ML&VL Assignment

2022-23

Group Name:

The Random Forest Rascals

Group Members:

Shafieha, Romina

Yao, Keyan

Jadhav, Siddhesh Suresh

Budhwal, Archisha

Bandara, Sandun Savinda

Table of Contents

1. Executive Summary

This report provides an overview of the insights and findings from HR attrition analytics The objective is to assist XYZ Company in tackling the issue of staff attrition and lowering attrition rates. The XYZ company can foster a positive work environment and achieve long-term success by comprehending the elements that affect attrition and putting focused initiatives into place.

The results of the analysis showed that some variables significantly affect attrition. These factors can aid in spotting possible risks and formulating strategies to reduce attrition. Machine learning techniques for predicting employee attrition, such as the random forest algorithm and neural network algorithm, have shown encouraging results.

2. Introduction

The success of any organization relies heavily on its workforce. Employee attrition is a key issue for HR managers since it can negatively impact productivity and overall organisational success.

In this business analytics report, we aim to analyse a comprehensive HR dataset to gain insights into the drivers of employee attrition within XYZ company. By leveraging machine learning techniques, we strive to identify patterns and underlying factors that may contribute to attrition. The ultimate goal is to provide actionable recommendations and strategies for mitigating attrition.

Constant attrition in a company can have several negative effects, including:

1. Increased costs: Increased expenditures for recruiting, training, and lost productivity can be caused by high attrition rates. The cycle of recruiting and training that might happen as a result of constant attrition can be expensive and time-consuming.

2. Loss of institutional knowledge: When experienced staff members depart from a company, they carry important information and experience with them. Losing institutional knowledge as a result of constant attrition could be harmful to the long-term success of the business.

3. Decreased productivity: High attrition rates can lead to a decrease in productivity, as remaining employees may need to take on additional responsibilities or spend time training new hires. This can lead to increased stress and burnout among employees, which can further decrease productivity.

4. Low morale: Constant attrition can create a sense of instability and uncertainty among employees, which can lead to low morale and decreased job satisfaction. This can result in decreased engagement and motivation, further reducing productivity.

5. Negative impact on company culture: High attrition rates can lead to a negative company culture, where employees feel undervalued and unsupported. This can make it difficult to attract and retain top talent, further perpetuating the cycle of attrition.

The dataset provided for this analysis contains a rich collection of employee-related variables, such as demographics, job characteristics and performance metrics. By exploring this dataset and applying suitable analytical methods, we can uncover hidden relationships and factors that may significantly influence attrition rates.

Provided data consists of 6 subsets. All the subsets are for 4410 employees.

1- General data: This dataset consists of 24 variables (including employee ID. Variables include the attributes of each employee and whether they have left the company I the previous year or not.
2- Employee survey data: The file has 4 variables, one which is employee ID and shows how satisficed each employee is.
3- Manager survey data: This file consists of 3 columns (including employee ID) and shows how satisfied managers are with the employees.
4- In time: Shows the daily attendance of the employees
5- Out time: Shows the daily departure of the employees
6- Data dictionary: includes brief explanation of each attribute

3. Data Collection and Pre-processing

In the context of analysing data provided by XYZ company, data preparation and manipulation play a vital role. These steps include combining data frames, dealing with missing values,

figuring out how many hours were worked, and changing data types. These procedures allow us to make sure the gathered data is appropriately arranged and prepared for analysis.

Merging Data: We first combined the “general_data”, “employee_survey_data”, and “manager_survey_data” data sets into a single merged data frame using the shared key "EmployeeID" in this process. The data can now be viewed uniformly for further analysis thanks to this consolidation.

Handling Null Values: To manage null values, the "in_time" and "out_time" columns are searched for any NA values. For these missing values, we use the usual string representation "2999-12-28 23:59:59" to handle them. By adding the missing values, we guarantee that the dataset is complete and prepared for further analysis, avoiding potential problems brought on by inaccurate data.

Calculating Worked Hours: We precisely calculate the amount of time worked on various days by deducting the in-time from the out-time for each employee. This data is useful for assessing productivity and detecting trends in working hours.

Integration and Analysis: Calculating employees working hours and days are useful gauges for examining attendance and productivity. All the tables are combined in a single file and then converted into a file, ready to run on machine learning algorithms.

4. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a vital stage in understanding the dataset and providing insights into the possible characteristics of employee attrition within the company. By observing and analysing different subsets of data, certain patterns, trends, and associations of the dataset are derived within before any complex modelling.

In addressing the main business issue of attrition, it was imperative to understand the attrition percentage from the total population, which also indicated the imbalanced nature of the dataset before any further diagnostics were performed. This was taken as the starting point and attrition related subset was further analysed separately for identification of central characteristics and trends.

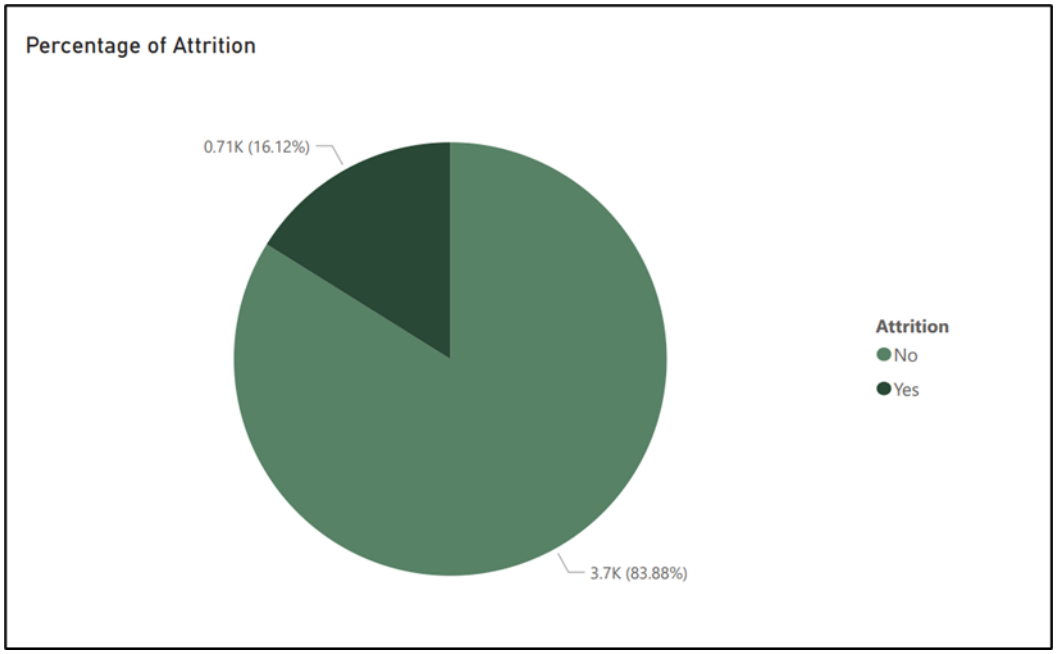


Figure 01: Attrition percentage of the total population.

When plotting certain demographic variables such as age, gender, education level and in observation of these distributions, certain patterns emerged, which acted as a good foundation for further drill down of the data. Below is a snapshot (Figure 02) of the key variable of the dataset and some high-level findings.

