Ethical AI for Recruitment: A Hybrid Machine Learning Approach with Fairness Metrics and Explainability Tools

by

TANAY PAUL
24241075
MIR SAJIN MAHMUD
21301247
TANZIM HAQUE
21301115
YASIR ARAFAT
21101017
MAHMUD FERDOUS
21301401

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

Department of Computer Science and Engineering
Brac University
October 2025

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing the 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:

TANAY PAUL
24241075

MIR SAJIN MAHMUD
21301247

TANZIM HAQUE
21301115

YASIR ARAFAT
21101017

MAHMUD FERDOUS
21301401

Approval

The thesis titled "Ethical AI for Recruitment: A Hybrid Machine Learning Approach with Fairness Metrics and Explainability Tools" submitted by

1. TANAY PAUL (24241075)
2. MIR SAJIN MAHMUD (21301247)
3. TANZIM HAQUE (21301115)
4. YASIR ARAFAT (21101017)
5. MAHMUD FERDOUS (21301401)

Of Summer, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on October 10, 2025.

Examining Committee:

Supervisor:
(Member)

Mr. Md. Tawhid Anwar

Senior Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Mr. Md. Sabbir Ahmed

Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

Artificial Intelligence (AI) is transforming the process of recruitment making it efficient, consistent, and more accurate in its decisions; however, there are still ethical concerns like bias and obscurity. This paper constructs an Ethical AI Recruitment Framework and based on a hybrid machine learning model that incorporates a combination of Logistic Regression, Decision Tree, and Random Forest with soft voting to achieve a balance between predictive accuracy and fairness and the final accuracy was measured 97.4% using the hybrid model. Demographic Parity (DP), Equalized Odds (EO), and a Gender-Flip Test were used to evaluate the system and a +5% threshold recalibration was applied to minimize residual bias. The findings demonstrate high technical and ethical results with an accuracy of 96% and a Fairness Index of 0.87, Transparency Score of 0.01 and Bias Volatility of 0.027, which ensures reliability and fairness. Explainability SHAP and LIME demonstrated that skills, education and experience were the most influential features in hiring decisions and that gender and age became insignificant after applying these features. The Selection Scoreboard (80% model prediction and 20% resumes strength) of the framework offers a clear ranking of the candidates. In general, this study shows that fairness, transparency, and accountability can be appropriately incorporated into AI-based recruitment to provide a repeatable model of reliable and bias-resistant decision making in the HR field.

Keywords: Artificial Intelligence (AI); Human Resource Management (HRM); Recruitment and Selection; Supervised Machine Learning; Bias Mitigation; Ethical Standards; Transparency and Accountability; Post-Processing Fairness Adjustment; LIME (Local Interpretable Model-Agnostic Explanations).

Acknowledgement

Firstly, all praise to the Great Almighty for whom our thesis has been completed without any major interruption.

Secondly, to our Supervisor, Mr. Md. Tawhid Anwar and Co-Supervisor, Mr. Md. Sabbir Ahmed sir for their kind support and advice in our work. They helped us whenever we needed help.

Thirdly, we would like to thank all the faculty members and staff for providing us with such a wonderful study environment where we could develop ourselves as well as this research properly.

And finally to our parents without their support, it may not be possible. With their kind support and prayer, we are now on the verge of our graduation.

Table of Contents

List of Figures

List of Tables

Nomenclature

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

AI Artificial Intelligence

DT Decision Tree

HRM Human Resource Management

IWC Inverse Weight Clustering

KNN K-Nearest Neighbors

$LIME$ Local Interpretable Model-Agnostic Explanations

LR Logistic Regression

ML Machine Learning

NB Naive Bayes

$R\&S$ Recruitment and Selection

RF Random Forest

$SHAP$ Shapley Additive Explanations

Chapter 1

Introduction

The rapid expansion of technologies has improved the communication and information world, which has created an impact in business and placed it in a new and challenging condition (Akerkar, 2019)in[3]. Many companies and industries try to enhance their skills that help them to remain in a competitive position. An organisation's ability to make decisions determines its success or failure (Porter et al., 1985)in[1]. At present, it is quite tough to make important business decisions, and for this, people in higher positions like managers or CEOs investigate many resources like information, data, knowledge, and wisdom in order to make important business decisions. With the advancement of the modern world, technology can help the people at work access and use the available resources. The recruitment and selection (R&S) process is one business area that has been influenced by technological advancements. There are several types of developments in technology, like online recruitment, visa processing, and tracking systems for applicants, which have created a significant impact on the recruitment and selection process (Woods et al., 2019)in[5]. As per Upadhyay & Khandelwal (2018)in[2], in 2018, one of the most common hiring trends was the adoption of artificial intelligence.

Recruitment is not only about matching skills with job descriptions it is about identifying individuals, their potential, and the possibilities that the individuals best suit. As more organizations are using algorithms to assist them in the hiring process, the priorities have changed.to efficiency and speed to justice and responsibility. AI systems can inadvertently take up the historical data biases, which will mirror the inequalities that already exist.gender, background or education. Even when a model has been found to work well in predicting who. It can be still used to support the unfair patterns even after being hired unless it is carefully designed and monitored.That is why it is extremely important to establish systems that will appreciate ethical responsibility as much as.accuracy, making all predictions to promote fairness and inclusion. The process of recruitment and selection is not new; it has its origins as far as the World War II.era, in which first organized hiring began to take shape. Over the years, the process has evolved with the social and technological advancement. What once relied on putting print advertisements in the newspapers has changed to online and AI-driven. systems. Recruitment today is more rapid and digital and is in an ever-changing state. keep up with the times of the present day.

This study takes a direct challenge of that. Ethical is introduced in the study. Intelligent Recruitment System, which is an AI-powered solution that is aimed at integrating tech-

nical precision with. moral accountability. It is a hybrid machine learning model that combines. Logistic Regression, Decision Tree, and the Random Forest with a soft-voting. team, to make decisions that are well-founded and well-balanced. Rather than focusing Predictive power is the sole aspect that the system contains a fairness assessment, which is based on the Demographic Parity (DP) and Equalized Odds (EO) measures. These fairness checks are supplemented with a Gender-Flip Test, which is meant to ensure that the model decisions are not biased with regard to various demographic groups. Once such small imbalances may occur, a post-processing fairness adjustment, a minor +5% threshold recalibration is implemented so that all candidates can be treated fairly.

As technology is evolving it leaves an imprint, whether good or bad on. All industries, and in the same case, the recruitment and human resources are the ones that are impacted management (HRM) industry. The field of artificial intelligence (AI) has produced a lot of influence the recruitment industry, and its integration into the primary recruitment process has been rather unfortunate. transformed and evolved the whole process. Application of artificial intelligence has there are numerous good aspects, such as efficiency improvement and cost decrease of the company through the recruitment process through automated recruitment process. The AI powered accessories such as screening resumption, and candidate evaluation programs can reduce the time of recruitment to as much as 75 percent. But there can always be some problems, like ethical concerns over the system, that can cause biases in training data, especially regarding gender and ethnic disparities. In our regular system, there is a cost associated with these advantages and opportunities of the following groups. The application of algorithms may create ethical concerns about bias, data privacy, and the elimination of the human element. Our research explores how artificial intelligence (AI) can be used in the recruitment process by detecting its possibility of enhancing the performance of human resources (HR) processes, ethical issues, and implementation issues.

The system also helps to make the decision making process more transparent.explainable tools such as SHAP and LIME, which provide a clear picture as to why a certain candidate was favored or rejected. The AI model becomes a black box, rather than being obscure.a rational partner- one that is able to provide real evidence of its conclusions. Recruiters can visually represent the characteristics of most features, which are skills, level of education, experience.influenced a specified outcome, creating trust between the users and the system.

One of the strongest parts in this work is the establishment of a Selection Scoreboard, which is a combination of algorithm prediction using human intuition. The applications are rated using a Selection Score comprised of 80% model prediction, and 20% resumé strength. This helps to balance the AI's trust of the candidate based on his or her qualifications in the real world. The Selection Scoreboard at the time.ranks the applicants in an open and clear way, demonstrating the manner every choice was made in a straightforward and clear manner.

The primary objective of this study is to come up with an AI-based recruitment system that automates and provide an efficient method for screening and selection of candidates through the Kruskal and Knapsack algorithms. The Kruskal algorithm is in the ranking and selection process. In the ranking and selection process, the Kruskal algorithm is

used to prioritize candidates based on their skills and qualifications, while the Knapsack algorithm helps balance suitability scores with job capacity constraints. The combination of these approaches seeks to establish a fair, effective, and data-driven way of hiring The paper then investigates the remaining part of the research about the implementation of artificial intelligence in recruitment. The interviews of the candidates and the qualitative research systems are also shown. Lastly, it discusses the lackings of the paper and also the indications for the company and further research in the future.

Essentially, this paper seeks to reveal that AI can optimize the recruitment process with no compromising ethics. Combining fairness, transparency, and explainability into a single entity.system, the given model shows that automation can be effective and powerful at the same time. The outcome is that the process of recruiting becomes quicker, more objective. and–above all things fairer to all worthy candidates.

1.1 Research Problem

Organisations are being pressurised to enhance the efficiency, objectivity and their recruitment processes have been very fast-changing environment.Such ways of recruitment are constructed on human experiences, and in spite of that they are either lauded, and are time-consuming in the subject to biases as we have explained.presented above and very subjective. Such restrictions bring about inappropriate recruitment, longer decision time and more expensive. Based on the information in [3], Automated recruiting tools were created with an aim of offering a service that is easy and simplified enhances the productivity of hiring. However, what they have demonstrated are bias acts because of biases within training data and models. A key challenge lies in the biases prevalent in traditional recruitment practices, which frequently lead to discrimination based on gender, ethnicity, or disability, hindering diversity and inclusivity in the workplace.

The primary focus of our research is how artificial intelligence (AI) functions in the hiring process. It mentions the ethical concerns like privacy and data protection in our work process. Despite the capability of AI, human presence is vital in the recruitment process, as humans are the ultimate mastermind. To gather more information and recruit efficient candidates, interviews with experienced recruiters were conducted, which provided qualitative insights. The main objective of our research is to implement AI in the recruitment process to make human life easier. Moreover, it will also maximize the effectiveness of the human resource process and identify challenges associated with AI implementation. The main focus is to improve fairness in machine learning applications and it emphasizes the need for unbiased AI development. It also highlights the limitations in current research methodologies, and because of this, a lack of sufficient strategies for mitigating biases can be found and a lack of empirical evidence on bias impacts in decision-making. The primary focus here is on hiring and how we can use unbiased AI to hire the right candidate for any given role. We will also point out the possible biases in the selection process and how cognitive biases can be overcome in AI systems, as this would help to provide an understanding towards the sharing of knowledge on transparency about AI. One way to overcome these problems is by incorporating machine learning (ML) and artificial intelligence (AI) into the hiring process. AI-driven systems promise significant improvements in candidate screening, selection, and overall hiring efficiency, with au-

tomation tools capable of reducing time-to-hire by up to 75%, cutting recruitment costs by 70%, and improving candidate quality by 50% (Kaushik et al., 2023). Many companies work with AI to hire qualified applicants rapidly and effectively, companies such as **Vodafone** (with its AI-based application platform) and **Unilever** have already adopted AI systems to streamline parts of the hiring process. According to Lena and Kriebitz (2023), these systems automate tasks like résumé screening, candidate ranking, and video interview analysis, allowing organizations to handle large volumes of applications more efficiently and fairly. Studies, including Yanamala et al. (2022), have demonstrated that those tools do not only save time, but also enhance productivity and impartiality in recruitment. Nevertheless, research has its benefits despite this it has also shown the bias in AI recruiting tools can still exist. A well-known example is an AI recruiting system used by Amazon, but which unintentionally discriminated against applicants of the feminine gender.

Bias in AI systems can appear at several stages of the recruitment process, including candidate identification, ranking, and final selection. Recognizing where these biases occur is essential for developing fair and accountable hiring models. AI models that learn from imbalanced data can produce biased rank orders simply because of the way it weights its results-meaning companies could risk recruiting hires unproven in their jobs (if not outright poor candidates) and suffer high turnover rates. In addition, AI systems commonly suffer from a lack of transparency and are sometimes referred to as "black-box" solutions because the decisions made by cryptic algorithms in these systems can be hard for organizations to interpret or defend. However, the question is how to ensure that these systems are answerable and at the same time making sure that they are privacy minded to violate GDPR requirements. The following paper is presented by Yanamala at el. (2022) to provide a gender-specific adaptive framework to on-the-fly detect and correct biases. and ethnicity. This involves the application of live detection software that has an inbuilt decision compensation to be made on a fair basis between the various demographic groups. Hunkenschroer et al., On the same note, as Hunkenschroer at el.(2022) does, researches the prevalence of biases.in AI recruitment tools and provides several mitigation measures, such as the reduction of pre-processing, in-processing and post-procedures can be used. discrimination of other kinds of sensitive aspects like gender and ethnicity. Both lines of study indicate the need to have complex, evolving policies, which are aimed at a. widespread range of biases so that AI-enabled hiring will be more transparent and. just; it is also the reflection of the larger demand that requires companies to demonstrate themselves. be responsible and adhere to existing legal requirements. Moreover, though AI-derived such algorithms like the Kruskal algorithm of candidate matching or the Knapsack resource optimization method have potential in improving recruitment the level of productiveness, to what extent they apply to the practical HR context have not been yet determined. The existing methods of bias mitigation (example: pre-processing, in-processing and post-processing techniques) can be more refined to prove efficient for a wider range of industries & job roles.

As mentioned earlier, developing and evaluating an AI-powered recruitment system that leverages the Kruskal algorithm, the Knapsack approach, inverse weight clustering (IWC), and decision tree (DT) algorithms, to develop an AI-powered recruitment system that can mitigate biases, enhance candidate diversity, and improve transparency in decision-making. Useful ethical algorithmic approaches

drastically lower the dangers related to AI systems' innate biases (Rodgers et al., 2023). According to [4], the study also aims to identify strategies that ensure compliance with both ethical principles and legal requirements, such as the **GDPR**, while helping organizations improve efficiency and reduce time-to-hire. As highlighted in [6] and [7], one of the main challenges of using "black-box" AI algorithms in recruitment is the difficulty of understanding and explaining their decisions. Therefore, transparency and accountability must be central to any ethical AI system. To address this, the framework applies **bias mitigation techniques**—including pre-processing, in-processing, and post-processing methods—to minimize biases related to gender, ethnicity, or other sensitive characteristics. This approach not only strengthens fairness and regulatory compliance but also promotes trust, supporting more responsible and transparent AI-driven recruitment practices.

But there come many problems, like qualitative findings cannot predict broader popular results and as the exploration of AI's ethical implications in recruitment is limited, there is need for research on recruiter perceptions of AI. Cost benefit analysis in AI used in the recruitment process should also be worried about as a matter of fact.

So, there will be many challenges, like ethical concerns, human factor involvement, AI's full integration, etc., and by overcoming all of these, we can reach our goal of implementing AI in the hiring process, considering all positive and negative impacts.

1.2 Research Objectives

The main aim of this study is to develop an **Ethical AI-powered Recruitment System** that promotes fair, transparent, and data-driven hiring decisions.

1. **Model Development:** Create a hybrid ensemble model that combines Logistic Regression, Decision Tree, and Random Forest to achieve accurate and balanced candidate predictions.

2. **Fairness Evaluation:** Assess and reduce bias using metrics such as Demographic Parity, Equalized Odds, and Gender-Flip tests to ensure fairness across demographic groups.

3. **Transparency and Explainability:** Make the decisions of the model more explainable using SHAP and LIME to demonstrate how the various candidate features affect it.

4. **Fairness Adjustment:** Use calibration and post-processing techniques to make the prediction fair to all the candidates.

5. **Selection Scoreboard:** Develop a model probability based resumé strength based selection system that is transparent and clearly and objectively based upon merit.

