EMPLOYEE ATTRITION PREDICTION USING SUPERVISED MACHINE LEARNING MODELS

A PROJECT REPORT

Submitted by

ROHITH SS	**(190801065)**
SAILESH BAABU S	**(190801068)**

in partial fulfilment for the award

of the degree

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

SRI VENKATESWARA COLLEGE OF ENGINEERING

(An Autonomous Institution; Affiliated to Anna University, Chennai-600025)

ANNA UNIVERSITY:: CHENNAI 600 025

JUNE 2023

SRI VENKATESWARA COLLEGE OF ENGINEERING

(An Autonomous Institution; Affiliated to Anna University, Chennai-600025)

ANNA UNIVERSITY, CHENNAI - 600 025

BONAFIDE CERTIFICATE

Certified that this project report **"EMPLOYEE ATTRITION PREDICTION USING SUPERVISED MACHINE LEARNING MODELS"** is the bonafide work of **"ROHITH SS (190801065) and SAILESH BAABU S (190801068)"** who carried out the project work under my supervision.

SIGNATURE	**SIGNATURE**
Dr. V. VIDHYA, M.E., Ph.D.,	**Ms. N. UMA, M.E.,**
HEAD OF THE DEPARTMENT	**SUPERVISOR**
	ASSISTANT PROFESSOR
INFORMATION TECHNOLOGY	INFORMATION TECHNOLOGY

Submitted for the project viva-voce examination held on ………………..

INTERNAL EXAMINER	**EXTERNAL EXAMINER**

ABSTRACT

In the era of data science and big data analytics, people analytics help organizations and their human resources (HR) managers to reduce attrition by changing the way of attracting and retaining talent. Employee retention is a major challenge for recruiters and employers alike, since employee attrition means not only the loss of skills, experiences and personnel but also the loss of business opportunities. Employee attrition presents a critical problem and a big risk for organizations as it affects not only their productivity but also their planning continuity. Employee attrition or voluntary turnover presents a key issue for organizations as it affects not only their productivity and work sustainability but also their long-term growth strategies. Firstly, we propose a people analytics approach to predict employee attrition that shifts from a big data to a deep data context by focusing on data quality instead of its quantity. In fact, this deep data-driven approach is based on a mixed method to construct a relevant employee attrition model in order to identify key employee features influencing his/her attrition. In this method, we started thinking 'big' by collecting most of the common features (an exploratory research) then we tried thinking 'deep' by filtering and selecting the most important features using survey and feature selection algorithms.

ACKNOWLEDGEMENT

We thank our Principal **Dr. S. Ganesh Vaidyanathan, M.E., Ph.D.,** Sri Venkateswara College of Engineering for being the source of inspiration throughout our study in this college.

We express our sincere thanks to **Dr. V. Vidhya, M.E., Ph.D.,** Head of the Department, Computer Science and Engineering for her encouragement accorded to carry this project.

We are also thankful to **Dr. G. Sumathi , M.E., PhD., & Dr. N. Gobalakrishnan, M.Tech., Ph.D.,** project coordinators for their continual support and assistance throughout the course of this project.

With profound respect, we express our deep sense of gratitude and sincere thanks to our guide **Ms. N. Uma**, **M.E.**, for her valuable guidance and suggestions throughout this project.

We also express our thanks to all Faculty members, Department of Information Technology, for rendering their support.

ROHITH SS
SAILESH BAABU S

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

ML	Machine Learning
XGBOOST	Extreme Gradient Boosting
HR	Human Resources
EDA	Exploratory Data Analysis
CNN	Convolutional Neural Network
SVM	Support Vector Machine
ECPR	Employee Prediction Retention
HRI	Human Resource Information
EWM	Entropy Weight Method
EIM	Employee Importance Model

CHAPTER 1

INTRODUCTION

1.1 DATA ANALYTICS

Data analytics is the science of analysing raw data to make conclusions about that information. Data analytics help a business optimize its performance, perform more efficiently, maximize profit, or make more strategically-guided decisions. The techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption. Various approaches to data analytics include looking at what happened (descriptive analytics), why something happened (diagnostic analytics), what is going to happen (predictive analytics), or what should be done next (prescriptive analytics). Data analytics relies on a variety of software tools ranging from spreadsheets, data visualization, and reporting tools, data mining programs, or open-source languages for the greatest data manipulation. Data analytics is a broad term that encompasses many diverse types of data analysis. Any type of information can be subjected to data analytics techniques to get insight that can be used to improve things. Data analytics techniques can reveal trends and metrics that would otherwise be lost in the mass of information. This information can then be used to optimize processes to increase the overall efficiency of a business or system. Data analytics is the process of exploring and analysing large datasets to find hidden patterns, unseen trends, discover correlations, and derive valuable insights to make business predictions. It improves the speed and efficiency of your business.

Data Analytics eliminates guesswork and manual tasks. Be it choosing the right content, planning marketing campaigns, or developing products.

Organizations can use the insights they gain from data analytics to make informed decisions. Thus, leading to better outcomes and customer satisfaction. Data analytics allows you to tailor customer service according to their needs. It also provides personalization and builds stronger relationships with customers. Analyzed data can reveal information about customers' interests, concerns, and more. It helps you give better recommendations for products and services. With the help of data analytics, you can streamline your processes, save money, and boost production. With an improved understanding of what your audience wants, you spend lesser time creating ads and content that aren't in line with your audience's interests. Data analytics gives you valuable insights into how your campaigns are performing. This helps in fine-tuning them for optimal outcomes. Additionally, you can also find potential customers who are most likely to interact with a campaign and convert into leads.

1.2 TYPES OF DATA ANALYTICS

1.2.1 Descriptive Analytics

Descriptive Analytics is the simplest type of analytics and the foundation the other types are built on. It allows you to pull trends from raw data and succinctly describe what happened or is currently happening. Descriptive analytics answers the question, "What happened?"

1.2.2 Diagnostic Analytics

Diagnostic analytics addresses the next logical question, “Why did this happen?” Taking the analysis a step further, this type includes comparing coexisting trends or movement, uncovering correlations between variables, and determining causal relationships where possible.

1.2.3 Predictive Analytics

Predictive analytics is used to make predictions about future trends or events and answers the question, “What might happen in the future?” By analyzing historical data in tandem with industry trends, you can make informed predictions about what the future could hold for your company.

1.2.4. Prescriptive Analytics

Finally, prescriptive analytics answers the question, “What should we do next?” Prescriptive analytics takes into account all possible factors in a scenario and suggests actionable takeaways. This type of analytics can be especially useful when making data-driven decisions.

1.3 MACHINE LEARNING

Machine Learning is said as a subset of artificial intelligence that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own. With the help of sample historical data, which is known as training data, machine learning algorithms build a mathematical model that helps in making predictions or decisions without being explicitly programmed. Machine learning brings computer science and statistics together for creating predictive models. Machine learning constructs or uses the algorithms that learn from historical data. The more we will provide the information, the higher will be the performance. A Machine Learning system learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it**.** The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately. The machine learning algorithms can be trained by providing them the huge amount of data and let them explore the data, construct the models, and predict the required output automatically. The performance of the machine learning algorithm depends on the amount of data, and it can be determined by the cost function. With the help of machine learning, we can save both time and money. The reason behind the need for machine learning is that it is capable of doing tasks that are too complex for a person to implement directly. As a human, we have some limitations as we cannot access the huge amount of data manually, so for this, we need some computer systems and here comes the machine learning to make things easy for us. In a complex problem, where the need to perform some predictions, so instead of writing a code for it, just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output. Machine learning has changed our way of thinking about the problem. The below block diagram explains the working of Machine Learning algorithm:

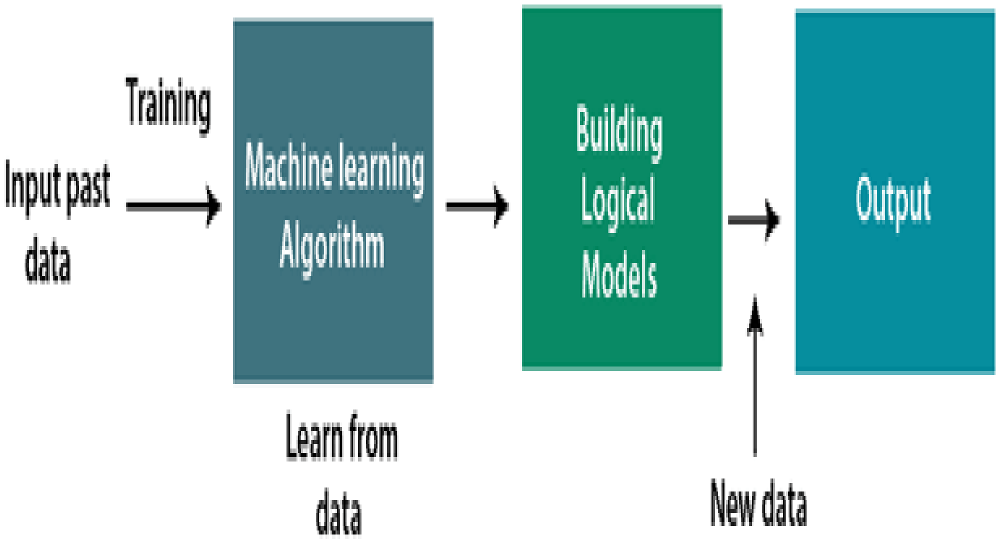


Figure 1.1 Workflow of a Machine Learning Model

1.4 CLASSIFICATION OF MACHINE LEARNING

1.4.1) Supervised Learning

Supervised learning, also known as supervised machine learning, is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately. It is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

This occurs as part of the cross-validation process to ensure that the model avoids overfitting or underfitting. Supervised learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox. In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher. Some methods used in supervised learning include neural networks, naïve bayes, linear regression, logistic regression, random forest, support vector machine (SVM), and more.

In the real-world, supervised learning can be used for Risk Assessment, Image classification, Fraud Detection, spam filtering, etc.

Supervised learning can be grouped further in two categories of algorithms:

- **Classification**
- **Regression**

1.4.2) Unsupervised Learning

Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. Unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Instead, models itself find the hidden patterns and insights from the given data. It can be compared to learning which takes place in the human brain while learning new things.

Its ability to discover similarities and differences in information make it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation, image and pattern recognition.

It’s also used to reduce the number of features in a model through the process of dimensionality reduction; principal component analysis (PCA) and singular value decomposition (SVD) are two common approaches for this. Other algorithms used in unsupervised learning include neural networks, k-means clustering, probabilistic clustering methods, and more. Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format**.**

It can be further classifieds into two categories of algorithms:

- **Clustering**
- **Association**

1.4.3) REINFORCEMENT LEARNING

Reinforcement learning works on a feedback-based process, in which an AI agent (A software component) automatically explore its surrounding by hitting & trail, taking action, learning from experiences, and improving its performance**.** Agent gets rewarded for each good action and get punished for each bad action; hence the goal of reinforcement learning agent is to maximize the rewards.

In reinforcement learning, there is no labelled data like supervised learning, and agents learn from their experiences only.

The reinforcement learning process is similar to a human being; for example, a child learns various things by experiences in his day-to-day life. An example of reinforcement learning is to play a game, where the Game is the environment, moves of an agent at each step define states, and the goal of the agent is to get a high score. Agent receives feedback in terms of punishment and rewards.

Due to its way of working, reinforcement learning is employed in different fields such as Game theory, Operation Research, Information theory, multi-agent systems.

A reinforcement learning problem can be formalized using Markov Decision Process (MDP). In MDP, the agent constantly interacts with the environment and performs actions; at each action, the environment responds and generates a new state.

1.5 PROBLEM STATEMENT

In the era of data science and big data analytics, people analytics help organizations and their human resources (HR) managers to reduce attrition by changing the way of attracting and retaining talent. The employee attrition presents a critical problem and a big risk for organizations as it affects not only their productivity but also their planning continuity.

The objective is to develop a relevant employee attrition model in order to identify key employee features influencing his/her attrition.

This analytics approach to predict employee attrition that shifts from a big data to a deep data context by focusing on data quality instead of its quantity. The deep data-driven approach is based on a mixed method to construct a relevant employee attrition model in order to identify key employee features influencing his/her attrition. This attrition prediction is based on supervised machine learning models.

