Predicting Employees Turnover in Factory using Machine Learning: The Case of Adama Industrial Park

Rediet Yonas Tamene

A Thesis Submitted to the Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Present in Partial Fulfillment of the Requirement for the Degree of Master's in

Computer Sciences and Engineering

Office of Graduate Studies

Adama Science and Technology University

February 2025

Adama, Ethiopia

Predicting Employees Turnover in Factory Using Machine Learning: The Case of Adama Industrial Park

By

Rediet Yonas Tamene

Advisor

Dr. Tilahun Melak

A Thesis submitted to the Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Office of Graduate Studies

Adama Science and Technology University

February 2025

Adama, Ethiopia

DECLARATION

I hereby declare that this Master Thesis entitled **"Predicting Employees Turnover in Factory Using Machine Learning: The Case of Adama Industrial Park"** is my original work. That is, it has not been submitted for the award of any academic degree, diploma or certificate in any other university. All sources of materials that are used for this thesis have been duly acknowledged through citation.

Rediet Yonas Tamene		
Name of student	**Signature**	**Date**

RECOMMENDATION OF ADVISORS

I, the advisor of this thesis, hereby certify that I have read the revised version of thesis entitled **"Predicting Employees Turnover in Factory Using Machine Learning: The Case of Adama Industrial Park"** prepared under my guidance by **Rediet Yonas** submitted in partial fulfillment of the requirements for the degree of masters of science in computer science and Engineering. Therefore, I recommend the submission of the revised version of thesis to the department following the applicable procedures.

Dr. Tilahun Melak	_____	_____
Major advisor	Signature	Date
_____	_____	_____
Co advisor	Signature	Date

APPROVAL PAGE

I the advisors of the thesis entitled "**Predicting Employees Turnover in Factory Using Machine Learning: The Case of Adama Industrial Park"** and developed by **Rediet Yonas,** hereby certify that the recommendation and suggestions made by the board of examiners and appropriately incorporated in to the final version of the thesis.

Dr Tilahun Melak		
Major advisor	Signature	Date
Co advisor	Signature	Date

We, the undersigned, members of the Board of Examiners of the thesis by **Rediet Yonas** have read and evaluated the thesis entitled **"Predicting Employees Turnover in Factory Using Machine Learning: The Case of Adama Industrial Park"** and examined the candidate during open defense. This is, therefore, to certify that the **thesis** is accepted for partial fulfillment of the requirement of the degree of Master of Science in Computer Science and Engineering.

Chairperson	Signature	Date

Finally, approval and acceptance of the thesis is contingent upon submission of its final copy to the Office of Postgraduate Studies (OPGS) through the Department Graduate Council (DGC) and School Graduate Committee (SGC).

Department Head	Signature	Date
School Dean	Signature	Date
Office of Postgraduate Studies, Dean	Signature	Date

ACKNOWLEDGEMENT

First of all, I would like to thank the Almighty of GOD and His mother, Saint Mary, for guiding and supporting me throughout this MSc thesis. I would like to express my appreciation to my advisor Tilahun Melak (PHD) for his commitment to supporting me and guiding me while conducting this Master's thesis work from beginning to end. I would like to thank my family for their unwavering support and encouragement throughout my MSc thesis journey. Their belief in my abilities has helped me overcome changes along the way. I am also grateful to my friends for their support. Finally, I would like to express my thanks to Adama Industrial Park for their support in providing all the important and necessary documents that are vital for conducting this MSc thesis.

TABEL OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONOMY AND ABBREVATION

ADASYN	Adaptive Synthetic Sampling
AI	Artificial intelligence
ANOVA	Analysis of Variance
AUC	Area under the curve
CNN	Conventional neural network
CPU	Central Processing unit
CSV	Comma Separated Values
DNN	Deep Neural Network
DT	Decision Trees
DTC	Decision Tree Classifier
EEDA	Employee Exploratory Data Analysis
ETC	Extra Trees Classifier
ERCA	Ethiopian Revenues and Customs Authority
FN	False Negative
FP	False Positive
GIC	Global Insurance Company
GNB	Gaussian Naïve Bayes
HR	Human Resource
HRM	Human Resource Management
IBM	International Business Machine Corporation
IPDC	Industrial Park Development Corporation
KNN	K-Nearest Neighbors
LR	Logistic Regression
LSTM	Long short term memory
ML	Machine Learning
MLP	Multi-layer perceptron
NB	Naive Bayes
PC	Personal Computer
RF	Random Forests
RFE	Recursive feature elimination

ROC	Receiver operating character
SBS	Sequential Backward Selection
SOMTE	Synthetic Minority oversampling technique
SPSS	Statistical Package for Social Science
SVM	Support vector
TN	True Negative
TP	True Positive
VC	Voting Classifier
XGBoost	Extreme Gradient Boosting

ABSTRACT

Employee turnover is a critical topic in the current human resource management literature. Turnover is the leaving of employees from their jobs and it occurs negatively influences not only financial and operational performance but also organizational performance and stability. Predicting employee turnover in this context is of great importance since it might indicate to the employer how and for what reasons employees may leave the organization and thus may allow steps to be taken to reduce costly employee turnover rates. This study aims to address the growing concern of employee turnover at Adama Industrial Park by focusing on understanding the reasons why employees leave and developing an appropriate machine learning model to predict turnover. To do so we begin by collecting data from the company HR office, resulting in 16078 raw data entries with 13 features collected for four years, from 2018-2022. Explanatory data analysis techniques are applied like handling missing values, encoding categorical data, feature selecting, data binning, and data normalization to make our data interpret for the machine learning model. We use the classification model. We split the data into 12862 samples which are 80% of the total data for training and 3216 samples which are 20% of the total data for testing. The models were trained using Logistic Regression, Random Forest, and K-Nearest-Neighbor. To enhance the model accuracy, we utilized a random search CV for the tuning parameter of our model. We evaluated the model's performance with accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Random forest model outperforms the high accuracy result of 88.87% compared with logistic regression and K-Nearest Neighbor. Also, Random forest scored an average accuracy of 88.31% result using 10-fold cross-validation it is the highest score compared with logistic regression and K-Nearest Neighbor. We use the Random Forest feature importance to identify which factor affects employees to leave and, in our study, we get salary is the main reason to leave employees from their jobs. The study's outcomes suggest that addressing salary-related issues could be pivotal in reducing turnover rates at Adama Industrial Park and similar industrial settings. This research aimed to address the gap by utilizing local data, which is less studied in the case of employee turnover prediction, to provide data-driven decision-making and organizational practices for employee retention. In conclusion, this research not only provides a predictive framework for employee turnover but also emphasizes the importance of data-driven strategies in HRM, offering a practical tool for improving organizational performance and employee satisfaction.

Key words: Machine learning, Logistic regression, Random forest, K-Nearest neighbor, Employees' turnover, Cross Validation, Random Search CV, AUC-ROC

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

One of an organization’s most valuable assets is its human capital, which can play a significant part in determining its strategic success (Ekhsan, 2019) Employee turnover refers to the process of leaving or resigning manpower from one current organization to another to pursue better opportunities such as improved salary, career advancement, and so on (Nahar et al., 2022). In recent years, employee turnover has become a major issue for all kinds of firms, and it has a significant impact on overall performance and operational efficiency. costly employee turnover disrupts regular work processes, lowers employee morale, and results in costly recruitment and training costs. Effective retention plans require an understanding of the factors that contribute to employee turnover.

Employee turnover can incur costs exceeding one hundred percent of an individual employee's annual salary, which can adversely affect the productivity and profits of the organization. A very large number of employee turnover also could determine community tax collections, social programs, and physical and mental health issues. Consequently, it is vital for organizational leaders to comprehend the factors that lead to increased employee turnover, considering both business and societal implications. (Skelton et al., 2020). Employee turnover has always been an issue of concern for organizations. A large number of employee turnover may be detrimental to the organization's and employees' interests. Turnover will affect the costs of hiring and selection, the induction process and personnel process, training of new personnel, and above all, the loss of expertise acquired by the employee during their work. Moreover, it puts them in an understaffed situation in which the remaining employees become less effective and productive.

Employees are an essential resource in most organizations. Staff turnover can be catastrophic for organizations in many aspects, for example, higher new employee recruitment and training cost, lower productivity, low staff morale, increased overtime cost, additional workload to the current staff, products cannot meet deadlines and affect long term organizational development plans. Some of the causes of employee turnover are employees do not like their salary, there is no career growth, there is no job satisfaction to work, etc.

Human resources are required to operate the technology, although most companies at present are destined to be technology-based. They are the most important and versatile assets of any company. Because of the general expansion of all economic sectors, the market is extremely competitive. With all of this growth and competitiveness, human resources have lots of options and opportunities open to them. Managing and keeping these resources is the largest problem that firms are currently facing. Because employees' skills and capacity are essential to a company's capacity to compete economically, getting and keeping talented workers is crucial for any organization. Additionally, another issue that firms are dealing with nowadays is keeping staff happy. (Lahkar Das & Baruah, 2013)

Organizations invest significant resources, including training and financial support, to develop their employees' skills. Therefore, losing skilled and experienced employees can be very expensive and greatly impacts productivity. Employees who have completed the learning curve are generally more productive. When these experienced individuals, who understand the organization's internal systems and structure, leave, it often results in decreased productivity, efficiency, and responsiveness. As a result, the opportunities lost due to employee turnover can be considerable.

According to the report by (Cepheus, 2020), the monthly labor turnover in Ethiopia's industrial parks exhibits fluctuations, with rates reaching over 20% during various months. Notably, there was a significant increase in turnover during early 2020. Ethiopian industrial parks were established to enable industrialization, job creation, and economic growth. The Human Resource Management (HRM) practices within the parks are, however, faced with several critical challenges. Low wages and lack of monetary rewards are named by the majority of workers as primary drivers for job resignation. Inadequate working conditions, with low standards of safety, long working hours, and low benefits, also contribute to high turnover. Employees are also demotivated by the absence of official career development programs, which prevents their long-term job attachment. Second, inadequate employee training and participation lead to most organizations not investing in continuous skill enhancement, causing job dissatisfaction. Finally, weak institutional policies suggest a lack of well-defined HRM regulations and retention procedures that align with Ethiopia's labor laws and industrial policies.

The Ethiopian government has implemented several labor and employment programs, including the National Employment Policy and Strategy and the Ethiopian Industrial Parks Human Resource Management Guidelines however their execution is still lacking. Ineffective workforce management results from the lack of formal employee retention tactics in many industrial parks. Many studies have

shown that recruiting new employees is much more difficult and costlier than retaining existing ones, making employee retention a better choice for organizations. To take action and know the reason why employees leave their jobs requires a model that can predict the employee turnover rate and support the company in taking necessary steps to prevent employee turnover by providing the necessary demands of employees. By predicting employee turnover using machine learning, organizations will be able to know why and where they must change employee retention measures. It can assist organizations in enhancing worker satisfaction, turnover cost savings, organizational performance improvements, and eventually, become a better performer in its line of business. This thesis deals with the problem of employer turnover prediction using machine learning in the Adama Industrial Park.

Machine learning is a sub-field of artificial intelligence that uses statistical and computational methods to enable machines to learn from data, improve their performance, and make predictions or decisions without being explicitly programmed. It involves training algorithms to recognize patterns and relationships in data through experience, and then allowing the algorithms to make predictions or take actions based on that learning. The application of machine learning has seen tremendous growth in recent years, driven by the increasing availability of big data and advancements in computational power. Industries are leveraging ML algorithms to refine decision-making processes, automate tasks, and obtain insights from massive datasets. For example, in human resources, machine learning is used to predict employee turnover, optimize recruitment processes, and analyze employee sentiment. (Qamar, 2022)

Machine learning can improve our understanding of employee turnover by seeing trends in employee data that can be used to predict which employees are most likely to leave the organization. This information can then be used to develop targeted retention strategies to retain high-performing employees and prevent high-potential employees from leaving. The model learns based on previous employees' data and gets re-trained as new data arise. This research uses labeled data from the human resource department to predict employee turnover so this research uses supervised machine learning.

The contribution of this research was to predict employee turnover using machine learning algorithms and to determine the cause that highly affect employee turnover. Additionally, it provides practical recommendations for HR professionals and organizational leaders at Adama Industrial Park, helping them implement effective interventions to enhance employee satisfaction and reduce turnover.

1.2 Motivation

Employee turnover is increasingly a critical issue over the past few years, particularly in the industrial sector where it is viewed as a significant concern for organizational growth. Increasing turnover rates have become a chief reason for concern regarding the stability of companies, particularly in industrial parks such as Adama Industrial Park. A significant number of employees are leaving their jobs and the traditional way of gathering data to know the reason behind employee turnover is inadequate. And the way of gathering information on how the cause of turnover is unreliable. Observing these trends has inspired me to undertake this research using machine learning techniques to improve the prediction of employee turnover, particularly at the Adama Industrial Park. This research aims not only to identify the determinants of employee turnover but also to enhance data-driven decision-making in human resource management. By utilizing data analytics, an organization will know the driving factors of employee turnover and thereby make better decisions. It can analyze complex datasets to bring out patterns that may not be identifiable from traditional methods; therefore, organizations will be in a position to address issues way in advance of the event of increased turnover. This research enhances the prediction process and also empowers HR professionals to implement targeted retention strategies, ultimately fostering a more stable and productive workforce. The research, therefore, helps to bridge the gap between advanced technology and its practical use in managing employees to improve organizational efficiency and enhance employee satisfaction in the case of Adama Industrial Park.

1.3. Statement of the Problem

Employees are a foundation and the greatest asset to a company because they drive a company to success. Turnover of employees is a crucial issue around the globe with affects the company due to several means' manly loss of expertise, the expense of recruitment, a decrease in productivity, products not being to arrive on deadline and so on. Workers are the most critical to the success or failure of any organization; they provide valuable contributions to the enterprise. Nowadays, employee turnover is becoming a serious organizational problem because it leads to a financial and moral impact on the organization of limited resources. Identifying the cause that causes employee turnover is necessary for organizations to prevent employees from leaving their jobs and it maintain a stable workplace. Individuals after being trained inclined to move to a different organization for better opportunities. Good pay, comfortable timings, good environment, and career prospects are some of the basic reasons that urge an employee to look for a different job opportunity. The instant a good employee expresses his intention to quit, it becomes the

responsibility of the management and the human resource department to step in immediately and find out the actual reasons for the move.(Venkatesan et al., 2022)

Employee turnover has a particularly negative effect on African nations, where many companies face challenges related to unstable economies, a shortage of skilled workers, and ineffective workforce management techniques. In Ethiopia, employee turnover is an increasing concern, particularly in emerging industrial zones and private sector enterprises. Recent studies show that the turnover rate in Ethiopian manufacturing industries firms can be more than 80% annually.(Halvorsen, 2021). The construction of industrial parks and the rapid pace of economic change have increased the need for qualified workers. Low wages, low job satisfaction, and lack of career prospects are some of the reasons that make it still difficult for employees. Adama Industrial Park, one of Ethiopia's major industrial hubs, faces a serious employee turnover problem, which has decreased operational efficiency and output. Workforce instability affects product delivery and market competitiveness. Reduced employee morale: The issue is compounded by high turnover, which makes the remaining employees wonder about their jobs. The lack of a comprehensive, data-driven strategy for predicting and mitigating turnover further complicates these challenges. Resolving this issue is crucial to guaranteeing industrial development and park commercial success.

Existing Literature, such as (Dwivedi, 2018) Study on Predicting Employees Attrition using XG Boost Machine learning approach and (Yadav et al., 2018) examination of Early Prediction of Employee Attrition using Data Mining Techniques, primarily use dataset from external sources like Kaggle. In contrast, there is a gap in research utilization of local data to identify factors affecting employs turnover. Despite the critical nature of this issue, there is a lack of data-driven approaches in Ethiopia to systematically analyze and predict employee turnover. Previous studies such as, an assessment of factors affecting employee's turnover intention in Ethiopian revenues and customs authority (Lemma, 2019) and Perceived cause of employee turnover: the case of Shinits ETP Garment PLC (Hailu, 2016) Utilize traditional statistical techniques to analyze and interpret data. In the existing local study, methods such as questionnaires are manual and cannot accurately identify the specific factors contributing to turnover or provide data-driven solutions.

In Ethiopia's context, there is a deficiency in research that utilizes a data-driven approach to pinpoint the factor influencing employee turnover, and also existing manual methods are not equipped to provide real-time insight. This research aimed to address the gap by utilizing local data, that has not been extensively

analyzed regarding employee turnover prediction, to provide insights to give worth decision-making and organizational practices for employee retention. In addition to that it serves as a baseline for academic purposes as well as it can be implemented practically for human resource management and also leads to improved employee retention and organizational performance.

1.4 Research Questions

The research questions for our research are:

RQ1: What is the most significant attribute affecting employee turnover prediction at Adama Industrial Park?

RQ2: Which supervised machine learning algorithm is the most appropriate for predicting employee turnover?

1.5 Objective of the Study

1.5.1 General Objective

The general objective of this research is to develop a machine-learning model that predicts employee turnover at Adama Industrial Park and to provide recommendations for reducing turnover rates.

1.5.2 Specific Objective

- ✓ To review the literature work done in the area of employee turnover prediction.
- ✓ To collect and prepare data effectively for training and testing purposes
- ✓ To organize the dataset with relevant attributes for prediction model.
- ✓ To clean the data by handling missing values and outliers to make it appropriate for modeling.
- ✓ To improve the model by applying the hyperparameter optimization technique.
- ✓ To evaluate the proposed employee' turnover prediction model using performance evaluation metrics.
- ✓ To select an appropriate classification algorithm for prediction.

1.6 Significance of the Study

Predicting employee turnover using machine learning in the context of Adama Industrial Park has significant importance for the following reasons. Firstly, it aims to identify the key factors influencing employee turnover within the park, allowing management to develop targeted retention strategies that can minimize attrition rates. Given the high costs associated with employee turnover, including recruitment, training, and lost productivity, this research can provide valuable insights to reduce these

expenses and enhance organizational efficiency. Moreover, by leveraging machine learning techniques, the study equips human resource departments with data-driven insights that facilitate informed decision-making regarding employee management. Understanding the predictors of turnover will enable Adama Industrial Park to proactively address employee concerns, thereby improving job satisfaction and fostering a more engaged and productive workforce.

Additionally, this research contributes to the existing body of knowledge in human resources and organizational behavior, particularly in the Ethiopian context, where such studies are relatively scarce. Ultimately, the findings can lead to enhanced organizational stability and a positive workplace culture, benefiting both employees and management at Adama Industrial Park.

1.7 Ethical Consideration

When conducting this research on predicting employee turnover using machine learning in the case of Adama Industrial Park, several ethical considerations have been taken into account. We communicate the purpose and implications of the research with HR managers ensuring they understand how their data is used. We protect the privacy and confidentiality of employee information by removing the column that has the names of the employees. The models we use are tested for accuracy and use all the preprocessing techniques to prevent algorithm bias. Moreover, the findings should be used to foster a supportive work environment for the HR department. By addressing these considerations, researchers can conduct their studies responsibly.

1.8 Scope of the Study

This study aims to predict employee turnover using machine learning in the case of Adama Industrial Park. This approach is designed to determine the reason for turnover and identify areas for improvement in employee retention efforts. This work involves a series of steps, including data collection, preparing the data, applying classification algorithms for modeling, and assessing the proposed model with test datasets. To conduct this study, we utilize 16078 employee's historical data. For this study supervised machine learning like Random forest, K Nearest Neighbors, and Logistic regression is used.

1.9 Limitation of the Study

In the beginning, the company was hesitant and not willing to provide the necessary information used for this research. This reluctance posed significant challenges, as access to relevant data is essential for the effectiveness of the study. Additionally, the company's data management practices were minimal,

resulting in poor data quality. Consequently, it takes a long time to preprocess the data and make it suitable for the machine learning algorithms used in the predictions. Furthermore, the data collected was limited to a specific time interval from 2018 to 2022. These factors combined created obstacles that impacted the overall research process and its outcomes.

1.10 Organization of the Study

The organization of the study is the overall content of the study which was conducted to reach the desired result. The organization of the employee turnover perdition using machine learning in the case of Adama Industrial Park is discussed below.

Chapter one is the introduction, background of the study, statement of the problem, motivation, objective of the study, research question, significance of the study, ethical consideration, delimitation of the study, and limitation of the study. Chapter two is an overview of relevant literature, summarizing existing studies related to employee turnover and the application of machine learning in human resources and also discusses different evaluation metrics use for measure the performance of classification algorithm. Chapter three is all about data preparation, data preprocessing, data description, data encoding, data conversion, materials used, and methodologies are discussed briefly. Chapter four presents a predictive model experiment with different algorithms and compares different classification algorithms using performance metrics and the results are discussed in detail. Finally, in chapter five conclusion and recommendations are stated including future suggestions.

CHAPTER TWO

LITERATURE REVIEW AND RELATED WORK

2.1 Literature Review

In this chapter, all of the necessary literature for this research is briefly presented. This includes a description of employee turnover and its various types, as well as an overview of machine learning, which helps to understand the issue well. Additionally, discuss the related work that has been done before.

