Arab American University

Faculty of Graduate Studies

A Machine Learning and Deep Learning Approach for Selecting the Most Suitable Teacher at the Ministry of Education

By

Yasmeen "Mohammad Khaled "Yaseen Salman

Supervisor

Dr. Ismail Khater

Co- Supervisor

Dr. Mohammad Hussein

This thesis was submitted in partial fulfillment of the requirements for the Master's degree in Data Science and Business Analytics.

10 /2024

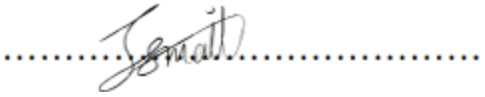
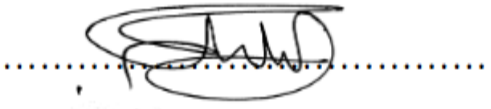
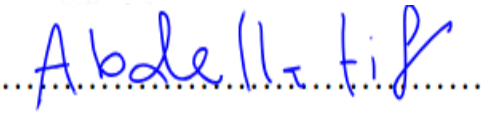
Thesis Approval

A Machine Learning and Deep Learning Approach for Selecting the Most Suitable Teacher at the Ministry of Education

By

Yasmeen "Mohammad Khaled "Yaseen Salman

This thesis was defended successfully on 6/10/2024 and approved by:

Committee members	Signature
1. Dr. Ismail Khater: Supervisor	
2. Dr. Mohammad Hussien: Co- Supervisor	
3. Dr. Ahmad Hassasneh: Internal Examiner	
4. Dr. Abdallatif Abuissa: External Examiner	

Declaration

I hereby declare that this thesis entitled "A Machine Learning and Deep Learning Approach for Selecting the Most Suitable Teacher at the Ministry of Education,2024 submitted in partial fulfillment of the requirements for the degree of Data Science and Business Analytics is my own original work and has not been submitted previously, in whole or in part, for any other degree or diploma at this or any other institution.

I confirm that all sources of information and data have been appropriately acknowledged and referenced. I further declare that the thesis does not contain any material previously published or written by another person except where due reference is made.

The Name of The Student: Yasmeen “Mohammad Khaled “Yaseen Salman

ID: 202112865

Signature: Yasmeen “Mohammad Khaled “Yaseen Salman

Date: 22.6.2025

Dedication

I express my gratitude to my family, whose unwavering love and support have been instrumental in my achievements, as I dedicate this thesis to them. To my parents, I express my gratitude for having faith in me even during moments of self-doubt, for fostering my inquisitiveness, and for motivating me to relentlessly chase after my aspirations. If it weren't for your steadfast direction, I wouldn't have had the bravery to begin this endeavor.

I would like to express my gratitude to my mentors, Dr. Ismail Khater and Dr. Mohammad Hussien, for their significant contribution to my intellectual and personal development. I greatly value your skills, insights, and encouragement, as they have successfully steered me through numerous hurdles. Lastly, I would like to express my gratitude to my friends for their unwavering support, dependable guidance, and motivation throughout this extensive undertaking. Your support has played a crucial role in maintaining my concentration and drive.

This thesis is the result of the combined assistance from all of you, and I am appreciative to each of you for playing a substantial part in my academic path. I offer this effort to you with sincere gratitude and profound admiration.

Acknowledgments

I wish to convey my profound appreciation to my thesis advisor, Dr. Ismail Khater, whose guidance and mentorship have been indispensable throughout this endeavor. Your astute observations, perseverance, and support have facilitated my navigation through the intricacies of my research. I express my gratitude to the instructors at Arab American University for their provision of crucial resources and support. I particularly value the sense of camaraderie among my colleagues in the Data Science department, whose teamwork and friendship enhanced the enjoyment of this process. I express my gratitude to my family and friends for their tolerance, understanding, and encouragement. I would like to express my gratitude to all individuals who have made valuable contributions to my thesis. Your assistance and encouragement have had a significant impact.

Abstract

There is a significant change that facing the recruitment of teachers for the Ministry of Education in Palestine, there is a limited number of vacancies and a higher number of applicants, however, the implants between the availability of opportunities and the supply of qualified candidates may create hard selection how is the suitable candidates, and leads to inefficiencies in the hiring process so addressing the issue is cru\tical for ensuring that the right person in the right, place.

Moreover, the Palestinian Ministry of Education offers a few positions for teachers every year, and many of the candidates who try to fill the empty positions in the research applications were more than 50000 applications for just almost 1000 positions Our study offers using ML to facilitate the selection criteria and choose the most appropriate candidate[1].

In this study we use ML models and compare imbalance and balanced datasets by training the models on the candidate variables such as academic performance, teaching experience gender, demographic factors, and more, as well as classification of the participants into three categories absent, not place, and place, we processed the datasets using support vector machines (SVM), decision tree (DT), XGBoost, Random forest (RF), gradient Boost (GB), Adaboost, Naive Bayes, logistic regression, KNN boost, and Multilayer Perceptron Neural Networks (MLPNNs)[2]The performance of each model was evaluated, and the random forced model was selected because it achieved the highest accuracy, 99%, for both imbalance and balance datasets.

Keywords: Education, Recruitment Module, Recommendation System, Classification Machine Learning Models, Synthetic Minority Over-sampling Technique (SMOTE), Random Forest, Advanced Algorithms, Decision-Making, Ministry of Education.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

Abbreviation	Explanation
AI	Artificial Intelligence
ML	Machine Learning
SVM	Support Vector Machines
RF	Random Forest
GB	Gradient Boosting
ROC	Receiver Operating Characteristic
AUC	Area Under Receiver
TP	True Positive
TN	True Negative
FP	False-Positive
FN	False-Negative
KNN	K-Nearest Neighbor
LR	Logistic Regression
NB	Naïve Bayes
NN	Neural Network
DL	Deep Learning
RMSE	Root Mean Square Error
DT	Decision Tree
EDA	Exploratory Data Analysis
SMOTE	Synthetic Minority Over-sampling Technique
IG	Information Gain

TPR	True Positive Rate
FPR	False Positive Rate

Chapter One

Introduction

1.1. Introduction

The Palestine Ministry of Education offers limited vacancies for hearing now teachers knowing that there is an urgent need for teachers in the state the limitations for those vacancies are due to defecates in the financial situation that the state through because of the occupation and poor economic conditions [3].

Data mining is the process of identifying patterns in various types of data and resolving problems from a large amount of data. Machine learning (ML) studies and creates algorithms that allow computers to learn from data [2]. Making predictions based on data accomplishes this by developing models and highlighting interrelationships by learning from historical relationships of the data. Machine learning uses a set of input variables (x) to forecast an output variable (y). [4]

Machine learning (ML) can quantify the relationship between input and output variables. One of the machine learning tasks used to predict a target variable in never-before-seen data is classification. The foundation of classification is the examination of an unclassified object and the placement of it in one of many predetermined classes [5]. According to how each database input has been assigned to one of the predefined categories, the database inputs that represent the items to be classified are sorted. By employing this technique, models can be produced that can classify newly discovered data that had not previously been classified. [6].

Many studies examine managing human resources (HR) and hiring new employees. These research works primarily concentrate on specific topics in the private sector, although there is very little research literature in the public sector. Modern tools like chat rooms, data aggregators from outside sources, and social media sites like LinkedIn, Facebook, and Twitter are used in

private sector processes. Evaluation and hiring procedures in the public sector should be based on unquestionable standards that apply to all candidates [6], [7].

Individual correct and strategic employment is crucial in strategic human resource management. Machine learning has been applied in employee analytics to predict a candidate's suitability for a particular job through classification techniques. HR managers should emphasize the use of appropriate machine learning techniques to highlight the significant benefits of integrating AI into the hiring process. A talent acquisition framework was created using a model based on performance indicators that monitor and assess talent development over time. This model considered a candidate's profile and the factors that led a candidate to decline a job offer, including salary, location, a particular job position, or an improved offer from a competitor. It was possible to predict whether a chosen candidate will accept or reject a job offer as well as the reason for a chosen candidate's rejection using machine learning techniques for classification, such as decision trees, support vector machines (SVM), and Nave Bayes, in conjunction with knowledge of the candidate's qualifications, wages, and performance indicators score [8]

The Proper selection and positioning of employees is an important issue for strategic human resources management. Within the framework, the research aimed to investigate the most efficient machine learning techniques to support employees' recruitment and positioning evaluation and conducted based on classification algorithms concerning employees of the public sector, seeking to predict the best fit in workplaces and allocation of employees. Based on the outcome of the administered tests, an algorithm model was built to assist the decision support system of employee recruitment and assessment. The present research could lead to the argument that adopting the Employees' Evaluation for Recruitment and Promotion Algorithm Model

(EERPAM) will significantly improve the objectivity of employees' recruitment and positioning procedures [9].

Artificial intelligence is one of the most innovative technologies widely used in organizations of all types to help them make decisions. The human resources (HR) department is one of the most important departments in any organization. This is because they work on quality indicators and employee skills, and employee quality raises the organization and creates an advantage in competitiveness. Employee selection is one of HR's complex tasks that requires a decision system to help them process the task quickly. Establishing a solid and stringent selection process assists the department and organization in general in reaching its goals, saving time, and redirecting effort to more important things.[10]

Our research proposes the utilization of Machine Learning (ML) and Deep Learning (DL) techniques to create a recommendation system for identifying the most suitable teachers at the Ministry of Education. By engaging algorithms including Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Gradient Boost (GB), AdaBoost, Multi-layer Perceptron Neural Networks (MLPNNs), Extra Tree Classifier (ETC), and XGBoost, the system will be able to categorize and recommend teachers based on a range of criteria such as qualifications, experience, teaching style, and student performance outcomes. On the other hand, in the absence of similar ML and DL-based recommendation systems within the Ministry, this initiative could significantly enhance the Ministry's ability to match teachers to the most suitable positions, thereby improving educational outcomes across Palestine. Implementing such a system would address a critical need within the Palestinian education system, ensuring more informed decision-making and better resource allocation.

1.2. Objectives

Applying machine learning (ML) and deep learning (DL) techniques to develop a recommendation system for selecting the most appropriate teachers at the Ministry of Education has shown great promise in achieving highly accurate results. The primary objective of this research is to propose ML and DL models that can effectively categorize and recommend teachers based on various factors such as qualifications, experience, teaching style, and student performance outcomes. Using a locally collected dataset from the education sector, these models can distinguish the suitability of candidates for specific teaching positions, which can assist education administrators in making more informed decisions, ultimately leading to improved educational outcomes across Palestine.

1.2.1. Specific Objectives

- Develop a recommendation system based on the applicants' data to select the right teacher in the Ministry of Education.
- Design a model that is based on machine and deep learning approaches.
- Compare the performance of the various machines and deep learning models leveraged for the teacher recommender system.
- Extract and design features for the machine learning approaches. Then, feature selection and dimensionality reduction techniques will be applied to avoid overfitting.
- Tune the various model parameters and evaluate the performance of the adopted models.

1.3. Contributions

The ML model that we tested and trained reflects a unique technique for candidate selection by developing a recommendation system specific to the Ministry of Education requirements. this new

approach is not similar to the traditional method of candidate selection it involves deep machine learning techniques that improve the accuracy and efficiency of teacher selection. However, also our research may have contributed to the educational data since the theoretical framework applies ML data analysis to current life problems by advancing the theory behind recommendation systems within the educational sector. Also, that provides a model that can be refined and expanded with additional data and increases the application of more ideas in the educational area.

On the other hand, the system may offer the Ministry of Education a tool that can lead to more information and transparent decision-making in teacher's recruitment.

Chapter Two

Background and Literature Review

2.1. Background

In the face of a rapidly evolving educational landscape, it is essential to prioritize recruiting highly educated and suitable teachers. The proficiency of educators directly impacts students' academic accomplishments and the general standard of educational excellence. The traditional methods of choosing teachers primarily rely on manual processes, which are susceptible to being time-consuming, biased, and less adaptable when assessing large volumes of data.

Given the pressing circumstances in Palestine, there is an urgent need for advanced technologies to improve the teacher selection process. Machine learning (ML) and deep learning (DL) are groundbreaking technologies that can analyze large datasets, identify patterns, and make precise predictions. These technologies have been extensively utilized in various domains, including human resources (HR) and talent acquisition in the commercial sector. However, their use in public sector human resources, particularly in education, has been limited. Data mining and machine learning empower companies to leverage vast data to identify significant relationships and make informed decisions. Machine learning utilizes algorithms that can extract information from historical data to make predictions and classify new data.[7].

Classification techniques are highly beneficial in HR procedures as they enable the methodical categorization of candidates based on many criteria, including abilities, experience, qualifications, and other relevant traits. This thesis explores the potential of machine learning and deep learning to create a comprehensive recommendation system for choosing appropriate professors inside the Ministry of Education. The goal is to develop a model to analyze teacher-related data to predict their suitability for specific positions. The study also investigates the challenges faced in the public sector, such as the need for fair and uniform evaluation methods and the relatively limited use of modern recruitment tools compared to the private sector. The importance of this study lies in its

ability to improve the effectiveness, accuracy, and fairness of teacher selection. The proposed recommendation system employs machine learning classification methods, including decision trees, support vector machines (SVM), and Naive Bayes, to aid HR departments in making educated decisions. Furthermore, the model considers other factors that influence candidates' decisions, such as their credentials, areas of expertise, test results, and other work-related conditions[2]. This guarantees a comprehensive method for the selection of educators. This thesis contributes to the growing body of research on the application of artificial intelligence in human resources and talent acquisition, with a specific focus on the field of education. An effective recommendation system for teacher selection can lead to enhanced recruiting outcomes, higher educational quality, and improved organizational efficiency [11].

Intensive research in electronic resources was conducted using different databases, including PUBMED, EBSCO, HINIRI, ScienceDirect, and Google Scholar. The researcher found a few similar studies conducted in Europe, the United States of America, and regional countries

2.2. Literature Review

A knowledge-driven decision support system (KDDSS) aids decision-making by incorporating techniques from artificial intelligence. It employs rules that enable inductive conclusion drawing based on previously stored facts. These systems can employ ad hoc procedures, which correspond to the system's behavior and the range of an expert.[13] . Recently, finding qualified candidates for open positions has been growing in popularity as of late. These programs employ complex algorithms to pair job seekers with employers who can make use of their unique set of skills, work history, and educational background. The need for efficient employment recommendation systems has grown alongside the number of online job boards and other recruitment hubs. On the other

hand, machine learning has become an extremely useful tool for data analysis and prediction in many fields, including HR recommendation systems. These systems can provide graduates with highly accurate and personalized job recommendations by employing machine learning techniques such as natural language processing, clustering, and classification. Recently, ML has been used in several studies to create a recommendation system for aiding hiring managers in making the best possible decisions. Using graduate-level academic data, previous research has examined how different ML techniques and input features affect students' academic data and student status. [12].

For instance, [14] E-recruitment recommender systems have emerged as a workable solution to the problem of professional allocation, reducing the time and expense of recruiting while increasing the volume of data that human resource management specialists must manage.

To illustrate, e-recruitment recommender systems have emerged as an effective solution to the challenges of professional allocation study used data from Kaggle, including gender, board of education, secondary school, and high school percentages. Using a Random Forest improves classification performance and yields a higher accuracy of 85% when compared to a Decision Tree. Consequently, Random Forest is the suggested method for their classification/prediction issue identifying the appropriate location/job using graduate features. [15] However, a hybrid environment using ML algorithms and statistical computations to address the issue of universities finding new master students. Studies showed that neither the offered salary nor the student's gender or entry test score had any impact on where they were placed. In contrast, placement status is related to MBA specialization and degree stream. The authors also found that the SVM model performed better than all other models in terms of 90% accuracy and F1-Score in predicting MBA students' placement status. With accuracy ratings of 88%, 82%, and 80%, respectively, RF, XGB, and GB are listed next. The study also discovered that factors such as work experience, MBA

percentage, and SSC percentage all affect placement status. In this study, the authors-built models that could predict high school placement test scores using a comprehensive dataset they were able to obtain from the High School Transition System in Turkey. [16]. In another study, researchers ran a sensitivity analysis on these models to determine the most important predictors. The results showed that the C5 decision tree algorithm, which had a 95% accuracy rate on hold-out samples, was the most reliable predictor, followed by support vector machines and artificial neural networks. With an overall accuracy of 82%, logistic regression models, on the other hand, were the least accurate. The sensitivity analysis revealed several predictors that significantly impacted placement test scores, including prior test experience, scholarship status, the number of siblings, and grade point average from the previous year. [17]. Additionally, various studies conducted by authors from South Asian University, based on an examination of historical placement data from 2014 to 2016, investigated the relationship between a dependent variable (placement) and multiple independent variables using linear regression, They discovered that a student's academic history, particularly 10th-grade score, followed by Current GPA University, plays a significant role in determining a student's chances of getting a job.[18] . Moreover, in the way to predicts whether a student will be placed in a Core IT business or a Consultancy business based on his academic performance, many algorithms were applied to students' data, including clustering with an accuracy of 95.52% followed by Classification DT with an accuracy of 82.39, and NB with an accuracy of 62.3%.To solve the student placement prediction problem, the paper recommends applying LR, KNN, DTR, XGBoost regression, GBR, light GBM regression models, and RFC. This task is completed in two stages. Phase 1 is run on a simple data set, and Phase 2 is run on a larger data set with more student features. The K-neighbor regression model outperforms other models in dataset-I, with a higher R2_score (0.94) and a lower RMSE value (103263.25).

However, for dataset II with a larger number of features, the linear regression model and the K-neighbor regression model have the highest R2_score (0.89). On the other hand, the RMSE value of the XGBoost regression model is very low for dataset II (12970.41) [19], [20].

In addition to these approaches, Python has been widely employed for data analysis and prediction in education the Role of Academic and Employability Skills is another study that used the same data sets as our research as a benchmark paper. Among the supervised classification algorithms evaluated were RF, GBLR, SVC, DTG, NB, KKN, and Linear Discriminant Analysis. Accuracy, F1 Score, G-Mean, and ROC AUC Score were used to evaluate performance and decide whether to hire a student. The most accurate algorithm tested was found to be Linear Discriminant Analysis Classification. The Linear Discriminant Analysis algorithm was fine-tuned using hyperparameter tuning to improve its performance. Linear Discriminant Analysis was improved by the addition of the parameter solver 'lsqr' and shrinkage 0.81. The results of Linear Discriminant Analysis with hyperparameter tuning are 86% more accurate [21].regarding the predict university recruitment results, the authors use supervised machine learning classifiers such as the MLP, SGC, DTC, Multinomial Naive Bayes Classifier, KNN, and Ensemble-based classifiers such as the RFC, Adaboost Classifier, Extra Trees Classifier, and Gradient Boost Classifier. These models take as input historical data from an organization's graduates, and the variables include the percentage of previous test results, major, and job experience. The Gradient Boosting Classifier produced the most accurate results (81.4%) of the classifiers tested.[22]. Machine learning algorithms can raise the caliber of teachers hired while lowering the likelihood of unfavorable results. The study used machine learning (ML) techniques to analyze the work histories of candidates for teaching positions and found that candidates with more relevant experience make better teachers and stay with the school district for longer periods of time. The ML model also decreased the likelihood of

discrimination against minority applicants during the hiring process by reducing recruiter bias. A country's economy, welfare, and human capital can all be significantly impacted by enhancing the selection of teachers at a low cost, claims the study. The researchers claim that these results can be applied to any industry and that using ML techniques to screen job applications can increase the accuracy of employee selection, accelerate the procedure, and reduce costs. [23]. Another Research seeks, as its primary objective, to answer the question of whether a job posting found online is fraudulent. If these phony advertisements for job openings can be singled out and eliminated, it will be much simpler for people looking for work to concentrate only on those postings that are for genuine employment opportunities. A dataset that was obtained from Kaggle is being utilized in the investigation. The process of identifying which job postings contain fraudulent information involves the application of several distinct classifiers. These classifiers include the Naive Bayes Classifier, the DTC, the MLPC, the KNN, the AdaBoost Classifier, the GBC, and the RF. The use of multiple classifiers to identify fraudulent employment opportunities is illustrated with the help of a supervised mechanism serving as an example. The results of the tests indicate that the Random Forest classifier is superior to its contemporaries in terms of its ability to perform classification. The methodology that was suggested was successful in meeting the accuracy goal that was set.[24]. Predictive educational data mining process is carried out using an organization's final-year student information to determine where students will be placed on campuses. Weka and R studio are used as the environments for the execution of machine learning algorithms. The results of applying the algorithms are tabulated and analyzed, which shows that the random tree algorithm gives a prediction accuracy of 100% both on the dataset and in the R environment. Partitioning in a Recursive Way and Regression Tree is more effective and has an

accuracy rate of 90%. The term "placement prediction system" refers to a system that makes projections about the levels of placements available to graduates of the final year.[25].

On the other hand, for the same purpose Python was used for data analysis and to make predictions, which included the implementation of a variety of machine learning algorithms. The accuracy of several different algorithms is analyzed by the authors, and the results are tabulated above. It should come as no surprise that SVM provides an accuracy of 95%. LR was also a superior technique that, depending on the dataset, offers an accuracy rate of up to 90%. It's possible that different datasets will result in different levels of precision for machine learning algorithms. Because all these methods have a precision that is higher than 87%, it is clear from looking at the results that KNN, RF, LR, and SVM, in the order that they are mentioned, are effective solutions for problems that involve binary classification[26]. Which also supported by incorporated twenty-seven thousand elements of information, which included three thousand observations and twelve features of students' mock job interview evaluation results, on-the-job training student's performance rating, and general point average (GPA) of students registered in the on-the-job training course of the school year 2015 to the school year 2018. The time covered by this study was from School Year 2015 to 2018. The authors used different types of analysis like DT and RF as well as SVM, KNN, and LR to understand how students find jobs. The performance matrix was used to evaluate the six algorithms about their accuracy, precision, and recall, as well as their f1-score and support measures. During the tests, the SVM achieved an accuracy level of 91.22%, which was significantly higher than that achieved by any of the learning algorithms, including DT (85%) and RF (84%). The primary goal of the study was to find the best algorithm to support selection and promotion decisions based on employee performance reviews. The survey responses of Greek public sector workers were used to create the dataset for the study, which also contains

details regarding their training and experience. When using supervised machine learning methods like support vector machines, decision trees, and Bayesian algorithms to predict how human resources should be distributed in the best way, the mean accuracy of classification was 97.03% for J48, 95.04% for Random Forest, 93.16% for Nave Bayes, and 92.41% for SMO.[27].

