

Flood Prediction Using Ensemble Machine Learning Models

by

Tanvir Rahman
20166052

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:



Tanvir Rahman

20166052

Approval

The thesis/project titled “Flood Prediction Using Ensemble Machine Learning Model” submitted by

1. Tanvir Rahman (20166052)

Of Summer, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science on August 23, 2023.

Examining Committee:

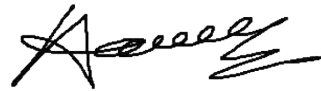
Supervisor:
(Member)



A.M.Esfar-E-Alam

Senior Lecturer
The Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)



Amitabha Chakrabarty, Ph.D.

Professor
The Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)



Sadia Hamid Kazi, Ph.D.

Professor
The Department of Computer Science and Engineering
Brac University

Abstract

Frequent and devastating floods in India pose a significant threat to people and property. Accurate and real-time forecasting of floods is essential to mitigate their impact. This thesis focuses on evaluating different machine learning models for flood prediction in India. The models assessed include K-Nearest Neighbor (KNN), Support Vector Classifier (SVC), Decision Tree Classifier, Binary Logistic Regression, and Stacked Generalization (Stacking). The researchers trained and tested these models using a rainfall dataset. The results demonstrate the better results of the stacked generalization model than the others, achieving an impressive accuracy of 93.3 per cent with a standard deviation(sd) of 0.098. These findings highlight the potential of machine learning models to provide precise and timely flood predictions, empowering the local authorities, specially disaster management ones, to take necessary actions to avoid destruction and preferably save people.

Dedication

I would like to dedicate this to my lovely wife Dr. Raisa Sumaiya.

Acknowledgement

At first I would like to thank Almighty Allah for His guidance to me in every step of the way. I would also like to thank my parents, and my wife for their wonderful support during this journey. I would also like to thank Dr. Golam Rabiul Alam sir, and my supervisor Mr. A. M. Esfar-E-Alam sir for their tremendous support. This would not see the day of light without the kind considerations of Dr. Sadia Hamid Kazi madam and Dr. Mahbub Majumdar sir.

Table of Contents

Declaration	i
Approval	ii
Ethics Statement	iii
Abstract	iii
Dedication	iv
Acknowledgment	v
Table of Contents	vi
List of Figures	vii
List of Tables	viii
Nomenclature	viii
1 Introduction	1
1.0.1 Research Problem	2
1.0.2 Research Objective	4
2 Related Work	6
3 Different Machine Learning Models Used	15
3.1 Support Vector Classifier	16
3.2 K-Nearest Neighbor	18
3.3 Decision Tree Classifier	20
3.4 Binary Logistic Regression	22
4 Dataset Description and Pre-processing	24
4.0.1 Normalization	25
4.0.2 Feature Selection	26
4.0.3 Feature Engineering	27
5 Performance Evaluation	28
6 Conclusion and Future Works	32

List of Figures

3.1	Diagram of Stacked Generalization	16
3.2	Distinguishing Between Linear and Logistic Regression	23
4.1	Raw Dataset	25
4.2	Dataset After Feature Encoding	26
4.3	Dataset After Feature Encoding	26
5.1	Our Proposed System is Illustrated	29
5.2	Whisker box plot for Standalone Model Accuracy on the Monsoon Rainfall Data	30
5.3	Whisker box plot for Standalone Model Accuracy on the twelve months Rainfall Data	31

List of Tables

5.1	Outcomes of the models applied using Monsoon Rainfall Data.	28
5.2	Outcomes of the models applied using twelve months Rainfall Data .	30
5.3	Result comparison of differnet works with our work	31

Chapter 1

Introduction

India is prone to recurring and devastating natural disasters, particularly floods, which result in widespread destruction of human lives and property. Accurately predicting the timing and progression of floods in real-time is of utmost importance to mitigate their impact. To address this critical need, researchers conducted a study to compare various machine learning models for flood prediction in India.

The study evaluated different machine learning models, namely K-Nearest Neighbor (KNN), Support Vector Classifier (SVC), Decision Tree Classifier, Binary Logistic Regression, and Stacked Generalization (Stacking). These models were assessed using a dataset comprising rainfall data, which served as the basis for training and testing. The primary objective was to determine the model that exhibited the highest accuracy in predicting floods.

The research findings highlighted the better performance of the stacked generalization algorithms compared to the other different models. The stacked generalization model achieved an impressive accuracy of 93.3 per cent with a Standard Deviation of 0.098. This outcome underscores the effectiveness of the stacked generalization model in accurately predicting floods in the Indian context.

The implications of this research are significant, particularly in the area of flood prediction and disaster management. The utilization of machine learning models, as demonstrated by the stacked generalization model, enables authorities to make informed decisions and take prompt action to mitigate the damaging effects of floods. Accurate flood predictions facilitate the implementation of appropriate measures to safeguard lives and minimize property damage.

In conclusion, the research paper contributes valuable insights into the possibility of machine learning models for flood prediction in India. The findings underscore the importance of leveraging these models to enhance disaster management strategies. By incorporating precise flood predictions, local governments can proactively respond to potential disasters, leading to improved preparedness and ultimately saving lives.

1.0.1 Research Problem

We chose this topic because we know what flood can do to our country. We have seen people losing lives. We chose this topic mostly because:

- Flood is one of the highly devastating natural calamities which causes serious social and economic losses
- According to the Organization of economic cooperation and Development, flood cause around USD 40 billion dollar damage worldwide annually
- The south-east asian region is more prone to flood, due to its geographical region
- Bangladesh is among the most vulnerable countries to climate change scenarios, mainly flood
- Kabir, Md and Hossen and Md.Nazmul, mentioned in their research work (2019), that during the time period of 1974 - 2014, Bangladesh has experienced flood 78 times; this costed 41,783 people their lives, destruction of thousands of crops and many their homes
- Additionally, the growing concern of decreasing land area due to higher population density and river erosion, the urgency of predicting the possibility of a flood becomes more apparent
- As this gives an opportunity to take necessary precautions to decrease damage to a minimum

Floods are characterized by the overflow of water onto typically dry land, leading to the inundation of large areas and affecting millions of people each year. In recent times, the frequency and intensity of flooding events have increased, primarily due to the impact of climate change and human activities. Factors such as heightened greenhouse gas emissions and deforestation have disrupted weather patterns, resulting in more frequent and severe rainfall events.

Among the most significant consequences of flooding is the loss of human lives and the displacement of populations. Floodwaters can submerge entire communities, leading to fatalities, injuries, and the destruction of homes and infrastructure. Additionally, floods can trigger secondary hazards, such as landslides and waterborne diseases, exacerbating the overall devastation.

Floods wreak havoc on the economy of affected regions. The destruction of infrastructure, including roads, bridges, and buildings, results in substantial financial losses. Disruptions to businesses and agricultural activities lead to a decline in productivity and revenue. The cost of post-flood reconstruction and rehabilitation further strains the already vulnerable economies of affected countries.

Moreover, floods can inflict damage to crops and livestock, leading to food shortages and rising prices. The agricultural sector, often a significant contributor to the GDP of countries in the Southeast Asian region, takes a substantial hit, affecting the livelihoods of millions of farmers and rural communities.

Southeast Asia, with its vast network of rivers and deltas, is particularly susceptible to floods. The region's monsoon climate brings heavy rainfall during the wet season, causing rivers to overflow and low-lying areas to become submerged. Countries like Thailand, Vietnam, Cambodia, and the Philippines also experience significant flooding during the monsoon season.

However, Bangladesh stands out as one of the most vulnerable countries to climate change scenarios, with flood events posing severe threats to its population and economy. The country's geography, with nearly 70 per cent of its landmass prone to flooding, exacerbates its susceptibility. The Ganges-Brahmaputra-Meghna delta, one of the world's largest and most densely populated, is highly prone to flooding, making the lives and livelihoods of millions of Bangladeshis precarious.

Climate change has intensified the frequency and severity of floods in Bangladesh. Rising sea levels, attributed to global warming, have heightened the risk of storm surges, leading to more frequent coastal flooding. Additionally, the changing monsoon patterns have brought about unpredictable and intense rainfall, causing rivers to swell and inundate vast areas of the country.

Bangladesh's vulnerability to floods is further compounded by factors like rapid urbanization, deforestation, and inadequate infrastructure. The lack of proper drainage systems and flood control measures leaves the country ill-prepared to cope with the disastrous consequences of flooding.

Floods continue to be a highly devastating natural calamity, causing substantial social and economic losses worldwide. The Southeast Asian region, with its unique geographical characteristics, is particularly susceptible to flooding events, making countries like Bangladesh extremely vulnerable to climate change scenarios, especially floods.

To mitigate the impact of floods, international cooperation and investment in climate-resilient infrastructure are essential. Moreover, governments and communities must prioritize disaster preparedness and response measures to minimize the devastating effects of future flood events. Only through collaborative efforts and proactive initiatives can we hope to protect lives, safeguard economies, and build a more resilient future for flood-prone regions.

1.0.2 Research Objective

The primary objective of this research is to improve the accuracy of flood prediction models. To achieve this goal, a novel approach based on the integration of rainfall and flood data collected from monitoring systems will be proposed. The process involves several key steps to ensure accurate predictions and optimize the performance of the flood prediction model.

