ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

PREDICT THE MAJOR FACTORS THAT HELPS TO PREDICT EMPLOYEE TURNOVER IN GOVERNMENT ORGANIZATION USING MACHINE LEARNING:- THE CASEOF ETHIOPIAN FEDERAL COURT

A Thesis Submitted to the School of Graduate Studies of Addis AbabaUniversity in Partial Fulfillment of the Requirements for the Degree of Master ofInformation Science

BY

ERISTIE ATINAF ASRESS

MAY 2020

Addis Ababa

Ethiopia

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCE

SCHOOL OF INFORMATION SCIENC

PREDICT THE MAJOR FACTORS THAT HELPS TO PREDICT EMPLOYEE TURNOVER IN GOVERNMENT ORGANIZATION USING MACHINE LEARNING:- THE CASE OF ETHIOPIAN FEDERAL COURT

A Thesis Submitted to the School of Graduate Studies of Addis AbabaUniversity in Partial Fulfillment of the Requirements for the Degree of Master ofInformation Science

BY

ERISTIE ATINAF ASRESS

Advisor:-

Dr.Wondwossen Mulugeta (PhD.)

MAY 2020

Addis Ababa

Ethiopia

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

PREDICT THE MAJOR FACTORS THAT HELPS TO PREDICT EMPLOYEE TURNOVER IN GOVERNMENT ORGANIZATION USING MACHINE LEARNING:- THE CASE OF ETHIOPIAN FEDERAL COURT

A Thesis Submitted to the School of Graduate Studies of Addis AbabaUniversity in Partial Fulfillment of the Requirements for the Degree of Master ofInformation Science BY

ERISTIE ATINAF ASRESS

Name and Signature of Members of the Examining Board

Name Title Date Signature

Wondwossen Mulugeta (PhD)Advisor_____ _____

Melkamu Beyene (PhD)Examiner _____ _____

Michael Melesse (PhD)Examiner _____ _____

DECLARATION

I declare that this thesis is my original work and has not been submitted as a partial requirement for a degree in any university

Name: _____

Signature: _____

Date: _____

This Thesis has been submitted for examination with my approval as a university advisor

Dr.Wondwossen Mulugeta

Name: _____

Signature: _____

Date: _____

DEDICATION

I would like to dedicate this thesis to my parents, my lovelyhusband, my lovely child Hemen, brothers and sisters for all their loveand support.

ACKNOWLEDGEMENTS

First of all, I would like to be thankful to my almighty God and His Mother Saint Marryfor giving me strength,hope andsupporting my thesis work from the starting to the end.

I would like to express my appreciation and deepest thanks to my advisor, Dr.Wondwossen Mulugeta for his keenguidance and advising. I am really thankful for hisoptimistic comments and critical readings of the study. His comments have helped me in maintaining the right way for my study and making itmeaningful.Inaddition my advisor, I would like to thank for Dr.Michael, his comment during our meetings acquire important contribution to my work and helps to enhance the quality of my work.

I would also thank my family who has been always with me through this study by encouraging and requesting me about the progress of my thesis work.Also my belovedhusband, Biruk Bayable, deserve lots of credits for he gives initiate and encouragelearningthe master program.

I am also really thankful to the Addis Ababa University Female Scholarship Departments for sponsoring full Admission and Scholarships for my master‘s studies. Moreover, I am gratefulto the organization (Ethiopian Federal Court)starting from manager to experts, especially to Ato Solomon and W/r Ababa who offered me all thenecessary information, data and documents for my thesis paper.

I would like to thank my classmate, for their friendly and supportive approach while doing group project works, assignments and general academic activities in the previous two years.

Last but not least I would like to express my gratefulness to my friends,colleagues’ and staffs especially, to Ato Mindaforsupporting me throughout my study. I amequally grateful to all who helped me in completing this study.

Thank you all.

ERISTIE ATINAF

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

HR	**Human Resource**
HRM	**Human Resource Management**
HRIS	**Human Resource Information System**
ICT	**Information Communication Technology**
EFC	**Ethiopian Federal Court**
ML	**Machine Learning**
AI	**Artificial Intelligence**
OLS	**Ordinary Least Square**
SVM	**Support Vector Machine**
DT	**Decision Tree**
NB	**Naïve Bayesian**
ANN	**Artificial Neural Network**
NN	**Neural Network**
GUC	**Graphical User Interface**
XGB	**Extreme Gradient Tree**
GBDT	**Gradient Boosting Decision Tree**
GB	**Gradient Boosting**
RF	**Random Forest**
SMOTE	**Synthetic Minority over Sampling**
DM	**Data Mining**

CSV	**Comma Separated Values**
TPR	**True Positive Rate**
FPR	**False Positive Rate**
TNR	**True negative Rate**
TP	**True Positive**
FP	**False Positive**
FN	**False Negative**
AUC	**Area under Curve**
ROC	**Receiver OperatingCharacter**

TABLE OF CONTENTS

Contents

ABSTRACT

Nowadays, Employee turnover is a serious issue in organizations. It affects the time, productivity, and stability of the given organizations. Employees are very important that helps the organization get success and gain revenue. So, Organizations need to know the key issues that the reason for employee turnover. Prediction models are highly associated with human resource management to identify the employee turnover patterns from employee previously recorded data. The objective of this research is to design a model and predicting staff turnover using a machine learning approach in the Ethiopian Federal court organization. For prediction three classification models namely, random forest, logistic regression and gradient boosting tree were used. The total datasets from the three federal court organizations were 3610 both active and terminated.For evaluate the prediction classification models the researcher was use confusion matrix, recall, precision and roc-curve to measure the performance of the classifiers. After evaluation, from the three classification models the finding shows that the best classification model is gradient boosting tree with an accuracy of 87.5%. Additionally, from the study it is found that the factors responsible for employee turnover are:-experience, salary, age and employee's number of year service are the most significant factors. The factors martial and gender were low predictor variables on employee turnover in the federal court organization.The study concludes that the most reliable and accurate classification model to predict employee turnover isan ensemble – based learning technique gradient boosting tree that was found as the most suitable classifier for building the predictive model.

CHAPTER ONE: Background and Justification of the Study

1.1 INTRODUCTION

According toGao(2017) the percentage of employees that a company wants to replace during a given period is called employee turnover rate. For any organization, employees are very essentialmainly those with full knowledgeable and skilled persons. Good employees will facilitate the organization get success and increaseprofits. Thus how to keep workers continue in the company becomes a serious problem to consider in human resource management. However, employee turnover has become a very well-known phenomenon. More and more employees leave their original companies for different reason. Some people may worry about the future job opportunity, some employees are not satisfied with their salary, some may assume the working hours are too long and working environment are not good, some don't like the quality of the supervisor and they think their work lack communication, etc. It will persuade time, productivity and continuity, etc. So it isof great importance for companies to search the reason that influence turnover and solve this problem in an effective and efficient way. This huge collection of employee data within an organization, especially those possessed by the Human Resource (HR) arm of the organization can be analyze for the effective prediction of employee turnover.

With the broad usage of computers and the internet, there has recently been a vastraise in publicly accessible data that can be analyzed. Be it online sales information, website traffic, or user habits, data is generate every day. Such huge amounts of data present both a problem and an opportunity. The problem is that it is hard for humans to analyze such big data. The opportunity is that this type of data is ideal for computers to process, because it is store digitally in a well-formatted way, and computers can process data much faster than humans. The concept of machine learning is to some extent coming to this environment. Computers can recognize digital data to discover patterns and rules to means that is too hard for a human to do. The basic thought of machine learning is that a computer can automatically learn from examples(Pojon, 2017).

Applications of machine learning cover a wide range of areas. One of the application areas of machine-learning human resource management (HRM) is the core department within

organizations that manages human capital. The function of HRM is to motivate employees and enhance workforce effectiveness.

Mitchell et al. (2001)said that “Machine Learning is a method used in Artificial Intelligence. Artificial Intelligence is the science and engineering of buildingsmart machines; especially intelligent computer program”. “Predictive analytics is the practice of extract information from current data set to determine pattern and predict future consequence and trend”. Depending on the type of input data, machine learning algorithms can be classified into Supervised and unsupervised learning. In supervised learning, input data comes with a Known class that is label data, whereas, in unsupervised learning, input data does not have a known class structure that means unlabeled data.

Supervised machine learning is getting to know is the searching for algorithms that reason from externally supplied instances to make common hypotheses, which then build predictions about future instances. Toanswer these problems, organizations use machine learning techniques to predict employee turnover. Exact predictions allow organizations to take action for retention or succession planning of employees. It can facilitate companyto build better decisions,handle new employees and also consider the turnover of old employees(Salunkhe, 2018).

So, the researcher plan to use supervised machine learning tools by using label data to analyze employee turnoverin federal court.In federal court organization there were no empirical studies conducted on professional employees’ turnover which were supported by formal and published research. The study's aim was predict the major factors of employees' turnover by using a machine learning algorithm and its retention mechanism which will help in developing appropriate policy and training programs at the federal court organization and other institutions.

1. 2 Motivation

The reason to motivate the researcher to work on employee turnover, at this time, there is a high employee turnover rate in government organizations especially the federal court and it wants to predict the causes of employee turnover. It attempt to develop retention mechanisms for a long period in the future.

Additionally, most researchersinvestigate the turnover employee's reasons by interview and questionnaire. However, different people have different feelings and they may not articulate their real opinion very well. So it's hard to identify the exact reasons whyemployees leave their organizations. But, using employee real data by applying machine learning techniques can simply predict the causes of employee turnover in government organizations particularly, in the federal court.

By doing so, the researcher believes, the work gives for innovative new knowledge in individual minds about the thought of machine learning and it motivate other researchers to do the best on Human Resource Management and other institutions in the future.

1.3. Statement of the problem

In today's competitive business world, it is necessary to control employee turnover for any organization toaccomplishorganizational goals. Employee turnover has become a serious organization problem.Because, it leads tofinancial and moral impact on the organization's limited resources.

Employee turnover is both expensive and disruptive to the organizational role. Today, organizations are finding difficult to keep well-performing, knowledgeable, and soundly educated employees as a result of turnover. High employee turnover has become a crisis for government organizations since knowledgeable and well-qualified experts leave the position they held in the offices(*An_Assessment_of_the_Causes_of_Employee.docx*, n.d.).

Gao (2017)also examined that Employee turnover will cause many problems. According to the research, if a company loses a good qualityworker, it will cause 1.5 times money cost than recruiting and teaching a new employee. It can also lower customer satisfaction and retention.

Additionally, in our country context, there is no data-driven approach research to identify the causes of employee turnover.

According to Tomassen (2017)by means of data to make organizational decisionreferred to as data-driven decision making and can direct to better organizational performance. But there has not been a lot of study done data-driven decision making. Additionally, He found an

optimisticcausal connection between data-driven decision making and organizational output and production. This positive relationship explain by the fact that human hasdifficulty to handlewith complexity, the huge amount of information, high-pressure time and simultaneously choices. To reduce the complexity, humans fall back on old behaviors and assumptions which lead to bias and error. Computers have the upper hand over humans here, since Computers almost unlimited processing power and are, essence not prone bias and subjectivity. Machine learning the process of performs tasks by looking at the previous historical data and from that generalized conclusions to respond to new situations. He also suggests that Turn over Prediction and automation of the retention decision can be done by applying supervised training methods and various classification algorithms. In Human Resource Management there is a vast amount of data that helps to analyze the employee turnover factors by using information technology tools such as machine learning algorithms.

If the company can predict employee turnover in the future, they can also work on retention avoid the loss of a valuable employee. With the Machine Learning technique, more particularly Predictive Analytics, we can predict employee turnover. Thus, this study was an attempt to fill this gap by identifying those factors using machine learning algorithms that cause employee turnover in the context of Ethiopia, specifically in Federal Court organizations. The study was focus on answering the following research questions.

1.4. Research questions

1. Which machine learning algorithm is most appropriate for predicting employee turnover?
2. To what extent would the machine learning approach be able to correctly predict staff turnover?
3. What are the significant predictors for causes for employee turnover?

1.5. Objectives

1.5.1. General Objective

The general objective of this research is to design a model and predicting staff turnover using a machine learning approach in government organizations taking the EthiopianFederal court as a case.

1.5.2 Specific Objectives

- To identify the major predictors of employee turnover in the organizations
- To see the effects of the sample dataset on the performance of the predictive model
- To review different kinds of literature related to employee turnover and machine learning algorithms.
- To collect data, feature selection, Preprocess, and visualize the datasets
- Comparingand Evaluate machine learning algorithm models and identify the best outperform Classifiers.
- To Design and develop a ML prototype
- Conclude the result and propose the recommendations.

1.6 Significance of the study

The researchersbelieve that this study can support the human resource managers in understanding the level of employee turnover which could in danger the progress of their organization's effectiveness and efficiency.

The study is also an attempt to identify the cause of employee turnover in the sectors and the related problems associated with it so that it helps the organization to be aware of the state of turnover.

Besides, the datasets which are gathered and analyzed in the study helps the organizations to understand the root cause of employee turnover.

Finally, the study helps to innovate new knowledge and gives awareness about the concept of machine learning algorithms for information technology experts, Human Resource

Managementsectors and others. Moreover, the study can be used as a baseline study for further studies on the topic.

1.7 Scope of the study

The study was conducted at the Ethiopian Federal Court (EFC) which is located in the Addis Ababa region. The study concentrated on the employees at the organization involving 3600 employees, who stay and employees leavefrom the organization within 20 departments. Some of the departments are; finance, e-procurement, ICT, CORT office,registrar office, human resources, general service, and other remaining departments are samples for the study.Besides, separate the departments from the attributes/variables to be used to investigate. The attributes are Age, Sex, Marital status, Payment, Education level, Work experience, department, andNumber of years stayed. From these sections, the study wasidentifying the core driving factors affecting employee turnover like payment and other demographic factors that have been reviewed.

1.8 Limitation of the study

The first limitation was the reluctance of the organization to give out their data for research purposes. Thisproblem negatively influenced the development of the research instruments and methodologies for the study.

Another challenge also the limitation of time some of the problems that the researcher encountered, due to HR data is often noisy, inconsistent and contains missing information it takes much time to preprocess and cleaning the datasets. Additionally, lack of research studies and the availability of sufficient related literature upon the Federal court organization and our countrywereother constraints. The researcher has overcome this problem by reviewing the existing limited literature to the possible extent and by discussing it with experienced researchers including the advisors and the examiners.

1.9. Ethical consideration

For this study different concerned persons starting from manager and employees, those works in human resource departments and another directorate that found in the federal court were

communicated. At this time the privacy, legal and confidentially matters were respected by the researcher.

1.10 Organization of the thesis

The paper was organized into six chapters. Chapter one introduction part including background of the study, statement of the problem, research questions, general and specific objectives of the study, limitations of the study, and significance of the study. Chapter two literature reviews related to the study. Chapter three was discusses the methodology and design. The fourth chapter deals with model building, the fifth chapterevaluation and data visualization chapter six experimental results and discussion. Finally, the conclusions and recommendations are treated in chapter seven.

2. CHAPTER TWO

LITERATURE REVIEW

Introduction

This chapter deals with the definition of terms and related works relevant to the study problem. It also carries out the ideas and concepts that other scholars have put forward concerning employee turnover and other related keywords. The study helps the researcher and other readers to have a clear understanding of the subject matter more. Besides, it will also help the readers to familiarize themselves well with employee turnover as a whole. It includes the definition of employee turnover, types of turnover, causes of employee turnover, strategy to minimize turnover, effects of employee turnover, machine learning algorithms, and related works are reviewed.

2.1 Definition of terms

2.1.1 Definition of Employee turnover

According toIqbal(2010)Organizational turnover has sometimes been defined as "the ratio of the number of organizational individuals who have left all through the period being regarded divided by means of the common wide variety of humans in that business enterprise at some point of the period" Additionally, a definition of organization turnover "turnover includes 'leaving any job of any duration' and is typically notion of as being observed through continued normal employment". Similarly, managers analyze employee turnover as the entire manner related to filling a vacancy. Each time position is vacated, both voluntarily and involuntarily, a new employee ought to be hired and trained. This substitute cycle is recognized as a turnover. This term, employee turnover, is additionally regularly utilized in efforts to measuringrelations of personnel in a company as they leave, regardless of the reason.

Another researcher also examined that, employee turnover "refers definitely to the motion of personnel out of an organization". It is a terrible aspect, which may lead to the failure of worker retention strategies in organizations. "Leaving of a job seems to reflect tremendous

administrative center problems, as a substitute than possibilities for development into higher Jobs". Turnover of employees disrupts teams, raises costs, reduces productivity, and outcomes in lost knowledge. So, it is crucial for the administration to recognize the significance of employee job satisfaction(Rehman, 2012).

Finally, Employee turnover as defined by Agyeman & Ponniah (2014) refers to the movement of employees out of an organization. Turnover of employees has both short and long-run negative effects for the organization. This affects teamwork, raises costs, reduces productivity, and results in lost knowledge.

2.1.2 Types of Turnover

Iqbal (2010)has examined four types of employee turnover. It can be seen that turnover is either voluntary being initiated by the employee, or involuntary, being initiated by the organization.

- Involuntary Turnover:-Involuntary turnover is split into discharge and downsizing types.
  - ✓ Discharge Turnover: Discharge turnover is aimed at the individual employee, due to discipline and job performance problems.
  - ✓ Downsizing Turnover: It occurs as part of an organizational restructuring or cost-reduction program to improve organizational effectiveness and increase shareholder value.
- Voluntary Turnover:-Voluntary turnover, in turn, is broken down into avoidable and unavoidable turnover.
  - ✓ Avoidable turnover: Avoidable turnover is that which potentially could have been prevented by certain organizational actions, such as pay raise or new job assignment.
  - ✓ Unavoidable turnover: A turnover that happens in unavoidable circumstances is called as unavoidable turnover. For instance, an Employee's death or a spouse's relocation.

In addition to this, Turnover can take several forms. It can be voluntary or involuntary, functional or dysfunctional, avoidable or unavoidable. Voluntary turnover, an Employee leaves the organization of his own free choice with some of the possible reasons being: Low salary, job dissatisfaction or better job opportunities elsewhere. Whereas, involuntary turnover takes effect when the organization decides to remove an employee due to poor performance or economic

crisis. However, most studies have focused on voluntary rather than involuntary turnover. This suggests that voluntary turnover is a critical issue for both employees and organizations(Adjei, 2014).

From the above, the researcherexamined that, most of the authors in their literature study turnover categorize as either voluntary or involuntary. Voluntary turnover initiated by the employee interest, whereas, involuntary is initiated by the organization, it may be either organization wants to terminate their relationship with employees or natural phenomena, like death, sick...Most studies focus on voluntary employee turnover and it is a serious problem for government organizations and other institutions. In Federal court highly affiliate with voluntary turnover types of employee turnover. Therefore, the study was focus on voluntary employee turnover.

2.1.3 Causes of employee turnover

It is essential to identify the reason for the turnover for organizations hence, that they can handle events to prevent it. Unless it consequences further cost of enrollment, training andpotentially lower revenue ability. Generally, the causes of employee turnover classify into three categories.

A. Demographic factors

According to Glu(2014) in a meta-analysis and assessment of voluntary turnover find out about conducted in 1987, it is visibly noted that some of the demographic attributes of an employee like age, gender combined with some of the work surroundings attribute like salary, tenure, job delight play the strongest position in predicting turnover.

Some different research indicates that variables inclusive of age, economic activity, tenure, working time in position and training are the strongest predictors of turnover(Cotton & Tuttle, 1986).

