PREDICTION OF THE FUTURE SUCCESS OF CANDIDATES BEFORE RECRUITMENT WITH MACHINE LEARNING: A CASE STUDY IN THE BANKING SECTOR

(İŞE ALIM ÖNCECİNDE ADAYLARIN GELECEK BAŞARILARININ MAKİNE ÖĞRENMESİYLE TAHMİNİ: BANKACILIK SEKTÖRÜNDE BİR VAKA ÇALIŞMASI)

by

Kaan AKSAÇ, B.S.

Thesis

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

in

INDUSTRIAL ENGINEERING

in the

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

of

GALATASARAY UNIVERSITY

March 2022

This is to certify that the thesis entitled

PREDICTION OF THE FUTURE SUCCESS OF CANDIDATES BEFORE RECRUITMENT WITH MACHINE LEARNING: A CASE STUDY IN THE BANKING SECTOR

prepared by **Kaan AKSAÇ** in partial fulfillment of the requirements for the degree of **Master of Science in Industrial Engineering** at the **Galatasaray University** is approved by the

Date:

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF SYMBOLS ..... **v**
LIST OF TABLES ..... **vi**
LIST OF FIGURES ..... **vi**
ABSTRACT ..... **viii**
ÖZET ..... **x**
1 INTRODUCTION **1**
2 HR ANALYTICS AND RECRUITMENT **4**
2.1 Why Recruitment Among HR Functions? 7
2.2 Why Pre-Hire Evaluation in Recruitment Processes? 9
3 LITERATURE REVIEW **10**
4 METHODOLOGY **21**
4.1 Machine Learning 23
4.1.1 Logistic Regression 25
4.1.2 K-Nearest Neighbors 26
4.1.3 Support Vector Machine 27
4.1.4 Decision Trees 29
4.1.5 Multi-Layer Perceptron 31
4.2 Evaluation Methods 32
5 REAL-LIFE APPLICATION **35**
5.1 Data Collection and Preparation 35
5.2 Machine Learning Modeling 40
5.3 Models Evaluation 41
5.4 Improving the Success of the Preferred Model 42
6 CONCLUSIONS **46**
REFERENCES **48**
APPENDICES **53**
BIOGRAPHICAL SKETCH **61**

LIST OF SYMBOLS

AHP	:	Analytic Hierarchical Process
ANN	:	Artificial Neural Network
ANP	:	Analytic Network Process
CART	:	Classification and Regression Trees
CRNN	:	Convolutional Recurrent Neural Network
DT	:	Decision Tree
HR	:	Human Resources
HRM	:	Human Resources Management
IV	:	Information Value
KNN	:	K-Nearest Neighbors
MADM	:	Multiple Attribute Decision Making
MCDM	:	Multi-Criteria Decision Making
ML	:	Machine Learning
MLP	:	Multi-Layer Perceptron
NN	:	Neural Network
ROI	:	Return on Investment
SQL	:	Structured Query Language
SVM	:	Support Vector Machine
TOPSIS	:	Technique for Order Preference by Similarity to An Ideal Solution

LIST OF FIGURES

Figure 2.2 Interest in Google Trend HR analytics by year 5
Figure 2.3 Steps of the recruitment process.....8
Figure 4.1 Proposed methodology of this study 21
Figure 4.2 Machine learning working scheme.....23
Figure 4.3 General classification of machine learning algorithms 24
Figure 4.4 Estimation probabilities using logistic regression (James et al., 2011).....26
Figure 4.5 K-nearest neighbors clustering (Zhang et al., 2020) 27
Figure 4.6 SVM hyperplane (James et al., 2011) 28
Figure 4.7 Cut of a decision tree.....30
Figure 4.8 Backpropagation network structure.....31
Figure 5.1 Categorization of the age attribute with IV 36
Figure 5.2 Learning curve for the proposed model.....44
Figure 5.3 Confusion matrix for the proposed model.....45

LIST OF TABLES

Table 3.1 A Summary of literature studies 16
Table 4.1 Binary confusion matrix.....32
Table 5.1 The attributes used in the model and their possible values.....37
Table 5.2 Example of the limited data set used in the model 39
Table 5.3 The results and averages for each fold of the executed algorithms 41
Table 5.4 Logistic regression hyperparameter tuning results (sorted by accuracy).....42

ABSTRACT

With the developing competitive environment and conditions, it has become important for companies to employ successful employees to survive. To create a strong workforce profile, companies first need to employ the right personnel. For this reason, the recruitment process is seen among the Human Resources (HR) processes with the highest return on investment (ROI).

The rapid digitalization brought by the Covid-19 pandemic process and the compliance requirements of this process have made new business models necessary in the recruitment processes. Companies are now advancing their recruitment processes through online platforms in parallel with their current processes in recruiting new personnel. In addition to this, the increase in job applications and the increasing number of position requirements have increased the complexity of the process and the candidate data that need to be evaluated. Wrong hiring decisions caused by the difficulty of evaluation and complexity can cause long-term and financial and non-financial losses for companies. For this reason, companies need more data-driven decision support systems to tackle these challenges. By analyzing complex data with machine learning (ML) approaches, companies can offer meaningful outputs to decision-makers in their recruitment processes.

Most of the analytical studies on recruitment processes focus on modeling the decisions of decision-makers. However, these models are sensitive to the subjective and biased evaluations of decision-makers. This study, on the other hand, focused on predicting the future performance of new candidates by learning the historical data of the employees who were evaluated as successful and unsuccessful in the recruitment process with machine learning algorithms. In this way, it is thought that decision-makers will be supported to

employ the right employee by reducing the difficulty and complexity of evaluating the increasing data in the recruitment process.

This study covers the data of 597 employees of a private bank serving in Turkey. In the creation of successful and unsuccessful output labels, the performance evaluations of the employees in the first two years have been taken into account. A three-stage methodology has been followed in the study. The first step of this study is to obtain and prepare the data set to be included in ML. In the second stage, the prepared data set has been divided into training and testing. Then, five-fold validation has been performed for Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNNs), Decision Trees (DTs), and Multi-Layer Perceptron (MLP) algorithms with the training data. According to the calculated evaluation criteria, a Logistic Regression model with an accuracy of 71.19% has been proposed. In the last stage, the prediction performance has been developed by optimizing the parameters for the proposed model. With the best parameter values of the developed model, a 73.14% accuracy rate has been obtained in the training data. Then, the model has been run with a test data set that it had not seen before, and a successful accuracy rate of 71.67% has been achieved.

ÖZET

Gelişen rekabet ortamında kurumların hayatta kalabilmeleri için başarılı çalışanları istihdam edebilmeleri önemli hale gelmiştir. Kurumların güçlü bir işgücü profili oluşturabilmek için en başta doğru personeli istihdam etmeleri gerekmektedir. Bu nedenle işe alım süreci yatırım getirisi olarak en yüksek insan kaynakları süreçleri arasında görülmektedir.

Covid-19 pandemi sürecinin getirdiği hızlı dijitalleşme ve bu sürecin uyum gereklilikleri, işe alım süreçlerinde yeni iş modellerini gerekli hale getirmiştir. Kurumlar, artık yeni personel istihdam edilmesinde mevcut süreçlerine paralel olarak online platformlar aracılığıyla da işe alım süreçlerini ilerletmektedir. Bunun paralelinde, iş başvuruları ve pozisyon gerekliliklerindeki artış, değerlendirilmesi gereken aday sayısını ve sürecin karmaşıklığını ciddi boyutlara çıkarmıştır. Değerlendirme güçlüğü ve karmaşıklığın neden olabileceği hatalı işe alım kararları, kurumlar için sonuçları uzun vadeli ve finansal ve finansal olmayan zararlara neden olabilmektedir. Bu nedenle kurumlar, bu zorluklarla mücadele etmek için veri odaklı karar destek sistemlerine daha fazla ihtiyaç duymaktadır. Bu noktada makine öğrenmesi yaklaşımları, karmaşık verileri analiz ederek işe alım süreçlerinde karar vericilere anlamlı çıktılar sunabilir.

İşe alım süreçlerine yönelik analitik çalışmaların pek çoğu karar vericilerin kararlarını modellemek üzerine odaklanmaktadır. Ancak bu modeller karar vericilerin sübjektif ve önyargılı değerlendirmelerine hassas olmaktadır. Bu çalışma ise, işe alınmış olan başarılı ve başarısız olarak değerlendirilen çalışanların işe alım sürecindeki tarihsel verilerini makine öğrenmesi algoritmaları ile öğrenerek, yeni adayların gelecek performanslarını tahmin etmeye odaklanmıştır. Bu sayede, işe alım sürecindeki artan verilerin değerlendirilmesine

ilişkin zorluk ve karmaşıklığı azaltarak, doğru çalışanın istihdam edilmesi için karar vericilere destek sağlanacağı düşünülmektedir.

Bu çalışma Türkiye' de hizmet veren özel bir bankanın 597 çalışanının verisini kapsamaktadır. Başarılı ve başarısız çıktı etiketlerinin oluşturulmasında çalışanların ilk iki yıldaki performans değerlendirmeleri dikkate alınmıştır. Çalışmada 3 aşamalı bir yöntem izlenmiştir. İlk aşama, makine öğrenmesine dâhil edilecek veri setinin elde edilmesi ve hazırlanmasıdır. İkinci aşamada, hazırlanmış veri seti eğitim ve test olarak bölümlendirilmiştir. Akabinde eğitim verisiyle Lojistik Regresyon, Destek Vektör Makinesi, K En Yakın Komşular, Karar Ağaçları ve Çok Katmanlı Algılayıcı algoritmaları için beş katlamalı doğrulama yapılmıştır. Hesaplanan değerlendirme kriterlerine göre 71.19% doğruluk elde edilen Lojistik Regresyon modeli önerilmiştir. Son aşamada, önerilen model için parametre optimizasyonu yapılarak tahmin performansı geliştirilmiştir. Geliştirilen modelin en iyi parametre değerleri ile eğitim verisinde 73.14% doğruluk oranı elde edilmiştir. Daha sonra bu model, daha önce görmediği test veri seti ile çalıştırılmış ve 71.67% oranında başarılı bir doğruluk oranına ulaşılmıştır.

1 INTRODUCTION

Human resources management (HRM) and studies in this field are developing day by day in number and quality in parallel with the development and trends of technology. While (HR) professionals still focus on the "human" aspects of running an organization, in just a decade they have become more dependent on technology and data analysis methods that did not exist before. Data analysis and the usage of technology have become an integral part of HR functions as a result of the need to manage HR functions sustainably, establish effective and agile decision mechanisms to adapt to changing conditions and examine existing and possible scenarios to discover strategic development areas. This situation has led to the development of HR analytics as a research and study area. However, situations, where data-related actions are planned continuously and precisely as a result of data processing (predictions, classifications, rankings, decisions, etc.), may disrupt the dynamics of HR functions.

The functions of HR can be listed as selection and placement, personnel rights management, wage management, performance management, training and development management, career management, as well as various classifications. Studies have shown that selective recruitment and placement is the HR function that improves corporate performance and has the highest return of investment (ROI) in HR (Vlachos, 2008; Ben-Gal, 2019). In addition to these, according to the 2017 Talent Acquisition Benchmark Report compiled by the Society for Human Resource Management, institutions fill an open position an average of 36 days, which costs about $ 4,500 per recruitment. All that time and expenditure spent conducting group discussions and interviews consume more than 90% of the total effort of the recruitment process (Sivaram and Kamar, 2011). Therefore, taking the time to improve

a company's recruiting and hiring processes through data analysis can provide significant savings. For this reason, the researches, proposed methodology, and case study examined in this study are specific to the selection and recruitment function of HR.

