# PREDICTING EMPLOYEE ATTRITION: A MACHINE LEARNING PERSPECTIVE

## DISSERTATION

*Submitted to the University of Kerala in partial fulfillment of the requirement for the award of Degree of Master of Science in Applied Statistics and Data Analytics*


*BY*
**ABIN E SAVIO**
**(85621615002)**

*Department of Statistics,*
*University of Kerala, Kariavattom Campus,*
*Thiruvananthapuram - 695581,*
*Kerala.*

**SEPTEMBER 2023**

**DEPARTMENT OF STATISTICS**
**UNIVERSITY OF KERALA**
**THIRUVANANTHAPURAM, KERALA - 695581**


## CERTIFICATE

This is to certify that this dissertation entitled ***Predicting Employee Attrition: A Machine Learning Perspective*** is the bonafide record of work carried out by **ABIN E SAVIO (85621615002)** in partial fulfillment of the requirements for the completion of Degree of Master of Science in Applied Statistics and Data Analytics, at the Department of Statistics, University of Kerala, Thiruvananthapuram.

Place
Date

*Dr. Manoj Chacko*
*Associate Professor*
*Department of Statistics*
*University of Kerala*

# ACKNOWLEDGEMENT

The success of anything needs co-operation and encouragement from different quarters. Words are inadequate to express my profound and deep sense of gratitude to those who helped me in bringing out this project successfully.

I wish to express my profound and deep sense of gratitude to **Dr. MANOJ CHACKO** , Associate Professor of the Department of Statistics, University of Kerala for suggesting the topic, his guidance, supervision and constant encouragement throughout the study. His valuable and inspiring suggestions, constructive criticism and moral support helped me in the successful completion of this work.

I would like to thank all the teachers, the administration officer, the librarian and research scholars of our department, my parents and all my friends for their help and encouragement that they have rendered to me during the course of this project work.

Above all I thank the Almighty for his blessing showered on us throughout the course of this project work.

**ABIN E SAVIO**

# DECLARATION

I hereby declare that the work presented in the Dissertation titled "Predicting Employee Attrition: A Machine Learning Perspective" is completed by me under the guidance of Dr. Manoj Chacko, Associate Professor, Department of Statistics, University of Kerala, Kariavattom Campus, Thiruvananthapuram and it has not been included in any other thesis submitted previously for the award of any degree.

Place: Kariavattom
Date: 01-11-2021

ABIN E SAVIO

# ABSTRACT

Employee attrition poses a significant challenge for organizations, affecting productivity, morale, and financial stability. In this study, we present an Employee Attrition Prediction Model designed to help organizations proactively identify and mitigate attrition risks. Leveraging a diverse dataset encompassing various employee attributes, job-related factors, and demographic information, we employ advanced machine learning techniques to develop an accurate predictive model.

The primary objectives of this research are threefold: first, to identify the key predictors of employee attrition through comprehensive statistical analysis; second, to validate these predictors using hypothesis testing methodologies; and third, to build, assess, and compare the performance of multiple machine learning models, including Decision Trees, Random Forest, XGBoost Classifier, and Support Vector Machine (SVM), to predict employee attrition accurately.

Our study's findings highlight the critical factors influencing attrition, providing actionable insights for organizations to address workforce stability and retention. Through rigorous model comparison, we identify the Support Vector Machine (SVM) as the most effective model, achieving an accuracy rate of 89%. The developed Employee Attrition Prediction Model empowers organizations to anticipate and intervene in potential attrition cases, fostering a more engaged and productive workforce, while contributing to overall organizational success.

# Contents

# Chapter 1

# Introduction

Employee attrition, or the rate at which employees leave an organization, is a critical concern for businesses across various industries. High attrition rates can lead to increased recruitment costs, loss of valuable talent, and disruptions in workflow. To proactively address this challenge and make data-driven decisions, many organizations turn to machine learning models. These models leverage historical data and a range of employee-related factors to predict and understand the likelihood of an employee leaving the company. By doing so, they provide valuable insights that can aid in developing retention strategies and improving overall workforce management.

In the realm of employee attrition prediction, machine learning models play a pivotal role in assisting HR departments and decision-makers. They harness the power of data analytics to identify patterns, correlations, and risk factors associated with employee turnover. These models take into account an array of variables, including but not limited to employee demographics, job roles, performance metrics, compensation, and job satisfaction surveys. By analyzing this diverse set of features, machine learning algorithms can identify the underlying factors that contribute to attrition and provide actionable insights.

The benefits of employing machine learning for attrition prediction are substantial. Organizations can use these models to proactively identify individuals or departments at a higher risk of attrition, enabling HR teams to tailor interventions and retention strategies accordingly. Moreover, these predictive models can continuously learn from new data,

adapting to changing workforce dynamics and improving their accuracy over time. By harnessing the power of machine learning, organizations can move beyond reactive responses to attrition and adopt a more proactive and strategic approach to talent retention, ultimately contributing to a more stable and productive work environment.

## 1.1 OVERVIEW OF EMPLOYEE ATTRITION

Employee attrition, often referred to as employee turnover, is a critical aspect of workforce management that measures the rate at which employees leave an organization over a specified period. It is a significant concern for businesses, as it can have substantial economic and operational impacts. Understanding the factors contributing to employee attrition is essential for organizations to develop effective strategies for talent retention and succession planning.

**Key Concepts and Definitions:**

**Attrition Rate:** This is the percentage of employees who leave an organization during a specific time frame. It is calculated by dividing the number of employees who have left by the average total number of employees during that period.

**Voluntary vs. Involuntary Attrition:** Employee attrition can be categorized into voluntary and involuntary attrition. Voluntary attrition occurs when employees choose to leave the organization, often due to personal or career-related reasons. Involuntary attrition, on the other hand, happens when employees are forced to leave due to factors such as layoffs or terminations.

**Importance of Attrition Prediction:**

Predicting employee attrition is a critical task for HR and management teams. It offers several benefits:

**Cost Reduction:** High attrition rates can be costly in terms of recruitment, training, and lost productivity. Predicting attrition allows organizations to proactively manage these costs.

**Talent Retention:** By identifying the key drivers of attrition, organizations can implement strategies to retain valuable employees.

**Succession Planning:** Attrition prediction aids in succession planning, ensuring that the

organization is prepared to fill key roles when employees depart.

**Employee Engagement:** Understanding attrition factors can lead to improvements in employee engagement and satisfaction.

## 1.2 MOTIVE OF STUDY

The primary motive of this study is to address the pressing challenges faced by organizations in managing their workforce effectively. Employee attrition, the rate at which employees voluntarily or involuntarily leave an organization, presents a significant challenge in the contemporary business landscape. In an era where human capital is a critical driver of competitive advantage, understanding and mitigating attrition has become paramount for sustaining organizational growth and success.

**The Significance of Attrition Prediction:**

**Cost Savings:** High employee attrition rates can be financially burdensome for organizations. The costs associated with recruiting, onboarding, and training new employees, coupled with the loss of productivity during transitions, can be substantial. By accurately predicting attrition, organizations can implement cost-saving measures, allocate resources efficiently, and strategically plan for talent acquisition.

**Enhanced Talent Retention:** Employee attrition not only leads to financial losses but also affects morale and productivity within the workforce. A predictive model for attrition allows organizations to proactively identify employees at risk of leaving and implement targeted retention strategies, such as professional development opportunities, improved working conditions, or competitive compensation packages, to retain valuable talent.

**Succession Planning:** Employee departures, especially from key positions, can disrupt workflow and hinder organizational performance. Effective succession planning relies on understanding attrition patterns and preparing potential successors in advance. Predictive modeling can help identify and groom future leaders within the organization.

**Improved Employee Satisfaction:** By uncovering the major factors driving attrition, organizations can take steps to address issues related to job satisfaction, work-life balance, and career growth. This not only reduces attrition but also enhances overall employee

satisfaction and engagement.

## 1.3 DATASET INFORMATION

In this project, supervised machine learning is employed on the Employee Attrition dataset obtained from Kaggle. The dataset contains 31 attributes including demographic details, work-related metrics, and attrition flag. These attributes include essential information such as employee age, department, education level, job satisfaction, and work-related factors like daily rates, travel commitments, and distance from home to work. The dataset also includes details on career progression, including years at the company, years in the current role, and years since the last promotion, among others. Moreover, it encompasses personal factors such as marital status and gender, along with indicators of job-related satisfaction, like work-life balance and performance ratings. With these diverse features, the dataset is a valuable resource for conducting predictive analysis to identify key factors contributing to employee attrition, a crucial challenge in workforce management and organizational sustainability.

## 1.4 OBJECTIVES OF THE STUDY

**1.Identify Key Predictors:** To determine and analyze the primary factors that significantly contribute to employee attrition within the organization. This involves conducting comprehensive statistical analysis to pinpoint the most influential variables affecting attrition.

**2.Predictive Modeling:** To develop, assess, and compare the performance of various machine learning models, including Decision Trees, Random Forest, XGBoost Classifier, Support Vector Machine (SVM). These models will be utilized for accurately predicting employee attrition, aiding in the identification of potential attrition cases.

**3.Feature Importance Analysis:** To determine the top ten most influential attributes within the predictive models. This feature importance analysis will assist in identifying the critical drivers of attrition and understanding their relative impact on employee turnover.

**4.Model Comparison:** To conduct a comparison of all employee attrition prediction models

based on their values for fitting the data and prediction errors. This comparative analysis will help to select the most effective model for predicting employee attrition.

**5.Predict Employee attrition** Predicting employee attrition using the most accurate and efficient predictive model.

## 1.5 LIMITATIONS

This project has certain limitations to keep in mind. Firstly, it relies on historical data, which means it looks at the past to predict the future. This approach may not capture all the new and evolving factors that could influence employee turnover. Secondly, while we're examining specific reasons why employees might leave, there could be other important factors we haven't considered. The accuracy of our predictions also hinges on the quality and completeness of the data we have. If the data isn't very good, our predictions might not be either. Additionally, organizations change over time, and what caused attrition in the past might not be the same in the future.

# Chapter 2

# Methodology

## 2.1 INTRODUCTION

Methodology refers to the procedure or technique adopted in research and it plays a crucial role in any type of research. It is the general research strategy that outlines the way in which research is to be undertaken and identifies the methods to be used in it. This chapter discusses the various analytical techniques we utilized for this assignment. We selected the most appropriate data analysis techniques based on the study's objective. The primary technique used in this study is machine learning. The chapter provides a comprehensive description of the machine learning techniques, models, graphical tools, and software used for the study.