In particular, we have adhered to the following classifiers: Decision Tree, Random Forest, Logistic Regression and Support Vector Machine (as machine learning models)

CHAPTER 2

LITERATURE REVIEW

[1] R. Punnoose et al. (2019) identified employee turnover as a key issue. To solve this problem, organizations use machine learning techniques to predict employee turnover. Accurate predictions enable organizations to take action for retention or succession planning of employees. However, the data for this modeling problem comes from HR Information Systems (HRIS); these are typically under-funded compared to the Information Systems of other domains in the organization which are directly related to its priorities. This leads to the prevalence of noise in the data that renders predictive models prone to over-fitting and hence inaccurate. The novel contribution of this paper is to explore the application of Extreme Gradient Boosting (XGBoost) technique which is more robust because of its regularization formulation. Data from the HRIS of a global retailer is used to compare XGBoost against six historically used supervised classifiers and demonstrate its significantly higher accuracy for predicting employee turnover. Pre-processing steps for the dataset used in this comparative study include data exploration, data visualization, data cleaning and reduction, data transformation, discretization, and feature selection. In this study, parameter tuning and regularization techniques to overcome overfitting issues are applied for optimization purposes.

[2] P. Likhitkar et **al. (2020)** explored the role of predictive analytics in human resource management domain. Human resource management undergoes a drastic change due to digitization.

In the current competitive environment, talented employee is undoubtedly the most valuable assets of the organization. The value of predictive analytics is proposed that uses analytics in effective decision-making process in the organization without any biasness.

The study also checked the mediating effect of work engagement in the relationships of Green HRM practices with employee retention. Following the current purpose, the study targeted employees in the pharmaceutical industry as the unit of analyses. Five hundred seventy-six respondents were selected through simple random sampling. Data were collected through a self-administrative questionnaire. Collected data was screened through SPSS23, which concluded with 349 usable questionnaires for data analysis and assessment. Structural equation modeling via Smart PLS 3.2.8 was employed to test the proposed model. The findings of the study show a positive but insignificant association of green HRM practices toward employee retention.

[3] Shikha N Khera et al. (2019) developed a model to predict employee attrition and provide the organizations opportunities to address any issue and improve retention. Predictive model was developed based on supervised machine learning algorithm, support vector machine (SVM). Archival employee data (consisting of 22 input features) were collected from Human Resource databases of three IT companies in India, including their employment status (response variable) at the time of collection. Accuracy results from the confusion matrix for the SVM model showed that the model has an accuracy of 85 per cent. Also, results show that the model performs better in predicting who will leave the firm as compared to predicting who will not leave the company. The aim of this research is to develop a model to predict employee attrition and provide the organizations opportunities to address any issue and improve retention. Predictive model was developed based on supervised machine learning algorithm, support vector machine (SVM). Archival employee data (consisting of 22 input features) were collected from Human Resource databases of three IT companies in India, including their employment status (response variable) at the time of collection. Accuracy results from the confusion matrix for the SVM model showed that the model has an accuracy of 85 per cent. Also, results show that the model performs better in predicting who will leave the firm as compared to predicting who will not leave the company.

[4] F. Fallucchi et al. (2020) intended to analyse how objective factors influence employee attrition, in order to identify the main causes that contribute to a worker's decision to leave a company, and to be able to predict whether a particular employee will leave the company. After the training, the obtained model for the prediction of employees' attrition is tested on a real dataset provided by IBM analytics, which includes 35 features and about 1500 samples.

Results are expressed in terms of classical metrics and the algorithm that produced the best results to identify the main causes that contribute to a worker's decision to leave a company, for the available dataset is the Gaussian Naïve Bayes classifier. It reveals the best recall rate (0.54), since it measures the ability of a classifier to find all the positive instances and achieves an overall false negative rate equal to 4.5% of the total observations. To this aim, we applied some machine learning techniques in order to identify the factors that may contribute to an employee leaving the company and, above all, to predict the likelihood of individual employees leaving the company. First, we assess statistically the data and then we classified them. To evaluate the algorithm's performance, the predicted results were collected and fed into the respective confusion matrices. From these it was possible to calculate the basic metrics necessary for an overall evaluation (precision, recall, accuracy, f1 score, ROC curve, AUC, etc.) and to identify the most suitable classifier to predict whether an employee was likely to leave the company. Results obtained by the proposed automatic predictor demonstrate that the main attrition variables are monthly income, age, overtime, distance from home.

[5] S. R. Ponnuru et al. (2020) presented a binary classification technique as the term to be predicted is whether a particular employee leaves company or not. This paper can help to predict the voluntary attrition of the employee in a company by considering some of the factors like Age, Job Satisfaction, Monthly Income, Years At Company. The data of the employee is sourced from Kaggle by IBM HR analytics. As employee attrition or voluntary turnover is a nonavoidable phenomenon, modelling it is a key issue for the process of attrition prediction. In addition, as we aim to adopt a deep data-driven approach, a research methodology that allows us to match theoretical models and experiments must

be adopted. That’s why we propose to conduct a mixed research method based on the combination of an exploratory research and a quantitative method where the aim is to understand and explain employee attrition phenomena. Unlike the past, in which there was a post-action response to deal with employee attrition, there is now the possibility of taking pre-emptive action by predicting the possibility of employee attrition through AI in advance.

[6] N. Jain et al. (2021) proposed a multi-attribute decision making (MADM) based scheme coupled with ML algorithms. The proposed scheme is referred as employee churn prediction and retention (ECPR). An accomplishment-based employee importance model (AEIM) that utilizes a two-stage MADM approach for grouping the employees in various categories. Preliminarily, we formulate an improved version of the entropy weight method (IEWM) for assigning relative weights to the employee accomplishments. Then, we utilize the technique for order preference by similarity to ideal solution (TOPSIS) for quantifying the importance of the employees to perform their class-based categorization. The CatBoost algorithm is then applied for predicting class-wise employee churn. Finally, we propose a retention policy based on the prediction results and ranking of the features. The proposed ECPR scheme is tested on a benchmark dataset of the human resource information system (HRIS), and the results are compared with other ML algorithms using various performance metrics. We show that the system using the CatBoost algorithm outperforms other ML algorithms. Employee attrition is a very critical issue from the organization's standpoint because it places a considerable burden on the organization for a wide range of issues: interruption of ongoing tasks, costs for employee re-employment and retraining (Yedida et al., 2018), risks of leaking core technologies and know-hows, etc. (Bennett et al., 1993). In particular, for organizational leaders, reducing employee attrition has a profound effect on organizational management and the improvement of organizational culture (Qutub et al., 2021).

[7] J. Gu et al. (2018) worked on deep neural networks, convolutional neural networks. Leveraging on the rapid growth in the amount of the annotated data and the great improvements in the strengths of processor units, the research on convolutional neural networks has been emerged swiftly and achieved state-of-the-art results on various tasks. In this paper, a broad survey of the recent advances in convolutional neural networks. We detailize the improvements of CNN on different aspects, including layer design, loss function, regularization, optimization and fast computation. Besides, we also introduce various applications of convolutional neural networks in speech and natural language processing. A review of the findings shows that logistic regression and random forest have had effective performance, with each being selected in various studies. At the same time, a lot of studies have considered boosting-based algorithms such as XGBoost, AdaBoost, and gradient boosting, too, showed an acceptable performance (Gabrani and Kwatra, 2018; Jain and Nayyar, 2018;Qutub et al., 2021;Srivastava and Eachempati, 2021;Yadav et al., 2018). The idea of the aggregation of classifiers was proposed by Breiman (1996), who believed that their combination could increase the overall accuracy of the model.

[8] X. Gao et al.(2019) proposed a weighted quadratic random forest algorithm which is applied to employee turnover data with high-dimensional unbalanced characteristics. First, the random forest algorithm is used to order feature importance and reduce dimensions. Second, the selected features are used with the random forest algorithm and the F-measure values are calculated for each decision tree as weights to build the prediction model for employee turnover. In the area of employee turnover forecasting, compared with the random forest, C4.5, Logistic, BP, and other algorithms, the proposed algorithm shows significant improvement in terms of various performance indicators, specifically recall and F-measure. In the area of employee turnover forecasting, compared with the random forest, C4.5, Logistic, BP, and other algorithms, the proposed algorithm shows significant improvement in terms of various performance indicators, specifically recall and F-measure. In the experiment using the employee dataset of a branch of a communications company in China, the key factors influencing employee turnover were identified as monthly income, overtime, age, distance from home, years at the company, and percent of salary increase. Among them, monthly income and overtime were the two most important factors. The study offers a new analytic method that can help human resource departments predict employee turnover more accurately and its experimental results provide further insights to reduce employee turnover intention.

[9] P. Ajit et al.(2019) identified employee turnover as a key issue. To solve this problem, organizations use machine learning techniques to predict employee turnover. Accurate predictions enable organizations to take action for retention or succession planning of employees. However, the data for this modeling problem comes from HR Information Systems (HRIS); these are typically under-funded compared to the Information Systems of other domains in the organization which are directly related to its priorities. This leads to the prevalence of noise in the data that renders predictive models prone to over-fitting and hence inaccurate. The novel contribution of this paper is to explore the application of Extreme Gradient Boosting (XGBoost) technique which is more robust because of its regularization formulation.

Employee's Attrition Prediction Using Machine Learning Approaches Timely delivery of any service or product is the primary goal of any organization in recent days due to high competition in industries. If a talented employee leaves unexpectedly, the company is not able to complete the task at defined times. It may become the reason for the loss of that company. Therefore, companies are interested in knowing the employee's attrition. They can make a proper substitute or ar-rangements earlier.There may be various reasons for employee attrition, which include less salary, job satisfaction, personal reasons, or environmental issues if the employer terminates an employee for any reason. It is known as involuntary attrition (Kaur & Vijay, 2016). On the other hand, voluntary attrition is known as the left of an employee by their side. This kind of attrition is a loss for the company if he or she is a talented employee. In the present scenario, everyone wants a higher salary and job security. Therefore, employees leave jobs immediately if they got a better chance in other places.

[10] M. Coladangelo et al.(2020) intended to analyse how objective factors influence employee attrition, in order to identify the main causes that contribute to a worker’s decision to leave a company, and to be able to predict whether a particular employee will leave the company. After the training, the obtained model for the prediction of employees’ attrition is tested on a real dataset provided by IBM analytics, which includes 35 features and about 1500 samples. Results are expressed in terms of classical metrics and the algorithm that produced the best results to identify the main causes that contribute to a worker’s decision to leave a company, for the available dataset is the Gaussian Naïve Bayes classifier. In this paper, we perform an analysis of the reasons or motivations that push an employee to leave the company and consequently allow the HR department to take timely appropriate countermeasures such as improving the work environment or production incentives. Starting from the dataset, we identify the main factors related to the employee’s attrition and we propose a real classification, based on the statistical evaluation of the data. The application of classification algorithms can support the HR management by allowing the adoption of staff management support tools in the company. The obtained model for the prediction of employees’ attrition is tested on a real dataset provided by IBM analytics, which includes 35 features and about 1500 samples. By analysing the correlations in the heatmap of 35 features, we derive the characteristics that have high correlations related to the reasons that an employee leaves the company. 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). The results obtained from the data analysis demonstrate that the adoption of machine learning systems can support the HR department in the company staff management. The paper is organised as follows.

[11] C. Zhang et al.(2020) proposed a weighted quadratic random forest algorithm which is applied to employee turnover data with high-dimensional unbalanced characteristics. First, the random forest algorithm is used to order feature importance and reduce dimensions. Second, the selected features are used with the random forest algorithm and the F-measure values are calculated for each decision tree as weights to build the prediction model for employee turnover. In the area of employee turnover forecasting, compared with the random forest, C4.5, Logistic, BP, and other algorithms, the proposed algorithm shows significant improvement in terms of various performance indicators, specifically recall and F-measure. The findings of the study help the practitioners including the firm and contribute to the literature in several ways. First, the firm will be able to identify the determinant of employee turnover during the present turbulent pandemic, and take suitable steps to mitigate the adverse effect of the issue. The present rate of turnover has drastically influenced the function and performance of the organization thus, the findings will be useful to the organization to take appropriate corrective measures to retain its competitiveness and ensure survival. Also, the retained skilled employees would help maintain higher productivity and profitability. Second, the findings will help understand the effect of certain factors over employee turnover in distinct contexts particularly, during a turbulent time and present pandemic. It will help resolve inconsistencies in the literature. Thereby, this will contribute to strengthening the present literature. Organizations that face a similar issue in similar environments may understand and come up with appropriate solutions to cope up with the issues. The rest of the paper is organized as below. The next section is allocated to review literature related to the factors influencing employee turnover, followed by the methodology adopted in the study. The next section presents the results while the final section is devoted to discussion and conclusion.

CHAPTER 3

PROPOSED WORK

3.1 OBJECTIVE

Organizations and their human resources (HR) managers try to reduce attrition by changing the way of attracting and retaining talent. The employee attrition presents a critical problem and a big risk for organizations as it affects not only their productivity but also their planning continuity. Employee attrition prediction is tackled as a supervised learning problem, and in particular, as a binary classification one. In other words, we are interested in detecting and confirming the existence or not of the employee's intention to leave.

The objective is to develop a relevant employee attrition model in order to identify key employee features influencing his/her attrition. The deep data-driven approach is based on a mixed method tno construct a relevant employee attrition model in order to identify key employee features influencing his/her attrition. This attrition prediction is based on supervised machine learning models.