An article designs and implements a complete prototype for a movie recommendation system using the KNN algorithm, a collaborative filtering algorithm, and recommendation system technology. The system is based on the KNN algorithm. The authors provide a comprehensive description of the design and development process, as well as the results of a professional test that evaluates the stability and high efficiency of the experimental system. (Cui 2020). [29] Critically explore in their article titled “Application of Machine Learning in Recommendation Systems” the essential role of recommender systems in handling the growing volume of information on the Internet. Focusing on filtering algorithms based on neighborhoods of users or objects and on content-based approaches, the study uses datasets from the MovieLens portal to practically implement and analyze various algorithms. The significant importance of recommender systems in addressing the challenges posed by the exponential growth of digital content is highlighted. The article also provides a comprehensive overview of machine learning classifications, including supervised, unsupervised, semi-supervised, and reinforcement learning, showing their adaptability to various situations. As recommender systems gain prominence, the conclusion suggests future directions, specifically exploring hybrid methods to improve prediction accuracy and mitigate errors in different systems.

Systematic Review for Machine Learning-Based Recommendation Systems for E-learning" produced by Khanal et al. systematically investigates recommender systems in the context of online education, addressing challenges such as data scarcity, startup problems in cold, scalability,

time consumption, and precision. The authors classify recommender systems into Content-Based, Collaborative Filtering, Knowledge-Based, and Hybrid Systems, presenting a taxonomy that includes essential components for the development of the system. The study reviews 101 articles published in recent years, highlighting the prevalence of Collaborative Filtering and underlining the competitive advantage of hybrid techniques. Evaluation metrics such as Mean, Precision, Recall, and F-measure are identified, with a call for future research to explore ranking metrics and user satisfaction. The article concludes by pointing out unresolved issues such as scalability, latency, privacy, and emerging challenges such as tampering attacks while suggesting that hybrid techniques show promise in addressing common problems in today's systems [30]. A Comprehensive Review of AI Deployment," conducted by Rusudan Tsiskaridze et al., analyzes the integration of artificial intelligence (AI) in human resource management (HRM), specifically focusing on AI applications in the selection process. Through a comprehensive analysis of 46 peer-reviewed journal articles spanning the years 2019 to 2023, the study categorizes its findings into five main areas: AI-driven HRM, ethical considerations of AI in HRM, advantages and drawbacks of AI-enabled recruitment tools and the application of AI in recruitment in various country contexts. Importantly, it highlights the potential for collaboration between AI and human workers in selection processes, while also addressing issues such as data privacy and transparency. The research identifies a notable lack of empirical studies in the European region, possibly influenced by GDPR regulations affecting the adoption of technology-based recruitment tools. Despite AI's ability to streamline screening and mitigate bias, ethical concerns remain. The article concludes by encouraging HR managers to view AI as a tool to improve productivity rather than as a substitute. It highlights continued advances in AI technology within HRM and suggests future research

directions, emphasizing key findings related to the collaborative, ethical, and regional aspects of AI implementation in personnel selection in HRM[31].

The benefits of ML extend beyond recruitment to other domains, such as education and HR analytics. For example, “A Machine Learning Approach for Automation of Resume Recommendation System” by Roy presents a machine learning-driven system to automate the resume recommendation process. It addresses the complex challenges faced by Human Resources (HR) in talent acquisition, especially in the context of India's large and dynamic labor market. The proposed model uses linear SVM as the main classifier, achieving an accuracy of 78.53%, outperforming alternatives such as Random Forest and Logistic Regression. The methodology covers pre-processing, classification and recommendation stages, with the aim of streamlining the laborious and manual resume selection process. The article suggests future improvements, including the integration of deep learning models and industry-specific customization, to further refine the system [32]. Also, the research article, written by Daryani et al., "Automation of Resume Selection "Automation of Resume Selection through Natural Language Processing and Similarity," presents an innovative solution to the difficulties that companies face when manually selecting many resumes. The proposed system leverages the power of Natural Language Processing (NLP) to extract relevant information from unstructured resumes, creating concise summaries to streamline the selection process. Using a vectorization and cosine similarity model, the system compares each resume to a provided job description, generating ranking scores to evaluate candidate suitability. The methodology incorporates various information extraction techniques, such as tokenization, stemming, lemmatization, part-of-speech (POS) tagging, chunking, and named entity recognition. Testing the system with an Amazon job description for a Software Developer Engineer position demonstrates successful information extraction, resulting in a ranked

list of candidates based on cosine similarity. The article concludes by highlighting the effectiveness of the system in overcoming challenges associated with manual shortlisting and improving the overall efficiency of the recruitment process[33] .

In their study "Predictive Analytics of HR - A Machine Learning Approach," Kakulapati, Vijayalakshmi, explore the field of human resources (HR) analytics, investigating its evolution from manual processes to predictive analytics using machine learning techniques. Focusing on improving HR functionality and results, the authors use random forest classification to analyze employee information, especially monthly income, with the goal of improving decision-making in HR processes. Furthermore, clustering techniques based on the similarity of performance metrics are used to evaluate employee performance. The introduction highlights the importance of HR management in addressing training needs, talent mapping, and organizational objectives. The article highlights the crucial role of HR analytics in improving employee productivity and fostering organizational growth. The transformative impact generated by predictive analytics in HR is discussed, driven by advances in computing, the online availability of HR data, and global competition for talent. In conclusion, the authors outline the application of unsupervised learning algorithms, specifically K-Means clustering, to automate HR decision-making related to employee placement and promotions based on identified clusters and attributes such as monthly income. The research highlights the potential of predictive analytics in HR to optimize business operations, improve employee satisfaction and facilitate accurate decision-making [34].

To investigate the current state of human resources in small and medium-sized enterprises, and identify the factors that hinder their development, the article by Zhu, Hong, titled "Research on Human Resources Recommendation Algorithm Based on Machine Learning," aims to We aim to address these obstacles by proposing the use of machine learning (ML) techniques for human

resources data management and analysis. Specifically, the article designs and implements a salary forecasting model using a gradient descent algorithm and a backpropagation neural network, with the intention of improving forecasting accuracy. The experiments carried out in the research demonstrate the effectiveness of the proposed model, achieving a training accuracy of 89.98% and a validation accuracy of 84.05%. The study highlights the importance of ML in improving human resource management, especially in tasks such as predicting employee turnover and retirement, sorting resumes, and streamlining human resources responsibilities. The article concludes by emphasizing the potential of ML to address the challenges arising from data growth in human resource management systems [35].

In the "Application of Machine Learning Algorithms to an Online Recruitment System," presents a novel approach to evaluating job applicants in online recruitment systems using machine learning algorithms. The proposed system implemented as a prototype, extracts objective criteria from applicants' LinkedIn profiles and infers their personality characteristics through linguistic analysis of their blog posts. The functionality of the system is demonstrated and evaluated in a real-world recruiting scenario. The study highlights the potential for automation in applicant classification and personality mining, to improve the efficiency of the hiring process. The system is designed to easily integrate into companies' Human Resources Management infrastructure, to assist recruiters in making decisions without replacing them. The conclusion of the article emphasizes the effectiveness of the proposed approach in accurately identifying the personality traits of applicants and classifying them appropriately [36]. As in the previous study Mr. Wang and colleagues present a recommendation system that integrates a gradient-boosting tree model and a convolutional network, emphasizing the importance of feature regularization. The study aims to improve personalized HR recommendations by exploring correlations between job seekers and job offers.

To address challenges such as gradient disappearance and feature loss, researchers optimize activation functions and clustering strategies, thereby improving the quality of recommendations. The evaluation results show a higher recall rate and F1 score compared to other algorithms, evidencing the effectiveness of the proposed approach in the field of HR recommendation [37] . also, the application of machine learning (ML) to improve productivity and decision-making in human resource selection is explored[38]. Differing from traditional economic models focused on production functions, the research highlights the importance of individual workers and variability in labor productivity, especially with heterogeneous labor inputs. The primary goal is to leverage ML to predict and select the most productive labor inputs, demonstrated in practical applications such as police hiring decisions and teacher tenure. The study highlights how ML, with its predictive emphasis, outperforms traditional tools in social policy applications, generating substantial gains in social well-being. The findings highlight the potential of ML to refine worker selection processes and its implications for decision making in various settings, encouraging future research in this evolving field.

To improve faculty selection processes and overcome the limitations of conventional recruitment methods, Ovallea, present an innovative approach in their article titled "Application of neural networks in the teacher selection process" The research aims to explore and address the limitations associated with traditional practices. The proposed model integrates neural networks, incorporating three criteria and 23 characteristics. Individual networks manage each criterion, and a final classification network synthesizes the information. Implemented using Matlab, the prototype effectively demonstrates the viability of the model. The evaluation with 30 candidates reveals a high sensitivity of 93% and a specificity of 86%, highlighting the effectiveness of the model in predicting the suitability of teachers. The article concludes by highlighting the adaptable

nature of the model, underlining its potential to be replicated in various educational institutions, demonstrating the ability of neural networks to revolutionize teacher selection processes and contribute to more informed hiring decisions [39].

In his article "Quality of Hire: Expanding Multi-Level Fit Employee Selection through Machine Learning," Shet presents an innovative concept of "Quality of Hire" in employee selection, fusing the theory of person-environment (P-E) fit. With machine learning techniques. Using the partition around medoid (PAM) clustering algorithm, the authors analyze a selected IBM data set, focusing on various dimensions of P-E fits. The proposal advocates a multilevel fit model rooted in P-E theory to improve hiring quality. By addressing gaps in the understanding of selection decisions and leveraging HR analytics with machine learning, the research seeks to contribute to organizational psychology, human resource management, and data science. The discussion highlights the importance of machine learning algorithms in human resources, presenting a novel application of PAM in this context. The proposed methodology offers valuable insights into the multilevel mechanisms of employee adjustment, potentially elevating the quality of hire in human resource management practices. The implications highlight the need to foster a data-driven culture and promote collaboration between HR specialists and machine learning professionals for successful implementation. Overall, the article highlights the relevance of adopting data-driven decision-making in the modern HR landscape [40].

Furthering the application of machine learning in recruitment, research has explored neural network-based models to improve the teacher selection process. An article titled "Applied Machine Learning Techniques on Selection and Positioning of Human Resources in the Public Sector" written by Pampouktsi, Panagiota. addresses the crucial issue of employee selection and placement in the context of management human resources strategy. Recognizing the importance of accurate

decision-making for human capital management, the research explores machine learning techniques to improve employee recruitment and assignment assessment in the public sector. The study conducts tests using classification algorithms on public sector employees, to predict optimal fit for workplaces and efficient employee allocation. The findings lead to the development of the Employee Appraisal Algorithm for Recruitment and Promotion Model (EERPAM), showing possible improvements in the objectivity of recruitment and positioning procedures. The article contributes to the evolving field of strategic human resource management by incorporating machine learning algorithms for effective decision support in HR processes. The research highlights the potential of technology, specifically artificial intelligence, to optimize the selection, placement, and promotion of personnel, ultimately supporting long-term organizational success. [41].

In their "Smart System Model for the Recruitment of Teachers" study, Auccahuasi and his colleagues employ a methodological approach centered on a multilayer neural network with 23 input criteria. The input layer incorporates numerical values that evaluate the candidate's performance in writing exercises, group interviews, and demonstration classes. An intermediate layer consisting of four neural network nodes leads to an output layer that produces a binary classification (0 or 1), indicating the candidate's suitability for the institution. In addition, the numerical output values also reflect the probability of suitability. The results highlight the effectiveness of the model in handling various criteria, showing its ability to adapt to evolving evaluation standards. The consistent performance of the neural network model in predicting teacher suitability underlines its potential for establishing reliable and efficient recruitment processes in educational institutions[42] .Also The report, authored by Brian A. Jacob, examines the teacher selection process in Washington, D.C. public schools, focusing on a centralized

application system established in 2009. The study reveals that measures such as subject-specific written assessments, Structured interviews, teaching auditions, and traditional indicators such as undergraduate grade point average significantly predict teacher performance in the district evaluation system. Notably, candidates in the top quartile outperform those in the bottom quartile by 0.6 standard deviations, equivalent to the improvement seen in the first three years of an average teacher. While highlighting the potential to improve teacher quality through improved selection, the report emphasizes the need for continuous improvement, noting that a sophisticated application system alone is insufficient without careful attention to how principals use the information generated. The study also highlights the relatively low cost of investing in teacher selection processes compared to the long-term benefits of hiring more effective teachers [43].

Another study uses database searches, including ERIC, BEI and Google Scholar, with a specific focus on literature translated into English published in the last decade. The study highlights the importance of selecting candidates for Initial Teacher Preparation (ITP) based on identified characteristics such as literacy, numeracy, interpersonal skills, willingness to learn, and motivation to teach. The selection process, which includes entry requirements, exams, interviews and portfolios, varies globally. Prominent examples from Singapore and Finland underline the importance placed on academic performance, communication skills and motivation in the selection process. The report concludes by recommending authentic measures beyond testing to assess teaching competence, such as work samples, portfolios, and ongoing assessment of student progress [44].

Another study led by Bin Yu and his team[45], developed a regression model to identify factors that influence annual school-level teacher turnover rates in North Carolina public schools. Using data from the Belk Endowment Educational Attainment Data Repository, North Carolina IRS tax

data, and county-level crime data, we employed an EXtreme Gradient Boosting (XGBoost) regression model to predict teacher turnover rates. Of the 252 characteristics analyzed, school-level factors took center stage, with the most influential being the percentage of teachers with 0-3 years of experience, short-term suspensions per 100 students, percentage of teachers achieving standard 1 competency, percentage of economically disadvantaged students and percentage of licensed teachers. These findings offer practical insights for educational administrators, suggesting ways to improve teacher retention by focusing on factors within their control, such as recruiting experienced teachers and addressing the needs of economically disadvantaged students. Although we recognize certain limitations, the study proposes future efforts to delve into more detailed data analysis at the teacher and student levels.

Personality prediction has also become an integral part of recruitment, as evidenced by ISudha Ganesh's groundbreaking study titled "Personality Prediction through CV Analysis using Machine Learning Algorithms for Automated E-Recruitment Process", we witnessed a significant advancement in e-recruitment practices. This study goes beyond conventional assessments by integrating personality assessment, using powerful machine learning algorithms such as Logistic Regression and Random Forest Classifier. Through online multiple-choice question (MCQ) tests, the research successfully predicts candidates' personalities, demonstrating the system's remarkable accuracy in predicting personality traits. The implications of this result are profound, offering valuable insights that can revolutionize HR decision-making in the recruiting arena. This points a promising path toward improving candidate screening efficiency, easing HR workloads and ultimately raising overall productivity in the recruiting space [46].

Charlene Grace's study, titled "An AI-based Model for Achieving High-Reliability Faculty Performance Using Various Machine Learning Algorithms," seeks to revolutionize teacher

performance assessment by leveraging machine learning. Focusing on factors such as seniority, tenure, academic rank, workload, and demographics, the research uses algorithms such as Support Vector Machine, Decision Tree - Fine Tree, and Ensemble-Bagged Tree through MATLAB R2021B. The Ensemble Bagged Tree algorithm, exhibiting superior precision, is presented as a powerful tool to predict and improve teacher performance. The scope of the study ranges from addressing technical requirements to employing data preprocessing techniques, highlighting the significant impact of demographic characteristics on faculty performance. In conclusion, the research provides an innovative predictive model, underlining its potential to elevate decision-making processes and improve the quality of teaching in educational institutions [47].

Deep Learning in Recruitment and Human Resource Management

Expanding on the role of machine learning, several studies have investigated deep learning approaches for improving job-resume matching and candidate-job fit predictions. In this study, the authors-built models that could predict high school placement test scores using a comprehensive dataset they were able to obtain from the High School Transition System in Turkey. The researchers ran a sensitivity analysis on these models to determine the most important predictors. The results showed that the C5 decision tree algorithm, which had a 95% accuracy rate on hold-out samples, was the most reliable predictor, followed by support vector machines and artificial neural networks. With an overall accuracy of 82%, logistic regression models, on the other hand, were the least accurate. The sensitivity analysis revealed several predictors that had a significant impact on placement test scores, including prior test experience, scholarship status, the number of siblings, and grade point average from the previous year [48]. authors of proposed the Topic-based Ability-aware Person-Job Fit Neural Network (TAPJFNN) in, a neural network model for reducing reliance on manual labor and improving fitting results interpretation in the process of matching

job seekers with suitable job openings. The TAPJFNN model, which is based on recurrent neural networks, generates semantic representations for job requirements and job seekers' experiences using rich information from historical job application data such as gender, specialization, and secondary and high school results. Additionally, there are hierarchical attention strategies for determining the importance of job requirements and job experience contributions to specific ability requirements, as well as a refinement strategy for predicting person-job fit based on historical recruitment records. The TAPJFNN model is used in two human-centered applications: talent sourcing and job recommendation. When tested on a large-scale real-world dataset, it outperforms several baselines in terms of effectiveness and interpretability. Academic and Employability Factors in Campus Recruitment [17]. However, paper describes a multi-view co-teaching network for improving the performance of job-resume matching algorithms, which are used to match job seekers with suitable job openings on online job boards. For the matching algorithm, the network considers two perspectives: text-based and relation-based models, which are integrated via representation enhancement and data enhancement strategies. On a real-world job-resume interaction dataset, the network outperforms several competitive baselines in terms of matching accuracy.

Summary of literature

The literature can present that machine learning has become a good tool in recent human resources management, particularly in the recruitment process by analyzing large databases. Machine learning algorithms can identify a pattern and make informed predictions about job susceptibility, improving efficiency and accuracy.

The literature shows that machine learning has significant promise in the job recruitment system by utilizing techniques like random forest. The systems can efficiently match job seekers with

suitable positions. Such studies have demonstrated that random forests outperformance decision trees in job classification while the SMV model has an accuracy rate of 90% and predicting student replacement outcomes.

So, by considering factors such as education, gender, and discord, the algorithms. Such as XG post and Random Forest in addition to the decision tree. Have been instrumental in streamlining the recruitment process and enhancing placement accuracy. The deep learning approach. Have advances in the field of job resume matching. Models such as Topic Based Ability, Aware Person, Job Fit, Natural Network, and Multiview Core Teaching Network leverage semantic representation and historical data to improve mismatching. Visions. However. The various machine learning models including random forest, SVM, and deep learning techniques have consistently demonstrated their potential to optimize and automate the recruitment process. This makes machine learning an essential tool for modern. Human resource practice.

Table 2. 1 Summary of literature.

Authors	Year	Title	Focus	Model/Algorithm	Results	Conclusion
Ovalle et al.	2023	Application of neural networks in the teacher selection process	Teacher selection	Neural networks with three criteria and 23 characteristics	Sensitivity: 93%, Specificity: 86%	Adaptable and effective model with replication potential in educational institutions
Shet and Nair	2023	Quality of Hire: Expanding Multi-Level Fit Employee Selection through Machine Learning	Employee selection	PAM clustering algorithm	Improved understanding of selection decisions, promotion of data-driven HR culture	Emphasizes machine learning in HR and collaboration between HR and data professionals
Pampouktsi et al.	2021	Applied Machine Learning Techniques on Selection and Positioning of Human Resources in the Public Sector	HR in public sector	Classification algorithms	Development of EERPAM model	AI and machine learning can optimize HR processes for long-term success
Rojas Romero et al.	2022	Smart System Model for the Recruitment of Teachers	Teacher recruitment	Multilayer neural network with 23 input criteria	Effective adaptation to evolving evaluation standards, consistent performance	Reliable, efficient recruitment processes in educational institutions can be established
Jacob	2016	Teacher Selection Process in Washington, D.C. Public Schools	Teacher selection	Subject-specific assessments, interviews, teaching auditions	Top-quartile candidates outperform bottom-quartile candidates	Continuous improvement in selection processes necessary for long-term benefits
Hobson et al.	2010	Candidate Selection for Initial Teacher Preparation	Teacher preparation	Literature review of global practices	Importance of academic performance, communication skills, motivation	Recommends authentic measures beyond testing to assess teaching competence
Yu, Rezsonya, and Drew	2019	Predicting Teacher Turnover Rates in North Carolina Public Schools	Teacher turnover	XGBoost regression model with multi-source data	Identification of key factors affecting turnover, like experience and economic disadvantage	Provides practical insights for improving teacher retention
Sudha et al.	2021	Personality Prediction through CV Analysis using Machine Learning	E-recruitment	Logistic Regression, Random Forest Classifier	High accuracy in predicting personality traits through online MCQ tests	Potential to revolutionize HR decision-making in recruitment
M. Garces	2024	AI-based Model for Achieving High Reliability Faculty Performance	Teacher performance	SVM, Decision Tree, Ensemble-Bagged Tree	Ensemble Bagged Tree algorithm shows superior precision	Innovative predictive model for improving teaching quality in educational institutions

Chapter Three

Proposed Method

3.1.Proposed Method

This section provides an overview of the technique utilized in this investigation. The expected results entail acquiring significant insights and doing model analysis to classify applicants into two unique categories: hired and not hired. The dataset utilized in this investigation was obtained from the Ministry of Education and was gathered in 2022. The first CSV file contained 25 columns, often known as features. The methodology will be implemented by strictly following three main phases. The first phase entails the analysis and depiction of the collected dataset. The second stage encompasses Feature Engineering and Preprocessing, which is commonly acknowledged as the most challenging process in classification. The procedure entails transforming and creating supplementary attributes using the preexisting features in the dataset. Furthermore, the procedure includes the normalizing of data by variance scaling and the encoding of categorical variables. To tackle the problem of imbalanced data, we proposed utilizing the class weighting technique to obtain a more equitable distribution across all categories. The benefit of equalizing all categories is to eradicate bias and mitigate poor performance across many models[48] .