- Attrition was higher among single males within the age category of 26 year to 34 years.
- Employees who have completed their secondary education and have worked less than 2 years are the most susceptible to attrition.
- Attrition is recoded higher among employees who had worked less than 2 years with the same manager and not received a promotion in at least 5 years.
- Significantly higher attrition rates are observed in the R&D department, which paves way for further investigation.

Apart from the above several other variables, deemed important for the study were plotted in figure 03 to understand the distribution and to observe possible outliers. This also aids in boundary visualisation for later model result interpretation.

Figure 02: Gives a synopsis of the key attributes in the considered subset of attrition.

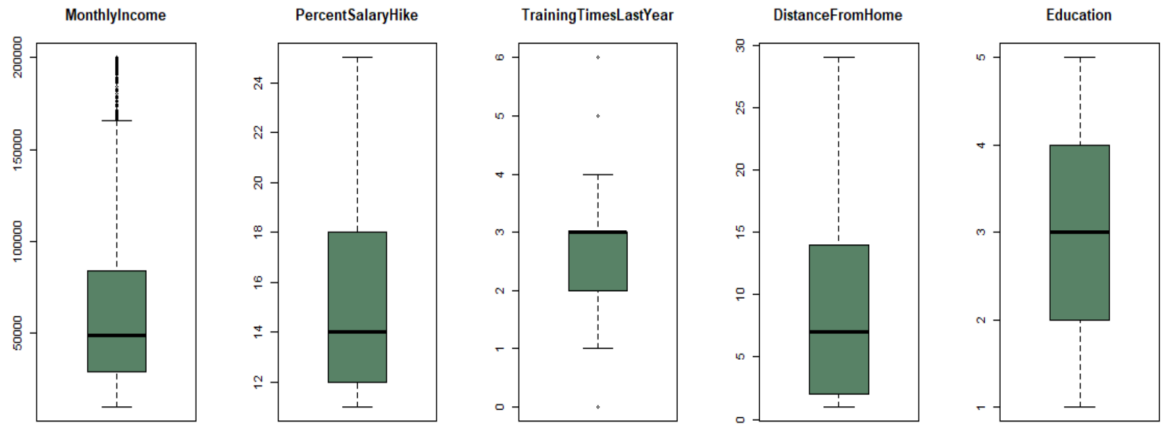


Figure03: Distribution and spread of other key variables.

The EDA has aided in familiarising the structure of the data and the key variables, this understanding is crucial for the interpretation of results of the model and algorithms, where any irregularities and outliers are easily identified in the correct business context. Further EDA also plays key role in feature selection, balancing of data, to understand the limitation and biases that can impact the final interpretation. In general, EDA acts as a good yardstick in

understanding the business problems, interpretation of modelled results and in arriving at pragmatic recommendations to the higher attrition rates of the company.

5. Feature Engineering

We did data balancing and feature selection on the dataset to improve the performance.

5.1 Data Balancing

One crucial aspect of the dataset used in this analysis is its inherent class imbalance. The goal of the machine learning models on this dataset is to identify employees that have left the company and use it to predict individuals who may be inclined to leave. However, the provided dataset exhibits a significant disparity in the distribution of the target variable, with most instances representing employees who have not left the company. This class imbalance can potentially affect the performance of the predictive models.

To address this issue, we employed data balancing techniques to mitigate the impact of class imbalance during the model training process. By balancing the training dataset, we aim to create a more representative and unbiased learning environment for the models. We used oversampling to adjust the class distribution. Oversampling involves duplicating minority class instances or generating synthetic samples.

The generated models will later be evaluated on the original test dataset.

Figure 04 clarifies the balancing impacts on attrition.

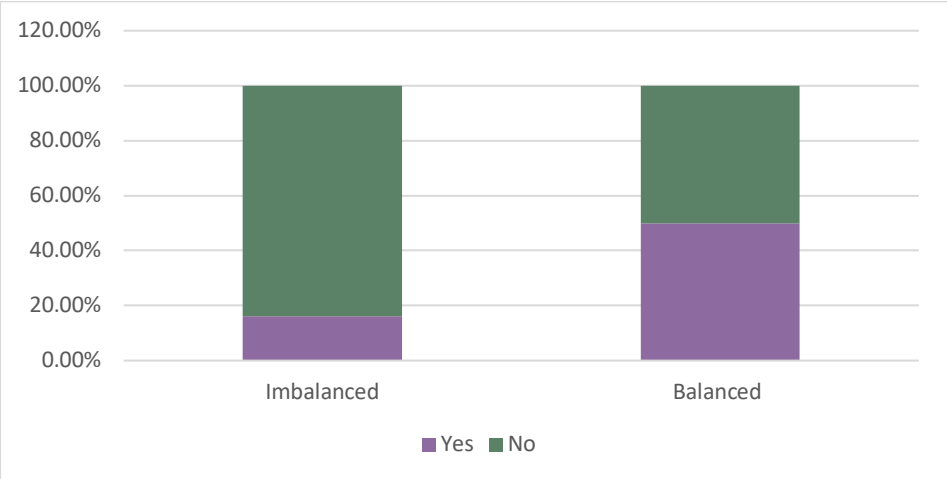


Figure 04 - Percentages for attrition before and after balancing

5.2 Feature Selection

In addition to balancing the dataset, we used feature selection to further improve the performance of the models. Feature selection involves identifying the most relevant and informative features from the available dataset, which can enhance model performance.

We used correlation matrix and the logistic regression for this purpose. The unbalanced data was used for feature selection.

The selected variables are determined in Table 01 and Figure 05.