## 2.2 MACHINE LEARNING

Machine Learning is a subset of Artificial Intelligence that focuses on developing algorithms that allow a computer to learn from data and past experiences independently. The term "Machine Learning" was first coined by Arthur Samuel in 1959. A Machine Learning system learns from historical data, also known as training data, to build prediction models. When it receives new data, it predicts the output without explicit programming. The accuracy of the predicted output depends on the amount of data available, as more data helps build a better model that can predict the output more accurately. Instead of

writing code to solve complex problems that require predictions, we can feed data to generic algorithms. With the help of these algorithms, the machine can build the logic based on the data and make predictions. Machine Learning has revolutionized our problem-solving approach. The block diagram below illustrates how Machine Learning algorithms work

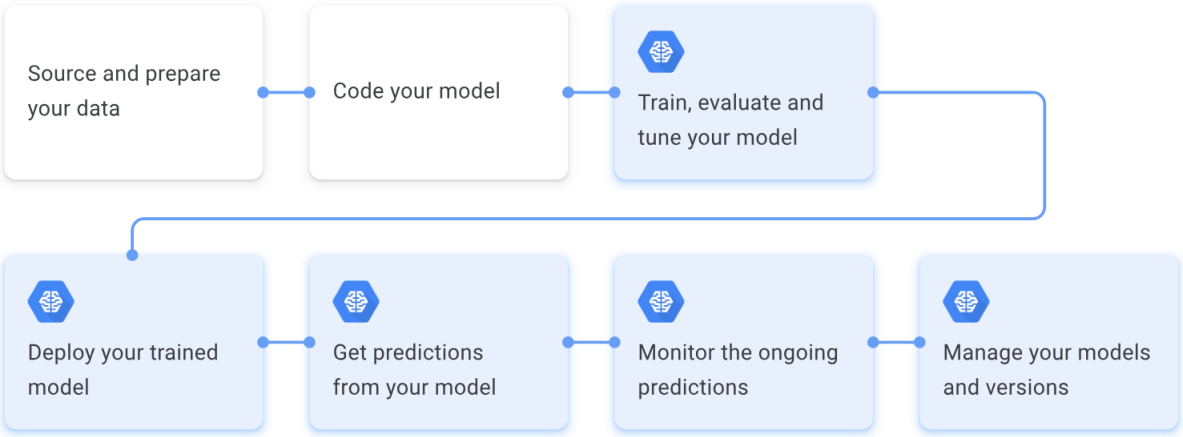


Figure 2.1: ML Workflow

## 2.3 TYPES OF MACHINE LEARNING

When it comes to training machine learning algorithms, there are various methods to choose from, each with its own advantages and disadvantages. To understand the benefits and drawbacks of each type of machine learning, let's first look at the kind of data they deal with. In machine learning, there are two types of data: labelled data and unlabelled data. Labelled data includes both input and output parameters in a machine-readable format, but labelling the data requires a lot of human effort. Unlabelled data has only one or none of the parameters in a machine-readable form. This makes it less labour-intensive but requires more complex solutions. Machine learning is primarily divided into three types based on the methods and learning approaches used. These types are:

### 2.3.1 SUPERVISED LEARNING

Supervised learning is the most basic type of machine learning. In this approach, the

algorithm is trained on labelled data. Despite requiring accurately labelled data, supervised learning is highly effective when used in the right circumstances. In supervised learning, the algorithm works with a small training dataset, which is a smaller part of the larger dataset. This dataset provides the algorithm with a basic understanding of the problem, solution, and data points to be dealt with. The algorithm then establishes a cause-and-effect relationship between the variables in the dataset, essentially finding relationships between the parameters given. At the end of the training, the algorithm has an idea of how the data works and the relationship between the input and the output. This solution is then deployed for use with the final dataset, which it learns from in the same way as the training dataset. As a result, supervised machine learning algorithms continue to improve even after deployment, discovering new patterns and relationships as they train themselves on new data.
Supervised machine learning can be classified into two types of problems: classification and regression.

- **Classification:** Classification algorithms are used to solve problems in which the output variable is categorical, such as "Yes" or "No," "Male" or "Female," "Red" or "Blue," etc. Classification algorithms predict the categories present in the dataset. Some real-world examples of classification algorithms include Spam Detection, Email filtering, etc.
- **Regression:** Regression algorithms are used to solve problems in which there is a linear relationship between input and output variables. They predict continuous output variables, such as market trends, weather predictions, etc.

### 2.3.2 UNSUPERVISED LEARNING

Unsupervised machine learning can work with unlabelled data, which means that it does not require human labour to make the dataset machine-readable. As a result, the program can work with much larger datasets. In supervised learning, the labels allow the algorithm to find the exact nature of the relationship between any two data points. However, unsupervised learning does not have labels to work off of, resulting in the creation of hidden structures. Relationships between data points are perceived by the algorithm in an abstract

manner, with no input required from human beings. The creation of these hidden structures makes unsupervised learning algorithms versatile. Instead of a defined and set problem statement, unsupervised learning algorithms can adapt to the data by dynamically changing hidden structures. This offers more post-deployment development than supervised learning algorithms.

### 2.3.3 REINFORCEMENT LEARNING:

Reinforcement learning draws inspiration from how humans learn from data in their lives. It features an algorithm that improves itself and learns from new situations using a trial-and-error method. Favourable outputs are encouraged or 'reinforced,' and non-favourable outputs are discouraged or 'punished.' Based on the psychological concept of conditioning, reinforcement learning works by putting the algorithm in a work environment with an interpreter and a reward system. In every iteration of the algorithm, the output result is given to the interpreter, which decides whether the outcome is favourable or not. If the program finds the correct solution, the interpreter reinforces the solution by providing a reward to the algorithm. If the outcome is not favourable, the algorithm is forced to reiterate until it finds a better result. In most cases, the reward system is directly tied to the effectiveness of the result. In typical reinforcement learning use-cases, such as finding the shortest route between two points on a map, the solution is not an absolute value. Instead, it takes on a score of effectiveness, expressed in a percentage value. The higher this percentage value is, the more reward the algorithm receives.

## 2.4 GRAPHICAL TOOLS USED

### 2.4.1 PIE CHART

A pie chart is a circular graphical representation commonly used to visually display data distribution or composition. It consists of a full circle divided into slices or sectors, with each slice representing a specific category or group within a dataset. The size of each slice is directly proportional to the quantity or percentage it represents relative to the entire dataset, making it a useful tool for showcasing the relative proportions of different categories.

Labels or legends are typically added to each slice to clarify what each segment signifies, ensuring that viewers can easily interpret the chart. Pie charts are particularly effective when you need to convey the composition of a dataset or the distribution of categorical variables, making them a popular choice in business reports, presentations, and educational materials for their ability to provide a clear and concise visual summary of data relationships.

### 2.4.2 BAR PLOT

The visual display of data (often grouped) in the shape of vertical or horizontal rectangular bars, with the length of the bars corresponding to the measure of the data, is called a bar graph. Bar charts are another name for them. One tool used in statistics for processing data is the bar graph.

### 2.4.3 HISTOGRAM

A frequency distribution shows how often each different value in a set of data occurs. A histogram is the most commonly used graph to show frequency distributions. It looks very much like a bar chart, but there are important differences between them. The histogram is used for variables whose values are numerical and measured on an interval scale. It is generally used when dealing with large data sets (greater than 100 observations). A histogram can also help detect any unusual observations (outliers) or any gaps in the data.

### 2.4.4 HEATMAP

Heatmaps visualize the data in a 2-dimensional format in the form of colored maps. The color maps use hue, saturation, or luminance to achieve colour variation to display various details. This colour variation gives visual cues to the readers about the magnitude of numeric values. Heatmaps can describe the density or intensity of variables, visualize patterns, variance, and even anomalies. Heatmaps show relationships between variables. These variables are plotted on both axes. We look for patterns in the cell by noticing the color change. It only accepts numeric data and plots it on the grid, displaying different data values by varying color intensity.

## 2.5 ML ALGORITHMS USED

### 2.5.1 DECISION TREE

A supervised learning method called a decision tree can be used to solve classification and regression problems, but it is typically favoured for doing so. It is a tree-structured classifier, where internal nodes stand in for a dataset's features, branches for the decision-making process, and each leaf node for the classification result.

The Decision Node and Leaf Node are the two nodes of a decision tree. While Leaf nodes are the results of decisions and do not have any more branches, Decision nodes are used to create decisions and have numerous branches.

The decisions or the test are performed on the basis of features of the given dataset. It is a graphical depiction that shows all options for solving a dilemma or making a choice in light of certain parameters.

Because it begins with the root node and develops on additional branches to form a structure resembling a tree, it is known as a decision tree.

We employ the CART algorithm, or Classification and Regression Tree algorithm, to construct a tree.

Simply said, a decision tree poses a question and divides the tree into subtrees based on the response (Yes/No).

**Assumptions while creating Decision Tree**

Below are some of the assumptions we make while using Decision tree:

- In the beginning, the whole training set is considered as the root.
- Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.
- Records are distributed recursively on the basis of attribute values.
- Order to placing attributes as root or internal node of the tree is done by using some statistical approach.

Below diagram explains the general structure of a decision tree

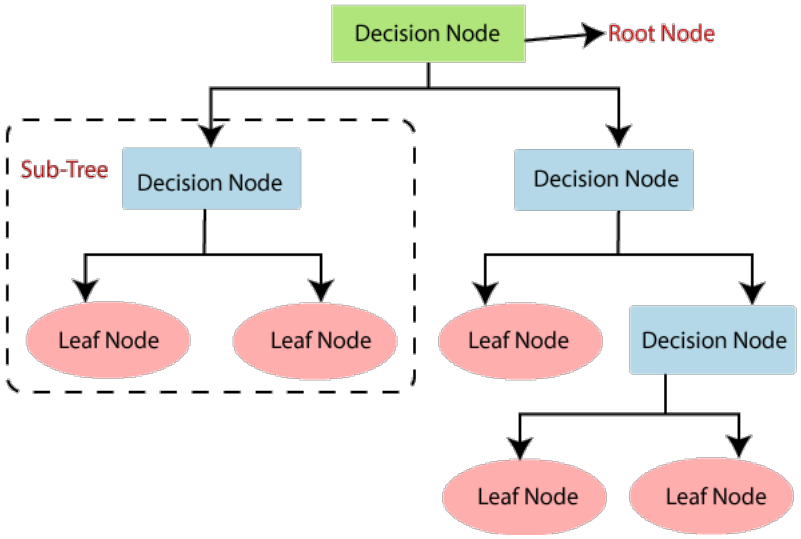


Figure 2.2: Decision Tree Classification

### 2.5.2 RANDOM FOREST CLASSIFICATION

A Random Forest Classifier stands as a robust and versatile machine learning algorithm renowned for its proficiency in both classification and regression tasks. Its strength lies in harnessing the wisdom of an ensemble of decision trees. Each decision tree is constructed during the training phase, contributing to the collective intelligence of the forest. This ensemble approach offers a superior capacity for making accurate predictions while mitigating the risk of overfitting.

Random Forest Classification is a versatile and powerful machine learning algorithm widely employed for solving classification problems in various domains. It belongs to the ensemble learning family, which means it builds a collection or "forest" of decision trees during training. Each tree in the forest is constructed using a subset of the training data and potentially considering a different subset of features. The primary advantage of this ensemble approach is that it mitigates the risk of overfitting, as the predictions are determined collectively by the multiple trees. When making predictions, each tree independently evaluates the input data, and the most frequent class prediction among the trees is chosen as the final

output.