Later on, another study conducted on demographic issue variables that have been discovered to have a stable relationship with retention and turnover intentions is age, gender, tenure, education and, income levels. These have influenced employee retention and turnover over time. Demographic elements have been chosen due to the fact they influence employee retention

strategies. The number of research in which demographic elements have been employed to inspect job satisfaction and job attitudes has shown that they are strong predictors of turnover intentions. Concerning years of service, mentioned that employees with higher tenure may additionally have familiarity with their work position and have reached a higher stage of career attainment than that personnel with lower tenure. Finally, the Author selects the following demographic traits chosen for the study primarily based on the literature review are; Gender, Age, Marital status, Qualification, Income, Years of service could be viewed that these have the causes of employee turnover(Agyeman & Ponniah, 2014).

B. Job-related factors

According to Ongori (2007) there are numerous reasons why people leave from one enterprise to another or why humans go away from the Organization. The ride of job-related stress (job stress), the range elements that lead to job-related stress (stressors), lack commitment in the organization; and job dissatisfaction make employees quit. If the roles of employees are not spelled out by using management/ supervisors, this would accelerate the degree of employees quitting their jobs due to lack of position clarity.

The relation between job satisfaction and employee turnover is reciprocal to each different and this relationship is high when the unemployment fee is low in society and similarly low when the unemployment fee is high. "Even although humans are now not satisfied with their jobs, they will be less possible to stop if there are few options (Adjei, 2014).

Organizational factors

Organizational instability has been proven to have a high degree of excessive turnover. Indications are that personnel are greater probably to continue to be when there is a predictable work environment and vice versa. In organizations where there was once an excessive level of inefficiency, there was also an excessive stage of staff turnover, Therefore, in conditions, the place organizations are not secure employees tend to cease and seem to be for secure groups because with stable organizations they would be able to predict their career advancement. Additionally, excessive labor turnover may suggest negative personnel policies, terrible

recruitment policies, terrible supervisory practices, negative complaint procedures, or lack of motivation. All these factors make a contribution to excessive worker turnover in the feeling that there is no perfect management practices and policies on personnel things hence personnel are now not recruited scientifically, promotions of personnel are no longer based on spelled out policies, no criticism approaches in place and therefore personnel decides to quit(Ongori, 2007).

Another researcher Abdali(2011)thinking the coordination between managers or supervisors with their sub-ordinates might also create an effect on employee turnover. It relies upon the employee's satisfaction with their supervisors and also the communication abilities of supervisors to manage their subordinates. Also, the organizational surroundings may additionally affect turnover. Employees favor staying remains with the organization just due to the fact of a smooth and healthy climate. The match between proportions of environment and employee values may figure out trustworthiness with the organization.

The researcher was understands that the causes of employee turn over the first one related to demographic factors such as -Age, Tenure, Work Experience, Gender, Marital status, income, and educations are the main significant factors. On the other hand from job-related factors: - job satisfaction, work satisfaction, motivation and job role, job stresses are causes of employee turnover. Finally, organizational factors examined that work environment, leadership style, promotion, organization policies, Compensation, and other job-related factors are the sources of employee turnover in government organizations and other institutions.

2.1.4 Effects of employee turnover

If employee turnover is now not managed appropriately, it would affect the organization adversely in phrases of personnel prices and the long run; it would affect its liquidity position. However, voluntary turnover incurs a sizeable cost, each in terms of direct fees (replacement, recruitment, selection, temporary staff, management time) and additionally possibly more extensively in phrases of indirect expenses (morale, pressure on last staff, prices of learning, product/service quality, organizational memory) and the loss of social capital (Dess & Shaw, 2001)

Stovel & Bontis(2002)stated that if employee turnover is now not managed excellent it would affect the organization adversely in phrases of personnel prices and the long run, it would affect its liquidity position. However, voluntary turnover incurs a sizeable cost, each in terms of direct fees (replacement, recruitment, selection, temporary staff, management time) and additionally possibly more extensively in phrases of indirect expenses (morale, pressure on last staff, prices of learning, product/service quality, organizational memory) and the loss of social capital.

Research conducted in our country by Blen (2018) on "Employee Turnover and Organization Performance: The Case of Shines ETB Garment PLC"the result indicated that turnover has a consequence in high cost of recruitment, training, productivity, and quality of production(Asegid, 2018).

Finally, Iqbal(2010)stated that the consequences of excessive turnover are each financial and non-financial. High turnover can be a serious hurdle to productivity, quality, and profitability at corporations of all sizes. For the smallest of companies, and excessive turnover fee can suggest that honestly having a sufficient group of workers to fulfill everyday functions is a challenge, even beyond the issue of how properly they work is achieved when personnel is available. On the other hand, Turnover can, however, be recommended for organizations. It can allow the organization to hire new employees with extra present-day education who are now not locked into current ways of doing things. Also, other high-quality penalties are higher quality, less highly-priced replacements, displacement of the negative performer, innovation, flexibility, adaptability, provide possibilities to promote talented, excessive performers. Despite their many potential benefits, voluntary (being initiated by using the employee) turnover, are typically high priced propositions. Therefore, both voluntary and involuntary turnover can be managed strategically to enable the organization to maximize the prices incurred with the process. Retention strategies ought to involve the evaluation of both retention prices and benefits. Retention strategies must focus now not only on how many employees are retained however precisely who is retained. An ineffective employee retention approach can disrupt the entire organizational productiveness and employee morale.

2.1.5 Strategies to minimize employee turnover

Turnover is all about leaving, whereas retention is all about staying. Low employee turnover fee means high worker retention. Employee retention connotes the means, design or set of decision-

making conduct put in place via the organization to hold their competent body of workers overall performance(Gberevbie, 2008).

Ongori(2007)Pointed that Strategies on how to decrease worker turnover, confronted with problems of worker turnover, management has several policy options, changing (improving existing) policies towards recruitment, selection, induction, training, job design, and wage payment. Policy choice, however, has to be appropriate to the unique analysis of the problem. Employee turnover attributable to bad resolution procedures, for example, is unlikely to enhance where the policy change to focus exclusively on the induction process. Equally, employee turnover attributable to wage fees that produce salaries that are no longer competitive with different firms in the local labor market is unlikely to reduce have been the policy adjustment only to decorate the organization's provision of on-the-job education opportunities. Given that there is an increase in direct and indirect expenses of labor turnover, therefore, management is regularly exhorted to discover the reasons why human beings leave the organizations so that suitable motion is taken by way of the management. Extensive research has proven that the following categories of human capital management factors provide a core set of measures that senior administration can use to expand the effectiveness of their funding in human beings and enhance universal company performance of the business: Employee engagement, the organization's capability to engage, retain, and optimize the price of its personnel hinges on how well jobs are designed, how employees' time is used, and the commitment and guide that is proven to personnel by the administration would motivate Employees to remain for longer duration in their organizations.

There are several methods via which an organization or commercial enterprise may enhance the effectiveness of their employees. This can be through periodic education programs, regular motivation in the workplace, worker empowerment, promotion, and bonus packages among others(Samuel & Chipunza, 2009).

Mitchell et al..,(2001) examined that Employee retention is a strategic and coherent manner that starts with an examination of the reasons that a worker joins an organization. Retention is a complex construct, no longer just one variable and it is affected through countless factors, amongst which are: job satisfaction, work overload, etc. The reasons for humans staying in an organization are multidimensional and lie in the fruits of a host of social, psychological and

organizational factors. In growing countries, due to common influence, compensation is necessary however it now not the solely critical retention factor. Reasons for staying are seen as the easy obverse of reasons for leaving. But that is simply an oversimplification of facts. Factors making employees remain is very frequently unique from what makes them leave. Apart from that, whilst making a retention strategy, one has to keep in mind "one dimension does not match all". Employees from different industries or departments or levels have different needs. One cannot expect that one single approach can make them all stay. Thus to create an environment the place all want to stay agency has to keep their character wants in their mind. Some of the normal tendencies in Employee Retention Strategies are:

Job Satisfaction: It encompasses the feelings, beliefs, and thoughts about the job. An employee who is comfortable with her/his job will most possibly stay. Hence, management wants to identify and apply excellent variables that will create job satisfaction.

Training: Training is referred to as a planned effort to facilitate the study of job-related knowledge, skills, and conduct through an employee. An organization that makes an effort to educate employees, which is vital for their personal and professional growth, is appreciated using employees. Hence, they tend to stay in the organization.

Reward: When employees receive an honest reward for their contribution in the direction of the organization, they tend to stay.

Supervisory Support: Supportive supervisor helps a worker not only to gain the person's goal however smooth relationship between them enhances the probabilities of personnel stay in the organization. However, a competitive market and a certified body of workers have been disturbing much greater than simple requirements. Thus to enhance advantageous retention approach businesses some add on have been made such as participatory selection making, work flexibility to provide personnel work-life balance, profession boom and the like.

2.1.6 Machine Learning Techniques on employee turn over

According to Rabbi, n.d (2019) the advent of computer and web technological know-how has led to several breakthroughs in the fields of science, medicine, law, and business, to identify just a few. Moreover, the rapid acceleration of break troughs in the area of web and computer technological know-how has also shifted how the future may seem to be in a very short period. One of the most considerable potential areas of future innovation is in artificial intelligence. In this area, the present-day trending subject is that of 'machine learning', whereby computer systems in actuality begin to analyze similarly to humans, with the aid of the use of experience as a teacher.

Mandal & Sairam (2012) havementioned Machine learning is a department of artificial intelligence that targets at solving real-life engineering problems. It offers the chance to analyze except being explicitly programmed and it is based on the idea of mastering from data. It is so a great deal all overthe place used a dozen instances a day that we may additionally not even know it. The advantage of computer mastering (ML) strategies is that it makes use of mathematical models, heuristic learning, information acquisitions, and choice trees for choice making. Thus, it gives controllability, observability, and stability.

According to Alpaydin(2009)Machine learning is programming computers to optimize an overall performance criterion the use of instance facts or experience. We have a model described up to some parameters, and gaining knowledge is the execution of a computer program to optimize the parameters of the model the usage of the training records or experience. The model can also be predictive to make predictions in the future, or descriptive to attain know-how from records or both. Machine getting to know uses the idea of information in building mathematical fashions due to the fact the core mission is making inferences from a sample. The role of computer science is two-fold: First, in training, we want environment-friendly algorithms to solve the problem, as well as to keep and procedure the big amount of records we normally have. Second, once a model is learned, its illustration and algorithmic solution for inference want to be environment-friendly as well..

2.1.6.1 Types of machine learning algorithms

Machine Learning shortly describes as ML is a type of Artificial Intelligence (AI) that compose reachable computer systems with the effect to be educated besides being veraciously programmed. ML gaining knowledge of hobby in the extensions of computer applications which is capable sufficient to adjust when unprotected to new-fangled data. ML algorithms are commonly classified into three divisions specifically supervised learning, unsupervised and reinforcement learning(Praveena & Jaiganesh, 2017).

Other researchers Mohammed et al.., (2016)also stated them the term machine learning means to enable machines to learn without programming them explicitly. The objectives of machine learning are to enables machines to make predictions, perform clustering, extract association rules, or make decisions from a given dataset. There are four types of machine learning methods. They are supervised, unsupervised, semi-supervised and reinforcement learning methods.

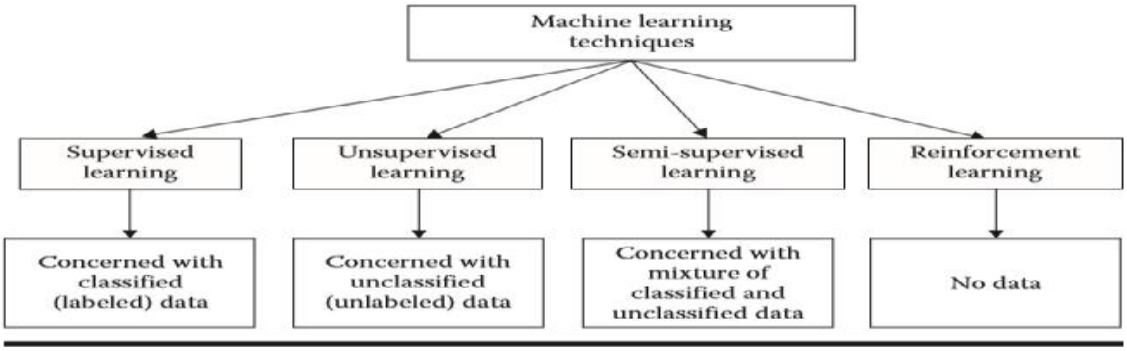


Figure 1: Different machine learning techniques **Source**:- **(Mohammed et al., 2016)**

- Supervised Learning: In supervised learning, the target is to infer a function or mapping from training data that is labeled. That means it works on class or target data in agiven dataset.New data is classified based on the training set.
- Unsupervised Learning: In unsupervised learning, we lack supervisors or training data. In other words, all that we have is unlabeled data.The class labels of training data is unknown.it classify the given data based on a set of measurements, observations with the aim of establishing the existence of classes or clusters in the data

- Semi-Supervised Learning: In semi-supervised learning, the given data are a mixture of classified and unclassified data. This combination of labeled and unlabeled data is used to generate an appropriate model for the classification of data. In this case apply small number of labeled data to label large amount of unlabeled data.
- Reinforcement Learning: The reinforcement learning method aims at using observations gathered from the interaction with the environment to take actions that would maximize the reward or minimize the risk.

Thus, the study focuses on supervised machine learning algorithms, because machine learning algorithms depend on the problem. In this study the human resource employee data mostly categorical, numerical and label datatype is related to supervise machine learning.

Supervised machine learning algorithms - is an algorithm which wishes external assistance. The input dataset is divided into train and test datasets. The training dataset has an output variable that desires to be estimated or classified. All algorithms learn some type of patterns from the training dataset and follow them to the check dataset for prediction or classification (Kotsiantis et al., 2007).

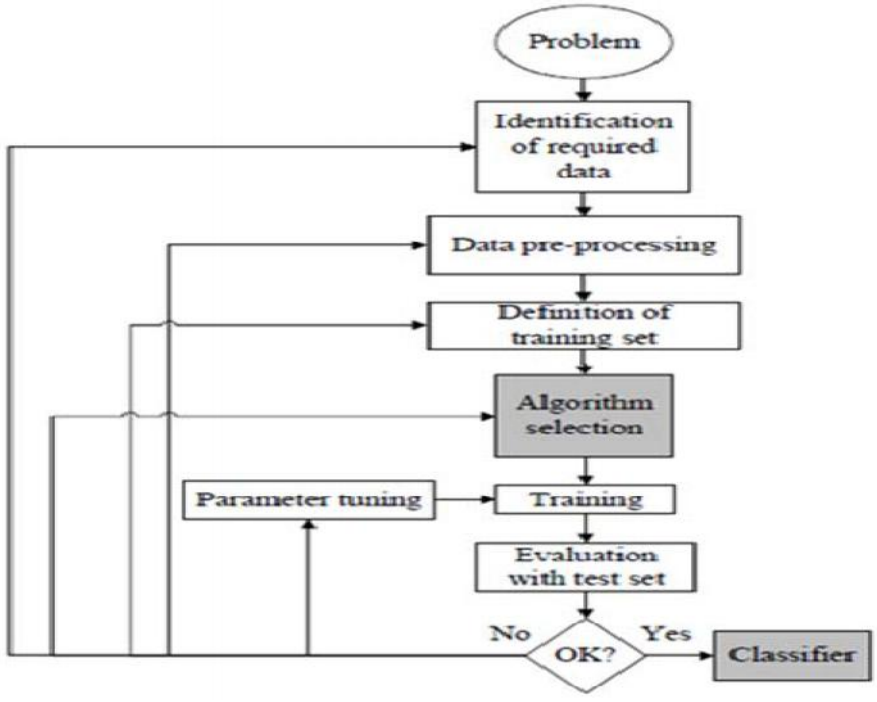


Figure 2**The workflow of a supervised machine learning algorithm**

(Kotsiantis et al., 2007)

Supervised machine learning - is the mission of conceiving a meaning from labeled training data which has a set of training examples. As far as supervised learning is concerned, every example is a mainstay containing an input object (which is usually a vector quantity) and an enforced output value (may also be referred to as a supervisory signal).

Steps performed in the Supervised Machine Learning Algorithms:-

Step – 1: Establish the type of training examples. The user needs to courage the type(s) of data that will be used as a training set.

Step – 2: Converge a training set. The training set the ambition to be a delegate of the real-world use of the function. As an effort, a set of input objects is collected that remains and analogous outputs are also collected.

Step – 3: Resolve the input feature illustration of the learned function / learned attribute. The accurateness of the learned function is securely based on the input object is representation.

Step – 4: Resolve the formation of the learned function and comparable machine learning algorithm.

Step – 5: Assimilate the design and execute the learning algorithm on the collected training set.

Step – 6: Evaluate the accurateness/correctness of the learned function. Then, parameter adapts and learning may be performed on the resulting function and needs to be measured on a test data set that is break up from the training set(Praveena & Jaiganesh, 2017).

2.1.6.2 Supervised machine learning algorithms

For the study, various supervised machine learning algorithms are demonstrated and assessed in their ability to predict employee turnover. This section provides a general overview of the theory behind these algorithms. Some supervised machine learning algorithms are the following.

Logistic regression: This is a classification characteristic that uses a category for building and uses a single multinomial logistic regression model with a single estimator. Logistic regression usually states the place the boundary between the instructions exists, additionally states the

category probabilities depend on distance from the boundary, in a precise approach. This moves in the direction of the extremes (0 and 1) greater swiftly when the statistics set is larger. These statements about possibilities that make logistic regression greater than just a classifier. It makes stronger, more unique predictions, and can be fit differently; however, those strong predictions may want to be wrong. Logistic regression is a method to predict, like Ordinary Least Squares (OLS) regression. However, with logistic regression, prediction outcomes in a dichotomous result(Newsom, 2015).

Decision Trees: Decision Trees (DT) are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values (Kotsiantis et al., 2007).Decision tree learning, used in data mining and machine learning, uses a decision tree as a predictive model that maps observations about an item to conclusions about the item's target value. More descriptive names for such tree models are classification trees or regression trees. Decision tree classifiers usually employ post-pruning techniques that evaluate the performance of decision trees, as they are pruned by using a validation set(Ye et al., 2009).

Naive Bayesian (NB) Networks: These are very simple Bayesian networks which are composed of directed acyclic graphs with only one parent (representing the unobserved node) and several children (corresponding to observed nodes) with a strong assumption of independence among child nodes in the context of their parent. Thus, the independence model (Naive Bayes) is based on estimates. Bayes classifiers are usually less accurate than other more sophisticated learning algorithms (such as ANNs)(Domingos & Pazzani, 1997) However, performed a large-scale comparison of the naive Bayes classifier with state-of-the-art algorithms for decision tree induction, instance-based learning, and rule induction on standard benchmark datasets, and found it to be sometimes superior to the other learning schemes, even on datasets with substantial feature dependencies. Bayes classifier has an attribute-independence problem which was addressed with Averaged One-Dependence Estimators(Hormozi et al., 2012)Other learning schemes, even on datasets with substantial feature dependencies. Bayes classifier has an attribute-independence problem which was addressed with Averaged One-Dependence Estimators.

Neural Networks: Neural Networks (NN) can function some regression and/or classification duties at once, even though oftentimes every network performs only one. In the large majority of cases, therefore, the network will have a single output variable, although, in the case of many-state classification problems, this might also correspond to various output devices (the post-processing stage takes care of the mapping from output devices to output variables). Artificial Neural Network (ANN) depends upon three fundamental aspects, input and activation functions of the unit, community architecture and the weight of each enter connection. Given that the first two aspects are fixed, the conduct of the ANN is defined through the present-day values of the weights. The weights of the internet to be educated are firstly set to random values, and then situations of the training set are repeatedly exposed to the net. The values for the entry of an occasion are positioned on the enter devices and the output of the internet is in contrast with the desired output for this instance. Then, all the weights in the net are adjusted slightly in the path that would bring the output values of the net closer to the values for the preferred output. There are several algorithms with which a community can be skilled(Osisanwo et al., 2017).