In particular, we have adhered to the following classifiers: Decision Tree, Random Forest, Logistic Regression and Support Vector Machine (as machine learning models).

3.2 PROPOSED ARCHITECTURE

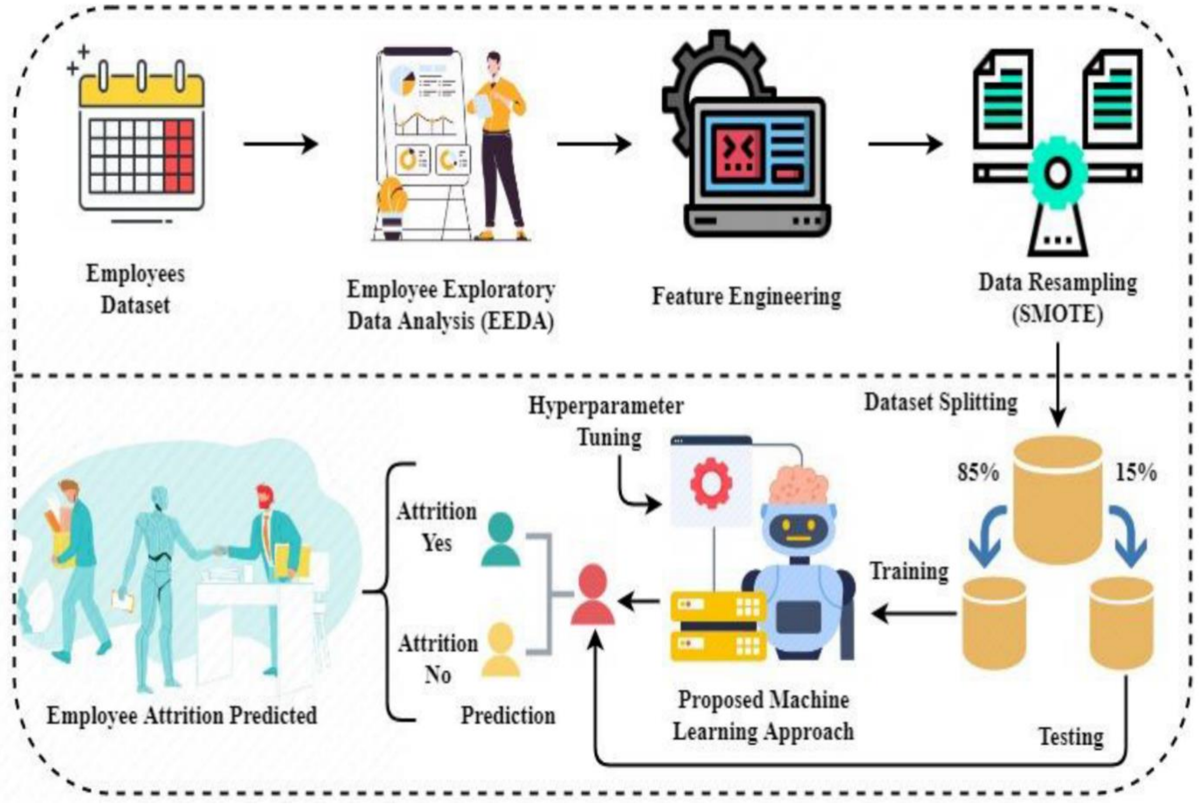


Figure 3.1 System Architecture

The data is collected and then moved to the data pre-processing stage to train the model. The pre-processing is done in the initial stages to remove missing data and Inconsistent data. The presence of missing data and inconsistent data can produce biased estimates leading to invalid conclusions. Data pre-processing include data cleaning, data integration, data transformation, and data reduction.

In data cleaning stage, the noisy and inconsistent data are removed. In the data integration stage, data from various sources are put into a coherent data store. In data transformation stage, normalizations are applied to improve the accuracy and efficiency of the algorithms.

After the pre-processing stage the dataset is ready for the descriptive analysis phase. Descriptive analytics are used to summarize or turn data into relevant information so investigate what has occurred. In other words, descriptive analytics have some meaningful impact by explaining what has already happened however, they are not much helpful in predicting what will happen or may happen in the future. Descriptive analysis can be categorized into four types which are measures of frequency, central tendency, dispersion or variation, and position. These methods are optimal for a single variable at a time.

Exploratory data analysis (EDA) is done on the dataset. EDA aims to spot patterns and trends, to identify anomalies, and to test early hypotheses. EDA focuses on understanding the characteristics of a dataset before deciding what we want to do with that dataset. EDA provides invaluable insights that an algorithm cannot. You can think of this a bit like running a document through a spellchecker versus reading it yourself.

While software is useful for spotting typos and grammatical errors, only a critical human eye can detect the nuance. An EDA is similar in this respect—tools can help you, but it requires our own intuition to make sense of it. This personal, in-depth insight will support detailed data analysis further down the line.

Initial data analysis (IDA) can help you spot any structural issues with your dataset. You may be able to fix these, or you might find that you need to reprocess the data or collect new data entirely. While this can be a nuisance, it’s better to know upfront, before you dive in with a deeper analysis. Before diving in with a full analysis, it’s important to make sure any assumptions or hypotheses you’re working on stand up to scrutiny. While an EDA won’t give you all the details, it will help you spot if you’re inferring the right outcomes based on your understanding of the data. If not, then you know that your assumptions are wrong, or that you are asking the wrong questions about the dataset.

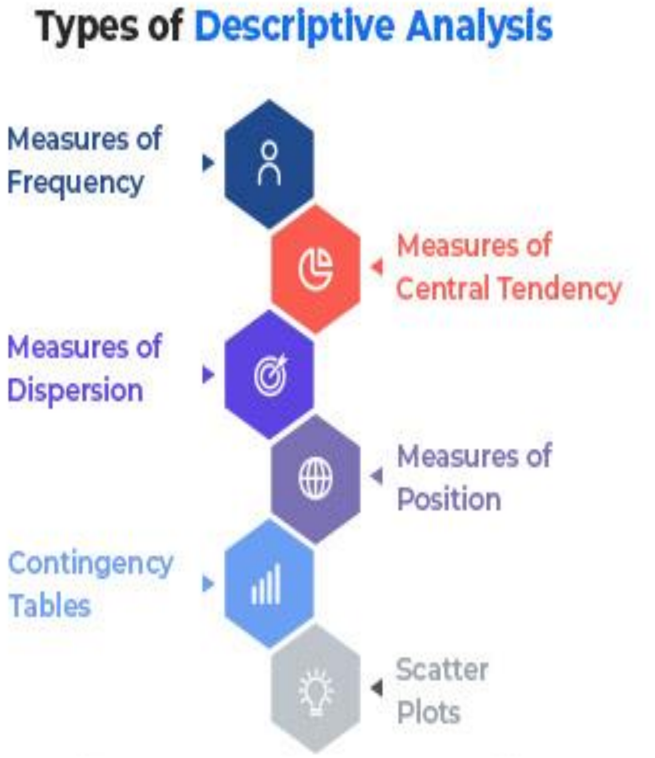


Figure 3.2 Types of Descriptive Analysis

Exploratory data analytics uses visual techniques, such as graphs, plots, and other visualizations. This is because our natural pattern-detecting abilities make it much easier to spot trends and anomalies when they’re represented visually. EDA isn’t just about finding helpful information. It’s also about determining which data might lead to unavoidable errors in your later analysis.

Knowing which data will impact your results helps you to avoid wrongly accepting false conclusions or incorrectly labeling an outcome as statistically significant when it isn't. As a simple example, outliers (or data points that skew a trend) stand out much more immediately on a scatter graph than they do in columns on a spreadsheet.

Perhaps the most practical outcome of EDA is that it will help you determine which techniques and statistical models will help you get what you need from your dataset. For instance, do you need to carry out a predictive analysis or a sentiment analysis? An EDA will help you decide.

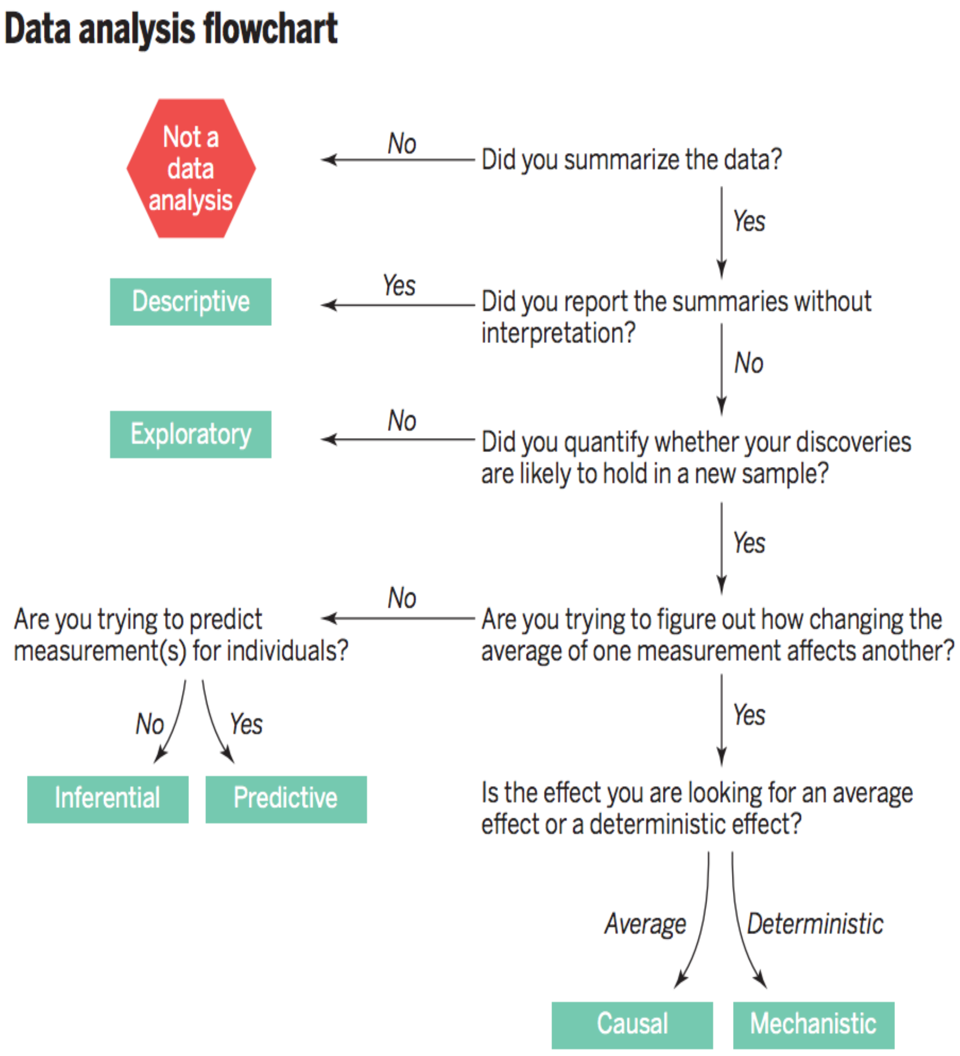

Data analysis flowchart
Not a data analysis
No
Did you summarize the data?
Yes
Descriptive
Yes
Did you report the summaries without interpretation?
No
Exploratory
No
Did you quantify whether your discoveries are likely to hold in a new sample?
Yes
Are you trying to predict measurement(s) for individuals?
No
Are you trying to figure out how changing the average of one measurement affects another?
No
Yes
Yes
Inferential
Predictive
Is the effect you are looking for an average effect or a deterministic effect?
Average
Deterministic
Causal
Mechanistic

Figure 3.3 Data Analysis Flowchart

Next step is building the machine learning model for prediction. Building an ML model requires splitting of data into two sets, such as 'training set' and 'testing set' in the ratio of 80:20 or 70:30. It is performed using library scikit-learn. A set of supervised (for labelled data) and unsupervised (for unlabeled data) algorithms are available to choose from depending on the nature of input data and business outcome to predict. The supervised learning models are used in this problem. In supervised learning, models are trained using labelled dataset, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output. With the help of supervised learning, the model can predict the output on the basis of prior experiences. In case of labelled data, it is recommended to choose logistic regression algorithm if the outcome to predict is of binary (0/1) in nature; choose decision tree classifier (or) Random Forest® classifier (or) KNN if the outcome to predict is of multi-class (1/2/3/...) in nature.