One of the remarkable features of Random Forests is their adaptability to a wide range of data types. They can effectively handle both categorical and continuous features, making them suitable for diverse datasets frequently encountered in real-world applications. Additionally, Random Forests are robust to noisy data and outliers, and they excel at capturing complex, nonlinear relationships within the data. This versatility makes them a popular choice for tasks such as image classification, sentiment analysis, and disease diagnosis.

Furthermore, Random Forests offer valuable insights into feature importance. They assign importance scores to each feature based on their contribution to the overall model's performance. This information is crucial for understanding which attributes drive the classification decisions and can guide feature selection and model interpretation efforts. The ability to assess feature importance makes Random Forests not only a powerful predictive tool but also a valuable resource for gaining deeper insights into the underlying patterns and factors influencing the classification outcomes.

In practice, Random Forest Classification is known for its ease of use and robustness. It requires minimal data preprocessing, handles missing values gracefully, and doesn't rely on strong assumptions about the data distribution. While Random Forests can perform well with default hyperparameters, fine-tuning the number of trees in the forest and the maximum depth of each tree can further optimize model performance. Overall, Random Forest Classification is a reliable and widely adopted algorithm that excels in complex classification tasks, offering accuracy, interpretability, and robustness to data variability.

### 2.5.3 XGBoost CLASSIFIER

XGBoost, which stands for eXtreme Gradient Boosting, is a cutting-edge machine learning algorithm renowned for its exceptional performance in both classification and regression tasks. It belongs to the ensemble learning family, specifically the gradient boosting method, which iteratively improves the model's accuracy by minimizing prediction errors. XGBoost's strength lies in its ability to handle a wide range of data types and deliver highly accurate predictions while mitigating overfitting.

The core idea behind XGBoost is to build an ensemble of decision trees sequentially,

where each tree corrects the errors made by its predecessors. Unlike traditional gradient boosting algorithms, XGBoost incorporates a unique regularization term in the objective function, which helps prevent overfitting by penalizing overly complex trees. Additionally, XGBoost employs a "gradient boosting" approach, where each tree is trained on the residual errors of the previous trees, allowing it to focus on the instances that were poorly predicted before. This iterative process continues until a predefined number of trees (a hyperparameter) is reached or no further improvement can be achieved.

XGBoost offers several compelling features, including speed and scalability. It has been meticulously optimized to handle large datasets efficiently and can take advantage of parallel processing, making it suitable for both small and big data applications. Furthermore, XGBoost provides built-in support for handling missing data, eliminating the need for extensive preprocessing. It also incorporates advanced techniques like tree pruning and early stopping, which enhance its ability to generalize well to unseen data.

Another key aspect of XGBoost is its ability to provide interpretable feature importance scores. These scores quantify the impact of each feature on the model's predictions, aiding in feature selection and model understanding. This feature importance information can guide data scientists in identifying the most influential variables and refining their models accordingly.

In conclusion, XGBoost has become a cornerstone algorithm in various machine learning competitions and real-world applications due to its robustness, scalability, and remarkable predictive accuracy. Its unique regularization techniques, handling of missing data, and feature importance analysis make it a versatile and powerful tool for tackling a wide range of machine learning problems.

### 2.5.4 SUPPORT VECTOR MACHINE

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in

the future. This best decision boundary is called a hyper plane. SVM chooses the extreme points/vectors that help in creating the hyper plane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. SVM algorithm can be used for Face detection, image classification, text categorization, etc.

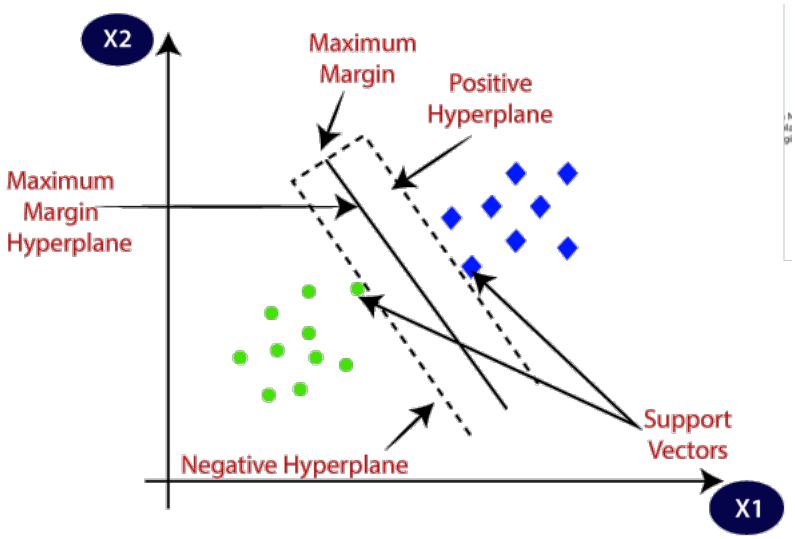


Figure 2.3: Classifying data using SVM

### TYPES OF SVM

- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
- **Non-Linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

### SVC

Support Vector Classification (SVC) is a specific implementation of the Support Vector Machine (SVM) algorithm, which is a powerful and widely used supervised machine learning technique. SVM is known for its effectiveness in handling both classification and regression tasks, but SVC is specifically designed to tackle classification problems. The fundamental

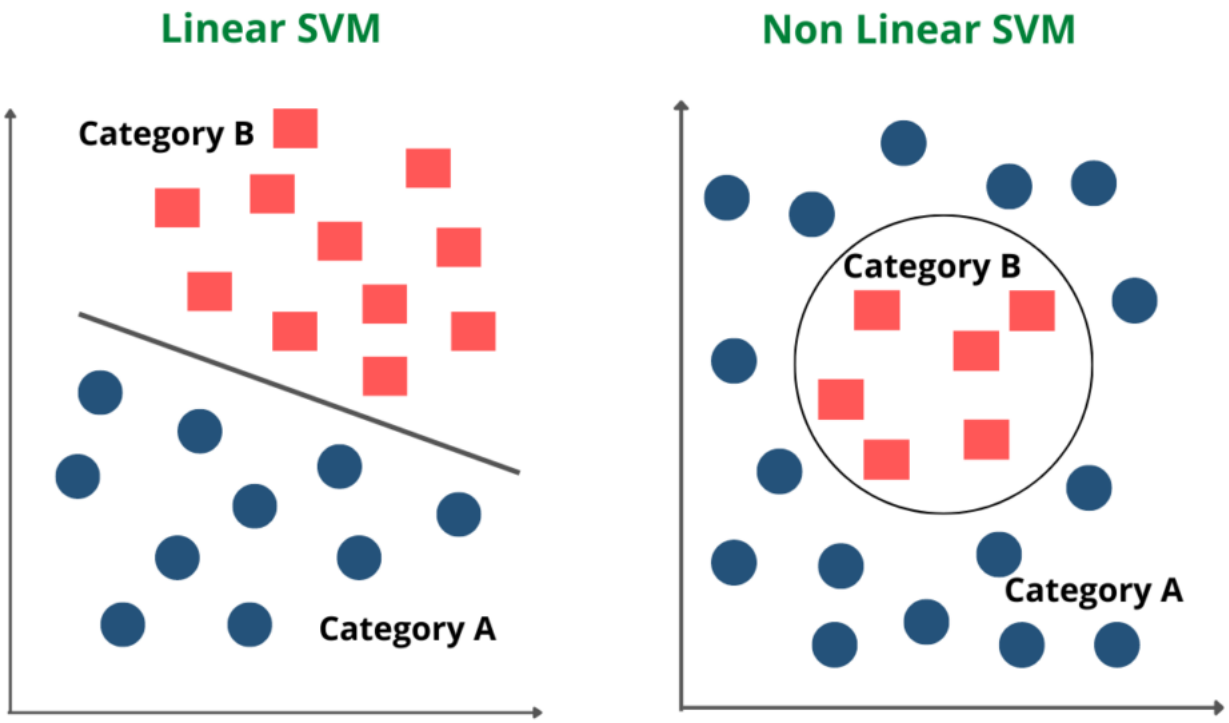


Figure 2.4:

goal of SVC is to find the optimal hyper plane that best separates data points belonging to different classes in the feature space. In binary classification, the hyper plane acts as a decision boundary, effectively dividing the data into two classes. The objective of SVC is to maximize the margin between the support vectors, which are the data points closest to the decision boundary, and this margin plays a crucial role in the algorithm's performance and generalization ability.The steps involved in SVC are as follows:

**Data Preparation:** Gather a labeled dataset containing input features and their corresponding class labels for the classification task.

**Feature Mapping:** If the data is not linearly separable, use kernel functions (e.g., polynomial, radial basis function) to map the data into a higher-dimensional space where separation becomes feasible.

**Margin Calculation:** Identify the optimal hyper plane that maximizes the margin between support vectors of different classes.

**Training:** Train the SVC model on the labeled dataset, adjusting the hyper plane's

parameters to find the best fit.

**Prediction:** Use the trained SVC model to make predictions on new, unseen data points, assigning them to the appropriate classes.

## 2.6 STATISTICAL TOOLS USED

### 2.6.1 CHI-SQUARE TEST FOR INDEPENDENCE

A chi – square test for independence is applied when you have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables. The test consists of four steps:

**1. State the hypothesis** A chi–square test for independence is conducted on two categorical variables. Suppose that variable A has r levels, and variable B has c levels. The null hypothesis states that knowing the level of variable A does not help you predict the level of variable B, i.e., the variables are independent. The alternative hypothesis states that the variables are not independent.

**2. Formulate an analysis plan** The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify a significance level and should identify the chi – square test for independence as the test method.

**3. Analyse sample data** Using sample data, find the degrees of freedom, expected frequencies, test statistic, and the p – value associated with the test statistic.

**4. Interpret results** Based on the sample findings, the researcher rejects the null hypothesis. This involves comparing the p – value with null hypothesis. We reject the null hypothesis when p – value is less than the significance level.

$$\chi^2 = \sum_{\text{all } r,c} \frac{(O_{r,c} - E_{r,c})^2}{E_{r,c}} \tag{2.1}$$

## 2.7 EVALUATION METRICS

Evaluation metrics are used to measure the quality of the statistical or machine learning model. Evaluating machine learning models or algorithms is essential for any project. There

are many different types of evaluation metrics available to test a model.

### 2.7.1 ACCURACY SCORE

Accuracy is the most straightforward metric. It calculates the ratio of correctly predicted instances to the total instances in the dataset. It's suitable when class distribution is roughly balanced. However, it can be misleading when dealing with imbalanced datasets.

$$\text{Accuracy} = \frac{\text{Number of Correct Prediction}}{\text{Total Number of Prediction}} \tag{2.2}$$

### 2.7.2 PRECISION

Precision measures the accuracy of positive predictions. It's the ratio of true positives to the total predicted positives. Precision is useful when the cost of false positives is high, and you want to minimize them.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{2.3}$$

### 2.7.3 RECALL

Recall calculates the ability of the classifier to find all the positive instances. It's the ratio of true positives to the total actual positives. Recall is essential when you want to minimize false negatives, such as in medical diagnoses.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{2.4}$$

### 2.7.4 F1 SCORE

The F1-Score is the harmonic mean of precision and recall. It balances precision and recall and is useful when there is an uneven class distribution.