Ultimately, the dataset will be prepared for modeling by carefully selecting appropriate machine-learning techniques. Some machine learning techniques are supporting vector machine (SVM), Random Forest (RF), Decision Tree (DT), XGBoost, Adaboost, Extra tree classifier, and Gradient Boost (GB). Each of these relevant machine learning models has its distinct parameters that allow for precise learning and promote intelligent decision support or classification[49]. Subsequently, the models will be employed using unknown datasets to attain the most accurate testing results. After training and testing the ML models, the next step is to evaluate their performance using metrics such as accuracy, precision, recall, true positive rate, false positive rate, and F1 score. These measurements can determine the strengths or weaknesses of the models. To confirm the

testing models' generalizability—the capacity of the model to adjust and provide an accurate response to a previously unknown dataset so that it may be applied in real-world scenarios[50].

The proposed model's structure is illustrated in Figure 3.1. Starting with the process of gathering data. Next, Exploratory data analysis is a valuable tool for data scientists to gain deeper insights into datasets and distributions. It allows them to identify patterns in the data, detect missing values, explore connections between characteristics, and uncover anomalies and outliers. Furthermore, scientists employ visualization libraries to acquire a robust comprehension of the gathered data. Subsequently, data scientists employ features engineering and preprocessing techniques to address any challenges that may arise during the exploratory data analysis (EDA) phase and to handle imbalanced data. Subsequently, the dataset will be partitioned into a training dataset including 75% of the data, and a testing dataset comprising 25% of the data. This division is crucial to enable researchers to accurately assess the performance of the models and prevent any inadvertent leakage of data. Subsequently, multiple machine learning models will be utilized to classify the employment status and the total.

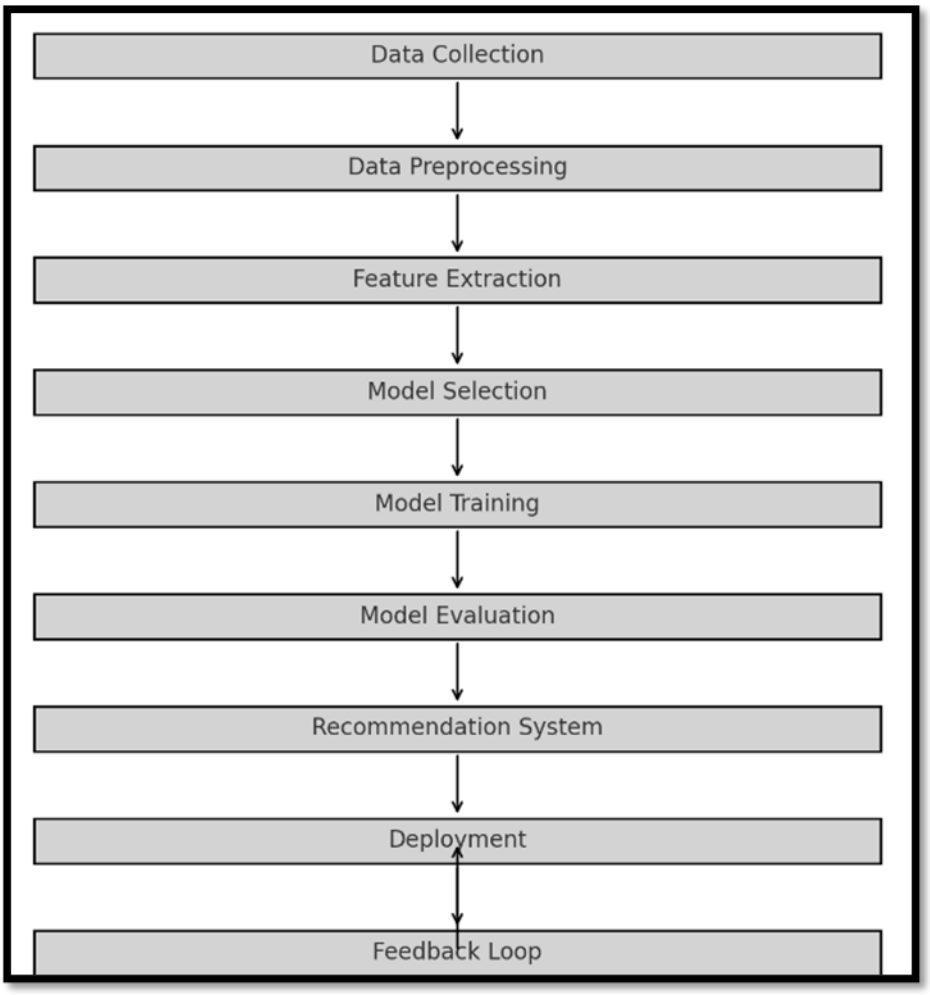


Figure 3. 1 The proposed model for developing a recommendation system to select the right teacher at the Ministry of Education.

3.2. Data Description

The Ministry of Education in 2022 provided this information through the submission of applications via their application portal. The candidate is tasked with completing the form, which includes their personal information, experiences, and academic qualifications. Furthermore, the ministry is accountable for administering interviews and assessments for employment purposes. The proportion of male applicants is 16%, whilst the proportion of female applicants is 84%.

Approximately 50,000 recordings have been estimated, each including 25 features. The ages of the applicants varied from the low 20s to the high 50s.

Description for each column in the dataset:

1) Year: This column represents the year in which the data was collected. In this dataset, all records are from the year 2022.
2) Administrative Unit: This column indicates the administrative unit to which the applicant belongs. Administrative units could represent different regions, districts, or departments within the Ministry of Education.
3) Gender: This column denotes the gender of the applicant, indicating whether they are male or female.
4) Exam Material: This column specifies the subject or field in which the applicant took the exam. It could include various subjects related to education and teaching.
5) Record Type: This column describes the type of record associated with the applicant. It could include information about the applicant's academic history, employment status, or other relevant details.
6) Order: This column represents the order or sequence number of the applicant. It may indicate the chronological order in which applications were received or processed.
7) Employment Status: This column indicates the employment status of the applicant, such as whether they are currently employed or not.
8) Employment Reason: This column provides information about the reason for the applicant's employment status, such as being placed, not placed, or at disposal.
9) College Degrees: This column specifies the highest level of education attained by the applicant, typically in terms of college degrees or qualifications.

10) Basic Qualification Mark (Max=25): This column represents the mark, or score obtained by the applicant in their basic qualification assessment, with a maximum possible score of 25.
11) Tawjihi Mark (Max=15): This column indicates the mark, or score obtained by the applicant in their Tawjihi examination, with a maximum possible score of 15.
12) Replacement Mark (Max=8): This column represents any replacement mark or additional score that the applicant may have obtained, with a maximum possible score of 8
13) Experience Mark (Max=8): This column indicates the mark, or score awarded to the applicant based on their prior experience, with a maximum possible score of 8.
14) Scientific Qualification Mark (Max=14): This column denotes the mark, or score obtained by the applicant in their scientific qualification assessment, with a maximum possible score of 14.
15) Special Case Mark (Max=6): This column represents any special case mark, or additional score awarded to the applicant, with a maximum possible score of 6
16) Exam Mark (Max=15): This column indicates the mark, or score obtained by the applicant in their exam, with a maximum possible score of 15.
17) Interview Mark (Max=17): This column denotes the mark, or score awarded to the applicant based on their performance in the interview, with a maximum possible score of 17
18) Total This column represents the total marks, or cumulative score obtained by the applicant, considering all relevant assessments and evaluations.

Table 3. 1 Data Description.

Feature Name	Feature Type	Description
Year	Integer	The academic year in which applicants applied for vacancies
Administrative Unit	Object	The city to which applicants applied for the vacancy
Gender	Object	Gender (Male, Female)
Exam Material	Object	The test subject was taken by the applicant
Record Type	Object	The record type of the applicant after screening
Order	Integer	The order of the applicant after screening
Employment Status	Object	Hired, not hired
Employment Reason	Object	Employment Reason if available
Full Name	Object	Full Name
ID	Object	National code
College Degrees	Object	College Degrees
Specialization	Object	Specialization
Basic Qualification	Float	Bachelor's grade out of 100 or 4
Basic Qualification Mark	Float	Bachelor's grade out of 25
Tawjihi	Float	Tawjihi Mark out of 100
Tawjihi Mark	Float	Tawjihi Mark out of 15
Replacement (Days)	Float	The number of days the applicant worked as a substitute in education
Replacement Mark	Float	The applicant's grade is in exchange for the number of days the applicant worked as a substitute in education
Experience (Days)	Float	Experience in Days
Experience Mark	Float	The applicant's mark is instead of transferring days of experience
Scientific Qualification Mark	Float	Diploma, PhD
Special Case Mark	Float	If the applicant has special cases, for example, a prisoner, a wounded person
Exam Mark	Float	Exam Mark
Interview Mark	Float	Interview Mark
Total	Float	Total mark for all the criteria

3.3. Data Preprocessing

The dataset utilized in this study originates from the Ministry of Education, collected in the year 2022. Initially comprising 25 columns (features) in CSV file, the dataset underwent a thorough

cleaning and preprocessing procedure supervised by HR and data science experts. Post-cleaning, it was refined to include 18 features, ensuring a standardized and high-quality dataset for analysis.

Preprocessing the data before training in machine learning can be considered as an important step. Preprocessing data includes many main steps: data cleansing or data scrubbing, transformation, aggregation, normalization, or feature scaling. [51].

Develop a recommendation system based on the applicants' data to select the right teacher in the ministry of education using machine and deep learning techniques, reduce features based on feature selection techniques, tune model parameters, and evaluate the performance of the model.

For numeric data, standardization is usually used, which is critical for regularization techniques to avoid overfitting many algorithms. In dealing with categorical data, we proposed using many techniques from many libraries such as sci-kit-learn, and we proposed to use a label encoder.

In data processing steps such as label encoding for categorical features were transformed into numerical values using label encoding. Each category within a feature was assigned as unique integer. This step makes the data compatible with algorithms that require numerical input. Moreover one-hot encoding. Another categorical encoding technique, one-hot encoding, is applied to create binary columns for each category. This approach avoids ordinality assumptions and ensures that categorical variables are represented accurately in the dataset. Additionally, a column transformer was used to apply these transformations to the appropriate columns while leaving numerical features unchanged. This preprocessing pipeline ensures that both categorical and numerical features are appropriately prepared for use in machine learning models.

In general, I took a number of procedures to ensure that the data was error-free, consistent, and focused on the most important elements. I converted the data into model-compatible forms, deleted

unneeded or duplicated columns, and addressed any outliers or inconsistencies. All of these stages contribute to ensuring that the dataset is clean, well-prepared, and suitable for the creation of a machine learning model.

Data preprocessing is a critical step in preparing data for analysis and developing machine learning models. The procedure begins with Data Cleaning, in which I meticulously verified each column to ensure that the data was error-free and consistent. I checked the unique numbers in each column to ensure their accuracy and that there were no inaccurate or outlier values. I also discovered and eliminated duplicate data to avoid skewed findings. I took necessary action in columns with missing values, either eliminating or imputing the missing data. In several number columns, I used scaling techniques to standardize values, During the scaling process, the dataset was initially partitioned to avoid data leakage. Min-Max scaling was employed as it useful when data is not Gaussian, while Standard scaling was not utilized. Standard scaling is typically applied to normally distributed data, making Min-Max scaling the preferred choice for this dataset. removing negative values and those more than 100. To preserve consistency, university grades were converted from a scale of 1 to 4 to a 100-point system. During Data Transformation, categorical variables were encoded into numerical representations, and I used One-Hot Encoding when needed to turn categorical data into a format that the machine learning model could understand. During the Feature Engineering step, I carefully chose the model's most relevant characteristics, deleting those that had little to no impact on performance. Features like "Name" and "ID Number" were removed because they did not add to model correctness. To reduce complexity, I deleted unnecessary columns such as "Substitute Duration" and "Substitute Grade," which provided comparable information. In addition, I will use SelectKBest to select some features and reduce the dimensionality of the model. First, I'll split the features into continuous and categorical. For the

categorical features, I will select a subset using Chi-square, as the input is categorical and the output is categorical. For the continuous features, I will select a subset using ANOVA, as these is appropriate when the input is numerical and output are categorical. I addressed outliers by eliminating any numbers that exceeded predicted ranges, ensuring that only genuine data was used. Following data cleaning and transformation, I split the dataset into training and testing sets to assess the model's performance on previously unseen data. Finally, during the Data Validation process, I verified the data's overall quality and consistency and used Cross-Validation to determine the model's ability to generalize to other subsets of the data. This procedure ensured that the data was clean, well-structured, and prepared for the subsequent phases in developing an efficient machine learning model. I compared different techniques for handling imbalanced classes using precision and recall metrics across all classification models to determine which technique and model yielded the highest precision. I used several techniques, including SMOTE, ADASYN, Oversampling, and Near Miss Down sampling, and ultimately selected Oversampling as the best approach due to its superior accuracy

Figure 3.2 presents the categories absent, not place and place. There is an imbalance in the dataset classes. Initially, I attempted to use the class weighting technique in this research, which is a common technique used in ML to treat the imbalance classes problem. The imbalanced data can impure the classification process. class weighting technique, could improve the performance of models which can lead to high accurate classification results.

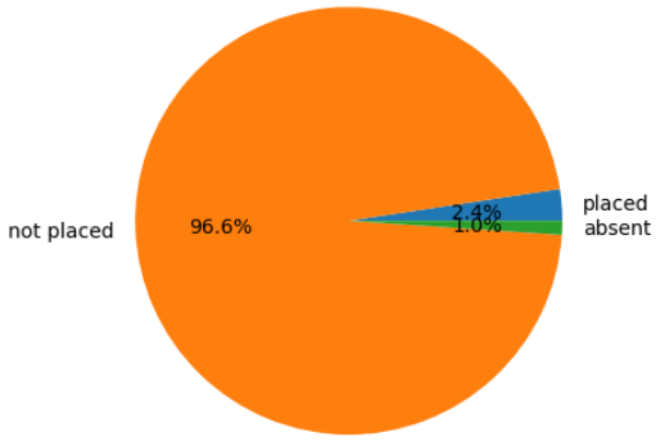


Figure 3. 2 Employment status categories for imbalance.

Figure 3.2 presents the employment status categories after using the class weighting technique, which achieves a balance between all classes.

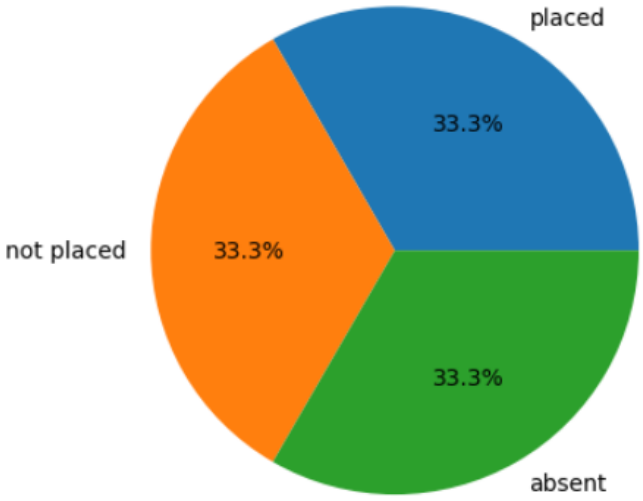


Figure 3. 3 Employment status categories after class weighting technique

3.4. Exploratory Data Analysis (EDA)

EDA is a significant stage in data analysis. It includes visualization, exploration, descriptive statistics for numeric values, pattern recognition, anomaly detection, data types, correlation between features, etc. A better understanding of the data can lead to initial insights to pick up the appropriate ML.

More overt the descriptive statistics that provided valuable insights into the numerical features of the dataset. Here's a summary of the key findings:

Year: All records are from the year 2022.

Order: The mean order is approximately 26.53, with a wide range of values indicating variability in the application order.

Basic Qualification Mark: The average mark is around 15.17 out of a maximum of 25, with a standard deviation of approximately 7.09.

Tawjihi Mark: Applicants have an average Tawjihi mark of about 11.68 out of 15, with a standard deviation of around 1.72.

Replacement Mark: The mean replacement mark is quite low at approximately 0.21, indicating that most applicants did not have replacement marks.

Experience Mark: On average, applicants have a low experience mark of about 0.50 out of 8, with a standard deviation of approximately 1.60.

Scientific Qualification Mark: The mean scientific qualification mark is about 0.31 out of 14, with a standard deviation of around 1.12.

Special Case Mark: Most applicants do not have a special case mark, as indicated by the low mean of approximately 0.06.

Exam Mark: The average exam mark is approximately 6.31 out of 15, with a standard deviation of about 3.72.

Interview Mark: On average, applicants scored around 3.26 out of 17 in the interview, with a wide standard deviation of approximately 5.53.

Total: The total marks have an average of approximately 37.51, with a standard deviation of around 11.18.

These statistics provide a comprehensive overview of the distribution and variability of numerical features in the dataset.

Figure 3.4 presents overall details for all features, such as data types, number of records, missing values, and duplicates.

Figure 3. 4 Overview of the Dataset

Figure 3.5 shows that males occupied (16.1%) of the data represented by an orange slice, while females occupied (83.9%) represented by a blue slice., in this sense we can claim that there is an accepted balance between both classes which signifies gender non-bias in the data.

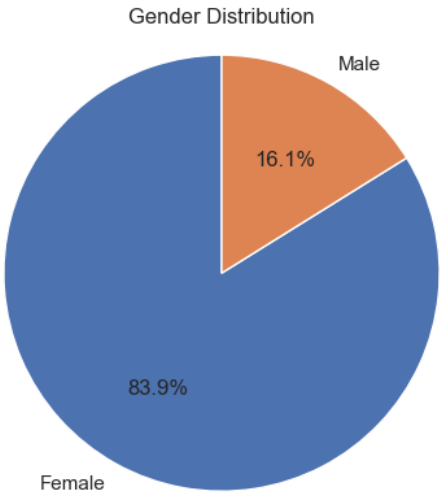


Figure 3. 5Gender Distribution for the sample.

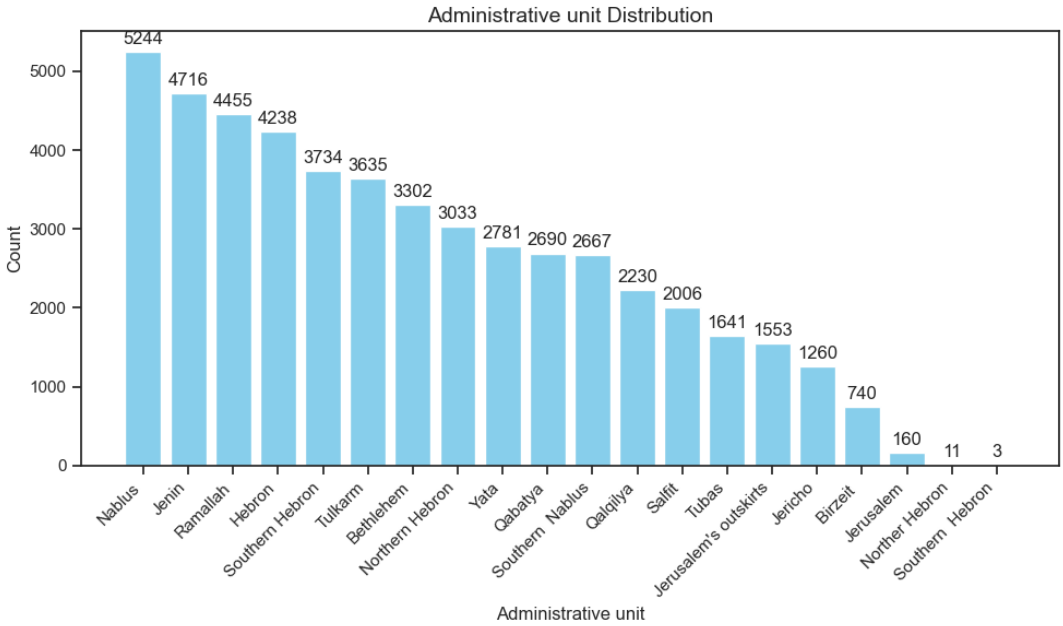


Figure 3. 6 Administrative unite distribution for the sample

In Figure 3.7, the histograms demonstrate that most candidates have total scores clustered within a specific range, resulting in a skewed distribution. The distribution of average scores also reveals concentrations around specific values, indicating standardized grading. Also, the Tawjihi, experience, and exam marks exhibit distinct clusters of values, suggesting that many candidates achieved similar scores in these categories.

