National College of Ireland

Levereging Machine Learning to Predict Employee Attrition: India

MSc Research Project
Data Analytics

Sangeeta Kumari
Student ID: x21141088

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sangeeta Kumari……
Student ID:	x21141088……..……
Programme:	Msc in Data Analytics…… **Year:** ……2022…….
Module:	Research Project……..……
Supervisor:	Dr.Catherine Mulwa…….……..……
Submission Due Date:	15th Dec 2022……..……
Project Title:	Leveraging Machine Learning to Predict Employee Attrition…
Word Count:	8337…… **Page Count**…29……..…….

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Sangeeta Kumari……

Date: ……31st Jan 2023……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Leveraging Machine Learning to Predict Employee Attrition

Sangeeta Kumari
X21141088

Abstract

Employee attrition is an important issue that affects businesses and organizations across the world. This can be attributed to the fact that hiring new employees takes time and consumes a lot of resources. Hence, predicting if an employee is unhappy at an organization can help the organization take necessary steps to avert attrition. Hence, developing a model that can predict employee attrition is very beneficial for the organization and for the employee as well.

This research studied the performances of four machine-learning models viz. decision tree, random forest, SVM, and artificial neural network. The models were implemented through two experiments in which the first experiment involved predicting the attrition for the whole organization whereas the second experiment involved predicting the attrition department-wise. The random forest model of machine learning model implemented in the model achieved the best performance with accuracy of 64.39% and f1-score of 63.96%.

1 Introduction

In the current business environment, it has been identified that companies are highly focused on implementing different technology in their operating and decision-making process to attain a higher level of business efficiency. Technologies, like machine learning, provided an opportunity for organisations to acquire relevant data, process the same and make appropriate predictions based on which the course of actions in an organisation could be determined in an appropriate and effective manner. Thus, machine learning processes also play a key contributing role in the process of identifying the potential events or events which could create an impact on the operating processes of the companies. In this regard, Zhou (2021) defined machine learning as a key part of artificial intelligence (AI) that shows the capability or potential of a machine to imitate the behaviour of human beings.

In the operating processes of the companies, employees could be considered as an essential factor to consider as efficiency and effective human resources support the organisations to develop plans, implement the same in the operating process and achieve the targeted goals in the long-term. Singh (2019) highlighted in a study that for attaining a competitive edge and long-term business sustainability organisations must focus on developing strategies to acquire and retain an efficient workforce. Thus, higher employee attrition is observed to be a key concerning factor for the companies in the current business situation. In this regard, this research paper will focus on identifying the potentiality and effectiveness of using machine learning processes to predict

employee attrition. This is because appropriate predictions of potential employee attrition level would be effective for the management of the companies to identify the issues present in the business or relevant other factors which are motivating the companies to leave the firm and also develop plans which could provide an opportunity to minimise the overall employee attrition in near future as well.

1.1 Motivation and Project Background

Tanwar and Prasad (2016) highlighted that employees are integral parts of the companies which support and ensure the proper implication of the developed managerial strategies. As a result of that, proper employee management could contribute to the long-term organisational goals and objectives. The availability of an efficient workforce in the organisational operating process also contributes to the process of developing innovative strategies that could support the proper utilisation of the other resources in an appropriate and effective manner.

This, in turn, supports minimising the overall costs of the businesses in this price-sensitive business environment. Apart from that, with the growth in the overall market competition, it has been identified that companies are highly focused on acquiring and managing an efficient workforce in their operating processes (Iqbal et al., 2017). For that purpose, they are focusing on identifying and utilising effective human resource management strategies to motivate employees and retain them. Even companies are also trying to acquire efficient employees from other organisations to add value to their processes of operations.

In this process, it has been observed that for companies employee attrition is a key matter of concern. This is because high employee attrition creates different operational difficulties for the organisations, like allocation of resources, the performance of required tasks, fulfilling orders and others. Recognising that, Frye *et al.* (2018) also highlighted that it is essential for the management of a firm to monitor its employee attrition level to develop its HRM plans accordingly. In this aspect, it is essential to state that employee attrition may occur for different reasons, such as termination, resignation, retirement, or death. In this regard, Showry and Manasa (2016) highlighted that a lack of knowledge of the underlying reasons or factors of employee attrition in an organisation often causes significant damage or complexities in HRM planning and decision-making.

Hence, it is essential for organisations to implement relevant strategies and tools which could support in assessing the present attrition situation, identifying the key reasons for attritions and determining the attrition level in the near future as well. In the present operating environment of the companies, it has been identified that the firms are highly focused on implementing or using different technologies, like artificial intelligence and others. The key focus of using such revolutionary technologies in business operations is to enhance overall decision-making, planning and performance efficiency. The implication of such technologies is observed to deliver and serve the companies in multidimensional aspects, like cost reduction, performance efficiency, efficient utilisation of the resources, decision-making support and others. As a result of that efficient

utilisation of the resources, companies are observed to attain a higher competitive edge in their process of operations.

In this process of higher focus on utilising different technologies in the operating processes of the companies, it has been identified that Machine Learning (ML) has gained significant market recognition. Fallucchi *et al.* (2020) highlighted that with the implication of appropriate ML techniques in the business processes the enterprises are now able to identify the trends in its different internal and external aspects, such as business operational patterns, behavioural trends of the customers and others. As a result of that, depending on the potential trends, the management of the firms is now able to develop suitable innovative strategies in their different decision-making aspects, like the development of products and services, strategies to enhance HRM practices and others. Considering the implication dynamics of ML, large multinational companies, such as Uber, Google, Amazon and others, have focused on incorporating suitable ML practices at the centre of their operating processes (Lee and Shin, 2020). Hence, appropriate utilisation or implication of ML techniques could be considered as a major competitive differentiator for the companies. Even the multidimensional usability of ML provides the management of the enterprises to add value for its internal and external stakeholders and attain higher growth in the long term.

In this aspect, proper consideration and evaluation of the implication and effectiveness of the ML approach in the employee attrition aspects could be considered to deliver a higher value for the organisations. This is because employees are crucial aspects for the companies and hold a strong influence on their operational efficiency as well. ML approaches hold a key characteristic of identifying the existing trend and making predictions based on that. Hence, it could be expected that the use of ML techniques in the employee attrition situation would be effective to deliver a new approach to the HRM practices of the organisations. In this situation, performing detailed research would be effective in determining whether ML approaches are effective in predicting employee attrition for companies. Apart from that, the most suitable ML approaches in this regard could also be identified along with the potential issues present in the system as well. This is because confirming the proper application requires the companies to develop a suitable framework, make financial investments and adjust the existing operating processes as well.

1.2 Research Question

The research question can help to understand the basic features related to the subject that can enhance the knowledge and information related to the topic of this study. It is necessary to identify the basic requirement to understand this research effectively.

"*How can the analytical statistics of the data which is evaluated based on Accuracy & f1-score, and machine learning several models, SVM(Support Vector Machines), Neural Networks, Decision tree, Random forest tree combined to reduce employee churn in an organization?"*

Sub RQ: What is the change in the results of the machine learning model when applied department-wise than with the organization?

Solving the research question requires the following objectives to be completed.

1.3 Research Objectives

Obj 1: A critical Review of Employee attrition in India
Obj 2: Exploratory Data Analysis
Obj 3: Data Pre-processing
Obj 4: Implemented model are evaluated based on accuracy & f1-score of analytical analysis
Obj 5: Implementation of Employee Attrition Prediction Models
Obj 5(i): Experiment-1performed on organisation level of attrition
Obj 5(ii): Experiment carried out on several department wise attrition rate
Obj 6: Evaluation and Experiment result
Obj 6(i): Evaluation and experiment result of experiment-1
Obj 6(ii): Evaluation and experiment result of experiment-2
Obj 7: Comparision of developed models
Obj 7(i): Experiment-1 which carried result for implemented models for whole dataset in organization lavel
Obj 7(ii): Experiment-1 evaluation of models in department wise

The rest of the research structure is as follows, chapter 2 presents the literature review conducted for identifying the state of the art and implementation of employee attrition prediction in businesses, this will help in designing the methodology presented in chapter 3 , chapter 4 presents the design specifications of the study, chapter 5 discusses the implementation, chapter 6 presents the results obtained from the study, and chapter 7 concludes the findings of the research and put forward the future aspects of the research.