2.1.1 Concept of Employees Turnover

In human resources, employee turnover denotes the rate at which employees leave an organization and are replaced by new employees. However, there are various expressions of employee turnover in different research studies depending on the aspects of the phenomenon that the authors aim to emphasize. Some of the definitions given by different researchers are presented as follows:

To put it more simply, employee turnover is the sequence of events that occurs when a person quits their position and is replaced. It can readily be seen as a negative reflection on the efficacy and efficiency of the organization and is frequently used as a gauge of business performance. (Glebbeek & Bax, 2004).In today's workplace, the term "employee turnover" refers to the percentage of departing employees who are replaced by new hires. There are some studies on this subject, but it is a very delicate and difficult sector for management to govern and regulate. It is observed that the rate of turnover fluctuates in different workplaces according to the circumstances but the motives behind are relatively comparable**.** (Panigrahi & Rout, 2020)

Employee turnover is the flow of employees around the labor market; between firms, jobs, and occupations; and between and between the states of employment and unemployment. (Abbasi & Hollman, 2000)**.**Whenever the job is left, either by choice or forcefully, a replacement employee has to be recruited and trained and this process of replacement is referred to as turnover, according to Woods, as cited in (Ongori, 2007)

Employees turnover can imply a situation where the workers exit the organization willingly due to various factors hence hurting the organization financially as well as in its capability to deliver the minimum services necessary,(Yankeelov et al., 2009)

Employee turnover is the rate at which a company hires and fires employees as well as the length of time that employees typically stay with the company. Employee turnover puts undue strain on current

employees, who then leave and join the company. The remaining employees also face a heavier workload, which lowers morale, increases stress levels and ultimately results in absenteeism. To make up for the work of the resigned employees, the staff members are also compelled to put in more hours. (Nelly anzazi, 2018)

2.1.2 Cause of Employees Turnover

Employees resign from their organizations for various reasons. Different research studies have identified various causes for employee turnover. According to a study by (Abdali, 2011) the major causes of employee turnover include

Demographic and personal characteristics of employees: refer to the various individual attributes and background factors that define an employee's profile and identity within an organization. Employee's personal and demographic traits may be the cause of their departure from the company. Age, gender, qualifications, marital status, experience, and tenure are some of these attributes. Various demographic variables were demonstrated by various investigations as the cause of departure. While the number of relatives in the community increased organizational exits, kinship associations and the number of children improved turnover. Long-term, older employees leave the organization at a higher rate than short-term, younger ones.

Job Satisfaction: Workers who are unhappy with their workplace and themselves will leave an organization. Mismatches in work expectations, restrained career development, and low earnings are generally the key reasons for job dissatisfaction among employees. In a society where unemployment is low, job satisfaction and employee turnover have a positive association; conversely, when unemployment is high, this link is inverse.

Organization and work environment: The internal culture, values, and work environment of the Company play a vital role in shaping employee turnover. Hierarchical rigidity, ineffective leadership, overwork, and restricted growth opportunities can all lead to an increased desire to exit the organization.

Job content and Intrinsic motivation: The intrinsic nature of the job itself, in terms of its range, complexity, and rework ability, can play a large role in worker turnover. Unusually routine or repetitive work will cause dissatisfaction and a higher turnover likelihood. In addition, the level of job pressure and tension often predicts increased turnover.

Poor Communication: Communication is vital in retaining employees. Employees tend to be dissatisfied and leave working for the organization if they feel they are excluded or their suggestions are not needed.

Management behaviors: Inefficient management behaviors like micro-management, lack of supportiveness, and poor management can result in high turnover. If managers provide assistance, guidance, and advice on the job and responsibility, workers feel valued and valued.

2.1.3 Types of Employees' Turnover

Employee turnover can be categorized into several types. These turnovers can happen in any companies here are some of them

Voluntary Vs Involuntary: Both voluntary and involuntary turnover can lead to employees departing a business. It is deemed voluntary when an employee leaves the company voluntarily; it is deemed involuntary when they are fired without their agreement. Termination, layoffs or redundancies, retirement, chronic sickness, physical or mental incapacity, relocation abroad, or death are some of the factors that may cause it. (Mbah & Ikemefuna, 2012)

Internal Vs External: There are two types of employee turnover: internal and external. Employees leaving their current assignment to take up new tasks or positions within the company is considered internal. Both good and bad emotions could result from this. If the change in task and supervisor boosts morale, the feeling might be good; if the new role is project-related or disrupts relationships, such as holding a brief for a colleague in a different location, it might be unpleasant. Like the external turnover, the impact of this internal turnover might be significant enough to warrant monitoring. Internal turnover can be managed through human resource strategies including succession planning and recruitment policies.(Mbah & Ikemefuna, 2012)

Skilled Vs Unskilled: The turnover rate for unskilled workers, sometimes referred to as "contract staff," is typically high. The explanation for their departure is not implausible. These employees leave the company at the first chance of a better position since they lack the status of a permanent contract and, as a result, do not enjoy the same conditions of service as their permanent counterparts. Due to the ease of hiring new employees, employers are not concerned about this type of turnover. However, a high turnover rate of talented workers can be detrimental to the company and ultimately the organization by resulting in the loss of human capital. These consist of acquired information, training, and abilities. In addition to the expense of replacing them, the departure of these specialist workers may put the company at a competitive disadvantage because their abilities are relatively rare and can be used again in the same industry. These expenses might be very high, particularly if the staff members hold important positions and contribute significantly to the company. (Mbah & Ikemefuna, 2012)

2.1.4 How to Reduce Employees Turnover

In order to reduce employee turnover, HRM can adopt several strategies for increasing recruitment, selection, induction, training, job design, and remuneration. Motivated employees will stick around. In return, an organization should engage employees by giving them an appropriate platform for decision-making and letting them feel valued. It has been researched that a high level of engagement guarantees lower levels of turnover rates.

(Ongori, 2007) has suggested that there are various ways of reducing employee turnover. First, organizations need to have appropriate recruitment and selection techniques, where proper interviews and psychometric tests must be conducted to identify the proper fit for job requirements, which minimizes labor mismatch. Proper induction and comprehensive training make the recruits adapt and feel valued; regular development opportunities would keep them engaged. In addition, job design which provides a variety of tasks, autonomy, and feedback increases job satisfaction and results in retention. Salary and benefits are major factors: Organizations offering performance-related pay and additional rewards in the form of bonuses retain employees

Employee engagement and empowerment through involvement in decision-making can be developed and lead to loyalty and commitment. Strong communication systems will also help in nurturing trust and openness within the organization. A proper work environment that takes care of ambiguities in roles and job stressors can reduce the turnover rate considerably since the employees feel that their well-being and dignity are respected. Making available clear avenues of career development, through mentorship programs and promotional pathways, can help encourage employees to see a future for themselves in the organization. It supports the employees with flexible working hours and work-family-friendly policies to allow them to balance their personal lives with professional ones. Lastly, addressing organizational factors by maintaining stability and focusing on qualitative management practices will further enhance satisfaction and retention among employees. These strategies provide an organization with a more engaged and loyal workforce.

2.1.5 Overview of Machine Learning

Machine learning can be defined as an advanced branch of Artificial intelligence (AI) where a computer system can learn and grow with experience without being programmed directly for such functions. Machine learning is a technique where it enables the computer system to learn to find hidden patterns and knowledge within data through predictive models based on inputted data.

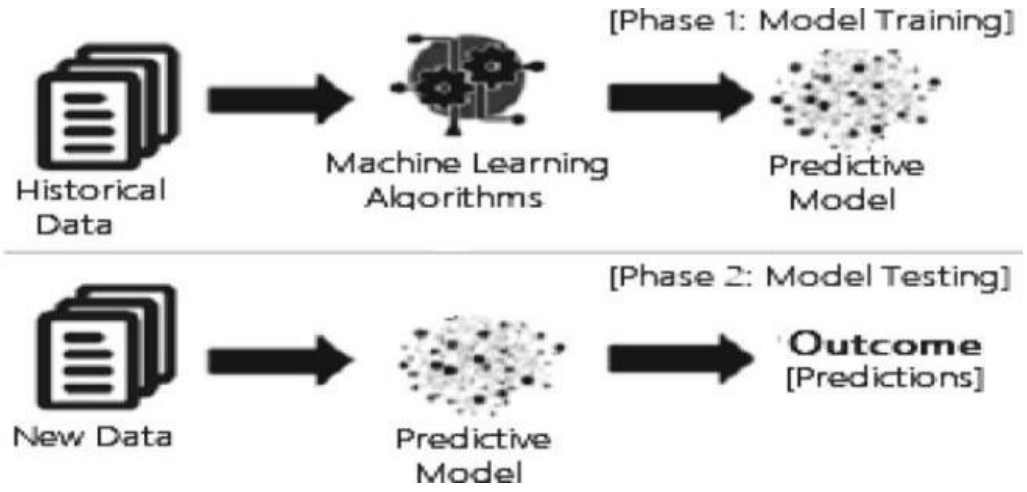


Figure 2. 1 Training and Testing Phase of Machine Learning Model (Sarker, 2021)

Machine learning is the study of how computers can learn better with experience automatically. It's one of the fastest-growing areas of technical study today, standing at the intersection of computer science and statistics and the core of artificial intelligence and data science. Recent progress in machine learning has been due both to the invention of new algorithms and theoretical work and also to an explosion in available online data and a related decline in computational costs. Applications for the data-intensive approaches to machine learning can now be found in science, technology, and commerce which empower evidence-based decisions in topics ranging from health care, manufacturing, and education, via financial modeling and policing, to advertising and marketing. (Horvitz & Mulligan, 2015)

Machine Learning is briefly described as the process of deriving useful information from data, achieved by devising reliable prediction algorithms. The algorithms have sufficient room for optimization, but their efficiency depends directly on the quality and amount of data gathered. Statistics and data analysis are thus routinely connected with machine learning. These algorithms are promising in terms of optimization but are very much dependent on the quality and quantity of data gathered. Statistics and data analysis thus often come with machine learning. (Misilmani, 2019)

2.1.4.1 Types of Machine Learning

Machine learning is a branch of artificial intelligence concerning the creation of algorithms enabling computers to learn and make predictions using data. Machine learning has numerous types, and each of them has varying purposes and uses. Knowing the types helps organizations and researchers choose the appropriate approach for the problem they face. The major types of machine learning include:

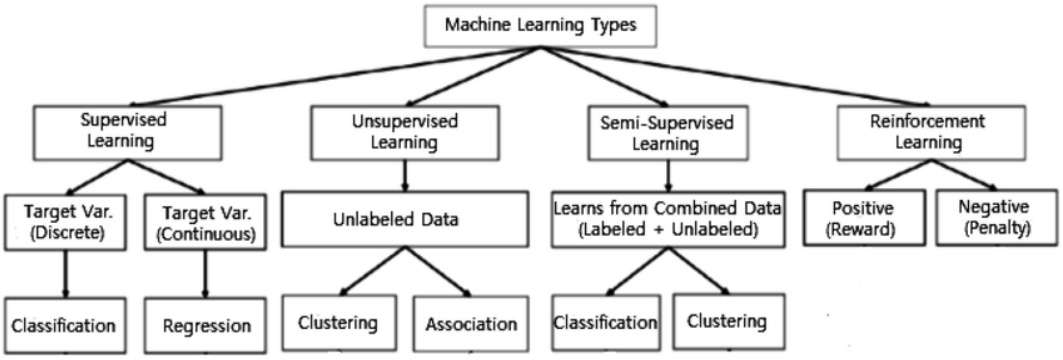


Figure 2. 2 Types of Machine Learning (Sarker, 2021)

Supervised machine learning: a type of machine learning where the model is trained from a labeled dataset, i.e., input-output pairs. Employing this method, the algorithm learns to convert the input features into their respective output labels by learning the relationships and patterns of the data. Learning is a process where the model is exposed to vast amounts of labeled data and then allowed to adjust its parameters in a way that its predictive error is minimized. Trained, the model can now be applied to new, unseen data to make predictions based on the learned relationships. The efficacy of supervised learning is dependent mainly on the quality and quantity of labeled data and the chosen algorithm applied. This renders it a significant tool in finance, healthcare, and marketing areas, where accurate predictions can make decision-making life possible.

Unsupervised Machine Learning: In unsupervised learning, the machine is trained using unlabeled data, meaning that there are no predefined outputs or targets. Rather than that, the machine infers patterns and structures from the data by clustering data points together or by dimensionality reduction to deduce the underlying attributes or distributions.

Semi-supervised Machine learning: is a type of machine learning in which an algorithm learns from partially labeled or unlabeled data, in addition to the labeled data it receives. In semi-supervised learning, the algorithm is trained on a smaller set of labeled data and a larger set of unlabeled data.

Reinforcement Machine Learning: In reinforcement learning, the machine learns by interacting with an environment through a sequence of actions. The machine receives feedback in the form of rewards or penalties based on the actions it takes. The machine learns to optimize its actions based on the rewards it receives and the goals it aims to achieve.

In this research historical employee data is collected from HR management and supervised machine learning algorithms for prediction models.

2.1.4.2 Supervised Machine Learning

Supervised machine learning: It is normally considered as a routine process of machine learning to acquire a function that transforms an input to an output using sample input-output pairs. It makes use of labeled training data with a set of training instances to draw conclusions. Supervised learning occurs when certain objectives are defined for their fulfillment from a given set of inputs, i.e., in task-oriented learning. The most usual operations of supervised are "classification"-divides the data, and "regression"-adapts the data.(Sarker, 2021)

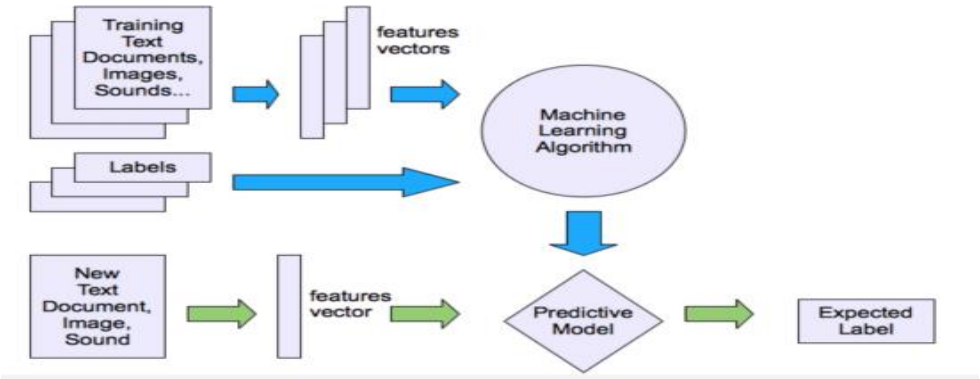


Figure 2. 3 Supervised Machine learning Model (Nasteski, 2017)

Logistic Regression (LR): Logistic regression is one of the machine learning algorithms and falls under the supervised learning type of classification problems with two classes for a dependent variable. Logistic regression is a statistical model utilized to make an estimate of the probability of an event, which is measured in terms of one or more influencing variables. Logistic regression establishes a relationship between a dependent variable and several predictor variables through fitting data in a logistic curve. A logistic regression model's output is a binary value of 0 or 1 for the chance of a point belonging to a particular class.

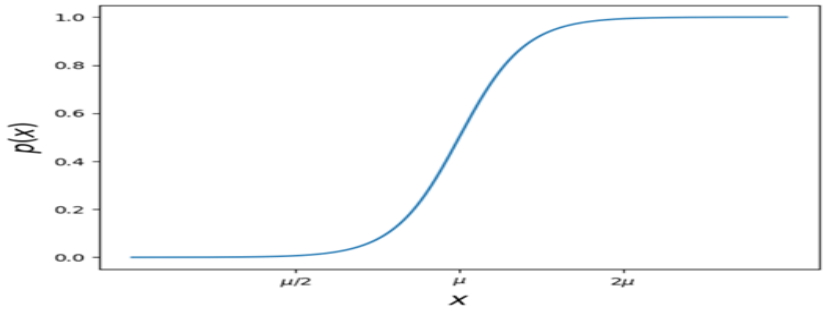


Figure 2. 4 Graph of logistic Regression curve (Omari, 2023)

Decision Trees (DT): Decision Trees is a supervised predictive model of machine learning that can be used for both classification and regression problems. It is a tree model that is utilized in predicting the result of a given input data through a sequence of if-else decisions. The tree model represents all possible results, decisions, and their impacts. Each of the internal nodes in the tree is a choice based on values of a chosen feature, and each leaf node is a class label or an integer value, depending on if the problem is a classification problem or regression problem. During training, the algorithm recursively divides data according to the best-split criteria.

Decision Tree is among the most widely used supervised machine learning methods applied to both classification and regression tasks. It's a tree with internal nodes as features and leaves as class labels or predicted values. Thus, predictions are made by following any path from the root to one of the leaves based on feature values of the input. (Awad & Fraihat, 2023)

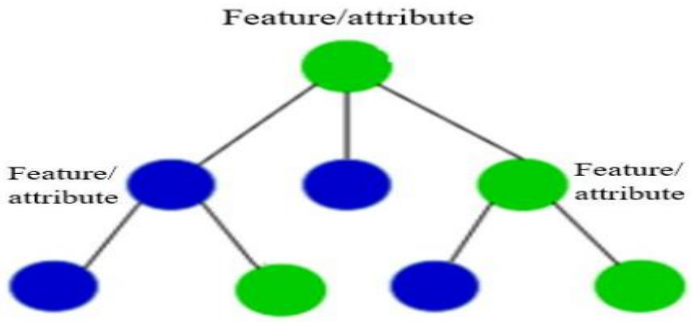


Figure 2. 5 Decision tree (Omari, 2023)

Random Forests (RF): Random forest is a machine learning method applied for classification, regression, and many more. Random forest is an ensemble method, which employs over one decision tree to predict the output more efficiently. The overall concept behind a random forest is to grow numerous decision trees on randomly sampled subsets of the original data, and for each decision tree in the forest, a different subset of the original features and observations is used. The randomness has the benefit of reducing overfitting and increasing generalization, resulting in a more robust model. In training, each tree of the forest is constructed on a bootstrapped sample of the training set, and for each split, a random subset of features is considered. The tree is incrementally constructed by splitting the data into binary segments based on the best split. The prediction for the last class is taken as the majority vote of all predictions of the forest of trees.

A random forest is an ensemble learning method consisting of n sets of de-correlated decision trees. It uses Bootstrap Aggregation through a resampling method to reduce variance in replacement. In prediction, a random forest is based on multiple trees of averaging-augur, for example, in regression or calculation of

majority votes in classification on the leaf nodes. Random forest models are a development of the idea of decision trees and have therefore attained much better improvement in prediction accuracy compared to a single tree by growing the 'n' number of trees in such a way that each tree is sampled at random without replacement from the training set. (Kirasich et al., 2018)

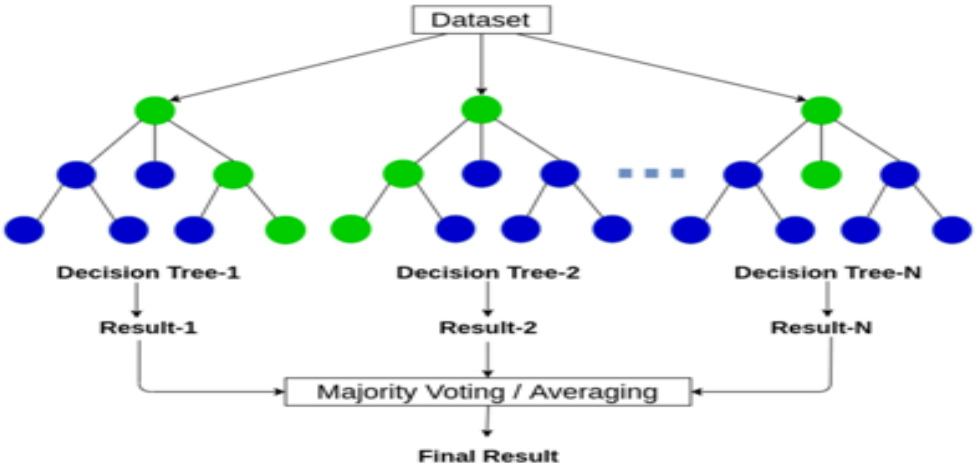


Figure 2. 6 Random Forest (Abhishek, 2024)

Naive Bayes (NB): Naive Bayes is One of the most commonly used machine learning algorithms for classifying tasks is Naive Bayes, with text classification and sentiment analysis as common applications. It applies Bayes' theorem and assumes independence between features since "naive" is what it is referred to as. The algorithm works by determining the probability for every class about the input features given. It does this by calculating the prior probability of each class (the chance of every class occurring regardless of the input features) and then calculating the probability of each feature given each class. Probability is calculated through the conditional probability of every attribute given a class. With the prior probability and probability calculated, the algorithm then calculates the posterior probability for each class given the input features. The output class is then set as the highest posterior probability across the classes.

Extreme Gradient Boosting (XGBoost): XGBoost is a highly advanced machine-learning algorithm that is optimized for speed and performance. It is an extension of gradient boosting that optimizes decision tree training using regularization techniques and parallel processing. XGBoost adds trees sequentially, where each additional tree corrects the errors made by the preceding trees, and it produces highly accurate prediction models. it is particularly effective on organized data and is very popular in competitive machine

learning due to its performance in Kaggle competitions.