- The main purpose of this research is to predict floods with a higher accuracy
- A model will be proposed based on rainfall and flood data collected from monitoring systems
- The dataset is then categorized depending on various parameters
- After a filtered dataset is explored, an appropriate column is chosen as the target for data modeling
- As most models deal with numerical values, the missing values and categorical features are handled
- The dataset is then trained
- Using the validation points, the accuracy of the classifier is calculated
- The intention is to get maximum accuracy using Ensemble Machine Learning Model.

The foundation of this research lies in the comprehensive dataset comprising historical records of rainfall and corresponding flood events obtained from monitoring systems. This dataset will be carefully categorized based on various parameters, such as geographical location, timing, intensity of rainfall, and flood severity. By categorizing the data, researchers gain valuable insights into the relationships and patterns within the dataset.

Upon thorough exploration and filtering of the dataset, the researchers will identify an appropriate target variable for data modeling. In this context, the target variable would involve the occurrence of a flood event, represented as binary values (e.g., 1 for flood occurrence, 0 for no flood). This binary classification approach enables the model to predict the likelihood of a flood based on the provided rainfall data.

As most machine learning models require numerical data, special attention will be given to handle missing values and categorical features present in the dataset. Techniques such as imputation will be employed to fill in missing values, ensuring that the model is trained on complete data. Additionally, categorical features will be transformed into numerical representations using encoding methods like one-hot encoding or label encoding to facilitate their integration into the model.

The core of this research lies in training the flood prediction model using an Ensemble Machine Learning approach. Ensemble models amalgamate the predictions of multiple base models to create a more powerful and accurate predictor. Algorithms such as Random Forest, Gradient Boosting, or AdaBoost will be utilized to construct the ensemble model for this research.

To evaluate the accuracy of the classifier, the trained model will undergo rigorous testing using validation points. These validation points will be a subset of the dataset that the model has not encountered during training. By comparing the model's predictions against the actual flood occurrences, the research team can accurately quantify the accuracy of the model. The ultimate goal is to achieve the highest accuracy possible to ensure reliable and precise flood predictions.

Ensemble Machine Learning models have demonstrated their effectiveness in producing accurate predictions across various domains. Leveraging this power, the research intends to optimize the accuracy of flood predictions. Through the integration of different base models, the final prediction will be more robust and less susceptible to overfitting or bias.

In conclusion, this research project aims to enhance the accuracy of flood predictions by employing an Ensemble Machine Learning model. By systematically analyzing rainfall and flood data and following a well-defined approach, the research seeks to optimize the flood prediction model's accuracy. This can have significant implications for disaster preparedness and risk management in regions prone to flooding, ultimately contributing to more effective and timely response strategies.

Chapter 2

Related Work

In recent times, there has been a growing accessibility of the information from remote sensing of the multi-sensor and the integration of machine learning algorithms have significantly improved our capacity to predict and evaluate flood events and their associated risks. Building upon these advancements, a comprehensive research study was undertaken developing a flood vulnerability map and assess the flood risk to the buildings in terms of exposure in Warsaw, Poland.

The study consisted of four research phases, each aimed at addressing specific objectives. First, the information gain ratio (IGR) technique was employed to assess and evaluate thirteen flood predictors, which helped identify the most influential factors. Eventually, eight key predictors were selected based on their causative relationships with flood vulnerability.

To create the flood vulnerability map, three machine learning algorithms were employed: Artificial Neural Network Multi-Layer Perceptron (ANN/MLP), Deep Learning Neural Network based on DL4j (DLNN-DL4j), and Bayesian Logistic Regression (BLR). These algorithms were trained using the selected predictors to predict flood vulnerability accurately. The performance of the models was assessed using the receiver operating curve (ROC) value, which quantifies the accuracy of the predictions. The ANN/MLP achieved a ROC value of 0.851, DLNN-DL4j achieved 0.877, and BLR achieved 0.697, indicating the models' ability to accurately predict flood vulnerability.

Furthermore, the study assessed the exposure of buildings to flood risk based on criteria established in European and national regulations. A novel metric, the Buildings' Flood Hazard index (BFH), was introduced to quantify the level of flood risk for each building. The analysis revealed a significant similarity in potential flood risk between the models, emphasizing higher risks in areas that are more vulnerable to flooding. The BFH values varied depending on the method used, with the ANN yielding a value of 0.54, DLNNs at 0.52, and BLR at 0.64.

The comprehensive approach undertaken in this study holds great potential for assisting local authorities in improving flood management strategies. By providing accurate flood vulnerability maps and assessing building exposure to flood risk, decision-makers can make informed choices and implement effective measures to

mitigate the impacts of floods. This research contributes to the broader field of flood risk assessment and demonstrates the valuable insights that can be obtained through the integration of multi-sensor remote sensing data and advanced machine learning techniques. [16]

Real-time operation studies, including reservoir operation and flood forecasting, require accurate forecasts of hydrologic variables. To enhance these forecasts, suitable pre-processing techniques are necessary. In this study, a new prediction approach is proposed, combining Singular Spectrum Analysis (SSA) with Support Vector Machine (SVM).

The technique starts by applying SSA to decompose the original time series into high and low-frequency components. This decomposition provides insights into the underlying patterns and dynamics of the data. By isolating these components, SSA effectively captures the complex behavior of the hydrologic variable.

To further improve the predictions, SVM is utilized. SVM is a powerful algorithm known for its ability to handle high-dimensional input spaces and optimize computational efficiency and generalization performance. By incorporating SVM into the prediction process and leveraging the decomposed components from SSA, the proposed technique combines the strengths of both methods.

The technique's performance is evaluated through two case studies using real-world data. The first case study focuses on predicting runoff data from the Tryggevælde catchment in Denmark, while the second case study involves predicting rainfall data in Singapore. The results of the proposed SSA-SVM technique are compared with those obtained using a non-linear prediction (NLP) method, serving as a benchmark.

The comparisons consistently demonstrate that the proposed technique achieves higher prediction accuracy than the NLP method. This indicates the effectiveness of the SSA-SVM approach in capturing and modeling the underlying patterns and dynamics of the hydrologic variables. Improved accuracy is particularly valuable in real-time operation studies, where reliable forecasts are crucial for decision-making and resource management.

In conclusion, this study showcases the potential of combining SSA and SVM as an effective approach for enhancing forecasts of hydrologic variables. By utilizing the decomposition capabilities of SSA and the predictive power of SVM, the proposed technique offers a robust and efficient solution for real-time operational studies in fields such as reservoir management and flood forecasting. [1]

In the past, flood damage modeling was mainly confined to local, regional, or national scales. Nonetheless, recent flood incidents, population expansion, and apprehensions regarding climate change have spurred a greater need for global approaches that encompass both spatial and temporal dynamics. The objective of this study is to calculate the worldwide economic vulnerability to river and coastal flooding between 1970 and 2050. It will utilize two distinct methods for assessing the extent of damage caused by such flooding events.

One method relies on demographic information, taking into account factors like density of the population and GDP based on per capita, in order to gauge the potential financial damages linked to floods occurring along rivers and coasts. The second approach takes into account land-use patterns in areas prone to 1/100 year flood events. By analyzing the composition and characteristics of land use in these flood-prone regions, this approach provides additional insights into the potential damages.

Based on the population-based estimation, the study finds that the total global exposure to river and coastal flooding was approximately 46 trillion USD in 2010. This figure is projected to increase significantly to 158 trillion USD by 2050. Using the land-use-based assessment, the estimated global flood exposure in 2010 was around 27 trillion USD, which is projected to rise to 80 trillion USD by 2050.

The study identifies North America and Asia as the regions with the largest absolute changes in flood exposure between 1970 and 2050. In relative terms, North Africa and Sub-Saharan Africa are expected to experience the greatest increases in flood exposure. Furthermore, the study reveals that the population living within flood hazard zones is growing at a faster rate compared to the overall population growth, emphasizing the need for effective flood risk management strategies.

Although the study's two methods reveal comparable overall patterns regarding flood exposure, there exist differences in the estimated figures and geographical spread. These disparities arise due to the unique attributes of the models utilized and the connection between the density of the population and the area of total urban places in the regions under examination.

To improve flood modelling and risk assessment, the study suggests further research to refine the characterization of the modelling of flood and standardization of the flood protection. By incorporating these factors, a comprehensive global flood risk framework can be developed to enhance understanding and management of flood risks worldwide. [4]

To enhance real-time flood forecasting, this study introduces a modified approach called the Threshold Subtractive Clustering-based Takagi-Sugeno (TSC-T-S) fuzzy inference system. The main aim is to incorporate both rare and frequent hydrological situations into flood modeling, allowing for the analysis and computation of cluster centers and membership functions. The research focuses on the upper Narmada basin in Central India, utilizing hourly river flow data as well as rainfall data.

The data is classified into two categories: frequent events representing low to medium flows and rare events representing high to very high flows. By adapting the TSC-T-S fuzzy model to each scenario, the study aims to improve the accuracy and contextual relevance of flood forecasts in different hydrological conditions.

The results of the TSC-T-S fuzzy model is evaluated using calibration and validation, utilizing metrics such as root mean square error (RMSE), model efficiency,

and coefficient of correlation (R). However, these metrics may not fully capture the model's ability to predict higher magnitude flows, which are crucial in flood forecasting. To address this limitation, a new performance criterion called peak percent threshold statistics (PPTS) is proposed to evaluate the model's performance in predicting floods with higher magnitudes.