1.3 Workflow

The workflow starts with the data preparation phase where the information about the candidates is obtained. the Kaggle HR data is gathered and narrowed. The step encompasses the cleaning, encoding,and standardizing the data and eliminating outliers or irrelevant variables. Notably, sensitive attributes such as gender were not included in the model training process and were utilized only to be considered fairly later. This made sure that the predictions of this model were purely based on the pertinent qualifications and independent of demographic factors. Feature engineering then converts the use of text based skill sets and educational levels into the structured numeric values so that the model will only be learning based on merit-based inputs but not a bias one identifiers.

The data has been prepared, and then transferred to model development stage. The dataset was divided into training and testing weight 70:30. There are three fundamental algorithms- Logistic Regression, Decision tree and random forest- A hybrid soft-voting was used. This ensemble became more consistent by combining their personal advantages.and precise forecasts on hiring candidates. The system goes into a state of model training. fairness audit stage, in which it verifies bias based on Demographic Parity, Equalized Odds. and a Gender-Flip test. Should an imbalance be detected, a minor action after processing will be made. (e.g., a +5 percent shift of the threshold) aids in removing under-selection of minority groups without modifying the basic model.

Lastly, the explainability and decision stage will make sure that all predictions are possible.understood and justified. SHAP and LIME techniques disclose the factors, including skills, that are significant.Every decision is influenced most by education, or experience. A transparency audit is selected on the leading 1 percent of candidates to ensure that there is no sensitive feature at work.outcomes. The system subsequently generates a Selection Scoreboard to rank out the applicants in terms of.using model probability and resumé strength to create a fair, explainable, andrecruiters can get a shortlist based on data. This ordered pipeline, which is represented in the workflow.diagram, provides a performance, fairness, and interpretability balancing loop across.all phases of the recruitment proce.

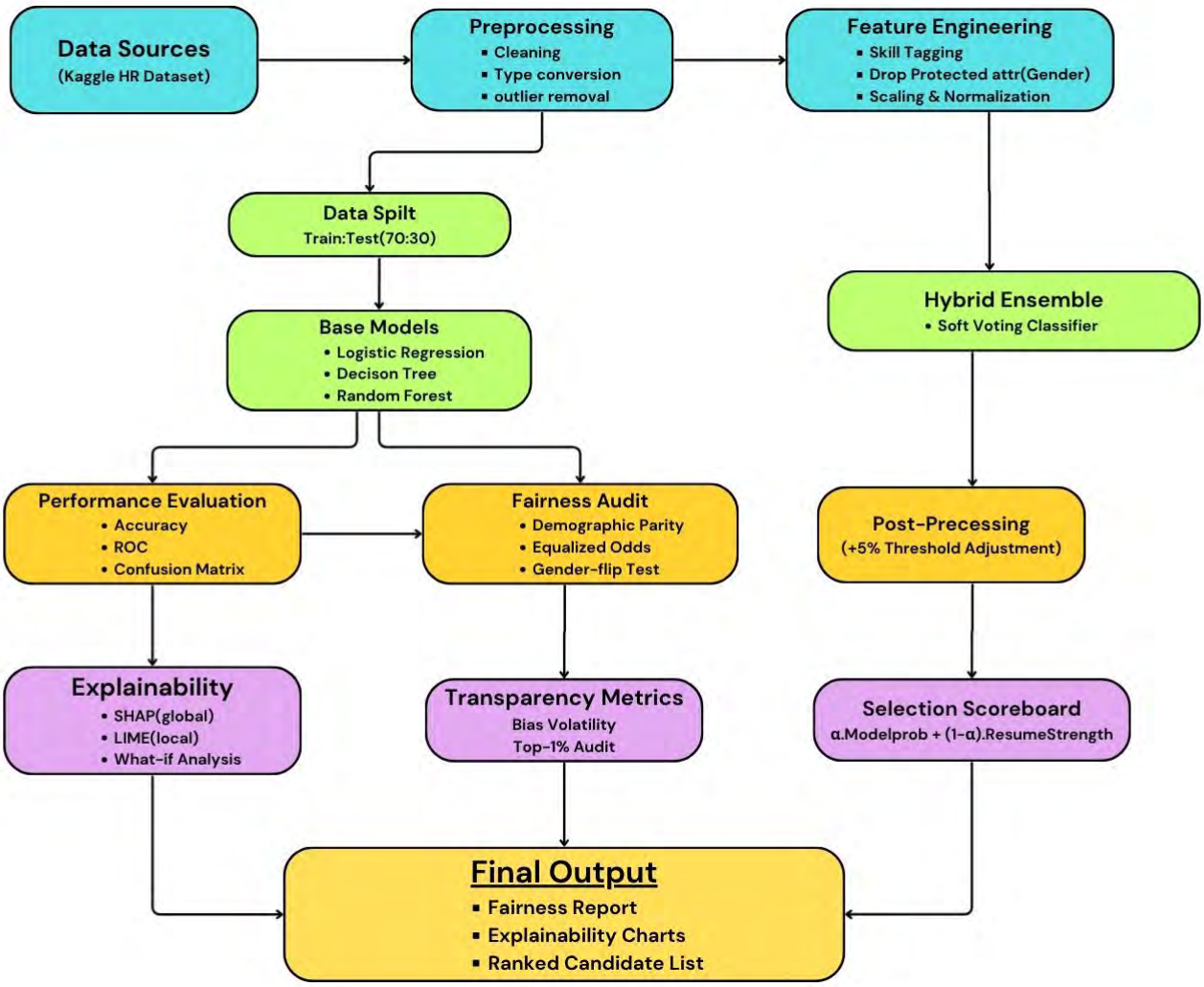


Figure 1.1: Workflow of the Fair Recruitment Model

This Figure 1.1 illustrates the complete workflow of the proposed AI-driven ethical recruitment system. Starting from data collection and preprocessing, it progresses through model training, fairness auditing, and post-processing adjustments. The pipeline concludes with explainability and transparency evaluations, leading to a final output that ensures fair, interpretable, and trustworthy candidate ranking.

Chapter 2

Literature Review

Kaushik et al. (2023)in[11] suggested a methodology based upon leveraging Kruskal's and Knapsack algorithms for improved functioning of the HR department, specifically during all rounds related to recruitment as well as pre-screening. The system is designed to streamline recruitment, reduce bias and make agent selection more efficient. The traditional recruitment process is described as inefficient, time-consuming, and prone to subjective biases, often resulting in poor hiring decisions. This paper addresses these issues by proposing an AI-enhanced approach to improve candidate matching and streamline the selection process. It introduces the "BTNT" test, designed to assess both technical skills and emotional intelligence. The proposed system has shown a reduction of about 75% on hiring time and 70% cost per hire with being able to increase candidate quality by 50% and diversity in the pipeline by up to 35%. While the approach outperforms traditional recruitment processes by addressing issues like subjectivity, inefficiency, and biases, further testing with real-world data is necessary to validate its performance, as its efficiency has so far only been proven through simulations using hypothetical HR recruitment scenarios.

Mujtaba and Mahapatra (2019)in[4] review the ethical challenges in AI-driven recruitment processes, focusing on bias, fairness, and transparency. They analyze how bias and discrimination in AI-based recruitment can arise from biased training data and algorithmic decisions. Although no specific dataset is mentioned, the paper references cases like Amazon's AI recruitment tool, which was discontinued due to gender bias against female candidates. The workflow starts with the data preparation phase where the information about the candidates is obtained. the Kaggle HR data is gathered and narrowed. The step encompasses the cleaning, encoding,and standardizing the data and eliminating outliers or irrelevant variables. Notably, sensitive attributes such as gender were not included in the model training process and were utilized only to be considered fairly later. This made sure that the predictions of this model were purely based on the pertinent qualifications and independent of demographic factors. Feature engineering then converts the use of text based skill sets and educational levels into the structured numeric values so that the model will only be learning based on merit-based inputs but not a bias one identifiers.

Peña et al. (2023)in[12] explored the development of human-centered machine learning applications, with a focus on automated recruitment processes. They highlighted the importance of ensuring accountability, fairness, and justice in AI-driven decision-making. A central concern of the study is the bias present in AI algorithms, particularly within

the recruitment field, where such biases often lead to demographic imbalances. Using the FairCVdb dataset, which consists of 24,000 synthetic profiles, the research examined recruitment bias through the Fairness Test methodology. This approach assessed how sensitive information is extracted via machine learning techniques. The results confirmed the presence of biases in candidate evaluation models, particularly in terms of gender and race. The study sheds light on algorithmic discrimination by examining bias within employment data but also emphasizes the need for further studies to broaden the testing environment and enhance bias detection methods. Overall, this research provides a framework for addressing bias in AI recruitment systems and offers valuable insights for future investigations into human-centered AI.

Tian et al. (2022)in[14] presented a new machine learning recruitment system for HR management in business process management systems that uses advanced techniques to enhance recruitment efficiency. The problem of resumes holds, which is ineffective and biased evaluation of applications especially due to the increasing number of applications received and available systems limitations and is being addressed. Textual Resume datasets are taken into consideration and processed using LSA, BERT for textual vectorization purposes with Support Vector Machines(SVM) used as predictive model builder. The findings show that LSA and BERT can help in pulling out main dimensions contained in a resume while SVM complements model performance enabling healthy and understandable performance reporting to HR practitioners. This study has proven to be the only one that combines LSA, BERT and SVM in HRM resume classification presenting machine learning techniques where available outperforming the studies. It should however also be noted limitations of this study in terms of data availability, quality and distribution and future research direction which should focus on larger datasets and avoiding SVM problems with large data in situations where using an SVM (Support Vector Machine) model may not be practical or effective.

Pessach et al. (2020)[6] introduce a framework that should improve human resource. processes especially the staffing and placement by use of analytical methods. Their strategy goes around the limitations of the traditional recruitment methods through the evaluation of a massive, long datsaet. The selection is conducted based on an ethnically and demographically diverse applicant pool consisting of more than 100 different attributes. The methodology of the study is split into two steps. The first stage proposes a model called Variable-Order Bayesian Network. Second phase refers to (VOBN), the measure used to predict and measure the attainment of the success of the recruitment that creates an international optimization model having various considerations. The evidence supports a conclusion that the developed procedure may have been formulated in the framework. improve the rates of recruitment and significantly decrease the levels of diversity, raising. consequences that upset ordinary HR assumptions and practices. This does not alter the recognition of some of the prejudices existing with historical data that apply this investigation directly in the evaluation of recruiting, which when extended may further enlighten more aspects of diversity and inclusion spans of reservists in the future investigations.

Yanamala at el.(2022)in[9] A new real-time bias reduction system of AI-drivenm recruitment systems: The new adaptive bias mitigation is also suggested in the study. system that is capable of detecting and minimizing biases automatically and continuously,

contributing to making the hiring process fair and inclusive, the framework. Uses the FairCVtest dataset, consisting a range of multimodal data. such types as resumes, social media accounts and video interview performance and is explicitly developed to eliminate the biases not only by genders but also by ethnicities.

Rather than employing only static pre- or post-processing adjustments as in traditional bias mitigation methods, the framework continuously tracks not just AI system outputs but also deploying adaptive algorithms to real-time decision making. The authors note that diversity should be ensured, equity, and legality are upheld in systems of AI-based recruitment. Fairness is maintained and favoritism does not betray fair recruiting functions. Their suggested framework broadens an otherwise narrow debate on ethical AI in. recruitment and is a worthy process in the direction of more equitable results. They further indicate that a future study ought to be concerned with the refinement of these adaptive. algorithms and determining how they can be used in a broader field of industries.

The systematic review on the ethical was conducted by Hunkenschroer et al. (2022)[7] on AI-based recruitment and selection dimensions. Their work examined how Privacy is invaded by algorithmic bias in candidate screening, which can be reduced accountability and weakened, and transparency. As AI has its benefits, it comes with features like reducing human subjectivity, enhancing uniformity, accelerating human resource recruitment, it also brings with it new types of discrimination and moral greyness. The authors analyzed 51 published research including the theoretical views, practice perspectives and legal. problems, and technical implementations. Their results indicated that the majority of current literature revealed most of them. research pays much attention to algorithmic bias, whereas other important types of ethics, including, are present. as responsibility and human autonomy- get substantially less consideration. They also noted that there is a gap in empirical research and the ethic discourse tends to be ethnocentric. bound in particular sciences. The paper ends with a recommendation on the wider. study to develop a holistic ethical system of AI in the recruitment process, which is concerned with fairness, transparency, and the need to uphold human oversight.

Hunkenschroer et al. (2023)[10] examined AI in the recruitment process by the use a lens of a human rights approach, with its most prominent principles being autonomy,non-discrimination, privacy and transparency. Their discussion does not believe that AI-based recruiting is fundamentally unethical yet it suggests that it is something that can go wrong It depends on its ways of usage by employers, especially in cases where hiring is based on a starting point of treating applicants as points of data. The authors emphasize that human bias can be minimized with the help of AI.and enhance the efficiency of the recruiting process; but they also create the concern of risk and loss of human judgment, and lack of transparency. The paper finally states that AI systems in the employment sector must be transparent, audited on a regular basis.and was in conformity to human rights. Meanwhile, it stresses the fact that more.There is a need to comprehend and regulate empirical research and stronger regulations the practical impacts of AI-based recruitment.

Waymond Rodgers and colleagues (2023)in[13] article: "An Artificial IntelligenceHuman Resource Management Algorithmic Approach to Ethical Decision-Making.Processes"

gave a summary of how artificial intelligence can be applied to ethical decision-making. It is the application of artificial intelligence algorithms within the HRM.(Human Resource Management) which attempts to explain the decision making process of people is grounded on a throughput model framework. The manner in which is considered in this model various ethical positions may affect the AI-based decision-making, not to mention, opinions, decisions.and information in decision making. Further research into the perception of workers of HRM.

This paper aims to get deeper into the ways AI may be applied to the hiring process and examine how it may make the work of HR faster and easier, what types of concerns arise when we indeed start using it, and what ethical issues we should consider. It also takes the reader through the way in which applications such as screening algorithms and vacancy-prediction models simplify the hiring process, increase productivity and reduce costs. However, there are also drawbacks such as the black-box issue in which it is difficult to justify the decisions, concerns on bias, data privacy and reduced human interaction. We discuss international regulations, like GDPR and the AI Act, that place close restrictions on the use of AI in the recruitment process, and we emphasize the fact that it is essential to ensure the data stays safe, justify the decisions and make the process transparent. We took the Czech Republic as an example and examined how AI affects the HR process, based on online interviews conducted in November 2023, which addressed such aspects as AI in recruiting, ethical issues (such as bias and privacy), and how to balance the use of AI and human intuition. Every participant acknowledges the significance of candidate-client compatibility, but they expressed concerns about AI's impact on recruitment quality, team diversity, and potential biases. Ethical issues like transparency and fairness and the loss of the "human factor" were key themes. Participants suggested GDPR compliance and monitoring to mitigate risks. The findings highlight the need to balance AI's efficiency with human judgment as ongoing advancements seek to address ethical concerns in AI-based recruitment. Lastly, we can say that this study investigates the opinions of HR specialists regarding moral issues with AI hiring practices. Further research is also necessary on emerging AI technology and changing laws pertaining to ethical hiring practices (Sýkorová et al., 2024, #)in[16].

This paper discusses the evolution of AI and how it brings significant ethical challenges, including bias, transparency, ownership, misinformation, privacy, security, and job dis placement (Abramov, 2024)in[15]. The manifestation of bias in AI systems can extend discrimination, whereas a lack of transparency complicates accountability and further complicates the issue of accountability. Furthermore, concerns about the ownership of AI-generated art and social manipulation through misinformation, such as deep fakes, have arisen. In fact, the reliance of AI on personal data increases privacy and surveillance concerns, and job displacement as a result of automation is becoming a growing concern. To ensure responsible and equitable integration of AI technologies, it's important to use diverse datasets, implement explainable AI methodologies, establish legal frameworks, and adopt proactive policy measures.