XGBoost is a scalable and efficient version of gradient boosting that optimizes prediction accuracy through regularization and parallelization, and is particularly suitable for big data and intricate models. (Chen & Guestrin, 2016)

K-Nearest Neighbors (KNN): K-Nearest Neighbors (KNN) is a machine learning algorithm used for classification as well as regression. In simple words, as described in "Pattern Recognition and Machine Learning" by Christopher Bishop, KNN makes use of the fact that like points occupy proximity in the feature space. The algorithm provides a target data point based on the identification of the K nearest neighbors on a predetermined measure of distance, i.e., Euclidean distance. In a classification problem, it returns the most common of these neighbors, and in a regression problem, it computes the average of the values of the neighbors. One key benefit of KNN is that it is powerful but uncomplicated without making any assumptions about the data's underlying distribution. But the value of K matters; a low K can lead to noise sensitivity, but a high K can over smooth the boundary of the decision. Despite its advantages, KNN can be computationally costly, particularly with big data sets, due to the need for all points to be calculated by distance. Overall, KNN remains a fundamental technique in machine learning, loved for its directness and use in everyday life in a variety of applications. (Jordan et al., 2006)

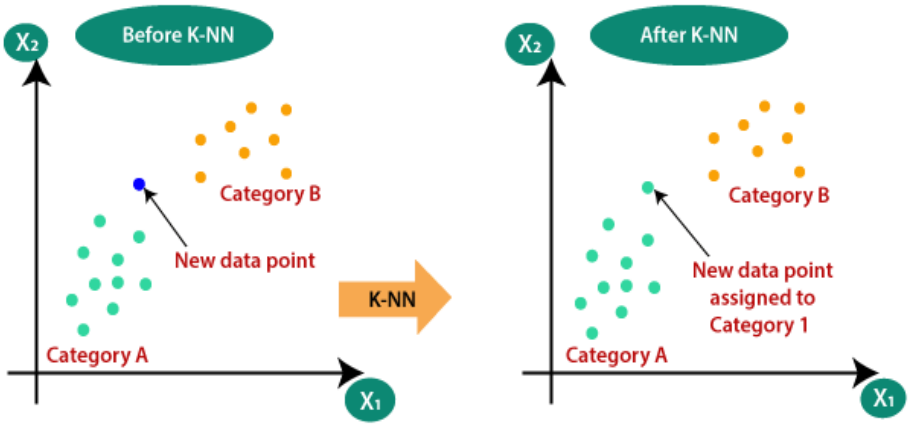


Figure 2. 7 K-Nearest Neighbors (Omari, 2023)

2.2 Related work

The related Work is based on the summary of various researchers in the field of employee turnover prediction by using different techniques.

Ensemble method-based architecture using random forest importance to predict employee's turnover (Hossen et al., 2021)**.** This paper utilized a sizable dataset obtained from Kaggle, 15000 rows, and 10 columns, to conduct their analysis. To predict employee turnover, the authors preprocess the row data. Sequential Backward Selection (SBS), Chi-square, and Random Forest Importance is used to find the most relevant features. The author applied a variety of machine learning algorithms including support vector (SVM), Decision tree (DT), Gaussian Naïve Bayes (GNB), Random forest (RF), Multi-layer perceptron (MLP) and K-Nearest Neighbor (KNN).

The author has suggested five various methodologies, the first methodology is to utilize applied classifier algorithms to estimate performance using all features and observed that RF, DT, SVM, LR, GNB, and KNN have achieved an accuracy of 98.49%, 97.44%, 94.45%, 75.78%, 79.91%, and 92.80% respectively and observed that RF has highest accuracy with 98.49%. The second methodology is to utilize applied classifier algorithms to estimate the performance with SBS feature selection k-Nearest Neighbor (KNN) achieved highest accuracy of 96.75%. The third one employed a classifier with feature reduction by chi-square and Random forest importance random forest 99.00%, SVM 97.00% and decision tree 97.00%, the KNN model 97.00%.MLP model 81.00% following the feature reduction which previously had an accuracy of 97.00%. The Gaussian NB accuracy was 81.00% previously. The four approaches utilized an ensemble technique named bagging and boosting to combine several learning methods the random forest classifier gives the best accuracy in instances only and it is 99.4% with boosting. The five approaches utilized classifier algorithms to compare the performance with 10-fold Cross-validation to remove the overfitting issue and get a good-trained model RF 99.4, DT 98.32, SVM 95.9, GNB 80.2.

The researcher concluded that run five different models to obtain better accuracy named as with all the features, with less feature with SBS, Chi-square and Random Forest Importance, with 10-fold cross-validation and with bagging, booting. The researcher gets better accuracy 99.4% using the reduced feature with 10-Fold Cross-validation by implementing the Random Forest classifier.

Predicting employee attrition using machine learning approaches (Raza et al., 2022) In this research, the author needs to implement a learning framework to predict employee attrition and to find the causes of employee turnover. The IBM HR employee attrition dataset was utilized for research findings and use by Employee Exploratory Data Analysis (EEDA) was done to obtain useful insight. SOMTE (Synthetic Minority Oversampling Technique) data resampling technique is applied to make data balance. After all data preprocessing is done the author uses 85:15 ratios to split a dataset.

Four sophisticated machine learning-based algorithms were utilized in this research study: Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree Classifier (DTC), and Extra Trees Classifier (ETC). The research process utilized to predict employee attrition is known as the ETC. The research compared the employee attrition forecasts based on the four sophisticated machine learning techniques: ETC, SVM, LR, and DTC. Precision measures of the employed machine learning methodologies were 87% for the SVM technique, 72% for the LR technique, and 83% for the DTC technique. Precision, recall, f1 score, accuracy, and ROC accuracy metrics of the proposed Extra Trees Classifier (ETC) were 93% respectively.

K-fold validation and previous applicable state-of-the-art investigations were used to validate the methods. The SVM approach produced an accuracy score of 88%, the LR technique a score of 74%, the DTC technique a score of 84%, and the proposed method a score of 93% utilizing datasets that were 10-fold. According to the EEDA application, age, work level, hourly rate, and monthly income are the main causes of employee attrition. Organizations are able to combat staff turnover thanks to the research findings.

Employee attrition prediction using machine learning algorithms(Lekan et al., 2022) Study the performance of different algorithm like Logistic Regression, Naive Bayes Classifier, Random Forest Classifier, and XGBoost to forecast an employee's position in a company based on IBM data set.

In this experiment, accuracy, recall, f1 score, and precision are chosen as measures to indicate the efficiency of the algorithm. The results using Logistic regression, XGBoost, Decision tree, and Radom Forest with accuracy from an imbalanced dataset are 82.01%, 85.4%, 77.7%, and 83.6%, respectively. In this respect, it could be asserted that logistic regression, XGBoost, Decision tree, and Random Forest in the case of the synthetic balance dataset obtained 68.48%, 84.58%, 71.66%, and 82.08% accordingly. The XGBoost classifier had a higher performance in the AUC of about 0.50 in the ROC curve. The XGBoost model was 85.5% with the original data, but about 84.6% using a synthetic dataset. The study concludes that the XGBoost Algorithm yields the best results for the dataset.

From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction (Yahia et al., 2021).The paper addresses a breakthrough approach to employee attrition prediction that breaks free from big data to align in a depth of data scenario by focusing on quality rather than quantity in data. The research does not gather all available features but is interested in applicable attributes for attribute prediction. By filtering and selecting the key features using two proven feature selection techniques—Recursive Feature Elimination, which is a wrapper method, and SelectKBest, a filter method—authors integrated experiential research (thinking "big") with quantitative methods (thinking "deep").

To obtain the features required for the employee attrition forecast, the author applies this feature selection method to a classification algorithm. The attrition prediction approach, which is based on machine, deep, and ensemble learning models, is tested on three different datasets: a small HR real dataset of 450 samples with 16 features, a medium-sized IBM HR simulated dataset of 1470 samples with 34 features, and a large-scale HR simulated dataset of 15,000 samples with 9 features. The author employs a range of algorithms, such as Random Forest, XGBoost, and Vote Classifier as ensemble learning models, Decision Tree, Logistic Regression, and Support Vector Machine as machine learning models, and three deep learning models (DNN, LSTM, and CNN). Each classifier's hyperparameters were adjusted using a grid search algorithm, and the dataset was divided into training, test, and validation sets in a 10:70:20 ratio. The training dataset was then used to train the various models in their optimal configuration. Accuracy is used as a standard parameter in this study to validate and assess the performance of the employed classifiers.

The researcher chose 11 essential features for the employee attrition prediction based on the combination of the two feature selection methods (RFE and SelectKBest). The predictors' ability has been enhanced by minimizing the number of attributes and keeping only the selected attribute. The researcher mentions the ensemble method VC accuracy has slightly increased from 0.93 before feature selection to 0.96 after feature selection. This supports the effectiveness of SelectKBest and RFE as feature selection algorithms for enhancing and validating the employee attrition model. The ensemble learning VC (voting classifier) achieved the highest accuracy, recording results of 0.96, 0.98, and 0.99 for large, medium, and small simulated human resources datasets, respectively.

This paper makes two key contributions: it proposes a deep data-driven predictive approach that emphasizes relevant data and impactful feature selection to enhance learning efficiency and accuracy. Additionally, it offers insights into understanding attrition phenomena and provides recommendations for effective employee retention. Overall, the study emphasizes the significance of data quality and uncovers unexpected motivators behind employee attrition.

Employee Attrition Prediction Using Deep Neural Networks (Al-Darraji et al., 2021) To enhance the prediction of employee attrition, the author applies deep learning in conjunction with a few preparation methods, such as data cleaning, categorical data encoding, and rescaling procedures. In the dataset analysis process, a correlation matrix is utilized to determine how the features in the dataset relate to one another. The IBM dataset, which includes 35 attributes for 1470 employees, was used by the author. The dataset

is biased in favor of the working employees because just 237 employees have left the company, while 1233 are still employed. The prediction model performs comparatively poorly as a result of this imbalance. To balance this, the author scales down the dataset to a balanced state by utilizing the Adaptive synthetic (ADASYN) sampling approach. To avoid overfitting or underfitting, the model hyperparameters such as the number of hidden layers, number of neurons, activation functions, etc., should be optimally selected. In this case, the hyperparameters of the prediction model are adjusted by performing a grid search with multi-core machines and multithreading programming. To evaluate the performance of the prediction model, the dataset is divided into two sets 70% of the dataset is used to train the model, while the remaining 30% is used to test the model. Train-test validation and k-fold cross-validation are used for validation techniques. Accuracy was 91.16% and 94.16% with the imbalanced and synthetic balanced sets. A comparative study is also conducted using 10-fold cross-validation, with an accuracy extracted which was 89.11%.

Hybrid GA–DeepAutoencoder–KNN Model for Employee Turnover Prediction (Lim et al., 2024) this research conduct on employee turnover prediction using a hybrid genetic algorithm–autoencoder–k-nearest neighbor (GA–DeepAutoencoder–KNN) model. To improve prediction precision, the suggested model is combined with the KNN model, an autoencoder, and a genetic algorithm. The IBM dataset, a medium-sized data set comprising 1,500 records and 35 features, was used in this research. The dataset underwent several preprocessing processes to remove noise, including missing values, empty columns, and meaningless constant features, to improve the accuracy and computation time of the suggested model.

For balancing the target variable, the resampling technique was implemented by the author. A confusion matrix for the classification model performance, Accuracy, recall, and precision have been implemented. A comparison of the experiment and comparison with the proposed model with the traditional DeepAutoencoder–KNN and k-nearest neighbor models was conducted. The outcome reveals that in comparison with the traditional models, the GA–DeepAutoencoder–KNN model yielded a substantially high accuracy score (90.95%). The precision for the DeepAutoencoder-KNN and the standalone KNN is 86.48% and 88.37%, respectively. For comparison and assessment of their efficacy, the GA–DeepAutoencoder model was also paired with other machine learning models like the RF, DT, SVM, and NB models. In comparing GA–DeepAutoencoder-based ML models the DeepAutoencoder–KNN model performed with the highest accuracy score of 90.95%. The second and the third best models in terms of accuracy were the GA–DeepAutoencoder–RF and GA–DeepAutoencoder–DT models with accuracy of 81.76% and 80.14%, respectively. The tree hybrid models (GA–DeepAutoencoder–RF and GA–

DeepAutoencoder–DT) experienced a reduction in accuracy performance in comparison with individual RF and individual DT models.

This explains that single RF and DT models experience less of the problem of high dimensionality. Moreover, the GA–DeepAutoencoder–SVM models realized a 6.35% improvement regarding accuracy compared with the use of the single SVM model. In the NB combination, accuracy fell from 68.24% to 62.23% compared with the single NB model. Concisely, the combination of DeepAutoencoder, KNN, and SVM attained accuracy enhancement for employee turnover prediction, where KNN and SVM suffer from high-dimensional problems and noises in the data.

Factors that affect Employees turnover in Ethiopia, in case of Hibret Bank sc.(Mandefro, 2022) conducted to determine the variables affecting staff turnover at hibret bank and evaluated the effect of this variable on employee retention. Data was gathered through open and close-ended questionnaires, as well as semi-structured interviews. A stratified sampling technique was used, resulting in a sample size of 358 permanent employees out of a total population of 4,433 employees. SPSS was used to analyze the data, enabling multiple regression analysis, correlation analysis, and descriptive statistics.

The study revealed that among the major turnover drivers are career development opportunities, supervisor relationships, appraisals, and compensation. More specifically, excellent training and development activities were found to enhance employees' competencies and commitment, while good supervisor relationships and equitable appraisals resulted in job satisfaction. Sub-standard pay and bonus schemes were also among the key turnover drivers, and they revealed that pay strategies need improvement. Overall, the research recognizes the necessity of addressing the above forces to reduce organizational instability and staff turnover in Hibret Bank.

Factors Affecting Employee Turnover at ERCA Large Taxpayers Brach Office (Meseret et al., 2020) examined how job satisfaction, salary, career growth opportunities, nature of the job, and organizational environment impact employee turnover. Its objective is to find out the main cause contributing to employee turnover, the level of turnover, and the associated challenges. Data were collected from 222 employees at the ERCA Large Taxpayers Branch Office using a questionnaire survey. Out of the 222 distributed questionnaires, 184 were fully completed and returned, making the response rate approximately 82.9%.

Statistical methods such as ANOVA, Pearson Correlation Coefficient, and Regression Analysis were utilized to examine the data. Outcomes indicated that job satisfaction, salary, career growth opportunities,

and whether the job was of a specialized nature significantly determine employee turnover. The study identifies that in improving overall organizational functioning, ERCA must adopt productive measures to create job satisfaction and retention.

Cause of employee turnover: The case of Global Insurance Company (Selhadin, 2019) examined the underlying reasons for employee turnover at Global Insurance Company (GIC). The study sample consisted of 70 employees current and former employees of GIC. Questionnaires were employed to collect numerical data about employee satisfaction, while interviews provided more information regarding personal experience and perception. Statistical Package for Social Science (SPSS) version 20 was used to analyze data with descriptive statistics. Results showed that GIC employee turnover was perceived to be high due to low compensation, poor working environments, and ineffective management. Compensation, benefits, and promotion opportunities were determined by this study as significant turnover intention determinants. Overall, the research pointed out that GIC needs to revise its pay and promotion plans to improve its employee retention.

2.3 Gap analysis

The research attempts to review both foreign and local literature related to employee turnover prediction. The gap identified in existing studies is that they utilized public data that did not identify the real-world underlying causes of employee turnover specific to individual organizations. Additionally, local studies are most likely to utilize manual methods for collecting data and also have very small data. These local studies methods in the literature utilize SPSS which is a traditional statistical technique to analyze and interpret data, which may limit its ability to uncover complex patterns and relationships.

In this research, the problem of employee's turnover prediction was addressed using specific Adama industrial park characteristics. In addition, this research used different approaches such as different attributes and preprocessing activity, a large amount of data set (16078) compared to the existing research mentioned here in this study. In our study, the application of machine learning facilitates the analysis of large datasets, significantly enhances predictive accuracy, and uncovers deeper insights into the factors influencing employee turnover, thereby addressing the limitations inherent in the use of traditional SPSS methods found in existing research.

Table 2. 1 Summary of the related work

No	*Title*	*Reference*	*Used algorithm/Methods*	*Obtained result*	*Limitation/future work (Gap)*
1	Ensemble method-based architectur using random forest importance to predict employee's turnover	(Hossen et al., 2021)	SVM, DT, GNB, RF, MLP, KNN	Random forest Classifier	✓ Analysis is predicated on a small number of features. ✓ Public data is use
2	Predicting employee attrition using machine learning approaches	(Raza et al., 2022)	SVM, LR, DT, ETC	Extra tree classifier	✓ The findings may not be applicable to all companies or sectors. ✓ Public data is use
3	Employee attrition Predicting using machine learning Algorithm	(Lekan et al., 2022)	LR, NB, RF, XGboost	XGboost	✓ The study did not fully identify the features that inspire employees to leave their job. ✓ Public data is use
4	From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction	(Yahia et al., 2021)	proposals of a deep data driven predictive approach focuses on the use of relevant data and the selection of impactful features	emphasizing data quality and revealing unexpected motivators	✓ Used simulated datasets (it may not accurately reflect real word complexities to misleading result). ✓ Public data is use
5	Employee Attrition Prediction Using Deep Neural Networks	(Al-Darraji et al., 2021)	DNN	Using imbalance 91.6 & using balance 94.16	✓ Prediction accuracy could still be improved further. ✓ Public data is use

6	Hybrid GA–DeepAutoencoder–KNN Model for Employee Turnover Prediction	(Lim et al., 2024)	Deep auto Encoder, KNN & GA-DeepAutoencoder -KNN	GA-DeepAutoEncoder-KNN	✓ Small sample size. ✓ Public data is use
7	Factors that affect Employees turnover in Ethiopia, in case of Hibret Bank Sc	(Mandefro, 2022)	Descriptive Statistics, Correlation, Regression Analysis	Identified significant factors influencing turnover, including career development, mentoring, and reward systems.	✓ Small data set ✓ Utilize traditional statistical techniques to analyze and interpret data
8	Factors Affecting Employee Turnover at ERCA Large Taxpayers Brach Office	(Meseret et al., 2020)	Descriptive & Explanatory Research Design	low salary, poor working environment, and management performance contribute to turnover.	✓ Small data set ✓ Utilize traditional statistical techniques to analyze and interpret data
9	Cause of employee turnover: The case of Global Insurance company	(Selhadin, 2019)	Questionnaires, Interviews, SPSS for analysis	High turnover due to low salary, poor working conditions, and management issues; job satisfaction is crucial.	✓ Small data set ✓ Utilize traditional statistical techniques to analyze and interpret data

2.4 Model Evaluation Method

This section presents the different evaluation performance Metrics that measure the model used for predicting employees' turnover using machine learning in case of Adama industrial part.

Accuracy: Accuracy refers to how well the model predictions approximate the actual values created. It represents the overall success of the performance of the model. (Jordan et al., 2006)

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \text{ ----------[2.1]}$$

Precision: It assesses the accuracy of positive predictions by showing how many of the cases predicted as positive were indeed correct.

$$\textbf{Precision} = \frac{TP}{TP+FP} \text{---------- [2.2]}$$

Recall (Sensitivity): Recall is a metric of the ratio of true positive predictions to all actual positive instances. It measures the ability of a classification model to recall all instances that are relevant**.**

$$\textbf{Recall} = \frac{TP}{TP+TN} \text{---------- [2.3]}$$

F1-Score: The F1 score is a performance indicator that balances precision and recall by combining both into a single value. It is especially helpful when class distributions are not uniform because it shows how well the model can identify relevant cases and reduce false positives. The model performs better if the F1 score, which goes from 0 to 1, is higher.

$$\textbf{F1-Score} = 2* \frac{precision*recall}{precision+recall} \text{---------- [2.4]}$$

Confusion Matrix: Confusion matrix is an evaluation matrix for performance that is used to quantify error in a problem of classification. It illustrates the entire misclassification specification. (Beauxis-Aussalet, 2014).The quantity of correct and wrong predictions a classifier presents is indicated in a confusion matrix, which is graphically depicted as a table. It provides richer information to data practitioners regarding the performance, errors, and exposures of the model. This makes it possible for data practitioners to fine-tune their models and perform more analysis.

Table 2. 2 Confusion Matrix

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

- ✓ **True Positive (TP)**: Models correctly predicted a positive outcome.
- ✓ **True Negative (TN):** Models correctly predicted a negative outcome.
- ✓ **False Positive (FP):** Models incorrectly predicted a positive outcome; it is also known as Type I error.
- ✓ **False Negative (FN):** Models incorrectly predicted a negative outcome., it is also known as Type II error.