Recruitment in an organization affects the quality of the employees, so choosing the right person for the right job by analyzing the data is very important in HRM. Using the knowledge obtained, these analyzes can help HR professionals better understand who will be best suited for both a specific role and the company's overall workplace culture. By hiring the "right" person at the right time the first time, HR professionals will focus most of their time on employee retention and spend less time with those they predict will not work in the future.

However, it is difficult to predict the job performance and retention of the candidates in the selection process. Traditional selection approaches, including personality tests, job sample tests, job knowledge tests, and interviews, have been widely used for many years. These approaches usually conclude based on the decision maker's subjective judgments. However, personnel selection is a complex process in which many factors must be evaluated simultaneously in the decision-making process.

When the competitive positions and strategies of an institution in the sectors in which it operates are evaluated, it is seen that the recruitment and placement function has quite different dynamics. The banking sector is one of the examples where recruitment and placement cannot be managed stably due to its dynamic business structure. As globalization and technology advance, multi-functional roles and cross-functional roles are increasing as new business needs are constantly being created. Recruiting high-potential talent in the banking sector has become more important. In addition to this, demands for the competencies and skills have been diverse and more complex. For this reason, it is considered that traditional recruitment and placement methods, which are valid for stable conditions, will no longer be sufficient and suitable for the banking sector (Lievens et al., 2002). As an innovative approach to assist decision-makers in the recruitment process in their task of identifying the best candidates, ML can provide huge savings in terms of financial resources and time.

In this study, a model has been proposed that can predict the future performance of new candidates by learning the historical data of current employees in the recruitment process with ML techniques. This study aims to create a foresight system that can assist decision-makers in recruiting candidates.

In the first part of the study, information about HR analytics and recruitment processes has been given and why this field has been chosen in the study is explained. In the second part, literature studies on performance evaluations of employees and candidates have been examined. In the last section, a three-stage methodology has been followed for the ML approach used in practice. The first stage of the application is the collection and preparation of the data set to be included in ML. In the second stage, five-fold validation has been performed for the prepared data set with Logistic Regression, SVM, KNNs, DTs, and MLP algorithms, which are widely used in this field. The best model has been proposed according to the calculated evaluation criteria. In the last stage, the parameters of the proposed model have been optimized and the prediction performance in the training data has been improved. Then, the performance of the developed model has been tested with the best parameter values and the accuracy score has been calculated.

2 HR ANALYTICS AND RECRUITMENT

HR analytics is a discipline that serves horizontally with all functions, with technology development and data management affecting HR functions. Although there are many definitions for HR analytics, which has recently been increasing, these definitions have some common points. HR analytics includes not only determining and measuring metrics, but also collecting, processing, analyzing, and interpreting HR data by data analytics and technologies. However, in HR analytics, not only the management and processing of data within the scope of HR are taken into account, but also studies are carried out to ensure the integration of these data with data outside of HR. Considering this integration, the fact that the effects of the decisions taken with the support of HR analytics have effects on company performance explains the role of HR analytics in strategic HRM (Marler and Boudreau, 2017).

Considering many definitions, Marler and Boudreau (2017) defined HR analytics as:
"An HR practice enabled by information technology that uses descriptive, visual, and statistical analyses of data related to HR processes, human capital, organizational performance, and external economic benchmarks to establish business impact and enable data-driven decision-making"

Although the studies on HR analytics have accelerated in recent years, these studies are mainly based on long-standing research and ideas in the field of HR. The metrics and key indicators to be used in this field, their determination, and evaluation have been the subject of discussion by HR professionals and researchers until the late 1970s, and many types of research have been carried out in light of these discussions. The following years focused on

the development of new techniques to calculate and analyze the return on human and intellectual capital. In the 2000s, techniques have been developed to measure the effects of HR practices on corporate performance. However, HR analytics has attracted great attention in business-oriented articles in journals such as Harvard Business Review, and in reports published by consulting organizations (Madsen and Slatten, 2017).

Google Trends is a tool that expresses interest in certain topics in terms of internet search numbers. The evolution of searches for HR analytics concepts and ideas, based on Google Trends searches, is shown in Figure 2.1 (Data source: https://trends.google.com/trends/, data queried on June 1, 2021).

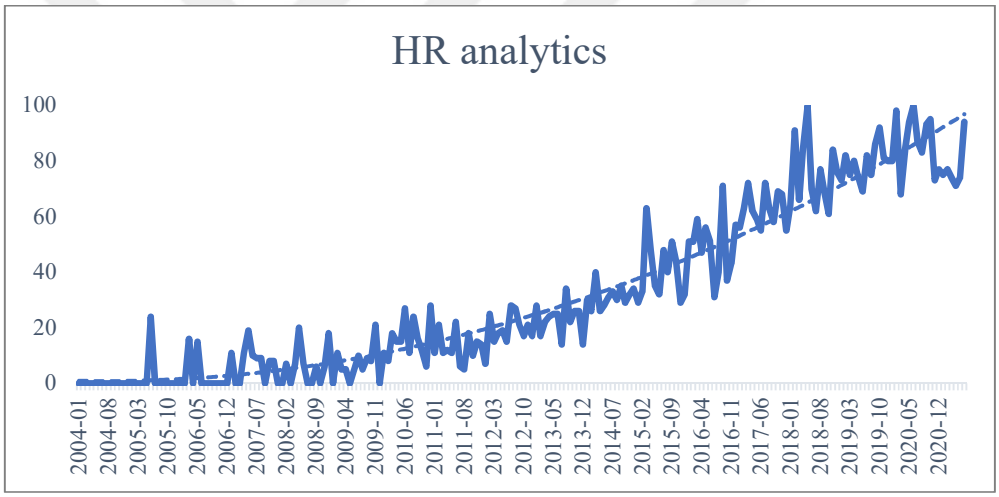

Figure 2.1 Interest in Google Trend HR analytics by year

Analyzing Figure 2.1, it is seen that the interest in HR analytics has increased significantly as of 2015. Recently, the common point of research on HR analytics is how it can be used as a decision support system and to predict future decisions or behaviors, called “predictive analytics”.

Predictive analytics in HR started gaining momentum in late 2010 as more data analytics tools to analyze HR data became available. However, according to Deloitte research, only

8% to 10% of companies use predictive analytics in HR (King Jr, 2018). This rate is expected to increase gradually in the coming years.

Isson and Harriott (2016) presented a framework in seven chapters, in which the issues and problems of HR analytics are discussed within the lifecycle of HR activities. They summarized the main issues of HR analytics as below:

- Workforce planning (processes that help the company determine what capabilities it needs to gain competitive advantage and achieve business goals)
- Talent resource (studies on which channels and recruiting resources will be most effective in acquiring potential candidates)
- Talent acquisition (the process of analyzing the relationship between the performance of the candidate to be recruited during the interview and the performance in the position sought)
- Onboarding, engagement, and culture fit (to do on ensuring the commitment, culture harmony, and loyalty of hired employees)
- Talent performance management and the lifetime value of talent (revealing that high or low performance)
- Talent retention (data analysis to address the bright talents at risk of leaving and why)
- Talent fitness, talent bodily health, and talent safety (efforts to create an environment that supports their employees' well-being, health, and safety to be successful)

The concept of talent, which is frequently mentioned here, refers to the employees, who are the most important capital of the companies. On account of this, hiring the right talent formed the framework of this study.

Many organizations invest in increasing their employees' success in the organization or in hiring the best employees. Employing the “best” employee is organizational psychology that is supposed to provide a competitive advantage. But hiring this “best” employee at a reasonable cost is an impossibility theory. After an employee is hired, companies are

involved in many training and development activities. A lot of expenses will be made until the efficiency of the employee begins to be obtained. After this point, this employee may leave the institution for many different reasons such as a career plan and not adapting to the institution. Subsequently, all these processes will be run again after the employee leaves. This is a failure of existing HR processes.

HR departments that keep up with technological progress have entered the process of transitioning to data-based decision support systems to support talent and recruitment strategies. Recently, terminologies such as data mining and data analytics have taken HR dynamics to a new level. After this stage, we start to encounter HR analytics frequently. (Jain and Jain, 2020).

2.1 Why Recruitment Among HR Functions?

Recruitment is an HR activity that deals with the selection of applicants to the institution. In a broader definition, it can be defined as the process of creating a talent pool consisting of qualified candidates applying to the positions opened according to the needs of the institution and selecting the right candidates from this pool.

The selection of candidates is carried out based on strategies structured and well-defined by the organization. This strategic process may differ according to many institutions. However, it can be summarized in three main stages in general. These are: identifying job requirements, attracting potential candidates, and selecting the most suitable and qualified candidates for the required position. A summary of the process is given in Figure 2.2.

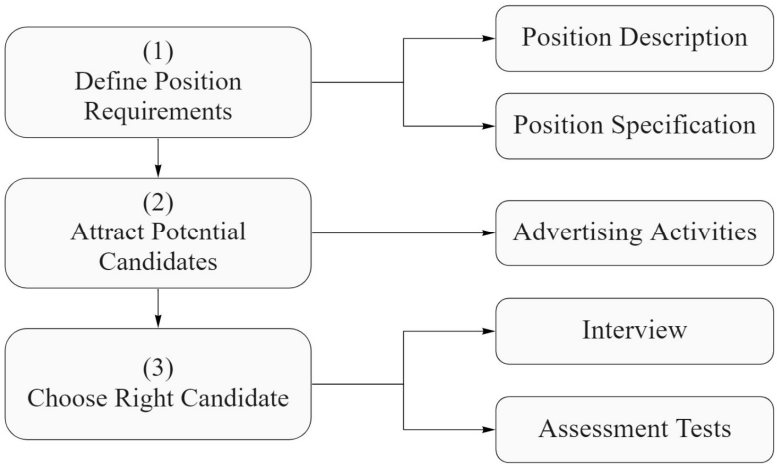

Figure 2.2 Steps of the recruitment process

Recruitment in an organization affects the quality of employees, so choosing the right person for the right job is very important in HRM.

Recent research points the different tasks of HR functions to the application areas of HR analytics and discusses how these areas can affect the ROI. The research shows that the areas of application of HR predictive analytics in workforce planning and recruiting have the highest impact on ROI. Additionally, other HR analytics applications such as “job analysis”, “industry analysis” and “performance management” have a low expected ROI. Areas such as “training and development”, “compensation” and “retention” have moderate expected ROI (Ben-Gal, 2019). From this point of view, it can be said that HR professionals' estimation-oriented studies on the recruitment and placement function and developing applications will provide a higher ROI.

2.2 Why Pre-Hire Evaluation in Recruitment Processes?

Post-hire information includes HR functions such as employee engagement, corporate engagement, corporate support, and retention. This information affects the success of the employees, but the cost of a wrong decision will be quite high as it requires post-employment evaluation. Simulating recruiters' decisions may not always be the best approach, as they are affected by highly subjective and potentially misjudgments that preserve rather than ameliorate hiring bias. Moreover, organizations often focus on turnover when assessing post-hire employee success. The turnover rate alone may not indicate recruitment success. As is often the case in practice, underperforming employees may not leave firms because of corporate policies to minimize layoffs and encourage high internal mobility. In addition, dismissals may be limited due to many regulations and conjunctures. For example, many examples of this have been seen in the Covid-19 pandemic period. In this process, the state prohibited the employees of companies from leaving their jobs, except for compulsory situations. Also, turnover is a reality for HR itself. The experience of recruiters, who perform an important function of HR, is of great importance. Considering the competition in the banking sector, it is seen that turnover is high among recruiters. This situation seriously damages the recruitment memory.