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.5}$$

### 2.7.5 CONFUSION MATRIX

A confusion matrix provides a tabular summary of the classifier's performance, including true positives, true negatives, false positives, and false negatives. These are some of the most common evaluation metrics for classification problems. The choice of metric(s) depends on the specific goals and requirements of your classification task. It is often a good practice to consider multiple metrics to gain a comprehensive understanding of your model's performance.

## 2.8 SOFTWARE USED

### 2.8.1 PYTHON

Python is a versatile and widely-used programming language that plays a pivotal role in this employee attrition prediction project. Its rich ecosystem of libraries and tools, such as Pandas, NumPy, Scikit-Learn, and Seaborn, enables efficient data manipulation, statistical analysis, and machine learning model development. Python's simplicity and readability make it accessible for data scientists and analysts to explore, visualize, and extract meaningful insights from the employee attrition dataset. Its flexibility and scalability also support the implementation of various machine learning algorithms, empowering us to build predictive models to forecast employee turnover accurately. Python's open-source nature fosters a collaborative and innovative environment, making it an indispensable tool in our quest to understand and mitigate employee attrition within the organization.

### 2.8.2 PYTHON STANDARD LIBRARY

The Python Standard Library contains the exact syntax, semantics, and tokens of Python. It contains built-in modules that provide access to basic system functionality like I/O and some other core modules. The Python standard library consists of more than 200 core modules. All these work together to make Python a high-level programming language. Some of the commonly used libraries are:

- **Matplotlib:** This library is responsible for plotting numerical data. And that's why it

is used in data analysis. It is also an open-source library and plots high-defined figures like pie charts, histograms, scatterplots, graphs, etc.

- **Pandas:** Pandas are an important library for data scientists. It is an open-source machine learning library that provides flexible high-level data structures and a variety of analysis tools. It eases data analysis, data manipulation, and cleaning of data. Pandas support operations like Sorting, Re-indexing, Iteration, Concatenation, Conversion of data, Visualizations, Aggregations, etc.

- **Numpy:** The name "Numpy" stands for "Numerical Python". It is the commonly used library. It is a popular machine learning library that supports large matrices and multidimensional data. It consists of in-built mathematical functions for easy computations. Even libraries like TensorFlow use Numpy internally to perform several operations on tensors. Array Interface is one of the key features of this library.

- **SciPy:** The name "SciPy" stands for "Scientific Python". It is an open-source library used for high-level scientific computations. This library is built over an extension of Numpy. It works with Numpy to handle complex computations. While Numpy allows sorting and indexing of array data, the numerical data code is stored in SciPy. It is also widely used by application developers and engineers.

- **Scikit-learn:** It is a famous Python library to work with complex data. Scikit-learn is an open-source library that supports machine learning. It supports variously supervised and unsupervised algorithms like linear regression, classification, clustering, etc. This library works in association with Numpy and SciPy.

# Chapter 3

# ANALYSIS OF THE DATA

## 3.1 EXPLORATORY DATA ANALYSIS

### 3.1.1 DESCRIPTIVE STATISTICS

| | Age | DailyRate | DistanceFromHome | Education | EnvironmentSatisfaction | HourlyRate | JobInvolvement | JobLevel | JobSatisfaction | MonthlyIncome | ... | PercentSalaryHike | RelationshipSatisfaction |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 1470.000000 | 1470.000000 | 1470.000000 | 1470.000000 | 1470.000000 | 1470.000000 | 1470.000000 | 1470.000000 | 1470.000000 | 1470.000000 | ... | 1470.000000 | 1470.000000 |
| **mean** | 36.923810 | 802.485714 | 9.192517 | 2.912925 | 2.721769 | 65.891156 | 2.729932 | 2.063946 | 2.728571 | 6502.931293 | ... | 15.209524 | 2.712245 |
| **std** | 9.135373 | 403.509100 | 8.106864 | 1.024165 | 1.093082 | 20.329428 | 0.711561 | 1.106940 | 1.102846 | 4707.956783 | ... | 3.659938 | 1.081209 |
| **min** | 18.000000 | 102.000000 | 1.000000 | 1.000000 | 1.000000 | 30.000000 | 1.000000 | 1.000000 | 1.000000 | 1009.000000 | ... | 11.000000 | 1.000000 |
| **25%** | 30.000000 | 465.000000 | 2.000000 | 2.000000 | 2.000000 | 48.000000 | 2.000000 | 1.000000 | 2.000000 | 2911.000000 | ... | 12.000000 | 2.000000 |
| **50%** | 36.000000 | 802.000000 | 7.000000 | 3.000000 | 3.000000 | 66.000000 | 3.000000 | 2.000000 | 3.000000 | 4919.000000 | ... | 14.000000 | 3.000000 |
| **75%** | 43.000000 | 1157.000000 | 14.000000 | 4.000000 | 4.000000 | 83.750000 | 3.000000 | 3.000000 | 4.000000 | 8379.000000 | ... | 18.000000 | 4.000000 |
| **max** | 60.000000 | 1499.000000 | 29.000000 | 5.000000 | 4.000000 | 100.000000 | 4.000000 | 5.000000 | 4.000000 | 19999.000000 | ... | 25.000000 | 4.000000 |

Figure 3.1: Descriptive Statistics

### 3.1.2 GRAPHICAL REPRESENTATION OF DATA

#### Pie Chart For EducationalField

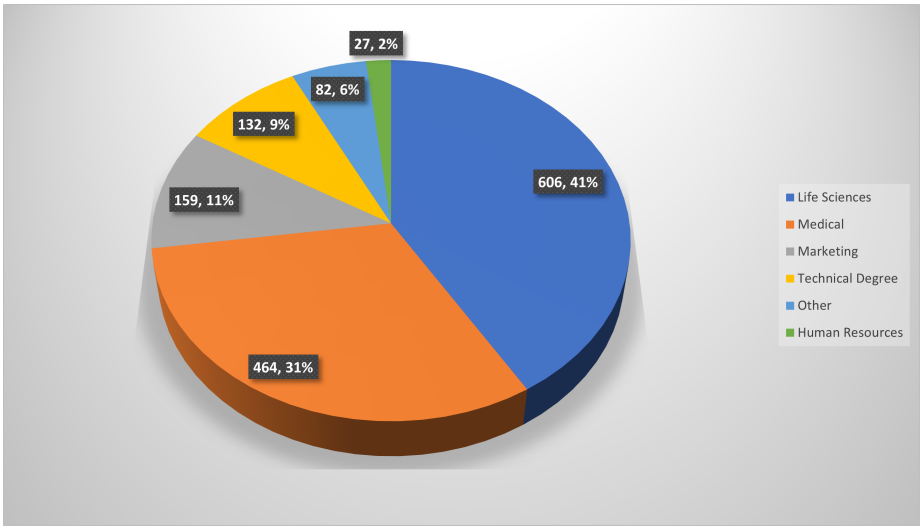


Figure 3.2: Pie Chart Representing EducationField

#### Inference

The pie chart representing the distribution of Education Fields shows that Life Sciences is the most prevalent category, constituting 41% of the total. Following closely is Medical, comprising 31%, while Marketing accounts for 11%. Technical Degree makes up 9% of the distribution, and there is a smaller category labeled as "Others" at 6%. Lastly, Human Resources represents the smallest proportion at 2%. These percentages provide a visual overview of the relative distribution of education fields within the dataset, with Life Sciences and Medical being the dominant categories.

### Pie chart for JobRole

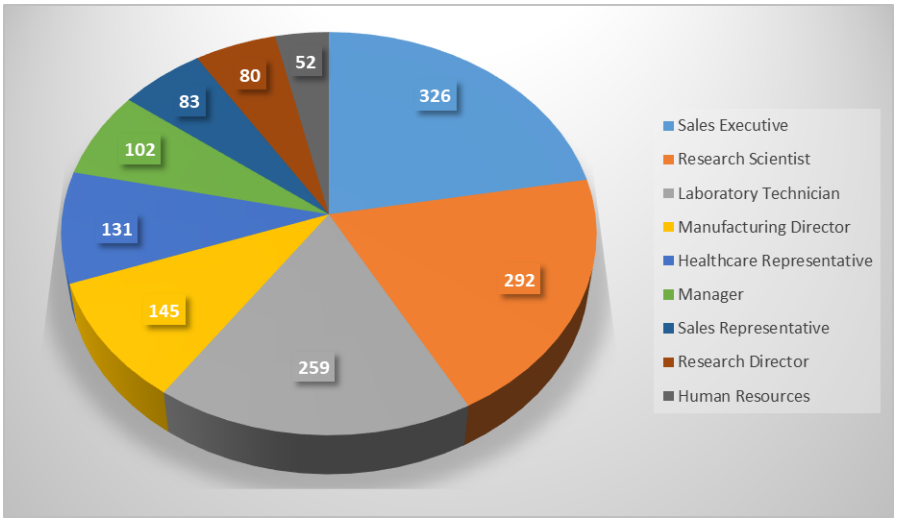


Figure 3.3: JobRole

### Inference

The pie chart illustrates the distribution of job roles within the organization, represented in percentages. Sales Executives occupy the largest portion at 22.8%, followed closely by Research Scientists at 20.5%. Laboratory Technicians make up 18.1% of the workforce, while Manufacturing Directors account for 10.1%. Healthcare Representatives constitute 9.1% of employees, Managers represent 7.1%, Sales Representatives make up 5.7%, Research Directors contribute 5.5%, and Human Resources round out the chart at 3.6%. This visualization provides a clear overview of the proportion of each job role, emphasizing the significant presence of Sales Executives and Research Scientists in the organization.

## Bar Chart Analysis of Variables

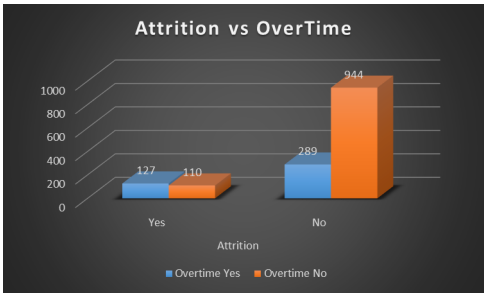


Figure 3.4: Bar Plot of Attrition vs Overtime

### Inference

The bar chart shows attrition cases categorized by overtime status. In 'Attrition: Yes,' more cases are observed in 'Overtime: Yes,' suggesting attrition occurs more frequently among overtime workers. Conversely, in 'Attrition: No,' the majority of cases are in 'Overtime: No,' indicating a lower attrition rate among non-overtime workers.

