4.5	Car Acceptability Prediction and Explanation	35
4.5.1	Data Collection	35
4.5.2	Dataset Preprocessing	36
4.5.3	Model Training	37
5	Result Analysis	39
5.1	Experimental Setup	39
5.2	Result Analysis on Maternal Health Dataset	40
5.2.1	Accuracy	40
5.2.2	Confusion Matrix	40
5.2.3	Sensitivity, Specificity, Precision and F1-Score	42
5.2.4	ROC Curve	44
5.2.5	AUC	44
5.2.6	Feature Importance Plot	45
5.2.7	Interpretability Through XAI Methods	47
5.3	Result Analysis on Iris Dataset	72
5.3.1	Accuracy	72
5.3.2	Confusion Matrix	73
5.3.3	Sensitivity, Specificity, Precision and F1-Score	74
5.3.4	ROC Curve	75
5.3.5	AUC	76
5.3.6	Interpretability Through XAI Methods	77
5.4	Result Analysis on Breast Cancer Wisconsin Dataset	91
5.4.1	Accuracy	91
5.4.2	Confusion Matrix	91
5.4.3	Sensitivity, Specificity, Precision and F1-Score	93
5.4.4	ROC Curve	94
5.4.5	AUC	94
5.4.6	Interpretability Through XAI Methods	96
5.5	Result Analysis on Car Acceptability Dataset	110
5.5.1	Accuracy	110
5.5.2	Confusion Matrix	111
5.5.3	Sensitivity, Specificity, Precision and F1-Score	112
5.5.4	ROC Curve	114
5.5.5	AUC	115
5.5.6	Interpretability Through XAI Methods	116
5.6	Time Complexity Analysis	127

6 Conclusion	129
Bibliography	134

List of Figures

4.1	Top Level Overview of the XOPSSIS Method	15
5.1	Train-Test Split Method Accuracy of 14 Algorithms	41
5.2	Confusion Matrix of Gradient Boosting Model	42
5.3	Comparison of Sensitivity, Specificity, Precision, and F1-Score Across Three Risk Levels	43
5.4	ROC Curve for Gradient Boosting	44
5.5	Area Under the Curve (AUC)	45
5.6	Feature Importance Ranking using SHAP Summary Plot with Ran- dom Forest	46
5.7	Feature Importance Ranking using SHAP Summary Plot with XGBoost	46
5.8	Feature Importance Ranking using TOPSIS Method	47
5.9	Explanation of high risk level by XOPSSIS (Case 01)	47
5.10	Normalized Feature Contributions of high risk level by XOPSSIS (Case 01)	48
5.11	Explanation of high risk level by LIME (Case 01)	48
5.12	Explanation of high risk level by SHAP (Case 01)	49
5.13	Explanation of high risk level by XOPSSIS (Case 02)	50
5.14	Normalized Feature Contributions of high risk level by XOPSSIS (Case 02)	51
5.15	Explanation of high risk level by LIME (Case 02)	51
5.16	Explanation of high risk level by SHAP (Case 02)	52
5.17	Explanation of high risk level by XOPSSIS (Case 03)	53
5.18	Normalized Feature Contributions of high risk level by XOPSSIS (Case 03)	53
5.19	Explanation of high risk level by LIME (Case 03)	54
5.20	Explanation of high risk level by SHAP (Case 03)	54
5.21	Explanation of mid risk level by XOPSSIS (Case 04)	55
5.22	Normalized Feature Contributions of mid risk level by XOPSSIS (Case 04)	56
5.23	Explanation of mid risk level by LIME (Case 04)	56
5.24	Explanation of mid risk level by SHAP (Case 04)	57
5.25	Explanation of mid risk level by XOPSSIS (Case 05)	58
5.26	Normalized Feature Contributions of mid risk level by XOPSSIS (Case 05)	58
5.27	Explanation of mid risk level by LIME (Case 05)	59
5.28	Explanation of mid risk level by SHAP (Case 05)	59
5.29	Explanation of mid risk level by XOPSSIS (Case 06)	61

5.30	Normalized Feature Contributions of mid risk level by XOPSSIS (Case 06)	61
5.31	Explanation of mid risk level by LIME (Case 06)	62
5.32	Explanation of mid risk level by SHAP (Case 06)	63
5.33	Explanation of low risk level by XOPSSIS (Case 07)	64
5.34	Normalized Feature Contributions of low risk level by XOPSSIS (Case 07)	64
5.35	Explanation of low risk level by LIME (Case 07)	65
5.36	Explanation of low risk level by SHAP (Case 07)	65
5.37	Explanation of low risk level by XOPSSIS (Case 08)	66
5.38	Normalized Feature Contributions of low risk level by XOPSSIS (Case 08)	67
5.39	Explanation of low risk level by LIME (Case 08)	67
5.40	Explanation of low risk level by SHAP (Case 08)	68
5.41	Explanation of low risk level by XOPSSIS (Case 09)	69
5.42	Normalized Feature Contributions of low risk level by XOPSSIS (Case 09)	70
5.43	Explanation of low risk level by LIME (Case 09)	70
5.44	Explanation of low risk level by SHAP (Case 09)	71
5.45	Train-Test Split Accuracy for Different Algorithms on Iris Dataset	73
5.46	Confusion Matrix for Iris Dataset	73
5.47	Performance Evaluation Metrics on Iris Dataset	74
5.48	ROC Curve for Gradient Boosting Model on Iris Dataset	75
5.49	AUC Curve for Gradient Boosting Model on Iris Dataset	76
5.50	Explanation of Iris-Setosa level by XOPSSIS (Case 01)	77
5.51	Normalized Feature Contributions of Iris-Setosa level by XOPSSIS (Case 01)	78
5.52	Explanation of Iris-Setosa level by LIME (Case 01)	78
5.53	Explanation of Iris-Setosa level by SHAP (Case 01)	79
5.54	Explanation of Iris-Setosa level by XOPSSIS (Case 02)	80
5.55	Normalized Feature Contributions of Iris-Setosa level by XOPSSIS (Case 02)	80
5.56	Explanation of Iris-Setosa level by LIME (Case 02)	81
5.57	Explanation of Iris-Setosa level by SHAP (Case 02)	81
5.58	Explanation of Iris-Versicolor level by XOPSSIS (Case 03)	82
5.59	Normalized Feature Contributions of Iris-Versicolor level by XOPSSIS (Case 03)	82
5.60	Explanation of Iris-Versicolor level by LIME (Case 03)	83
5.61	Explanation of Iris-Versicolor level by SHAP (Case 03)	83
5.62	Explanation of Iris-Versicolor level by XOPSSIS (Case 04)	84
5.63	Normalized Feature Contributions of Iris-Versicolor level by XOPSSIS (Case 04)	85
5.64	Explanation of Iris-Versicolor level by LIME (Case 04)	85
5.65	Explanation of Iris-Versicolor level by SHAP (Case 04)	86
5.66	Explanation of Iris-Virginica level by XOPSSIS (Case 05)	86
5.67	Normalized Feature Contributions of Iris-Virginica level by XOPSSIS (Case 05)	87
5.68	Explanation of Iris-Virginica level by LIME (Case 05)	87

5.69	Explanation of Iris-Virginica level by SHAP (Case 05)	88
5.70	Explanation of Iris-Virginica level by XOPSSIS (Case 06)	88
5.71	Normalized Feature Contributions of Iris-Virginica level by XOPSSIS (Case 06)	89
5.72	Explanation of Iris-Virginica level by LIME (Case 06)	89
5.73	Explanation of Iris-Virginica level by SHAP (Case 06)	90
5.74	Train-Test Split Accuracy for Different Algorithms on Breast Cancer Dataset	92
5.75	Confusion Matrix for Breast Cancer Dataset	92
5.76	Performance Evaluation Metrics on Breast Cancer Dataset	93
5.77	ROC Curve for Gradient Boosting Model on Breast Cancer Dataset	95
5.78	AUC Curve on Breast Cancer Dataset	95
5.79	Explanation of Malignant level by XOPSSIS (Case 01)	97
5.80	Normalized Feature Contributions of Malignant level by XOPSSIS (Case 01)	97
5.81	Explanation of Malignant level by LIME (Case 01)	98
5.82	Explanation of Malignant level by SHAP (Case 01)	99
5.83	Explanation of Malignant level by XOPSSIS (Case 02)	100
5.84	Normalized Feature Contributions of Malignant level by XOPSSIS (Case 02)	101
5.85	Explanation of Malignant level by LIME (Case 02)	102
5.86	Explanation of Malignant level by SHAP (Case 02)	103
5.87	Explanation of Benign level by XOPSSIS (Case 03)	104
5.88	Normalized Feature Contributions of Benign level by XOPSSIS (Case 03)	104
5.89	Explanation of Benign level by LIME (Case 03)	105
5.90	Explanation of Benign level by SHAP (Case 03)	106
5.91	Explanation of Benign level by XOPSSIS (Case 04)	107
5.92	Normalized Feature Contributions of Benign level by XOPSSIS (Case 04)	108
5.93	Explanation of Benign level by LIME (Case 04)	109
5.94	Explanation of Benign level by SHAP (Case 04)	109
5.95	Train-Test Split Accuracy for Different Algorithms on Car Acceptability Dataset	111
5.96	Confusion Matrix for Car Acceptability Dataset	112
5.97	Performance Evaluation Metrics on Car Acceptability Dataset	113
5.98	ROC Curve for Gradient Boosting Model on Car Acceptability Dataset	114
5.99	AUC Curve for Gradient Boosting Model on Car Acceptability Dataset	115
5.100	Explanation of Unacc level by XOPSSIS (Case 01)	116
5.101	Normalized Feature Contributions of Unacc level by XOPSSIS (Case 01)	117
5.102	Explanation of unacc level by LIME (Case 01)	118
5.103	Explanation of unacc level by SHAP (Case 01)	118
5.104	Explanation of Acc level by XOPSSIS (Case 02)	120
5.105	Normalized Feature Contributions of Acc level by XOPSSIS (Case 02)	120
5.106	Explanation of acc level by LIME (Case 02)	121
5.107	Explanation of acc level by SHAP (Case 02)	122
5.108	Explanation of Good level by XOPSSIS (Case 03)	123
5.109	Normalized Feature Contributions of Good level by XOPSSIS (Case 03)	123
5.110	Explanation of good level by LIME (Case 03)	124
5.111	Explanation of good level by SHAP (Case 03)	124

5.112	Explanation of Vgood level by XOP SIS (Case 04)	125
5.113	Normalized Feature Contributions of Vgood level by XOP SIS (Case 04)	125
5.114	Explanation of vgood level by LIME (Case 04)	126
5.115	Explanation of vgood level by SHAP (Case 04)	126

List of Tables

4.1	Sample Dataset	18
4.2	Sample Train Dataset	18
4.3	Sample Test Dataset	18
4.4	Euclidean distance to the ideal solution and the anti-ideal solution . .	20
4.5	Euclidean distance to the ideal solution and the anti-ideal solution with TOPSIS Score	21
4.6	TOPSIS Ranking	21
4.7	Filtered Similar Instances for High Predicted Risk Level	22
4.8	Correlation between Features and Risk Level	23
4.9	Weighted Feature Contributions to Risk Level	23
4.10	Data Distribution By Risk Level	25
5.1	Accuracy of 14 Algorithms	41
5.2	Performance Evaluation Metrics	43
5.3	Accuracy of 3 Algorithms	72
5.4	Performance Evaluation Metrics on Iris Dataset	74
5.5	Accuracy of 3 Algorithms	91
5.6	Performance Evaluation Metrics on Breast Cancer Dataset	93
5.7	Accuracy of 3 Algorithms	110
5.8	Performance Evaluation Metrics on Car Acceptability Dataset	113

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

AUC Area Under the Curve

BodyTemp Body Temperature

BS Blood Sugar

DeepLift Deep Learning Important FeaTures

DiastolicBP Diastolic Blood Pressure

DL Deep Learning

ELI5 Explain Like I'm 5

GridSearchCV Grid Search Cross-Validation

LIME Local Interpretable Model-agnostic Explanations

ML Machine Learning

ROC Receiver Operating Characteristic

SHAP Shapley Additive Explanations

SystolicBP Systolic Blood Pressure

TOPSIS Technique for Order of Preference by Similarity to Ideal Solution

XAI Explainable Artificial Intelligence

XGBoost Extreme Gradient Boosting

Chapter 1

Introduction

Explainable AI (XAI) has emerged as a critical component in the field of predictive modeling, aiming to address the growing concern regarding the lack of transparency and interpretability in machine learning models. With the widespread implementation of these models in diverse domains, including healthcare and beyond, understanding the factors that contribute to predictions becomes essential for ensuring accountability and enabling informed decision-making. The lack of transparency and interpretability in machine learning models [1] has been a major hurdle in their widespread adoption across diverse domains. Traditional black-box models often provide accurate predictions but fail to provide meaningful explanations behind their decisions. This lack of interpretability poses challenges for practitioners, researchers, and end-users, who rely on a comprehensive understanding of the factors influencing predictions to make accurate choices[2] and place faith in the outcomes of these models. Furthermore, it raises concerns about potential biases and fairness in decision-making processes, particularly in sensitive domains where the consequences of predictions are far-reaching.

Existing XAI methods such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) have made significant contributions to providing interpretability [3] in machine learning models. LIME approximates the model locally and assigns feature importances based on local perturbations, while SHAP uses Shapley values to assign importances through a game-theoretic approach. However, they have certain limitations in capturing the nuanced decision-making processes of complex models and providing detailed explanations across various domains beyond healthcare.

In this study, we propose XOPSIS, a novel XAI method that revolutionizes the generation of explanations for machine learning model predictions. XOPSIS surpasses the limitations of existing methods by incorporating advanced features such as calculating the highest and lowest possible feature values, performing instance ranking based on the TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) score, and considering the average feature values of similar instances. Moreover, XOPSIS introduces the concept of filtering instances based on predicted target, enabling a more tailored and focused explanation generation process.

By incorporating these additional features, XOPSIS delivers comprehensive and

granular explanations that capture the intricate relationships and interactions among the features in the data. The calculation of the highest and lowest possible feature values provides insights into the range of influence each feature can have on the prediction outcome. The instance ranking based on TOPSIS score ensures that the most representative and significant instances are considered during the explanation generation process. Furthermore, filtering instances based on predicted target allows XOPSIS to focus on the specific class or label of interest, enhancing the relevance and specificity of the generated explanations. Finally, considering the average feature values of similar instances adds an additional layer of context and accuracy to the explanations, enabling a more nuanced understanding of the contributing factors behind the predictions.

These innovative features set XOPSIS apart from existing XAI methods, enabling it to generate more insightful and accurate explanations. XOPSIS promotes transparency and interpretability across several domains by giving consumers an in-depth knowledge of how machine learning models make decisions.

To evaluate the effectiveness of XOPSIS, we employ our approach to multiple datasets, including a real-world Maternal Health dataset, the benchmark Iris dataset, the Breast Cancer diagnosis dataset, and the Car Acceptability dataset. Through these applications, XOPSIS demonstrates its practical applicability and performance in generating specific instance prediction explanations. Maternal health is a domain where accurate risk prediction plays a crucial role in facilitating timely intervention and tailored care. By applying XOPSIS in this context, our aim is to enhance the interpretability of machine learning models and gain a comprehensive understanding of the factors influencing maternal health risk, including the benchmark Iris dataset, breast cancer diagnosis, and the Car Acceptability dataset.

1.1 Contributions

In a nutshell, the following contributions are made by this study:

- We introduce XOPSIS as an innovative XAI method for generating comprehensive explanations of machine learning model predictions.
- We demonstrate the practical applicability of XOPSIS by leveraging it in various datasets, including a real-world Maternal Health dataset, the benchmark Iris dataset, the benchmark Breast Cancer diagnosis dataset, and the Car Acceptability dataset.
- We assess the performance of XOPSIS in generating specific instance prediction explanations, which exhibit similarities to explanations produced by other established XAI methods.

By integrating XOPSIS into the field of health risk prediction and leveraging XAI techniques, our approach aims to enhance transparency, interpretability, and accountability in machine learning models. The insights provided by XOPSIS empower healthcare providers, researchers, and stakeholders across various domains to make informed decisions, ultimately contributing to improved outcomes and fostering trust in AI technologies.

Chapter 2

Background Study

In the evolving landscape of machine learning and artificial intelligence, achieving transparency and interpretability in complex models has become pivotal for their widespread adoption and credibility across various domains. Interpretability methods [4] play a crucial role in demystifying the decision-making processes of these models, allowing stakeholders to understand and trust their outcomes. This section delves into the background study of three significant techniques: LIME, SHAP, and TOPSIS, which contribute to the explainability and insightfulness of machine learning models.

2.1 LIME

LIME (Local Interpretable Model-agnostic Explanations) revolutionizes the landscape of interpretability by offering a practical solution [5] to one of the most formidable challenges in machine learning: explaining the predictions of complex models. As machine learning models have advanced in complexity, their decision-making processes have become increasingly difficult to decipher. LIME addresses this issue by providing transparent, instance-specific explanations, making it an indispensable tool for model transparency and accountability.

At its core, LIME operates on the principle of approximating a black-box model's behavior around a specific instance of interest. This approach is rooted in the realization that while global explanations might be intricate, local interpretations can be more understandable [6]. To achieve this, LIME generates a dataset of perturbed instances by introducing controlled noise to the original instance's features. These perturbed instances are then used to create a surrogate interpretable model that mimics the behavior of the complex model around the instance in question.

LIME follows a systematic workflow encompassing several pivotal stages. Initially, an instance requiring an explanation is selected. This forms the basis for subsequent analysis. Subsequently, [7] minor alterations are introduced to the features of the chosen instance, thereby creating a diverse dataset through perturbation. Predictions are then derived for these perturbed instances, effectively capturing the output behavior of the underlying black-box model. A surrogate model is subsequently established, trained using the perturbed instances and their corresponding predictions. This surrogate model is deliberately constructed to be both interpretable and lo-

cally accurate. Leveraging this surrogate model, the next step involves interpreting the model’s decisions, thereby extracting critical insights and feature importances. These insights contextualize the rationale behind the black-box model’s specific prediction for the chosen instance. By interpreting the surrogate model’s explanations, local insights emerge, unveiling the intricate interplay between features and predictions specific to the selected instance. This comprehensive workflow not only demystifies the predictions of complex models but also empowers users to make informed decisions based on the extracted insights.

LIME’s true strength lies in its versatility. It’s model-agnostic [8], meaning it can be applied to a wide array of machine learning models without requiring knowledge of their internal architectures. This makes LIME an invaluable tool for understanding the decision-making of models like deep neural networks, random forests, support vector machines, and more. This adaptability makes LIME suitable for addressing the opacity of models ranging from image classifiers to natural language processing algorithms.

However, like any technique, LIME has its limitations. The accuracy of the surrogate model heavily depends on the quality and diversity of the perturbed instances. Also, the explanations generated by LIME are localized and may not capture global model behavior accurately.

Nonetheless, LIME’s significance is undeniable. It bridges the gap between complex, high-performance models and human understanding, making it a cornerstone of explainable AI. It empowers domain experts, regulators, and end-users with the ability to validate, trust, and even enhance machine learning systems by providing a comprehensible rationale for individual predictions.

2.2 SHAP

SHAP, an acronym for SHapley Additive exPlanations, a pioneering technique in the field of Explainable AI (XAI), offers a sophisticated approach to understanding the influence of individual features on the predictions made by complex machine learning models. Anchored in cooperative game theory, SHAP provides a robust framework for attributing contributions to each feature, thereby demystifying the decision-making process of black-box models.

At the heart of SHAP lies the concept of Shapley values [9], a central notion in cooperative game theory used to distribute the value of a cooperative endeavor among its participants fairly. SHAP cleverly adapts this concept to machine learning models, where features collaborate to predict an outcome. In this context, SHAP quantifies the average marginal contribution of a feature by considering its impact on predictions across all possible feature combinations.

The SHAP algorithm operates by comparing the model’s prediction for a specific instance with a reference prediction, typically the average prediction of the training dataset. This difference is referred [10] to as the Shapley value, representing the contribution of each feature to the variation in prediction. The crux of SHAP’s

power lies in its ability to disentangle the complex interplay of features, enabling us to discern which features amplify or dampen a particular prediction.

In practice, SHAP generates a comprehensive set of explanations, each elucidating the role of a feature in influencing the prediction for a specific instance. These explanations manifest as positive or negative values, signifying whether a feature positively or negatively affects the prediction. Notably, the summation of SHAP values across all features equates to the disparity between the model’s prediction for the instance and the reference prediction.

One of SHAP’s compelling features is its universality – it is applicable to various model types, including those that are not inherently interpretable. By providing insights into feature contributions, SHAP bridges the gap between complex models and human comprehension. This attribute has far-reaching implications, from understanding the factors driving an individual prediction to identifying bias and fairness concerns within the model’s decision-making process.

In essence, SHAP stands as a critical tool for fostering transparency, accountability, and interpretability in the realm of machine learning. By attributing contributions to features, SHAP empowers practitioners, researchers, and end-users to trust and comprehend the decisions made by complex models, even in scenarios where the model’s internal workings remain elusive.

2.3 TOPSIS

TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution), a key technique in Multi-Criteria Decision Making (MCDM) [11], addresses the challenge of ranking alternatives when considering multiple conflicting criteria. It operates by measuring the similarity of each alternative to the ideal solution and the dissimilarity to the anti-ideal solution. This method is valuable in scenarios where decisions involve a range of criteria, each with varying degrees of significance. While not exclusively designed for XAI, TOPSIS finds application in model interpretation [12] by facilitating nuanced comparisons between instances and shedding light on their distinctive attributes.

The procedure of TOPSIS, or Technique for Order of Preference by Similarity to Ideal Solution [13], encompasses several sequential steps. The first step involves the normalization of a decision matrix, comprising criterion values for each alternative. This normalization process ensures equitable evaluation across all criteria, preventing undue influence from disproportionately large values. Following this, the weighted normalization stage ensues, wherein each normalized value is multiplied by its corresponding criterion weight. These weights convey the relative significance of the criteria, introducing nuanced effects on the final ranking. Ideal and anti-ideal solutions are then defined for every criterion, with the ideal solution characterized by maximal values for beneficial criteria and minimal values for non-beneficial ones. Conversely, the anti-ideal solution represents [14] the least desirable values. Subsequently, the calculation of distances comes into play, involving the computation of Euclidean distances from both the ideal and anti-ideal solutions for each alterna-

tive. These distances provide insights into the relative proximity or divergence of each alternative concerning these reference points in the multidimensional criterion space. Determination of a relative closeness score follows, reflecting each alternative’s performance in relation to the criteria. This score is ascertained by evaluating the ratio of the distance to the anti-ideal solution to the summation of distances to both the ideal and anti-ideal solutions. Finally, based on their relative closeness scores, alternatives are ranked, with higher scores signifying superior overall performance and more favorable rankings. The fundamental principle of TOPSIS revolves around the concept of an “ideal solution” and an “anti-ideal solution.” The ideal solution represents the characteristics that an alternative should ideally possess to be considered optimal, while the anti-ideal solution embodies the exact opposite. The distance between an alternative and these two reference points forms the crux of the TOPSIS method.

When applied to XAI, TOPSIS can be employed to provide insights into instance-level predictions by assessing the proximity of instances to the ideal and anti-ideal solutions. This approach enables a comprehensive understanding [15] of the relationships between instances and the criteria they are evaluated upon. By quantifying these distances, TOPSIS can highlight instances that align closely with the ideal solution and those that deviate significantly.

Comparing TOPSIS to AHP (Analytic Hierarchy Process), another popular MCDM method [16], reveals distinct characteristics. In AHP, decision-makers assign pairwise comparisons and weights to criteria, resulting in a complex weighting process. Conversely, TOPSIS uses preset weights [17] and focuses on assessing the similarity of alternatives to the ideal and anti-ideal solutions.

The selection of TOPSIS for the XOPSIS XAI method is driven by specific reasons. One key factor is the use of uniform weights for each feature in TOPSIS, simplifying the process compared to AHP, where feature weights are ranked. Furthermore, TOPSIS is well-suited for scenarios where transparency and interpretability [18] are paramount, as it produces clear rankings based on a set of predefined criteria. In contrast, AHP’s complex weight assignment may hinder transparency. Additionally, TOPSIS aligns with the goal of XOPSIS to enhance interpretability in machine learning models across various domains. Its structured and comprehensible approach facilitates meaningful explanations, promoting trust and informed decision-making.

TOPSIS is particularly useful in scenarios where model predictions are influenced by a multitude of factors, [19] each with varying degrees of importance. The method facilitates the systematic comparison of instances across these diverse criteria, offering a holistic view of how individual attributes contribute to the final decision. This can be especially valuable in fields like healthcare, [20] finance, and risk assessment, where decisions are often influenced by multiple conflicting considerations. One of the defining characteristics of TOPSIS is its flexibility in handling both quantitative and qualitative data. This adaptability is vital when dealing with real-world data [21] that can vary widely in terms of format and meaning. By converting data into a standardized form, TOPSIS enables fair comparisons between different attributes and instances, regardless of their initial nature.

In conclusion, while not a traditional XAI method, TOPSIS complements the interpretability [22] landscape by providing a structured approach to understanding multi-criteria decision-making. By assessing the distance between instances and reference solutions, TOPSIS contributes to uncovering the rationale [23] behind predictions in complex models. Its ability to accommodate diverse data types and facilitate comprehensive comparisons makes it a valuable tool for understanding the intricate relationships within predictive models.

Chapter 3

Related Work

Concerns about the transparency and interpretability of machine learning models for health risk prediction, notably in the area of maternal health, have been raised. Explainable AI (XAI) approaches have come to light as a viable solution to these challenges in response to these worries. Existing XAI methods such as LIME, SHAP, and ELI5 have been widely studied and applied in various domains, including healthcare. However, there is a need for novel XAI methods that can provide specific instance prediction explanations with high granularity. By presenting XOPSIS, an entirely new XAI method that provides thorough insights into how to make choices of machine learning models, this work seeks to narrow this gap.

3.1 Explainable AI

Explainable Artificial Intelligence (XAI) techniques, which seek to shed light on the inner workings of machine learning models and make their predictions more transparent and intelligible, have gained popularity in recent years. One such method is the use of the multi-criteria decision-making technique TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution), which is an explainable AI method for ranking and prioritizing various risk factors according to their relative importance in the context of maternal health. When utilized for decision-making in a variety of fields, TOPSIS has proven to be efficient in managing several factors and producing outcomes that are easy to understand. Several studies have applied LIME and SHAP for model interpretation in various domains, including healthcare. These methods are being applied to clarify how feature importance, feature interactions, and overall model behavior relate to predictions made by machine learning models. To the finest of the information we have, however, there hasn't been much study on the use of TOPSIS as an explainable AI technique for interpreting maternal health risk predictions and explaining specific data point predictions, along with the complementary use of LIME and SHAP for model interpretation.

In order to explain the predictions of any classifier, this paper [24] introduces a technique called LIME (Local Interpretable Model-Agnostic Explanations). The interpretability issue in black-box machine learning models is addressed by LIME by offering local explanations that highlight the key factors influencing each prediction. The method generates simplified, interpretable models around specific instances and quantifies the importance of each feature. By providing clear explanations, this

method increases consumers' faith in the predictions by allowing them to see how complex classifiers make decisions.

SHAP (SHapley Additive exPlanations), a standardized method for unraveling model predictions, is presented in the study [25]. Assigning relevance levels to each feature and measuring their contribution to the prediction is done by SHAP using game theory principles. The behavior of the model is explained globally by SHAP by computing Shapley values, which reflect the minimal impact of each feature across various combinations. This approach allows for better understanding of the model's decision-making process and enables insights into individual feature contributions, leading to improved interpretability and trust in machine learning models.

The paper [26] introduces DeepLIFT, a method for learning important features in deep neural networks by propagating activation differences. By comparing the difference in neuron activations when a feature is present versus missing, DeepLIFT calculates the impact of every characteristic to the outcome. This approach allows for fine-grained feature attribution and provides insights into the important features driving the model's predictions. By quantifying feature importance, DeepLIFT enhances interpretability and understanding of deep neural networks, facilitating trust and further analysis of these complex models.

In [27], the area of explainable AI for trees is examined, with a focus on both local and global comprehension. The authors propose an integrated framework that combines SHAP (SHapley Additive exPlanations) values with tree ensembles to provide interpretable explanations for individual predictions. This method provides explanations for the overall performance of the model and allows for the discovery of key characteristics impacting certain predictions. By bridging the gap between local interpretability and global understanding, the paper contributes to the development of explainable AI techniques for tree-based models, enhancing trust and transparency in decision-making processes.

In [28], the author introduces RuleFit, a method for predictive learning through rule ensembles. RuleFit combines the interpretability of rule-based models with the predictive power of ensemble methods. The approach creates a set of rules that capture both linear and non-linear relationships in the data, and then integrates them into an ensemble model. RuleFit provides interpretable explanations by identifying important features and the corresponding rules that contribute to predictions. The method offers a balance between model complexity and interpretability, making it suitable for various domains where transparency is desired.

The paper [29] presents a comprehensive approach for explainable AI with a focus on tree-based models. The authors introduce a unified framework, called TreeExplainer, that allows for the interpretation of both local and global model behavior. TreeExplainer provides insights into how individual features contribute to predictions and how different interactions between features influence model outcomes. By bridging the gap between local and global interpretability, the method offers a deeper understanding of tree-based models and enhances transparency in their decision-making processes. This framework provides valuable tools for practitioners

seeking to gain global insights and explanations from tree-based models.

A thorough examination and analysis of the state of interpretable machine learning are provided in the study [30]. In order to establish more rigorous and standardized methods for clarity, the paper addresses the difficulties and limitations of present approaches to clarity in machine learning models. It also suggests a framework for doing so. In addition to offering suggestions for developing the field of interpretable machine learning, the paper highlights the necessity of scientifically grounded, replicable, and validated interpretable machine learning. Scientists and practitioners interested in the topic of interpretable machine learning will find the paper to be a useful resource.

The authors of [31] propose a method for evaluating each feature’s importance to a prediction generated by a model, offering insights into the significance of various features for the predictions made by the model. The use of this method in numerous domains is discussed in the study, which also emphasizes its potential to make prediction models more comprehensible and reliable. The paper serves as a valuable contribution to the field of interpretable machine learning and provides insights into the explanation of model predictions.

A method called “Anchors” has been introduced in [32] that provides interpretable explanations for individual predictions, identifying the most important features that influence a model’s output. The paper indicates that Anchors performs better than conventional explanation techniques both in terms of accuracy and comprehensibility. It also illustrates the usefulness of Anchors in numerous trials. The study offers a potential method for producing precise and understandable justifications for model predictions, which makes it an important contribution to the subject of interpretable machine learning.

3.2 Maternal Health

Maternal health risk prediction has become a critical area of research to improve maternal health outcomes and reduce maternal morbidity and mortality. Machine learning approaches have been used in several research to forecast threats to maternal health, such as gestational diabetes, preeclampsia, premature birth, and other issues. Although the predictive accuracy of these studies has shown encouraging results, the black-box nature of machine learning models has constrained their interpretability, making it difficult to comprehend the underlying causes influencing the risk estimations.

Research on the connection between low birth weight in the initial pregnancy and the risk of preeclampsia in the subsequent pregnancy is presented in this work [33]. According to the study, preeclampsia in a subsequent pregnancy may be predicted by low birth weight in the initial pregnancy, necessitating thorough monitoring and early interventions in women who have a history of giving birth to low birth weight babies.

Pre-eclampsia, a potentially dangerous pregnancy condition that involves excessive blood pressure and organ damage, is thoroughly discussed in this study [34]. The paper discusses the pathophysiology, risk factors, diagnosis, and management of pre-eclampsia, including the role of regular antenatal care, monitoring, and timely interventions. It also emphasizes how critical early detection and effective treatment are to preventing the harmful effects of pre-eclampsia on both the mother and the fetus.

The authors of [35] provide an analysis of pre-eclampsia rates in the US from 1980 to 2010 by age-period-cohort. The study looks at variations and patterns in the occurrence of pre-eclampsia across time, taking into account the impact of age, time, and birth cohort. The findings provide insights into the changing rates of pre-eclampsia and contribute to a better understanding of this pregnancy complication in the US context during the specified time period.

In [36], the authors examine the incidence of preeclampsia, a hypertensive disorder during pregnancy, and associated risk factors and outcomes. In order to distinguish between early-onset and a late-onset preeclampsia, the paper covers each subtype's risk factors and undesirable consequences. For proper management and care throughout pregnancy, the report emphasizes the need of comprehending the distinctions between initial and postpartum preeclampsia and stresses the need for more research in this area.

This paper [37] provides an overview of the existing research on the importance of prenatal care and the diagnosis of high-risk pregnancies for effective maternal health care. It highlights the use of machine learning and deep learning algorithms in predicting risk levels based on pregnancy risk factors. Additionally, Explainable AI techniques such as LIME and SHAP are explored for providing interpretable explanations. The review emphasizes the significance of early diagnosis and appropriate treatment in reducing maternal mortality and improving maternal and fetal well-being.

A study [38] conducted in 2020 implemented a modified decision tree algorithm for diagnosing high-risk pregnancies. The study comprised six independent variables and one objective variable, and it used data from six hospitals from 2018 to 2020. IoT technology was used to gather the research's data. The results showed that, when predicting the model, the improved decision tree approach showed a 97% gain in accuracy over comparable methods. WEKA and Python software were also utilized in the algorithm's implementation.

In their study [39], Caruana et al. (2015) address the critical task of predicting pneumonia risk and hospital 30-day readmission in healthcare settings. They emphasize the importance of intelligible models for medical decision-making and propose approaches to enhance model interpretability. Their research contributes to the growing field of interpretable machine learning in healthcare, aiming to improve the transparency and trustworthiness of predictive models for medical applications.

3.3 IRIS Species

In their work [40], Swain et al. (2012) present an approach for the classification of iris plants using neural networks. Their study focuses on leveraging neural network techniques to effectively classify iris plant species. By applying advanced machine learning methods, the authors contribute to the field of plant species classification and demonstrate the potential of neural networks in improving accuracy and efficiency in this domain. The research underscores the significance of machine learning applications in botanical studies.

In their study presented at the 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), Pachipala et al. utilize a Random Forest approach for the classification of iris flowers within an AWS environment. The authors [41] highlight the application of Random Forest as a powerful technique in the context of flower classification, taking advantage of AWS infrastructure to enhance computational capabilities. This work emphasizes the integration of cloud computing and machine learning methodologies for effective and scalable flower species classification, contributing to the advancement of automated botanical analysis.

3.4 Breast Cancer

In their recent work [42] published by MDPI, Naeem and Ali delve into the domain of breast cancer diagnosis through the application of machine learning techniques. The authors explore the utilization of various machine learning methodologies for the purpose of breast cancer detection, emphasizing the potential of these techniques in contributing to accurate and efficient diagnostic procedures. This study contributes to the ongoing efforts in leveraging machine learning advancements for medical diagnosis, particularly in the critical area of breast cancer detection and classification.

In a recent preprint published on arXiv, [43] Zuluaga-Gomez delves into the realm of breast cancer diagnosis through the lens of machine learning techniques. The author's work revolves around exploring the potential of machine learning methodologies for enhancing breast cancer diagnosis accuracy and efficiency. By investigating the application of these techniques to breast cancer data, this study aims to contribute valuable insights into the field of medical diagnosis, particularly focusing on the critical area of breast cancer detection and classification. The work adds to the growing body of research dedicated to leveraging machine learning advancements for improved medical diagnostic procedures.

In a comprehensive article published in the journal *Diagnostics*, the authors conduct a systematic review [44] to explore the diverse landscape of artificial intelligence (AI) applications in breast cancer diagnosis. The authors delve into the realm of AI and its multifaceted role in enhancing breast cancer diagnostic processes. Through a meticulous examination of existing literature, this study critically evaluates various AI techniques and their effectiveness in aiding medical professionals in accurately diagnosing breast cancer. By providing a consolidated overview of AI applications in breast cancer diagnosis, the research contributes to the understanding of the trans-

formative potential of AI in the field of medical imaging and diagnosis.

In the esteemed journal [45] *Expert Systems with Applications*, Mehmet Fatih Akay presents a research article which delves into the integration of support vector machines (SVMs) with feature selection techniques to enhance the accuracy of breast cancer diagnosis. Akay's research explores the intricate interplay between advanced machine learning algorithms and the identification of relevant features to improve the diagnostic process. By leveraging SVMs and feature selection methods, the study contributes to the ongoing efforts to develop more accurate and efficient diagnostic tools for breast cancer. The paper's findings offer valuable insights into the potential of combining sophisticated algorithms with intelligent feature selection strategies in the realm of medical diagnosis.

3.5 Car Acceptability

In the *International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE)*, the authors [46] present an article titled "Predicting Overall Car Performance Using Artificial Neural Network." This research article investigates the application of artificial neural networks (ANNs) to predict the overall performance of cars. The authors delve into the realm of machine learning by utilizing ANNs to model and forecast the intricate and multifaceted aspects of car performance. Through this work, Al-Mubayyed, Abu-Nasser, and Abu-Naser contribute to the development of predictive models in the automotive domain, enhancing the understanding and evaluation of car performance characteristics. The study showcases the potential of artificial neural networks in analyzing and forecasting complex automotive parameters, offering insights into the practical utilization of machine learning techniques in the automotive industry.

In their article [47], the authors delve into the realm of predictive modeling for car performance. Published in 2020, the study explores the application of Artificial Neural Networks (ANNs), specifically Jittered Neural Networks (JNNs), to forecast various aspects of car performance. The authors harness the power of machine learning and neural networks to analyze and predict intricate performance parameters of automobiles. By focusing on JNNs, the research aims to enhance the accuracy and efficiency of predictions, contributing to the field of automotive engineering and predictive modeling. Through this work, Al-Mobayed, Al-Madhoun, Al-Shuwaikh, and Abu-Naser offer insights into the potential of neural network-based methodologies in the analysis and prediction of car performance characteristics.

Chapter 4

Methodology

4.1 Proposed Explainable AI Method

TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) is a widely used multi-criteria decision-making (MCDM) method that involves ranking alternatives based on their proximity to an ideal solution, providing insights into the model's prediction behavior by representing how each feature contributes towards the prediction for a specific instance. TOPSIS is widely used in various fields such as finance, marketing, engineering, and environmental management to support decision-making processes when multiple criteria are involved. It provides a simple yet effective approach for decision-makers to compare and rank alternatives based on their performance against multiple criteria simultaneously, making it a valuable tool in decision analysis and decision support systems. The XOPSSIS method is executed as per the following steps. The flowchart presented in Figure 4.1 outlines the key steps, focusing on the initial stages of Data Preprocessing, Classification/Prediction, Topsis score calculation and XOPSSIS Explanation Generation.

In the Data Preprocessing step, the flowchart begins by initializing the dataset and performing target labeling to prepare the data for analysis. Subsequently, the formulation of training and testing datasets is conducted using two techniques: train-test split and k-fold cross validation. These techniques ensure the availability of suitable data subsets for model training and evaluation, promoting robustness and generalizability of the subsequent analysis.

Moving to the next step, the flowchart transitions to the classification/prediction phase. Here, the three main algorithms utilized for risk assessment, namely gradient boosting, random forest, and XGBoost, are highlighted. These algorithms are among the 14 employed in the study, chosen for their proven effectiveness in predictive modeling. By applying these algorithms to the preprocessed data, the flowchart facilitates the generation of predicted risk labels, offering valuable insights into the classification of risk levels in the target population.

Subsequently, the flowchart progresses to the third step: Topsis score calculation. This step involves the computation of Topsis scores, a multi-criteria decision-making method, which further refines the risk assessment process. The Topsis score calculation enables the evaluation of alternatives based on multiple attributes, assisting in the identification of the most optimal options.

Assuming that, we start with an input dataset consisting of m alternatives and n criteria, represented by a matrix X with dimensions $m \times n$. Each row of X corre-

sponds to an alternative, and each column represents a criterion. Next, we normalize the dataset by dividing each element by the sum of its corresponding column. This normalization step ensures that each criterion is treated equally and comparable across alternatives. After normalization, we calculate the weighted normalized decision matrix, where each element of the normalized matrix is multiplied by the weight assigned to its corresponding criterion. Then, we determine the ideal and anti-ideal solutions. The ideal solution represents the best possible values for each criterion, while the anti-ideal solution represents the worst possible values. These solutions are derived by finding the maximum and minimum values for each criterion in the dataset, respectively. Using the ideal and anti-ideal solutions, we calculate the Euclidean distance of each alternative to both solutions. The Euclidean distance measures the similarity or dissimilarity between an alternative and the ideal or anti-ideal solution. With the calculated Euclidean distances, we can determine the TOPSIS score for each alternative.

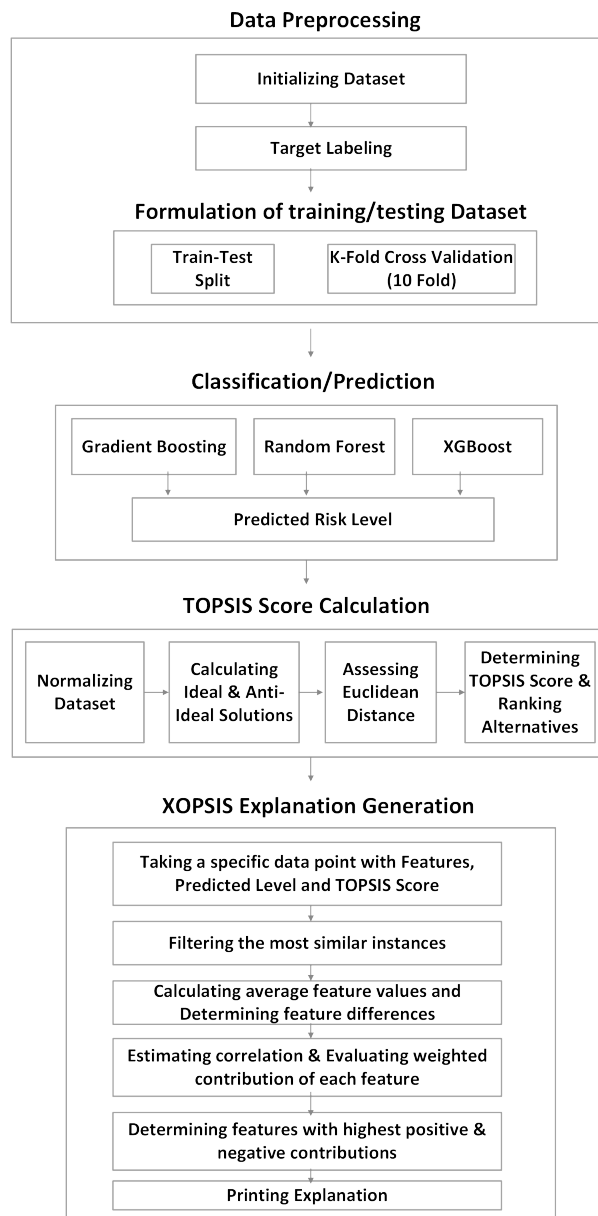


Figure 4.1: Top Level Overview of the XOPSIS Method

Algorithm 1 XOPSIS Based Explanation Method

procedure XOPSISEXPLANATION(Dataset X, Weight vector w_i , Data Point)

1. Normalize the dataset:

$$X' \leftarrow \frac{X}{\sum X}$$

2. Calculate the weighted normalized decision matrix:

$$X'' \leftarrow X' \cdot w_i$$

3. Calculate the Euclidean distance to ideal and anti-ideal solutions:

$$D^+ \leftarrow \sqrt{\sum (X'' - A^+)^2}$$

$$D^- \leftarrow \sqrt{\sum (X'' - A^-)^2}$$

4. Determine the TOPSIS score for each alternative:

$$\text{TOPSIS Score} \leftarrow \frac{D^-}{D^+ + D^-}$$

5. Rank alternatives based on TOPSIS Scores in descending order
6. Take a specific Data Point as input and filter the most similar instances based on TOPSIS Score and predicted risk level
7. Calculate average feature values for these instances:

$$\text{Avg. Features} \leftarrow \frac{\sum \text{features of instances}}{\text{number of instances}}$$

8. Determine the feature differences:

$$\text{Feature Diff} \leftarrow \text{DataPoint Features} - \text{Avg. Features}$$

9. Estimate correlation between each feature and target variable:

$$\text{Corr} \leftarrow \text{CalculateCorrelation}(df, \text{'RiskLevel'})$$

10. Evaluate the weighted contribution of each feature:

$$\text{Weighted Contrib} \leftarrow \text{Feature Diff} \cdot \text{Corr}$$

11. Detect features with positive and negative contributions:

if Weighted Contrib ≥ 0 **then**

 Positive Contrib \leftarrow Pos. Weighted Contrib.

else

 Negative Contrib \leftarrow Neg. Weighted Contrib.

end if

12. Print the Explanation: Data Point index, actual risk level, predicted risk level, features with highest positive and negative contributions

end procedure

The TOPSIS Score is calculated as the ratio of the distance to the anti-ideal solution divided by the sum of the distances to both the ideal and anti-ideal solutions. Finally, based on the TOPSIS scores, the alternatives are ranked in descending order. Higher TOPSIS scores indicate better rankings, reflecting the alternatives' proximity to the ideal solution and their distance from the anti-ideal solution.