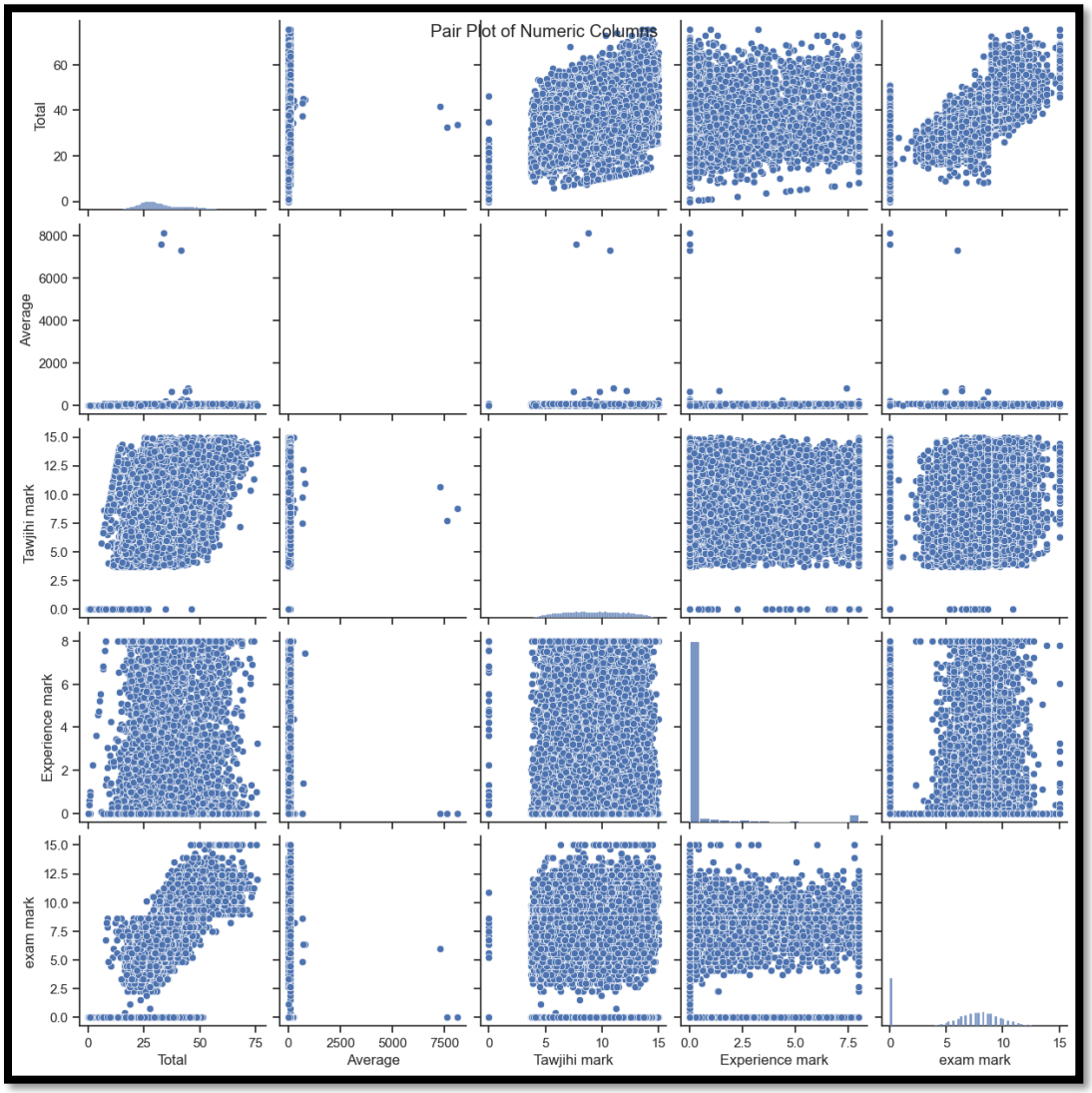


Figure 3. 7 Pair Plot

Figure 3.8 displays a heatmap providing valuable insights into the relationships between different variables in our dataset. Understanding these correlations is crucial for optimizing the feature selection process of our machine learning models. In this visual representation, the correlation between variables ranges from -1 to 1. A value of 1 signifies a perfect positive correlation, -1 represents a perfect negative correlation, and 0 indicates no correlation. The color intensity helps in quickly identifying the strength and direction of the correlations.

A significant correlation exists between Tawjihi and Tawjihi marks (0.92), indicating a strong positive correlation, which aligns with the idea that these two variables likely measure related aspects of educational achievement. Similarly, there is a strong relationship between replacement and replacement marks (0.93), showing the link between the presence of replacements and their associated marks. Additionally, the exam mark and interview mark (0.60) have a moderate positive correlation, suggesting that good performance in exams also tends to translate to good performance in interviews. The exam mark and total (0.69), and interview mark and total (0.83) are strongly correlated with the total score, indicating that performance in exams and interviews significantly impacts the final evaluation.

On the other hand, the negative correlation between Year and Total (-0.26) suggests that the total score tends to decrease slightly as the years increase, indicating changes in criteria or performance over time. Furthermore, many variables show low or near-zero correlations with others (e.g., Experience and other marks), suggesting they don’t have a strong linear relationship with those aspects.

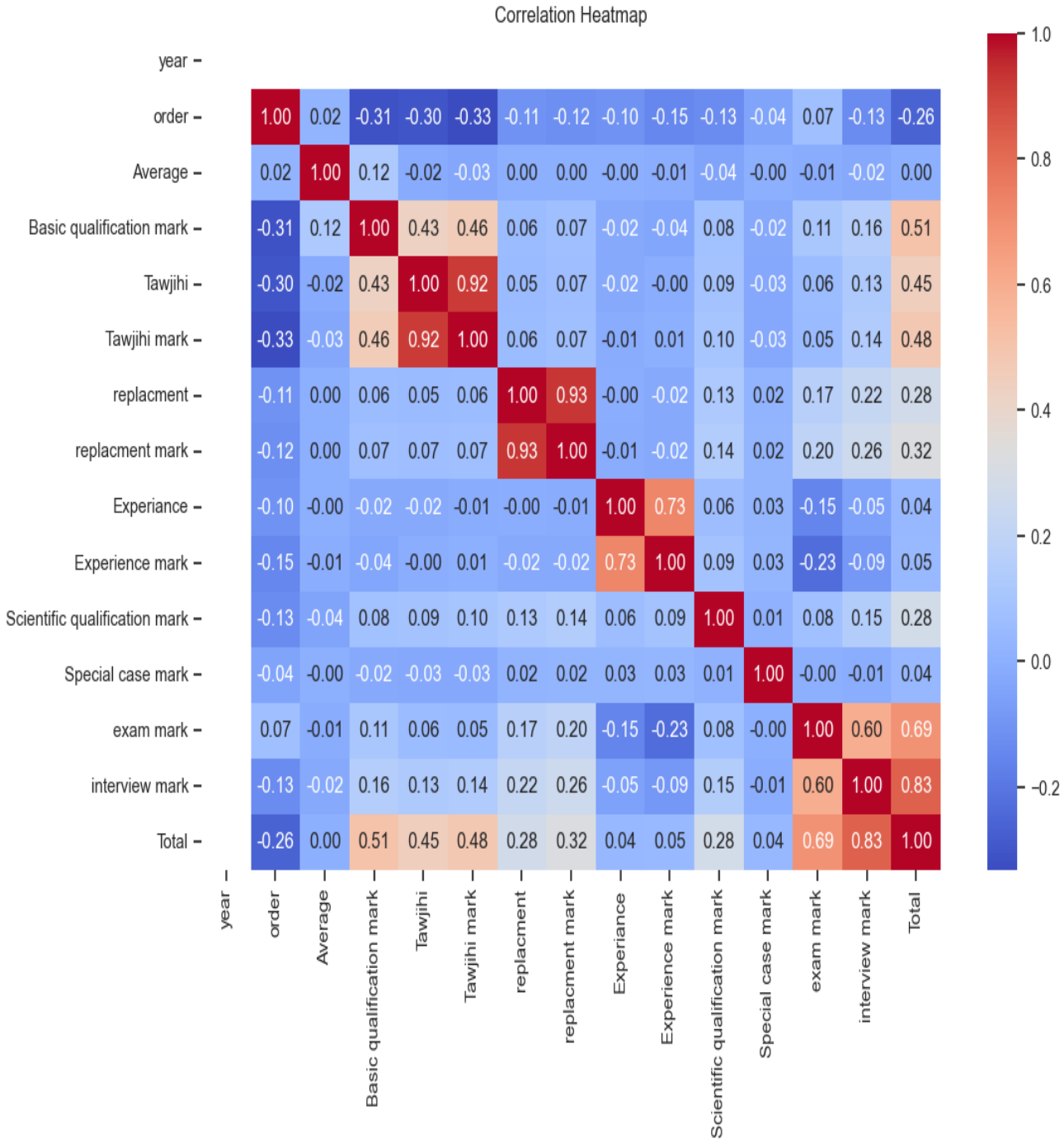


Figure 3. 8 Correlation Heatmap

Figure 3.9 is a scatter plot showing the relationship between the average scores, and Tawjihi marks for candidates with average scores below 100. It shows a prominent cluster of candidates with high averages (60-100) and varying Tawjihi (6-14); however, those with low averages (around 0) have Tawjihi marks under 6. That may imply a gap between these clusters, suggesting distinct candidate categories, which means there is no clear linear relationship between the two variables.

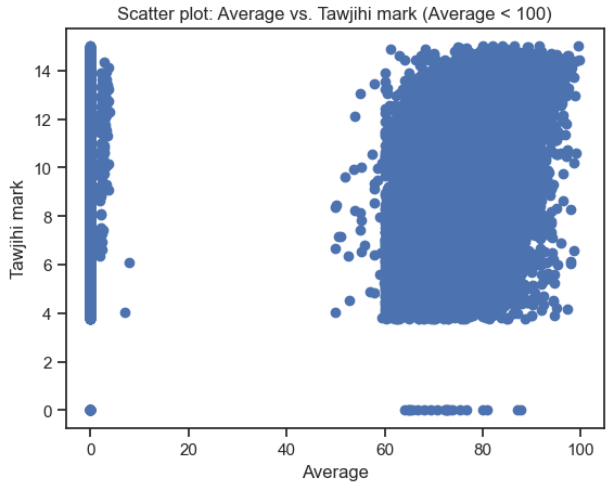


Figure 3. 9 Box plot presents the distribution of the essential qualifications.

Figure 3.10 box plot shows the distribution of the essential qualification marks across various educational qualifications and illustrates differences in performance. First, the PhD holders achieve the highest median marks around 18-19 with consistent scores and few outliers, which may reflect their advanced expertise. For the second and third educational degrees, the candidates with master's and bachelor's degrees showed more variability, with median marks around 16-17 and 13-14, respectively, and significant outliers, particularly in the Bachelor's group. Fourthly, a median of 12-15 was for the candidiasis holding a diploma, and the lower median was around 10 for the vocational diploma holders.

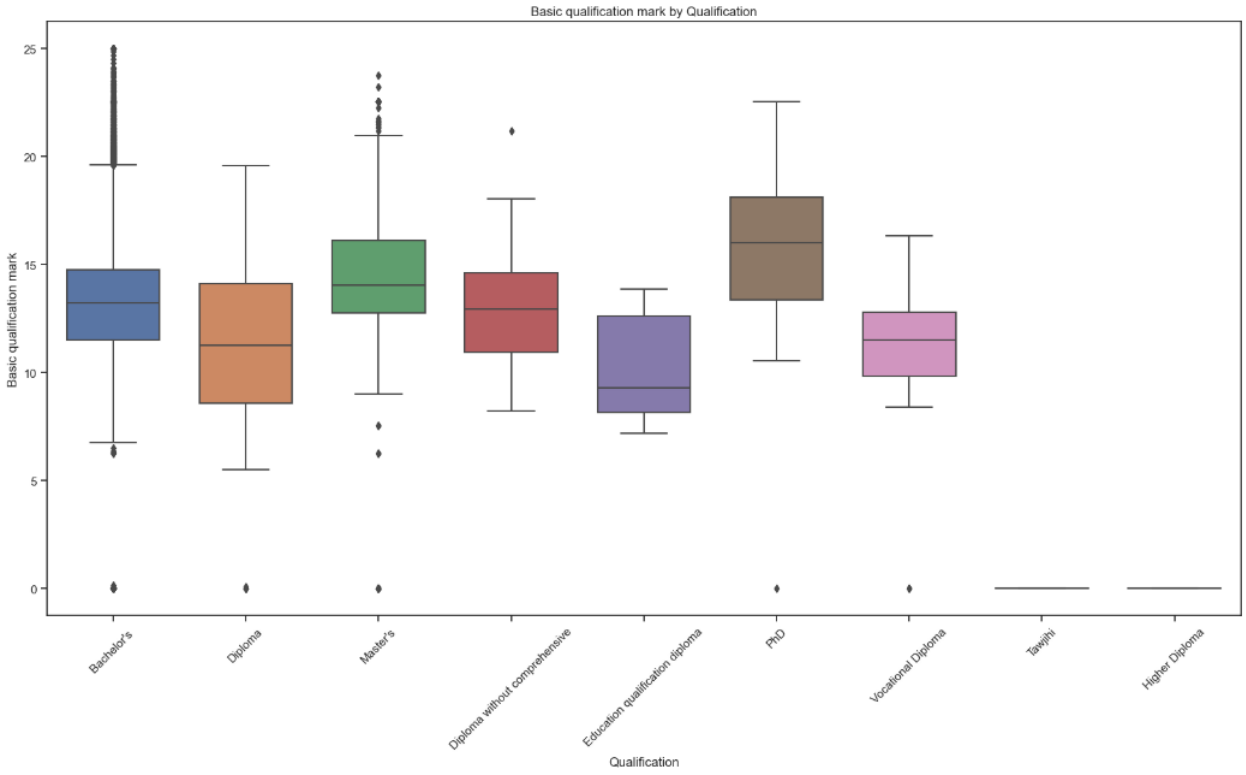


Figure 3. 10 Essential qualification mark by qualification.

3.5. Proposed Models

As part of the research, a recommendation system for selecting the appropriate teacher was developed, and different machine-learning algorithms were employed to train the data and identify the most effective models. The models aim to predict the best candidates based on various factors, ensuring accurate recommendations. We utilize supportive machine learning algorithms from diverse architectures like Random Forest, Support Vector Machines, K-Nearest Neighbors, XGBoost, and Multi-layer Perceptron Neural Networks. However, each one of the algorithms is discussed individually in the following sections, highlighting its role in optimizing the teacher selection process.

3.5.1. Support Vector Machines

The Support vector machines SVM is considered as a supervised machine learning algorithm for classification-based tasks. It creates a decision boundary between different classes, which enables the model to predict the datapoint class depending on the feature vectors. These decision boundaries, known as the hyperplane, are orientated so that they are as far as possible from the closest data points from each class. These closest points are called support vectors.[55].

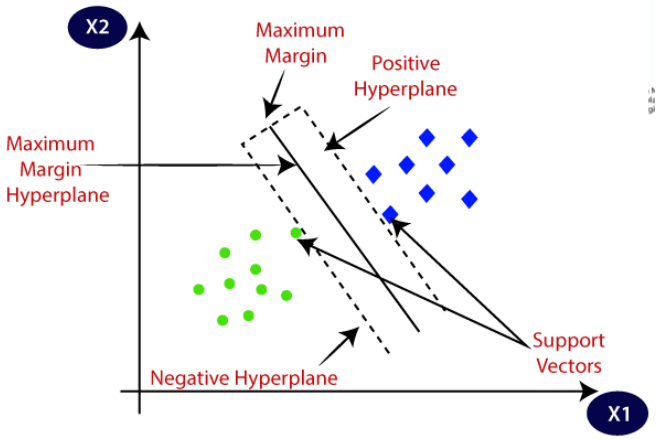


Figure 3. 11 The Basics of SVM. [54].

Figure 3.11 presents the basics of SVM, and how it works to separate data points into two classes and easily predict new data points based on the maximum marine hyperplane between support vectors that identify the best hyperplane. However, The types of hyperplane maximum margin classifier (hard margin), support vector classifier (soft margin), and support vector machines (non-linear) [55].

The SVM equation is a linear equation to make a classification for data points and try to find the optimal hyperplane and the equation is shown in the figure below.

$$W * X + b = 0 \qquad \textbf{(1)}$$

Where x is the input vector, W is the weight vector and b are bias.

3.5.2. Decision Trees

Decision trees (DTs) are considered non-parametric supervised machine learning methods used in data mining for creating models to predict the target variable by learning rules from the data attributes [54] .A decision tree's hierarchical structure includes a root node, branches, internal nodes, and leaf nodes (terminal nodes). It aims to assign the optimal decision based on the probability of the target.[55], [56]

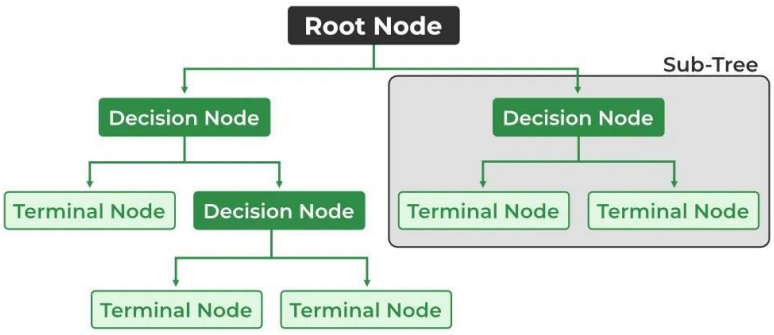


Figure 3. 12 The basics of DTs [57]

The Gini formula is:

$$1 - P(Class\ 1)^2 - P(Class\ 2) \quad \textbf{(2)}$$

The entropy formula is:

$$P(class\ 1)\ x\ log2(P(class\ 1)) - P(class\ 2)\ x\ log2\ (P(class\ 2)) \quad \textbf{(3)}$$

Where P is the probability and probability of the class by logarithm base= 2.

Information Gain (IG)

IG or reduction of Entropy, which evaluates the uncertainty reduction for a given feature and determines which attribute should be chosen as a decision node or root node [58].

3.5.3. Ensemble Methods

Strong learners by utilizing multiple weak learners are known as ensemble learning. which involves using multiple trees instead of just one and combining these weak learners to arrive at the most optimal solution for a strong learner. Voting classifiers employ two main approaches to achieve the best outcome: Bagging and Boosting. Bagging involves creating independent samples from the training dataset, each taking a different subset of the dataset with replacement. The primary aim of bagging is to reduce variance and overfitting by aggregating all the votes and picking the majority. However, Boosting takes a different approach without replacement, and the process involves sequential sampling. Each tree in the sequence aims to address the issues encountered by the previous tree. The objective of Boosting is to reduce the training error at each iteration [58], [59].

3.5.3.1.Random Forests

Random Forests are supervised machine learning algorithms where hundreds or thousands of decision trees are built, and they gather all the results to make predictions or classifications. RF used bagging techniques to find the optimal solution, which is a strong tool to avoid overfitting, extract complex interactions between features, and provide the features with importance. build hundreds or thousands of decision trees and gather all the results to make predictions or classifications. RF uses Bagging techniques to find the optimal solution, which is a strong tool to avoid overfitting, extract complex interactions between features, and provide.[59], [60].

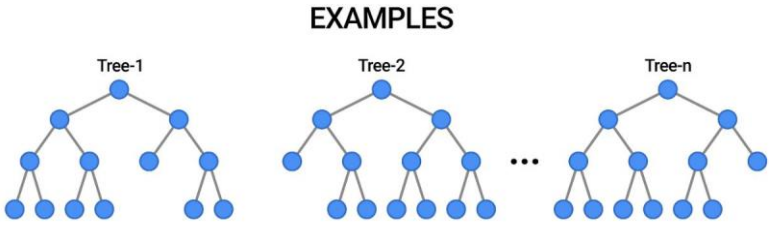


Figure 3. 13 The basics of RF [60]

The equation of RF is:

$$F(x) = 1/N \sum_{i} ((fi(x))$$

(4)

Where F(x) is the output of the RF algorithm, N = the number of decision trees in the forest.

3.5.3.2.Extra Tree Classifier

The extra tree classifiers and RF both belong tom bagging technology in the same ensemble methods in introducing the additional randomness during the tree construction and process. [55], [61].

Table 3. 2 The difference between the RF and ETC combining the main categories.[61].

	RF	**ETC**
Randomness	Random feature subsampling at each split.	Random feature and threshold selection at each split
Handling of Randomness	Majority voting or averaging of predictions	Majority voting of predictions
Computational Efficiency	Evaluate multiple thresholds for each feature	Selects random thresholds without evaluating multiple candidates
Variance	Lower variance due to reduced randomness	Slightly higher variance due to increased randomness
Hyperparameter Tuning	Common hyperparameters (estimators, max_depth, etc.)	Different default values for some hyperparameters (e.g., larger max_features)
Training Speed	Slightly slower due to evaluating multiple thresholds	Faster due to not evaluating multiple candidates

3.5.3.3.Gradient Boosting

Another machine learning algorithm is Gradient Boosting (GB). I am using GB, which uses a sequential approach. GB's main aim is to lessen errors from previous predictors. According to the literature, the "gradient" in Gradient Boosting refers to the technique's optimization process, which minimizes the loss function by repeatedly adding weak models to the ensemble. [61].

The GB could be applied to non-linear relationships and extract the pattern in the data giving the privilege of using GB to prevent overfitting, and, like other algorithms that could be used for classification or regression, GB can handle complex datasets and obtain appropriate accuracy.

Also, GB can address missing data. GB can handle any kind of data, numerical or categorical, easily. Finally, GB could be applied to non-linear relationships and extract the pattern in the data [62], [63].

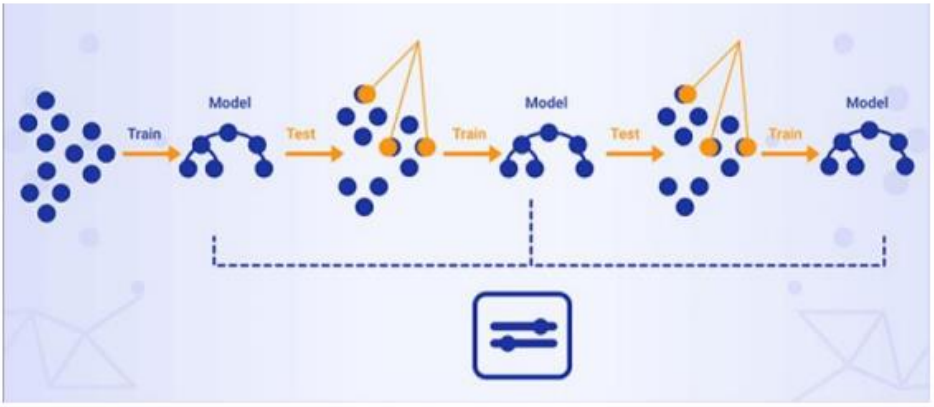


Figure 3. 14 Gradient Boosting Flowchart [62]

3.5.3.4.XGBoost

The XGBoost is considered the development outcome of GB, which refers to Extreme Gradient Boosting. The aim of using XGBoost is to have a better version of the GB algorithm's ability. The advantage of XGBoost, like GB, is that it prevents overfitting in datasets. One of its uses is classification or regression, and it can handle complex datasets and obtain appropriate accuracy. Also, XGBoost can handle missing data. XGBoost can handle any data, numerical or categorical, efficiently. Finally, XGBoost could be applied to non-linear relationships and extract the pattern in the data. [63]

Table 3. 3 GB vs XGBoost

Type	Gradient Boosting	XGBoost
Optimization	Ensemble Learning	Ensemble Learning
Regularization	An iterative process, adjusting residuals	An iterative process adjusting gradients
Handling Missing Values	No	regularization techniques to control overfitting
Parallel Computing	No	Yes, learned the best direction for missing values
Hyperparameter Tuning	Moderate	Yes, column block Parallelization
Pruning	No	Tree pruning
Performance	High performance, handles complex data	High performance, handles complex data
Popular Uses	Various domains such as finance, healthcare, N	Various domains such as finance, healthcare, NLP
Implementation	More complex implementation	More optimized and efficient implementation

3.5.4. Multilayer Perceptron Neural Networks

Can deal with huge datasets to recognize hidden patterns and solve complex problems. An ANN sonatinas interconnected nodes called artificial neurons or units. Artificial Neural Networks (ANNs) is a supervised machine learning model. ANNs are computational models inspired by the functionality of biological neural networks in our brains. [64]. These units settle in layers, typically including an input layer as a first layer in the architecture, one or more hidden layers, and an output layer. The connections between these neurons are represented by weights and biases, which are modified weights and biases to minimize error during the learning process in every iteration. [11]. The Multi-Layer Perceptron is one type of Neural Network that consists of the input layer, at least one hidden layer, and an output layer.

