Project Report

On

EMPLOYEE ATTRITION PREDICTION

Submitted

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

APPLIED STATISTICS AND DATA ANALYTICS

by

NAJEEBA K A

(Register No. SM21AS014)

(2021-2023)

Under the Supervision of

ANAKHA KURIAKOSE

DEPARTMENT OF MATHEMATICS AND STATISTICS

ST. TERESA'S COLLEGE (AUTONOMOUS)

ERNAKULAM, KOCHI - 682011

APRIL 2023

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

CERTIFICATE

This is to certify that the dissertation entitled, **EMPLOYEE ATTRITION PREDICTION** is a bonafide record of the work done by Ms. **NAJEEBA K A** under my guidance as partial fulfillment of the award of the degree of **Master of Science in Applied Statistics and Data Analytics** at St. Teresa's College (Autonomous), Ernakulam affiliated to Mahatma Gandhi University, Kottayam. No part of this work has been submitted for any other degree elsewhere.

Date:
Place: Ernakulam

ANAKHA KURIAKOSE
Assistant Professor,
Department of Mathematics and Statistics,
St. Teresa's College(Autonomous),
Ernakulam.

Smt Betty Joseph
Associate Professor,
Department of Mathematics and Statistics,
St. Teresa's College(Autonomous),
Ernakulam.

External Examiners

1:..............................

2: ....

DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of **ANAKHA KURIAKOSE**, Assistant Professor, Department of Mathematics and Statistics, St. Teresa's College(Autonomous), Ernakulam and has not been included in any other project submitted previously for the award of any degree.

Ernakulam. **NAJEEBA K A**

Date: **SM21AS014**

ACKNOWLEDGEMENTS

I express my deep gratitude to Ms. Anakha Kuriakose, Assistant Professor, Department of Mathematics and Statistics, St Teresa's College, Ernakulam for always motivating and encouraging me to learn and research everything independently.

I must mention several individuals who encouraged me to carry out this work. Their continuous invaluable knowledgeable guidance throughout the course of this study helped me to complete the work up to this stage

In addition, the very energetic and competitive atmosphere of the Department had much to do with this work. I acknowledge with thanks to faculty, teaching and non-teaching staff of the department, and Colleagues.

Above all, I must thank God almighty and my parents for giving me the blessings to take over this project.

Ernakulam. NAJEEBA K A

Date: **SM21AS014**

ABSTRACT

Employees are the most important assets of an organization. Hiring new will always take more effort and cost rather than retaining the old ones. High employee turnover can hurt an organization's performance, productivity, and bottom line. This is why organizations need to understand the reasons why their employees leave and develop strategies to reduce turnover. In this project, I developed a model to predict employee attrition using machine learning algorithms. The goal of the model is to predict the likelihood of future employees leaving the company, and thus help the company identify at-risk employees and take proactive measures to retain them. By providing organizations with the insights and tools to address the dissatisfaction factors of their employees, this system can help organizations proactively manage employee turnover and improve overall organizational performance. By reducing overall employee turnover, the model aims to improve business performance and support the company's growth. The project will explore different algorithms and techniques to identify the best-performing model and fine-tune it for optimal results. In this paper, I try to build a system that will predict employee attrition based on the Employee dataset from the Kaggle website. I generated a heat map to show the relations between the attributes. For prediction purposes, used four different machine learning algorithms KNN (K-Nearest Neighbor), SVM (Support Vector Machine), and Random Forest.

Contents

Chapter 1

INTRODUCTION

1.1 EMPLOYEE ATTRITION

Employee attrition refers to the loss of employees through resignations, retirements, and terminations. High levels of attrition can hurt a company's productivity, bottom line, and morale among remaining employees. Employers can take various steps to reduce employee attrition, such as offering competitive compensation and benefits, improving working conditions, and investing in employee development and training programs. Additionally, it is important to understand the reasons behind employee attrition in order to take appropriate actions to address the issues and improve retention. Employers can conduct exit interviews, employee surveys, or other forms of research to gather information on why employees are leaving. By identifying the root causes of attrition, organizations can develop effective strategies to improve employee satisfaction, reduce turnover, and ultimately increase productivity and profitability.

Employee attrition prediction is important for several reasons, including:

Cost reduction: Employee turnover can be a costly process for organizations. It involves expenses such as recruiting, training, and onboarding new employees. By predicting which employees are likely to

leave, organizations can proactively work on retaining them, reducing the cost of turnover.

Talent retention: Employee turnover can result in a loss of valuable knowledge, skills, and experience. Predicting which employees are likely to leave can help organizations identify the factors that contribute to turnover and take steps to retain valuable employees.

Succession planning: Knowing which employees are likely to leave in the near future can help organizations plan for succession and ensure that critical positions are not left vacant.