Figure 4.1 showcases a comprehensive flowchart representing the step-by-step execution of our proposed XOPSIS methodology. This methodology, displayed in this figure, encompasses the sequential steps involved in generating explanations for the predicted risk level of a given data point. By utilizing TOPSIS scores and weighted contributions, the flowchart provides valuable insights into the specific features that contribute positively or negatively to the prediction. After calculating TOPSIS Score, the process takes a specific Data Point as an input, which contains several features along with Ideal solutions, anti-ideal solutions, TOPSIS Score, and Predicted Risk Level.

The next step in the process is to filter out the most similar instances to the given data point based on their TOPSIS Score in descending order and the predicted risk level. This is done to identify instances that are most similar to the given data point and can provide useful insights into predicting the risk level.

Once the most similar instances are identified, the process calculates the average feature values for these instances. This step helps in determining the average feature values for instances that are similar to the given data point. The next step is to determine the feature differences between the given data point and the average feature values. This step helps in identifying the features that are different between the given data point and the average feature values.

The process then estimates the correlation between each feature and the target variable (Risk Level). This step helps in evaluating the correlation between each feature and the target variable and identifying the features that are most correlated with the target variable. The process then evaluates the weighted contribution of each feature to the prediction by multiplying the feature differences with the correlations. This step helps in determining the contribution of each feature towards predicting the risk level. The next step is to detect features with the highest positive and negative contributions based on their weighted values. This step helps in identifying the features that have the most significant positive and negative impact on predicting the risk level.

Finally, the process prints an "Explanation" that includes the data point index, actual risk level, predicted risk level, and the features with the highest positive and negative contributions indicating the completion of the process.

4.1.1 Algorithm Application on a Small Dataset

In this section, we will apply the XOPSIS-based explanation algorithm step by step on a sample small dataset. The XOPSIS algorithm is designed to provide explanations for data points by identifying the contributions of individual features towards the predicted outcome. The sample dataset shown in Table 4.1 we will be using consists of various health-related measurements, such as age, blood pressure, body temperature, and heart rate, along with the corresponding risk level assigned to each individual.

The goal of our algorithm is to analyze the dataset, calculate the weighted contribu-

tions of each feature, and identify the features that have the most positive or negative impact on the risk level prediction. By doing so, we aim to gain insights into which factors are influential in determining the risk level and provide an explanation for the prediction outcome. Now, we will apply the XOPSIS algorithm meticulously, dissecting each step, to unravel the intricate web of feature contributions.

Serial	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	Risk Level
1	25	130	80	15	98.6	86	high
2	30	140	85	12	98	70	high
3	23	130	70	7.5	98	78	mid
4	20	120	75	7.3	100	70	mid
5	50	140	90	15	98	90	high
6	21	90	65	6.8	98	76	low
7	22	100	65	7.2	98	70	low
8	17	85	60	9	102	86	mid
9	23	90	60	6.4	98	76	low
10	23	120	80	7	98	66	low

Table 4.1: Sample Dataset

After train-test split (80:20) and preprocessing, we obtain the following sample train-set.

Serial	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	Risk Level
6	21	90	65	6.8	98	76	0
1	25	130	80	15	98.6	86	2
8	17	85	60	9	102	86	1
3	23	130	70	7.5	98	78	1
10	23	120	80	7	98	66	0
5	50	140	90	15	98	90	2
4	20	120	75	7.3	100	70	1
7	22	100	65	7.2	98	70	0

Table 4.2: Sample Train Dataset

After train-test split (80:20) and preprocessing, we obtain the following sample test-set.

Serial	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	Risk Level
9	23	90	60	6.4	98	76	0
2	30	140	85	12	98	70	2

Table 4.3: Sample Test Dataset

Step 01

In this step, we divide each element by the maximum value of its corresponding column of the sample dataset. For example, to normalize the first element of the “Age” column, we divide 25 by the maximum value of all the values in the “Age” column ($25 / 50 = 0.50$). Similarly, we perform this calculation for each element in the dataset. After Normalizing, we obtain the dataset as follows.

$$X' = \begin{bmatrix} 0.50 & 0.928571 & 0.888889 & 1.000000 & 0.966667 & 0.955556 \\ 0.60 & 1.000000 & 0.944444 & 0.800000 & 0.960784 & 0.777778 \\ 0.46 & 0.928571 & 0.777778 & 0.500000 & 0.960784 & 0.866667 \\ 0.40 & 0.857143 & 0.833333 & 0.486667 & 0.980392 & 0.777778 \\ 1.00 & 1.000000 & 1.000000 & 1.000000 & 0.960784 & 1.000000 \\ 0.42 & 0.642857 & 0.722222 & 0.453333 & 0.960784 & 0.844444 \\ 0.44 & 0.714286 & 0.722222 & 0.480000 & 0.960784 & 0.777778 \\ 0.34 & 0.607143 & 0.666667 & 0.600000 & 1.000000 & 0.955556 \\ 0.46 & 0.642857 & 0.666667 & 0.426667 & 0.960784 & 0.844444 \\ 0.46 & 0.857143 & 0.888889 & 0.466667 & 0.960784 & 0.733333 \end{bmatrix}$$

Step 02

After normalizing the dataset in step 1, we proceed to step 2, which involves calculating the weighted normalized decision matrix. In this step, we multiply the normalized dataset by the weight vector using equal weights. Here, we consider a sample weight vector obtained using equal weights: $w_i = [\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}]$. Let us apply this step to the sample dataset. For example, to calculate weighted normalized decision matrix, the first element of the “Age” column, we obtain 0.0874 from Step 01 and then multiply it by 0.167 as we obtain from w_i ($0.50 * 0.167$) = 0.083333. Similarly, we perform this calculation for each element in the dataset. Finally, we obtain the dataset as follows.

$$X'' = X' \times w_i = \begin{bmatrix} 0.083333 & 0.154762 & 0.148148 & 0.166667 & 0.161111 & 0.159259 \\ 0.100000 & 0.166667 & 0.157407 & 0.133333 & 0.160131 & 0.129630 \\ 0.076667 & 0.154762 & 0.129630 & 0.083333 & 0.160131 & 0.144444 \\ 0.066667 & 0.142857 & 0.138889 & 0.081111 & 0.163399 & 0.129630 \\ 0.166667 & 0.166667 & 0.166667 & 0.166667 & 0.160131 & 0.166667 \\ 0.070000 & 0.107143 & 0.120370 & 0.075556 & 0.160131 & 0.140741 \\ 0.073333 & 0.119048 & 0.120370 & 0.080000 & 0.160131 & 0.129630 \\ 0.056667 & 0.101190 & 0.111111 & 0.100000 & 0.166667 & 0.159259 \\ 0.076667 & 0.107143 & 0.111111 & 0.071111 & 0.160131 & 0.140741 \\ 0.076667 & 0.142857 & 0.148148 & 0.077778 & 0.160131 & 0.122222 \end{bmatrix}$$

In this step, each element in the normalized dataset (X') is multiplied by its corresponding weight value from the weight vector (w_i). For example, the first element in the weighted normalized decision matrix (X'') is calculated by multiplying the first element of (X') by the first weight value from w_i .

Step 03

To determine the ideal solution and anti-ideal solution, we need to find the maximum and minimum values for each criterion in the dataset. The ideal solution represents the best possible values for each criterion, so the ideal solution is, $A^+ = [0.166667, 0.166667, 0.166667, 0.166667, 0.166667, 0.166667]$. On the other hand, the anti-ideal solution represents the worst possible values, so the anti-ideal solution is, $A^- = [0.056667, 0.101190, 0.111111, 0.071111, 0.160131, 0.122222]$.

Step 3 of the algorithm involves calculating the Euclidean distance to the ideal and anti-ideal solutions for each data point in the dataset. Let's calculate the Euclidean distances for the given dataset using the ideal solution and anti-ideal solution we determined earlier. For each data point, we calculate the Euclidean distance to the ideal solution (D^+) and the Euclidean distance to the anti-ideal solution (D^-).

Data Point 1:

- Features: $[0.083333, 0.154762, 0.148148, 0.166667, 0.161111, 0.159259]$

- Euclidean distance to the ideal solution (D^+):

$$\begin{aligned} &= \sqrt{((0.083 - 0.167)^2 + (0.155 - 0.167)^2 + (0.148 - 0.167)^2 \\ &\quad + (0.167 - 0.167)^2 + (0.161 - 0.167)^2 + (0.159 - 0.167)^2)} \\ &= \sqrt{0.007056 + 0.000144 + 0.000361 + 0 + 0.000036 + 0.000064} \\ &= \sqrt{0.007661} \approx 0.086688 \end{aligned}$$

- Euclidean distance to the anti-ideal solution (D^-):

$$\begin{aligned} &= \sqrt{((0.083 - 0.057)^2 + (0.155 - 0.101)^2 + (0.148 - 0.111)^2 \\ &\quad + (0.167 - 0.071)^2 + (0.161 - 0.160)^2 + (0.159 - 0.122)^2)} \\ &= \sqrt{0.000676 + 0.002916 + 0.001369 + 0.009216 + 0.000001 + 0.001369} \\ &= \sqrt{0.015547} \approx 0.124323 \end{aligned}$$

Similarly, we calculate the Euclidean distances for the remaining data points in Table 4.4.

Data Point	D+	D-
1	0.086688	0.124323
2	0.083999	0.110611
3	0.130746	0.065238
4	0.141565	0.052662
5	0.006536	0.174871
6	0.155071	0.025722
7	0.148484	0.028570
8	0.154832	0.047424
9	0.156765	0.027899
10	0.137583	0.059601

Table 4.4: Euclidean distance to the ideal solution and the anti-ideal solution

Step 04

Step 4 of the algorithm involves determining the TOPSIS score for each data point. The TOPSIS score represents the relative closeness of each alternative (data point) to the ideal and anti-ideal solutions. Let's calculate the TOPSIS score for each data point using the Euclidean distances we obtained in Step 3. For each data point, we calculate the TOPSIS score as follows:

$$\text{TOPSIS Score} = \frac{D^-}{D^+ + D^-} \quad (4.1)$$

For example, for Data Point 1, we obtain the TOPSIS Score = $\frac{0.124323}{0.086688+0.124323} \approx 0.58917823$. Similarly, we calculate the TOPSIS Score for the remaining data points as follows in Table 4.5.

Data Point	D+	D-	TOPSIS Score
1	0.086688	0.124323	0.58917823
2	0.083999	0.110611	0.56837344
3	0.130746	0.065238	0.33287573
4	0.141565	0.052662	0.27113549
5	0.006536	0.174871	0.96397072
6	0.155071	0.025722	0.1422742
7	0.148484	0.028570	0.16136506
8	0.154832	0.047424	0.23447515
9	0.156765	0.027899	0.15108087
10	0.137583	0.059601	0.30226133

Table 4.5: Euclidean distance to the ideal solution and the anti-ideal solution with TOPSIS Score

Step 05

To apply Step 5 of the algorithm, we need to rank the alternatives (data points) based on their TOPSIS scores in descending order. Let's sort the data points in descending order of their TOPSIS scores and create a table to display the results.

Rank	Data Point	TOPSIS Score
1	5	0.96397072
2	1	0.58917823
3	2	0.56837344
4	3	0.33287573
5	10	0.30226133
6	4	0.27113549
7	8	0.23447515
8	7	0.16136506
9	9	0.15108087
10	6	0.1422742

Table 4.6: TOPSIS Ranking

In Table 4.6, the data points have been ranked based on their TOPSIS scores, with the highest score being ranked first.

Step 06

To filter the most similar instances based on TOPSIS Scores and predicted risk level from the test sample dataset for the specific data point (30, 140, 85, 12.0, 98.0, 70, High), we will compare the TOPSIS scores and predicted risk levels of each instance in the dataset. For the TOPSIS scores, we can see from the previous step that the instances are ranked in descending order. Now, we will filter the instances that have the highest TOPSIS scores and the same predicted risk level as the specific data point.

From the Table 4.7, we can see that instances 5 and 1 have the highest TOPSIS scores (0.949, 0.651 and 0.4999, respectively) and the predicted risk level of "high." Therefore, we will filter instances 5 and 1 as the most similar instances to the specific data point based on TOPSIS scores and predicted risk level.

Rank	Data Point	Risk Level	TOPSIS Score
1	5	High	0.963971
2	1	High	0.589178

Table 4.7: Filtered Similar Instances for High Predicted Risk Level

In this Table 4.7, the instances that have the highest TOPSIS scores and match the predicted risk level (High) are Data Point 5 and Data Point 1 respectively.

Step 07

To calculate the average feature values for the instances that match the predicted risk level, we sum up the feature values of those instances and divide by the number of instances. Here are the average feature values: Average Features for Risk Level High:

- Age: $(50 + 25) / 2 = 37.5$
- SystolicBP: $(140 + 130) / 2 = 135.0$
- DiastolicBP: $(90 + 80) / 2 = 85.0$
- BS: $(15 + 15) / 2 = 15.0$
- BodyTemp: $(98 + 98.6) / 2 = 98.3$
- HeartRate: $(90 + 86) / 2 = 88.0$

Step 08

Step 8 of the algorithm involves determining the feature differences between the specific data point and the average feature values calculated in step 7. Let's calculate the feature differences for the provided specific data point (30, 140, 85, 12.0, 98.0, 70) and the average feature values:

Feature Differences:

- Age: $30 - 37.5 = -7.5$
- SystolicBP: $140 - 135 = 5.0$
- DiastolicBP: $85 - 85 = 0.0$
- BS: $12 - 15 = -3.0$

- BodyTemp: $98 - 98.3 = -0.3$
- HeartRate: $70 - 88 = -18.0$

Step 09

Table 4.8 displays the correlation values between each feature and the Risk Level in the dataset. Positive correlation values indicate a positive relationship with the Risk Level, while negative correlation values indicate a negative relationship. The higher the correlation value (closer to 1 in absolute value), the stronger the correlation with the Risk Level. The features include Age, SystolicBP, DiastolicBP, BS (Blood Sugar), BodyTemp (Body Temperature), and HeartRate.

Feature	Correlation with Risk Level
Age	0.566361
SystolicBP	0.738184
DiastolicBP	0.694773
BS	0.893124
BodyTemp	0.119903
HeartRate	0.540133

Table 4.8: Correlation between Features and Risk Level

Step 10

In this step, we need to calculate the weighted feature contributions. For this, We will multiply the feature differences by their respective correlations.

Here are the calculations:

- Age: $-7.5 * 0.566361 = -4.247705$
- SystolicBP: $5.0 * 0.738184 = 3.690919$
- DiastolicBP: $0 * 0.694773 = 0.0000$
- BS: $-3.0 * 0.893124 = -2.679373$
- BodyTemp: $-0.3 * 0.119903 = -0.035971$
- HeartRate: $-18.0 * 0.540133 = -9.722396$

These values represent the weighted feature contributions.

Step 11

Feature	Normalized Weighted Contribution	Contribution Type
Age	-0.21	Negative
SystolicBP	0.18	Positive
DiastolicBP	0.00	Positive
BS	-0.13	Negative
BodyTemp	-0.00	Negative
HeartRate	-0.48	Negative

Table 4.9: Weighted Feature Contributions to Risk Level

Table 4.9 displays the weighted contributions of each feature to the Risk Level. The contributions are calculated by multiplying the correlation value of each feature

with the Risk Level by the respective weight as shown in the previous step. If the weighted contribution is more than or equal to zero, it is classified as a positive contribution, indicating a positive impact on the Risk Level. Conversely, if the weighted contribution is less than zero, it is classified as a negative contribution, indicating a negative impact on the Risk Level.

Step 12

Finally, we will print the explanation and plot for this specific data point.

4.2 Maternal Health Risk Prediction and Explanation

4.2.1 Data Collection

We collected data from the publicly available dataset “Maternal Health Risk Data” obtained from Kaggle [48]. The dataset contains 1014 rows and includes the following six features: Age, SystolicBP, DiastolicBP, BS, BodyTemp, and HeartRate. These features are potential risk factors associated with maternal health.

- Age: The age of the patient in years
- SystolicBP: Upper value of the blood pressure and the measurement in millimeters of mercury (mmHg)
- DiastolicBP: Lower value of the blood pressure and the measurement in millimeters of mercury (mmHg)
- BS: The blood sugar (glucose) level measurement in millimoles per liter (mmol/L)
- BodyTemp: The human body temperature measurement in degrees Fahrenheit (°F)
- HeartRate: The heart rate measurement in beats per minute (bpm)

The dataset includes a target variable called ‘Risk Level’ which is categorized into three values: Low, Mid, and High Risk. This variable represents the risk level of maternal health based on the given features.

4.2.2 Dataset Preprocessing

Checking for Null Values

The dataset was checked for any missing or null values. It was found that there were no null values present in the dataset, indicating that the dataset is complete in terms of data availability.

Encoding of Categorical Variable

The target variable, which contains categorical values (Low, Mid, High), was encoded using the ordinal encoder. Ordinal encoding is used to convert categorical variables with ordered categories into numerical representations. In this case, the "Risk Level" categories were encoded as (0, 1, 2) respectively, based on the order of risk levels (Low, Mid, High). Encoding categorical variables is required to convert them into numerical representations that can be utilized as input to machine learning algorithms. Ordinal encoding was used in this case to maintain the order of categories and represent the risk levels as numerical values. Table 4.10 shows the distribution of the dataset.

Risk Level	Frequency	Percentage
Low Risk	406	40.04
Mid Risk	336	33.14
High Risk	272	26.82
Total	1014	100.0

Table 4.10: Data Distribution By Risk Level

Feature Scaling

Standard Scaling (also known as z-score normalization) was applied to the numerical features in the dataset, including Age, Systolic BP, Diastolic BP, Blood Sugar level (BS), Body Temperature, and Heart Rate. Standard scaling transforms the features to have zero mean and unit variance, which helps in bringing all the features to a similar scale, making them comparable and avoiding any dominance of one feature over the others during model training. Feature scaling is essential to normalize the numerical features and bring them to a similar scale, avoiding any potential bias towards features with larger values. Standard scaling helps in centering the features around zero with unit variance, which can improve the model's training and prediction accuracy.

Train-Test Split

The dataset was divided into training and testing sets in an 80:20 ratio using the train-test split approach. This means that 80% of the dataset was used to train the machine learning model, with the remaining 20% left aside for testing and evaluating the performance of the trained model. This aids in determining the generalization capacity and performance of the model on previously encountered data. The train-test split is critical for evaluating the performance of the model on unknown data and its generalization capacity. We may obtain an unbiased estimate of the performance and identify any potential overfitting issues by evaluating a subset of the dataset.

K-fold Cross-Validation

This technique divides the dataset into k equally sized folds and uses them for training and testing iteratively in order to assess the performance of a machine learning

model. The dataset is divided into k equally sized folds (e.g., $k=10$, meaning 10 folds). The model is trained on $k-1$ folds (i.e., $k-1$ parts of the dataset) and tested on the remaining 1 fold (i.e., the remaining part of the dataset). This process is repeated k times, with each fold being used once as the test set and the remaining $k-1$ folds used as the training set in each iteration. The performance metrics, such as accuracy, precision, recall, F1 score, etc., are calculated for each fold, and the average performance is reported as the final evaluation of the model.

The goal of utilizing k -fold cross-validation is to achieve a more trustworthy estimate of the performance by leveraging the whole dataset for both training and testing. It aids in decreasing model performance variability due to variable train-test splits and enables a more robust evaluation of the performance of the model.

Hyperparameter Tuning

Hyperparameters are the parameters of a machine learning model that are not learned during the training process and need to be set manually. Examples of hyperparameters include the learning rate, regularization strength, number of estimators, etc. Hyperparameter tuning is the process of finding the optimal values for these hyperparameters to optimize the performance of the model.

Grid Search CV is a popular hyperparameter tuning technique that exhaustively searches for the best combination of hyperparameter values from a predefined grid of possible values. A grid of hyperparameter values is defined, specifying the possible values for each hyperparameter. The model is trained and evaluated using k -fold cross-validation for each combination of hyperparameter values from the grid. The performance metrics are recorded for each combination of hyperparameter values. The combination of hyperparameter values that results in the best performance is selected as the optimal set of hyperparameter values. The purpose of using Grid Search CV is to systematically search for the optimal hyperparameter values to maximize the outcome of the model. It helps in finding the best hyperparameter values that result in the best model performance on the given dataset, improving the accuracy, robustness, and generalization ability of the model.

4.2.3 Model Training

In the next step of our methodology, we conducted model training using a total of 14 algorithms, including 13 machine learning algorithms and 1 deep learning algorithm. These algorithms were applied to the preprocessed dataset that underwent encoding, feature scaling, and train-test splitting. The accuracy of all these algorithms is presented in Table 5.1 Furthermore, this section discusses some of these algorithms in detail.

Gradient Boosting

Gradient boosting is a machine learning ensemble approach which integrates the predictions of numerous base models to produce a more accurate and resilient final model. It operates by fitting weak learners, often decision trees, to the residuals of previous model predictions iteratively. The residuals are the difference between the

actual target values and the predicted values from the prior model, and the next weak learner is trained to detect patterns in the residuals that the previous models missed.

The algorithm starts with an initial prediction for each data point and then computes the negative gradient of the loss function with respect to these initial predictions. This gradient represents the direction in which the model's predictions need to be adjusted to minimize the loss function. The next weak learner, typically a decision tree, is trained to predict the negative gradient, which is added to the initial predictions to update the overall model. The algorithm continues this process for a specified number of iterations or until a predefined stopping criterion is met. Each iteration of gradient boosting adjusts the predictions of the previous models by adding new predictions from the weak learner, with the learning rate controlling the magnitude of the update. The final prediction of the gradient boosting model is obtained by summing the predictions from all the weak learners, weighted by the learning rate.

Mathematically, the update formula for the gradient boosting algorithm can be represented as follows:

For iteration $m = 1$ to M :

1. Compute the negative gradient of the loss function with respect to the current predictions:

$$rim = -\partial L(y_i, y^i) / \partial y^i \quad (4.2)$$

2. Train a weak learner, typically a decision tree, to predict the negative gradient:

$$hm(xi) = WeakLearner(X, rim) \quad (4.3)$$

3. Update the current predictions with the predictions from the weak learner, weighted by the learning rate:

$$y^i = y^i + \alpha hm(xi) \quad (4.4)$$

4. Repeat steps 1-3 for M iterations or until a predefined stopping criterion is met. Here,

- $L(y_i, y^i)$ is the loss function that measures the error between the actual target values y_i and the current predictions y^i .
- rim is the negative gradient of the loss function with respect to the current predictions for the i -th data point at iteration m .
- $hm(xi)$ is the prediction from the weak learner (e.g., decision tree) at iteration m for the i -th data point.
- α is the learning rate, a hyperparameter that controls the magnitude of the update at each iteration.
- M is the total number of iterations or boosting rounds.

The gradient boosting algorithm iteratively improves the model predictions by fitting weak learners to the residuals and updating the predictions based on the negative gradient of the loss function. This process continues until a stopping criterion is met, resulting in an ensemble model that can capture complex patterns in the data and achieve high predictive accuracy.

4.2.4 Random Forest

Random Forest is an ensemble method that combines multiple decision trees to form a robust and accurate predictive model. It uses a combination of decision trees to overcome the limitations of individual trees, such as overfitting and bias. Random Forest uses a bagging technique, where each tree is trained on a randomly sampled subset of the training data with replacement. This helps to reduce the risk of overfitting and improve the generalization performance of the model. Random Forest provides interpretable outputs in the form of decision trees, which can be visualized and easily understood. This allows for model interpretation and explanation, making it suitable for applications where model transparency is important. Random Forest provides a feature importance score, which indicates the relative importance of each feature in the model's decision-making process. This can be used for feature selection, interpretation, and model explanation. The functioning of Random Forest is as follows.

1. Initialize the model with the number of decision trees N and the maximum depth d for each tree.
2. For each decision tree $t=1,2,\dots,N$:
 - a. Randomly select a subset of the training data with replacement, typically called a "bootstrap" sample. Let D_t denote the bootstrap sample.
 - b. Train a decision tree T_t on the bootstrap sample D_t with a maximum depth of d . The decision tree is trained using a feature subset that is randomly selected at each split. Let $T_t(x)$ denote the prediction of the t -th decision tree for the input data point x .
3. Predict the class label of a new data point x using the majority vote of the predictions from all the decision trees: $\text{Prediction}(x)=\text{mode}(T_1(x),T_2(x),\dots,T_N(x))$ where the mode function returns the most frequent prediction among all the decision trees.

In this notation, D_t represents the bootstrap sample used to train the t -th decision tree, and $T_t(x)$ represents the prediction of the t -th decision tree for the input data point x .

The Random Forest algorithm uses a collection of decision trees, each trained on a randomly selected subset of the training data with replacement. This introduces randomness into the model and helps to reduce overfitting. The majority vote of the individual tree predictions is used as the final prediction for a new data point, making Random Forest a powerful and robust ensemble learning algorithm for classification tasks. Overall, Random Forest is a powerful and flexible algorithm that offers several useful features beyond its mathematical aspects, making it widely used in various machine learning tasks and applications.

4.2.5 XGBoost

XGBoost, short for Extreme Gradient Boosting, is a popular gradient boosting algorithm that offers several additional features. XGBoost provides built-in support for L1 (Lasso) and L2 (Ridge) regularization techniques, which help to prevent

overfitting and improve model generalization performance. Regularization can be controlled through hyperparameters, allowing for fine-tuning of the model. XGBoost supports early stopping, where the model training can be automatically stopped if no improvement in model performance is observed on a validation set over a certain number of iterations. This helps to prevent overfitting and reduces training time. XGBoost provides a feature importance score, which indicates the relative importance of each feature in the model’s decision-making process. This can be used for feature selection, interpretation, and model explanation. Mathematically, the XGBoost algorithm can be described as follows.

1. Given a training dataset with input features denoted as X and corresponding target labels denoted as y .
2. Initialize the model with a constant prediction \hat{y} for all instances in the training set, usually set to the mean of the target labels.
3. For $t=1,2,\dots,T$, where T is the number of boosting iterations:
 - a. Compute the negative gradient of the loss function with respect to the predicted values, denoted as

$$gt = -\frac{\partial \text{loss}(y, \hat{y})}{\partial \hat{y}} \quad (4.5)$$

where $\text{loss}(y, \hat{y})$ is the chosen loss function.

- b. Fit a weak learner (e.g., decision tree) to the negative gradient gt with respect to the input features X , and obtain a prediction denoted as $ht(X)$.
 - c. Update the model by adding the prediction $ht(X)$ scaled by a learning rate η , denoted as $\hat{y} \leftarrow \hat{y} + \eta ht(X)$.
4. Once the desired number of boosting iterations T is reached, the final model \hat{y} is obtained.

The XGBoost algorithm uses a gradient boosting framework to sequentially train decision trees based on the gradients and Hessians of the loss function, and combines their predictions with a learning rate to update the predicted values. The sigmoid function is used to convert the final predicted values into class probabilities for classification tasks. XGBoost is known for its accuracy, efficiency, and scalability, making it a popular choice for many machine learning tasks. Overall, XGBoost combines the power of gradient boosting with additional features for efficient and effective model training, making it a widely used and popular algorithm in machine learning competitions and real-world applications.

4.3 IRIS Species Classification and Explanation

4.3.1 Dataset Description

We chose the popular “IRIS Dataset” [49] as a part of our experimental evaluation to assess the performance and effectiveness of our XOPSIS algorithm. This dataset offers a suitable testing ground for our algorithm’s capabilities in accurately classifying iris flowers into their respective species. By applying XOPSIS to the IRIS dataset, we aim to validate the algorithm’s performance and demonstrate its potential for generating comprehensive explanations and insights in a well-established and widely used dataset.

The IRIS dataset is a well-known benchmark dataset in the field of machine learning and consists of samples from three different species of iris flowers: Iris-setosa, Iris-versicolor, and Iris-virginica. It serves as an excellent example for classification tasks.

The dataset comprises four features, namely sepal length, sepal width, petal length, and petal width. These features provide measurements in centimeters and represent different aspects of the iris flowers' morphology. The sepal length denotes the length of the iris flower's sepal, while the sepal width represents its width. Similarly, the petal length and petal width indicate the length and width of the iris flower's petals, respectively.

The target variable in the IRIS dataset is the "Species" column, which classifies each sample into one of the three iris species mentioned earlier: Iris-setosa, Iris-versicolor, and Iris-virginica. This categorical variable enables the use of supervised learning techniques for classification tasks.

4.3.2 Dataset Preprocessing

To ensure the reliability and consistency of the data, a series of preprocessing steps were applied to the IRIS dataset. The following subsection describes each step in detail.

Checking for Null Values

Prior to analysis, the dataset was examined for any missing or null values. Fortunately, no null values were found, indicating that the dataset is complete and no imputation or data filling was required. This ensures the integrity and reliability of the data used for the subsequent analysis.

Encoding of Categorical Variable

Since the target variable "Species" contains categorical values (Iris-setosa, Iris-versicolor, Iris-virginica), an ordinal encoder was employed to convert these categorical labels into numerical representations. The ordinal encoding method was chosen to maintain the order of the categories, with Iris-setosa, Iris-versicolor, and Iris-virginica encoded as 0, 1, and 2, respectively. This transformation enables the utilization of categorical variables as input to machine learning algorithms, facilitating the classification task.

Feature Scaling

To bring the numerical features of the dataset onto a similar scale and avoid any dominance of one feature over others during model training, standard scaling (z-score normalization) was applied to the following features: sepal length, sepal width, petal length, petal width. Standard scaling transforms the features to have zero mean and unit variance, enhancing the comparability and training process of the model. This normalization step ensures that all features contribute equally and prevents bias towards features with larger values.

Train-Test Split

To evaluate the performance of the XOPSIS algorithm on the IRIS dataset, the dataset was divided into training and testing sets using an 80:20 ratio. The training set, comprising 80% of the dataset, was used to train the machine learning model, while the remaining 20% served as an independent test set for evaluating the model's performance. This partitioning enables the assessment of the model's generalization capacity and its ability to accurately classify iris flowers into their respective species.

K-fold Cross-Validation

In order to obtain a more reliable estimate of the model's performance and minimize the impact of variable train-test splits, k-fold cross-validation was employed. The dataset was divided into k equally sized folds, where each fold was used as the test set once, while the remaining k-1 folds were used for training. This process was repeated k times, with performance metrics such as accuracy, precision, recall, and F1 score calculated for each fold. The average performance across all folds was reported as the final evaluation of the XOPSIS algorithm on the IRIS dataset. This technique ensures a robust evaluation of the model's performance and reduces the variability associated with different train-test splits.

Hyperparameter Tuning

Hyperparameters play a crucial role in determining the performance of a machine learning model. Grid Search CV, a popular hyperparameter tuning technique, was employed to identify the optimal set of hyperparameter values for the XOPSIS algorithm. A predefined grid of hyperparameter values was specified, and the model was trained and evaluated using k-fold cross-validation for each combination of hyperparameters. The performance metrics were recorded, and the combination of hyperparameter values yielding the best performance was selected as the optimal set. This rigorous optimization process enhances the accuracy, robustness, and generalization ability of the XOPSIS algorithm on the IRIS dataset.

By following these dataset preprocessing steps, we ensure the reliability and suitability of the IRIS dataset for evaluating the performance and effectiveness of the XOPSIS algorithm.

4.3.3 Model Training

In the next step of our methodology, we conducted model training using three different algorithms: Gradient Boosting, Random Forest, and XGBoost. These algorithms were applied to the preprocessed dataset that underwent encoding, feature scaling, and train-test splitting. The accuracy of all these algorithms is presented in Table 5.3.

It is important to note that while we previously provided detailed descriptions of these algorithms in an earlier section, this subsection focuses on mentioning the application of these algorithms without repeating the detailed model descriptions. Instead, we refer readers back to the earlier section for comprehensive explanations of each algorithm.

The results and performance metrics of these three algorithms applied to the Iris dataset are presented in Table 5.3 for reference.

4.4 Breast Cancer Prediction and Explanation

4.4.1 Dataset Description

The dataset [50] used in this study is a well-known benchmark dataset commonly used for training and evaluating machine learning algorithms. It comprises biopsy features of 569 breast masses classified as malignant (cancer) or benign (not cancer). The features in the dataset were extracted computationally from digital images of fine needle aspirate biopsy slides. These features correspond to various properties of cell nuclei, including size, shape, and regularity. Specifically, the dataset provides the mean, standard error, and worst values of 10 nuclear parameters, resulting in a total of 30 features.

Each feature in the dataset is associated with a specific description:

- radius_mean: Mean radius of the tumor cells
- texture_mean: Mean texture of the tumor cells
- perimeter_mean: Mean perimeter of the tumor cells
- area_mean: Mean area of the tumor cells
- smoothness_mean: Mean smoothness of the tumor cells
- compactness_mean: Mean compactness of the tumor cells
- concavity_mean: Mean concavity of the tumor cells
- concave_points_mean: Mean number of concave portions of the contour of the tumor cells
- symmetry_mean: Mean symmetry of the tumor cells
- fractal_dimension_mean: Mean “coastline approximation” of the tumor cells
- radius_se: Standard error of the radius of the tumor cells
- texture_se: Standard error of the texture of the tumor cells
- perimeter_se: Standard error of the perimeter of the tumor cells
- area_se: Standard error of the area of the tumor cells
- smoothness_se: Standard error of the smoothness of the tumor cells
- compactness_se: Standard error of the compactness of the tumor cells
- concavity_se: Standard error of the concavity of the tumor cells
- concave_points_se: Standard error of the number of concave portions of the contour of the tumor cells
- symmetry_se: Standard error of the symmetry of the tumor cells

- `fractal_dimension_se`: Standard error of the “coastline approximation” of the tumor cells
- `radius_worst`: Worst (largest) radius of the tumor cells
- `texture_worst`: Worst (most severe) texture of the tumor cells
- `perimeter_worst`: Worst (largest) perimeter of the tumor cells
- `area_worst`: Worst (largest) area of the tumor cells
- `smoothness_worst`: Worst (most severe) smoothness of the tumor cells
- `compactness_worst`: Worst (most severe) compactness of the tumor cells
- `concavity_worst`: Worst (most severe) concavity of the tumor cells
- `concave_points_worst`: Worst (most severe) number of concave portions of the contour of the tumor cells
- `symmetry_worst`: Worst (most severe) symmetry of the tumor cells
- `fractal_dimension_worst`: Worst (most severe) “coastline approximation” of the tumor cells

The target variable in the dataset indicates whether a mass is malignant (“M”) or Benign (“B”). The predictors consist of a matrix with the mean, standard error, and worst values of 10 nuclear measurements, resulting in a total of 30 features per biopsy.

These measurements include factors such as nucleus radius, texture, perimeter, area, smoothness, compactness, concavity, number of concave portions, symmetry, and fractal dimension.

Overall, this dataset provides a comprehensive set of features derived from biopsy images to facilitate the development and evaluation of machine learning models for breast cancer classification.

4.4.2 Dataset Preprocessing

To ensure the reliability and consistency of the data, a series of preprocessing steps were applied to the breast cancer Wisconsin dataset. The following subsection describes each step in detail.

Checking for Null Values

Prior to analysis, the dataset was examined for any missing or null values. The column “Unnamed: 32” was found to have null values, and therefore it was removed from the dataset. Additionally, the column “id” was also removed from the dataset, although it did not contain any null values. By removing these columns, we ensure that the dataset is clean and free from missing values, maintaining the integrity and reliability of the data for subsequent analysis.

Encoding of Categorical Variable

Since the target variable “diagnosis” contains categorical values (“M” for malignant and “B” for benign), an ordinal encoder was employed to convert these categorical labels into numerical representations. In the ordinal encoding process, the label “M” was encoded as 0, indicating malignant, and the label “B” was encoded as 1, indicating benign. This transformation allows the utilization of the categorical variable as input to machine learning algorithms, facilitating the classification task while preserving the inherent order of the categories.

Feature Scaling

To bring the numerical features of the dataset onto a similar scale and avoid any dominance of one feature over others during model training, standard scaling (z-score normalization) was applied to the 30 features. Standard scaling transforms the features to have a zero mean and unit variance, enhancing the comparability and training process of the model. This normalization step ensures that all features contribute equally and prevents bias towards features with larger values.

Train-Test Split

To evaluate the performance of the XOPSIS algorithm on the breast cancer Wisconsin dataset, the dataset was divided into training and testing sets using an 80:20 ratio. The training set, comprising 80% of the dataset, was used to train the machine learning model, while the remaining 20% served as an independent test set for evaluating the model’s performance. This partitioning enables the assessment of the model’s generalization capacity and its ability to accurately classify breast masses as malignant or benign.

K-fold Cross-Validation

In order to obtain a more reliable estimate of the model’s performance and minimize the impact of variable train-test splits, k-fold cross-validation was employed. The dataset was divided into k equally sized folds, where each fold was used as the test set once, while the remaining k-1 folds were used for training. This process was repeated k times, with performance metrics such as accuracy, precision, recall, and F1 score calculated for each fold. The average performance across all folds was reported as the final evaluation of the XOPSIS algorithm on the breast cancer Wisconsin dataset. This technique ensures a robust evaluation of the model’s performance and reduces the variability associated with different train-test splits.

Hyperparameter Tuning

Hyperparameters play a crucial role in determining the performance of a machine learning model. Grid Search CV, a popular hyperparameter tuning technique, was employed to identify the optimal set of hyperparameter values for the XOPSIS algorithm. A predefined grid of hyperparameter values was specified, and the model was trained and evaluated using k-fold cross-validation for each combination of hyperparameters. The performance metrics were recorded, and the combination of hyperparameter values yielding the best performance was selected as the optimal

set. This rigorous optimization process enhances the accuracy, robustness, and generalization ability of the XOPSIS algorithm on the breast cancer Wisconsin dataset. By following these dataset preprocessing steps, including the removal of the “Un-named: 32” column and ordinal encoding of the “diagnosis” variable, we ensure the reliability and suitability of the breast cancer Wisconsin dataset for evaluating the performance and effectiveness of the XOPSIS algorithm.

4.4.3 Model Training

In the next step of our methodology, model training was performed using three different algorithms: Gradient Boosting, Random Forest, and XGBoost. These algorithms were applied to the preprocessed breast cancer Wisconsin dataset, which had undergone encoding, feature scaling, and train-test splitting. The accuracy of each algorithm is presented in Table 5.5.

It is important to note that while detailed descriptions of these algorithms were provided earlier in the paper, this subsection focuses on highlighting their application without repeating the comprehensive model explanations. Readers are referred back to the earlier section for a thorough understanding of each algorithm.

The results and performance metrics of these three algorithms applied to the Breast Cancer Wisconsin Dataset are presented in Table 5.5 for easy reference.

4.5 Car Acceptability Prediction and Explanation

4.5.1 Data Collection

The Car Acceptability dataset [51] used in this research was collected from Kaggle. It is based on a simple hierarchical decision model originally developed for the demonstration of the DEX expert system for decision making (M. Bohanec, V. Rajkovic: Expert system for decision making. *Sistemica* 1(1), pp. 145-157, 1990). The Car Acceptability dataset used in this research consists of 1729 unique rows and contains a target variable, Car_Acceptability, along with six feature variables. The goal of the analysis is to predict the car acceptability based on the values of these features. The dataset evaluates cars according to the following features:

- Buying_Price: Categorical Data [vhigh, high, med, low]
- Maintenance_Price: Categorical Data [vhigh, high, med, low]
- No_of_Doors: Categorical Data [2, 3, 4, 5more]
- Person_Capacity: Categorical Data [2, 4, more]
- Size_of_Luggage: Categorical Data [small, med, big]
- Safety: Categorical Data [low, med, high]

The Car_Acceptability variable serves as the target variable and is also categorical, with the following values: [unacc, acc, good, vgood]. The objective is to develop a model using XOPSIS that accurately predicts the car acceptability based on the given features.

The Car Acceptability Classification Database is derived from this decision model and is available under the CC-BY-NC-SA 4.0 license for non-commercial usage. The dataset is provided by the UCI Machine Learning Repository, a widely recognized and reliable source for machine learning datasets.

This dataset offers a valuable opportunity to evaluate the effectiveness of XOPSIS in the classification of car acceptability based on various features. By applying XOPSIS to this dataset, we aim to gain insights into the decision-making process of the model and generate comprehensive explanations for the predictions. The analysis of this dataset will contribute to the understanding of XOPSIS's performance in a new domain and its potential for providing interpretable insights in the field of car acceptability classification.

4.5.2 Dataset Preprocessing

To ensure the reliability and consistency of the Car Acceptability dataset, several preprocessing steps were applied. The following subsection outlines each step in detail.

Checking for Null Values

The dataset was first examined for any missing or null values. No missing values were found, ensuring that the dataset is complete and does not require imputation or further handling of missing data.

