BINDURA UNIVERSITY OF SCIENCE EDUCATION

FACULTY OF SCIENCE AND ENGINEERING

COMPUTER SCIENCE DEPARTMENT

PREDICTION OF EMPLOYEE ATTRITION USING ENSEMBLE MODEL BASED ON MACHINE LEARNING ALGORITHMS

By

TERRENCE ZAKWA

REG NUMBER: B1851379

SUPERVISOR: MR NDUMIYANA

A RESEARCH PROJECT SUBMITTED TO THE COMPUTER SCIENCE DEPARTMENT IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE BACHELOR OF SCIENCE HONOURS DEGREE IN COMPUTER SCIENCE.

Approval Form

The undersigned certify that they have supervised the student Terrence Zakwa in the research dissertation entitled, "**Prediction Of Employee Attrition Using Ensemble Model Based On Machine Learning Algorithms**" submitted in partial fulfilment of the requirements for a Bachelor of Science Honors Degree in Computer Science at Bindura University of Science Education.

STUDENT: **DATE:**

…… **……...**

SUPERVISOR: **DATE:**

…… **...…….**

CHAIRPERSON: **DATE:**

……..…… **...…….**

EXTERNAL EXAMINER: **DATE:**

……..…… **...…….**

Abstract

Employee attrition denotes the nonstop drop in the number of workers in an organization by the process of withdrawal, abdication, or death (Dutta et al,. 2020). Employee attrition is expressed as the normal process by which the employees leave the organization due to some reasons, such as the resignation of employees. There are many factors that can cause employee attrition (Peng, 2021). The employees leave the organization faster than they are hired. When the employee leaves the organization, the vacancies remain unfilled, resulting in a loss for the organization. The first objective is to analyse different machine learning algorithms used in HR datasets to predict attrition. The second objective is to design and implement a machine learning model based on Ensemble Modelling which predicts employee attrition using artificial neural network. The final and last objective is to evaluate the effectiveness of the Ensemble Modelling method and machine learning in employee attrition prediction.Therefore, to this end, the researcher managed to develop a system or model that uses ensemble modelling for various ML algorithms which are Logistic Regression, SVM, Linear SVC, KNN, Naive Bayes, Decision Tree, Gradient Boosting Trees and Random Forest algorithm to predict which employee is likely to leave using a dataset of an unknown organization found on Kaggle. All the machine learning algorithms were combined using ensemble modelling and the results were satisfying. The employee attrition rate helps to understand the progress level of an organization. The researcher performed all the necessary black, white box tests and performance tests using the confusion matrix, the author found that the system had satisfactory performance. The system was tested in accuracy, misclassification error/error rate and it achieved 91.2% and 0.8% respectively. The model attained an overall precision of 86% and a sensitivity or recall of 84%. An F1 score of 91.3% was achieved with a specificity or true negative rate of 96%.

Key words: deep learning, ensemble modelling, employee attrition

Acknowledgements

I would like to extend my gratitude and sincere thanks to my supervisor Mr Ndumiyana for his constant motivation and support during the course of my work. I truly appreciate and value his esteemed guidance and encouragement from the beginning. I also want to thank Mr Matombo, Mr. Chaka for their co-supervision, I really appreciate all the time and efforts they put with the bid of helping me to come up with a quality research. Furthermore, I would like to mention my father, who made this all possible and also played a supporting role which contributed positively to my welfare.

Contents

Chapter 1: Problem Identification

1.1 Introduction

Employee attrition denotes the nonstop drop in the number of workers in an organization by the process of withdrawal, abdication, or death (Dutta et al,. 2020). Employee attrition is expressed as the normal process by which the employees leave the organization due to some reasons, such as the resignation of employees. There are many factors that can cause employee attrition (Peng, 2021). The employees leave the organization faster than they are hired. When the employee leaves the organization, the vacancies remain unfilled, resulting in a loss for the organization. The employee attrition rate helps to understand the progress level of an organization. The high attrition rate shows that the employees are frequently leaving. The results of the high attrition rate are the loss of organizational benefits (CSAMCS 2021). In order to keep the organization in progress, the attrition rate must be controlled.

Many types of employee attrition help us to understand the attrition process. The attrition type is whether an employee chooses to leave the company voluntarily. The involuntary attrition type is when the organization ends the employment process. The external attrition type is referred to when an employee leaves an organization to work for another organization. Internal attrition occurs when an employee is given another position within the same organization as a promotion. The employee attrition rate is the measure of people who leaves the organization. By measuring the attrition rate, we can identify the causes and factors that need to be solved to eliminate employee attrition. The attrition rate is calculated by dividing the number of employees who have left the company by the average number of employees over some time. The attrition rate helps us find the company's progress over a specific period. Therefore, using machine learning in predicting employees who are likely to leave/quit is one of the best methods that helps companies in taking proactive measures in reducing the attrition rate.

1.2 Background Of The Study

According to Peng(2021), after six months of job duration, 1/3 of new employees leave the organization. The 3 to 4.5 million employees leave their job every month in the United States, according to the Job Openings and Labor Turnover Survey (JOLTS) (CSAMCS,2020). The employee attrition rate is 57.3% in 2021 to the report of the Bureau of Labor Statistics . The report

also suggests that in many industries, the employee attrition rate is close to 19% (CSAMCS,2020). The cost per hire of new employees is USD 4129 by SHRM [4]. Ninety percent of employee retention rate is considered suitable for a company, and the attrition rate must be less than 10%.

Employee attrition has a positive impact on making an organisation successful (Punnoose and Ajit, 2016). There are a lot of effects on attrition that might have an impact on a company. These effects include the fact that worker turnover is expensive, the employees who quit on their want to be replaced. It is time-consuming to hire new workers and this quality time is supposed to be distributed to other important tasks. It is not very easy for a new employee to start working at the pace of others(Norrman, 2020).

Investments in training the new worker isn't always an easy task and productiveness may be affected. Experienced people convey understanding and experience which contributes to productiveness .(Alduayj and Rajpoot, 2019). Thirdly, the employee morale of the remaining employees gets affected negatively. The reason for that is when employees leave a specific group of employees need to cover the gap and do the work that was supposed to be done by those that left the company. The range of exertions will increase if worker turnover will increase, this indicates a mild growth in workload that could lower motivation and morale. Employee attrition also affects the profit of the company, the effects of the above have an impact on organizational overall performance which also affects the capability to perform at the preferred level. Expertise loss and decreased productiveness moreover affect the profit. The above effects on attrition are the reasons why companies look for strategies to lower attrition. The capability of expecting worker turnover should upload the tool to the toolbox even as constructing such strategies(Punnoose and Ajit, 2016). These techniques have to broaden the possibilities of stopping worker turnover and restrict the threat of having all of the terrible outcomes that have been said earlier.

1.3 Statement Of The Problem

It is difficult to predict attrition and often introduces noticeable voids in an organization’s skilled workforce (Alao & Adeyemo,2013). Service firms recognize that the timely delivery of their services can become compromised, overall firm productivity can decrease significantly and, consequently, customer loyalty can decline when employees leave unexpectedly (Sexton et al,2005). As a result, it is imperative that organizations formulate proper recruitment, acquisition

and retention strategies and implement effective mechanisms to prevent and diminish employee turnover, while understanding its underlying, root causes (Al-Radaideh & AlNagi, 2012), Chang,2009).Basing on this problem, the author seeks to develop a machine learning based attrition system.