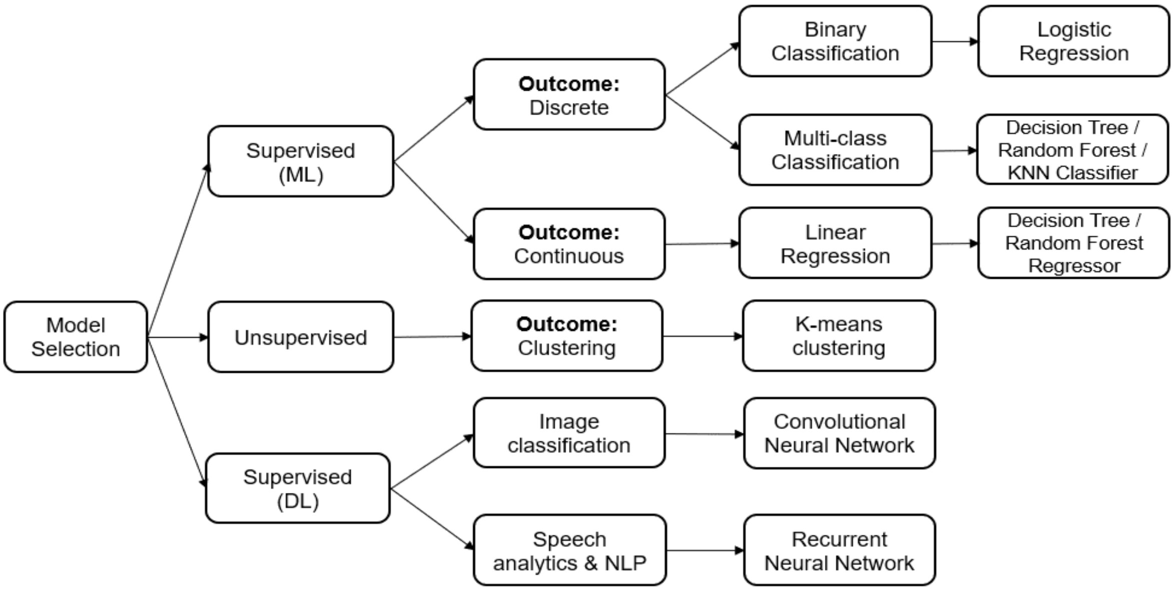


Figure 3.4 Types of Machine Learning Models

Since our problem requires classification of whether an employee may leave or not we will be using supervised machine learning algorithms

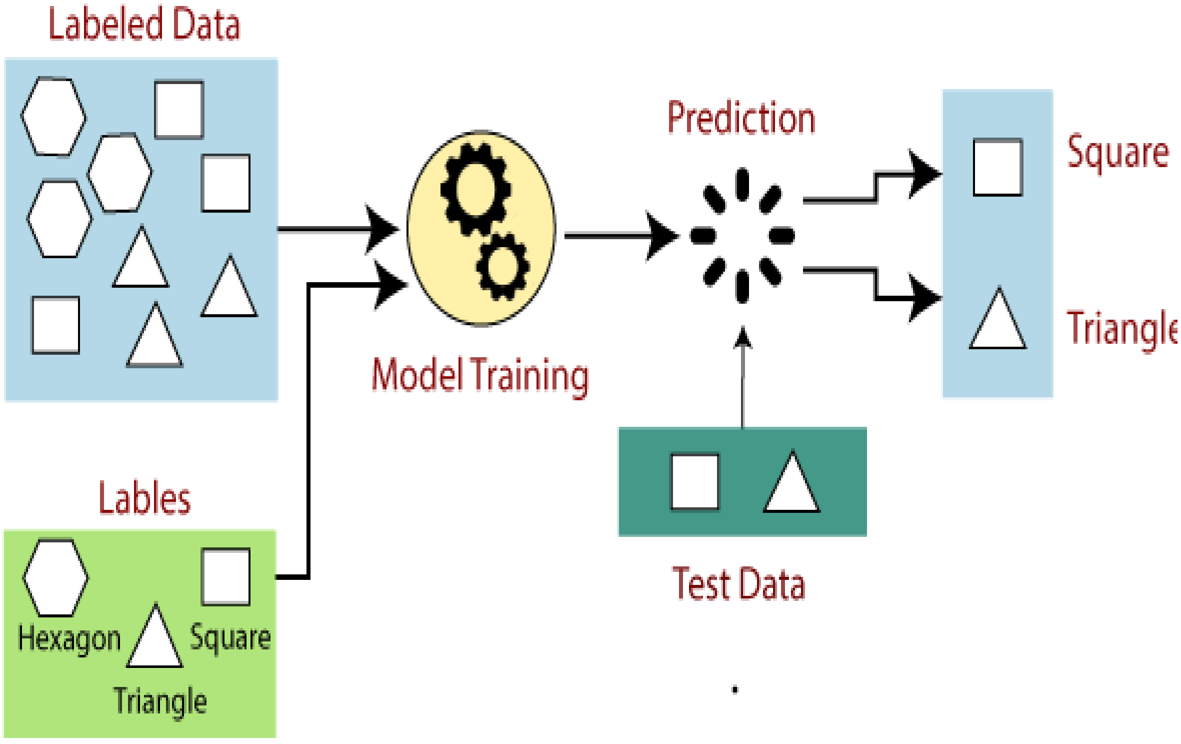


Figure 3.5 Working of Supervised Learning

The models are created from the supervised machine learning algorithms and the data is tested, thus classifying whether the employee has the intention to leave or not. In supervised learning, models are trained using labelled dataset, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output.

First Determine the type of training dataset. Collect /Gather the labelled training data. Split the training dataset into training **dataset, test dataset, and validation dataset**. Determine the input features of the training dataset, which should have enough knowledge so that the model can accurately predict the output. Determine the suitable algorithm for the model, such as support vector machine, decision tree, etc.

Execute the algorithm on the training dataset. Sometimes we need validation sets as the control parameters, which are the subset of training datasets. Evaluate the accuracy of the model by providing the test set. If the model predicts the correct output, which means our model is accurate.

Different supervised machine learning models are used such as logistic regression, support vector machines, decision tree and random forest.

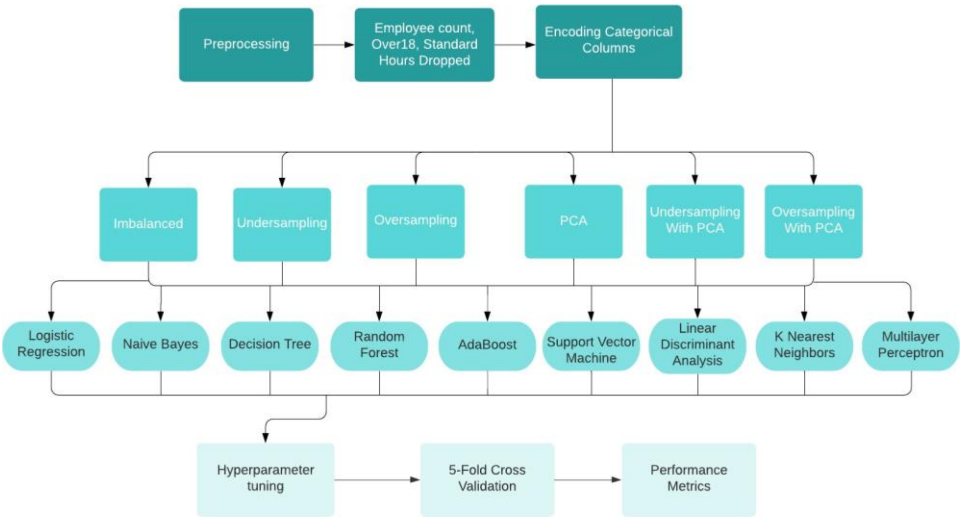


Figure 3.6 Supervised Learning Flowchart

CHAPTER 4

REQUIREMENT SPECIFICATIONS

4.1 HARDWARE SPECIFICATION

- 4GB RAM
- DUAL CORE PROCESSOR
- 64 BIT WINDOWS OPERATING SYSTEM

4.2 SOFTWARE SPECIFICATION

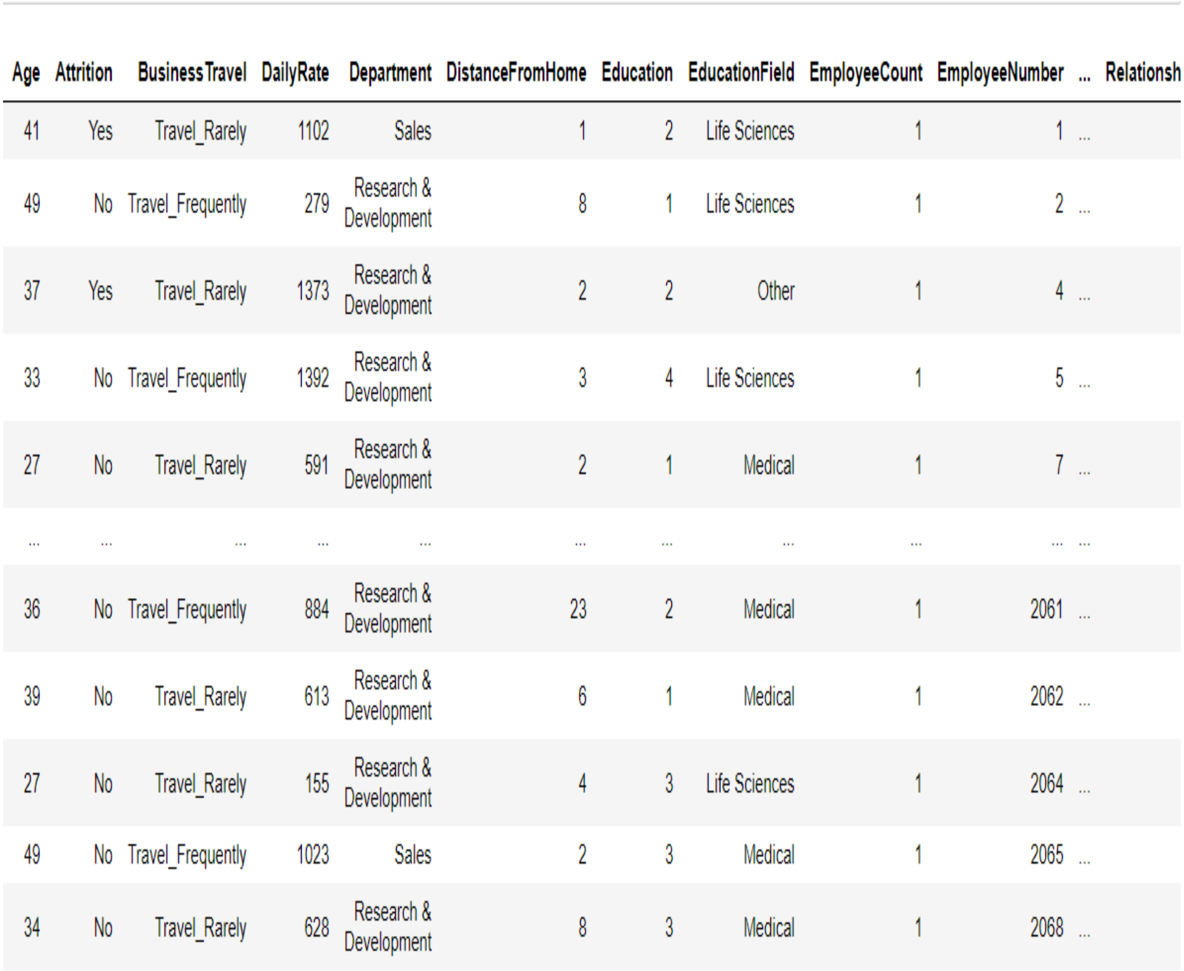
- PYTHON
- PANDAS
- JUPYTER NOTEBOOK
- MATPLOTLIB
- SCIKIT-LEARN
- PLOTLY

CHAPTER 5

IMPLEMENTATION MODULES

5.1 DATASET

The dataset contains 34 input features.



Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	...	Relationsh
41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	...	
49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	...	
37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	...	
33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	...	
27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	...	
...	...	...	...	...	...	...	...	...	...	...	
36	No	Travel_Frequently	884	Research & Development	23	2	Medical	1	2061	...	
39	No	Travel_Rarely	613	Research & Development	6	1	Medical	1	2062	...	
27	No	Travel_Rarely	155	Research & Development	4	3	Life Sciences	1	2064	...	
49	No	Travel_Frequently	1023	Sales	2	3	Medical	1	2065	...	
34	No	Travel_Rarely	628	Research & Development	8	3	Medical	1	2068	...	

Figure 5.1 Data Collection

5.2 EXPLORATORY DATA ANALYSIS

Distribution graphs for features were analyzed. Some inferences are discussed below.