Extreme Gradient Boosting (XGB):- Extreme Gradient Boosting is a tree-based method that was introduced in 2014 by Chen(Chen & Jeong, 2007)It is also commonly referred to as XGBoost. It is a scalable and accurate implementation of gradient boosted trees, explicitly designed for optimizing the computational speed and model performance. (Ajit, 2016)Compared to gradient boosting, XGBoost utilizes a regularization term to reduce the overfitting effect, yielding a better prediction and much faster computational run times.

Gradient Boosting Trees (GBT): -a Gradient boosting tree is an ensemble machine learning technique proposed in 2001 by using Friedman. It is used to for regression and classification purposes. The distinction between RF and GBT is the gradient boosted tree models research sequentially. In GBT, a series of trees are constructed and every tree attempts to right the errors of the preceding tree in the series. Trees are introduced sequentially until no further enhancement can be achieved. Making predictions in GBT is quick and memory-efficient; boosting may want to be seen as a form of regularization to reduce over fitting(Friedman, 2001).

Random Forests (RF):-Random forests take an ensemble strategy that offers an improvement over the basic decision tree structure by combining a group of weak learners to form a better learner. Ensemble methods make use of a divide-and-conquer method to enhance algorithm

performance. In random forests, several Decision trees, i.e., weak learners, are constructed on bootstrapped education sets, and a random sample of m predictors are chosen as split candidates from the full set P predictors for Each selection tree. As m _ P, the majority of the predictors are not considered. In this case, all of the individual trees are not likely to be dominated by a few influential predictors. By taking the common of these uncorrelated trees, a reduction invariance can be attained (Friedman, 2001), making the final result less variable and more reliable.

2.2 Related Works

Related works help the researcher familiarize and understanding different related works to your work. In this section review related research and practical findings related to the subject of the study.

Zhao et al., (2018)conducted on the topic "Employee Turnover Prediction with Machine Learning: A Reliable Approach" investigates that the performance of ten supervised machine learning Methods(1) a decision tree method; (2) a random forest method; (3) a gradient boosting trees method; (4) an extreme gradient boosting method; (5) a logistic regression method; (6) support vector machines; (7) neural networks; (8) linear discriminate analysis; (9) a Naïve Bayes method; and (10) a K-nearest neighbor method was evaluated on various HR datasets. In addition to some data preprocessing tasks were conducted, including Missing Value Imputation based on their data type(for numerical data types, the missing entries are replaced by the median and for categorical data, the missing entries were replaced by the mode), Data Type Conversion, Feature Selection and Feature Scaling were introduced and used in this study. The data set used in this research is from two different sources. The first dataset originates from a regional bank in the United States of America, collected from 2013 to 2016, of 14,322 Employee entries and 24 features which of 28% of the bank's employees had left. The second dataset is a simulated dataset created by IBM Watson Analytics contains 1,470 employee entries and 38 features, in which 237 employees (16%) left.

Following these basic data cleaning procedures, the final datasets consisted of 9,089 employees with 19 features for Bank data and 1,470 employees with 31 features for IBM data. Both datasets were contained common HR features like age, compensation, gender, and education. For evaluate the study, the researcher use accuracy, precision, recall-measure, and Roc curve were

applied. After evaluation, the result indicated that small HR datasets may contain high variance and randomness. For medium and large HR datasets, the data variance decreases and a more reliable model may be built. Best practice would be using tree-based ensemble methods such as extreme gradient boosting and gradient boosted trees. The numerical experiment results indicate that Tree-based classifiers (XGB, GBT, RF, and DT) and logistic regression were worked well in general, tree based classifiers requires minimal data preprocessing, has decent predictive power, and ranks the feature importance automatically and reliably. However, due to the complexity of employee turnover prediction, one should try to find the classifier that best fits the underlying data before taking this approach.

The study conducted on **"Improving Employee Retention by Predicting Employee Attrition using Machine Learning Techniques"**from Dissertation study the researcher was applied confusion matrix for accuracy measurement. In this study, different data preprocessing techniques were conducted, including - Removing columns with the same values, remove all the rows containing NA, Move Attrition to last column and, removing all unwanted columns. After preprocessing and accuracy evaluation, it is found that the factors responsible for attrition and retention are -: Percent Salary like Monthly Income, Years Since Last Promotion, Distance From Home, Job Role, Performance Rating, Job Level, Environment Satisfaction, Years In Current Role, Relationship Satisfaction, Years With Current Manager, Job Satisfaction, Work-Life Balance, Number of Companies Worked, Years At Company, Over-Time, Total Working Years, Marital Status, Age and Gender. the dataset source 1470 collected from the IBM and selected six data mining algorithms (Decision Tree, SVM, Random Forest, XGBoost Tree, KNN and Logistic Regression), and from this algorithms, Logistic Regression which provided the most accurate output, was selected for developing an application(Salunkhe, 2018).

Research conducted in India on **"Prediction of Employee Turnover in Organizations using Machine Learning Algorithms":**by (Ajit, 2016)suggests that Employee turnover has been identified as a key issue for organizations because of its adverse impact on workplace productivity and long term growth strategies. To solve this problem, organizations use machine learning techniques to predict employee turnover. Accurate predictions enable organizations to take action for retention or succession planning of employees. The dataset had 73,115 data points with each labeled active or terminated. The data was gathered from 2 sources: the HRIS database

of the organization, as well as the BLS (Bureau of Labor Statistics). The HRIS database of the organization provided some key features like demographics features e.g. age etc.; compensation-related features like pay etc.; team related features like peer attrition etc. The BLS data provided key features like the unemployment rate, household income, etc. Overall there were 33 features of which 27 were numeric while 6 were categorical. The population chosen is distributed across various locations in the US. The data was pulled at a Quarterly level in 18 months. There are 2 Class labels - Active and Terminated labeled 0 and 1 respectively. Each employee would have a record for every quarter of being active in the organization, until the quarter of turnover (if it occurs).In this study like the above study some data preprocessing and preparation techniques were introduced. For model validation technique the researcher has split the dataset 80:20 into training and testing. Additionally, The Area under the receiver operating characteristic curve (ROC-AUC) is the measure chosen here to compare classification accuracies. After comparing the classifier accuracy from XGBoost, Logistic Regression, Naïve Bayesian, Random Forest, KNN and Linear Discriminates Analysis (LDA) It is seen that the two tree-based classifiers in Random Forest and XGBoost performs better than the other classifiers during training and that XGBoost is significantly better than Random Forest during testing. The XGBoost classifier outperforms the other classifiers in terms of accuracy and memory utilization.

Hossain was conducted on the Title **"EMPLOYEE ATTRITION: A STUDY ON FEATURES OF SIGNIFICANCE & PREDICTION MODEL BUILDING"**: he examined that Employee attrition is a major problem for any organization interms of cost to replace awell-trained employee with good qualityperformance. For the study apply some data preprocessing tasks and feature selection to select the significant employeeattributes. Forthe model building the researcher was choice gradient boosting classifier. In addition, evaluate the mode by using recall, precision and accuracy, after the evaluation gradient boosting tree was best to outperform 96%.Finally,the finding shows that, employees less satisfied are tends to leave the organization and Employees who have been with the company longer and have involved with work are less likely to leave in the organization(Hossain, 2019).

A research done by Sabbeh (2018) in Egypt, **analyzed the performance of ten machine learningTechniques** such as ensemble methods, Regression methods, SVM, Decision Trees, Instance-Based learning and Discriminates analysis on 3,333 with 17 features telecommunication

dataset to predict customer retention. The first step before applying the selected analytical models on the dataset, data was preprocessed follows three steps data transformation, data cleaning, and feature selection to be more suitable for analysis. Additionally, 10-fold cross-validations were used for model training and testing. He observed that the ensemble learning techniques both Random forest and Ad boost models of machine learning outperformed the other methods with an overall accuracy of 96% and logistic regression had the least performance with an accuracy of 86.7%. Along with the classification algorithms, clustering customers according to their behavioral patterns were done in some research for better predictive accuracy.

Yousaf & Bhulai (2016)"**Analyzing which factors are of influence in predicting the employee turnover"** The original data was provided in the single CSV file and it consists of 39 attributes and 10,616 instances. 10 attributes have 100% missing values so they are removed before exploring anything in data. After preprocessing the clean dataset was 9086. To evaluate the prediction of the selected models, the researcher was applied k-fold cross-validation were applied to reduce the bias of sampling data and ensuring model error randomness. Additionally, to estimate the model's performance, different evaluating measures can be considered (Overall accuracy, Kappa, Sensitivity, Specificity). To overcome the class imbalanced problem, the 'SMOTE' techniques were used. The results of the predicting models indicate that from the selected model's logistic regression, artificial neural networks, and random forest, the random forest works best on these datasets. The factors which have the highest influence on employee turnover are Age, Location, Currency, Base salary, Business level, In-position, and In-service. The factors which have less impact in predicting employee turnover are Hire type, Gender, Contract type, and Functional area.

Another research conducted in 2016 in South Africa bySchlechter et al., (2016)the title on" **Predicting voluntary turnover in employees using demographic characteristics: A South African case study".**In the study the source of data was extracted from an existing human resources database, 2592 employees in a general insurance company based in South Africa and Namibia formed the basis for the analysis. Company records included an observation of the database of the years 2008–2011; the organization provided 22 data fields for each employee record, of which only 14 variables were selected for analysis. The researchers were choosing Logistic regression analysis was employed to predict employee turnover using various

demographic characteristics available within the database. The model tested 14 demographic variables and among those, the five variables were found to have statistically significant predictive value: age, years of service, cost center, performance score and the interaction between several dependants and years of service. It is proposed that these five demographic variables be used as a model to help identify employees at risk of a turnover or termed as indicators of voluntary turnover variables.

Another research also conducted on the title **"A Comparative Study of Employee Churn Prediction Model"**in Indonesia: the researcher used three machine-learning algorithms namely:- decision tree, random forest, and naïve Bayes by using 16,649 instances of employee personal data from 2 years period (2015-2017) in one Indonesia's telecommunication company. Similar to the above researches data preprocessing tasks were introduce and the researcher used a confusion matrix for users to compare the performance of the selected algorithms additionally, for model construction he uses classification technique classifies the model into 70:30, 70 for training and 30 for testing. After evaluation, the findings from the study suggest that the best classification model is random forest due to its immense accuracy of 97.5%. The second-best method is naïve Bayes with 96.6%, and the lowest accuracy of the classification model is a decision tree with 88.7%. The study concludes that the most reliable and accurate classification model to predict employee churn is a random forest(Alamsyah & Salma, 2018).

Jose(2019) on the Title **"Predicting Customer Retention of an App-Based Business Using Supervised Machine Learning":** The data was 80751 records with 10 features. It contains the customer activity data from July 1, 2018, to September 30, 2018. The research employed a quantitative and inductive approach to build a supervised machine learning model that addresses the class imbalance problem and efficiently predict customer retention. The Cross-Industry Standard Process for Data Mining (CRISP-DMDM) methodology is followed in the research lifecycle and Python programming with Jupyter Notebooks interface was used to implement the experiments of the research. Additionally, retention Precision is used as the evaluation metrics for the research. The research evaluated the performance of different sampling methods to tackle the class imbalance (Random Under – Sampling, Random Over – Sampling, SMOTE) on different single and ensemble machine learning models, Random Forest, Logistic Regression, SVM and XGBoost algorithms are used in the experiment to predict the retention of the

customers. The results show that Random Under-Sampling used along with the XGBoost classifier yields the best precision in identifying the retention class.

Research done byAlao&Adeyemo(2013)on the Title **"ANALYZING EMPLOYEE ATTRITION USING DECISION TREE ALGORITHMS"** indicated that the researcher used Three hundred and nine (309) complete records of employees of one of the Higher Institutions in Nigeria who worked in and left the institution between 1978 and 2006 were used for the study. The demographic and job-related records of the employee were the main data that were used to classify the employee into some predefined attrition classes. Waikato Environment for Knowledge Analysis (WEKA) and See5 for Windows was used to generate decision tree models and rule-sets. The data was prepared, pre-processed and cleansed using the preprocess tab of the Explorer window of the WEKA GUI Chooser. Additionally, Classification techniques were used to develop the prediction models used in the study.

The results indicate that employee demographics and job-related attributes as important factors that affect employee turnover within an organization. The most important attributes were the Salary and Length of service of the employee. Employees who have worked longer in the organization with no reasonable increase in income are likely to be more discouraged which influences their attrition. Also, low ranking employees with very few years of service put in are likely to turnover when they realize the income may not improve given their low ranks, they, therefore, leave in search of better-paying jobs.

In our country Ethiopia, there is no machine learning study on employee turnover. But in social science from management departments research done by Mamuye(2018)on **Statistical Assessment of Employee's Turnover and Its Causes; In the Case of Moret and Jiru Wereda, North Shoa, Amhara,** Ethiopia: the result indicated that Age, income, experience, educational level, and satisfaction were predictor factors for employee turnover and other factors such as gender, working hour and stress are found to be insignificant predictors of the turnover intention of the employees.

Additionally another study on the title "**FACTORS INFLUENCING EMPLOYEE INTENTION TO TURNOVER AT COMMERCIAL BANK OF ETHIOPIA**" by Birknesh Gemechu in 2017: the researcher was using both primary and secondary data sources using semi-

structured questionnaire and interview. The result indicated that salary and work environment become significant factors in the institution(Gemechu, 2017).

Finally, research conducted by Odiro(2017)on **"Assessment of Employee Turnover and Its Impact on Three Selected Government TVET Colleges in Addis Ababa**: in the methodology part use interview and questionnaire methods for data collection .the result shows that lack of promotion, dissatisfaction with pay, and lack of training and development, dissatisfaction with management style and nature of work is accepted as a reason to leave the organization.

From the above study, the researchers understand that those related researches it tells before model building and evaluation collecting data and preprocessing is the main task of activities in the study. After preprocessing select the algorithms that fit the study depends upon the problem is the second important task. Finally, building the model and evaluate the model; it helps to identify the best outperform classifier and the predictor variable that cause employee turnover and understand how much the algorithms predict the result; it enhances to decide for the future.

From the literature examined that mostly demographic factors such as; age, years of service, income, gender, marital status, education, job role, and experiences are the most predictor variables for the cause of employee turnover. Tree-based classifiers (XGB, GBT, RF, and DT) worked well in general, and were found to be best outperformed compare to other supervised machine learning algorithms. This outcome indicated that those algorithms best fit the employee's numerical and categorical data. Moreover, those algorithms were found to be they could handle the HR datasets which contained noise, missing values and imbalanced. Lastly, in supervised machine learning the data is a label and uses classification methods thus, the classifiers are using classification methods. Therefore, in human resource directorate the employee profile data mostly, contain numeric and categorical and it should fit the best classifiers to achieve the best accuracy. Another reason to select those algorithms in machine learning according toZhao et al., (2018), due to the complexity of employee turnover prediction, one should try to find the classier that best fits the underlying data before taking the approach. This indicates the selection of classifier, it depends on our problems. Also, the researchers review different works of literature that examined those selected classifiers outperformed best performances on employee turnover predictions. For the study, the researcher used three different prediction models based on the literature output. The selected models were logistic

regression, Gradient Boosting, and random forest. These models were trained and tested using the same dataset. From the three take the best one that gives good accuracy performance.

2.3 Research gaps

the researcher examines from the previous research mostly, the authors apply both real data set and personal opinion; from IBM Watson and only use personal opinion data sources through questionnaires, interviews, open and closed-ended questions, and other data collection methods. The data sources were not real data sets, because personal opinion may differ from person to person and it is difficult to gain accurate information .additionally, machine learning techniques do not work on opinion mind data rather work on real data sets. Thus, based on the above research gap the study works only employee real data sets.

Table 1: Summary Related Works

No	*Author and title*	*Statement of the problem*	*purpose of the study*	*Finding*
1	Employee Turnover Prediction with Machine Learning: A Reliable Approach: By (Zhao1,et.al 2018)	➢ Previous research on machine learning methods is often problem-specific and has narrow evaluation metrics across various models.	➢ to provide a comprehensive description and assessment of supervised machine learning	➢ From ten supervised machine learning Tree-based classifiers (XGB, GBT, RF, and DT) were best outperformed.
2	Improving Employee Retention by Predicting Employee Attrition using Machine Learning Techniques: (Salunkhe, 2018)	✓ Employee Turnover causes a huge loss to the company and skilled employees.	✓ The aim was to find the attributes responsible for employee attrition.	✓ Factors responsible for attrition Years At Company, Marital Status, Age and Gender. ✓ Logistic Regression which provided the most accurate output.
3	Prediction of Employee Turnover in Organizations using Machine Learning Algorithms":by (Ajit, 2016)	❖ Noise data from HRIS and affects the ongoing work on existing employees.	❖ Topresented the importance of predicting employee turnover.	❖ From XGBoost, Logistic Regression, Naïve Bayesian, Random Forest, andKNN:-Random Forest and XGBoost performs the best.

4	Employee Attrition: a Study onFeatures of Significance& Prediction Model building (Hossian,2019)	❖ The cost of replacing a well-trained employee can be difficult.	❖ To discover which attributes contribute from the given datasets the reason for employee turnover.	❖ Employees less satisfied are tends to leave the organization. ❖ Gradient boosting tree were best outperform of 96%.
5	"Analyzing which factors are of influence in predicting the employee turnover":(Yousaf & Bhulai, 2016)	❖ Employee turnover is a serious issue for organizations.	❖ The purpose of the study was to analyze the factors which influence the employee turnover in an organization.	❖ From the selected models random forest works the best on these datasets. ❖ The factors on employee turnover are Age, Location, salary, Business level, and position.
6	Predicting voluntary turnover in employees using demographic characteristics: A South African case study".(Schlechter et al., 2016)	❖ Turnover causes reduction in customer satisfaction, loss of organizational knowledge, and reduced profit.	❖ The purposes of this research construct a model that could be used to predict employee turnover by using demographic variables.	❖ Age, no of year service, salary, and performance score demographic variables were found to have statistically significant predictive values.

CHAPTER THREE

DESIGN AND METHODOLOGY

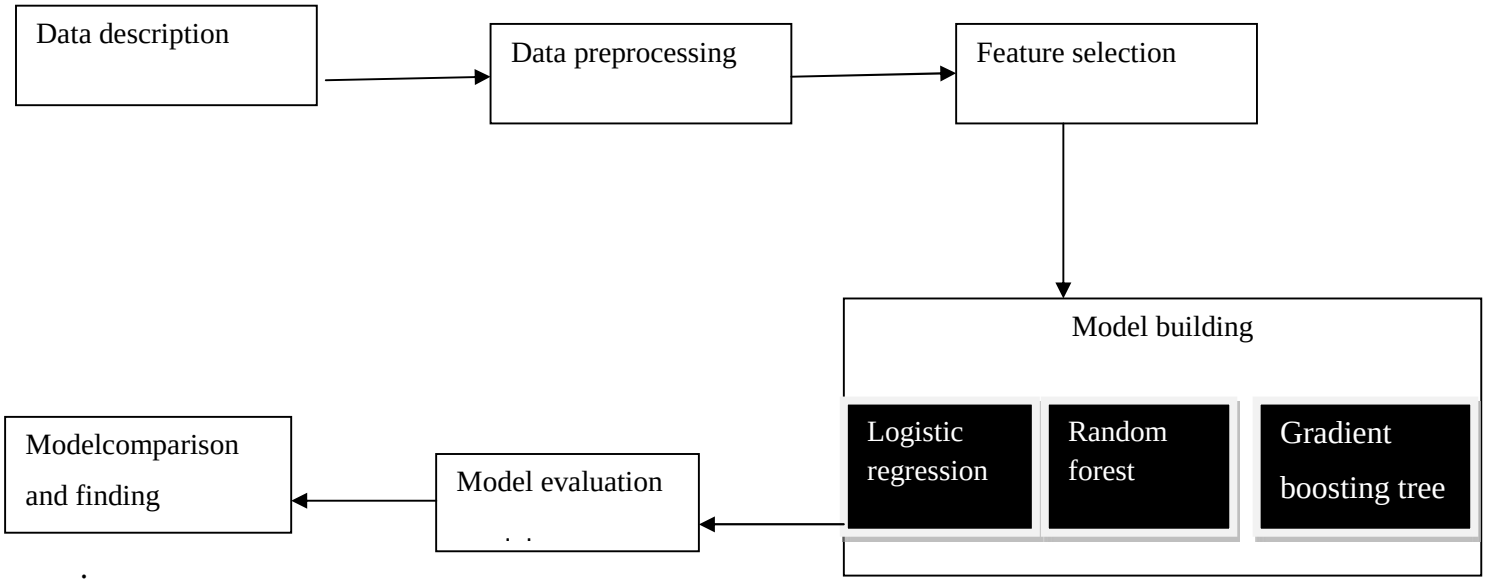
3.1 Introduction

This chapterpresents a detailed overview of the design and methodology used in the research toanswer the research question. Generally,the researcher divides the study into 5 phases. The first phase is data collection from the Human Resource Information System (HRIS) on the organization. The second phase is data preprocessing includes many steps such as removing noisy profiles, completing missing data information, removing some unnecessary column or row data, etc. This phase aims to eliminate imperfect information and mistakes in consequence of users' erroneous data entry and collect suitable data for our problem. In the third phase, feature selection selects the best attributes that give good accuracy for the classifier model. In the fourth phase model construction, the researcher used 3 models to predict employee turnover which are logistic regression, random forest, and extreme gradientboosting treeand construct the model into training and testing. The last phase is model evaluation and comparison to find the best method of employee turnover. Evaluate the classifier models by using performance metrics confusion matrix, recall, precision, f-measure, and other performance metrics.