1.4 Research Aim

The essential goal of this dissertation is to develop a framework that could help in predicting whether or not an employee will stay in a company using machine learning methods.

1.5 Research Objectives

- To design a machine learning model based on Ensemble Modelling which predicts employee attrition using artificial neural network
- To implement a machine learning model based on Ensemble Modelling which predicts employee attrition using artificial neural network
- Evaluate the effectiveness of the Ensemble Modelling method and machine learning in employee attrition prediction.

1.6 Research Questions

- What are the different machine learning algorithms used in HR datasets to predict attrition?
- How to design and implement a machine learning model based on Ensemble Modelling which predicts employee attrition using artificial neural network?
- Is the use of Ensemble Modelling method and machine learning in employee attrition prediction effective?

1.7 Research Propositions/Hypothesis

- H_0: The system will be able to predict employee attrition.
- H_1: The system will fail to predict employee attrition.

1.8 Justification/Significance Of The Study

This research contributes to knowledge by managing to use machine learning algorithms for human resource datasets and being able to predict the future attrition of employees and it's a way to get deep information about the overall performance of an organisation. Moreover, the research

also strengthens some records on how the companies recognise that employee attrition affects general performance and company competitive advantage.

1.9 Limitations/challenges

- Time needed to carry out the research is limited
- The most crucial issue is measuring the impact of worker appraisal and satisfaction to the employer which permits the organisation to lessen the attrition rate of personnel and to do the financial improvement by reducing their human resource rate in the company.

1.10 Scope/Delimitation Of The Research

This research will only look at the attrition of employees no other issues that relate to employees for example salary hikes, motivation, sick leave, and many more. This research is only limited to the machine learning models but there are other models that can be used like the deep learning models. It takes time to be able to use all of the models.

1.11 Definition Of Terms

Employee- a person employed for wages or salary, especially at non-executive level.

Employee Attrition- is the departure of employees from the organization for any reason (voluntary or involuntary), including resignation, termination, death or retirement.

Organisation- an organized group of people with a particular purpose, such as a business or government department.

Ensemble Modelling- is the process of running two or more related but different analytical models and then synthesizing the results into a single score or spread in order to improve the accuracy of predictive analytics and data mining applications

Machine learning- the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyse and draw inferences from patterns in data.

Chapter 2: Literature Review

A literature review, according to (Puebo, 2020), is a scientific study prepared from published sources that summarizes current understanding on a given issue. In this chapter, the researcher concentrates on answering the research questions and reveals previous and current systems that are similar to the research project at hand that have been done by other authors. This will be extremely valuable to the author because it will serve as a guide to identifying solutions, strategies, and techniques utilized by prior writers to solve earlier research problems. It is a tool that informs the researcher if the study proposal is possible based on the findings of previous researchers in that field. This chapter, in accordance with the definition of a literature review, presents all of the related work for this research. We are going to describe the use of machine learning in human resource and a little description of what feature selection is according to other researchers.

2.1 Human resource management

All businesses would desire to have a competitive advantage over rivals. The most crucial way to acquire a competitive advantage is through effective management of human resources (Human resource management, HRM). Employee contentment and happiness have an impact on their performance (Lee, 2007). Effective human resource management will improve employee morale, which will boost output and the competitive edge of the business (Brockett et al., 2019).

2.1.1 Employee attrition/ staff turnover

The goal of HRM is to maintain employee satisfaction. According to a study by Ryan, Schmit, and Johnson from 1996, employee turnover and satisfaction are related. High levels of satisfaction reduce staff turnover (Davis, 2013). Numerous empirical studies have demonstrated that employee turnover can and frequently does have an impact on organizational effectiveness, which affects competitive advantage. The following are some ways that employee turnover affects organizational effectiveness: (1) whilst a worker leaves their tacit and specific information is misplaced which decreases the general information inside that particular agency, the productiveness of the agency might be affected negatively, (2) expanded workload on different employees, (3) harm enterprise morale, (4) decrease turnover prices suggest that there may be

much less hiring with a purpose to decrease prices which includes activities for training and employment process (Ojakaa, Olango and Jarvis, 2014).

Worker turnover has both immediate and long-term effects. It's crucial to remember them both if you want to fully understand the true effects and how employee turnover affects competitive advantage. (Vasa and Masrani, 2019) contend that while some employee rotations may be advantageous for a company, in the majority of cases, turnover can be very expensive and can interrupt workflow. We will turn our attention to the American fast-food industry as an illustration of how expensive worker churn works. Recognizing the causes of the phenomena in the first place is incredibly important if you want to be able to reduce worker turnover. This is due to the fact that knowing the causes of the issue presents agencies with the opportunity to act.

2.1.2 Retention controls

Retention control, a component of HRM, is the study of how businesses should operate to keep their employees for the longest possible period of time. Companies must recognise and assess the turnover situation in order to be able to acquire effective retention control. The information provided regarding the turnover situation is insufficient. Research has shown that companies with retention techniques in the area will increase the likelihood of maintaining their personnel and that maintaining personnel will create an elegance that may appeal to new personnel as well. In order for businesses to take action, they also need a strategic retention plan with concrete steps to adopt. Furthermore, James & Mathew (2012) argue that the business enterprise must actively work to maintain their quality personnel and that a failure in that regard will cause their quality appearing personnel can be lost. This is why retention control has developed to a vital part of HRM and why the effects ought to have a sort of huge effect on competitive advantage.

2.2 Machine learning in HRM

The time it has taken for machine learning to enter the realm of human resources has been long and it hasn't been used until just a few years back. According to LinkedIn, only 22% of organizations have initiated analytics into their human resource departments and it is uncertain how well these analytics are implemented. Fields such as marketing have rich access to big data, take an example the sales of a product in a country. There will be data such as how many products have been sold when they have been sold, and how many observations the product has gotten. With those conditions, ML is a very effective tool to use to analyse big sets of data.(Tambe,

Cappelli and Yakubovich, 2019) The use of machine learning in HR presents several problems and challenges that have to be accounted for. First of all, the data stored in human resources are very small in contrast to what data-sets used in data science are, and often data is not stored at all(Tambe, Cappelli and Yakubovich, 2019). Even a big organization with 10000 samples is nowhere close to containing as much HR data as product sales data, for example. Secondly, the decisions within human resource management could get big. For example, a decision of whom gets fired and who gets hired. Thirdly, measuring the performance of individuals is often hard due to those complex roles depending on other roles - teamwork - and individual performance can be hard to break down from group performance. Fourth, deciding whether or not a person should be hired is not only based on tangible qualities but also on social and psychological relationships between employees (Tambe, Cappelli and Yakubovich, 2019). Lastly, a machine learning algorithm needs to be trained and the algorithm will most likely be characterized by that. This could, for example, be a problem when hiring new employees if a machine learning algorithm is trained on the current workforce and say that the majority of the workforce is white men, the algorithm could be biased towards white men when looking for new candidates (Tambe, Cappelli and Yakubovich, 2019). This exact example happened Amazon in 2018 and they 10 specifically had removed gender from their model as a criterion, but even then, the model learned in a way so that it was biased towards one specific group of people. However, the potential for machine learning within HR is huge. That is due to that HR operations and outcomes affect organizational performance in different ways. These are operations such as onboarding for new employees, training new and old employees, identifying good and bad performance, determining who should be promoted, employee retention, and employee benefits. (Tambe, Cappelli and Yakubovich, 2019)