Area under the curve & Receiver Operating Characteristics (AUC-ROC): ROC is the curve formed when "Sensitivity" is plotted concerning "1-Specificity" for varying instances of the classification threshold "T" above and outside the range of index values. ROC-AUC is one of the widely used techniques to calculate classifier performance. The AUC describes the overall discriminative power of the model for distinguishing between the positive and negative classes. The AUC values are between 0 and 1. 0.5 is a value representing no discriminative power (i.e., random prediction), and a perfect classification is 1 (Fawcett, 2006).

2.5 Cross Validation

Cross-validation is a commonly utilized data resampling technique for estimating the actual prediction error of models and optimizing their parameters.(Berrar, 2024) It consists of dividing a dataset into various subsets, or folds, which allows for a structured approach to evaluating models. In methods like k-fold cross-validation, the model undergoes training several times—utilizing a different fold as the validation set each time, while the other folds act as the training set. This method produces multiple performance metrics that, when averaged, offer a more dependable estimate of the model's performance on new data, helping to alleviate problems such as overfitting. Moreover, cross-validation is vital for hyperparameter tuning because it permits practitioners to assess the impact of various hyperparameter configurations on model performance. By examining multiple setups through cross-validation, it is possible to determine the combination that provides the best average results, ensuring that the final model is both solid and generalizable. In summary, cross-validation strengthens the reliability of both model evaluation and optimization processes in machine learning workflows.

2.6 Hyperparameter Tuning

Hyperparameter tuning is a process of coming up with an effective set of hyperparameters that maximizes the performance of a given model of machine learning. This stage is very critical since these hyperparameters will dictate how the model learns its structure: from the learning rate to the number of neurons in a neural network and the kernel size in support vector machines. Unlike model parameters, which are learned from the data, hyperparameters are set before training and have to be very well-selected. Effective hyperparameter tuning can be very useful in improving both the performance and generalization of a model. Its importance in machine learning is pretty high; if done right, this really can help make the model that much more effective. It helps balance bias and variance, thus improving the generalization of the model on new, unseen data, which is very important for its robustness and reliability in real-world applications. Besides, the best hyperparameters are identified, probably leading to better use of computational resources, faster training time, and cost reduction. Efficiency becomes more critical when dealing with large-scale models and datasets. (A Ilemobayo et al., 2024)

Various techniques have been developed to automate the process of hyperparameter tuning in an optimized way. The brute-force technique involves a grid search over predefined sets of hyperparameters. While easy and straightforward to implement, this can be computationally expensive for large hyperparameter spaces. Random search is an alternative to grid search, sampling the hyperparameters randomly from a distribution. This has turned out to be much more effective at finding the optimal hyperparameters since a much larger and more diverse range of combinations can be explored. Bayesian optimization is based on probabilistic models. It builds up a surrogate model to approximate the objective function and iteratively constructs the most promising hyperparameters to be evaluated, which creates a balance between exploration and exploitation. Hence, this optimization technique becomes handy in the case of expensive optimization. (A Ilemobayo et al., 2024)

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Overview

This chapter discusses the methodology adopted for the study in this thesis through the systematic discussion of how the study arrived at a clear understanding of employee turnover prediction. The work commences with a detailed explanation of the overall data collection process, emphasizing the sources of the employee dataset obtained from the HR office at Adama Industrial Park. It highlights the underlying significance of collecting comprehensive and relevant data to ensure the accuracy of the analysis. The methodology proceeds to data preprocessing as a very vital step in preparing the data for machine learning. This includes missing value handling, where various techniques are considered-including imputation. Data binning and grouping are also discussed within the chapter, where continuous variables are segmented into discrete bins to enable model interpretability and good performance.

Another important aspect that this chapter covers is the encoding of data. The text has described how categorical variables are transformed into numerical formats using one-hot encoding techniques, for instance, to make them algorithmically proper. That is an important step since this will ensure the machine learning models understand the data in an appropriate manner. The feature selection describes the most appropriate selection and retention of important predictors of employee turnover. This simplifies not only the model itself but also increases predictive accuracy by removing the noisy and irrelevant features.

The chapter further elaborates on data normalization, specifically Min-Max scaling, a process that rescales the range of feature values to fall between 0 and 1. Min-Max scaling becomes important in algorithms sensitive to the data scale so that no single feature becomes dominant in the model. Finally, discusses the material used for this research.

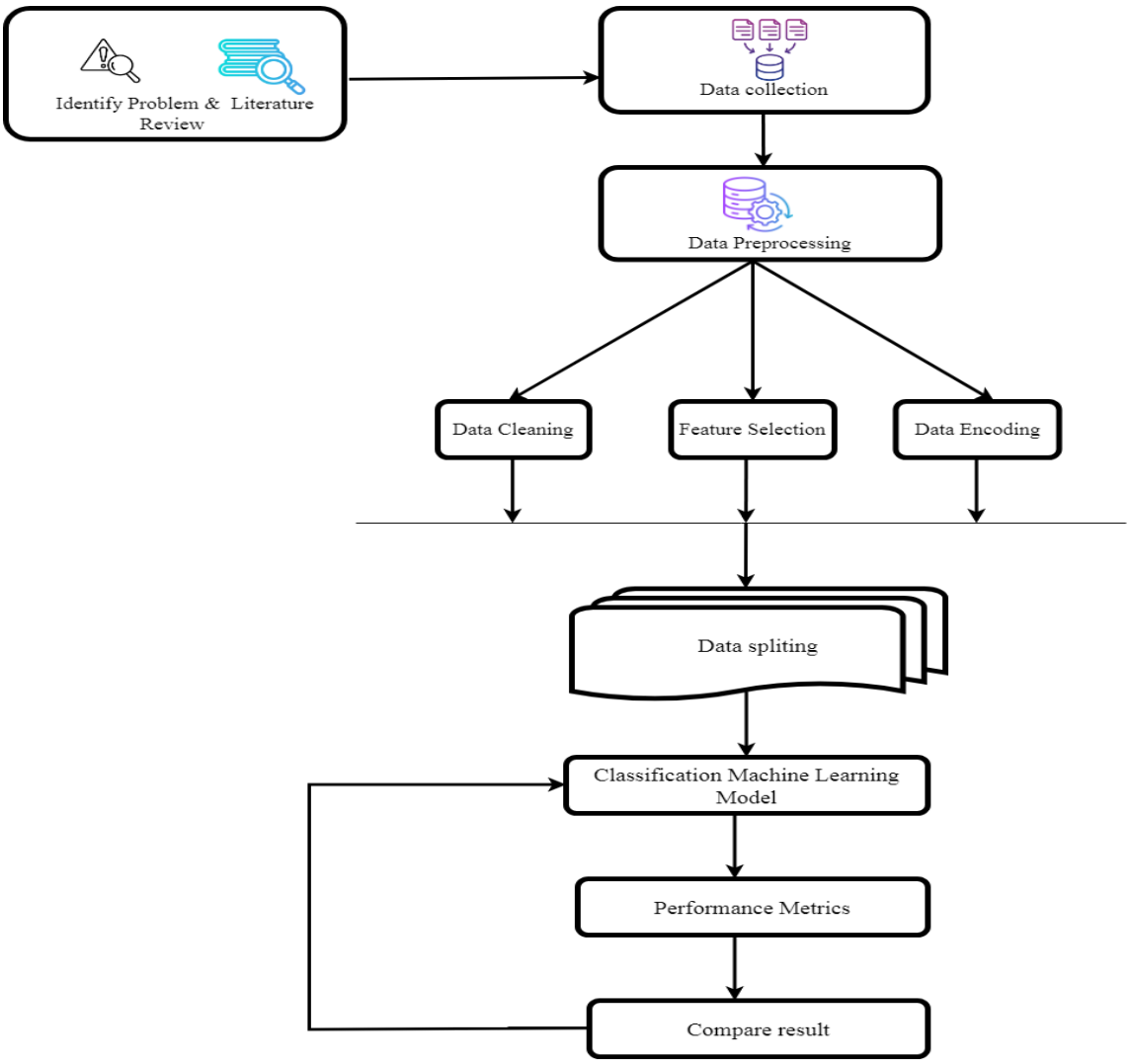


Figure 3. 1 Methodology workflow Diagram

3.2 Data Source

The data for this research was collected from industrial park development corporation HR office at Adama Industry Park. Adama Industry Park is one of the largest industries with a lot of manpower in our city. The historical data of employees is collected to predict employees' turnover from 2018-2022 excel dataset. The dataset consists of 16,078 rows, each representing an individual employee, and 13 columns, which include a mix of demographic and employment-related features.

3.3 Data Collection

The data set was created by gathering information from the organization. Data collection involves gathering the necessary information from relevant sources to address the existing problem or research

question. The more relevant data that we collect, the better our model will be. This data is having 16078 rows and 13 Columns. The dataset has one dependent variable called status type with contain two class namely active employees and withdrawn employees. The remaining 12 columns are independent variables that provide additional information about each employee, such as their age, salary, years of service, gender, education level, etc. These independent variables serve as the input features for the machine learning model, which will analyze how these attributes influence an employee's likelihood of staying or leaving the company.

Table 3. 1 Sample dataset in CSV format

Personnel Number	Age	Gender	Marital status	Working Years	Department	Position	Hire Date	Status type	Date of birth	Education Level	Leaving Reason	Salary
50000002	35	Female	Single	5	Production planning and	Merchandiser	20/09/2018	Active	1/1/1989	Degree	Exist	6700
50000004	38	Male	Single	5	Production planning and	Senior Supervisor	14/10/2018	Active	10/1/1986	Diploma	personal reasons	7300
50000008	26	Female	Single	3		Planner leader	1/6/2020	Withdrawn	2/9/1997	Degree	personal reasons	6500
50000009	39	Female	Single	5		supervisor	15/10/2018	Active	1/1/1985	Diploma	Exist	7000
50000011	30	Female	Marr.	5	Warehouse		10/10/2018	Withdrawn	1/1/1994	Diploma	personal reasons	6000
50000016	33	Female	Single	5	Warehouse	Workshop Director	8/10/2018	Active	1/1/1970	Diploma	Exist	7000
50000017	35	Male	Single	5	Administration	Assistant Deputy Genera	8/10/2018	Active	1/1/1989	Degree	Exist	11255
50000019	38	Male	Marr.	2	Iron	Ironing worker	22/01/2021	Withdrawn	1/1/1986		Exist	2900
50000023	33	Female	Marr.	5	Administration	Administration assistanc	8/10/2018	Active	1/1/1994	Certificate	Exist	7300
50000031	30	Male	Single	5	Maintenance	Technical support	10/10/2018	Active	14/01/1993	Certificate	Exist	3500
50000036	44	Female	Marr.	2	Maintenance	Workshop Director	13/02/2021	Active	1/1/1980	Diploma	Exist	6000
50000038	36	Female	Marr.	3	Administration	Administration assistanc	25/05/2020	Withdrawn	1/1/1988	Certificate	personal reasons	5500
50000042	24	Male	Single	2	security	security	8/7/2021	Active	11/10/1999	Certificate	Exist	4000
50000062	40	Male	Marr.	5	Finance	accountant	10/10/2018	Active	1/1/1984	Degree	Exist	7500
50000064	39	Female	Marr.	2	Administration	Administration assistanc	22/04/2021	Active	1/1/1985	Degree	Exist	8500
50000065	50	Male	Marr.	5	Finance	Payroll leader	1/12/2018	Active	1/1/1974	Degree	Exist	6000
50000066	30	Male	Marr.	5	Maintenance	electrician	10/10/2018	Active	1/1/1970	Diploma	personal reasons	7500
50000068	41	Female	Marr.	5	Sewing	Operator	20/10/2018	Active	1/1/1983	Certificate	Exist	3500
50000080	25	Female	Marr.	5	Occupational Health and	nurse	10/10/2018	Active	1/1/1970	Degree	personal reasons	8000
50000084	28	Female	Marr.	5	Sewing	Operator	20/10/2018	Withdrawn	1/1/1996	Certificate	personal reasons	3500

3.4 Data Description

The data description presents a detailed overview of the dataset structure and key attributes. The dataset used in this thesis was taken from the IPDC HR office in Excel format. The dataset consists of 16078 raw with 13 columns that include categorical and numeric attributes. The attributes can be explored and yield insightful information about organizational structure, employee demographics, and retention that guide HR strategy and decision-making. Table 3.2 provides a concise glance at each attribute description in the data set as well as what comprises each column. This table is essential for understanding the role each feature plays in analyzing employee behavior and predicting turnover. As an HRM expert, I leverage my knowledge of these aspects to guide the selection of features that are most likely to influence turnover.

Table 3. 2 Data Description

NO	Attribute name	Datatype	Description of attributes
1	Personal Number	int64	Unique numerical identity of employee
2	Age	int64	Length of time the employee lived
3	Gender	Object	Category of woman & men employee
4	Marital Status	Object	The state of being of employee married, divorce, not married
5	Year of Service	int64	The number of service year the employee work on
6	Department	Object	Sub class of the general organization
7	Position	Object	A part where employee is assigned to work
8	Hire date	Object	The date an employee starts his/her career
9	Status type	Object	Indication where an employee leave or stay
10	Date of birth	Object	The date on which employee was born
11	Education Level	Object	The level of education the employee has
12	Leaving reason	Object	The reason why an employee leaves the company
13	Salary	int64	Payment of an employee per month

3.5 Data preparation and preprocessing

The raw data that we collected from the IPDC industry park is in Excel format and contains missing values this can affect the model and is not suitable for implementation. The dataset also with different range and most of the columns is categorical variables. Applying algorithms to this noisy data is unlikely to yield accurate results, as they may struggle to identify patterns effectively. Inconsistent data points can interfere with the model's learning process and result in inaccurate predictions while missing values can skew the total statistics. Fixing problems like missing values, improving data quality, and making sure the data is appropriate for machine learning applications are the main goals of data preprocessing. This process may involve converting data from Excel to CSV, identifying and addressing missing values, and encoding data.

The step of importing libraries is essential for bringing in the necessary tools for this thesis. Python libraries are required for executing data processing tasks. Libraries such as NumPy, pandas, matplotlib, and sci-kit-learn, along with various Python functions and classes, are imported to conduct experiments

using the following code. The subsequent step is to load the dataset after importing the libraries. Employee data collected in Comma Separated Values (CSV) format is imported using the read_csv() function, and the dataset is uploaded to Jupyter Notebook within Anaconda. The read_csv() function from pandas is utilized to read the CSV file. The command data. head displays the dataset along with the number of rows and columns. If a specific number is not provided, the head() method returns the first five rows of the dataset.

```
df = pd.read_csv(r'C:\Users\Mitiku\Desktop\ipdc employees.csv')
```

```
df.head()
```

	Personnel Number	Age	Gender	Marital status	Working Years	Department	Position	Hire Date	Status type	Date of birth	Education Level	Leaving Reason	Salary
0	50000002	35	Female	Single	5	Production planning and control	Merchandiser	20/09/2018	Active	1/1/1989	Degree	Exist	6700
1	50000004	38	Male	Single	5	Production planning and control	Senior Supervisor	14/10/2018	Active	10/1/1986	Diploma	personal reasons	7300
2	50000008	26	Female	Single	3	NaN	Planner leader	1/6/2020	Withdrawn	2/9/1997	Degree	personal reasons	6500
3	50000009	39	Female	Single	5	NaN	supervisor	15/10/2018	Active	1/1/1985	Diploma	Exist	7000
4	50000011	30	Female	Marr.	5	Warehouse	NaN	10/10/2018	Withdrawn	1/1/1994	Diploma	personal reasons	6000

```
df.shape
```

```
(16078, 13)
```

Figure 3. 2 Load employees CSV data

3.6 Data conversion

This procedure helps transform the data into a consistent form that machine learning can quickly process. This may involve converting excel files to CSV, json or other suitable formats. In our case, we convert excel data to CSV file. This can provide several key benefits that enhance data management and processing, reduce file size, minimize complexity, enhance portability, and improve compatibility making it a valuable step in facilitating data-driven workflows and analyses.

3.7 Data Cleaning

Data cleaning is a necessary part of the data preprocessing step, it mainly focusses on enhancing the quality and reliability of data by identifying and correcting errors and inconsistencies within the dataset. Handling missing values is the mean task using imputation techniques such as filling in gaps or dropping columns with excessive missing data. Additionally, data cleaning includes removing duplicates, and irrelevant and detecting outliers to increase data quality and to get a better prediction. In our case we have 3278 missing

value excite in the dataset, which could potentially affect the quality of the model. To ensure the integrity of the dataset and avoid bad results, we apply the data imputation method.

```
# Check for missing values
df.isnull().sum()
```

	0
Personnel Number	0
Age	0
Gender	15
Marital status	15
Working Years	0
Department name	663
Position Name	474
Hire Date	137
Status type	0
Date of birth	13
Education Level	807
Leaving Reason Text	1154
Salary	0

Figure 3. 3 Missing Value

3.8 Feature selection

Feature Selection is one of the most critical machine learning notions that impact models' performance heavily. The data variables utilized to train models have a significant contribution towards the performance can be accomplished. Variable selection and data cleansing should be the very first task in model designing.

Most of the data to be mined contains a lot of irrelevant attributes, such that removal of such attributes shall be necessary. Besides, many algorithms do not work well with a large number of attributes, which is why feature selection techniques need to be applied. The main use of feature selection is to improve the model and avoidance of overfitting. Domain knowledge is used in our case, to drop the hiring and birth date, as both features are redundant. They are effectively represented by the year of service and age columns respectively. Additionally, the Personnel Number column does not provide any predictive power regarding turnover; it can also lead to overfitting and increased dimensionality. Moreover, using unique identifiers like personnel numbers raises privacy concerns, as they may expose sensitive employee information and compromise confidentiality. After preprocessing, the sample dataset consists of 16,078

rows of data, each with 10 carefully selected columns that provide important features for modeling employee turnover predictions.

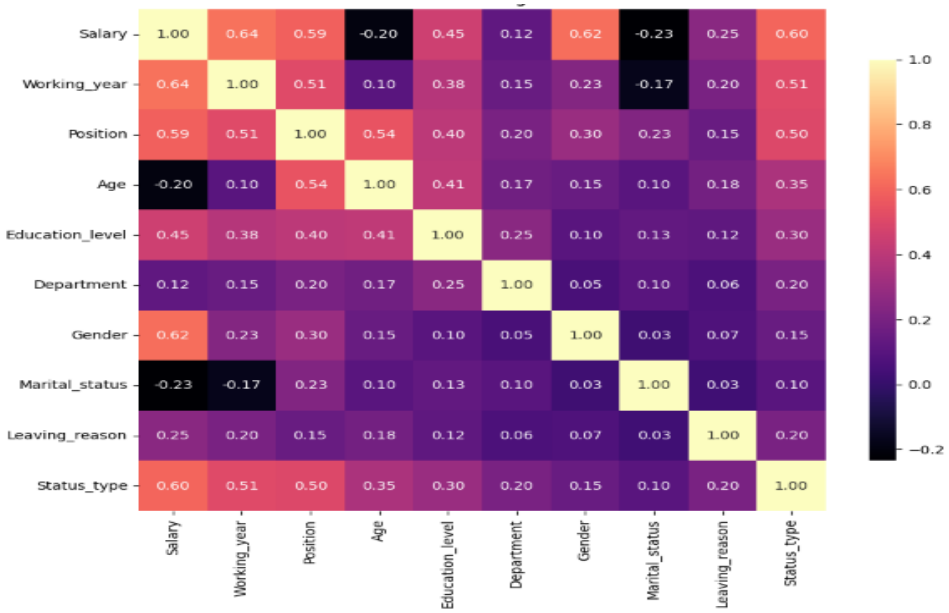


Figure 3. 4 Correlation matrix of features

To fully visualize the dataset and the relationships between features, the correlation matrix is essential for identifying patterns and dependencies. The correlation analysis indicated in Figure 3.4 reveals that the target variable has a strong correlation with some features and a weak correlation with others. Additionally, there are some weak correlations among the features. This approach identifies the important correlations in the dataset and ensures that the strongest predictors are used effectively while considering the weaker associations.

3.9 Data Binning

Data binning is a valuable method that helps in preparing data for analysis or machine learning applications, making it easier to identify trends and patterns. Which consists of grouping continuous or discrete values into intervals or bins. It is used to reduce the complexity of data analysis, noise reduction, and improve data visualization. In this research, we use the binning method for numerical variables and the grouping method for categorical data to make our data easier to analyze and visualize.

3.10 Data Encoding

Machine learning algorithms can’t read and interpret data that exists in categorical form. To address this issue, using encoding techniques is a necessary step. Encoding data is the process of changing Categorical

data into numerical format. In this study, we have many categorical variables, such as gender, marital status, educational level, salary, and so on. To convert these categorical variables to numerical variables, we use the One-Hot Encoding technique, which helps maintain consistency, clarity, and effectiveness in the model. This ensures that the dataset is well-formatted for analysis and that the model properly interprets all features. The dataset used in this research contains a target variable that possesses two categorical categories, namely 'active' and 'withdrawn,' and these are encoded to 0 and 1 via the Label Encoder function in sklearn.