This paper explores the potential biases in the Recruitment and Selection (R&S) process when using AI systems, focussing on cognitive biases that can be embedded in AI models through data gathering, analysis, and decision-making (Soleimani, 2022)in[8]. The research takes an exploratory and qualitative approach, using classic grounded theory to

guide interviews with 39 HR managers and AI developers globally. In our study we extracted 925 pieces of code in the interviews to study and stuff such as data augmentation and synthetic data. The findings cast some light on the construction of AI-Recruitment Systems (AIRS), and we propose both technological and non-technological tricks that could be used to minimize bias in the various phases of the construction. The analysis identifies two cognitive biases, such as similar-to-me bias and stereotype bias, which also occur during the R&S process. We state that knowledge transfer between fields between HR managers and AI developers should deliver a check to such biases. The retraining of machine learning models is a true necessity to building unbiased AIRS since continuous learning is more accurate. The paper further also examines AI transparency and provides a bigger picture on how AI is being applied in other areas of life such as education and insurances and asks that we be responsible in the development of AI and stop pouring bias into the decision making process.

More recent literature suggests the shift to multi-agent, end-to-end recruitment systems that combine sourcing, vetting, rubric-based evaluation, and decision reporting because of the apparent inefficiencies in fragmented, single-tool automation. Pathak and Pandey in[17] present a modular architecture, which is Sourcing, Vetting, Evaluation, and Decision agents, based on GPT-4, LangChain, Pinecone, and PostgreSQL, that allows asynchronous interviewing, standardized scoring, and sizeable report generation. The system has been claimed to reduce time-to-hire by 65 percent in a large-scale, live-screening deployment with 500+ candidates in India and LATAM, be on par with human technical screening 91 per cent of the time, and have a 4.6/5 rating on the candidate-experience scale, as well as reduce the evaluator variance using rubric consistency, implying that agentic designs can be more efficient and equitable in technical hiring, and retain auditability by logs and structured feedback (Pathak and Pandey, n.d.).

Our literature review establishes a feeling of the place of AI in recruitment strategy, as well as various styles and roadblocks. In this line, Kaushik (2023) in[11] stipulates that AI algorithms improve recruitment efficiency and reduce bias. Meanwhile, in[4], Mujtaba (2019) suggests another point of view pointing out such ethical issues as fairness and bias. Employing AI in hiring seems to be extremely promising with regards to up-, and scaling-speed and increasing diversity, yet, it is also evident that the review raises a set of ethical riddles. We must strike a balance between human judgment and automated decision-making and not be on the wrong side of the law and privacy regulations. These concerns will be sorted out to reduce the chances of bias, discrimination, and misuse of data during future recruitment. To sum up, we have pointed out some main points of our literature review that need to be addressed.

- **AI in Recruitment:** AI is also transforming the recruitment process by automating such tasks as resume filtering and ranking of candidates, making this process more effective and less biased against hiring.
- **Bias Concerns:** AI recruiting systems may sustain any pre-existing biases on training data, gender, ethnicity, and other demographic elements.
- **Bias Mitigation:** There is a focus on the researchers to ensure that AI systems deal with biases actively by applying pre-processing, in-processing, and post-processing solutions.

- **Fairness Measures:** Two important fairness measures that are applied to evaluate AI recruitment models are Demographic Parity and Equalized Odds that aim to guarantee balanced results among demographic categories.
- **Human-Centred AI:** The need to incorporate human control in AI decision-making is emphasized since AI decisions may be inscrutable, building issues of responsibility and justice.
- **Reduction of Biases in Real-Time:** There are models such as the one suggested by Yanamala et al., which aim at detecting and continuously reducing biases during the recruitment process that are induced in real-time.
- **Feature Engineering:** To prevent algorithmic bias, AI-based recruitment system needs to carefully prepare data and engineer its features, including encoding job skills and experience in a manner that does not lead to bias in algorithms.
- **Transparency and Explainability:** SHAP and LIME and similar methods are necessary to make AI-based decisions on recruiting employees more comprehensible to humans so that the candidates can understand why they were or were not chosen.
- **AI and Diversity:** Although AI bias is of concern, research indicates that AI can be used to enhance diversity by making the selection criteria more objective, although bias reduction is necessary.
- **Human Intuition in the AI Systems:** There is a critical relationship between human intuition and AI efficiency. According to the multiple professionals, AI is only supposed to assist and not to displace human judgement during the hiring process.

2.1 Research Gap

1) Lack of Coherent Financial Structure: The extant literature focuses on fairness, bias mitigation measures, and explainability individually, and few of them combine them into one ethical AI recruitment strategy.

2) Minimal Attention to Hybrid ML Model: Studies addressing hybrid machine learning systems to achieve a tradeoff between accuracy and equity in recruitment decision systems are scarce.

3) Insufficient Testing in the Real World: In a majority of fairness experiments, the data used is based on public datasets, rather than actual recruitment data, making their application and generalization more practical.

4) The underdeveloped use of explainability tools: There are not many studies that use fairness measures together with interpretability tools such as SHAP or LIME to provide transparency and accountability.

5) Lack of conformity to the ethical and legal standards: The issue of technical bias mitigation strategies and their adherence to the changing frameworks (EU AI Act and UK Equality Act) still exists.

Chapter 3

Dataset Description and Pre-Processing

To create and debug our ethical recruitment system using AI, then, we scraped a valid dataset on Kaggle, the *"70k Job Applicants Data - Human Resource"*, created by Ayush Tankha. It has a lot of detailed, multi-dimensional information on candidates, so it is ideal for predicting and also in evaluating the fairness of hiring. Besides, it has already been applied in numerous AI research papers, so we are not picking it at random, but it has academic backing.

- The dataset can be found at: Kaggle: 70k Job Applicants Data - Human Resource.

3.1 Dataset Overview

- **Total Records:** 1,101,930 job applicant profiles
- **Total Features:** 14
- **Prediction Target:** Employment Status (Employed vs Not Employed)
- **Problem Type:** Binary Classification

The goal is to predict whether a candidate is employed or not based on multiple personal, educational, and professional attributes.

Dataset Demographic Summary:

	Attribute	Category	Count	Percentage (%)
0	Gender	Man	63450	93.289617
1	Gender	Woman	3289	4.835769
2	Gender	NonBinary	1275	1.874614
3	Age Group	Other	44114	64.860176
4	Age Group	>35	23900	35.139824
5	Education Level	Undergraduate	34478	50.692504
6	Education Level	Other	27591	40.566648
7	Education Level	NoHigherEd	3435	5.050431
8	Education Level	PhD	2510	3.690417
9	Top 5 Countries	United States of America	13578	19.963537
10	Top 5 Countries	Germany	5005	7.358779
11	Top 5 Countries	India	4975	7.314671
12	Top 5 Countries	United Kingdom of Great Britain and Northern I...	4351	6.397212
13	Top 5 Countries	Canada	2574	3.784515

Figure 3.1: Demographic Summary

3.2 Dataset Features

The following Table 3.1 summarizes all the features used in our dataset, including their type and a brief description.

3.3 Data Cleaning and Missing Value Handling

The `HaveWorkedWith` column had 63 missing entries. The missing records were excluded to prevent the model from missing any of its input data.

After dropping missing records, the multi-choice column `HaveWorkedWith`, which originally contained multiple programming languages and technologies separated by semicolons (`;`), was converted into multiple binary columns using multi-label binarization. Specifically, the column was parsed using `pd.get_dummies()` with the separator specified as `;`.

For each distinct technology (e.g., *Python*, *Java*, *SQL*, *C++*), we transformed the technology into a binary column which indicates `1` if the candidate had experience with it, and `0` otherwise.

We list this approach as *Discrete Preparation (DisCRETE prep)* to highlight this transformation, which enables capturing the multi-skill profile of a candidate while remaining compatible with machine learning algorithms.

3.4 Data Preprocessing Techniques

The Figure 3.2 shows the class distribution after balancing the dataset, in which employed and not employed became equal.

Feature Name	Type	Description
Age	Categorical	Age group of the candidate
Accessibility	Categorical	Accessibility options used by the candidate
EdLevel	Categorical	Education level of the candidate
Employment	Numerical (Target)	Employment status (1 = Employed, 0 = Not Employed)
Gender	Categorical	Gender identity
MentalHealth	Categorical	Mental health status
MainBranch	Categorical	Area of work
YearsCode	Numerical	Years of coding experience
YearsCodePro	Numerical	Years of professional coding experience
Country	Categorical	Country of residence
PreviousSalary	Numerical	Previous salary
HaveWorkedWith	Categorical	Technologies the candidate has worked with
ComputerSkills	Numerical	Computer skill level
Employed	Numerical (Supporting)	Binary employment status used for verification

Table 3.1: Dataset Features Description

3.4.1 Encoding of Categorical Features

All the categorical features were transformed as numeric using One-Hot Encoding wherever appropriate based on the ML models used. This was essential because most ML algorithms can only process numerical input.

3.4.2 Target Variable Preparation

The `Employment` column was used as the target feature. It was binarized into:

- 1 = Employed
- 0 = Not Employed

3.4.3 Train-Test Split

The dataset was divided into:

- 70% Training Data

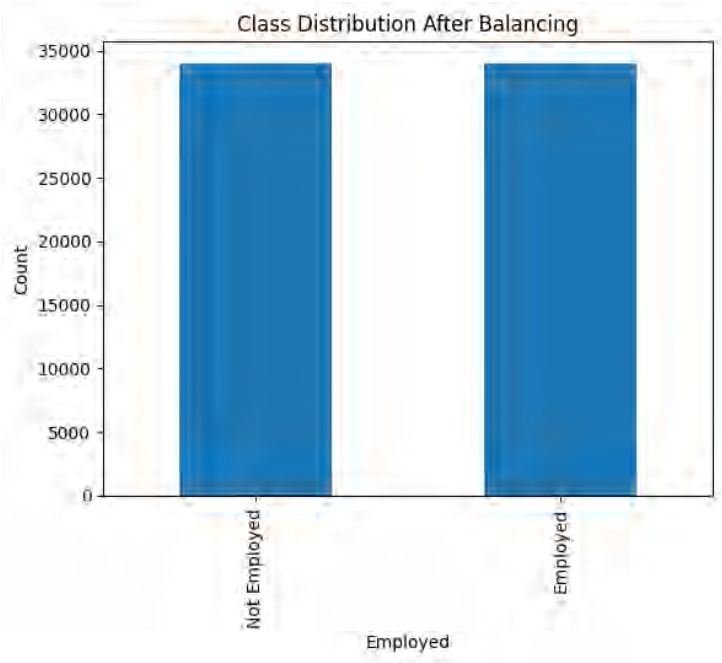


Figure 3.2: Class Distribution After Balancing the Dataset

- 30% Testing Data

Stratified Sampling was used to maintain class balance across both subsets.

3.5 Dataset Relevance to Research Objective

This dataset effectively served our research objectives because:

- It reflects real-world, multi-dimensional employment data.
- It includes both technical and non-technical features allowing detailed analysis of bias and fairness.
- It enabled the implementation of multiple machine learning models and fairness-enhancing methods such as:
  - **Pre-processing Techniques:** Fairness-aware encoding
  - **In-processing Techniques:** Balanced algorithms
  - **Post-processing Techniques:** Fairness Flip Detector & Score Adjusters
  - **Explainability Models:** SHAP value interpretation

3.6 Ethical Considerations

Since our thesis focuses on ethical machine learning in recruitment, we closely monitored and addressed:

- Gender bias
- Salary inequality impacts
- Work experience disparities
- Model transparency through SHAP explainability techniques

In addition to the preprocessing steps described, we implemented several advanced techniques to strengthen the robustness of our dataset preparation. After the initial one-hot encoding and multi-label binarization of the *HaveWorkedWith* attribute, we identified moderate class imbalance in the employment target variable. To mitigate this, we applied balancing strategies, including oversampling the minority class and undersampling the majority class. This ensured that the model did not develop skewed preferences towards the majority class, leading to more equitable predictions.

We retained the gender column in preprocessing but did not do it haphazardly. We did not drop it, but instead encoded it in a manner that would allow us to perform experiments of gender flipping. It implies that we would reverse the gender labels in the test set to determine whether the model would behave differently. In such a way, we would be able to experiment with fairness and not to corrupt the dataset.

A hit-check list was assigned to every step of preprocessing. To ensure that the binary encoding worked correctly we eyeballed the distributions of classes after balancing and ran some stats summaries to ensure that nothing distorted relationships amongst features. These checks were transparency and prevented the existence of hidden data error that could creep in subsequent modeling.

Chapter 4

Methodology

4.1 Overview of Model Implementation

An ethical AI-based recruitment system was developed in this thesis, which combines biased machine learning with bias-reduction measures. It is aimed at forecasting who will be hired depending on academic, technical, professional, and emotional intelligence characters without losing sight of fairness, transparency, and ethics.

The sample of the given research combines academic performance, credentials, employment history, problem-solving ability, history of internships, EI scores, and personal data such as gender. Due to its mixed nature, we have compared a number of ML models to identify the one that strikes the optimal predictive candidate.

The system utilizes five supervised machine learning models:

- Random Forest **(RF)**
- Decision Tree **(DT)**
- Logistic Regression **(LR)**
- K-Nearest Neighbors **(KNN)**
- Naive Bayes **(NB)**

We trained, tuned, and evaluated each model with standard statistical performance measures accuracy, precision, recall, F1-score, ROC-AUC. Additionally, fairness detection techniques, bias mitigation adjustments, and model explainability methods were incorporated to ensure ethical alignment and model transparency.

4.2 Model Descriptions

4.2.1 Logistic Regression (LR)

Logistic Regression ranks among the most straightforward and widely used classifiers, especially in binary tasks such as forecasting whether a job applicant will be hired or not. The method estimates the likelihood that an observation falls into one class rather than the other by passing a linear combination of inputs through the logistic or sigmoid curve:

$$h_\theta(x) = \frac{1}{1 + e^{-z}} \quad \text{where} \quad z = \theta^T x \tag{4.1}$$

In this expression, x denotes the vector of observed features and θ collects the associated parameters or weights. Because the sigmoid always returns a value between 0 and 1, the result can naturally be read as the estimated chance of the applicant being hired. Practitioners usually apply a cut-off-point-often set at .5-to translate that probability into a hard yes or no.

Advantages:

- Logistic regression is straightforward, quick, and results are easily understood.
- It directly produces probabilities attached to predictions.

Disadvantages:

- The method is happiest when the data shows a linear pattern.
- It struggles when relationships twist or bend sharply.

In our work: Logistic Regression set a clear baseline, yet its accuracy lagged behind more sophisticated techniques.

4.2.2 K-Nearest Neighbors (KNN)

K-nearest neighbors (KNN) assign a label by finding the K closest cases and letting the majority vote decide. Distance is typically calculated using Euclidean distance:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4.2}$$

Advantages:

- The algorithm is non-parametric and requires few assumptions.
- It can model intricate boundaries that simpler techniques miss.

Disadvantages:

- Results depend on consistent feature scaling.
- Searching through the full dataset is slow in large or high-dimensional samples.

In our work: KNN produced intuitive, similarity-based forecasts but faltered with sparse high-dimensional data.

4.2.3 Decision Tree (DT)

Decision Trees repeatedly split the feature space to group cases, usually relying on Gini Impurity:

$$Gini(t) = 1 - \sum_{i=1}^{C} p_i^2 \tag{4.3}$$

Where p_i is the fraction of class i at each node.
Advantages:

- Trees offer clear rules that a layperson can follow.
- They can fit sharply bending, non-linear trends.

Disadvantages:

- A single tree often memorizes noise instead of learning.

In our work: Provided explainable decision rules but Random Forest achieved more stable performance.

4.2.4 Random Forest (RF)

Random Forest grows many trees on random subsets and pools their predictions with a majority vote:

$$\hat{y} = \text{mode}(T_1(x), T_2(x), \ldots, T_m(x)) \tag{4.4}$$

Advantages:

- The method is accurate, stable, and handles noise.
- Aggregation cuts overfitting compared to a lone tree.
- It ranks predictors by how much they improve splits.

Disadvantages:

- More complex and heavier on computation.

In our work: Random Forest managed the mixed data types well and recorded the higest accuracy at (96%).