2 Literature Review on Prediction of Employee Attrition

2.1 Introduction

Machine learning is the subject of the IT ground of organisations that helps enhance the performance of the technological ground effectively. The perception and ideas regarding machine learning are very crucial and vital that help the developer to follow the structure and pattern effectively. The aim of the literature review will focus on the ideas and roles regarding machine learning and benefits of machine learning in organisations. Overall research will be based on the implementation of machine learning and its beneficial results in the organisation. The key of the literature review will illustrate the list of the articles through which the researchers have discussed will be represented and the literature gap will show the barriers which have occurred in the research regarding the information.

2.2 Benefits and Barriers Present to Implementing Machine Learning in Predicting Employee Attrition

Machine Learning Algorithms

According to Ajit (2016), the machine learning algorithm is a crucial part of machine learning. This shows the capability of the system effectively. A proper algorithm of machine learning can help the individual to fulfil their official work without adding any subject properly and without any type of hassle. Machine learning can help the individual to gather the basic information regarding inclusion subject video chat with business organisations. Official meetings where employees are able to gather data or details regarding the product or any other crucial subject related to the business procedure that can be presented in the meetings are also shared with other employees for developing new ideas and plans as well as structures. Machine learning algorithms can help the employees Understand their performance and therefore in the organisation. Machine learning is a very effective subject when it comes to analysing the inputs and outputs of the increase in the enterprise because through the algorithm the employees can understand the basic requirement to enhance their performance in the company. It is very much required that workers of the business organisation are able to input the proper information in the system so that they can get the exact details properly without any type of disruption.

Impact on the Organisational Policies

According to Ribes (2017), organisation policies are very much important to maintain the system and the structure of the business firm. A company is able to make proper decisions based on the policies of the company. Business policies are a crucial subject that helps to create a disciplinary environment in the workplace of the business organisation. Machine learning the employees are able to rectify mistakes properly and effectively to avoid any type of turmoiled situation in the business organisation. Machine learning will be responsible for the development of the performance of workers that can positively impact the policies of the organisation because the workers will be aware of the basic needs which can be noted with the help of machine learning. Through the machine learning process, employees are able to understand the importance of the organisation's policy to maintain discipline in the workplace so that they can achieve their personal goals in the enterprise to enhance their career field effectively. required to prioritise the machine Derby because this subject provides the proper and required information that can give beneficial and terrible results to any individual in the business organisation. It is very much required for the members of the organisation to enhance their knowledge of technology.

Lack of Technological Knowledge

According to Spanoudes and Nguyen (2017), knowledge regarding technology is very much important for every employee in the organisation. A lack of knowledge regarding the technology can prevent the employees of the company to reach the higher goals in the business organisation. It is one of the most noticeable behaviours that usually prevents machine learning in business organisations. Machine learning needs proper knowledge regarding it so that the workers are able

to get proper beneficial results from this subject. The IT ground of the company needs to hire proper developers who are skilled and experienced regarding the technological experiments so that they can analyse the requirements which need to be implemented in a system that can help the company to achieve its goals effectively. Moreover, it is very required that workers of the organisation must have proper ideas to understand the system and the result of the implementation of this subject in the business so that they can get proper results in the marketing presidio in the markets. The members of the company need to understand that there are certain grounds which need proper details and information so that they can proceed effectively. dealing with the customers is one of the most important tasks for the members of the organisation.

Poor Cognition

According to Chekroud *et al* (2017), poor cognition can prevent the workers of the business organisation from inviting machine learning into the enterprise. The employees must understand this situation so that they can able to take better opportunities and advantages of this procedure can help the organisation to enhance performance in the corporate sector full stop the business organisation can go to a lot of competition that can help them to take a lot of opportunities but due to lack of prediction about the future result, it can be very challenging for companies to perform remarkably against the competitors.

There are a lot of senior members in the organisation who a piece of lack knowledge regarding the technology and give the reason why it can be very hard for machine learning in the enterprise. It is very important for every employee of the company to enhance their cognitive skills to understand the proper result of machine learning in the enterprise that can help to improve the growth of the organisation.

2.3 Different Machine Learning Algorithms and their Implication in Business Process

According to Bishop (2022), machine learning algorithms are a crucial subject during the development of systems in the technical field. However, the developers need to be aware of the important data which programs are worthy to be implemented in the system through programming procedure. Every business organisation is increasingly applying machine algorithms to make proper and accurate decisions. Proper information can help the members of the company to get proper details and data regarding the business procedures. can help individuals to understand the data properly because degrees of organisations expect accurate information regarding the scenarios of the marketing status. The developers need to be knowledgeable regarding subjects related to the companies through their own personal research so that they can implement all the necessary information in the devices and the systems. Machines or computer systems are responsible to provide division and perfect as less appropriate information. There are four types of machine learning algorithms.

Supervised

According to Ray (2019), the machine is taught by examples.
The operator is concerned about the algorithm of the system. Implementations of the data help the system to protect and deliver important messages or information to the individuals that can help them to get the proper data regarding any type of subject. Under the supervised algorithm classification, forecasting and regression are the main features.

Semi-supervised
According to Shang *et al* (2022), Semi-supervised is quite different from supervised learning. In semi-supervised learning, mixing machine learning prefers to deliver the unlabeled data in the learning procedure of the machine. The system is able to deliver their details and information by combining the additional information. This subject is very much remarkable that enhances the knowledge of the developer regarding current affairs and other important topics related to other subjects.

Unsupervised
According to Guo and Li, (2022), the machine starts delivering the information through their own analysis in the answer-providing learning. This shows that the developer has input the basic and essential requirements where the Machines are able to deliver the information through their own analysis skills. Sometimes it can be very challenging for the developer to find The activated information regarding any type of subject which can be implemented in the machine therefore they try to improve the performance of the machine so that the system is able to deliver the basic or self-analysed information.

Reinforcement
According to Mason and Grijalva, (2019), reinforcement is one of the most interesting sections of machine learning. Developers search for advanced implementations that help the machine to analyse, evaluate or take action regarding any kind of important matter.

2.4 Implementation of Machine Learning in Employee Management

Leading Digital Transformation
According to Sahaja (2021), digital transformation is a crucial subject in the business organisation to enhance the performance of the companies in an effective manner. The employees of the company can be aware of the changes which are very much required in the business organisation through the implementation of machine learning. Digital transformation can help the workers of the organisation to enhance their performance in an effective manner due by conducting research on the subjects related to business management with the help of machine learning that can provide important data related to the performance or within a view of the markets.

Enhancing the Pieces of Information

According to García-Peñalvo (2018), With the help of machine learning employees are able to get the required data details regarding a subject which are related to marketing that will positively impact the inputs and outputs of the business organisation. Machine learning the workers of the company are able to check their own collected information appropriately the help of machine learning to get confirmation of what they have collected regarding the marketing session. It is very much required to be confirmed the collected data before representing those details to other workers of the enterprise. Enhancing information through machine learning is very important.

Improvement of Strategic Management

According to Choudhury *et al* (2021), strategic management is very much important in the business organisation subject that helps enterprises to grow rapidly and effectively in the corporate sector. Machine learning can help employees to create new ideas regarding strategy development which are worthy to be implemented for the growth of the company. Through the help of machine learning, the workers will be able to enhance their creative thinking and analysis which can help them to share their thoughts regarding the development of the marketing for the studio of the enterprises. Improvement in strategic development can be more effective if the employees are able to improve their knowledge regarding machine learning and the best outcomes from this subject.