	Status type	age_group_adult	age_group_elder	age_group_matured	age_group_young	Gender_Female	Gender_Male	Marital status_Marr.	Marital status_Single	M
0	1	1	0	0	0	1	0	0	1	
1	0	0	0	1	0	0	1	1	0	
2	1	1	0	0	0	1	0	0	1	
3	0	0	0	1	0	1	0	1	0	
4	0	1	0	0	0	1	0	1	0	

5 rows × 34 columns

Figure 3. 5 Encoded data

3.11 Data Normalization

Normalization is a technique widely applied in machine learning data preparation. It is often one of the initial preprocessing steps performed when working with data, particularly tabular data. It is crucial because each variable may use a different scale for its values. Best practice dictates that data should be normalized to ensure all values are transformed to a common scale and also ensure that each feature is equally important during the distance calculation and model training processes for better interpretation and performance of predictive analytics. We apply Min-Max to maintain the relationships between the data points while ensuring consistent scaling across all features.

3.12 Model Selection

To select an appropriate model for our research, we draw upon insight from the previous literature. Employees attrition estimation using random forest algorithm (Pratt et al., 2021) compares Random Forest, Logistic Regression, Gaussian Naive Bayes, Decision Tree, K-Nearest Neighbors, and Support Vector Machine using evaluation performance and achieve the best accuracy of 85.12 random forest and 75.42, 76.42, 75.51, 80.67, and 83.76 respectively.(Yedida et al., 2018) presents a study focused on employee's attrition prediction. The authors evaluate the KNN algorithm alongside various other machine learning methods, such as Naive Bayes, Logistic Regression, and Multi-Layer Perceptron (MLP). The results indicate that the KNN classifier outperforms the other techniques, attaining an accuracy of 94.32%

and an AUC of 0.9697. The study concludes by recommending the KNN classifier as a robust approach. (Guerranti & Dimitri, 2023) performed employee attrition prediction by considering a wide range of machine learning models: Logistic Regression, Classification Trees, Random Forest, Naive Bayes, and a simple Neural Network (NN). Accuracy, AUC-ROC, and F-score metrics are reported as the results of the models. As this dataset is highly imbalanced, most weight was given to the AUC-ROC metric. Logistic Regression reached an accuracy of about 88% with an AUC-ROC of 85%, making it also one of the best models in this work. Besides, Random Forest showed quite robust performance since it not only yielded high accuracy but also gave very valuable feature interpretation. The study concluded that both LR and RF are effective for employee attrition prediction, with a note on the interpretability of a model in HR applications. Study and prediction analysis of employee's turnover using machine learning approach (Chakraborty et al., 2021) This paper proposed Logistic Regression, Decision Trees, Random Forest, Naïve Bayes, Support Vector Machine, and K-Nearest Neighbors for employee turnover prediction. Accuracy, precision, recall, and F1-score were the performance metrics used to evaluate each of the several models. With an accuracy of 90.20% and a precision rate of 99%, the Random Forest model outperformed the others, making it appropriate for forecasting whether employees will quit or remain with a company. The confirmation that the Random Forest model is effective for the prediction of attrition makes this a no-brainer for the most useful tool in formulating HR strategies for organizations. By refer papers from different literature and seeing the interpretability and complexity of the model we select Logistic Regression, Random Forest, and K-Nearest Neighbors for our research. Initially, we used the default parameter values for the algorithms. To enhance the results, we implemented hyperparameter tuning techniques, specifically using Randomized Search with cross-validation (Randomized Search CV). This approach allows us to systematically explore different parameter combinations to find the optimal settings for our models.

3.13 Materials use for study

3.13.1 Hardware Tools

Some specific requirements of hardware that will be used for designing and implementing employees' turnover prediction are mentioned below.

- ✓ PC OR Processor: Intel® Core™ i5-3340M CPU @ 2.70GHz
- ✓ 4.00GB RAM
- ✓ 1TB Hard disk
- ✓ System Type: 64-bit operating system, x64-based processor.

3.13.2 Software Tools

Some specific requirements of the software that will be used for designing and implementing employees turnover prediction is mentioned below.

- ✓ Programming Languages: Python's numerous libraries and frameworks, such scikit-learn, make it a popular choice for machine learning.
- ✓ Machine Learning Libraries: Scikit-learn is a widely Python library for machine learning tasks.
- ✓ Anaconda: Anaconda is a free and open-source software program for creating machine-learning models.
- ✓ Jupyter Notebooks: An interactive platform for creating and recording machine-learning models
- ✓ Data Analysis and Visualization Tools: NumPy and Pandas are essential data libraries manipulation and analysis. Data visualization can also be done with programs like Seaborn and Matplotlib.
- ✓ Draw.io: A tool for illustrating the architecture and diagram in the study.
- ✓ Google Colab: is cloud-based, it is simple and quick to work and doesn't depend on a local computer.

CHAPTER FOUR

RESULT AND DISCUSSION

4.1 Overview

This chapter outlines the process of conducting experiments to construct a predictive model. It evaluates the creation of machine learning models, including Logistic Regression, K-Nearest Neighbor, and Random Forest. The chapter details the major experiments conducted, along with interpretations and performance evaluations of the prediction models. All preprocessing activities performed on the dataset, as described in Chapter Three, highlight key tasks undertaken during this phase. This section emphasizes and summarizes the significant experiments conducted to identify the optimal model in line with the objectives set in Chapter One. A series of experiments were carried out, resulting in prediction models with varying accuracies, recalls, precisions, Fl-score, and area under the receiver operating characteristic curve (AUC-ROC). We use the random search CV to improve the results of the algorithms. Cross-validation is utilized to ensure robust evaluation and accuracy of models. Additionally, this section discusses various development activities, including the evaluation of the models built and the selection of the best-performing model.

4.2 Proposed Architecture

The proposed architecture below summarizes the steps followed for the prediction of employee turnover at Adama Industrial Park. The process begins with the obtaining of the dataset for employees, which is the preprocessing to make it ready for modeling. Such preprocessing would include handling missing values, normalizing the data, and encoding categorical variables. After pre-processing, the resultant data is then split into training and testing sets on which the various selected models of machine learning are to be trained. We have used three different models in this work, namely Logistic Regression, K-Nearest Neighbors, and Random Forest. Later, once the training is complete, the respective performance metric scores would be computed in terms of accuracy, precision, recall, F1-score, and AUC-ROC. Additionally, hyperparameter tuning and cross-validation are used to ensure the robustness and accuracy of the model. At this point, each model's results will be compared to identify the best-performing predictive algorithm. Then, the best-performing model will be used for prediction. This would bring out insights related to employee turnover, which would be helpful for the organizations to know and take up the necessary retention measures.

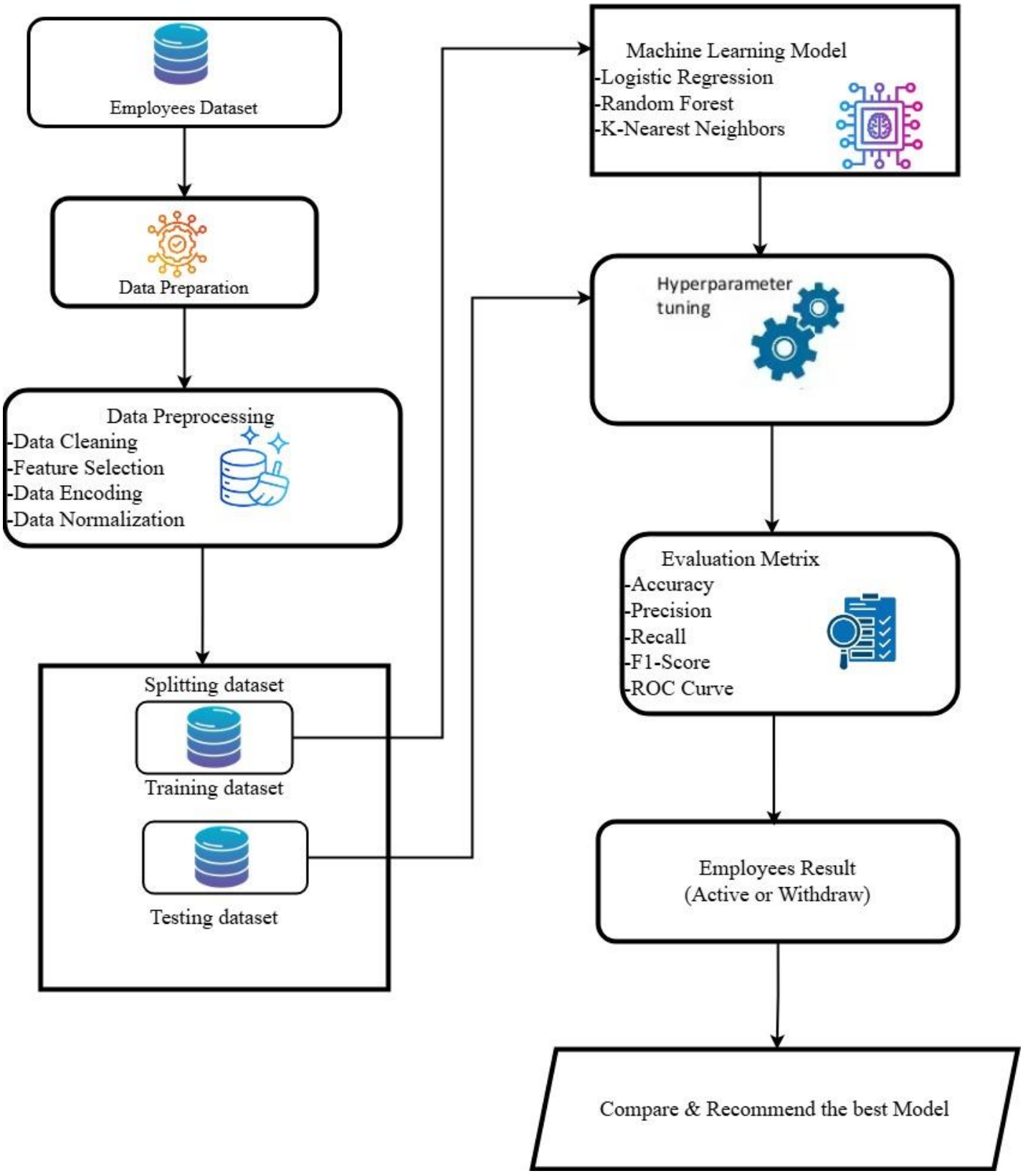


Figure 4. 1 The Propose Architecture for Employees' Turnover Prediction Model

4.3 Predictive Model Building

In predictive modeling, a model that forecasts future events is constructed utilizing both historical and present data. Predictive modeling involves gathering and preprocessing data, following which a model is developed, predictions are produced, and the model is verified (or updated) in light of new information. Our research is a classification problem therefore in the context of our study the term prediction is refers to determine whether the employees are active or withdrawn. The mean focus of the thesis was that whether we can identify the important features behind employee turnover and use them to build a predictive model that can help the organization management with turnover prediction.

4.4 Train-test split

Train-test split is a popular technique in machine learning where one splits a data set into two sets: one to learn the model and another to verify the accuracy of the model. The resulting end of doing that is that one has four sets: X-train and y-train for the data to learn the model, where the features and the relative target labels are kept, and X-test and y-test for the test set, which one uses in model validation. By training on X-train and y-train, the model learns to identify patterns and relationships within the data. When making predictions, the model uses an X-test, and its performance is assessed by comparing these predictions against the true labels in the y-test.

In this research, we have a total of 16078 employees' data. We split the data in 80:20 ratios meaning 80% (12862) we use for training purposes and 20% (3216) for testing purposes. For each selected algorithm, the model is trained on the training dataset and evaluated on the test dataset created.

Table 4. 1 Train-test split of dataset

Dataset split	Amount	Description
Training set (80%)	12862	To train the model
Testing set (20%)	3216	To test the model
Total	16078	Total amount of entire data

4.5 Build Logistic Regression Model

In this instance, the logistic regression approach is used to generate a prediction model. Logistic regression analysis looks at the link between a set of independent variables and a categorical dependent variable. It estimates the probability that a given input belongs to a particular category by applying a logistic function to a linear combination of the input features. The logistic function is appropriate for representing binary outcomes because its output falls between 0 and 1.

Experiment 1

In this experiment, we used a logistic regression model from scikit-learn to classify data. The model was set up with default settings, meaning it uses standard options that work well for many situations. First, we trained the model on our12862 training data using model.fit (X_train, y_train). This step helps the model learn the patterns in the data. After training, we checked how well the model performed on a separate 3216 test dataset by calculating its accuracy with `model. Score (X_test, y_test). Finally, we used model. predict(X_test) to get the model's predictions for the test data. This whole process helps us understand how effectively the model can classify new data based on what it learned from the training set. Overall, it's a straightforward approach to building and evaluating a logistic regression model**.**

Evaluating Model Performance

In this part, we discuss the evaluation and analysis of the logistic regression-based employee turnover prediction model. The performance of this predictive model is assessed using accuracy, a confusion matrix, and a classification report.

```
model = LogisticRegression(penalty='l2', dual=False, tol=0.0001,
                           C=1.0, fit_intercept=True, intercept_scaling=1,
                           class_weight=None, random_state=None,
                           solver='lbfgs', max_iter=100, multi_class='auto',
                           verbose=0, warm_start=False, n_jobs=None, l1_ratio=None)
model.fit(X_train, y_train)
accuracy = model.score(X_test, y_test)
print("Accuracy:", accuracy)
y_pred = model.predict(X_test)
```

```
Accuracy: 0.8753109452736318
```

```
percentage_accuracy = accuracy * 100
print(f"Accuracy: {percentage_accuracy:.2f}%")
```

```
Accuracy: 87.53%
```

Figure 4. 2 Accuracy of Logistic Regression model

Accuracy: The report above shows that the overall accuracy of the model is 87.53, i.e., it correctly classified 87.53% out of total of 3,216 instances.

Classification report: It provides a complete picture of a model's performance on a classification problem. It includes key metrics such as precision, recall, and F1 score for each class in a dataset

	precision	recall	f1-score	support
Active	0.86	0.97	0.91	2103
Withdrawn	0.92	0.70	0.79	1113
accuracy			0.88	3216
macro avg	0.89	0.83	0.85	3216
weighted avg	0.88	0.88	0.87	3216

Figure 4. 3 Classification Report for the Logistic Regression Model

The classification report in Fig 4.3 provides a comprehensive evaluation of the logistic regression model in predicting employee's turnover. The performance metrics for a model predict two classes called Active employees and Withdrawn employees. In the case of active class, the result shows 86% of precision ,97% of recall, and 91% of f1-score. In the case of the withdrawn class the result is 92% precision,70% recall, and 79% f1-score.

Confusion Matrix

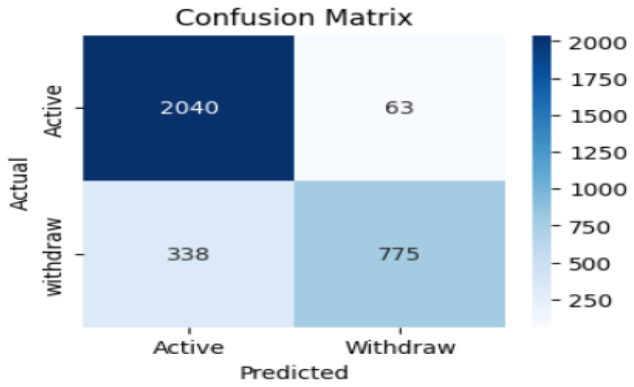


Figure 4. 4 Confusion Matrix for Logistic Regression model

The Confusion matrix in fig 4.4 shows the performance of logistic regresstion in classifaying employees turnover prediction as active or withdrawn.The model correctly predicted 2040 instance of active class and 775 instance of withdrawn class. However it incorrectly predict 63 instance of active class as withdrawn and 338 instance of withdrawn class as active .

Pseudocode representation of Logistic Regression model:

1. Import necessary library – NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn, Logistic Regression
2. Load the dataset containing employee features and turnover labels
3. Data preprocessing:
   a. Handle missing values
   b. Convert categorical variables in to numerical variables
   c. Normalize numerical variables
3. Split the data into training and testing sets
4. Initialize the Logistic Regression model with specified parameters
5. Fit the model using the training data:
   a. Call Randomized search cv to optimize hyperparameter
6. Evaluate the model on the test set:
   a. Compute accuracy, precision, recall, and F1-score, ROC
7. Output the model performance metrics
   a. accuracy, precision, recall, and F1-score, ROC

4.6 Build Random Forest Model

The random forest classification algorithm is utilized in this experiment to develop a prediction model, and Python is used to forecast our model. Multiple decision trees are used by Random Forest to generate predictions. A random sample of the data is used to train each tree, and the Random Forest aggregates the output from all the trees when it comes time to make a forecast. This method lowers the possibility of errors that could arise from using just one decision tree and helps to increase accuracy.

Experiment 2

For the random forest model, we use 12862 data to train our model and 3216 data to test our model. The model is built using the default parameter of random forest.

Evaluating Model Performance

By using the classification performance evaluation Metric namely accuracy, classification report includes (precision, recall, f1-score) and confusion matrix we evaluate our model.

```
model = RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,criterion="gini",
                               max_depth=None, max_features='sqrt',
                               max_leaf_nodes=None, max_samples=None,
                               min_impurity_decrease=0.0, min_samples_leaf=1,
                               min_samples_split=2, min_weight_fraction_leaf=0.0,
                               n_estimators=100, n_jobs=None, oob_score=False,
                               random_state=None,
                               verbose=0, warm_start=False)
model.fit(X_train, y_train)
accuracy = model.score(X_test, y_test)
print("Accuracy:", accuracy)
y_pred = model.predict(X_test)
```

```
Accuracy: 0.882773631840796
```

```
percentage_accuracy = accuracy * 100
print(f"Accuracy: {percentage_accuracy:.2f}%")
```

```
Accuracy: 88.28%
```

Figure 4. 5 Accuracy of Random forest model

Accuracy: By using the performance measurement of accuracy we achieve an accuracy of 88.28%.

Classification report:

	precision	recall	f1-score	support
Active	0.87	0.97	0.92	2103
Withdrawn	0.93	0.72	0.81	1113
accuracy			0.88	3216
macro avg	0.90	0.84	0.86	3216
weighted avg	0.89	0.88	0.88	3216

Figure 4. 6 Classification report of Random Forest Model

The classification report in fig 4.6 shows a detail summary of the random forest model evaluation in prediction of employee's turnover. The performance metrics for a model predict two classes called Active employees and Withdrawn employees. In the case of active class, the result shows 87% of precision, 97% of recall and 92% of f1-score. In the case of withdrawn class, the result shows 93% of precision,72% of recall, and 81% of f1-score

Confusion Matrix:

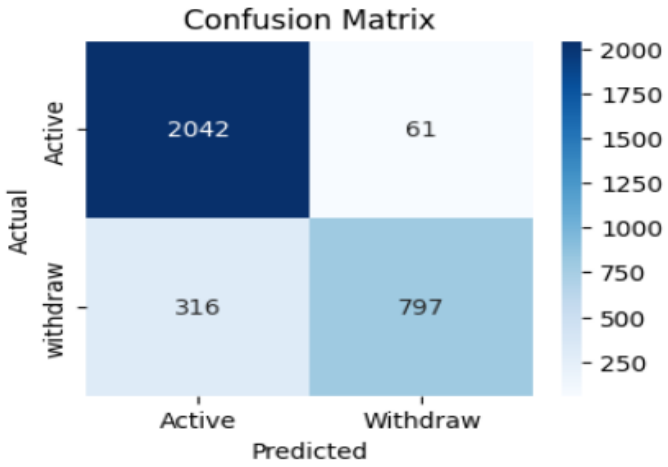


Figure 4. 7 Confusion Matrix for Random Forest Model

The Confusion matrix in fig 4.7 shows the performance of random forest in classifaying employees turnover prediction as active or withdrawn.The model correctly predicted 2042 instance of active class or true positive and 797 instance of withdrawn class or true negative. However it incorrectly predict 61 instance of active class as withdrawn and 316 instance of withdrawn class as active.

Pseudocode representation of Random Forest model:

1. Import necessary library – NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn, Random forest
2. Load the dataset containing employee features and turnover labels
3. Data preprocessing:
   a. Handle missing values
   b. Convert categorical variables in to numerical variables
   c. Normalize numerical variables
3. Split the data into training and testing sets
4. Initialize the Logistic Regression model with specified parameters
5. Fit the model using the training data:
   a. Call Randomized search cv to optimize hyperparameter
6. Evaluate the model on the test set:
   a. Compute accuracy, precision, recall, and F1-score, ROC
7. Output the model performance metrics
   a. accuracy, precision, recall, and F1-score, ROC

4.7 Build K-Nearest Neighbors Model

The K-Nearest Neighbor technique has become widely used in machine learning and data mining because of its unique performance and ease of use. Following training sample data, classification is used to predict the labels of test data points. (Pandey & Jain, 2017)

Experiment 3

In this experiment the model is build using default parameter and train the model using the train data and test using test data.

Evaluating Model Performance

To measure our model, we use different performance evaluations namely accuracy, and classification reports including (precision, accuracy, and recall) Metrices like we did in the above two experiments.