Organizational performance: High levels of employee turnover can have a negative impact on organizational performance. By predicting which employees are likely to leave, organizations can take steps to address the underlying causes of turnover and improve overall performance.

It is possible to use machine learning algorithms to predict employee attrition and identify potential reasons for an employee's decision to leave the organization. By analyzing data such as job performance, job satisfaction, and demographic information, machine learning models can identify patterns and trends that may indicate a higher likelihood of an employee leaving the organization. These predictions can then be used to inform employee retention strategies, such as targeted interventions or incentives to improve job satisfaction and engagement.

1.2 OBJECTIVES

1. **Identify the factors that affect employee attrition in an organization.**

2. **Build a model that predicts whether an employee stays or leaves the organization using machine learning algorithms such as Random Forest, Support Vector Machine, and K-Nearest Neighbors.**

Chapter 2

LITERATURE REVIEW

Scholars have examined employee attrition from a variety of angles.Some of them are discussed here:

1. **To determine the factors that led employees to choose to remain with or quit the company, some research examined their behaviors. Some investigations employed algorithms for machine learning to forecast employee attrition based on their history. The machine learning models utilized by Alduayj and Rajpoot included random forests, k-nearest neighbors, and support vector machines with various kernel functions. The original class-imbalanced dataset, artificial over-sampled datasets, and artificial under-sampled datasets were all used. Although their approach with the synthetic dataset demonstrated great accuracy, its accuracy with the genuine dataset was not sufficient.**

2. **The same dataset was used by Usha and Balaji to assess the performance of the decision tree, naive Bayes, and k-means for prediction. They used 70%:30% splits for train and test sets and 10-fold cross-validation to validate the algorithms. Their work is less accurate than other works in comparison. This is because they did not use the data preprocessing stage in their work.**

3. **In their study of the factors that influence an employee to leave**

an organization, Fallucchi et al. used a variety of machine learning algorithms to find the most effective classifier. These methods include Naive Bayes, logistic regression, k-nearest neighbor, decision trees, random forests, and support vector machines. They cross-validated their work and split the train-test set, however, they didn't mention cross-validation in their results, which only included the 70%:30% split train-test set. The test accuracy, nevertheless, is higher than the training accuracy, which is a positive sign but might yet be enhanced.

4. An attrition prediction paradigm with three stages was presented by Zangeneh et al. They first reduced the amount of data by using the "max-out" feature selection method. They trained a logistic regression model for prediction in the second step. The third stage then includes confidence analysis to validate the prediction model. Because of the preprocessing and postprocessing, the system has considerable complexity in addition to its poor accuracy. Classification trees and random forests were employed by Pratt to predict attrition. They preprocess the data using Pearson correlation to remove undesirable features prior to categorization. Nonetheless, when compared to other machine learning techniques, their work demonstrates a small boost in accuracy.

Chapter 3

METHODOLOGY

3.1 DATA SOURCE

For this study, the HR dataset named 'IBM HR ANALYTICS EMPLOYEE ATTRITION and PERFORMANCE', has been picked, which is available on the IBM website.

3.2 DATA DESCRIPTION

The dataset used in this study was produced by IBM Analytics. It contains 35 features for 1470 employees. The table given below displays the dataset features and their corresponding types. The "Attrition" feature represents an employee's choice: Either Yes (quit the firm) or No (stay at the company)
Table 3.1 shows IBM dataset features.

Table 3.1: Dataset features

Feature Name	Type
Age	Number
Business Travel	Category
DailyRate	Number
Department	Category
Distance from Home	Number
Education	Category
Education Field	Category
Employee Count	Number
employee number	Number
Environment Satisfaction	Category
Gender	Category
HourlyRate	Number
JobInvolvment	Category
JobLevel	Category
Educationfeild	Category
JobRole	Category
Jobsatisfaction	Category
MaritalStatus	Category
MonthlyIncome	Number
MonthlyRate	Number
NumCompaniesWorked	Number
Over18	Category
Overtime	Category
PercentSalaryHike	Number
PerformanceRating	Number
RelationshipSatisfaction	Category
StandardHours	Number
StockOptionLevel	Category
TotalWorkingYears	Number
TrainingTimesLastYear	Number
WorklifeBalance	Category
YearsAtCompany	Number
YearsInCurrentRole	Number
YearsSinceLastPromotion	Number
YearsWithCurrentManager	Number
Attrition	Category

3.3 EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a statistical approach used to analyze and summarize datasets. The goal of EDA is to gain an understanding of the data, identify patterns and relationships, and discover any anomalies or outliers.