Encoding of Categorical Variables

As the dataset contains categorical variables, including the target variable “Car_Acceptability” and the feature variables, an encoding process was performed to transform these categorical labels into numerical representations. Specifically, an ordinal encoder was used to assign numerical values to each category. The encoding mappings are as follows:

- Buying_Price: [vhigh = 3, high = 2, med= 1, low = 0]
- Maintenance_Price: [vhigh = 3, high = 2, med= 1, low = 0]
- No_of_Doors: [2 = 0, 3 = 1, 4 = 2, 5more = 3]
- Person_Capacity: [2 = 0, 4 = 1, more = 2]
- Size_of_Luggage: [small = 0, med = 1, big = 2]
- Safety: [low = 0, med = 1, high = 2]
- Car_Acceptability (Target): [unacc=0, acc=1, good=2, vgood=3]

This encoding process enables the utilization of the categorical variables as input to machine learning algorithms, allowing for effective classification tasks while preserving the inherent order of the categories.

Feature Scaling

To ensure that the numerical features are on a similar scale and avoid dominance of any particular feature during model training, feature scaling was performed. Specifically, standard scaling (z-score normalization) was applied to the feature variables. This transformation centers the data by subtracting the mean and scales it by dividing by the standard deviation, resulting in features with a mean of 0 and a standard deviation of 1. This normalization step enhances the comparability and training process of the model, ensuring that each feature contributes equally to the model's predictions.

Train-Test Split

To evaluate the performance of the XOPSIS algorithm on the Car Acceptability dataset, the dataset was split into training and testing sets using an 80:20 ratio. The training set, comprising 80% of the dataset, was used to train the machine learning model, while the remaining 20% served as an independent test set for evaluating the model's performance. This division enables the assessment of the model's generalization capacity and its ability to accurately classify car acceptability.

K-fold Cross-Validation

To obtain a more reliable estimate of the model's performance and reduce the impact of variable train-test splits, 10-fold cross-validation was employed. The Car Acceptability dataset was divided into 10 equally sized folds. The model was trained and evaluated using each fold as the test set once, while the remaining nine folds were used for training. This process was repeated 10 times, with performance metrics calculated for each fold using the k-fold cross-validation method.

Hyperparameter Tuning

Hyperparameters play a crucial role in determining the performance of a machine learning model. To optimize the XOPSIS algorithm's performance on the Car Acceptability dataset, hyperparameter tuning was performed using Grid Search CV. A predefined grid of hyperparameter values was specified, and the model was trained and evaluated using k-fold cross-validation for each combination of hyperparameters. The performance metrics were recorded, and the combination of hyperparameter values yielding the best performance was selected as the optimal set. This rigorous optimization process enhances the accuracy, robustness, and generalization ability of the XOPSIS algorithm on the Car Acceptability dataset.

By following these dataset preprocessing steps, including the encoding of categorical variables, feature scaling, and k-fold cross-validation, we ensure the reliability and suitability of the Car Acceptability dataset for evaluating the performance and effectiveness of the XOPSIS algorithm.

4.5.3 Model Training

In the model training phase, we employed three different algorithms: Gradient Boosting, Random Forest, and XGBoost. These algorithms were trained on the

preprocessed Car Acceptability dataset, which had undergone preprocessing steps including encoding of categorical variables, feature scaling, and train-test splitting. Each algorithm was trained on the training set of the dataset, using the labeled data to learn the underlying patterns and relationships between the features and the target variable. The goal of model training is to optimize the algorithm's parameters and adjust the model's internal mechanisms to make accurate predictions on unseen data.

During the training process, the algorithms iteratively adjusted their internal parameters based on the provided features and their corresponding target labels. This iterative optimization process aimed to minimize the discrepancy between the predicted outputs and the true labels in the training set.

By training multiple algorithms on the preprocessed Car Acceptability dataset, we aimed to evaluate their performance and effectiveness in capturing the underlying patterns and making accurate predictions. The trained models will be further evaluated and compared using various performance metrics in the subsequent analysis.

Chapter 5

Result Analysis

This section of the study encompasses three key subsections such as Accuracy, Feature Importance Plot, and Interpretability through XAI Methods. In the Accuracy subsection, we present and elaborate on the accuracy plots, providing comprehensive details on the performance of the developed models. Moving on to the Feature Importance Plot subsection, we showcase three different plots, namely the feature importance ranking using SHAP summary plot with Random Forest, feature importance ranking using SHAP summary plot with XGBoost, and feature importance ranking using TOPSIS methods with Gradient Boosting. Lastly, in the Interpretability through XAI Methods subsection, we demonstrate the explanations generated by XOPSIS, LIME, and SHAP for specific instances, offering multiple insightful cases to illustrate the interpretability of our proposed approach and compare it with existing methods.

5.1 Experimental Setup

The experiments were conducted using the Google Colab platform, which provided a robust and scalable environment for coding and analysis. The computational resources offered by Google Colab were leveraged to efficiently execute the code and algorithms. The experiments were performed on a computer with the following specifications:

- Processor: Intel Core i5-8265U CPU @ 1.60GHz, 1800Mhz, 4 Core(s), 8 Logical Processor(s)
- RAM: 8 GB DDR4
- GPU: Intel(R) UHD Graphics 620
- Storage: ST1000LM035-1RK172 (1 TB)
- Operating System: Windows (Microsoft Windows 11 Pro)

Python, a widely used programming language in machine learning research, was employed for coding. The primary libraries utilized include scikit-learn for machine learning implementations, matplotlib for data visualization, and numpy for numerical computations. The experimentation process involved the following key steps:

1. **Data Preprocessing:** The datasets were loaded, and preprocessing steps such as feature scaling and splitting into training and testing sets were performed.
2. **Model Training:** The machine learning models, including Gradient Boosting and XGBoost, were trained on the training data using optimized hyperparameters.
3. **Prediction and Evaluation:** The trained models were used to make predictions on the testing data. Accuracy, confusion matrix, sensitivity, specificity, and ROC curves were computed for model evaluation.
4. **Explainability Techniques:** LIME, SHAP, and TOPSIS-based XAI methods (XOPSIS) were applied to interpret the model predictions. Explanations were generated for specific instances, highlighting feature importance and decision rationales.
5. **Analysis and Interpretation:** The results obtained from different models and explainability techniques were analyzed and compared to gain insights into the models' performance and interpretability.

The experimental setup ensured consistency and reproducibility of the results, allowing for an in-depth analysis of the models' behavior and the effectiveness of the XAI techniques.

5.2 Result Analysis on Maternal Health Dataset

5.2.1 Accuracy

As demonstrated in Table 5.1, the train-test split technique outperforms the k-fold cross-validation method in terms of accuracy for all 14 algorithms. The resulting accuracy scores are depicted in a bar plot in Figure 5.1.

From Table 5.1, it is evident that the accuracy varies significantly across different algorithms. The top-performing algorithms based on the train-test split technique are Random Forest, Gradient Boosting, and XGBOOST, while the least accurate algorithms are Logistic Regression, Gaussian Naive Bayes, and AdaBoost. Furthermore, the k-fold cross-validation technique generally produces lower accuracy scores compared to the train-test split technique, likely due to the use of multiple subsets of data for training and testing, which can result in a more generalized model but with slightly lower accuracy.

The bar plot in Figure 5.1 visually represents the accuracy scores for each algorithm using the train-test split technique, revealing that the top-performing algorithms have accuracy scores above 90%, while the least accurate algorithms have scores below 75%. Notably, the Gradient Boosting algorithm achieves the highest accuracy score of 90.64% among all the tested algorithms.

5.2.2 Confusion Matrix

The confusion matrix presented in Figure 5.2 is a 3x3 matrix that represents the classification results for three risk levels: low risk, mid risk, and high risk. The matrix

Serial No	Algorithms	Accuracy	
		Train-Test Split	K-Fold Cross Validation
01	KNN	89.163%	85.389%
02	Random Forest	90.148%	84.402%
03	Gradient Boosting	90.640%	84.990%
04	XGBoost	89.655%	81.345%
05	CatBoost	87.192%	81.839%
06	Bagging Classifier	89.163%	82.434%
07	AdaBoost	75.862%	62.713%
08	Decision Tree	88.670%	83.023%
09	Extra Tree Classifier	89.655%	83.032%
10	Logistic Regression	72.906%	64.793%
11	Support Vector Classifier	84.729%	82.338%
12	Gaussian Naive Bayes	71.921%	64.577%
13	Voting Classifier	89.655%	84.011%
14	Multi Layer Perceptron	83.744%	77.708%

Table 5.1: Accuracy of 14 Algorithms

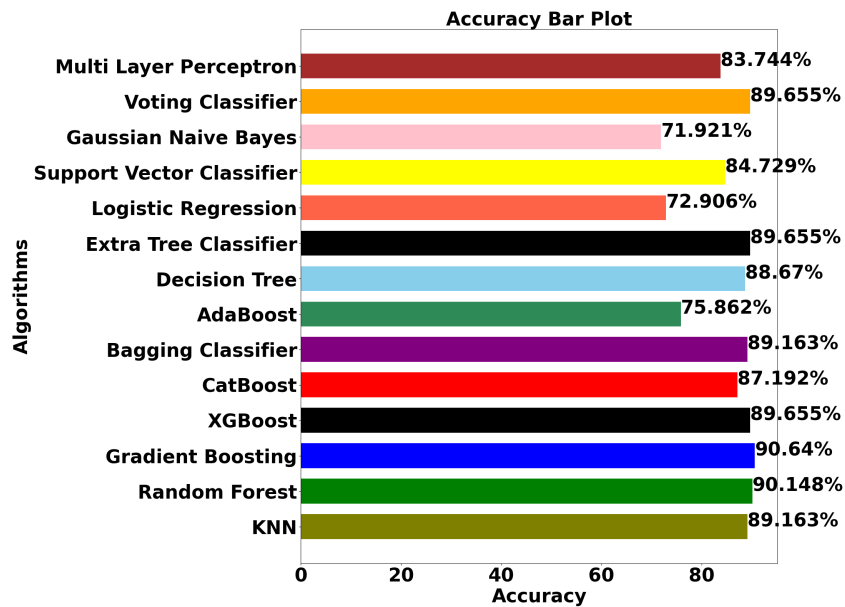


Figure 5.1: Train-Test Split Method Accuracy of 14 Algorithms

provides a comprehensive overview of the performance of the model by displaying the true labels on the left side and the predicted labels at the bottom.

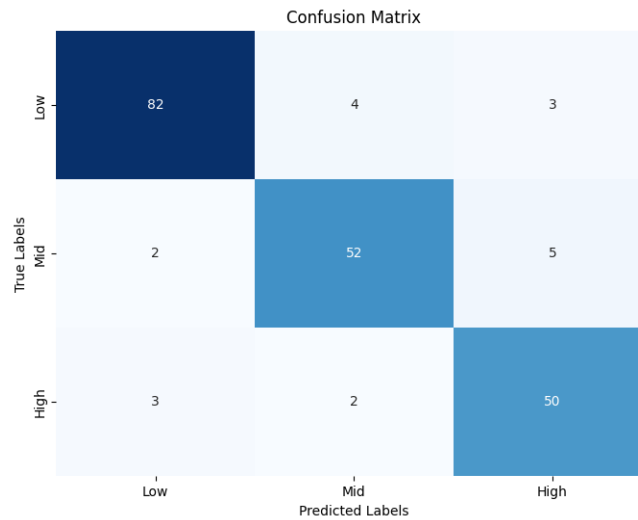


Figure 5.2: Confusion Matrix of Gradient Boosting Model

In the first row of the matrix, it is observed that out of the instances labeled as low risk, 82 were correctly predicted as low risk, while 4 instances were misclassified as mid risk and 3 instances as high risk. Moving to the second row, which represents instances labeled as mid risk, the model correctly predicted 52 instances as mid risk, misclassified 2 instances as low risk, and 5 instances as high risk. Lastly, in the third row representing high-risk instances, the model correctly classified 50 instances as high risk, misclassified 3 instances as low risk, and 2 instances as mid risk.

This confusion matrix provides a detailed breakdown of the classification results, enabling a deeper analysis of the model's performance for different risk levels. By examining the matrix, it becomes evident that the model exhibits a higher accuracy in predicting low-risk and mid-risk instances compared to high-risk instances. The presented confusion matrix serves as a valuable tool for evaluating the performance of the classification model and provides insights into the effectiveness of the risk level prediction.

5.2.3 Sensitivity, Specificity, Precision and F1-Score

We focus on evaluating the performance of the Gradient Boosting model, which we selected as our final model based on its highest accuracy score obtained earlier. We assess its effectiveness by analyzing several important performance metrics, including precision, recall, F1-score, sensitivity, and specificity. These metrics provide valuable insights into the ability of the model to accurately classify instances and its performance across different classes. Table 5.2 shows the values of performance evaluation metrics.

The plot showed in Figure 5.3 illustrates the performance metrics for a multiclass classification task across three classes: Low, Mid, and High Risk. The x-axis represents the classes, with Low on the left side, Mid in the middle, and High on the right side. The y-axis represents the scores of the performance metrics. The plot includes four performance metrics: Sensitivity, Specificity, Precision, and F1-Score.

	Sensitivity	Specificity	Precision	F1-Score
Low Risk	0.92	0.96	0.94	0.93
Mid Risk	0.88	0.96	0.90	0.89
High Risk	0.91	0.95	0.86	0.88

Table 5.2: Performance Evaluation Metrics

These metrics provide valuable insights into the model’s performance in correctly classifying instances and its overall effectiveness across different classes.

Sensitivity, also known as the true positive rate, measures the proportion of correctly classified positive instances for each class. A higher sensitivity indicates that the model is effective at identifying instances belonging to a specific class.

Specificity, also known as the true negative rate, measures the proportion of correctly classified negative instances for each class. A higher specificity indicates that the model is effective at correctly classifying instances not belonging to a specific class.

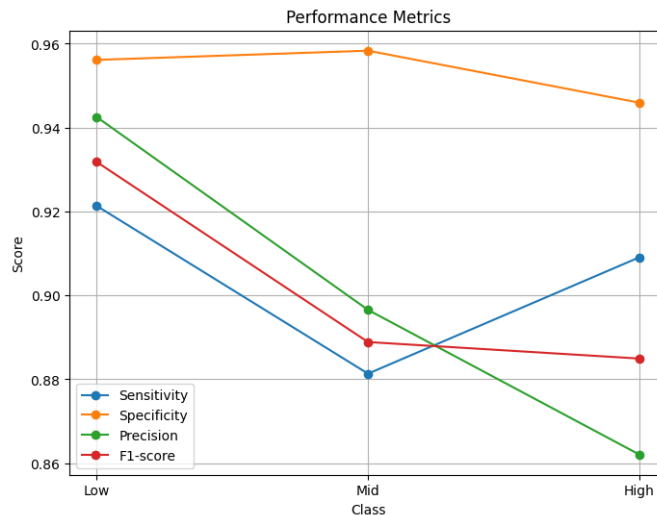


Figure 5.3: Comparison of Sensitivity, Specificity, Precision, and F1-Score Across Three Risk Levels

Precision measures the proportion of correctly classified instances for each class among all instances predicted to belong to that class. A higher precision indicates that the model has a low rate of falsely classifying instances into a specific class.

The F1-Score is a harmonic mean of precision and recall (which is equivalent to sensitivity in this case). It provides a balanced measure of the model’s performance by considering both precision and recall. A higher F1-Score indicates a better trade-off between precision and recall for each class.

By plotting these performance metrics together, we can observe the relative performance of the model across the Low, Mid, and High classes. We can see that the model performs consistently well across the classes, with generally high scores for Sensitivity, Specificity, Precision, and F1-Score. However, there are some variations in the scores across the classes, indicating potential differences in the model’s performance in correctly classifying instances among the different classes.

5.2.4 ROC Curve

The Receiver Operating Characteristic (ROC) curve provides valuable insights into the model's trade-off between sensitivity (true positive rate) and specificity (true negative rate), allowing for the evaluation of its performance across different thresholds. The area under the ROC curve (AUC) is often used as a summary metric to quantify the overall performance of the model, with a higher AUC indicating better classification performance.

Figure 5.4 illustrates the ROC curves generated by the Gradient Boosting model for the Maternal health risk dataset. The ROC curve showcases the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) for each risk level: low, mid, and high. The x-axis represents the FPR, while the y-axis represents the TPR. We depict three ROC curves, each corresponding to a different risk level, in distinctive colors. This consolidated visualization allows us to discern the model's ability to discriminate between different health risk levels. Notably, the ROC curves exhibit a characteristic upward trend, indicative of the model's skill in distinguishing positive instances.

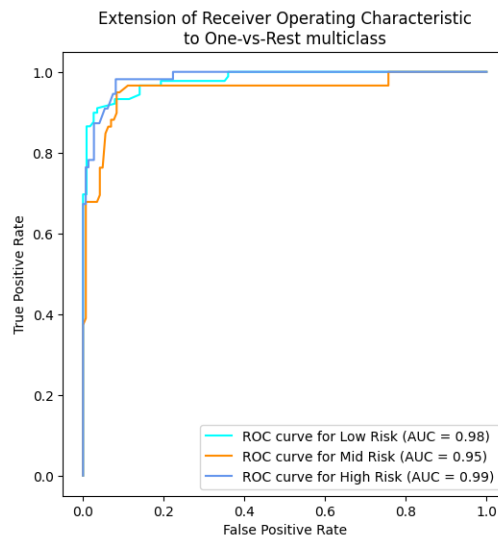


Figure 5.4: ROC Curve for Gradient Boosting

5.2.5 AUC

Figure 5.5 showcases the AUC scores obtained from the Gradient Boosting model for the Maternal health risk dataset. The AUC score quantifies the discriminative power of the model for each risk level: low, mid, and high. On the x-axis, we denote the three risk levels, while the y-axis indicates the corresponding AUC scores. The obtained AUC scores are as follows: 0.98 for low risk, 0.95 for mid risk, and 0.99 for high risk. These scores provide a succinct summary of the model's ability to rank instances correctly, further substantiating its effectiveness in distinguishing between varying maternal health risk levels. Notably, the consistently high AUC scores underscore the model's robust performance across all risk categories.

- For the Low risk level, the AUC score is 0.98

- For the Mid risk level, the AUC score is 0.95
- For the High risk level, the AUC score is 0.99

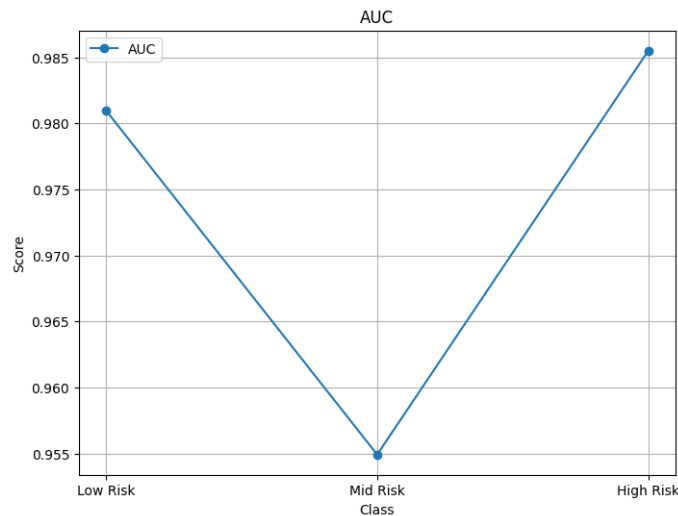


Figure 5.5: Area Under the Curve (AUC)

In summary, the ROC and AUC analyses collectively demonstrate the Gradient Boosting model’s efficacy in predicting maternal health risk levels across diverse scenarios. The ROC curves highlight the model’s true positive rate across various false positive rates, while the AUC scores quantitatively substantiate its discriminative capability for each risk level.

5.2.6 Feature Importance Plot

The SHAP summary plot is a visual representation of the SHAP values for all features in a machine learning model. It provides a comprehensive overview of how each feature contributes to model predictions across all instances in a dataset. By examining the summary plot, patterns and trends in feature contributions can be identified, helping to understand the importance and directionality of the impact of each feature on model predictions. This can aid in model interpretation and decision-making by providing a holistic view of feature importance and their effects on model outcomes.

Figure 5.6 displays that the Random Forest model ranks the features based on their importance. The most important feature is BS (Blood Sugar), which is ranked as the top feature. The second most important feature is SystolicBP (Systolic Blood Pressure), followed by Age as the third most important feature. DiastolicBP (Diastolic Blood Pressure) is ranked as the fourth most important feature, HeartRate (Heart Rate) as the fifth, and BodyTemp (Body Temperature) as the least important feature.

Similar to the Random Forest model, Figure 5.7 shows that the XGBoost model ranks the features based on their importance. In this case, SystolicBP (Systolic Blood Pressure) is ranked as the most important feature, followed by BS (Blood Sugar) as the second most important feature, and Age as the third most important feature. DiastolicBP (Diastolic Blood Pressure) is ranked as the fourth most

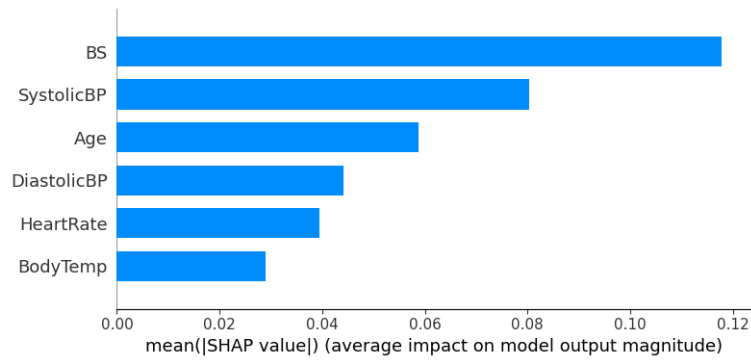


Figure 5.6: Feature Importance Ranking using SHAP Summary Plot with Random Forest

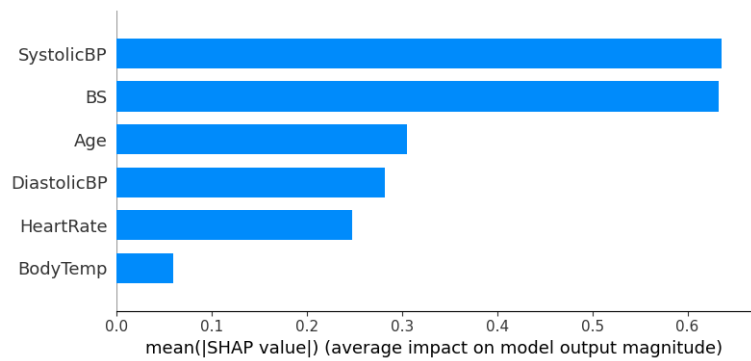


Figure 5.7: Feature Importance Ranking using SHAP Summary Plot with XGBoost

important feature, HeartRate (Heart Rate) as the fifth, and BodyTemp (Body Temperature) as the least important feature.

The Topsis method calculates the feature importance based on the similarity of feature values between the given data points and similar instances. In this method, SystolicBP (Systolic Blood Pressure) is ranked as the most important feature, followed by Age as the second most important feature, and BS (Blood Sugar) as the third most important feature. BodyTemp (Body Temperature) is ranked as the fourth most important feature, DiastolicBP (Diastolic Blood Pressure) as the fifth, and HeartRate (Heart Rate) as the least important feature, shown in Figure 5.8.

In summary, based on the three methods used in this research work, the top three important features are SystolicBP, Age, and BS, although the order of their importance may vary depending on the method and model used. It appears that SystolicBP (Systolic Blood Pressure) is consistently ranked as the most important feature by two of the methods used, namely the SHAP summary plot with XGBoost and the Topsis method. This suggests that SystolicBP may have a higher impact on the target variable or outcome compared to other features in our dataset according to these methods. In conclusion, based on our comprehensive analysis of feature importance using multiple methods, SystolicBP, Age, and BS consistently emerged as the most important features in our research work.

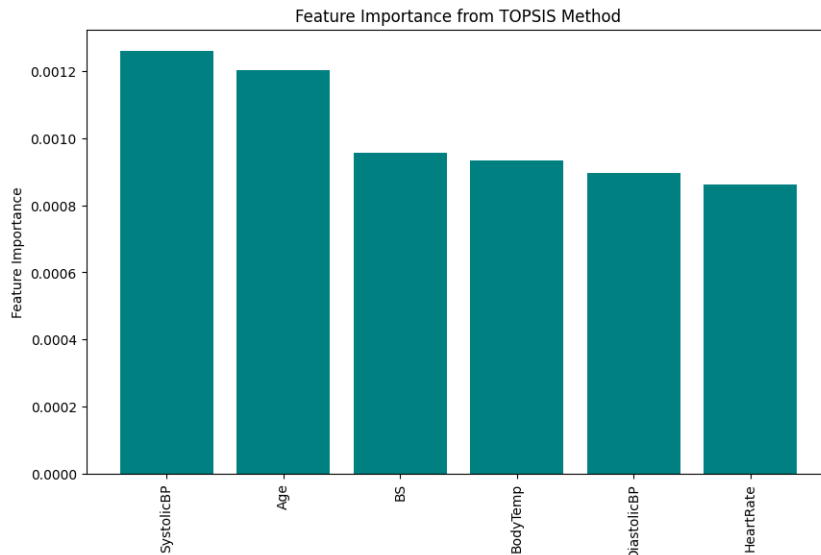


Figure 5.8: Feature Importance Ranking using TOPSIS Method

5.2.7 Interpretability Through XAI Methods

1. High Risk Explanation

Case 01

From Figure 5.9, it appears to be a summary of a prediction result for a specific data point. It starts by showing the actual risk level and the predicted risk level, both of which are 2.0 indicating a high risk level.

```

----- TOPSIS Explanation -----
Data point index: 954
----- Data Point Features -----
Age          25.00
SystolicBP   140.00
DiastolicBP  100.00
BS           7.01
BodyTemp     98.00
HeartRate    80.00
Name: 954, dtype: float64
Actual risk level: 2.0
Predicted Risk Level: 2.0

Explanation:
- The TOPSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  • DiastolicBP
  • SystolicBP
- The features that contribute negatively to the prediction are:
  • Age
  • BS
  • HeartRate
  • BodyTemp

```

Figure 5.9: Explanation of high risk level by XOPSIS (Case 01)

Next, it mentions that the explanation is based on topsis score, which is calculated by comparing the similarity of feature values between the given data point and similar instances. The explanation then proceeds to mention which features contribute positively and negatively to the prediction. According to the explanation, the features that have a positive impact on the prediction are Diastolic BP and Systolic BP. On the other hand, the features that have a negative impact on the prediction are Age, BS, HeartRate and Body Temperature.

Overall, the explanation provides insight into how different features are influencing the prediction of high risk level for the given data point, with some features con-

tributing positively and others negatively to the prediction.

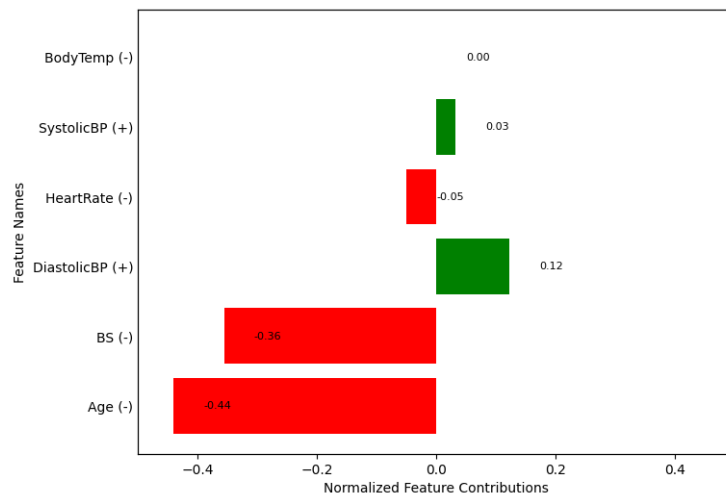


Figure 5.10: Normalized Feature Contributions of high risk level by XOPSSIS (Case 01)

The plot showed in Figure 5.10, represents normalized feature contributions on the x-axis and feature names on the y-axis. Feature contributions are represented as numerical values, with positive contributions in green and negative contributions in red. The feature contributions are as follows: Age (-0.44), BS (-0.36), BodyTemp (0.0), HeartRate (-0.05), SystolicBP (0.03), and DiastolicBP (0.12).

Figure 5.11 illustrates that the probabilities predicted by the model for different risk categories, such as “low risk,” “mid risk,” and “high risk.” In this case, the model predicts a probability of 1.0 for “high risk,” and 0.0 for both “low risk” and “mid risk.”

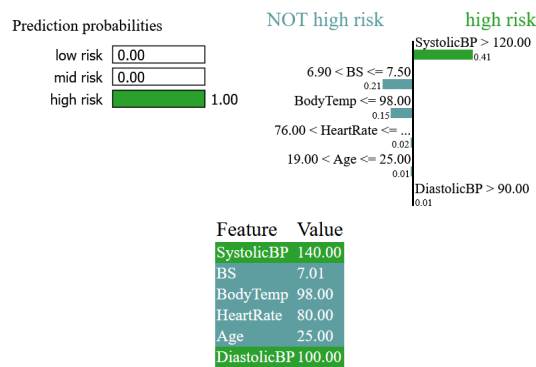


Figure 5.11: Explanation of high risk level by LIME (Case 01)

The plot generated by LIME visually presents the features that contribute to the predicted risk category. It is divided into two sections, with “high risk” on the right side and “not high risk” on the left side. On the right side of the plot, LIME identifies the features that positively contribute to the prediction of “high risk.” These features are SystolicBP > 120 with a weight of 0.41 and DiastolicBP > 90 with a weight of 0.01. The weights indicate the relative importance of these features in

determining the “high risk” prediction. On the left side of the plot, LIME identifies the features that negatively contribute to the prediction of “not high risk” These features are $6.90 < BS \leq 7.50$ with a weight of 0.21, $BodyTemp \leq 90$ with a weight of 0.15, $76 < HeartRate$ with a weight of 0.02 and $19 < Age \leq 25$ with a weight of 0.01. Similar to the features for “high risk”, these features are shown with their corresponding weights.

The LIME output also displays the actual values of the features for the specific data point that was explained. These feature values provide context and reference to better understand how the model arrived at its prediction.

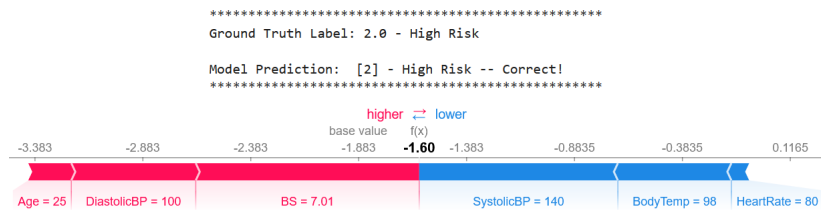


Figure 5.12: Explanation of high risk level by SHAP (Case 01)

The plot in Figure 5.12, starts by showing the Ground Truth Label, which is the actual label or class of the data point, and the Model Prediction, which is the predicted label or class by the machine learning model. In your case, it shows “Ground Truth Label: 2.0” or “High Risk” and “Model Prediction: 2.0 - High Risk - Correct!” indicating that the data point belongs to the “High Risk” class and the model has made a correct prediction.

The red colored features in the plot, such as Age=25, DiastolicBP=100, and BS=7.01, are driving up the predicted probability of belonging to the “high risk” class. These features have a positive impact on the prediction, meaning that higher values of these features are associated with a higher predicted probability of high risk. The blue colored features in the plot, such as SystolicBP=140, BodyTemp=98, and HeartRate=80, are driving down the predicted probability of belonging to the “high risk” class. These features have a negative impact on the prediction, meaning that higher values of these features are associated with a lower predicted probability of high risk.

The “higher – >< – lower” notation indicates the direction of the impact of the features on the prediction. “Higher” in red color indicates that higher values of the red features are associated with a higher predicted probability of high risk, while “lower” in blue color indicates that lower values of the blue features are associated with a lower predicted probability of high risk. In this case, higher values of the features associated with red color are driving up the predicted probability of the “high risk” class, while blue color are driving down the predicted probability of the “high risk” class. The value of -1.60 that is displayed in bold text in the middle of the red and blue colored features represents the contribution of the combined set of features associated with the red and blue data points on the predicted probability of belonging to the “high risk” class. It is the aggregated impact of all the features considered together, taking into account their respective SHAP values.

Based on the explanations provided by LIME, SHAP, and XOPSIS for the same data point, it can be concluded that features such as Diastolic BP and Systolic BP

contribute positively to the prediction of high risk, indicating that higher values of these features are associated with a higher predicted probability of high risk. Conversely, features such as Age, BS, HeartRate and Body Temperature contribute negatively to the prediction of high risk, meaning that higher values of these features are associated with a lower predicted probability of high risk. Notably, LIME and XOPIS show consistent results in terms of the direction of feature contributions, where both methods highlight the same features as positive or negative contributors to the prediction. SHAP explanation provides specific feature values that drive up or down the predicted probability of high risk. Moreover, XOPIS explanation emphasizes the similarity of feature values between the given data point and similar instances in the calculation of the TOPSIS score. Overall, these explanations offer valuable insights into how different features impact the prediction of high risk and can aid in a better understanding of the behavior of the model.

Case 02

The actual and predicted risk level of this data point is given as 2.0 in Figure 5.13 which indicates that the prediction of the model matches the actual risk level in this case.

```

----- TOPSIS Explanation -----
Data point index: 597
----- Data Point Features -----
Age                22.0
SystolicBP        90.0
DiastolicBP       60.0
BS                 7.5
BodyTemp          102.0
HeartRate         60.0
Name: 597, dtype: float64
Actual risk level: 2.0
Predicted Risk Level: 2.0

Explanation:
- The TOPSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  • BodyTemp
- The features that contribute negatively to the prediction are:
  • SystolicBP
  • DiastolicBP
  • Age
  • BS
  • HeartRate

```

Figure 5.13: Explanation of high risk level by XOPIS (Case 02)

The TOPSIS score, which is a measurement of similarity between the given data point and similar instances, is calculated based on the feature values. The explanation further provides information on which features contribute positively and negatively to the prediction. The feature “Body Temp” is mentioned to contribute positively to the prediction. This implies that higher values of “Body Temp” are associated with a higher predicted risk level. The features “SystolicBP”, “DiastolicBP”, “Age”, “HeartRate”, and “BS” are mentioned to contribute negatively to the prediction. The plot showed in Figure 5.14, represents normalized feature contributions for this data point, with positive contributions in green and negative contributions in red.

The initial section of the output from Figure 5.15, displays the prediction probabilities for different risk classes. In this case, the predicted probability for the “low risk” class is 0.0, the predicted probability for the “mid risk” class is 0.0, and the

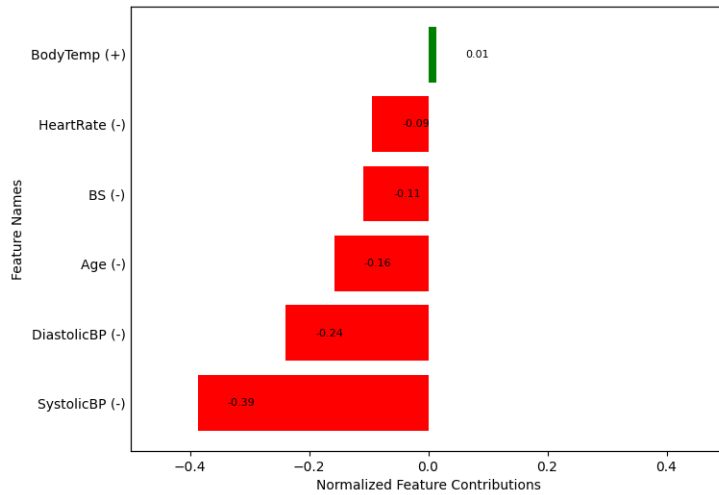


Figure 5.14: Normalized Feature Contributions of high risk level by XOPSSIS (Case 02)

predicted probability for the “high risk” class is 1.0. This indicates that the model has predicted a high risk probability for the given data point.

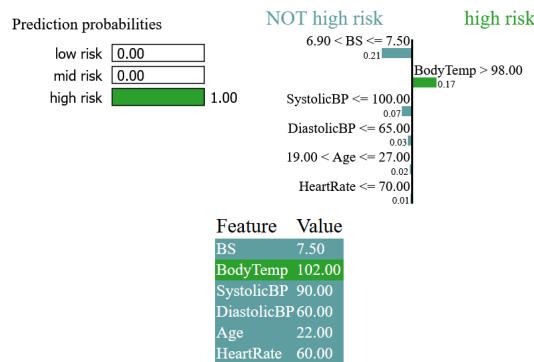


Figure 5.15: Explanation of high risk level by LIME (Case 02)

The plot in the output provides further insights into the features that are contributing to the high risk prediction. On the right side of the plot, it shows that the feature “Body Temp > 98.0” has a contribution of 0.17 towards the high risk prediction. This means that a higher value of “Body Temp” greater than 98.0 is positively impacting the prediction towards high risk. On the left side of the plot, it shows several features and their corresponding contributions towards the high risk prediction. These features and their contributions are as follows.

“BS <= 7.50” has a contribution of 0.21, indicating that a value of “BS” less than or equal to 7.50 is negatively impacting the prediction towards high risk. “SystolicBP <= 100” has a contribution of 0.07, indicating that a value of “SystolicBP” less than or equal to 100 is also negatively impacting the prediction towards high risk. Similarly, “DiastolicBP <= 65” has a contribution of 0.03, “19 < Age <= 27” has a contribution of 0.02, “HeartRate <= 70” has a contribution of 0.01. The output also displays the feature value for the given data point, which can provide additional context for the explanations above.

In summary, the LIME output suggests that a higher body temperature above 98.0 is positively contributing to the high risk prediction, while lower values of blood sugar

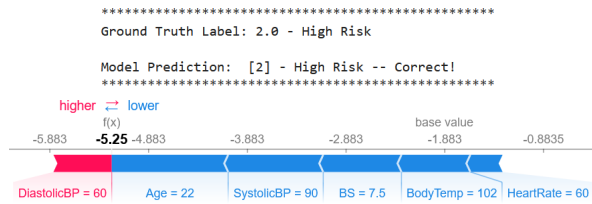


Figure 5.16: Explanation of high risk level by SHAP (Case 02)

(BS), systolic blood pressure (SystolicBP), diastolic blood pressure (DiastolicBP), Age, and HeartRate are negatively contributing to the high risk prediction.

Figure 5.16 illustrates that the predicted risk label by the model, which matches the ground truth label, indicating that the prediction is correct. The SHAP force plot shows the contributions of different features towards the prediction of high risk. It appears that higher values of (DiastolicBP = 60) have a positive impact on the prediction, driving up the probability of high risk. On the other hand, lower values of (Age=22, SystolicBP=90, BS=7.5, Body Temp=102) have a negative impact on the prediction, pulling it towards the low risk class. Therefore, DiastolicBP can be considered to have a positive impact, while the other features have a negative impact on the prediction. That means, if the DiastolicBP is greater than 60, it would have a positive impact on the prediction, pulling it towards the high risk class, as indicated by the red color and the “higher >” text. On the other hand, if the values of the other features (Age, SystolicBP, BS, and Body Temp) are lower than their current values, it would have a negative impact on the prediction, pulling it towards the low risk class, as indicated by the blue color and the “< –lower” text.

In conclusion, LIME and XOPIS explanations consistently suggest that higher body temperature (Body Temp) is positively associated with a higher predicted risk level, while SHAP explanation indicates that higher values of DiastolicBP are positively associated with a higher predicted risk level. Additionally, higher values of features such as Age, SystolicBP, BS, and HeartRate compared to the current values observed in the given data point are associated with a higher predicted risk level. This information can be useful in making informed decisions and understanding the factors that contribute to the prediction of risk level for the given data point.

Case 03

From Figure 5.17, we observe that the actual risk level of the given data point is 2.0 (high risk), and the model has predicted the same risk level of 2.0 (high risk) for this data point. The features that contribute positively to the prediction are BS, which means that higher values of BS are associated with a higher predicted risk level. On the other hand, the features that contribute negatively to the prediction are SystolicBP, Age, HeartRate, DiastolicBP, and BodyTemp. Figure 5.18 shows the normalized feature contributions to this particular data point.

Based on the LIME explanation for the new data point provided in Figure 5.19, the prediction probabilities indicate that the model has predicted a high risk with a probability of 1.0, and low and mid risk with probabilities of 0.0 each. The ac-

```

----- TOPSIS Explanation -----
Data point index: 1012
----- Data Point Features -----
Age          43.0
SystolicBP   120.0
DiastolicBP  90.0
BS           18.0
BodyTemp     98.0
HeartRate    70.0
Name: 1012, dtype: float64
Actual risk level: 2.0
Predicted Risk Level: 2.0

Explanation:
- The TOPSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  • BS
- The features that contribute negatively to the prediction are:
  • SystolicBP
  • HeartRate
  • Age
  • DiastolicBP
  • BodyTemp

```

Figure 5.17: Explanation of high risk level by XOPSIS (Case 03)

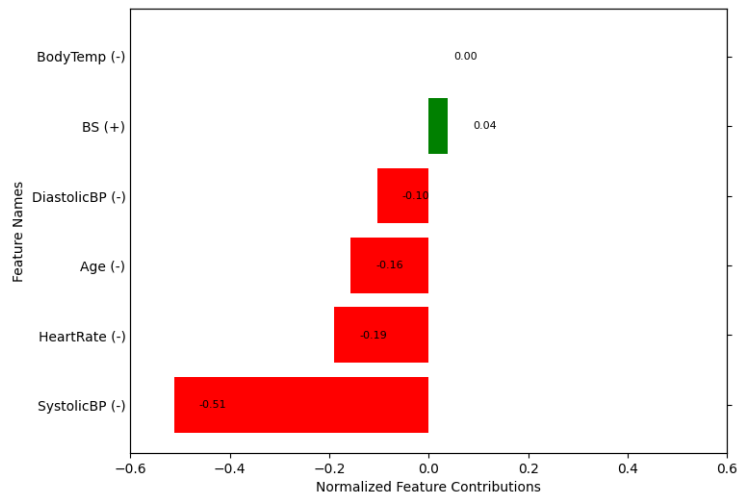


Figure 5.18: Normalized Feature Contributions of high risk level by XOPSIS (Case 03)

companying plot shows the contribution of different features towards the prediction of high risk. It appears that the feature “BS > 8.0” is the only one shown on the right side of the plot and is positively contributing to the prediction of “High Risk” with a contribution value of 0.74. The other features are shown on the left side of the plot and are contributing negatively to the prediction of “Not High Risk” with their respective contribution values.

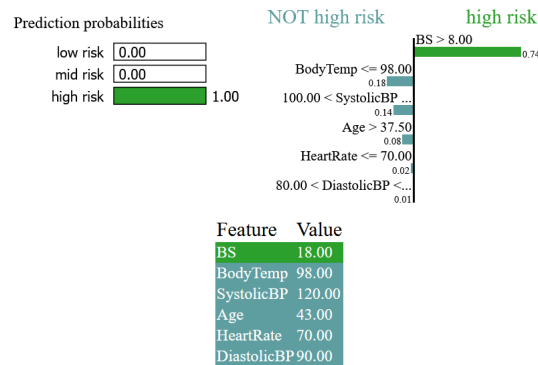


Figure 5.19: Explanation of high risk level by LIME (Case 03)

Lastly, the LIME explanation also provides the feature values for this particular data point, which can help in interpreting the contributions of the features towards the high risk prediction for this specific instance.