Accuracy: The KNN measure is 85.88 % accurate.

```
knn = KNeighborsClassifier( n_neighbors=5,  weights='uniform',
                            algorithm='auto',  leaf_size=30,
                             p=2, metric='minkowski',  metric_params=None,
                             n_jobs=None,)

# Fit the model to the training data
knn.fit(X_train, y_train)

# Make predictions on the test data
y_pred = knn.predict(X_test)

# Evaluate the accuracy of the model
accuracy = knn.score(X_test, y_test)
print("Accuracy:", accuracy)
```

```
Accuracy: 0.8588308457711443
```

```
percentage_accuracy = accuracy * 100
print(f"Accuracy: {percentage_accuracy:.2f}%")
```

```
Accuracy: 85.88%
```

Figure 4. 8 Accuracy of K-Nearest-Neighbor Model

Classification model:

	precision	recall	f1-score	support
Active	0.87	0.97	0.92	2103
Withdrawn	0.93	0.72	0.81	1113
accuracy			0.88	3216
macro avg	0.90	0.84	0.86	3216
weighted avg	0.89	0.88	0.88	3216

Figure 4. 9 Classification report of K-Nearest-Neighbor

The classification report in fig 4.9 shows a detail summary of the K-Nearest Neighbor model evaluation in prediction of employee's turnover. The performance metrics for a model predict two classes: Active employees and Withdrawn employees. In case of active class, the result shows 87% of precision, 97% of recall and 92% of f1-score. In case of withdrawn class, the result shows 93% of precision, 72% of recall and 81% of f1-score.

Confusion Matrix:

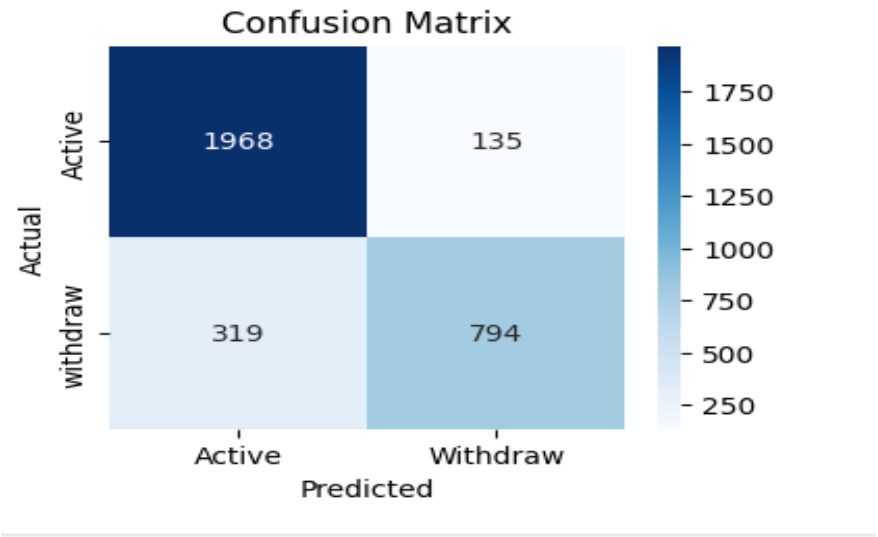


Figure 4. 10 Confusion Matrix of K-Nearest-Neighbor Model

The Confusion matrix in fig 4.10 shows the performance of K-Nearest-Neighbour in classifaying employees turnover prediction as active or withdrawn.The model correctly predicted 1968 instance of active class or true positive and 794 instance of withdrawn class or true negative.However it incorrectly predict 135 instance of active class as withdrawn and 319 instance of withdrawn class as active.

Pseudocode representation of K-Nearest Neighbors model:

1. Import necessary library – NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn, K-Nearest Neighbors
2. Load the dataset containing employee features and turnover labels
3. Data preprocessing:
   a. Handle missing values
   b. Convert categorical variables to numerical variable
   c. Normalize numerical variable
3. Split the data into training and testing sets
4. Initialize the Logistic Regression model with specified parameters
5. Fit the model using the training data:

a. Call Randomized search cv to optimize hyperparameter

6. Evaluate the model on the test set:

a. Compute accuracy, precision, recall, and F1-score, ROC

7. Output the model performance metrics

a. accuracy, precision, recall, and F1-score, ROC

4.8 Implementation of Hyperparameter Optimization

Hyperparameter tuning is the process of selecting the optimal value for the model to obtain the best possible performance by using different hyperparameter technique. In our study, we use random search for tuning the parameter. Random search is easier to implement, requires less computational time and efficient not all hyperparameters are equally important to tune. (Bergstra et al., 2012)

4.8.1 Hyperparameter tuning for LR Model

Logistic Regression is one of the supervised machine learning models that is used to solve binary classification tasks, making predictions or decisions based on historical data. It is used to predict binary response corresponding for a given set of independent variables. (Vaidya, 2017)

In Logistic Regression, various hyperparameters influence mode performance but the main parameter used for tuning are

- ✓ **Penalty:** Is a regularization term added to the loss function in an attempt to prevent overfitting by preventing overly complex models.
- ✓ **Inverse Regulation Strength (C):** Inverse of regularization strength in logistic regression; smaller values indicate stronger regularization.
- ✓ **Solver:** The algorithm used for optimization.
- ✓ **Max-iter:** maximum number of iterations allowed for the optimization algorithm to converge to a solution.
- ✓ **Class weight:** places weights on classes in imbalanced data to adjust the model sensitivity to every class.

Table 4. 2 Hyperparameter Tuning of LR model with Random Search CV

Parameters	Hyperparameter tuning	Optimized value
Penalty	L1, L2, Elastic net, None	L1
Inverse Regulation Strength(C)	0.001,0.01,0.1,1,10,100,1000	100
Solver	newton-cg, lbfgs, liblinear, sag, saga	Saga
Max-iter	100,200,300,400,500	300
Class Weight	None, balanced	None

As indicated in Figure 4.11 below, the optimized parameter value of Logistic Regression after using random search CV is presented.

```
param_dist = {
    'penalty': ['l1', 'l2', 'elasticnet', 'none'],
    'C': [0.001, 0.01, 0.1, 1, 10, 100, 1000],
    'solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'],
    'max_iter': [100, 200, 300, 400, 500],
    'class_weight': [None, 'balanced']
}

# Create a Logistic Regression model
lr = LogisticRegression()

# Create a RandomizedSearchCV object
random_search = RandomizedSearchCV(estimator=lr,param_distributions=param_dist, n_iter=10, cv=5,  random_state=42)

# Fit the RandomizedSearchCV object to the data
random_search.fit(X_train, y_train)

# Print the best parameters
print("Best parameters:", random_search.best_params_)

# Get the best model
best_lr = random_search.best_estimator_

# Evaluate the best model on the test set
accuracy = best_lr.score(X_test, y_test)
```

```
Best parameters: {'solver': 'saga', 'penalty': 'l1', 'max_iter': 300, 'class_weight': None, 'C': 100}
```

Figure 4. 11 Logistic Regression tuning result

4.8.2 Hyperparameter tuning for RF Model

Random Forest is one of the supervised machine learning algorithms that is used both for regression as well as classification problems. Random Forest extends the fundamental concept of a single classification tree by creating multiple trees to perform classification while the model becomes trained. In classifying an instance, each tree in the forest gives out its result (casts its vote in a class), and the algorithm takes the one with the highest votes among all trees in the forest.(Breiman, 2001)

Random forest hyperparameter tuning is important to improve the model performance. There is a common parameter used in tuning

- ✓ **n_estimation:** Number of trees in the forest. More trees improve performance but it will take to longer time to process.
- ✓ **Maximum Depth (Max-Depth):** Maximum depth of the tress. It is utilized to avoid overfitting.
- ✓ **Min_Sample_Spilt:** Minimum number of samples required to split an internal node
- ✓ **Min_Sample_Leaf:** Minimum number of samples that must be present in a leaf node
- ✓ **Max_Features:** Number of features to consider when looking for the best split

Table 4. 3 Hyperparameter Tuning of RF Model with Random Search CV

Parameters	Hyperparameter tuning	Optimized value
n_estimation	100,200,300	200
Max_Depth	None,10,20,30	20
Min_Sample_Split	2,5,10	10
Min_Sample_leaf	1,2,4	1
Max_Features	auto, sqrt, log2	Sqrt

As indicated in Figure 4.12 below, the optimized parameter value of Random Forest after using Random Search CV is presented.

```
param_dist_rf = {
    'n_estimators': [100, 200, 300],
    'max_depth': [None, 10, 20, 30],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'max_features': ['auto', 'sqrt', 'log2']
}
# Create a Logistic Regression model
rf_classifier = RandomForestClassifier(random_state=42)


# Create a RandomizedSearchCV object
random_search = RandomizedSearchCV(estimator=rf_classifier,param_distributions=param_dist_rf,n_iter=10,
              cv=5,scoring='accuracy',random_state=42
)

# Fit the RandomizedSearchCV object to the data
random_search.fit(X_train, y_train)

# Print the best parameters
print("Best parameters:", random_search.best_params_)

# Get the best model
best_rf_classifier = random_search.best_estimator_
```

```
Best parameters: {'n_estimators': 200, 'min_samples_split': 10, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': 20}
```

Figure 4. 12 Random Forest tuning result

4.8.3 Hyperparameter tuning for KNN

A supervised machine learning approach for classification and regression applications is K-Nearest Neighbors (KNN). To make predictions, it looks at the 'k' data points that are closest to a new data point in the training set. These predictions are then based on the average value (for regression) or the most common class (for classification).

KNN hyperparameter tuning is essential to improve the model performance. There is a common parameter used in tuning

- ✓ **n_neighbors:** Define how many nearest neighbors are considered when making predictions
- ✓ **Weights:** Determines how the neighbors influence the prediction.
- ✓ **Metric:** distance measure to use for calculating the distance between data points.
- ✓ **P:** type of distance metric used in Minkowski distance, with p=1 for Manhattan distance and p=2 for Euclidean distance.

Table 4. 4 Hyperparameter Tuning of KNN Model with Random search CV

Parameters	Hyperparameter tuning	Optimized value
n_neighbours	1to31	16
Weights	Uniform, distance	Uniform
Metric	Euclidean , Manhattan, Minkowski	Minkowski
P	1,2,	2

As indicated in Figure 4.13 below, the optimized parameter value of K-Nearest Neighbors after using Random Search CV are presented.

```
param_dist = {
    'n_neighbors': list(range(1, 31)),
    'weights': ['uniform', 'distance'],
    'metric': ['euclidean', 'manhattan', 'minkowski'],
    'p': [1, 2]
}

knn = KNeighborsClassifier()
random_search = RandomizedSearchCV(knn, param_distributions=param_dist, n_iter=10, cv=5, scoring='accuracy', random_state=42)


random_search.fit(X_train, y_train)

print("Best parameters:", random_search.best_params_)
best_knn = random_search.best_estimator_
```

```
Best parameters: {'weights': 'uniform', 'p': 2, 'n_neighbors': 16, 'metric': 'minkowski'}
```

Figure 4. 13 K-Nearest-Neighbor tuning result

4.9 Analysis of Hyperparameter tuning results for Models

The main purpose of hyperparameter tuning is to find the optimal combination of parameters used to reduce the error in the classification algorithm and refine the model performance. In this study, we use the random search CV technique for the tuning purpose of the selected model. The hyperparameter tuning used in Logistic Regression is the penalty, Inverse regulation strength, solver, max-iter, and class weight. The hyperparameter tuning used in Random forest is n_estimation, Max_Depth, Min_sample_split, Min_ sample_leaf, and Max_features. The hyperparameter tuning used in K-Nearest Neighbors (KNN) are n_neighbours, weights, metric, and p. The optimized values for each parameter are shown in table 4.2, table 4.3, and table 4.4. The results of hyperparameter tuning for each model, along with various evaluation metrics, are shown in Table 4.5.

Table 4. 5 Hyperparameter tuning result of models

Models	**Target value**	**Hyperparameter tuning result**			
		Precision	Recall	Fl-score	Accuracy
LR	Active (0)	0.86	0.97	0.91	87.59%
	Withdrawn (1)	0.92	0.70	0.80	
RF	Active (0)	0.87	0.98	0.92	88.87%
	Withdrawn (1)	0.94	0.72	0.82	
KNN	Active (0)	0.85	0.98	0.91	87.62%
	Withdrawn (1)	0.95	0.68	0.79	

Hyperparameter tuning result of accuracy for models: In logistic regression model, with penalty = l1, c =100, solver = saga, max iter = 300 and class weight = none parameter values the accuracy of the model is 87.59%, with shows a slight improvement from the previous result before tuning.

```
model = LogisticRegression(penalty='l1', dual=False, tol=0.0001,
                           C=100, fit_intercept=True, intercept_scaling=1,
                           class_weight=None, random_state=None,
                           solver='saga', max_iter=300, multi_class='auto',
                           verbose=0, warm_start=False, n_jobs=None, l1_ratio=None)
model.fit(X_train, y_train)
accuracy = model.score(X_test, y_test)
print("Accuracy:", accuracy)
y_pred = model.predict(X_test)
```

```
Accuracy: 0.8759328358208955
```

+ Code

```
percentage_accuracy = accuracy * 100
print(f"Accuracy: {percentage_accuracy:.2f}%")
```

```
Accuracy: 87.59%
```

Figure 4. 14 Accuracy result of Logistic Regression after tuning

For the Random forest model, with n-estimation = 200, max-depth = 20, min-sample-split = 10, , min-sample-leaf = 1 and max feature = sqrt parameter value used the accuracy of the model improved to 88.87% showing a slight enhancement similar to the observed model of logistic regression model.

```
model = RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,criterion="gini",
                               max_depth=20, max_features='sqrt',
                               max_leaf_nodes=None, max_samples=None,
                               min_impurity_decrease=0.0, min_samples_leaf=1,
                               min_samples_split=10, min_weight_fraction_leaf=0.0,
                               n_estimators=200, n_jobs=None, oob_score=False,
                               random_state=None,
                               verbose=0, warm_start=False)
model.fit(X_train, y_train)
accuracy = model.score(X_test, y_test)
print("Accuracy:", accuracy)
y_pred = model.predict(X_test)
```

```
Accuracy: 0.888681592039801
```

```
percentage_accuracy = accuracy * 100
print(f"Accuracy: {percentage_accuracy:.2f}%")
```

```
Accuracy: 88.87%
```

Figure 4. 15 Accuracy result of Random forest after tuning

The last model which is K-Nearest Neighbors which n-neighbor = 16, weights = uniform, metric = minkowski and p =2 parameter value recorded an accuracy of 87.62% indicating an improvement over its previous accuracy of 85.88%.

```
knn = KNeighborsClassifier(
    n_neighbors=16,  weights='uniform',  algorithm='auto',  leaf_size=30,
       p=2, metric='minkowski', metric_params=None, n_jobs=None,
)

# Fit the model to the training data
knn.fit(X_train, y_train)

# Make predictions on the test data
y_pred = knn.predict(X_test)

# Evaluate the accuracy of the model
accuracy = knn.score(X_test, y_test)
print("Accuracy:", accuracy)
```

```
Accuracy: 0.8762437810945274
```

```
percentage_accuracy = accuracy * 100
print(f"Accuracy: {percentage_accuracy:.2f}%")
```

```
Accuracy: 87.62%
```

Figure 4. 16 Accuracy result of K-Nearst Neighbour after tuning

Hyperparameter tuning result of classification report for models: In the classification report, the evaluation metrics of precision, recall, and f1 score after tuning parameter of each model are as follows: In logistic regression, the model achieves 86% of precision, 97% of recall and 91% of fl-score in case of active employees. In the case of withdrawn employees, the result shows 92% of precision, 70% of recall, and 80% of f1-score and there is an improvement in the value of f1 score for withdrawn employees before tuning which is 79%.

	precision	recall	f1-score	support
Active	0.86	0.97	0.91	2103
Withdrawn	0.92	0.70	0.80	1113
accuracy			0.88	3216
macro avg	0.89	0.83	0.85	3216
weighted avg	0.88	0.88	0.87	3216

Figure 4. 17 Classification report of Logistic regression after tuning

	precision	recall	f1-score	support
Active	0.87	0.98	0.92	2103
Withdrawn	0.94	0.72	0.82	1113
accuracy			0.89	3216
macro avg	0.91	0.85	0.87	3216
weighted avg	0.89	0.89	0.88	3216

Figure 4. 18 Classification report of Random Forest After tuning

The classification report in Fig 4.18 shows a detailed summary of the Random forest model evaluation in the prediction of employee's turnover after tuning. The performance metrics for a model predict two classes: Active employees and Withdrawn employees. In the case of active class, the result shows 87% of precision, 98% of recall, and 92% of f1-score. In the case of withdrawn class, the result shows 94% of precision, 72% of recall, and 82% of F1-score. The model enhances the value of recall from 97% to 98% in active employees. Also, improve the precision and f1-score result of withdrawn employees.

	precision	recall	f1-score	support
Active	0.85	0.98	0.91	2103
Withdrawn	0.95	0.68	0.79	1113
accuracy			0.88	3216
macro avg	0.90	0.83	0.85	3216
weighted avg	0.89	0.88	0.87	3216

Figure 4. 19 Classification repot of K-Nearst Neighbour after tuning

Fig 4.19 shows K-Nearest Neighbors result after tuning is, In the case of active class, the result shows 85% of precision, 98% of recall, and 91% of f1-score. In the case of withdrawn class, the result shows 95% of precision, 68 % of recall and 79 % of F1-score. The model enhances the value of recall from 97% to 98% After tuning the recall value for active employees and the precision value for withdrawn employees are enhanced.

Hyperparameter tuning result of confusion matrix for models:

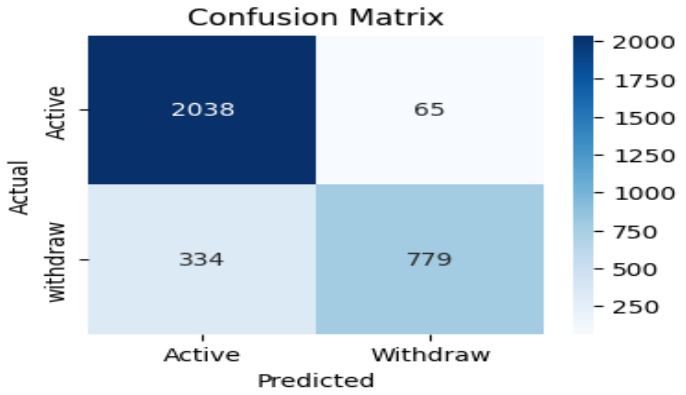


Figure 4. 20 Confusion matrix of Logistic Regression after tuning

The Confusion matrix in fig 4.20 shows the performance of Logistic Regresstion in classifaying employees turnover prediction as active or withdrawn.The model correctly predicted 2038 instance of active class or true positive and 779 instance of withdrawn class or true negative. However it incorrectly predict 65 instance of active class as withdrawn and 334 instance of withdrawn class as active.

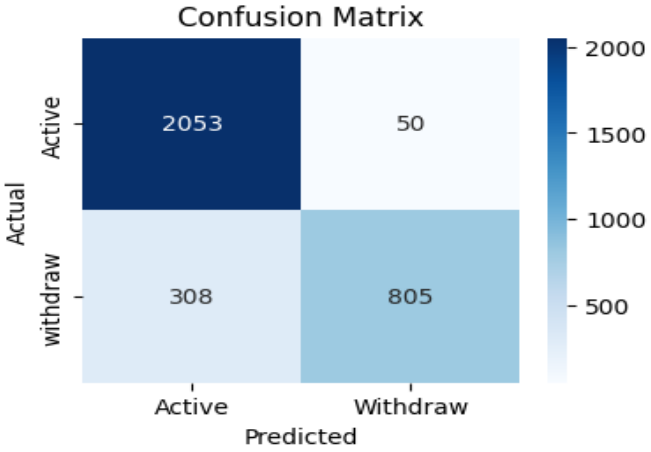


Figure 4. 21 Confusion matrix of Random forest after tuning

The Confusion matrix in fig 4.21 shows the performance of Random Forest in classifaying employees turnover prediction as active or withdrawn.The model correctly predicted 2053 instance of active class or true positive and 805 instance of withdrawn class or true negative. However it incorrectly predict 50 instance of active class as withdrawn and 308 instance of withdrawn class as active.

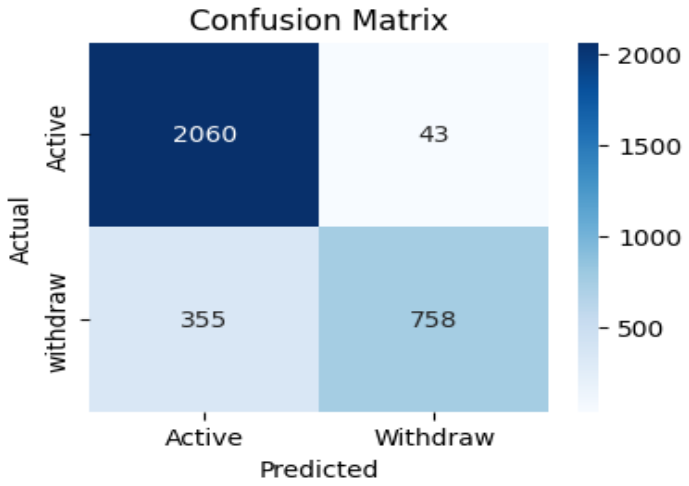


Figure 4. 22 Confusion matrix of K-Nearest-Neighbor after tuning

The Confusion matrix in fig 4.21 shows the performance of Random Forest in classifaying employees turnover prediction as active or withdrawn.The model correctly predicted 2060 instance of active class or true positive and 758 instance of withdrawn class or true negative. However it incorrectly predict 43 instance of active class as withdrawn and 355 instance of withdrawn class as active.