2.3 Machine learning

Machine learning is the study of how to make computers learn from experience (Jordan and Mitchell, 2015; Tomassen, 2016). That is, to teach the computer to predict outcomes, based on examples. The data which a machine learning algorithm uses plays a big part in its success. The possibility to solve bigger and more complex problems grows when the amount of available data grows (Pedro, 2012). Machine learning techniques are widely used today, and they can be found in areas such as cars, the stock market, or health care (Pedro, 2012). However, the usage of ML models is limited to people’s will to use them (Ribeiro, Singh and Guestrin, 2016)There are several

types of machine learning techniques available: supervised, unsupervised, and reinforcement learning (Ayodele, 2010). The goal is the same but the approach and what prerequisites are to be fulfilled are different.

2.3.1 Types of Machine learning

There are three types of machine learning which are namely supervised, unsupervised and reinforcement learning which can also be called monitored, unattended and strengthening learning.

2.3.1.1 Supervised Learning

This type is the machine teaching task of delivering a function that maps an input to an output depending on an instance of duos of input-outputs (Stuart, Peter, 2010). It infers a function from marked training data constituting of a set of inputs objects and desired output values (Mehryar, Afshin & Ameet, 2012). A monitored/Supervised learning algorithm analyses the learning information and create an inferred function that can be used to map fresh instances. An ideal situation will enable the algorithm to generalize in a sensible manner from the learning information to the unseen circumstances.

2.3.1.2 Unsupervised Learning

It is a type of machine learning technique where the users do not need to supervise the model. The term unsupervised refers to Hebbian teaching allied with teacher-free, it is a method of modeling input probability density (Hinton & Sejnowski, 1999). A central framework of unmonitored learning is statistical density estimation, although unsupervised teaching involves many other fields involving the summary and explanation of data characteristics.

2.3.1.3 Reinforcement Learning

Is a machine learning zone involved with how software officials should act in an area to maximize some cumulative compensation concept. It varies from supervised teaching in that marked input / output duos do not need to be present and sub-optimal activities do not need to be clearly fixed. The focus is instead on discovering equilibrium between exploring and exploiting present understanding (Kaelbling, Littman, & Moore, 2011).

2.4 Machine Learning Algorithms for Employee Attrition Forecasting

2.4.1 Neural Networks

Neural networks refer to a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates (James Chen, Michael J

Bolye, 2020). According to Berry and Linoff, 2004, neural network can learn by example in much the same way that human experts gain from experience. The neural network can adjust to changing inputs thus the network can be able to generate the results without needing redesign the output criteria.

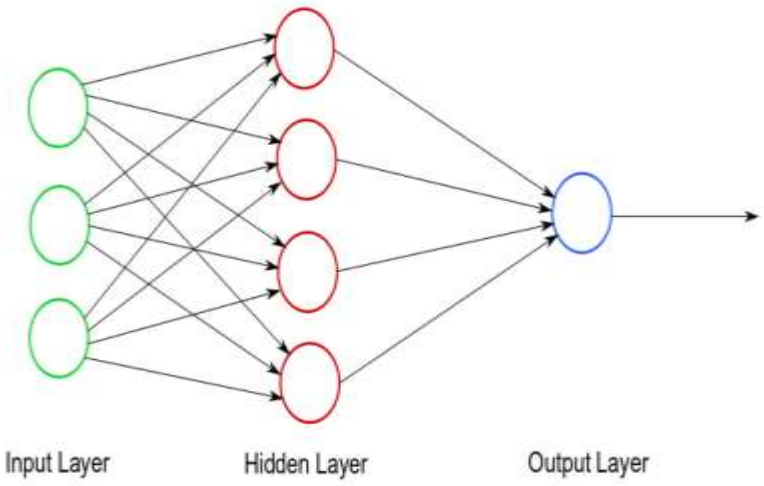


Figure 1: Neural network framework

It consists of at least three layers of nodes, the input layer which consist of one node for each independent attribute. Output layer consist of nodes for the class attributes and connecting these layers is one or more intermediate layers of nodes that transform the input into an output.

2.4.2 K-Nearest Neighbour

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

2.4.3 Logistic Regression

It is an algorithm used for binary classification problems, thus it can predict the likelihood of an event happening or not, measuring the relationship between a dependent variable and one or more independent variables. It can also be called a parametric classification model meaning it have a fixed number of parameters that depends on certain input features. This algorithm as compare to Mantel-Haenszel has an advantage of handling more than two explanatory variables simultaneously when classifying, this is according to Biochem Med(Zagreb) , 2014.

Logistic regression model can produce an outcome based on the individual characteristic. Therefore, this kind of an algorithm is a simple way of classifying variable in machine learning. Using attrition prediction, it will be easier as the algorithm can be trained based on the individual characteristics and therefore produce best results.

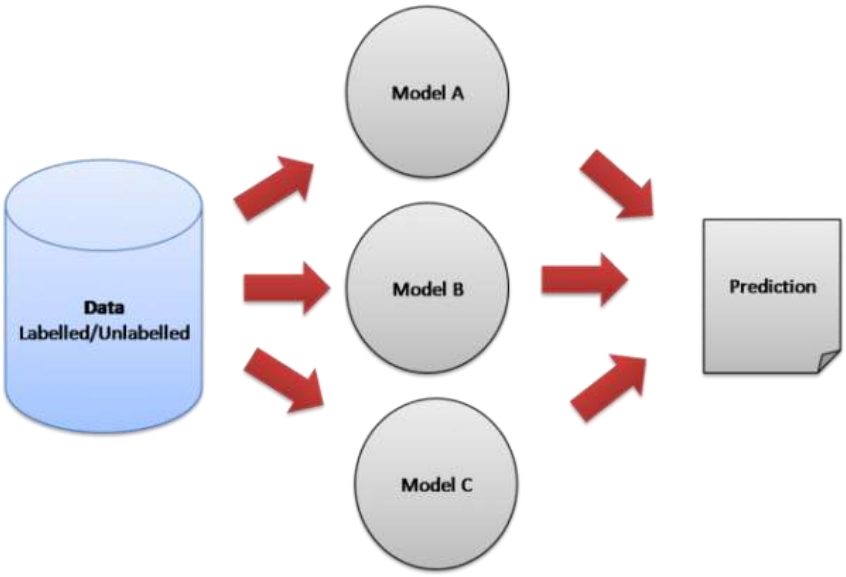
When measuring the chance of an outcome the logistic regression uses a logarithm equation of the chance because the chance is a ratio.