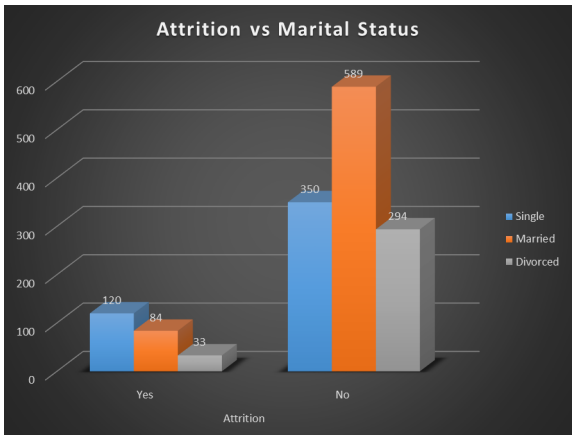


Figure 3.5: Bar Plot of Attrition vs Marital Status

### Inference

Attrition cases are distributed differently among marital status groups. Specifically, among employees who experienced attrition, there appears to be a higher proportion of 'Single' individuals compared to 'Married' or 'Divorced' individuals. Conversely, among employees who did not experience attrition, 'Married' individuals make up the largest group, followed by 'Single' and 'Divorced' individuals.

## NUMERICAL VARIABLE ANALYSIS

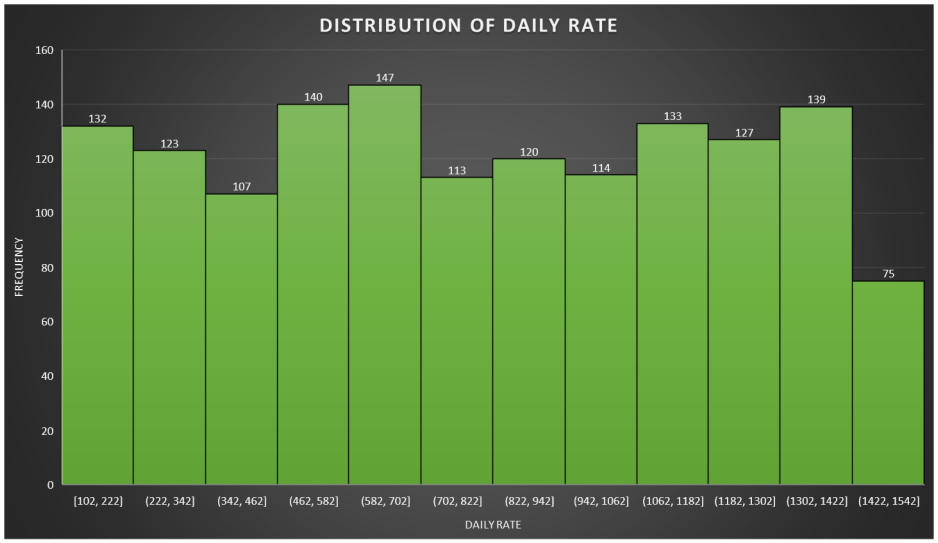


Figure 3.6: Distribution of DailyRate

### Inference

The histogram shows that the majority of daily rates fall within the lower range, with the highest frequency occurring in the range of 500 to 700. There are also noticeable peaks in the 1000 to 1182 and 1300 to 1400 ranges. This histogram provides insights into the distribution of daily rates among employees in the organization, which can be valuable for understanding compensation structures and potentially identifying patterns related to employee attrition..

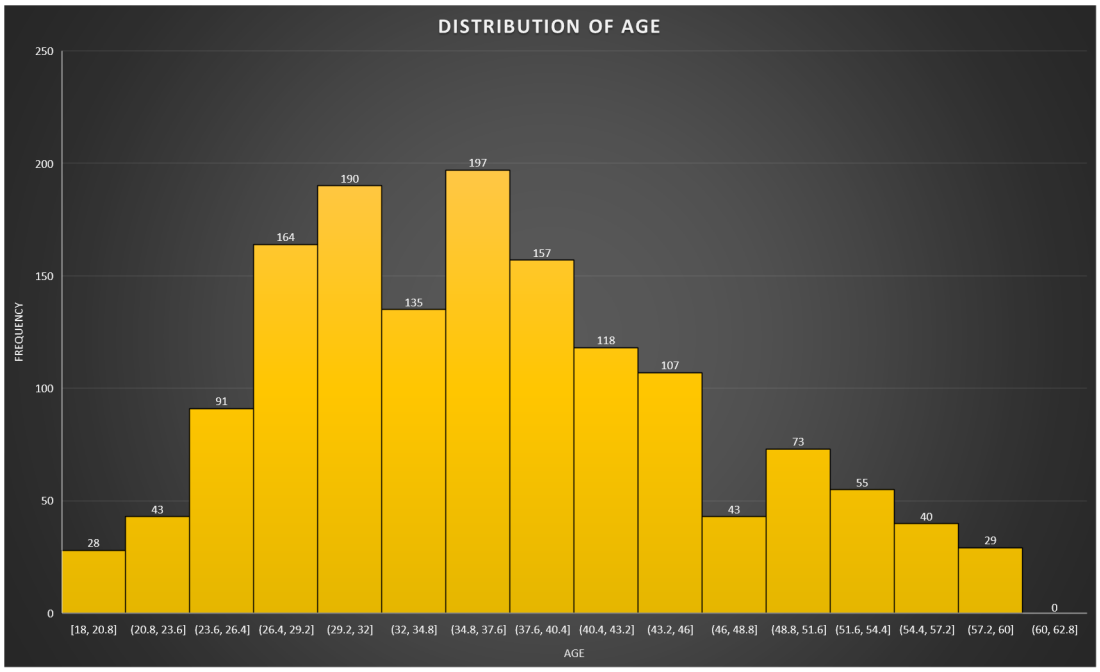


Figure 3.7: Distribution of Age

### Inference

The histogram shows that the majority of employees fall within the age range of 29 to 38 years, with a peak around the ages of 34 and 37. There is also a noticeable presence of employees in their early 40s and a smaller number of employees in their mid to late 50s. This histogram provides a clear representation of the age demographics within the organization, which can be valuable for various analytical purposes, including understanding workforce dynamics and employee attrition patterns.

## Heatmap

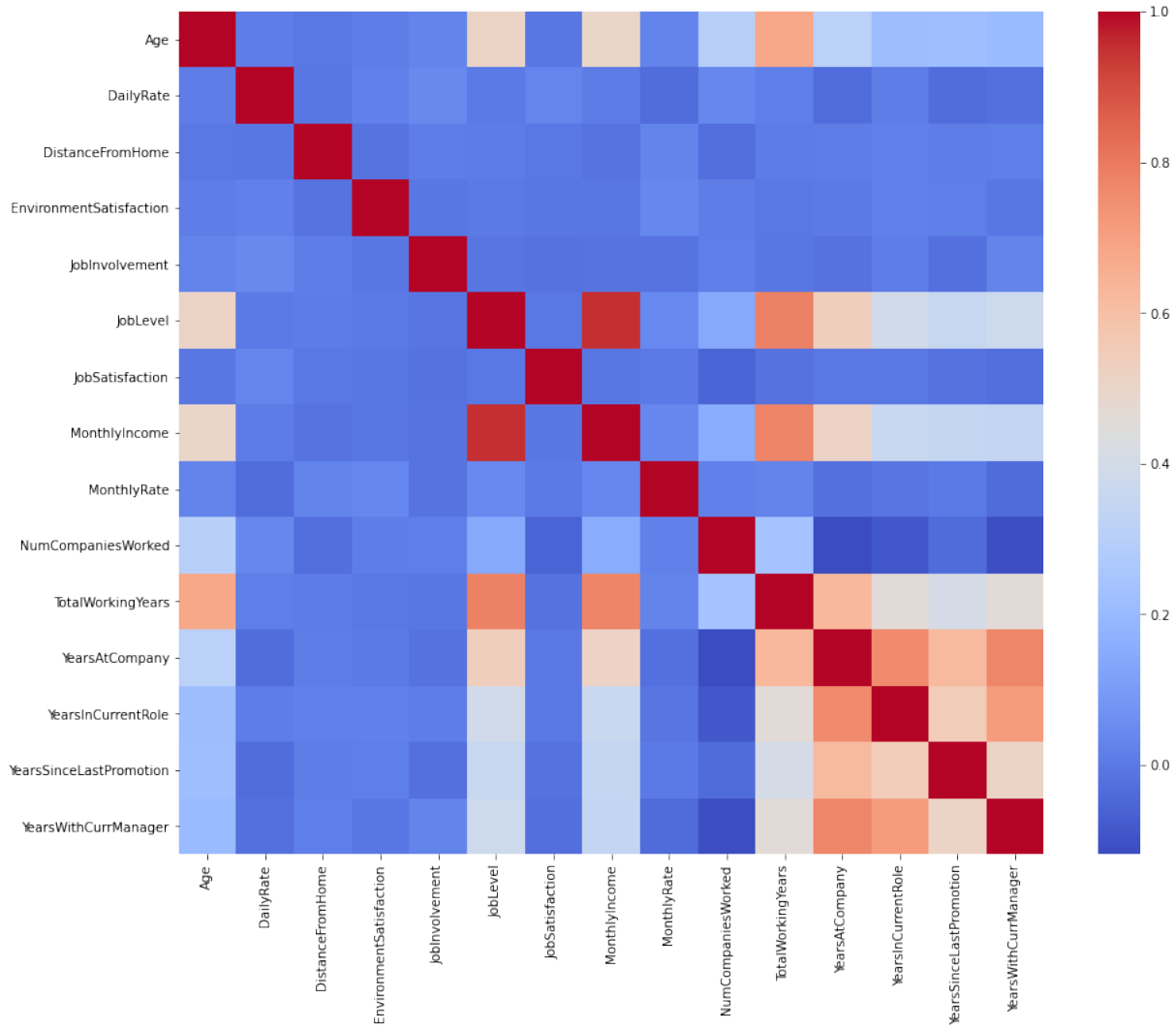


Figure 3.8: Heatmap

Heatmaps visualize the data in a 2-dimensional format in the form of coloured maps. It shows relationships between attributes.

# Chapter 4

# IMPLEMENTATION

## 4.1 DECISION TREE

### 4.1.1 CLASSIFICATION REPORT

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.88 | 0.77 | 0.82 | 247 |
| 1 | 0.28 | 0.47 | 0.35 | 47 |
| accuracy | | | 0.72 | 294 |
| macro avg | 0.58 | 0.62 | 0.58 | 294 |
| weighted avg | 0.79 | 0.72 | 0.74 | 294 |

Figure 4.1: CLASSIFICATION REPORT OF DECISION TREE

### 4.1.2 CONFUSION MATRIX

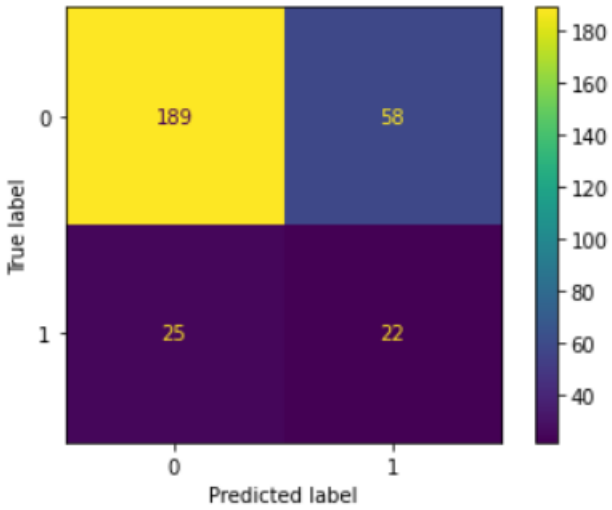


Figure 4.2: CONFUSION REPORT OF DECISION TREE

### 4.1.3 INFERENCE

The confusion matrix shows 189+22 = 211 correct predictions and 58+25 =83 incorrect ones.