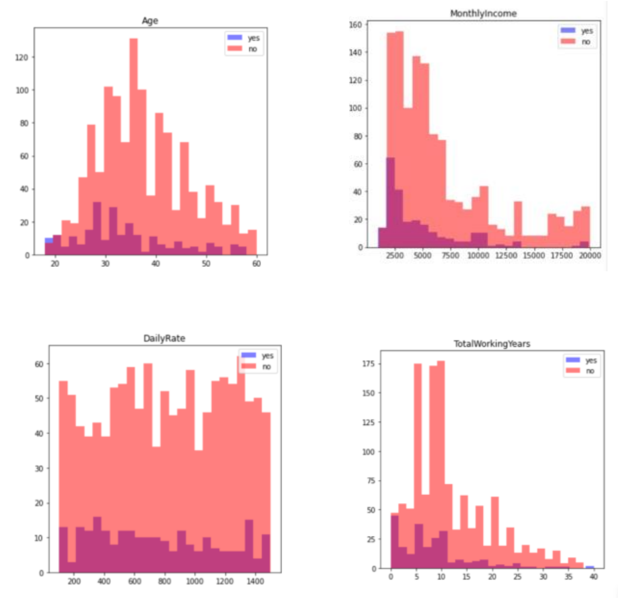


Figure 5.2. Graphs for Exploratory Data Analysis

Employees around the age of 28 appear to be more prone to leave the organisation, as shown in Fig. 3. Higher attrition rates were correlated with lower monthly revenue. Although employees with less than 10 years of service had higher attrition rates, newer employees had the greatest attrition rates. If they work above their scheduled hours, they are more likely to quit. More personnel who travel regularly experience attrition. Compared to other positions, sales executives are more likely to depart the organization. No gender-based differences in attrition were found to be significant.

5.3 DATA PRE-PROCESSING

Data pre-processing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms.

Pre-processing of data is mainly to check the data quality. The quality can be checked by the following

- **Accuracy**: To check whether the data entered is correct or not.
- **Completeness**: To check whether the data is available or not recorded.
- **Consistency:** To check whether the same data is kept in all the places that do or do not match.
- **Timeliness**: The data should be updated correctly.
- **Believability**: The data should be trustable.
- **Interpretability**: The understandability of the data.

Major Tasks in Data Pre-processing:

1. Data cleaning
2. Data integration
3. Data reduction
4. Data transformation

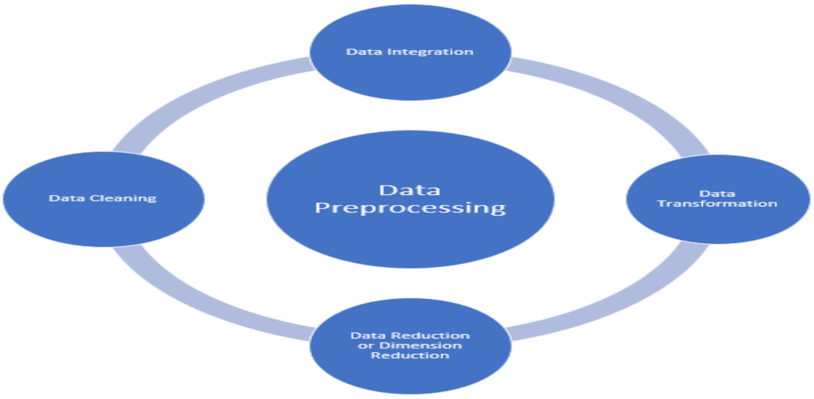


Figure 5.3 Data Pre-processing Modules

Handling missing values:

- Standard values like “Not Available” or “NA” can be used to replace the missing values.
- Missing values can also be filled manually but it is not recommended when that dataset is big.
- The attribute’s mean value can be used to replace the missing value when the data is normally distributed.

There are no missing/null values in the dataset. To visualize the distribution of different features, we plot bar graphs. Using these, we observe that the features ’Employ- eeCount’, ’Over18’, and ’StandardHours’ have only one unique value and hence add no value to attrition prediction. Thus, they are dropped. Employee number is varying for each row and is not related to the attrition column and is also dropped.

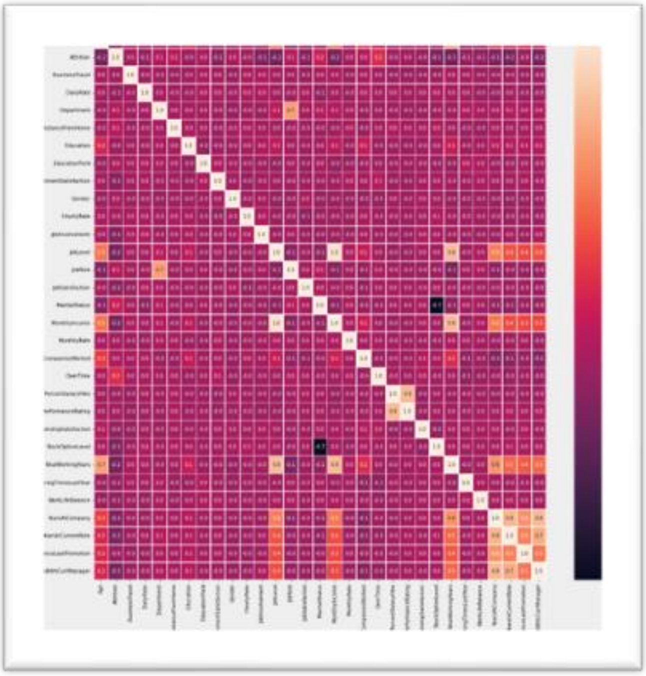


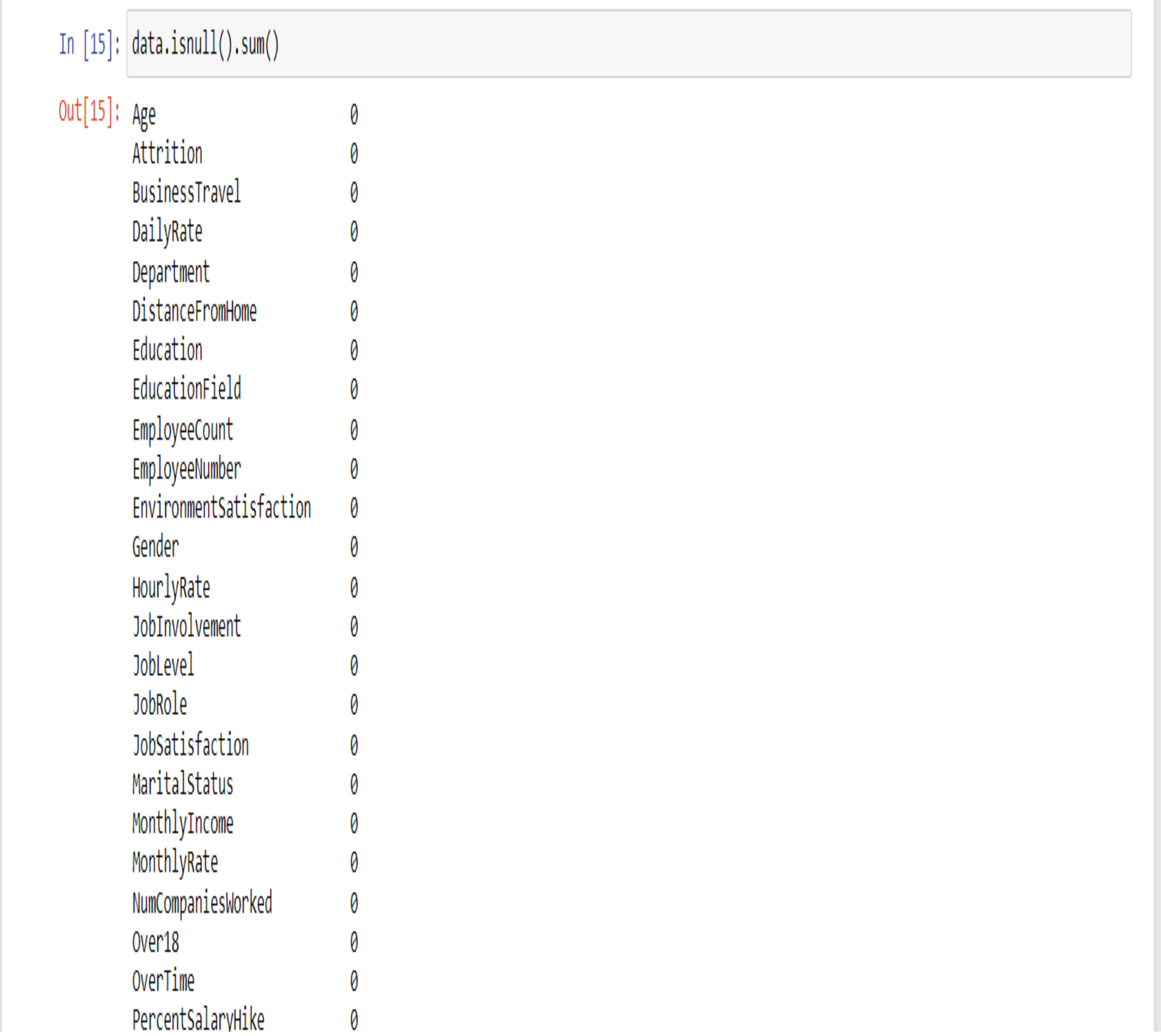
Figure 5.4. Correlation Matrix

Data Transformation:

The change made in the format or the structure of the data is called data transformation. This step can be simple or complex based on the requirements. There are some methods in data transformation.

- **Smoothing**: With the help of algorithms, we can remove noise from the dataset and helps in knowing the important features of the dataset. By smoothing we can find even a simple change that helps in prediction.
- **Aggregation**: In this method, the data is stored and presented in the form of a summary. The data set which is from multiple sources is integrated into with data analysis description. This is an important step since the accuracy of the data depends on the quantity and quality of the data.

The dataset does not have any missing values or duplicate values. If present, it can be handled by either removing it or apply the mean or mode methods. Pre-processing of data is mainly to check the data quality.



```
In [15]: data.isnull().sum()

Out[15]: Age                         0
         Attrition                   0
         BusinessTravel              0
         DailyRate                   0
         Department                  0
         DistanceFromHome            0
         Education                   0
         EducationField              0
         EmployeeCount               0
         EmployeeNumber              0
         EnvironmentSatisfaction     0
         Gender                      0
         HourlyRate                  0
         JobInvolvement              0
         JobLevel                    0
         JobRole                     0
         JobSatisfaction             0
         MaritalStatus               0
         MonthlyIncome               0
         MonthlyRate                 0
         NumCompaniesWorked          0
         Over18                      0
         OverTime                    0
         PercentSalaryHike           0
```

Figure 5.5 Data Pre-processing

5.4 DATA VISUALIZATION

Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from. Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics, and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

Data visualization is one of the steps of the data science process, which states that after data has been collected, processed and modeled, it must be visualized for conclusions to be made. Data visualization is also an element of the broader data presentation architecture (DPA) discipline, which aims to identify, locate, manipulate, format and deliver data in the most efficient way possible.

Data visualization is important for almost every career. It can be used by teachers to display student test results, by computer scientists exploring advancements in artificial intelligence (AI) or by executives looking to share information with stakeholders.

As businesses accumulated massive collections of data during the early years of the big data trend, they needed a way to quickly and easily get an overview of their data. Visualization tools were a natural fit.

Visualization is central to advanced analytics for similar reasons. When a data scientist is writing advanced predictive analytics or machine learning (ML) algorithms, it becomes important to visualize the outputs to monitor results and ensure that models are performing as intended.

The dataset contains 34 features. With respect to these features univariate and multivariate analysis has been done and the visualizations have been made. Different types of charts are used to show the insights in a better way.

1) The frequency of the attrition is counted and plotted as a pie chart for a better visualization.