At first, Figure 5.20 indicates that the ground truth label for the data point is High Risk, and the prediction of the model is also High Risk, which is correct. DiastolicBP=90 being shown in red in the plot indicates that it has a positive contribution towards the prediction of high risk, pushing the prediction towards a higher risk level. The rest of the features, including Age=43, SystolicBP=120, BS=18, BodyTemp=98, and HeartRate=70, are shown in blue, indicating that they have negative contributions towards the prediction of high risk.

The explanations from LIME and XOPSS are consistent in terms of associating higher values of BS with a higher predicted risk level. SHAP explanation is consistent with LIME and XOPSS in terms of associating higher values of DiastolicBP, Age, SystolicBP, and BS with a lower predicted risk level, but it differs in not assigning any significant contribution to HeartRate. The similarity between the explanations lies in the direction of contribution, where higher values of certain features are associated with a higher predicted risk level in LIME and XOPSS, while lower values of certain features are associated with a lower predicted risk level in SHAP.

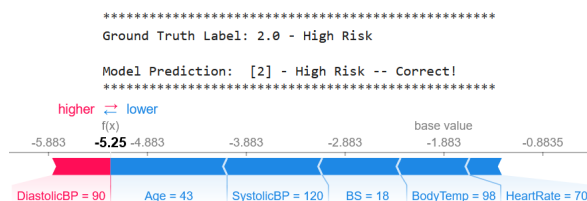


Figure 5.20: Explanation of high risk level by SHAP (Case 03)

2. Mid Risk Explanation

Case 04

```
----- TOPSIS Explanation -----
Data point index: 706
----- Data Point Features -----
Age          23.00
SystolicBP   130.00
DiastolicBP  70.00
BS           7.01
BodyTemp     98.00
HeartRate    78.00
Name: 706, dtype: float64
Actual risk level: 1.0
Predicted Risk Level: 1.0

Explanation:
- The TOPSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  * SystolicBP
  * HeartRate
- The features that contribute negatively to the prediction are:
  * Age
  * DiastolicBP
  * BS
  * BodyTemp
```

Figure 5.21: Explanation of mid risk level by XOPSIS (Case 04)

In Figure 5.21, the first section provides the feature values of the given data point. It includes values for Age, DiastolicBP, SystolicBP, HeartRate, BodyTemp, and BS. The next section shows the actual risk level associated with the given data point, which is given as “1.0”. This indicates the true risk level of the data point in the dataset. The predicted risk level for the given data point, which is also given as “1.0”. This indicates the risk level predicted by the model for the data point.

The next section provides the explanation of the TOPSIS Score, which is calculated based on the similarity of feature values between the given data point and other similar instances in the dataset. This section lists the features that have a positive contribution towards the prediction of the risk level. In this case, it includes SystolicBP and HeartRate. On the other hand, the features that have a negative contribution towards the prediction of the risk level. In this case, it includes Age, DiastolicBP, BS, and BodyTemp. Figure 5.22 shows the normalized feature contribution of this specific data point by XOPSIS method.

Based on the LIME output provided in Figure 5.23, it appears that the predicted risk level for the given data point is associated with a high probability of being classified as “Mid Risk” (1.0) and low probabilities of being classified as “Low Risk” (0.0) or “High Risk” (0.0). The plot in the output shows the feature values of the data point, with features associated with “Mid Risk” shown on the right side of the plot, and features associated with “Not Mid Risk” (likely indicating “Low Risk” or “High Risk”) shown on the left side of the plot.

On the right side of the plot (associated with “Mid Risk”), the features that are highlighted and their corresponding values are SystolicBP > 120 (with a contribution score of 0.05) and HeartRate > 76 (with a contribution score of 0.00). On the left side of the plot (associated with “Not Mid Risk”), the features that are highlighted and their corresponding values are BS between 6.90 and 7.50 (with a contribution score of 0.16), BodyTemp <= 98.0 (with a contribution score of 0.12), Age between 19 and 27 (with a contribution score of 0.05), and DiastolicBP > 65 (with a contribution score of 0.00).

These feature values and their associated contribution scores provide insights into how different features contribute to the predicted risk level of “Mid Risk” versus

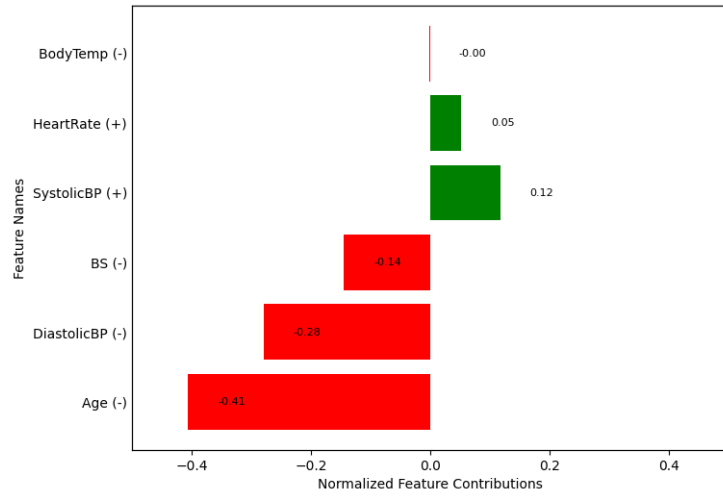


Figure 5.22: Normalized Feature Contributions of mid risk level by XOPSIS (Case 04)

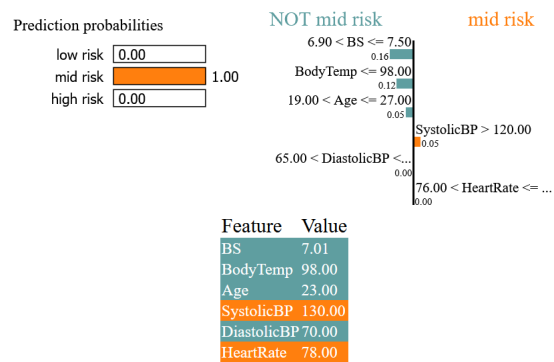


Figure 5.23: Explanation of mid risk level by LIME (Case 04)

“Not Mid Risk” for the given data point, as interpreted by the LIME model. It is important to note that the contribution scores represent the relative importance of each feature in determining the predicted risk level, as approximated by the LIME model.

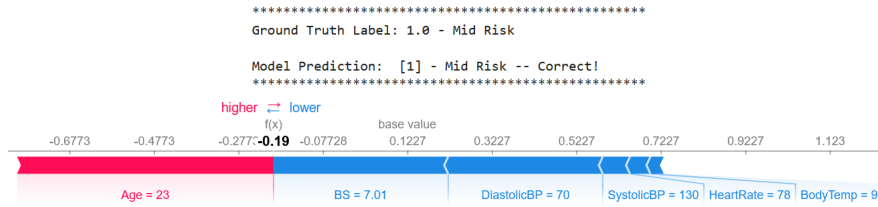


Figure 5.24: Explanation of mid risk level by SHAP (Case 04)

Based on the information provided in the SHAP force plot in Figure 5.24, the feature “Age” appears to have a positive contribution towards the prediction of “Mid Risk”, as higher values of “Age” are associated with a higher probability of “Mid Risk” prediction. On the other hand, the features BS, DiastolicBP, SystolicBP, HeartRate, and BodyTemperature appear to have a negative contribution towards the prediction of “Mid Risk”, as higher values of these features are associated with a lower probability of “Mid Risk” prediction.

In summary, the similarity between the LIME and XOPSIS explanations is that they both identify “SystolicBP” and “HeartRate” as features that positively contribute to the prediction of “Mid Risk” for the given data point. Additionally, both LIME and XOPSIS indicate that features such as “Age,” “DiastolicBP,” “BS,” and “BodyTemp” have a negative contribution to the prediction of “Mid Risk.” This consistent observation across LIME and XOPSIS suggests that these features may have a similar impact on the predicted risk level for the given data point. It is important to note that while there is similarity between LIME and XOPSIS explanations, there may be inconsistencies with SHAP explanation regarding the contribution of “Age.” It is crucial to thoroughly review and validate the explanations to ensure accuracy and consistency when interpreting and comparing results from different explanation methods.

Case 05

The XOPSIS method starts by showing in Figure 5.25, the features of the data point that are being considered for the prediction. These features are likely to be the input variables or attributes that are used in the prediction model. The XOPSIS method then displays the actual risk level and predicted risk level for the data point is “Mid Risk”. These are the risk levels associated with the specific data point being analyzed, as determined by the prediction model.

The explanation suggests that the method assesses how close the feature values of the given data point are to the feature values of other instances in the dataset, in order to determine the similarity. This explanation further mentions that certain features contribute positively to the prediction. These features are identified as BodyTemp and HeartRate. The explanation also indicates that certain features contribute negatively to the prediction. These features are identified as Age, BS, SystolicBP and DiastolicBP.

```

----- TOPSIS Explanation -----
Data point index: 811
----- Data Point Features -----
Age          18.0
SystolicBP   120.0
DiastolicBP   80.0
BS           7.9
BodyTemp     102.0
HeartRate    76.0
Name: 811, dtype: float64
Actual risk level: 1.0
Predicted Risk Level: 1.0

Explanation:
- The TOPSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  • BodyTemp
  • HeartRate
- The features that contribute negatively to the prediction are:
  • Age
  • BS
  • SystolicBP
  • DiastolicBP

```

Figure 5.25: Explanation of mid risk level by XOPSIS (Case 05)

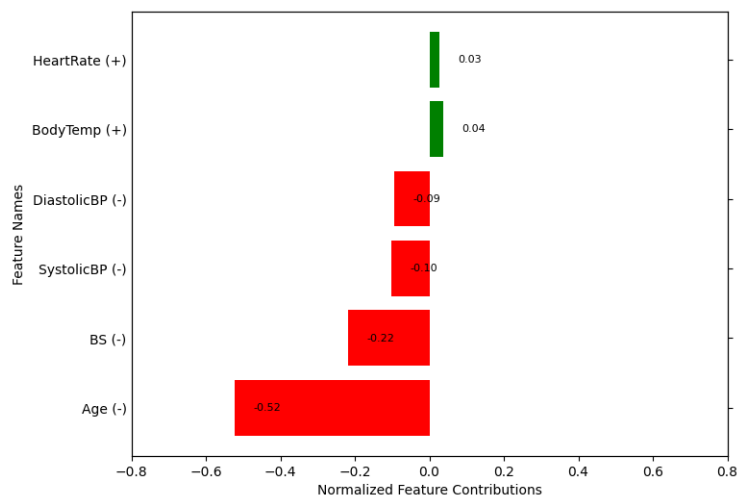


Figure 5.26: Normalized Feature Contributions of mid risk level by XOPSIS (Case 05)

In summary, the XOP SIS method is being used to calculate a score for a specific data point based on the similarity of its feature values with other instances. The features that contribute positively or negatively to the prediction are identified, indicating their impact on the predicted risk level for the given data point in Figure 5.26.

Based on the LIME output provided in Figure 5.27, the predicted risk level for the given data point is “Mid Risk” with a probability of 1.0, as indicated by the Prediction Probabilities section.

The plot in the output provides further insights into the features that contribute to the prediction of “Mid Risk.” On the right side of the plot, features such as “BodyTemp” greater than 98.0, “SystolicBP” greater than 100, and “HeartRate” greater than 70 are identified as positively contributing to the prediction of “Mid Risk,” with corresponding weights or importance scores of 0.10, 0.02, and 0.00 respectively. This means that higher values of these features are associated with a higher probability of the prediction being “Mid Risk”. On the left side of the plot, features such as “Age” less than or equal to 19.0, “BS” (Blood Sugar) greater than 7.50 and less than or equal to 8.0, and “DiastolicBP” greater than 65 are identified as negatively contributing to the prediction of “Mid Risk”, with corresponding weights or importance scores of 0.11, 0.08, and 0.00 respectively. The feature values of the data point are also provided in the output, which can be used to further understand the specific values of each feature for this particular prediction.

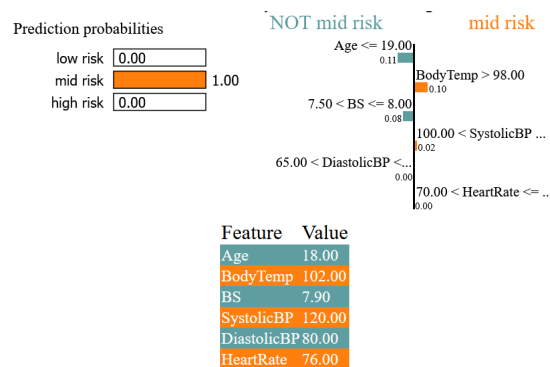


Figure 5.27: Explanation of mid risk level by LIME (Case 05)

The SHAP force plot in Figure 5.28, is a visualization that helps to explain the prediction of a machine learning model for a specific data point. In this case, the prediction is for the “Mid Risk” level, as indicated by the Ground Truth Label and Model Prediction, both showing a value of 1.0 for “Mid Risk”. The SHAP force plot consists of two parts: the left side, shown in red color, and the right side, shown in blue color. The labels “higher – >” on the left side and “< – lower” on the right side indicate the directionality of the feature values in those regions.

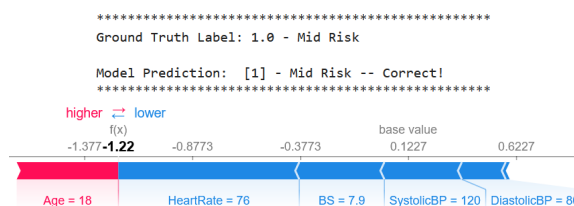


Figure 5.28: Explanation of mid risk level by SHAP (Case 05)

On the left side (red color), the label "higher – >" indicates that values of the feature in that region have a positive contribution towards the prediction of "Mid Risk" for the specific data point being explained in the plot. The red color represents features that push the model score higher, meaning they increase the output raw value, and thus have a positive contribution towards the prediction of "Mid Risk". On the right side (blue color), the label "< – lower" indicates that values of the feature in that region have a negative contribution towards the prediction of "Mid Risk". The blue color represents features that push the model score lower, meaning they decrease the model output value, and thus have a negative contribution towards the prediction of "Mid Risk" for this particular data point.

Additionally, the numerical values shown in the plot represent the feature values for the specific data point being explained. For example, Age=18 is shown in red color with a value of -1.377, indicating that a higher value of Age would have a positive contribution towards the prediction of "Mid Risk". The rest of the features, HeartRate, BS, SystolicBP, and DiastolicBP, are shown in blue color with corresponding values, indicating their negative contribution towards the prediction of "Mid Risk" for this data point.

The base value of 0.1227, shown in bold text, represents the expected model output value for the "Mid Risk" prediction. The value of $f(x)$ -1.22, positioned in the middle of the red and blue colored features, indicates the overall contribution of all the features towards the prediction of "Mid Risk" for the specific data point being explained in the plot.

In summary, the findings from the XOPIS and LIME plots are aligned, as they both identify almost similar positively and negatively contributing features for predicting "Mid Risk" in the given data point. However, the SHAP force plot provides more comprehensive and detailed information with numerical values and color-coded directionality, which adds to the overall consistency and interpretability of the explanations across different methods, albeit not fully aligning with XOPIS and LIME.

Case 06

The features of the given data point are displayed in Figure 5.29. The actual and predicted risk level of the data point is shown, with a value of 1.0. This indicates that the model has predicted a "Mid Risk" level for the given data point. The features that positively contribute to the prediction are HeartRate and BodyTemp, while the features that negatively contribute to the prediction are SystolicBP, DiastolicBP, Age, and BS.

A plot is shown in Figure 5.30, with normalized feature contributions. Positive contributions are displayed in green color, while negative contributions are displayed in red color. The normalized contributions for each feature, displayed in parentheses next to the feature name, are as follows: BodyTemp (0.02), BS (-0.04), HeartRate (0.07), DiastolicBP (-0.23), Age (-0.24), and SystolicBP (-0.41). These values indicate the strength and directionality of the contributions of each feature towards the predicted "Mid Risk" level for the given data point. Positive contributions suggest that higher values of those features positively contribute to the "Mid Risk" prediction, while negative contributions suggest that lower values of those features negatively contribute to the "Mid Risk" prediction.

In Figure 5.31, the initial section displays the predicted probabilities for each risk category. In this case, the predicted probability for "Mid Risk" is 1.0, while the

```

----- TOPSIS Explanation -----
Data point index: 66
----- Data Point Features -----
Age          17.0
SystolicBP   85.0
DiastolicBP  60.0
BS           9.0
BodyTemp     102.0
HeartRate    86.0
Name: 66, dtype: float64
Actual risk level: 1.0
Predicted Risk Level: 1.0

Explanation:
- The TOPSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  • HeartRate
  • BodyTemp
- The features that contribute negatively to the prediction are:
  • SystolicBP
  • Age
  • DiastolicBP
  • BS

```

Figure 5.29: Explanation of mid risk level by XOPSIS (Case 06)

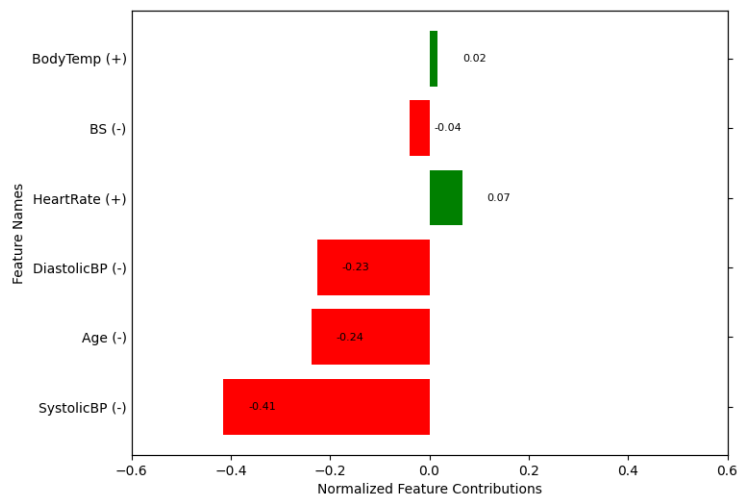


Figure 5.30: Normalized Feature Contributions of mid risk level by XOPSIS (Case 06)

predicted probabilities for “Low Risk” and “High Risk” are both 0.0. This suggests that the model’s prediction for this data point is a “Mid Risk” classification with high confidence.

The plot is divided into two sections, with “Mid Risk” on the right side and “Not Mid Risk” on the left side. Each section represents the features that are contributing to the respective classification. On the right side of the plot, the features that are positively contributing to the prediction of “Mid Risk” are displayed. These features include “BodyTemp” greater than 98.0 with a contribution weight of 0.11, and “HeartRate” greater than 80.0 with a contribution weight of 0.01. These higher values of “BodyTemp” and “HeartRate” are influencing the model to predict “Mid Risk” for this data point. On the left side of the plot, the features that are negatively contributing to the prediction of “Mid Risk” are displayed. These features include “BS greater than 8.0 with a contribution weight of 0.15, “Age” less than or equal to 19.0 with a contribution weight of 0.11, “SystolicBP” less than or equal to 100 with a contribution weight of 0.05, and “DiastolicBP” less than or equal to 65 with a contribution weight of 0.01. These lower values of “BS”, “Age”, “SystolicBP”, and “DiastolicBP” are influencing the model to predict “Not Mid Risk” for this data point.

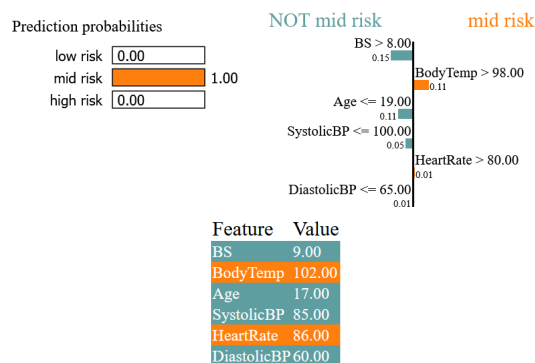


Figure 5.31: Explanation of mid risk level by LIME (Case 06)

Finally, the LIME output also provides the feature values for the specific data point being explained. These feature values are likely the values of the input features of the data point for which the explanation is being generated.

Figure 5.32 displays the SHAP force plot which provides an explanation of the prediction for the data point being analyzed, which is labeled as “Mid Risk” with a correct model prediction. Features such as “BodyTemp=102”, “DiastolicBP=60”, and “HeartRate=86” are shown with red arrows pointing to the right, indicating that higher values of these features have a positive contribution towards the prediction of “Mid Risk”. The corresponding SHAP values (-1.377, -0.8773, -0.3773) quantify the magnitude of the negative contribution for each feature. Features such as “SystolicBP=85”, “Age=17”, and “BS=9” are shown with blue arrows pointing to the left, indicating that lower values of these features have a negative contribution towards the prediction of “Mid Risk”. The corresponding SHAP values (0.1227, 0.6227, 1.123) quantify the magnitude of the negative contribution for each feature. The base value of 0.1227, written in bold letters, represents the expected prediction value without considering any specific feature contributions. The value of -0.12, written in the middle of the red and blue colored features, represents the final prediction value after considering the contributions of all the features together.

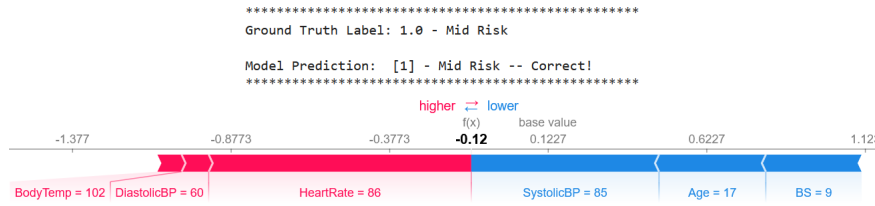


Figure 5.32: Explanation of mid risk level by SHAP (Case 06)

Overall, the SHAP force plot helps in understanding the contribution of each feature towards the prediction of “Mid Risk” for the specific data point, with red indicating positive contributions and blue indicating negative contributions. The magnitude of the contributions is quantified by the SHAP values, and the plot provides a visual representation of the feature contributions, aiding in the interpretability of the prediction of the model.

Finally, all three explanations indicate that the predicted risk level for the given data point is “Mid Risk”. They identify similar positively contributing features, such as BodyTemp and HeartRate, as significant for predicting “Mid Risk” in the given data point. They also highlight similar negatively contributing features, such as SystolicBP, DiastolicBP, Age, and BS, as influential in predicting “Not Mid Risk” for the given data point. One key difference between the explanations provided by LIME, SHAP, and XOPSIS is that SHAP identifies DiastolicBP as a positive contributing feature for predicting “Mid Risk” in the given data point, whereas LIME and TOPSIS do not.

LIME provides probability-based explanations and importance scores for positively and negatively contributing features. SHAP uses a force plot to represent the magnitude and direction of contributions for each feature, with red indicating positive and blue indicating negative contributions. XOPSIS provides normalized contributions for each feature, with positive contributions shown in green and negative contributions shown in red.

The explanations provided by LIME, SHAP, and XOPSIS are aligned in terms of identifying similar positively and negatively contributing features for predicting “Mid Risk” in the given data point. However, they differ in terms of the representation and visualization of the contributions. LIME provides probability-based explanations, SHAP uses a force plot, and XOPSIS provides normalized contributions. The choice of explanation method may depend on the specific use case, context, and user preference.

3. Low Risk Explanation

Case 07

In the given XOPSIS explanation provided in Figure 5.33, the data point features are shown initially, followed by the actual risk level and predicted risk level, which are both 0.0. Then, the explanation mentions that the TOPSIS score is calculated based on the similarity of feature values between the given data point and similar instances.

The explanation further specifies that there are features that contribute positively to the prediction, which are SystolicBP, Age, BS, and BodyTemp. These positive

```

----- TOPSIS Explanation -----
Data point index: 30
----- Data Point Features -----
Age          20.0
SystolicBP   100.0
DiastolicBP   90.0
BS           7.1
BodyTemp     98.0
HeartRate    88.0
Name: 30, dtype: float64
Actual risk level: 0.0
Predicted Risk Level: 0.0

Explanation:
- The TOPSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  • SystolicBP
  • Age
  • BS
  • BodyTemp
- The features that contribute negatively to the prediction are:
  • DiastolicBP
  • HeartRate

```

Figure 5.33: Explanation of low risk level by XOPSIS (Case 07)

contributions are shown in green color in the plot in Figure 5.34. Additionally, there are features that contribute negatively to the prediction, which are DiastolicBP and HeartRate. These negative contributions are shown in red color in the plot. It is important to note that in the context of the given XOPSIS explanation, green color indicates positive contributions, while red color indicates negative contributions. This color scheme is used to visually represent the impact of each feature on the prediction of the risk level.

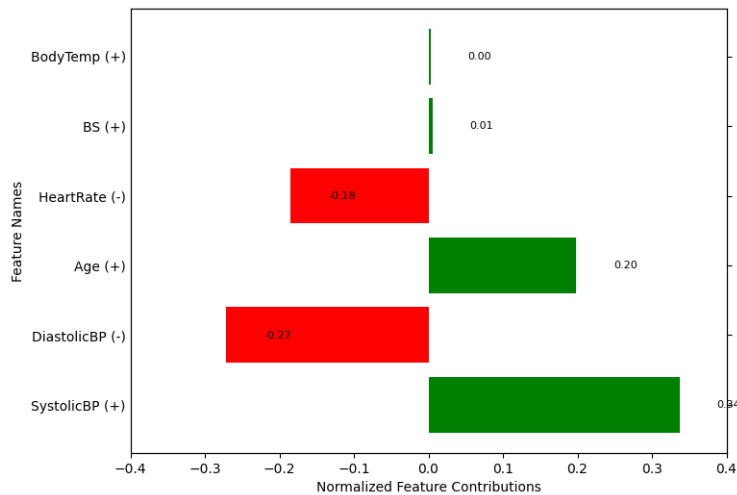


Figure 5.34: Normalized Feature Contributions of low risk level by XOPSIS (Case 07)

Figure 5.35 shows the predicted probabilities for each risk level as (low risk 1.0, mid risk 0.0, high risk 0.0), indicating that the model has predicted the data point to be in the low risk category with a probability of 1.0. The plot is divided into two sections, the right side showing the features that contribute to the prediction of low risk, and the left side showing the features that contribute to the prediction of not being in the low risk category.

The features that contribute positively to the prediction of low risk are: “BS” (Blood Sugar) with a value range of 6.90 to 7.50, with a contribution weight of 0.39; “BodyTemp” (Body Temperature) with a value less than or equal to 98.0, with a contribution weight of 0.25; “SystolicBP” (Systolic Blood Pressure) with a

value less than or equal to 100.0, with a contribution weight of 0.16; and “Age” with a value range of 19 to 27, with a contribution weight of 0.09. On the other hand, the features that contribute negatively to the prediction of being in the low risk category are: “HeartRate” (Heart Rate) with a value greater than 80, with a contribution weight of 0.06; and “DiastolicBP” (Diastolic Blood Pressure) with a value greater than 80, with a contribution weight of 0.01. The actual feature values of the data point are explicitly mentioned in this figure, and they are used in the calculation of the contribution weights for each feature in determining the predicted risk level.

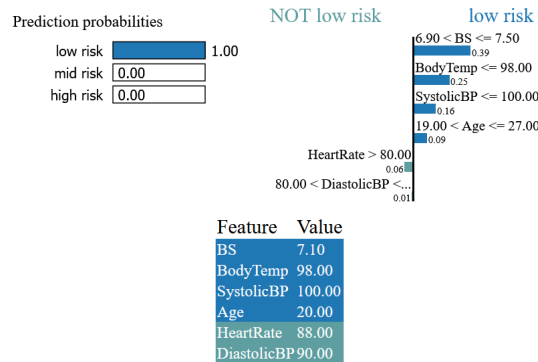


Figure 5.35: Explanation of low risk level by LIME (Case 07)

In summary, the LIME output indicates that the model has predicted the data point to be in the low risk category based on the contributions of features such as BS, BodyTemp, SystolicBP, and Age, while features such as HeartRate and DiastolicBP have lesser contributions to the prediction of not being in the low risk category. The SHAP force plot shown in Figure 5.36, appears to be for a classification model with a ground truth label of 0.0, indicating a low risk prediction, and a model prediction of 0, which is also classified as low risk and is correct.

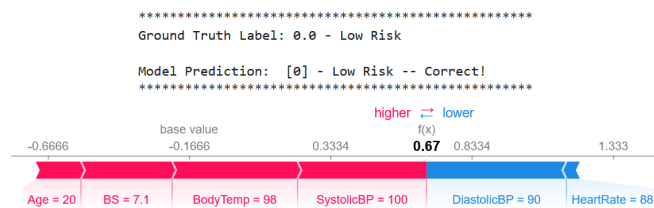


Figure 5.36: Explanation of low risk level by SHAP (Case 07)

The plot uses red and blue colors to represent the direction and magnitude of feature contributions. Features with positive contributions are shown in red with the label ‘higher— >’, while features with negative contributions are shown in blue with the label ‘<—lower’. The features and their corresponding contributions shown on the plot are as follows: Age = 20, BS=7.1, BodyTemp=98, SystolicBP=100 in red color, indicating a positive contribution towards the model prediction and an increase the predicted risk level. DiastolicBP = 90, HeartRate = 88 are shown in blue color, indicating a negative contribution towards the model prediction and a reduction in the predicted risk level. The base value is -0.1666, which serves as a reference point for the contributions. The value “f(x) 0.67” in bold letters and in the middle of

the red and blue color features represents the final predicted risk level based on the accumulated contributions of the features.

Overall, the SHAP force plot provides a visual representation of how each feature contributes towards the prediction of the model of low risk, with red indicating positive contributions and blue indicating negative contributions. The magnitude of the contributions is shown with values, and the base value provides a reference point.

In conclusion, all three explanations provided by LIME, SHAP, and XOPSIS for the same data point share similarities in terms of identifying the features that contribute positively (BS, BodyTemp, SystolicBP, and Age) and negatively (HeartRate and DiastolicBP) to the prediction of low risk, using visual representations to convey the contributions, providing information about actual and predicted risk levels, explaining the calculation methodology or scoring system, and mentioning specific feature values used in the calculation. These explanations provide valuable insights into the factors that influence the prediction of risk level for the given data point, and can aid in interpreting the predictions of the model and building trust in the decision-making process of the model.

Case 08

The first section shows the values of the features for the given data point in Figure 5.37. It provides an overview of the specific feature values that are being considered for the prediction. This section shows the actual risk level and the predicted risk level for the data point is 0.0. It helps to compare the true risk level with the predicted risk level generated by the XOPSIS model.

```

----- TOPSIS Explanation -----
Data point index: 305
----- Data Point Features -----
Age          22.0
SystolicBP   100.0
DiastolicBP   65.0
BS           6.9
BodyTemp     98.0
HeartRate    80.0
Name: 305, dtype: float64
Actual risk level: 0.0
Predicted Risk Level: 0.0

Explanation:
- The TOPSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  • SystolicBP
  • DiastolicBP
  • Age
  • BS
  • BodyTemp
- The features that contribute negatively to the prediction are:
  • HeartRate

```

Figure 5.37: Explanation of low risk level by XOPSIS (Case 08)

The next section provides an explanation of how the Topsis Score is calculated. The Topsis Score is based on the similarity of feature values between the given data point and similar instances in the dataset. It indicates which features contribute positively and negatively to the prediction. This section lists the features that contribute positively to the prediction. These features have a positive impact on the predicted risk level. In this case, the features that contribute positively are SystolicBP, DiastolicBP, Age, BS, and BodyTemp. These features for the given data point, increase the predicted risk level. This section also lists the features that contribute negatively to the prediction. These features have a negative impact

on the predicted risk level. In this case, the feature that contributes negatively is HeartRate. This feature for the given data point, decreases the predicted risk level.

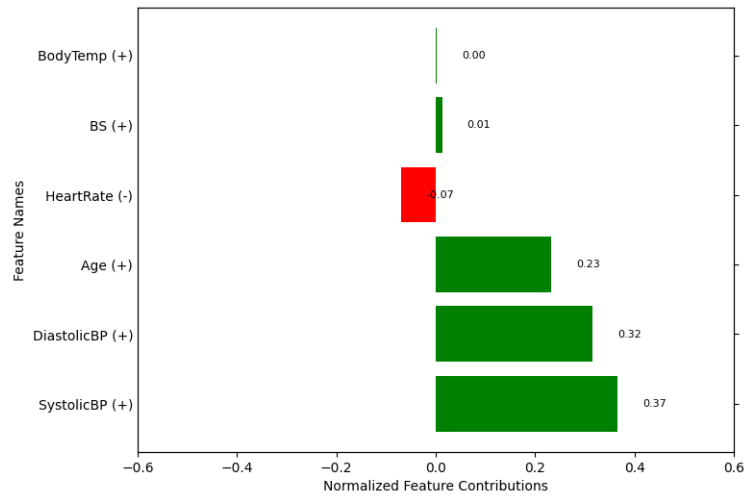


Figure 5.38: Normalized Feature Contributions of low risk level by XOP SIS (Case 08)

Figure 5.38 presents a plot that shows the normalized feature contributions visually. Positive contributions are shown in green color, while negative contributions are shown in red color. This plot provides a visual representation of how each feature contributes to this particular prediction, either positively or negatively. Overall, the XOP SIS explanation provides insights into how the features of the given data point contribute to the predicted risk level. It highlights which features have a positive or negative impact on the prediction and provides a visual representation of the normalized feature contributions.

The predicted probabilities in Figure 5.39, for each risk level are shown as (low risk 1.0, mid risk 0.0, high risk 0.0). This indicates that the model has predicted the data point to be in the low risk category with a probability of 1.0, and there are no predicted probabilities for mid risk or high risk categories. The plot is divided into two sections, with the right side showing the features that contribute to the prediction of low risk, and the left side showing the features that contribute to the prediction of not being in the low risk category.

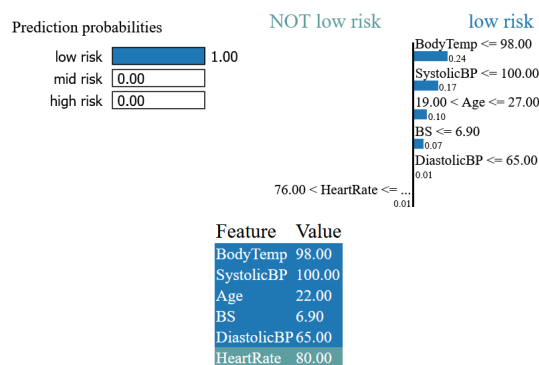


Figure 5.39: Explanation of low risk level by LIME (Case 08)

BodyTemp<=98.0 indicates that if the body temperature is less than or equal to

98.0, it contributes positively to the prediction of low risk with a contribution weight of 0.24. SystolicBP \leq 100 implies that if the systolic blood pressure is less than or equal to 100, it contributes positively to the prediction of low risk with a contribution weight of 0.17. 19 $<$ Age \leq 27 demonstrates that if the age is between 19 and 27, it contributes positively to the prediction of low risk with a contribution weight of 0.10. BS \leq 6.90 represents that if the blood sugar is less than or equal to 6.90, it contributes positively to the prediction of low risk with a contribution weight of 0.07. DiastolicBP \leq 65 reveals that if the diastolic blood pressure is less than or equal to 65, it contributes positively to the prediction of low risk with a contribution weight of 0.01. On the other hand, 76 $<$ HeartRate suggests that if the heart rate is greater than 76, it contributes negatively to the prediction of low risk with a contribution weight of 0.01. The actual feature values of the data point are explicitly mentioned in the provided LIME output.

In summary, the LIME output indicates that the model has predicted the data point to be in the low risk category based on the contributions of features such as BodyTemp, SystolicBP, Age, BS, and DiastolicBP, while HeartRate has a lesser contribution to the prediction of not being in the low risk category. The contribution weights of each feature indicate the magnitude of their impact on the predicted risk level.

The SHAP force plot shown in Figure 5.40, starts by showing the ground truth label, which is 0.0 representing “Low risk”, and the correct prediction of the model is “Low risk” with a prediction value of [0].



Figure 5.40: Explanation of low risk level by SHAP (Case 08)

The plot then displays a range of features, such as BS (6.9), Age (22), SystolicBP (100), DiastolicBP (65), BodyTemp (98), and HeartRate (80), in red color. The red color indicates that these features have higher values compared to the base value. The arrows on the plot represent the direction of the contribution of each feature towards the prediction. If an arrow is pointing to the right (labeled as “higher – >”) and is in red color, it indicates a positive contribution towards the prediction of “Low risk”. This means that higher values of these features are positively correlated with the prediction of “Low risk” for this data point.

The base value, which is -0.1666, serves as the reference point for the prediction. The value of $f(x)$, which is 2.84 and shown in bold text and in the middle of the red and blue color features, represents the overall contribution of all features towards the prediction. Since the value is positive, it suggests that the higher values of the features collectively contribute positively to the prediction of “Low risk” for this particular data point.

In conclusion, all three explanations mention the predicted risk level for the data point, with LIME and XOPSSIS showing the prediction probabilities, and SHAP showing the ground truth label and model prediction. They also mention the features

that contribute positively or negatively to the predicted risk level. They mention similar features, such as SystolicBP, DiastolicBP, Age, BS, and BodyTemp as positive contributors, while HeartRate is mentioned as a negative contributor in LIME and Topsis explanations, but as a positive contributor in the SHAP explanation. Based on the similarities, it can be concluded that LIME, SHAP, and XOPSIS explanations provide insights into the features that contribute to the predicted risk level for the given data point. They mention similar features as positive or negative contributors, although the representation and format of the explanations may vary. While LIME and XOPSIS consistently identify HeartRate as a negative contributor, SHAP represents it as a positive contributor. These explanations can help in interpreting and understanding the prediction of the risk level for the given data point in a transparent and interpretable manner, allowing for better decision-making and model evaluation.

Case 09

```

----- TOPSIS Explanation -----
Data point index: 39
----- Data Point Features -----
Age           35.0
SystolicBP   100.0
DiastolicBP   70.0
BS            6.1
BodyTemp     98.0
HeartRate     66.0
Name: 39, dtype: float64
Actual risk level: 0.0
Predicted Risk Level: 0.0

Explanation:
- The TOPSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  • SystolicBP
  • DiastolicBP
  • HeartRate
  • BS
  • BodyTemp
- The features that contribute negatively to the prediction are:
  • Age

```

Figure 5.41: Explanation of low risk level by XOPSIS (Case 09)

Based on the provided XOPSIS explanation from Figure 5.41, it appears to provide an overview of the actual and predicted risk level for the given data point, followed by an explanation of how the Topsis Score is calculated based on similarity of feature values with similar instances. The explanation also highlights the features that contribute positively and negatively to the prediction. The positively contributing features are identified as SystolicBP, DiastolicBP, HeartRate, BS, and BodyTemp, indicating that higher values of these features are associated with a higher predicted risk level. On the other hand, the negatively contributing feature is identified as Age, suggesting that a higher value of Age is associated with a lower predicted risk level.

The explanation is further supported by a plot in Figure 5.42 that visualizes the normalized feature contributions for this data point. Positive feature contributions are shown in green, while negative feature contributions are shown in red. This plot provides a graphical representation of how each feature contributes to the predicted risk level for the given data point, based on their normalized contribution values. Overall, the XOPSIS explanation provides insights into the features that are contributing positively or negatively to the prediction of the risk level for the given

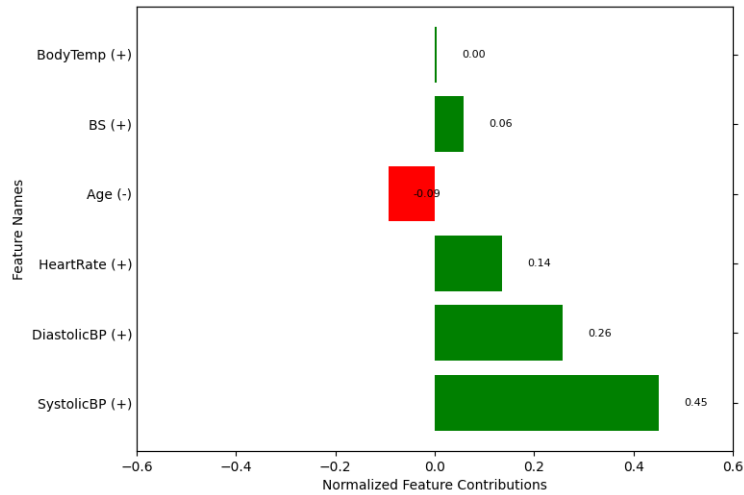


Figure 5.42: Normalized Feature Contributions of low risk level by XOPSIS (Case 09)

data point, and how these contributions are normalized and visualized in a plot for better understanding.

Figure 5.43 shows the predicted probabilities for each risk level. In this case, the model predicts a low risk with a probability of 1.0, and zero probabilities for mid risk and high risk. This indicates that the model is highly confident in predicting a low risk for the given data point. The plot is divided into two sides, with the “low risk” category on the right side and the “not low risk” category on the left side.

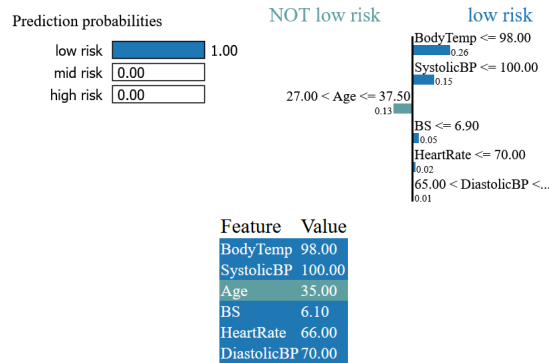


Figure 5.43: Explanation of low risk level by LIME (Case 09)

The features that contribute positively to the prediction of low risk are listed on the right side of the plot. These features including BodyTemp ≤ 98.0 with a contribution weight of 0.26, SystolicBP ≤ 100 with a contribution weight of 0.15, BS ≤ 6.90 with a contribution weight of 0.05, HeartRate ≤ 70 with a contribution weight of 0.02, DiastolicBP > 65 with a contribution weight of 0.01. The contribution weights represent the magnitude of the contribution of each feature towards the predicted low risk. Higher positive weights indicate stronger positive contribution to the prediction. The feature $27 < \text{Age} \leq 37.50$ is listed on the left side of the plot with a contribution weight of 0.13. This indicates that this feature is contributing negatively towards the prediction of low risk. The output also displays the actual feature values of the data point being explained. This allows for a complete

understanding of the feature values used by the model in making the prediction. In summary, based on this LIME output, the model predicts a low risk for the given data point mainly based on features such as BodyTemp, SystolicBP, BS, HeartRate, and DiastolicBP, while Age is contributing negatively towards the prediction. It provides insights into how different features influence the prediction of the model and helps in interpreting the prediction of low risk for the given data point in an interpretable manner.

Based on the information provided in Figure 5.44, the SHAP force plot begins by showing the ground truth label and model prediction, which in this case are both 0.0 - Low risk, and the prediction is correctly labeled.