True Positives = 22

True Negative = 189

False Positive = 58(Type 1 error)

False Negative = 25 (Type 2 error)

Accuracy = 72%

### 4.1.4 TOP 10 FEATURE IMPORTANCE ATTRIBUTES(DECISION TREES)

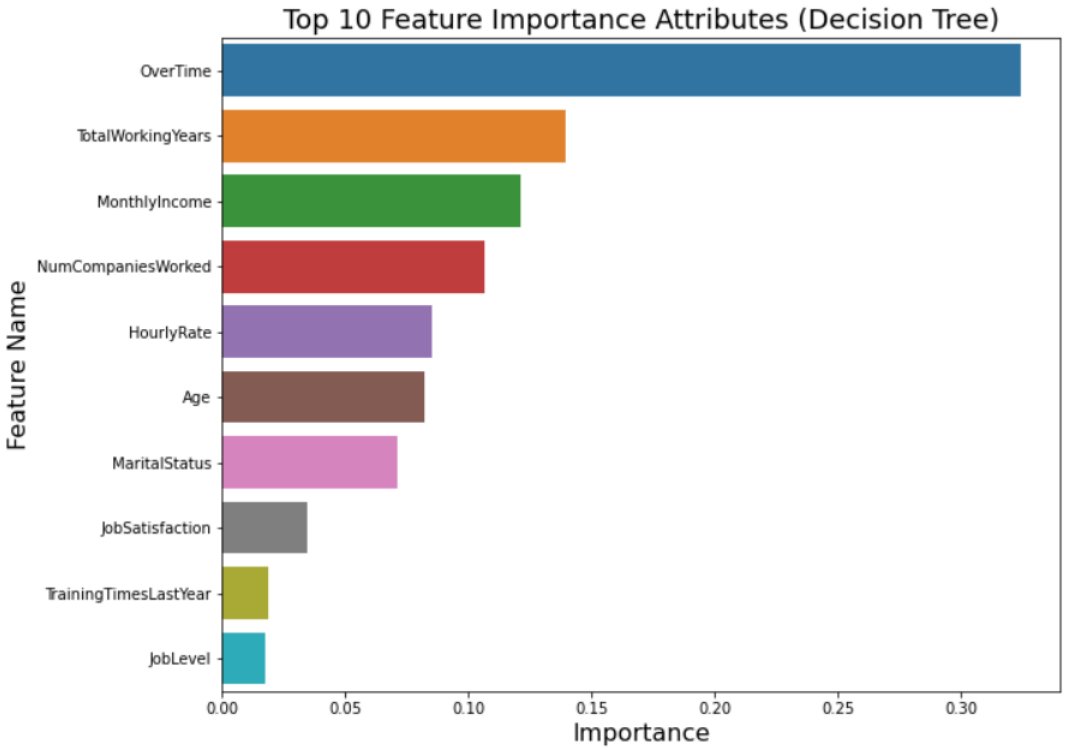


Figure 4.3:

## 4.2 RANDOM FOREST CLASSIFIER

### 4.2.1 CLASSIFICATION REPORT

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.87 | 0.99 | 0.92 | 247 |
| 1 | 0.77 | 0.21 | 0.33 | 47 |
| accuracy | | | 0.86 | 294 |
| macro avg | 0.82 | 0.60 | 0.63 | 294 |
| weighted avg | 0.85 | 0.86 | 0.83 | 294 |

Figure 4.4: CLASSIFICATION REPORT OF RANDOM FOREST

### 4.2.2 CONFUSION MATRIX

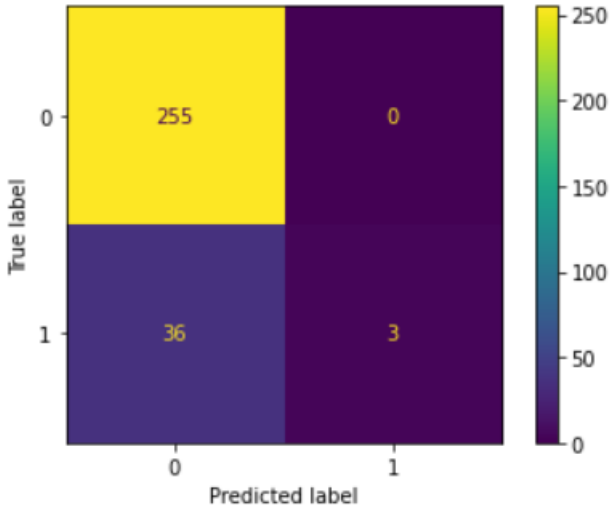


Figure 4.5: CONFUSION REPORT OF RANDOM FOREST

### 4.2.3 INFERENCE

The confusion matrix shows 255+3 = 258 correct predictions and 36+0 =36 incorrect ones.

True Positives = 3

True Negative = 255

False Positive = 0(Type 1 error)

False Negative = 36 (Type 2 error)

Accuracy = 86%

### 4.2.4 TOP 10 FEATURE IMPORTANCE ATTRIBUTES(RANDOM FOREST)

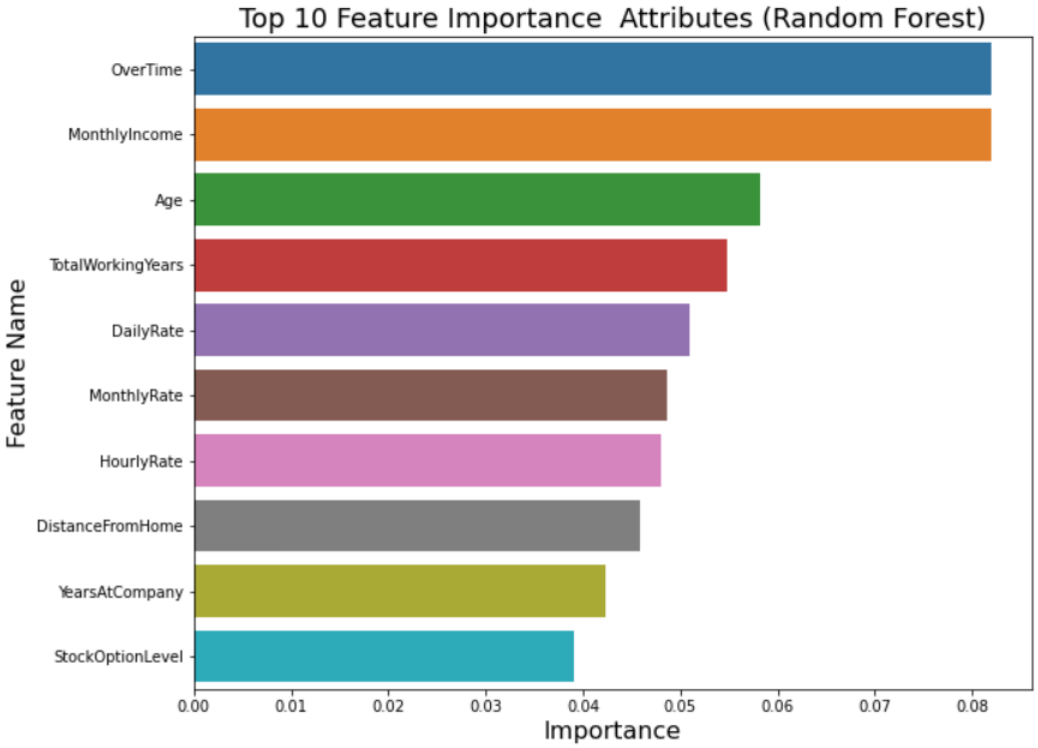


Figure 4.6:

## 4.3 XGBOOST CLASSIFIER

### 4.3.1 CLASSIFICATION REPORT

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| No | 0.88 | 1.00 | 0.93 | 255 |
| Yes | 1.00 | 0.08 | 0.14 | 39 |
| accuracy | | | 0.88 | 294 |
| macro avg | 0.94 | 0.54 | 0.54 | 294 |
| weighted avg | 0.89 | 0.88 | 0.83 | 294 |

Figure 4.7: CLASSIFICATION REPORT OF XGBOOST CLASSIFIER

### 4.3.2 CONFUSION MATRIX

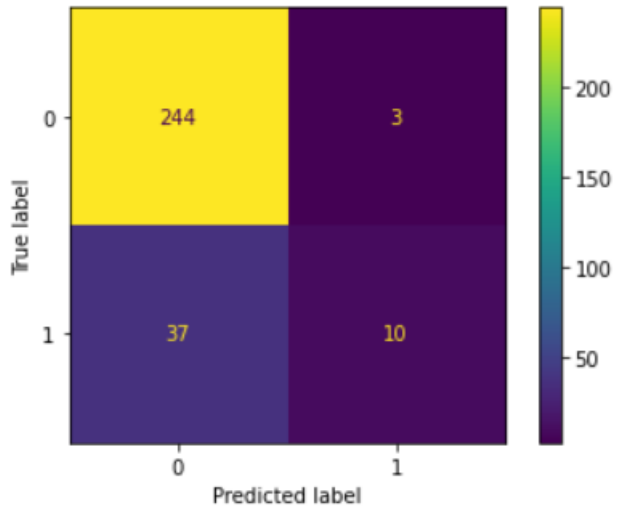


Figure 4.8: CONFUSION REPORT OF XGBOOST CLASSIFIER

### 4.3.3 INFERENCE

The confusion matrix shows 244+10 = 254 correct predictions and 37+3 =40 incorrect ones.