EDA involves various techniques such as data visualization, summary statistics, and hypothesis testing. It is typically used in the early stages of data analysis before formal statistical modeling is performed.
Some of the key steps in EDA include:

- Data collection and cleaning: This involves acquiring the data and checking for any errors or missing values.
- Data visualization: This involves creating visual representations of the data, such as scatter plots, histograms, and box plots, to identify patterns and relationships.
- Summary statistics: This involves calculating measures such as mean, median, and standard deviation to summarize the data.
- Hypothesis testing: This involves testing hypotheses about the data using statistical tests such as t-tests and ANOVA.

Overall, EDA provides a way to gain insight into the data and identify any potential issues that may need to be addressed before performing more advanced analysis.

3.4 MACHINE LEARNING

Machine learning is a subfield of artificial intelligence (AI) that involves building computer algorithms that can automatically learn from data and improve their performance on a given task over time. The goal of machine learning is to develop systems that can identify patterns and relationships in data, and use that knowledge to make predictions or take actions.

There are three main types of machine learning: Supervised learning, Unsupervised learning, and Reinforcement learning.

3.4.1 Supervised Machine Learning

Supervised machine learning is a subfield of machine learning where a model is trained on labeled data to make predictions or classify new data. In supervised learning, the input data is paired with the correct output, and the goal is to learn a mapping function from inputs to outputs that can generalize to new, unseen data.

The labeled data is typically divided into two sets: a training set and a test set. The model is trained on the training set, and its performance is evaluated on the test set. The test set is used to measure the accuracy of the model and ensure that it can generalize well to new, unseen data.

Some popular algorithms used in supervised learning include linear regression, logistic regression, decision trees, random forests, support vector machines (SVMs), and artificial neural networks (ANNs). These algorithms are used in various applications such as image and speech recognition, natural language processing, recommendation systems, fraud detection, and predictive analytics.

3.4.2 Unsupervised Machine Learning

Unsupervised machine learning is a subfield of machine learning where a model is trained on unlabeled data to discover patterns and relationships without any prior knowledge of what the output should be. In unsupervised learning, the goal is to identify hidden structures or clusters in the data that can help us gain insights or make better decisions.

Unlike supervised learning, there is no specific output variable to predict in unsupervised learning. Instead, the algorithm tries to find meaningful patterns in the data that can be used for further analysis or decision-making.

Various algorithms used in unsupervised learning are clustering algorithms, dimensionality reduction algorithms and anomaly detection algorithms.

3.4.3 Reinforcement Machine Learning

Reinforcement learning is a type of machine learning where an agent learns to make decisions by interacting with an environment and receiving feedback in the form of rewards or penalties. In reinforcement learning, the goal is to learn a policy that can maximize the expected reward over time.

The agent takes actions in the environment based on the current state and receives feedback in the form of a reward signal. The reward signal tells the agent whether its actions were good or bad, and the goal is to learn a policy that can maximize the cumulative reward over time.

Reinforcement learning can be used in various applications such as robotics, game-playing, and autonomous vehicles. Some popular algorithms used in reinforcement learning include Q-learning, deep Q-networks, and policy gradient.

3.5 MODEL BUILDING

3.5.1 Random Forest

Random Forest is a machine-learning algorithm that can be used for both classification and regression problems. It is based on the concept of ensemble learning, which combines multiple classifiers to solve a complex problem and improve the performance of the model. The algorithm creates multiple decision trees and takes the prediction from each tree. The final output is the prediction that receives the most votes. As the number of trees increases, the accuracy of the model also increases and it prevents an overfitting problem. Random Forest algorithm can also handle high dimensional data and categorical variables well. It is also relatively easy to interpret, making it a popular algorithm for various applications.

ALGORITHM

- **Step-1: Select random K data points from the training set.**
- **Step-2: Build the decision trees associated with the selected data points (Subsets).**
- **Step-3: Choose the number N for the decision trees that you want to build.**
- **Step-4: Repeat Step 1 and 2.**
- **Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.**

3.5.2 K-Nearest Neighbors

K-Nearest Neighbor (KNN) is a type of lazy learning algorithm that classifies data based on the similarity of the data points to its neighbors. It is a simple and fundamental classification method that is often used as a starting point for classification studies. The algorithm works by determining the k closest data points (neighbors) to a given point and then classifying the point based on the majority class of the k neighbors. KNN is known for its simplicity, but can be computationally expensive for large data sets.
The K-NN working can be explained based on the below algorithm:

- **Step-1: Select the number K of the neighbors.**
- **Step-2: Calculate the Euclidean distance of K number of neighbors.**
- **Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.**
- **Step-4: Among these k neighbors, count the number of the data points in each category.**
- **Step-5: Assign the new data points to that category for which the neighbor is maximum.**
- **Step-6: Our model is ready.**

3.5.3 Support Vector Machine

Support Vector Machine (SVM) is a type of supervised learning algorithm that can be used for both classification and regression problems. It is particularly useful for problems with a large number of features and/or a complex, non-linear decision boundary. The SVM algorithm seeks to find a decision boundary (a line or hyperplane) that maximally separates the different classes in the training data. This boundary is chosen so as to maximize the margin or the distance between the boundary and the closest data points from each class (called support vectors). During the classification, when new unseen data is given as input, it predicts to which class it belongs by checking on which side of the hyperplane it falls. SVM is known for its ability to handle high-dimensional data and its robustness to overfitting.