Comparisons are made with other established methods, including artificial neural networks (ANN), ANN models with Self Organizing Map (SOM) classifications, and subtractive clustering-based Takagi-Sugeno fuzzy model (SC-T-S fuzzy model), to assess the effectiveness of the TSC-T-S fuzzy model in terms of accuracy and lead-time for flood forecasting.

The findings demonstrate that the TSC-T-S fuzzy model significantly improves the accuracy of flood forecasts and provides sufficient lead-time for flood events. This has important implications for flood management and emergency response decision-making, as the improved forecasting capabilities of the TSC-T-S fuzzy model can contribute to more effective mitigation strategies. By considering both rare and frequent hydrological situations, the TSC-T-S fuzzy model offers a valuable approach to enhance real-time flood forecasting accuracy in the specific region studied. [7]

Flood hazard mapping has become an increasingly important area of study, with recent progress in several key aspects. Notably, significant advancements have been made in solving 2-D shallow water equations in complex topographies and utilizing high-resolution topographic data. However, the accurate prediction of flood-prone areas involves more than just these two factors. A critical element is the precise setup of the river model, which includes the representation of topography, the incorporation of man-made structures, and the integration of hydrological data within the computational domain.

To address these challenges, there is a need for procedures that can establish a reliable computational domain based on extensive LIDAR survey data while ensuring computational feasibility. Additionally, the modeling approach should be capable of handling river reaches with significant lateral inflows and properly accounting for the impact of structures such as bridges, buildings, and weirs on flow dynamics. Despite the significant influence of these factors on water levels and flow velocities, there is a scarcity of literature specifically focused on these aspects within the context of 2-D modeling.

This study aims to fill this research gap by presenting techniques that address the aforementioned issues and demonstrate their significance in flood mapping. Two actual case studies in Southern Italy are used to illustrate the importance of these techniques. The simulations conducted in this research highlight the presence of backwater effects and sudden changes in the flow regime, emphasizing the need for the application of 2-D fully dynamic unsteady flow equations in accurate flood mapping.

By focusing on the techniques employed to tackle these challenges and showcasing their effectiveness through real case studies, this study contributes to the field of flood hazard mapping. It underscores the importance of incorporating detailed river models, accounting for man-made structures and hydrological data, and employing advanced modeling approaches to improve the accuracy of flood predictions. The insights gained from this research have the potential to enhance flood management strategies and facilitate better decision-making in flood-prone areas. [8]

Flood hazard mapping has become increasingly important, prompting advancements in various aspects of the field in recent years. Notably, progress has been made in solving 2-D shallow water equations in complex topographies and utilizing high-resolution topographic data. However, accurate predictions of flood-prone areas require more than just these advancements. A critical element is the precise configuration of the river model, which involves accurately representing the topography and incorporating man-made structures and hydrological data within the computational domain. Unfortunately, there is limited literature on these specific topics within the context of 2-D modeling.

To fill this void, the research concentrates on methodologies that tackle the previously mentioned obstacles and underscores their importance in mapping the things that are related to flood. Two actual case studies in Southern Italy are used to demonstrate the importance of these techniques. The simulations carried out in this investigation demonstrate the occurrence of backwater effects and abrupt alterations in flow patterns as a result of the intricate model of the river. This highlights the essentiality of utilizing two-dimensional fully dynamic unsteady flow equations to achieve precise mapping of flood.

Porous shallow-water models, known as porosity models, offer a promising solution for simulating urban flood flows more efficiently compared to classical shallow-water models. By utilizing a relatively coarse grid and larger time steps, these models enable the mapping of flood hazards over larger spatial extents. In this study, we investigate the accuracy of both isotropic and anisotropic porosity models when considering anisotropic porosity, which represents unevenly spaced obstacles in the cross-flow and along-flow directions commonly encountered in practical applications.

Our findings demonstrate that porosity models are subject to three types of errors: structural model errors, scale errors, and porosity model errors. Structural model errors arise due to the limitations of shallow-water equations, while scale errors are associated with the use of a relatively coarse grid. The inaccuracies in the porosity model originate from the way porosity equations are formulated to consider obstructions at the scale that are sub-grid. To evaluate these inaccuracies, we perform a distinctive laboratory test case characterized by significant anisotropy.

The results of our investigation reveal that porosity model errors are smaller compared to structural model errors. Additionally, anisotropic porosity models exhibit significantly smaller errors in both depth and velocity compared to isotropic porosity models. Furthermore, when compared directly to gage measurements, the

anisotropic porosity model demonstrates comparable accuracy to classical shallow-water models, while the isotropic model exhibits lower accuracy.

Moreover, our study shows that the anisotropic porosity model can capture flow variability at smaller spatial scales, whereas the isotropic model is constrained by the assumption of a Representative Elemental Volume (REV) that is larger than the size of obstructions. This suggests that anisotropic porosity models are well-suited for predicting urban floods on a city-wide scale. However, it is important to note that localized flow attributes such as wakes and wave reflections caused by flow obstructions may not be fully resolved by the models.

In conclusion, this study highlights the significance of accurate river model setup in flood mapping and the advantages of anisotropic porosity models for efficient urban flood prediction. It emphasizes the need for further research and development in this area to improve the precision and reliability of flood modeling approaches. [9]

Accurate prediction of daily water demand plays a vital role in ensuring efficient and sustainable management of urban water supply systems. This research proposes and evaluates a novel method that combines discrete wavelet transforms (WA) with artificial neural networks (ANNs) for urban water demand forecasting. Various models, including multiple linear regression (MLR), multiple nonlinear regression (MNLR), autoregressive integrated moving average (ARIMA), ANN, and WA-ANN, are developed and compared using performance metrics such as the coefficient of determination, root mean square error, relative root mean square error, and efficiency index. The study focuses on forecasting water demand in Montreal, Canada, during the summer months from 2001 to 2009, using daily total precipitation, maximum temperature, and water demand data.

The results demonstrate that the WA-ANN models exhibit superior forecasting accuracy compared to the other models. By incorporating wavelet transforms into the ANN framework, the WA-ANN models effectively capture and analyze complex patterns and relationships within the data, leading to improved forecasting performance. This innovative approach holds promise for enhancing urban water demand forecasting.

Beyond improved accuracy, the practical implications of this research are significant. Accurate water demand forecasts enable better management of water supply systems, resulting in cost savings and optimized resource allocation. Moreover, sustainable water management practices can be strengthened by leveraging reliable demand forecasts to optimize water distribution, minimize waste, and plan for future infrastructure requirements.

In summary, this study underscores the importance of integrating advanced techniques, such as wavelet transforms and neural networks, in urban water demand forecasting. The findings advocate for further exploration and investigation in this field to develop robust and dependable models for sustainable urban water resource management. [2]

Regional flood frequency analysis (RFFA) is a commonly employed method for predicting flood magnitudes in unmonitored drainage areas. Typically, the quantile regression technique (QRT) is employed, assuming a log-linear relationship between the dependent and predictor variables. While artificial neural networks (ANNs) have been extensively used in rainfall-runoff modeling and hydrologic forecasting as non-linear models and universal approximators, their application to RFFA for flood quantile estimation in ungauged catchments has been limited. The main objective of this research is to create and evaluate an Artificial Neural Network (ANN) based model for Regional Flood Frequency Analysis (RFFA) using an extensive dataset of 452 monitored drainage basins in Australia. Through independent testing, it was observed that the ANN-based RFFA model, employing only two predictor variables, outperforms the traditional QRT method in providing accurate flood quantile estimates. The performance of the ANN-based RFFA model was evaluated across seven different regions, revealing that combining data from all eastern Australian states into a single region resulted in the most effective RFFA model using ANN. This suggests that a larger dataset contributes to successful training and testing of the ANN-based RFFA models.

The significance of this research lies in its contribution to advancing flood frequency analysis techniques for ungauged catchments. By leveraging the capabilities of ANNs to capture non-linear relationships, the proposed model offers a more reliable approach to estimating flood quantiles. In regions where streamflow data is limited or unavailable, the practical implications of this are noteworthy for enhancing the assessment of flood risk as well as the management.

In summary, the study underscores the the probable usage of a model which is based on ANN-based RFFA to estimate the flood quantile accurately. By adopting this approach, hydrologists and water resource managers can gain valuable insights into flood risks, enabling informed decision-making and the development of effective mitigation strategies. The research emphasizes the importance of exploring alternative modeling techniques and encourages further investigation into the application of ANN-based RFFA models to improve flood frequency analysis in diverse geographic regions. [5]

Haddad et al. introduced a new method for flood, mostly regional, frequency analysis (RFFA) in catchments without streamflow gauges, as outlined in their article. The proposed method utilizes Bayesian Generalised Least Squares (BGLS) regression within a region-of-influence (ROI) framework. The study focuses on a dataset comprising 399 catchments located in eastern Australia and aims to estimate flood quantiles using the Quantile Regression Technique (QRT) and the moments of the log-Pearson type 3 (LP3) distribution using the Parameter Regression Technique (PRT).