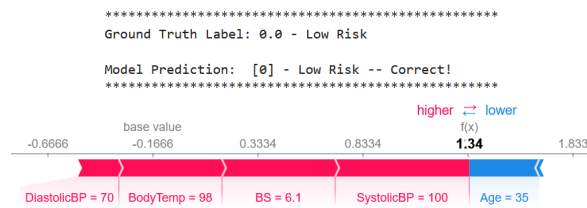


Figure 5.44: Explanation of low risk level by SHAP (Case 09)

The plot then displays a series of arrows indicating the direction of the contribution for different features. The arrow pointing to the right (in red color) represents features with higher values that contribute positively towards the prediction, while the arrow pointing to the left (in blue color) represents features with lower values that contribute negatively towards the prediction. The features DiastolicBP=70, BodyTemp=98, BS=6.1, and SystolicBP=100 are labeled in red color on the side of the “higher – >” arrow, indicating that higher values of these features positively contribute towards the prediction of low risk. On the other hand, the features Age=35 and HeartRate=66 are labeled in blue color on the side of the “< – lower” arrow, indicating that lower values of these features negatively contribute towards the prediction of low risk.

The base value of -0.1666 is the expected value of the prediction. The value written in bold letter and in the middle of the red and blue color features, which is $f(x)$ 1.34, represents the total estimated contribution of all the features towards the prediction of low risk for the given data point. Positive contributions (red) from features with higher values and negative contributions (blue) from features with lower values are summed up to arrive at the total estimated contribution of 1.34, which is added to the base value to obtain the final predicted value for the data point.

In summary, the SHAP force plot provides a visual representation of the contributions of different features towards the prediction of low risk for the given data point, with red arrows indicating positive contributions and blue arrows indicating negative contributions. The estimated contributions of each feature are shown in bold letter, and the base value is provided as a reference point for interpreting the overall prediction.

In summary, the predicted risk level in all three explanations is consistent for the given data point, indicating that the prediction of the model aligns with the actual risk level for the data point. All three explanations highlight the positive or negative contributions of features towards the predicted risk level. For example, in LIME and XOPIS, features like SystolicBP, DiastolicBP, BS, HeartRate and BodyTemp are

mentioned as positive contributors, while Age is mentioned as a negative contributor. Similarly, in SHAP, some features like DiastolicBP, BS, BodyTemp, and SystolicBP are shown as positively contributing, while Age, HeartRate is shown as negatively contributing. Based on these similarities, we can conclude that LIME, SHAP, and XOPIS explanations provide consistent insights into the features that contribute to the predicted risk level for the given data point. Although there may be some differences in representation and format, these explanations can collectively help in interpreting and understanding the prediction of risk level for the given data point in a transparent and interpretable manner.

5.3 Result Analysis on Iris Dataset

5.3.1 Accuracy

The accuracy of the trained models on the Iris dataset was evaluated using two different techniques: the train-test split and k-fold cross-validation with 10 folds. The accuracy scores for three algorithms, namely Gradient Boosting, XGBoost, and Random Forest, are presented in Table 5.3.

Serial No	Algorithms	Accuracy	
		Train-Test Split	K-Fold Cross Validation
01	Gradient Boosting	96.667%	95.333%
02	XGBoost	93.333%	95.299%
03	Random Forest	90.000%	94.667%

Table 5.3: Accuracy of 3 Algorithms

From the accuracy scores in Table 5.3, we observe that all three algorithms achieve high accuracy on the Iris dataset. The Gradient Boosting algorithm achieves the highest accuracy of 96.667% using the train-test split technique, while XGBoost and Random Forest achieve accuracies of 93.333% and 90.000%, respectively.

Comparing the train-test split accuracy with the 10-fold cross-validation accuracy, we notice a slight decrease in accuracy for all algorithms when using cross-validation. This difference may be attributed to the use of multiple subsets of data for training and testing in cross-validation, leading to a more generalized model but with slightly lower accuracy.

Additionally, Figure 5.45 illustrates the train-test split accuracy scores for the three algorithms. The x-axis represents the algorithm names (Gradient Boosting, XGBoost, and Random Forest), while the y-axis represents the accuracy. Each bar in the plot represents the accuracy score achieved by the corresponding algorithm.

The bar plot visually demonstrates the variation in accuracy across different algorithms. It is evident that Gradient Boosting achieves the highest accuracy score of 96.667%, followed by XGBoost with an accuracy score of 93.333%. Random Forest achieves an accuracy score of 90.000%.

This bar plot provides a clear comparison of the accuracy achieved by each algorithm, highlighting the superior performance of Gradient Boosting on the Iris dataset.

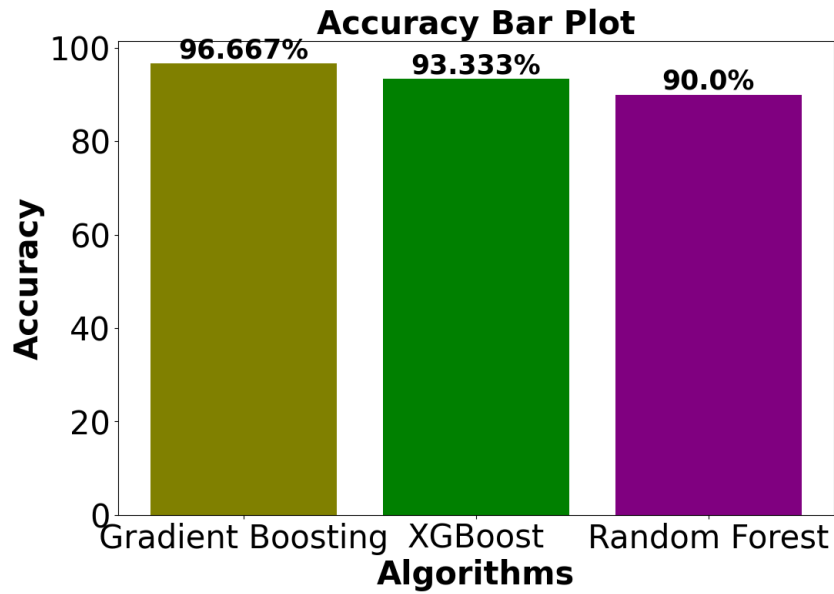


Figure 5.45: Train-Test Split Accuracy for Different Algorithms on Iris Dataset

These accuracy results demonstrate the effectiveness of the Gradient Boosting, XGBoost, and Random Forest algorithms in accurately classifying the Iris dataset.

5.3.2 Confusion Matrix

The Figure 5.46 presents the confusion matrix for the classification of Iris dataset using the trained model (Gradient Boosting). The true labels are displayed vertically, representing Iris-Setosa, Iris-Versicolor, and Iris-Virginica from top to bottom. The predicted labels are shown horizontally. The values in the cells indicate the count of samples classified into each class.

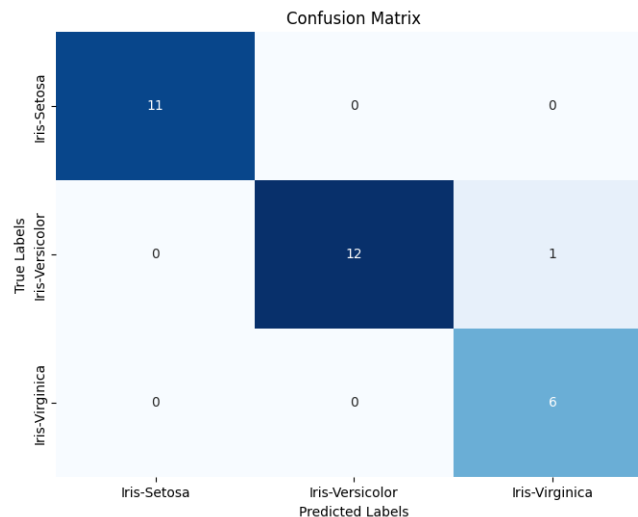


Figure 5.46: Confusion Matrix for Iris Dataset

The values in the confusion matrix indicate the counts of samples that belong to each class. Specifically, the values in the first row represent the samples that are

truly labeled as Iris-Setosa. In this case, there are 11 samples correctly classified as Iris-Setosa, while there are no misclassifications for this class.

Moving to the second row, the values represent the samples that are truly labeled as Iris-Versicolor. Out of the 13 samples of this class, 12 are correctly classified as Iris-Versicolor, while 1 sample is misclassified.

Finally, the third row represents the samples that are truly labeled as Iris-Virginica. All 6 samples of this class are correctly classified.

By analyzing the confusion matrix, we can gain insights into the performance of the classification model. In this case, the model shows high accuracy in predicting the Iris-Setosa and Iris-Virginica classes, but there is a single misclassification in the Iris-Versicolor class.

5.3.3 Sensitivity, Specificity, Precision and F1-Score

	Sensitivity	Specificity	Precision	F1-Score
Iris Setosa	1	1	1	1
Iris Versicolor	0.92	1	1	0.96
Iris Virginica	1	0.96	0.86	0.92

Table 5.4: Performance Evaluation Metrics on Iris Dataset

The Table 5.4 presents the performance evaluation metrics, including sensitivity, specificity, precision, and F1-score, for the classification of the Iris dataset using the trained model. The metrics are calculated for each class: Iris Setosa, Iris Versicolor, and Iris Virginica.

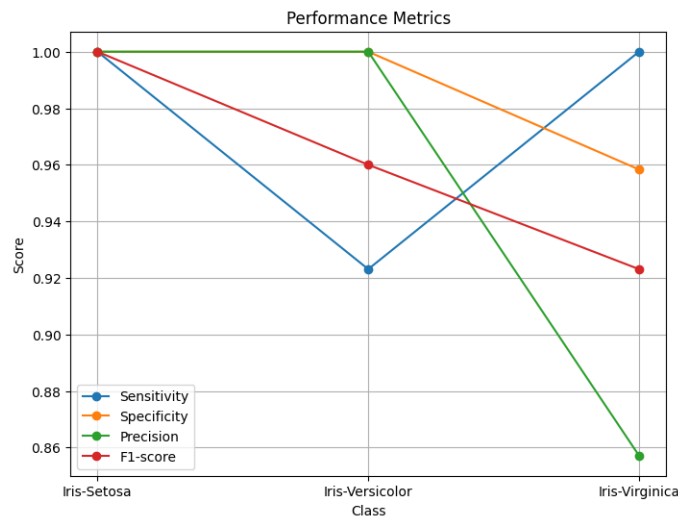


Figure 5.47: Performance Evaluation Metrics on Iris Dataset

For Iris Setosa, the model achieves perfect sensitivity, specificity, precision, and F1-score, with a score of 1 for each metric. This indicates that all samples of Iris Setosa are correctly classified.

For Iris Versicolor, the model demonstrates a sensitivity of 0.92, meaning that 92% of the Iris Versicolor samples are accurately identified. The model also achieves

perfect specificity, precision, and an F1-score of 0.96, indicating high performance in correctly classifying Iris Versicolor.

For Iris Virginica, the model achieves perfect sensitivity, correctly identifying all samples of Iris Virginica. It demonstrates a specificity of 0.96, indicating a high ability to distinguish Iris Virginica from other classes. However, the precision for Iris Virginica is 0.86, implying that there may be some misclassifications. The F1-score for Iris Virginica is 0.92, reflecting a balance between precision and recall.

Furthermore, Figure 5.47 displays a plot representing the performance evaluation metrics for each class. The x-axis denotes the three labels: Iris Setosa, Iris Versicolor, and Iris Virginica, while the y-axis represents the score. This plot allows for a visual comparison of the sensitivity, specificity, precision, and F1-score across the different classes of the Iris dataset.

5.3.4 ROC Curve

In this section, we delve into the ROC and AUC analyses conducted on the Iris dataset using the Gradient Boosting model. These analyses are pivotal in evaluating the model's performance across its three distinct classes: Iris-Setosa, Iris-Versicolor, and Iris-Virginica.

Figure 5.48 illustrates the ROC curves generated by the Gradient Boosting model for the Iris dataset. Similar to previous analyses, the ROC curves provide insights into the model's ability to discriminate among the different classes. The x-axis represents the False Positive Rate (FPR), while the y-axis represents the True Positive Rate (TPR). We present three ROC curves, each corresponding to one of the three Iris species: Iris-Setosa, Iris-Versicolor, and Iris-Virginica. These curves are distinctively colored, facilitating visual differentiation. As the curves ascend, they underscore the model's aptitude in classifying instances across diverse classes.

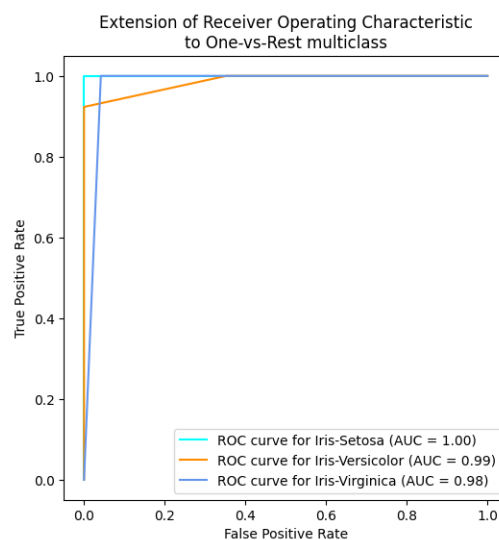


Figure 5.48: ROC Curve for Gradient Boosting Model on Iris Dataset

Overall, the ROC curve provides insights into the trade-off between the true positive rate and the false positive rate for the Gradient Boosting model on the Iris dataset. It demonstrates the model's ability to discriminate between different classes and highlights the varying performance for each class.

5.3.5 AUC

Figure 5.49 showcases the Area Under the Curve (AUC) scores obtained from the Gradient Boosting model for the Iris dataset. The AUC scores serve as quantitative indicators of the model's predictive prowess for each Iris species. On the x-axis, we label the three Iris species, while the y-axis represents the corresponding AUC scores. The calculated AUC scores are as follows: 1.0 for Iris-Setosa, 0.99 for Iris-Versicolor, and 0.98 for Iris-Virginica. These scores affirm the model's proficiency in accurately ranking instances across the different species. The high AUC scores are emblematic of the model's robust performance and its capacity to effectively distinguish between the Iris classes.

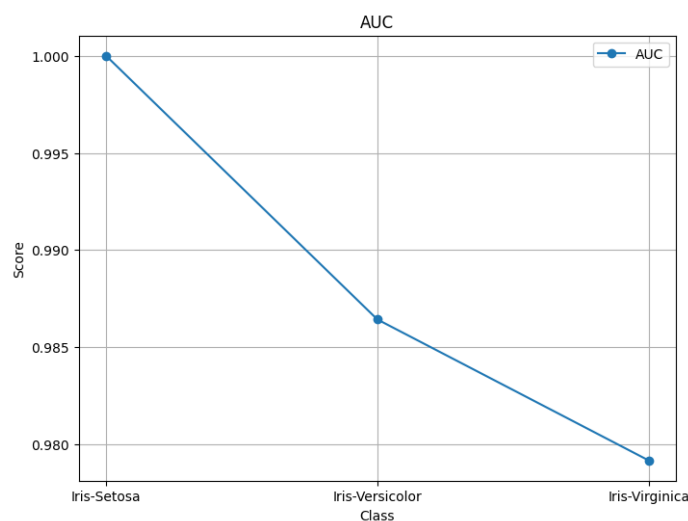


Figure 5.49: AUC Curve for Gradient Boosting Model on Iris Dataset

The AUC scores for each class are as follows:

- The AUC score of 1.0 signifies flawless separation between positive and negative instances. This implies that the model accurately distinguishes Iris Setosa from other species.
- With an AUC score of 0.99, there is a strong level of separation between positive and negative instances. This indicates that the model effectively discriminates Iris Versicolor from other species, though not as perfectly as in the case of Iris Setosa.
- Achieving an AUC score of 0.98, there is a notable degree of separation between positive and negative instances. This suggests that the model reliably differentiates Iris Virginica from other species, though slightly less distinctly than Iris Setosa and Iris Versicolor.

The AUC curve represents the performance of the Gradient Boosting model in terms of its ability to rank instances correctly across different classes. The single line on the plot connects the three points corresponding to the AUC scores for each class, illustrating the varying levels of separation achieved by the model for the different Iris classes.

The AUC curve is a valuable metric for evaluating the overall performance of a model. It provides a summarized view of the model's ability to discriminate between classes, with higher AUC scores indicating better performance.

In essence, the ROC and AUC analyses conducted on the Iris dataset using the Gradient Boosting model substantiate its efficacy in classifying instances into the

distinct Iris species. The ROC curves visually underscore the model’s discrimination ability, while the AUC scores provide a comprehensive quantification of its predictive accuracy across the varied classes.

5.3.6 Interpretability Through XAI Methods

In this subsection, we provide explanations for the predictions made by the Gradient Boosting model on the Iris dataset using three XAI (Explainable Artificial Intelligence) methods: XOPSIS, LIME, and SHAP. These methods aim to shed light on the important features that contribute to the model’s decision-making process.

1. Iris Setosa Explanations

Case 01

In Figure 5.50, we present the XOPSIS explanation for Case 01 in the Iris dataset. The figure starts by displaying the actual feature values of the data point under consideration. It provides insights into the specific values of the features, such as sepal length, sepal width, petal length, and petal width.

```

----- TOPSIS Explanation -----
Data point index: 44
----- Data Point Features -----
sepal_length    5.1
sepal_width     3.8
petal_length    1.9
petal_width     0.4
Name: 44, dtype: float64
Actual Species: 0.0
Predicted Species: 0.0

Explanation:
- The TOPSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  • petal_length
  • petal_width
- The features that contribute negatively to the prediction are:
  • sepal_length
  • sepal_width

```

Figure 5.50: Explanation of Iris-Setosa level by XOPSIS (Case 01)

Furthermore, the figure reveals that the actual species of the data point is Iris Setosa, while the model predicted the species as 0, which also corresponds to Iris Setosa. The XOPSIS explanation highlights the contributing features to the model’s prediction. It identifies positive contributing features, which in this case are petal length and petal width. These features positively influenced the model’s decision towards predicting the data point as Iris Setosa.

On the other hand, the XOPSIS explanation also identifies negative contributing features, which are sepal length and sepal width. These features had a negative impact on the model’s decision, suggesting that they were not significant in classifying the data point as Iris Setosa.

In Figure 5.51, we present the normalized feature contributions by XOPSIS for Case 01 in the Iris dataset. The figure illustrates the relative importance and direction of each feature’s contribution to the model’s prediction. The normalized feature contributions are visualized using a color scheme, where green represents positive contributions and red represents negative contributions. The intensity of the color indicates the magnitude of the contribution.

These normalized feature contributions provide insights into the relative importance of each feature in determining the model’s prediction for Case 01. The positive

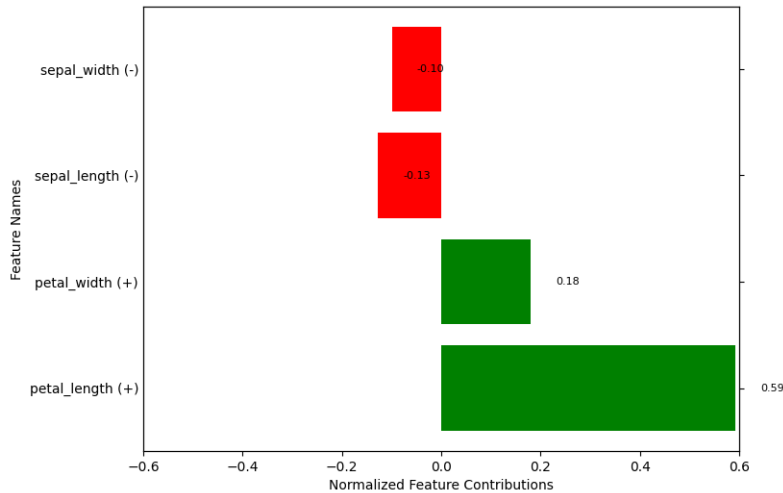


Figure 5.51: Normalized Feature Contributions of Iris-Setosa level by XOPSSIS (Case 01)

contributions of petal width and petal length suggest that higher values of these features favor the prediction of Iris Setosa. Conversely, the negative contributions of sepal width and sepal length imply that lower values of these features contribute towards the prediction of Iris Setosa.

By providing such explanations, XOPSSIS enhances our understanding of how individual features influence the model’s decision-making process for this specific data point in the Iris dataset.

In Figure 5.52, we present the LIME explanation for Case 01 in the Iris dataset. LIME provides an interpretable understanding of the model’s prediction by highlighting the important features and their contributions.

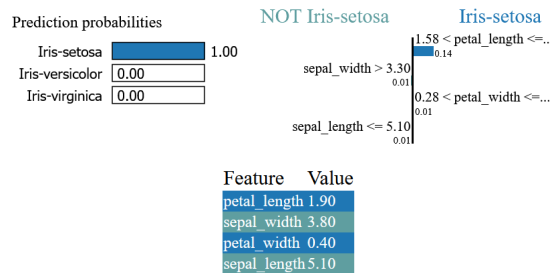


Figure 5.52: Explanation of Iris-Setosa level by LIME (Case 01)

The left section of the figure displays the prediction probabilities for each class. In this case, the prediction probability for Iris-Setosa is 1.0, indicating a high certainty in the model’s prediction for this class. The probabilities for the remaining classes (Not Iris-Setosa) are all 0, suggesting a clear distinction in the model’s prediction. In the right section, LIME highlights the features that contribute to the prediction of Iris-Setosa. Specifically, it indicates that a petal length greater than 1.58 and a petal width greater than 0.28 favor the prediction of Iris-Setosa. Conversely, it suggests that a sepal width greater than 3.30 and a sepal length less than or equal to 5.10 contribute to the prediction of Not Iris-Setosa.

The bottom section of the figure displays the actual feature values for Case 01. This provides a context for understanding how the feature values align with the LIME

explanation. By comparing the actual feature values with the highlighted features in the middle section, we can observe the consistency or discrepancy between the two.

Through the LIME explanation, we gain insights into the specific feature values that drive the model’s prediction for Case 01. This helps us understand the decision-making process of the model and provides interpretability to the predictions made for the Iris dataset.

Figure 5.53 presents the SHAP (SHapley Additive exPlanations) explanation for Case 01 in the Iris dataset. SHAP provides a unified framework for interpreting the predictions of machine learning models by assigning importance values to each feature.

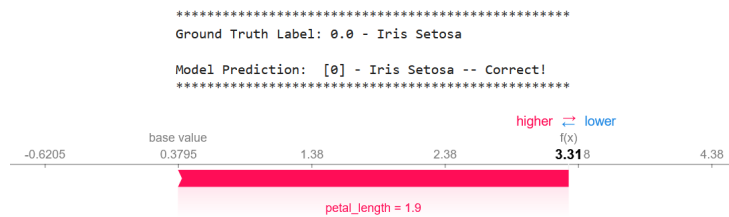


Figure 5.53: Explanation of Iris-Setosa level by SHAP (Case 01)

The top section of the figure displays the ground truth label and the model’s prediction. In this case, the ground truth label is Iris-Setosa, and the model’s prediction is also Iris-Setosa (predicted label: 0).

The bottom section of the figure showcases the SHAP force plot. The plot reveals the contribution of each feature to the model’s prediction for the given data point. The base value is indicated as 0.3759, representing the expected model output for the dataset.

For Case 01, the plot emphasizes the contribution of the Petal Length feature. The feature value of 1.9 is highlighted, and its impact on the prediction is depicted as a horizontal line connecting the base value to the feature value. The plot indicates that a higher value of Petal Length leads to a higher prediction value, while a lower value results in a lower prediction value.

The color scheme employed in the plot conveys the impact of the feature on the prediction. The region from the base value to the feature value (0.3759 to 3.31) is shown in red, indicating a positive influence, while the region below the base value is shown in blue, suggesting a negative influence. The force plot also provides additional information on the contribution of other features. However, in this particular case, the plot only includes Petal Length. This suggests that Petal Length is the primary driver behind the model’s prediction for Case 01.

By utilizing the SHAP force plot, we gain insights into the relative importance and influence of individual features on the model’s predictions. This enhances our understanding of the decision-making process of the model for the Iris dataset and facilitates the interpretation of its predictions for specific data points.

In summary, these explanations reveal commonalities in identifying important features such as Petal Length and Petal Width contributing positively to the prediction of Iris-Setosa. They also highlight the influence of Sepal Length and Sepal Width, which have negative contributions to the prediction.

While XOPSIS provides feature-level contributions, LIME offers local approximations, and SHAP assigns importance values. Together, these explanations provide a comprehensive understanding of how the model reaches its prediction for Case 01 in the Iris dataset.

Case 02

```

----- TOPSIS Explanation -----
Data point index: 14
----- Data Point Features -----
sepal_length  5.8
sepal_width   4.0
petal_length  1.2
petal_width   0.2
Name: 14, dtype: float64
Actual Species: 0.0
Predicted Species: 0.0

Explanation:
- The TOPSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  • petal_length
  • sepal_width
  • petal_width
- The features that contribute negatively to the prediction are:
  • sepal_length

```

Figure 5.54: Explanation of Iris-Setosa level by XOPSIS (Case 02)

Figure 5.54 represents the actual data point features, actual species (Iris-Setosa), and predicted species (Iris-Setosa). The explanation highlights the positively contributed features, including Petal Length, Petal Width, and Sepal Width, while Sepal Length is identified as a negatively contributed feature.

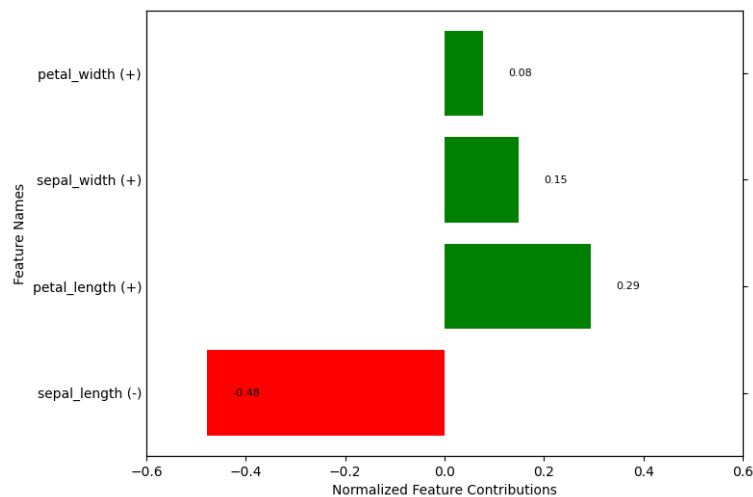


Figure 5.55: Normalized Feature Contributions of Iris-Setosa level by XOPSIS (Case 02)

In the Figure 5.55, the normalized feature contributions are displayed. The feature contributions are as follows: Petal Width (0.08), Petal Length (0.29), Sepal Width (0.15), and Sepal Length (-0.48). Positive contributions are depicted in green color, while negative contributions are represented in red color.

The LIME explanation for Case 02 in Figure 5.56 provides insights into the model's prediction. The prediction probabilities indicate a high confidence for Iris-Setosa (1.0), while the probabilities for the other species are 0. The LIME explanation

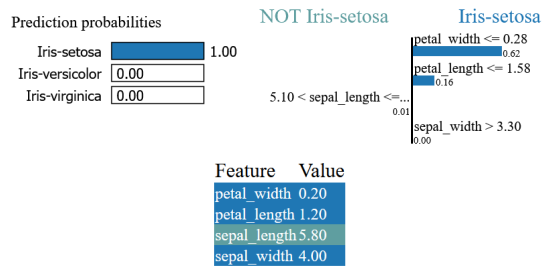


Figure 5.56: Explanation of Iris-Setosa level by LIME (Case 02)

highlights the conditions that contribute to the prediction of Iris-Setosa: petal width ≤ 0.28 (0.62), petal length ≤ 1.58 (0.16), and sepal width > 3.30 (0.0). Additionally, it identifies that a sepal length of more than 5.1 (0.01) is associated with a prediction of Not Iris-Setosa. The actual feature values of the data point are also provided in the explanation.



Figure 5.57: Explanation of Iris-Setosa level by SHAP (Case 02)

The SHAP explanation for Case 02 in the Figure 5.57 sheds light on the model’s prediction. It begins by stating the ground truth label and the model’s prediction: 0 (Iris Setosa). The explanation includes a plot where the base value is 0.3795. The feature importance is represented by $f(x)$ on a scale of 0.3795 to 3.31. The plot is annotated with “higher – $>$ – lower,” where “higher – $>$ ” is highlighted in red and “ $<$ – lower” is displayed in blue. In this case, the Petal length feature (1.2) is prominently shown in red, indicating its positive contribution to the model’s prediction.

In summary, all three explanations (XOPSIS, LIME, SHAP) provide insights into the model’s prediction for the Iris Setosa species. According to XOPSIS and LIME, the positively contributed features are Petal length, Petal width, and Sepal width, while Sepal length has a negative contribution. In SHAP, it reveals that the Petal length feature (1.2) has a positive contribution towards the prediction of Iris Setosa. These explanations consistently highlight the importance of features such as Petal length in determining the prediction of Iris Setosa in Case 02. Additionally, XOPSIS and LIME align in identifying Petal width and Sepal width as a contributing feature to the prediction as well.

Iris Versicolor Explanations

Case 03

In Case 03, the XOPSIS explanation provides insights into the prediction for the Iris Versicolor species. The explanation begins by displaying the actual feature values

```

----- TOPSIS Explanation -----
Data point index: 56
----- Data Point Features -----
sepal_length  6.3
sepal_width   3.3
petal_length  4.7
petal_width   1.6
Name: 56, dtype: float64
Actual risk level: 1.0
Predicted Risk Level: 1.0

Explanation:
- The TOPSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  • petal_width
  • petal_length
  • sepal_length
- The features that contribute negatively to the prediction are:
  • sepal_width

```

Figure 5.58: Explanation of Iris-Versicolor level by XOPSIS (Case 03)

of the data point in Figure 5.58. It then reveals that the actual species is 1 (Iris Versicolor) and the predicted species is also 1, indicating a correct prediction. The XOPSIS explanation further highlights the positively contributed features, which include Petal length, Petal width, and Sepal length. On the other hand, Sepal width is identified as a negatively contributed feature.

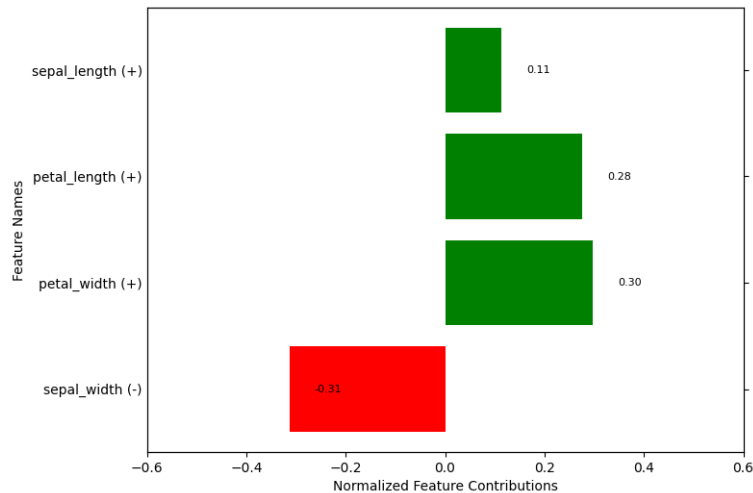


Figure 5.59: Normalized Feature Contributions of Iris-Versicolor level by XOPSIS (Case 03)

Figure 5.59 displays the normalized feature contributions, where the positive contributions (Petal length: 0.28, Petal width: 0.30, Sepal length: 0.11) are depicted in green color, and the negative contribution (Sepal width: -0.31) is shown in red color.

Overall, the XOPSIS explanation for Case 03 emphasizes the importance of features such as Petal length, Petal width, and Sepal length in predicting the Iris Versicolor species.

The LIME explanation from Figure 5.60 provides insights into the prediction of the Iris Versicolor species for a specific data point.

The explanation starts by displaying the prediction probabilities, indicating that the predicted probability for Iris Versicolor is 1, while the probabilities for other species are 0.

The LIME explanation further highlights the specific feature ranges that contribute to the prediction of Iris Versicolor. It states that for Iris Versicolor, the feature

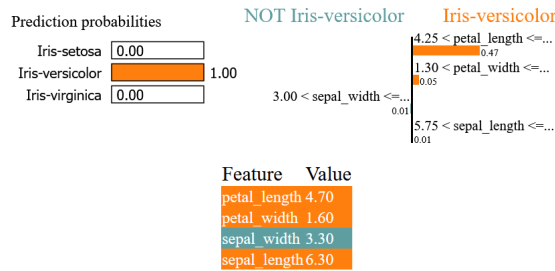


Figure 5.60: Explanation of Iris-Versicolor level by LIME (Case 03)

values of Petal length are greater than 4.25, Petal width is greater than 1.30, and Sepal length is greater than 5.75. Additionally, it indicates that for a prediction of Not Iris Versicolor, the feature value of Sepal width is greater than 3.

In the LIME explanation, these details are presented numerically, with corresponding importance weights assigned to each feature. For Iris Versicolor, Petal length has an importance weight of 0.47, Petal width has an importance weight of 0.05, Sepal length has an importance weight of 0.01, and Sepal width has an importance weight of 0.01.

Finally, the actual feature values of the data point are shown, providing a clear understanding of the values associated with each feature.

Overall, the LIME explanation for Case 03 highlights the specific feature ranges and their importance in predicting the Iris Versicolor species. It emphasizes that higher values of Petal length, Petal width, and Sepal length contribute to the prediction of Iris Versicolor, while a higher value of Sepal width is associated with a prediction of Not Iris Versicolor.

The SHAP explanation for Case 03 provides insights into the prediction of the Iris Versicolor species for a specific data point.

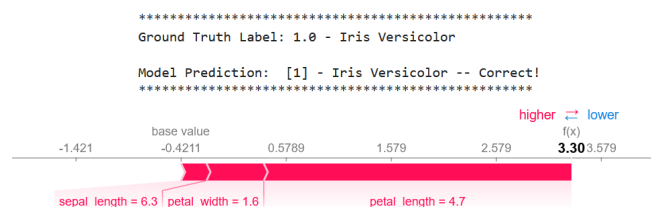


Figure 5.61: Explanation of Iris-Versicolor level by SHAP (Case 03)

In Figure 5.61, the explanation begins by displaying the ground truth label and model prediction, indicating that the true label is 1 (Iris Versicolor) and the model predicts the same.

The SHAP explanation further illustrates the contribution of different features through a force plot. The base value is -0.4211, and the force plot shows the progression of the $f(x)$ value. On top of the plot, it includes the notation “higher – > < – lower,” where “higher – >” is highlighted in red, and “< – lower” is highlighted in blue.

In the force plot, three features are shown in red color, indicating their positive contributions to the prediction of Iris Versicolor. These features are Sepal length (from -0.4211 to -0.1211), Petal width (from -0.1211 to 0.5789), and Petal length (from 0.5789 to 3.30). The red portion representing the contribution of Sepal length

is relatively smaller compared to the contribution of Petal width. By examining the force plot, it becomes apparent that higher values of Sepal length, Petal width, and Petal length positively contribute to the prediction of Iris Versicolor.

Overall, the SHAP explanation for Case 03 highlights the contribution of specific features and their ranges in predicting the Iris Versicolor species. It emphasizes that higher values of Sepal length, Petal width, and Petal length have a positive impact on the prediction, providing a clearer understanding of the importance of these features in the model's decision-making process.

In summary, these three explanations consistently emphasize the importance of features such as Petal length, Petal width, and Sepal length in determining the prediction of Iris Versicolor in Case 03. The features Sepal width are shown to have negative contributions in XOPSIS, LIME and SHAP explanations. The similarity among the explanations provides confidence in the significance of these features for the model's decision-making process.

Case 04

```

----- TOPSIS Explanation -----
Data point index: 85
----- Data Point Features -----
sepal_length  6.0
sepal_width   3.4
petal_length  4.5
petal_width   1.6
Name: 85, dtype: float64
Actual Species: 1.0
Predicted Species: 1.0

Explanation:
- The TOPSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  • petal_width
  • petal_length
- The features that contribute negatively to the prediction are:
  • sepal_width
  • sepal_length

```

Figure 5.62: Explanation of Iris-Versicolor level by XOPSIS (Case 04)

In the XOPSIS explanation for Case 04, it first displays the actual feature values of the data point. It then reveals the actual species and predicted species, with the predicted species being 1. The explanation highlights the positively contributed features, which include Petal length and Petal width. On the other hand, it identifies the negatively contributed features as Sepal length and Sepal width. These details are presented in Figure 5.62.

In Figure 5.63, the XOPSIS explanation provides the normalized feature contributions. The positive contributions are represented by Petal length (0.14) and Petal width (0.35), which are shown in green color. Conversely, the negative contributions are depicted by Sepal length (-0.10) and Sepal width (-0.42), which are displayed in red color.

Overall, the XOPSIS explanation for Case 04 consistently emphasizes that Petal length and Petal width have positive contributions to the predicted species, while Sepal length and Sepal width have negative contributions. These insights shed light on the importance of these features in the model's decision-making process.

The Lime explanation for the same data point in Figure 5.64, begins by presenting the prediction probabilities, with Iris-Versicolor having a probability of 1 and the

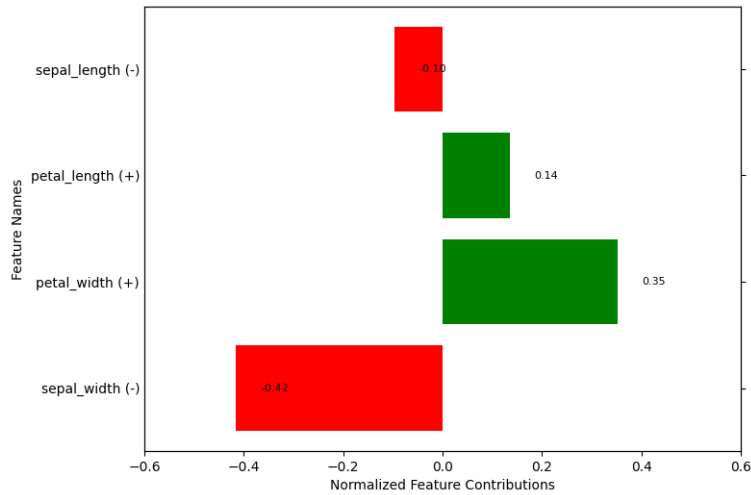


Figure 5.63: Normalized Feature Contributions of Iris-Versicolor level by XOP SIS (Case 04)

remaining classes having probabilities of 0. It then highlights the criteria for classifying the data point as Iris-Versicolor, indicating that the values for Petal width (0.23) and Petal length (0.18) are above the specified thresholds.

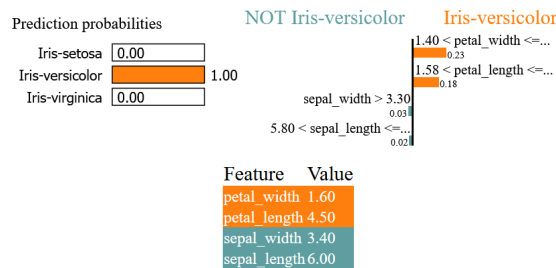


Figure 5.64: Explanation of Iris-Versicolor level by LIME (Case 04)

Additionally, it mentions that for the data point to be classified as not Iris-Versicolor, the values for Sepal width should be greater than 3.30 (0.03) and the Sepal length should be larger than 5.8 (0.02). Finally, it presents the actual feature values of the data point.

This Lime explanation provides insights into the specific features that contribute to the classification of the data point as Iris-Versicolor. It emphasizes the importance of Petal width and Petal length in determining the predicted class, while also considering the values of Sepal width and Sepal length in the decision-making process. The SHAP explanation for the same data point in Figure 5.65, begins by displaying the ground truth label and the model prediction, both of which are 1.0, indicating that the predicted class is Iris-Versicolor. It then presents a force plot, where the base value is -0.08492 and $f(x)$ is shown at 3.20. The plot includes annotations indicating “higher – >” in red color and “< – lower” in blue color.

Within the force plot, three features are highlighted in red color: Sepal width, Petal width, and Petal length. The red section is smallest for Sepal width, indicating a relatively lower contribution to the prediction. On the other hand, the red section is larger for Petal width and largest for Petal length, suggesting that these two features have a more substantial positive impact on the predicted class of Iris-Versicolor.

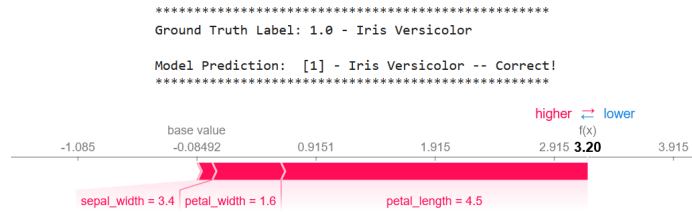


Figure 5.65: Explanation of Iris-Versicolor level by SHAP (Case 04)

This SHAP explanation provides insights into the relative contributions of different features in determining the predicted class of Iris-Versicolor for the given data point. It highlights the importance of Petal width and Petal length, with Petal length being the most influential feature in this particular case. The contribution of Sepal width is comparatively lower, indicating a less significant role in the prediction.

In summary, the explanations for Case 04 consistently highlight the significance of Petal length and Petal width as positively contributing features for predicting the class of Iris-Versicolor. Both XOPSIS and LIME explanations identify these features as important, while XOPSIS also highlights the negative contributions of Sepal length and Sepal width. The SHAP explanation reinforces the importance of Petal length and Petal width, showing them as the primary factors influencing the prediction. Overall, these explanations consistently emphasize the importance of Petal length and Petal width in determining the prediction of Iris-Versicolor in Case 04.

Iris-Virginica Explanations

Case 05

In the XOPSIS explanation for Case 05, the actual feature values of the data point are displayed in the Figure 5.66. It is revealed that the actual species is Iris-Virginica and the predicted species is also Iris-Virginica (2). The explanation highlights the positively contributed features, which include Petal length, Petal width, Sepal length, and Sepal width.

```

----- TOPSIS Explanation -----
Data point index: 118
----- Data Point Features -----
sepal_length  7.7
sepal_width   2.6
petal_length  6.9
petal_width   2.3
Name: 118, dtype: float64
Actual Species: 2.0
Predicted Species: 2.0

Explanation:
- The TOPSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  • petal_length
  • sepal_length
  • sepal_width
  • petal_width

```

Figure 5.66: Explanation of Iris-Virginica level by XOPSIS (Case 05)

In Figure 5.67, the normalized feature contributions are shown. Petal width has a contribution of 0.06, Petal length has a contribution of 0.52, Sepal length has a contribution of 0.32, and Sepal width has a contribution of 0.10. Positive contribu-

tions are depicted in green color, while negative contributions are represented in red color.