3.6 MODEL EVALUATION

Model evaluation is the process of measuring the performance of a machine learning model on a set of data. The goal of model evaluation is to determine how well the model can generalize to new, unseen data. There are several metrics that can be used to evaluate the performance of a model, depending on the type of problem and the specific goals of the model.

Here are some common metrics used for model evaluation:

- Confusion matrix: This is a table that summarizes the actual and predicted class labels for a set of data. It is useful for evaluating the performance of a classification model in each class.
- Accuracy: This measures the proportion of correctly classified instances over the total number of instances. It is a commonly used metric, but it can be misleading in some cases, especially when the classes are imbalanced.
- Precision: This measures the proportion of true positives (TP) over the total number of predicted positives (TP+FP). Precision is useful when we want to minimize the number of false positives.

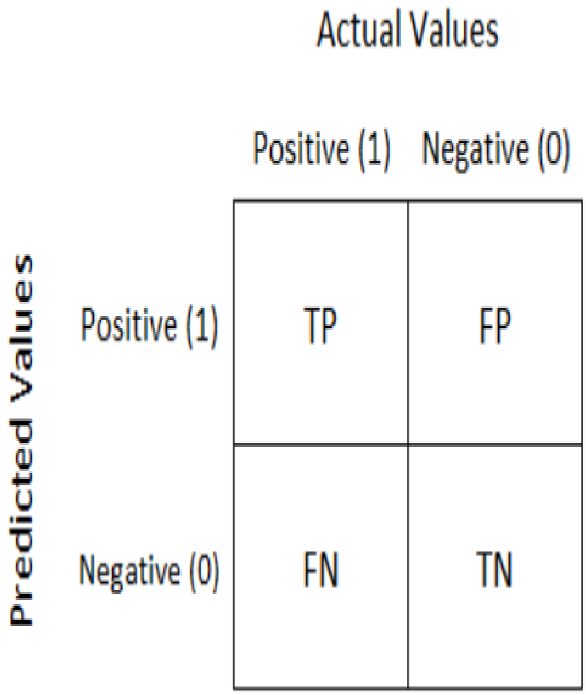

Figure 3.1: cofusion matrix

- **Recall: This measures the proportion of true positives (TP) over the total number of actual positives (TP+FN). The recall is useful when we want to minimize the number of false negatives.**
- **F1 score: This is the harmonic mean of precision and recall. It is a useful metric when we want to balance precision and recall.**

Chapter 4

DATA ANALYSIS

4.1 DATA EXPLORATION

Data collection is the process of gathering and compiling information from various sources to be used for analysis. In this case, the data used for employee attrition analysis was obtained from the Kaggle website and contains 1470 records and 35 attributes. The dataset was explored and identified patterns/relationships that come relevant to the problem were. This involved visualizing the data using techniques such as histograms. The key target attribute is "Attrition." The main features included in the data set are various demographic and employment details such as age, business travel, salary, department, education, job role, and years of experience. To make the classification algorithm more effective, categorical values are converted to numeric values. For example, the 'Business Travel' attribute containing the values 'Travel-Rarely', 'Travel-Frequently', and 'Non-Travel' is converted to 1, 2, and 3 respectively.

4.1.1 Finding Correlation Between Variables

A correlation matrix is a table that shows the correlation coefficients between pairs of variables. Correlation coefficients are a measure of the strength and direction of the linear relationship between two variables. The figure depicts the correlation matrix of the dataset features. The cell colors vary from black to cream color. white cells represent no cor-