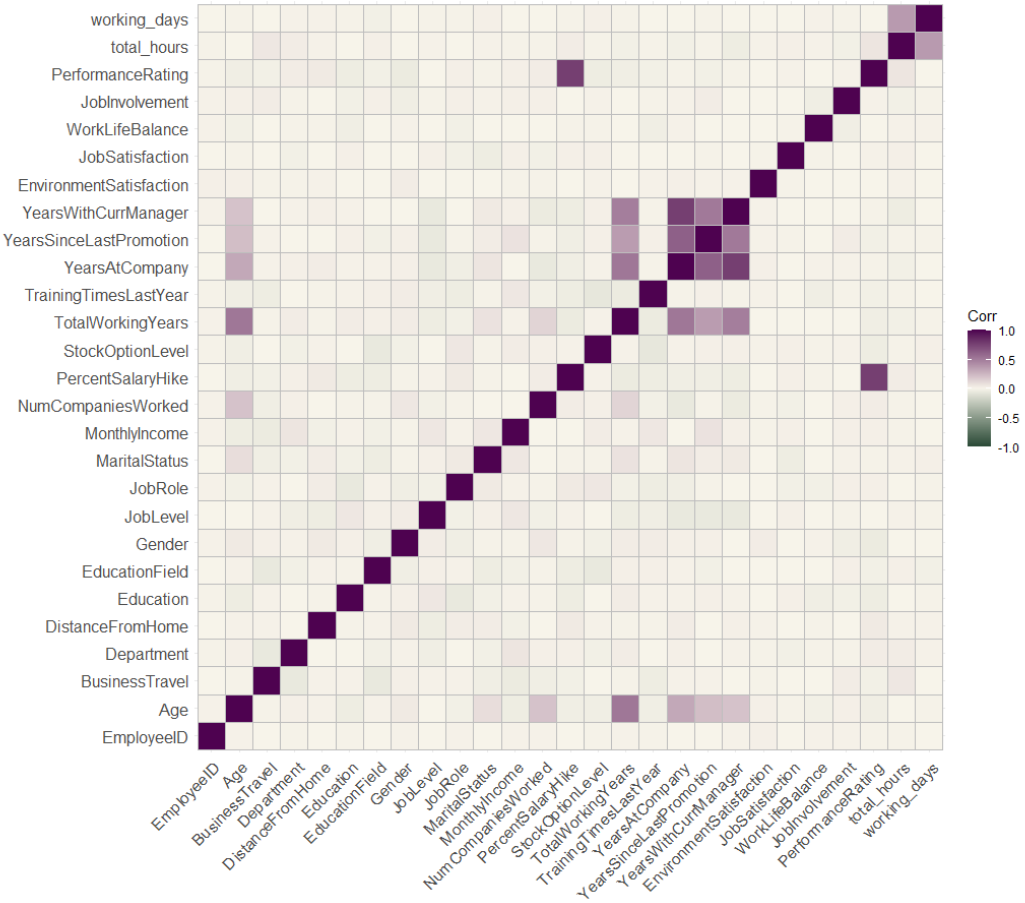


Figure 05- Correlation plot on the unbalanced data

To ensure reliability, different combinations of variables with high correlation were tested to find the best combination.

The final variables are as follows:

- Total hours
- Total working years
- Marital status

- Age
- Number of years with the current manager
- Years at company
- Business travel

Table 01 - Logistic regression on the combined data frame

	Coefficients	Estimate	Std. Error	z value	Pr(>|z|)	
	(Intercept)	3.73E+00	2.48E+00	1.508	0.13147	
	Employee ID	-3.21E-08	3.51E-05	-0.001	0.99927	
•	Age	-4.05E-02	6.36E-03	-6.365	1.95E-10	***
•	Business Travel	1.59E-03	1.98E-04	8.016	1.09E-15	***
	Department	4.38E-05	6.37E-05	0.688	0.49172	
	Distance From Home	-1.54E-03	5.61E-03	-0.275	0.78347	
	Education	-4.42E-02	4.41E-02	-1.003	0.31577	
	Education Field	1.28E-04	1.46E-04	0.874	0.382	
	Employee Count	NA	NA	NA	NA	
	Gender	-5.48E-04	4.57E-04	-1.199	0.23061	
	Job Level	-2.77E-02	4.09E-02	-0.678	0.49746	
	Job Role	-9.40E-05	1.01E-04	-0.933	0.3508	
•	Marital Status	-6.02E-03	6.46E-04	-9.32	2.00E-16	***
	Monthly Income	-1.41E-06	9.90E-07	-1.421	0.15528	
	Number of Companies Worked	5.22E-02	1.64E-02	3.184	0.00145	**
	Over18	NA	NA	NA	NA	
	Percent Salary Hike	2.00E-02	1.94E-02	1.03	0.3028	
	Standard Hours	NA	NA	NA	NA	
	Stock Option Level	-3.56E-02	5.33E-02	-0.669	0.50363	
•	Total Working Years	-7.80E-03	2.58E-03	-3.027	0.00247	**
	Training Times Last Year	-1.44E-01	3.61E-02	-3.992	6.56E-05	***
•	Years At Company	-2.26E-02	1.62E-02	-1.397	0.16246	
	Years Since Last Promotion	1.36E-01	2.09E-02	6.495	8.30E-11	***
•	Years With Current Manager	-1.32E-01	2.30E-02	-5.761	8.38E-09	***
	Environment Satisfaction	-1.914e-02	9.15E-03	-2.092	0.03647	*
	Job Satisfaction	-1.63E-02	1.07E-02	-1.519	0.12868	
	Work Life Balance	-1.90E-03	9.52E-03	-0.199	0.84201	
	Job Involvement	-6.49E-02	6.21E-02	-1.047	0.29528	
	Performance Rating	-1.48E-01	1.95E-01	-0.758	0.44837	
•	Total hours	1.72E-03	1.39E-04	12.397	2.00E-16	***
	Working days	-1.63E-02	8.75E-03	-1.866	0.06208	

The signs of the correlations indicate the direction of the relationship between the features and the likelihood of attrition.

1. Total Hours (positive correlation): This suggests that employees who work longer hours may be more likely to experience attrition.

2. Total Working Years (negative correlation): This suggests that employees who have been working for a longer period may be less likely to experience attrition.

3. Marital Status (negative correlation): This suggests that employees who are married may be less likely to experience attrition compared to those who are not married.

4. Age (negative correlation): This suggests that younger employees may be more likely to experience attrition compared to older employees.

5. Years With Current Manager (negative correlation): This suggests that employees who have been working with their current manager for a shorter period may be more likely to experience attrition. And the people who are working with their current manager for longer period

6. Years At Company (negative correlation): This suggests that employees who have been working at the company for a shorter period may be more likely to experience attrition.

7. Business Travel (positive correlation): This suggests that employees who travel frequently for business may be more likely to experience attrition.

6. Model Development & Evaluation

In general, the key elements in machine learning to use in performance evaluation are:

1. True Positives (TP): The number of employees who are correctly classified as having left the company.

2. False Positives (FP): The number of employees who are incorrectly classified as having left the company (i.e., the model predicted that they left, but they did not).

3. True Negatives (TN): The number of employees who are correctly classified as having not left the company.

4. False Negatives (FN): The number of employees who are incorrectly classified as having not left the company (i.e., the model predicted that they did not leave, but they did).

From these four elements, various metrics can be derived that provide insights into the performance of the classification model. Some commonly used metrics include:

1. Accuracy: The proportion of correct predictions out of all predictions.
2. Precision: The proportion of true positives out of all positive predictions.
3. Sensitivity: The proportion of true positives out of all actual positives.
4. Specificity: The proportion of true negatives out of all actual negatives.
5. Kappa: A statistical measure that is used to evaluate how well two classifiers agree with each other when assigning labels to a set of data.

The interpretation and importance of each metric will depend on the specific objectives and requirements of the analysis.

As it is the target of this project is identified attrition, it is essential to reduce number of FN. For that purpose, Accuracy and Sensitivity are selected as the best measures for evaluating the models.

To be able to train and further evaluate the model, we have divided the dataset into two parts. First part, training data, will be trained to teach the model any underlying patterns. Then, the model will be run on the second part, test dataset to evaluate the performance of the model.

We used a repeated k-fold cross-validation to mitigate impacts of random partitioning

6.1 KNN

We deployed KNN algorithm as one of our machine learning models to learn patterns and relationships from the dataset and uncover insight. KNN is a supervised algorithm and a popular technique used for classification.

During the training phase, KNN stores the feature vectors and corresponding class labels of the training data. No explicit model is built during this phase.