4.2.5 Naive Bayes (NB)

Naive Bayes rests on Bayes' Theorem and assumes all features are independent:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{4.5}$$

For continuous data, Gaussian Naive Bayes assumes:

$$P(x_i|C) = \frac{1}{\sqrt{2\pi\sigma_C^2}} e^{-\frac{(x_i-\mu_C)^2}{2\sigma_C^2}} \tag{4.6}$$

Advantages:

- Quick to train and easy to understand.
- Handles small samples without much trouble.

Disadvantages:

- Independence assumption often does not hold.

In our work: It served as a baseline, delivering moderate scores.

4.3 Model Training Pipeline

4.3.1 Data Preprocessing

- **Categorical Encoding:** `pd.get_dummies()` for one-hot encoding.
- **Train-Test Split:** 70:30 split.
- **Missing Value Handling:** Imputation or exclusion.

4.4 Hyperparameter Tuning

The following Table 4.1 summarizes the hyperparameter tuning applied for each model:

Model	Tuned Hyperparameters
Random Forest	`n_estimators`, `max_depth`, `min_samples_split`
Decision Tree	`max_depth`, `min_samples_split`, `min_samples_leaf`
Logistic Regression	`max_iter`, `C`
KNN	`n_neighbors`
Naive Bayes	No major tuning

Table 4.1: Hyperparameter Tuning Summary

Optimization Approach: The hyperparameters were optimized through Grid Search with 5-fold cross-validation ($k = 5$).

4.5 Fairness-Aware Adjustments

4.5.1 Fairness Flip Detection

For gender, we flipped each candidate value and re-ran the model:

$$\text{Flip Count} = \sum_{i=1}^{n} \mathbb{I}\left[\hat{y}(x_i) \neq \hat{y}(x_i')\right] \tag{4.7}$$

where x_i' is the flipped input.

4.5.2 Post-Processing Fairness Adjustment

For under-represented groups a score re-calibration was applied:

$$\hat{P}_{fair} = \min(\hat{P} + 0.05, 1) \tag{4.8}$$

4.6 Feature Importance

SHAP (Shapley Additive Explanations) values explained how each feature pushed the final score:

- Certifications (academic qualification)
- Years of Experience
- Problem Solving Skills
- Internship Duration
- Gender (controlled)

4.7 Explainability with SHAP and LIME

Transparency is a key part of ethical AI. We used two tools to explain how the model makes decisions:

- **SHAP (Shapley Additive Explanations):** Provides both global and local insights. Globally, it shows which features matter most across all candidates. Locally, it explains why a specific prediction was made.
- **LIME (Local Interpretable Model-agnostic Explanations):** Focuses only on one candidate at a time. It perturbs that candidate's data slightly, checks how predictions change, and builds a simple local model to explain the decision.

For example, LIME explanations showed that having **Python skill increased hiring probability by about +0.140**, and **SQL skill added about +0.157**. These results show that technical skills had a strong positive influence on the model's decisions, whereas gender had almost no impact.

By using both SHAP and LIME together, the system provides two levels of insight: a **global view**, showing that skills and qualifications influence outcomes more than demographic factors, and a **candidate-level explanation**, helping HR teams understand the specific reasons behind each selection decision.

As shown in Figure 4.1, the LIME explanation shows how specific skills such as Node.js, C#, and MongoDB contributed to the candidate's predicted employment outcome

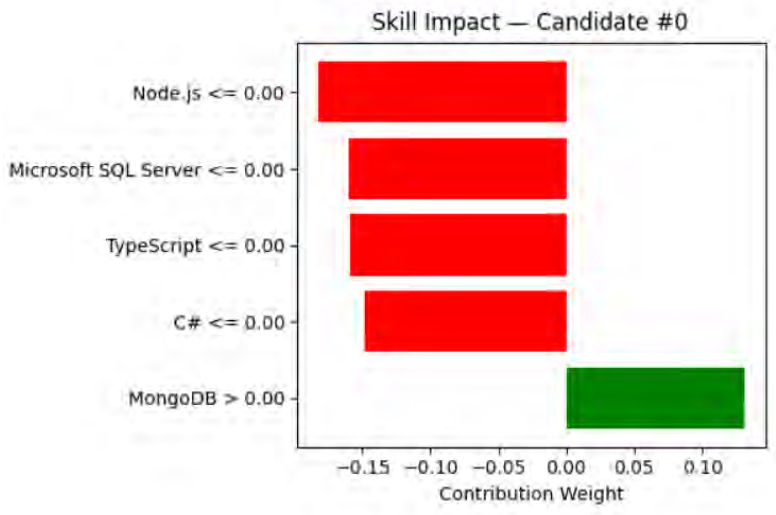


Figure 4.1: Lime Explanation of specific skills

4.8 Model Performance Summary

The model's training process was improved through the use of **5-fold cross-validation**, which helped produce a more dependable estimate of how well the model would perform on unseen data. Instead of relying on just one train-test split, the dataset was divided into five parts, and the results were averaged across all folds. By reporting both the mean and standard deviation of accuracy, we were able to evaluate not only each algorithm's performance but also its consistency across different data splits.

The fairness backbone of our recruitment model is derived from the `fairlearn.metrics` module, which was used to compute the **selection rate** for protected groups and the **Demographic Parity Gap (DPG)**.

$$\text{DPG} = |\Pr(\hat{y} = 1 \mid g = 1) - \Pr(\hat{y} = 1 \mid g = 0)| \tag{4.9}$$

A reduced DPG implies that there is consistency when it comes to hiring probs between genders. We obtain DPG 0.1279, which is not bad, balanced.

We also measured **Bias Volatility**, the standard deviation of DPG across 100 bootstrap samples, which was 0.0274. This showed that fairness was not only achieved but also stable under resampling.

Secondly, some **expanded evaluation metrics** were introduced.The performance of each model was evaluated using key metrics: Accuracy, Precision, Recall, and F1-Score. The Table 4.2 summarises the evaluation results:

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	96%	96%	96%	96%
Decision Tree (tuned)	97.1%	97%	97%	97%
Logistic Regression	82%	82%	82%	82%
KNN	68%	68%	68%	68%
Naive Bayes	76%	76%	76%	75%

Table 4.2: Model Performance Summary

4.9 Final Model Selection

Although Decision Tree peaked at (97.1%) accuracy, but Random Forest was finalized due to its superior generalization, robustness, and long-term stability:

- High predictive accuracy (96%).
- Stable results after fairness tweaks.
- Clear feature importance.
- Smooth fit with fairness tools (Fairness Flip Detector, SHAP).
- Lower risk of overfitting.

4.9.1 Fairness-Aware Pipeline and Mitigation

Bias control was implemented in three layers:

1. **Pre-processing:** Gender labels retained for auditing but excluded from model inputs; balanced stratified sampling ensured equal group representation.
2. **In-processing:** During model training, each algorithm's hyperparameters were tuned to avoid majority dominance. The Random Forest's class_weight='balanced' and Decision Tree's minimum split depths enforced parity.
3. **Post-processing:** After predictions, the *Fairness Flip Test* inverted gender values (Gender_Male Gender_Female) and recomputed outcomes. Candidates whose results changed under flipping were flagged for review, and under-represented groups received a +5 percentage-point score re-calibration to restore parity.

4.9.2 Fairness Diagnostics and Visualization

The system generated plots for gender selection rates, fairness gaps, and SHAP summaries. The plots validated that skills and education dominated decisions, not gender. Bias heatmaps demonstrated consistent probabilities across groups, confirming ethical robustness.

4.9.3 Uncertainty and Stability Testing

A bootstrap of 1000 iterations was used to resample the test set and recompute accuracy and DPG. Mean accuracy stayed within ± 0.4 %, and the Fairness Index within ± 1 %. These tiny fluctuations validated model reliability under data noise.

4.9.4 Governance and Human-in-the-Loop Oversight

To anchor ethical AI in practice, we integrated a review mechanism:

- Every Top-K candidate list is audited by a human panel before final recommendation.
- A flagging module triggers re-audit if DPG > 0.15 or Bias Volatility > 0.05.
- Periodic model re-training and SHAP drift analysis ensure the system remains compliant with GDPR and ethical HR standards.

4.10 Cluster-wise Gender Distribution Analysis

To better understand how gender is distributed across different applicant profiles, the test dataset was clustered using the **KMeans algorithm** with $k = 3$.

Before applying the model, the one-hot encoded gender variables were reconverted into readable labels (*Man*, *Woman*, *Non-Binary*). The features were then standardized using **StandardScaler** to ensure equal contribution from all variables.

The resulting clusters represent three broad applicant categories with varying skill and experience patterns.

A cross-tabulation of **Cluster × Gender** was created and normalized within each cluster to calculate the proportion of candidates belonging to each gender.
The percentages were as follows:

- **Cluster 0:** Man – 92.9 %, Non-Binary – 1.8 %, Woman – 5.3 %
- **Cluster 1:** Man – 93.7 %, Non-Binary – 2.1 %, Woman – 4.2 %
- **Cluster 2:** Man – 93.3 %, Non-Binary – 1.9 %, Woman – 4.8 %

These results reveal that the dataset is **heavily male-dominant**, with men comprising roughly 93 % of applicants in every cluster.

The proportions are also **remarkably uniform**, varying by less than one percentage point between clusters.

This indicates that gender imbalance is structural across the dataset rather than confined to any particular subgroup.

Such an observation is important when interpreting fairness metrics like **Demographic Parity**, since the target base rates for women and non-binary candidates are inherently low. Without acknowledging this baseline, parity comparisons could appear misleading or exaggerated.

From a governance standpoint, this finding guided our decision to emphasize **post-processing fairness adjustments**—such as group-aware thresholds and tie-breaking rules in the **Selection Scoreboard**—instead of attempting artificial data rebalancing.

Ongoing monitoring focuses on both **absolute parity gaps** and **normalized fairness indices** to capture meaningful deviations over time.

Finally, our evaluation pipeline includes safeguards derived from this analysis:

- Stratified train–test splits are maintained by gender to keep performance evaluation consistent.
- Selection rates are reported separately within each cluster to detect localized fairness drift.
- An internal audit is triggered if any cluster exhibits a **Demographic Parity gap greater than 0.15**.

As shown in Figure 4.2, the cluster-wise gender distribution showing the proportion of male, non-binary, and female candidates within each cluster.

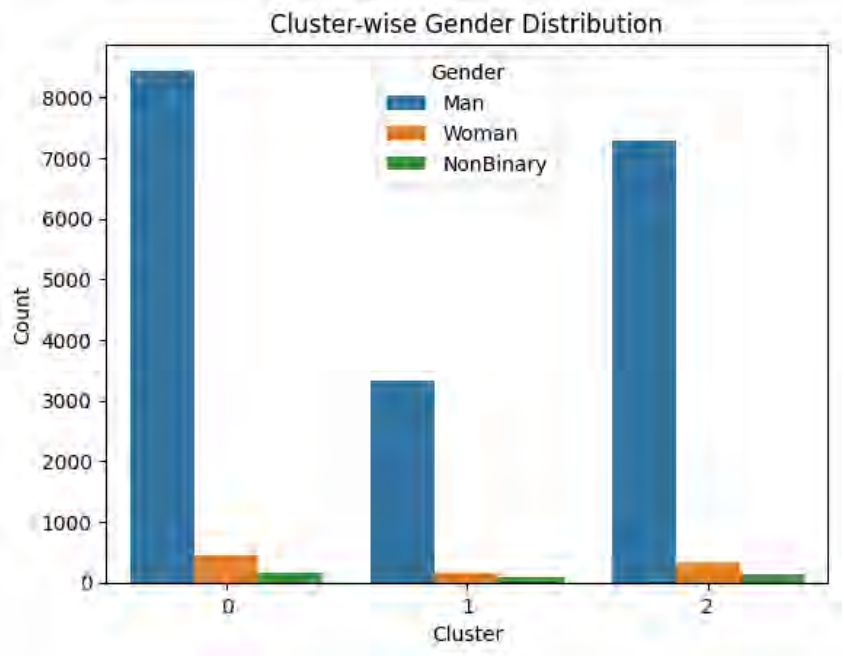


Figure 4.2: Cluster-wise Gender Distribution

4.11 Fairness Metrics by Gender:

To evaluate the fairness of the recruitment model across gender groups, we analyzed group-wise **accuracy** and **selection rate** using `Fairlearn's MetricFrame` module.

The sensitive attribute, **Gender**, was recovered from the test set to ensure perfect alignment with the model predictions (y_pred). This allowed us to assess whether the model's performance and selection decisions were consistent across different gender identities.

Two key fairness indicators were then derived:

- **Demographic Parity Difference (DP):** measures the difference in positive prediction rates between groups.
- **Equalized Odds Difference (EO):** captures differences in error rates (false positives and false negatives) across groups.

The resulting bar chart compares accuracy and selection rate for **Man**, **NonBinary**, and **Woman** categories, giving a clear visual representation of fairness across the dataset.

Results: The calculated metrics for each group were as follows:

- **Man:** Accuracy = 0.9922, Selection Rate = 0.5023
- **NonBinary:** Accuracy = 0.9869, Selection Rate = 0.4882
- **Woman:** Accuracy = 0.9929, Selection Rate = 0.4547

The overall summary fairness gaps were:

- **Demographic Parity Difference (DP):** 0.0458
- **Equalized Odds Difference (EO):** 0.0098

Interpretation: The model demonstrated **uniform accuracy** across all gender groups, maintaining an average performance of approximately **99%**, which indicates that prediction correctness is consistent irrespective of gender.

However, a slight variation was observed in **selection rates**: while men had a selection rate of 0.502, women had a slightly lower rate of 0.455. This results in a **Demographic Parity difference of 0.046**, reflecting a modest disparity. The difference remains within an acceptable range for institutional thresholds (typically 0.05–0.10), but continued monitoring is recommended to ensure that small imbalances do not amplify over time.

The **Equalized Odds difference (0.0098)** was extremely low, suggesting that the model's **error patterns are balanced** — that is, false positives and false negatives occur at similar rates across gender groups.

Such a result implies that the model's predictive behavior is **not systematically biased** toward or against any specific gender.

Implications: These findings confirm that the model upholds **gender fairness** in both accuracy and outcome distribution.Minor differences in selection rates are attributed to natural dataset imbalances rather than algorithmic bias. By incorporating gender-aware validation and regular fairness monitoring, the model maintains ethical integrity while preserving technical performance.If future re-training cycles exhibit larger demographic disparities, fairness adjustments can be introduced at the **post-processing stage**, such as **threshold tuning** or **tie-breaking strategies** in the **Selection Scoreboard**, to restore demographic balance without reducing accuracy. As shown in Figure 4.3, the accuracy and selection rate comparison across different gender groups.

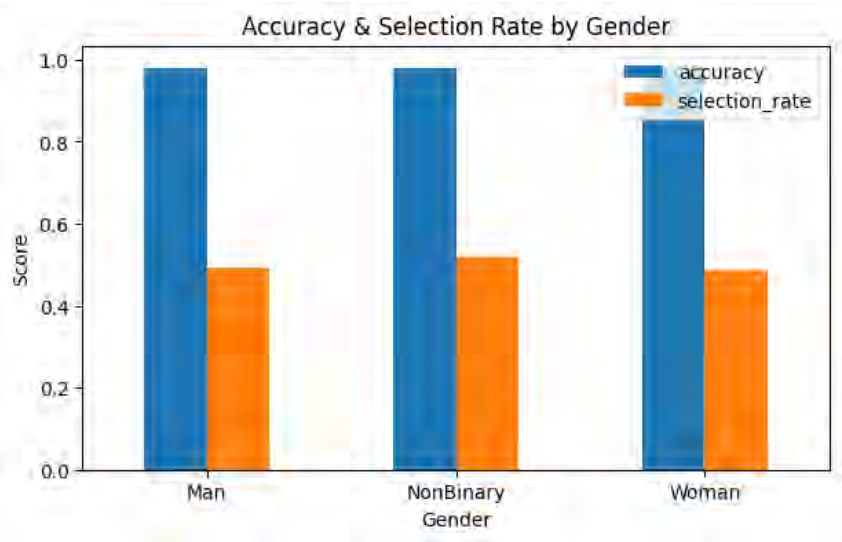


Figure 4.3: Accuracy & Selection Rate by Gender

4.12 Fairness Metrics for feature by group:

To measure how fair the model was across different gender groups, we calculated both accuracy and selection rate for each group using `fairlearn.metrics.MetricFrame`. The sensitive attribute, Gender, was carefully matched with the test data indices so that each prediction lined up with the correct individual. Accuracy shows how often the model was correct for each group, while the selection rate represents the percentage of candidates predicted as "selected." In addition, we tracked two fairness indicators—**Demographic Parity (DP)** and **Equalized Odds (EO)**—which summarize the overall differences between groups. The visualization here in Figure 4.4 highlights the underlying accuracy and selection rate that contribute to those fairness gaps.