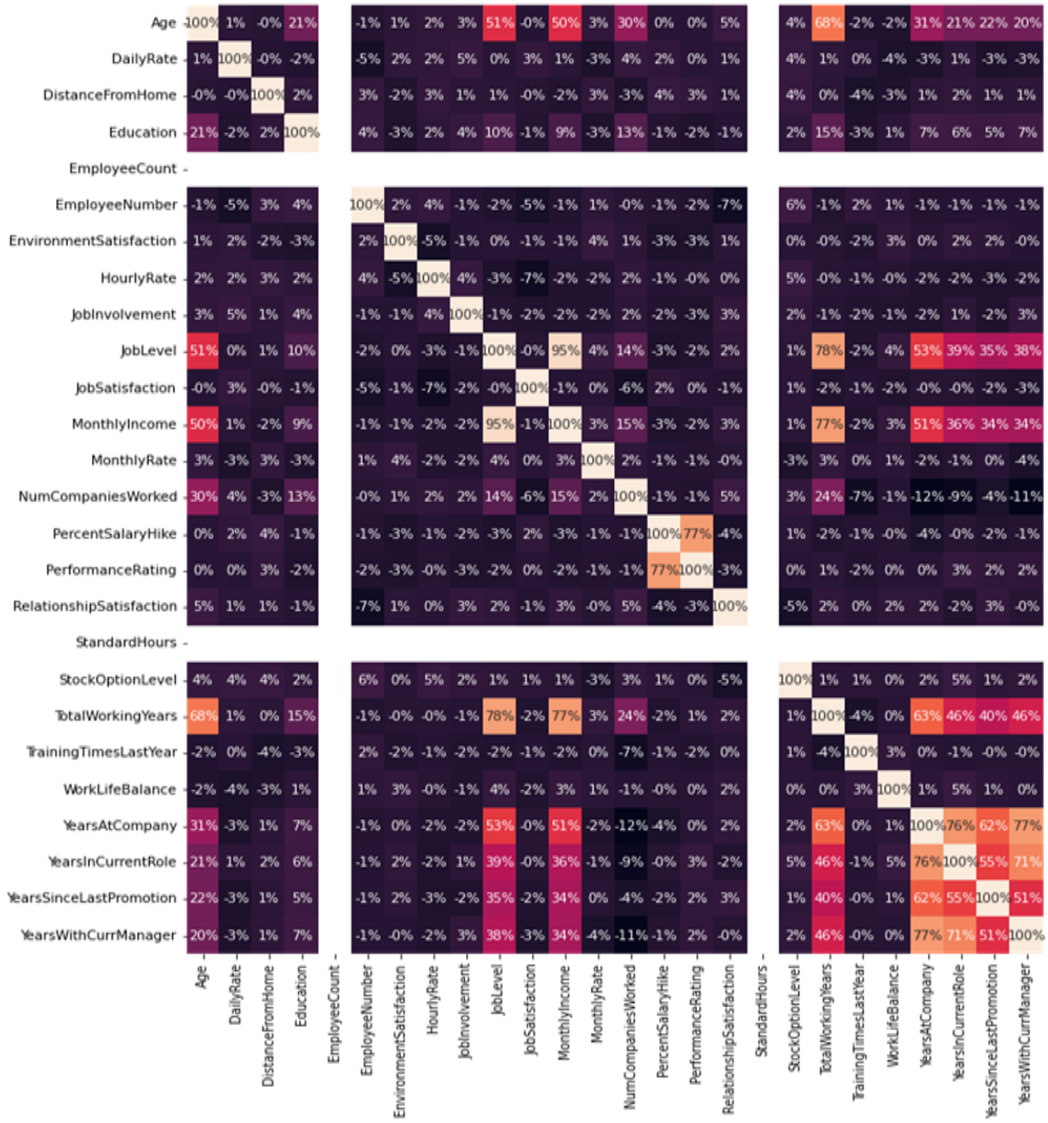

Figure 4.1: Correlation Matrix

relation, while red variations represent a high correlation. Black variations represent a negative correlation among dataset features. Analysis of correlation results :

- **Monthly income is highly correlated with Job level.**
- **The job level is highly correlated with total working hours.**
- **Monthly income is highly correlated with total working hours.**
- **Age is also positively correlated with Total working hours.**
- **Marital status and stock options level are negatively correlated.**

4.2 DATA VISUALISATION

Here are some visualizations to obtain some basic ideas from the dataset.

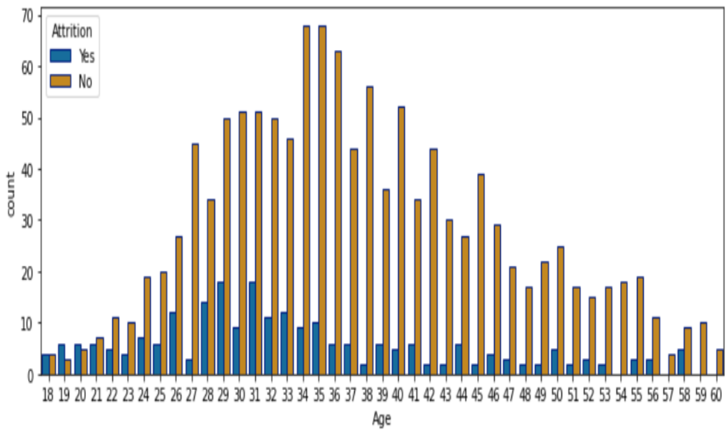

Figure 4.2: Age vs Attrition

Figure 4.2 shows that workers between the ages of 30 and 40 are more likely to leave their jobs. The threat to companies is likewise equally imposed by workers between the ages of 20 and 30.