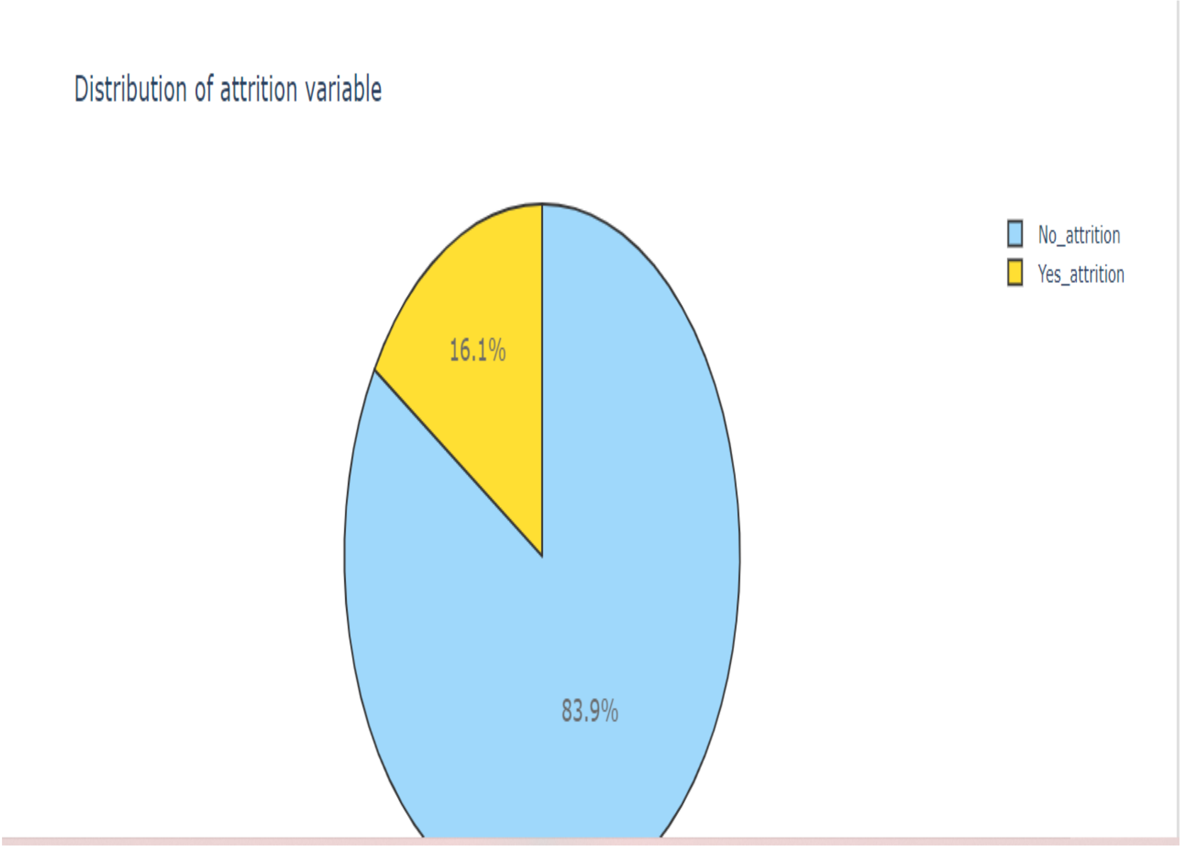
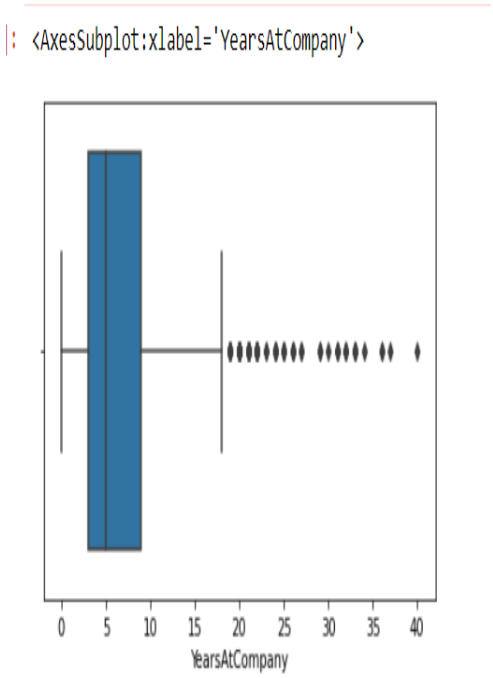


Figure 5.6 Distribution of Attrition

2) The outliers with respect to the 'yearsatcompany' are shown with the help of a boxplot.



```
: <AxesSubplot:xlabel='YearsAtCompany'>
```

Figure 5.7 Distribution of Outliers

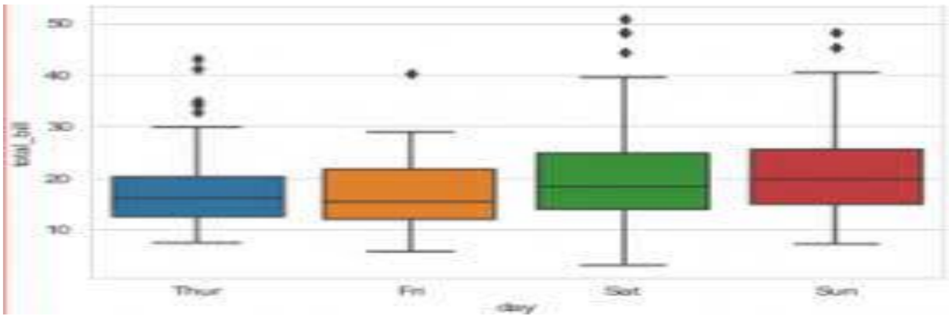


Figure 5.7 Architecture of Box Plot

3) The effect of age on attrition is shown with the help of a bar graph and a line graph.

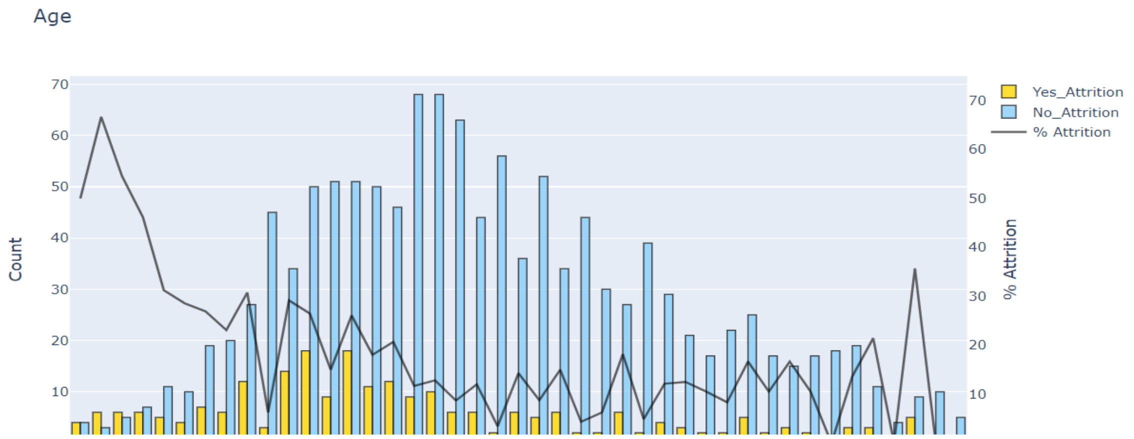
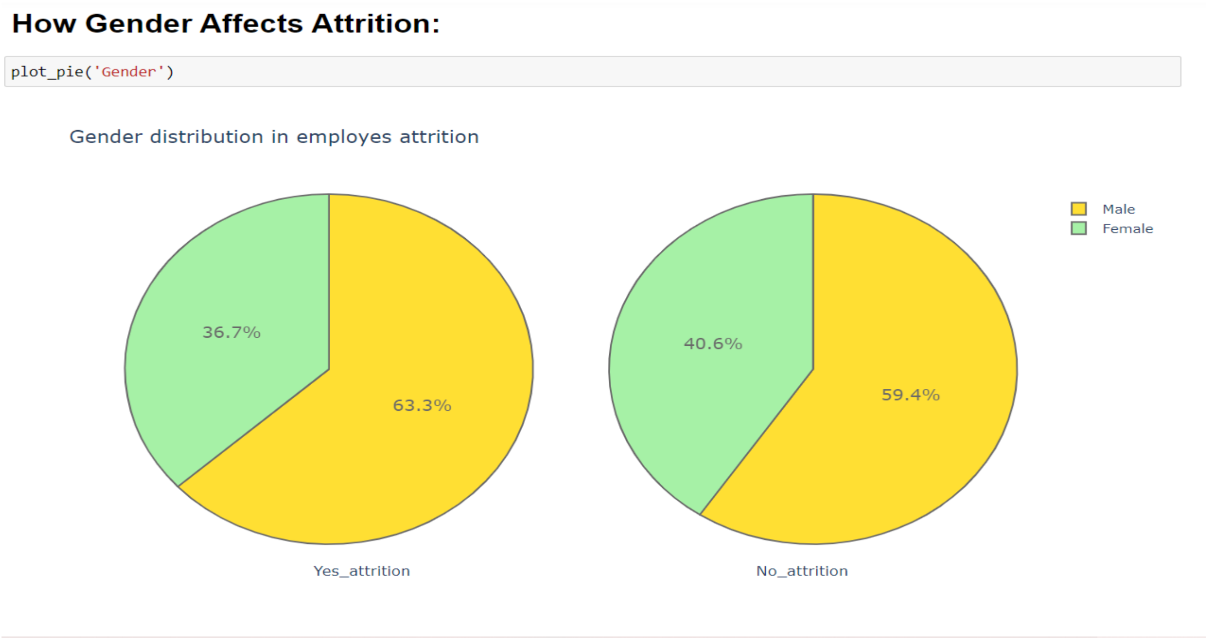


Figure 5.8 Distribution of Age on Attrition

4) The effect of gender on attrition is shown with the help of pie charts as comparison becomes easier.



How Gender Affects Attrition:

```
plot_pie('Gender')
```

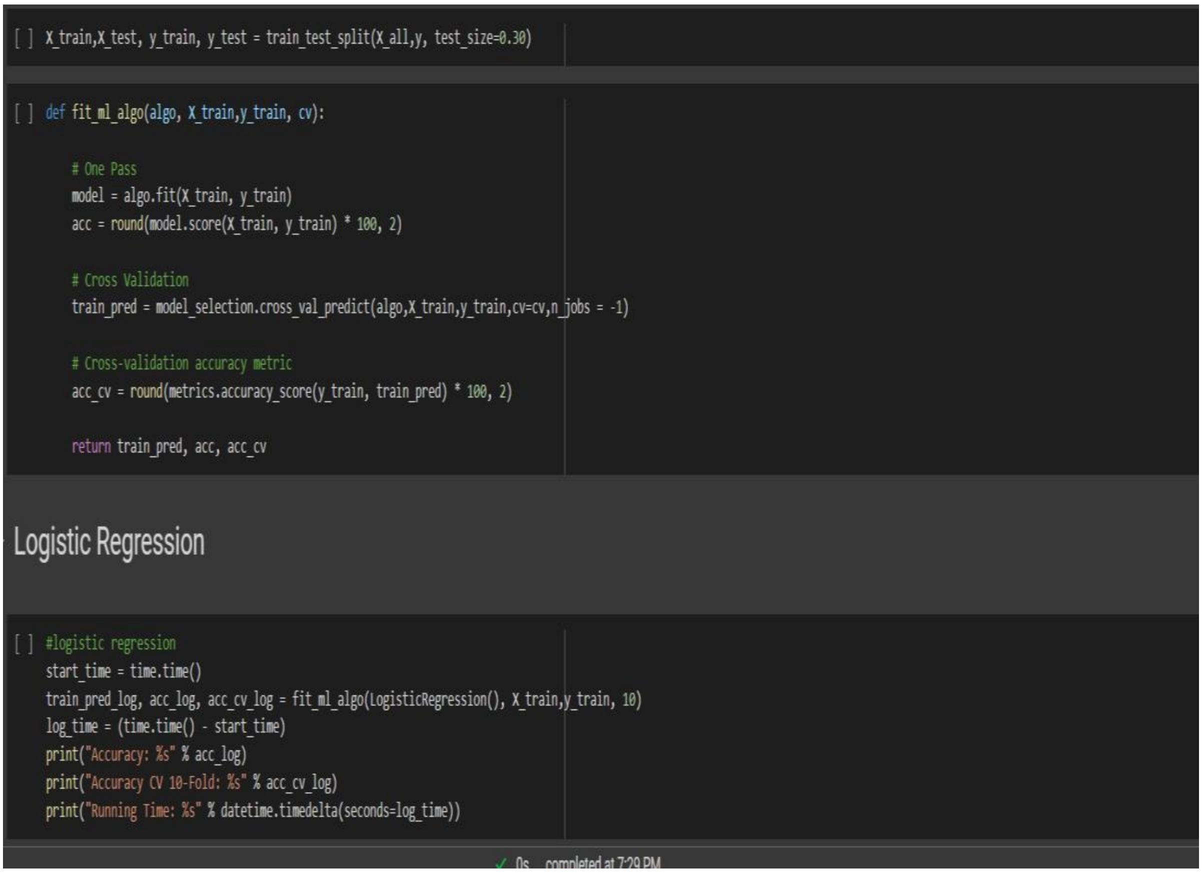
Figure 5.9 Effect of Gender on Attrition

5.4 MODEL TRAINING

Model training is the primary step in machine learning, resulting in a working model that can then be validated, tested and deployed. The model’s performance during training will eventually determine how well it will work when it is eventually put into an application for the end-users. Both the quality of the training data and the choice of the algorithm are central to the model training phase. In most cases, training data is split into two sets for training and then validation and testing. Both the quality of the training data and the choice of the algorithm are central to the model training phase. In most cases, training data is split into two sets for training and then validation and testing.

The dataset is splitted to a 70-30 ratio where 70 percent of data is used for training the model and the rest is used for testing. The data is prepared and the model’s hyperparameters have been determined, it’s time to start training the models.