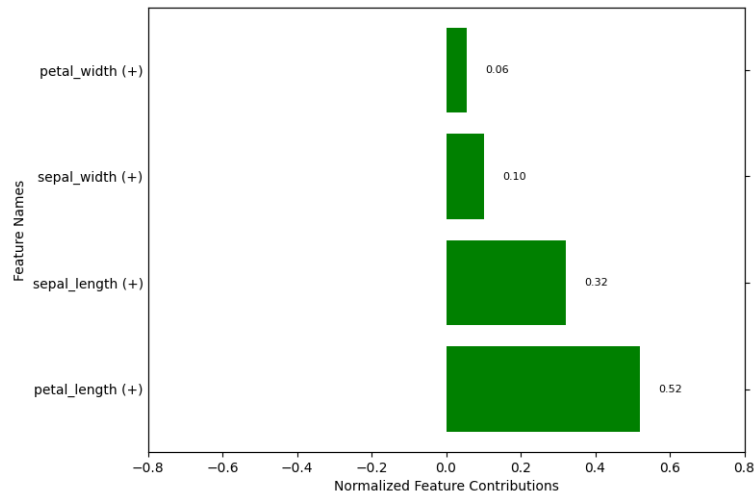


Figure 5.67: Normalized Feature Contributions of Iris-Virginica level by XOPSIS (Case 05)

Overall, the XOPSIS explanation for Case 05 suggests that Petal length, Petal width, Sepal length, and Sepal width play important roles in predicting the class of Iris-Virginica.

In Figure 5.68, for the same datapoint, the prediction probabilities are displayed. It shows that the prediction for Iris-Virginica is 1.0, indicating high confidence in the predicted class, while the probabilities for the other classes are all 0.

The explanation further highlights the specific threshold values for the features that contribute to the prediction of Iris-Virginica. It reveals that the Petal length should be greater than 5.10 (contribution: 0.64), the Petal width should be greater than 1.80 (contribution: 0.61), the Sepal length should be greater than 6.40 (contribution: 0.02), and the Sepal width should be less than or equal to 2.80 (contribution: 0.01) for the classification of Iris-Virginica. All these features positively contribute to the prediction of Iris-Virginica, indicating that higher values of Petal length, Petal width, Sepal length, and lower values of Sepal width are indicative of the Iris-Virginica class.

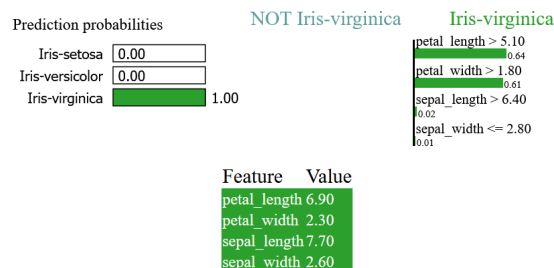


Figure 5.68: Explanation of Iris-Virginica level by LIME (Case 05)

Lastly, the actual feature values for the datapoint are presented, allowing for a direct comparison with the LIME explanations.

Overall, the LIME explanation for Case 05 provides insights into the threshold values of key features that contribute to the prediction of Iris-Virginica, emphasizing

the importance of Petal length, Petal width, Sepal length, and Sepal width in determining the classification outcome.

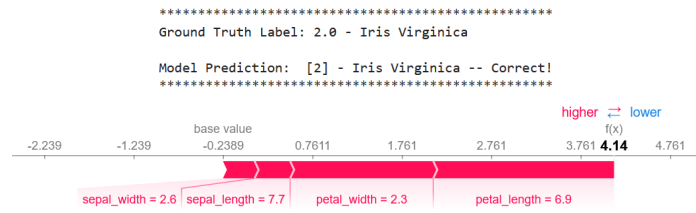


Figure 5.69: Explanation of Iris-Virginica level by SHAP (Case 05)

The Shap explanation in Figure 5.69, for the given datapoint indicates that the ground truth label and model prediction are both Iris-Virginica. The corresponding force plot displays a base value of -0.2389 and $f(x)$ value at 4.14. The plot is annotated with “higher - >” in red color and “< - lower” in blue color.

Examining the contributions, we observe that sepal width has the lowest red part, suggesting a relatively smaller positive impact on the prediction. Sepal length shows a lower red part, indicating a slightly stronger positive contribution. Moving further, petal width exhibits a larger red part, implying a more significant positive influence. Finally, petal length displays the largest red part, signifying the highest positive contribution among the features considered in this explanation.

In summary, the three explanations (XOPSIS, LIME, SHAP) consistently highlight the importance of all four features (petal length, petal width, sepal length, and sepal width) in predicting Iris-Virginica in Case 05. Each explanation identifies these features as positively contributing to the prediction, emphasizing their significance in different degrees. This consistency underscores the robustness of these features in determining the classification of Iris-Virginica.

Case 06

```
----- TOPSIS Explanation -----
Data point index: 143
----- Data Point Features -----
sepal_length  6.8
sepal_width   3.2
petal_length  5.9
petal_width   2.3
Name: 143, dtype: float64
Actual Species: 2.0
Predicted Species: 2.0

Explanation:
- The TOPSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  • petal_length
  • petal_width
- The features that contribute negatively to the prediction are:
  • sepal_width
  • sepal_length
```

Figure 5.70: Explanation of Iris-Virginica level by XOPSIS (Case 06)

In Figure 5.70, the XOPSIS explanation reveals the importance of features in predicting Iris-Virginica. The actual feature values of the data point are displayed, followed by the actual and predicted species, which are both Iris-Virginica. The explanation identifies the positively contributing features as petal length and petal width, indicating their significance in the prediction. On the other hand, the negatively contributing features are identified as sepal length and sepal width, suggesting

that higher values in these features are associated with a lower likelihood of the Iris-Virginica classification.

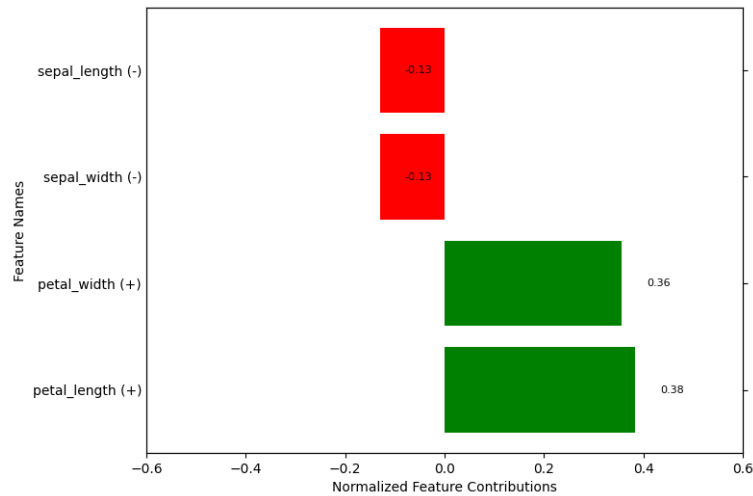


Figure 5.71: Normalized Feature Contributions of Iris-Virginica level by XOPIS (Case 06)

The normalized feature contributions in Figure 5.71, further illustrate the importance of these features. Petal length and petal width have positive contributions, indicated by their green color, suggesting that higher values in these features contribute positively to the prediction of Iris-Virginica. In contrast, sepal length and sepal width have negative contributions, represented by their red color, indicating that higher values in these features have a negative impact on the prediction.

Overall, this XOPIS explanation provides insights into the specific features that play a crucial role in determining the classification of Iris-Virginica in Case 06.

The Lime explanation in Figure 5.72 provides insights into the prediction of Iris-Virginica. The prediction probabilities indicate a high confidence of 1.0 for Iris-Virginica, with the probabilities of the remaining classes being 0. The explanation highlights the following conditions for Iris-Virginica: petal length greater than 5.2 (0.72) and petal width greater than 1.80 (0.32). These conditions suggest that larger values in these features contribute to the classification of Iris-Virginica.

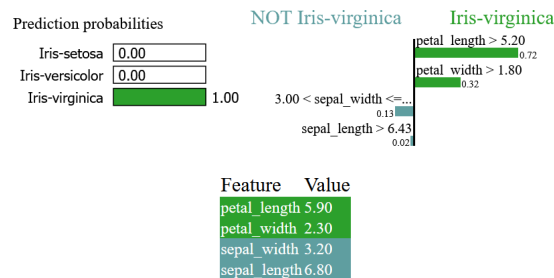


Figure 5.72: Explanation of Iris-Virginica level by LIME (Case 06)

Additionally, the explanation shows that a sepal width greater than 3 (0.13) and a sepal length greater than 6.43 (0.02) are associated with the prediction of not being Iris-Virginica. The actual feature values of the data point are also presented.

Overall, the Lime explanation for Case 06 emphasizes the importance of petal length and petal width in predicting Iris-Virginica, while suggesting that larger sepal width and sepal length values are indicative of other classes.

The Shap explanation in Figure 5.73 provides insights into the prediction of Iris-Virginica. The ground truth label and model prediction indicate a classification of 2, corresponding to Iris-Virginica. The force plot reveals a base value of -0.7175 and a $f(x)$ value of -2.24. The plot is divided into sections marked with “higher – >” in red and “< – lower” in blue.

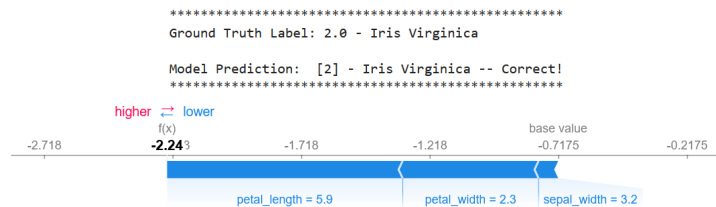


Figure 5.73: Explanation of Iris-Virginica level by SHAP (Case 06)

The blue sections represent features with lower contributions, while the largest blue section corresponds to petal length=5.9. This suggests that a lower value of petal length has a negative impact on the prediction of Iris-Virginica. Similarly, the blue section for petal width=2.3 indicates that a lower value of petal width also contributes negatively to the prediction.

On the other hand, the smaller blue section for sepal width=3.2 suggests that a slightly lower value in sepal width has a comparatively smaller negative impact on the prediction. It’s worth noting that the force plot only shows the negative contributions, as indicated by the blue color.

Overall, the Shap explanation for Case 06 highlights the importance of petal length and petal width in predicting Iris-Virginica, emphasizing that lower values in these features contribute to a lower likelihood of Iris-Virginica being the predicted class. The smaller negative contribution of sepal width suggests that it has a relatively lesser impact on the prediction.

In summary, the explanations for Case 06 consistently highlight the importance of petal length and petal width in determining the prediction of Iris-Virginica. Both XOPSSIS and LIME explanations identify these features as positively contributing factors to the prediction, while sepal length and sepal width are identified as negatively contributing factors. In the Shap explanation, the force plot indicates that petal length = 5.9 and petal width = 2.3 have larger blue sections, suggesting that lower values in these features negatively impact the prediction of Iris-Virginica. The smaller blue section for sepal width = 3.2 indicates a relatively smaller negative impact on the prediction. Overall, these explanations consistently emphasize the importance of petal length and petal width in determining the prediction of Iris-Virginica in Case 06. They highlight that higher values in these features positively contribute to the likelihood of Iris-Virginica being the predicted class, while sepal length and sepal width have negative contributions.

5.4 Result Analysis on Breast Cancer Wisconsin Dataset

5.4.1 Accuracy

The results of the accuracy analysis for the breast cancer Wisconsin dataset using three different algorithms, namely Gradient Boosting, XGBoost, and Random Forest, are presented in Table 5.5. The accuracy values are provided for both the train-test split and k-fold cross-validation scenarios.

Serial No	Algorithms	Accuracy	
		Train-Test Split	K-Fold Cross Validation
01	Gradient Boosting	94.737%	92.278%
02	XGBoost	92.982%	93.333%
03	Random Forest	90.000%	91.917%

Table 5.5: Accuracy of 3 Algorithms

For Gradient Boosting, the train-test split accuracy is reported as 94.737%, indicating that the model achieved a high level of accuracy when evaluated on the independent test set. Similarly, the k-fold cross-validation accuracy for Gradient Boosting is reported as 92.278%, which demonstrates the model's consistent performance across different folds.

XGBoost also yielded promising results, with a train-test split accuracy of 93.860%. This indicates the model's ability to accurately classify breast masses into malignant or benign categories. The k-fold cross-validation accuracy for XGBoost is reported as 93.333%, suggesting that the model's performance remains stable across different train-test splits.

Random Forest achieved a train-test split accuracy of 92.982%, showing its effectiveness in accurately predicting the breast cancer outcomes. The k-fold cross-validation accuracy for Random Forest is reported as 91.917%, indicating the model's reliability and generalization capability across multiple folds.

To visually represent the accuracy of these algorithms, a bar plot was generated in Figure 5.74. The x-axis of the bar plot represents the names of the three algorithms (Gradient Boosting, XGBoost, and Random Forest), while the y-axis represents the corresponding accuracy values. This plot provides a quick and intuitive comparison of the accuracy performance among the algorithms, further supporting the findings presented in Table 5.5.

Overall, the accuracy analysis demonstrates the effectiveness of Gradient Boosting, XGBoost, and Random Forest in accurately classifying breast masses and highlights their potential for assisting in breast cancer diagnosis and decision-making processes.

5.4.2 Confusion Matrix

The confusion matrix, presented in Figure 5.75, provides a detailed analysis of the model's performance by comparing the true labels and predicted labels for the breast

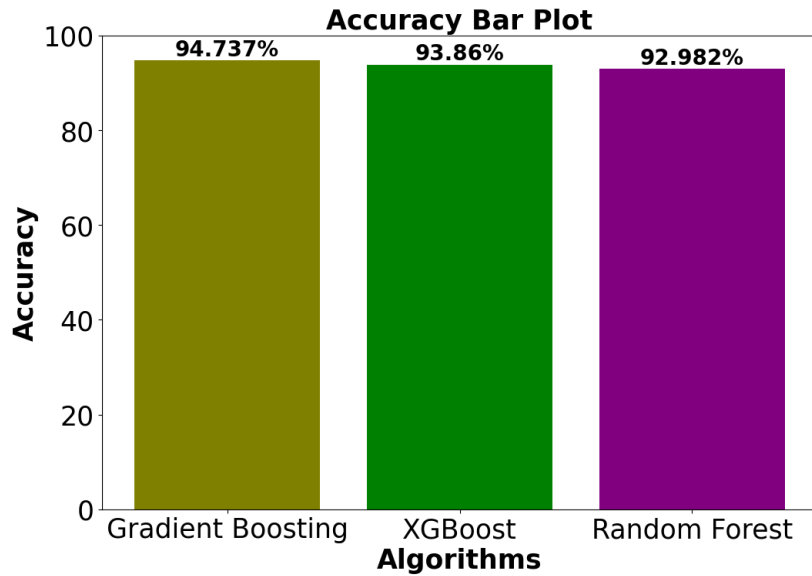


Figure 5.74: Train-Test Split Accuracy for Different Algorithms on Breast Cancer Dataset

cancer Wisconsin dataset. The matrix is organized vertically based on the true labels and horizontally based on the predicted labels.

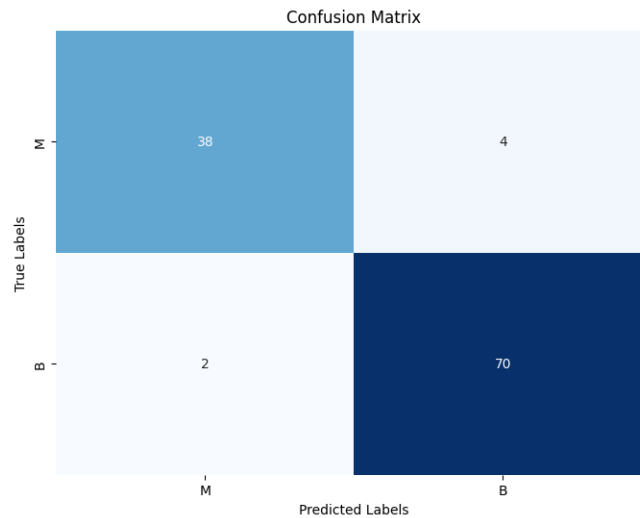


Figure 5.75: Confusion Matrix for Breast Cancer Dataset

In the first row of the confusion matrix, the label “M” represents malignant breast masses. The value (38,4) indicates that out of the total 42 instances of malignant masses, the model correctly predicted 38 cases as malignant (true positives), while 4 cases were incorrectly classified as benign (false negatives).

In the second row, the label “B” represents benign breast masses. The value (2, 70) shows that out of the total 72 instances of benign masses, the model correctly predicted 70 cases as benign (true negatives), while 2 cases were incorrectly classified as malignant (false positives).

The confusion matrix provides valuable insights into the model’s performance in terms of correctly identifying malignant and benign breast masses. It highlights the

trade-offs between false positives and false negatives, which are crucial considerations in breast cancer diagnosis. The model’s ability to accurately classify malignant masses (high true positive rate) and benign masses (high true negative rate) is essential for minimizing misdiagnosis and ensuring appropriate medical interventions. The presented confusion matrix assists in evaluating the model’s performance metrics such as accuracy, precision, recall, and F1 score, which provide a comprehensive understanding of the model’s effectiveness in breast cancer classification. These metrics help gauge the model’s ability to correctly identify both malignant and benign cases and serve as important benchmarks for assessing the model’s performance in real-world clinical settings.

5.4.3 Sensitivity, Specificity, Precision and F1-Score

	Sensitivity	Specificity	Precision	F1-Score
M	0.905	0.972	0.95	0.93
B	0.972	0.905	0.95	0.96

Table 5.6: Performance Evaluation Metrics on Breast Cancer Dataset

The subsection on Sensitivity, Specificity, Precision, and F1-Score presents a comprehensive evaluation of the model’s performance on the breast cancer Wisconsin dataset. The performance metrics are summarized in Table 5.6 and visualized in a plot. This table provides an overview of the metrics for both malignant (M) and benign (B) classes.

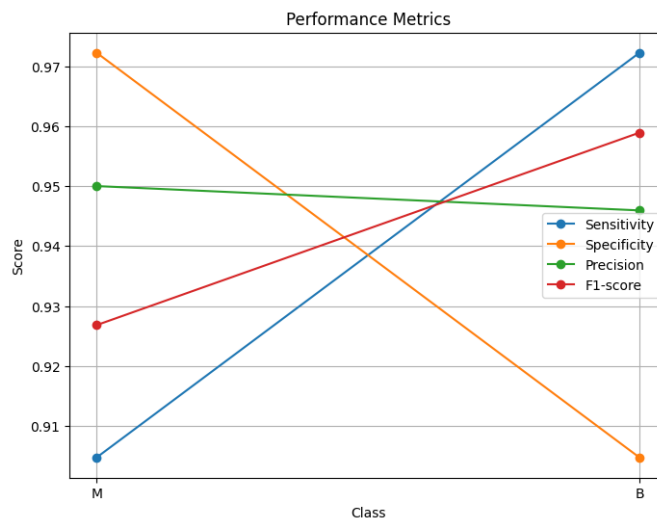


Figure 5.76: Performance Evaluation Metrics on Breast Cancer Dataset

For the malignant class (M), the model achieved a sensitivity (true positive rate) of 0.905, indicating that it correctly identified 90.5% of the actual malignant cases. The specificity (true negative rate) for the malignant class is 0.972, indicating that the model correctly classified 97.2% of the actual benign cases as benign. The precision score, which represents the proportion of correctly classified positive predictions, is

0.95, indicating that 95% of the instances predicted as malignant were indeed malignant. The F1-score, which considers both precision and recall, is 0.93, providing an overall measure of the model's performance for the malignant class.

For the benign class (B), the model achieved a sensitivity of 0.972, indicating that it correctly identified 97.2% of the actual benign cases. The specificity for the benign class is 0.905, indicating that the model correctly classified 90.5% of the actual malignant cases as malignant. The precision score for the benign class is 0.95, indicating that 95% of the instances predicted as benign were indeed benign. The F1-score for the benign class is 0.96, providing an overall measure of the model's performance for the benign class.

These metrics play a crucial role in evaluating the model's performance and assessing its effectiveness in breast cancer classification. Sensitivity and specificity provide insights into the model's ability to correctly identify positive and negative cases, respectively. Precision represents the proportion of true positive predictions among all positive predictions, while the F1-score combines precision and recall into a single metric that balances the trade-off between them.

To visually depict these performance metrics, a plot is presented in Figure 5.76. The plot shows the class labels (malignant and benign) on the x-axis and the scores on the y-axis. Four performance metrics (sensitivity, specificity, precision, and F1-score) are displayed for both classes. This plot provides a concise and intuitive representation of the model's performance across different evaluation metrics, allowing for easy comparison and interpretation.

5.4.4 ROC Curve

The ROC curves provide valuable insights into the classification performance of the Gradient Boosting model on the Breast Cancer dataset. In Figure 5.77, the ROC plot exhibits a familiar pattern, with two distinct curves representing the two classes within the dataset. The classes, namely Malignant and Benign, are plotted along the x-axis (False Positive Rate) and y-axis (True Positive Rate). The plot showcases how the model's sensitivity to true positives and specificity in true negatives trade off as the classification threshold varies. Remarkably, the ROC curves for both classes in this dataset exhibit a similar pattern, suggesting balanced performance in differentiating between Malignant and Benign cases.

The ROC curve provides a graphical representation of the model's trade-off between true positive rate and false positive rate across different classification thresholds. It allows us to assess the model's discrimination ability and select an appropriate threshold based on the desired balance between sensitivity and specificity.

5.4.5 AUC

Figure 5.78, displaying the AUC scores, complements the insights obtained from the ROC curves. AUC (Area Under the Curve) is a vital metric indicating the overall discriminatory power of the model. In this specific context, an AUC score of 0.97 is achieved for both Malignant and Benign classes. This implies a high level of separation between the positive and negative instances within each class. Such AUC scores highlight the effectiveness of the Gradient Boosting model in distinguishing between Malignant and Benign cases. The similarity in AUC scores further reinforces

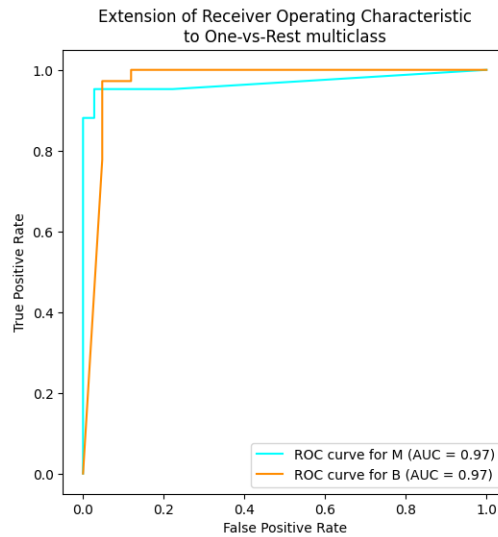


Figure 5.77: ROC Curve for Gradient Boosting Model on Breast Cancer Dataset

the consistent and reliable performance of the model across the two distinct classes. This information is crucial for evaluating the model’s diagnostic capabilities and its suitability for real-world medical applications.

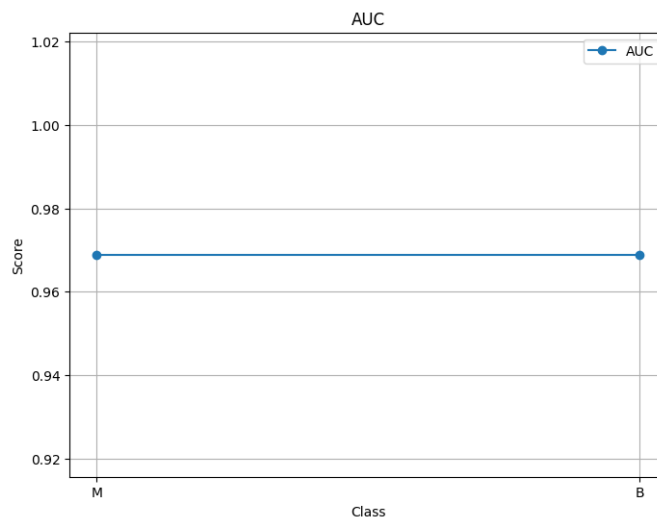


Figure 5.78: AUC Curve on Breast Cancer Dataset

For the Gradient Boosting model, the AUC scores are as follows:

- Malignant (M) class: 0.97
- Benign (B) class: 0.97

These AUC scores indicate the ability of the Gradient Boosting model to correctly classify instances from both the malignant (M) and benign (B) classes. The scores range from 0 to 1, with a higher score indicating a better discriminatory performance. The AUC score provides a quantitative measure of the model’s ability to distinguish between the classes, where a score of 1 represents a perfect classifier and a score of 0.5 indicates a random classifier.

The AUC score for the Gradient Boosting model offers valuable insights into its overall performance, allowing us to assess its effectiveness in classifying breast cancer

instances.

In summary, the ROC and AUC analyses demonstrate the robustness and effectiveness of the Gradient Boosting model in classifying Malignant and Benign cases within the Breast Cancer dataset. The similarity in performance across both classes underscores the model's capability to provide accurate and balanced predictions in this critical medical domain.

5.4.6 Interpretability Through XAI Methods

In this subsection, we provide explanations for the predictions made by the Gradient Boosting model on the Breast Cancer dataset using three XAI (Explainable Artificial Intelligence) methods: XOPSSIS, LIME, and SHAP. These methods aim to shed light on the important features that contribute to the model's decision-making process.

1. Malignant Explanations

Case 01

The XOPSSIS analysis was performed on a specific data point to provide insights into its diagnosis prediction. The analysis includes an examination of the actual feature values, the actual and predicted diagnosis, as well as the contribution of each feature to the prediction.

Figure 5.79 displays the actual feature values of the data point, providing a visual representation of its characteristics. The analysis reveals that the actual diagnosis of the data point is 'M' (malignant). The predicted diagnosis is also 'M', indicating that the model correctly identified the data point as malignant.

The XOPSSIS analysis further delves into the features that positively and negatively contribute to the prediction. The positive contributions are the features that support the prediction of malignancy, while the negative contributions are the features that suggest a benign diagnosis.

The following features positively contribute to the prediction: `area_mean`, `area_worst`, `area_se`, `perimeter_mean`, `perimeter_worst`, `texture_worst`, `radius_mean`, `perimeter_se`, `texture_mean`, `concavity_worst`, `compactness_worst`, `concavity_mean`, `compactness_mean`, `radius_se`, `concave_points_worst`, `symmetry_worst`, `concave_points_mean`, `symmetry_mean`, `smoothness_worst`, `fractal_dimension_worst`, `concavity_se`, `compactness_se`, `smoothness_mean`, `concave_points_se`, `fractal_dimension_se`. The following features negatively contribute to the prediction: `radius_worst`, `texture_se`, `smoothness_se`, `fractal_dimension_mean`, `symmetry_se`.

Figure 5.80 depicts the normalized feature contributions, with positive contributions shown in green and negative contributions shown in red. It is observed that most of the normalized feature contributions are 0, indicating that they have minimal impact on the prediction for this specific data point.

The XOPSSIS analysis provides valuable insights into the importance of different features in predicting the diagnosis of the data point. By understanding the contribution of each feature, we can gain a better understanding of the factors influencing the model's prediction and make informed decisions based on this analysis.

Figure 5.81 represents the LIME explanation by showing the prediction probabilities for the two classes: M (malignant) with a probability of 1.0 and B (benign) with a


```

----- TOPSIS Explanation -----
Data point index: 70
----- Data Point Features -----
radius_mean          18.940000
texture_mean         21.310000
perimeter_mean      123.600000
area_mean           1130.000000
smoothness_mean     0.000090
compactness_mean    0.102900
concavity_mean      0.108000
concave_points_mean 0.079510
symmetry_mean       0.158200
fractal_dimension_mean 0.054610
radius_se           0.788800
texture_se          0.797500
perimeter_se        5.486000
area_se            96.050000
smoothness_se       0.004444
compactness_se      0.016520
concavity_se        0.022690
concave_points_se   0.013700
symmetry_se         0.013860
fractal_dimension_se 0.001698
radius_worst        24.860000
texture_worst       26.580000
perimeter_worst     165.900000
area_worst          1866.000000
smoothness_worst    0.119300
compactness_worst   0.233600
concavity_worst     0.268700
concave_points_worst 0.178900
symmetry_worst      0.255100
fractal_dimension_worst 0.065890
Name: 70, dtype: float64
Actual Diagnosis: 0.0
Predicted Diagnosis: 0.0

Explanation:
- The TOPSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  * area_mean
  * area_worst
  * area_se
  * perimeter_mean
  * perimeter_worst
  * texture_worst
  * radius_mean
  * perimeter_se
  * texture_mean
  * concavity_worst
  * compactness_worst
  * concavity_mean
  * compactness_mean
  * radius_se
  * concave points_worst
  * symmetry_worst
  * concave points_mean
  * symmetry_mean
  * smoothness_worst
  * fractal_dimension_worst
  * concavity_se
  * compactness_se
  * smoothness_mean
  * concave points_se
  * fractal_dimension_se
- The features that contribute negatively to the prediction are:
  * radius_worst
  * texture_se
  * smoothness_se
  * fractal_dimension_mean
  * symmetry_se

```

Figure 5.79: Explanation of Malignant level by XOPSIS (Case 01)

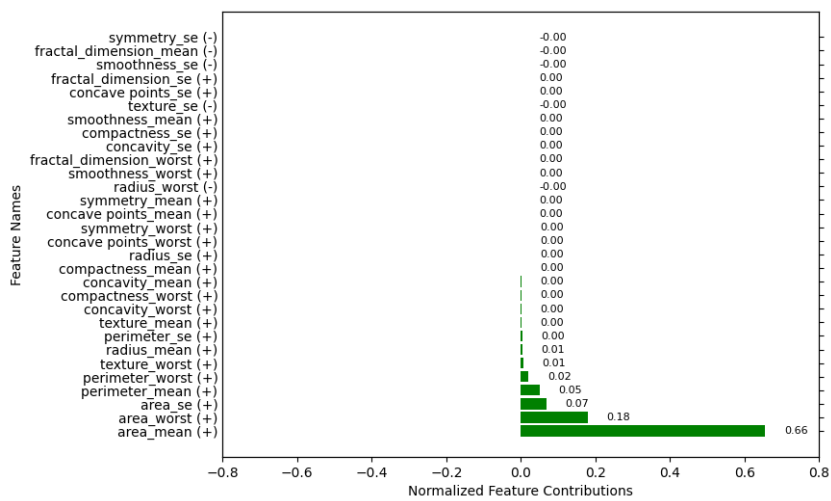


Figure 5.80: Normalized Feature Contributions of Malignant level by XOPSIS (Case 01)

probability of 0.0.

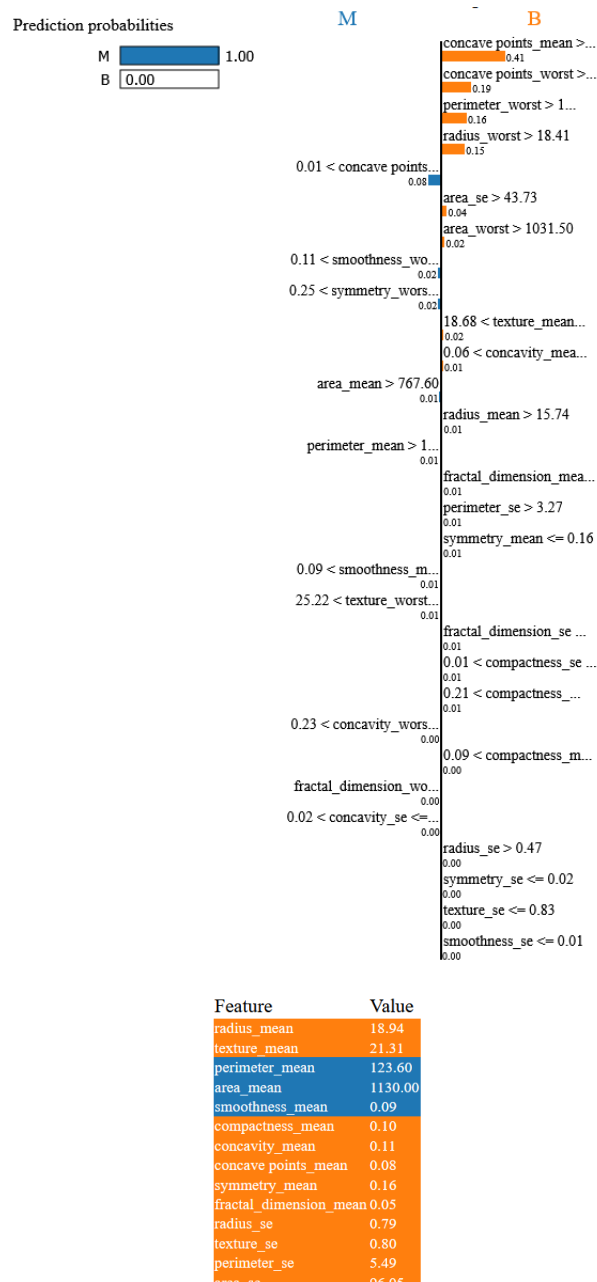


Figure 5.81: Explanation of Malignant level by LIME (Case 01)

Next, a plot is displayed, divided into two sections: M (left side) and B (right side).

In the M section, the following features are highlighted:

- The concave points feature has a value greater than 0.01 and contributes to the prediction as 0.08.
- The smoothness worst feature has a value greater than 0.11 and contributes as 0.02.
- The symmetry worst feature has a value greater than 0.25 and contributes as 0.02.
- The area mean feature has a value larger than 767.60 and contributes as 0.01.
- The perimeter mean feature has a value larger than 1 and contributes as 0.01.

- The smoothness mean feature has a value greater than 0.09 and contributes as 0.01.
- The texture worst feature has a value greater than 25.22 and contributes as 0.01.
- The concavity worst feature has a value greater than 0.23 and contributes as 0.00.
- Other features like fractal dimension worst, concavity se, and compactness se also contribute as 0.00.

In the B section, the following features are highlighted:

- The concave points mean feature has a value greater than 0.41 and contributes to the prediction as 0.19.
- The concave points worst feature has a value greater than 0.19 and contributes as 0.16.
- The perimeter worst feature has a value larger than 1 and contributes as 0.15.
- The radius worst feature has a value larger than 18.41 and contributes as 0.15.
- Other features like area se, area worst, texture mean, concavity mean, radius mean, and fractal dimension mean also contribute to the prediction with varying values.

Finally, the actual feature values of the data point are displayed, providing a comprehensive view of the attributes associated with the prediction.

This LIME explanation highlights the specific features and their corresponding contributions that influence the prediction of the given data point, providing insights into the decision-making process of the model.

The SHAP explanation for the given data point in Figure 5.82 starts by indicating the Ground Truth label, which is 0, and the Model prediction, which is also 0, indicating a correct prediction.

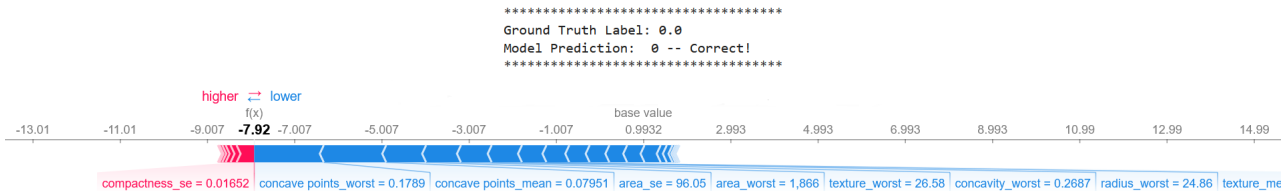


Figure 5.82: Explanation of Malignant level by SHAP (Case 01)

Next, a force plot is displayed, representing the SHAP values. The base value is shown as 0.9932, and the $f(x)$ value is -7.92. The force plot is divided into two sections: “*higher* - >” shown in red color and “< - *lower*” shown in blue color.

In the “*higher* - >” section, the following features are highlighted as having a positive impact on the prediction: - Compactness se - Compactness mean - Fractal dimension mean - Symmetry mean

In the “< - *lower*” section, the following features are highlighted as having a negative impact on the prediction: - Concave points worst - Concave points mean - Area se - Area worst - Texture worst - Concavity worst - Radius worst

The SHAP explanation provides insights into the contribution of each feature to the prediction outcome. The features in the “*higher* - >” section positively influence the prediction, while those in the “< - *lower*” section have a negative influence.

By analyzing the SHAP values, we can understand how each feature contributes to the final prediction and gain a deeper understanding of the model’s decision-making process for the given data point.

Case 02

In Figure 5.83, we begin by displaying the actual feature values of the data point. The actual diagnosis is predicted to be Malignant (0.0) based on our model. The XOPSSIS explanation for this data point is as follows:

```
----- TOPSSIS Explanation -----
Data point index: 83
----- Data Point Features -----
radius_mean      19.100000
texture_mean     26.290000
perimeter_mean   129.100000
area_mean        1132.000000
smoothness_mean  0.121500
compactness_mean 0.179100
concavity_mean   0.193700
concave_points_mean 0.146900
symmetry_mean    0.163400
fractal_dimension_mean 0.072240
radius_se        0.519000
texture_se       2.910000
perimeter_se     5.801000
area_se          67.100000
smoothness_se    0.007545
compactness_se   0.060500
concavity_se     0.021340
concave_points_se 0.018430
symmetry_se      0.030560
fractal_dimension_se 0.010390
radius_worst     20.330000
texture_worst    32.720000
perimeter_worst  141.300000
area_worst       1298.000000
smoothness_worst 0.139200
compactness_worst 0.281700
concavity_worst  0.223200
concave_points_worst 0.184100
symmetry_worst   0.231100
fractal_dimension_worst 0.092030
Name: 83, dtype: float64
Actual Diagnosis: 0.0
Predicted Diagnosis: 0.0

Explanation:
- The TOPSSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  • area_worst
  • area_mean
  • area_se
  • perimeter_worst
  • perimeter_mean
  • radius_worst
  • radius_mean
  • perimeter_se
  • concavity_worst
  • radius_se
  • compactness_worst
  • concavity_mean
  • symmetry_worst
  • concave_points_worst
  • compactness_mean
  • symmetry_mean
  • texture_se
  • concavity_se
  • smoothness_worst
  • fractal_dimension_worst
  • concave_points_se
  • fractal_dimension_mean
  • symmetry_se
- The features that contribute negatively to the prediction are:
  • texture_mean
  • texture_worst
  • concave_points_mean
  • smoothness_mean
  • compactness_se
  • fractal_dimension_se
  • smoothness_se
```

Figure 5.83: Explanation of Malignant level by XOPSSIS (Case 02)

- The TOPSSIS score is computed by comparing the feature values of the given data point with those of similar instances.
- The features that contribute positively to the prediction of Malignant diagnosis include area_worst, area_mean, area_se, perimeter_worst, perimeter_mean, radius_worst, radius_mean, perimeter_se, concavity_worst, radius_se, compactness_worst, concavity_mean, symmetry_worst, concave_points_worst, compactness_mean, symmetry_mean, texture_se, concavity_se, smoothness_worst, fractal_dimension_worst, concave_points_se, fractal_dimension_mean, and symmetry_se. - Conversely, the features that contribute negatively to the prediction are texture_mean, texture_worst, concave_points_mean, smoothness_mean, compactness_se, fractal_dimension_se, and smoothness_se.

In Figure 5.84, we illustrate the normalized feature values. Positive feature contributions are highlighted in green, indicating their influence in the prediction, while negative feature contributions are shown in red. It is worth noting that most of

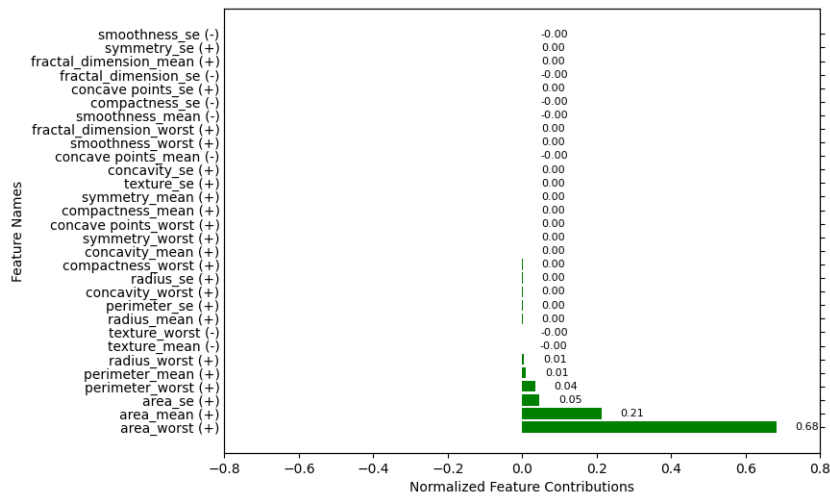


Figure 5.84: Normalized Feature Contributions of Malignant level by XOPSSIS (Case 02)

the normalized feature contributions are close to 0, indicating limited impact on the prediction outcome.

This XOPSSIS explanation sheds light on the specific features that drive the model’s prediction of Malignant diagnosis for the given data point, providing insights into the factors influencing the outcome.

In the LIME explanation for the given data point, we start by displaying the prediction probabilities: Malignant (M) with a probability of 1.0 and Benign (B) with a probability of 0.0. The LIME explanation further utilizes a plot in Figure 5.85 where the left side represents the M class and the right side represents the B class. In the left side (M), the plot highlights the following features and their corresponding contributions to the M prediction: $\text{concave_points_se} > 0.01$ (0.10), $0.13 < \text{smoothness_worst}$ (0.01), $\text{symmetry_se} > 0.02$ (0.01), $\text{perimeter_se} > 3.27$ (0.01), $\text{perimeter_mean} > 1$ (0.01), $\text{concavity_mean} > 0.13$ (0.01), $0.16 < \text{symmetry_mean}$ (0.00), $0.01 < \text{smoothness_se}$ (0.00), $\text{compactness_mean} >$ (0.00), and $\text{fractal_dimension_se} >$ (0.00).

In the right side (B), the plot showcases the features and their contributions to the B prediction: $\text{concave_points_mean} >$ (0.42), $\text{concave_points_worst} >$ (0.19), $\text{perimeter_worst} > 1$ (0.17), $\text{radius_worst} > 18.41$ (0.16), $\text{texture_worst} > 29.69$ (0.12), $\text{texture_mean} > 21.59$ (0.07), $\text{area_worst} > 1031.50$ (0.03), $\text{symmetry_worst} \leq 0.25$ (0.02), $\text{area_se} > 43.73$ (0.02), $\text{fractal_dimension_mean}$ (0.01), $\text{radius_se} > 0.47$ (0.01), $\text{compactness_se} > 0.03$ (0.01), $\text{radius_mean} > 15.74$ (0.01), $0.21 < \text{compactness}$ (0.01), $\text{area_mean} > 767.60$ (0.01), $0.23 < \text{concavity_worst}$ (0.01), $0.02 < \text{concavity_se} \leq$ (0.01), $\text{texture_se} > 1.44$ (0.00), $0.08 < \text{fractal_dimension}$ (0.00), and $\text{smoothness_mean} > 0.10$ (0.00).

The actual feature values of the data point are then presented to provide a complete understanding of the instance. The LIME explanation offers valuable insights into the significant features contributing to the Malignant and Benign predictions, enabling a better understanding of the decision-making process of the model.