Improvement of Finances Management

According to Ammar *et al* (2022), finance management is a crucial part of the business organisation. Machine learning can help the members of the finance department to provide significant information and data regarding the income and investments of the company with additional information. Collecting the data regarding the finances is very much important to know the status of the profitability of business organisations in marketing. Machine learning can provide the data in a qualitative manner where the employees are able to identify the negative and positive outcomes in a sequenced way which can be represented to the authority of the business organisation.

Delivering Proper Prediction

According to Subhashini and Gopinath (2021), delivering the appropriate predictions through machine learning can help the employees of the organization to take the required precautions to prevent any type of student that can be negative. Property auctions can help the members of the company to analyse the positive side of the performance of the business organization to get motivated to start enhancing the planet which can help to maintain the outputs from the marketing session (Hellström, 2016).

2.5 Conclusion

From the reviewed literature, it can be seen that there is a clear gap that exists in the field of employee attrition prediction that needs to be addressed in order to develop a system that answers

the research question (section 1.2) presented and fulfill the research objectives (section 1.3). The next chapter presents the methodology to develop an employee attrition prediction system.

3 Methodology

The presented research is conducted by following the methodology presented in figure 1 below. The modules present in the methodology are discussed in detail below. The methodology presented below is Knowledge Discovery in Databases (KDD) methodology. This methodology aims to extract knowledge from the selected dataset. The presented methodology consists of four stages involving viz. a) Data Selection, b) Data Preparation/Transformation, c)Data Modeling, and d) Evaluation. The stages involved are explained in detail in upcoming sections.

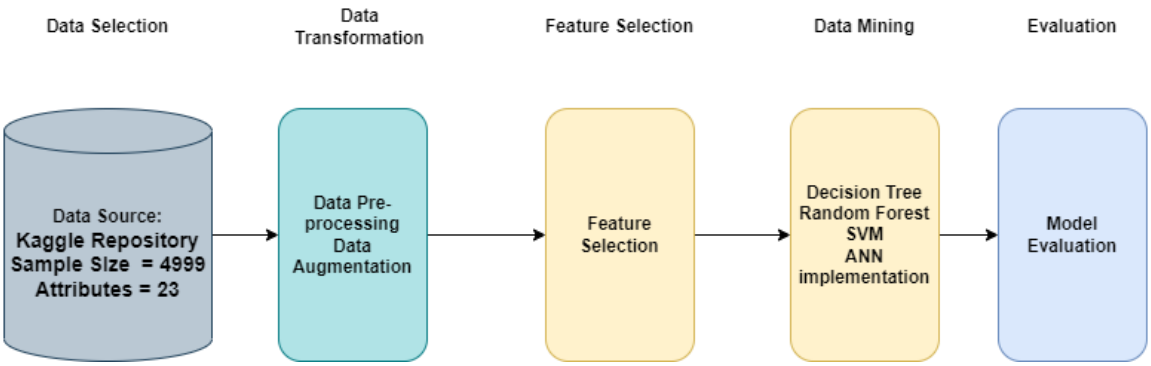


Figure 1: Methodology incorporated in the study

3.1 Data Selection

The dataset that has been used in the study is obtained from the Kaggle website and is created by Hacker Earth. The dataset consists of 23 attributes, their datatypes and names are presented in table 1 below. It consists of a total of 4999 samples pertaining to each of the employees under consideration.

https://www.kaggle.com/datasets/blurredmachine/hackerearth-employee-attrition

Table 1: Dataset attributes and their description

Column Name	Description
Employee_ID	Unique ID of each employee
Age	Age of each employee
Unit	Department under which the employee work
Education	Rating of Qualification of an employee (1-5)
Gender	Male-0 or Female-1
Decision_skill_possess	Decision skill that an employee possesses
Post_Level	Level of the post in an organization (1-5)
Relationship_Status	Categorical Married or Single
Pay_Scale	Rate in between 1 to 10
Time_of_service	Years in the organization
growth_rate	Growth rate in percentage of an employee
Time_since_promotion	Time in years since the last promotion
Work_Life_balance	Rating for work-life balance given by an employee.
Travel_Rate	Rating based on travel history(1-3)
Hometown	Name of the city
Compensation_and_Benefits	Categorical Variabe
VAR1 - VAR5	Anominised variables
Attrition_rate(TARGET VARIABLE)	Attrition rate of each employee

3.2 Data Preparation

This is one of the important steps in any data mining methodology and is responsible for making the data suitable for modeling. This stage of the methodology consists of five substages.

(i) Exploratory Data Analysis – In this substage the data is visualized thoroughly to identify any missing values in the dataset. Some of the attributes in the dataset contain null values such as 'work_life_balance' which are replaced with 0. It also helps an analyst to decide on the required operations that need to be performed on the dataset for analysis.

(ii) Encoding – Encoding is a process to replace categorical (string) variables in numerical form. This is necessary for machine learning algorithms as some of them can not classify categorical variables. Most of the variables in the dataset are categorical and hence are converted to numerical form.

(iii) Class Balancing – It is a process of balancing the classes in an unbalanced classification problem such as employee attrition. In a machine learning algorithm, class balancing is necessary so as to avoid overfitting, and obtain a reliable classification. The dataset consists of 4611 samples from the non-attrition group and the remaining 388 from the attrition group. This begs for the class balancing of the dataset. The dataset is balanced such that the samples pertaining to both groups

equal 4611 in number. Class balancing has added advantage that it acts data augmentation process increasing the total number of samples in the dataset.

(iv) Feature Selection – Feature selection in the study is done using the chi2 statistical test. The chi2 test allows for the identification of the most important attributes in the dataset. The test lists the attributes with their statistical significance. This study selects the top 20 best attributes for model training and the remaining are neglected.

(v) Data Splitting – Testing a supervised machine learning model requires the test data to be labeled. Generating new data is not a suitable process as it is manual and might require a lot of time to perform. Hence, the dataset at hand is divided into two sets viz. Training and testing data. This is generally done in the ratio of 80:20. Meaning, that 80% of the data is used as the training data, while the remaining 20% is the test data.

3.3 Data Modeling

The models that have been implemented in the study are discussed in detail below.

3.3.1 Decision Tree

The decision tree algorithm of machine learning works by building a sequential tree by comparing the information gained through each of the splits. This avoids the misclassification of the data. The process is sequential and each attribute is traversed with the others to identify their effects on the classification of the sample. The traversing of the decision tree starts at the decision node which is also called a root node. Each leaf node represents a class. The generation of a tree is a recursive process to distribute the attributes. The implementation of the model is further included in chapter 5 of this literature.

3.3.2 Random Forest Tree

The random forest classifier is a collection of decision tree classifiers that work collectively to classify the given data reliably. The classification of the random forest is obtained from voting or averaging the results from constituent decision trees. Random forest hence belongs to the ensemble class of classifiers that boost the performance of underlying classifiers and avoid overfitting of the data.

3.3.3 Support Vector Machines

Support vector machines are a classification model that defines a hyperplane such that it efficiently separates the training data. The training data point closest to the hyperplane are called support vectors. SVM is generally considered to be a linear binary classifier, but its use can be extended to the non-linear problem through the process of kernel trick wherein the normal linear data is mapped to higher dimension. The kernels that can be used for this are ‘Radial Basis Function (RBF)’, ‘Polynomial’ etc. It can also be used for multilabel classification through the use of ‘One vs One’ or ‘One vs All’ strategies.

3.3.4 Neural Networks

Artificial neural networks have the ability to mimic the pattern recognition abilities of a human brain. Similar to a brain, a neural network consists of neurons that are connected to each other through connections called synapses. The training on the neural network happens with the change of the weights contributing to the output of the network. The weight updation is performed through an algorithm known as backpropagation in which an error is found that corresponds to the deviation of the network output with respect to the expected result. This information is then passed on to the network. The weight updation takes place through equation 1 below.

$$wi(p+1) = wi(p) + alpha * xi * e(p)$$

Equation 1

Where wi(p) denotes the network weights at p^{th} stage, alpha is the learning rate, and e(p) is the error showing the difference between the actual output and the expected output.