In this chapter experiment design; data description, data preprocessing and feature selection is discussed. In chapter four model building is discussed; from this, the selected classification models, the concept of classification and the implementation phases are discussed.

In chapter five the evaluation and finding part performance metrics, area under roc curve, cross-validation and data visualizationwere discussed.In chapter six experimental results and discussion was discussed. In the last chapter discussed conclusion and recommendation. Python programming with the Jupiter Notebooks interface was used to implement and analyze the experiments of the study.

General information about the methodology is as follows



.

Figure 3 Research processes

3.2 Experiment Design

The design of the numerical experiments was performed on the research that has been created with the intent to comprehensively measure the effectiveness of supervised machine learning algorithms. The data belong to class 0 (employee stay), and the second refers to the probability that the data belong to class 1 (employee turnover). Predicting probabilities of a particular label provides us with a measure of how likely an employee is to leave the company.

3.3 Data collection and Description

Collecting data is one of the most important steps in this process. Meaningful of the data is the heart for creating good prediction models.

The electronically data wascollected from the organization's federal court humanresources information system. Three electronically datasets were collected with all personally identifiableinformation from three different federal court organizationsand integrate them.The first dataset originates from the federal Supreme Court the data was collected from 2009 to

December 2012, during which time 25% of the organization'semployees have left. The original employee data attributes contain 820 instances both active and terminated and consist of 11 features. The second data originate from the federal first instance court. The dataset collects from 2010 to December 2012, during which time 15% of the organization's employees have left. The original employee data contain 1949 both active and terminated instances and 14 features. The third and the last data set collected from the federal higher court the data was collected from 2009 to December 2012, during which time 25 % of the organization's employees have left. The original employee data contain 843 both active and terminated instances and 12 variables. To sum up, from three federal courts 3003 active and 607 terminated 3610 instances during which 21 % of the organization's employees have left. The data sources were recorded in different format form and unstructured it difficult to preprocesses.

General original employee information's below the table.

Table 2: Attributes containing from three datasets employee information

Attributes	Type	Description
First name	categorical	First name of an employee
Last name	categorical	The family name of an employee
Gender	Boolean	Gender of employee
Date of birth	date	Date of birth of an employee
Date in service	date	Date when the employee joined the company
Date in hire	date	Date when an employee started a current job
Position title	categorical	Title of employee
salary	numerical	The income of employee per annum
Reason for turnover	categorical	The reason why an employee left the organization
Existing employee	categorical	Employee stay in the organization
Education	Categorical	The education level of employee
Work experience	Numerical	No of year service in employee job
Departments	Categorical	The working class in the organization

Pension no	Numerical	No of pension
Marital status	Categorical	Family background of employee
Job role	Categorical	Position of employee job working
No of year service	Numerical	duration of time the employee has worked in the Institution
Turnover	categorical	Indication whether the employee leaves or stay.This is the target class

From the above variables Based on feature selection values from the python code result; the remaining variables like position title there is no see any change in the accuracy result.

A. Gender
B. Age
C. education
D. Work experience
E. marital status
F. No of year service
G. Salary
H. departments
I. Turnover

3.4 Data Preprocessing and preparation

Once the data is collected, it's time to observe the condition of outliers, incorrect, missing, or irrelevant information.Data preprocessing is the main task and it takes much time in the study helps tobuilding the given predictive model.We want to clean the datasets to acquire accurate outcome from the given model.Without preprocessing the data set we cannot get accurate results.

According to Pant (2019)data pre-processing is a method of cleaning the raw facts i.e. the facts is accumulated in the actual world and are transformed into a clean data set. Certain steps are performing to convert the facts into small clean records set; this phase of the manner is called statistics pre-processing. Data originates from some raw sources, which potential it may include null values as properly as inappropriate information. It is necessary to clean this fact so that it can

be suitable for the models and generate better outcomes. Before the use of the information for training in the Machine Learning Algorithm and checking out the test data, data is preprocessed to be more appropriate for analysis. We need data pre-processing to obtain accurate results from the utilized model in the machine getting to know the research.

Most of the real-world data is messy, some of these types of data are:

(a) Missing data: Missing data can be found when it is not continuously created or due to technical issues in the application.

(b) Noisy data: This type of data is also called outliers, this can occur due to human errors (human manually gathering the data).

(c) Inconsistent data: This type of data might be collected due to human errors (mistakes with the name or values) or duplication of data.

Three Types of data:

- Numeric (income, age)
- Categorical (gender, nationality)
- Ordinal (low/medium/high).

These are some of the basic pre-processing techniques that can be used to convert raw data

- Conversion of data: As we know that Machine Learning models can only handle numeric features, hence categorical and ordinal data must be somehow converted into numeric features.

- Ignoring the missing values: Whenever we encounter missing data in the data set then we can remove the row or column of data depending on our needs.

- Filling the missing values: Whenever we encounter missing data in the data set then we can fill the missing data manually, most commonly the mean, median or highest frequency value is used

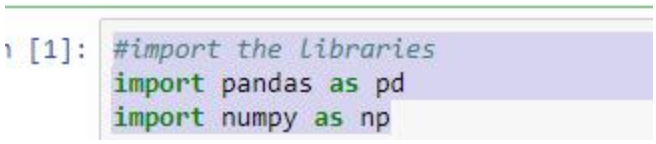
- Outlier detection:Some error data might be present in our data set that deviates drastically from other observations in a data set. [Example: human weight = 800 Kg; due to mistyping of extra 0]

Zhao et al., (2018)suggests that Missing values had been imputed to guarantee that all the algorithms would be in a position to handle them. Nevertheless, some algorithms may want to deal with missing values automatically without imputation, such as XGBoost. To limit the evaluation complexity, the lacking values have been imputed based on their records type. For numerical data types, the missing entries are changed by the median value of the whole entries. For categorical data, the missing entries were replaced through the mode value of the completeEntries. Data Type Conversion and Feature Selection One of the fundamental data preprocessing methods are to convert categorical variables into a numerical format. Because, some algorithms, such as logistic regression, neural networks, and K-nearest neighbors, are not capable to work directly with categorical variables.

Additionally, the researcher was handles the missing value through the use of Fillna, in place, inference, and mean Average techniques to eliminate the given errors. The researcher has preprocessed the data using python Jupiter notebook and excel via making use of the following activities:

1) Translate Amharic employee profile to English:The original employee information most of their profiles are recorded in the Amharic language. So it needs to converts to the English language used to easily understand and interpret the given rawdata. By using Googletranslator software the researcher convert the given employee Amharic profiles into the English language.

2) Import the libraries



This above figure shows that the imported libraries in Python using import keyword and this is the most popular libraries which any Data Scientist used.

- NumPy, short for Numerical Python, is one of the most important foundational pack ages for numerical computing in Python.

- A panda is used to working with tabular data. It consists of data structures and data manipulation tools designed to make data cleaning and analysis fast and easy in Python.

3) Import the data sets

In this section import the three courts employee data individually and preprocesses the datasets each of them. For example, takethe first data is the first instance court employee profiles.This data was before preprocess.

```
employeedata =pd.read_csv('C:/Users/biruk/Desktop/thesis python code/final 1 st instance court  unprocess.CSV')
```

```
employeedata.head(5)
```

	No.	Gender	Birth date	Education	experience	No of year service	marital status	department	Salary	job role	reason to leave	Turnover
0	1	M	6/10/1972	Diploma	NaN	13 years 9 months	NaN	NaN	5310.0	Senior Change Management Professional I	exist	No
1	2	M	15/8/1975	certificate	12.0	1 year	rubbed	Cort Manager Office	1087.0	Court Writer II	exist	No
2	3	SE	17/10/1982	Diploma	3.0	1 year 9 months	rubbed	Registrar's Office	3145.0	Taipei Secretariat	exist	No
3	4	SE	5/12/1982	Diploma	3.0	1 year	single	Registrar's Office	1353.0	Case Message Worker	exist	No
4	5	M	12/8/1986	Degree	12.0	1 year	single	Registrar's Office	3145.0	Case Message Worker	exist	No

```
employeedata.shape # to know the size of the datasets
```

```
(1951, 12)
```

```
col_names = employeedata.columns.tolist()    #to list the colu
print("Column names:")
print(col_names)
```

```
Column names:
['No.', 'Gender', 'Birth date', 'Education ', 'experience', 'Nc
b role', 'reason to leave', 'Turnover']
```

4) Check out the Missing ,noisy and inconsistent type data Values:

```
employeedata.isnull().any() # to the null,noisy,and missing data se
```

```
No.                    False
Gender                 False
Birth date             False
Education              False
experience              True
No of year service      True
marital status          True
department              True
Salary                  True
job role                True
reason to leave        False
Turnover               False
dtype: bool
```

```
employeedata.dtypes # to know the data type used to convert the dat
```

```
No.                     int64
Gender                 object
Birth date             object
Education              object
experience            float64
No of year service     object
marital status         object
department             object
Salary                float64
job role               object
reason to leave        object
Turnover               object
dtype: object
```

from the above we observe that the data values No,Gender,Birth date,Education,Reason to leave and turnover have ‘false’ Boolean values: means that the data is not missing where as the data valuesExperience, No of year service,Department,and Salary are have ‘true’ Boolean values the data set has the missing values and needs to be preprocessed.Additionally, the data type except ‘No’ is an object and float it needs to be converted into numeric; because machine learning is not accept object and float type data rather than numeric.

5) Data type conversions: by using excel sheets and python the categorical data is converting into numeric. As this dataset has a lot of categorical variables, like, salary is converting into ordinal (high, medium and low) and the ordinal variables changes to numeric (0, 1, 2) byusingordinal encoder functions. the variable turnover similarly, the Yes and No values, convert to 0 and1, the variable martial also has single and married variables have converted to 1(‘married) and 2(‘single’)the date format of (dd/mm/yyyy) and the amount format of (0,000,000.00) has to be converting to (yyyy) and (0000000) and similarly the remaining other variables are converted to numeric data values by using excel sheets and Jupiter python.

   For example let us see salary variables how to change ordinal and numeric vice versa.

```
employeedata = {'salary range' : ['Low', 'Medium', 'High']}
df_ordinal = pd.DataFrame(employeedata)
```

```
df_ordinal.head()
```

:

	salary range
0	Low
1	Medium
2	High

```
from sklearn.preprocessing import OrdinalEncoder
ordinalencoder = OrdinalEncoder()
ordinalencoder.fit_transform(df_ordinal[['salary range']])
```

```
array([[1.],
       [2.],
       [0.]])
```

6) After converting the variables RemoveNULL and unnecessary missing values:
   - ✓ Fill the missed value by using median or mode: the missing values were imputed based on their data type by calculating the mean, median or mode of the feature and replace it with the missing values. For example,take education or schoolingvariables; diploma (1),degree (2),master (3), and certificate (0).from this diploma variable is the highest values from the others, so, the missing values filling by the mode value by1.
   - ✓ Removed unnecessary column: remove nonsignificant columns that are not important for prediction purposes such as name, pension no, Job roles, reason to leave, and others.
   - ✓ In placeis used to replace the ordinal and categorical values into numericvalues and dropna used to remove unnecessary variables. For Example 'No' column drops below the figure and similarly, other unnecessary columns are removed.

```
employeedata.drop(['No.'], axis=1)
```

	Gender	Birth date	Education	experience	No of year service	marital status	department	Salary	
0	M	6/10/1972	Diploma	NaN	13 years 9 months	NaN	NaN	5310.0	S
1	M	15/8/1975	certificate	12.0	1 year	rubbed	Cort Manager Office	1087.0	C
2	SE	17/10/1982	Diploma	3.0	1 year 9 months	rubbed	Registrar's Office	3145.0	Taip
3	SE	5/12/1982	Diploma	3.0	1 year	single	Registrar's Office	1353.0	Case Mes
4	M	12/8/1986	Degree	12.0	1 year	single	Registrar's Office	3145.0	Case Mes
5	SE	1/2/1976	Diploma	14.0	7 years 5 months	rubbed	President's Office	4316.0	C
					1 year 8				

```
employeedata.martial.replace('single','0', inplace=True)
```

```
employeedata.martial.replace('married','1', inplace=True)
```

```
employeedata.schooling.replace('certificate','0', inplace=T
```

```
employeedata.schooling.replace('Diploma','1', inplace=True)
```

```
employeedata.schooling.replace('Degree','2', inplace=True)
```

```
employeedata.schooling.replace('master','3', inplace=True)
```

After preprocess the employee variables their data types value is' false 'and numeric type data indicated that thereis no messing value.

```
employeedata.isnull().any()
```

```
Gender                 False
Age                    False
schooling              False
experience             False
No of year service     False
martial                False
department             False
Salary                 False
 Rleave                False
Turnover               False
dtype: bool
```

```
employeedata.dtypes
```

```
Gender                 int64
Age                    int64
schooling              int64
experience             int64
No of year service     int64
martial                int64
department             int64
Salary                 int64
 Rleave                int64
Turnover               int64
dtype: object
```

7) Transform the excel format to CSV (Comma Separated Values) format:for easy loading into the machine learning software; Rearrange the columns:.column name ‘turnover’ is the shift to on the right side in the last column CSV files.

8) Merge the three federal court data sets into one single CSV file and save the clean data in a CSV format file.

	Gender	Age	schooling	experience	No of year service	martial	department	Salary	Turnover
0	1	37	0	12	1	1	1	0	0
1	2	30	1	3	7	1	2	2	0
2	2	26	1	3	1	2	2	0	0
3	1	24	2	12	1	2	2	1	0

```
employeedata.shape
```

```
(3607, 9)
```

```
col_names = employeedata.columns.tolist()    #to list the columns
print("Column names:")
print(col_names)
```

```
Column names:
['Gender', 'Age', 'schooling', 'experience', 'No of year service', 'martial', 'department', 'Salary', 'Turnover']
```

3.5Data Exploration/analysis

Let us take two employee features from ten features and explore each.Primarily, find out the number of employees who left the organization and those who didn't:

```
employeedata['Turnover'].value_counts()
```

```
0    2943
1     664
Name: Turnover, dtype: int64
```

```
employeedata['Turnover'].value_counts()/len(employeedata)*100 # to know the percentage of employee who stay and le
```

```
0    81.59135
1    18.40865
Name: Turnover, dtype: float64
```

```
employeedata['martial'].value_counts()
```

```
2    2661
1     946
Name: martial, dtype: int64
```

There are 664 employees left and 2943 employees stayed in the federal court organizations.

There are 2661 employees are single and 946 employees are married.

```
employeedata.groupby('Turnover').mean()
```

Turnover	Gender	Age	schooling	experience	No of year service	martial	department	Salary
0	1.529392	35.22528	0.869861	7.423378	5.67754	1.754332	4.994903	3478.329596
1	1.293675	36.40512	0.835843	4.528614	3.88253	1.664157	5.210843	3383.207831

From the above Snapshootwe observe that

- ✓ The average gender of employees who stay with the organization is higher than that of the employees who left.
- ✓ The average no of year service of employees who left the company is less than that of the employees who stayed.
- ✓ The employees who had high salary payments are less likely to leave than that of the employee who did not have a high salary payment.
- ✓ The average experience of employees who stay is greater than that of the employee left. From this more experienced employees are staying in the organization, and who employees have low experience is left from the institutions.

3.6Feature Selection

Feature selection is one of the main concepts of DM and Machine Learning.It is a process of selecting necessary useful variables in a dataset to improve the results of machine learning and make it more accurate. At which, Using too many numbers of variables in a dataset reduce predictive performance. The data set may contain too many features; some of them do not promote the prediction accuracy, and thus make the predictive model excessively complicated. Therefore, unnecessary useless variables must be avoided to make the model efficiently works.

Deciding which unnecessary variable to avoid can be done in a manual manner using domain knowledge or it can be done automatically(Nasr et al., 2019).

Another researcher Chen & Jeong(2007) elimination (RFE) the technique for feature selection in small training pattern classification. Pointed out that Feature selection is to choose a subset of applicable features from a larger set of original ones in terms of some pre-defined criteria such as classification performance or category separability. It performs a sizeable position in machine learning applications. Feature selection is especially vital in small pattern classification problems, where the number of available training samples is very small compared to the number of features. The primary benefit of feature selection in small sample classification issues is to overcome overfitting issues to improve prediction performance. Among several feature selection methods, Recursive Feature Elimination (RFE) is a currently developed feature selection technique for small pattern classification problems. RFE tends to discard "weak" features, which may additionally provide a full-size enhancement in overall performance when combined with other features. In the paper, they endorse more desirable recursive feature.

After pre-processing, the researcher was applying a choice variable to keep only the most applicable features to achieve the best accuracy. Limiting the number of features used in machine learning problems is to be sure vital to right manage the complexity of the learning phase and to keep away from over fitting. In the study, Recursive Feature Elimination (RFE) is proposing for function selection helps to identify which variables contribute the most to predicting the target attribute. Feature selection helps us determine which variables are significant that can predict employee turnover with excellent accuracy statistical records analysis.

Apply feature selection help us come to a decision in which variables are important that can predict employee turnover with the best accuracy. For example there are a total of 9columns, how about select 6?

```
from sklearn.feature_selection import RFE
rfe = RFE(model,6)
rfe = rfe.fit(X_train,y_train)
print(rfe.support_)
print(rfe.ranking_)
```

```
C:\Users\brook\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:432: FutureWarning: Default solver will be changed to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\brook\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:432: FutureWarning: Default solver will be changed to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\brook\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:432: FutureWarning: Default solver will be changed to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)
```

```
[ True  True False  True  True  True False  True]
[1 1 3 1 1 1 2 1]
```

The figure shows that how feature selection select the best variables.You can examine that RFE select the 6 variables for us, which are marked true in the support_ array and marked with a choice “1” in the ranking array. They are:

‘Gender’,’age’,’experience’,’no year service’,’martial’and ‘salary’.