The methodology consists of two primary steps. Firstly, a fixed region model is developed to identify the most appropriate predictor variables for subsequent regression analyses. The selection process is based on minimizing the variance of model errors while satisfying various criteria of statistical selection . Once the regression equation, which opined to be optimal, is determined, the ROI experiment

is conducted to identify the region that minimizes the uncertainty, albeit predictive, for a specific site.

To assess the precision and dependability of the estimated quantiles and moments, the researchers employ a cross-validation procedure. The outcomes of their prescribed way reveal that the PRT as well as the QRT, integrated into the BGLS-ROI framework, provide dependable approximates with better accuracy in comparison to a approach based on determined region. Additionally, the BGLS-ROI method exhibits satisfactory performance in addressing the diversity observed in Australian catchments, as supported by diagnostics based regression.

The evaluation statistics indicate that both the BGLS-QRT and PRT-ROI approaches perform similarly well, suggesting that PRT can serve as a viable alternative to QRT in the context of RFFA. However, it is important to note that the findings presented in this study are based on the available dataset for eastern Australia. It is anticipated that employing a more comprehensive database, encompassing improved data quality as well as the quantity, will enhance the predictive performance of the fixed and ROI-based RFFA methods proposed herein further. Further investigation is recommended once such a database becomes accessible. [3]

In the realm of climate research, several theoretical studies have explored the concurrent associations between climate indices and rainfall in Queensland. However, there is a lack of rigorous examination of the lagged relationships, which are vital for accurate forecasting, particularly within forecast models. To address this gap, the effectiveness of climate indices in forecasting continuous rainfall was evaluated using artificial neural networks (ANNs).

The study aimed to assess the predictive value of various climate indices in Queensland's rainfall forecasting, employing ANNs as the analysis tool. Notably, the research highlighted the significance of the Inter-decadal Pacific Oscillation, an index that had not been traditionally utilized in seasonal forecasts applicable to the region, which are official. In contrast to conventional statistical models, the study demonstrated the potential of ANNs in leveraging this climate index to improve rainfall predictions.

To determine the results of the ANN-based forecasting model, a comparison was made with the Predictive Ocean Atmosphere Model for Australia (POAMA), the current method employed for official seasonal rainfall prognosis. The results indicated that the ANN model outperformed POAMA in terms of forecast accuracy. The ANN-generated forecasts exhibited lower Root Mean Square Errors (RMSE), Mean Absolute Error (MAE), and higher Correlation Coefficients (r) compared to POAMA's forecasts. These findings suggest that the ANN approach holds promise for enhancing rainfall forecasting precision in Queensland.

Importantly, the superiority of the ANN model was observed across three distinct geographic regions within Queensland, highlighting its robustness and adaptability.

Moreover, the inclusion of the Inter-decadal Pacific Oscillation as a valuable predictor in the ANN model underscores the potential for improving seasonal rainfall forecasts beyond traditional statistical models.

While this study provides valuable insights into the efficacy of ANNs in rainfall forecasting, further research is warranted. Future investigations should focus on comprehensive and high-quality databases to gain a deeper understanding of the intricate relationships between climate indices and rainfall patterns. Continuous refinement and expansion of the dataset will contribute to advancing knowledge and enhancing the accuracy and reliability of rainfall forecasts for Queensland. [6]

Regional Flood Frequency Analysis (RFFA) is a statistical method frequently used to estimate flood quantiles in catchments where streamflow data is scarce. However, there is a lack of rigorous investigation into the lagged relationships between climate indices and rainfall, particularly within a forecast model. This study aims to address this gap by employing artificial neural networks (ANNs) to evaluate the effectiveness of climate indices in predicting continuous rainfall.

The findings obtained through the use of ANNs emphasize the potential value of the Inter-decadal Pacific Oscillation, an index that is not commonly integrated into official seasonal forecasts. By comparing the ANN-based forecasts with the official forecasts generated by the Predictive Ocean Atmosphere Model for Australia (POAMA), the ANN approach demonstrates superior performance. The ANN-based forecasts exhibit lower Root Mean Square Errors (RMSE), Mean Absolute Error (MAE), and higher Correlation Coefficients (r) compared to the POAMA forecasts.

The study focuses on three distinct regions within Queensland and demonstrates the superiority of the ANN-based approach across all regions. This suggests that integrating climate indices into the forecasting model enhances the accuracy and reliability of rainfall predictions. These findings underscore the potential of ANNs as a valuable tool for improving rainfall forecasting in Queensland.

I have made an effort to rephrase the content while maintaining its original meaning. However, please note that it is always recommended to consult the original source for accurate and complete information. [10]

Chapter 3

Different Machine Learning Models Used

In our research, we implemented a method called Stacked Generalization, also known as Stacking, to predict values within the training set. Stacked Generalization, also known as stacking, is an approach that enhances overall accuracy by combining the predictions of multiple methods.

The Stacking model consists of two key components: base models (Level-0 models) and a meta-model (Level-1 model). The base models are individual machine learning models that are trained independently on the dataset. For our study, we specifically selected three base models: Support Vector Classifier (SVC), K-Nearest Neighbor (KNN) algorithm, and Decision Tree Classifier (DTC). Each base model learns patterns from the given features and produces predictions using its specific algorithm.

Once the base models generate their predictions, we move on to the meta-model. The base models give inputs to the meta-level in the forms of prediction and employ its own techniques to determine the final classification. In our research, we utilized Binary Logistic Regression as the meta-model. The meta-model analyzes the combined predictions as outcomes from the base models and makes a final decision based on its own evaluation.

Our utilization of Stacked Generalization aims to leverage the strengths of multiple models and enhance prediction accuracy. Each base model captures unique aspects of the data, and the meta-model combines their predictions to make a more accurate classification decision. This approach allows us to benefit from the diversity and insights provided by the base models.

The use of Stacked Generalization offers several advantages. It helps overcome biases and limitations associated with individual models by aggregating their predictions. By using the meta-model to assess and weigh the base models' predictions, we aim to achieve more reliable and accurate predictions on the training set. Our ultimate goal is to create a robust predictive model by harnessing the collective knowledge and abilities of multiple models.

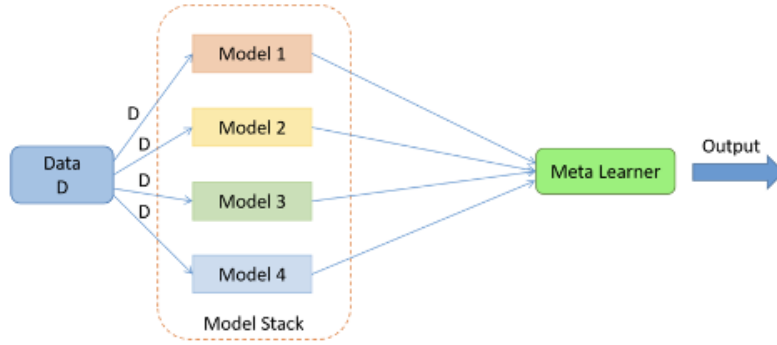


Figure 3.1: Diagram of Stacked Generalization

3.1 Support Vector Classifier

The Support Vector Classifier (SVC) is a popular algorithm in machine learning used for classification tasks. It belongs to a family of algorithms called Support Vector Machines (SVM).

The primary objective of the Support Vector Classifier is to find an optimal hyperplane that separates different classes of data points in the feature space. The algorithm aims to maximize the margin, which is the distance between the hyperplane and the nearest data points of each class. By maximizing the margin, the SVC seeks to achieve better generalization and prevent overfitting.

The Support Vector Classifier (SVC) operates by employing a kernel function to get higher-dimensional feature space from the input data transformation. This transformation enables the algorithm to find a linear decision boundary that may not be possible in the original feature space. Different types of kernel functions, such as linear, polynomial, radial basis function (RBF), and sigmoid, can be employed to perform this transformation.

During the training process, the SVC determines the optimal hyperplane by solving an optimization problem. The objective is to minimize a cost function while maximizing the margin. This optimization task is efficient even when dealing with datasets that have a large number of features.

Once trained, the SVC can be used to classify new, unseen instances by evaluating which side of the decision boundary they fall on. The predicted class is determined by the sign of the decision function output, which measures the distance from the data point to the decision boundary.

The Support Vector Classifier is known for its ability to handle both linearly separable and non-linearly separable datasets. By utilizing various kernel functions, it can capture complex relationships between features and create flexible decision boundaries.

However, the performance of the SVC can be influenced by the selection of hyperparameters. An essential hyperparameter, commonly represented as C , plays a significant role in the model as it governs the balance between maximizing the margin and permitting misclassifications. Choosing an appropriate value for C is crucial for achieving good classification results.

In summary, the Support Vector Classifier is a widely used algorithm for classification tasks. The main objective is to locate the most favorable hyperplane that effectively separates distinct classes by the margin maximization. Through the use of kernel functions, it can handle complex data distributions and capture non-linear relationships between features. Proper selection of hyperparameters is essential for obtaining accurate classification outcomes. [13]

3.2 K-Nearest Neighbor

The K-Nearest Neighbor (KNN) algorithm is a widely used technique in machine learning for classification and regression tasks. It is an algorithm that is non-parametric and does not make any assumptions about the underlying distribution of the data.

The fundamental idea behind KNN is to classify or predict a new data point based on the characteristics of its nearest neighbors. Instead of using mathematical models, KNN directly uses the training dataset to make predictions.