3.4 Evaluation

The models that are implemented in the study are evaluated based on Accuracy, and F1-score.

3.4.1 Accuracy

Accuracy is the proportion of correct predictions among a number of predictions performed in the classification. Mathematically it can be written as below equation 2.

$$accuracy = \frac{(tp + tn)}{(tp + fp + fn + tn)}$$

Equation 2

$$\text{Where, } tp = true\ positives, tn = true\ negatives, fp = false\ positives, fn = false\ negatives$$

3.4.2 Precision

Precision is the proportion of all true positive samples among the samples classified as positive as per equation 3.

$$precision = \frac{tp}{tp + fp}$$

Equation 3

3.4.3 Recall

Similar to precision, recall is also a relevance-based metric. Recall is a proportion of true positive samples among the samples denoting the quality of a classification, which is shown in equation 4 below.

$$recall = \frac{tp}{tp + fn}$$

Equation 4

1 https://www.scikit-learn.org/

3.4.4 F1-score

F1-score is the harmonic mean of precision and recall in Equation 5 below.

$$F_1 = 2.\frac{precision.recall}{precision + recall}$$

Equation 5

3.5 Conclusion

The methodology to predict the employee attrition has been presented in the chapter using four machine learning techniques. The source of data collection as well as the necessary data transformation steps of the KDD methodology are discussed. The models presented are discussed in detail along with the evaluation metrics that are used to evaluated the presented models in the study. The upcoming chapters will discuss the design specifications of the system along with its implementation and the results obtained for the same.

4 Design Specifications

The process flow of the presented study is shown in figure 2 below. It shows the steps taken in obtaining the knowledge from the dataset. Python programming is the language of choice that is selected for the implementation of the study because of its simplicity and availability of a wide variety of libraries for the implementation of important aspects of the study.

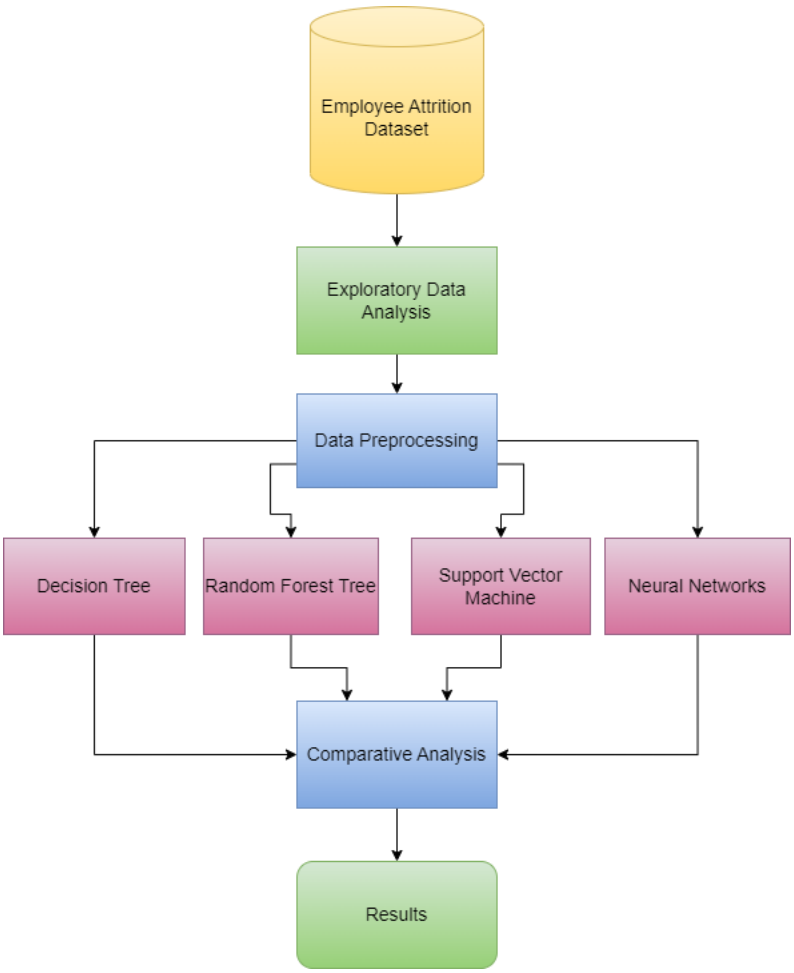


Figure 2: Process design flow for the study

Conclusion

The chapter presented the design process flow of the study that involves vital components necessary to fulfil the objectives of the study.

5 Implementation of Models in Employee Attrition Prediction

5.1 Introduction

The implementation of the methodology discussed in the previous chapter is discussed in this chapter. The study undertaken has been implemented in Anaconda Environment's Jupyterlab as a Jupyter Notebook using Python programming language using a number of functions from the libraries mentioned below.

1. pandas
2. matplotlib

3. sklearn
4. numpy

5.2 Implementation

5.2.1 Reading the dataset

The dataset chosen for the study is read using the *read_csv()* function of the pandas library. The function takes the filename as input along with the number of lines that need to be read. For the study, all the samples present in the dataset are read by not providing the number of lines that need to be read. Once the dataset is read, the data in the file is stored in python memory as a pandas dataframe object.

The data is then viewed using the *head()* function of the pandas library. By default, the function prints the first 5 rows of the dataframe. To view more lines, the number is given as an argument to the *head()* function. The data read using the *read_csv()* function is shown in figure 3 below.

	Employee_ID	Gender	Age	Education_Level	Relationship_Status	Hometown	Unit	Decision_skill_possess	Time_of_service	Time_since_promotion	...	Compensation_and_Benefits	Work_Life_balance	VAR1	VAR2	VAR3	VAR4	VAR5	VAR6	VAR7	Attrition_rate
0	EID_23371	F	42.0	4	Married	Franklin	IT	Conceptual	4.0	4	...	type2	3.0	4	0.7516	1.8688	2.0	4	5	3	0.1841
1	EID_18000	M	24.0	3	Single	Springfield	Logistics	Analytical	5.0	4	...	type2	4.0	3	-0.9612	-0.4537	2.0	3	5	3	0.0670
2	EID_3891	F	58.0	3	Married	Clinton	Quality	Conceptual	27.0	3	...	type2	1.0	4	-0.9612	-0.4537	3.0	3	8	3	0.0851
3	EID_17492	F	26.0	3	Single	Lebanon	Human Resource Management	Behavioral	4.0	3	...	type2	1.0	3	-1.8176	-0.4537	NaN	3	7	3	0.0668
4	EID_22534	F	31.0	1	Married	Springfield	Logistics	Conceptual	5.0	4	...	type3	3.0	1	0.7516	-0.4537	2.0	2	8	2	0.1827
5	EID_2278	M	54.0	3	Married	Lebanon	Purchasing	Conceptual	19.0	1	...	type2	1.0	3	-1.8176	1.8688	2.0	2	8	3	0.7613
6	EID_18588	F	21.0	4	Married	Springfield	Purchasing	Directive	2.0	1	...	type2	2.0	3	-0.9612	0.7075	2.0	3	7	3	0.2819
7	EID_1235	F	NaN	3	Married	Springfield	Sales	Directive	34.0	4	...	type3	2.0	3	-0.1048	-0.4537	2.0	3	9	3	0.1169
8	EID_10197	M	40.0	4	Single	Springfield	Production	Analytical	13.0	1	...	type0	4.0	1	NaN	1.8688	2.0	5	6	3	0.1968
9	EID_21262	M	45.0	3	Married	Lebanon	IT	Directive	21.0	4	...	type3	4.0	3	0.7516	-0.4537	2.0	4	8	3	0.2870

10 rows × 24 columns

Figure 3: Reading the data using head() function

5.2.2 Exploratory Data Analysis

Exploratory data analysis is an important step in the development of a data analysis study. EDA is a tool in which the data is visualized to get important insights into the data. For the presented study, the EDA is undertaken to identify, important attributes, and nulls present in the data and get an overview of the dataset.