True Positives = 10

True Negative = 244

False Positive = 3(Type 1 error)

False Negative = 37 (Type 2 error)

Accuracy = 88%

### 4.3.4 TOP 10 FEATURE IMPORTANCE ATTRIBUTES(XGBOOST CLASSIFIER)

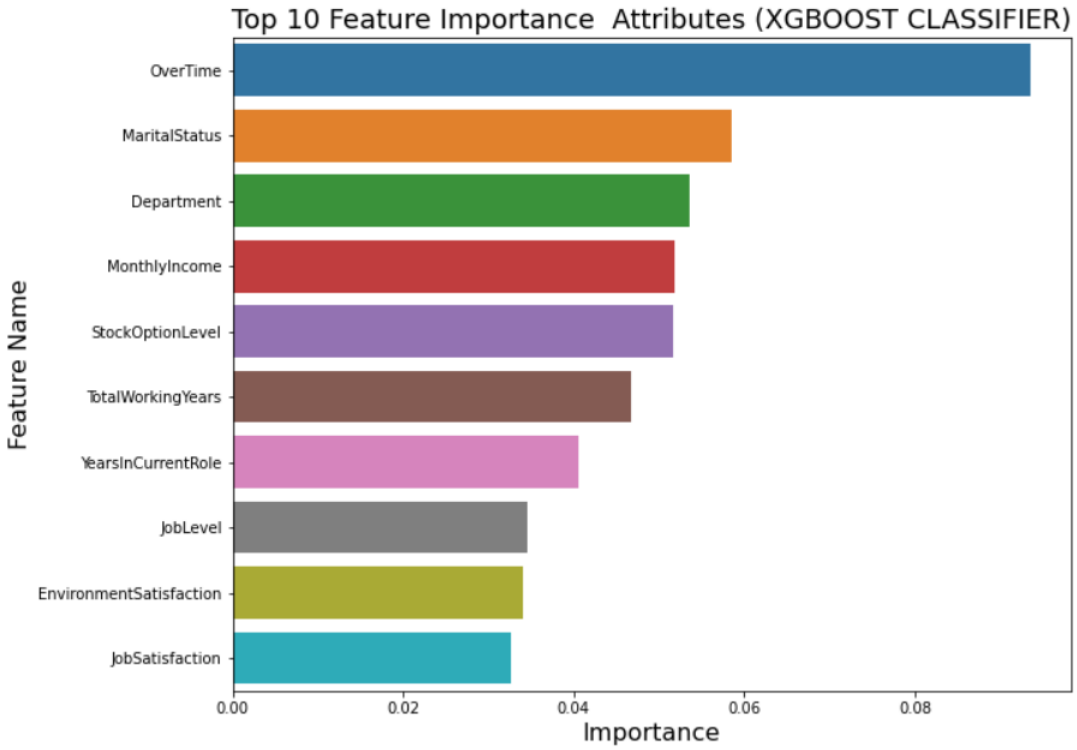


Figure 4.9:

## 4.4 SUPPORT VECTOR MACHINE

### 4.4.1 CLASSIFICATION REPORT

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| No | 0.91 | 0.96 | 0.94 | 255 |
| Yes | 0.62 | 0.41 | 0.49 | 39 |
| accuracy | | | 0.89 | 294 |
| macro avg | 0.76 | 0.69 | 0.71 | 294 |
| weighted avg | 0.87 | 0.89 | 0.88 | 294 |

Figure 4.10: CLASSIFICATION REPORT OF SVM

### 4.4.2 CONFUSION MATRIX

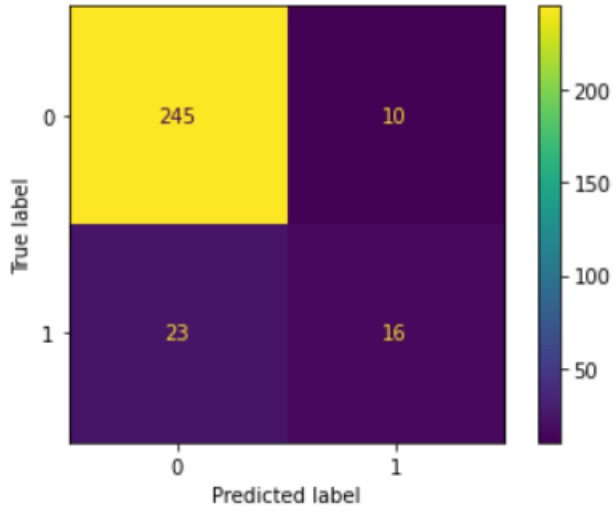


Figure 4.11: CONFUSION REPORT OF SVM

### 4.4.3 INFERENCE

The confusion matrix shows 245+16 = 261 correct predictions and 10+23 =33 incorrect ones.

True Positives = 16

True Negative = 245

False Positive = 10(Type 1 error)

False Negative = 23 (Type 2 error)

Accuracy = 89%

### 4.4.4 TOP 10 FEATURE IMPORTANCE ATTRIBUTES(SVM)

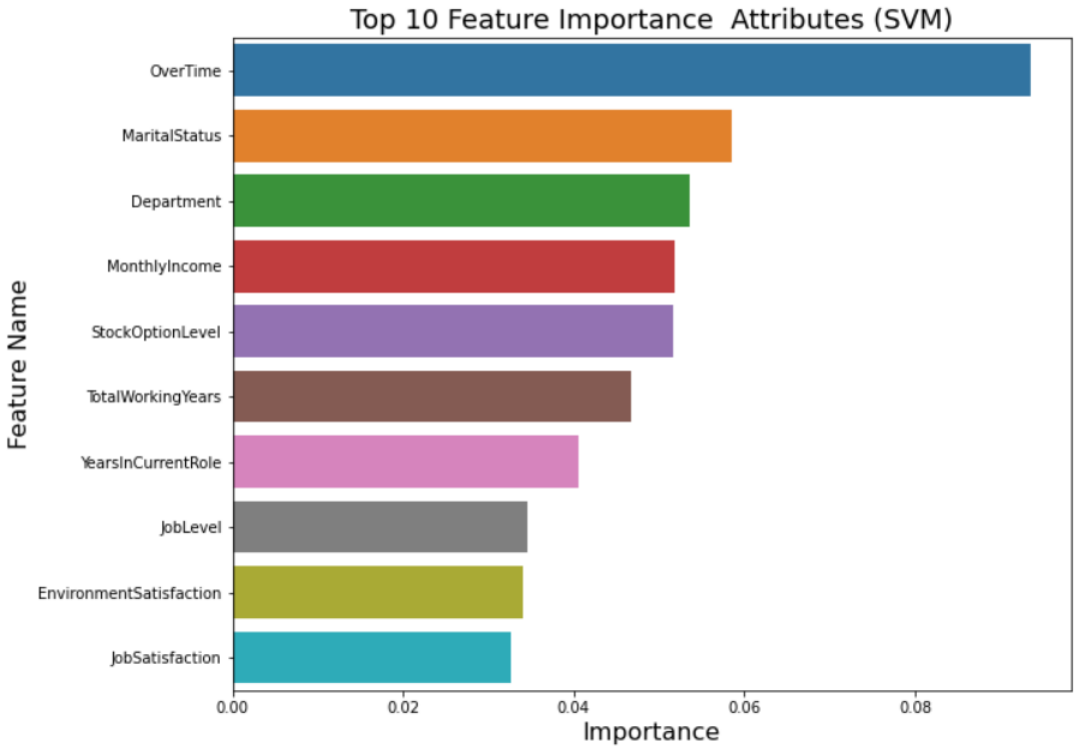


Figure 4.12:

# Chapter 5

# COMPARISON OF MODELS AND PREDICTION

## 5.1 INTRODUCTION

The optimal classification model is chosen in this chapter by comparing metrics like accuracy, precision, recall, and f1-score. Then, using this model, employee attrition predictions of certain randomly chosen samples are made.

## 5.2 COMPARISON OF ACCURACY

The comparison of accuracies obtained from various algorithms is as shown in the figure.

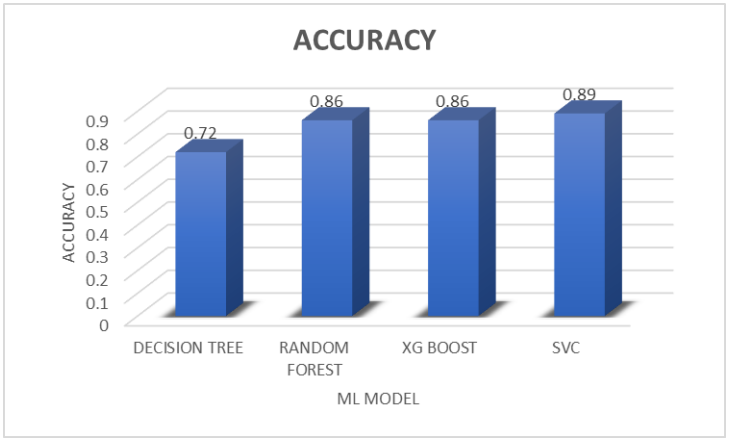


Figure 5.1: Accuracy vs ML Models

**INFERENCE**

Out of all the algorithms chosen, Support Vector Machine Classifier performs best with an accuracy of 89 %.

## 5.3 COMPARISON OF PRECISION

The comparison of precision obtained from various algorithms for each class is as shown in the figure.

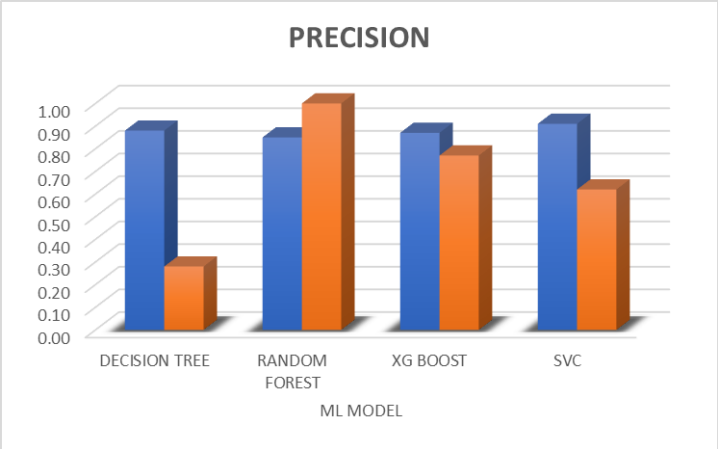


Figure 5.2: Precision vs ML Models

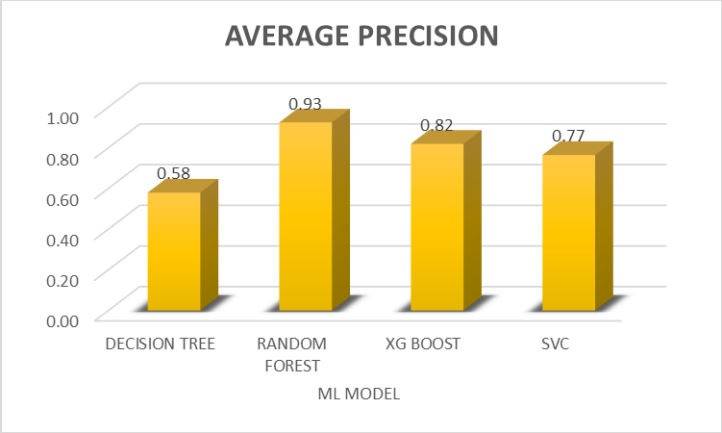


Figure 5.3: Average Precision vs ML Models

**INFERENCE**

Out of all the algorithms chosen, Random Forest performs best with an average precision of 93 %.

## 5.4 COMPARISON OF RECALL

The comparison of recall obtained from various algorithms for each class is as shown in the figure.