3.7Tool

As earlier described the above activities are done by using python Jupiter notebook. The reason to select this tool:-

as pointed byPedregosa et al., (2011)Scikit-learn harnesses a wealthy environment to provide state-of-the-art implementations of many typical computing device mastering algorithms while

maintaining an easy-to-use interface tightly integrated with the Python language. These solutions the developing want for statistical records evaluation using non-specialists in the software and web industries, as well as in fields outdoor of computer science, such as biology or physics. According to the researcher, Scikit-learn differs from other machine learning toolboxes in Python for a variety of reasons: the first it contains compiled code for efficiency, the second it relies upon only on numpy and scipy to facilitate easy distribution that has optionally available dependencies such as R and shogun, and the third focuses on essential programming, which makes use of a data-flow framework. Binary packages are reachable on a wealthy set of platforms along with Window platforms. Additionally, Scikit-learn exposes a wide range of computing device studying algorithms, both supervised and unsupervised, the use of a consistent, task-oriented interface, thus enabling easy assessment of techniques for a given application. Since it relies on the scientific Python ecosystem, it can without difficulty be built-in into applications outside the traditional range of statistical data analysis.

CHAPTER FOUR

BUILDING THE PREDICTIVE MODEL

Our main goal is to train the best performing model possible using the pre-processed data.In this section the pre-processed data is used to build machine learning models to predict employee turnover. The label data is available for training and supervised machine learning models are used for modeling in this study.

4.1. Classification techniques

Soofi &Awan (2017)examined that supervised machine learning techniques try to discover out the relationship between input attributes (independent variables) and a target attribute (dependent variable). Supervised methods can similarly be categorized into two primary categories; classification and regression. In the regression, the output variable takes non-stop values while in classification output variable takes class labels. Classification is a data mining (machine learning) approach used to forecast team membership for records instances. Classification is an admired task in machine learning specifically in plans and knowledge discovery.
Additionally,Pant(2019)knows that Classification is the process of predicting the type of given data points. Classes are now and then called as targets/ labels or categories. Classification predictive modeling is the assignment of approximating a mapping function (f) from input variables (X) to discrete output variables (y).A classification method is used to strengthen the prediction fashions in the study. Classification is used to classify each object in a set of statistics into one of a predefined set of lessons or groups.

For the study, the researcher classified employee profiles into two classes; employees who leave their current organizations (for turnover) as"1" and employees who proceed to work for the current organization (stayed) as "0".

4.2 Splitting the data-set into Training and Test Set

For training a model the researcher initially split the model into two sections which are 'Training data**',** and 'Testing data**'.** Train the classifier using 'training data set**',**' and then test the performance of the classifier on unseen 'test data set**'**.

- Training set: Training set is used to build a model.

According to data science,80% of the data of the dataset is taken for training data.

- ❖ Test set: Testing data is used to test the system. It is the set of data that is used to verify whether the system is producing the correct output after being trained or not.

Generally, we split the data-set into 70:30 ratios or 80:20 what does it mean, 70/80 percent data take in train and 30/20 percent data take in the test. However, this Splitting can be varies according to the data-set shape and size.The researcher want to do here is to divide the dataset into a training set and a testing set using function train_test_split(). It requires passing 3 parameters - features, target, and test-size. The dataset is split into two sections in ratio of 70:30. It means 70% data is use for model training and 30% is used for model testing.

```
#splitting the data set into training and testing
import sklearn
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.3,random_state=42)
```

```
len(X_train)
```

```
2524
```

```
len(y_test)
```

```
1083
```

Figure 4splitting Training and Testing Set data

- ✓ X_train is the training part of the matrix of features.
- ✓ X_test is the test part of the matrix of features.
- ✓ y_train is the training part of the dependent variable that is associated to X_train .
- ✓ y_test is the test part of the dependent variable that is associated to X_train .

From the above snapshot, it can understood that train-test –split classify the dataset into training and testing. It requires features, target, and test size. The size of the training dataset is 2524 and 1083the length of testing data set in the ratio of 70: 30 data sets.

4.3 Selected Classification Prediction Models

As earlier the researcher tries to describe on the literature part there are many different classification techniques or classifiers, for the study using three classification classifiers; logistic regression, random forest, and gradient boosting tree is select for model building. These models aretrainand test using the same dataset and accuracy is determined using the cross-validation technique. After that, fit the model on train set using fit () and performed prediction on the test set usingpredict ().finally,A model has the bestaccuracyis selected from the experiments and is used to answer the research questions.

4.3.1 Random forest

Random forests are an effective tool in prediction. Because of the Law of Large Numbers, they do no longer overfit. Injecting the proper kind of randomness makes them correct classifiers and regressors. Furthermore, the framework in phrases of the strength of the individual predictors and their correlations offers insight into the capacity of the random forest to predict for a while; the conventional thinking was that forests should now not compete with arcing kind algorithms in phrases of accuracy. Boosting and arcing algorithms can decrease bias as well as variance. Random forests take an ensemble approach that offers an improvement over the fundamental choice tree structure with the aid of combining a group of weak learners to form a better learner(Breiman, 2001).

According toSabbeh(2018)stated that Random Forests (RF) are an ensemble learning technique that can support classification and regression. It extends the fundamental thinking of a single classification tree by increasing many classification trees in the training phase. To classify an instance, each tree in the forest generate its response (vote for a class), the model chooses the class that has obtained the most votes basic the trees in the forest. One predominant benefit of RF

over traditional decision trees is the protection in opposition to overfitting which makes the model in a position to deliver a high performance.

4.3.2 Logistic regression

Logistic regression is a specialized form of regression used to predict and provide an explanation for a categorical based variable. It works high-quality when the structured variable is a binary specific variable. One distinctive advantage of logistic regression is that it is not restricted with the aid of the normality assumption which is a fundamental assumption in the regression analysis. This method can also accommodate non-metric variables such as nominal or express variables utilizing coding them into dummy variables. Another benefit of logistic regression is that it immediately predicts the chance of an event occurring. To make sure that the dependent variable, which is the probability, is bounded between zero and one, the logistic regression defines a relationship between the dependent and independent variables that resembles an S-shaped curve, which uses an iterative manner to estimate the 'most likely' values of the coefficients. This results in the use of a 'likelihood' function in becoming the equation rather than using the sum of squares approach of the regression analysis(Yousaf & Bhulai, 2016).

The primary output in logistic regression is a probability that the given input point belongs to a certain class. Based on the value of the probability, the model creates a linear boundary separating the input space into two regions. Logistic regression is easy to implement and work well on linearly separable classes, which makes it one of the most widely used classifiers(Zhao et al., 2018). Lastly Logistic regression classifier is one of the basic linear models for classification. Logistic regression is a specific category of regression best used to predict for binary or categorical dependent variables. It’s often used with regularization in the form of penalties to avoid over-fitting(Ajit, 2016).

4.3.3 Gradient Boosting tree

"Boosting" is a widespread method for enhancing the performance of any learning algorithm. Boosting can appreciably reduce the error of any "weak" studying algorithm that consistently

generates classifiers that want only to be a little bit better than random guessing. By again and again going for walks a given susceptible learning algorithm on various distributions of the training data, and then combining the weak learner classifiers into a single composite classifier (Freund & Schapire, 1996)**.**

Gradient boosting trees is an ensemble machine learning method proposed in 2001 by Friedman 2001 for regression and classification purposes. The algorithm is iteratively constructed and boosts a sequence of decision trees, every being trained and pruned on examples that have been filtered by way of earlier skilled trees. The incorrectly categorized examples using the preceding trees are resample with a higher likelihood to supply a new probability distribution for the next ace in the ensemble to train onDrucker & Cortes(1996),Breiman(1997) and Friedman(2001).Unlike highly interpretable single DT, GBT is harder to visualize and interpret.

Gradient boosted decision tree (GBDT) is a powerful machine-learning technique that has a wide range of commercial and academic applications and produces state of-the-art results for many challenging machine learningproblems. The algorithm builds one decision tree at a time to fit the residual of the trees that precede it. GBDT has been widely used recently mainly due to its high accuracy, fast training and prediction time, and small memory footprint(Si et al., 2017).

The difference between RFand GBT is the gradient boosted tree models learn sequentially. In GBT, a series oftrees are built and each tree attempts to correct the mistakes of the previous tree in the series. Trees are added sequentially until no further enhancement can be achieved. Making predictions in GBT is fast and memory-efficient; boosting could be viewed as a form of regularization to reduce overfitting(Murphy, 2012).

According to Hossain(2019)Gradient boosting is known as one of the most accepted and effective methods used for predictive model building. Boosting refers to renovating weak learners into strong learners. The usual thought is to training a choice tree so that each new tree is a fit on a modified model of the original data set. This procedure is repeated for a particular range of iterations. Subsequent trees assist us to categorize observations that are no longer nicely categorized via the previous trees. Forecasts of the remaining collaborative model are consequently the weighted sum of the predictions made using the previous three models.

- By using different experiments, that means by using the random-state and parameter tuningvalues the researcher is trying to evaluate the accuracy of the three classification model..

- The first experiment is By using random-state

The reason to select random state it helps tomake fixed same data on training and testing data.

First Experiment (E1): Using the random-state values=0

Logistic regression model

```
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
model= LogisticRegression()
model.fit(X_train,y_train)
```

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='l2', random_state=None, solver='warn',
          tol=0.0001, verbose=0, warm_start=False)
```

```
from sklearn.metrics import accuracy_score
accuracy_score(y_test,modelpred )
```

```
0.8190212373037857
```

Random forest

```
from sklearn.ensemble import RandomForestClassifier
rf=RandomForestClassifier()
rf.fit(X_train,y_train)
```

```
C:\Users\user\Documents\anaconda2\lib\site-packages\sklearn\ensemble\forest.py:246: FutureWarning: The default value of n_e
stimators will change from 10 in version 0.20 to 100 in 0.22.
  "10 in version 0.20 to 100 in 0.22.", FutureWarning)
```

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=None,
            oob_score=False, random_state=None, verbose=0,
            warm_start=False)
```

```
accuracy_score(y_test,modelpred )
```

```
0.8190212373037857
```

Figure 5 experment1 random state=0

From the above figure the first logistic regression and random forest classification model 81 % and gradient boosting tree is 85 % accuracy are outperformingwhen the random state value is =0

Table3: accuracy percentage for classification models in E1

No	Classifiers	Random-state number	Prediction accuracy
1	Random forest	0	81%
2	Logistic regression	0	81%
3	Gradient boosting tree	0	85%

Second Experiment (E2): Using the random-state values=10

Figure 6experment2 random state=10

In the second experment the two classfication models ;random forest and logistic regression is 63% and gradient boosting tree is 83% is outperforming when random-state =10

Table 3: accuracy percentage for classification models in E2

No	Classifiers	Random-state number	Prediction accuracy
1	Random forest	10	63%
2	Logistic regression	10	63%
3	Gradient boosting tree	10	83%

Third Experiment (E3): Using the random-state values=30

Figure 7 experment3 random state= 30

In the third experiment random forest and logistic regression classification models outperforms the same 66% and gradient boosting tree 87% accuracy are outperforms when random-state=30.

Table 4: accuracy percentage for classification models in E3

No	Classifiers	Random-state number	Prediction accuracy
1	Random forest	30	81%
2	Logistic regression	30	81%
3	Gradient boosting tree	30	84%

FourthExperiment (E4): Using the random-state values=20

Figure 8 experiment 4 random-state=20

In the fourthexperiment random forest and logistic regression classification models outperforms the same 83% and gradient boosting tree 86% accuracy are outperforms when random-state=20.

Table 5: accuracy percentage for classification models in E4

No	Classifiers	Random-state number	Prediction accuracy
1	Random forest	20	83%
2	Logistic regression	20	83%
3	Gradient boosting tree	20	86%

Fifth Experiment (E5): Using the random-state values=42

Figure 9 experment5 random state=42

In the fifthexperiment random forest and logistic regression classification models outperforms the same 83% and gradient boosting tree 87.5% accuracy are outperforms when random-state=42.

Table 6: accuracy percentage for classification models in E5

No	Classifiers	Random-state number	Prediction accuracy
1	Random forest	42	83%
2	Logistic regression	42	83%
3	Gradient boosting tree	42	87.5%

- The second is by changing the classifiers parameter(parameter-tuning)

Logistic regression

The first classifier is logistic regression, primarily run the algorithm by default and the accuracy result is 81%.secondly,by changing the parameter 'solver = ' warn' to "lbfgs" and the multi-class parameter = "warn" to "multinomial" the accuracy is similar to the previous one that is 81%.finally ,change the parameter of the algorithm "solver"='' Newton-cg'' , 'class-weight ' ='' balanced'',max-iter=1000,multi-class='' ovr'' after running the code the accuracy also the same to the previous accuracy.

Logistic regression model

```
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
model= LogisticRegression()
model.fit(X_train,y_train)
```

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='12', random_state=None, solver='warn',
          tol=0.0001, verbose=0, warm_start=False)
```

```
from sklearn.metrics import accuracy_score
accuracy_score(y_test,modelpred )
```

```
0.8107109879963066
```

```
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
model= LogisticRegression(random_state=0, solver='lbfgs', multi_class='multinomial')
model.fit(X_train,y_train)
```

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='multinomial',
          n_jobs=None, penalty='12', random_state=0, solver='lbfgs',
          tol=0.0001, verbose=0, warm_start=False)
```

```
from sklearn.metrics import accuracy_score
accuracy_score(y_test,modelpred )
```

```
0.8107109879963066
```

```
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
model= LogisticRegression(random_state=1, solver='newton-cg', multi_class='ovr',class_weight='balanced',max_iter=1000)
model.fit(X_train,y_train)
```

```
LogisticRegression(C=1.0, class_weight='balanced', dual=False,
          fit_intercept=True, intercept_scaling=1, max_iter=1000,
          multi_class='ovr', n_jobs=None, penalty='12', random_state=1,
          solver='newton-cg', tol=0.0001, verbose=0, warm_start=False)
```

```
from sklearn.metrics import accuracy_score
accuracy_score(y_test,modelpred )
```

```
0.8107109879963066
```

Random forest

The second classifier is the random forest, firstly run the algorithm by default and the accuracy result is the same as logistic regression 81%.secondly,by changing the parameter ‘max-depth= ‘ none’ to 2 and the random-state = “none” to 0 the accuracy is similar to the previous one that is 81%.finally ,change the parameter of the algorithm max_depth=3, max_leaf_nodes=2, n_estimators = 100 after running the code the accuracy also the same to the previous classifiers accuracy.

Random forest

```
from sklearn.ensemble import RandomForestClassifier
rf=RandomForestClassifier()
rf.fit(X_train,y_train)
```

```
C:\Users\user\Documents\anaconda2\lib\site-packages\sklearn\ensemble\forest.py:246: FutureWarning: The default value of n_e
stimators will change from 10 in version 0.20 to 100 in 0.22.
  "10 in version 0.20 to 100 in 0.22.", FutureWarning)
```

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=None,
            oob_score=False, random_state=None, verbose=0,
            warm_start=False)
```

```
accuracy_score(y_test,modelpred )
```

```
0.8107109879963066
```

```
from sklearn.ensemble import RandomForestClassifier
rf=RandomForestClassifier(max_depth=2, random_state=0)
rf.fit(X_train,y_train)
```

```
C:\Users\brook\Anaconda3\lib\site-packages\sklearn\ensemble\forest.py:245: FutureWarning: The default value of n_estimators wil
l change from 10 in version 0.20 to 100 in 0.22.
  "10 in version 0.20 to 100 in 0.22.", FutureWarning)
```

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                       max_depth=2, max_features='auto', max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=10,
                       n_jobs=None, oob_score=False, random_state=0, verbose=0,
                       warm_start=False)
```

```
accuracy_score(y_test,modelpred )
```

```
0.8107109879963066
```

```
from sklearn.ensemble import RandomForestClassifier
rf=RandomForestClassifier(max_depth=3,max_leaf_nodes=2,n_estimators=100)
rf.fit(X_train,y_train)
```

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                       max_depth=3, max_features='auto', max_leaf_nodes=2,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       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)
```

```
accuracy_score(y_test,modelpred )
```

```
0.8107109879963066
```

Gradient boosting tree

The third classifier is gradient boosting tree, firstly from the previous classifier run the algorithm by default and the accuracy result is the is 83%.secondly,by changing the parameter ‘max-depth= 3,’max-features’=2, mini-sample-split=2 and n-estimators=20 the accuracy is similar to the

previous classifiers that is 81%.thirdly, the parameter values,'n-estimators'=2000 and 'random-state=0 the accuracy is better than the previous accuracy is 84.fourthly, change the parameter 'learning-rate=0.06 and 'n-estimators' =1000 the accuracy is also the best one from the previous outcome is 87%.finally ,by selecting the best parameter that has the best accuracy,learning rate ='0.06,'max-depth=3 and 'n-estimators=1000 the accuracy is 87.%.this indicates that by changing the good parameters that have good accuracy we can get good prediction outcome. Fromgradient boosting parameters, n-estimators, learning rate and max depth are affect the outcome of the given classification models.Therefore, the gradient boosting tree classifier is a good model among the three classification models.

Gradient boosting tree

```
from sklearn.ensemble import GradientBoostingClassifier
gbc = GradientBoostingClassifier()
gbc.fit(X_train, y_train)
```

```
GradientBoostingClassifier(criterion='friedman_mse', init=None,
              learning_rate=0.1, loss='deviance', max_depth=3,
              max_features=None, max_leaf_nodes=None,
              min_impurity_decrease=0.0, min_impurity_split=None,
              min_samples_leaf=1, min_samples_split=2,
              min_weight_fraction_leaf=0.0, n_estimators=100,
              n_iter_no_change=None, presort='auto', random_state=None,
              subsample=1.0, tol=0.0001, validation_fraction=0.1,
              verbose=0, warm_start=False)
```

```
gbc.score(X_test, y_test)
```

```
0.8320716520162512
```

```
from sklearn.ensemble import GradientBoostingClassifier
gbc = GradientBoostingClassifier( learning_rate=0.05, max_depth=2,max_features=2, min_samples_split=2,n_estimators=20,)
gbc.fit(X_train, y_train)
```

```
GradientBoostingClassifier(criterion='friedman_mse', init=None,
              learning_rate=0.05, loss='deviance', max_depth=2,
              max_features=2, max_leaf_nodes=None,
              min_impurity_decrease=0.0, min_impurity_split=None,
              min_samples_leaf=1, min_samples_split=2,
              min_weight_fraction_leaf=0.0, n_estimators=20,
              n_iter_no_change=None, presort='auto',
              random_state=None, subsample=1.0, tol=0.0001,
              validation_fraction=0.1, verbose=0,
              warm_start=False)
```

```
gbc.score(X_test, y_test)
```

```
0.8116343490304709
```

```
from sklearn.ensemble import GradientBoostingClassifier
gbc = GradientBoostingClassifier(  n_estimators=2000,random_state=0)
gbc.fit(X_train, y_train)
```

```
GradientBoostingClassifier(criterion='friedman_mse', init=None,
                           learning_rate=0.1, loss='deviance', max_depth=3,
                           max_features=None, max_leaf_nodes=None,
                           min_impurity_decrease=0.0, min_impurity_split=None,
                           min_samples_leaf=1, min_samples_split=2,
                           min_weight_fraction_leaf=0.0, n_estimators=2000,
                           n_iter_no_change=None, presort='auto',
                           random_state=0, subsample=1.0, tol=0.0001,
                           validation_fraction=0.1, verbose=0,
                           warm_start=False)
```

```
gbc.score(X_test, y_test)
```

```
0.8494921514312096
```

```
from sklearn.ensemble import GradientBoostingClassifier
gbc = GradientBoostingClassifier( learning_rate=0.06,n_estimators=1000)
gbc.fit(X_train, y_train)
```

```
GradientBoostingClassifier(criterion='friedman_mse', init=None,
                           learning_rate=0.06, loss='deviance', max_depth=3,
                           max_features=None, max_leaf_nodes=None,
                           min_impurity_decrease=0.0, min_impurity_split=None,
                           min_samples_leaf=1, min_samples_split=2,
                           min_weight_fraction_leaf=0.0, n_estimators=1000,
                           n_iter_no_change=None, presort='auto',
                           random_state=None, subsample=1.0, tol=0.0001,
                           validation_fraction=0.1, verbose=0,
                           warm_start=False)
```