The EDA module of the study is implemented mostly using the pandas and the matplotlib libraries. The pandas library is used to obtain insights into the dataset using the functions such as *describe()* and *info()*. The *describe()* function provides the statistical information about the dataframe which includes, count, mean, standard deviation, min value, max values, and 25, 50, and 75 percentile values. The *info()* function on the other hand provides information about the datatypes present under each of the attributes.

The visualization part of the EDA is performed using the matplotlib library of Python. This visualization is done using plots such as bar plot and pie plot. To implement the visualization using matplotlib, a figure object has to be initialized by providing the type of graph and the attribute to be visualized.

The results obtained using the above functions are shown in the below figures. Figure 4 below shows the output for *describe()* function. It shows that there are 17 numerical attributes and hence remaining 6 attributes are categorical meaning their values are not represented as numbers.

	Age	Education_Level	Time_of_service	Time_since_promotion	growth_rate	Travel_Rate	Post_Level	Pay_Scale	Work_Life_balance	VAR1	VAR2	VAR3	VAR4	VAR5	VAR6	VAR7	Attrition_rate
count	4713.000000	4999.000000	4903.000000	4999.000000	4999.000000	4999.000000	4999.000000	4992.000000	4989.000000	4999.000000	4595.000000	4999.000000	4524.000000	4999.000000	4999.000000	4999.000000	4999.000000
mean	39.632930	3.169834	13.378748	2.356071	46.964593	0.815163	2.799560	6.005208	2.382241	3.104221	-0.007690	-0.021416	1.891247	2.844569	7.086617	3.247650	0.187242
std	13.715201	1.074162	10.363130	1.150377	15.844159	0.648699	1.164119	2.050351	1.126703	0.835476	0.990397	0.997814	0.532078	0.939148	1.172540	0.925324	0.182154
min	19.000000	1.000000	0.000000	0.000000	20.000000	0.000000	1.000000	1.000000	1.000000	1.000000	-1.817600	-2.776200	1.000000	1.000000	5.000000	1.000000	0.000000
25%	27.000000	3.000000	5.000000	1.000000	33.000000	0.000000	2.000000	5.000000	1.000000	3.000000	-0.961200	-0.453700	2.000000	2.000000	6.000000	3.000000	0.070850
50%	37.000000	3.000000	10.000000	2.000000	47.000000	1.000000	3.000000	6.000000	2.000000	3.000000	-0.104800	-0.453700	2.000000	3.000000	7.000000	3.000000	0.142500
75%	52.000000	4.000000	21.000000	3.000000	61.000000	1.000000	4.000000	8.000000	3.000000	4.000000	0.751600	0.707500	2.000000	3.000000	8.000000	4.000000	0.234600
max	65.000000	5.000000	43.000000	4.000000	74.000000	2.000000	5.000000	10.000000	5.000000	5.000000	1.608100	1.868800	3.000000	5.000000	9.000000	5.000000	0.995900