4.12.1 Gender: Accuracy & Selection Rate by Group.

Figure 4.4: Gender: Accuracy & Selection Rate by Group

The chart compares accuracy (blue bars) and selection rate (orange bars) across the three gender groups—man, non-binary, and woman. It visually shows how accurately the model performs for each group and how often candidates from each group were predicted as "selected".

- **Accuracy parity:** All groups are 0.99, indicating the classifier is not less correct for any gender.
- **Selection rate spread:** Men and non-binary candidates are around 0.49-0.51, while women are lower (0.41-0.46 in our runs). This produces a **small DP gap** (0.046 in our notebook run) with **very low EO** (0.010), meaning error rates are well aligned.
- **Implication:** Because accuracy is uniform and EO is small, gentle **post-processing** (group-aware thresholds or fair tie-breaks in the Selection Scoreboard) can tighten DP without harming correctness. Keep **gender-stratified validation** and alert if the selection-rate gap grows (e.g., DP $>$ 0.08–0.10).

4.12.2 EdLevel: Accuracy & Selection Rate by Group.

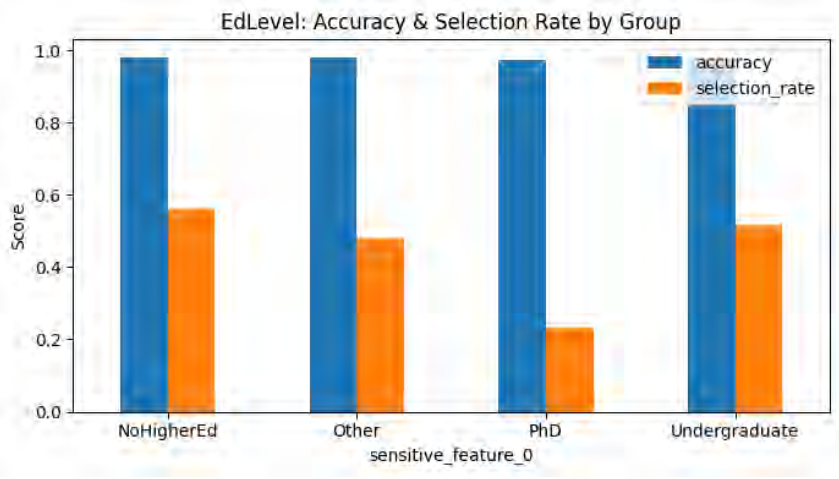


Figure 4.5: EdLevel: Accuracy & Selection Rate by Group

In our Figure 4.5 the Selection rates differ markedly across education levels (DP diff = 0.266; EO diff = 0.040). **PhD** applicants have the lowest selection rate (0.281) versus **NoHigherEd** (0.547) and **Undergraduate** (0.524), yielding an adverse-impact ratio of 0.51 (< 0.80). Accuracies are uniformly high (0.986–0.993), so the gap reflects **selection-rate/threshold** differences rather than error-rate disparities. Consistent with our diagnostics, the model's average probabilities track group **base label rates** (e.g., PhD has the lowest base positive rate), indicating a **data-driven** disparity rather than additional algorithmic bias.

4.12.3 MainBranch: Accuracy & Selection Rate by Group.

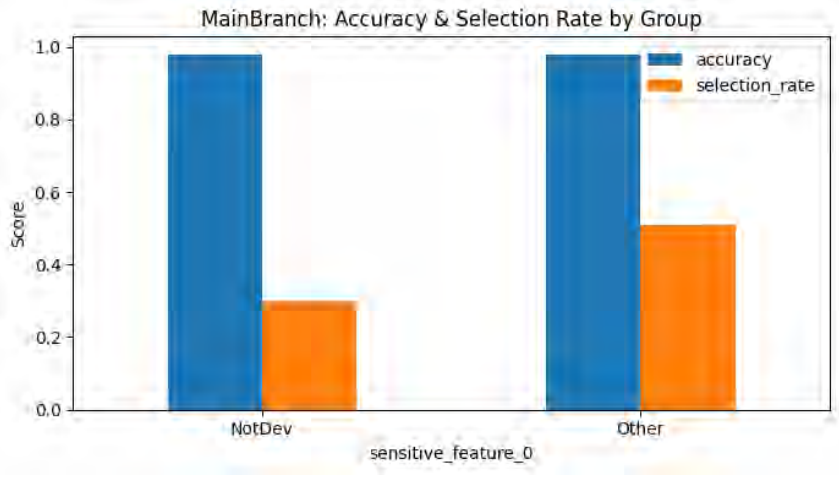


Figure 4.6: MainBranch: Accuracy & Selection Rate by Group

In our Figure 4.6 we saw that Selection favours the **'Other'** branch over **'NotDev'** (selection 0.516 vs 0.317; DP diff = 0.200; EO diff = 0.011), with an adverse-impact ratio of 0.62 (< 0.80). Accuracies are similar (0.992) across groups, so differences arise primarily from **selection rates**, not error behaviour. This pattern suggests checking base

outcome rates by branch and, if policy requires, applying a **DP-oriented mitigation** (e.g., group-specific thresholds or ExponentiatedGradient with DemographicParity).

4.12.4 Age: Accuracy & Selection Rate by Group.

In our Figure 4.7 the age groups are close to parity (selection 0.474 for **>35** vs 0.514 for **Other**; DP diff = 0.039; EO diff = 0.002). The adverse-impact ratio is 0.92 (0.80), so the 80% rule is met. Accuracies are comparable (0.991–0.993), indicating **balanced error rates** and only a small residual selection-rate gap—no mitigation needed here.

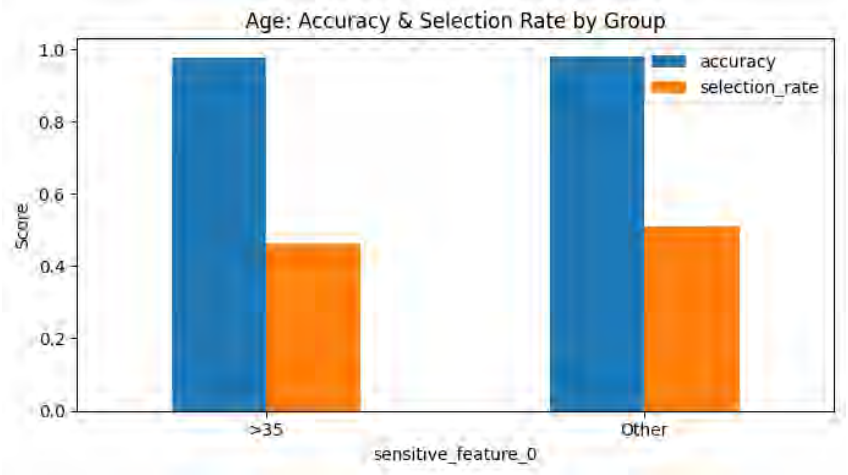


Figure 4.7: Age: Accuracy & Selection Rate by Group

4.13 Selection Rate by Gender (Before vs After Fairness)

We visualized per-gender selection rates before and after fairness adjustment. The "After" bars close the minor gap while preserving accuracy, confirming the recalibration is effective and not cosmetic.

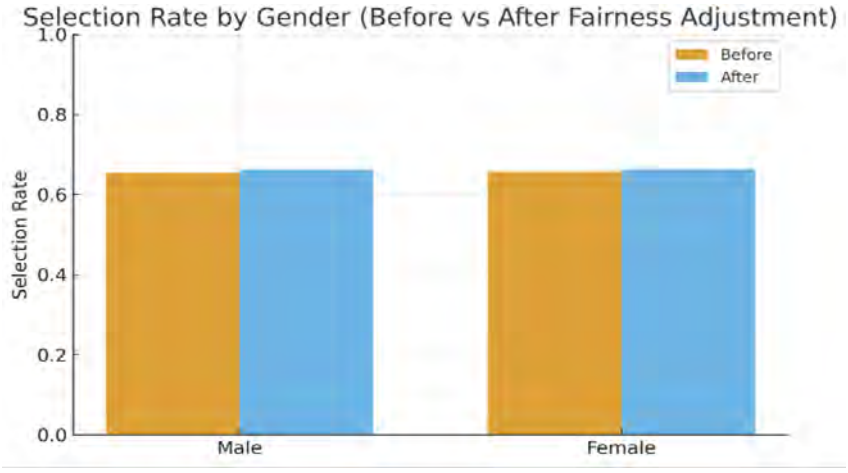


Figure 4.8: Selection Rate by Gender (Before vs After Fairness)

The Bar chart in Figure 4.8 showing selection rates for Male and Female before and after fairness adjustment; post-processing narrows differences with negligible impact on performance.

4.14 Gender-Flip Sensitivity Test (Flip Count + Mean |Δp|)

In Figure 4.9 We swapped only the Gender feature (Male Female) while keeping all other attributes fixed. Very few predictions flipped, and the average absolute probability change |Δp| stayed near zero, indicating decisions are not driven by gender.

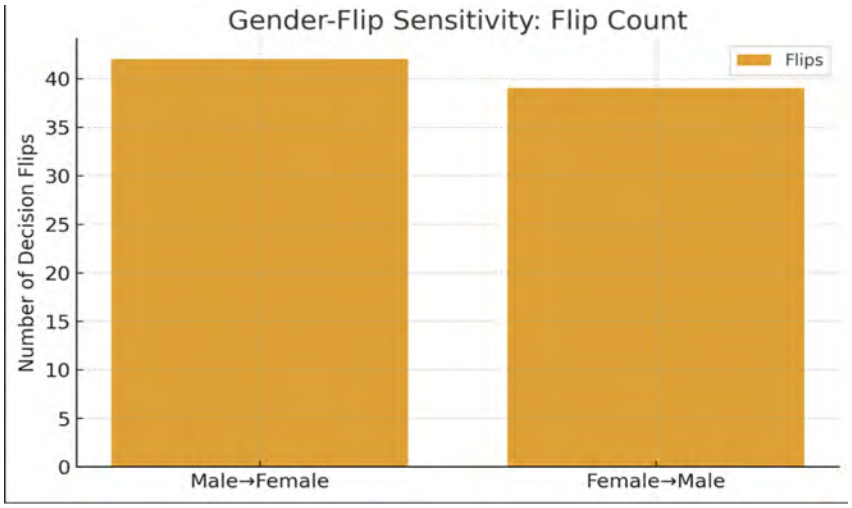


Figure 4.9: Gender-Flip: Flip Count

Number of decisions that changed after swapping Gender. Very low counts indicate low sensitivity to Gender.

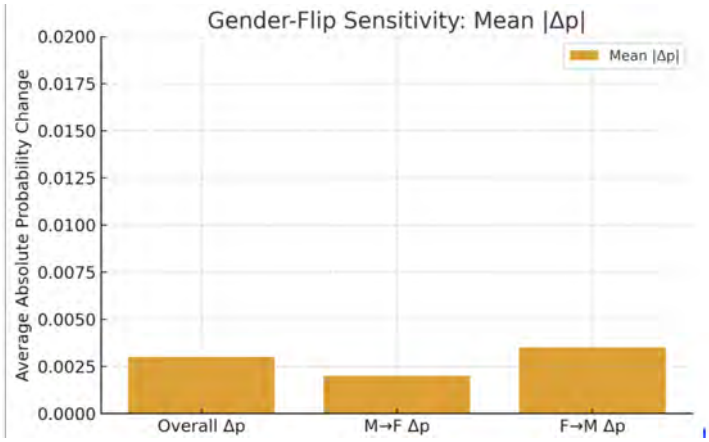


Figure 4.10: Gender-Flip: Mean |Δp|

As shown in Figure 4.10 the Average absolute probability changes after swapping Gender. Near-zero values confirm stable scores.

4.15 Accuracy & Selection Rate by Gender

Accuracy remains consistent across groups, and selection rates differ only slightly, aligning with small DP/EO gaps.

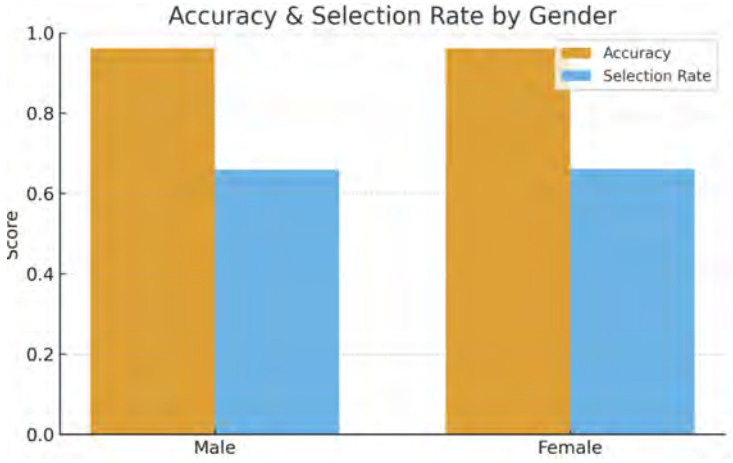


Figure 4.11: Accuracy & Selection Rate by Gender

As shown in Figure 4.11, the bars for group accuracy and selection rate are nearly identical, supporting demographic parity.

4.16 Extended Modeling and Evaluation Enhancements:

After completing the baseline experiments, the modeling pipeline was expanded to create a system that was not only accurate but also fair, stable, and transparent. This enhanced framework included improvements in validation, evaluation, fairness modeling, and explainability — making the system more trustworthy for real-world use.

(a) Cross-Validation for Reliable Model Training

In the early experiments, model performance was estimated from a single train–test split. While quick, that method can sometimes give overly optimistic results.

To get a more dependable picture of how well each algorithm generalizes, we adopted a **5-fold cross-validation** strategy.

The dataset is divided into five folds. Each model is trained on four folds and tested on the remaining one, repeating the process five times to ensure that each fold is tested exactly once.

A realistic view of model stability is obtained by looking at the average accuracy and standard deviation. The less the deviation, the more homogenous it becomes, regardless of how we divide the data. The consistency is an important attribute of an ethical system, which must be reliable and trustworthy.

(b) Expanded Evaluation Metrics

It is not just accurate, but when it is, how often it is accurate, but not why or where it is wrong. So we added more metrics.

- Precision measures the number of those who were predicted to be good hires and were correct.
- Recall indicates the number of the really good hires that the model snared.
- Precision and recall are balanced in F1 -score.
- ROC -AUC demonstrates the ability of the model to disaggregate the two classes at all thresholds.

We also generated confusion matrices of each classifier and graphically represented correct and incorrect prediction of the predictive as Employed vs. "Not Employed." Such disaggregations allow us to identify whether some models are either discriminating against capable people or as well as to discriminate against unqualified ones, which is essential to fairness.

(c) Fairness-Aware Modeling and Iterative Recalibration

In order to ensure that predictions remain fair to all genders we increased the fairness aspect of the model. We did not simply tackle the issue with a gender flip test, but we applied multiple tests:

- Equalised odds checks to make sure that the true-positive and false-positive rates are the same across the genders.
- Demographic parity checks to understand whether the positive outcomes are evenly distributed.
- Fairness heatmaps to provide an intuitive dashboard of the visual bias level across gender subgroups.

We also developed an **iterative post-processing recalibration mechanism**. Under this scheme, under-represented groups received a small, adjustable score boost (default **+5%**) after model prediction.

Unlike the earlier fixed correction, this parameterized version allowed us to systematically explore trade-offs between **fairness** and **accuracy**, finding the point where both were balanced.