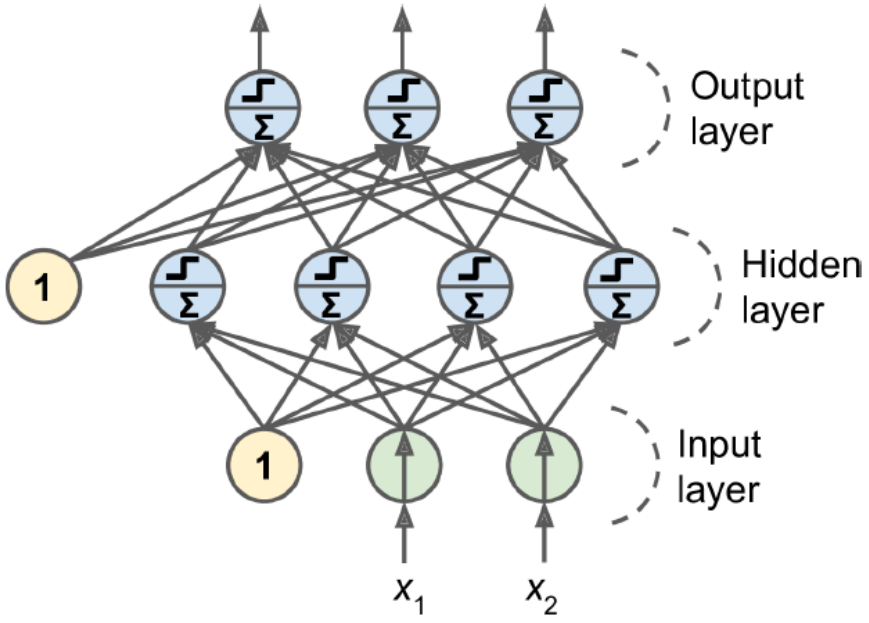


Figure 3. 15 The architecture of Multilayer Perceptron NNs [58]

3.6.Tuning Parameter

In this section, we are trying to explain the tuning parameters in ML that indicate the most appropriate set of parameters to boost the performance of models. The parameters must be defined in the training dataset, or hyperparameters, which are the parameters tuned to achieve the most accurate results. However, Hyperparameter tuning can be selected manually by an expert or can use automatic techniques to determine the hyperparameter such as grid search evaluating all possible combinations, random search such as sampling random combinations, or more advanced optimization techniques like Bayesian optimization or genetic algorithms. [58]. In this research, we propose using the Grid search technique to determine the best hyperparameter to achieve an appropriate result for our model.

SVM contains a few parameters that demand to be tuned for optimal performance for the model. There are parameters that, when tuned, could achieve appropriate results, such as the kernel function types that permit SVMs to deal with nonlinearity relationships between features and target variables, the regularization parameter (C) and gamma, (C) try to make the balance between training error and decision boundary complexity. Otherwise, gamma decides the influence of individual training examples[65]. Using the grid search technique provides the most appropriate choice to determine the best parameter value to obtain appropriate results. Grid search technique will be used for all upcoming models to achieve the best values and results.

DT tuning parameters is an important step to enhancing the effectiveness of the model. There are many parameters that, when tuned, could achieve appropriate results, such as the maximum depth controls the depth of the tree, preventing overfitting when set too high. Otherwise, A low value may result in underfitting.[65]. The minimum number of samples for splitting and leaf nodes control the stopping criteria for further partitioning in the tree. The criterion parameter specifies

the quality measure used for splitting the nodes. The two commonly used criteria are Gini impurity and entropy to figure out the IG to determine the root of the tree and all nodes.[66].

For tuning the parameters of bagging algorithms like Random Forest and Extra Trees classifiers, there are diverse parameters that can be tuned. Such as the number of trees (n_estimators), which controls the depth of trees (max_depth), minimum samples necessary to split an internal node (min_samples_split), minimum samples required at a leaf node (min_samples_leaf), the number of features considered for each split (max_features), and the use of bootstrap sampling (bootstrap)[67].

For tuning the parameters of boosting algorithms like Adaboost, Gradient Boosting, and XGBoost, there are diverse parameters that can be tuned. Such as the learning rate (step size), the number of estimators (iterations), and weak learner parameters (e.g., maximum depth). In GB and XGBoost the sampling[68], [69]. In the tuning of the parameters of a Multilayer Perceptron (MLPNN), many parameters can be tuned. Such as the hidden layer of the architecture (the total number of layers and neurons in each layer), activation functions, learning rate, regularization techniques, batch size, number of epochs, and the optimization algorithm[69]. ML algorithms can be fine-tuned to obtain the most appropriate performance, highest accuracy, and agile convergence, and to avoid overfitting or underfitting in our model.

3.7.Performance Evaluation

Evaluating the performance in ML is the technique to test the model results. It is considered an important stage in developing machine learning models. Also, the evaluation model's performance can help the data scientist to assess the model's behavior in with dealing the unseen dataset. To evaluate the model's performance there are many metrics to help the data scientist. Such as Confusing Matrix, precision, recall, F-1 score, and AUC and ROC curve. The cross-validation

techniques can shuffle the dataset to avoid the pattern in the dataset which can help the models to achieve appropriate results for each model. After that, experts will compare the results between the model's performance by experts to determine which model could be the best one.[49].

3.8.Confusion Matrix

We have the basic concepts in a confusion matrix. TP is the number of true outcomes, and the model classifies it as true. While FP is the number of false outcomes, and the model classifies it as true. On the other hand. TN is the number of false outcomes, and the model classifies it as false. FN is the number of true outcomes, and the model classifies it as false. The important formulas based on the Confusion Matrix are Accuracy, Precision, recall, and F-1 score. [58], [65], [70]

Table 3. 4 Confusion matrix [71]

	Predicted: Yes	**Predicted: No**
Actual: Yes	TP	FN
Actual: No	FP	TN

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

(5) [72]

The accuracy formula can provide the percentage of correct predictions on all expectations.

$$Precision = \frac{TP}{(TP + FP)}$$

(6) [11]

Precision can provide the percentage of correct values that are correct.

$$Recall = \frac{TP}{(TP + FN)}$$

(7)[72]

The recall can provide the percentage of correct values that are predicted correctly.

$$F - 1\ score = \frac{2(precision * recall)}{(precision + recall)}$$

(8) [72]

The f-1 score is a metric that seeks to measure both precision and recall.

3.9.ROC & AUC Curve

We used the Receiver Operating Characteristic (ROC) and according to the definition, the ROC is a curve that plots sensitivity on the y-axis against specificity on the x-axis. That's to explain the relationship between True Positive Rate (TPR) and False Positive Rate (FPR). Aera Under Curve (AUC) points out the model's performance, more cover area under the curve is more accurate. [11], [64], [72]

Chapter Four

Result and Outcome

In this chapter, we represent the results of all ML models we propose to use in this research, the results will appear in two phases using the imbalance and balance data sets also we used ROC and AUC to determine the best module that we should use as the final one.

4.1. Classification Result of SVM

The SVM model covers balanced and imbalanced datasets, so first, dealing with the data imbalanced, the SVM model's performance on this imbalanced dataset is heavily skewed towards the dominant class, "not placed." The model achieves high precision (0.97), recall (1.00), and F1-score (0.99) for this class, which constitutes most of the data. Also, it completely fails to correctly predict the minority classes, "absent" and "placed," as indicated by their precision, recall, and F1-scores of 0.00. This is further reflected in the confusion matrix, where all instances of "absent" and "placed" are misclassified as "not placed." The overall accuracy is 97%, but this metric is misleading due to the significant class imbalance. The macro averages (precision, recall, and F1-score, all at 0.33) highlight the model's poor performance across different classes. This underscores the challenges of using SVM on highly imbalanced data, where the model may overly favor the majority class, leading to poor predictive power for the minority classes.

In contrast, dealing with the balanced dataset, the model demonstrates overall solid performance with an accuracy of 99%, as appears in Figure 4.2, primarily driven by its excellent prediction of the dominant class, which is not placed, which achieves perfect precision, recall, and F1-score. However, the model struggles with the minority class, "absent," showing low precision (0.13) and moderate recall (0.36), resulting in a poor F1-score of 0.19. The "placed" class fares better, with high precision (0.94) and rational recall (0.81), leading to an F1-score of 0.87. The weighted

averages are very high due to the prevalence of "not placed." Macro averages highlight the disparity in performance across different classes, indicating that the model's effectiveness diminishes significantly for the less frequent categories. Figure 4.3 also presents the ROC for SVM.

	precision	recall	f1-score	support
absent	0.13	0.36	0.19	25
not placed	1.00	1.00	1.00	12142
placed	0.94	0.81	0.87	319
accuracy			0.99	12486
macro avg	0.69	0.72	0.69	12486
weighted avg	1.00	0.99	0.99	12486

Figure 4. 1 SVM Balance

	precision	recall	f1-score	support
absent	0.00	0.00	0.00	25
not placed	0.97	1.00	0.99	12142
placed	0.00	0.00	0.00	319
accuracy			0.97	12486
macro avg	0.32	0.33	0.33	12486
weighted avg	0.95	0.97	0.96	12486

```
Confusion Matrix:
[[    0    25     0]
 [    0 12142     0]
 [    0   319     0]]
```

Figure 4. 2 SVM imbalance

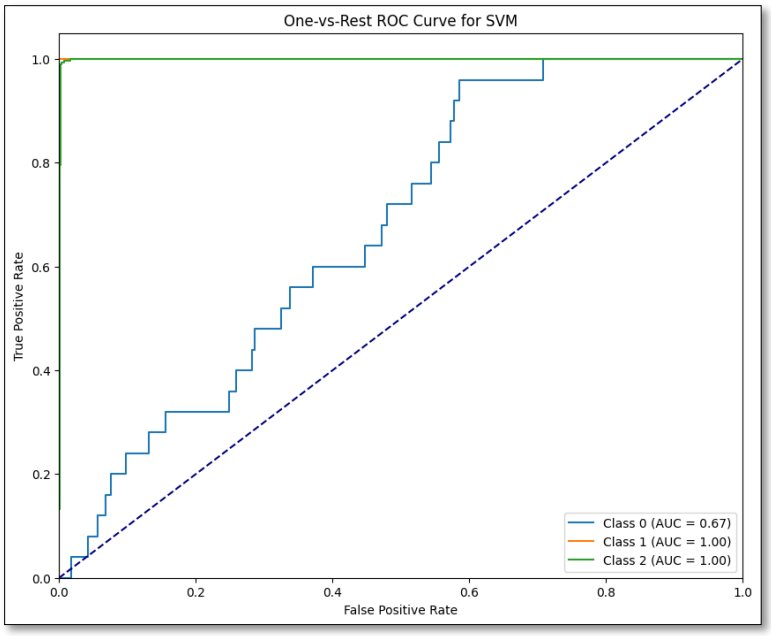


Figure 4. 3 ROC for SVM

4.2.Decision Trees

In this phase, the Decision Tree deals with an imbalanced dataset. Figure 4.4 represents the performance metric used to evaluate the effectiveness of applying the machine learning algorithm for the Decision Tree, which includes accuracy, precision, recall, and F1 score. Figure 4.6 represents the AUC Roc curve. These metrics will be used to analyze the model's performance comprehensively. As we can see in the figures, the result was for the class absent, not placed, and the place was the accuracy (99), and for the absent class, the Precision 12, Recall 12, F1-Score 12. The second class, Not Placed the Precision 100, Recall 100, F1-Score 100. The third class Placed the Precision 93, Recall 93, and F1-Score 93.

Conversely, the second phase of the Decision Tree deals with balanced datasets. Figure 4.5 represents the performance metric used to evaluate the effectiveness of applying the machine learning algorithm for the Decision Tree, which includes accuracy, precision, recall, and F1 score.

Figure 4.6 represents the AUC Roc curve. These metrics will be used to analyze the model's performance comprehensively. As we can see in the figure, the result was for the class absent, not placed, and the place had an accuracy of 99, and for the absent class, the precision was 21, recall 20, and F1-Score 20. The second class, Not Placed the Precision 100, Recall 100, F1-Score 100. The third class Placed the Precision 94, Recall 94, and F1-Score 94. In conclusion, the Decision Tree model demonstrates excellent performance, particularly for classes "not placed" and "placed." However, its performance in the "absent" class could be improved. The Decision Tree model appears to be a strong candidate for classification tasks, given its high accuracy and balanced performance across multiple metrics.

```
Accuracy (Decision Tree): 0.9963959634790965
Classification Report (Decision Tree):
              precision    recall  f1-score   support

      absent       0.12      0.12      0.12        25
  not placed       1.00      1.00      1.00     12142
      placed       0.93      0.93      0.93       319

    accuracy                           1.00     12486
   macro avg       0.68      0.68      0.68     12486
weighted avg       1.00      1.00      1.00     12486
```

Figure 4. 4 Decision Tree imbalanced

Classification Report:

	precision	recall	f1-score	support
0	0.93	0.91	0.92	367
1	1.00	1.00	1.00	14582
2	0.17	0.21	0.19	34
accuracy			1.00	14983
macro avg	0.70	0.70	0.70	14983
weighted avg	1.00	1.00	1.00	14983

Figure 4. 5 Decision Tree Balanced

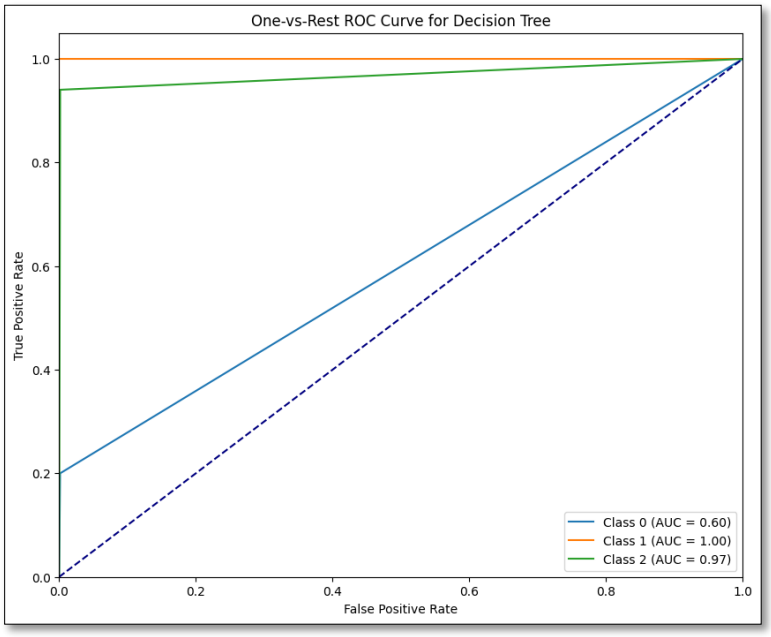


Figure 4. 6 ROC Decision Tree

4.3 Random Forests

In this phase, Random Forests deals with an imbalanced dataset. Figure 4.7 represents the performance metric used to evaluate the effectiveness of applying the machine learning algorithm for RF, which includes accuracy, precision, recall, and F1 score. Figure 4.9 represents the AUC Roc curve. These metrics will be used to analyze the model's performance comprehensively. As we can see in the figure, the result was for the class absent, not placed, and the place was the accuracy (100), and for the absent class, the Precision 0, Recall 0, F1-Score 0. The second class, Not Placed the Precision 100, Recall 100, F1-Score 100. The third class Placed the Precision 93, Recall 100, and F1-Score 96.

On the other hand, the second phase of Random Forests deals with balanced datasets. Figure 4.8 represents the performance metric used to evaluate the effectiveness of applying the machine learning algorithm for RF, which includes accuracy, precision, recall, and F1 score. Figure 4.9 represents the AUC Roc curve. These metrics will be used to analyze the model's performance comprehensively. As we can see in the figure, the result was for the class absent, not placed, and the place had an accuracy of 100, and for the absent class, the precision was 0, recall 0, and F1-Score 0. The second class, Not Placed the Precision 100, Recall 100, F1-Score 100. The third class Placed the Precision 93, Recall 100, and F1-Score 96. Overall, the model performs well in classifying "not placed" instances but struggles with "absent" and "placed" instances, particularly in terms of precision and recall.

	precision	recall	f1-score	support
absent	0.00	0.00	0.00	25
not placed	1.00	1.00	1.00	12142
placed	0.93	1.00	0.96	319
accuracy			1.00	12486
macro avg	0.64	0.67	0.65	12486
weighted avg	1.00	1.00	1.00	12486

Figure 4. 7 RF imbalance

	precision	recall	f1-score	support
absent	0.00	0.00	0.00	25
not placed	1.00	1.00	1.00	12142
placed	0.93	1.00	0.96	319
accuracy			1.00	12486
macro avg	0.64	0.67	0.65	12486
weighted avg	1.00	1.00	1.00	12486

Figure 4. 8 RF Balanced

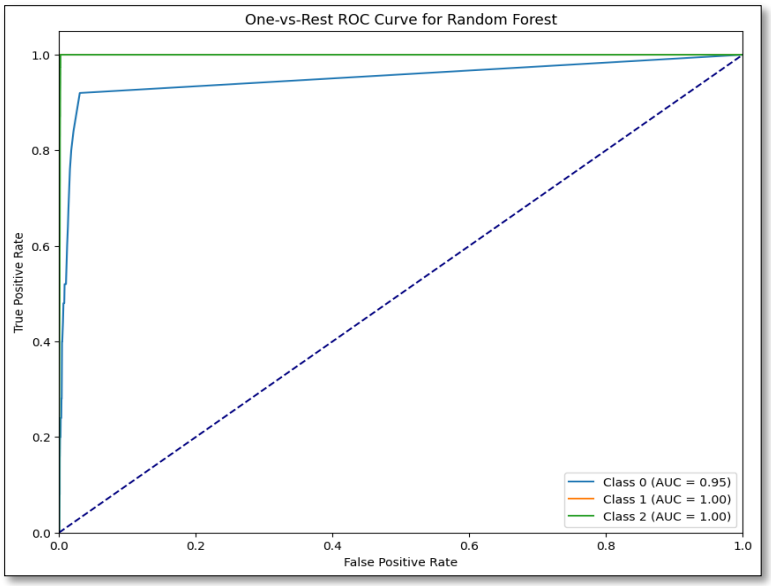


Figure 4. 9 ROC for Random Forest

4.3.AdaBoost

In this phase, AdaBoost deals with an imbalanced dataset. Figure 4.10 represents the performance metric used to evaluate the effectiveness of applying the machine learning algorithm for AdaBoost, which includes accuracy, precision, recall, and F1 score. Figure 4.12 represents the AUC Roc curve. These metrics will be used to analyze the model's performance comprehensively. As we can see in the figure, the result was for the class absent, not placed, and the place was the accuracy (99), and for the absent class, the Precision 20, Recall 4, F1-Score 7. The second class, Not Placed the Precision 100, Recall 100, F1-Score 100. The third class Placed the Precision 93, Recall 99, and F1-Score 96.

On the other hand, the second phase of AdaBoost deals with balanced datasets. Figure 4.11 represents the performance metric used to evaluate the effectiveness of applying the machine learning algorithm for AdaBoost, which includes accuracy, precision, recall, and F1 score. Figure

4.12 represents the AUC Roc curve. These metrics will be used to analyze the model's performance comprehensively. As we can see in the figure, the result was for the class absent, not placed, and the place had an accuracy (99), and for the absent class, the Precision 20, Recall 4, F1-Score 7. The second class, Not Placed the Precision 100, Recall 100, F1-Score 100. The third class Placed the Precision 93, Recall 99, and F1-Score 96. Moreover, it correctly classified all instances for the majority class ('not placed') but struggled to classify instances for the minority classes ('absent' and 'placed'). Candidate for classification tasks, given its high accuracy and balanced performance across multiple metrics.

```
Accuracy (AdaBoost): 0.9977574883869934
AdaBoost Classifier:
              precision    recall  f1-score   support

      absent       0.20      0.04      0.07        25
  not placed       1.00      1.00      1.00     12142
      placed       0.93      0.99      0.96       319

    accuracy                           1.00     12486
   macro avg       0.71      0.68      0.67     12486
weighted avg       1.00      1.00      1.00     12486

Confusion Matrix:
[[    1     0    24]
 [    0 12142     0]
 [    4     0   315]]
```

Figure 4. 10 AdaBoost Imbalance

```
Accuracy (AdaBoost - Balanced): 0.9977574883869934

AdaBoost Classifier - Balanced:
Classification Report:
              precision    recall  f1-score   support

      absent       0.20      0.04      0.07        25
  not placed       1.00      1.00      1.00     12142
      placed       0.93      0.99      0.96       319

    accuracy                           1.00     12486
   macro avg       0.71      0.68      0.67     12486
weighted avg       1.00      1.00      1.00     12486
```

Figure 4. 11 AdaBoost Balanced

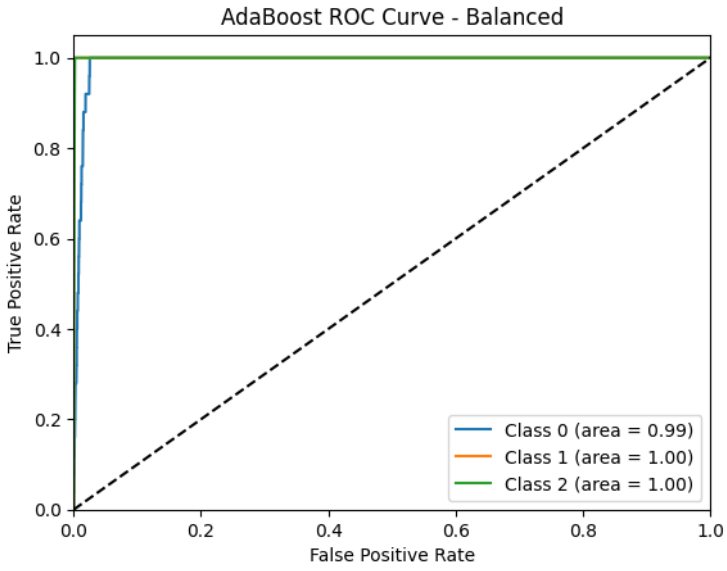


Figure 4. 12 ROC for AdaBoost

4.4.Logistic Regression

In this phase, Logistic Regression deals with an imbalanced dataset. Figure 4.13 represents the performance metric used to evaluate the effectiveness of applying the machine learning algorithm for Logistic Regression, which includes accuracy, precision, recall, and F1 score. Figure 4.15 represents the AUC Roc curve. These metrics will be used to analyze the model's performance

comprehensively. As we can see in the figure, the result was for the class absent, not placed, and the place was the accuracy (99), and for the absent class, the Precision 0, Recall 0, F1-Score 0. The second class, Not Placed the Precision 100, Recall 100, F1-Score 100. The third class Placed the Precision 93, Recall 98, and F1-Score 95.