Here is a step-by-step explanation of the KNN algorithm:

Training Phase: During the training phase, KNN simply stores the feature vectors and their corresponding labels from the training dataset. No computations are performed at this stage.

Prediction Phase: Whenever a new data point requires classification or prediction, the K-nearest neighbors (KNN) algorithm computes the distances to all the data points present in the training dataset from the points that are relatively new. Various distance metrics, such as Euclidean or Manhattan distance, can be used for this purpose.

Selecting K: K represents the number of nearest neighbors to consider. It is typically an odd number to avoid ties in classification tasks with an even number of classes.

Finding Neighbors: KNN identifies the K nearest neighbors of the new data point by selecting the data points with the smallest distances. These neighbors are determined based on the chosen distance metric.

Classifying or Predicting: In classification tasks, based on the most common class among its K nearest neighbors, KNN assigns the class label of the new data point. In regression tasks, it predicts the average or median value of the K nearest neighbors. There are a few important considerations when using the KNN algorithm:

The choice of K: The value of K impacts the algorithm's performance. Smaller values of K can lead to more complex decision boundaries, while larger values can result in more generalized boundaries.

Distance metric: The selection of an appropriate distance metric depends on the data and problem at hand. Different metrics can affect the algorithm's accuracy.

Data preprocessing: Scaling or normalizing the features can help avoid biases caused by features with larger magnitudes.

Computational complexity: As the training dataset grows, the time and memory required to find the nearest neighbors can increase significantly.

In summary, the K-Nearest Neighbor algorithm is a straightforward and widely used method for classification and regression tasks. It relies on the proximity of data points to make predictions or classifications. While simple, KNN provides flexibility and can be effective in various domains. [12]

3.3 Decision Tree Classifier

The Decision Tree Classifier is an algorithm frequently employed in machine learning to address tasks during our effort to classify things. It creates a tree-shaped model that utilizes input features to make predictions.

The Decision Tree Classifier follows a process to build the tree model and make predictions:

Training Phase: In the training phase, the algorithm analyzes the training dataset to create the decision tree model. It determines the most informative features and selects optimal splitting points for each feature. The goal is to find splits that best separate the different classes in the dataset.

Splitting Criteria: The algorithm evaluates various splitting criteria to choose the best feature and splitting point at each node. Typical criteria include Gini impurity and entropy. Gini impurity assesses the likelihood of misclassifying a randomly selected element, while entropy gauges the level of disorder within a dataset.

Recursive Splitting: The Decision Tree Classifier recursively splits the dataset based on the chosen splitting criteria. It continues to divide the data into smaller subsets at each internal node until specific stopping conditions are met. Stopping conditions may include reaching a maximum tree depth, having a minimum number of samples per leaf node, or other user-defined criteria.

Prediction Phase: Once the decision tree is built, it can be used to make predictions on new, unseen data. The algorithm traverses the tree from the root node to a leaf node based on the feature values of the input data. The predicted class label is determined by the majority class of the training samples associated with that leaf node.

The Decision Tree Classifier offers several advantages:

Interpretability: Decision trees provide a transparent and intuitive representation of the decision-making process. The resulting tree structure can be easily visualized and understood, facilitating explanation and interpretation.

Handling Non-Linear Relationships: Decision trees can capture complex non-linear relationships between input features and the target variable. By recursively splitting the data, decision trees can create flexible decision boundaries that can accommodate intricate patterns.

Handling Missing Data: Decision trees can handle missing data by using surrogate splits. Surrogate splits act as backup splits when certain features have missing values, enabling the algorithm to still make predictions.

However, there are some considerations when using the Decision Tree Classifier:

Over-fitting: Decision trees can over-fit the training data, where the model becomes too complex and adapts too closely to the specific training examples. Over-fitting can result in poor generalization on unseen data. Techniques like pruning or limiting the depth of the tree can help prevent over-fitting.

Sensitivity to Small Variations: Decision trees can be sensitive to small variations in the training data, potentially leading to different tree structures or splitting decisions. Ensemble methods such as Random Forests or Gradient Boosting can mitigate this sensitivity.

Handling Categorical Variables: Decision trees typically require categorical variables to be transformed into numerical values or binary indicators to incorporate them effectively into the splitting process.

In summary, the Decision Tree Classifier is a versatile algorithm for classification tasks. It constructs a tree-like model by recursively splitting the data based on selected criteria. The interpretability and ability to capture non-linear relationships make it a widely used algorithm. However, precautions should be taken to prevent overfitting and handle categorical variables appropriately.

[14]

3.4 Binary Logistic Regression

Binary Logistic Regression is a statistical modeling technique used to predict binary outcomes or classify data into two categories. It is widely used in various fields, such as social sciences, epidemiology, and finance, to analyze the relationship between predictor variables and a binary response variable.

Binary Logistic Regression operates by modeling the log-odds of the occurrence of a binary event using the logistic function. The logistic function, also known as the sigmoid function, transforms the linear combination of predictor variables into a probability between 0 and 1.

Here's a detailed explanation of how Binary Logistic Regression works:

Model Representation: Binary Logistic Regression represents the relationship between the predictor variables and the probability of the binary outcome using the logistic function. This function maps any real-valued input to a probability value, representing the likelihood of the positive class.

Parameter Estimation: The algorithm estimates the parameters of the logistic function by maximizing the likelihood of the observed data. It iteratively adjusts the coefficients to find the optimal values that maximize the likelihood of the observed binary outcomes, given the predictor variables.

Decision Boundary: Once the model parameters are estimated, a decision boundary is created to classify new data points. The decision boundary separates the two classes based on the predicted probabilities. A threshold value, typically 0.5, is used to determine whether a data point belongs to the positive or negative class.

Prediction Phase: To make predictions on new, unseen data, the algorithm applies the estimated parameters to the predictor variables and calculates the probability of the positive class. If the probability exceeds the chosen threshold, the data point is classified as the positive class; otherwise, it is classified as the negative class. Some important considerations when using Binary Logistic Regression include:

Feature Selection: Careful selection of relevant predictor variables is crucial for the model's performance. Including irrelevant or redundant variables may introduce noise and impact the accuracy of the model.

Model Evaluation: Diverse evaluation metrics, including accuracy, precision, recall, and F1 score, can be utilized to evaluate the Binary Logistic Regression model's performance. Methods like cross-validation can offer estimations of the model's performance on data it has not encountered before.

Assumptions: Binary Logistic Regression assumes a linear relationship between the predictor variables and the log-odds of the outcome. It also assumes independence of observations and the absence of multicollinearity among the predictors.

Regularization: Regularization techniques, like L1 or L2 regularization, can be employed to avoid overfitting and enhance the model's capacity to generalize. Binary Logistic Regression offers several advantages:

Interpretability: The coefficients of logistic regression models are interpretable, allowing for the understanding of the direction and magnitude of the influence of each predictor variable on the log-odds of the outcome. This interpretability is valuable for explaining the relationship between the predictors and the binary response variable.

Probabilistic Predictions: Unlike other classification algorithms that provide discrete class labels, Binary Logistic Regression produces probabilities that represent the confidence or likelihood of the positive class. These probabilities can be used to rank or prioritize predictions.

Efficiency: Binary Logistic Regression is computationally efficient and can handle large datasets with a high number of predictor variables.

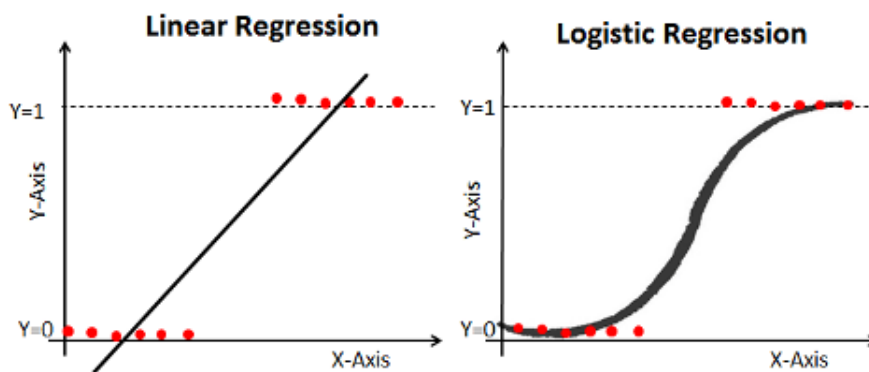


Figure 3.2: Distinguishing Between Linear and Logistic Regression

However, there are some considerations when using Binary Logistic Regression:

Linearity Assumption: Logistic regression postulates a linear connection between the predictor variables and the log-odds of the outcome. If the relationship is non-linear, additional transformations or interactions may be necessary.

Imbalanced Data: Logistic regression can be sensitive to imbalanced datasets where one class is significantly more prevalent than the other. Methods like over-sampling, under-sampling, or adjusting class weights can assist in dealing with this problem

Outliers: Outliers in the data can have a significant impact on logistic regression. [11]

Chapter 4

Dataset Description and Pre-processing

The dataset for Kerala is widely utilized and highly valuable for researchers, scientists, and policymakers. It covers a significant time period from 1907 to 2017, providing detailed information on monthly and annual rainfall measurements.