On the other hand, with pre-hire information, it is possible to have a lot of data to predict the post-hire behavior of the employees and support the hiring decision. Thus, recruiters can have an objective measure of the true success of employees and be supported by meaningful insights and decision support systems for their recruitment and performance. As a result, it is highly beneficial to focus on the early pre-hire forecasts that have the greatest impact on ROI for an organization. In this way, it will also be of great benefit to anticipate toxic employees who may not leave the job and significantly damage the corporate culture.

Our pre-hire performance estimation approach, which we focused on in our study, seems to be compatible with all these inferences.

3 LITERATURE REVIEW

Literature studies on the estimation of employee performance in recruitment in HR have been examined in two dimensions subjects and methods.

Early studies of the literature have been seen to focus on performance forecasting with post-employment personnel data. It may be useful to use post-hire personnel data (performance, attendance, etc.) for some processes such as retention, personnel satisfaction, and career planning. However, the use of post-hire personnel data for turnover tendency and future performance prediction may lead to recruitment errors such as hiring prone to quit or low-performing employees, and late detection of these errors (Valle et al., 2012). Predictions made using personnel data before recruitment is a decision support mechanism that allows the right person to be placed in the right job. In addition to this, these predictions can employ people who are thought to be successful in the long run, thus providing many financial and non-financial benefits for institutions. However, very few of the studies in the literature on predicting future performance and turnover tendency focus on pre-employment analysis (Chien and Chen, 2008). So, the literature study has been limited to the estimation of employee performance before recruitment.

When the methods are examined for recruitment analytics; traditional statistical approaches such as descriptive statistics, hypothesis testing, analysis of variance, regression, multi-criteria decision making, and correlation analysis are frequently used. On the other hand, data mining approaches have been widely used after 2008. Among these approaches, ML-related applications are seen as innovative techniques.

Cho and Ngai (2003) focused on the characteristics of data warehousing and the appropriate data mining techniques that can be used to support agent selection in the insurance industry.

Chien and Chen (2007) proposed a framework based on rough set theory to take out possibly useful rules from historical human resource data to increase candidate selection effectiveness in terms of retention and performance of new talent.

Chien and Chen (2008) developed a data mining framework to predict their work performance and retention by using their demographical data with 3825 employees' samples in the semiconductor industry. Moreover, they recommended useful rules for personnel selection by applying DT algorithms.

Dagdeviren (2010) proposed a model for the problem of employee selection using both the Analytic Network Process (ANP) and modified Technique For Order Preference By Similarity To An Ideal Solution (TOPSIS) methods. ANP method for obtaining selection indicator of the considered dependence weight in the ranking of applicants used and modified TOPSIS method has been adopted.

Jantan et al. (2010) conducted a study on the application of a data mining approach to employee development regarding their future performance. They introduced classification rules to assist HR professionals in determining whether an employee has the opportunity to be promoted.

Chaudhry and Usman (2011) investigated the relationship between employees' emotional intelligence and their performance. They stated that evaluating the emotional intelligence of employees is an effective method that can be used by recruiters to predict their performance.

Delgado-Gomez et al. (2011) examined the possibility of making better the accuracy of the existing expert system for salesperson selection by applying SVMs. The system estimated the performance of these applicants based on some points that are measures of cognitive traits, personalities, sales skills, and biological data.

Sivaram and Kamar (2011) focused on exploring key criteria and leveraging existing patterns by using DTs to minimize effort in the recruiting process in the information technology industry.

Jantan et al. (2011) tried to use this approach in talent management to determine current capabilities by predicting their performance with experience data using classifying algorithms.

Rouyendegh and Erkan (2012) studied a fuzzy Analytic Hierarchical Process (AHP) to choose the most appropriate academic staff.

Azar et al. (2013) provided a decision-making method for managers to use in the recruitment process in a commercial bank. The effective components in the performance of the employees were determined by exploring the hidden patterns of the connection between the test results and their job performance.

Bach et al. (2013) explored whether it is possible to predict employee success in recruiting in the banking sector using a data mining approach. The variables used as predictors in this study are psychological test results. The variables used as the criterion variable are employee performance and employee promotion rate.

Thakur et al. (2015) proposed an ideal selection framework for hiring the right candidate and to recenter on the selection criteria by applying the random forest algorithm in the software industry.

Li et al. (2016) proposed a model to deal with the employee performance estimation problem which can be a significant part to estimate and enhance the performance of a manufacturing system using an improved KNN algorithm.

Shehu and Saeed (2016) improved an adaptable personnel selection model to assist personnel selection for hiring and adapt to changes in personnel selection strategy.

Kirimi and Moturi (2016) used the data mining classification practice to extract information that is important to estimate employee performance using former assessment records, from a public administration development institute in Kenya.

Menon and Rahulnath (2017) offered an efficient approach to evaluating and ranking job seekers in a recruitment process by predicting their emotional intelligence using social media knowledge. They designed the personality predictor based on ML techniques like supervised classification.

Harris (2018) looked at ways in which human feedback could be used to better train ML algorithms, paying special attention to native risks such as data overfitting and bias avoidance.

Sarker et al. (2018) showed how data clustering and DTs can be used for predicting the employee's performance for the following year.

Xue et al. (2019) presented a hybrid convolutional recurrent neural network (CRNN) with the KNN model, a dataset with 22 attributes, which is used to predict personnel performance in the future and help decision-makers to select the most competent candidates.

Mahmoud et al. (2019) asserted a procedure that can help decision-makers select the best candidate by estimating his\her performance according to produced performance patterns by applying Machine-learning techniques.

Santiago and Gara (2019) proposed a model using a naive Bayes classifier to help HR personnel understand the psychological climate and supported their decision to solve turnover by selecting desired applicants who are likely to stay longer in an organization.

Lather et al. (2019) predicted the performance of employees in an organization through various factors, including individual, field-specific and socio-economic. They indicated that

the predicted performance could be the basis for deciding whom to be hired and what type of project to place on.

Nasr et al. (2019) created a model with DM techniques (DT, naive Bayes, and SVM) to estimate employee performance by applying a real data set in Egypt.

Karam et al. (2020) proposed a framework that presents an established tool to use cognitive abilities associated with systems thinking skills as a selection tool for hiring potential applicants using fuzzy Multi-Criteria Decision-Making (MCDM) methods in a US large-scale organization. The purpose of this study was to rank all applicants based on their systems thinking skills and afterward to select the candidates most in line with the company's strategy.

Du and Li (2020) presented a data-driven approach to the student selection process in a business graduate program.

Pessach et al. (2020) proposed a hybrid decision support system for HR professionals to improve pre-recruitment processing using a variable-order Bayesian network (for interpretability) and ML algorithm.

Calixto and Ferreira (2020) used a naive Bayes model to classify 594 salespeople into pre-defined categories based on a global company.

Chuang et al. (2020) improved a data-driven Multiple Attribute Decision Making (MADM) method, which integrates MADM techniques and ML, to assist personnel selection more objectively and their competency progress in a Chinese food company.

A study by Arora et al. (2020), proposed a predictive study based upon linear and logistic regression. This research has aimed to estimate the educationist's performance based on Educationist Evaluation Index and experience in Delhi.

Dhliwayo and Coetzee (2020) examined cognitive intelligence, ability emotional intelligence, trait emotional intelligence, and personality types as predictors of job performance to recommend a valid selection model with 299 samples for various organizations in Zimbabwe.

Santhosh and Mohanapriya (2020) proposed a framework to support HR to observe employee performance by applying the C4.5 classifier, naive Bayes classifier, and generalized fuzzy c-means clustering.

The literature studies are summarized in Table 3.1 in terms of the methods used and their results.

Table 3.1 A Summary of literature studies

Authors	Study	Method	Results
Karam et al. (2020)	Multi-Criteria Decision-Making Model to Recruit Employee Candidates	Fuzzy MCDM	The proposed model has been ordered all potential employees based on the systems thinking skills and afterward classified the most appropriate candidates most in line with the company's strategy.
Du and Li (2020)	A Data-Driven Approach to High-Volume Selection	Linear Regression Analysis	The results have shown that the interview score is statistically significant to predict performance, but the screening score is not.
Pessach et al. (2020)	Variable-Order	Variable-Order Bayesian Network with Gradient Boosting	It has been proposed a framework capable of providing a balanced hiring process while findings both diversity and recruitment success (Prediction performance rate = 0.73)
Calixto and Ferreira (2020)	Salespeople Performance Appraisal with Predictive Analytics	Naive Bayes	The results have been successful to predict the accuracy of 92.50% for the proposed model.
Chuang et al. (2020)	A Data-Driven MADM Approach for Employee Selection and Improvement	DEMATEL, PROMETHEE-AS	The proposed model can assist managers to select the right employee for a position. Moreover, teamwork and ethics have been shown that two significant factors are acting the employees' job efficiency.
Arora et al. (2020)	Estimation of Educationist's Performance Using A Regression Model	Linear and Logistic Regression	The experiment results have shown that there is a correlation between EEI and experience with an accuracy of 71.47% with linear and 85.71% with logistic regression.

Authors	Study	Method	Results
Dhliwayo and Coetzee (2020)	Cognitive Intelligence, Emotional Intelligence, And Personality Types as Predictors of Job Performance	Survey and Statistical Analysis	The results have shown that cognitive intelligence is the best predictor of job performance comes to second ability emotional intelligence and is followed by personality types.
Santhosh and Mohanapriya (2020)	Performance Prediction with Fuzzy Logic	Generalized Fuzzy C-Means	The proposed model is more effective on account of accurately predicting employee performance. This model has helped to predict an employee's rise, career growth, and promotion.
Xue et al. (2019)	A Hybrid Framework for Personnel Performance Prediction	CRNNs with KNN	With the proposed framework, prediction performance has been improved with an accuracy of 96.49%
Mahmoud et al. (2019)	Performance Predicting in Hiring Using ML	J48 Algorithm	The proposed model has provided the highest information benefit for the feature "Job Title" by predicting an accuracy of 70%.
Santiago and Gara (2019)	A Model-Based Prediction of Desirable Candidates	Naive Bayes Classifier	The results have indicated that over the age of 20 and living away from the workplace has a higher possibility of staying longer than other candidates.
Lather et al. (2019)	Prediction of Personnel Performance Using ML Techniques	SVMs, Random Forest, Naive Bayes, NNs, and Logistic Regression	SVM has been verified to be the most accurate in terms of accuracy for the proposed framework.
Nasr et al. (2019)	Predicting Employees' Performance Applying Data Mining Techniques	DT, Naïve Bayes, and SVM.	The SVM technique has the highest prediction rate using up to five influencing factors with an accuracy of 86.90%. Education has been a more influential factor in performance than experience.
Harris (2018)	Making Better Recruitment Decisions Using "Human In The Loop"	C4.5 DT Classification	The results have shown that human involvement in the loop to help establish rules can improve the sequencing made by the machine learning algorithms.

Authors	Study	Method	Results
Sarker et al. (2018)	Employee's Performance Analysis and Prediction	K-Means DT	It has been estimated the number of employees selected for promotion or appointment and fired based on their performance.
Menon and Rahulnath (2017)	A Novel Approach to Appraise and Rank Job Seekers in A Recruitment Process	Linear Regression and Support Vector Regression (SVR).	A positive correlation of approximately 65% has been found between the scores assigned by the recruiter and the scores predicted by the system by using the Pearson correlation coefficient.
Li et al. (2016)	Employee Performance Modeling for Manufacturing	KNN	The results have indicated a high estimation accuracy when the weighted distance between the predicted individual and samples has been small.
Shehu and Saeed (2016)	Personnel Selection Model For Recruitment	C4.5, Random Tree, REPTree and CART	The proposed model has been ranked the highest after comparing it with selection models developed using four DTs.
Kirimi and Moturi (2016)	A Data Mining Classification Approach in Employee Performance Estimation	C4.5 Algorithm	Employee performance is highly influenced by experience, age, academic qualification, professional education, gender, marital status, and previous performance evaluation scores.
Thakur et al. (2015)	Data Mining for Prediction of Human Performance in The Recruitment of New Personnel	Random-Forest	It has been shown that the academic scores of students cannot predict their performance in the software industry. Other features like Programming Skills, Domain-Specific Knowledge, and Analytical Skills must be tested as a significant prediction.