```
gbc.score(X_test, y_test)
```

```
0.8725761772853186
```

finally select the best parametre have good accuracy

```
from sklearn.ensemble import GradientBoostingClassifier
gbc = GradientBoostingClassifier(  n_estimators=1000, learning_rate = 0.05, max_depth = 3,  random_state = 0)
gbc.fit(X_train, y_train)
```

```
GradientBoostingClassifier(criterion='friedman_mse', init=None,
                           learning_rate=0.05, loss='deviance', max_depth=3,
                           max_features=None, max_leaf_nodes=None,
                           min_impurity_decrease=0.0, min_impurity_split=None,
                           min_samples_leaf=1, min_samples_split=2,
                           min_weight_fraction_leaf=0.0, n_estimators=1000,
                           n_iter_no_change=None, presort='auto',
                           random_state=0, subsample=1.0, tol=0.0001,
                           validation_fraction=0.1, verbose=0,
                           warm_start=False)
```

```
gbc.score(X_test, y_test)
```

```
0.8707294552169099
```

4.3 cross-validation

To evaluate the prediction of the selected models, k-fold cross-validation is applied to reduce the bias of sampling data and ensuring model error randomness. K-fold cross-validation randomly divides data into k subsets and one subset is used as testing data and k-1 subsets are used as training data. This process is repeated k times to cover all data(Yousaf & Bhulai, 2016).

According to Arlot & Celisse(2010)multiple strategies can be used to stop over-fitting and get a more realistic performance estimate, including, but no longer restricted to k-fold cross-validation. Validation is the procedure of splitting the records into education and a validation

phase had been the training statistics is used for training and the validation data for evaluation. Cross-validation is a method that can be used to determine how well the predictive ability of an algorithm can be generalized to an independent dataset. The dataset is randomly broken up into ok mutually exclusive subsets (the folds) of approximately equal size. Each subset is used as the "test" set once and used to consider the model that was fit the usage of all different subsets as training data. This procedure is repeated so that all folds are used as the "test" set once. The cross-validation estimate of accuracy is the general number of correct classifications, divided using the variety of instances in the dataset.(Kohavi, 1995) demonstrated that ten-fold cross-validation is often better, even if the data allows for more folds.

Therefore, the researcher is trying to use ten-fold cross-validation as an evaluation scheme. The evaluation scheme is used to estimate the accuracy of the given models.To see the correct accuracy result.Let us train the classification models and see the result.

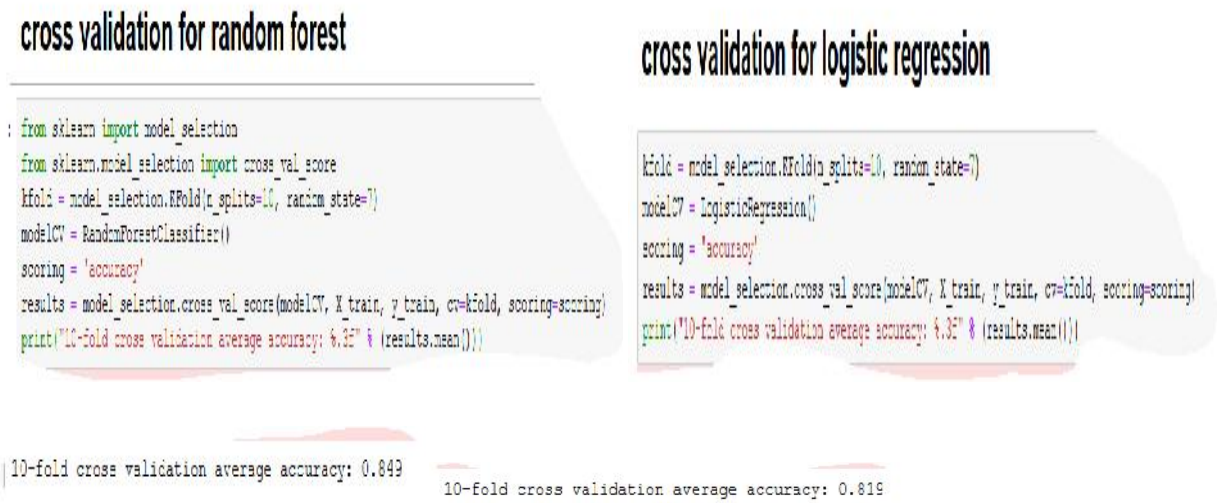


Figure 10 cross- validation accuracy for random forest and logistic regression classifiers

cross validation for gradient boosting tree

```
kfold = model_selection.KFold(n_splits=10, random_state=7)
modelCV = GradientBoostingClassifier()
scoring = 'accuracy'
results = model_selection.cross_val_score(modelCV, X_train, y_train, cv=kfold, scoring=scoring)
print("10-fold cross validation average accuracy: %.3f" % (results.mean()))
```

```
10-fold cross validation average accuracy: 0.859
```

Figure 11 cross -validation accuracy for gradient boosting tree classifier

As we can see from the cross validation results the average accuracy remains very close to the gradient tree model accuracy; hence, we can conclude that the model generalizing well.

Table 7: Accuracy percentages for classification models using 10 fold cross- validation

No	Classifiers	Prediction accuracy
1	Random forest	85%
2	Logistic regression	82%
3	Gradient boosting tree	86%

CHAPTER FIVE

EVALUATION AND DATA VISUALIZATION

Model Evaluation is an essential part of the model development process. It helps to find the best model that represents the data and how well the chosen model will work in the future. Inthis section the researchers try to evaluate the three classification models by using different evaluation metrics.Developing the prediction models with the above algorithms and training the models with training data set, ran the test data for each model and validate by using confusion matrix, recall, precision, f1-measure, and Roc curve. After evaluation discusses the result of the classification models and tries to answer the research questions.

5.1 Evaluation metrics

There are different metrics to evaluate models in machine learning, like TPR, FPR, and TNR, accuracy, precision, recall and F-measure and so on. For the study, the researcher was applied those evaluation metrics to evaluate the given classifier performance.

5.1.2 Confusion matrix

Aconfusion matrix is used to measure the accuracy of a classification model the number of correct and incorrect predictions made by a classifier.

Maria Navin & Pankaja,R (2016)Stated that the evaluation of document classification techniques can be obtained in terms of correctness by computing statistical measures namely the True Positives (TP), True Negatives (TN), False Positive (FP) and False Negatives (FN). These components form the Confusion Matrix as shown in the table.

A confusion matrix is a table that can be generated for a classifier on a binary data set and can be used to describe the performance of the classifier.

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

Table 8: confusion matrix

This matrix is based on the terms

- True Positives (TP) - prediction and actual both are yes.
- True Negatives (TN) - prediction is and actual is no.
- False Positives (FP) - prediction is yes and actual is no.
- False Negatives (FN) - prediction is no and actual is yes.

Other metrics for performance evaluation are Precision, Recall & F-Measure

- Precision — what proportion of positive predictions were correct? A model that produces no false positives has a precision of 1.0
- Recall — what proportion of actual positives were predicted correctly? A model that produces no false negatives has a recall of 1.0.
- F1 -score — a combination of precision and recall. The closer to 1.0, the better.
- Receiver operating characteristic (ROC) curve & Area under the curve (AUC) — The ROC curve is a plot comparing true positive and false-positive rates. It is a performance metric for binary classification problems is used. The ROC represents **a** model's ability to discriminate between positive and negative classes and is better suited to this project. An area of 1.0 represents a model that made all predictions perfectly. An area of 0.5 represents a model as good as random.

Let us createa confusion matrixbelow to analyze predictions made by a classifier and evaluate the accuracy of the classification.

Random Forest

```
from sklearn.metrics import classification_report
print(classification_report(y_test, rf.predict(X_test)))
```

	precision	recall	f1-score	support
0	0.86	0.97	0.91	879
1	0.71	0.30	0.43	204
accuracy			0.85	1083
macro avg	0.79	0.64	0.67	1083
weighted avg	0.83	0.85	0.82	1083

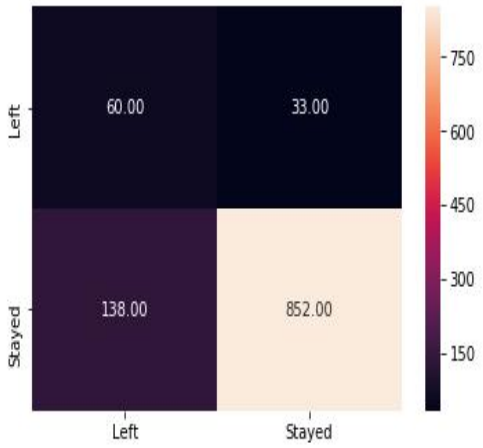


Figure 12confusion matrix for random forest

Logistic regression


```
print(classification_report(y_test, model.predict(X_test)))
```

	precision	recall	f1-score	support
0	0.89	0.62	0.73	879
1	0.29	0.66	0.40	204
accuracy			0.63	1083
macro avg	0.59	0.64	0.57	1083
weighted avg	0.77	0.63	0.67	1083

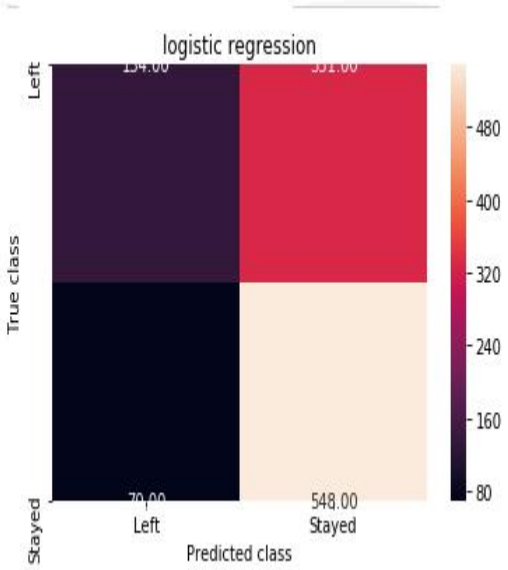


Figure 13 confusion matrix for logistic regression

Gradient boosting tree

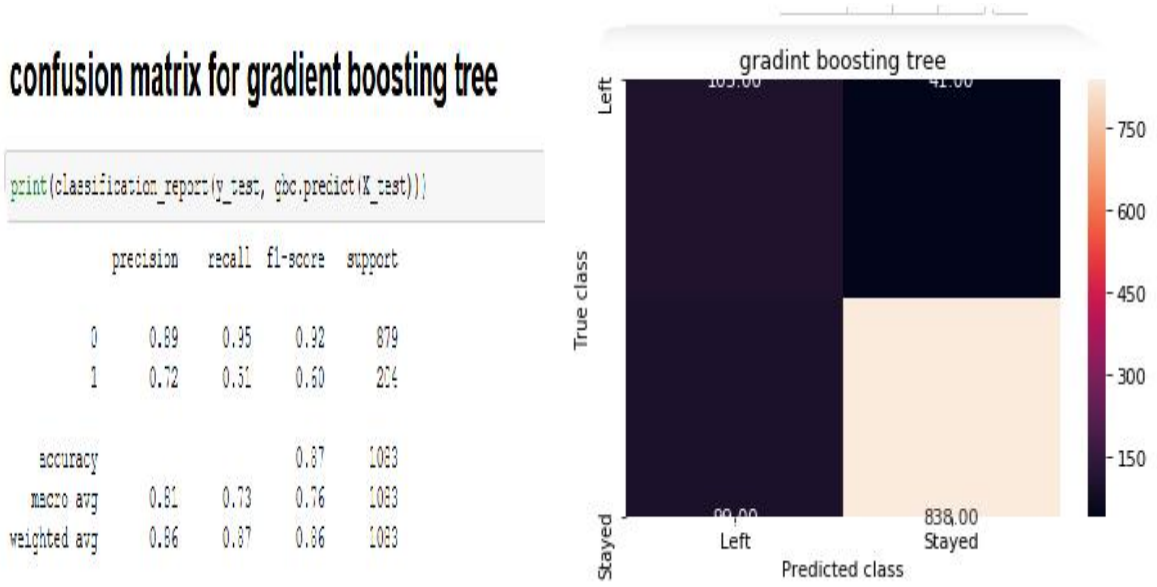


Figure 14confusion matrix for gradient boosting tree

When an employee left, what proportion of actual positive werethe classifiers predict correctly? This measurement is called recall.From the above confusion matrixlogistic regression is best for this measurement**.** Out of all the turnover cases, the logistic regression tree correctly retrieved 134 out of 204. This translates to a turnover "recall" of about 66% (134/204) classifierfar better than random forests (30%) classifier and gradient boosting tree (51%) classifier .additionally,When a model predicts an employee leave, What percentage of positive identifications the classifiers was actually correct? This measurement is called precision. From this measurement gradient boosting treethe classifier is best to outperform to the remaining two classifiers 72% precision (103/ out of 144) with logistic regression at about 29% (134 out of 465) and random forest 71%.

Table 9: performancemeasure for classification models

no	Classifier	Recall	Precision	F1-score
1	Random forest	30%	71%	43%
2	Logistic regression	66%	29%	40%
3	Gradient boosting tree	51%	72%	60%

5.2Evaluation measure using ROC curve

The performance of classification classifiers is also evaluated using ROC curve. The datasetsare divided into training and testing sets random-state and each model istrained on a training set and evaluated on a testing set. ROC curves found from theseclassifiers for every dataset are given in below figures.

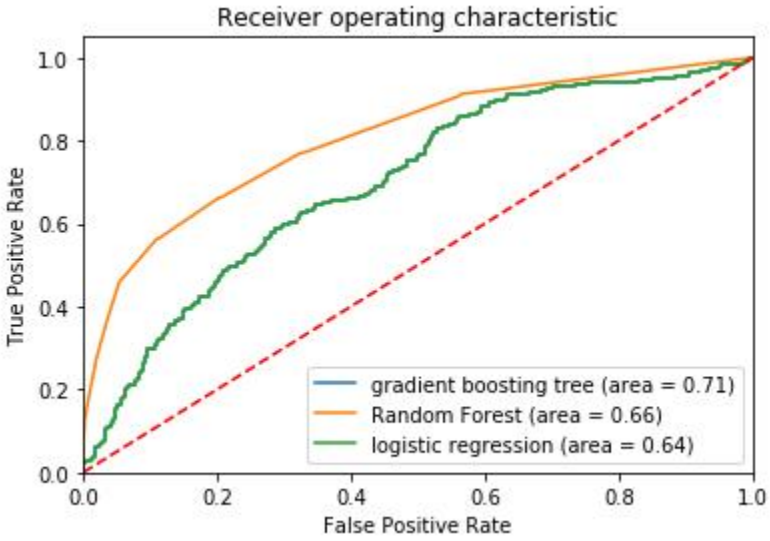


Figure 15 the Roc-Curve accuracy result

The ROC curves show from the three classification modelsgradient boosting tree worked the best on the given datasets.

5.3. Data visualization

Data Visualization: visualize the data to get a much clearer picture of the data and helps to identify the important feature values. Various graphs and plots were plotted for analyzing the effect of the attributes on employee turnover.

- matplotlib is a desktop plotting package designed for creating (mostly two dimensional) publication-quality plots..By using matplotlib let us visualize the data below the figure. Analyze the four significant factors for employee turnover by using a bar chart.

Bar chart for employee age and the frequency of turnover

```
%matplotlib inline
import matplotlib.pyplot as plt
pd.crosstab(employeedata.Age,employeedata.Turnover).plot(kind='bar')
plt.title('Turnover Frequency for Age')
plt.xlabel('Age')
plt.ylabel('Frequency of Turnover')
plt.savefig('Age_bar_chart')
```

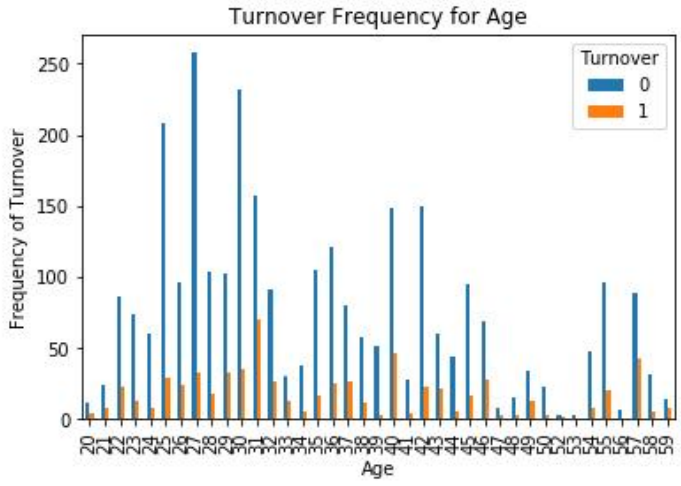


Figure 16Turnover frequency of age

From this we examine that the frequency of employee turnover depends a great deal on the employee's age. As we observe from the bar chart most of the organization employee's age is between twenty- five and fifty- seven. From this between twenty five and fourty -two ages were left from the organization. It indicates that no of young employees who left with the organization is higher than the no of adults employees who.Thus, Age is a good predictor of the employee turnover.

Bar chart for employee salary level and the frequency of turnover

```
table=pd.crosstab(employeedata.Salary, employeedata.Turnover)
table.div(table.sum(1).astype(float), axis=0).plot(kind='bar', stacked=True)
plt.title('Stacked Bar Chart of Salary Level vs Turnover')
plt.xlabel('Salary Level')
plt.ylabel('Proportion of Employees')
plt.savefig('Salary_bar_chart')
```

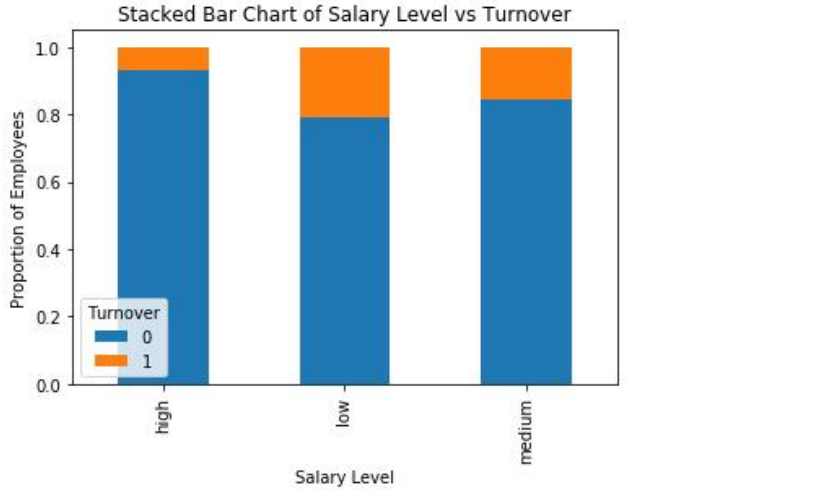


Figure 17 Turnover frequency of salary

From the above we can observe that employees have low salary tends to leave from the organizations. Additionally, the proportion of the employee turnover depends on their salary level; hence, the salary level can be a good predictor in predicting employee turnover.

Bar chart for employee experience and the frequency of turnover