Figure 4: Output of the describe() function for pandas dataframe

Figure 5 below shows the output for *info()* function. It shows that there are a number of attributes that contain some null values. Examples of such attributes are Age, Time_of_service, Pay_scale, Work_life_balance, VAR2, and VAR4. This can be seen from the non-null count in the figure. The null values in the dataset hence require to be operated upon to be useful in the processing.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4999 entries, 0 to 4998
Data columns (total 24 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Employee_ID               4999 non-null   object
 1   Gender                    4999 non-null   object
 2   Age                       4713 non-null   float64
 3   Education_Level           4999 non-null   int64
 4   Relationship_Status       4999 non-null   object
 5   Hometown                  4999 non-null   object
 6   Unit                      4999 non-null   object
 7   Decision_skill_possess    4999 non-null   object
 8   Time_of_service           4903 non-null   float64
 9   Time_since_promotion      4999 non-null   int64
 10  growth_rate               4999 non-null   int64
 11  Travel_Rate               4999 non-null   int64
 12  Post_Level                4999 non-null   int64
 13  Pay_Scale                 4992 non-null   float64
 14  Compensation_and_Benefits 4999 non-null   object
 15  Work_Life_balance         4989 non-null   float64
 16  VAR1                      4999 non-null   int64
 17  VAR2                      4595 non-null   float64
 18  VAR3                      4999 non-null   float64
 19  VAR4                      4524 non-null   float64
 20  VAR5                      4999 non-null   int64
 21  VAR6                      4999 non-null   int64
 22  VAR7                      4999 non-null   int64
 23  Attrition_rate            4999 non-null   float64
dtypes: float64(8), int64(9), object(7)
memory usage: 937.4+ KB
```

Figure 5: Output for the info() function for the pandas dataframe

The bar plot for the visualization of '*Gender'* attribute is presented in figure 6 below.

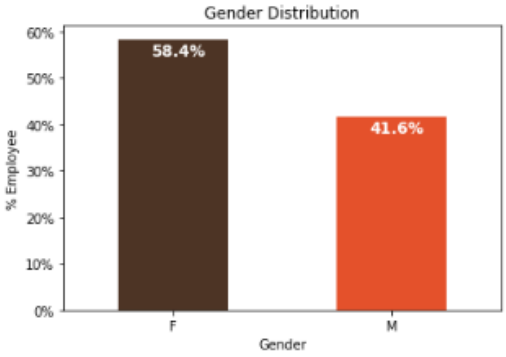


Figure 6: Distribution of values in the 'Gender' attribute

The visualization above shows that there are 58.4% of samples in the data corresponding to female employees whereas there are 41.6% of samples that correspond to male employees. The distribution of attribute values of 'Work_life_balance' can be visualized from figure 7 below.

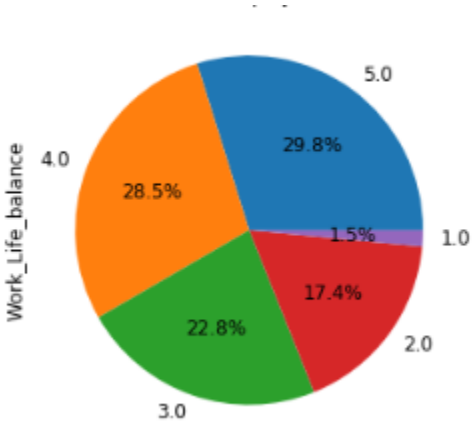


Figure 7: Pie-chart showing the distribution of values for the 'Work_life_balance' attribute.

From the above visualization figure 7, it can be seen that the values corresponding to the attribute are almost equally distributed, except the value 1 with a low percentage of 1.5%.

5.2.3 Data Transformation

Data pre-processing is another important step in any data analytics study. It is responsible for making the data adequately well-structured in order to get reliable and reproducible results. There are 4 important steps that have been performed in the pre-processing part of the study. These are,

Label Encoding:

Label encoding is a process of converting categorical variables into numerical form. This is essential for the machines to understand the data. The label encoding in the study is performed using the sklearn library available for python. *LabelEncoder()* function of the library creates a

label encoder object which is then applied to the attribute that need to be encoded. This is done using the *fit_transform()* function for the object. This function, first identifies the number of unique values in an attribute, and then replaces these non-integer values into integer values before mapping them respectively.

Labeling the Samples:
The attrition rate in the dataset gives a probability of attrition. To process the data, the attrition rate is rounded off to be either 0 or 1. This will give a label to each of the samples present in the dataset making it trainable by machine learning algorithm. This is done using the *round()* function of the Numpy library.

Oversampling:
Oversampling is a data augmentation technique, in which a dataset with imbalanced classes is modified such that the number of samples belonging to each class is equal. The oversampling in the study is performed using the *sample()* function of the pandas dataframe object. This function generates random samples of a given size by learning from the dataframe that is to be resampled. This process hence helps to obtain a balanced dataset which helps machine learning algorithms to learn better.

Feature Selection:
Feature selection is a process of identifying the most prominent features present in a dataset. It can be done through various processes, but the chi-square test is one of the most used ones. This study implements chi-square using the *chi2()* function available in the sklearn library's 'feature_selection' module. It takes the data from which the features need to be selected and the class labels associated with its samples.

Once the data is ready for processing, the data is modeled using the machine learning models mentioned in section 3. The modeling in the study is performed through two experiments. These are discussed below.

5.3 Experiment 1

This experiment involves modeling the data for the whole of the dataset and finding the best model suitable for predicting employee attrition shown in below Table 2. In this experiment, the models presented are used for modeling the whole dataset.

Table 2: Model parameters chosen for experiment 1

Model	Function (sklearn library)	Parameters
Decision Tree	DecisionTreeClassifier()	Default values specified by sklearn library
Random Forest	RandomForestClassifier()	N_estimators = 100, criterion = gini

Support Vector Machine	SVC()	C =1, kernel = rbf, degree = 3, gamma = ‘scale’
Artificial Neural Network	MLPClassifier()	Hidden_layer_sizes = (150,100,50); Max_iter = 600; activation = ‘tanh’; solver = ‘adam’

5.4 Experiment 2

This experiment involves modeling the data department-wise. The models mentioned are applied to model the data such as to predict employee attrition department-wise. In this experiment, the data is selected based on the department mentioned in the dataset. The class balancing and feature selection stages of data transformation are applied to the department-wise data. Table 3 below shows the number of samples obtained per department from the dataset after class balancing.

Table 3: Number of samples department-wise

Department	Number of Samples per class
IT	946
Logistics	778
Quality	129