Authors	Study	Method	Results
Azar et al. (2013)	A Model for Personnel Recruitment with A Data Mining Approach	DT (CHAID, QUEST, CART and C 5.0)	It has been concluded that only five variables out of the 26 effective variables have the highest effect on the 'promotion score' goal, such as the province of employment, education level, exam result, interview result, and work experience.
Bach et al. (2013)	Predicting Employees' Success While Recruiting at A Bank	Multiple Regression	It has been shown that the following attributes are significant in predicting a personnel's achievement at work in the banking sector: cognitive ability, judgment, dominance, social courage, sensitivity, openness to change, warmth, and emotional stability.
Rouyendegh and Erkan (2012)	Choosing of Academic Staff Using the Fuzzy Analytic Hierarchy Process	Fuzzy Analytical Hierarchy Process (FAHP)	According to the FAHP approach, a model has been presented where the staff with the highest score is the most appropriate selection for the job.
Chaudhry and Usman (2011)	Examining the Relationship Between Employees' Emotional Intelligence and Their Performance	Pearson Correlation and Linear Regression Analysis	It has been revealed that using the employees' job performance can be predicted emotional intelligence scores.
Delgado-Gomez et al. (2011)	Making Better Sale Performance Prediction	SVMs	It has been shown that the proposed system achieves 5% higher accuracy (82%) than the state-of-the-art systems.
Sivaram and Kamar (2011)	Helping A Sector's Recruitment Process by Determining Superior Selection Criteria	ID3, C4.5, and CART Algorithms	It has been observed that the performance of the C4.5 algorithm is higher than the others. Examinations have been made with DTs and 20 rules were drawn to support recruitment decisions.
Jantan et al. (2011)	Applying Data Mining Techniques to Ensure the Right Person for The Right Job	C4.5, J48	It has been shown that the feasibility (accuracy = 95.14%) of the Data Mining techniques recommended for employee performance data.

Authors	Study	Method	Results
Dagdeviren (2010)	A Hybrid Multi-Criteria Decision-Making Method for Applicant Selection	ANP and Modified TOPSIS	It has been shown the effectiveness and applicability of the proposed model. Company management practices and the results of the candidate selection method have been found acceptable.
Jantan et al. (2010)	Performing Data Mining Classification Methods for Employee Performance Prediction	C4.5, J4.8	The C4.5 / J4.8 classifiers are the algorithm with the highest accuracy (79.49%) in the sample study.
Chien and Chen (2008)	Data Mining to Improve Personnel Selection	DT with Chi-Squared Automatic Interaction Detection	It has supported the development of HRM actions including job rotation, job redesign, mentoring, and career path development offering a receivable level of accuracy.
Chien and Chen (2007)	Hire and Acquire High-Potential Talents for Semiconductor Manufacturing	Rough Set Theory	29 rule sets have been released that could serve as a reference for hiring the right talent.
Cho and Ngai (2003)	An Approach Based on Data Mining for Recruitment of Insurance Sales Agents	Discriminant Analysis, DTs, and ANNs	It has been found that the system would assist the recruiters in the selection of new candidates with the rule sets and impact analysis it produces.

4 METHODOLOGY

This study aims to present an analytical framework that can be applied to HR recruiters as a decision support tool to recruit new candidates accurately and efficiently by revealing the relationship between the historical data of the employees in the recruitment processes and their job performance after they are hired. Figure 4.1 shows the framework with the following steps:

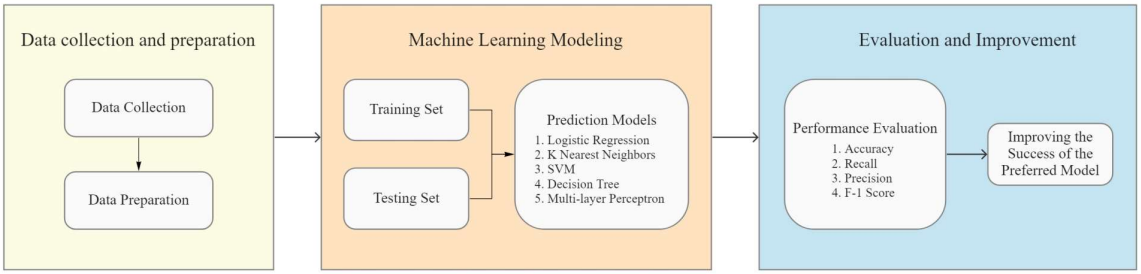

Figure 4.1 Proposed methodology of this study

In this context, the proposed methodology defined in Figure 4.1 consists of three main components:

- **Data Collection and Preparation:** Analyzing the data sources and types to be used in the study, reaching the right data, and preparing these data for ML form the basis of this section. After these data are obtained, they need to be combined and prepared. However, the collected data can often be dirty, incomplete, and inaccurate. Generally, data preparation processes are the most time-consuming part of ML studies. To increase the quality and efficiency of this data in the data preparation process and to train algorithms correctly; It may be necessary to control the data

distribution and outliers, remove or correctly derive blank or missing data, perform correlation analysis, enrich the data, and convert it into suitable formats in which the algorithms can work. These studies are given in detail in Section 5.1.

- **Machine Learning Modeling:** After the data collection and preparation processes are completed, the data set to be used is divided into training and test sets before the models are run. Five-fold validation has been performed for Logistic Regression, SVM, KNNs, DTs, and MLP algorithms, which are observed to be widely used in the estimation of candidate performance in the literature. Details of these algorithms have been given in the following sections.

- **Evaluation and Improvement:** After the models have been trained with the training data, their performance has been measured according to the evaluation criteria. The model with the best training performance has been proposed. Then, the performance of the proposed model has been improved by parameter optimization. Finally, the learning curve of the developed model has been examined and tested with a data set that it had not seen before. The output results have been interpreted by examining the confusion matrix.

In this section, the ML approach, the algorithms used in the application, and the evaluation criteria are given respectively.

4.1 Machine Learning

The concept of ML was introduced by Arthur Samuel in the 1950s. In those years, Samuel has programmed a computer that plays checkers, which constantly improves itself by learning from its mistakes.

ML is a discipline that aims to obtain certain results by using statistical methods via computers. The computers are expected to reveal certain rules by learning from the data using ML algorithms. There is an important difference between ML and traditional approaches. Unlike traditional approaches, ML is not the creating outputs from the data by programming certain rules using computers; it is programming the computers within these rules. According to ML algorithms; the computer compares data using mathematical models and learns from them. Then, thanks to the pattern it has created, it predicts the outputs when a new data set arrives. The general operation of ML is depicted in Figure 4.2.

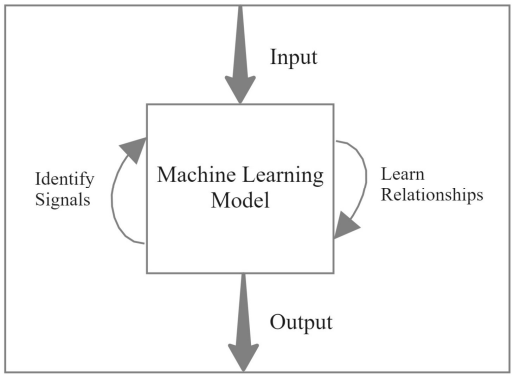

Figure 4.2 Machine learning working scheme

However, the ML algorithms that will guide this study are examined in three classes supervised learning, unsupervised learning, and reinforcement learning as the most frequently used classification method, as seen in Figure 4.3.

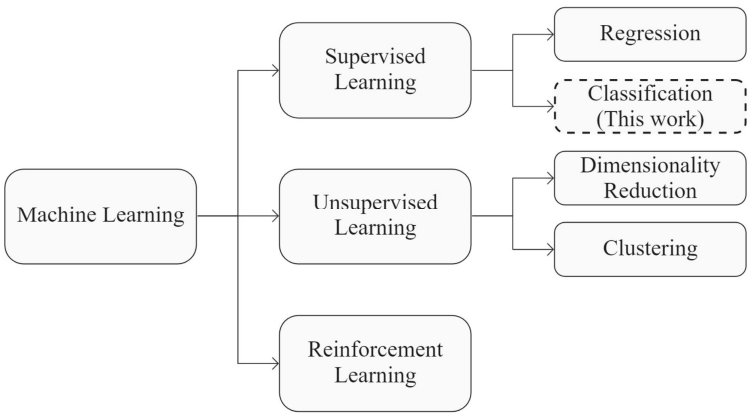

Figure 4.3 General classification of machine learning algorithms

- Supervised learning is a type of learning where inputs and outputs are combined into the ML model. The system generates a prediction model by establishing a cause-effect relationship between inputs and outputs with the help of mathematical algorithms. Thanks to the estimation model produced, the outputs of new inputs that have not been seen before can be predicted. Although supervised learning is expressed with different subsets in different sources, it produces solutions to two problems of class and regression type.

- Unlike supervised learning, there are no outputs in unsupervised learning. It is tried to get results from the data entered into the ML model without output information. In clustering analysis, which is the most common type, the system groups the data it examines with the help of algorithms. However, the system does not care about which group is what, it only creates clusters from the inputs. Afterward, the evaluators examine the clusters and reach meaningful information.

- The reinforcement learning method is similar to the learning method of living things that can think like humans. Reinforcement learning is learning what needs to be done to get the highest gain. The trainee is not guided as to what behaviors to follow. Instead, the trainee discovers by a trial method which behavior will yield the highest reward.

In this study, classification methods based on supervised learning will be used. The training data used in this study consists of historical data of employees before recruitment. A set of supervised ML algorithms will be used with performance grades from the company's performance management system as the target variable. This training data includes many personal data of employees before recruitment. This helps us build a robust system that can predict how a newly hired employee will perform.

Although there are many models used in solving classification problems based on the supervised learning approach; in this study, Logistic Regression, K-Means, SVM, DTs, and MLP models, are frequently used in the literature and meet the assumptions about the data set, will be used.

4.1.1 Logistic Regression

Linear regression is a technique that can use to estimate numerical outputs. However, there are some assumptions and drawbacks to using linear regression in classification outputs. There are many demographic variables in real-life examples (as in our model). Therefore, as an ML technique logistic regression is used for class predictions. In logistic regression, instead of modeling the output value directly, the probability that the output belongs to a class is modeled.

The mathematical interpretation of logistic regression is shown in Eq.1. X_i $(i = 1, \dots n)$, Y $(Y \in \{0,1\})$, $p(X)$, α describes the defining variables, category variables, class probability of belongings, and maximum eigenvalue of two populations respectively (Delgado-Gomez et al., 2011).

$$log\left(\frac{p(x)}{1-p(x)}\right) = \alpha + \sum_{i=1}^{n} X_i \qquad (1)$$

The probability of belonging to a class according to the given attributes is taken as a value between 0-1 using the sigmoid function. Then the probability threshold is decided. The estimation function used in logistic regression is as in Figure 4.4.