Figure 5.86 displays the SHAP explanation for the given data point begins by indicating the Ground Truth label and the Model prediction, both of which are correctly

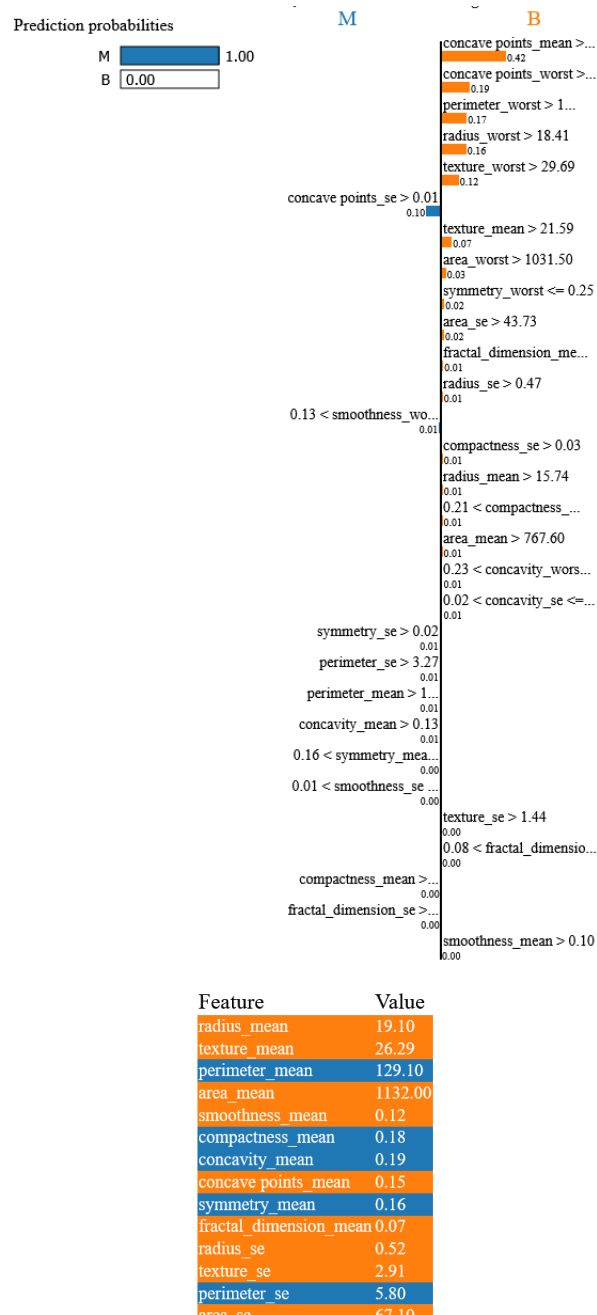


Figure 5.85: Explanation of Malignant level by LIME (Case 02)

classified as 0.0. The explanation then utilizes a force plot with a base value of 0.9932 and an $f(x)$ value of -7.86. The force plot visually represents the contributions of different features to the prediction, highlighting the direction of their impact.

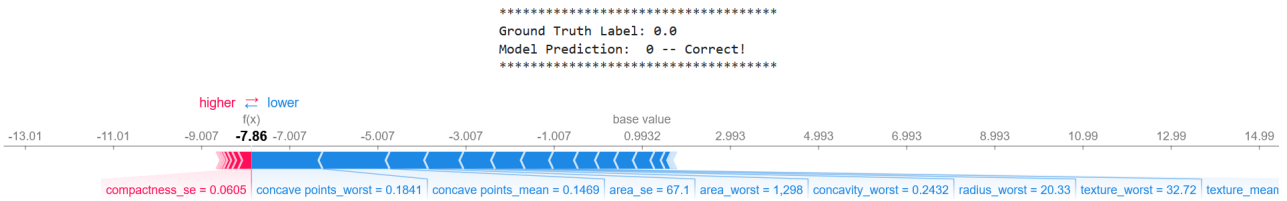


Figure 5.86: Explanation of Malignant level by SHAP (Case 02)

On the “*higher* – $>$ ” side of the force plot, marked in red, we observe the following features: compactness_se, compactness_mean, fractal_dimension_mean, smoothness_se, and texture_se. These features contribute positively to the prediction, indicating that higher values in these features increase the likelihood of the predicted class.

On the “ $<$ – *lower*” side of the force plot, marked in blue, we see the following features: concave_points_worst, concave_points_mean, area_se, area_worst, concavity_worst, radius_worst, and texture_worst. These features contribute negatively to the prediction, implying that lower values in these features increase the likelihood of the predicted class.

The SHAP explanation provides valuable insights into the feature contributions, allowing us to understand which features drive the model’s prediction towards a specific class. By analyzing the force plot, we can gain a better understanding of the relationship between the features and the prediction, enhancing interpretability and transparency.

2. Benign Explanations

Case 03

In Figure 5.87, the actual feature values of the data point are presented. It is then revealed that the actual and predicted diagnosis for this data point is 1.0, indicating a benign classification. The explanation for this prediction is provided as follows:

- The TOPSIS score is calculated based on the similarity of feature values between the given data point and similar instances.
- The features that positively contribute to the prediction are listed, including area_se, perimeter_se, texture_mean, concavity_worst, compactness_worst, concavity_mean, radius_se, concave_points_worst, compactness_mean, concavity_se, concave_points_mean, compactness_se, fractal_dimension_worst, perimeter_mean, concave_points_se, symmetry_mean, texture_se, and fractal_dimension_se.
- On the other hand, there are features that negatively contribute to the prediction, including area_worst, area_mean, texture_worst, radius_worst, radius_mean, perimeter_worst, symmetry_worst, smoothness_worst, smoothness_mean, fractal_dimension_mean, smoothness_se, and symmetry_se.

Figure 5.88 presents the normalized feature contributions, where positive contributions are depicted in green, and negative contributions are shown in red. It is observed that most of the normalized feature contributions are close to 0.

```

----- TOPSIS Explanation -----
Data point index: 394
----- Data Point Features -----
radius_mean          12.100000
texture_mean         17.720000
perimeter_mean       78.070000
area_mean            446.200000
smoothness_mean      0.102900
compactness_mean     0.097580
concavity_mean       0.047830
concave_points_mean  0.033260
symmetry_mean        0.193700
fractal_dimension_mean 0.061610
radius_se            0.284100
texture_se           1.652000
perimeter_se         1.869000
area_se              22.220000
smoothness_se        0.008146
compactness_se       0.016310
concavity_se         0.018430
concave_points_se    0.007513
symmetry_se          0.020150
fractal_dimension_se 0.001798
radius_worst         13.560000
texture_worst        25.800000
perimeter_worst      88.330000
area_worst           559.500000
smoothness_worst     0.143200
compactness_worst    0.177300
concavity_worst      0.160300
concave_points_worst 0.062660
symmetry_worst       0.304900
fractal_dimension_worst 0.070810
Name: 394, dtype: float64
Actual Diagnosis: 1.0
Predicted Diagnosis: 1.0

Explanation:
- The TOPSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  * area_se
  * perimeter_se
  * texture_mean
  * concavity_worst
  * compactness_worst
  * concavity_mean
  * radius_se
  * concave_points_worst
  * compactness_mean
  * concavity_se
  * concave_points_mean
  * compactness_se
  * fractal_dimension_worst
  * perimeter_mean
  * concave_points_se
  * symmetry_mean
  * texture_se
  * fractal_dimension_se
- The features that contribute negatively to the prediction are:
  * area_worst
  * area_mean
  * texture_worst
  * radius_worst
  * radius_mean
  * perimeter_worst
  * symmetry_worst
  * smoothness_worst
  * smoothness_mean
  * fractal_dimension_mean
  * smoothness_se
  * symmetry_se

```

Figure 5.87: Explanation of Benign level by XOPSIS (Case 03)

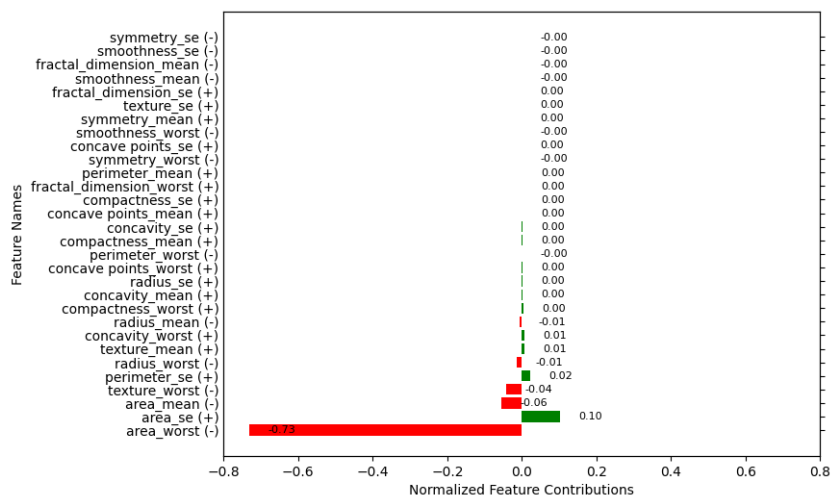


Figure 5.88: Normalized Feature Contributions of Benign level by XOPSIS (Case 03)

This XOPSSIS explanation sheds light on the factors influencing the prediction of a benign diagnosis for the given data point, providing valuable insights into the decision-making process.

The LIME explanation for the given data point starts by presenting the prediction probabilities in Figure 5.89, where the model predicts the probability of M (malignant) as 0.0 and the probability of B (benign) as 1.0.

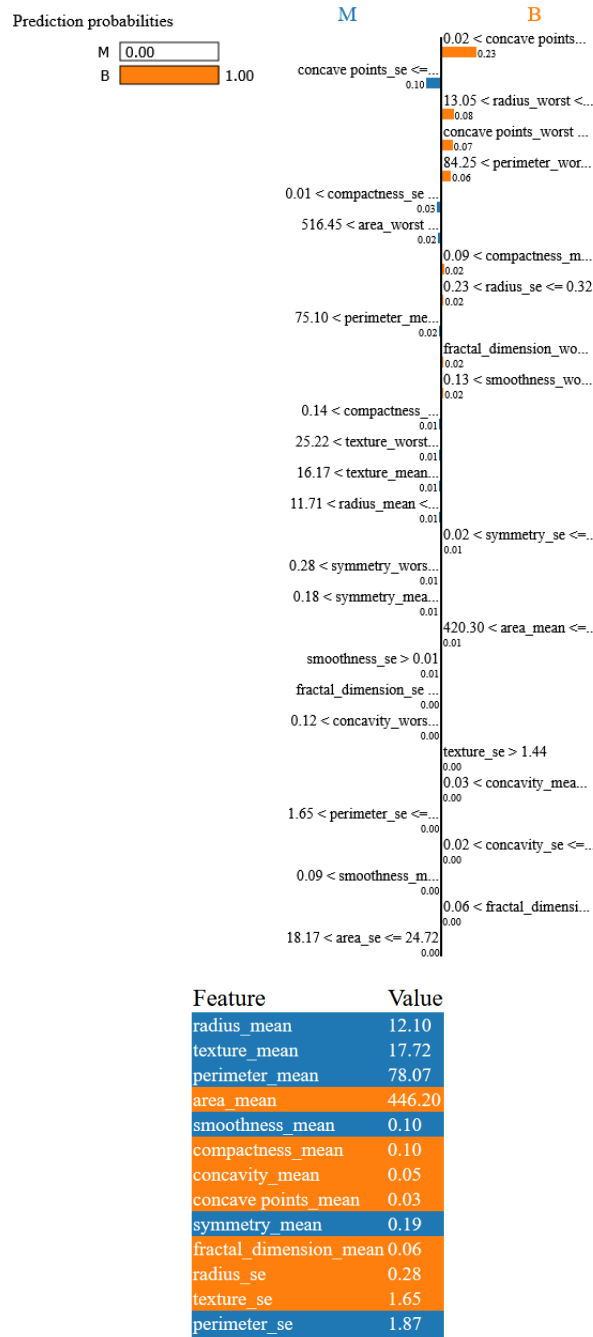


Figure 5.89: Explanation of Benign level by LIME (Case 03)

A plot is then displayed, with the left side representing the explanation for the M class and the right side representing the explanation for the B class. In the M class (left side), the following observations are made:

- concave_points_se is less than or equal to a certain threshold (0.10).
- compact-

ness_se is greater than a specific value (0.01). - area_worst is greater than a certain threshold (516.45). - perimeter_mean falls within a specific range (75.10). - compactness_mean and texture_worst satisfy certain conditions. - Other features such as texture_mean, radius_mean, symmetry_worst, symmetry_mean, smoothness_se, fractal_dimension_se, concavity_worst, perimeter_se, smoothness_mean, and area_se also contribute to the explanation.

In the B class (right side), the following observations are made: - concave_points_mean falls within a certain range (0.02). - radius_worst falls within a specific range. - concave_points_worst is present. - perimeter_worst falls within a specific range (84.25). - compactness_mean, radius_se, fractal_dimension_worst, smoothness_worst, symmetry_se, area_mean, texture_se, concavity_mean, concavity_se, and fractal_dimension fall within certain thresholds.

Finally, the actual feature values of the data point are shown, completing the LIME explanation.

This LIME explanation provides insights into the important features and their values that contribute to the model’s prediction for the given data point, offering transparency and interpretability for the decision-making process.

Figure 5.90 displays the SHAP explanation also mentions the ground truth label and the model’s prediction, which is 1.0 (correct). This indicates that the model accurately classified the data point as the predicted class.

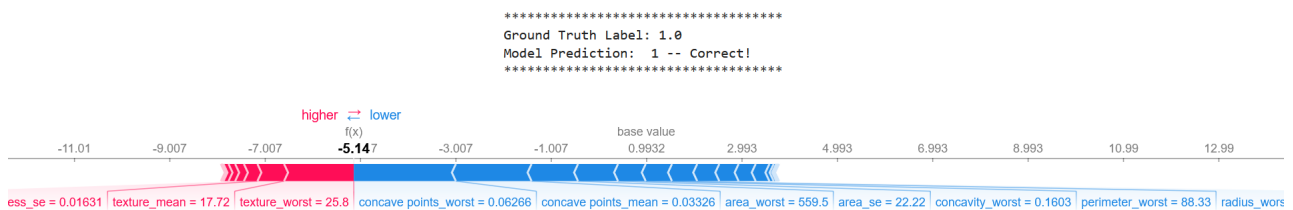


Figure 5.90: Explanation of Benign level by SHAP (Case 03)

The SHAP explanation for the given data point begins by showing the base value, which is 0.9932, and the value of $f(x)$, which is observed to be -5.14. On top of the plot, there is a label that reads “higher – > < – lower”. The text “higher – >” is displayed in red color, while “< – lower” is shown in blue color.

In the “higher – >” part of the explanation, the following features are highlighted: texture_worst, texture_mean, symmetry_se, compactness_mean, fractal_dimension_mean. These features indicate higher values that contribute positively to the prediction for the given data point.

In the “< – lower” part of the explanation, the following features are highlighted: concave_points_worst, concave_points_mean, area_worst, area_se, concavity_worst, perimeter_worst.

These features indicate lower values that contribute negatively to the prediction for the given data point.

The SHAP explanation provides insights into the individual feature contributions and their impact on the model’s prediction. It helps to understand which features have a positive or negative effect on the predicted outcome and contributes to the interpretability of the model’s decision-making process.

Case 04

In Figure 5.91, we begin with the actual feature values of the data point. The actual and predicted diagnosis is provided as 1.0, indicating a benign condition. The explanation of the XOPSIS method for this data point is as follows:

```
----- TOPSIS Explanation -----
Data point index: 551
----- Data Point Features -----
radius_mean      11.130000
texture_mean     22.440000
perimeter_mean   71.490000
area_mean        378.400000
smoothness_mean  0.095660
compactness_mean 0.081940
concavity_mean   0.048240
concave_points_mean 0.022570
symmetry_mean    0.203000
fractal_dimension_mean 0.065520
radius_se        0.280000
texture_se       1.467000
perimeter_se     1.994000
area_se          17.850000
smoothness_se    0.003495
compactness_se   0.030510
concavity_se     0.034450
concave_points_se 0.010240
symmetry_se      0.029120
fractal_dimension_se 0.004723
radius_worst     12.020000
texture_worst    28.260000
perimeter_worst  77.800000
area_worst       436.600000
smoothness_worst 0.108700
compactness_worst 0.178200
concavity_worst  0.156400
concave_points_worst 0.064130
symmetry_worst   0.316900
fractal_dimension_worst 0.080320
Name: 551, dtype: float64
Actual Diagnosis: 1.0
Predicted Diagnosis: 1.0

Explanation:
- The TOPSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  * area_worst
  * area_mean
  * perimeter_worst
  * area_se
  * perimeter_mean
  * radius_worst
  * radius_mean
  * perimeter_se
  * concavity_worst
  * compactness_worst
  * concavity_mean
  * radius_se
  * concave_points_worst
  * compactness_mean
  * concave_points_mean
  * concavity_se
  * smoothness_worst
  * compactness_se
  * fractal_dimension_worst
  * concave_points_se
  * smoothness_mean
  * fractal_dimension_se
  * symmetry_se
- The features that contribute negatively to the prediction are:
  * texture_worst
  * texture_mean
  * symmetry_worst
  * symmetry_mean
  * smoothness_se
  * texture_se
  * fractal_dimension_mean
```

Figure 5.91: Explanation of Benign level by XOPSIS (Case 04)

- The TOPSIS score is calculated based on the similarity of feature values between the given data point and similar instances.
- The features that contribute positively to the prediction of a benign diagnosis are listed, including area_worst, area_mean, perimeter_worst, area_se, perimeter_mean, radius_worst, radius_mean, and several others. These features have a positive impact on the model's prediction.
- On the other hand, there are features that contribute negatively to the prediction, such as texture_worst, texture_mean, symmetry_worst, and others. These features have a negative impact on the model's prediction.

In Figure 5.92, we visualize the normalized feature contributions. Positive feature contributions are represented in green, indicating their positive impact on the prediction, while negative feature contributions are shown in red, indicating their negative impact. It is worth noting that most of the normalized feature contributions are close to 0, indicating their limited influence on the prediction.

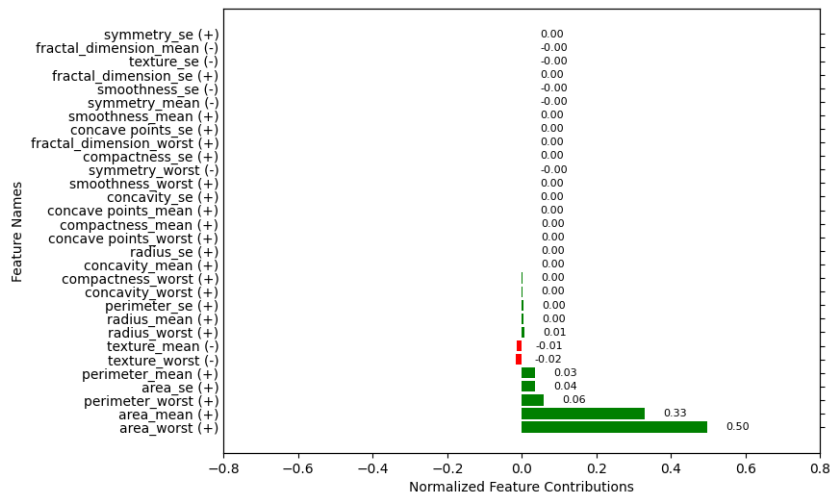


Figure 5.92: Normalized Feature Contributions of Benign level by XOPSSIS (Case 04)

Overall, the XOPSSIS explanation provides insights into the contribution of different features in predicting a benign diagnosis for the given data point. It helps us understand the important factors considered by the model and their relative influence.

In the LIME explanation for the given data point in Figure 5.93, we begin with the prediction probabilities, which indicate a high confidence in the benign diagnosis (B) with a probability of 1.0 and a negligible probability for the malignant diagnosis (M).

The explanation is presented through a plot, with the left side representing the features contributing to the malignant prediction (M) and the right side representing the features contributing to the benign prediction (B). Here are the key findings:

In the left side (M):

- Texture_mean is identified as a contributing factor, with a value greater than 21.59.
- Concave_points.se is another influential feature, with a value greater than 0.01.
- Other features such as perimeter_se, smoothness_mean, symmetry_worst, and more are also considered in the malignant prediction, but to a lesser extent.

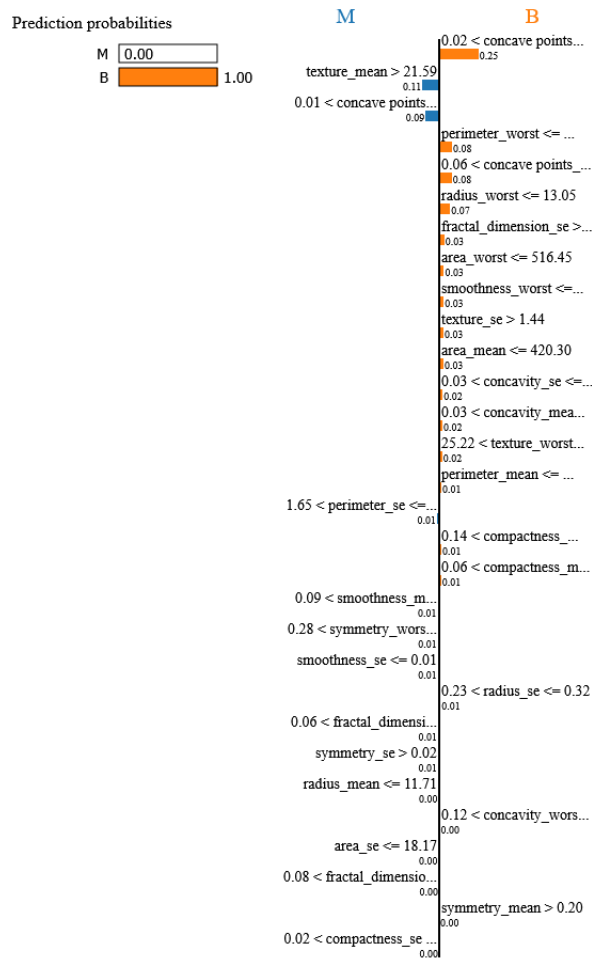
In the right side (B): - Concave_points_mean plays a significant role in the benign prediction, with a value greater than 0.02. - Other important features include perimeter_worst, radius_worst, fractal_dimension.se, and more, which contribute to the likelihood of a benign diagnosis.

The actual feature values of the data point are provided, allowing for a comprehensive understanding of the data point's characteristics.

The LIME explanation helps us identify the specific features that contribute to the model's prediction of a benign diagnosis. It provides insights into the importance and directionality of these features, allowing us to interpret and validate the model's decision-making process.

The SHAP explanation in Figure 5.94 for the given data point starts by stating the ground truth label and the model's correct prediction, which in this case is a benign diagnosis with a confidence of 1.0.

The explanation is provided through a force plot, which showcases the base value of 0.9932 and $f(x)$ value of -7.31. On top of the plot, the text "higher- >< -lower" is displayed, with "higher- >" highlighted in red color and "< -lower" in blue color.



Feature	Value
radius_mean	11.13
texture_mean	22.44
perimeter_mean	71.49
area_mean	378.40
smoothness_mean	0.10
compactness_mean	0.08
concavity_mean	0.05
concave points_mean	0.02
symmetry_mean	0.20
fractal dimension_mean	0.07
radius_se	0.28
texture_se	1.47
perimeter_se	1.99

Figure 5.93: Explanation of Benign level by LIME (Case 04)

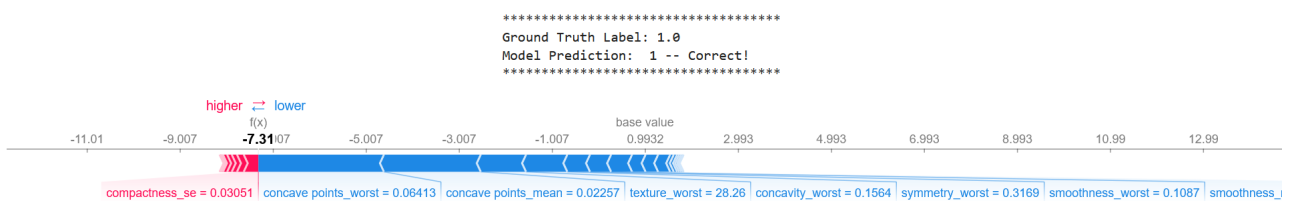


Figure 5.94: Explanation of Benign level by SHAP (Case 04)

In the “*higher* – >” section, the following features are identified as having a positive impact on the prediction: Compactness_se, Compactness_mean, Fractal_dimension_mean, Area_mean, Smoothness_se.

These features contribute to a higher likelihood of a benign diagnosis. Conversely, in the “< *lower*” section, the following features are highlighted as having a negative impact on the prediction: Concave_points_worst, Concave_points_mean, Texture_worst, Concavity_worst, Symmetry_worst, Smoothness_worst.

These features are associated with a lower likelihood of a benign diagnosis.

The SHAP explanation provides valuable insights into the contribution of individual features towards the model’s prediction. By visually representing the positive and negative influences of each feature, it enhances our understanding of the decision-making process.

In conclusion, the explanations provided by XOP-SIS, LIME, and SHAP offer valuable insights into the predictions made by the model for a specific data point. XOP-SIS uses TOPSIS scores to identify features that positively or negatively contribute to the prediction. LIME highlights features that have higher or lower values, indicating their influence on the prediction. SHAP provides a comprehensive view of feature contributions through a force plot. These explanations enhance our understanding of the model’s decision-making process and shed light on the importance of different features in making accurate predictions. By examining these explanations, we can gain valuable insights into how the model arrives at its predictions and improve our trust and interpretability of the model’s outputs.

5.5 Result Analysis on Car Acceptability Dataset

5.5.1 Accuracy

The accuracy of the trained models was evaluated using two different approaches: train-test split and k-fold cross-validation. The results are summarized in Table 5.7.

Serial No	Algorithms	Accuracy	
		Train-Test Split	K-Fold Cross Validation
01	Gradient Boosting	98.266%	87.447%
02	XGBoost	97.688%	88.021%
03	Random Forest	95.954%	87.276%

Table 5.7: Accuracy of 3 Algorithms

From the train-test split analysis, the Gradient Boosting algorithm achieved an accuracy of 98.266%, followed by XGBoost with an accuracy of 97.688%, and Random Forest with an accuracy of 95.954%. These accuracies represent the models’ performance in predicting car acceptability on the test set. It is worth noting that all three algorithms performed remarkably well, demonstrating their ability to capture the underlying patterns and make accurate predictions.

Furthermore, the k-fold cross-validation accuracy was evaluated for each algorithm. The results indicated that the Gradient Boosting algorithm achieved an accuracy of

87.447%, followed by XGBoost with an accuracy of 88.021%, and Random Forest with an accuracy of 87.276%. These accuracies reflect the models' performance in predicting car acceptability across different folds of the dataset. The consistency of accuracies across the folds demonstrates the robustness of the models in generalizing to unseen data.

To provide a visual comparison of the train-test split accuracy, we created a barplot (Figure 5.95) showcasing the performance of each algorithm. The x-axis represents the different algorithms (Gradient Boosting, XGBoost, and Random Forest), while the y-axis represents the accuracy. The barplot clearly illustrates the superior performance of the Gradient Boosting algorithm, followed closely by XGBoost, and then Random Forest.

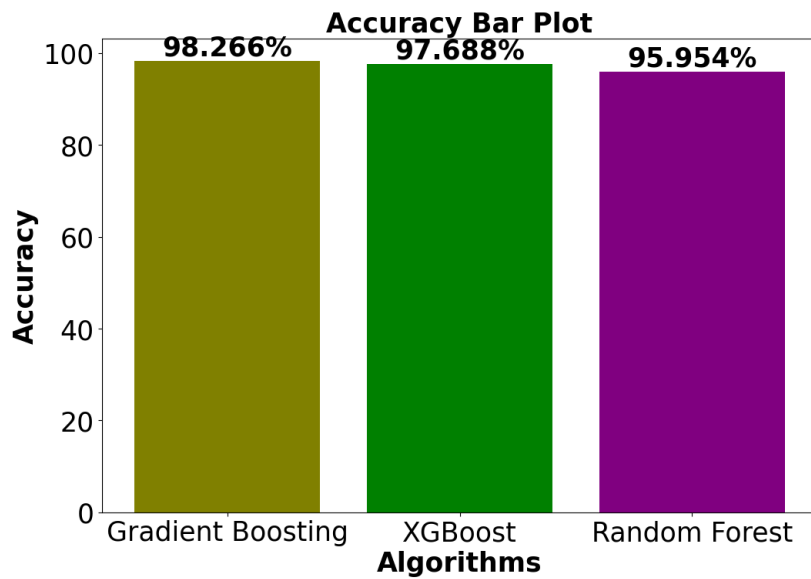


Figure 5.95: Train-Test Split Accuracy for Different Algorithms on Car Acceptability Dataset

The high accuracies obtained by the trained models in both the train-test split and k-fold cross-validation analyses indicate their effectiveness in predicting car acceptability based on the provided features. The superior performance of the Gradient Boosting and XGBoost algorithms highlights their potential for accurate car acceptability classification, while Random Forest also exhibits competitive performance. In summary, the accuracy analysis of the trained models showcases their ability to accurately predict car acceptability. The Gradient Boosting and XGBoost algorithms demonstrate superior performance, while Random Forest also exhibits commendable accuracy. These results highlight the potential of these algorithms for practical applications in car acceptability classification tasks.

5.5.2 Confusion Matrix

The confusion matrix provides detailed insights into the performance of the trained models in predicting car acceptability. The matrix represents the relationship between the true labels and the predicted labels. In our analysis, the confusion matrix is presented in Figure 5.96.

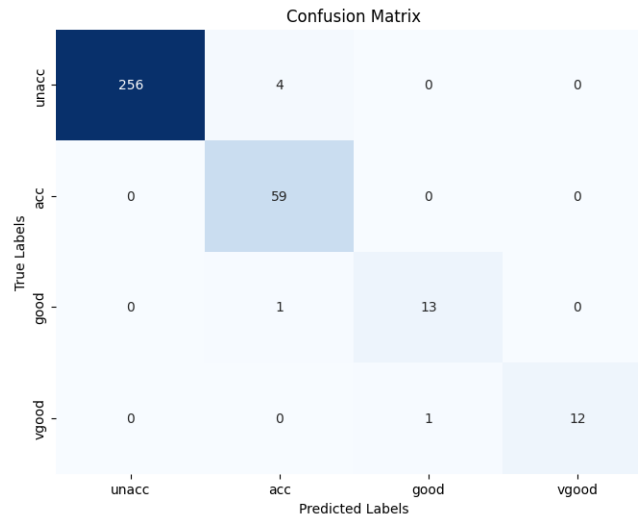


Figure 5.96: Confusion Matrix for Car Acceptability Dataset

In the confusion matrix, the rows represent the true labels, while the columns represent the predicted labels. The four car acceptability categories, namely “unacc” (unacceptable), “acc” (acceptable), “good”, and “vgood” (very good), are arranged in the same order both vertically and horizontally.

The values within the matrix indicate the number of instances that fall into each category. For example, in the first row, the model correctly predicted 256 instances as “unacc” (unacceptable) while misclassifying 4 instances as “acc” (acceptable). Similarly, in the second row, all instances were correctly classified as “acc”, resulting in 59 accurate predictions. The remaining rows follow a similar pattern, reflecting the model’s performance for each car acceptability category.

The confusion matrix provides a comprehensive overview of the model’s performance, allowing us to assess its accuracy and identify any potential misclassifications. The diagonal elements represent the true positive predictions, indicating instances that were correctly classified. Off-diagonal elements represent misclassifications or false positive predictions.

By analyzing the confusion matrix, we can evaluate the model’s performance for each car acceptability category and identify any imbalances or discrepancies in the predictions. These insights can guide further improvements in the model or aid in making informed decisions based on the specific requirements of car acceptability classification.

In summary, the confusion matrix presents a detailed breakdown of the model’s predictions for car acceptability. It allows us to assess the accuracy and performance of the trained models across different car acceptability categories, providing valuable insights for further analysis and evaluation.

5.5.3 Sensitivity, Specificity, Precision and F1-Score

To comprehensively evaluate the performance of the trained models on the car acceptability dataset, we calculated several performance metrics, including sensitivity, specificity, precision, and F1-score. Table 5.8 presents the performance evaluation metrics for each car acceptability class.

The sensitivity metric measures the ability of the model to correctly identify positive

	Sensitivity	Specificity	Precision	F1-Score
unacc	0.985	1.0	1.0	0.99
acc	1.0	0.983	0.92	0.96
good	0.929	0.997	0.93	0.93
vgood	0.923	1.0	1.0	0.96

Table 5.8: Performance Evaluation Metrics on Car Acceptability Dataset

instances, while specificity measures the model’s ability to correctly identify negative instances. Precision indicates the proportion of correctly predicted positive instances out of all instances predicted as positive. The F1-score provides a balanced measure of precision and recall, considering both the false positive and false negative rates.

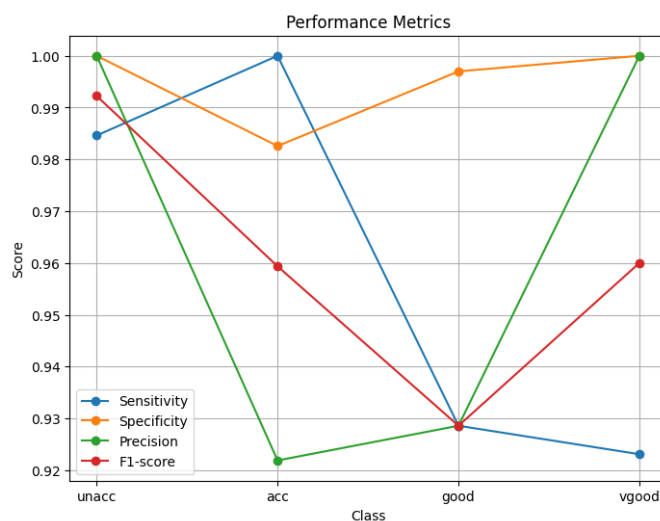


Figure 5.97: Performance Evaluation Metrics on Car Acceptability Dataset

In Figure 5.97, we visualize the performance metrics for each car acceptability class on the same plot, allowing for easy comparison. The x-axis represents the car acceptability classes, namely “unacc”, “acc”, “good”, and “vgood”, while the y-axis represents the performance scores for sensitivity, specificity, precision, and F1-score. Each class is represented by a line in the plot, enabling a visual assessment of the variations in the metrics across different car acceptability categories.

By examining the performance metrics and the corresponding plot, we can gain insights into the strengths and weaknesses of the models in classifying car acceptability. The high sensitivity values across all classes indicate the models’ ability to correctly identify positive instances. The specificity values demonstrate the models’ capacity to accurately identify negative instances. The precision scores indicate the models’ effectiveness in correctly predicting positive instances, while the F1-scores provide a balanced evaluation of precision and recall.

In summary, the sensitivity, specificity, precision, and F1-score metrics provide a comprehensive evaluation of the models’ performance on the car acceptability dataset. The visual representation of these metrics in the plot allows for easy comparison and interpretation of the models’ capabilities in classifying different car acceptability categories. These performance metrics contribute to a deeper un-

derstanding of the models' performance and can guide further improvements or decision-making processes in the context of car acceptability classification.

5.5.4 ROC Curve

The ROC curves offer a comprehensive view of the Gradient Boosting model's classification performance on the Car Acceptability dataset. Figure 5.98 depicts the ROC plot, which adheres to the familiar pattern seen before. However, in this instance, the plot showcases four distinct curves, each representing a different class within the dataset. These classes include unacc (unacceptable), acc (acceptable), good, and vgood (very good). The plot is structured with the x-axis representing the False Positive Rate and the y-axis representing the True Positive Rate. The presence of multiple curves emphasizes the model's ability to distinguish between the various car acceptability categories based on the chosen classification threshold.

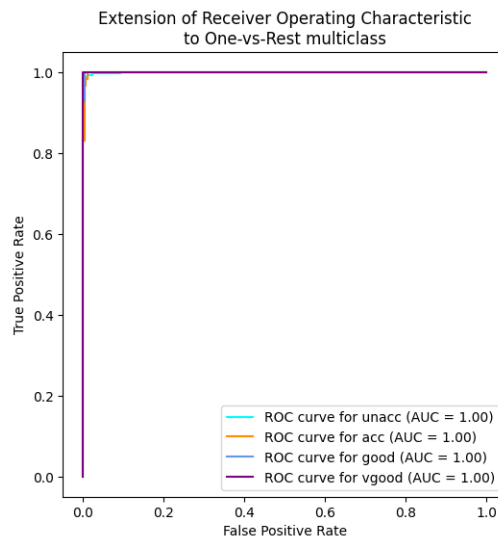


Figure 5.98: ROC Curve for Gradient Boosting Model on Car Acceptability Dataset

The ROC curve provides valuable insights into the performance of the Gradient Boosting algorithm across different car acceptability classes. By analyzing the curve, we can observe the balance between the true positive rate and the false positive rate for each class, allowing us to assess the algorithm's ability to discriminate between the classes. The shape and position of the curve reflect the algorithm's overall classification performance and its ability to achieve high true positive rates while minimizing false positive rates.

In summary, the ROC curves demonstrated that Gradient Boosting algorithm performed well in accurately classifying instances across the car acceptability classes. The curves for each class showed varying degrees of discriminative ability, with low false positive rates and high true positive rates observed. By analyzing the ROC curves, we gained a comprehensive understanding of the discrimination capabilities of the algorithm, enabling us to assess their suitability for car acceptability classification tasks. The ROC curves serve as a visual representation of the algorithm's performance and provide valuable information for selecting the most appropriate algorithm for car acceptability prediction.

5.5.5 AUC

The AUC (Area Under the Curve) is a metric used to assess the overall performance of a classifier algorithm by measuring the area under the Receiver Operating Characteristic (ROC) curve. In this subsection, we present the AUC values for each car acceptability class: unacc, acc, good, and vgood.

Figure 5.99, housing the AUC scores, serves as a complementary analysis to the ROC curves. The AUC (Area Under the Curve) scores offer insights into the model's overall discriminatory power across different classes. For the Car Acceptability dataset, the AUC scores are as follows: unacc (0.9995), acc (0.9992), good (0.9998), and vgood (1.0). These exceptional AUC scores signify an outstanding ability of the Gradient Boosting model to accurately classify the car acceptability levels. Particularly noteworthy is the perfect AUC score of 1.0 achieved for the vgood class, indicating a flawless distinction between the “very good” cars and others. The consistently high AUC scores across all classes underscore the model's remarkable performance in classifying car acceptability with a high degree of precision.

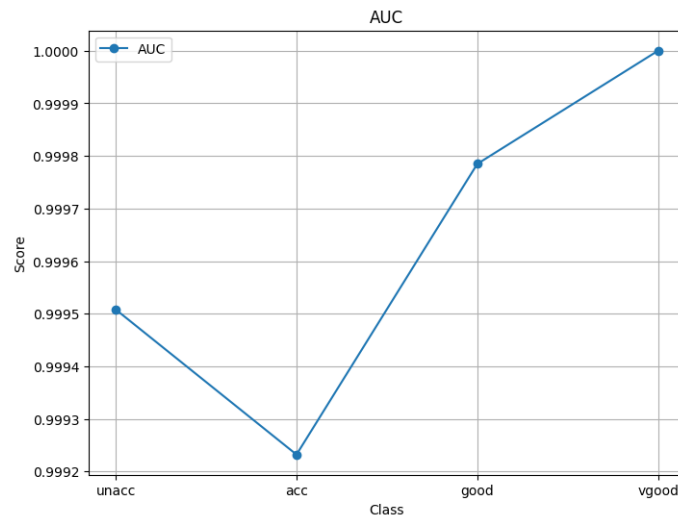


Figure 5.99: AUC Curve for Gradient Boosting Model on Car Acceptability Dataset

The AUC values provide a quantitative measure of the algorithm's ability to distinguish between different car acceptability classes. A higher AUC value indicates better discrimination and classification performance, with values closer to 1 representing a more accurate and reliable classifier.

For the unacc class, the AUC value was calculated as 0.9995, indicating that the algorithm achieved a high level of discrimination performance in distinguishing unacc instances from other classes.

For the acc class, the AUC value was determined to be 0.9992, suggesting that the algorithm exhibited good discrimination capabilities in correctly classifying acc instances.

The good class achieved an AUC value of 0.9998, indicating a very high level of discrimination accuracy for identifying instances belonging to the good class.

Lastly, the vgood class obtained the highest possible AUC value of 1.0, demonstrating excellent discrimination capabilities and accurate classification of vgood instances.

The AUC curve, representing the AUC values for each car acceptability class, exhibits a linear trend from the unacc class to the vgood class. This trend illustrates the varying levels of discrimination and classification performance across the different car acceptability categories.

The AUC values provide valuable insights into the overall performance of the classifier algorithm in distinguishing between car acceptability classes. Higher AUC values indicate a stronger classifier with better discrimination capabilities, while lower AUC values suggest the need for further improvements in classification accuracy.

In summary, the ROC and AUC analyses highlight the effectiveness of the Gradient Boosting model in differentiating between different car acceptability categories within the Car Acceptability dataset. The presence of multiple ROC curves emphasizes the model's versatility in handling the multi-class nature of the problem. The exceptionally high AUC scores further reinforce the model's robustness and reliability in making accurate classifications across all car acceptability levels, which has significant implications for decision-making in the automotive domain.

5.5.6 Interpretability Through XAI Methods

In this subsection, we provide explanations for the predictions made by the Gradient Boosting model on the Car Acceptability dataset using three XAI (Explainable Artificial Intelligence) methods: XOPSSIS, LIME, and SHAP. These methods aim to shed light on the important features that contribute to the model's decision-making process.

1. Unacc Explanation

Case 01

In the XOPSSIS explanation for a specific data point, Figure 5.100 presents the actual feature values. It showcases the features of the data point and provides insights into the actual values recorded. Additionally, it displays the predicted Acceptability Level, which in this case is 0.0 (Unacc).

```

----- TOPSSIS Explanation -----
Data point index: 344
----- Data Point Features -----
Buying_Price      3.0
Maintenance_Price 0.0
No_of_Doors       0.0
Person_Capacity   2.0
Size_of_Luggage   0.0
Safety            2.0
Name: 344, dtype: float64
Actual Acceptability Level: 0.0
Predicted Acceptability Level: 0.0

Explanation:
- The TOPSSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  • Size_of_Luggage
  • No_of_Doors
  • Buying_Price
- The features that contribute negatively to the prediction are:
  • Maintenance_Price
  • Safety
  • Person_Capacity

```

Figure 5.100: Explanation of Unacc level by XOPSSIS (Case 01)

The explanation proceeds by introducing the TOPSSIS score, which is calculated based on the similarity of feature values between the given data point and similar

instances. It highlights the features that contribute positively to the prediction, including `Size_of_Luggage`, `No_of_Doors`, and `Buying_Price`. These features have a favorable impact on the prediction outcome, suggesting that higher values or specific characteristics in these features increase the likelihood of the car being classified as “Unacc” (Unacceptable).

Conversely, the explanation identifies the features that contribute negatively to the prediction. These features, namely `Maintenance_Price`, `Safety`, and `Person_Capacity`, have a detrimental effect on the prediction outcome. Lower values or certain characteristics in these features are associated with a higher probability of the car being classified as “Unacc.”