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta 0 + \beta 1 x1 + \beta 2 x2 + \cdots \beta m x m$$

In this equation π indicates the probability of the event for example churning or not and the other are coefficients associated with the reference group and the explanatory variables.

2.4.4 Ensemble Modelling

Ensemble modeling is a process where multiple diverse models are created to predict an outcome, either by using many different modeling algorithms or using different training data sets. The ensemble model then aggregates the prediction of each base model and results in once final prediction for the unseen data. The motivation for using ensemble models is to reduce the generalization error of the prediction. As long as the base models are diverse and independent, the prediction error of the model decreases when the ensemble approach is used. The approach seeks the wisdom of crowds in making a prediction. Even though the ensemble model has multiple base models within the model, it acts and performs as a single model. Most of the practical data mining solutions utilize ensemble modeling techniques.



2.5 Related Work

A lot of researchers have proved (Fallucchi *et al.*, 2020; Shah *et al.*, 2020) the usefulness of human resource management (HRM) in organisations for production, management, and some relations. HRM's effect on productiveness has effective results on increasing a business's capital and its intensity (Guthrie and Wright, 2005). Most research attention is on analyzing and tracking clients and their behavior (Gordini and Veglio, 2017). Many research analysed worker attrition. Existing research(Joseph *et al.*, 2021) confirmed that work ethics and job-related issues are the elements that usually affect worker attrition, these issues include income and time taken working in an organisation. Another research (Fallucchi *et al.*, 2020)decided how objective factors decide worker attrition, classical metrics were used to express the effects and one of the classification algorithms that had the best results for the IBM HR dataset is the Gaussian Naïve Bayes classifier. It had the best recall rate of (0.54) and an accuracy of 85%. The authors only targeted the simplest work-specific elements.

Authors in (Yuan, 2021) in contrast a Naïve Bayes classifier and the decision tree algorithm J48 in predicting the chance of an employee departing from the company. In particular, methodologies have been evaluated for every set of rules: tenfold cross-validation and percentage cut up 60. The results confirmed the accuracy of 86.4% and an incorrect category of 15.6% with J48 the usage of tenfold cross-validation, at the same time as there has been accuracy of 83.7% and incorrect

classification of 16.3% using percent cut up 60. In contrast, the Naïve Bayes classifier received an accuracy of 79.9% and a wrong classification of 22.1% using tenfold cross-validation, at the same time the s accuracy of 81% and the wrong classification of 18% was received using percent cut up 70. Authors explored the software of Logistic Regression whilst predicting worker turnover and received an accuracy of 85% and a false negative rate of 14%.

Other authors in (Najafi-Zangeneh *et al.*, 2021)targeted an improved Machine Learning-Based Employees Attrition Prediction Framework, the trouble in this became the need to enhance worker attrition prediction that specializes in every function. An HR dataset is selected because of the case study. The wrapper methods of feature selection technique where is proposed for measurement in the pre-processing stage, the validity of parameters is checked through training the version for a couple of bootstrap datasets and it gave an accuracy of 78%.

(Norrman, 2020) targeted predicting worker attrition with system studying on an individual level, and the results it may have on an organization. The problem became to expect worker turnover for individuals. Random forest model, Support vector machine model, and a logistic regression model are in comparison in terms of accuracy in predicting worker attrition with using large human resource information sets, and each random forest and support vector machine gave an accuracy of 97% which might be a bias.

A recent study by (Yahia, Hlel and Colomo-Palacios, 2021) looked at Big Data to Deep Data to Support People Analytics for Employee Attrition. Prediction employee attrition presents a critical problem and a big risk for organizations as it affects not only their productivity but also their planning. This attrition prediction approach is based on machine, deep and ensemble learning models and experiments on large-sized and medium-sized simulated human resources datasets and then a real small-sized dataset from a total of 450 responses it had an accuracy of (0.96, 0.98, and 0.99 respectively).

(Chourey, 2019) a research scholar did a survey paper on employee attrition prediction using machine learning techniques. This paper focused on identifying factors affecting employee attrition like salary hikes, growth opportunities, work environment, business travel opportunities, superior-subordinate relationships, recognition and appreciation, years since last promotion. Machine learning models: support vector machine (SVM) with many kernel functions, random

forest, and K-nearest neighbor (KNN) were used for this research. KNN (K = 3) achieved the very best performance, with a 0.93 F1- score.

An Improved Random Forest Algorithm for Predicting Employee Turnover was done by (Gao, Wen and Zhang, 2019) . Employee turnover is considered a major problem for many organizations and enterprises. The problem is critical because it affects not only the sustainability of work but also the continuity of enterprise planning and culture, and an improved RF algorithm, the WQRF based on the weighted F-measure, is proposed. The random forest algorithm is used to order feature importance and reduce dimensions. RF had an accuracy of 92.65% whereas WQRF had an accuracy of 92.80%.

2.5.1 Performance Comparison of The Reviewed Models

AUTHOR	TITTLE	YEAR	PROBLEM	SOLUTION	ACCURACY
Francesca Fallucchi 1,2, Marco Coladangelo 1, Romeo Giuliano 1, and Ernesto William De Luca	Predicting Employee Attrition Using Machine Learning Techniques	Published: 3 November 2020	Determining how objective factors determine employee attrition	Results are expressed in terms of classical metrics and the algorithm that produced the best results for the available dataset is the Gaussian Naïve Bayes classifier. It reveals the best recall rate (0.54)	85%
Rahul Yedida PESIT-BSC, Rakshit Vahi PESIT-BSC, Abhilash PESIT-BSC Deepti Kulkarni PESIT-BSC,	Employee Attrition Prediction	Published: April 2019	predicting whether an employee of a company will leave or not, using the k-Nearest Neighbors algorithm.	evaluation of employee performance, average monthly hours at work and number of years spent in the company, among others, as our features. Other approaches to this problem include the use	94.32%

				of ANNs, decision trees and logistic regression.	
FREDRIK NORRMAN	Predicting employee attrition with machine learning on an individual level, and the effects it could have on an organization	2020	Employee turnover prediction for individuals	random forest model, support vector machine model and a logistic regression model are compared in terms of accuracy in predicting employee attrition with the usage of large human resource data sets.	97% on both
Anjali Chourey Prof. Sunil Phulre Dr. Sadhna Mishra MTech. Research Scholar, Associate Prof., Dept. of CSE, Prof. & Head of Dept. (CSE),	A SURVEY PAPER ON EMPLOYEE ATTRITION PREDICTION USING MACHINE LEARNING TECHNIQUES	2018	identifying factors affecting employee attrition like salary hikes, growth opportunities, work environment, business travel opportunities, superior – subordinate relationship, recognition and appreciation, years since last promotion	machine learning models: support victor machine (SVM) with many kernel functions, random forest and Knearest neighbour (KNN). KNN (K = 3) achieved the very best performance, with 0.93 F1-score.	93%