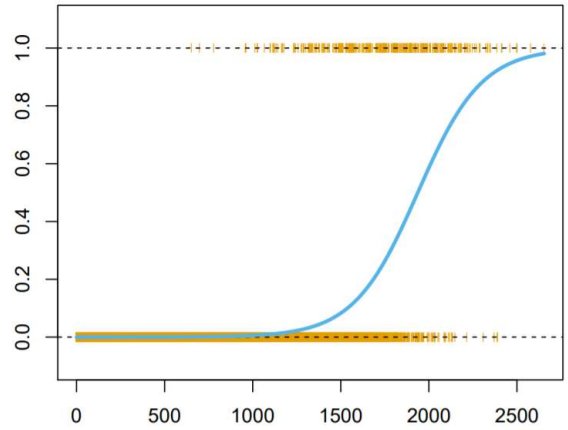

Figure 4.4 Estimation probabilities using logistic regression (James et al., 2011)

4.1.2 K-Nearest Neighbors

With K-means clustering, firstly proposed by MacQueen (1967), to divide a data set into k clusters, k initial cluster centers are selected, the related items are assigned to the nearest cluster centers (d_i) and the cluster centers are updated (C_j). These iterations are repeated until the optimum result is reached (Wagstaff et al., 2001). The mathematical interpretation of K-Means clustering is shown in Eq.2.

$$lSSE = \sum_{i}^{k} \sum_{x \in C_j} dist^2(d_i, x) \qquad (2)$$

Similar to the K-means algorithm, in ML created using KNNs, new values entered for prediction are classified by finding their nearest neighbors in the training set. It assumes that

data points that are close to each other will be in the same class as each other. Figure 4.5 gives an example of dividing classes according to the k value.

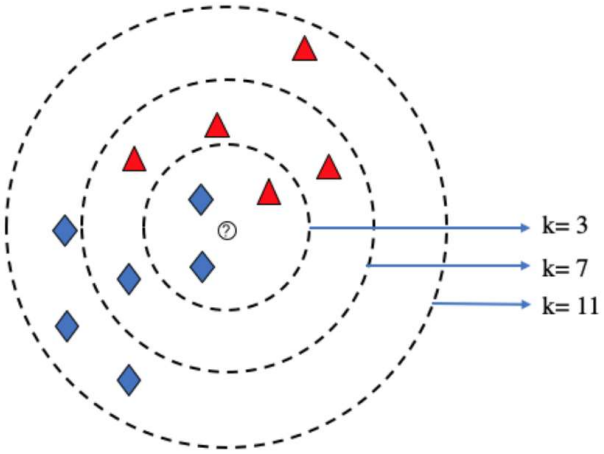

Figure 4.5 K-nearest neighbors clustering (Zhang et al., 2020)

4.1.3 Support Vector Machine

The SVM is a classification approach that was developed in the computer science community in the 1990s and has grown in popularity ever since. SVMs have been shown to perform well in a variety of settings. It is generally considered one of the best "out of the box" classifiers. SVM states the optimum hyperplane allocating the classes to make a classification (James et al., 2011).

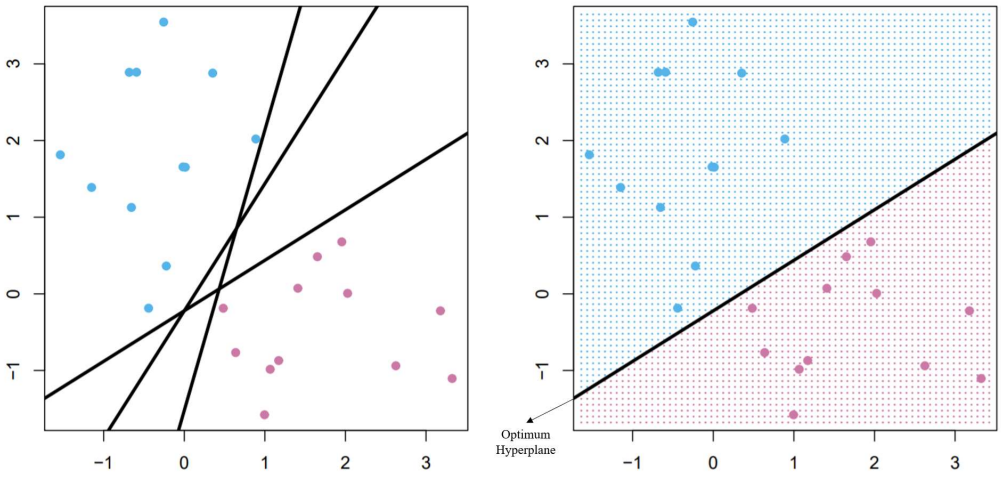

Figure 4.6 SVM hyperplane (James et al., 2011)

If the M m-dimensional data $x_i, i = 1,\ldots,M$, which belongs to the is assumed to be linearly sectional, they find the decision function $D(x) = w^t x + b$ where w is an m-dimensional vector, b is a bias term, and for $i = 1,\ldots,M$

$$w^t x_i + b = \begin{cases} \geq 1, & y_i = 1 \\ \leq -1, & y_i = -1 \end{cases} \tag{3}$$

However, the assumption that the data are linearly separable is rarely achievable. When this assumption is not met, the SVM optimization needs to be extended. Here, the non-negative slack variables $\delta_i,\ i = 1,\ldots,M$ is shown. In this case, the optimization problem is the minimization of the equation.

$$\frac{1}{2}\|W\|^2 + C\sum_{i=1}^{M}\delta_i \tag{4}$$

s.t.

$$y_i(w^t x_i + b) \geq 1 - \delta_i, i = 1, \ldots, M$$

SVM can perform mathematically non-linear transformations with the help of a Kernel function $K(x_i, x_j)$, and in this way, it allows the data to be separated linearly in high dimensions. In this case, the optimization is based on maximizing the following Eq. 5 (Delgado-Gomez et al., 2011).

$$\sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i,j}^{M} \alpha_i \alpha_j y_i y_j \, K(x_i, x_j) \tag{5}$$

s.t.

$$\sum_{i=1}^{M} y_i \, \alpha_i = 0$$

SVM Radial Basis Function kernel has been used in this study.

4.1.4 Decision Trees

DTs are a successful technique that is widely used in supervised learning problems with its high accuracy and simple operation logic.

Although other algorithms such as neural networks (NN) can also be used for classification, decision trees have the advantages of easy interpretation and understanding, and fast learning, as decision-makers compare with real data for validation. In addition, DTs allow us to analyze various categorical data without the need for assumptions about the underlying distribution (Chien and Chen, 2008). In general, the main advantages of using DTs over other ML approaches are:

- faster calculation
- higher level of accuracy
- easier to learn
- generating more logical rules

In the proposed approach, a DT, which is a data mining method, is used to predict the future performance of the employee and to create a rule set from the trained data. Thus, it is thought that it will contribute to the interpretation of the ML approach to be designed for personnel selection. We can briefly summarize the working logic as follows. DTs are automatically updated each time the algorithm is run, with all historical trees stored as a result of the ML example tree in Figure 4.7. There are various DT algorithms. In this study, the Classification and Regression Trees (CART) algorithm has been used.

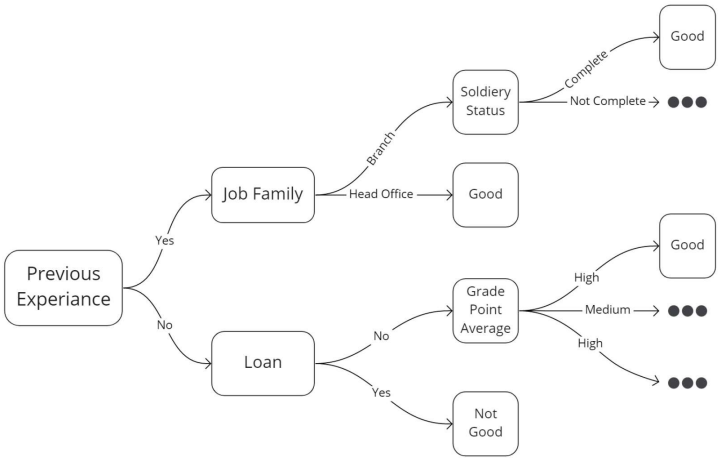

Figure 4.7 Cut of a decision tree

4.1.5 Multi-Layer Perceptron

MLP is a supervised learning algorithm that learns the behaviors in the dataset. If this perceptron structure consists of input and output layers, it can be expressed as single-layer NNs. The NN architecture, which is formed when hidden layers are added between the first and last layer, is expressed as multi-layer NNs. Although this structure is similar to logistic regression, it differs because there are many hidden layers between input and output variables.

The MLP classifier used in the study focuses on training the model using the back-propagation algorithm. The back-propagation algorithm starts with the input variables of the data set from the input layer to the output layer. Secondly, the error between the label value in the output data and the value calculated in the model is calculated, and this error value is propagated backward and the parameter (weights and bias) values between the layers are updated. This iterative structure ends when the number of iterations or the error value specified in the parameter is reached. The model seen in Figure 4.8 is an example of a backpropagation network structure.

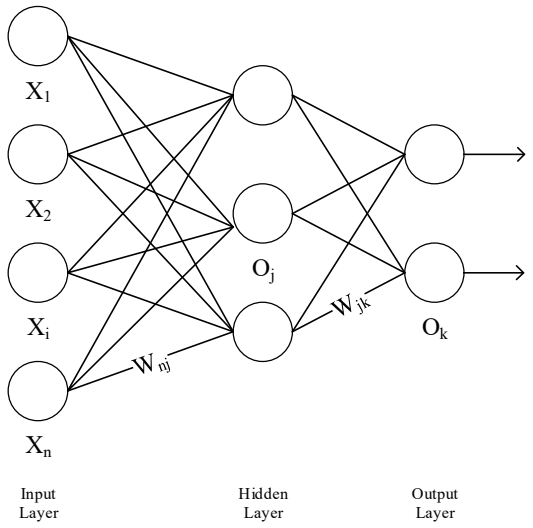

Figure 4.8 Backpropagation network structure

4.2 Evaluation Methods

Before moving on to classification analysis evaluation criteria, we need to examine the confusion matrix used in the criteria. The confusion matrix allows the evaluation of the predictions of classification models.

Suppose a dataset contains a total of N numbers of data evaluated by the M model. In a binary confusion matrix that can be seen in Table 4.1, the set of values correctly predicted for the positive class is called true positives, and the set of values correctly predicted for the negative class is called true negatives. The set of values of the positive class that is falsely predicted to be negative is called false negatives, and the set of values of the negative class that is falsely predicted to be positive is called false positives. The numbers of true-positive, true-negative, false-positive, and false-negative observations are denoted by the abbreviations TP, TN, FP, and FN respectively (Ruuska et al., 2018).

Table 4.1 Binary confusion matrix

		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP	FP
	Negative	FN	TN

Accuracy:

Accuracy assessment is the most basic and simple assessment criterion. The accuracy value is obtained by dividing the number of correctly predicted values by the number of all predictions.

$$Accuracy\ \%\ (M) = \left(\frac{TP + TN}{N}\right) * 100 \tag{6}$$

The accuracy model is useful if there is an even distribution among the number of classes. If there are large inequalities between the ratio of classes, checking using different criteria will provide a more accurate assessment.

Precision:

The precision assessment criterion is the ratio of the number of correctly predicted positives to the total number of positive predictions. Precision gives the ratio of the truly positive predictions of the classification model to the total positive predictions. In other words, it shows what percentage of positive predictions are correct.

$$Precision\ \%\ (M) = \left(\frac{TP}{TP + FP}\right) * 100 \quad (7)$$

Recall:

The recall evaluation criterion is the ratio of the number of true positives to the total number of true positives. The total number of true positives is obtained by adding the true positives to the false negatives. It shows the percentage of positive values that were predicted correctly.

$$Recall\ \%\ (M) = \left(\frac{TP}{TP + FN}\right) * 100 \quad (8)$$

F1-Score:

The F1-Score is a classification model evaluation criterion obtained by taking the harmonic mean of the sensitivity and recall criteria. This criterion is widely used in uneven distributions where a class is small.