```
%matplotlib inline
import seaborn as sns
import matplotlib.pyplot as plt
pd.crosstab(employeedata.experience,employeedata.Turnover).plot(kind='bar')
plt.title('Turnover Frequency for experience')
plt.xlabel('experience')
plt.ylabel('Frequency of Turnover')
plt.savefig('experience_bar_chart')
```

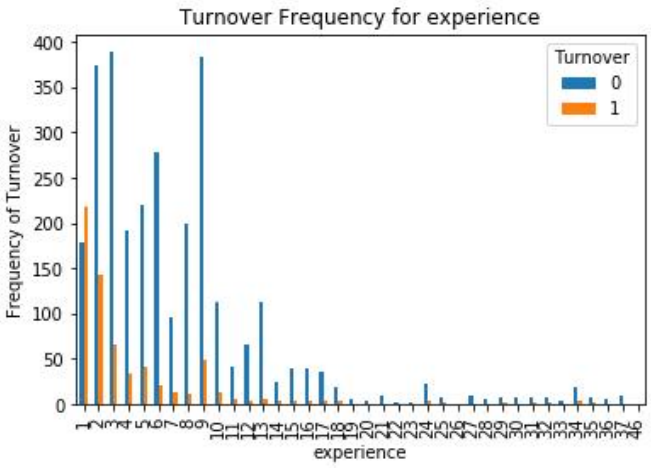


Figure 18 Turnover frequency of experience

From the experience bar chart we examine that the employees experience in the organization mostly between zero and eighteen years has stayed in the organization. From this employees who have between zero and three year of experience have highly left from the organization.Employee's have between four and nine years of experiences have moderately left within the organization. Additionally, employee's have more than ten year of experience have not left from the organization. This indicates that freshman employees were highly left from the organization, whereas, employee's have more experience have exist within the organization. Hence, employee's experience can be a good predictor in predicting employee turnover.

Bar chart for employee's departments and the frequency of turnover

```
%matplotlib inline
import matplotlib.pyplot as plt
pd.crosstab(employeedata.department,employeedata.Turnover).plot(kind='bar')
plt.title('Turnover Frequency for department')
plt.xlabel('department')
plt.ylabel('Frequency of Turnover')
plt.savefig('department_bar_chart')
```

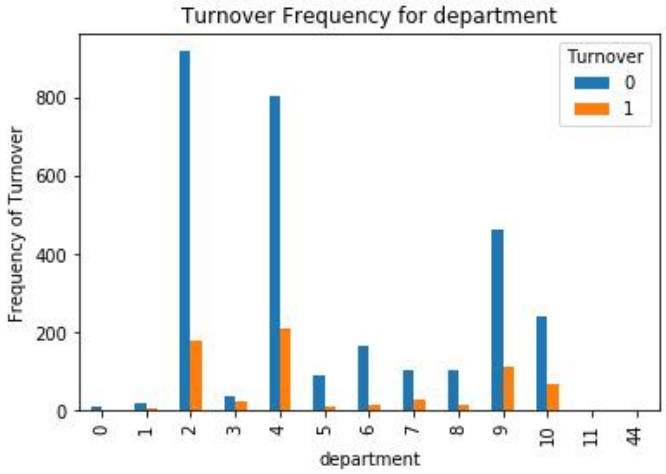


Figure 19Turnover frequency of department

From the above 1 indicates 'office president', '2 indicate that 'cort manager', 3 indicates 'registrar' and 4 indicates 'general service departments. Totally, the federal courts have ten departments, from this registrar and general service departments have a higher no of additional employees. These departments have higher employee turnover than the rest.The frequency of employee turnover depends on the department they work in. Thus, the department can be a good predictor of employee turnover.

Histograms chart for employee's profile and the frequency of turnover

Histograms are often one of the most helpful tools we can use for numeric variables during the exploratory phrase. Histograms arethe most helpful tools we can use for numeric variables during the visualization phrase.

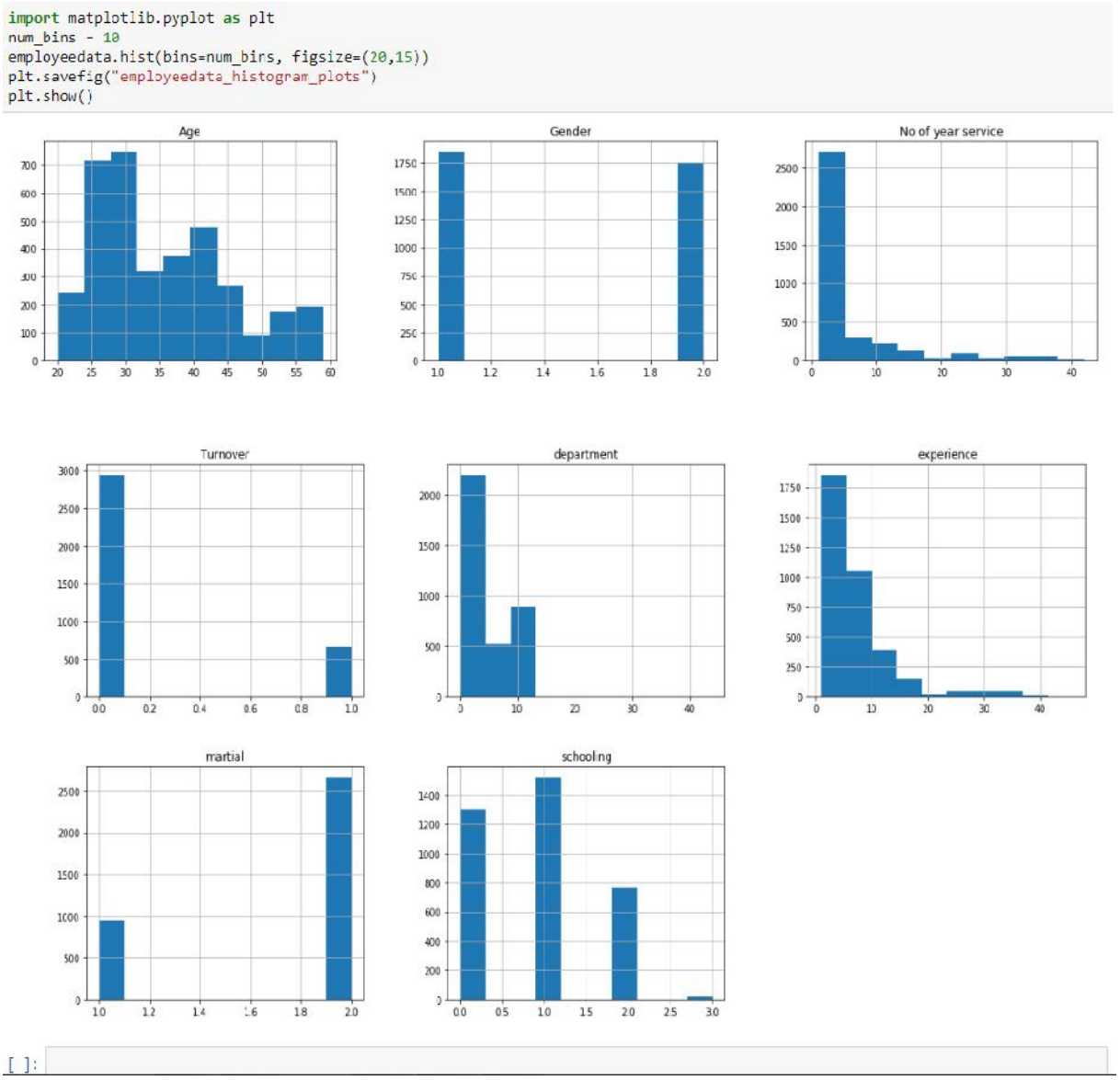


Figure 20Histogram chart for employee turnover frequency

5.4 Significant factors

In the performance analysis that the gradient boosting tree classifiers give the highest ac-curacy, precision, Roc-curve and f1-measure measures. So it is better to check the significant factor which helps an important role to predict the reason for employee turnover. The significant factors identified by using the feature importance, Feature importance helps to select the most significant factors which, cases of employee turnover in the organization and other institutions in the given feature of the data. AHigh score means the feature is more significant to the output variable. Feature importance is an inbuilt class, coming with Tree-Based Classifiers. For the study the researcher was apply random feature importance classifier on the given dataset to identify the columns in the dataset that has the most influence in determining the values of the column 'Turnover' which indicates whether the employee has left the organization (1) nor not (0).

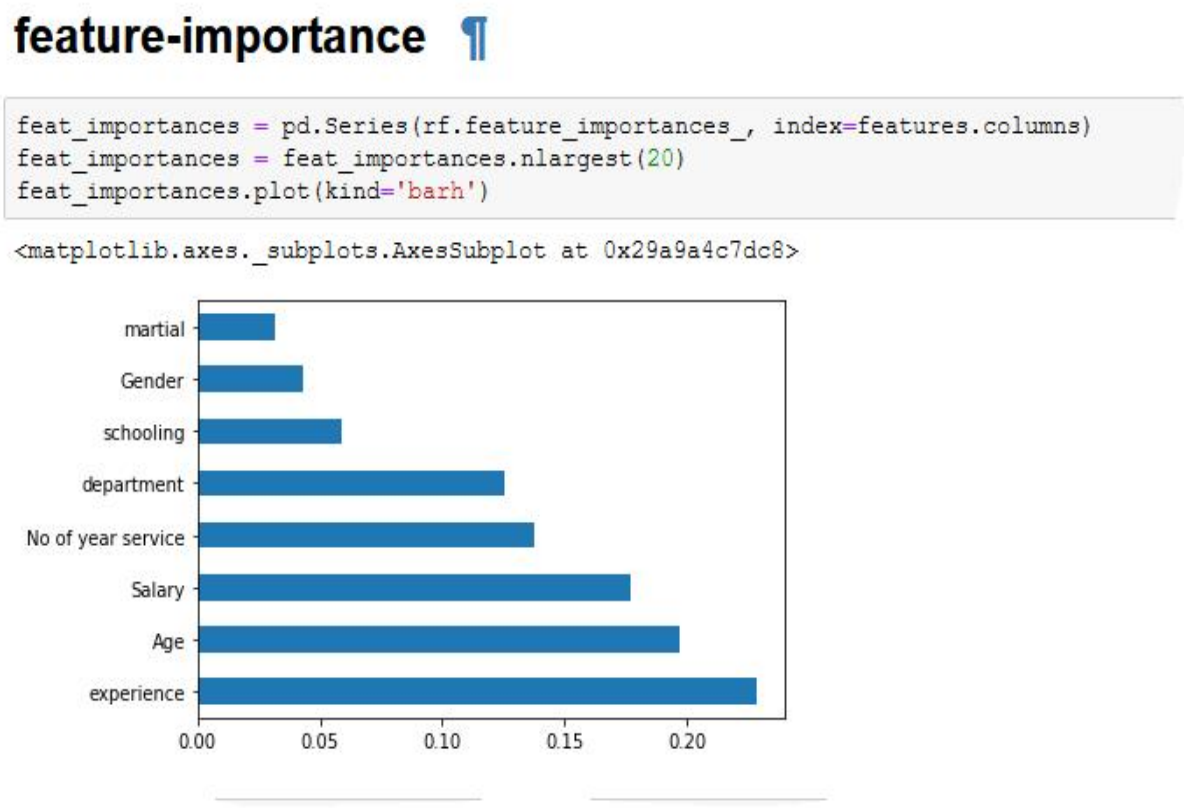


Figure 21 feature importance

The above figure,the importance’ of factors shows that experience is one of the most important factors in deciding the model accuracy for all three datasets. Age is the second-best for the given datasets salary is the third predictor variables in which have a high impact on model predictions. However, the marital status, gender and schooling have low status on the model accuracy. It indicates that experience, age, salary and no of year service are the most predictor variables on the given dataset.

CHAPTER SIX

EXPERMENTAL RESULT AND DISCUSSION

6.1 Evaluation of the results

These sections analyze the results obtained from the experiments done in the study. Also the performances of classifiers arediscussed in this section. The accuracy of the machine learningclassification techniques was measured through using the random -state, parameter tuning and the averaging accuracy of 10-fold Cross-validation dataset that supported by sklearn jupyter python tool.From the first experimentby using random-state the result shows that it affects on the classification model's performance .when we change the random-state values the classification models outperform different accuracy results. Fromthe five random-state experiments the result indicated thatthe random state numbers generate was increase the classification model accuracy was also increase. From those experiments in the fifth experiment gradient booting treeclassifier was bestoutperformed 87.5% when random-state=42.

The second experiment was using parameter-tuning by changing the parameters of the classifiers. The first classifier was logistic regression;from the above experiments firstly, run the algorithm by default and the accuracy result is 81%.secondly, thirdly and finally the researcher was trying to change the parameter of the classifier the accuracy was also the same to the previous one. The second classifier was random-forest; in this classifier the same to logistic regression change the parameters different times there is no change in the performance accuracy. The classifier also outperformed the accuracy of 81% the same to logistic regression.

The third classifier is gradient boosting tree, the same to from the previous classifier run the algorithm by default the accuracy result is 83%.secondly, by changing the parameter 81%.thirdly, the accuracy is better than the previous accuracy is 84.fourthly, the accuracy is also the best one from the previous outcome is 87%.finally, by selecting the best parameter that has the best accuracy, the accuracy is 87%. As we observe from the result by changing the parameters of the classifiers that have good accuracy we can get a good prediction outcome. From gradient boosting tree classifier parameters: n-estimators, learning rate and max depth are

affecting the outcome of the given classification models. Therefore, from the experimentsgradient boosting tree classifier is the good model among the three classification models.

The next experiment by using 10 fold cross-validation to check the predictive ability of classifier how well generalized on the given an independent datasets. From the cross-validation result the average accuracy close to the gradient tree model accuracy. Therefore, the gradient boosting tree model is generalizing well.

In the evaluation metrics result from the three prediction models the gradient boosting tree classifier was also best outperformedin terms of precision result of 72%.additionally, logistic regression classifier was best outperformed in terms of recall result of 66%.finally, the researcher try to measure the accuracy result by using Roc-Curve .as shown the roc curve result gradient boosting treewas best outperformed.

From the evaluation metrics result the result shows that there is class imbalance that means the terminated class is low. To solve this problem the researcher was apply SMOTE class imbalance methods. But; there is no see any change in the accuracy result from the three classification models. The results of the above experiments indicated that all of the three techniques had good and moderate accuracy, which is greater than 60%. In all three experiments, the dataset produced suitable models for each of the threeselected classification techniques.From the three classification models gradient boosting tree is to best outperform. As a result of the above, the gradient boosting tree classifier was the most appropriate classifier for thegiven datasets.
This agree to the finding of (Zhao et al., 2018).Tree-based classifiers best outperform due to its minimal data preprocessing, has decent predictive power, and, reliably.

From the feature importance figure result the study has found that several variables hadgreatly influenced the reason for employee turnover in federal court organizations .One of the significant factor variables that had the highest effect was the employee’s experience. The result indicated that fresh employees mostly, from zero to three years of experience were left from the organization. On the otherhand employee’s have above ten years of experience exist for a long

period of time in the organizations. The evidence shows that the existence a condition of the regularity of employees' treatment and supporting them with professional training in the human resource department and other concernedbodies is enhancing their performance and can minimize employee turnover.The second significant factor is age had positively affected employee's left from the organization.As shown in the results of feature importance demonstration. Where the age factor has an important role in the employees turnoverin the organization. The demonstration examined that young employee's have left the organization.The younger employees from age 25-40 were more likely to turnover than the olderemployee.But with the existence of different employee's treatments like promotion, training, wages and good payment system within the organization employees tends to stay a long period of time.

The other important factor was employees'-'salary. The result shows that employee's who have low salary tends to leave from the organization. Employee'swho have high salary exists in the organization for a long period. But, the fair salary payment system in the organization helps to reduce the employee's turnover.Last but not least, the result has shown that the employee's number of year service variableshad the other impact on the reason for employee's leave from the federal court organizations. As we see from the figureemployee's has low no of year service have left an employee'-s has more no of year service exist in the organization for a long period. From the result the above remedy in experience, age and salary hadapply for the variable no of year service increase the employees satisfaction to exist a long time in the organization. Some Personal Variables such as department and education hadslightly affected the employee turnover in federal court organization. The remaining factors martial and gender were low predictor variables on employee turnover in the organization.

6.2Research questions andevaluation Results

The first research question was:

1. Which machine learning algorithm is most appropriate for predicting employee turnover?

As we see in the experiments result from the three classification models gradient boosting tree was best out performed. The second questions were

2. To what extent would the machine learning approach be able to correctly predict staff turnover?

The answer to this question is gradient boosting tree was 87.5% of accuracyoutperformed.

3. What are the significant predictor variables for causes for employee turnover?

From the evaluation result the researcher found that the first predicator causes for employee turnover variable were employee's experience. Employees have low number of years experience were left from the organization, whereasemployees have high no of years experiences exist for long period of time within the organization. The second predicator variable was employees' age. Employees have between 25-40 ages young employees leave from the organization. Employees have greater than 40 year old employees less likely to leave from the organizations. The third significant factor variable was employee'ssalary. Employees have low payment tends to leave from the organization; on the other hand employees have high salary less likely to leave the organization. The last not the least employees' number of years service was the important predicator variable for the cause of employee turnover.Employees have between zero and three years of experience were high likely to leave from the organizations. Whereas, employees have more experience less likely to leave from the organizations.

CHAPTER SEVEN

CONCLUSION AND RECOMMANDATION

This chapter concludes the researchwork done in the above activities and give suggestions and clue on the future research.

7.1 Conclusion

Nowadays, employee turnover has become a serious issue in government organizations and other institutions. Employees leave from one organization to another for different reasons, It influences time, productivity and continuity of the organizations. Today, the emergence of artificial intelligence can helps to predict the future activities of employees by using employee's real dataset. One of the branches of artificial intelligence Applying machine learning techniques in the different problem domains in the HRM field is considered as an important and urgent issue. Especially, at the organization sector in Federal court, furthermore, increasing the academic research on machine learning in HR for reaching the organization sector with a high performance.Thus, thisstudy has set out to be a real-time application in the organizationswhere the management can predict the future actions of the employees based on the given employees data. The main objective of this paper was to build a model whichcan efficiently predictemployees that might turnover in future. Considering the realsituation, the management will be more interested to know the potential employeeswho might actually leave so that they can set their attention on retention mechanisms to preventemployee turnovers.Steps of Machine learning experiment methodology is followed throughout the research and Jupiter notebook pythonprogramming tool was used to implement the research.The research workflow is studied primarily by contacting the organization of Federal court. Anoverview of the current employee turnover and its effects has been studied. In the first phase, the employee data set of the organization is collected, and statistical analysisis done on the dataset to study the properties of the data. The data exploration visualization was also done to help to have a clear understanding.In the data preparation phase, the data has been cleaned to avoid the data errors, outliers, missing values….etc. The categorical variables have been encoded into numeric values,

and the data is identical. Then the pre-processed data is then used for modeling in the machine learning data modelingphase. The final phase was evaluating the machine learning classifiers and compared by using different evaluation metrics.All the three classifier performances were evaluated using 10 Fold cross-validation is used as the metrics to measure the performance of the classifiers and to avoid over fitting.After evaluation from the study it is found that the factors responsible for employee turnover are:-experience, salary, age and employees no of year service are the most significant factors. Some Personal Variables such as department and education had moderately affected the employees' turnover in federal court organization. The remaining factors martial and gender were low predicator variables on employee's turnover in the organization.The ensemble–based learning technique gradient boosting tree was found as the most suitable classifier for building the predictive model, where it had the greatestprediction accuracy through all the three experiments that had executed with the highest percentage 87.5%.

The researcher was facing different challenges in conducting this research. The first major challenge of the study wasthe unwillingness of the organization to give out their data for research purposes.Other constraints due to unstructured, noisy, missing, inconsistent data and limited employee dataset were takes much time to preprocess affect the accuracy of the given classifier model.Lack of machine learning research studies and the availability of sufficient related literature upon the Federal courtorganization and our country was another challenge during the study.

The result indicates that For HRM department, thisapplying machine learning techniques can be used in predictingthe performance of the potential talents that will be promoted, predicting the performance of the recentapplicant employees where various actions can be taken for avoiding any risk related to hiring employees with alow performance, enhancing development and training strategies, and so on. In addition, it helpshigher managements and HRM takes retention mechanisms for employees to continue within the organizations for a long period.

In general, recognitions, appraisals, encouragement, education, and growth of employees can go a long way to ensure employees join a particular organization. A government organization that has good wishes and has the interest of its employees at heart will keep its most experienced employee. Lastly, it can be finalized that employees who are happy with their work will not look the need to leave or move to anotherorganization.

7.2Future works and Recommendation

The researcher made the following recommendations based on the findings of this study:

- The first recommendation is future research work on this area may focus on containing higher number of employees and include more significant essential factors and features; toobtain the best accuracy for the predictive classification model. This gives contribution to accurately predict employee turnover for anyorganization in future research.
- Secondly, the accessible factors to understand the employee turnover reasonswere small. The factors likerace, religionand other remaining factors will includecould be helpful to know the turnover phenomena further. More data will help to do a more precise analysis and purify the prediction model. Additionally, future research should be conducted by including other prediction supervised machine learning models such as:-support vector, extreme gradient boosting tree, neural network etc models can also be tested to compare the performance the classifier and helps to identify the best-outperformed accuracy classifier on the given dataset.
- Thirdly, the dataset used in this research is imbalanced. It is typically known that classifierstrained on imbalanced datasets create partiality in prediction and over fitting in the test set. While parameter tuning, feature selection methods and SMOTE used for this research.It is very important that other class imbalance handling methods, like randomoversampling and random under sampling, should be applied to balancethe dataset and observe if this mightget better the prediction accuracy of the classifiers.
- Fourthly, thefederal court information directorate keeps employees profile ina different arrangement such as in hard-ware, soft ware format, MS-word, and MS-excel, there is inconsistency in recording employees,which are being difficult to datapreprocessand time-consuming. So itwould be important to have kept the employees profile in software format by using technology tools to keep the data in a distributed system; improves data integrity and continuity a long period.Today's withthe development of technology,data is more important to solve different problems; so it is important to keep the data in such manner .moreover, it would be representative of the federal court organization. It is recommended that the HR department shouldhave to build a well-recorded mechanism of employees profile data which are critical.

- Fifthly, the study findings reported in this study show an important role to create awareness ofthe significant factorsthe reason foremployee's turnover. The study also contribute to the federal court organizationi.e. Human Resource Management and the management bodies at a different level to be responsible of the determinant factors of employeeturnover intention and allow them to work togetherset certain essential mechanism andretention plan to minimize turnover and maximize organizational commitment through keeping the professional and more skilled employees to achieve better to theintended objectives and vision of the organization. Since the Human resource is heart to every function of the organization, the management of the organization should give consideration and develop retention mechanisms. Ifthe organization has well-developed growth a strategy that includes a variety ofdevelopment opportunities such as job planning, encouragement, training, education, transparency, integrity and personaldevelopment. The management should plan a jobexpansion program foremployees.
- Finally, for this study, we use only federal court organizations found in Addis Ababa region.However, furtherinvestigations needed by including other regional federal courts andgovernmentorganizations to comprehensively to other determinant factors ofemployees turnover issue can be also a future research area.

Reference

Abdali, F. (2011). Impact of Employee Turnover on Sustainable Growth of Organization in Computer Graphics Sector of Karachi, Pakistan. *Afro Asian Journal of Social Sciences*, *2*(2.4), 1–27.

Adjei, A. (2014). *Evaluating the factors that contribute to employee turnover at Toase medical centre*.

Agyeman, C. M., & Ponniah, V. M. (2014). Employee demographic characteristics and their effects on turnover and retention in MSMEs. *International Journal of Recent Advances in Organizational Behaviour and Decision Sciences*, *1*(1), 12–29.

Ajit, P. (2016). Prediction of employee turnover in organizations using machine learning algorithms. *Algorithms*, *4*(5), C5.

Alamsyah, A., & Salma, N. (2018). A Comparative Study of Employee Churn Prediction Model. *2018 4th International Conference on Science and Technology (ICST)*, 1–4.

Alao, D., & Adeyemo, A. B. (2013). Analyzing employee attrition using decision tree algorithms. *Computing, Information Systems, Development Informatics and Allied Research Journal*, *4*.

Alpaydin, E. (2009). *Introduction to machine learning*. MIT press.

An_Assessment_of_the_Causes_of_Employee.docx. (n.d.).

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79.

Asegid, B. (2018). *Employee Turnover and Organization Performance: The Case of Shintes ETB Garment PLC*. Addis Ababa University.

Breiman, L. (1997). *Arcing the edge*. Technical Report 486, Statistics Department, University of California at ….

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Chen, X., & Jeong, J. C. (2007). Enhanced recursive feature elimination. *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, 429–435.

Cotton, J. L., & Tuttle, J. M. (1986). Employee turnover: A meta-analysis and review with implications for research. *Academy of Management Review*, *11*(1), 55–70.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, *29*(2–3), 103–130.

Drucker, H., & Cortes, C. (1996). Boosting decision trees. *Advances in Neural Information Processing Systems*, 479–485.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Icml*, *96*, 148–156.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.

G, Y. (2017, September 7). *The 7 Steps of Machine Learning*. Medium. https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e

Gao, Y. (2017). *Using decision tree to analyze the turnover of employees*.

Gberevbie, D. E. (2008). *Staff recruitment, retention strategies and performance of selected public and private organizations in Nigeria*. Covenant University, Ota.

Gemechu, B. (2017). *Factors Influenecing Employee Intention To Turnover At Commercial Bank Of Ethiopia*. Addis Ababa University.

Glu, Z. Ö. K. (2014). Employee turnover prediction using machine learning based methods. *Doctoral Dissertation*.

Hormozi, H., Hormozi, E., & Nohooji, H. R. (2012). The classification of the applicable machine learning methods in robot manipulators. *International Journal of Machine Learning and Computing*, *2*(5), 560.

Hossain, M. (2019). *A Study on Features of Significance & Prediction Model Building*.

Iqbal, A. (2010). Employee turnover: Causes, consequences and retention strategies in the Saudi organizations. *The Business Review, Cambridge*, *16*(2), 275–281.

Izeboud, E. (2017). *Prediction of job transition using publicly available professional profiles*. Tilburg University.

Jose, J. (n.d.). *Predicting Customer Retention of an App-Based Business Using Supervised Machine Learning*. 85.

Jose, J. (2019). *Predicting Customer Retention of an App-Based Business Using Supervised Machine Learning*.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, *14*, 1137–1145.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, *160*, 3–24.

Mamuye, N. (2018). Statistical assessment of employee's turnover and its causes; In the case of Moret and Jiru Wereda, North Shoa, Amhara, Ethiopia. *American Journal of Theoretical and Applied Statistics*, *7*(4), 139–146.

Mandal, I., & Sairam, N. (2012). Accurate prediction of coronary artery disease using reliable diagnosis system. *Journal of Medical Systems*, *36*(5), 3353–3373.

Maria Navin, J. R., & Pankaja, R. (2016). *Performance Analysis of Text Classification Algorithms using Confusion Matrix*.

Mitchell, T. R., Holtom, B. C., & Lee, T. W. (2001). How to keep your best employees: Developing an effective retention policy. *Academy of Management Perspectives*, *15*(4), 96–108.

Mohammed, M., Khan, M. B., & Bashier, E. B. M. (2016). *Machine learning: Algorithms and applications*. Crc Press.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.

Nasr, M., Shaaban, E., & Samir, A. (2019). A proposed Model for Predicting Employees' Performance Using Data Mining Techniques: Egyptian Case Study. *International Journal of Computer Science and Information Security (IJCSIS)*, *17*(1).

Newsom, I. (2015). *Data Analysis II: Logistic Regression*.

Odiro, T. T. (2017). Assessment of Employee Turnover and Its Impact on Three Selected Government TVET Colleges in Addis Ababa. *Assessment*, *29*.

Ongori, H. (n.d.). *A review of the literature on employee turnover*. 6.

Ongori, H. (2007). *A review of the literature on employee turnover*.

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: Classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*(3), 128–138.

Pant, A. (2019, January 23). *Workflow of a Machine Learning Project*. Medium. https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Pojon, M. (2017). *Using machine learning to predict student performance*.

Praveena, M., & Jaiganesh, V. (2017). A literature review on supervised machine learning algorithms and boosting process. *International Journal of Computer Applications*, *169*(8), 32–35.

Punnoose, R., & Ajit, P. (2016). Prediction of Employee Turnover in Organizations using Machine Learning Algorithms. *International Journal of Advanced Research in Artificial Intelligence*, *5*(9). https://doi.org/10.14569/IJARAI.2016.050904

Rabbi, F. (2019). *A review of the recent trends in the use of machine learning in business*.

Raut, P. P., Borkar, N. R., & Student, M. E. (2017). Machine Learning Algorithms: Trends, Perspectives and Prospects. *International Journal of Engineering Science*, *4884*.

Rehman, M. S. (2012). Employee turnover and retention strategies: An empirical study of public sector organizations of Pakistan. *Global Journal of Management and Business Research*, *12*(1), 1–8.

Sabbeh, S. F. (2018). Machine-learning techniques for customer retention: A comparative study. *(IJACSA) International Journal of Advanced Computer Science and Applications*, *9*(2).

Salunkhe, T. P. (2018). *Improving employee retention by predicting employee attrition using machine learning techniques*. Dublin Business School.

Samuel, M. O., & Chipunza, C. (2009). Employee retention and turnover: Using motivational variables as a panacea. *African Journal of Business Management*, *3*(9), 410–415.

Schlechter, A. F., Syce, C., & Bussin, M. (2016). Predicting voluntary turnover in employees using demographic characteristics: A South African case study. *Acta Commercii*, *16*(1), 1–10.

Si, S., Zhang, H., Keerthi, S. S., Mahajan, D., Dhillon, I. S., & Hsieh, C.-J. (2017). Gradient boosted decision trees for high dimensional sparse output. *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3182–3190.

Soofi, A. A., & Awan, A. (2017). Classification techniques in machine learning: Applications and issues. *Journal of Basic and Applied Sciences*, *13*, 459–465.

Stovel, M., & Bontis, N. (2002). Voluntary turnover: Knowledge management–friend or foe? *Journal of Intellectual Capital*, *3*(3), 303–322.

Tomassen, M. (n.d.). *Exploring the Black Box of Machine Learning in Human Resource Management*. 52.

Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, *36*(3), 6527–6535.

Yousaf, H. M. N., & Bhulai, S. (2016). *Analysing which factors are of influence in predicting the employee turnover*.

Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018). Employee turnover prediction with machine learning: A reliable approach. *Proceedings of SAI Intelligent Systems Conference*, 737–758.

Zhu, X. (n.d.). *Forecasting Employee Turnover in Large Organizations*. 126.

List of Appendixes

Appendix A

The Jupiter python Source code for data preprocessing and exploration