The value of K is a hyperparameter that determines the number of neighbours to consider when making predictions. It is important to select an appropriate value for K, as it can greatly influence the algorithm's performance. A smaller value of K can lead to overfitting, while a

larger value may result in oversimplification. We do not specify the k in this report but rather evaluate the performance directly.

The KNN model was trained on the balanced training data with feature selection and evaluated on the original test data. The optimal results are demonstrated in Figure 06.

Based on our measures of interest, one of the folds is selected as the optimal result for the KNN model.

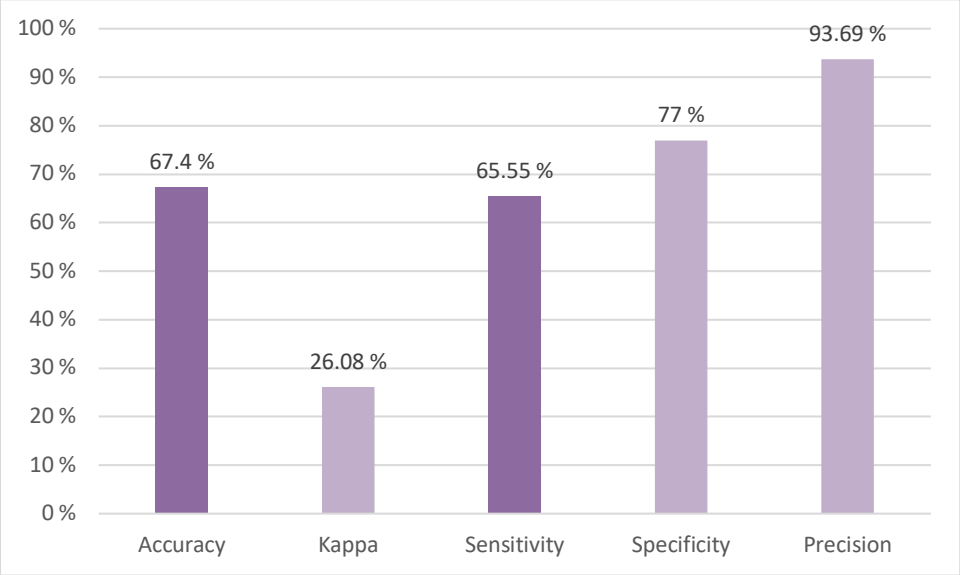


Figure 06 – Optimum results for KNN

6.2 Random Forest

One of the key benefits of the random forest algorithm is its capacity to improve forecast accuracy. The method uses an ensemble of decision trees to aggregate predictions from several trees to generate a final forecast. The model's overall accuracy is much increased by using an ensemble technique, which enables us to make more precise predictions about potential employee attrition scenarios.

Additional benefits of the random forest method are scalability and efficiency. Large HR datasets with a variety of numerical and categorical variables can be handled by it without the requirement for laborious pre-processing. The random forest algorithm is useful due to its scalability.

Looking at the random forest model's performance metrics further demonstrates its capacity to forecast attrition. The model was successful in its predictions in about 89.41% of the dataset's circumstances, earning a high accuracy score of 89.41%. By examining its performance metrics, the random forest model's capacity to forecast staff attrition is further demonstrated. The model accurately predicted 89.41% of the instances in the dataset, earning an accuracy score of 89.41%.

The kappa value of 67.17% demonstrates high agreement between the model's predictions and the projected results, beyond what could be attributable to chance alone. The sensitivity score of 98.11% is a crucial factor in attrition prediction. It demonstrates how the model may successfully identify employees who are most likely to quit the company.

On the other hand, the specificity score of 61.59% shows that the model's ability to recognise employees who are likely to stay with the company is comparatively strong. Last but not least, the precision score of 89.09% shows how accurately the model predicts when a worker is going to retire.

The optimal results for random forest are demonstrated in Figure 07.

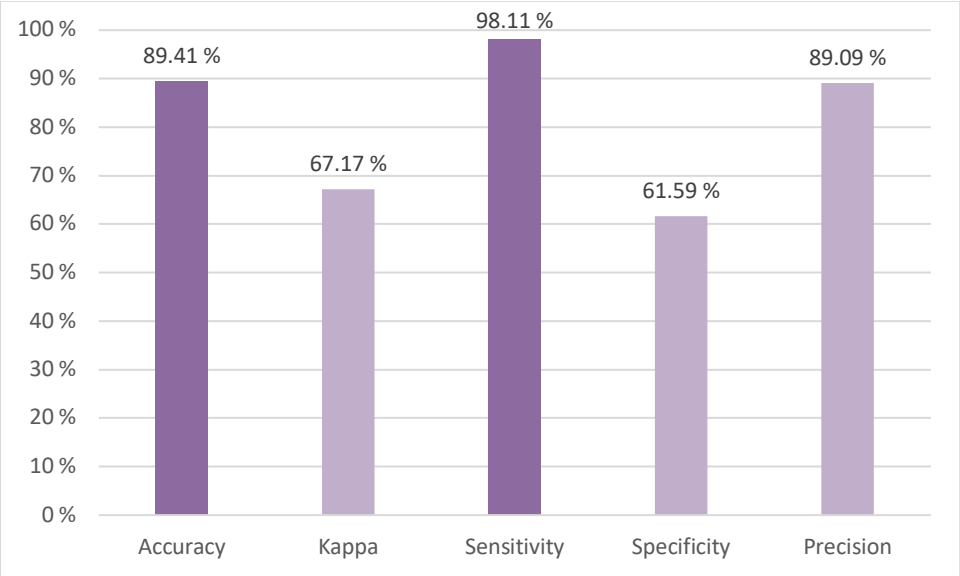


Figure 07 - Optimal results of Random Forest

6.3 Neural Network

With an accuracy of 81.62%, the neural network model developed to analyse the HR attrition dataset showed excellent findings. This shows that the model was accurate in its predictions in about 81.62% of the cases, offering a solid foundation for predicting staff attrition. Additionally, the kappa value of 22.1% indicates better agreement than could be attributed to pure random chance between the model's predictions and the predicted results.

The model can successfully identify employees who are most likely to leave the organisation, as evidenced by its sensitivity score of 86.82%. This high sensitivity is essential for attrition prediction since it guarantees that a sizable fraction of possible attrition instances are identified, lowering the chance of missing employees who could be on the verge of leaving.

In particular, the model's precision score of 92.06% deserves attention. This score shows that when the model predicted that an employee was likely to quit, it was generally 92.06% correct in its forecasts.

The optimal results for Neural Network are demonstrated in Figure 08.