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (9)$$

5 REAL-LIFE APPLICATION

The purpose of this study is to create a supervised ML-based classification model to predict the future performance of candidates after they are hired based on a real dataset to support recruiters and HR decision-makers of a private bank serving in the financial system in Turkey. Within the scope of this study, there has been a change in the employee performance evaluation methodology of the bank, valid in 2018 and beyond. For this reason, data for 2017 and before has not been considered. Since the employee performance data for 2021 has not been published at the time of the study, these data have also been excluded from the scope. So only 2018, 2019, and 2020 data have been taken into account. In addition, while the performances of the employees in the study have been labeled as successful and unsuccessful, they have been determined according to the evaluation forms within the first two years to see the effect of historical data on their future performance. Structured Query Language (SQL) has been used to obtain the dataset. Python programming language has been used to analyze data, run ML models, calculate evaluation criteria and perform parameter optimization. The outputs of the proposed methodology are given in the following sections.

5.1 Data Collection and Preparation

This section describes the steps taken from collecting data to making it ready for processing for models. These steps are usually the one that takes the longest time in ML processes and needs the most attention in the process. The better the quality and validity of the data, the better the modeling results. In addition, it requires a high amount of expertise. In our study, this process is structured in five steps described below:

Step 1: Meetings have been held with business partners to understand the needs in the recruitment process and to provide information about the objectives of the study. In these meetings, the data records of the personnel have also been evaluated. Expert opinions have been received about which data are requested from the candidates during the recruitment process and which of these data may affect the performance of the employee after hiring.

Step 2: Input and output attributes have been defined in cooperation with the bank's business partners. In addition, some literature studies have also been used to determine the attributes (Chien and Chen 2007; Chien and Chen 2008; Sivaram and Kamar 2011; Azar et al., 2013; Kirimi and Moturi 2016; Nasr et al., 2019; Mahmoud et al., 2019; Lather et al., 2019). While categorizing the numerical variables of age and school importance, the information value (IV) calculation has been used. An example categorization for the age attribute is as in the following Figure 5.1.

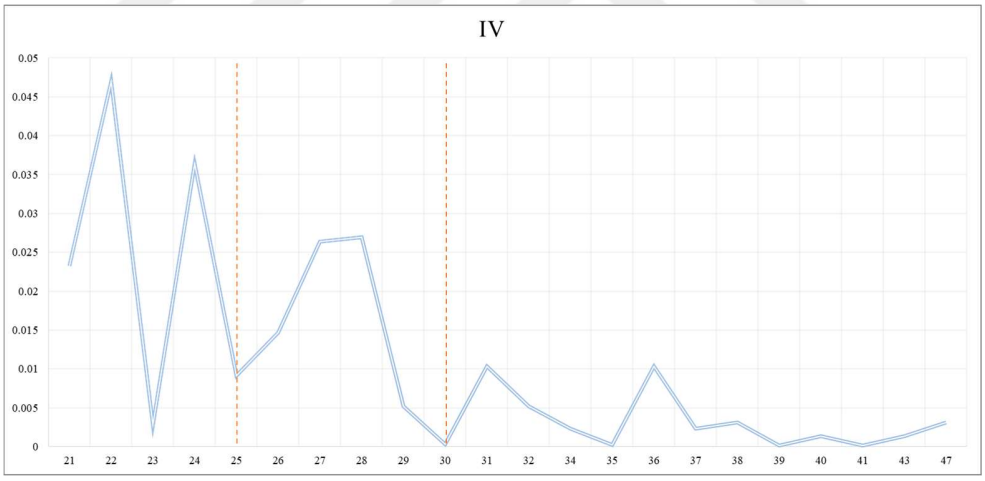

Figure 5.1 Categorization of the age attribute with IV

All these attributes have been used to predict whether the target class (performance of employees of the bank) will unsuccessful or successful. These attributes, their definitions, expected values, and data types are compiled in detail in Table 5.1.

Table 5.1 The attributes used in the model and their possible values

Attribute	Description	Expected Values	Type
Age	It refers to the age of the employee at the time of employment. This data is divided into 4 categories.	{<25 = A, 25-29 = B, >30 = C}	Input
Gender	It indicates the gender of the employee.	{Male, Female}	Input
Marital Status	It refers to the status of the employee as married or single.	{Married, Single}	Input
Child Status	It tells you whether you have children or not.	{Yes, No}	Input
Military Service Status	It shows whether the employee has done his military service or not. Women employees are exempted.	{Completed, Exempted, Not Completed}	Input
Job Family	There are business families in the company on which each position depends. It is the job family information of the position from which the employee was hired.	{Branch_Sales, Branch_ Operations, Head Office, IT}	Input
Previous Experience	It refers to the previous work experience of the employee. If it is more than 1 year, it is labeled as "yes", otherwise "no".	{Yes, No}	Input
Intern Status	It is the information on whether the employee works part-time in the company before starting a full-time job.	{Yes, No}	Input
Grade Point Average	University undergraduate graduation grade point average.	{Low, Medium, High, Excellent}	Input
University Importance	It is the information that the undergraduate graduation university is scored according to its academic success.	{Low, Medium, High}	Input
Foreign Language Certificate	Whether it has an English or Arabic language certificate	{Yes, No}	Input
Reference Information	Whether the employee has a reference from the company when applying to the company.	{Yes, No}	Input

Attribute	Description	Expected Values	Type
Loan	Whether the employee was a customer before applying for a job with the company.	{0, 1}	Input
Customer Status	Whether the employee was a customer before applying for a job with the company.	{Yes, No}	Input
Performance	Performance score.	{Unsuccessful, Successful}	Output

Step 3: According to the bank's HR historical records, the past candidate forms of the employees hired in 2018, 2019, and 2020 have been examined. The attributes that have been determined in Step 2 have been obtained from the database with the SQL programming language. Relevant records and performance scores have been evaluated and examined. The records that could be thought to be incorrect have been cleared. Attributes containing blank data (such as school importance) have been filled in by looking at the data of employees with similar profiles to prevent data loss.

In addition, the performance score of the employees who left without performance evaluation has been produced by considering the way they left the job. The company has three types of layoffs. These are resignation, termination, and others. To increase the significance of the model, resignation and other types of employees have been excluded from the data set. Because employees can be successful but quit their jobs without any reason. The performance scores have been considered unsuccessful if the employees are dismissed due to termination.

Finally, a data set consisting of 14 attributes and 597 employee data has been used to create a model. An example section of the dataset can be viewed in Table 5.2. The detailed table for the types and frequency values of the variables related to the selected data is given in Appendix A.

Table 5.2 Example of the limited data set used in the model

#	Age	Gender	Marital Status	Child Status	Military Service Status	Job Family	…	Intern Status	Grade Point Average	Foreign Language Certificate	Reference Information	Loan	Performance
0	B	Male	Single	No	Exempted	Branch_Operation	…	No	Medium	No	Yes	No	Unsuccessful
1	A	Male	Single	No	Not Completed	Branch_Operation	…	No	High	No	No	No	Unsuccessful
2	B	Male	Single	No	Not Completed	Branch_Operation	…	No	Medium	No	Yes	No	Unsuccessful
3	C	Male	Married	Yes	Completed	IT	…	Yes	High	No	No	No	Successful
4	B	Male	Single	No	Not Completed	Head_Office	…	No	Medium	No	No	No	Successful
5	A	Male	Single	No	Not Completed	Head_Office	…	No	Excellent	No	Yes	No	Successful
6	A	Male	Married	No	Not Completed	Branch_Sales	…	Yes	Medium	No	Yes	No	Successful
7	B	Male	Single	No	Exempted	Head_Office	…	No	High	No	Yes	No	Successful
8	B	Male	Single	No	Not Completed	Head_Office	…	No	Medium	Yes	Yes	Yes	Successful
9	B	Male	Single	No	Not Completed	Head_Office	…	No	Medium	Yes	No	Yes	Unsuccessful
10	A	Female	Single	No	Exempted	Head_Office	…	No	High	Yes	No	No	Successful

Step 4: Non-numeric categorical data columns are digitized so that the data can be recognized by the ML model and mathematical operations can be performed on it. Some algorithms can work directly with categorical data. For example, a DT can be learned directly from categorical data (but the Scikit learn library used does not support this yet) without the need for data transformation. However, many ML algorithms cannot work directly on non-numeric data labels. For this reason, 0-1 labeling for binary attributes and one hot encoding method for other categorical variables was used with the codes given in Appendix B.

Step 5: Correlation is a widely applied technique in ML for data analysis. Highly correlated features can harm the learning and generalization of models. For this reason, correlation analysis has been performed between the features by using the codes in Appendix C. The "age" attribute, which was observed to be highly correlated with other variables, has not been included in ML.

By applying all these steps, the data set has been made ready for training in models. In addition to this, some data has been added for information purposes and will not be trained in the ML model. These are the "gender" and "marital status" columns in the dataset. The aim is to prevent prejudice, positive or negative distinctions that may occur in recruitment processes, as has been widely observed in the literature recently.

5.2 Machine Learning Modeling

After the preparation and preprocessing of the data, the necessary python libraries have been loaded first to run the ML models used in the application and to evaluate them. The basic library used for ML in this study is Scikit-learn. After library uploads, the data set has been divided into two as inputs and outputs for ML. Inputs are data that the model will use to make sense of the output (performance) in the decision column.

70% of the data partitioned as input and output have been reserved for training and validation. 30% of the remaining dataset has been used for the final performance evaluation.

The performances of the algorithms to be trained in the application have been calculated by performing five-fold validation. The average of the Accuracy, Precision, Recall, and F1-score evaluation criteria have been calculated for each layer. In the analysis of the data, K Neighbors Classifier, Random Forest Classifier, Logistic Regression, DT Classifier, MLP, Categorical Naive Bayes, and SVM - Linear Kernel models have been used. The codes for the calculations are shared in Appendix D.

5.3 Models Evaluation

The labeled training data has been trained with supervised ML models such as Logistic Regression, SVM, K Neighbors Classifier, DT Classifier, and MLP. The outputs of the five-fold calculated model are presented in Table 5.3.

Table 5.3 The results and averages for each fold of the executed algorithms

Model	Fold	Accuracy	Precision	Recall	F1-score
Logistic Regression	1	0.7417	0.8750	0.6269	0.7304
	2	0.6667	0.6522	0.7377	0.6923
	3	0.8067	0.8784	0.8228	0.8497
	4	0.6807	0.7000	0.6034	0.6481
	5	0.6639	0.5862	0.9273	0.7183
	Mean	**0.7119**	**0.7384**	**0.7436**	**0.7278**
SVM	1	0.7167	0.8511	0.5970	0.7018
	2	0.7167	0.6957	0.7869	0.7385
	3	0.7479	0.8889	0.7089	0.7887
	4	0.6723	0.6727	0.6379	0.6549
	5	0.6891	0.6023	0.9636	0.7413
	Mean	**0.7085**	**0.7421**	**0.7389**	**0.7250**
K Neighbors Classifier	1	0.6917	0.7778	0.6269	0.6942
	2	0.6917	0.6667	0.7869	0.7218
	3	0.7311	0.8507	0.7215	0.7808
	4	0.5882	0.5882	0.5172	0.5505
	5	0.7059	0.6351	0.8545	0.7287
	Mean	**0.6817**	**0.7037**	**0.7014**	**0.6952**
Decision Tree Classifier	1	0.7417	0.8462	0.6567	0.7395
	2	0.6667	0.6721	0.6721	0.6721
	3	0.7143	0.8358	0.7089	0.7671

Model	Fold	Accuracy	Precision	Recall	F1-score
	4	0.6555	0.6889	0.5345	0.6019
	5	0.6471	0.5844	0.8182	0.6818
	Mean	**0.6968**	**0.7312**	**0.6976**	**0.7060**
Multilayer Perceptron	1	0.7500	0.8364	0.6866	0.7541
	2	0.6667	0.6615	0.7049	0.6825
	3	0.7983	0.8873	0.7975	0.8400
	4	0.6303	0.6346	0.5690	0.6000
	5	0.6639	0.5824	0.9636	0.7260
	Mean	**0.7034**	**0.7303**	**0.7415**	**0.7237**

As a result of five-fold validation, the logistic regression model with the best training performance (Accuracy = 71.19%) has been selected.