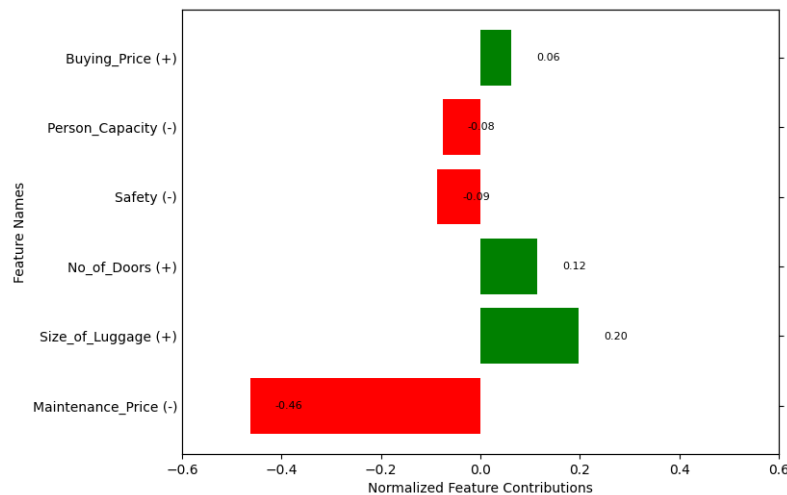


Figure 5.101: Normalized Feature Contributions of Unacc level by XOPSS (Case 01)

Figure 5.101 illustrates the plot of normalized feature contributions. Each feature’s contribution to the prediction is represented by a bar in the plot. The positive contributions, denoted in green, indicate the features’ positive influence on the prediction, while the negative contributions, shown in red, represent the features’ negative impact.

According to the plot, `Buying_Price` has a positive contribution (+0.06), suggesting that higher values or specific characteristics in this feature contribute favorably to the prediction. Similarly, `No_of_Doors` and `Size_of_Luggage` also have positive contributions, indicating their importance in the prediction. On the other hand, `Maintenance_Price` has a notable negative contribution (-0.46), suggesting that lower values or certain characteristics in this feature have a strong negative influence on the prediction outcome. Additionally, `Safety` and `Person_Capacity` also exhibit negative contributions (-0.09 and -0.08, respectively), indicating their adverse effects on the prediction.

The visualization of the normalized feature contributions provides a comprehensive understanding of how each feature contributes to the prediction outcome. The distinct colors (green and red) aid in easily identifying the positive and negative contributions, enabling users to interpret the impact of individual features on the car acceptability classification.

In Figure 5.102, the LIME explanation for the given data point starts by displaying the prediction probabilities for each class. In this case, the prediction probabilities

are as follows: unacc=1.0, acc=0.00, good=0.00, vgood=0.0. This indicates a high confidence in classifying the data point as “unacc” (Unacceptable) and negligible probabilities for the other classes.

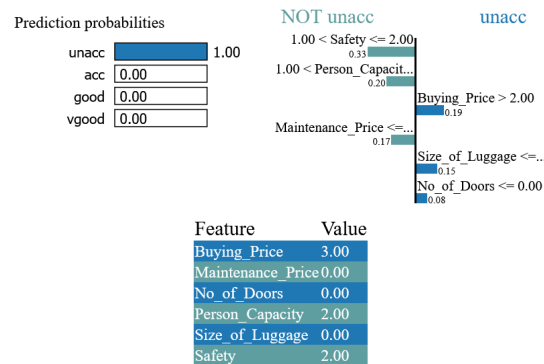


Figure 5.102: Explanation of unacc level by LIME (Case 01)

The explanation proceeds with a plot that is divided into two sections. On the right side of the plot, it focuses on the features that contribute to the classification of “unacc,” while on the left side, it explores the features that indicate a deviation from the “unacc” class.

On the right side of the plot, it is evident that the feature Buying_Price has a positive contribution (0.19), indicating that a value greater than 2.00 contributes to the classification of “unacc.” Similarly, the features Size_of_Luggage (0.15) and No_of_Doors (0.08) also have positive contributions, suggesting that certain threshold values or higher values contribute to the prediction of “unacc” class.

On the left side of the plot, representing the features associated with “Not unacc” classes, all features exhibit negative contributions. The feature Safety has a positive contribution (0.33) when the value falls within the range of 1.00 to 2.00. Likewise, the feature Person_Capacity has a positive contribution (0.20) when the value exceeds 1.00. Furthermore, the feature Maintenance_Price demonstrates a negative contribution (0.17) when the value is below a certain threshold, contributing to the classification of “Not unacc” categories.

The explanation concludes by presenting the actual feature values of the data point. These values provide insight into the specific characteristics of the data point that led to its classification as “unacc” and the contribution of each feature to this prediction. The provided SHAP explanation in Figure 5.103 indicates that the ground truth label for the given data point is 0.0, which corresponds to the “unacc” class. The model’s prediction for this data point is also 0, which aligns with the ground truth label. Thus, the model prediction is correct in classifying the data point as “unacc.”

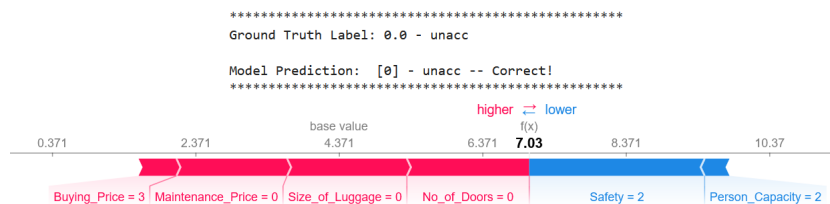


Figure 5.103: Explanation of unacc level by SHAP (Case 01)

The SHAP explanation for the given datapoint reveals the following:

In the accompanying force plot, the base value is 4.371, and the value of $f(x)$ is seen as 7.03. The force plot is divided into two sections, with “higher– >” displayed in red color and “< –lower” displayed in blue color.

In the red-colored “higher– >” section, the features `Buying_Price=3.0`, `Maintenance_Price=0.0`, `Size_of_luggage=0`, and `No_of_doors=0` contribute positively to the prediction. These features indicate that higher values for `Buying_Price`, `Maintenance_Price`, and `Size_of_luggage`, as well as having `No_of_doors` equal to 0, contribute to the classification of “unacc.”

In the blue-colored “< –lower” section, the features `Safety=2.0` and `Person_Capacity=2.0` contribute negatively to the prediction. This suggests that lower values for `Safety` and `Person_Capacity` are associated with the classification of “unacc.”

Overall, the SHAP explanation provides insights into the feature contributions that lead to the prediction of “unacc” for the given data point.

In summary, the explanations provided by XOPSSIS, LIME, and SHAP for the specific data point consistently support the classification of the data point as “unacc.” XOPSSIS and LIME highlight the positive contributions of features such as `Buying_Price`, `Size_of_Luggage`, and `No_of_Doors`, indicating their influence in favor of the “unacc” class. It is evident that both XOPSSIS and LIME provide identical feature contributions for the data point, reinforcing the classification of “unacc.” This consistency in feature importance across XOPSSIS and LIME strengthens the interpretability and trustworthiness of the model’s prediction for the car’s acceptability. The alignment of explanations between XOPSSIS and LIME underscores the agreement in identifying the key features influencing the classification decision, further enhancing our understanding of the model’s decision-making process. SHAP further reinforces these findings, with `Buying_Price`, `Maintenance_Price`, and `Size_of_Luggage` exhibiting positive contributions, and `Safety` and `Person_Capacity` demonstrating negative contributions.

These consistent explanations from multiple XAI methods strengthen our confidence in the classification of the data point as “unacc” and provide valuable insights into the features that contribute to this prediction. Such interpretability aids in understanding the decision-making process of the model and enhances transparency in the classification of car acceptability.

2. Acc Explanation

Case 02

In the XOPSSIS explanation for a new data point, the Figure 5.104 presents the actual feature values. It then displays the actual and predicted Acceptability Level, which in this case is predicted as 1.0 (Acc). The explanation provided by XOPSSIS includes the following details. The features that contribute positively to the prediction of “Acc” are `Buying_Price`, `Maintenance_Price`, `Safety`, and `No_of_Doors`. Their positive contributions indicate that higher values or specific ranges of these features favor the classification of “Acc.”

Conversely, the XOPSSIS analysis identifies `Size_of_Luggage` and `Person_Capacity` as features that negatively contribute to the prediction. This means that lower values or certain thresholds of these features are associated with the classification of “Acc.” Figure 5.105 illustrates the normalized feature contributions. Positive contributions are represented in green, while negative contributions are shown in red. The nor-

```

----- TOPSIS Explanation -----
Data point index: 1172
----- Data Point Features -----
Buying_Price      1.0
Maintenance_Price 1.0
No_of_Doors      3.0
Person_Capacity  1.0
Size_of_Luggage  0.0
Safety           2.0
Name: 1172, dtype: float64
Actual Acceptability Level: 1.0
Predicted Acceptability Level: 1.0

Explanation:
- The TOPSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  * Buying_Price
  * Maintenance_Price
  * Safety
  * No_of_Doors
- The features that contribute negatively to the prediction are:
  * Size_of_Luggage
  * Person_Capacity

```

Figure 5.104: Explanation of Acc level by XOPSIS (Case 02)

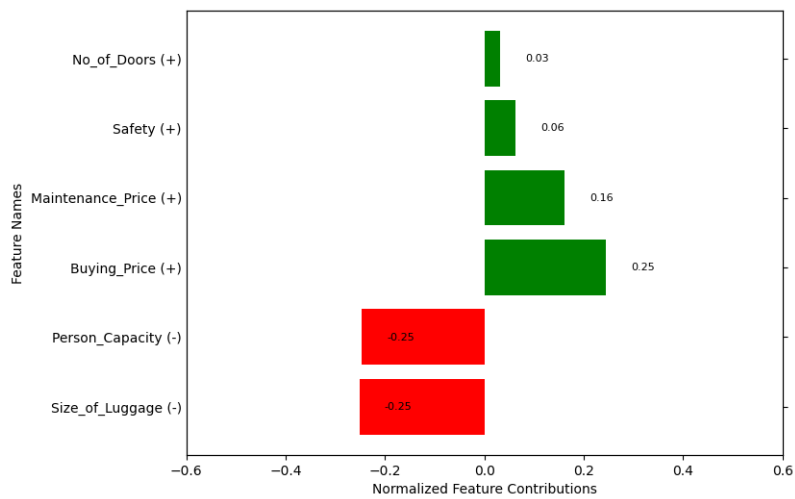


Figure 5.105: Normalized Feature Contributions of Acc level by XOPSIS (Case 02)

malized feature contributions for the data point are as follows: No_of_Doors (+0.03), Safety (+0.06), Maintenance_Price (+0.16), Buying_Price (+0.25), Person_Capacity (-0.25), and Size_of_Luggage (-0.25). These values indicate the degree to which each feature influences the classification of the data point.

By uncovering these specific feature contributions, XOPSIS provides valuable insights into the decision-making process of the model, shedding light on the factors that drive the prediction of “Acc.” This information enhances our understanding of the underlying dynamics of the dataset and reinforces the interpretability of the XOPSIS algorithm.

In Figure 5.106, the LIME explanation for the same data point begins by displaying the prediction probabilities: unacc=0.00, acc=1.00, good=0.00, vgood=0.00. This indicates that the model predicts the data point to belong to the “acc” (Acceptability Level: 1.0) class with a probability of 1.00.

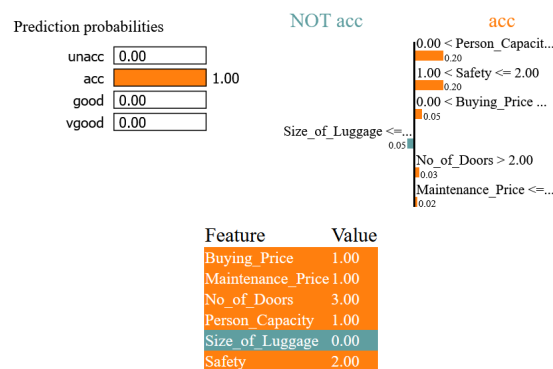


Figure 5.106: Explanation of acc level by LIME (Case 02)

Next, the LIME explanation presents a plot that showcases the feature contributions. On the right side of the plot, it displays the features associated with the “acc” class. These features include Person.Capacity (0.20), Safety (0.20), Buying_Price (0.05), No_of_Doors (0.03), and Maintenance_Price (0.02). These positive contributions suggest that higher values or specific ranges of these features support the classification of “acc.”

On the left side of the plot, it represents the features associated with the “Not acc” classes. In this case, it shows that the Size_of_Luggage feature has a contribution of 0.05. This negative contribution suggests that lower values or certain thresholds of Size_of_Luggage favor the classification of “Not acc.”

Lastly, the LIME explanation provides the actual feature values of the data point, providing a comprehensive understanding of the specific characteristics that contribute to the model’s prediction.

Figure 5.107 displays the SHAP explanation for the same data point begins by indicating the ground truth label as 1.0 - acc (Acceptability Level: 1.0) and the model’s correct prediction of [1] - acc.

Next, it presents a force plot where the base value is -1.534. The $f(x)$ value is observed at 2.37, and on top of the plot, the annotation “higher– >< –lower” is displayed. The text “higher– >” is highlighted in red, while “< –lower” is highlighted in blue.

In the plot, the features Buying_Price, Person_Capacity, Maintenance_Price, and Safety are displayed in red, representing their positive contributions (higher– >).

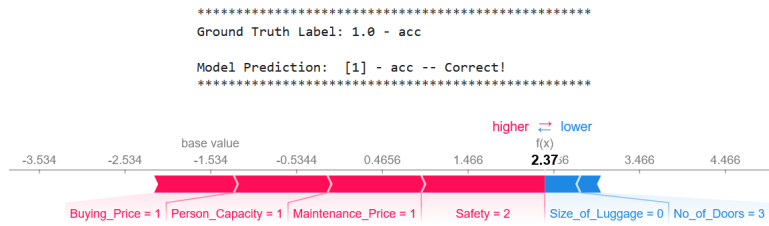


Figure 5.107: Explanation of acc level by SHAP (Case 02)

These features contribute positively to the classification of “acc,” indicating that higher values or specific ranges of these features support the prediction of “acc.” On the other hand, the features Size_of_Luggage and No_of_Doors are shown in blue, representing their negative contributions ($< -$ lower). These features have a negative impact on the prediction of “acc,” suggesting that lower values or specific thresholds of these features favor the classification of “Not acc.”

By visualizing the feature contributions in the force plot, the SHAP explanation provides insights into how each feature influences the model’s prediction for the given data point.

Analyzing the explanations provided by XOPSSIS, LIME, and SHAP for the same data point, it is evident that all three methods consistently highlight the features that contribute to the prediction of “acc” (Acceptability Level: 1.0). According to XOPSSIS, Buying.Price, Maintenance.Price, Safety, and No_of_Doors have positive contributions, indicating that higher values or specific ranges of these features support the classification of “acc.” On the other hand, Size_of_Luggage and Person.Capacity are identified as features with negative contributions, suggesting that lower values or specific thresholds of these features favor the classification of “Not acc” while LIME shows only Size_of_Luggage as negative contributing feature.

Similarly, the SHAP explanation aligns with XOPSSIS and LIME, showcasing the positive contributions of Buying.Price, Person.Capacity, Maintenance.Price, and Safety. Additionally, SHAP identifies Size_of_Luggage and No_of_Doors as features with negative contributions. This consistency across the three explanations enhances our understanding of the influential features and reinforces the model’s prediction for the acceptability level of the car.

The agreement in feature importance between XOPSSIS, LIME, and SHAP provides a robust interpretation of the model’s decision-making process. It instills confidence in the identified features and their impact on the classification outcome, ultimately increasing the transparency and interpretability of the model’s predictions for the given data point.

3. Good Explanation

Case 03

The XOPSSIS explanation for another new data point consistently identifies the features that contribute to the prediction of “Good” (Acceptability Level: 2.0) in Figure 5.108. According to XOPSSIS, the features Safety, Buying.Price, Maintenance.Price, and Person.Capacity have positive contributions, while Size_of_Luggage and No_of_Doors have negative contributions. This insight provides a deeper understanding of the factors influencing the model’s decision and highlights the impor-

```

----- TOPSIS Explanation -----
Data point index: 1667
----- Data Point Features -----
Buying_Price      0.0
Maintenance_Price 0.0
No_of_Doors      1.0
Person_Capacity   2.0
Size_of_Luggage   0.0
Safety            2.0
Name: 1667, dtype: float64
Actual Acceptability Level: 2.0
Predicted Acceptability Level: 2.0

Explanation:
- The TOPSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  • Safety
  • Buying_Price
  • Maintenance_Price
  • Person_Capacity
- The features that contribute negatively to the prediction are:
  • Size_of_Luggage
  • No_of_Doors

```

Figure 5.108: Explanation of Good level by XOPSIS (Case 03)

tance of safety, pricing, maintenance, and passenger capacity in determining the acceptability level of a car. The normalized feature contributions further reinforce the significance of these features in Figure 5.109, with Safety exhibiting the highest positive contribution, followed by Buying_Price and Maintenance_Price. Conversely, Size_of_Luggage has the highest negative contribution, indicating its impact on the prediction of “Good.”

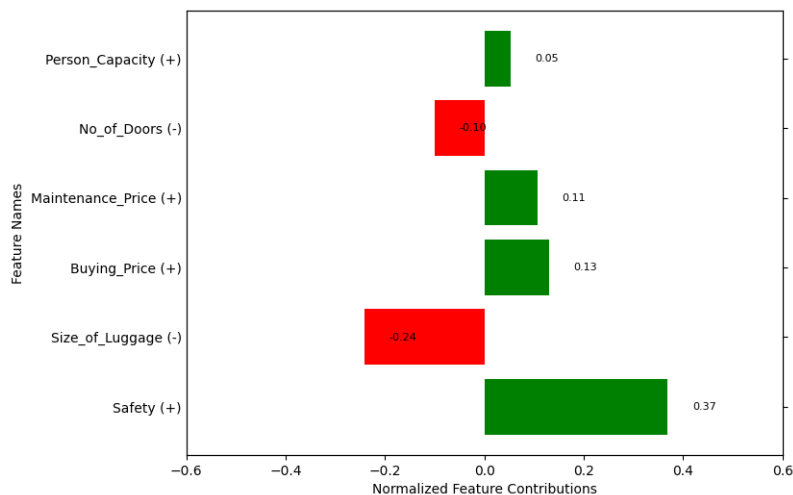


Figure 5.109: Normalized Feature Contributions of Good level by XOPSIS (Case 03)

The LIME explanation in Figure 5.110 for the same data point reveals the feature contributions that support the prediction of “Good” (Acceptability Level: 2.0). LIME assigns probabilities to each class, and in this case, the model predicts a high probability of “Good” while assigning negligible probabilities to other classes. The LIME plot illustrates that certain feature conditions on the right side, such as $\text{Buying_Price} \leq 0.00$, $\text{Maintenance_Price} \leq \dots$, $1.00 < \text{Person_Capacity} \dots$, and $1.00 < \text{Safety} \leq 2.00$, contribute to the classification of “Good.” Conversely, features on the left side, such as $\text{Size_of_Luggage} \leq \dots$, and $0.00 < \text{No_of_Doors} \dots$, are associated with “Not good” categories. The actual feature values of the data point are also provided for reference.

Figure 5.111 represents the SHAP explanation for the data point reveals that the ground truth label and model prediction both correctly identify it as “good.” The

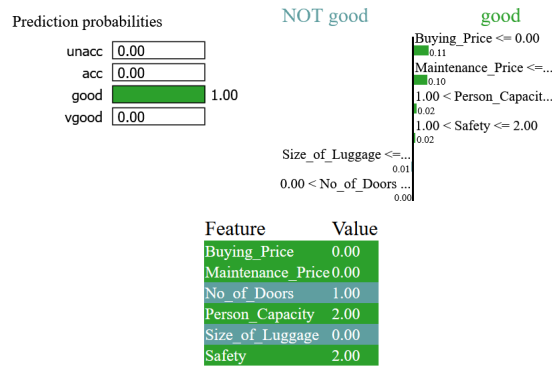


Figure 5.110: Explanation of good level by LIME (Case 03)

force plot demonstrates the feature contributions, with a base value of -4.879 and a $f(x)$ value of -5.76 . The plot is divided into two sections, denoted by “higher $>$ ” in red and “ $<$ -lower” in blue. In the red section, the features `Buying_Price`, `Size_of_luggage`, and `Safety` exhibit positive contributions. On the other hand, the blue section shows negative contributions from the features `Person_Capacity`, `Maintenance_Price`, and `No_of.doors`.

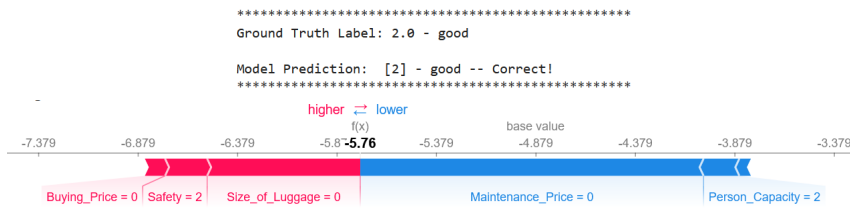


Figure 5.111: Explanation of good level by SHAP (Case 03)

After reviewing the XOPIS, LIME, and SHAP explanations for the same data point, we can observe a consistent pattern in feature importance. All three methods highlight `Safety`, and `Buying_Price` as positive contributors and `No_of.doors` as negative contributors to the prediction of “good” acceptability. However, there is a discrepancy in the treatment of `Size_of_Luggage`. While XOPIS and LIME consider it as a negative contributor, SHAP does not highlight it as a significant factor. This inconsistency suggests that `Size_of_Luggage` may have varying degrees of impact on the model’s decision across different explanation methods. Nonetheless, the consensus on the positive contributions of `Safety`, and `Buying_Price` and negative contributions of `No_of.doors` enhances our understanding of the key features driving the prediction of “good” acceptability.

3. Vgood Explanation

Case 04

In the XOPIS explanation for a new data point, the analysis begins by displaying the actual feature values. Figure 5.112 then reveals the actual and predicted acceptability level, indicating a prediction of “Vgood” (very good). The explanation continues by identifying the features that contribute positively and negatively to the prediction.

```

----- TOPSIS Explanation -----
Data point index: 1481
----- Data Point Features -----
Buying_Price      0.0
Maintenance_Price 2.0
No_of_Doors      2.0
Person_Capacity   2.0
Size_of_Luggage   1.0
Safety            2.0
Name: 1481, dtype: float64
Actual Acceptability Level: 3.0
Predicted Acceptability Level: 3.0

Explanation:
- The TOPSIS score is calculated based on the similarity of feature values
  between the given data point and the similar instances.
- The features that contribute positively to the prediction are:
  • Buying_Price
  • Person_Capacity
  • No_of_Doors
- The features that contribute negatively to the prediction are:
  • Maintenance_Price
  • Size_of_Luggage
  • Safety

```

Figure 5.112: Explanation of Vgood level by XOPSIS (Case 04)

The features that positively contribute to the prediction of “Vgood” are Buying_Price, Person_Capacity, and No_of_Doors. This suggests that higher values or specific thresholds for these features align with a higher likelihood of the car being classified as “Vgood.” On the other hand, the features Maintenance_Price, Size_of_Luggage, and Safety demonstrate negative contributions to the prediction. Lower values or specific thresholds for these features are associated with a higher likelihood of the car being classified as “Vgood.”

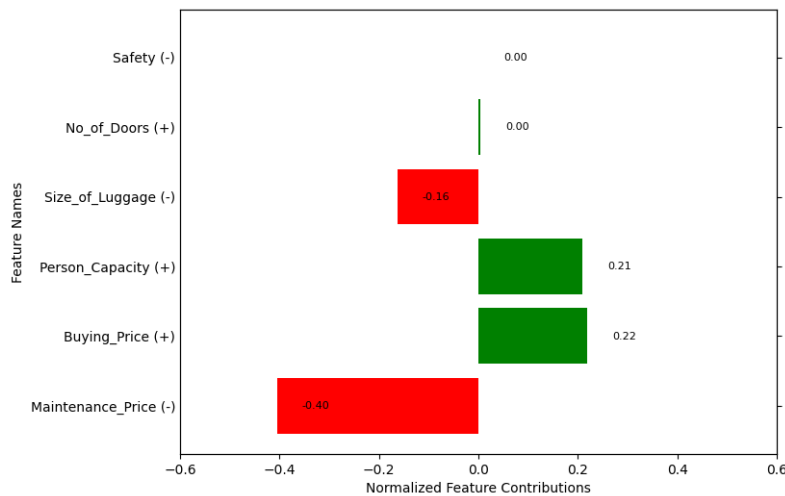


Figure 5.113: Normalized Feature Contributions of Vgood level by XOPSIS (Case 04)

Figure 5.113 presents the normalized feature contributions, visually represented by green and red colors. Safety exhibits a negligible contribution (close to 0), indicating that it does not strongly influence the prediction. No_of_Doors has a neutral contribution (close to 0), implying that its value does not significantly impact the prediction. Size_of_Luggage demonstrates a negative contribution (-0.16), suggesting that values lower than a certain threshold contribute to the classification of “Vgood.” Conversely, Person_Capacity and Buying_Price show positive contributions (0.21 and 0.22, respectively), indicating that higher values or specific thresholds for these features contribute to the prediction. Lastly, Maintenance_Price exhibits a substantial negative contribution (-0.40), suggesting that lower values or specific thresholds

for this feature strongly influence the classification of “Vgood”. Figure 5.114 represents the LIME explanation for the same data point begins by displaying the prediction probabilities, indicating a high confidence prediction of “Vgood” (very good) with a probability of 1.0. The explanation continues with a plot that visually represents the contributions of different features to the prediction.

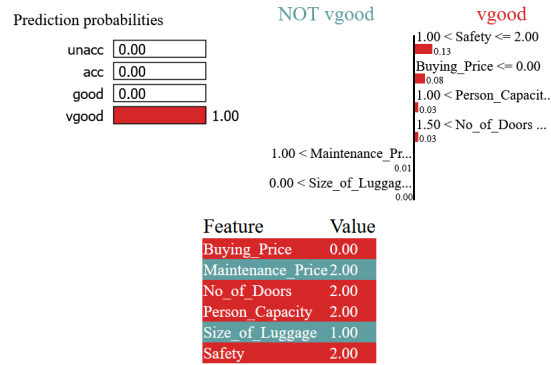


Figure 5.114: Explanation of vgood level by LIME (Case 04)

On the right side of the plot, which corresponds to the “Vgood” class, the features that positively contribute to the prediction are Safety, Buying_Price, Person_Capacity, and No_of_Doors. This suggests that specific thresholds or higher values for these features are indicative of the car being classified as “Vgood.” On the left side of the plot, representing the “Not vgood” class, the feature Maintenance_Price demonstrates a negative contribution, implying that lower values or specific thresholds for this feature are associated with a lower likelihood of the car being classified as “Vgood.” Additionally, Size_of_Luggage exhibits a negligible contribution, indicating that its value does not significantly influence the prediction. The LIME explanation concludes by presenting the actual feature values of the data point, allowing for a comprehensive understanding of the specific characteristics that led to the prediction of “Vgood” for this particular instance.

Figure 5.115 displays the SHAP explanation for the same data point begins by stating the Ground Truth Label and the Model Prediction, both of which correctly identify the class as “Vgood.”

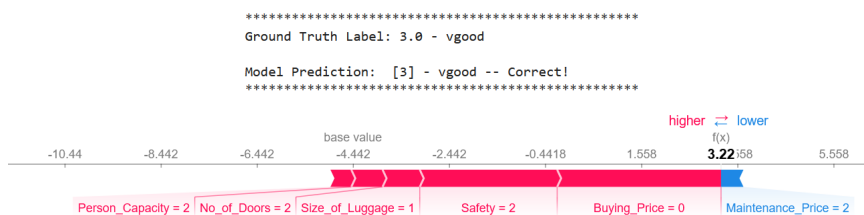


Figure 5.115: Explanation of vgood level by SHAP (Case 04)

The explanation further presents a force plot that visualizes the contributions of different features to the prediction. The base value is indicated as -4.442, and the plot shows that the overall contribution leads to a value of 3.22. The plot is divided into two sections, with “higher >” denoted in red color and “< -lower” denoted in blue color.

On the “higher – >” side, features such as Person_Capacity, Buying_Price, Size_of_Luggage, and Safety demonstrate positive contributions, suggesting that higher values or specific thresholds for these features positively impact the classification as “Vgood.” On the “< – lower” side, the feature Maintenance_Price exhibits a negative contribution, indicating that lower values contribute to the classification as “Vgood.” By visualizing the feature contributions in the force plot, the SHAP explanation provides insights into how each feature influences the prediction of “Vgood” for the given data point.

Upon analyzing the explanations provided by XOPSSIS, LIME, and SHAP for the same data point, it becomes evident that all three methods consistently identify the key features contributing to the prediction of “Vgood” as the acceptability level for the car.

According to XOPSSIS, the features Buying_Price, Person_Capacity, and No_of_Doors positively contribute to the classification, while Maintenance_Price, Size_of_Luggage, and Safety have negative contributions.

Similarly, LIME highlights the importance of Safety, Buying_Price, and Person_Capacity as positive contributors to the prediction of “Vgood.” On the other hand, Size_of_Luggage and Maintenance_Price are identified as negative contributors.

In line with the other methods, SHAP indicates that Person_Capacity, Buying_Price, Size_of_Luggage, Safety and No_of_Doors positively impact the classification of “Vgood,”. Maintenance_Price is identified as a negative contributor.

Overall, the consistency in the identified features across XOPSSIS, LIME, and SHAP explanations reinforces the significance of Buying_Price, Person_Capacity, and Size_of_Luggage as influential factors in predicting the “Vgood” acceptability level for the given data point. The alignment in feature importance enhances our understanding of the model’s decision-making process and strengthens the reliability and interpretability of the predictions.

5.6 Time Complexity Analysis

The time complexity analysis of the XOPSSIS algorithm is crucial to understanding its computational efficiency. The algorithm’s time complexity is evaluated step by step, considering the number of instances (n) and the number of features (m) in the dataset.

1. **Normalization of the Dataset:** The initial step involves normalizing the dataset. This process requires iterating through each instance and feature, resulting in a time complexity of $O(n)$.
2. **Calculating Weighted Normalized Decision Matrix:** Similar to the normalization step, the computation of the weighted normalized decision matrix necessitates iterating through the dataset, contributing to a time complexity of $O(n)$.
3. **Euclidean Distance Calculation:** The calculation of Euclidean distances to ideal and anti-ideal solutions involves iterating through instances and features. Consequently, this step has a time complexity of $O(n * m)$.
4. **TOPSIS Score Determination:** The calculation of the TOPSIS score for each alternative entails iterating through instances, yielding a time complexity of $O(n)$.
5. **Ranking Alternatives:** The ranking process, based on TOPSIS scores, involves sorting the alternatives. This step has a time complexity of $O(n * \log(n))$, where $\log(n)$ arises from the sorting operation.

6. **Instance Filtering:** Filtering instances based on TOPSIS scores and predicted risk level entails iterating through instances, resulting in a time complexity of $O(n)$.
7. **Average Feature Value Calculation:** The calculation of average feature values involves iterating through features, leading to a time complexity of $O(m)$.
8. **Feature Difference Determination:** Determining feature differences requires iterating through features and computing the differences, leading to a time complexity of $O(m)$.
9. **Correlation Estimation:** Estimating the correlation between features and the target variable involves iterating through instances and features, resulting in a time complexity of $O(n * m)$.
10. **Weighted Contribution Evaluation:** Evaluating the weighted contribution of each feature involves iterating through features, leading to a time complexity of $O(m)$.
11. **Feature Classification:** Detecting features with positive and negative contributions entails iterating through features and classifying them, resulting in a time complexity of $O(m)$.
12. **Explanation Printing:** The final step of printing the explanation has a constant time complexity of $O(1)$.

Considering the aforementioned complexities, the overall time complexity of the XOPSIS algorithm predominantly hinges on the steps involving iteration through instances and features. As a result, the algorithm's time complexity can be approximated as $O(n * m)$, where n represents the number of instances and m denotes the number of features.

Chapter 6

Conclusion

In conclusion, this research presents XOP-SIS, a novel explainable AI (XAI) method, and evaluates its performance across multiple datasets, including maternal health, breast cancer, the benchmark Iris dataset, and the Car Acceptability dataset. XOP-SIS demonstrates its effectiveness in providing interpretable insights into predictions, particularly in maternal health risk prediction, iris species prediction, and car acceptability prediction.

Our analysis using XOP-SIS consistently highlighted the importance of specific features in predicting maternal health risks and breast cancer diagnoses. These findings were in line with previous research and demonstrated the ability of XOP-SIS to uncover meaningful insights. The positive and negative feature contributions identified by XOP-SIS provided valuable information about the factors influencing the predictions.

Furthermore, our research revealed the convergence of results between XOP-SIS and established explainable AI methods such as LIME and SHAP. In most cases, we observed exact similar explanations in terms of the positive and negative contributing features for XOP-SIS and LIME. SHAP also showed similar findings, providing additional support to our findings. This convergence not only validated the reliability and interpretability of XOP-SIS but also emphasized the consistency in the importance of features across different explainable AI techniques.

While the evaluation primarily focused on maternal health and breast cancer datasets, the successful application of XOP-SIS to the benchmark Iris dataset and the Car Acceptability dataset suggests its potential for broader applicability across diverse domains. This versatility and effectiveness of XOP-SIS as an XAI method underscore its value in providing interpretable insights.

The availability of limited data for the maternal health dataset may have influenced the generalizability of our findings. Future studies should consider larger and more diverse datasets to validate and enhance the robustness of XOP-SIS in different healthcare contexts.

Our research has significant implications for maternal health care and the advancement of the field of explainable AI. By providing interpretable insights into maternal health risk prediction, XOP-SIS can aid healthcare professionals in making informed decisions and implementing targeted interventions. The consistent and robust findings from our study, particularly the similarities observed with LIME, underscore the importance of utilizing explainable AI methods in healthcare research.

In summary, our research introduces and evaluates the novel explainable AI method,

XOPSIS, showcasing its effectiveness in providing interpretable insights into predictions across multiple datasets. The convergence of results with established methods, such as LIME, and its potential for broader applicability highlight the effectiveness and versatility of XOPSIS in providing interpretable insights. These findings contribute to the advancement of the field of explainable AI and its application in various domains, including maternal health care, breast cancer diagnosis, and beyond.

Bibliography

- [1] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.
- [2] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.
- [3] S. M. Lundberg, G. G. Erion, and S.-I. Lee, “Consistent individualized feature attribution for tree ensembles,” *arXiv preprint arXiv:1802.03888*, 2018.
- [4] M. Craven and J. Shavlik, “Extracting tree-structured representations of trained networks,” *Advances in neural information processing systems*, vol. 8, 1995.
- [5] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, “How to explain individual classification decisions,” *The Journal of Machine Learning Research*, vol. 11, pp. 1803–1831, 2010.
- [6] J. Chen, L. Song, M. Wainwright, and M. Jordan, “Learning to explain: An information-theoretic perspective on model interpretation,” in *International conference on machine learning*, PMLR, 2018, pp. 883–892.
- [7] M. R. Zafar and N. Khan, “Deterministic local interpretable model-agnostic explanations for stable explainability,” *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 525–541, 2021.
- [8] X. Zhao, W. Huang, X. Huang, V. Robu, and D. Flynn, “Baylime: Bayesian local interpretable model-agnostic explanations,” in *Uncertainty in artificial intelligence*, PMLR, 2021, pp. 887–896.
- [9] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, “Explanation of machine learning models using improved shapley additive explanation,” in *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2019, pp. 546–546.
- [10] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.
- [11] C.-H. Yeh, “A problem-based selection of multi-attribute decision-making methods,” *International Transactions in Operational Research*, vol. 9, no. 2, pp. 169–181, 2002.
- [12] S. Zeng, J. Chen, and X. Li, “A hybrid method for pythagorean fuzzy multiple-criteria decision making,” *International Journal of Information Technology & Decision Making*, vol. 15, no. 02, pp. 403–422, 2016.

- [13] H. Zhang, G. Wei, and C. Wei, “Topsis method for spherical fuzzy magdm based on cumulative prospect theory and combined weights and its application to residential location,” *Journal of Intelligent & Fuzzy Systems*, vol. 42, no. 3, pp. 1367–1380, 2022.
- [14] E. K. Zavadskas and Z. Turskis, “Multiple criteria decision making (mcdm) methods in economics: An overview,” *Technological and economic development of economy*, vol. 17, no. 2, pp. 397–427, 2011.
- [15] J. Černevičienė and A. Kabašinskas, “Review of multi-criteria decision-making methods in finance using explainable artificial intelligence,” *Frontiers in artificial intelligence*, vol. 5, p. 827584, 2022.
- [16] M. R. Galankashi, E. Nasri, S. A. Helmi, and M. M. Arjmand, “Hospital selection problem: An integrated analytic hierarchy process (ahp) and fuzzy-topsis approach,” in *Proceedings of the 12th Annual International Conference on Industrial Engineering and Operations Management (IEOM)*, 2022, pp. 7–10.
- [17] D. L. Olson, “Comparison of weights in topsis models,” *Mathematical and Computer Modelling*, vol. 40, no. 7-8, pp. 721–727, 2004.
- [18] M. Behzadian, S. K. Otaghsara, M. Yazdani, and J. Ignatius, “A state-of the-art survey of topsis applications,” *Expert Systems with applications*, vol. 39, no. 17, pp. 13 051–13 069, 2012.
- [19] Y.-J. Lai, T.-Y. Liu, and C.-L. Hwang, “Topsis for modm,” *European journal of operational research*, vol. 76, no. 3, pp. 486–500, 1994.
- [20] Ž. Stević, D. Pamučar, A. Puška, and P. Chatterjee, “Sustainable supplier selection in healthcare industries using a new mcdm method: Measurement of alternatives and ranking according to compromise solution (marcos),” *Computers & industrial engineering*, vol. 140, p. 106 231, 2020.
- [21] S. Liu, “Research on the teaching quality evaluation of physical education with intuitionistic fuzzy topsis method,” *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 5, pp. 9227–9236, 2021.
- [22] S. Nădăban, S. Dzitac, and I. Dzitac, “Fuzzy topsis: A general view,” *Procedia computer science*, vol. 91, pp. 823–831, 2016.
- [23] N. Y. Seçme, A. Bayrakdaroğlu, and C. Kahraman, “Fuzzy performance evaluation in turkish banking sector using analytic hierarchy process and topsis,” *Expert systems with applications*, vol. 36, no. 9, pp. 11 699–11 709, 2009.
- [24] M. T. Ribeiro, S. Singh, and C. Guestrin, ““why should i trust you?”” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. DOI: 10.1145/2939672.2939778.
- [25] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, vol. 30, 2017.
- [26] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning important features through propagating activation differences,” in *International conference on machine learning*, PMLR, 2017, pp. 3145–3153.

- [27] S. M. Lundberg, G. Erion, H. Chen, *et al.*, “From local explanations to global understanding with explainable ai for trees,” *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020.
- [28] J. H. Friedman and B. E. Popescu, “Predictive learning via rule ensembles,” *The annals of applied statistics*, pp. 916–954, 2008.
- [29] S. M. Lundberg and S.-I. Lee, “Consistent feature attribution for tree ensembles,” *arXiv preprint arXiv:1706.06060*, 2017.
- [30] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” *arXiv preprint arXiv:1702.08608*, 2017.
- [31] E. Štrumbelj and I. Kononenko, “Explaining prediction models and individual predictions with feature contributions,” *Knowledge and information systems*, vol. 41, pp. 647–665, 2014.
- [32] M. T. Ribeiro, S. Singh, and C. Guestrin, “Anchors: High-precision model-agnostic explanations,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.
- [33] S. Rasmussen, L. M. Irgens, S. Albrechtsen, and K. Dalaker, “Predicting preeclampsia in the second pregnancy from low birth weight in the first pregnancy,” *Obstetrics & Gynecology*, vol. 96, no. 5, pp. 696–700, 2000.
- [34] B. Sibai, G. Dekker, and M. Kupferminc, “Pre-eclampsia,” *The Lancet*, vol. 365, no. 9461, pp. 785–799, 2005.
- [35] C. V. Ananth, K. M. Keyes, and R. J. Wapner, “Pre-eclampsia rates in the united states, 1980-2010: Age-period-cohort analysis,” *Bmj*, vol. 347, 2013.
- [36] S. Lisonkova and K. Joseph, “Incidence of preeclampsia: Risk factors and outcomes associated with early-versus late-onset disease,” *American journal of obstetrics and gynecology*, vol. 209, no. 6, 544–e1, 2013.
- [37] A. Rahman and M. G. Rabiul Alam, “Explainable ai based maternal health risk prediction using machine learning and deep learning,” in *2023 IEEE World AI IoT Congress (AIIoT)*, 2023, pp. 0013–0018. DOI: 10.1109/AIIoT58121.2023.10174540.
- [38] M. Ahmed and M. A. Kashem, “Iot based risk level prediction model for maternal health care in the context of bangladesh,” in *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)*, IEEE, 2020, pp. 1–6.
- [39] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, “Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission,” in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 1721–1730.
- [40] M. Swain, S. K. Dash, S. Dash, and A. Mohapatra, “An approach for iris plant classification using neural network,” *International Journal on Soft Computing*, vol. 3, no. 1, p. 79, 2012.
- [41] Y. Pachipala, H. C. Maddipati, P. S. Udimudi, V. Thota, V. N. Sree, and L. R. Burra, “Iris flower classification by using random forest in aws,” in *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, 2022, pp. 1772–1777.

- [42] S. NAEEM and A. ALI, “Breast cancer diagnosis using machine learning techniques,” 2022.
- [43] J. Zuluaga-Gomez, “Breast cancer diagnosis using machine learning techniques,” *arXiv preprint arXiv:2305.02482*, 2023.
- [44] D. Uzun Ozsahin, D. Ikechukwu Emegano, B. Uzun, and I. Ozsahin, “The systematic review of artificial intelligence applications in breast cancer diagnosis,” *Diagnostics*, vol. 13, no. 1, p. 45, 2022.
- [45] M. F. Akay, “Support vector machines combined with feature selection for breast cancer diagnosis,” *Expert systems with applications*, vol. 36, no. 2, pp. 3240–3247, 2009.
- [46] O. M. Al-Mubayyed, B. S. Abu-Nasser, and S. S. Abu-Naser, “Predicting overall car performance using artificial neural network,” 2019.
- [47] A. A. Al-Mobayed, Y. M. Al-Madhoun, M. N. Al-Shuwaikh, and S. S. Abu-Naser, “Artificial neural network for predicting car performance using jnn,” 2020.
- [48] Csafrít, *Maternal health risk data*, Dec. 2021. [Online]. Available: <https://www.kaggle.com/datasets/csafrít2/maternal-health-risk-data>.
- [49] U. M. Learning, *Iris species*. [Online]. Available: <https://www.kaggle.com/datasets/uciml/iris>.
- [50] U. M. Learning, *Breast cancer wisconsin (diagnostic)*. [Online]. Available: <https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>.
- [51] B. Z. Marko Bohanec, *Car acceptability classification dataset*. [Online]. Available: <https://www.kaggle.com/datasets/subhajeetdas/car-acceptability-classification-dataset>.