An important feature of this dataset is its inclusion of data on flood occurrences. This information allows for a comprehensive analysis of the relationship between rainfall patterns and the frequency of floods in each corresponding year. By studying these patterns over time, researchers can gain insights into the contributing factors behind flood events in Kerala.

Kerala, located in southern India, is the specific geographic focus of this dataset. With its monsoon climate and picturesque landscapes, the region experiences substantial rainfall throughout the year. Therefore, this dataset is particularly valuable for understanding the hydrological dynamics and flood occurrences specific to Kerala.

The Kerala dataset serves as a crucial resource for various studies and analyses related to rainfall trends, climate change, and flood risk assessment. Through the examination of this dataset, researchers can develop a deeper understanding of long-term rainfall patterns and variations in Kerala. This knowledge is essential for formulating effective strategies in flood mitigation, water resource management, and disaster preparedness in the region.

The Kerala dataset is highly significant as a widely used and comprehensive collection of data. Its coverage of rainfall indices, along with information on flood occurrences, makes it an invaluable resource for studying the hydrological dynamics and flood risks in Kerala.

The raw dataset is visually depicted in figure 4.1, as shown in the image below:

1	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL RAINFALL	FLOOD
2	KERALA	1901	28.7	44.7	51.6	160	174.7	824.6	743	357.5	197.7	266.9	350.8	48.4	3248.6	YES
3	KERALA	1902	6.7	2.6	57.3	83.9	134.5	390.9	1205	315.8	491.6	358.4	158.3	121.5	3326.6	YES
4	KERALA	1903	3.2	18.6	3.1	83.6	249.7	558.6	1022.5	420.2	341.8	354.1	157	59	3271.2	YES
5	KERALA	1904	23.7	3	32.2	71.5	235.7	1098.2	725.5	351.8	222.7	328.1	33.9	3.3	3129.7	YES
6	KERALA	1905	1.2	22.3	9.4	105.9	263.3	850.2	520.5	293.6	217.2	383.5	74.4	0.2	2741.6	NO
7	KERALA	1906	26.7	7.4	9.9	59.4	160.8	414.9	954.2	442.8	131.2	251.7	163.1	86	2708	NO
8	KERALA	1907	18.8	4.8	55.7	170.8	101.4	770.9	760.4	981.5	225	309.7	219.1	52.8	3671.1	YES
9	KERALA	1908	8	20.8	38.2	102.9	142.6	592.6	902.2	352.9	175.9	253.3	47.9	11	2648.3	NO

Figure 4.1: Raw Dataset

The Kerala dataset’s rainfall index is calculated based on the weather data collected and maintained by reputable institutions such as the Indian Meteorological Department and the Ministry of Earth Sciences. This index represents the amount of precipitation recorded in a specific area during a particular timeframe, relative to the long-term average. It serves as a valuable tool for assessing rainfall patterns and deviations from the expected norm.

In addition to rainfall measurements, the dataset includes categorical values, specifically 'True' and 'False', indicating whether a flood occurred in each corresponding year. These categorical values provide insights into the occurrence of floods in relation to the recorded rainfall levels.

To determine the normal monsoon conditions in Kerala, a range of +/- 19 per cent is applied to the average rainfall of 2039.6mm recorded during the June to September period. Rainfall exceeding this threshold is considered a potential indicator of flood occurrence.

The dataset preprocessing phase consists of three key stages:

4.0.1 Normalization

To prepare the dataset for analysis, normalization techniques, also known as feature encoding, are utilized. In this study, there are two attributes in the dataset that contain string-type data: 'sub-division' and 'flood'. These attributes undergo encoding procedures to transform their values.

For the 'flood' attribute, which has only two distinct values ('True' and 'False'), binary encoding is employed. This encoding method replaces 'True' with 1 and 'False' with 0, effectively representing the presence or absence of a flood.

Similarly, binary encoding is applied to represent the values of the 'sub-division' feature. Since 'sub-division' has only one unique value ('Kerala'), the entire feature is substituted with 1 to denote its presence.

The resulting dataset, after feature encoding, can be observed in Figure 4.2.

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL RAINFALL	FLOODS
0	1	1901	28.7	44.7	51.6	160.0	174.7	824.6	743.0	357.5	197.7	266.9	350.8	48.4	3248.6	1
1	1	1902	6.7	2.6	57.3	83.9	134.5	390.9	1205.0	315.8	491.6	358.4	158.3	121.5	3326.6	1
2	1	1903	3.2	18.6	3.1	83.6	249.7	558.6	1022.5	420.2	341.8	354.1	157.0	59.0	3271.2	1
3	1	1904	23.7	3.0	32.2	71.5	235.7	1098.2	725.5	351.8	222.7	328.1	33.9	3.3	3129.7	1
4	1	1905	1.2	22.3	9.4	105.9	263.3	850.2	520.5	293.6	217.2	383.5	74.4	0.2	2741.6	0
...
112	1	2013	3.9	40.1	49.9	49.3	119.3	1042.7	830.2	369.7	318.6	259.9	154.9	17.0	3255.4	1
113	1	2014	4.6	10.3	17.9	95.7	251.0	454.4	677.8	733.9	298.8	355.5	99.5	47.2	3046.4	1
114	1	2015	3.1	5.8	50.1	214.1	201.8	563.6	406.0	252.2	292.9	308.1	223.6	79.4	2600.6	0
115	1	2016	2.4	3.8	35.9	143.0	186.4	522.2	412.3	325.5	173.2	225.9	125.4	23.6	2176.6	0
116	1	2017	1.9	6.8	8.9	43.6	173.5	498.5	319.6	531.8	209.5	192.4	92.5	38.1	2117.1	0

117 rows x 16 columns

Figure 4.2: Dataset After Feature Encoding

4.0.2 Feature Selection

From this point onwards, the dataset no longer includes the sub-division attribute as it did not contribute directly to flood prediction. According to reference [15], the monsoon season in Kerala spans from June to October. Initially, the study identified the months corresponding to the monsoon season in Kerala. Subsequently, the Logistic Regression algorithm, along with SVC, KNN, and Decision Tree, was applied to the selected features. To evaluate the performance of the models using multiple features, feature set of the dataset for its entirety was utilized in the purpose of training. The target was to assess how well the models performed under different feature sets. This idea establishes a baseline to compare with other models trained on subsets of features, enabling a thorough analysis of the impact of feature selection on model accuracy.

The resulting dataset, after feature selection, can be observed in 4.3:

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL RAINFALL	FLOODS
0	KERALA	1901	28.7	44.7	51.6	160.0	174.7	824.6	743.0	357.5	197.7	266.9	350.8	48.4	3248.6	YES
1	KERALA	1902	6.7	2.6	57.3	83.9	134.5	390.9	1205.0	315.8	491.6	358.4	158.3	121.5	3326.6	YES
2	KERALA	1903	3.2	18.6	3.1	83.6	249.7	558.6	1022.5	420.2	341.8	354.1	157.0	59.0	3271.2	YES
3	KERALA	1904	23.7	3.0	32.2	71.5	235.7	1098.2	725.5	351.8	222.7	328.1	33.9	3.3	3129.7	YES
4	KERALA	1905	1.2	22.3	9.4	105.9	263.3	850.2	520.5	293.6	217.2	383.5	74.4	0.2	2741.6	NO
...
112	KERALA	2013	3.9	40.1	49.9	49.3	119.3	1042.7	830.2	369.7	318.6	259.9	154.9	17.0	3255.4	YES
113	KERALA	2014	4.6	10.3	17.9	95.7	251.0	454.4	677.8	733.9	298.8	355.5	99.5	47.2	3046.4	YES
114	KERALA	2015	3.1	5.8	50.1	214.1	201.8	563.6	406.0	252.2	292.9	308.1	223.6	79.4	2600.6	NO
115	KERALA	2016	2.4	3.8	35.9	143.0	186.4	522.2	412.3	325.5	173.2	225.9	125.4	23.6	2176.6	NO
116	KERALA	2017	1.9	6.8	8.9	43.6	173.5	498.5	319.6	531.8	209.5	192.4	92.5	38.1	2117.1	NO

117 rows x 16 columns

Figure 4.3: Dataset After Feature Encoding

4.0.3 Feature Engineering

In order to eliminate biases and ensure the dataset's suitability for model usage, the Standard Scaler technique was employed. In this method, the data is standardized by centering it around its mean and scaling it to have a unit variance.

To maintain training data without any bias, we used in this study a dataset that was split into testing and training sets using an 20:80 ratio. Additionally, the attributes underwent standardization using the Standard Scaler technique. This process ensures that every feature is transformed to a uniform scale, enabling the models to be trained on a fair and unbiased dataset.

Chapter 5

Performance Evaluation

The rainfall dataset utilized in this project has been consolidated into a CSV file. The analysis focuses on the rainfall patterns observed in Kerala, India, spanning from 1901 to 2017.

Prediction Model Classification

In this thesis, we employed four distinct classifiers, namely K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Decision Tree Classifier (DTC), and Binary Logistic Regression, for prediction purposes. Additionally, we implemented a stacked model that utilized the four aforementioned classifiers as base models. The meta-model used in the stacked model was Binary Logistic Regression. The stacked model combines the outputs of the base models to make the final prediction, leveraging the strengths and diverse perspectives of each classifier to improve accuracy.