4.10 Cross validation result analysis

The cross-validation key approach is to divide the dataset into k times or fold then train the model on certain folds and test them on the rest of the folds. This technique ensures that every data point has a chance to be represented in both the training set and the testing set, thereby proving a robust evaluation of the performance model. In this study, we use 10-fold cross-validation which may better capture the overall performance of the model.(Qiu, 2024) Models prediction by using 10-fold cross-validation is shown in Fig 4.23. We use 10-fold for the selected model (LR, RF, & KNN) to improve the accuracy of the model. In this analysis, Random Forest archives an 88.31% average accuracy score which is the best, logistic regression achieves the second-best accuracy of 87.33% and K Nearest Neighbor archives the last score of 83.84%.

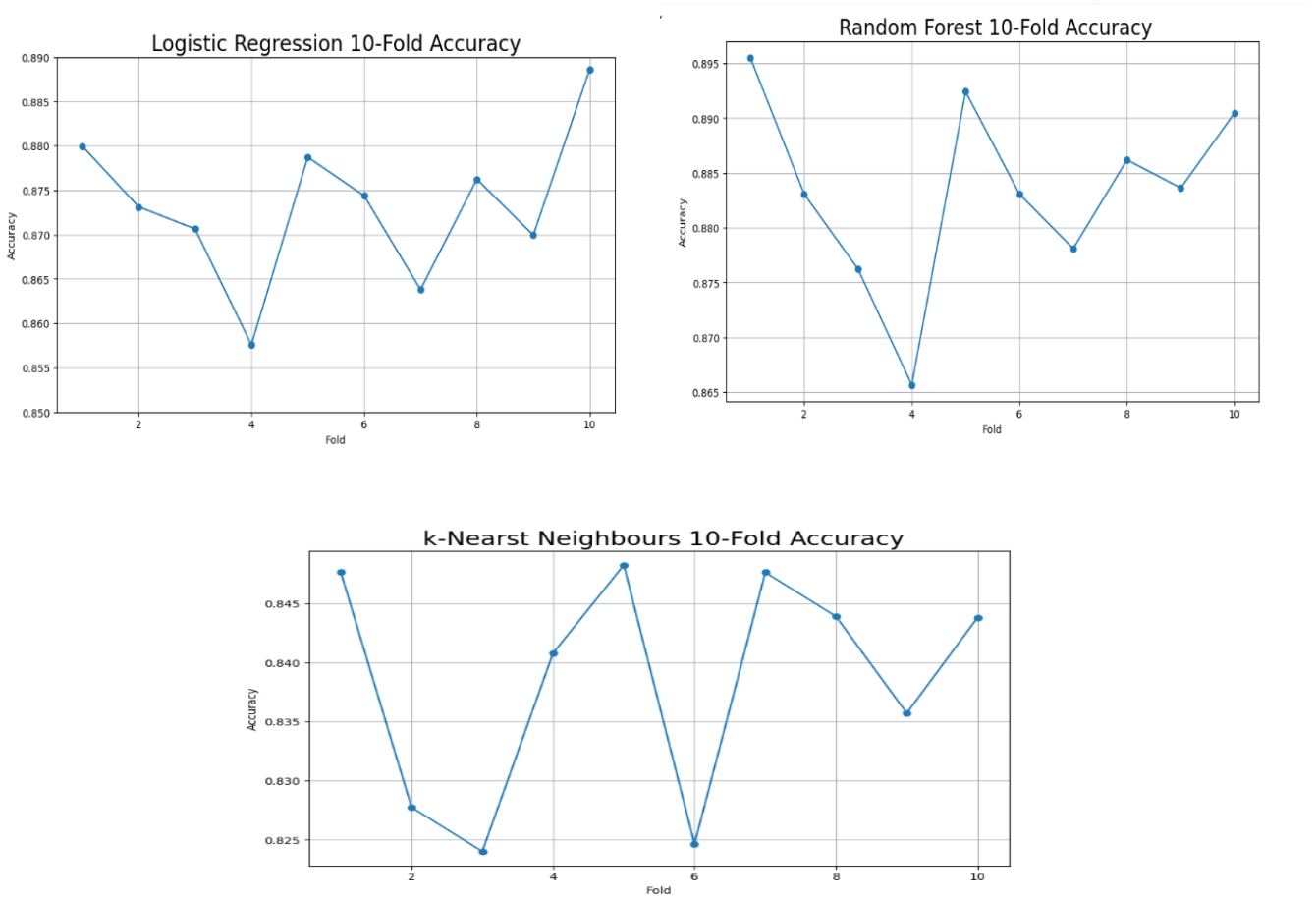


Figure 4. 23 Accuracy of the selected modes with 10-Fold Cross validation

4.11 Models Performance evaluation and analysis through AUC-ROC curve

ROC curve is used as a performance evaluation matrix for a different classification model. It represented in the graph that plots true positive rate and false positive rest. In our study, the models are evaluated using AUC-ROC curve. As indicated in fig 4.24, both Logistic Regression and Random Forest achieve AUC of 90%, While K-Nearest Neighbor achieves an AUC 87%.

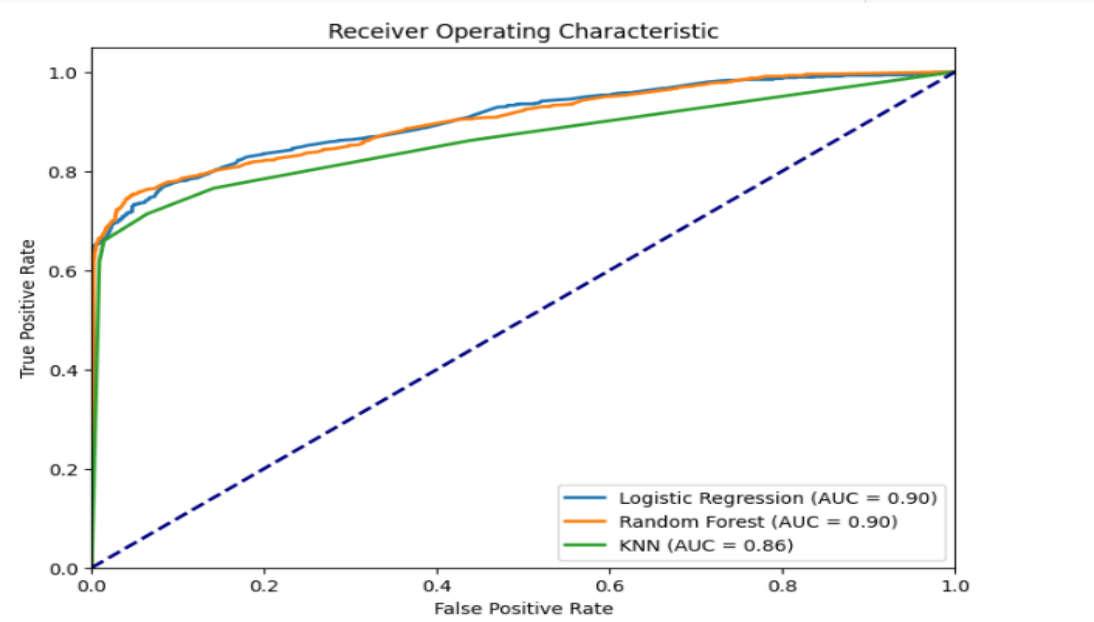


Figure 4. 24 Results of ROC Curve analysis

4.12 Comparison of Models

In this section, we evaluate and compare each model using key performance metrics, such as accuracy, precision, recall, and F1-score to predict employee's turnover. Figure 4.25 presents the comparison of the Logistic regression, Random forest, and K-Nearest-Neighbor with the performance evaluation matrix of accuracy. As indicated in Figure 4.25 Random forest achieved a higher accuracy result compared to the other model. KNN scores the second-best accuracy while logistic regression comes at last.

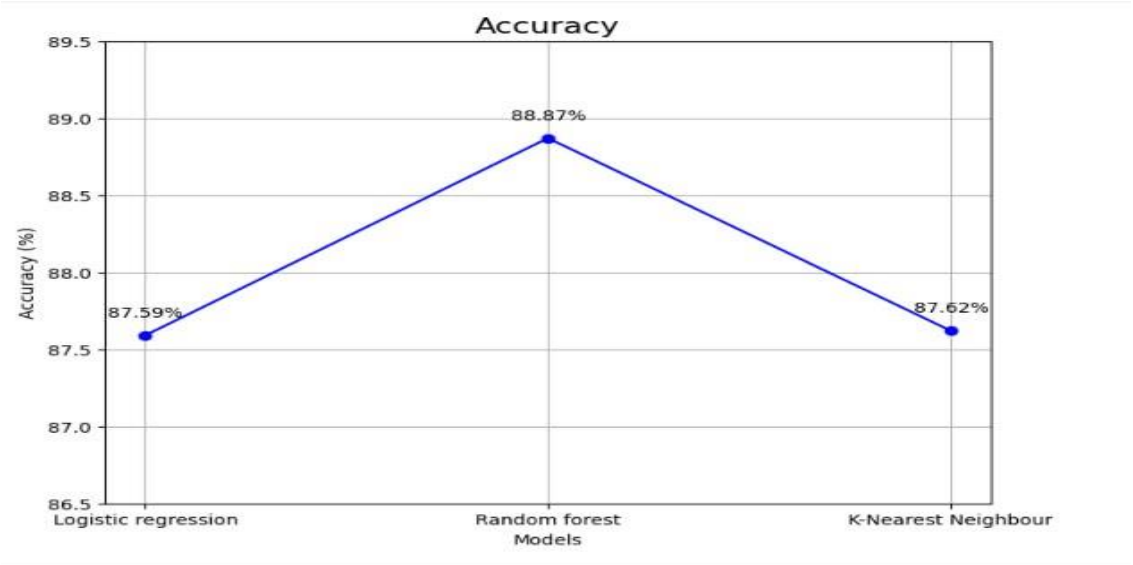


Figure 4. 25 Accuracy result of models

In precision the result is measure in both classes, in our case the two class active employees and turnover are measure with the selected models. As it indicates in figure 4.26 the result of random forest is score high comparing with Logistic Regression and K-Nearest Neighbor in case of active class. Conversely in case of withdrawn class KNN archive high result comparing with Random forest and Logistic Regression.

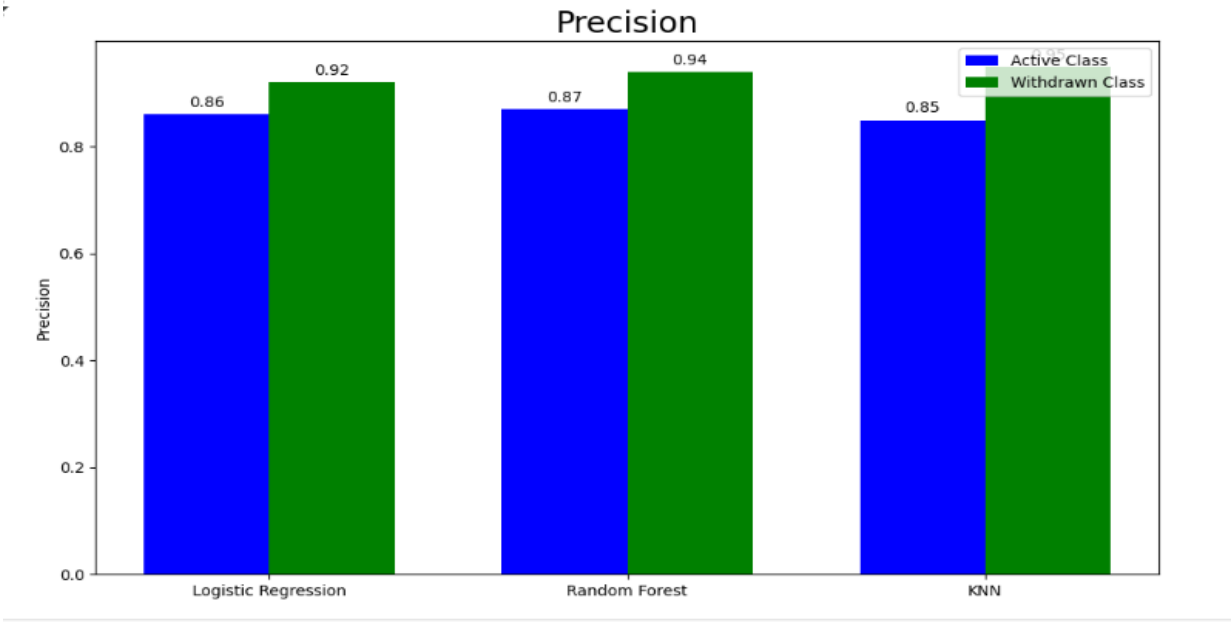


Figure 4. 26 Precision result of models in both classes

In recall, the results of the models are shown for both the active and withdrawn class. As it indicates in figure 4.27 the Random forest model and K –Nearest Neighbor archive similar performance levels for the active class. In case of withdrawn class Random forest score high result comparing with logistic regression and random forest.

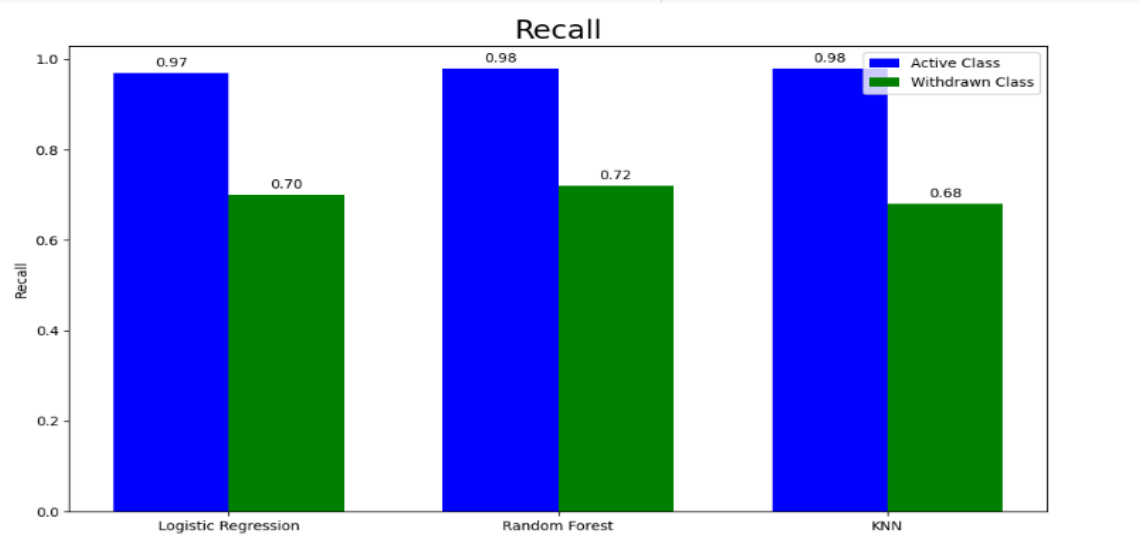


Figure 4. 27 Recall result of models in both classes

Figure 4.28 shows the F1-score performance of the selected models with both active and withdrawn classes. As indicated in Figure 4.28 Random forest performed high results in both active and withdrawn classes compared with other selected models. Both logistic regression and k-Nearest-Neighbor archive similar result in case of active class.

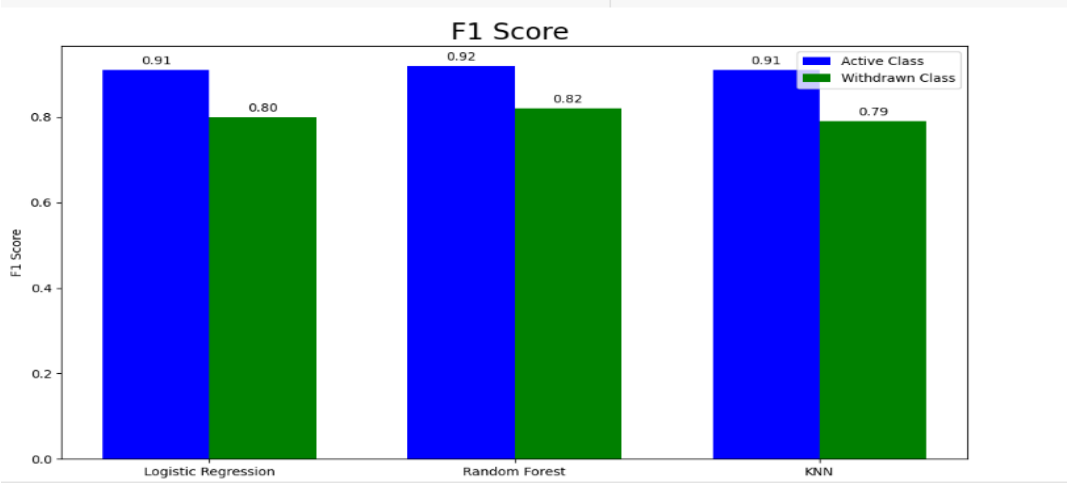


Figure 4. 28 F1-Score result of models in both classes

4.13 Identify Significate factor

To identify the significant factors that play a crucial role in predicting employee turnover, we utilize random forest feature importance. This technique helps to select the most relevant factors associated with employee turnover.

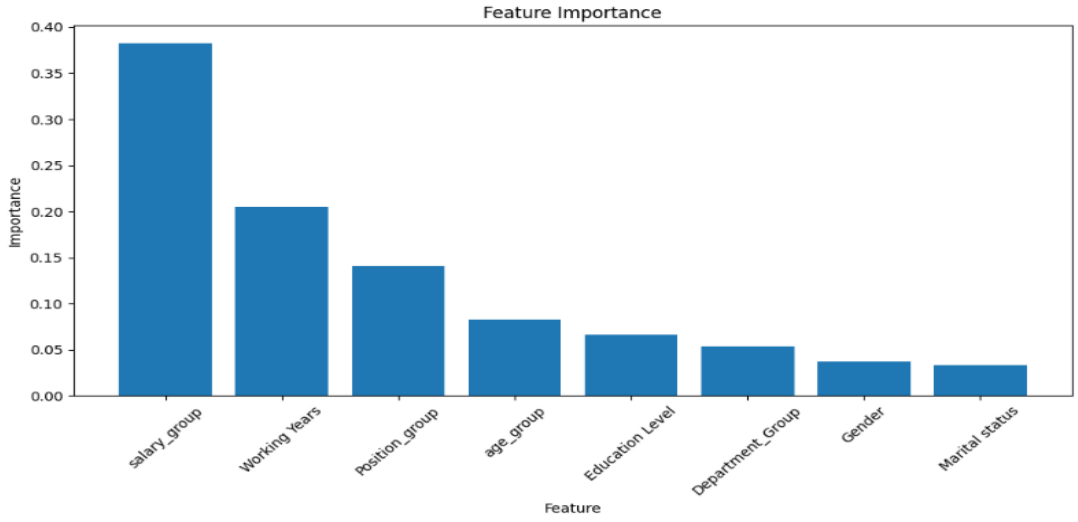


Figure 4. 29 Feature Importance

Figure 4.29 indicates the importance of features from the highest to the lowest. In this case, the most important factors that affect employee turnover is Salary. Year of service, position, and age are the second third, and fourth most important factors. However, education level, department, gender, and marital status are the low factors that affect employee turnover.

4.14 Discussion of Result

The main objective of this research is to develop a machine-learning model that predicts employee turnover at Adama Industrial Park and to provide recommendations for reducing turnover rates. The machine learning models used to conduct this research are LR, RF, and KNN. In this research, we have a total of 16078 employees' data out of these 65.6 % or 10545 employees are active and 34.4% or 5533 employees are turnover. We split the data in 80:20 ratios meaning 80% (12862) we use for training purposes and 20% (3216) for testing purposes. For each selected algorithm, the model is trained on the training dataset and evaluated on the test dataset created. All experiments, including logistic regression, k nearest neighbor, and random forest are performed separately on the same training and test data.