HR	229
Purchasing	329
Sales	618
Production	138
Operations	438
Accounting	341
Marketing	142
RnD	429
Security	94
Total	*4611*

5.5 Conclusion

This section of the report presented the implementation of the system through two experiments in which the first experiment involves the implementation of the study for the prediction of employee attrition along the organization and the second experiment involves the implementation for the prediction of the employee attrition along the departments present in the organization. The upcoming evaluation section presents the results obtained for the experiments and evaluate them.

6 Evaluation and Results

This section of the study discusses the evaluation of the implemented methodology. The models presented are evaluated based on two parameters viz. Accuracy and f1-score. The evaluation of the models is performed for both the experiment performed in the study.

6.1 Evaluation and Results of Experiment 1

Table 4 below enlists the results obtained for the data modeling part of the study. The table lists the model name along with the data that has been used to perform the experiment, the accuracies obtained for each of the implemented model as well as the f1-score evaluation metric.

Table 4: Results for implemented models for whole dataset

Model	Data	Accuracy (%)	F1-score (%)
Decision Tree	Whole dataset	64.17	63.62
Random Forest	Whole dataset	64.39	63.96
Support Vector Machines	Whole dataset	50.02	65.72
Artificial Neural Network	Whole dataset	54.25	62.62

The table 4 clearly shows that the best results have been obtained for the random forest classifier with an accuracy of 64.39% and an f1-score of 63.96%. The decision tree classifier shows an accuracy of 64.17% and an f1-score of 63.62% which are almost the same as that of the random forest classifier. The lowest values of the accuracy metric have been obtained for the SVM classifier with a value of 50.02%, however, it has the highest F1-score of 65.72% among all the implemented models. This shows that the model is sensitive to the samples more than any other model. Having accuracy higher compared to the f1-score does not mean the model is better, it is the f1-score that needs to be higher. From the results above, hence it can be said that the best model for the prediction of attrition for the organization is SVM.

Figure 8 below shows the visual representation of the accuracies obtained for experiment 1.

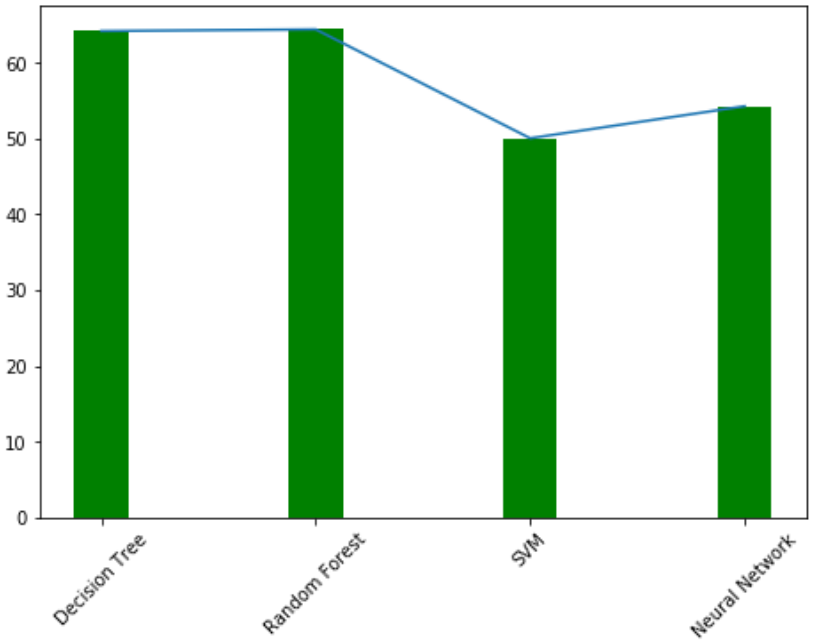


Figure 8: Accuracy comparison for the implemented models

Figure 9 below shows the visual representation of the f1-scores obtained for experiment 1.

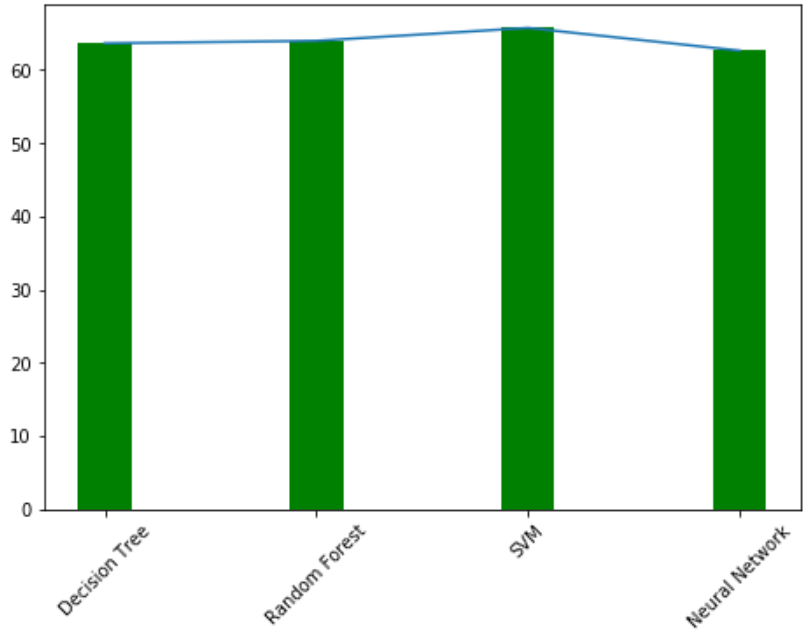


Figure 9: F1-scores of models implemented in the study

6.2 Evaluation and Results of Experiment 2

In this experiment, the models are used to predict employee attrition in each of the departments included in the dataset. This is done by the data from the dataset where the employees belong to

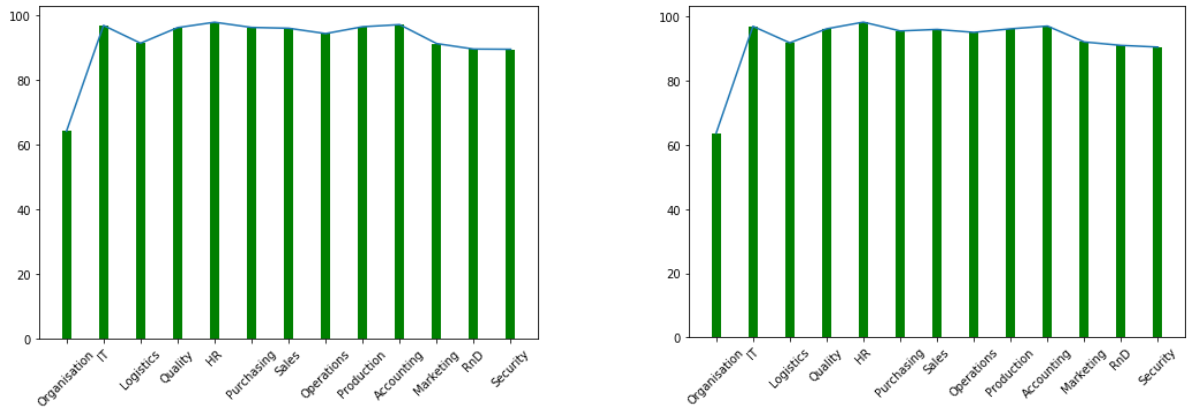
the selected department. There are a total of 13 departments in the dataset. Results obtained for the experiment are listed in table 5.

Table 5: Evaluation of models department-wise

Department	**Decision Tree**		**Random Forest**		**SVM**		**Neural Network**	
Metrics	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
IT	96.83	96.93	98.15	98.19	55.40	59.66	77.83	81.89
Logistics	91.34	91.79	91.67	92.07	62.82	63.97	85.25	86.39
Quality	96.15	96.15	**100**	**100**	50	53.57	98.07	98.03
HR	97.82	98.24	**100**	**100**	72.82	80.62	98.91	99.11
Purchasing	96.21	95.49	97.72	97.24	55.30	59.31	89.39	87.03
Sales	95.96	95.96	97.58	97.54	59.27	62.45	73.38	65.26
Production	96.42	96.15	98.21	98.03	51.78	55.73	84.65	86.15
Operations	94.32	95.05	96.59	96.96	54.54	43.66	100	100
Accounting	97.08	97.01	97.08	97.01	54.74	52.30	94.89	94.30
Marketing	91.22	92.06	98.24	98.30	59.64	68.49	94.73	94.73
RnD	89.53	91	91.27	92.38	58.72	59.88	73.25	75.78
Security	89.47	90.47	**100**	**100**	60.52	66.67	94.73	95

The models implemented are then trained on the selected data and tested on the test set of data. The results from the table 5 show that the random forest has achieved the best performance among the classifiers considering the accuracy and f1-score. The model achieved a 100% performance in the prediction of attrition in the departments of Quality, HR, and Security. This can be attributed to the ensemble architecture of the model that enhances the performances of the based models through aggregation or voting.

Figure 10 below shows the results for decision tree classifiers through graphs.

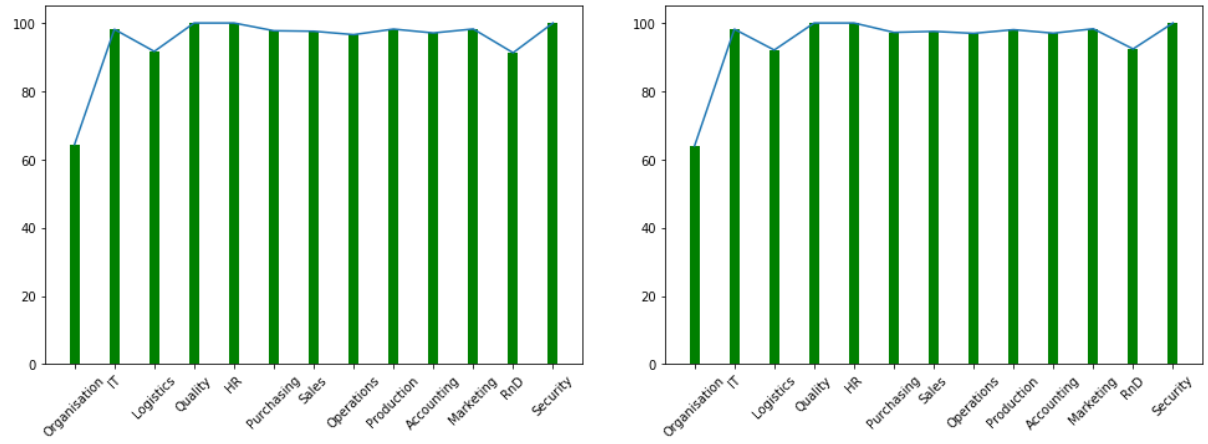


a) Accuracies of the decision tree department-wise

b) F1-scores for the decision tree department-wise

Figure 10: Results for decision tree classifier for department-wise attrition prediction

The decision tree classifier has achieved the best results for the accounting department. Figure 11 shows the results for the random forest classifier through graphs. The classifier performed very well across the departments except for the Logistics and RnD departments.

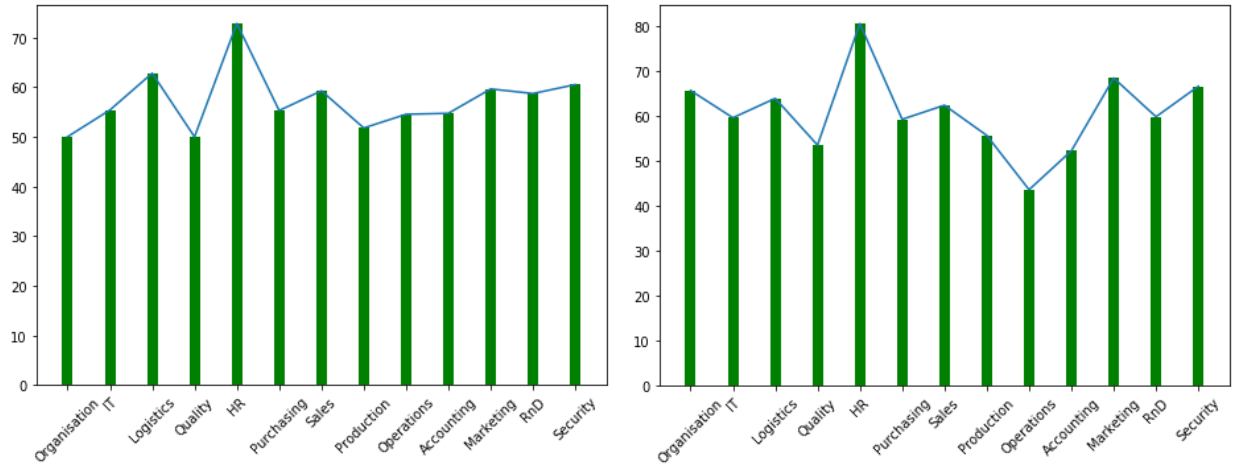


a) Accuracies of the random forest classifier department-wise

b) F1-scores for the random forest classifier department-wise

Figure 11: Results for random forest classifier

The SVM classifier has not performed well across the departments as seen in figure 12. It achieved the highest performance in the HR department.

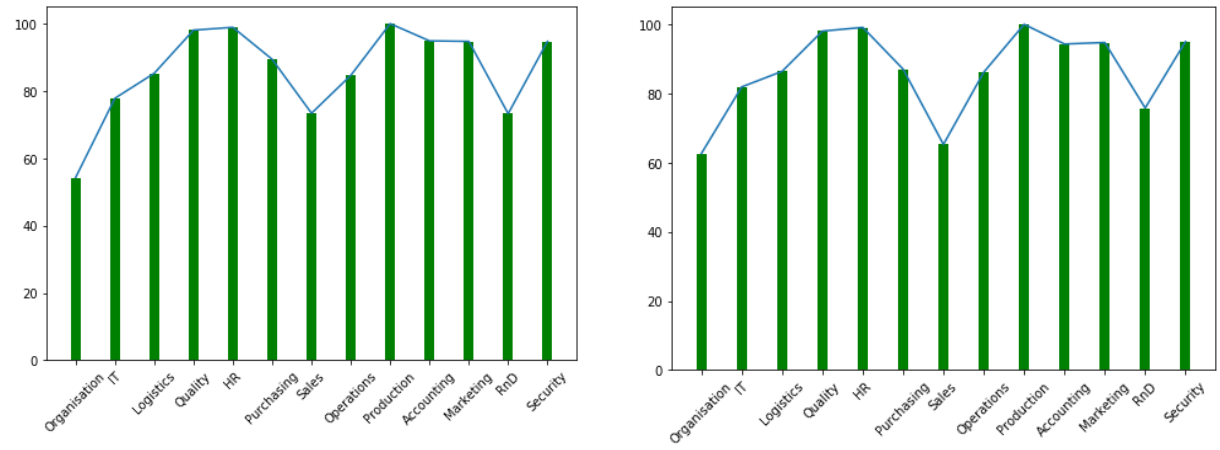


a) Accuracies of the SVM classifier department-wise

b) F1-scores for the SVM classifier department-wise

Figure 12: Results for SVM classifier

The performance of the neural network varies across the departments as seen in figure 13. The figure clearly shows that the model performance was the worst in the sales and RnD departments.



a) Accuracies of the neural network classifier department-wise

b) F1-scores for the neural network classifier department-wise

Figure 13: Results for neural network classifier

6.3 Conclusion

The results obtained for the study has answered the research question presented in section 1.2 of the report and also fulfilled the research objectives presented in section 1.3. The results obtained for the developed models significantly add to the knowledge in the field of employee attrition prediction leveraging the machine learning modalities.

7 Conclusion and Future Work