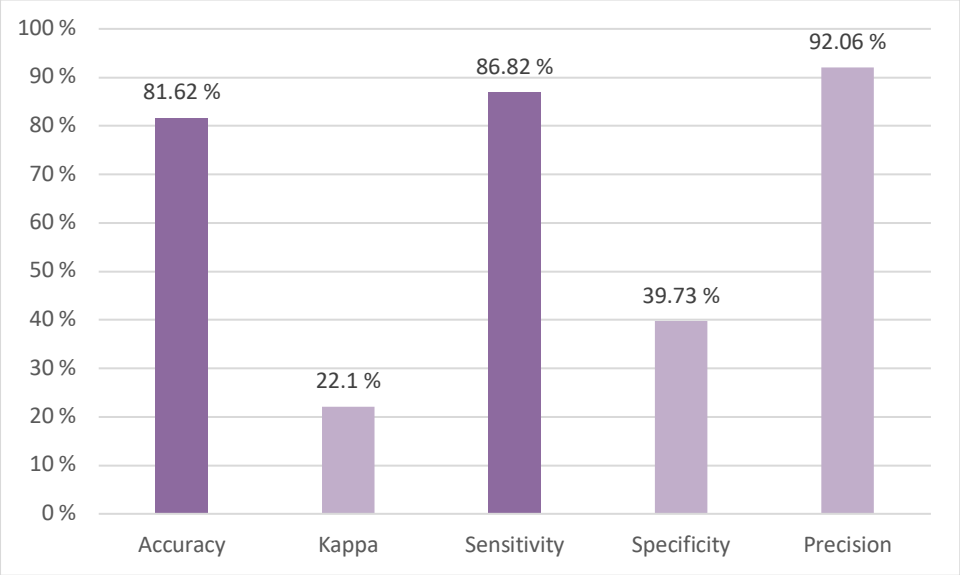


Figure 08 - Optimal results of Neural Network

7. Insights and Findings

We found that certain variables have a significant impact on attrition likelihood. These variables include the total hours worked, total working years, marital status, age, years with the current manager, years at the company, and business travel. Understanding how these factors relate to attrition can help us to identify potential risks and develop targeted strategies to address them.

The Neural network method and the Random Forest algorithm among the machine learning models we used both performed well in forecasting employee attrition. Additionally, the KNN model produced precise predictions by contrasting each employee's attributes with those of their closest neighbours.

We can better assess attrition risks by paying attention to variables including total hours worked, total years worked, marital status, age, years with the current management, years with the company, and business travel. we can take action to promote work-life balance for employees who put in more hours, for example, based on the characteristics that have been discovered. To keep workers with shorter tenures under their existing supervisors, they might also put an emphasis on mentorship programmes and professional development chances. Companies can also create retention plans that are particular to age groups and aid staff members who frequently travel for work.

8. Recommendations

Focusing on the total hours worked is one of the main recommendations. Employees can better manage their personal and professional life by introducing flexible work arrangements. The company can show their commitment to work-life integration by providing flexible scheduling or remote work choices, which will ultimately lead to greater employee happiness. Increasing the amount of required paid holidays also gives workers the much-needed rest and renewal they need, lowering the risk of burnout and improving general job satisfaction.

Another key issue to think about is total working years. Giving employees career development chances not only improves their knowledge and abilities but also shows the company's commitment to their professional development. Businesses invest in the potential of their employees and foster a sense of loyalty by providing both internal and external training programmes. Employee motivation is increased when exceptional performance is

acknowledged through employee recognition programmes, which also build a positive workplace environment and raise job satisfaction. Additionally, implementing mentoring programmes can facilitate new hires' transition into their positions and support the retention of their expertise within the company.

Another important issue to consider when thinking about attrition reduction is marital status. The implementation of family-friendly policies, such as flexible work schedules, parental leave, family insurance coverage, and childcare assistance, recognises and supports employees' obligations to their families. This not only fosters a healthy work atmosphere but also draws in and keeps workers who value work-life balance and obligations to their families.

Attrition rates can be considerably impacted by the effect of years spent under the current boss. Company should invest in training programmes for managers to make sure they can interact and communicate with their workforce in an effective manner. Managers can establish good relationships with their staff members and increase job satisfaction by fostering a positive and encouraging work environment. Setting up a system for quarterly employee feedback on their managers and the overall workplace atmosphere offers insightful information and enables rapid resolution of any issues or concerns

finally, dealing with the problem of business travel can have a big impact on attrition rates. Reduces the need for frequent travel by providing alternate work arrangements like teleconferencing or virtual meetings. Employees can maintain a better work-life balance as a result of spending less time away from their families and other responsibilities, which also lowers the costs connected with travel. Additionally, offering help to family members while an employee is abroad or providing travel reimbursement shows the organisation cares about their well-being and lessens any potential stress brought on by the need to travel. From a company standpoint, these actions improve general productivity, save travel-related costs, and increase employee happiness.

9. Limitation

The difference between causation and correlation is a significant constraint. Although there is a correlation between the discovered variables and attrition, it's crucial to remember that correlation does not imply causation.

In order to promote justice and inclusivity, we should be mindful of potential biases in the data or models and take action to reduce them. Even if the models offer insightful information, it is crucial to acknowledge the significance of human variables in comprehending attrition. Interpersonal dynamics, job happiness, and employee motivation are complicated issues that cannot be fully understood by data alone. To fully understand the underlying causes of attrition, the company should combine quantitative analysis with qualitative research techniques and employee feedback.

Finally, it's critical to recognise that both the corporate environment and personnel dynamics are dynamic. As a result, the company should periodically assess attrition rates and modify its tactics as necessary. The company can proactively address new elements that may have an impact on staff retention by keeping up with evolving trends and issues in employee attrition.

10. Conclusion

By focusing on variables such as total hours worked, total working years, marital status, age, years with the current manager, years at the company, and business travel, companies can gain a deeper understanding of attrition risks and develop targeted approaches.

Implementing the recommendations that we mentioned can significantly reduce attrition rates and create a more supportive and engaging work environment. By prioritizing flexible work arrangements, career development opportunities, family-friendly policies, effective management practices, and alternative work arrangements, businesses can foster employee loyalty, satisfaction, and retention. From a business perspective, these efforts contribute to higher productivity, reduced turnover costs, and the cultivation of a skilled and motivated workforce. Ultimately, investing in employees' well-being and growth is crucial for long-term success and organizational resilience.

By considering these limitations and continuously refining their approaches, businesses can leverage the insights to make informed decisions and develop effective attrition reduction strategies. Prioritizing employee well-being, growth, and creating a supportive work environment can contribute to higher productivity, reduced turnover costs, and the long-term success of the organization. Ultimately, investing in employees is a critical factor for organizational resilience and sustainable business growth.

11. References

Apokalypsepartyteam, 2022. *R-bloggers.* [Online]
Available at: https://www.r-bloggers.com/2022/02/beginners-guide-to-machine-learning-in-r-with-step-by-step-tutorial/
[Accessed (May 2023].

Boehmke, B., 2020. *http://github.io/.* [Online]
Available at: https://bradleyboehmke.github.io/HOML/
[Accessed 9 May 2023].

Choudhary, V., 2018. *kaggle.com.* [Online]
Available at: https://www.kaggle.com/datasets/vjchoudhary7/hr-analytics-case-study
[Accessed 9 May 2023].

Fred Nwanganga, M. C., 2020. *Practical Machine Learning in R.* s.l.:John Wiley & Sons.