On the other hand, the second phase of Logistic Regression deals with balanced datasets. Figure 4.14 represents the performance metric used to evaluate the effectiveness of applying the machine learning algorithm for Logistic Regression, which includes accuracy, precision, recall, and F1 score. Figure 4.15 represents the AUC Roc curve. These metrics will be used to analyze the model's performance comprehensively. As we can see in the figure, the result was for the class absent, not placed, and the place had an accuracy of 99, and for the absent class, the Precision was 12, Recall 63, and F1-Score 75. The second class, Not Placed the Precision 100, Recall 100, F1-Score 100. The third class Placed the Precision 90, Recall 66, and F1-Score 76. In summary, the RF model performance did not resolve the issue in the absent class.

```
Accuracy (Logistic Regression): 0.9975172192855999
Classification Report (Logistic Regression):
              precision    recall  f1-score   support

      absent       0.00      0.00      0.00        25
  not placed       1.00      1.00      1.00     12142
      placed       0.93      0.98      0.95       319

    accuracy                           1.00     12486
   macro avg       0.64      0.66      0.65     12486
weighted avg       1.00      1.00      1.00     12486
```

Figure 4. 13 Logistic Regression imbalance

```
Accuracy (Logistic Regression): 0.9975172192855999
Classification Report (Logistic Regression):
              precision    recall  f1-score   support

      absent       0.00      0.00      0.00        25
  not placed       1.00      1.00      1.00     12142
      placed       0.93      0.98      0.95       319

    accuracy                           1.00     12486
   macro avg       0.64      0.66      0.65     12486
weighted avg       1.00      1.00      1.00     12486
```

Figure 4. 14 Logistic Regression balance

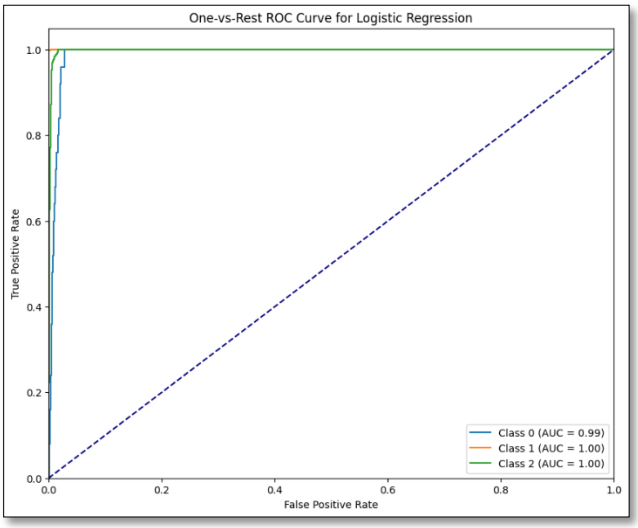


Figure 4. 15 ROC Logistic Regression

4.5.Gradient Boosting

Gradient Boosting in phase one deals with an imbalanced dataset. Figure 4.16 represents the performance metric used to evaluate the effectiveness of applying the machine learning algorithm for Gradient Boosting, which includes accuracy, precision, recall, and F1 score. Figure 4.18 represents the AUC Roc curve. These metrics will be used to analyze the model's performance comprehensively. As we can see in the figure, the result was for the class absent, not placed, and the place was the accuracy (99%), and for the absent class, the Precision 0%, Recall 0%, and F1-

Score 0%. The second class was Not Placed the precision 100%, recall 100%, and F1-Score 100%. The third class Placed the Precision at 93%, Recall at 98%, and F1-Score at 95%.

On the other hand, the other phase of Gradient Boosting deals with balanced datasets. Figure 4.17 represents the performance metric used to evaluate the effectiveness of applying the machine learning algorithm for Gradient Boosting, which includes accuracy, precision, recall, and F1 score. Figure 4.18 represents the AUC Roc curve. These metrics will be used to analyze the model's performance comprehensively. As we can see in the figure, the result was for the class absent, not placed, and the place had an accuracy of 99%, and for the Absent class, the Precision was 0%, Recall 0%, and F1-Score 0%. The second class is Not Placed the Precision 100%, Recall 100%, and F1-Score 100%. The third class Placed the Precision at 93%, Recall at 98%, and F1-Score at 95%. In summary, The Gradient Boosting classifier achieved high accuracy; its performance varied across different classes. Further optimization or exploration of various algorithms may be necessary to improve performance, especially for the minority classes.

```
Accuracy (Gradient Boosting): 0.9975172192855999

Gradient Boosting Classifier:
              precision    recall  f1-score   support

      absent       0.00      0.00      0.00        25
  not placed       1.00      1.00      1.00     12142
      placed       0.93      0.98      0.95       319

    accuracy                           1.00     12486
   macro avg       0.64      0.66      0.65     12486
weighted avg       1.00      1.00      1.00     12486

Confusion Matrix:
[[    0     0    25]
 [    0 12142     0]
 [    6     0   313]]
```

Figure 4. 16 Gradient boosting imbalanced

```
Accuracy (Gradient Boosting - Balanced): 0.9975172192855999

Gradient Boosting Classifier - Balanced:
Classification Report:
              precision    recall  f1-score   support

      absent       0.00      0.00      0.00        25
  not placed       1.00      1.00      1.00     12142
      placed       0.93      0.98      0.95       319

    accuracy                           1.00     12486
   macro avg       0.64      0.66      0.65     12486
weighted avg       1.00      1.00      1.00     12486


Confusion Matrix:
[[    0     0    25]
 [    0 12142     0]
 [    6     0   313]]
```

Figure 4. 17 Gradient boosting Balanced

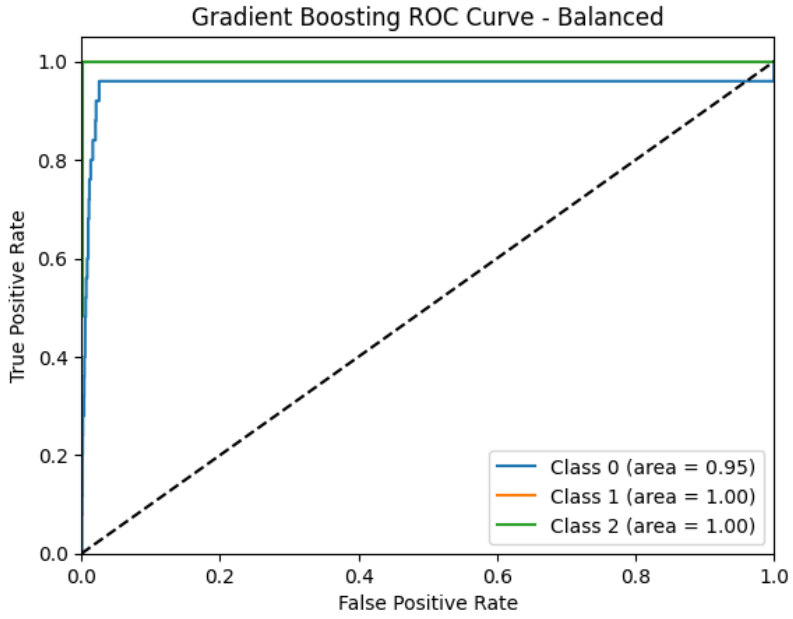


Figure 4. 18 ROC Gradient Boosting

4.6.Naive Bayes

In this phase of Naïve Bayes deals with an imbalanced dataset. Figure 4.19 represents the performance metric used to evaluate the effectiveness of applying the machine learning algorithm for Naïve Bayes, which includes accuracy, precision, recall, and F1 score. Figure 4.21 represents the AUC Roc curve. These metrics will be used to analyze the model's performance comprehensively. As we can see in the figure, the result was for the class absent, not placed, and the place was the accuracy (99), and for the absent class, the Precision 0, Recall 0, F1-Score 0. The second class, Not Placed the Precision 100, Recall 100, F1-Score 100. The third class Placed the Precision 93, Recall 93, and F1-Score 93.

On the other hand, the second phase of Naïve Bayes deals with balanced datasets. Figure 4.20 represents the performance metric used to evaluate the effectiveness of applying the machine learning algorithm for Naïve Bayes, which includes accuracy, precision, recall, and F1 score. Figure 4.21 represents the AUC Roc curve. These metrics will be used to analyze the model's performance comprehensively. As we can see in the figure, the result was for the class absent, not placed, and the place had an accuracy of 100, and for the absent class, the precision was 5, recall 12, and F1-Score 7. The second class, Not Placed the Precision 100, Recall 100, F1-Score 100. The third class Placed the Precision 93, Recall 93, and F1-Score 100. Overall, the Naive Bayes Classifier with balanced class weights performs well in identifying instances belonging to the "not placed" class but struggles with the "absent" class, likely due to imbalanced data.

	precision	recall	f1-score	support
absent	0.05	0.12	0.07	25
not placed	1.00	1.00	1.00	12142
placed	0.93	0.93	0.93	319
accuracy			0.99	12486
macro avg	0.66	0.68	0.67	12486
weighted avg	1.00	0.99	0.99	12486

Figure 4. 19 Naive Bayes imbalance

	precision	recall	f1-score	support
absent	0.05	0.12	0.07	25
not placed	1.00	1.00	1.00	12142
placed	0.93	0.93	0.93	319
accuracy			0.99	12486
macro avg	0.66	0.68	0.67	12486
weighted avg	1.00	0.99	0.99	12486

Figure 4. 20 Naive Bayes balance

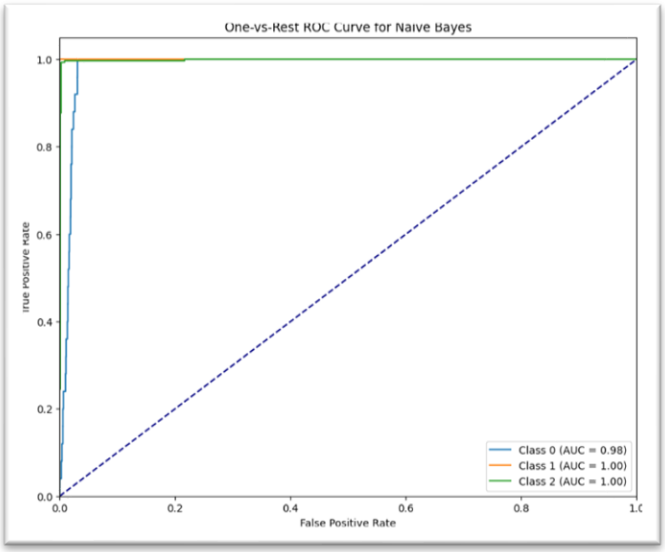


Figure 4. 21 ROC Naive Bayes

4.7.KNN Boost

KNN deals with an imbalanced dataset. Figure 4.21 represents the performance metric used to evaluate the effectiveness of applying the machine learning algorithm for KNN, which includes accuracy, precision, recall, and F1 score. Figure 4.23 represents the AUC Roc curve. These metrics will be used to analyze the model's performance comprehensively. As we can see in the figure, the result was for the class absent, not placed, and the place was the accuracy of 98%, and for the Absent class, the Precision was 0%, Recall 0%, and F1-Score 0%. The second class, Not Placed the Precision 98%, Recall 100%, and F1-Score 99%. The third class Placed the Precision at 72%, Recall at 45%, and F1-Score at 55%.

Then again, the second phase of KNN deals with balanced datasets. Figure 4.22 represents the performance metric used to evaluate the effectiveness of applying the machine learning algorithm for KNN, which includes accuracy, precision, recall, and F1 score. Figure 4.23 represents the AUC Roc curve. These metrics will be used to analyze the model's performance comprehensively. As

we can see in the figure, the result was for the class absent, not placed, and the place had an accuracy of 98%, and for the Absent class, the Precision was 0%, Recall 0%, and F1-Score 0%. The second class, Not Placed the Precision 98%, Recall 100%, and F1-Score 99%. The third class Placed the Precision at 72%, Recall at 4\\5%, and F1-Score at 55%. Overall, the Naive Bayes Classifier with balanced class weights performs well in identifying instances belonging to the "not placed" class but struggles with the "absent" class, likely due to imbalanced data.

```
Accuracy (KNN): 0.9805382027871216
Classification Report (KNN):
              precision    recall  f1-score   support

      absent       0.00      0.00      0.00        25
  not placed       0.98      1.00      0.99     12142
      placed       0.72      0.45      0.55       319

    accuracy                           0.98     12486
   macro avg       0.57      0.48      0.51     12486
weighted avg       0.98      0.98      0.98     12486
```

Figure 4. 22 KNN Imbalance

```
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.92      0.92       367
           1       1.00      1.00      1.00     14582
           2       0.18      0.21      0.19        34

    accuracy                           1.00     14983
   macro avg       0.70      0.71      0.70     14983
weighted avg       1.00      1.00      1.00     14983
```

Figure 4. 23 KNN Balanced

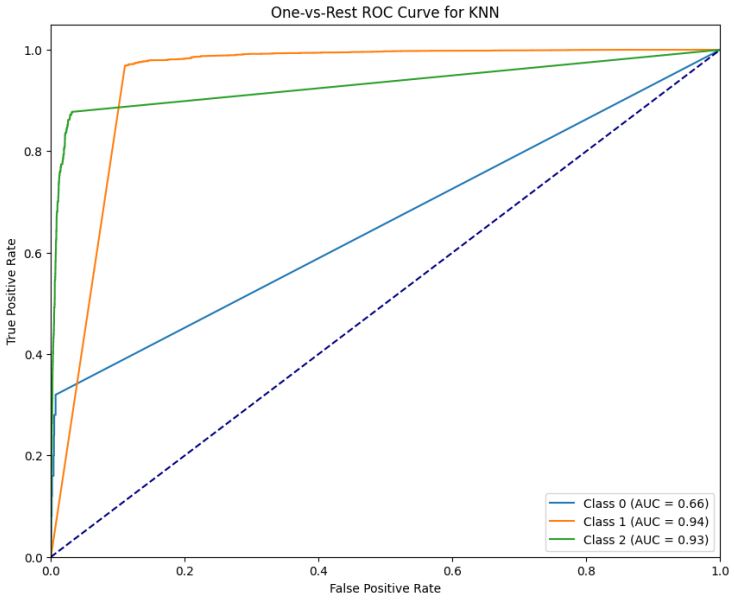


Figure 4. 24 ROC KNN

4.8.XGBoost

Another Model where use is XGBoost the figure 4.24 represents the performance metric used to evaluate the effectiveness of applying the machine learning algorithm for XGBoost, which includes accuracy, precision, recall, and F1 score. Figure 4.26 represents the AUC Roc curve. These metrics will be used to analyze the model's performance comprehensively. As we can see in the figure, the result was for the class absent, not placed, and the place was the accuracy of 98%, and for the Absent class, the Precision was 0%, Recall 0%, and F1-Score 0%. The second class, Not Placed the Precision 98%, Recall 100%, and F1-Score 99%. The third class Placed the Precision at 72%, Recall at 45%, and F1-Score at 55%.

Then again, the second phase of XGBoost deals with balanced datasets. Figure 4.25 represents the performance metric used to evaluate the effectiveness of applying the machine learning algorithm for XGBoost, which includes accuracy, precision, recall, and F1 score. Figure 4.26 represents the AUC Roc curve. These metrics will be used to analyze the model's performance comprehensively.

As we can see in the figure, the result was for the class absent, not placed, and the place had an accuracy of 98%, and for the Absent class, the Precision was 0%, Recall 0%, and F1-Score 0%. The second class, Not Placed the Precision 98%, Recall 100%, and F1-Score 99%. The third class Placed the Precision at 72%, Recall at 45%, and F1-Score at 55%. Overall, the Naive Bayes Classifier with balanced class weights performs well in identifying instances belonging to the "not placed" class but struggles with the "absent" class, likely due to imbalanced data.

	precision	recall	f1-score	support
absent	0.14	0.04	0.06	25
not placed	1.00	1.00	1.00	12142
placed	0.93	0.98	0.95	319
accuracy			1.00	12486
macro avg	0.69	0.67	0.67	12486
weighted avg	1.00	1.00	1.00	12486

Figure 4. 25 XGBoost Imbalance

Classification Report:

	precision	recall	f1-score	support
0	0.92	0.97	0.94	367
1	1.00	1.00	1.00	14582
2	0.20	0.09	0.12	34
accuracy			1.00	14983
macro avg	0.71	0.69	0.69	14983
weighted avg	1.00	1.00	1.00	14983

Figure 4. 26 XGBoost Balance

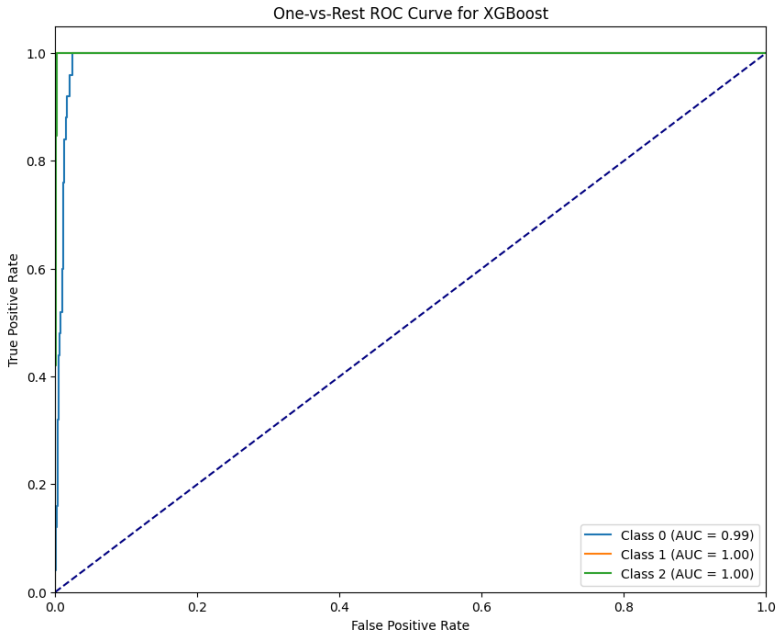


Figure 4. 27 ROC For XGBoost

4.9.Discussion of Results Comparison of Each model with Imbalanced and balanced Class

Selecting the model will depend on the specific characteristics of dataset and the problem we are trying to solve. In the figure 3.45 and table 3.5 below presents the comparison of each model with an implanted and balanced class and The Random Forest model achieved the highest accuracy of 0.998 in balanced and imbalanced class scenarios. The SVM model showed a significant drop in accuracy from 0.994 in the balanced class to 0.972 in the imbalanced class. The Logistic Regression model performed better in the imbalanced class scenario with an accuracy of 0.9975 compared to 0.9894 in the balanced class. The KNN model had the lowest accuracy, with 0.9810 in the balanced class and 0.9805382027871216 in the imbalanced class. The ensemble models, such as XGBoost, Gradient Boosting, and AdaBoost, maintained high accuracy in both the balanced and imbalanced class scenarios.