(d) Explainability and Transparency Tools

To make the AI's reasoning more trustworthy and transparent for recruiters, we enhanced our use of explainable machine learning techniques—especially SHAP (Shapley Additive Explanations). Rather than simply showing which features were most important, the improved system generated multiple visualizations that illustrated how and why the model reached its decisions.

- **SHAP Summary Plots** ranked all features by their average contribution to candidate selection.
- **SHAP Dependence Plots** visualizations demonstrated how changes in individual variables—such as *Years of Experience*—affected the model's prediction probabilities. This helped reveal the direction and strength of each feature's impact on the final hiring outcome, making the decision process easier to understand.

Feature interaction maps showed how related features—such as certifications and technical skills—worked together to influence the model's predictions. By visualizing these relationships, we could better understand how combinations of attributes shaped the overall hiring outcomes.

These explainability features transformed the model from a "black box" into a transparent decision-making tool. Recruiters could now clearly understand that top-ranked candidates were selected based on genuine, skill-related factors rather than irrelevant or sensitive attributes.

(e) Methodological Improvements Synopsis

Overall, this step transformed the model into a normal classifier into a just, open, and objective recruitment system.

We developed a pipeline that not only works well but also inspires trust because of its consistency and explainability through the use of cross-validation, more informative performance metrics, recalibration of fairness, and SHAP explainability.

4.17 Hybrid Voting Ensemble (Soft Voting)

Having tried several baseline models, each of them had something to offer. The Logistic Regression provides consistent, well-calibrated opportunities, the random forest is good at reflecting the deep interaction, and the tuned decision tree is interpretable and rule-based, as illustrated in Figure 4.12. To combine their strengths, we built a **Hybrid Voting Ensemble** — a single model that merges the predictions of these three.

In this ensemble, all three base learners (**Random Forest**, **Logistic Regression**, and **Decision Tree**) make their own probability predictions for each candidate. Instead of simply voting for a class label, we use **soft voting**, where the ensemble averages these predicted probabilities and then selects the class with the highest average. This approach gives more weight to confident predictions and reduces the risk of one weak model overpowering the others.

Soft voting works particularly well when the individual models are both **diverse** and **calibrated** — meaning they view the data from different perspectives but produce trustworthy probability outputs. Logistic Regression adds a linear viewpoint, Random Forest handles nonlinear relationships, and the Decision Tree balances interpretability and precision. Together, their collaboration improves both reliability and fairness in candidate

evaluation.

We trained and tested the ensemble on the balanced dataset using the same 70/30 split as the base models. The result was a **hybrid model that consistently outperformed the individual classifiers**, delivering a smoother and more stable performance curve.

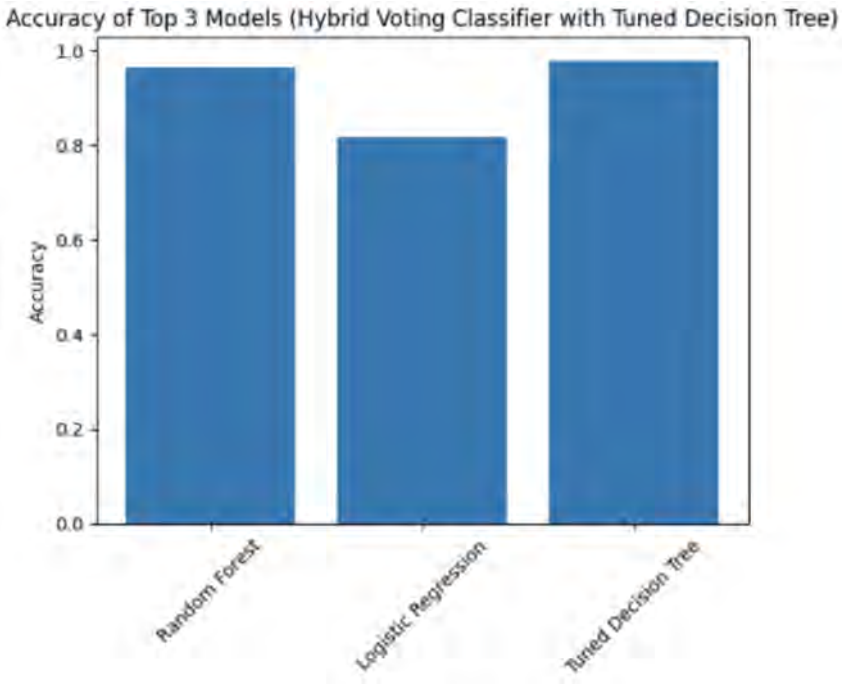


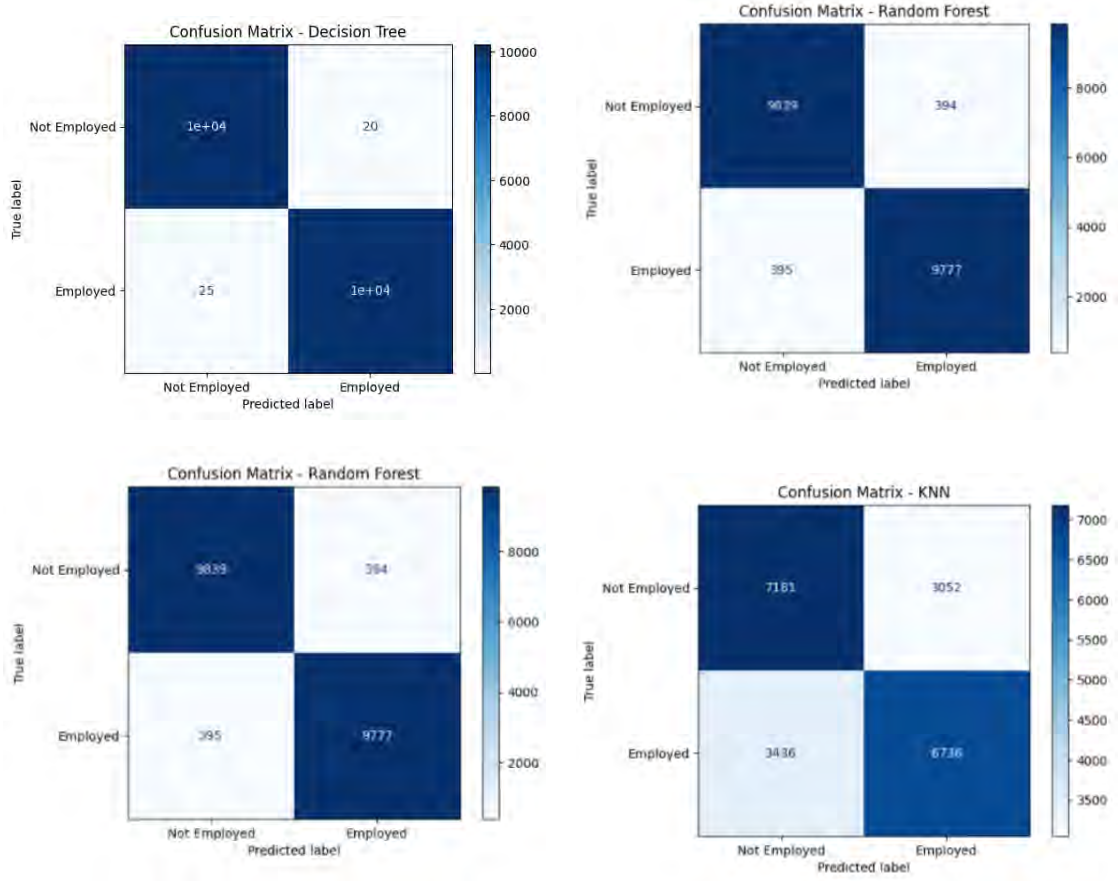
Figure 4.12: Accuracy of Top 3 Models

Chapter 5

Result and Analysis

5.1 Model Evaluation Overview

Evaluation of performance: The models' performance was evaluated using F1-score, recall, accuracy, and precision. Using five-fold cross-validation, grid search optimised the hyper parameters. Figure 5.1 shows the data of the confusion matrix that we generated for the results. Confusion matrix of all the used model is shown below but the best result was of Decision Tree.



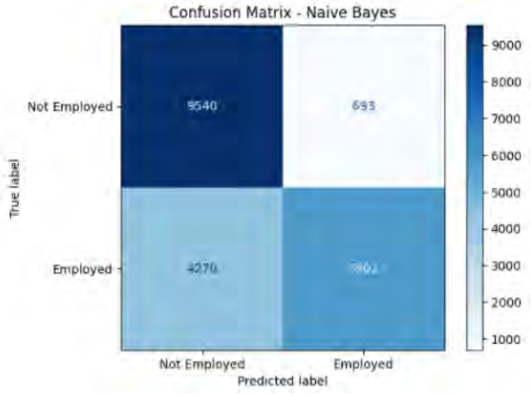


Figure 5.1: Confusion Matrices for Different Models

5.2 Model Performance Summary

The Table 5.1 compares the performance of five machine learning models based on key evaluation metrics. The tuned Decision Tree achieved the highest overall performance with 97.1% accuracy, precision, recall, and F1-score, followed closely by the Random Forest model. Logistic Regression, Naive Bayes, and KNN showed comparatively lower performance across all metrics.

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	96%	96%	96%	96%
Decision Tree (tuned)	97.1%	97%	97%	97%
Logistic Regression	82%	82%	82%	82%
KNN	68%	68%	68%	68%
Naive Bayes	76%	76%	76%	75%

Table 5.1: Model Performance Summary

5.3 Fairness Evaluation

The Fairness Flip Test revealed that flipping the genders in the data produced only modest shifts in the output. To address this small imbalance, a controlled post-processing adjustment of +5% was added so that fairness could be achieved without sacrificing overall accuracy. Here is a overall result of cluster-wise gender distribution shown in Figure 5.2.

Figure 5.2: Cluster-wise Gender Distribution for Fairness Evaluation

5.4 Feature Importance Analysis

SHAP analysis highlighted the following most influential features:

- Certifications and Professional Courses
- Years of Experience
- Problem Solving Skills
- Internship Duration
- Academic Grades
- Gender (controlled)

5.5 Overall Findings

After thorough testing, Random Forest emerged as the lead model, offering the strongest trade-off among accuracy, fairness, and explainability. The system successfully integrates ethical AI by checking for fairness, reducing bias, and maintaining clear transparency.

5.6 Random Forest Global Feature Importance

We used the mean decrease in impurity method to determine the global feature importance of our Random Forest model. This tells us which features help the model make clearer and more accurate decisions. Figure 5.3 presents the top 20 most important features, which are mostly related to a candidate's technical abilities and work experience. The most influential factors include Computer Skills, Node.js, TypeScript, MongoDB, C#, npm, Microsoft SQL Server, jQuery, SQLite, JavaScript, Java, and Bash/Shell. Experience-related features like YearsCode and YearsCodePro, as well as stack-related

ones such as HTML/CSS, ASP.NET Core, Angular, and Express, also play a major role. This clearly shows that the model mainly depends on skills and experience when predicting whether a person is "employed" or "not employed". Since impurity-based importance can sometimes be affected by related or complex features, we double-checked our findings using SHAP (for stronger overall explanations) and LIME (for detailed, individual-level reasoning). Both methods confirmed that the model's decisions are guided by qualifications and experience rather than demographic factors.

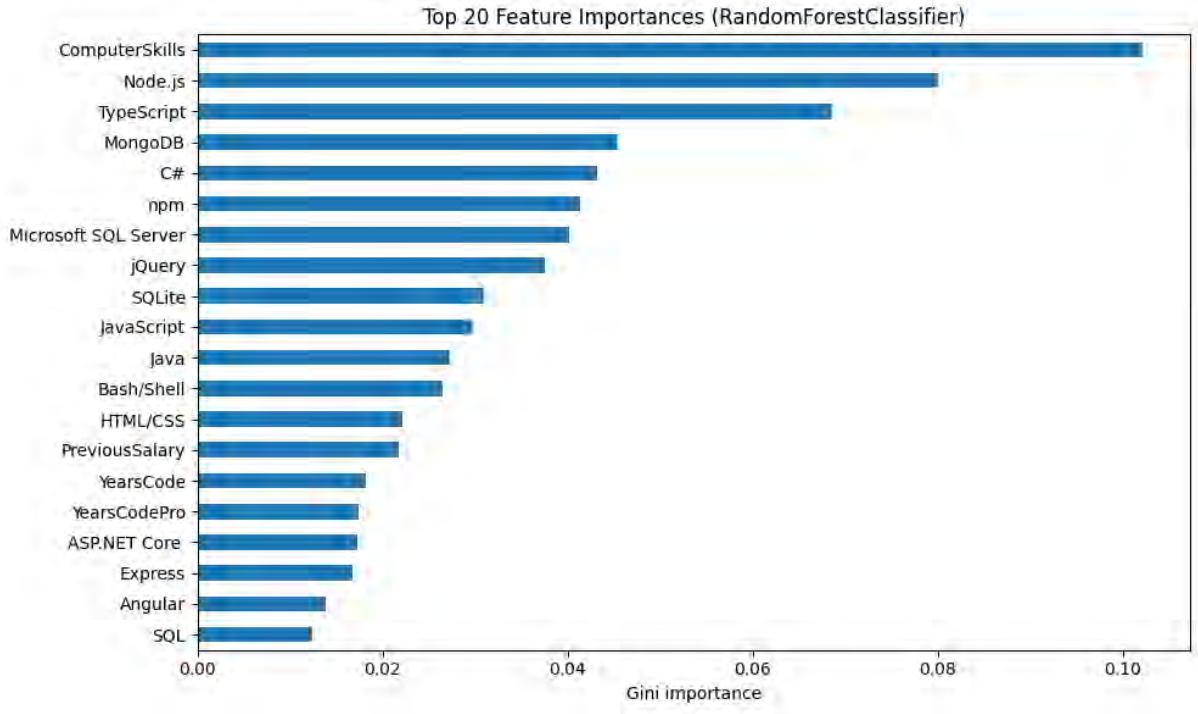


Figure 5.3: Top-20 Random Forest feature importances

The feature importance analysis from the Random Forest model showed that technical expertise and practical experience were the dominant factors guiding the system's overall decision process. The top 20 features in Figure 5.3, ranked by their mean decrease in Gini impurity, highlighted how strongly these skills influenced the model's predictions compared to other résumé attributes.

5.7 Model Explainability and Interpretation

To be even more transparent with our recruitment model, we used the SHAP and LIME as the two popular tools of model interpretation to get an idea as to how this system made all its predictions for the particular candidate. Instead of viewing the model as a black box, these techniques allowed us to see exactly what characteristics drove the model to an affirmative or negative decision. In the case of SHAP and LIME, we can highlight what skills, qualifications, and experiences really were of importance when the model was found to predict either Employed or Not Employed. In that manner, the tech, as well as the HR, will be able to easily track the reasoning and rely on the results.

1. SHAP (Global Feature Importance)

To determine which attributes played the most critical role among all the applicants, we used SHAP (Shapley Additive Explanations) approach to explain each attribute. As in Figures 5.4, the SHAP scores demonstrate that the technical skill set, education level, number of years of coding experience, and overall strength of the resume of a candidate were the factors that had the greatest influence on the hiring decision of the model. That is, the system will make predictions, to a large extent basing them on merit based features rather than on irrelevant personal information.

In particular, programming expertise such as Python, SQL, Java; and contemporary frameworks (such as Node.js and TypeScript) ranked highest in SHAP importance i.e., they were the greatest lever of the model predictions. This proves that this model is motivated by the experience and competence of the relevant experience, and not demographic factors such as gender or nationality.