Xiang Gao , 1 Junhao Wen , 2 and Cheng Zhang 1	An Improved Random Forest Algorithm for Predicting Employee Turnover	17 April 2019	Employee turnover is considered a major problem for many organizations and enterprises. The problem is critical because it affects not only the sustainability of work but also the continuity of enterprise planning and culture.	an improved RF algorithm, the WQRF based on the weighted F-measure, is proposed. The random forest algorithm is used to order feature importance and reduce dimensions.	RF 92.65% WQRF 92.80%
Xinlei Wang & Jianing Zhi	A machine learning-based analytical framework for employee turnover prediction	24 AUG 2021	Employee turnover (ET) can cause severe consequences to a company, which are hard to be replaced or rebuilt. It is thus crucial to develop an intelligent system that can accurately predict the likelihood of ET	we propose a machine learning-based analytical framework that adopts a streamlined approach to feature engineering, model training and validation, and ensemble learning towards building an accurate and robust predictive model.	, the model achieved 0.98 on D1 and 0.865 on D2

NESRINE BEN YAHIA 1 , JIHEN HLEL1 , AND RICARDO COLOMO-PALACIOS 2 , (Senior Member, IEEE)	From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction	27 APRIL 2021	employee attrition presents a critical problem and a big risk for organizations as it affects not only their productivity but also their planning continuity	this attrition prediction approach is based on machine, deep and ensemble learning models and is experimented on a large-sized and a medium-sized simulated human resources datasets and then a real small-sized dataset from a total of 450 responses	(0.96, 0.98 and 0.99 respectively)
Shawni Dutta The Bhawanipur Education Society College, Kolkata,India	Employee attrition prediction using neural network cross validation method	June 2020	to detect the feasibility of utilising related parameters and determine the probability of being affected by attrition process	feed-forward neural network and 10-fold cross validation procedure is provided under a single platform that can determine the attrition process beforehand.	87,01%
Shravan Shah, Soham Alatekar, 2Yash Bhangare, 3Bhavesh Kasar, 4Prof. Rahul Patil	Analysis Of Employee Attrition and Implementing A Decision Support System Providing Personalized Feedback And Observations	8 July 2020	Employees are every organization's precious asset. So, if they left jobs suddenly, any company might incur huge costs.	IBM dataset is predicted using three machine learning algorithms viz. Logistic Regression, Random Forest	87.9%

Abbas Heiat, College of Business, Montana State University, Billings, MT 59102,	PREDICTING EMPLOYEE ATTRITION THROUGH DATA MINING	2019	investigating individual employee characteristics and organizational variables that may lead to employee attrition.	Two classification methods used to develop models for predicting employee attrition. Artificial Neural Network (ANN) model predicted the employee attrition more accurately (85.33%) than Decision Tree (C&R Tree) model (80.89%).	85.33% and 80.89%

2.6 Chapter Summary

The author was successful in obtaining and collecting relevant information and data for the research topic. Some of the concepts employed by the researcher came from a variety of places, including academic papers, textbooks, and the internet, which revealed holes that needed to be filled. The information gathered from all of these sources will be utilised in the preceding chapters of the study to meet the research project's objectives. The method utilized in the design and development of the proposed solution is discussed in the following chapter.

CHAPTER 3: METHODOLOGY

3.0 Introduction

Research is a fact-finding activity that includes scientific research or an in-depth analysis of a particular issue of interest. Depending on whether the research is exploratory, descriptive, or diagnostic, quantitative or qualitative methodologies are used. When it comes to making economic decisions, research has shown to be a significant tool for government institutions and policymakers. (Mackey & Gass, 2013). Methodology is defined as the systematic, theoretical analysis of the methods or procedures applied to a particular field of study. In this chapter, the author will describe the methods used to achieve the stated study and system goals. Using the information gathered in the previous chapter, the author will establish the necessary procedures for building a solution and will be able to pick among competing strategies to achieve the research's targeted goals. Both primary and secondary data were analyzed to make the study process easier. Official sources, interviews, questionnaires, and observations were used to compile the data for this study.

3.1 Research Design

Research design is the architectural backbone of the study (Moule & Goodman, 2013). According to Polit and Hungler (2014), research design refers to the plan used to address research questions as well as controlling the challenges during the research process. One of four research models can be used by a researcher: observational, experimental, simulation, or generated. The researcher chose to employ both experimental methodologies since the application must be built and tested on a regular basis to ensure that it is producing the desired result. An experimental strategy is the best choice because this is a trial, trial, or preliminary technique. The researcher gathers experimental data via active intervention when a variable is changed in order to cause and measure change or create difference.

3.2 Requirements Analysis

The requirements must be realistic, documented, tested, executable, traceable, and quantifiable, as well as linked to known business demands and explicit enough to make system design easier (Abram Moore, Bourque, & Dupuis 2004). As a result, all of the functional and non-functional

specifications for the required system must be documented at this point. To develop uniform and unambiguous requirements, the collected requirements are analysed, changed, and inspected.

3.2.1 Functional Requirements

These can be characterized as a system's or component's function. A function is made up of three parts: inputs, behaviour, and outputs. "Functional requirements are those acts that a system must be able to accomplish, without regard for physical limits," Bittner explained. Computations, specialized subtle components, data control and preparation, and other specific capabilities that define what a system should achieve are examples. Use cases depict the behavioural conditions that apply to the great majority of instances in which the system applies the functional requirements.

The proposed system must be able to meet the following requirements:

i. to predict the churn or employee turnover using data provided
ii. to accept the data as supplied by user

3.2.2 Non-Functional Requirements

They're also known as quality requirements, and they're used to assess a system's performance rather than its intended behavior. The suggested system should be capable of meeting the following requirements:

Performance requirements

i. Flexibility requirements
ii. Accessibility requirements
iii. Quick response time

3.2.3 Hardware Requirements

Core i5 processor or better

3.2.4 Software Requirements

- Windows 10 Operating system
- Visual Studio Community Edition IDE
- Python
- Jupiter Notebook
- Flask Library

- Anaconda IDE

3.3 System Development

The overall design of the system and how it was constructed to achieve the desired goals are discussed in this system. It contains a list of all of the software tools and models that were used in the creation of the system.

3.3.1 System Development tools

The author had to figure out a suitable methodology to use during the development phase of the proposed solution in this section; however, the author has identified many frameworks for various projects, each of which has its own set of benefits and drawbacks based on the system to be designed and how it is able to produce accurate results aligned to the set objectives. The waterfall model, spiral model, and progressive (prototyping) model are examples of these frameworks. Because the model must be constructed and tested repeatedly to get at a final functional system, the author picked prototyping as a strategy for the suggested solution.

3.5.1 Prototyping Model

Prototyping model is a software development model in which a prototype is built, tested, and reworked until an acceptable prototype is achieved. It also creates base to produce the final system or software. The basic idea here is that instead of freezing the requirements before a design or coding can proceed, a throwaway prototype is built to understand the requirements. This prototype is developed based on the currently known requirements. By using this prototype, the client can get an "actual feel" of the system, since the interactions with prototype can enable the client to better understand the requirements of the desired system. Prototyping is an attractive idea for complicated and large systems for which there is no manual process or existing system to help determining the requirements. The prototype is usually not a complete system and many of the details are not built in the prototype. The goal is to provide a system with overall functionality (Lewallen, 2005).