Employee attrition is an important issue that affects businesses and organizations across the world. This can be attributed to the fact that hiring new employees takes time and consumes a lot of resources. Hence, predicting if an employee is unhappy at an organization can help the organization take necessary steps to avert attrition. Hence, developing a model that can predict employee attrition is very beneficial for the organization and for the employee as well.
This research studied the performances of four machine-learning models viz. decision tree, random forest, SVM, and artificial neural network.

The models were implemented through two experiments in which the first experiment involved predicting the attrition for the whole organization whereas the second experiment involved predicting the attrition department-wise. The dataset that was chosen for the study was class imbalanced dataset hence, the dataset was class balanced by over-sampling the minority samples. The models implemented were evaluated based on accuracy score and f1-score and it was observed that the random forest model was the best model for predicting employee attrition across the organization as well as department-wise.

The limitations associated with the research have been that the dataset is small and the models implemented in the study are not hyperparameter tuned. The results obtained from the study however show a high misclassification rate, and hence the models need to be improved before implementing them in real-world scenarios.

Future Work

In future implementations of the study, a larger dataset can be selected in order to improve the reliability of the results obtained for the models. The models that have been used can be tuned properly meaning the parameters chosen for the models can be selected as optimum parameters which might improve the performances of the models. Modern machine learning modalities can also be tested with the dataset selected in the future work of the implemented study.

8 Acknowledgement

The research would not have been possible without the supervision, guidance, and support of Dr. Catherine Mulwa. For their trust in me, I would like to thank my mother, father, and my husband.

References

Ammar, M., Haleem, A., Javaid, M., Bahl, S. and Verma, A.S., 2022. Implementing Industry 4.0 technologies in self-healing materials and digitally managing the quality of manufacturing. Materials Today: Proceedings, 52, pp.2285-2294

Bishop, N., 2022, May. Manipulation of Machine Learning Algoirhtms. In Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (pp. 1833-1835).

Ebrahimi, P., Basirat, M., Yousefi, A., Nekmahmud, M., Gholampour, A. and Fekete-Farkas, M., 2022. Social Networks Marketing and Consumer Purchase Behavior: The Combination of SEM and Unsupervised Machine Learning Approaches. Big Data and Cognitive Computing, 6(2), p.35.

Guo, L.Z. and Li, Y.F., 2022, June. Class-imbalanced semi-supervised learning with adaptive thresholding. In International Conference on Machine Learning (pp. 8082-8094). PMLR.
Hellström, A., 2016. Machine learning in finance management: Case OpusCapita.

Shang, M., Li, H., Ahmad, A., Ahmad, W., Ostrowski, K.A., Aslam, F., Joyklad, P. and Majka, T.M., 2022. Predicting the Mechanical Properties of RCA-Based Concrete Using Supervised Machine Learning Algorithms. Materials, 15(2), p.647.

Sahija, D., 2021. Critical review of machine learning integration with augmented reality for discrete manufacturing. Independent Researcher and Enterprise Solution Manager in Leading Digital Transformation Agency, Plano, USA.

Chekroud, A.M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z., Belgrave, D., DeRubeis, R., Iniesta, R. and Dwyer, D., 2021. The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry, 20(2), pp.154-170.

Choudhury, P., Allen, R.T. and Endres, M.G., 2021. Machine learning for pattern discovery in management research. Strategic Management Journal, 42(1), pp.30-57.

Subhashini, M. and Gopinath, R., 2021. Employee Attrition Prediction in Industry using Machine Learning Techniques.

Zhou, Z.H., 2021. *Machine learning*. Springer Nature.

Kühl, N., Goutier, M., Hirt, R. and Satzger, G., 2020. Machine learning in artificial intelligence: Towards a common understanding. arXiv preprint arXiv:2004.04686.

Lee, I. and Shin, Y.J., 2020. Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons*, *63*(2), pp.157-170.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M. and Eckersley, P., 2020, January. Explainable machine learning in deployment. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 648-657).

Fallucchi, F., Coladangelo, M., Giuliano, R. and William De Luca, E., 2020. Predicting employee attrition using machine learning techniques. *Computers*, *9*(4), p.86.

Aziz, S. and Dowling, M., 2019. Machine learning and AI for risk management. In Disrupting finance (pp. 33-50). Palgrave Pivot, Cham.

Mason, K. and Grijalva, S., 2019. A review of reinforcement learning for autonomous building energy management. Computers & Electrical Engineering, 78, pp.300-312.

Ray, S., 2019, February. A quick review of machine learning algorithms. In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon) (pp. 35-39). IEEE.

Singh, D., 2019. A literature review on employee retention with focus on recent trends. *International Journal of Scientific Research in Science and Technology*, *6*(1), pp.425-431.

Frye, A., Boomhower, C., Smith, M., Vitovsky, L. and Fabricant, S., 2018. Employee Attrition: What Makes an Employee Quit?. *SMU Data Science Review*, *1*(1), p.9.

García-Peñalvo, F.J., Cruz-Benito, J., Martín-González, M., Vázquez-Ingelmo, A., Sánchez-Prieto, J.C. and Therón, R., 2018. Proposing a machine learning approach to analyze and predict employment and its factors.

Spanoudes, P. and Nguyen, T., 2017. Deep learning in customer churn prediction: unsupervised feature learning on abstract company independent feature vectors. arXiv preprint arXiv:1703.03869.

Iqbal, S., Guohao, L. and Akhtar, S., 2017. Effects of job organizational culture, benefits, salary on job satisfaction ultimately affecting employee retention. *Review of Public Administration and Management*, *5*(3), pp.1-7.

Ribes, E., Touahri, K. and Perthame, B., 2017. Employee turnover prediction and retention policies design: a case study. arXiv preprint arXiv:1707.01377.

Showry, M. and Manasa, K.V.L., 2016. Attrition Among the New Hires: A Soft Skill Perspective. *IUP Journal of Soft Skills*, *10*(4).

Tanwar, K. and Prasad, A., 2016. Exploring the relationship between employer branding and employee retention. *Global business review*, *17*(3_suppl), pp.186S-206S.

Wang, H., Lei, Z., Zhang, X., Zhou, B. and Peng, J., 2016. Machine learning basics. Deep learning, pp.98-164.

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