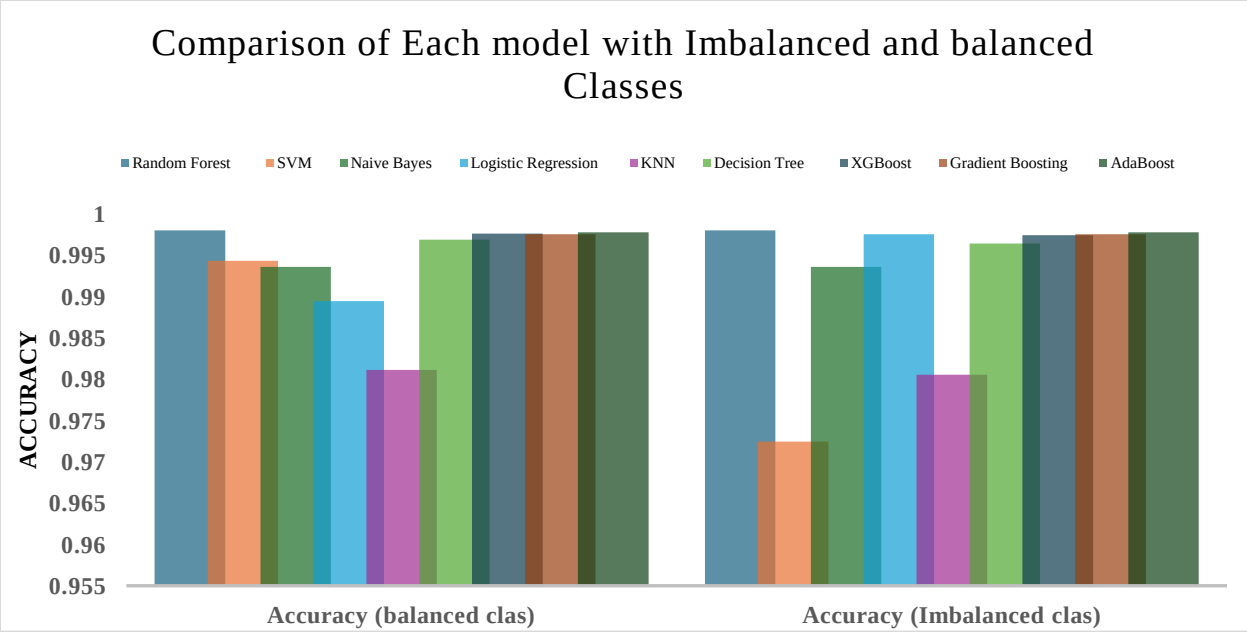


Figure 4. 28 Comparison of Each model with Imbalanced and balanced Classes

Table 4. 1 Comparison of Each model with Imbalanced and balanced Classes

Model	Accuracy (balanced class)	Accuracy (Imbalanced class)
Random Forest	v	0.997997757
SVM	0.994340177	0.972449143
Naive Bayes	0.993592824	0.993592824
Logistic Regression	0.98942816	0.997517219
KNN	0.981098831	0.980538203
Decision Tree	0.996876502	0.996395963
XGBoost	0.997597309	0.997410368
Gradient Boosting	0.997517219	0.997517219
AdaBoost	0.997757488	0.997757488

Table 4. 2 Best models hyperparameters for each module

Model	
Random Forest	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}
SVM	{'C': 0.27425083650459836, 'degree': 3, 'gamma': 0.26685430556951095, 'kernel': 'rbf'}
KNN	{'metric': 'euclidean', 'n_neighbors': 3, 'weights': 'uniform'}
Decision Tree	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2}
XGBoost	Best Parameters: {'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 150, 'subsample': 0.6}
Gradient Boosting	{'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 100
AdaBoost	{'learning_rate': 0.01, 'n_estimators': 50}

4.10. Techniques of Linear Regression

Liner regression techniques have been used the figures present that All models perform exceptionally well, with shallow MSE and high R-squared values. The differences between the training and testing metrics (MSE Difference and R-squared Difference) are also minimal for all models, indicating minimal overfitting. Ridge Regression has the most minor differences, followed closely by Elastic Net Regression. However, all models demonstrate strong performance and low risk of overfitting.

Figure 4.29 shows linear regression, which indicates that the OLS linear regression model is performing exceptionally well, with predicted values nearly identical to the actual values. Moreover, the Mean Squared Error (MSE): .6147766655527175e-28, R-squared (R2): 1.0, MSE Difference (Overfitting Check): 1.4017343970602236e-30 R-squared Difference (Overfitting Check): 0.0.

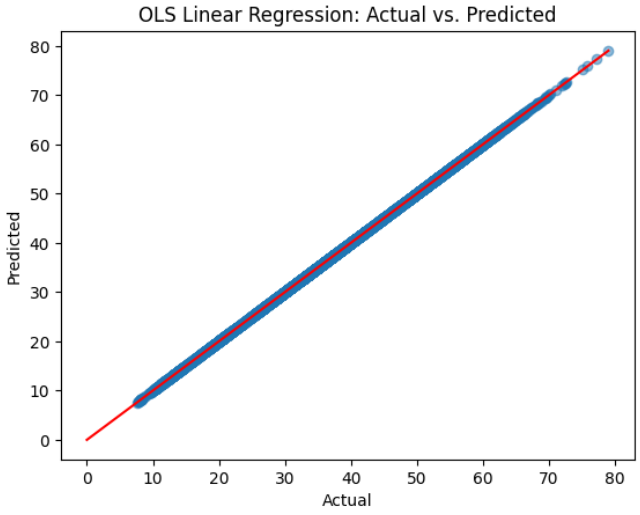


Figure 4. 29 Linear Regression

Figure 4.30 shows the Lasso regression, which compares the actual and predicted values. Moreover, the Lasso Regression model is practical and efficient, particularly in maintaining model simplicity while delivering accurate predictions. The result distributed Mean Squared Error (MSE): 0.008098804397854981, R-squared (R2): 0.9999356646014774 MSE Difference (Overfitting Check): -7.787682648354062e-05, R-squared Difference (Overfitting Check): 1.1941736488552124e-10.

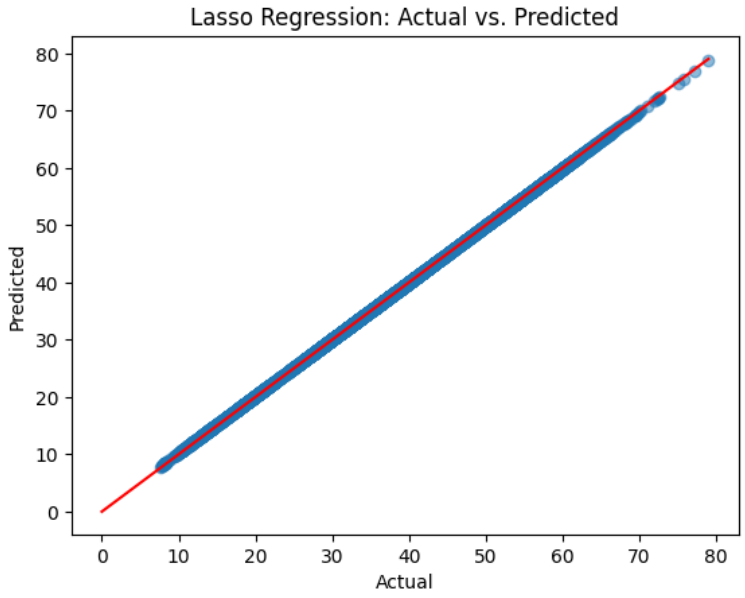


Figure 4. 30 Lasso Regression

Figure 3.31 shows elastic Net regression. The Elastic Net Regression model demonstrates excellent predictive accuracy, as evidenced by the close alignment of the predicted values with the actual values along the red line. Elastic Net Regression Results Mean Squared Error (MSE): 0.008034233174314736, R-squared (R2): 0.999936177543289, MSE Difference (Overfitting

Check): -7.725591977627254e-05 and R-squared Difference (Overfitting Check): 1.184652376196027e-10.

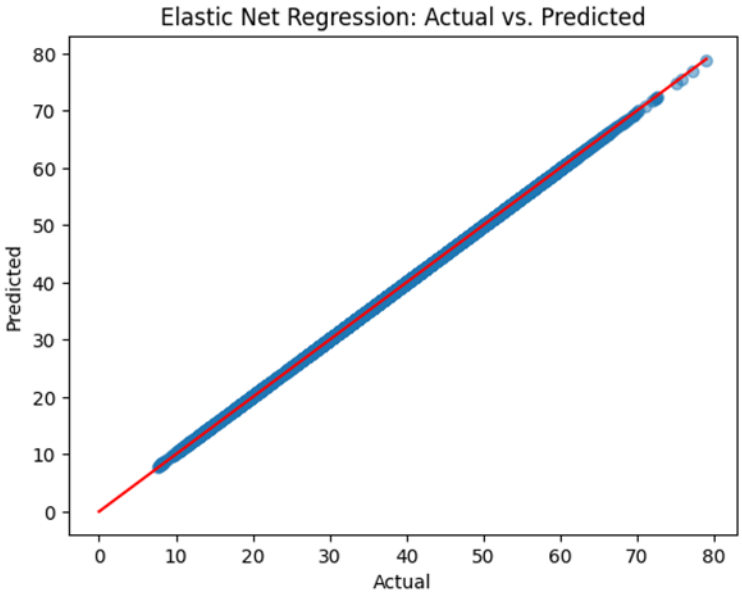


Figure 4. 31 Elastic Net Regression

Figure 4.32 presents the important features of the Random Forest classifier. Employment Reason is the most critical feature; however, other features contribute to varying degrees, but none come close to the importance of Employment Reason. The model's decisions or predictions are heavily driven by this single feature, which could imply that it captures a significant portion of the variance in the target variable.

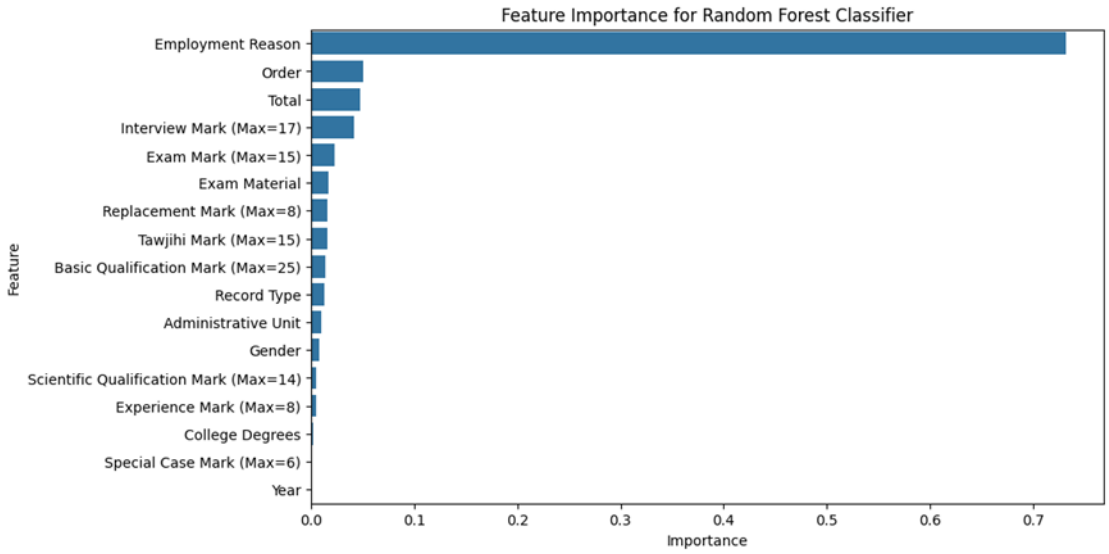


Figure 4. 32 Feature importance for random forest classifier.

4.11. Application

After the selection of the model, we decide to establish an application for the users that named Employment Status Prediction Streamlit App, and to access the user should have the minimum requirement such as Python 3.7 or later , Streamlit: For building the web interface (`pip install streamlit`), Pandas: For data manipulation (`pip install pandas`), Joblib: For loading the machine learning model (`pip install joblib`), Scikit-learn: Required for the model and label encoders (`pip install scikit-learn`).

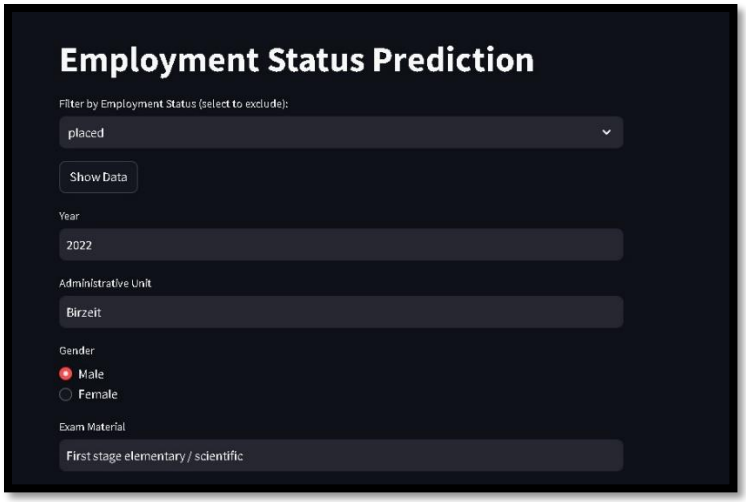


Figure 4. 33 Tthe front page of the model application screen.

Figure 4. 34 The system input screen

Figure 4. 35 The system input screen.

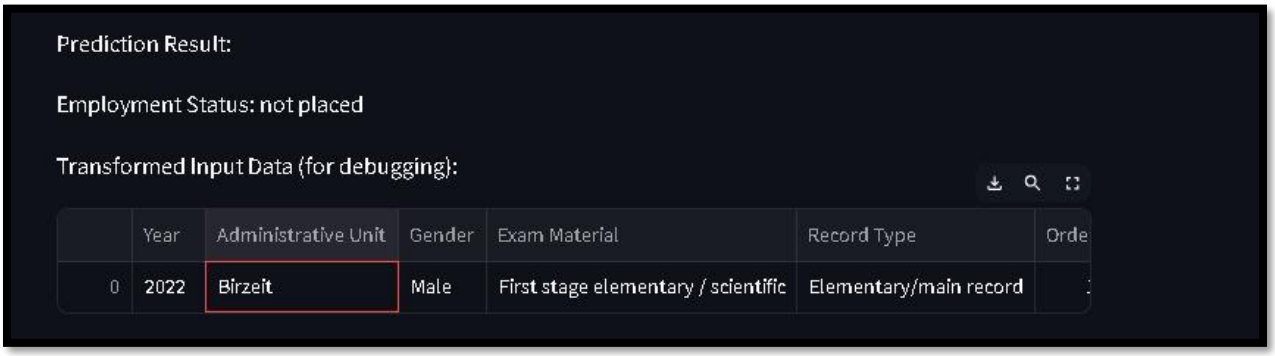


Figure 4. 36 The system Prediction Result screen.

Also, regarding that an instruction for the users how to use the system attached in attachment number 1.

4.12. Challenges and Limitations

Throughout my research, I met various significant challenges that made the process much more difficult. One of the primary challenges was an imbalanced dataset, which resulted in skewed results for the majority class. This made it difficult to produce accurate and fair estimates because the model tended to overlook the minority population. The large number of columns and numerical data made scaling and normalization difficult. I struggled with feature engineering and had to spend a lot of effort cleaning and changing data to make it usable. Due to the complexity of this approach, I sought professional aid in refining the data.

Another significant constraint was the difficulty of collecting a larger dataset for better generalization. Obtaining enough relevant data proved to be a tough and time-consuming effort, especially because I had to rely on a government ministry to provide it. I spent months attempting to obtain the necessary data, and the extended wait caused unnecessary delays in my project.

In terms of technological constraints, the model implementation required significant computational resources, creating hardware and processing power challenges. The complexity of the models I was working with, together with the size of the dataset, required me to optimize the code and experiment with several approaches to reach the desired results.

Furthermore, I observed a lack of previous study in this area, particularly in the public sector and teacher recruitment in Arabic-speaking countries. Due to a dearth of relevant research, I struggled to find standards or techniques to guide my work. Without similar case studies or reference points, I had to develop and adapt existing procedures to match the unique requirements of my project.

Chapter Five

Conclusion

5.1. Conclusions

Machine learning applications have become essential in our lives due to their value in enhancing life tasks. Our study takes the responsibility of exploring the role of ML in Selecting the Appropriate Teacher at the Ministry of Education Using Machine and Deep Learning Techniques. The comparative analysis of machine learning models on balanced and imbalanced datasets underscores the importance of class distribution in model performance. While models like Random Forests and Decision Trees show robust overall performance, they struggle with minority class prediction, particularly in imbalanced datasets. SVM, Logistic Regression, and ensemble models like Gradient Boosting and AdaBoost also exhibit challenges with minority classes, though their performance improves with balanced datasets. Naïve Bayes and KNN models perform the worst, indicating their unsuitability for this classification task.

However, balancing the dataset generally improves the performance across all models, but it is not a panacea. Findings suggest that for tasks involving significant class imbalance, additional strategies, such as data resampling, the use of different performance metrics, or specialized algorithms designed to handle imbalanced data, may be necessary to achieve more accurate and reliable predictions. In the end, the conclusion is that while some models, like Random Forests and Decision Trees, provide robust performance in general, careful consideration of dataset characteristics, particularly class distribution, is crucial in selecting the appropriate model for classification tasks.

References

Palestine Ministry of Education, “State of Palestine Ministry of Education LABOR MANAGEMENT PROCEDURES For: Supporting an Education Reform Agenda for Improving Teaching, Assessment and Career Pathways Project-SERATAC,” Jun. 2022.

Databricks.com, “What are Machine Learning Models?” Accessed: Sep. 01, 2024. [Online]. Available: https://www.databricks.com/glossary/machine-learning-models

“THE WORLD BANK-SOCIAL IMPACT ASSESSMENT (SIA) FOR EDUCATIONAL MULTIPHASE PROGRAMMATIC APPROACH (MPA) FOR WEST BANK & GAZA PROJECT: JUNE”, Accessed: Oct. 20, 2024. [Online]. Available: https://ontheworldmap.com/palestine/west-bank

M. Yağcı, “Educational data mining: prediction of students’ academic performance using machine learning algorithms,” Smart Learning Environments, vol. 9, no. 1, Dec. 2022, doi: 10.1186/S40561-022-00192-Z.

Seldon, “Machine Learning Regression Explained - Seldon.” Accessed: Sep. 01, 2024. [Online]. Available: https://www.seldon.io/machine-learning-regression-explained

Javatpoint, “Data Mining vs Machine Learning - Javatpoint.” Accessed: Aug. 01, 2024. [Online]. Available: https://www.javatpoint.com/data-mining-vs-machine-learning

S. Ayyat, Md. F. Abdullah, B. Binti, and A. Halim, “Challenges of Human Resource Management in Palestine,” The Journal of Management Theory and Practice (JMTP), vol. Volume-2, pp. 2716–7089, Nov. 2021, doi: 10.37231/jmtp.2021.2.4.122.

S. B. Kulkarni and X. Che, “Intelligent Software Tools for Recruiting,” Journal of International Technology and Information Management, vol. 28, no. 2, pp. 2–16, Jul. 2019, doi: 10.58729/1941-6679.1398.

P. Pampouktsi, S. Avdimiotis, M. Maragoudakis, and M. Avlonitis, “Applied Machine Learning Techniques on Selection and Positioning of Human Resources in the Public Sector,” Open Journal of Business and Management, vol. 09, no. 02, 2021, doi: 10.4236/ojbm.2021.92030.

X.-J. Yao, X.-F. Shi, Y.-P. Wang, G.-Y. Gan, and B.-Y. Tang, “The mechanical properties of high entropy (-like) alloy Wx (TaTiVCr) 1-x via first-principles calculations,” Fusion Engineering and Design, vol. 137, pp. 35–42, 2018.

E. Kariri, H. Louati, A. Louati, and F. Masmoudi, “Exploring the Advancements and Future Research Directions of Artificial Neural Networks: A Text Mining Approach,” Applied Sciences, vol. 13, p. 3186, Mar. 2023, doi: 10.3390/app13053186.

J. Wang, Y. Zhao, B. Perumal, and P. Selvaraj, “Managerial decision support system using an integrated model of AI and big data analytics,” Ann Oper Res, vol. 326, pp. 1–18, Jan. 2022, doi: 10.1007/s10479-021-04359-8.

A. Wael, “The Power of Artificial Intelligence in Recruitment: An Analytical Review of Current AI-Based Recruitment Strategies,” International Journal of Professional Business Review, vol. 8, p. e02089, Jun. 2023, doi: 10.26668/businessreview/2023.v8i6.2089.

M. N. Freire and L. N. de Castro, “e-Recruitment recommender systems: a systematic review,” Knowl Inf Syst, vol. 63, no. 1, pp. 1–20, Jan. 2021, doi: 10.1007/s10115-020-01522-8.

J. Nagaria and S. Velan, “Utilizing Exploratory Data Analysis for the Prediction of Campus Placement for Educational Institutions,” 2020.

D. Kumar, C. Verma, P. Singh, M. Raboaca, R. Felseghi, and K. Ghafoor, “Computational Statistics and Machine Learning Techniques for Effective Decision Making on Student’s Employment for Real-Time,” vol. 9, pp. 1–30, May 2021, doi: 10.3390/math9111166.

B. Şen, E. Uçar, and D. Delen, “Predicting and analyzing secondary education placement-test scores: A data mining approach,” Expert Syst Appl, vol. 39, no. 10, pp. 9468–9476, Aug. 2012, doi: 10.1016/j.eswa.2012.02.112.

A. Ojha, U. Pattnaik, and S. R. Sankar, “Data analytics on placement data in a South Asian University,” in 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017, pp. 2438–2442. doi: 10.1109/ICECDS.2017.8389888.

A. Giri, M. V. V Bhagavath, B. Pruthvi, and N. Dubey, “A Placement Prediction System using k-nearest neighbors classifier,” in 2016 Second International Conference on Cognitive Computing and Information Processing (CCIP), 2016, pp. 1–4. doi: 10.1109/CCIP.2016.7802883.

K. Pruthi and P. Bhatia, “Application of Data Mining in predicting placement of students,” in 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), IEEE, 2015, pp. 528–533.

D. Muhajir, M. Akbar, A. Bagaskara, and R. Vinarti, “Improving classification algorithm on education dataset using hyperparameter tuning,” Procedia Comput Sci, vol. 197, pp. 538–544, 2022, doi: https://doi.org/10.1016/j.procs.2021.12.171.

S. Dutta and S. K. Bandyopadhyay, “Forecasting of Campus Placement for Students Using Ensemble Voting Classifier,” Asian Journal of Research in Computer Science, pp. 1–12, May 2020, doi: 10.9734/ajrcos/2020/v5i430138.

Y. F. Yeh, M. H. Liu, Y. S. Hsu, and Y. H. Tseng, “From Student Teachers to Hired Educators: Exploring the Attributes of Successful Teacher Certification and Placement Through Data Mining,” in Handbook of Research on Teacher Education: Innovations and Practices in Asia, Springer Nature, 2022, pp. 589–606. doi: 10.1007/978-981-16-9785-2_29.

S. Dutta and S. K. Bandyopadhyay, “Fake job recruitment detection using machine learning approach,” International Journal of Engineering Trends and Technology, vol. 68, no. 4, pp. 48–53, Apr. 2020, doi: 10.14445/22315381/IJETT-V68I4P209S.