- These models were first used to analyze the monsoon data specifically for the monsoon season in Kerala, which encompasses the months of June, July, August, September, and October [17]. Consequently, only the rainfall data pertaining to these months was considered for the analysis. The results obtained from applying the models are presented in the table below.

Predictive	Models Accuracy	Standard Deviation
K-Nearest Neighbors (KNN)	83.4	0.158
Support Vector Classifier (SVC)	86.4	0.146
Decision Tree Classifier (DTC)	78.3	0.217
Binary Logistic Regression	83.6	0.164
Stacked Generalization	84.8	0.159

Table 5.1: Outcomes of the models applied using Monsoon Rainfall Data.

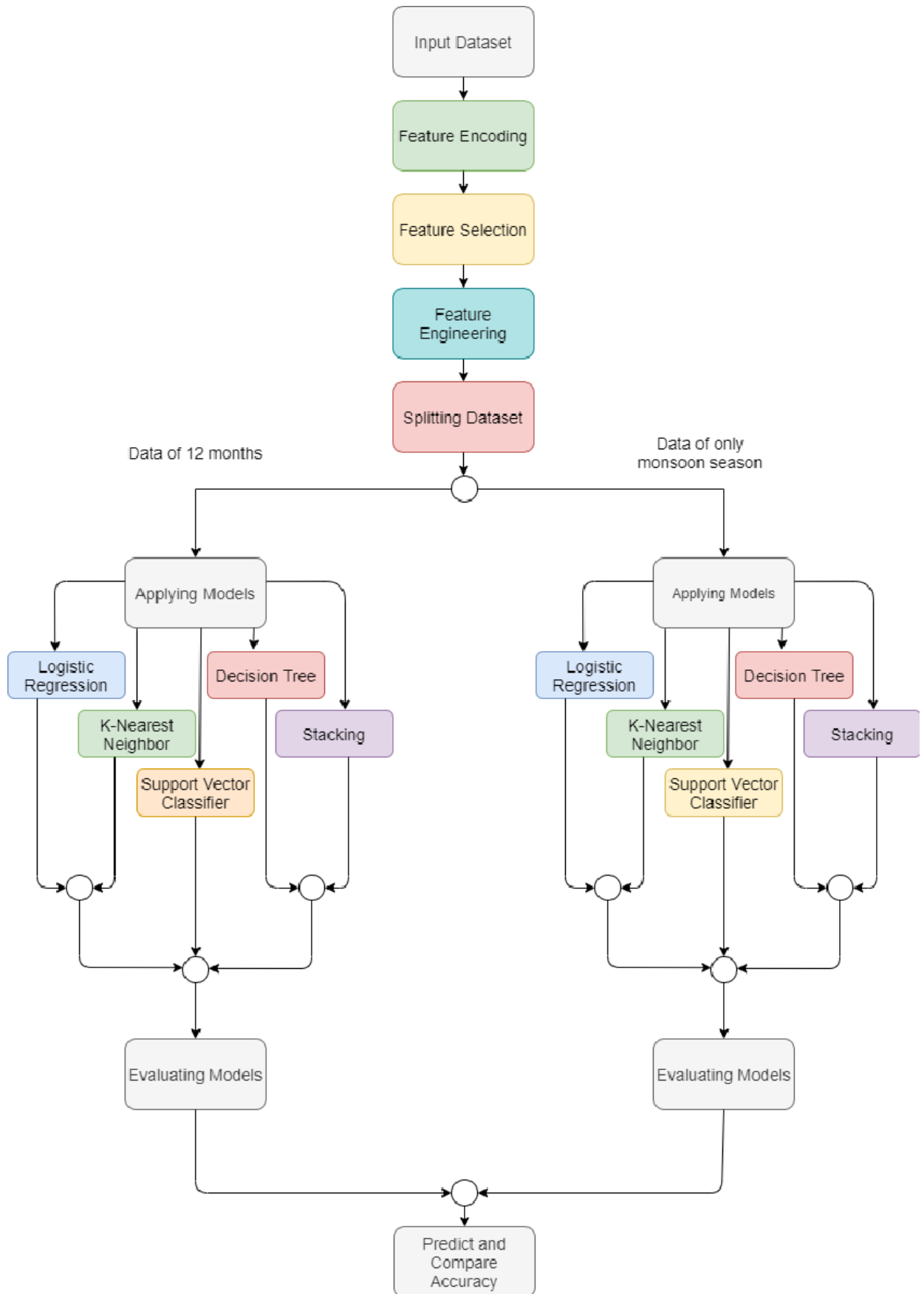


Figure 5.1: Our Proposed System is Illustrated

Standalone Model Accuracies on the Monsoon Rainfall Data

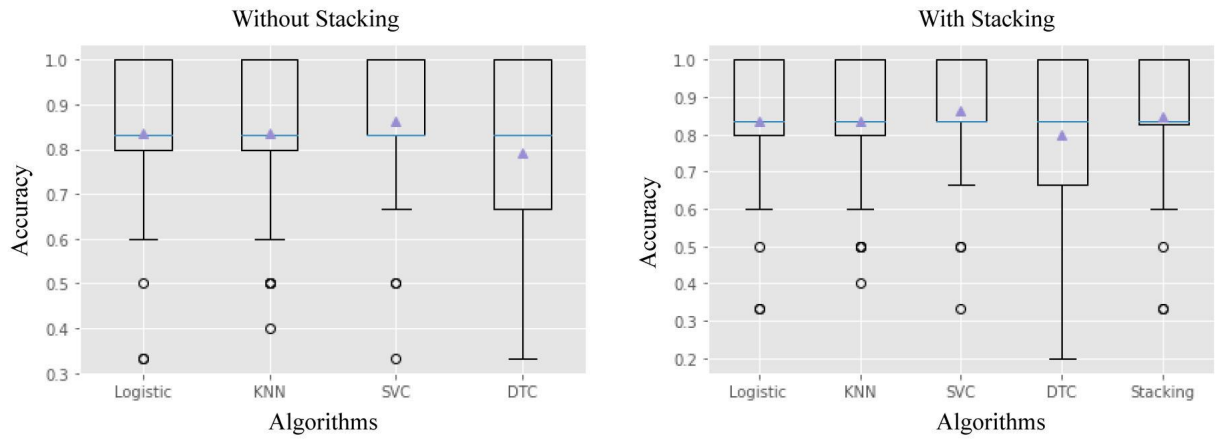


Figure 5.2: Whisker box plot for Standalone Model Accuracy on the Monsoon Rainfall Data

- Below is the table presenting the results obtained by utilizing the models on monsoon data from every month of the year. The objective is to improve accuracy by employing the same set of models. Shown in table 5.2:

Predictive	Models Accuracy	Standard Deviation
K-Nearest Neighbors (KNN)	74.6	0.172
Support Vector Classifier (SVC)	90.6	0.111
Decision Tree Classifier (DTC)	77.2	0.772
Binary Logistic Regression	93.0	0.103
Stacked Generalization	93.3	0.098

Table 5.2: Outcomes of the models applied using twelve months Rainfall Data

Upon examining the data presented in both tables, it is evident that utilizing rainfall data from all 12 months of the year yielded better accuracy for all models, except the Decision Tree Classifier (DTC), in comparison with using only the rainy season months. This suggests that predicting floods relies not only on rainfall during the monsoon season but also on precipitation throughout the entire year, alongside other contributing elements.

Among the individual models, the Binary Logistic Regression demonstrated the highest level of accuracy. Notably, Stacked Generalization outperformed all the individual models in terms of accuracy, exhibiting a lower standard deviation when

Standalone Model Accuracies on 12 Months Rainfall Data

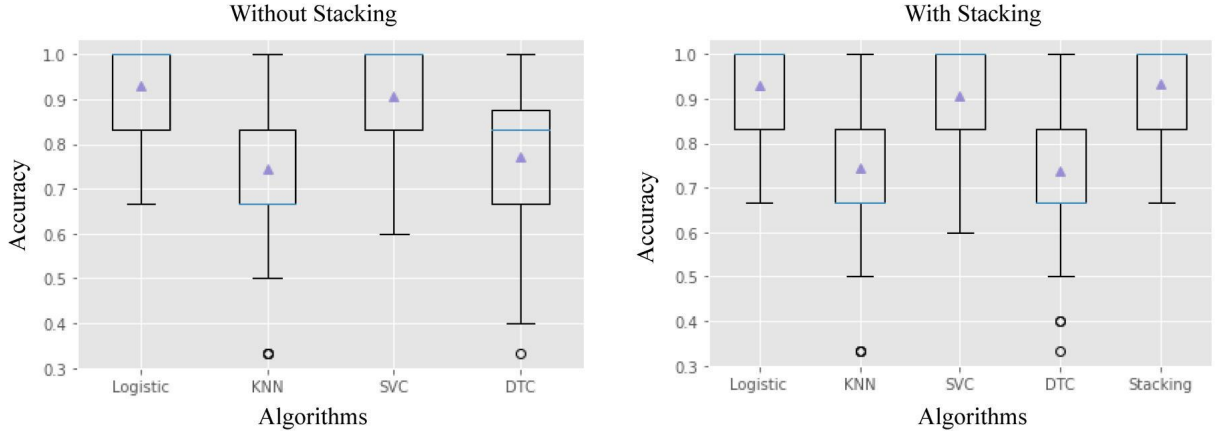


Figure 5.3: Whisker box plot for Standalone Model Accuracy on the twelve months Rainfall Data

Authors	Model Used	Accuracy
Ruslan et al.	NNARX	73.54
Ruslan et al.	NARX	87
S. Sankaranarayanan et al.	SVM	85.57
S. Sankaranarayanan et al.	KNN	85.73
S. Sankaranarayanan et al.	Naive Bayes	87.01
S. Sankaranarayanan et al.	DNN	91.18
Tanvir et al.	Stacked Generalization	93.3

Table 5.3: Result comparison of different works with our work

considering a year of data in terms of rainfall. Consequently, it is to be concluded that Stacked Generalization proves to be a more effective approach for flood prediction than any single algorithm models.