12. Appendix.

Code Snippets

```r
# ********************************************************************************
# *** 1.1 Loading libraries ***************
# ********************************************************************************
R.Version()$version.string
install.packages("pacman")
library(pacman)
p_load(dplyr)
p_load(lubridate)
# p_load(ggplot)     # visualization----------????
p_load(ggplot2)    # visualization
p_load(caret)  #featurePlot,
p_load(lares)  #corelationplot
#install.packages("psych", dependencies = TRUE)
#install.packages(c("mnormt", "foreign", "minqa", "nloptr", "numDeriv", "psych"), dependencies = TRUE)
p_load(psych)  #describe,
p_load(readxl)
p_load(lattice)
p_load(ggcorrplot)  # ggcorrplot
p_load(neuralnet)   #neuralnet
p_load(ROSE)    #oversample Balancing
#install.packages('BiocManager')
p_load(randomForest) # Load the randomForest package
# p_load(randomForestExplainer) # random forest explainer
p_load(partykit) # plot random forest tree
p_load(car)
p_load(cluster)# Clustering
p_load(factoextra) # clustering algorithms & visualization
p_load(tidyr)
p_load(rpart)     # desicion tree Rpart
p_load(rpart.plot) ## plot Desicion tree for Rpart
p_load(irr)
p_load(GGally)
# remotes::install_github("cran/DMwR", force = TRUE)
#remotes::install_github("cran/ggpubr")

p_load(DMwR)


```

```r
# ********************************************************************************
# *** 1.3 Summary of Data ********************
# ********************************************************************************
head(general_data)
head(employee_survey_data)
head(in_time)
head(out_time)
head(manager_survey_data)

## finding NA values in Dataset
any(is.na(general_data))
any(is.na(employee_survey_data))

any(is.na(manager_survey_data))
any(is.na(in_time))
any(is.na(out_time))

## replacing NA values with character
if (any(is.na(employee_survey_data))) {
  # replace NA values with a specified string
  employee_survey_data[is.na(employee_survey_data)] <- as.character("")
}

if (any(is.na(general_data))) {
  # replace NA values with a specified string
  general_data[is.na(general_data)] <- as.character("")
}
```

```r
# ********************************************************************************
# ********** 1.4 Merging Dataset ***********
# ********************************************************************************

merge_data <- merge(general_data,employee_survey_data,by = "EmployeeID")
merge_data <- merge(merge_data,manager_survey_data, by = "EmployeeID")

#convert null to time "2999-12-28 23:59:59"

if (any(is.na(in_time))) {
  # replace NA values with a specified string
  in_time[is.na(in_time)] <- "2999-12-28 23:59:59"
}

if (any(is.na(out_time))) {
  # replace NA values with a specified string
  out_time[is.na(out_time)] <- "2999-12-28 23:59:59"
}

any(is.na(in_time))
any(is.na(out_time))

# str for structure of Data
str(in_time)

# Function to calculate worked hours
worked_hours_fn <- function(in_col, out_col) {
  as.numeric(difftime(ymd_hms(out_col), ymd_hms(in_col), units = "hours"))
}

# Get the date column names
date_cols <- colnames(in_time)[-1]

# Join the in_time and out_time data frames
combined_data <- inner_join(in_time, out_time, by = "X", suffix = c("_in", "_out"))

# Initialize the worked_hours data frame with the X column and the same number of rows as combined_data
worked_hours <- data.frame(id = combined_data$X, stringsAsFactors = FALSE)
```

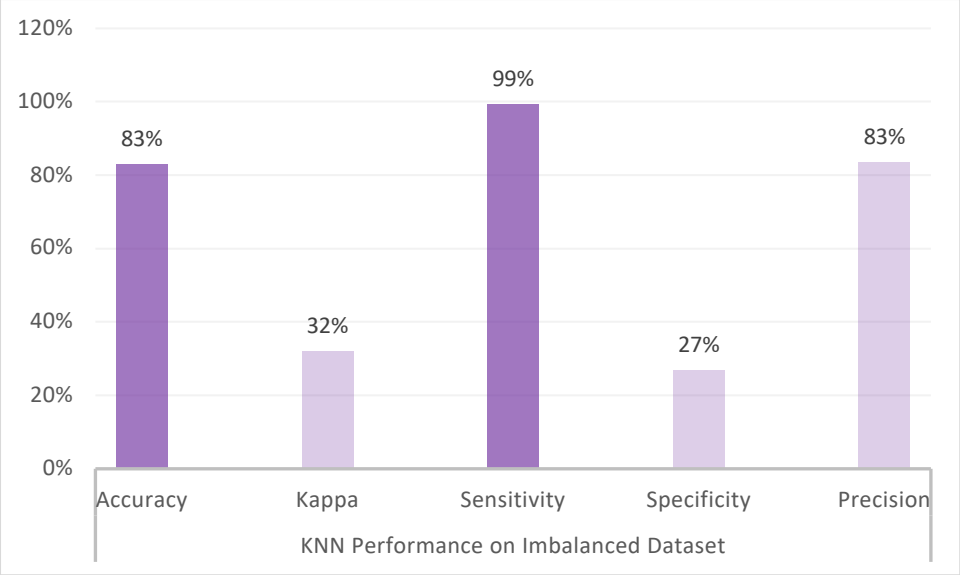
```r
# Loop through each date column  calculate worked hours
for (date_col in date_cols) {
  in_col_name <- paste0(date_col, "_in")
  out_col_name <- paste0(date_col, "_out")
  worked_hours_col_name <- paste0("worked_hours_", gsub("\\.", "_", date_col))

  worked_hours[[worked_hours_col_name]] <- worked_hours_fn(combined_data[[in_col_name]], combined_data[[out_col_name]])
}
{
  # calculate total hours and working days for each id
  total_hours <- apply(worked_hours[, 2:ncol(worked_hours)], 1, sum)
  working_days <- apply(worked_hours[, 2:ncol(worked_hours)] != 0, 1, sum, na.rm = TRUE)

  # combine results into a data frame
  worked_hours_new <- data.frame(id = worked_hours$id,
                                 total_hours = total_hours,
                                 working_days = working_days)
}

{
  #merge the new generated df into the main df by EmployeeID = id

  merge_data <- merge(merge_data,worked_hours_new, by.x = "EmployeeID", by.y="id")

  # move Attrition from middle to the end

  new_merge_data <- select(merge_data, -Attrition, everything())

  Data <- new_merge_data

  any(is.na(Data))
}

# ************************************************************************
```

```r
abstracted_training_data <- training_data
abstracted_training_data <- subset(abstracted_training_data, select = c("total_hours", "TotalWorkingYears","MaritalStatus", "Age",
                                                                        "YearsWithCurrManager", "YearsAtCompany",  "BusinessTravel" , "Y" ))

abstracted_balanced_training_data <- balanced_training
abstracted_balanced_training_data <- subset(abstracted_balanced_training_data, select = c("total_hours", "TotalWorkingYears","MaritalStatus",
                                    "YearsWithCurrManager", "YearsAtCompany",  "BusinessTravel" , "Y" ))
abstracted_testing_data <- testing_data
abstracted_testing_data <- subset(abstracted_testing_data, select = c("total_hours", "TotalWorkingYears","MaritalStatus", "Age",
                                                                      "YearsWithCurrManager", "YearsAtCompany",  "BusinessTravel" , "Y" ))

# abstracted_training_data$Y <- as.numeric(abstracted_training_data$Y) -1
# abstracted_testing_data$Y <- as.numeric(abstracted_testing_data$Y) -1
```