```
import pandas as pd #used for tabular data
import numpy as np # used for numeric computation
employeedata=pd.read_csv('C:/Users/biruk/Desktop/thesis python code/thesis pythoncode2/MERGE COURT FILES - Copy.csv') # used to read the file
employeedata.head(4)#method to view the first few records of the data set
employeedata.shape(3607, 9)# to know the size of the datasets
col_names = employeedata.columns.tolist()    #to list the columns
print("Column names:")
print(col_names)

employeedata.isnull().any() #to know the missing and clean data sets

employeedata.dtypes # to know the type of the datasets

employeedata['schooling'].value_counts()  # used to count the number of employees includein education variables

employeedata['Turnover'].value_counts() # used to know the number of employees stay and left

employeedata['Turnover'].value_counts()/len(employeedata)*100 # to know the percentage of employee who stay and leave

employeedata.groupby('Turnover').mean() # to know the mean average of employees turnover in the given employee variables.
```

Appendix B

Source code for Data visualization phase

I. Bar chart turnover frequency for gender variable

```
%matplotlib inline

import matplotlib.pyplot as plt

pd.crosstab(employeedata.Gender,employeedata.Turnover).plot(kind='bar')

plt.title('Turnover Frequency for Gender')

plt.xlabel('Gender')

plt.ylabel('Frequency of Turnover')

plt.savefig('Gender_bar_chart')
```

II. Bar chart turnover frequeny for education variable

```
%matplotlib inline
import matplotlib.pyplot as plt
pd.crosstab(employeedata.schooling,employeedata.Turnover).plot(kind='bar')
plt.title('Turnover Frequency for schooling')
plt.xlabel('schooling')
plt.ylabel('Frequency of Turnover')
plt.savefig('schooling_bar_chart')
```

III. Bar chart Turnover frequency for department variable

```
%matplotlib inline
import matplotlib.pyplot as plt
pd.crosstab(employeedata.department,employeedata.Turnover).plot(kind='bar')
plt.title('Turnover Frequency for department')
plt.xlabel('department')
plt.ylabel('Frequency of Turnover')
```

```
plt.savefig('department_bar_chart')
```

IV. Bar chart Turnover frequency for salay variables

```
table=pd.crosstab(employeedata.Salary, employeedata.Turnover)
table.div(table.sum(1).astype(float), axis=0).plot(kind='bar', stacked=True)
plt.title('Stacked Bar Chart of Salary Level vs Turnover')
plt.xlabel('Salary Level')
plt.ylabel('Proportion of Employees')
plt.savefig('Salary_bar_chart')
```

V. Bar chart turnover frequency for martial status variables

```
%matplotlib inline
import matplotlib.pyplot as plt
pd.crosstab(employeedata.martial,employeedata.Turnover).plot(kind='bar')
plt.title('Turnover Frequency for martial')
plt.xlabel('martial')
plt.ylabel('Frequency of Turnover')
plt.savefig('martial_bar_chart')
```

VI. Histogram chart turnover frequency for all employee dataset variables

```
import matplotlib.pyplot as plt
num_bins = 10
employeedata.hist(bins=num_bins, figsize=(20,15))
plt.savefig("employeedata_histogram_plots")
plt.show()
```

Appendix C

Source code for Model building path

```
target = employeedata['Turnover']
features = employeedata.drop('Turnover', axis = 1) # to know the the class and the colums/features
import sklearn
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.3,random_state=10) # split the datasets into train and test sets
len(X_train) # to know the size of train set
len(y_test) # to know the size of test set
performance = pd.DataFrame({'Actual':y_test,'modelpred':modelpred}) # to know the actual and the pridicted outcome
from sklearn.feature_selection import RFE
rfe = RFE(model,6)
rfe = rfe.fit(X_train,y_train)
print(rfe.support_)
print(rfe.ranking_) # it shows how to select the best attribute by using recursive feature elimination
```

a) **# Logistic regression model**

```
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
model= LogisticRegression()
model.fit(X_train,y_train) # to import the accuracy metrics
from sklearn.metrics import accuracy_score
accuracy_score(y_test,modelpred )# to know the accuracy result of logistic regression model
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
```

model= LogisticRegression(random_state=1, solver='newton-cg', multi_class='ovr',class_weight='balanced',max_iter=1000)

model.fit(X_train,y_train) # to know the accuracy result by using parametre tuning

b) # **Random forest model**

from sklearn.ensemble import RandomForestClassifier

rf=RandomForestClassifier()

rf.fit(X_train,y_train)

accuracy_score(y_test,modelpred)

c) # **Gradient boosting tree model**

from sklearn.ensemble import GradientBoostingClassifier

gbc = GradientBoostingClassifier()

gbc.fit(X_train, y_train)

from sklearn.ensemble import GradientBoostingClassifier

gbc = GradientBoostingClassifier(learning_rate=0.05, max_depth=2,max_features=2, min_samples_split=2,n_estimators=20,)

gbc.fit(X_train, y_train)

from sklearn.ensemble import GradientBoostingClassifier

gbc = GradientBoostingClassifier(learning_rate=0.04)

gbc.fit(X_train, y_train)

Appendix D

Source code for Cross validation

1. Cross validation for random forest

```
from sklearn import model_selection
from sklearn.model_selection import cross_val_score
kfold = model_selection.KFold(n_splits=10, random_state=7)
modelCV = RandomForestClassifier()
scoring = 'accuracy'
```

```
results = model_selection.cross_val_score(modelCV, X_train, y_train, cv=kfold, scoring=scoring)

print("10-fold cross validation average accuracy: %.3f" % (results.mean()))
```

2. cross validation for logistic regression

```
kfold = model_selection.KFold(n_splits=10, random_state=7)

        modelCV = LogisticRegression()
        scoring = 'accuracy'
        results = model_selection.cross_val_score(modelCV, X_train, y_train, cv=kfold,
        scoring=scoring)
        print("10-fold cross validation average accuracy: %.3f" % (results.mean()))
```

3. Cross validation for gradient boosting tree

```
        kfold = model_selection.KFold(n_splits=10, random_state=7)
        modelCV = GradientBoostingClassifier()
        scoring = 'accuracy'
        results = model_selection.cross_val_score(modelCV, X_train, y_train, cv=kfold,
        scoring=scoring)
        print("10-fold cross validation average accuracy: %.3f" % (results.mean()))
```

Appendix E

Source code for confusion matrix, recall&precision

A. Confusion matrix for random forest

```
from sklearn.metrics import classification_report

print(classification_report(y_test, rf.predict(X_test)))

import matplotlib.pyplot as plt

y_pred = rf.predict(X_test)

from sklearn.metrics import confusion_matrix

import seaborn as sns

forest_cm = metrics.confusion_matrix(y_pred, y_test, [1,0])

sns.heatmap(forest_cm, annot=True, fmt='.2f',xticklabels = ["Left", "Stayed"] , yticklabels = ["Left",
"Stayed"] )
```

```
plt.ylabel('True class')
plt.xlabel('Predicted class')
plt.title('Random Forest')
plt.savefig('random_forest')
```

B. # confusion matrix for logistic regression
print(classification_report(y_test, model.predict(X_test)))
y_pred = model.predict(X_test)
from sklearn.metrics import confusion_matrix
import seaborn as sns
forest_cm = metrics.confusion_matrix(y_pred, y_test, [1,0])
sns.heatmap(forest_cm, annot=True, fmt='.2f',xticklabels = ["Left", "Stayed"] , yticklabels = ["Left", "Stayed"])
plt.ylabel('True class')
plt.xlabel('Predicted class')
plt.title('logistic regression')
plt.savefig('logistic regression')

C. # confusion matrix for gradient boosting tree
print(classification_report(y_test, gbc.predict(X_test)))
gbc_y_pred = gbc.predict(X_test)
gbc_cm = metrics.confusion_matrix(gbc_y_pred, y_test, [1,0])
sns.heatmap(gbc_cm, annot=True, fmt='.2f',xticklabels = ["Left", "Stayed"] , yticklabels = ["Left", "Stayed"])
plt.ylabel('True class')
plt.xlabel('Predicted class')
plt.title('gradint boosting tree')
plt.savefig('gradint boosting tree')

Appendix F

Source code for Roc-Curve metrics

from sklearn.metrics import roc_auc_score

```
from sklearn.metrics import roc_curve
gbc_roc_auc = roc_auc_score(y_test, gbc.predict(X_test))
fpr, tpr, thresholds = roc_curve(y_test, gbc.predict_proba(X_test)[:,1])
rf_roc_auc = roc_auc_score(y_test, rf.predict(X_test))
rf_fpr, rf_tpr, rf_thresholds = roc_curve(y_test, rf.predict_proba(X_test)[:,1])
model_roc_auc = roc_auc_score(y_test, model.predict(X_test))
fpr, tpr, thresholds = roc_curve(y_test, model.predict_proba(X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='gradient boosting tree (area = %0.2f)' % gbc_roc_auc)
plt.plot(rf_fpr, rf_tpr, label='Random Forest (area = %0.2f)' % rf_roc_auc)
plt.plot(fpr, tpr, label='logistic regression (area = %0.2f)' % model_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc=''lower right'')
plt.savefig('ROC')
plt.show()
```

Appendix G

Source code for feature importance

```
feat_importances = pd.Series(rf.feature_importances_, index=features.columns)
feat_importances = feat_importances.nlargest(20)
feat_importances.plot(kind='barh')
```