Figure 4.3 depicts that there is a positive correlation between employee retention and job satisfaction.

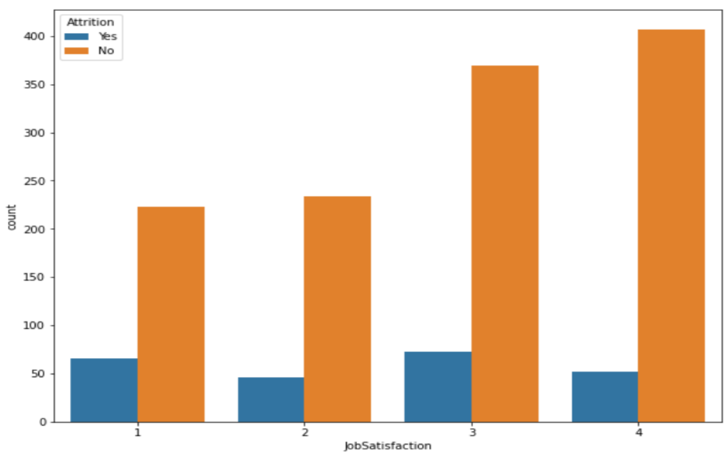

Figure 4.3: Jobsatisfaction vs Attrition

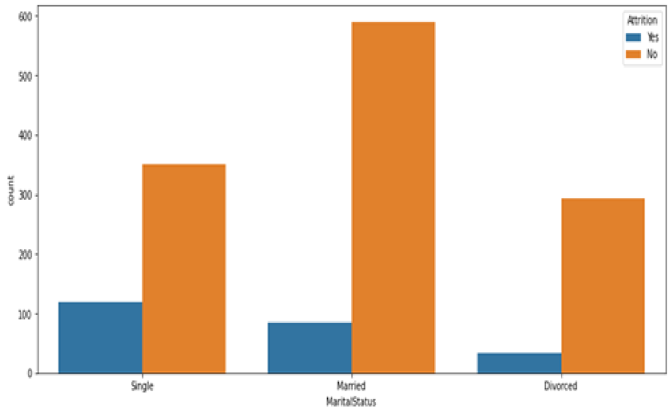

Figure 4.4: Marital Status vs Attrition

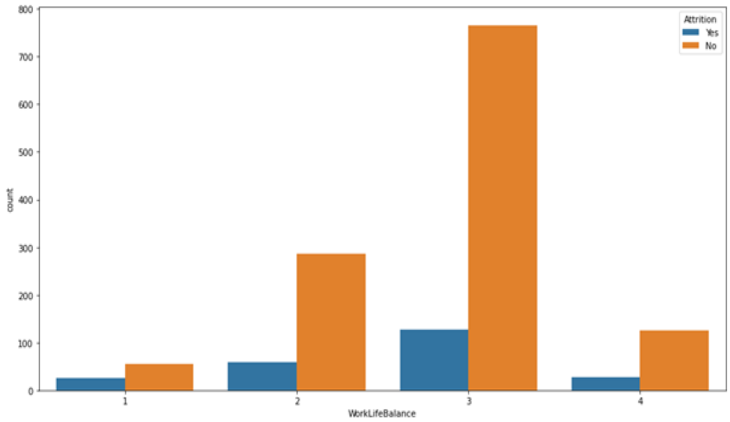

Figure 4.5: Worklifebalance vs Attrition

Figure 4.4 shows that single employees are more likely to quit the company and married employees stay back at the company.
Figure 4.5 depicts that the chance for the employee not to leave increases with increasing work-life balance and the chance to leave is very minimal with a high work-life balance.