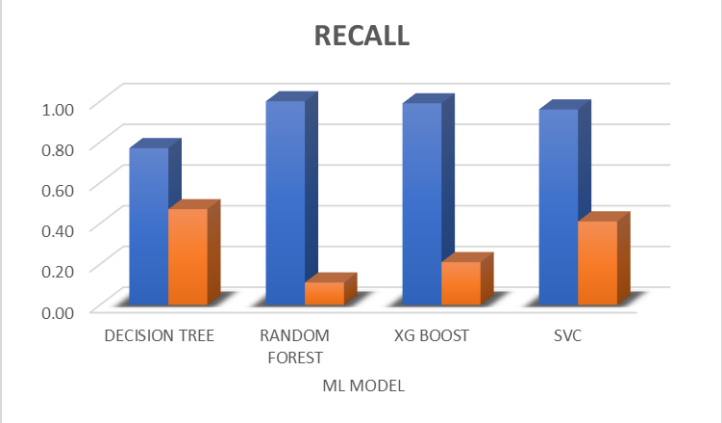


Figure 5.4: Precision vs ML Models

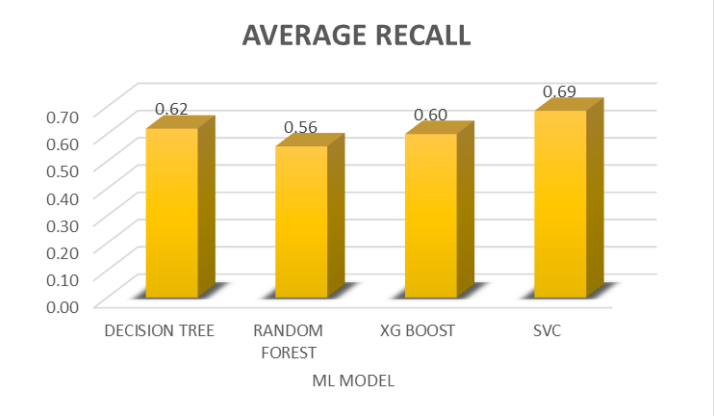


Figure 5.5: Average Precision vs ML Models

**INFERENCE**

Out of all the algorithms chosen, Support Vector Machine Classifier performs best with an average recall of 69 %.

## 5.5 COMPARISON OF F1 SCORE

The comparison of recall obtained from various algorithms for each class is as shown in the figure.

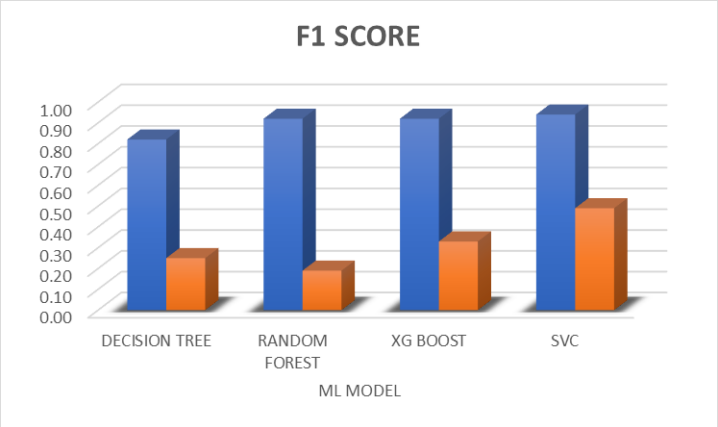


Figure 5.6: Precision vs ML Models

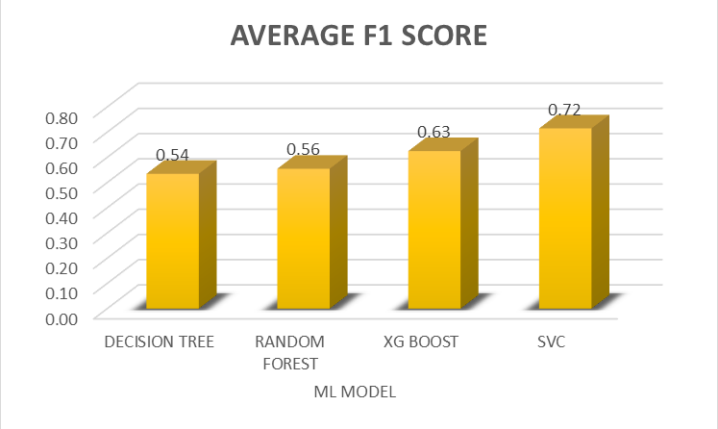


Figure 5.7: Average Precision vs ML Models

**INFERENCE**

Out of all the algorithms chosen, Support Vector Machine Classifier performs best with an average recall of 72 %.

## 5.6 PREDICTION

Out of all the algorithms chosen Support Vector Machine Classifier performs best. We make a prediction by giving new variables to our model.

**INPUT**

| FEATURES | INPUT VALUES |
|---|---|
| 'Age' | 41 |
| 'BusinessTravel' | 'Travel_Rarely' |
| 'DistanceFromHome' | 1 |
| 'Education' | 2 |
| 'EnvironmentSatisfaction' | 2 |
| 'JobSatisfaction' | 4 |
| 'Department' | 'Sales' |
| 'MonthlyIncome' | 5993 |
| 'DailyRate' | 1102 |
| 'EducationField' | 'Life Sciences' |
| 'EmployeeCount' | 1 |
| 'EmployeeNumber' | 1 |
| 'Gender' | 'Female' |
| 'HourlyRate' | 94 |
| 'JobInvolvement' | 3 |
| 'JobLevel' | 2 |
| 'JobRole' | 'Sales Executive' |
| 'MaritalStatus' | 'Single' |
| 'MonthlyRate' | 19479 |
| 'NumCompaniesWorked' | 8 |
| 'Over18' | 'Y' |
| 'OverTime' | 'Yes' |
| 'PercentSalaryHike' | 11 |
| 'PerformanceRating' | 3 |
| 'RelationshipSatisfaction' | 1 |
| 'StandardHours' | 80 |
| 'StockOptionLevel' | 0 |
| 'TotalWorkingYears' | 8 |
| 'TrainingTimesLastYear' | 0 |
| 'WorkLifeBalance' | 1 |
| 'YearsAtCompany' | 6 |
| 'YearsInCurrentRole' | 4 |
| 'YearsSinceLastPromotion' | 0 |
| 'YearsWithCurrManager' | 5 |

Figure 5.8: INPUT

**OUTPUT**

```
In [84]: print("Predicted Attrition:", predicted_attrition)
         Predicted Attrition: Yes
```

Figure 5.9: OUTPUT

# Chapter 6

# CONCLUSION

In conclusion, the Employee Attrition Prediction project has provided valuable insights into the critical factors influencing attrition within our organization. Through a rigorous analysis that involved graphical representations and the application of the chi-square technique, we successfully identified the primary predictors of employee attrition. This project's primary objective was to determine and analyze these key factors, which can significantly impact our workforce stability.

Our study also goes into hypothesis testing, using two-sample t-test to validate the existence of significant differences in working years between employees who choose to leave and those who decide to stay. This step helped confirm the importance of these demographic variables in the context of attrition.

Furthermore, we ventured into predictive modeling, employing a variety of machine learning algorithms, including Decision Trees, Random Forest, XGBoost Classifier, and Support Vector Machine (SVM). These models were rigorously evaluated and compared to select the best-performing one for predicting employee attrition. Based on the comprehensive model comparison, it is evident that the Support Vector Machine (SVM) emerged as the most accurate model, achieving an impressive accuracy rate of 89%.

| Model | Accuracy | Precision | Recall | f1-score |
|---|---|---|---|---|
| DECISION TREE | 72% | 0.58 | 0.62 | 0.58 |
| RANDOM FOREST | 86% | 0.93 | 0.55 | 0.56 |
| XG BOOST | 86% | 0.82 | 0.60 | 0.63 |
| SUPPORT VECTOR MACHINE | 89% | 0.76 | 0.69 | 0.71 |

The findings of this project not only enhance our understanding of the dynamics behind employee attrition but also provide a practical tool for identifying and addressing potential attrition cases within our organization. By utilizing the SVM model, we can proactively implement targeted retention strategies and human resource initiatives, ultimately contributing to a more stable and engaged workforce. This project underscores the importance of data-driven decision-making in talent management and aligns with our ongoing commitment to improving employee satisfaction and organizational success.

## 6.1 FUTURE SCOPE OF THE STUDY

The future scope of the Employee Attrition Prediction study is promising and opens doors to several avenues for further research and application:

- Enhanced Prediction Models: Continuously improving machine learning algorithms and techniques offer opportunities to develop even more accurate prediction models. Researchers can explore advanced ensemble methods, deep learning architectures, and hybrid models to achieve higher predictive performance.

- Real-time Monitoring: Extending the model to real-time attrition monitoring can provide organizations with timely alerts when employees exhibit attrition risk factors. This can enable proactive intervention strategies to retain valuable talent.

- Longitudinal Analysis: Future studies can focus on longitudinal data analysis, tracking employee behavior and satisfaction over time. This can help identify trends and patterns that precede attrition, enabling organizations to intervene early.

- Feature Engineering: Investigating novel features and attributes that contribute to attrition can lead to a more comprehensive understanding of the problem. Exploring sentiment analysis of employee feedback, social network analysis, or external economic indicators may provide additional insights.

- Interpretable Models:Enhancing the interpretability of predictive models can aid in understanding the reasons behind predictions. Developments in explainable AI and model interpretability techniques can make the decision-making process more transparent .

- Cultural and Industry Variations:Recognizing that factors influencing attrition can vary across industries and cultures, future research can focus on building customized models for specific sectors or regions.

Overall, the future scope of the study lies in continuous refinement and innovation in predictive models, as well as their ethical and practical application in human resource management. As organizations increasingly recognize the importance of talent retention, attrition prediction models will play a vital role in shaping HR strategies and workforce dynamics.

## 6.2 BIBILIOGRAPHY


- Andreas C Muller and Sarah Gudio, 2016, Introduction to Machine Learning with Python, O Reilly Media,Inc
- Shai Shalev-Shwartz and Shai Ben-David,2014, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press
- Ethem Alpaydin,2014 , Introduction to Machine Learning, Third edition, Cambridge,Massacgusetts Lundon,England,The MIT Press.
- Norman Matloff, Stastistical Regression and Classification From Linear Models to Machine Learning,Boca Raton London New York , CRC Press.
- Peter Bruce and Andrew Bruce,2017, Practica Statistics for Data Scientists 50 Essential Concepts, O Reilly Media, Inc.
- Subhashini, T. S., Hemalatha, M. (2014). A survey on employee attrition prediction. International Journal of Science and Research (IJSR), 3(12), 1157-1160.
- Sasirekha, M., Kumar, S. G. (2015). Employee attrition prediction: A hybrid approach. Procedia Computer Science, 47, 52-58.
- Dataset:https://www.kaggle.com/datasets/patelprashant/employee-attrition