The phases of the prototyping model:

- **Requirement Identification**: Here identification of product requirements is cleared in details. During this process, the users of the system are interviewed to know what their expectation from the system is.

- **Design Stage**: n this stage, a simple design of the system is created. However, it is not a complete design. It gives a brief idea of the system to the user. The quick design helps in developing the prototype.
- **Build the Initial Prototype**: An initial prototype the target software is built from the original design. Working off all the product components may not be perfect or accurate. The first sample model is tailored as per the comments were given by the users and based on that the second prototype is built.
- **Review of the Prototype**: After the product completes all the iterations of the update, it is presented to the customer or other stakeholders of the project. The response is accumulated in an organized way so that they can be used for further system enhancements.
- **Iteration and Enhancement of Prototype**: Once the review of the product is done, it is set for further enhancement based on factors like - time, workforce as well as budget. Also, the technical feasibility of actual implementation is checked. Features in the context full methodologies such as extreme programming or rapid application development.

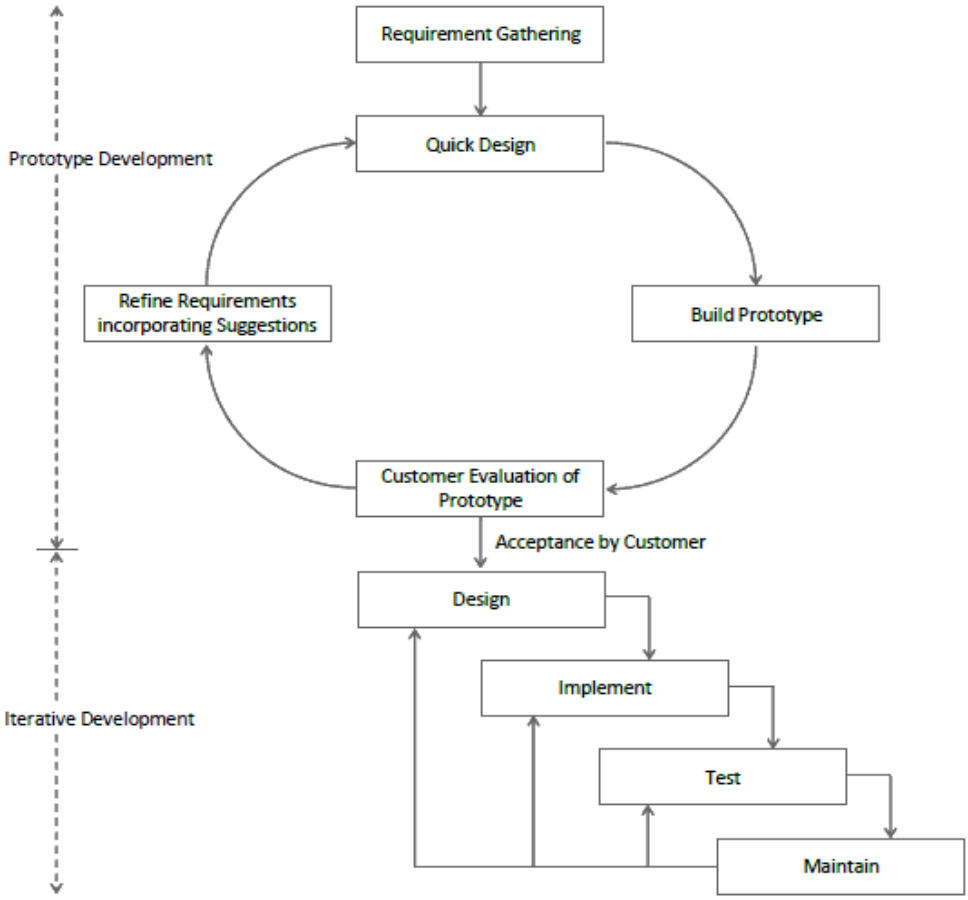


Figure 3: Prototype Model of Software Development

Figure 8: Prototype Method

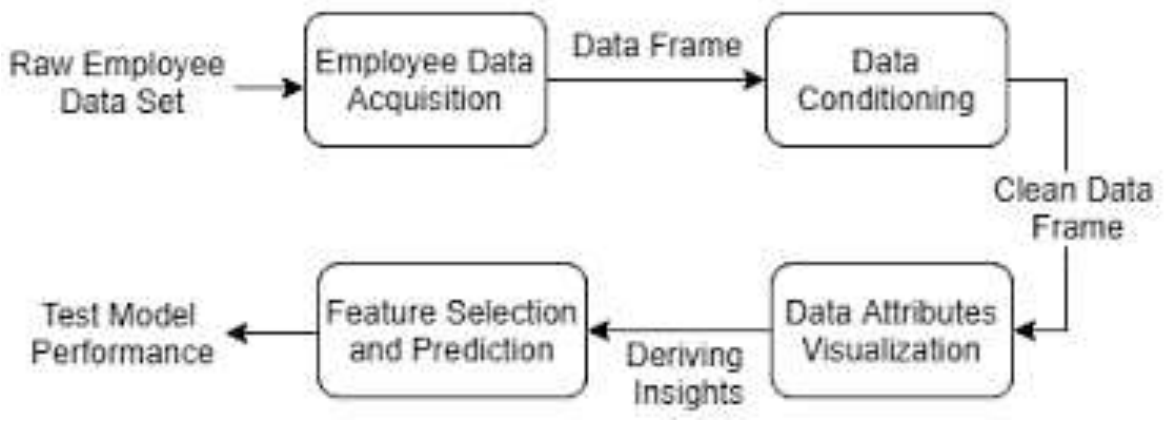
3.4 Summary of how the system works

The system allows the user ton input values relating to the parameters

3.5 System Design

3.5.1 System Dataflow diagrams (DFDs)

The Data Flow Diagram shows how the system's components are connected. To show component connections in the proposed system, the author used two distinct DFDs.



3.6 Implementation

```
# Function that runs the requested algorithm and returns the accuracy metrics
def fit_ml_algo(algo, X_train,y_train, cv):

    # One Pass
    model = algo.fit(X_train, y_train)
    acc = round(model.score(X_train, y_train) * 100, 2)

    # Cross Validation
    train_pred = model_selection.cross_val_predict(algo,X_train,y_train,cv=cv

    # Cross-validation accuracy metric
    acc_cv = round(metrics.accuracy_score(y_train, train_pred) * 100, 2)

    return train_pred, acc, acc_cv
```

Support Vector Machine

```
# SVC
start_time = time.time()
train_pred_svc, acc_svc, acc_cv_svc = fit_ml_algo(SVC(),X_train,y_train,10)
svc_time = (time.time() - start_time)
print("Accuracy: %s" % acc_svc)
print("Accuracy CV 10-Fold: %s" % acc_cv_svc)
print("Running Time: %s" % datetime.timedelta(seconds=svc_time))
```

```
Accuracy: 87.76
Accuracy CV 10-Fold: 86.1
Running Time: 0:00:00.207994
```