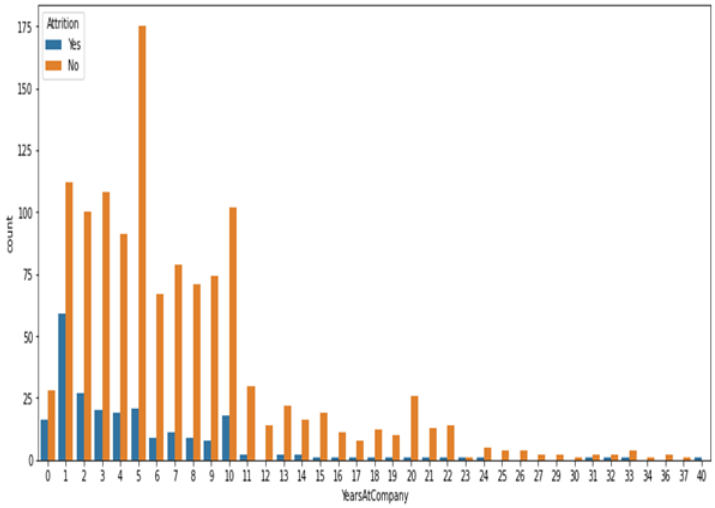

Figure 4.6: Years at Company

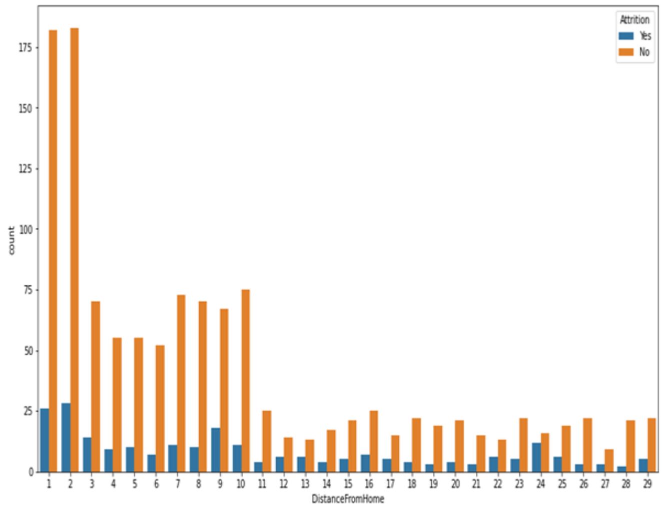

Figure 4.7: Distance From Home vs Attrition

From figure 4.6 it is clear that It is very less likely for the employee to leave after 10 years of employment at the company. Also, it is clear that people with experience are more likely to leave the company.

Figure 4.7 shows that Employee attrition possibly increases with the distance from their home. Employees within a 2 Km distance are the least likely to leave the company .

4.3 ACCURACY COMPARISON

In summary, the HR Employee-Attrition dataset contains various attributes such as department, gender, overtime, and business travel. Machine learning algorithms were used to build a model to predict whether employees will leave the organization or not. The predicted values were compared with test values to calculate the accuracy of each algorithm, and the table shows that Random Forest gives the highest accuracy while the KNN algorithm gives the lowest accuracy for the same dataset.

4.3.1 Random Forest

Classification Report:

	precision	recall	f1-score	support
0	0.86	1.00	0.93	310
1	0.90	0.16	0.26	58
accuracy			0.86	368
macro avg	0.88	0.58	0.59	368
weighted avg	0.87	0.86	0.82	368

Figure 4.8: Classification Report

Confusion Matrix:

```
[[309   1]
 [ 49   9]]
```

Figure 4.9: Confusion Matrix

Accuracy Score:0.8641

4.3.2 K-Nearest Neighbor

Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	34
1	0.20	0.15	0.17	60
2	0.74	0.86	0.80	274
accuracy			0.67	368
macro avg	0.32	0.34	0.32	368
weighted avg	0.59	0.67	0.62	368

Figure 4.10: Classification Report

Confusion Matrix:

```
[[  0   3  31]
 [  0   9  51]
 [  5  32 237]]
```

Figure 4.11: Confusion Matrix

Accuracy Score:0.6785

4.3.3 Support Vector Machine

Classification Report:

```
              precision    recall  f1-score   support

           0       0.87      1.00      0.93       245
           1       0.92      0.22      0.36        49

    accuracy                           0.86       294
   macro avg       0.89      0.61      0.64       294
weighted avg       0.87      0.87      0.83       294
```

Figure 4.12: Classification Report

Confusion Matrix:

```
[[244   1]
 [ 38  11]]
```

Figure 4.13: Confusion Matrix

Accuracy Score:0.8623

4.4 MODEL COMPARISON

MODEL	ACCURACY	BENEFITS
Random forest	86.41%	capable of handling large datasets, accuracy of the model is high.
SVM	86.23%	Memory efficient
KNN	66.85%	No training period, easy to implement

Figure 4.14: Model Comparison

The predicted values were compared with test values to calculate the accuracy of each algorithm, and the table shows that Random Forest gives the highest accuracy with 86.41% while the KNN algorithm gives the lowest accuracy with 66.85% for the same dataset.

4.5 FEATURE IMPORTANCE

The Random Forest is the most effective machine learning algorithm for the data set, and we used this ML technique to identify the key factors causing employee churn. From figure, it is clear that the key components of the proposed algorithm are Monthly Income, Age, Day Rate, Total Working Years, and Monthly Rate.