Chapter 6

Conclusion and Future Works

The objective of this research was to enhance flood prediction using meteorological data by developing an ensemble machine learning model. The study compared the performance of different models, including KNN, SVC, decision trees, and logistic regression, to identify the most accurate and precise approach for flood predictions.

The results indicated that the ensemble model demonstrated superior accuracy and precision compared to the individual models. The ensemble model combines the predictions of multiple models to produce a final prediction, leveraging the strengths of each component. This integration of multiple models contributed to improved flood prediction performance.

Furthermore, the study found that the ensemble model outperformed previous studies in flood prediction that employed machine learning models. This suggests that the ensemble approach presented in this research represents a significant advancement in flood prediction compared to existing literature.

Moreover, this study highlights the advantages of utilizing an ensemble machine learning model for flood prediction. The ensemble model, by combining multiple models, achieved better accuracy and precision than individual models and surpassed the performance of previous studies in the field of flood prediction using machine learning.

Future research in flood prediction can explore several avenues to enhance the accuracy of the proposed flood prediction model.

One potential area for improvement is the incorporation of additional data sources. Researchers can consider integrating data on soil moisture and land use to provide a more comprehensive understanding of the hydrological processes influencing floods. By including these factors, the flood prediction model can achieve better accuracy and precision in its predictions.

Another important aspect for future investigations is the evaluation of the model's performance across different temporal and spatial scales. Assessing the model's effectiveness under various time periods and geographical regions will help determine

its robustness and generalizability. This analysis is crucial to ensure the model's applicability in diverse flood-prone areas and varying climatic conditions.

Exploring alternative ensemble techniques represents another promising direction for future research. Researchers can explore ensemble methods like bagging and boosting to improve the accuracy of the flood prediction model. Comparing the performance of different ensemble techniques with the existing model can identify the most effective approach for enhancing prediction accuracy.

Lastly, future research should focus on developing practical applications based on the proposed model. This could involve the creation of an online flood prediction system that utilizes real-time meteorological and hydrological data. Such a system would provide timely and accurate flood warnings to local communities and authorities, enabling proactive decision-making and effective disaster response.

In summary, future research in flood prediction can involve the incorporation of additional data sources, evaluation across different temporal and spatial scales, exploration of alternative ensemble techniques, and the development of practical applications such as an online prediction system. By pursuing these avenues, researchers can advance the accuracy and effectiveness of flood prediction models, leading to improved flood management strategies and better preparedness for flooding events.

Bibliography

- [1] C. Sivapragasam, S.-Y. Liong, and M. F. K. Pasha, “Rainfall and runoff forecasting with SSA–SVM approach,” *Journal of Hydroinformatics*, vol. 3, no. 3, pp. 141–152, Jul. 2001. DOI: 10.2166/hydro.2001.0014. [Online]. Available: <https://doi.org/10.2166/hydro.2001.0014>.
- [2] J. Adamowski, H. F. Chan, S. O. Prasher, B. Ozga-Zielinski, and A. Sliusarieva, “Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in montreal, canada,” *Water Resources Research*, vol. 48, no. 1, Jan. 2012. DOI: 10.1029/2010wr009945. [Online]. Available: <https://doi.org/10.1029/2010wr009945>.
- [3] K. Haddad and A. Rahman, “Regional flood frequency analysis in eastern australia: Bayesian GLS regression-based methods within fixed region and ROI framework – quantile regression vs. parameter regression technique,” *Journal of Hydrology*, vol. 430-431, pp. 142–161, Apr. 2012. DOI: 10.1016/j.jhydrol.2012.02.012. [Online]. Available: <https://doi.org/10.1016/j.jhydrol.2012.02.012>.
- [4] B. Jongman, P. J. Ward, and J. C. Aerts, “Global exposure to river and coastal flooding: Long term trends and changes,” *Global Environmental Change*, vol. 22, no. 4, pp. 823–835, Oct. 2012. DOI: 10.1016/j.gloenvcha.2012.07.004. [Online]. Available: <https://doi.org/10.1016/j.gloenvcha.2012.07.004>.
- [5] K. Aziz, A. Rahman, G. Fang, and S. Shrestha, “Application of artificial neural networks in regional flood frequency analysis: A case study for australia,” *Stochastic Environmental Research and Risk Assessment*, vol. 28, no. 3, pp. 541–554, Jul. 2013. DOI: 10.1007/s00477-013-0771-5. [Online]. Available: <https://doi.org/10.1007/s00477-013-0771-5>.
- [6] J. Abbot and J. Marohasy, “Input selection and optimisation for monthly rainfall forecasting in queensland, australia, using artificial neural networks,” *Atmospheric Research*, vol. 138, pp. 166–178, Mar. 2014. DOI: 10.1016/j.atmosres.2013.11.002. [Online]. Available: <https://doi.org/10.1016/j.atmosres.2013.11.002>.
- [7] A. K. Lohani, N. Goel, and K. Bhatia, “Improving real time flood forecasting using fuzzy inference system,” *Journal of Hydrology*, vol. 509, pp. 25–41, Feb. 2014. DOI: 10.1016/j.jhydrol.2013.11.021. [Online]. Available: <https://doi.org/10.1016/j.jhydrol.2013.11.021>.

- [8] P. Costabile and F. Macchione, “Enhancing river model set-up for 2-d dynamic flood modelling,” *Environmental Modelling & Software*, vol. 67, pp. 89–107, May 2015. DOI: 10.1016/j.envsoft.2015.01.009. [Online]. Available: <https://doi.org/10.1016/j.envsoft.2015.01.009>.
- [9] B. Kim, B. F. Sanders, J. S. Famiglietti, and V. Guinot, “Urban flood modeling with porous shallow-water equations: A case study of model errors in the presence of anisotropic porosity,” *Journal of Hydrology*, vol. 523, pp. 680–692, Apr. 2015. DOI: 10.1016/j.jhydrol.2015.01.059. [Online]. Available: <https://doi.org/10.1016/j.jhydrol.2015.01.059>.
- [10] M. S. Gizaw and T. Y. Gan, “Regional flood frequency analysis using support vector regression under historical and future climate,” *Journal of Hydrology*, vol. 538, pp. 387–398, Jul. 2016. DOI: 10.1016/j.jhydrol.2016.04.041. [Online]. Available: <https://doi.org/10.1016/j.jhydrol.2016.04.041>.
- [11] Y. Wu, Q. Zhang, Y. Hu, *et al.*, “Novel binary logistic regression model based on feature transformation of XGBoost for type 2 diabetes mellitus prediction in healthcare systems,” *Future Generation Computer Systems*, vol. 129, pp. 1–12, Apr. 2022. DOI: 10.1016/j.future.2021.11.003. [Online]. Available: <https://doi.org/10.1016/j.future.2021.11.003>.
- [12] A. H. Ali, M. A. Mohammed, R. A. Hasan, M. N. Abbod, M. S. Ahmed, and T. Sutikno, “Big data classification based on improved parallel k-nearest neighbor,” *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 21, no. 1, p. 235, Feb. 2023. DOI: 10.12928/telkomnika.v21i1.24290. [Online]. Available: <https://doi.org/10.12928/telkomnika.v21i1.24290>.
- [13] J. Park, Y. Choi, J. Byun, J. Lee, and S. Park, “Efficient differentially private kernel support vector classifier for multi-class classification,” *Information Sciences*, vol. 619, pp. 889–907, Jan. 2023. DOI: 10.1016/j.ins.2022.10.075. [Online]. Available: <https://doi.org/10.1016/j.ins.2022.10.075>.
- [14] R. Singh, M. K. Singh, and D. K. Jhariya, “New framework for implementation of decision tree classifier,” *Intelligent Systems and Smart Infrastructure: Proceedings of ICISSI 2022*, p. 358, 2023.
- [15] [Online]. Available: <https://www.rma.usda.gov/en/Policy-and-Procedure/Insurance-Plans/Rainfall-Index>.
- [16] *Flood vulnerability and buildings’ flood exposure assessment in a densely urbanised city: Comparative analysis of three scenarios using a neural network approach - Natural Hazards — doi.org*, <https://doi.org/10.1007/s11069-022-05336-5>, [Accessed 18-Jun-2023].
- [17] *Monsoon likely to arrive in Kerala on May 31, says IMD — hindustantimes.com*, <https://www.hindustantimes.com/india-news/monsoon-likely-to-arrive-in-kerala-on-may-31-says-imd-101622366991176.html>, [Accessed 25-Jun-2023].