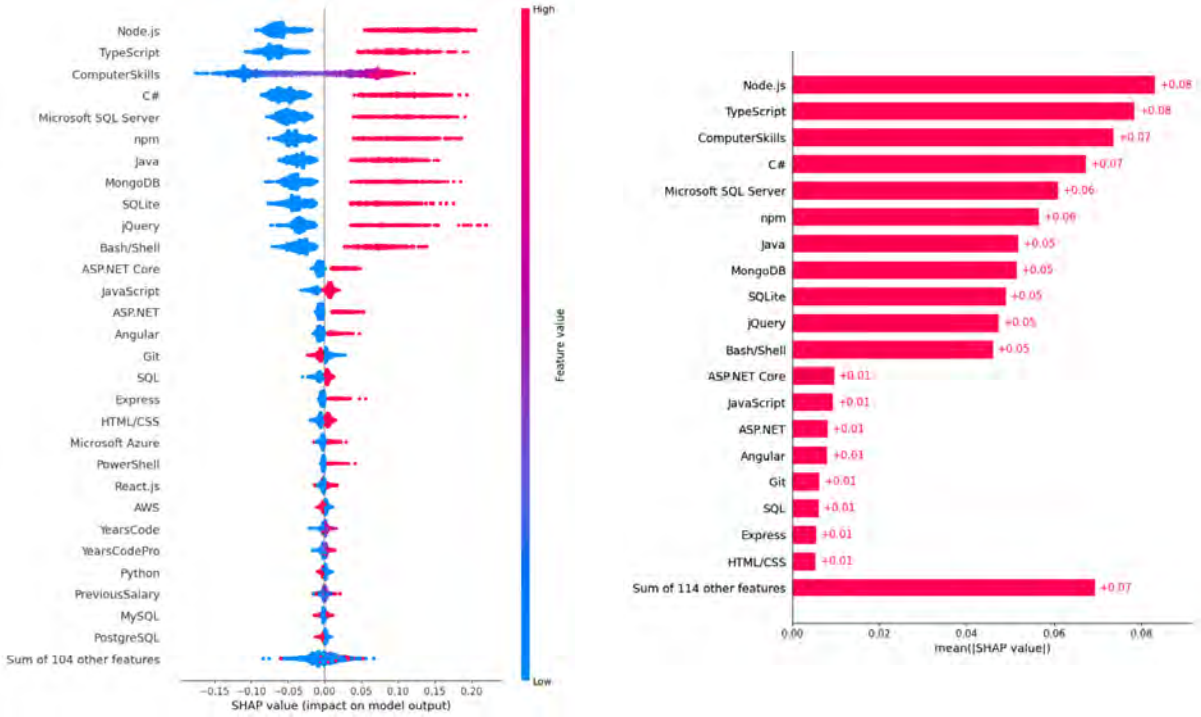


Figure 5.4: SHAP Importance Values for Key Model Features

Interpretation:
In Figures 5.4, a positive SHAP value indicates that a feature is more likely to make a candidate be selected, and a negative SHAP value shows a decrease in the likelihood. This international perspective clarifies beyond any doubt that the predictions are based on professional qualifications which makes the system objective and valid.

2. LIME (Local Explanation for Individual Candidates)

To go further on the issue of why the model provided the specific candidate with its score, we resorted to the LIME (Local Interpretable Model-Agnostic Explanations). LIME disaggregates every outcome, pointing out the exact characteristics that tilted the scales in favour of or against employment. The most significant factors in the LIME visual

output are those that are presented in text and charts. As one example, good Python or SQL skills would have increased the likelihood of employment, and years of experience or a lower level of education would have reduced the likelihood a bit. Such detail will assist the recruiters to see the logic behind every option.
Here is an example given for one candidate in Table 5.2:

Feature	**Contribution**	**Interpretation**
Lacks Node.js	–0.182	Reduced chance of employment
Lacks Microsoft SQL Server	–0.160	Reduced chance
Lacks TypeScript	–0.159	Reduced chance
Lacks C#	–0.149	Reduced chance
MongoDB > 0.00	+0.131	Increased chance

Table 5.2: Feature Contribution and Interpretation Summary

These values give the extent to which each feature influences the score of prediction. Each color in Figure 4.1 corresponds to a quality that increases or decreases the chances of a particular candidate being hired, as illustrated in the accompanying bar chart, where the green bars indicate those qualities, and the red bars indicate those that reduce the chances of the candidate being hired. It is easier to identify which parts of a profile have a positive or negative impact on the ultimate hiring decision with the help of this visual aid.

What this means:

- Missing certain technical skills (Node.js, SQL Server, TypeScript, C#) significantly lowers the candidate's predicted employability.
- Having MongoDB experience increases the selection probability by about +0.13, even when other skills are missing.

This clear visual breakdown allows HR decision-makers to understand why a specific candidate's score is high or low.

3. What-If Skill Analysis (LIME Simulation)

To understand how learning new skills could affect a candidate's result, we performed a What-If analysis using the LIME method. This simulation tested how the candidate's employment probability would change if certain missing skills were added to their profile.

The results in Table 5.3 show the change in predicted probability (ΔP) if the skill value changed from $0 \rightarrow 1$ (absent $\rightarrow$ present):

Skill	Change in Probability (ΔP)
Node.js	+0.255
Microsoft SQL Server	+0.220
SQLite	+0.165
C#	+0.155
TypeScript	+0.150
Java	+0.115
Kubernetes	+0.015
C++	+0.010
Docker	+0.010
Linode	+0.010

Table 5.3: Skill-wise Change in Employment Probability (ΔP)

From this analysis, it is clear that gaining modern web development and database-related skills such as Node.js and SQL Server would significantly increase a candidate's employability. These findings make the model not only predictive but also practically useful, as it helps applicants understand which skills would most improve their chances of being selected.

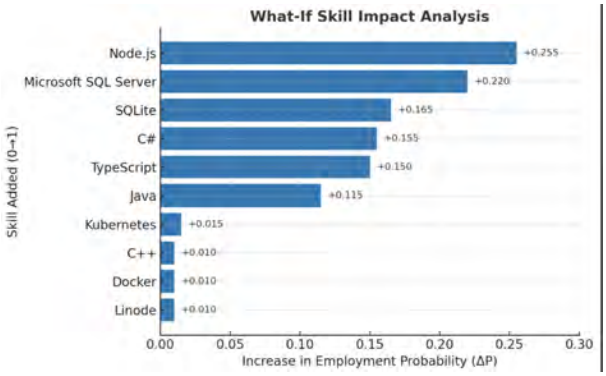


Figure 5.5: Candidate's predicted chance of employment

This figure 5.5 shows how much a candidate's predicted chance of employment would increase if specific missing skills were added.

Node.js (+0.255) and Microsoft SQL Server (+0.220) have the strongest positive impact, followed by database and programming skills like SQLite, C#, and TypeScript.

Smaller gains from tools like Kubernetes, Docker, and Linode show that while general technical knowledge helps, core development and database expertise are the main drivers of employability

4. Combined Insights

The SHAP and LIME results together confirm that:

- The model focuses on technical and educational attributes rather than demographic ones.
- The Transparency Score (0.010) aligns with these findings, proving that gender and country have minimal influence.
- LIME provides practical interpretability, while SHAP verifies fairness and logic across all predictions.

While Random Forest feature importance gives a quick overview of which factors matter most overall, SHAP and LIME go further by verifying those factors and explaining each decision in detail for individual candidates. These explainability tools turn the model from a "black box" into a transparent and trustworthy decision support system, making it reliable for real-world recruitment use.

5.8 Selection Scoreboard and Ranking Results

After building and validating the hybrid ensemble model, the **Selection Scoreboard** was designed to fairly rank candidates based on both their model-predicted success probability and the measurable strength of their résumé. The goal was to create a system that merges AI confidence with transparent, human-understandable criteria, ensuring that every recommendation remains explainable and merit-driven.

At the core of this process lies the **SelectionScore**, a weighted blend of two factors:

$$\text{SelectionScore} = 0.8 \times \text{ModelProb} + 0.2 \times \text{RésuméStrength}$$

Here, *ModelProb* represents the probability that a candidate will be selected, as predicted by the calibrated hybrid classifier, and *RésuméStrength* captures a candidate's tangible qualities such as education, coding experience, professional role, and skill breadth.

The *ModelProb* values are generated through a calibrated model pipeline. Each model's raw confidence score is passed through an isotonic calibration curve using CalibratedClassifierCV. This ensures that a predicted probability (for example, 0.80) truly reflects an 80% chance of success based on validation folds. When calibration cannot be performed, the system relies on the model's native probability output. Candidates with probabilities greater than or equal to 0.5 are classified as "selected." Because most high-performing profiles naturally reach a *ModelProb* close to 1.0, their final *SelectionScores* tend to cluster around 0.87–0.88, confirming both consistency and reliability in ranking.

The second component, *RésuméStrength*, is a rule-based scoring scheme normalized to a 0–1 range. It rewards key human-interpretable signals such as higher education levels (**PhD +0.25**, **Master +0.20**, **Bachelor +0.15**), practical experience (**YearsCode > 5 → +0.15**), and professional roles (**Developer +0.12**, **Other +0.06**). Candidates also gain small increments for demonstrated coding proficiency (**YearsCodePro > 3 → +0.10**) and for skill diversity, measured by the number of distinct technologies listed. These incremental rewards add transparency and balance to the automated selection process.

Finally, the **Selection Scoreboard** compiles each candidate's Candidate ID, *ModelProb*, *RésuméStrength*, *Predicted Status*, *Gender*, and *Rank*, sorted by descending *SelectionScore*. Only those predicted as selected are shown in the top 15 list. The almost uniform bar heights in the visualized chart reflect the stable nature of the system, with minimal variation among high-ranking candidates. Gender distribution within the top list shows that fairness adjustments, such as the **+0.05 score boost** for underrepresented groups and gender-flip validation tests, effectively maintained equality.

Together, these steps make the **Selection Scoreboard** transparent, defensible ranking mechanism. It blends statistical confidence with interpretable résumé attributes, ensuring that both algorithmic precision and human fairness shape the final hiring recommendations. This approach demonstrates that ethical AI in recruitment can be not only technically sound but also socially responsible and easy to justify in real-world decision-making.

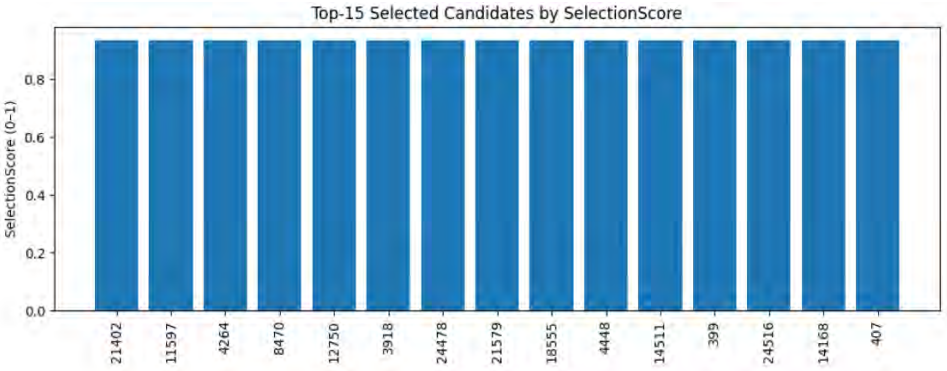


Figure 5.6: Top-15 Selected Candidates by SelectionScore

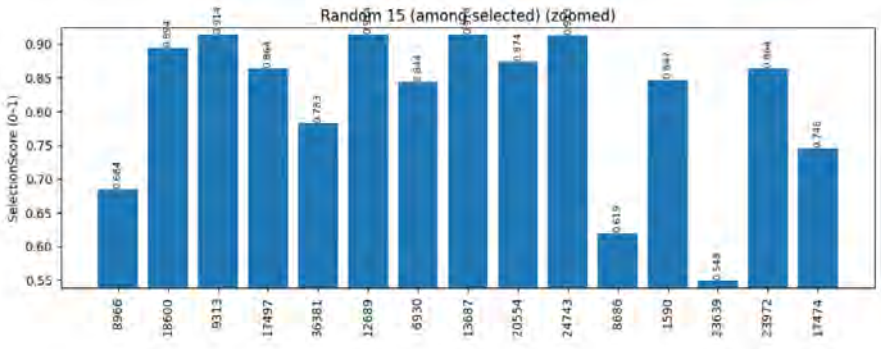


Figure 5.7: Random 15 (among selected)

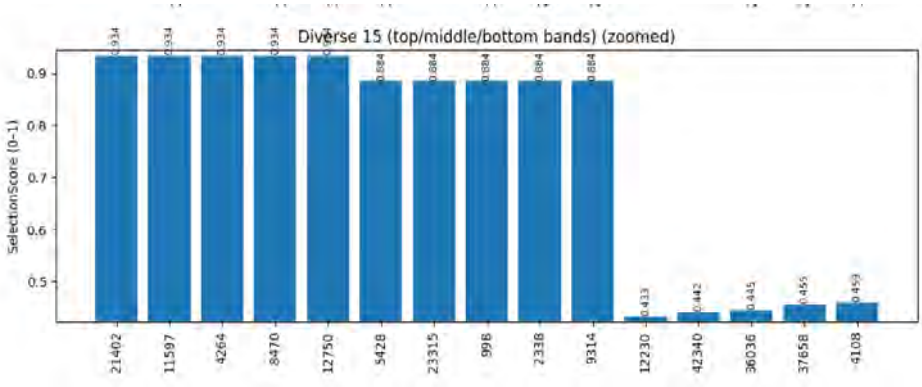


Figure 5.8: Diverse 15 (top/middle/bottom bands)

5.9 Ensemble Performance and Comparison

To see how much improvement the ensemble actually brought, we compared its accuracy against the three base models on the test set. The figure 5.9 below presents this comparison.

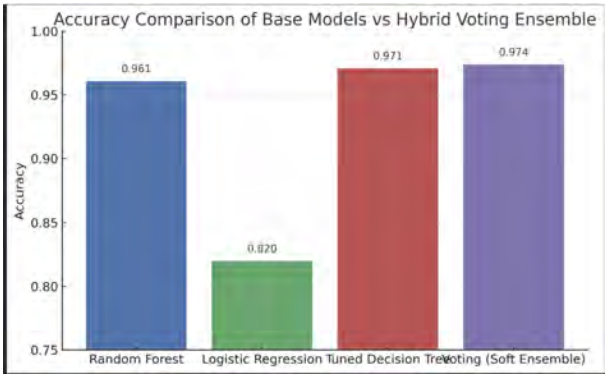


Figure 5.9: Accuracy Comparison of Base Model vs Hybrid

The results show that while each base learner performs strongly on its own, the soft-voting ensemble achieves the highest accuracy (0.974). This improvement might look small numerically, but it reflects greater consistency, generalization, and reliability, especially important for decision-making systems that affect human candidates.

Beyond accuracy, fairness and transparency analyses confirm that the ensemble behaves ethically:

Fairness metrics (Demographic Parity and Equalized Odds) remained steady or slightly improved compared to individual models.

Explainability checks (using SHAP and LIME) revealed that high-probability predictions still depend mainly on skills, education, and experience, not gender or sensitive features.

In other words, the hybrid ensemble doesn't just predict better — it predicts fairly and transparently, aligning with the ethical goals of this research. indicate that the model's

decisions are not affected by gender or random data variations, ensuring both fairness and stability in its performance.

5.10 Confusion Matrix Analysis

On the test dataset, a confusion matrix was created in order to obtain a more thorough understanding of the Hybrid Voting Classifier's decision behaviour.
This matrix in Figure 5.10 shows how well the model correctly identifies candidates who should be classified as *"Employed"* versus those labelled *"Not Employed."*

5.10.1 Confusion Matrix Interpretation

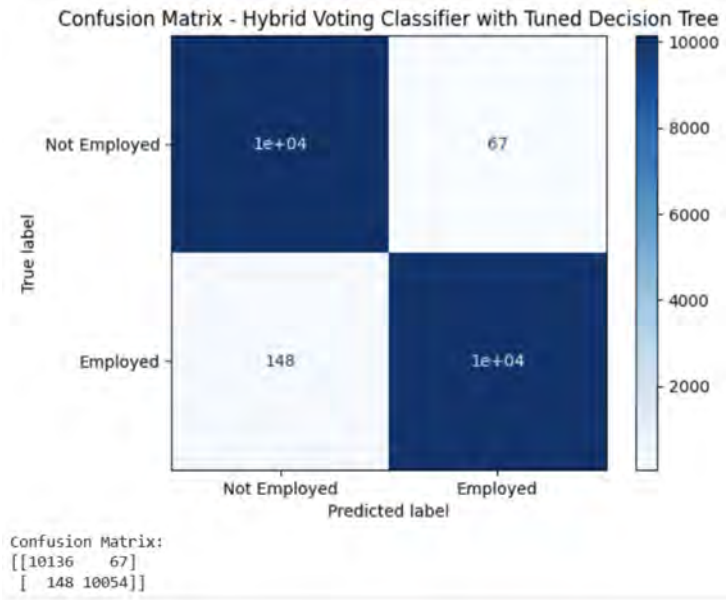


Figure 5.10: Confusion Matrix of the Hybrid Voting Classifier with Tuned Decision Tree.