The selected models are tuned using a random search CV to enhance the result of the model. The results obtained from the three classification models are compared with each other using the performance evaluation metrics, namely accuracy and classification report, which include precision, recall, F1-score, and confusion metrics. The analysis revealed, as discussed in the above section, that the Random forest model outperforms 88.87% accuracy score which is a high result compared to Logistic Regression and K-Nearest-Neighbor. The accuracy results for Logistic regression and K-Nearest-Neighbor are 87.59% and 87.62% respectively. This shows that Random Forest provides the most accurate predictions among the models tested. In the classification report of model results that is discussed in the above section, the precision score of the random forest result is 87% in the active class and the K-Nearest-Neighbor result is 95% in the withdrawn class. Precision measures the correctness of positive predictions; in this case, KNN is more accurate when predicting withdrawn employees as compared to Random Forest in the withdrawn class. However, Random Forest shows competitive precision in the active class (87%), indicating that it performs well when predicting active employees as well. Looking at recall, both Random Forest and K-Nearest Neighbor achieve a high recall of 98% for the active class, meaning that both models correctly identify 98% of all active employees. For the withdrawn class, Random Forest has a 72% recall. In terms of the F1-score, which balances precision and recall, the Random Forest model outperforms both Logistic Regression and K-Nearest Neighbor with a 92% F1-score in the active class and an 82% F1-score in the withdrawn class. The high F1 scores for both classes indicate that Random Forest provides a strong

balance between precision and recall, making it a highly reliable model overall. By using the dataset 10-fold cross-validation Random forest outperforms an average accuracy of 88.31% which is the highest score while Logistic regression and K-nearest neighbor performance are 87.33 and 83.84 respectively. Cross-validation helps evaluate a model's ability to generalize to unseen data, and Random Forest's higher performance in this context suggests it is more robust than the other models. In the Case of the AUC-ROC curve Random forest and logistic regression both achieve an AUC of 0.90 while K-Nearest Neighbor achieves AUC of 0.86 which has a slightly lower comparing with the other. In conclusion, the Random Forest model consistently outperforms the other models across multiple performance metrics.

The second finding of this study is to identify the main important factor that affects employee turnover. To identify the main cause of employee turnover we use the Random Forest feature importance and the result displays that salary is the main importance factor. These findings suggest that employees are more likely to leave an organization due to dissatisfaction with their pay. Year of service, position, and age are the second third, and fourth most important factors. However, education level, department, gender, and marital status are the low factors that affect employee turnover.

CHAPTER FIVE

CONCLUSION, CONTRIBUTION AND RECOMMENDATION

5.1 Conclusion

To conduct this research, the necessary data for the experiment was obtained from the Adama Industrial Park IPDC (Industrial Park Development Corporation) HR office in Excel format which is a total dataset of 16078. The researcher faced several challenges while conducting this study. The first major obstacle was the organizations are not willing to share their data for research purposes. Additionally, issues such as unstructured, noisy, missing, and inconsistent data, required significant time for preprocessing, which ultimately affected the accuracy of the classification models. Furthermore, the lack of machine learning research studies and the limited availability of relevant literature are another challenge during the study. The necessary preprocessing activities were applied to the dataset that we early explain in chapter three. After that classification algorithms are used for building our model. By using the three-classification algorithm namely logistic regression, k-nearest neighbor, and random forest experiments are done. Random Search CV is used for tuning the parameters of the model to enhance the model performance to achieve a high score. Python programming languages are used to execute the code. Models are measured using the performance evaluation Metrics of accuracy, precision, recall, f1score, and confusion matrix. Models are also compared using the ROC-AUC curve .10-fold cross-validation is used to compare the average accuracy value of each model. The three classification models are compared to get appropriate for predicting employee's turnover. According to our research random forest model outperforms 88.77% accuracy which is a high score compared with logistic regression and random forest. To obtain the factors that affect employee turnover feature importance is implemented and salary is the main factor that affects employee turnover.

In this research, we use local data which have quite dirty, and take all the preprocessing technique use for make suitable for algorithm. It is recommended to use advanced technology in real life HR management system to take a proactive step of employees' retention strategy. This study can be used as a starting point for any academic research work in this area.

5.2 Contribution of the study

- ✓ The study develops a predictive framework for the early detection of employee turnover risk based on historical employee data
- ✓ The study ensures comprehensive preprocessing and hyperparameter tuning for optimal performance
- ✓ The study Provides data-driven decision-making and organizational practices for employee retention.
- ✓ The study Evaluates machine learning models to determine the most effective approach.
- ✓ The study explores and identifies the most significant factors influencing employee turnover in the specific organization.
- ✓ The study will serve as a baseline for benchmarking local data on employee turnover in Ethiopia and similar developing economies for future research.

5.3 Recommendation

We recommend that every institution and organization utilize the products of machine learning methods for predicting employee turnover because they are more accurate and effective than traditional methods.

5.4 Future Work

Although this investigation is primarily for academic purposes, it will significantly contribute to human resources and benefit other researchers in related fields. While the findings of this study are encouraging, some areas require further exploration to develop a more comprehensive model and take it to an operational level. Consequently, the researcher suggests the following topics as future research directions according to this study:

- ✓ To gain a deeper understanding of turnover dynamics, it would be beneficial to include additional variables such as overtime, promotion history, performance rating, attendance, workload, and other relevant aspects. Expanding the dataset will enable more accurate analyses and enhance the prediction model
- ✓ By leveraging deep learning, researchers can uncover subtle patterns and interactions that machine learning methods might miss.
- ✓ Explore time series forecasting to predict the upcoming year.
- ✓ Explore other parameter tuning techniques to enhance model performance.
- ✓ Explore hybrid machine learning models to improve predictive performance.

REFERENCES

Abbasi, S. M., & Hollman, K. W. (2000). *Turnover: The Real Bottom Line*.

Abdali, F. (2011). Impact of Employee Turnover on Sustainable Growth of Organization in Computer Graphics Sector of Karachi, Pakistan. In *Afro Asian Journal of Social Sciences* (Vol. 2, Issue 2).

Abhishek, J. (2024, September). *Everything about Random Forest. Random Forest is one of the most… | by Abhishek Jain | Medium*. https://medium.com/@abhishekjainindore24/everything-about-random-forest-90c106d63989

A Ilemobayo, J., Durodola, O., Alade, O., J Awotunde, O., T Olanrewaju, A., Falana, O., Ogungbire, A., Osinuga, A., Ogunbiyi, D., Ifeanyi, A., E Odezuligbo, I., & E Edu, O. (2024). Hyperparameter Tuning in Machine Learning: A Comprehensive Review. *Journal of Engineering Research and Reports*, *26*(6), 388–395. https://doi.org/10.9734/jerr/2024/v26i61188

Al-Darraji, S., Honi, D. G., Fallucchi, F., Abdulsada, A. I., Giuliano, R., & Abdulmalik, H. A. (2021). Employee attrition prediction using deep neural networks. *Computers*, *10*(11). https://doi.org/10.3390/computers10110141

Awad, M., & Fraihat, S. (2023). Recursive Feature Elimination with Cross-Validation with Decision Tree: Feature Selection Method for Machine Learning-Based Intrusion Detection Systems. *Journal of Sensor and Actuator Networks*, *12*(5). https://doi.org/10.3390/jsan12050067

Beauxis-Aussalet, E. (2014). *Simplifying the Visualization of Confusion Matrix*. https://www.researchgate.net/publication/302412429

Bergstra, J., Ca, J. B., & Ca, Y. B. (2012). Random Search for Hyper-Parameter Optimization Yoshua Bengio. In *Journal of Machine Learning Research* (Vol. 13). http://scikit-learn.sourceforge.net.

Berrar, D. (2024). *Cross-validation 1*.

Breiman, L. (2001). *Random Forests* (Vol. 45).

Cepheus. (2020). *Ethiopia's Industrial Parks: A Data Pack on recent performance*.

Chakraborty, R., Mridha, K., Shaw, R. N., & Ghosh, A. (2021, September 24). Study and Prediction Analysis of the Employee Turnover using Machine Learning Approaches. *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies, GUCON 2021*. https://doi.org/10.1109/GUCON50781.2021.9573759

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785

Dwivedi, R. Kumar. (2018). *Proceedings of the 2018 International Conference on System Modeling & Advancement in Research Trends : SMART-2018 : (23rd-24th November, 2018)*. Prof. Rakesh Kumar Dwivedi : Principal College of Computing Sciences & Information Technology, Teerthanker Mahaveer University.

Ekhsan, M. (2019). The influence job satisfaction and organizational commitment on employee turnover intention. In *Management, and Accounting* (Vol. 1). http://e-journal.stie-kusumanegara.ac.id

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/J.PATREC.2005.10.010

Glebbeek, A. C., & Bax, E. H. (2004). *Is high employee turnover really harmful? an empirical test using company records*.

Guerranti, F., & Dimitri, G. M. (2023). A Comparison of Machine Learning Approaches for Predicting Employee Attrition. *Applied Sciences (Switzerland)*, *13*(1). https://doi.org/10.3390/app13010267

Hailu, S. (2016). *Perceived cause of employee turnover: the case of shints etp garment plc*.

Halvorsen, S. K. (2021). *Labour turnover and workers' well-being in the Ethiopian manufacturing industry* (Vol. 2021). UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2021/974-7

Horvitz, E., & Mulligan, D. (2015). Data, privacy, and the greater good. *Science*, *349*(6245), 253–255. https://doi.org/10.1126/science.aac4520

Hossen, M. A., Hossain, E., Ishwar, A. K. Z., & Siddika, F. (2021). Ensemble method based architecture using random forest importance to predict employee's turn over. *Journal of Physics: Conference Series*, *1755*(1). https://doi.org/10.1088/1742-6596/1755/1/012039

Jordan, M., Kleinberg, J., & Schölkopf, B. (2006). *Pattern Recognition and Machine Learning*.

Kirasich, K. ;, Smith, T. ;, & Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. In *SMU Data Science Review* (Vol. 1, Issue 3). https://scholar.smu.edu/datasciencereviewAvailableat:https://scholar.smu.edu/datasciencereview/vol1/iss3/9http://digitalrepository.smu.edu.

Lahkar Das, B., & Baruah, M. (2013). *Employee Retention: A Review of Literature* (Vol. 14, Issue 2). www.iosrjournals.orgwww.iosrjournals.org

Lekan, A. J., Akinode, L., & Bada, O. (2022). *The Federal Polytechnic, Ilaro, 16 th & 17 th*. https://www.researchgate.net/publication/364322002

Lemma, B. (2019). *School of graduate studies an assessment of factors affecting employees' turnover intention in ethiopian revenues and customs authority addis ababa, ethiopia*.

Lim, C. S., Malik, E. F., Khaw, K. W., Alnoor, A., Chew, X. Y., Chong, Z. L., & Akasheh, M. Al. (2024). Hybrid GA–DeepAutoencoder–KNN Model for Employee Turnover Prediction. *Statistics, Optimization and Information Computing*, *12*(1), 75–90. https://doi.org/10.19139/soic-2310-5070-1799

Mandefro, T. S. (2022). *College of business and economics department of management master of science in management factor that affect employees turnover in ethiopia, incases of hibret bank s.c*.

Mbah, S. E., & Ikemefuna, C. O. (2012). Job Satisfaction and Employees' Turnover Intentions in total Nigeria plc. in Lagos State. In *International Journal of Humanities and Social Science* (Vol. 2, Issue 14). www.ijhssnet.com

Meseret, W., Negash, R., & Mesfin Mekonnin, M. (2020). *Factors affecting employee turnover (at erca large taxpayers branch office) Under the Supervision of*.

Misilmani, H. M. E. , & N. T. (2019). *2019 International Conference on High Performance Computing & Simulation (HPCS)*. IEEE.

Nahar, L., Tasnim, F., Sultana, Z., & Tuli, F. A. (2022). Employee Turnover Prediction Model for Garments Organizations of Bangladesh Using Machine Learning Technique. *2022 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2022*. https://doi.org/10.1109/IEMTRONICS55184.2022.9795701

Nasteski, V. (2017). An overview of the supervised machine learning methods. *HORIZONS.B*, *4*, 51–62. https://doi.org/10.20544/horizons.b.04.1.17.p05

Nelly anzazi. (2018). *Employee turnover on organizational performance in the telecommunication industry in kenya nelly anzazi A Research Project Submitted to the School of management and leadership in partial Fulfillment of the requirement for the Degree of Bachelor in Leadership CORE View metadata, citation and similar papers at core*.

Omari, K. (2023). Comparative Study of Machine Learning Algorithms for Phishing Website Detection. *International Journal of Advanced Computer Science and Applications*, *14*(9), 417–425. https://doi.org/10.14569/IJACSA.2023.0140945

Ongori, H. (2007). A review of the literature on employee turnover. *African Journal of Business Management*, 49–054. http://www.academicjournals.org/ajbm

Pandey, A., & Jain, A. (2017). Comparative Analysis of KNN Algorithm using Various Normalization Techniques. *International Journal of Computer Network and Information Security*, *9*(11), 36–42. https://doi.org/10.5815/ijcnis.2017.11.04

Panigrahi, T., & Rout, M. (2020). *"Cause, Effect and Remedies of Employee Turnover": A Critical Literature Review First Author Second Author*.

Pratt, M., Boudhane, M., & Cakula, S. (2021). Employee attrition estimation using random forest algorithm. *Baltic Journal of Modern Computing*, *9*(1), 49–66. https://doi.org/10.22364/BJMC.2021.9.1.04

Qamar, N. (2022). The impact of machine learning on human. *Neuro Quantology*, *20*(13), 2424–2429. https://doi.org/10.14704/nq.2022.20.13.NQ88301

Qiu, J. (2024). An Analysis of Model Evaluation with Cross-Validation: Techniques, Applications, and Recent Advances. *Advances in Economics, Management and Political Sciences*, *99*(1), 69–72. https://doi.org/10.54254/2754-1169/99/2024OX0213

Raza, A., Munir, K., Almutairi, M., Younas, F., & Fareed, M. M. S. (2022). Predicting Employee Attrition Using Machine Learning Approaches. *Applied Sciences (Switzerland)*, *12*(13). https://doi.org/10.3390/app12136424

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. In *SN Computer Science* (Vol. 2, Issue 3). Springer. https://doi.org/10.1007/s42979-021-00592-x

Selhadin, I. (2019). *School of graduate studies causes of employee turnover: the case of global insurance company addis ababa, ethiopia*.

Skelton, A. R., Nattress, D., & Dwyer, R. J. (2020). Predicting manufacturing employee turnover intentions. *Journal of Economics, Finance and Administrative Science*, *25*(49), 101–117. https://doi.org/10.1108/JEFAS-07-2018-0069

Vaidya, A. (2017). *Predictive and probabilistic approach using logistic regression: application to prediction of loan approval*.

Venkatesan, R., Ponsudhakar, R., & Scholar, R. (2022). An emprical study of employees turnover and retention strategies of diraa hr services-an overview study. In *International Journal of Creative Research Thoughts* (Vol. 10, Issue 6). www.ijcrt.org

Yadav, S., Jain, A., & Singh, D. (2018). Early Prediction of Employee Attrition using Data Mining Techniques. *Proceedings of the 8th International Advance Computing Conference, IACC 2018*, 349–354. https://doi.org/10.1109/IADCC.2018.8692137

Yahia, N. Ben, Hlel, J., & Colomo-Palacios, R. (2021). From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction. *IEEE Access*, *9*, 60447–60458. https://doi.org/10.1109/ACCESS.2021.3074559

Yankeelov, P. A., Barbee, A. P., Sullivan, D., & Antle, B. F. (2009). Individual and organizational factors in job retention in Kentucky's child welfare agency. *Children and Youth Services Review*, *31*(5), 547–554. https://doi.org/10.1016/j.childyouth.2008.10.014

Yedida, R., Reddy, R., Vahi, R., & Kulkarni, D. (2018). *Employee Attrition Prediction*.

APPENDEX A

Import the necessary library

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

```
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.datasets import make_classification
from sklearn.metrics import accuracy_score, precision_score, recall_score,f1_score
```

Data Cleaning

```
imputer = SimpleImputer(strategy='most_frequent')
df_imputed = pd.DataFrame(imputer.fit_transform(df), columns=df.columns)
df_imputed.head()
```

Data Binning

```
# Define the bins and labels
bins = [16, 25, 35, 45, float('inf')]  # Define bins
labels = ['young', 'adult', 'matured', 'elder']  # Define labels

# Create a new column 'age_group' based on the 'Age' column
dta1['age_group'] = pd.cut(dta1['Age'], bins=bins, labels=labels, right=True)
```

Data Encoding

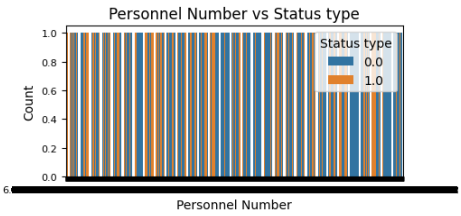
```
# One-hot encode other categorical columns and combine with encoded 'Status type'
categorical_cols = ['age_group', 'Gender', 'Marital status', 'Working Years','Education Level', 'Leaving Reason ', 'salary_group', 'Departmen
ohe = OneHotEncoder(sparse_output=False, handle_unknown='ignore')
encoded_data = ohe.fit_transform(new_dataframe[categorical_cols])
encoded_df = pd.DataFrame(encoded_data, columns=ohe.get_feature_names_out(categorical_cols)).astype(int)

final_df = pd.concat([new_dataframe[['Status type']], encoded_df], axis=1)
```

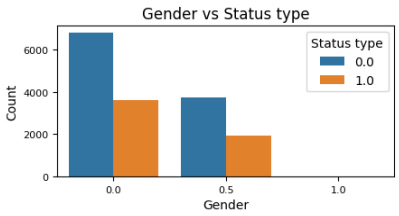
Train-test split

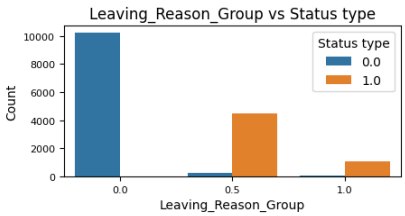
```
# spearate target variable & features
X = df.drop('Status type', axis=1)
y = df['Status type']
```

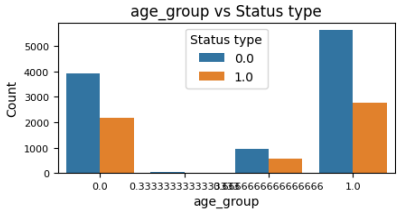
APPENDIX B

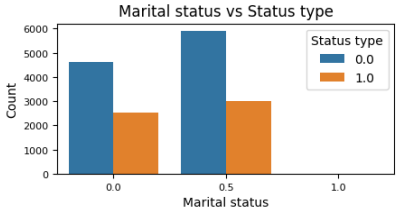
Frequency distribution of target variable with each independent features

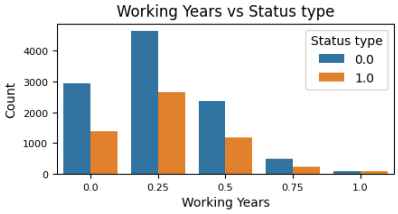


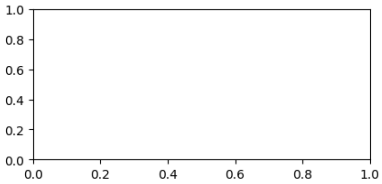


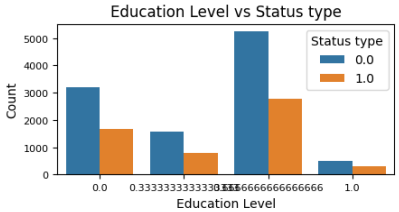


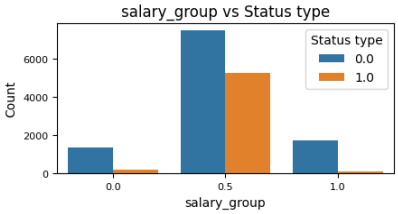


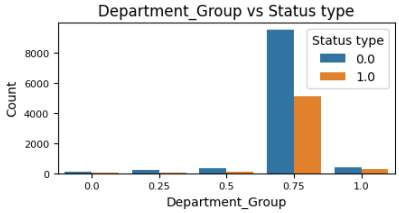


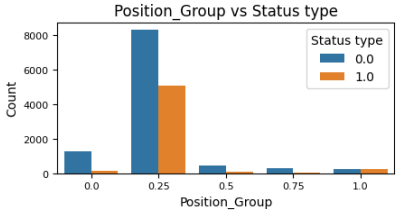


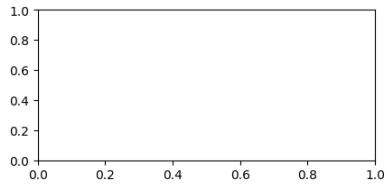












APPENDEX C

Random Forest feature importance

```
# Split the data into train and test
X = df.drop('Status type', axis=1)
y = df['Status type']

X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2,random_state=123,)

# Train the model
rf_clf = RandomForestClassifier()
rf_clf.fit(X_train , y_train)

# Get feature importances
importances = rf_clf.feature_importances_

# Create a list of feature names
feature_names = X_train.columns

# Create a DataFrame to store feature importances
feature_importances = pd.DataFrame({'feature': feature_names, 'importance': importances})

# Sort the DataFrame by importance
feature_importances = feature_importances.sort_values('importance', ascending=False)

# Print the feature importances
print(feature_importances)

# Plot the feature importances
plt.figure(figsize=(10, 6))
plt.bar(feature_importances['feature'], feature_importances['importance'])
plt.xticks(rotation=45)
plt.xlabel('Feature')
plt.ylabel('Importance')
plt.title('Feature Importance')
plt.tight_layout()
plt.show()
```