Logistic Regression

```
# Logistic Regression
start_time = time.time()
train_pred_log, acc_log, acc_cv_log = fit_ml_algo(LogisticRegression(), X_trair
log_time = (time.time() - start_time)
print("Accuracy: %s" % acc_log)
print("Accuracy CV 10-Fold: %s" % acc_cv_log)
print("Running Time: %s" % datetime.timedelta(seconds=log_time))
```

```
Accuracy: 89.89
Accuracy CV 10-Fold: 87.66
Running Time: 0:00:02.534987
```

K Nearest Neighbour ¶

```
# K Nearest Neighbour
start_time = time.time()
train_pred_knn, acc_knn, acc_cv_knn = fit_ml_algo(KNeighborsClassifier
knn_time = (time.time() - start_time)
print("Accuracy: %s" % acc_knn)
print("Accuracy CV 10-Fold: %s" % acc_cv_knn)
print("Running Time: %s" % datetime.timedelta(seconds=knn_time))
```

```
Accuracy: 89.21
Accuracy CV 10-Fold: 83.28
Running Time: 0:00:00.239998
```

Linear Support Vector Machines

```
# Linear SVC
start_time = time.time()
train_pred_svc, acc_linear_svc, acc_cv_linear_svc = fit_ml_algo(LinearSV
linear_svc_time = (time.time() - start_time)
print("Accuracy: %s" % acc_linear_svc)
print("Accuracy CV 10-Fold: %s" % acc_cv_linear_svc)
print("Running Time: %s" % datetime.timedelta(seconds=linear_svc_time))
```

```
Accuracy: 89.5
Accuracy CV 10-Fold: 87.27
Running Time: 0:00:00.269995
```

3.7 Conclusion

The chapter mainly focused on the methods and tool that were used to develop the model. Thus, different techniques and methods were used in developing the model up to the end, as mentioned above, the model was developed using different machine learning algorithms. Since the author is using ensemble modeling, these algorithms were combined to produce the most efficient performance. The next chapter presents the evaluation of the results of the model.

CHAPTER 4: RESULTS

4.0 Introduction

The effectiveness and efficiency of the established system solution had to be evaluated at the time the system was completed. The measures employed to assess the effectiveness and efficiency of the created system were performance, accuracy, and response time. Data gathered for the previous chapter was analyzed to produce results. White box, black box, and unit testing were used to determine the various behaviors the system displayed under various scenarios.

4.1 System Testing

Testing is essential in system development, when a system has been developed it has to be tested. This chapter shows tests that where undertaken and the results that were produced, test conducted mainly focus on functional and non-functional requirements of the proposed solution.

4.1.1 Black Box Testing

Black box testing is a technique where the internal organization, layout, and use of the product are not taken into account. In other words, the tester is unaware of how it operates within. Only the system's exterior behavior is assessed by the Black Box. Both the system's inputs and its outputs, or responses, are put to the test. The findings of the black box test the author did on the model are as follows. The system will therefore be evaluated to see how well it predicts the likelihood of an employee leaving an organization. The following are the findings of the author's black box testing of the model:

Over Time : ○Yes ○No

Performance Rating : 1-4

Relationship Satisfaction : 1-4

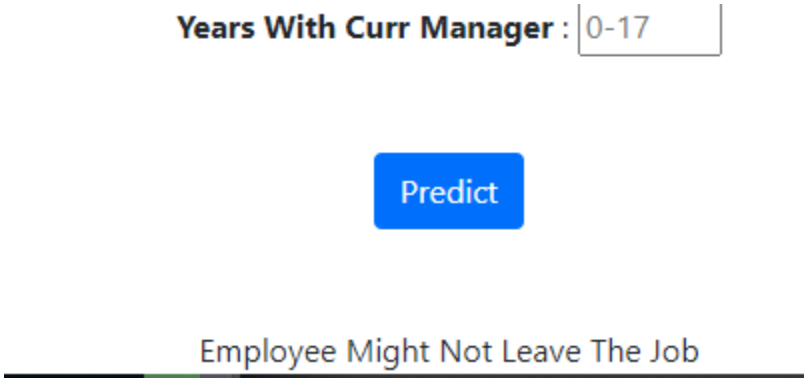
Stock Option Level : 0-3

Total Working Years : 0-40

Training Times Last Year : 0-6

Work Life Balance : 1-4

Years At Company : 0-40



4.1.2 White Box Testing

White box testing is a software testing technique where the underlying structure of the software is known to the tester before the software is tested. This kind of testing is typically done by software developers. White box testing requires understanding of programming and implementation. The lower levels of testing, such as unit and integration testing, are applicable to testing. White box testing focuses primarily on testing the computer code of the system being tested, including its branches, conditions, loops, and code structure. White box testing's primary objective is to evaluate the system's functionality. The developer tested the model as shown below;

```
# Function that runs the requested algorithm and returns the accuracy metrics
def fit_ml_algo(algo, X_train,y_train, cv):

    # One Pass
    model = algo.fit(X_train, y_train)
    acc = round(model.score(X_train, y_train) * 100, 2)

    # Cross Validation
    train_pred = model_selection.cross_val_predict(algo,X_train,y_train,cv=cv

    # Cross-validation accuracy metric
    acc_cv = round(metrics.accuracy_score(y_train, train_pred) * 100, 2)

    return train_pred, acc, acc_cv
```

4.2 Evaluation Measures and Results

A performance evaluation metric is used to gauge a model's effectiveness (Hossin & Sulaiman, 2015). According to Hossin & Sulaiman (2015), model assessment metrics can also be divided

into three categories: threshold, probability, and ranking. Performance is measured by how successfully the model can predict employee attrition. The confusion table in table 1 below was used by the author to evaluate the accuracy of the system.

4.2.1 Confusion Matrix

The confusion matrix is a table that shows the number of categories that have been assigned and those that have been anticipated. The table is used to define the model's performance.

True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are the four terminologies used.

- TP denotes situations that are genuinely true and that the test has yielded accurate result, whereas TN denotes numbers that are untrue and that the test has resulted as false.
- FP-Are those that the test shows as true, but which are actually untrue.
- FN-Numbers that the test shows as false but are really correct.

Type	Returned number of correct predictions	Returned number of incorrect predictions
1	True Positive	False Negative
2	False Positive	True Negative

Table 1 Confusion Matric

The technology was put to the test in terms the returned number of correct and incorrect predictions. For the purpose of observing the system's findings, three scenarios and a test environment were developed. The system was observed 40 times on each scenario using different testing input. All of the scene analysis was done to ensure that the answer was accurate and that false predictions were identified. The tables below indicate the outcomes of the tests that were conducted.

Table 2 Confusion matrix for employee attrition prediction

Test cases	Predictions	Number of tests	Correct predictions	False predictions	Classification
1	Yes	40	36	3	True positive
2	No	40	35	4	True negative

4.3 Accuracy

The number of correct predictions divided by the total number of tests in each category equals accuracy. The percentage of accuracy is then calculated by multiplying it by 100. The following equation is used to compute it:

Equation 1: Accuracy calculation as adopted from Karl Pearson (1904)

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} * 100$$

$$\text{Accuracy rate } = \frac{37 + 36}{36 + 35 + 3 + 4} * 100$$

$$\text{Accuracy} = \frac{73}{80} * 100$$

$$\text{Accuracy} = 91{,}2\%$$

4.4 Misclassification Rate/ Error Rate

- Overall, how often is it wrong?
- It tells you what fraction of predictions were incorrect. It is also known as Classification Error.
- **Error rate = (FP+FN)/(TP+TN+FP+FN) or (1-Accuracy**

= (3+4)/(37+36+3+4)

=0.8%

4.5 Precision

- When it predicts yes, how often is it correct?