5.4 Improving the Success of the Preferred Model

Hyperparameter tuning severely affects the performance and speed of ML models. In the selected model, the default parameters of the logistic regression algorithm have been used. To improve the performance of the model, the parameter values of the algorithm have been changed and an improvement has been achieved. The default parameters of the selected model are “penalty =l2”, “C=1”, and “solver=lbfgs”.

The suggested parameter set for the logistic regression model is as follows: solvers = {'newton-cg', 'lbfgs', 'liblinear'}, penalty = {'l2', 'l1'} and c = {100, 10, 5, 3, 1.0, 0.1, 0.01}. The accuracy values have been calculated by applying five folds to all combinations of the parameter set. Incorrect combinations are not considered as some penalties do not work with some solvers. The codes used in the calculations in Table 5.4 are shared in Appendix E.

Table 5.4 Logistic regression hyperparameter tuning results (sorted by accuracy)

Accuracy	C	Penalty	Solver
0.7314	3	l1	liblinear
0.7291	100	l2	newton-cg
0.7291	100	l2	lbfgs
0.7291	100	l2	liblinear

Accuracy	C	Penalty	Solver
0.7291	100	l1	liblinear
0.7291	1	l2	newton-cg
0.7291	1	l2	lbfgs
0.7291	1	l2	liblinear
0.7290	3	l2	newton-cg
0.7290	3	l2	lbfgs
0.7290	3	l2	liblinear
0.7266	10	l2	newton-cg
0.7266	10	l2	lbfgs
0.7266	10	l2	liblinear
0.7266	10	l1	liblinear
0.7266	5	l2	newton-cg
0.7266	5	l2	lbfgs
0.7266	5	l2	liblinear
0.7266	5	l1	liblinear
0.7266	0.1	l2	liblinear
0.7242	1	l1	liblinear
0.7194	0.1	l2	newton-cg
0.7194	0.1	l2	lbfgs
0.7123	0.1	l1	liblinear
0.6930	0.01	l2	liblinear
0.6809	0.01	l2	newton-cg
0.6809	0.01	l2	lbfgs
0.4750	0.01	l1	liblinear

The parameters with the best accuracy are “c = 3”, “penalty = l1”, and “solver =liblinear”. By applying five folds in these parameters, an accuracy value of 73.14% has been achieved. With the default parameter values of the logistic regression algorithm, 71.19% accuracy has been obtained. 1.95% improvement has been achieved with the best parameter values of the developed model.

Figure 5.2 shows the learning curve of the Logistic Regression algorithm on the training data with the relevant parameters to monitor the learning performance of the model. It is seen that the model continues to learn with the training data. It is seen that increasing the number of samples can provide further improvement in the score of the model.

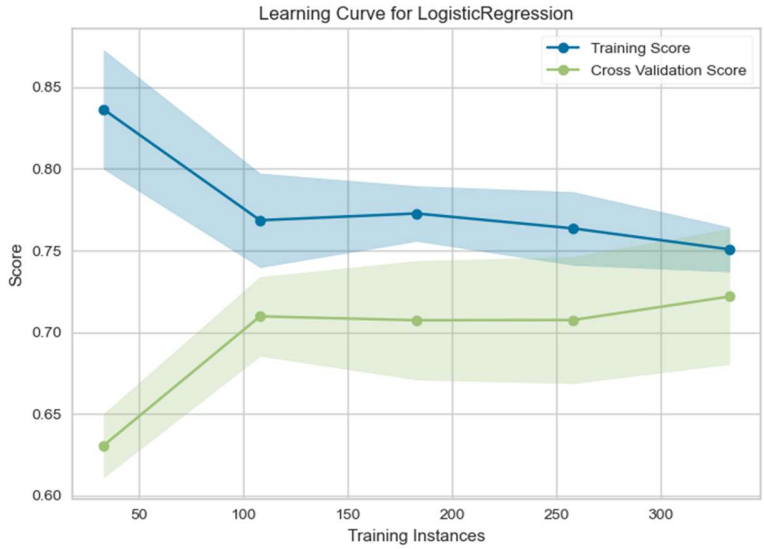

Figure 5.2 Learning curve for the proposed model

Parameter optimization has been performed on the best algorithm selected by training and validation operations on the training data. Finally, the model has been tested with a data set that it has not seen before. So, the test data set that was partitioned before training and validation has been used. The success of the developed model in the test data set has been obtained with good accuracy of 71.67%. Related codes are included in Appendix F.

Figure 5.3 shows the confusion matrix for the developed model. This allows us to gain more detailed information about the prediction mechanism of the model to evaluate the behavior and understand the effectiveness of the proposed model.

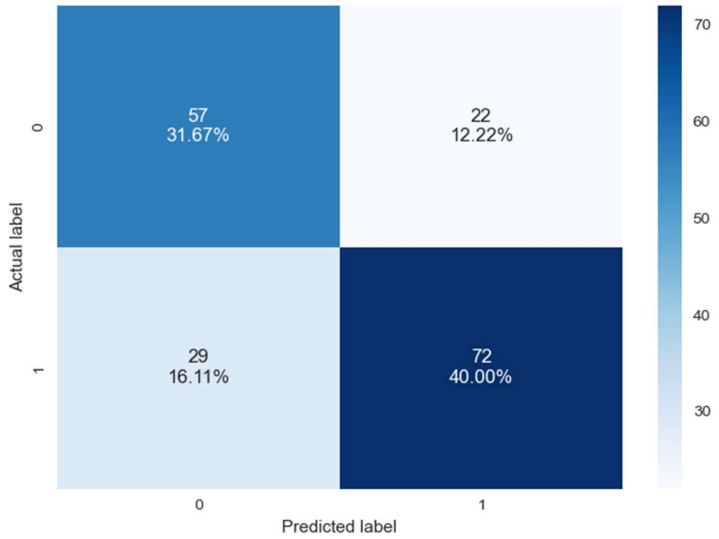

Figure 5.3 Confusion matrix for the proposed model

According to the confusion matrix in Figure 5.3, 57 employees who failed in the proposed model have been predicted as unsuccessful (TN). 72 employees who succeeded have been predicted to be successful (TP). There have been 29 employees who actually succeeded but the model mistakenly predicted (FN) as unsuccessful. There have been 22 employees who succeeded and the model mistakenly predicted them as successful (FP).

The proposed model's accuracy, precision, recall, and F1-score values are 71.67%, 76.6%, 71.27%, and 73.84%, respectively.

In addition, the precision value is especially important when the cost of false-positive estimation is high. It would be a costlier mistake for the model to predict a truly unsuccessful employee for recruitment as successful. In the proposed model, predicting a successful employee to be unsuccessful with a wrong prediction is preferable to evaluating an unsuccessful employee as successful with a wrong prediction. In this context, the high precision value (76.6%) has been a supportive measure for the proposed model.

6 CONCLUSIONS

Companies need to employ successful employees to gain a competitive advantage and high profitability. It has been supported by many studies that the ROIs made in recruitment processes are quite high. For this reason, the recruitment processes, which are handled with the traditional approach, should be supported by approaches with technology and data-oriented processes, and focus on employing the right employee. Studies in this area have gained momentum with the interest in HR analytics.

In recruitment processes, the development of ML-based systems can provide great savings in terms of financial resources and time. In this study, an analytical framework has been presented that can be used as a decision support tool to predict the post-hire success of the employees by analyzing the historical data in the recruitment processes and recruiting suitable candidates efficiently.

Within the scope of the study, the anonymous data of 597 employees of a private bank serving in Turkey have been examined. Past recruitment data of successful and unsuccessful employees have been evaluated together with the company's business partners. Subsequently, possible attributes that can predict the performance of the employee after recruitment have been determined. A supervised ML approach has been applied in this study. Preparation processes have been applied for the data set to be trained in ML. 70% of the data partitioned as input and output have been reserved for training and validation. 30% of the remaining dataset has been used for the final performance evaluation.

Firstly, the labeled training data has been trained with supervised ML models such as Logistic Regression, SVM, K Neighbors Classifier, DT Classifier, and MLP. As a result of five-fold validation, the logistic regression model with the best training performance (Accuracy = 71.19%) has been selected. Secondly, parameter optimization has been performed to improve the proposed model. With the best parameter values of the developed model, an accuracy rate of 73.14% has been achieved. Finally, the developed model has been tested with the data set that the model had not seen before, which has been separated as the test set. The proposed model achieved a successful accuracy score of 71.67% on the test data.

This study presents the first step in efforts to predict the future performance of pre-employment candidates. The proposed approach is a decision support system and it aims to support the final decision-makers in reducing the complexity of the recruitment processes and hiring the right employees.

In future studies, the accuracy score of the model can be increased by expanding the data set and adding new features (such as personality tests) that can predict performance. The study includes a scope based on the predictive analysis approach. This scope can be expanded with the prescriptive analysis approach and interpretable outputs can be presented for decision-makers.

REFERENCES

Arora, S., Agarwal, M., & Kawatra, R. (2020, March). Prediction of educationist's performance using a regression model. In 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 88-93). IEEE.

Azar, A., Sebt, M. V., Ahmadi, P., & Rajaeian, A. (2013). A model for personnel selection with a data mining approach: A case study in a commercial bank. SA Journal of Human Resource Management, 11(1), 1-10.

Bach, M. P., Simic, N., & Merkac, M. (2013). Forecasting employees' success at work in banking: could psychological testing be used as the crystal ball?. Managing Global Transitions, 11(3), 283.

Ben-Gal, H. C. (2019). An ROI-based review of HR analytics: practical implementation tools. Personnel Review.

Calixto, N., & Ferreira, J. (2020). Salespeople performance evaluation with predictive analytics in B2B. Applied Sciences, 10(11), 4036.

Chaudhry, A., & Usman, A. (2011). An investigation of the relationship between employees' emotional intelligence and performance. African Journal of Business Management, 5(9), 3556-3562.

Chien, C. F., & Chen, L. F. (2007). Using rough set theory to recruit and retain high-potential talents for semiconductor manufacturing. IEEE Transactions on Semiconductor Manufacturing, 20(4), 528-541.

Chien, C. F., & Chen, L. F. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. Expert Systems with applications, 34(1), 280-290.

Chuang, Y. C., Hu, S. K., Liou, J. J., & Tzeng, G. H. (2020). A data-driven MADM model for personnel selection and improvement. Technological and Economic Development of Economy, 26(4), 751-784.

Dağdeviren, M. (2010). A hybrid multi-criteria decision-making model for personnel selection in manufacturing systems. Journal of Intelligent manufacturing, 21(4), 451-460.

Delgado-Gómez, D., Aguado, D., Lopez-Castroman, J., Santacruz, C., & Artés-Rodriguez, A. (2011). Improving sale performance prediction using support vector machines. Expert systems with applications, 38(5), 5129-5132.

Dhliwayo, P., & Coetzee, M. (2020). Cognitive intelligence, emotional intelligence and personality types as predictors of job performance: Exploring a model for personnel selection. SA Journal of Human Resource Management, 18, 13.

Du, L., & Li, Q. (2020). A Data-Driven Approach to High-Volume Recruitment: Application to Student Admission. Manufacturing & Service Operations Management, 22(5), 942-957.