Now it’s time to test the best versions of each algorithm to determine which gives you the best model overall. Once the testing is done, you can compare their performance to determine which are the better models. The overall winner should have performed well (if not the best) in training as well as in testing.



```
[ ] X_train,X_test, y_train, y_test = train_test_split(X_all,y, test_size=0.30)
```

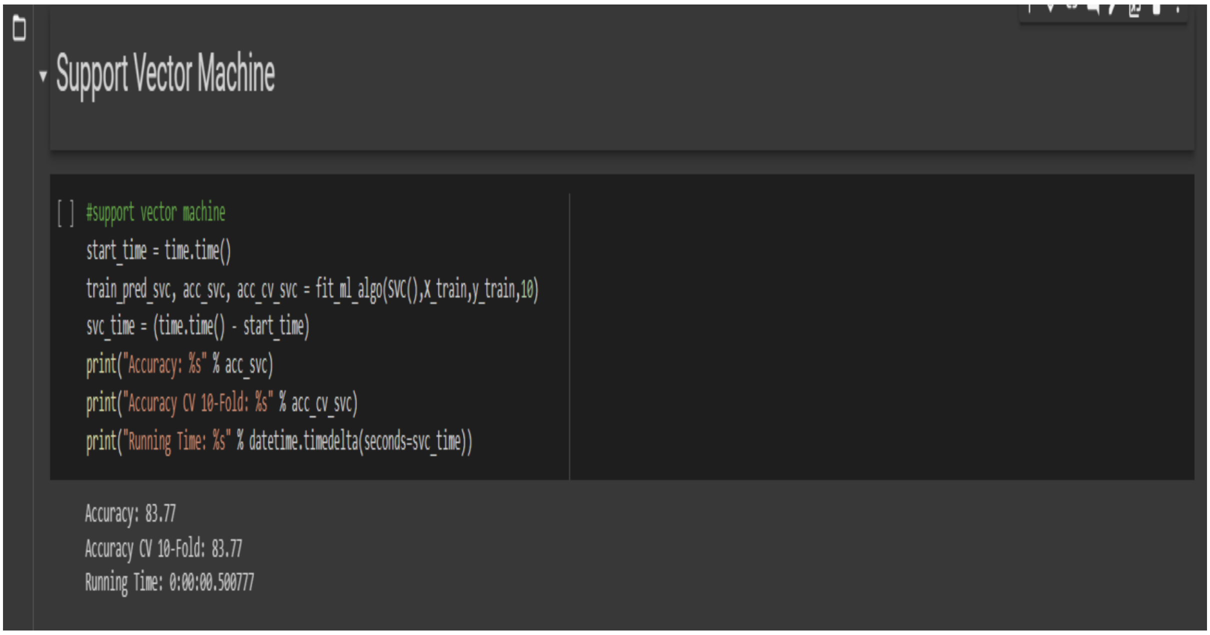
```
[ ] def fit_ml_algo(algo, X_train,y_train, cv):

        # 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,n_jobs = -1)

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

        return train_pred, acc, acc_cv
```

Logistic Regression

```
[ ] #logistic regression
    start_time = time.time()
    train_pred_log, acc_log, acc_cv_log = fit_ml_algo(LogisticRegression(), X_train,y_train, 10)
    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))
```

✓ 0s completed at 7:29 PM

Figure 5.9 Model Training Using Logistic Regression

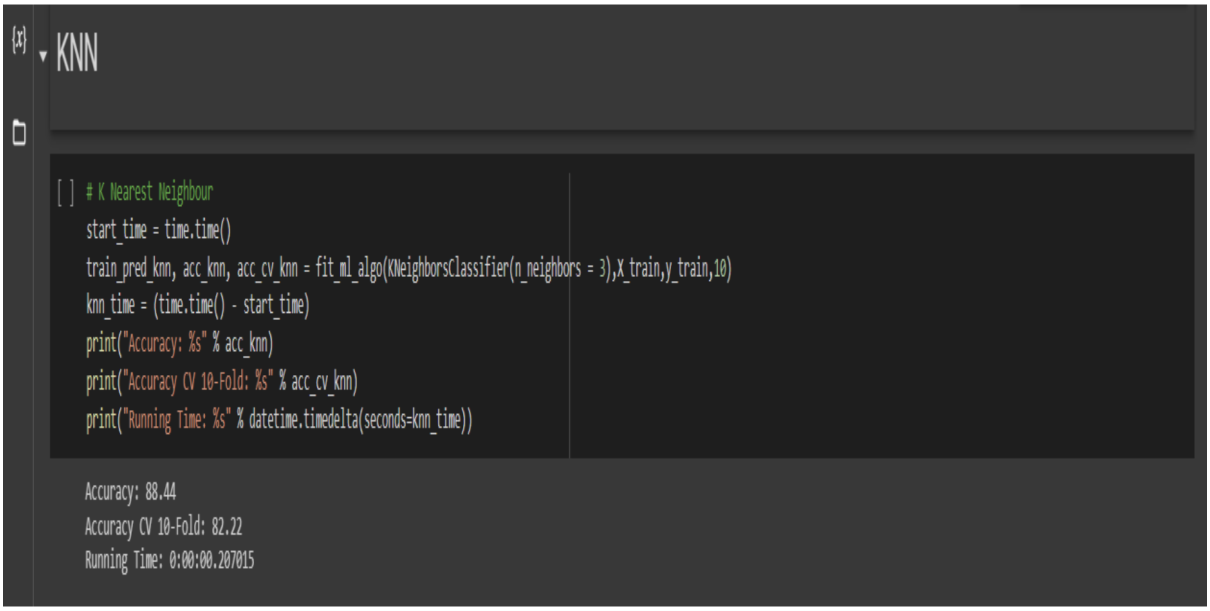


Support Vector Machine

```
[ ] #support vector machine
    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: 83.77
Accuracy CV 10-Fold: 83.77
Running Time: 0:00:00.500777
```

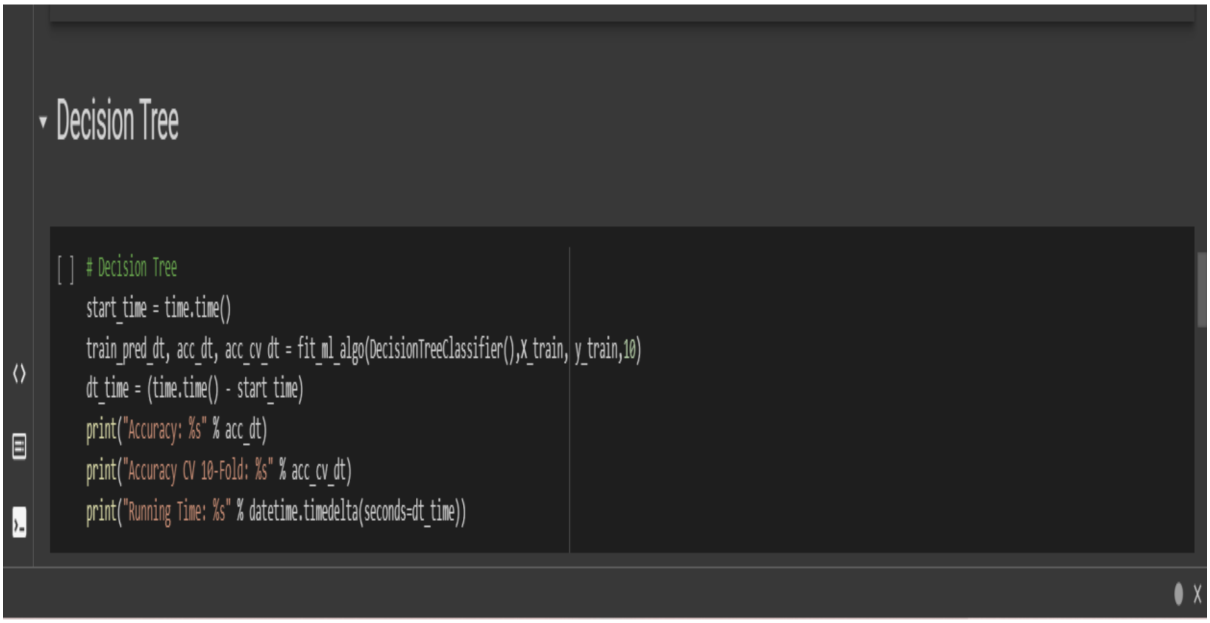
Figure 5.10 Model Training Using SVM



```
# K Nearest Neighbour
start_time = time.time()
train_pred_knn, acc_knn, acc_cv_knn = fit_ml_algo(KNeighborsClassifier(n_neighbors = 3),X_train,y_train,10)
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: 88.44
Accuracy CV 10-Fold: 82.22
Running Time: 0:00:00.207015
```

Figure 5.11 Model Training Using KNN

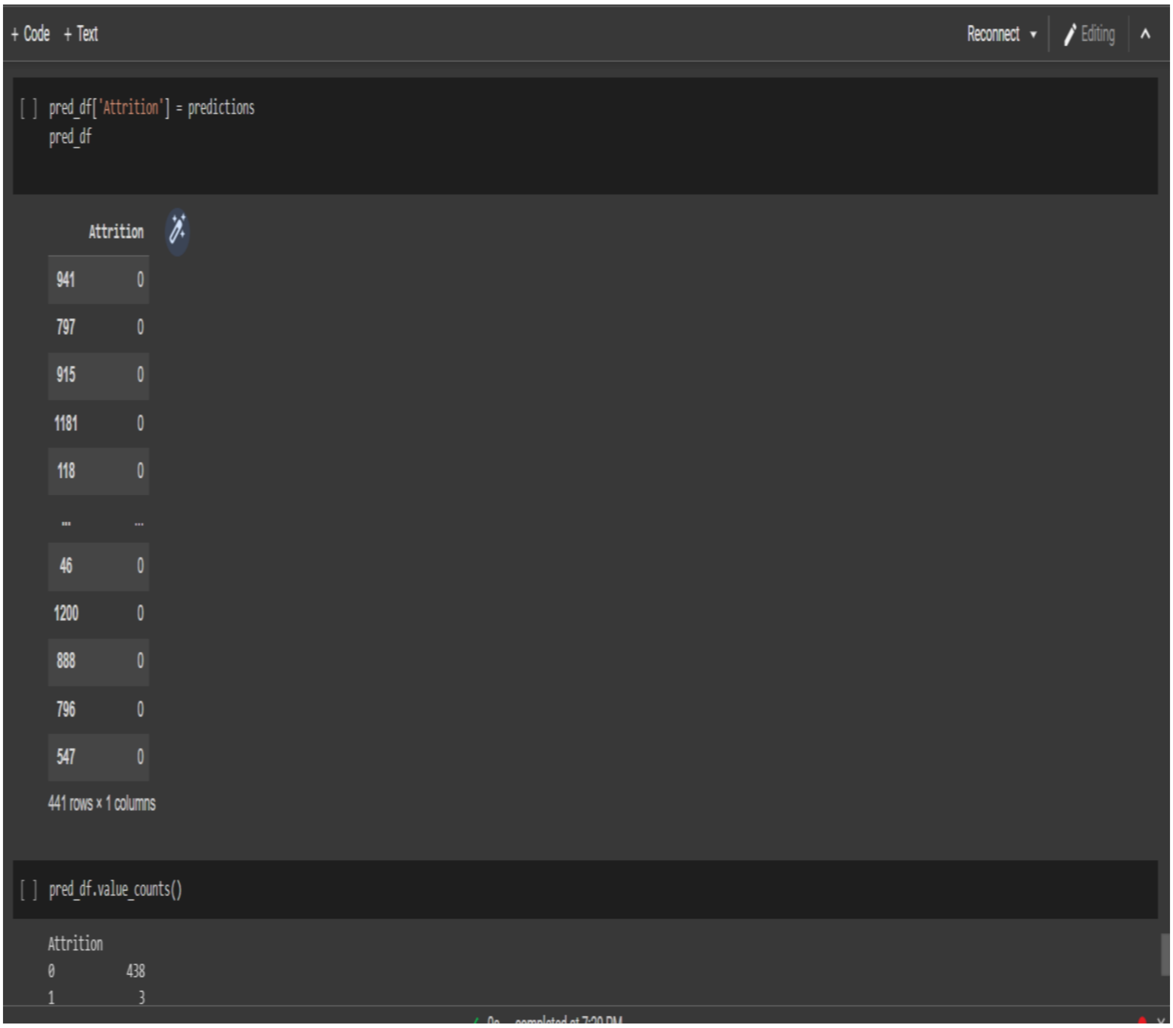


```
# Decision Tree
start_time = time.time()
train_pred_dt, acc_dt, acc_cv_dt = fit_ml_algo(DecisionTreeClassifier(),X_train, y_train,10)
dt_time = (time.time() - start_time)
print("Accuracy: %s" % acc_dt)
print("Accuracy CV 10-Fold: %s" % acc_cv_dt)
print("Running Time: %s" % datetime.timedelta(seconds=dt_time))
```

Figure 5.12 Model Training Using Decision Tree

5.5 MODEL TESTING

Model testing is a process in which a fully trained model is tested and validated on a test set. This is a process to verify the performance of the model using data which is not part of the training dataset.



```
[ ] pred_df['Attrition'] = predictions
    pred_df
```

```
      Attrition
941         0
797         0
915         0
1181        0
118         0
...       ...
46          0
1200        0
888         0
796         0
547         0
441 rows × 1 columns
```

```
[ ] pred_df.value_counts()
```

```
Attrition
0         438
1           3
```

Figure 5.13 Model Testing

CHAPTER 6

RESULTS AND DISCUSSIONS

The results of various classification metrics for all models and imbalanced/balanced data are summarised in the tables 2, 3 and 4.