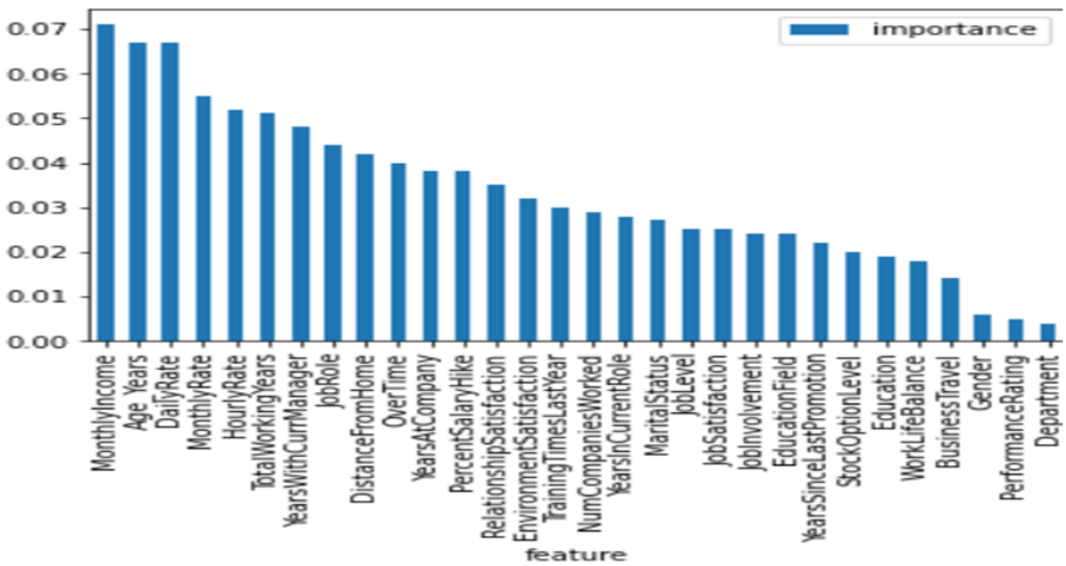

Figure 4.15: Feature Importance Estimated by Random Forest

Chapter 5

CONCLUSION

The employee attrition prediction by using the three machine learning techniques Random Forest, SVM, and KNN were applied in comparison in this study. The applied machine learning techniques achieved accuracy scores of 86.23% by the SVM technique, 66.85% by the KNN technique, and 86.41% by the Random forest technique. Hence Random Forest is an efficient machine learning algorithm for detecting the most important features responsible for employee attrition (turnover) rate in a given data set. Understanding the internal and external factors that contribute to employee attrition can help organizations take appropriate measures to minimize the rate of attrition. According to the proposed Random Forest algorithm, the most important features for employee attrition are Monthly Income, Age, Daily Rate, Total Working Years, and Monthly Rate. The study also highlighted the importance of using appropriate data sets, feature selection, and fine-tuning the parameters of the algorithm, to have better performance of the algorithm. Overall, this study suggests that machine learning can be a valuable tool for organizations looking to improve their retention rates and maintain a productive workforce.

REFERENCES

[1] Madara PRATT, Mohcine BOUDHENE, Sarma CAKULA. "Employee Attrition Estimation Using Random Forest Algorithm ", Baltic J. Modern Computing, , Vol. 9 , No. 1, 49-66Vol, May 2021.

[2] Adarsh Patel, Nidhi Pardeshi, Shreya Patil, Sayali Sutar, Rajashri Sadafule, Suhasini Bhat, ") Employee Attrition Predictive Model Using Machine Learning " International Research Journal of Engineering and Technology (IRJET) Volume: 07 Issue: 05, 2020.

[3] Zhang Heng, Xu Lexi, Cheng Xinzhou, Chao Kun, Zahoa Xueqing, "Analysis and Prediction of Employee Turnover Characteristics based on Machine Learning "The 18th International Symposium on Communications and Information Technologies (ISCIT),2018

[4] Pavansubhash,"https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset.

[5] Rupa Chatterjee Das, A.D.S."Conceptualizing the Importance of HR Analytics in Attrition Reduction."Int. Res. J. Adv. Sci. Hub 2, 40–48,2020

[6] Usha P.M.; Balaji N,"Analysing employee attrition using machine learning. "Karpagam J. Comput. Sci. 13, 277–282,2017

[7] Alduayj, S.S.; Rajpoot, K."Predicting Employee Attrition using Machine Learning."International Conference on Innovations in Information Technology (IIT), Al Ain, United Arab Emirates, 18–19 November 2018.