Harris, C. G. (2018). Making Better Job Hiring Decisions using" Human in the Loop" Techniques. In HumL@ ISWC (pp. 16-26).

Isson, J. P., & Harriott, J. S. (2016). People analytics in the era of big data: Changing the way you attract, acquire, develop, and retain talent. John Wiley & Sons.

Jain, P., & Jain, P. (2020). Understanding the Concept of HR Analytics. International Journal on Emerging Technologies, 11(2), 644-652.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Jantan, H., Hamdan, A. R., & Othman, Z. A. (2011). Towards applying data mining techniques for talent management. In International Conference on Computer Engineering and Applications, IPCSIT (Vol. 2, p. 2011).

Jantan, H., Puteh, M., Hamdan, A. R., & Ali Othman, Z. (2010). Applying data mining classification techniques for employee's performance prediction.

Karam, S., Nagahi, M., Dayarathna, V. L., Ma, J., Jaradat, R., & Hamilton, M. (2020). Integrating systems thinking skills with multi-criteria decision-making technology to recruit employee candidates. Expert Systems with Applications, 160, 113585.

Kirimi, J. M., & Moturi, C. A. (2016). Application of data mining classification in employee performance prediction. International Journal of Computer Applications, 146(7), 28-35.

Lather, A. S., Malhotra, R., Saloni, P., Singh, P., & Mittal, S. (2019, November). Prediction of employee performance using machine learning techniques. In Proceedings of the International Conference on Advanced Information Science and System (pp. 1-6).

Li, N., Kong, H., Ma, Y., Gong, G., & Huai, W. (2016). Human performance modeling for manufacturing based on an improved KNN algorithm. The International Journal of Advanced Manufacturing Technology, 84(1-4), 473-483.

Lievens, F., Van Dam, K., & Anderson, N. (2002). Recent trends and challenges in personnel selection. Personnel review.

Madsen, D. Ø., & Slåtten, K. (2017). The rise of HR analytics: A preliminary exploration. In Global Conference on Business and Finance Proceedings (Vol. 12, No. 1, pp. 148-159).

Mahmoud, A. A., Shawabkeh, T. A., Salameh, W. A., & Al Amro, I. (2019, June). Performance predicting in the hiring process and performance appraisals using machine learning. In 2019 10th International Conference on Information and Communication Systems (ICICS) (pp. 110-115). IEEE.

Marler, J. H., & Boudreau, J. W. (2017). An evidence-based review of HR Analytics. The International Journal of Human Resource Management, 28(1), 3-26.

Menon, V. M., & Rahulnath, H. A. (2016, September). A novel approach to evaluate and rank candidates in a recruitment process by estimating emotional intelligence through social media data. In 2016 International Conference on Next Generation Intelligent Systems (ICNGIS) (pp. 1-6). IEEE.

Nasr, M., Shaaban, E., & Samir, A. (2019). A proposed Model for Predicting Employees' Performance Using Data Mining Techniques: Egyptian Case Study. no. February.

Pessach, D., Singer, G., Avrahami, D., Ben-Gal, H. C., Shmueli, E., & Ben-Gal, I. (2020). Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. Decision Support Systems, 134, 113290.

Rouyendegh, B. D., & Erkart, T. E. (2012). SELECTION OF ACADEMIC STAFF USING THE FUZZY ANALYTIC HIERARCHY PROCESS(FAHP): A PILOT STUDY. Tehnicki vjesnik, 19(4), 923-929.

Ruuska, S., Hämäläinen, W., Kajava, S., Mughal, M., Matilainen, P., & Mononen, J. (2018). Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behavior of cattle. Behavioral processes, 148, 56-62.

Santhosh Kumar, S., Mohanapriya, G., & Shanmugapriya, M. M. (2020). A STUDY ON SOME PROPERTIES OF Q-FUZZY NORMAL SUBGROUPS. Journal of Critical Reviews, 7(12), 2818-2821.

Santiago, E. B., & Gara, G. P. P. A Model-Based Prediction of Desirable Applicants through Employee's Perception of Retention and Performance. In 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM) (pp. 1-6). IEEE.

Sarker, A., Shamim, S. M., Zama, M. S., & Rahman, M. M. (2018). Employee's performance analysis and prediction using K-means clustering & decision tree algorithm. Global Journal of Computer Science and Technology.

Shehu, M. A., & Saeed, F. (2016). An adaptive personnel selection model for recruitment using domain-driven data mining. Journal of Theoretical and Applied Information Technology, 91(1), 117.

Sivaram, N., & Ramar, K. (2011). Knowledge Engineering to aid the recruitment process of an Industry by identifying superior selection criteria. ICTACT Journal on soft computing.

Thakur, G. S., Gupta, A., & Gupta, S. (2015). Data mining for prediction of human performance capability in the software-industry. arXiv preprint arXiv:1504.01934.

Valle, M. A., Varas, S., & Ruz, G. A. (2012). Job performance prediction in a call center using a naive Bayes classifier. Expert Systems with Applications, 39(11), 9939-9945.

Vlachos, I. (2008). The effect of human resource practices on organizational performance: evidence from Greece. The international journal of Human resource management, 19(1), 74-97.

Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001, June). Constrained k-means clustering with background knowledge. In Icml (Vol. 1, pp. 577-584).

Xue, X., Feng, J., Gao, Y., Liu, M., Zhang, W., Sun, X., ... & Guo, S. (2019). Convolutional recurrent neural networks with a self-attention mechanism for personnel performance prediction. Entropy, 21(12), 1227.

Zhang, W., Chen, X., Liu, Y., & Xi, Q. (2020). A distributed storage and computation k-nearest neighbor algorithm based cloud-edge computing for cyber-physical-social systems. IEEE Access, 8, 50118-50130.

APPENDICES

Appendix A.

Attribute	Data Type	Non-Null Count	Unique Value Count	Top Category	Top Category Frequency
Age	Object	597	3	A	285
NumberOfChild	Object	597	2	No	566
Soldiery_Status	Object	597	3	Not Completed	303
Job_Family	Object	597	4	Branch_Operation	250
Previous_Experience	Object	597	2	No	538
Intern_Status	Object	597	2	No	364
Grade_Point_Average	Object	597	4	Medium	237
School_Importance	Object	597	3	High	257
Language_Certificate	Object	597	2	No	502
Reference	Object	597	2	Yes	302
Loan	Object	597	2	No	426
Customer_Status	Object	597	2	No	300
Performance	Object	597	2	Successful	320

Appendix B.

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.neural_network import MLPClassifier
from yellowbrick.model_selection import learning_curve
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import KFold
from sklearn.metrics import precision_score, recall_score,
confusion_matrix,ConfusionMatrixDisplay, accuracy_score, f1_score

df = pd.read_excel('Raw_Data.xlsx')
data = pd.DataFrame(df)

data = data.concat([data, data['Soldiery_Status'].str.get_dummies(sep="",
").add_prefix("Soldiery_Status_")], axis=1)
data.drop(columns=['Soldiery_Status'],axis = 1,inplace=True)

data = data.concat([data, data['Job_Family'].str.get_dummies(sep="",
").add_prefix("Job_Family_")], axis=1)
data.drop(columns=['Job_Family'],axis = 1,inplace=True)

data = pd.concat([data, data['Grade_Point_Average'].str.get_dummies(sep="",
").add_prefix("Grade_Point_Average_")], axis=1)
```

```
data.drop(columns=['Grade_Point_Average'],axis = 1,inplace=True)

data = pd.concat([data, data['School_Importance'].str.get_dummies(sep='',
').add_prefix("School_Importance_")], axis=1)
data.drop(columns=['School_Importance'],axis = 1,inplace=True)
```

Appendix C.

```
sns.set(font_scale=0.75)
correlation_X = X.corr()
mask = np.triu(correlation_X.corr())
plt.figure(figsize=(20, 20))
sns.heatmap(correlation_X,vmin=-1, vmax=1, center=0,
    annot=True,
    fmt='.1f',
    cmap='coolwarm',
    square=True,
    mask=mask,
    linewidths=1)
plt.show()
```

Appendix D.

```
def model_get_scores(model, X_train, X_test, y_train, y_test):
    model.fit(X_train, y_train)
    y_predict = model.predict(X_test)
    accuracy = accuracy_score(y_test, y_predict)
    precision = precision_score(y_test, y_predict)
    recall = recall_score(y_test, y_predict)
    f1score = f1_score(y_test, y_predict)
    return accuracy, precision, recall, f1score

folds = KFold(n_splits=5)
logistic_scores= []
SVM_scores = []
random_forest_scores = []
k_neighbors_scores = []
decision_tree_scores = []
MLP_scores = []


for train_index, test_index in folds.split(X_train,y_train):
    X_train, X_test, y_train, y_test = X.iloc[train_index,:], X.iloc[test_index,:],
y.iloc[train_index], y.iloc[test_index]
    logistic_scores.append(model_get_scores(LogisticRegression(), X_train, X_test, y_train,
y_test))
    SVM_scores.append(model_get_scores(SVC(), X_train, X_test, y_train, y_test))
    random_forest_scores.append(model_get_scores(RandomForestClassifier(), X_train,
X_test, y_train, y_test))
    k_neighbors_scores.append(model_get_scores(KNeighborsClassifier(), X_train, X_test,
y_train, y_test))
```

```
    decision_tree_scores.append(model_get_scores(DecisionTreeClassifier(), X_train,
X_test, y_train, y_test))
    MLP_scores.append(model_get_scores(MLPClassifier(), X_train, X_test, y_train,
y_test))

logistic_scores = pd.DataFrame (logistic_scores, columns =
['accuracy','precision','recall','f1score'])
SVM_scores = pd.DataFrame (SVM_scores, columns =
['accuracy','precision','recall','f1score'])
random_forest_scores = pd.DataFrame (random_forest_scores, columns =
['accuracy','precision','recall','f1score'])
k_neighbors_scores = pd.DataFrame (k_neighbors_scores, columns =
['accuracy','precision','recall','f1score'])
decision_tree_scores = pd.DataFrame (decision_tree_scores, columns =
['accuracy','precision','recall','f1score'])
MLP_scores = pd.DataFrame (MLP_scores, columns =
['accuracy','precision','recall','f1score'])
```

Appendix E.

```
solvers_list = ['lbfgs', 'newton-cg', 'liblinear']
penalty_list = ['l2','l1']
c_list = [100, 10, 5, 3, 1.0, 0.1, 0.01]

grid = dict(solver=solvers_list,penalty=penalty_list,C=c_list)
cv = KFold(n_splits=5)
grid_search = GridSearchCV(estimator=LogisticRegression(), param_grid=grid, n_jobs=-1,
cv=cv, scoring='accuracy',error_score=0)
result = grid_search.fit(X_train, y_train)
means = result.cv_results_['mean_test_score']
parameters = result.cv_results_['params']

for mean, parameters in zip(means, parameters):
   print("%f (%f) with: %r" % (mean, parameters))
```

Appendix F

```
LR_with_best_parameters = LogisticRegression(C=3,penalty='l1',solver='liblinear')
learning_curve(LR_with_best_parameters, X_train, y_train, cv=5, scoring='accuracy')

model = LogisticRegression(C=3,penalty='l1',solver='liblinear')
model.fit(X_train, y_train)
y_predict = model.predict(X_test)
accuracy = accuracy_score(y_test, y_predict)
precision = precision_score(y_test, y_predict)
recall = recall_score(y_test, y_predict)
f1score = f1_score(y_test, y_predict)

confusion_matrix = confusion_matrix(y_test, y_predict)
sns.heatmap(confusion_matrix, annot=True,cmap='Blues')
plt.show()
```

BIOGRAPHICAL SKETCH