Model	**Acc**	**Prec**	**Rec**	**F1**
LR	**0.875**	0.753	0.346	**0.472**
NB	0.787	0.394	**0.586**	0.471
DT	0.836	0.498	0.279	0.351
RF	0.862	0.841	0.181	0.296
AdaBoost	0.858	0.769	0.181	0.293
SVM	0.867	0.741	0.283	0.406
LDA	0.867	0.683	0.325	0.439
KNN	0.848	**0.771**	0.089	0.157
MLP	0.866	0.677	0.338	0.447

Table 2. Classification Scores for Imbalanced Data

Model	**Acc**	**Prec**	**Rec**	**F1**
LR	0.706	0.714	0.733	0.723
NB	0.653	0.624	**0.776**	0.691
DT	0.664	0.673	0.645	0.655
RF	**0.724**	0.726	0.726	**0.724**
AdaBoost	0.722	0.730	0.708	0.718
SVM	0.719	0.729	0.696	0.712
LDA	0.550	0.539	0.696	0.607
KNN	0.704	**0.734**	0.646	0.686
MLP	0.715	0.720	0.705	0.712

Table 3. Classification Scores for Undersampled Data with PCA

Model	**Acc**	**Prec**	**Rec**	**F1**
LR	0.758	0.751	0.771	0.761
NB	0.680	0.649	0.782	0.709
DT	0.841	0.796	0.917	0.852
RF	**0.992**	**0.986**	**0.998**	**0.992**
AdaBoost	0.989	0.982	**0.998**	0.990
SVM	0.951	0.912	**0.998**	0.953
LDA	0.609	0.610	0.736	0.667
KNN	0.894	0.826	**0.998**	0.904
MLP	0.946	0.905	**0.998**	0.949

Table 4. Classification Scores for Oversampled Data with PCA

The logistic regression model performed best for imbalanced data with an accuracy of 87.5%. For undersampled data with PCA, Random Forest model had best metric values with 72.4% accuracy and F1 score and 72.6% precision and recall. In the case of oversampled data with PCA, tree-based models performed best out of which Random Forest had the highest accuracy and F1 score of 99.2%, precision of 98.6%. As expected, the tree-based models performed well as they are known to work with nonlinear data. They can make more complex decision boundaries that fit very well on non-linear data. Decision Tree was able to achieve an accuracy score of 84% and recall of 91%. We also tried other complex models such as the SVC and MLP. SVC with a nonlinear kernel ’rbf’ and MLP also performed great on the testing data.

PCA, over and under sampling to balance data were also performed separately for these models. Among these, highest metric scores were observed for oversampled data while there was some improvement in performance for undersampled data. No considerable improvement in performance of models was obtained due to PCA alone.

Overall, all models performed better for oversampled data with PCA as compared to imbalanced data. The exceptions were LR and NB. Logistic regression didn’t perform well as it assumes that the data is linearly separable which was not the case as was seen in the EDA. Naive Bayes also didn’t perform well as many of the features are not conditionally independent such as the job role and the monthly income, education and job level as well as daily rate, hourly rate etc. This may also be because these classifiers were predicting the majority class most of the time and due to the imbalanced data scored high ac- curacies which was no longer the case for oversampled data.

There was no improvement in accuracy for any model for undersampling with PCA. Higher precision, recall and F1 scores were obtained for some due to the balancing. This is because undersampling caused downsizing of data in the majority to around 16% from the earlier 84% leading to loss of valuable information on the way as proposed by [5]. The unsupervised model KNN had good metric scores yet there were many supervised models like RF, AdaBoost, SVM which performed better.

We also realize that in a real-world scenario, the data will inherently be imbalanced as employees leaving a workforce will generally be fewer than those staying in the organisa- tion. Thus the above methods and results provide a good starting point for attrition prediction. Detailed, model-wise analysis is below.

6.1. Logistic Regression (LR)

The best performing logistic regression model was for oversampling with PCA; the hyperparamters obtained were: C as 0.1, penalty as l2 and solver as liblinear. It had higher accuracy for standardized data.

6.2. Naive Bayes (NB)

The Gaussian Naive Bayes achieved the best recall with imbalanced and undersampled data, 58.6% and 77.6% re- spectively. There was an increase in precision, recall and F1 scores in oversampled and undersampled data with PCA but a decrease in the accuracy.

6.3. Decision Tree (DT)

The decision Tree model trained on 30 features and un- scaled data as shown in Fig. 8 had the following tuned pa- rameters: criterion as gini, maximum depth as 13, maxi- mum features as one-third of total features, the maximum number of leaf nodes as 100 and the minimum number of samples in leaf as 1. According to this tree, OverTime, JobLevel, HourlyRate, TotalWorkingYears, MaritalStatus, MonthlyIncome and Age had higher importance. The lack of stability of decision tress was responsible for lower accu-racy, precision, recall, F1 score than other tree based coun-terparts.

6.4. Random Forest (RF)

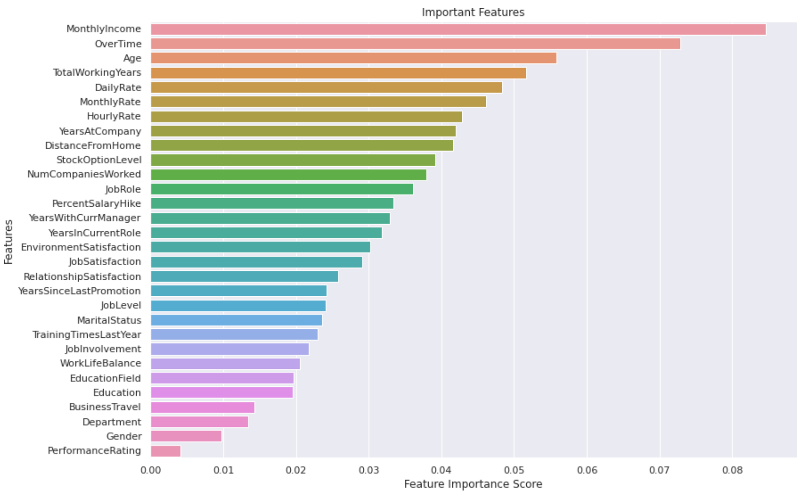


Figure 6.1. Feature Importance w.r.t. Random Forest with Oversampling

The best performance was obtained by setting the hyper- parameters Bootstrap to False, max depth to 100, min sam- ples leaf to 1, min sample required to split to 3 and the to- tal number of decision trees in the random forest estimator to 250. From Fig. 9, we observe that the most important features were Monthly-Income followed by OverTime and Age, while the least important features were Performance Rating, Gender and BusinessTravel. This ensemble model offered stability, lower bias and variance and thus had the best performance.

6.5. AdaBoost

An improvement was seen compared to the decision tree results and the model achieved the best recall or 99.8% using under-sampled data with PCA. The best hyper- parameters were the learning rate set to 1.0 and n estimators set to 1000. It is also the second-best performing model with high accuracy, precision and F1 score values.

6.6. Support Vector Machine (SVM)

The best performance was obtained by setting the hyper- parameter ’C’ to

100 and kernel to ’rbf’. Its ROC-AUC curve is in Fig. 10. As expected, model trained on over- sampled data performed the best and the model trained on the imbalanced data performed the worst.

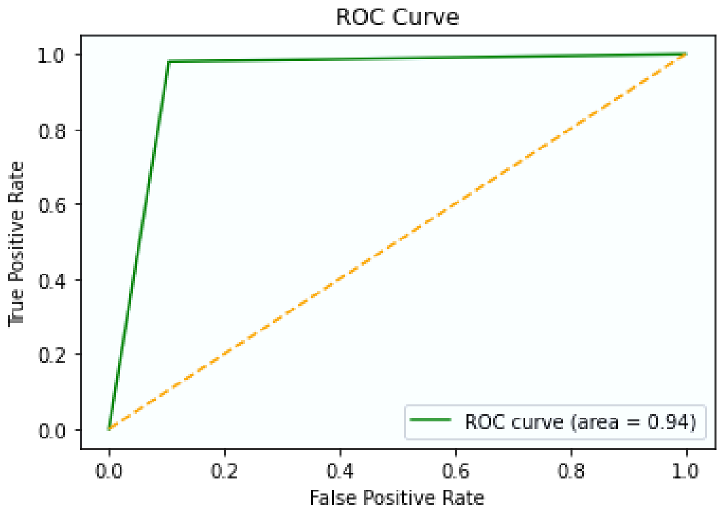


Figure 6.2. ROC-AUC Curve for SVC

6.7. Linear Discriminant Analysis (LDA)

The best performing LDA model was for oversampled data with PCA with hyper-parameters: shrinkage as auto and solver as ’lsqr’. LDA had higher accuracy and precision for imbalanced data but higher recall and F1 scores were observed for balanced data.

6.8.K-Nearest Neighbours (KNN)

KNN model performed best for oversampling with PCA and had hyperparameters: leaf size as 1, number of neigh- bors as 17 and weights as distance. It had a lower accuracy and precision for imbalanced data than undersampled data with PCA. This was the only unsupervised model and had the highest recall as well as good accuracy, precision, F1 Score.

6.9.MultiLayer Perceptron (MLP)

With hyperparameters: logistic activation function, al- pha as 0.05 and lbfs solver, the best performance with high- est metrics was obtained for oversampled data with PCA followed by imbalanced and undersampled data.

CHAPTER 7

CONCLUSION AND FUTURE WORK

The main goal of this research is to help HR managers to detect as soon as possible an employee’s intention to leave using predictive analytics methods and so to fight this attrition. The contributions can be summarized into three points. The proposal of a new employee attrition model that contains the required features necessary and sufficient to detect intention to leave and to predict positive attrition using a mixed research methodology. The proposal of machine, deep and ensemble learning predictive models and their experimentation in a variety of different settings to best assess their performance. The interpretation and the explication that enables HR managers to understand what makes an employee want to leave and to help them in adopting key policies to retention.

In terms of study limitations, considering dynamic features that deal with employees' behaviour and their emotional states will be promising to study their impact on employee attrition. In this case, the predictive models training must be on-line as data will be dynamic and new data can be added whenever required.

We acknowledge that other features to be considered and that can cause voluntary turnover and so can be integrated into our future study. In fact, they have proposed to consider health issues, job security and the use of new technologies in the company. Finally, in future research, considering unbalanced data is a real challenge especially for organizations and companies with high turnover rate because the adopted predictive models are experimentally not suitable for unbalanced data.

REFERENCES

1. D. Angrave, A. Charlwood, I. Kirkpatrick, M. Lawrence, and M. Stuart (2016) “HR and analytics: Why HR is set to fail the big data challenge” Hum. Resource Manage. J, Vol. 26, No. 1, pp. 1_11.

2. R. Colomo-Palacios, C. Casado-Lumbreras, S. Misra, and P. Soto-Acosta” (2014) Career abandonment intentions among software workers” Hum. Factors Ergonom. Manuf. Service Industries, Vol. 24, No. 6, pp. 641_655.

3. P. Likhitkar and P. Verma (2020) “HR value proposition using predictive analytics: An overview” in New Paradigm in Decision Science and Management. Singapore: Springer, pp. 165_171.

4. S. N. Mishra, D. R. Lama, and Y. Pal (2016) “Human resource predictive analytics(HRPA) for HR management in organizations” Int. J. Sci. Technol.Res., Vol. 5, No. 5, pp. 33_35.

5. T. Pape (2016) “Prioritising data items for business analytics: Framework and application to human resources” Eur. J. Oper. Res., Vol. 252, No. 2,pp. 687_698.

6. R. Punnoose and P. Ajit (2016) “Prediction of employee turnover in organizations using machine learning algorithms” Int. J. Adv. Res. Artif. Intell., Vol. 5,No. 9, pp. 376_271.

7. A. Tursunbayeva, S. D. Lauro, and C. Pagliari (2018) “People analytics_A scoping review of conceptual boundaries and value propositions” Int.J. Inf. Manage., vol. 43, pp. 224_247.