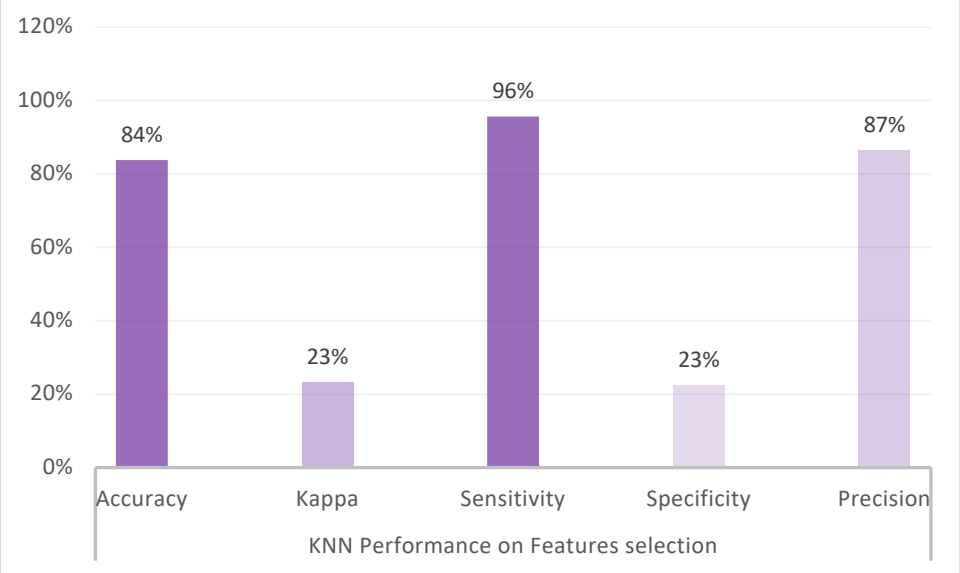
```r
{
  backup <- Data_for_abstract
  backup$Y <- ifelse(backup$Y == "Yes", 1, 0)
  Data_new <- backup
  Data_new <- subset(Data_new, select = c("total_hours", "TotalWorkingYears","MaritalStatus", "Age",
                                          "YearsWithCurrManager", "YearsAtCompany",  "BusinessTravel" , "Y" ))

  str(Data_new)
  # Data_new$Y <- ifelse(Data_new$Y == "Yes", 1, 0)
  # Data_new <- subset(Data_new, select = c("total_hours", "TotalWorkingYears","MaritalStatus", "Age",
  #           "YearsWithCurrManager", "YearsAtCompany",  "BusinessTravel" ,
  #          "TrainingTimesLastYear", "working_days", "EnvironmentSatisfaction",
  #          "YearsSinceLastPromotion", "PercentSalaryHike", "MonthlyIncome",
  #         "PerformanceRating", "NumCompaniesWorked", "Gender", "JobInvolvement",
  #        "Education", "JobRole", "EducationField", "Y"))


}
```

Exploring relationships among features – the correlation matrix

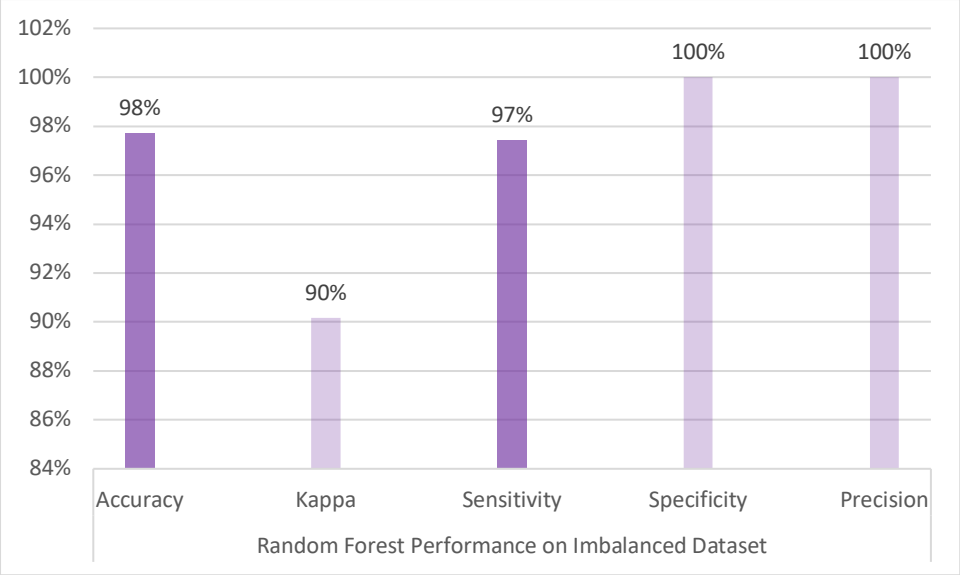
```r
Feature_Data <- Data_pre[-ncol(Data_pre)]
correlation_matrix <- cor(Feature_Data)
# Visualize the correlation matrix
ggcorrplot(correlation_matrix)
```



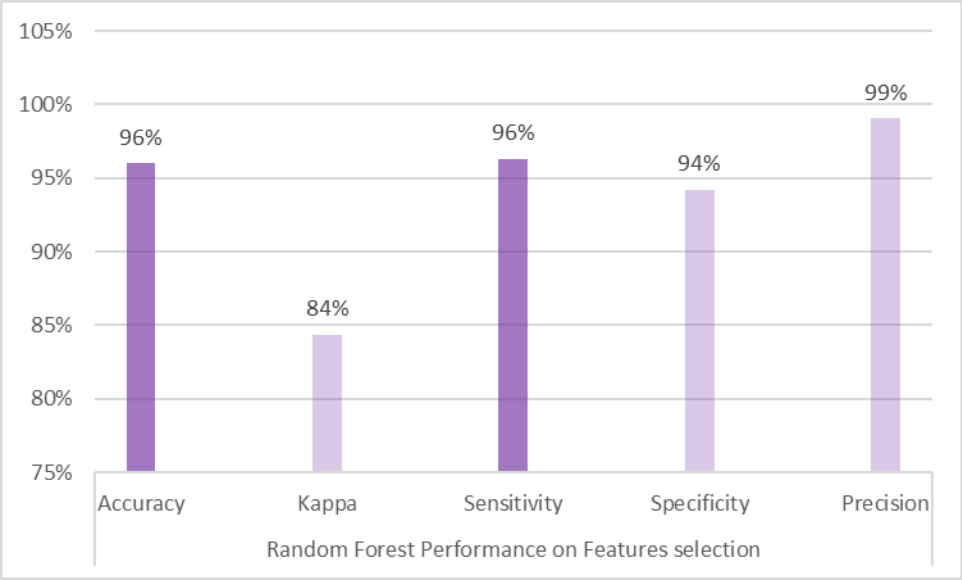
Appendix 01 - Result of KNN on Imbalanced Dataset