K. Sreenivasa Rao, N. Swapna, and P. P. Kumar, “Educational data mining for student placement prediction using machine learning algorithms,” 2018. [Online]. Available: www.sciencepubco.com/index.php/IJET

L. Sathish and T. Sudha Rani, “Student Placement Prediction Using Machine Learning Models (KNN, SVM, RF, Logistic Regression),” in Advances in Transdisciplinary Engineering, IOS Press BV, Jan. 2023, pp. 289–295. doi: 10.3233/ATDE221271.

C. D. Casuat, “Predicting Students’ Employability using Support Vector Machine: A SMOTE-Optimized Machine Learning System,” International Journal of Emerging Trends in Engineering Research, vol. 8, no. 5, pp. 2101–2106, May 2020, doi: 10.30534/ijeter/2020/102852020.

B.-B. Cui, “Design and Implementation of Movie Recommendation System Based on Knn Collaborative Filtering Algorithm,” 2020.

A. Nawrocka, A. Kot, and M. Nawrocki, “Application of machine learning in recommendation systems,” in 2018 19th International carpathian control conference (ICCC), IEEE, 2018, pp. 328–331.

S. S. Khanal, P. W. C. Prasad, A. Alsadoon, and A. Maag, “A systematic review: machine learning based recommendation systems for e-learning,” Educ Inf Technol (Dordr), vol. 25, no. 4, pp. 2635–2664, 2020.

R. Tsiskaridze, K. Reinhold, and M. Jarvis, “Innovating HRM recruitment: a comprehensive review of AI deployment,” Marketing i menedžment innovacij, vol. 14, no. 4, pp. 239–254, 2023.

P. K. Roy, S. S. Chowdhary, and R. Bhatia, “A Machine Learning approach for automation of Resume Recommendation system,” Procedia Comput Sci, vol. 167, pp. 2318–2327, 2020.

C. Daryani, G. S. Chhabra, H. Patel, I. K. Chhabra, and R. Patel, “An automated resume screening system using natural language processing and similarity,” ETHICS AND INFORMATION TECHNOLOGY [Internet]. VOLKSON PRESS, pp. 99–103, 2020.

V. Kakulapati, K. K. Chaitanya, K. V. G. Chaitanya, and P. Akshay, “Predictive analytics of HR-A machine learning approach,” Journal of Statistics and Management Systems, vol. 23, no. 6, pp. 959–969, 2020.

H. Zhu, “Research on human resource recommendation algorithm based on machine learning,” Sci Program, vol. 2021, no. 1, p. 8387277, 2021.

E. Faliagka, K. Ramantas, A. Tsakalidis, and G. Tzimas, “Application of machine learning algorithms to an online recruitment system,” in Proc. International Conference on Internet and Web Applications and Services, 2012, pp. 215–220.

H. Wang, G. Liang, and X. Zhang, “Feature regularization and deep learning for human resource recommendation,” IEEE Access, vol. 6, pp. 39415–39421, 2018.

A. Chalfin et al., “Productivity and selection of human capital with machine learning,” American Economic Review, vol. 106, no. 5, pp. 124–127, 2016.

C. Ovalle et al., “Application of neural networks in the teacher selection process,” Procedia Comput Sci, vol. 218, pp. 1132–1143, 2023.

S. Shet and B. Nair, “Quality of hire: expanding the multi-level fit employee selection using machine learning,” International Journal of Organizational Analysis, vol. 31, no. 6, pp. 2103–2117, 2023.

P. Pampouktsi, S. Avdimiotis, M. Maragoudakis, and M. Avlonitis, “Applied machine learning techniques on selection and positioning of human resources in the public sector,” Open Journal of Business and Management, vol. 9, no. 2, pp. 536–556, 2021.

K. C. Rojas Romero et al., “Smart system model for the recruitment of teachers,” 2022.

B. A. Jacob, “The power of teacher selection to improve education,” Evidence Speaks Reports, vol. 1, no. 12, pp. 1–7, 2016.

A. J. Hobson, P. Ashby, J. McIntyre, and A. Malderez, “International approaches to teacher selection and recruitment,” 2010.

B. Yu, N. Rezsonya, and J. Drew, “Teacher retention: Identifying areas to improve teacher recruitment and retention in North Carolina public schools,” SMU Data Science Review, vol. 2, no. 2, p. 10, 2019.

G. Sudha, K. K. Sasipriya, D. Nivethitha, and S. Saranya, “Personality prediction through CV analysis using machine learning algorithms for automated e-recruitment process,” in 2021 4th international conference on computing and communications technologies (ICCCT), IEEE, 2021, pp. 617–622.

C. G. M. Garces, “A AI based Model for Achieving High Reliability Faculty Performance Using Various Machine Learning Algorithms,” International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 11s, pp. 605–618, Jan. 2024, [Online]. Available: https://ijisae.org/index.php/IJISAE/article/view/4481

S. Bian et al., “Learning to Match Jobs with Resumes from Sparse Interaction Data using Multi-View Co-Teaching Network,” in International Conference on Information and Knowledge Management, Proceedings, Association for Computing Machinery, Oct. 2020, pp. 65–74. doi: 10.1145/3340531.3411929.

G. Varoquaux, O. Colliot, and G. Varoquaux, “Evaluating machine learning models and their diagnostic value,” Jun. 2023, Accessed: Sep. 01, 2024. [Online]. Available: https://hal.science/hal-03682454v4

“Confusion Matrix in Machine learning, Scikit-learn in Python.” Accessed: Sep. 01, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/

Md. Mubasir, “Don’t Overfit! II — How to avoid Overfitting in your Machine Learning and Deep Learning Models | by Md. Mubasir | Towards Data Science,” Towards Data Science. Accessed: Aug. 16, 2024. [Online]. Available: https://towardsdatascience.com/dont-overfit-ii-how-to-avoid-overfitting-in-your-machine-learning-and-deep-learning-models-2ff903f4b36a

A. Mohammed Salisu, “COMBINE HOLTS WINTER AND SUPPORT VECTOR MACHINES IN FORECASTING TIME SERIS,” 2013.

Sandeep Kumar, “svm Archives - AITUDE,” aitude.com. Accessed: Aug. 17, 2024. [Online]. Available: https://www.aitude.com/tag/svm/

A. Saini, “Support vector machine (svm): A complete guide for beginners,” analyticsvidhya. com, 2021.

A. Géron, Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. “ O’Reilly Media, Inc.,” 2022.

M. Al-khassaweneh, M. Bronakowski, and E. Al-Sharoa, “Multivariate and Dimensionality-Reduction-Based Machine Learning Techniques for Tumor Classification of RNA-Seq Data,” Applied Sciences 2023, Vol. 13, Page 12801, vol. 13, no. 23, p. 12801, Nov. 2023, doi: 10.3390/APP132312801.

“Decision Tree - GeeksforGeeks.” Accessed: Aug. 17, 2024. [Online]. Available: https://www.geeksforgeeks.org/decision-tree/

B. Farnham, S. Tokyo, B. Boston, F. Sebastopol, and T. Beijing, “Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems SECOND EDITION,” 2021.

“Ensemble Methods for Machine Learning - Gautam Kunapuli - Google Books.” Accessed: Aug. 18, 2024. [Online]. Available: https://books.google.ps/books?id=wXGazgEACAAJ&printsec=frontcover#v=onepage&q&f=false

Serafeim Loukas, “How Random Forests & Decision Trees Decide: Simply Explained With An Example In Python | by Serafeim Loukas, PhD | Towards Data Science.” Accessed: Aug. 18, 2024. [Online]. Available: https://towardsdatascience.com/how-do-random-forests-decision-trees-decide-simply-explained-with-an-example-in-python-6737eb183604

P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” Mach Learn, vol. 63, pp. 3–42, 2006.

“Top 10 Interview Questions on Gradient Boosting Algorithms -.” Accessed: Aug. 18, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2022/11/top-10-interview-questions-on-gradient-boosting/

D. Sharma, “A Machine Learning Methodology for Classifying Chronic Kidney Diseases,” 2023.

D. Borg, M. Suetake, and D. Brandão, “A neural network developed in a foundation fieldbus environment to calculate flow rates for compressible fluid,” Flow Measurement and Instrumentation, vol. 40, pp. 142–148, 2014, doi: 10.1016/j.flowmeasinst.2014.09.007.

S. Yin and J. Yin, “Tuning kernel parameters for SVM based on expected square distance ratio,” Inf Sci (N Y), vol. 370–371, pp. 92–102, 2016, doi: https://doi.org/10.1016/j.ins.2016.07.047.

“Decision Tree Hyperparameter ☑ ☑.” Accessed: Sep. 01, 2024. [Online]. Available: https://www.kaggle.com/code/avadhutvarvatkar/decision-tree-hyperparameter

Hyperparameters of Random Forest Classifier - GeeksforGeeks.” Accessed: Sep. 01, 2024. [Online]. Available: https://www.geeksforgeeks.org/hyperparameters-of-random-forest-classifier/

“XGBoost Parameters — xgboost 2.1.1 documentation.” Accessed: Sep. 01, 2024. [Online]. Available: https://xgboost.readthedocs.io/en/stable/parameter.html

"Tune Learning Rate for Gradient Boosting with XGBoost in Python - MachineLearningMastery.com." Accessed: Sep. 01, 2024. [Online]. Available: https://machinelearningmastery.com/tune-learning-rate-for-gradient-boosting-with-xgboost-in-python/

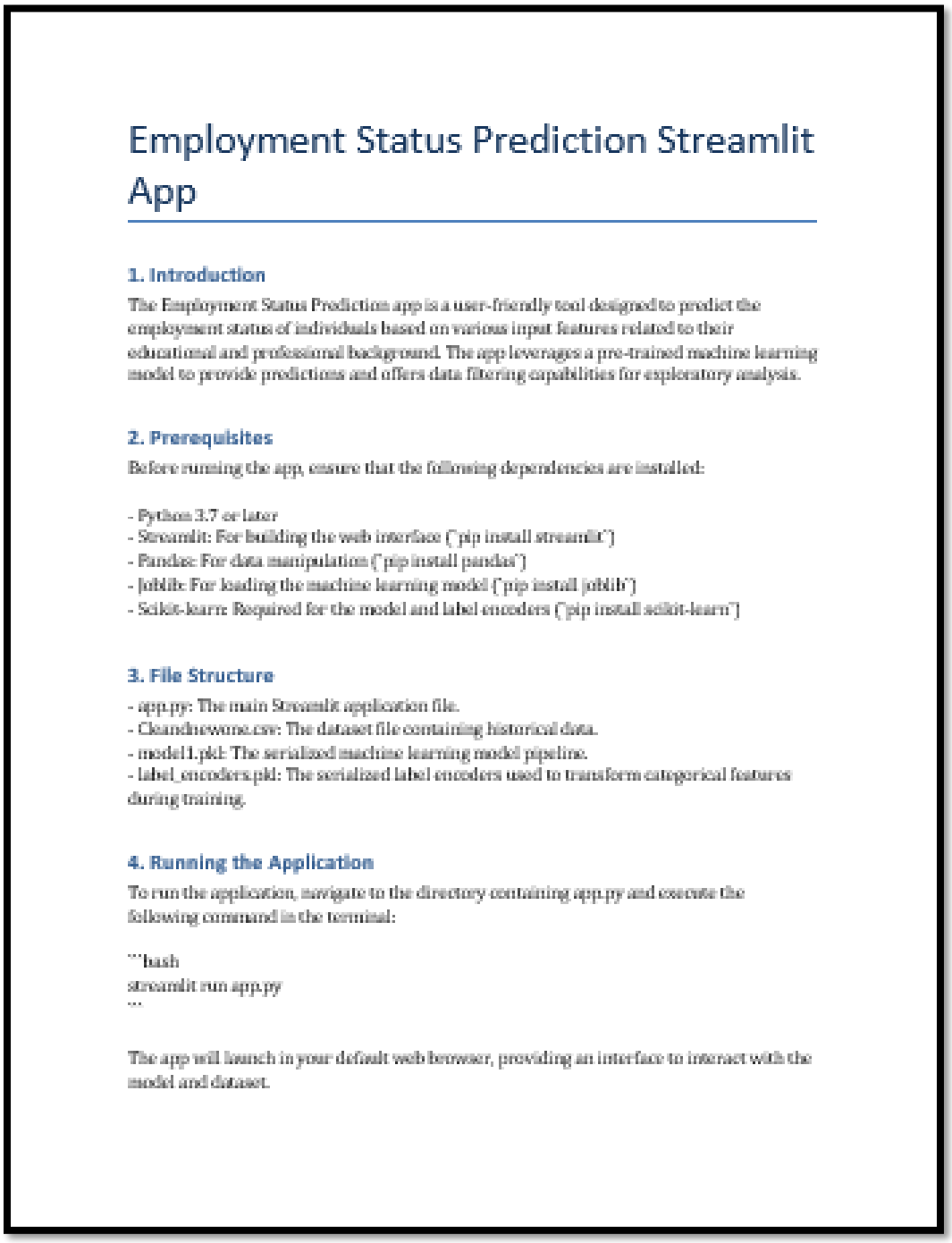
P. K. Roy, S. S. Chowdhary, and R. Bhatia, "A Machine Learning approach for automation of Resume Recommendation system," Procedia Comput Sci, vol. 167, pp. 2318–2327, 2020.

P. Dangeti, Statistics for Machine Learning. Packt Publishing Ltd, 2017.

P. Dangeti, Statistics for Machine Learning. Packt Publishing Ltd, 2017.

Appendices

Appendix 1 User Instruction



Employment Status Prediction Streamlit App

1. Introduction

The Employment Status Prediction app is a user-friendly tool designed to predict the employment status of individuals based on various input features related to their educational and professional background. The app leverages a pre-trained machine learning model to provide predictions and offers data filtering capabilities for exploratory analysis.

2. Prerequisites

Before running the app, ensure that the following dependencies are installed:

- Python 3.7 or later
- Streamlit: For building the web interface (`pip install streamlit`)
- Pandas: For data manipulation (`pip install pandas`)
- Joblib: For loading the machine learning model (`pip install joblib`)
- Scikit-learn: Required for the model and label encoders (`pip install scikit-learn`)

3. File Structure

- app.py: The main Streamlit application file.
- Cleandnewone.csv: The dataset file containing historical data.
- model1.pkl: The serialized machine learning model pipeline.
- label_encoders.pkl: The serialized label encoders used to transform categorical features during training.

4. Running the Application

To run the application, navigate to the directory containing app.py and execute the following command in the terminal:

```bash
streamlit run app.py
```

The app will launch in your default web browser, providing an interface to interact with the model and dataset.

5. Application Features

5.1. Data Filtering

Purpose: Allows users to explore the dataset by excluding a specific employment status.

Input: A dropdown menu allows users to select an "Employment Status" to exclude from the dataset.

Output: The filtered dataset is displayed in a table format when the user clicks the "Show Data" button.

5.2. Employment Status Prediction

Purpose: Predict the employment status of an individual based on user inputs.

Input: The app provides a form with the following fields:
- Year: Numeric input (2020-2023).
- Administrative Unit: Text input (e.g., "Birzeit").
- Gender: Radio button selection (Male or Female).
- Exam Material: Text input (e.g., "First stage elementary / scientific").
- Record Type: Text input (e.g., "Elementary/main record").
- Order: Numeric input (1-1000).
- Employment Status: Radio button selection ("placed" or "at disposal").
- Employment Reason: Text input (e.g., "at disposal").
- College Degree: Radio button selection (Bachelor's or Master's).
- Various Marks: Sliders for "Basic Qualification Mark", "Tawjihi Mark", "Replacement Mark", "Experience Mark", "Scientific Qualification Mark", "Special Case Mark", "Exam Mark", "Interview Mark", and "Total".

Output: The predicted employment status is displayed after clicking the "Predict" button.

5.3. Model Integration

Model Loading: The app loads the pre-trained model (model1.pkl) and label encoders (label_encoders.pkl) using the joblib library.

Prediction: The app processes the input data, applies necessary transformations using the label encoders, and passes the data to the model for prediction.

6. Technical Details

6.1. Data Preprocessing

Label Encoding: Categorical variables are encoded using pre-fitted label encoders (label_encoders.pkl).

Model Pipeline: The model pipeline (model1.pkl) includes both preprocessing steps and the predictive model.

6.2. Error Handling

The app includes basic error handling for missing or invalid inputs. If the model fails to predict due to unexpected input, an error message is displayed to the user.

7. Testing the Application

7.1. Test Cases

Here are some test cases you can use to ensure the functionality of the app:

1. **Test Case 1: Load Application**
   - **Action**: Run 'streamlit run app.py'.
   - **Expected Result**: The application loads successfully, displaying the title "Employment Status Prediction" and providing input fields.

2. **Test Case 2: Filter Dataset**
   - **Action**: Select "placed" from the dropdown and click "Show Data".
   - **Expected Result**: The dataset is displayed, excluding all rows where the employment status is "placed".

3. **Test Case 3: Predict Employment Status**
   - **Action**: Fill in all the input fields and click "Predict".
   - **Expected Result**: The predicted employment status is displayed below the input form.

4. **Test Case 4: Handle Missing Input**
   - **Action**: Leave a required field (e.g., "Year") empty and click "Predict".
   - **Expected Result**: The app displays an error message prompting the user to fill in the missing input.

5. **Test Case 5: Validate Input Types**
   - **Action**: Enter text in a numeric field (e.g., enter "abc" in "Year") and click "Predict".
   - **Expected Result**: The app should either prevent the input or show an error message indicating an invalid input type.

6. **Test Case 6: Model Prediction Accuracy**
   - **Action**: Use known data from Cleandrowcsv.csv to input values and compare the prediction with the actual employment status in the dataset.
   - **Expected Result**: The predicted employment status should match the actual status for a majority of the cases, reflecting the model's accuracy.

8. Future Improvements

Model Explainability: Integrate SHAP or LIME to provide users with explanations for predictions.

Feedback Mechanism: Allow users to provide feedback on the prediction accuracy to further train and improve the model.

Input Validation: Implement stricter validation rules to ensure that inputs fall within expected ranges.

9. Conclusion

The Employment Status Prediction app provides a powerful tool for predicting employment status based on a user's input. With a user-friendly interface and integration of a machine learning model, it enables users to explore the dataset and generate predictions quickly. The provided test cases ensure the app's functionality is reliable and robust.

If you encounter any issues or have suggestions for further improvement, please feel free to contribute or report them.

Appendix 2 Recruitment Request the Mistry of Education For 2022/2023 Teaching Year.

بسم الله الرحمن الرحيم

State of Palestine
Ministry of Education
Assistant Deputy Minister for Education Affairs

دولة فلسطين
وزارة التربية والتعليم
الوكيل المساعد للشؤون التعليمية

الرقم: و ت / ٣٠ د/ 28829
التاريخ: 2021/11/11م.

السادة المديرون العامون لمديريات التربية والتعليم المحترمون،

تحية طيبة وبعد،

الموضوع: استقطاب وتنقلات الموارد البشرية التعليمية والمساندة

للعام الدراسي 2023/2022م

أهديكم أطيب التحيات؛ وبعد مطالعة توصيات اللجان المختصة ومناقشة ملحوظات الميدان التربوي، وإيجاد رؤيا متشابهة من خلال تطبيق معايير الحوكمة الرشيدة، ووضع أسس واضحة تساعد على استقطاب واختيار الموظفين الأكثر كفاءة، وإرساء قيم الديمقراطية والعدالة، والمساواة في الفرص، والشفافية التي تضمن النزاهة على قاعدة المساءلة والمحاسبة وفق مقومات الجودة والنوعية.

وعلى ما تقدم ارفق اليكم التعليمات الخاصة بهذا المجال في الوثائق الآتية:

- أسس اختيار الوظائف التعليمية والمساندة في المدارس للعام الدراسي 2023/2022م.
- أسس النقل والتعيين وأولوياتهما للعام الدراسي 2023/2022م.
- التخصصات المعتمدة للمسار الأكاديمي والمهني والمساند للعام الدراسي 2023/2022م.

مع الاحترام والمودة،،،

أ.ثروت زيد

الوكيل المساعد للشؤون التعليمية

نسخة/ معالي وزير التربية والتعليم حفظه الله.
/ عطوفة وكيل وزارة التربية والتعليم المحترم.
/ الإخوة الوكلاء المساعدين المحترمين.

الملخص

يواجه توظيف المعلمين لوزارة التربية والتعليم في فلسطين تغيير كبيرا، فهناك عدد محدود من الوظائف الشاغرة وعدد أكبر من المتقدمين، ومع ذلك، فإن الفرز بين توافر الفرص وعرض المرشحين المؤهلين قد يخلق صعوبة في اختيار المرشحين المناسبين، ويؤدي إلى عدم كفاءة في عملية التوظيف، لذا فإن معالجة هذه القضية أمر بالغ الأهمية لضمان وجود الشخص المناسب في المكان المناسب.

وعلاوة على ذلك، تقدم وزارة التربية والتعليم الفلسطينية عددًا قليلاً من الوظائف للمعلمين كل عام، وكان العديد من المرشحين الذين يحاولون ملء الوظائف الشاغرة في طلبات البحث أكثر من 50000 طلب لما يقرب من 1000 وظيفة تقدمها دراستنا باستخدام التعلم الآلي لتسهيل معايير الاختيار واختيار المرشح الأكثر ملاءمة.[1]

وفي هذه الدراسة، استخدمنا نماذج التعلم الآلي وقارنا بين مجموعات البيانات المتوازنة وغير المتوازنة من خلال تدريب النماذج على المتغيرات المرشحة مثل الأداء الأكاديمي وخبرة التدريس والجنس والعوامل الديموغرافية والمزيد، بالإضافة إلى تصنيف المشاركين إلى ثلاث فئات وهي الغائب، وغير الموجود، والمكان، وقمنا بمعالجة مجموعات البيانات باستخدام آلات الدعم المتجهة(SVM) ، وشجرة القرار(DT) ، وXGBoost، والغابة العشوائية(RF) ، وتعزيز التدرج(GB) ، وAdaboost، وNaive Bayes، والانحدار اللوجستي، وتعزيزKNN ، وشبكات الأعصاب متعددة الطبقات [2] (MLPNNs) وتم تقييم أداء كل نموذج، وتم اختيار النموذج القسري العشوائي لأنه حقق أعلى دقة، 99٪، لكل من مجموعات البيانات المتوازنة وغير المتوازنة.