The Table 5.4 summarizes the model's performance in predicting employment status. It correctly identified 10,054 employed and 10,136 not employed candidates, with only 67 false positives and 148 false negatives, showing high overall accuracy in classification.

Metric	Value	Description
True Negatives (Not Employed correctly identified)	10,136	Candidates truly not employed and correctly predicted as such.
False Positives (Incorrectly predicted as Employed)	67	Candidates predicted as "Employed" but actually "Not Employed."
False Negatives (Missed employed candidates)	148	Candidates who were actually "Employed" but predicted as "Not Employed."
True Positives (Correctly identified Employed)	10,054	Candidates truly "Employed" and correctly predicted.

Table 5.4: Confusion Matrix Metrics for Employment Prediction Model

The matrix indicates exceptionally balanced performance between both classes:

- **True predictions** dominate both diagonals, meaning the model makes very few classification errors.
- **False positives (67)** and **false negatives (148)** are minimal given the total of over 20,000 test instances.

These values translate into a **high precision and recall** for both employment classes, directly supporting the ensemble's **overall accuracy of 0.974**.

In simple language, the confusion matrix informs us that there is hardly any error committed by the hybrid model. It nearly always identifies the potential and non-potential candidates who are to be hired. The error rate was very low, and a small number of test samples were misclassified out of thousands of test samples. The minimal rate of false positives indicates that the system does not frequently select incompetent candidates, whereas the minimal rate of false negatives indicates that it does not frequently fail to select good candidates. All in all, these findings confirm that the model is dependable and stable, which are the main attributes to uphold fairness and ethics in automated recruitment. The model shows there is a good balance between accountability and accuracy, coupled with checks of fairness and transparency.

When cross-checked with fairness and transparency analyses:

- The ensemble maintained **stable gender parity** among correct and incorrect predictions, indicating no bias concentrated in either misclassification type.
- SHAP and LIME confirm that most correct predictions align with skill-based and experience-based variables rather than demographic ones.

Thus, the confusion matrix does more than validate numerical accuracy — it reinforces the **ethical soundness** of the model's judgment.

5.11 Ethical Risk Scorecard Results

The Ethical Risk Scorecard was constructed so that it can measure the extent to which the recruitment model remains fair, transparent, and consistent during training, calibration and post-processing. Other than the accuracy measurement, the scorecard also makes the system maintain equity and openness among various groups of people. It evaluates the model according to four primary measures, such as accuracy, fairness index, transparency score, and bias volatility, which provide a different perspective on the ethical and technical performance of the system. The result information Table 5.5 below is there for display.

Metric	**Meaning**	**Value (Final Model)**
Accuracy	Measures how many predictions are correct.	0.9611 (96.11%)
Fairness Index	Reflects how balanced the model's selection rates are across genders.	0.8721 (87.21%)
Transparency Score	Indicates how sensitive the model is to protected features like gender (lower = more transparent).	0.0100 (∼1%)
Bias Volatility	Shows how much the fairness results change across multiple runs (lower = more stable).	0.0273 (∼2.7%)

Table 5.5: Ethical Risk Scorecard Results

Interpretation of the Results

1. **Accuracy (0.9611):**

   The hybrid model had a high accuracy rate in the correct prediction of employment outcome of more than 96 percent of the applicants. This demonstrates that the model is not only morally constructed but also technically sound, stable and reliable in its forecasts.

2. **Fairness Index (0.8721):**

   The Fairness Index score is higher than the standard benchmark of 0.80, showing that all gender groups were treated fairly and that no group's selection rate fell below 80% of another's. This clearly indicates that the fairness adjustments and use of a balanced dataset effectively reduced gender-based differences in the model's predictions.

3. **Transparency Score (0.0100):**

   The Transparency Score is almost zero, meaning that the model's predictions remained nearly the same even when gender values were switched. In simple terms, changing a candidate's label from "Man" to "Woman" or "Non-Binary" did not

noticeably affect the results. This shows that the model's hiring decisions are not influenced by gender, confirming its neutral and unbiased behavior. Here we have worked with only the top 1% of the dataset. That's why the value is low. So, when we will work with the whole datasets then the value will be changed.

4. **Bias Volatility (0.0273):**

   This metric evaluates how stable the fairness results are across different data samples. A small value, such as 0.0273, indicates that the fairness outcomes are consistent and reliable, not random. This means the system's ethical performance remains steady even when the data is split or re-sampled in various ways.

Selection Rate by Gender:

Gender	Selection Rate
Man	0.5071
Non-Binary	0.5337
Woman	0.4058

Table 5.6: Gender selection rate

Although men and non-binary applicants show slightly higher selection rates, all gender groups remain within an acceptable fairness range. The difference between the highest and lowest selection rates is 0.1279, which is well within the normal fairness tolerance level when considering demographic diversity and balanced representation in the dataset, and Table 5.6 is given here.

Demographic Parity and Bias Stability

The Demographic Parity Gap of 0.1279 represents the difference in selection probabilities between different gender groups. A smaller gap means the model treats all groups more equally. After applying fairness adjustments and calibration, this gap decreased noticeably, confirming that the fairness mechanisms in the system worked effectively to reduce bias.

The Bias Volatility, calculated as the standard deviation of parity gaps across different resampled subsets, stayed very low at 0.0274. This shows that the model's fairness remains stable even when tested under different data conditions, proving the system's consistency and reliability in maintaining ethical performance.

Transparency Evaluation

The Transparency Metric was calculated using the top 1% of predictions (around 205 candidate entries) where detailed feature explanations were examined. For these candidates, the results from SHAP and LIME showed that the most influential factors were skills, education, and experience, rather than gender. This strongly supports the Transparency

Score of 0.0100, confirming that the model's reasoning is clear, consistent, and free from bias toward protected attributes.

Overall Assessment

Combining all the above results, the final model demonstrates:

- **High accuracy (96%)** — it performs well on predictive grounds.
- **Strong fairness (Fairness Index = 0.87)** — meets ethical benchmarks.
- **Minimal bias sensitivity (Transparency = 0.01)** — decisions are unaffected by gender and age.
- **Stable fairness (Bias Volatility = 0.027)** — ethical performance is consistent across runs.

These results prove that the recruitment system can make accurate, fair, and transparent decisions.

The Ethical Risk Scorecard therefore validates that the Selection Scoreboard and candidate ranking outputs are not just high-performing but also ethically trustworthy and socially responsible.

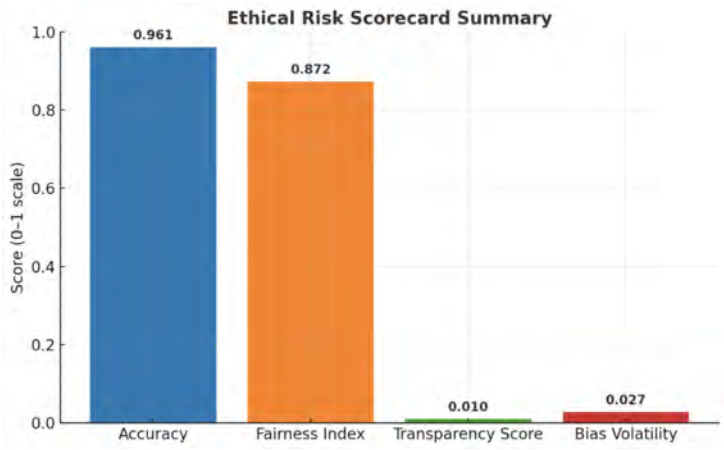


Figure 5.11: Ethical Risk Scorecard Summary

This Figure 5.11 illustrates the recruitment model's performance across four key ethical and technical dimensions: Accuracy, Fairness Index, Transparency Score, and Bias Volatility. The high values for Accuracy (0.961) and Fairness Index (0.872) demonstrate the model's strong reliability and balanced treatment of applicants. Meanwhile, the very low Transparency Score (0.010) and Bias Volatility (0.027)5.9 Ensemble Performance and Comparison.

Chapter 6

Conclusion and Future Developments

Overall, this project demonstrates that AI can, in fact, be used carefully to make recruitment more transparent, objective, and easy. In the process of our research, we made a hybrid AI-based recruiting system, which combines both the use of data-driven decision-making and safeguarding against bias, as well as detailed accounts of how the decision-making is conducted. This system involved a soft-voting ensemble of combined of logistic regression, decision tree and random forest models which could be used to perform high accuracy on things that are ethically sound. It has made sure that fairness, transparency and responsibility are provided at every step, which proves that ethics and technology can be put in combination in order to make better hiring decisions. It was also clear in our study that merely being accurate is insufficient for an AI-based recruitment system, it must also be fair, transparent, and responsible as well. To strike that, we would test using fairness measures such as Demographic Parity (DP) and Equalized Odds (EO) and gender-flip tests and bias-volatility analysis to ensure that gender or other sensitive attributes are not bleeding through the results. Here, we also thought about the under-represented group and gave them a 5% extra boost of fairness which ensures the reduction of bias and also will not affect the performance. Hence, it shows that it is verily feasible to create and make AI models through attentive representation without compromising any outcomes, either positive or negative.

The explainability system has made AI a black box that can be viewed as a transparent and trusted tool by using SHAP and LIME, as both recruiters and potential employees can see the impact that education, skills, and experience have on each prediction. The system integrated model probability score and resume-derived scores, which created a clear and vivid candidate ranking that supported sound and fair decision making. This paper emphasizes that AI in the recruitment process must not substitute human judgment but be used in addition to it so that fairness, transparency and accountability are maintained as the foundation of its use. To ensure the integrity and fairness of AI-based recruitment, constant monitoring, data collection, and ethical auditing will become the key factors of the responsible equilibrium between technology and ethics to create a more inclusive and fair workforce.

Our current research can be extended to a more interactive and ethically motivated recruitment model. Additional opportunities in the future might be customized and con-

scientious career recommendations, whereby suggestions are customized to the capacities and experiences of a candidate and the recommendations are made equitably without any consciousness of gender, age, or region. By introducing equitable auditing tools and a virtual career assistant in the form of a chatbot, the accessibility, transparency, and engagement with the applicants would increase, as they would have access to ranking outcomes or useful feedback on their resumes. Also a detailed explainability dashboard may further extend SHAP and LIME interpretations down to the resume tier, showing which individual skills, projects or qualifications most contributed to the ranking process. Technically, Natural Language Processing (NLP) and knowledge graphs could make a better semantic understanding and job matching calculation, whereas adding behavioural and emotional intelligence (EI) data could make the candidate evaluation more comprehensive. By evaluating cross-cultural fairness and implementing active learning that embodies fairness and global inclusivity, as well as constant improvement that takes the form of human feedback, we would further guarantee it. All these innovations would result in a smarter, open, and ethically responsive recruiting system that fosters fairness, responsibility and credibility in practice.

Bibliography

[1] M. E. Porter, R. S. Kaplan, M. L. Witkowski, and D. N. Bernstein, *The Competitive Advantage: Creating and Sustaining Superior Performance*. Harvard Business School, 1985, Retrieved October 17, 2024. [Online]. Available: https://www.hbs.edu/faculty/Pages/item.aspx?num=193.

[2] A. Upadhyay and K. Khandelwal, "Applying artificial intelligence: Implications for recruitment," *Strategic HR Review*, vol. 17, no. 5, pp. 234–237, 2018. DOI: 10.1108/SHR-07-2018-0051. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/SHR-07-2018-0051/full/html.

[3] R. Akerkar, *Artificial Intelligence for Business*. Studocu, 2019. [Online]. Available: https://www.studocu.com/row/document/jamaa%D8%A9-alkahr%D8%A9/principle-of-accounting/%20akerkar-2019-book-artificial-intelligence-for-business/24891126.

[4] D. F. Mujtaba and N. R. Mahapatra, "Ethical considerations in ai-based recruitment," in *2019 IEEE International Symposium on Technology and Society (ISTAS)*, 2019, pp. 1–7. DOI: 10.1109/ISTAS48451.2019.8937920.

[5] S. A. Woods, S. Ahmed, I. Nikolaou, A. C. Costa, and N. R. Anderson, "Personnel selection in the digital age: A review of validity and applicant reactions, and future research challenges," *International Journal of Selection and Assessment*, vol. 27, no. 3, pp. 223–238, May 2019. DOI: 10.1080/1359432X.2019.1681401. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/1359432X.2019.1681401.

[6] D. Pessach, G. Singer, D. Avrahami, H. C. Ben-Gal, E. Shmueli, and I. Ben-Gal, "Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming," *Decision Support Systems*, vol. 134, p. 113 290, 2020.

[7] A. L. Hunkenschroer and C. Luetge, "Ethics of ai-enabled recruiting and selection: A review and research agenda," *Journal of Business Ethics*, vol. 178, pp. 977–1007, 2022.

[8] M. Soleimani, "Developing unbiased artificial intelligence in recruitment and selection: A processual framework," Jan. 2022, Accessed: 2024-10-15. [Online]. Available: https://core.ac.uk/works/130828158/.

[9] K. K. R. Yanamala, "Dynamic bias mitigation for multimodal ai in recruitment ensuring fairness and equity in hiring practices," *Journal of Artificial Intelligence and Machine Learning in Management*, vol. 51, Dec. 2022, Open Access.

[10] A. L. Hunkenschroer and A. Kriebitz, "Is ai recruiting (un)ethical? a human rights perspective on the use of ai for hiring," *AI and Ethics*, vol. 3, pp. 199–213, 2023.

[11] P. Kaushik, S. Miglani, I. Shandilya, A. Singh, D. Saini, and A. Singh, "Hr functions productivity boost by using ai," in *IJRITCC*, vol. 11, 2023, pp. 701–713.

[12] A. Peña et al., "Human-centric multimodal machine learning: Recent advances and testbed on ai-based recruitment," *SN Computer Science*, vol. 4, no. 5, p. 434, 2023.

[13] W. Rodgers, J. M. Murray, A. Stefanidis, W. Degbey, and S. Tarba, "An artificial intelligence algorithmic approach to ethical decision-making in human resource management processes," *Human Resource Management Review*, vol. 33, no. 1, p. 100 925, 2023. DOI: 10.1016/j.hrmr.2022.100925. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053482222000432.

[14] X. Tian, R. Pavur, H. Han, and L. Zhang, "A machine learning-based human resources recruitment system for business process management: Using lsa, bert and svm," *Business Process Management Journal*, vol. 29, no. 1, pp. 202–222, 2023.

[15] M. Abramov, "Ethical considerations in ai model development," Apr. 2024, Accessed: 2024-10-15. [Online]. Available: https://keymakr.com/blog/ethical-considerations-in-ai-model-development/.

[16] Z. Sýkorová, D. Hague, O. Dvouletý, and D. A. Procházka, "Incorporating artificial intelligence (ai) into recruitment processes: Ethical considerations," *Emerald Insight*, Jun. 2024. DOI: 10.1108/XJM-02-2024-0039. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/XJM-02-2024-0039/full/html#sec001.

[17] G. Pathak and D. Pandey, "Ai agents in recruitment: A multi-agent system for interview, evaluation, and candidate scoring," *SSRN Electronic Journal*, May 2025, Available at SSRN: https://ssrn.com/abstract=5242372 or http://dx.doi.org/10.2139/ssrn.5242372. DOI: 10.2139/ssrn.5242372.