Precision = TP/(TP+FP)

37/(40+3)

=86%

4.6 Sensitivity/Recall/True Positive Rate

- When it's actually yes, how often does it predict yes?
- It tells what fraction of all positive samples were correctly predicted as positive by the classifier. It is also known as True Positive Rate (TPR), Sensitivity, Probability of Detection.

Recall = TP/(TP+FN)

=37(40+4)

=84%

4.7 Specificity/True Negative Rate

- When it's actually no, how often does it predict no?
- It tells what fraction of all negative samples are correctly predicted as negative by the classifier. It is also known as True Negative Rate (TNR).
- Equivalent to 1 minus False Positive Rate
- **Specificity = TN/(TN+FP) or 1-FP rate**

=1-4

=96%

4.8 Prevalence

- How often does the yes condition actually occur in our sample?
- It shows how often does the yes condition actually occur in our sample
- **Prevalence=Actual YES/(TP+TN+FP+FN)**

4.9 F1-Score/F1 Measure

- It combines precision and recall into a single measure.

- **F1-score=2 x (Precision x Recall/ Precision + Recall)**

  =2TP/(2TP+FP+FN)

  =2(37)/(2 x 37 + 3+4)

  =91.3%

4.5 Summary of Research Findings

The researcher performed all the necessary black, white box tests and performance tests using the confusion matrix, the author found that the system had satisfactory performance. The system was tested in accuracy, misclassification error/error rate and it achieved 91.2% and 0.8% respectively. The model attained an overall precision of 86% and a sensitivity or recall of 84%. An F1 score of 91.3% was achieved with a specificity or true negative rate of 96%.

4.6 Conclusion

To conclude this chapter, the author used different metrics for performance measurement of the system. Among them being, accuracy, specificity, recall precision, error rate f1-score and true positive rate. The next chapter presents the conclusion, objective realization and recommendations for further development.

Chapter 5: Recommendations and Future Work

5.0 Introduction

In the previous chapter, the researcher focused on presentation and analysis of obtained data. This chapter covers the research and development of the solution in line with the set objectives and examine the difficulties encountered by the researcher in designing and carrying out the research.

5.1 Aims and Objectives Realization

In summary, the objective of the research was to predict using a dataset if an employee is likely to leave an organization or not. Also, the other objective was to evaluate the effectiveness of ensemble modelling in employee churn prediction. Therefore, to this end, the researcher managed to develop a system or model that uses ensemble modelling for various ML algorithms which are Logistic Regression, SVM, Linear SVC, KNN, Naive Bayes, Decision Tree, Gradient Boosting Trees and Random Forest algorithm to predict which employee is likely to leave using a dataset of an unknown organization found on Kaggle.

The ensemble model, with the dataset given, as mentioned in the previous chapters, datasets comprised of testing and training data, thus, the system trained and then was able to be tested for accurate

predictions. It managed to give an accuracy rate of 91% after more test on data were given, making it more likely to predict the churn of employee given the required dataset. Therefore, as compared to other system the with a lesser accuracy percentage, the improvements meant a reduction of errors hence, making it easy for companies to take proactive measures on the employee attrition. This therefore, shows that the researcher managed to meet all the objectives.

5.2 Conclusion

The use of ensemble modelling in employee attrition managed to make quite great prediction as the model was able to train and predict churn or not. The author managed to detect the factors that influence employee churn. Thereby making it easy to determine which aspects need improvements and which does not.

5.3 Recommendations

There is need for greater coverage in the employee churn using some other machine learning algorithms and deep learning techniques. This makes it easier for comparisons and make decisions on which algorithm might produce good results based on the employee churn predictions.

5.4 Future Work

The researcher did not have enough time and proper machinery to carry out the research on a production level. Future work involves developing and the model on a production level scale to see if the system and the proposal might work in predicting the employee churn of telecoms and health care organization.

References

1. Peng, B. Statistical analysis of employee retention. In Proceedings of the International Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS 2021), Nanjing, China, 26–28 November 2021; Volume 12163, pp. 7–15.
2. Mathematics, and Computing Science (CSAMCS 2021), Nanjing, China, 26–28 November 2021; Volume 12163, pp. 7–15. https://doi.org/10.1117/12.2628107.
3. Employee Retention Statistics That Will Surprise You. 2022. Available online: https://www.apollotechnical.com/employeeretention-statistics/ (accessed on 6 September 2022).
4. Here's What Your Turnover and Retention Rates Should Look Like. Available online: https://www.ceridian.com/blog/turnoverand-retention-rates-benchmark (accessed on 6 September 2022)
5. Gandomi, A.H.; Chen, F.; Abualigah, L. Machine Learning Technologies for Big Data Analytics. Electronics 2022, 11, 421. [CrossRef]
6. Jia, X.; Cao, Y.; O'Connor, D.; Zhu, J.; Tsang, D.C.W.; Zou, B.; Hou, D. Mapping soil pollution by using drone image recognition and machine learning at an arsenic-contaminated agricultural field. Environ. Pollut. 2021, 270, 116281. [CrossRef] [PubMed]
7. Alao, D., Adeyemo, A.B.: Analyzing employee attrition using decision tree algorithms. Comput. Inf. Syst. Dev. Inform. Allied Res. J. 4 (2013)
8. Al-Radaideh, Q.A., Al Nagi, E.: Using data mining techniques to build a classification model for predicting employees performance. Int. J. Adv. Comput. Sci. Appl. 3, 144–151 (2012)
9. Chang, H.Y.: Employee turnover: a novel prediction solution with effective feature selection. WSEAS Trans. Inf. Sci. Appl. 6, 417–426 (2009)
10. Chien, C.F., Chen, L.F.: Data mining to improve personnel selection and enhance human capital: a case study in high-technology industry. Expert Syst. Appl. 34, 280–290 (2008)
11. Li, Y.M., Lai, C.Y., Kao, C.P.: Building a qualitative recruitment system via SVM with MCDM approach. Appl. Intell. 35, 75–88 (2011)
12. Nagadevara, V., Srinivasan, V., Valk, R.: Establishing a link between employee turnover and withdrawal behaviours: application of data mining techniques. Res. Pract. Hum. Resour. Manag. 16, 81–97 (2008)

13. Quinn, A., Rycraft, J.R., Schoech, D.: Building a model to predict caseworker and supervisor turnover using a neural network and logistic regression. J. Technol. Hum. Serv. 19, 65–85 (2002)
14. Sexton, R.S., McMurtrey, S., Michalopoulos, J.O., Smith, A.M.: Employee turnover: a neural network solution. Comput. Oper. Res. 32, 2635–2651 (2005)