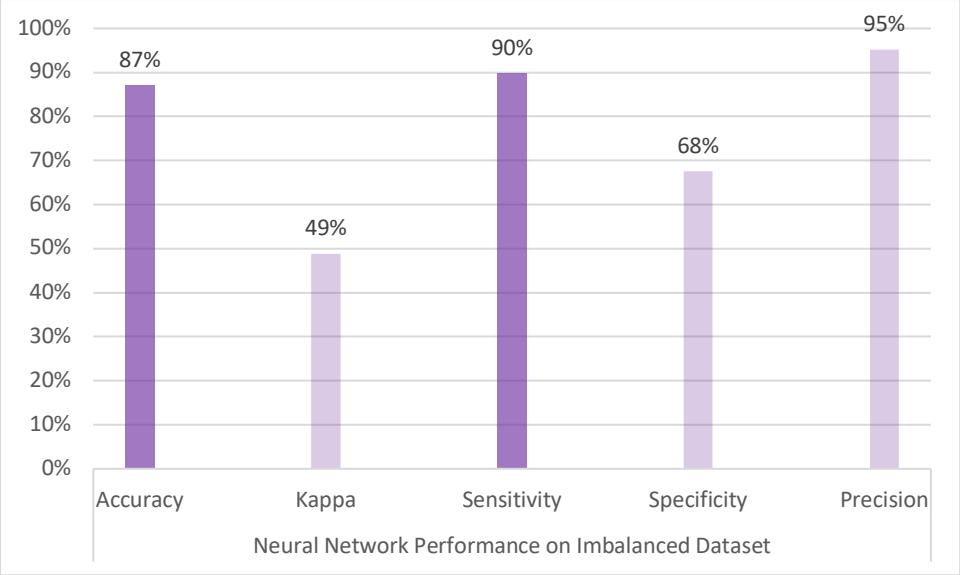
Appendix02 - Result of KNN on Features Selection



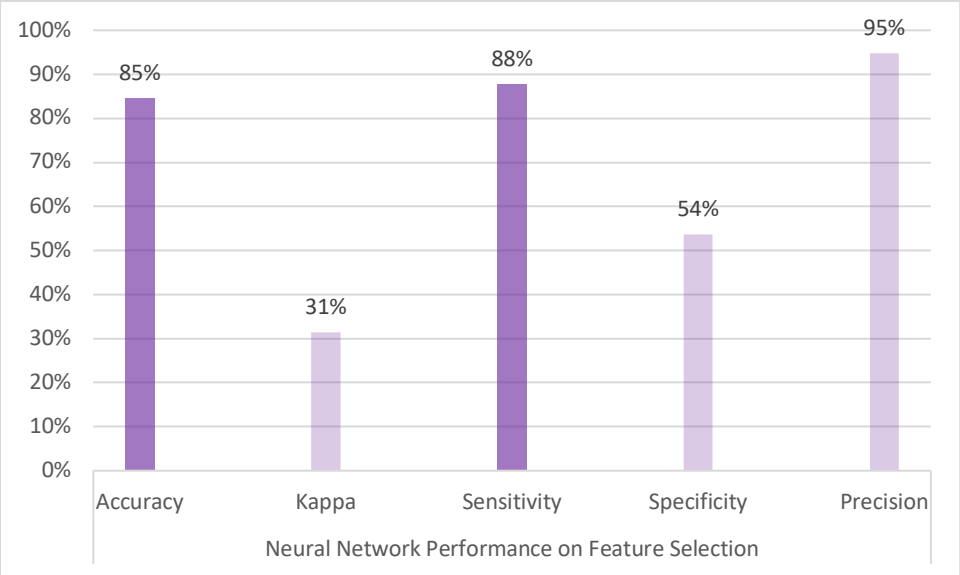
Appendix03 - Result of Random Forest on Imbalanced Dataset



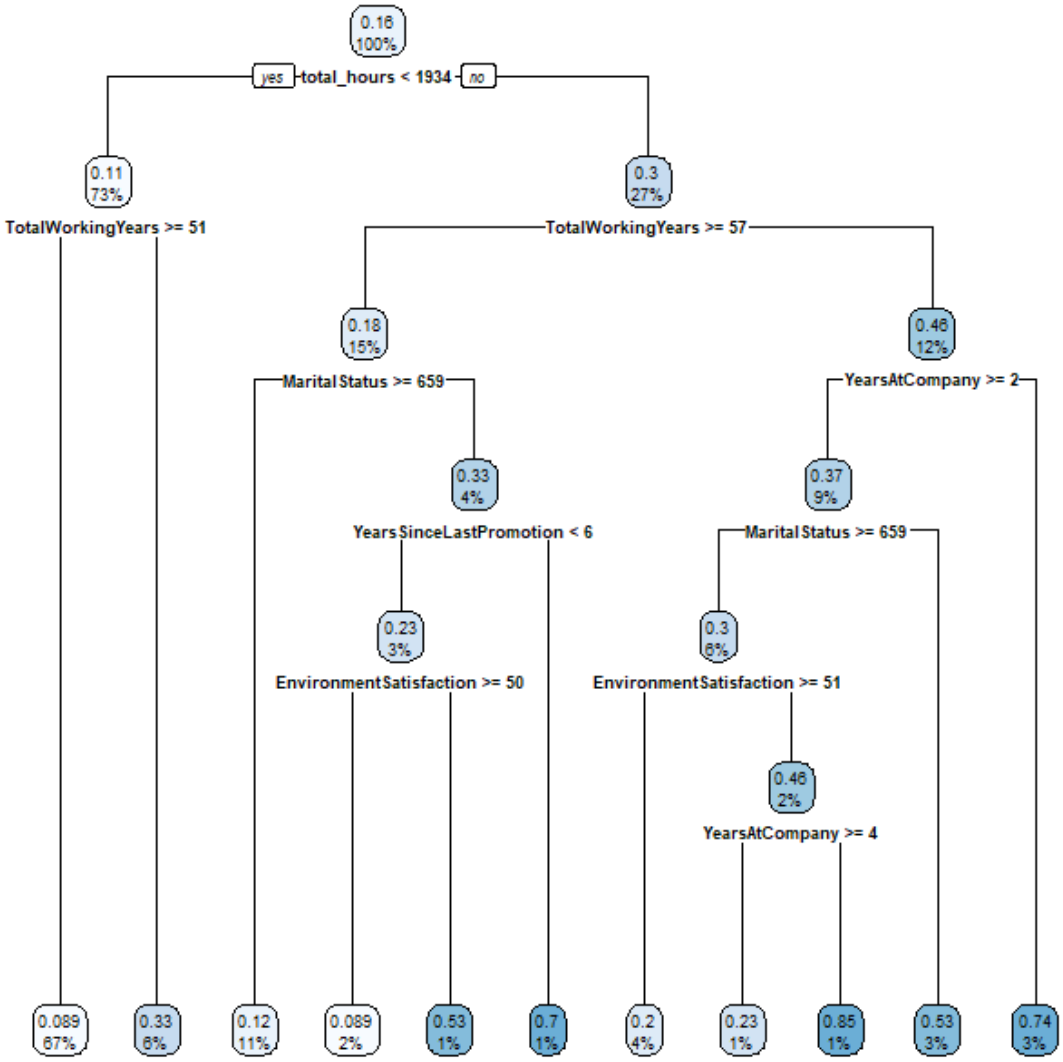
Appendix04 - Result of Random Forest on Feature Selection



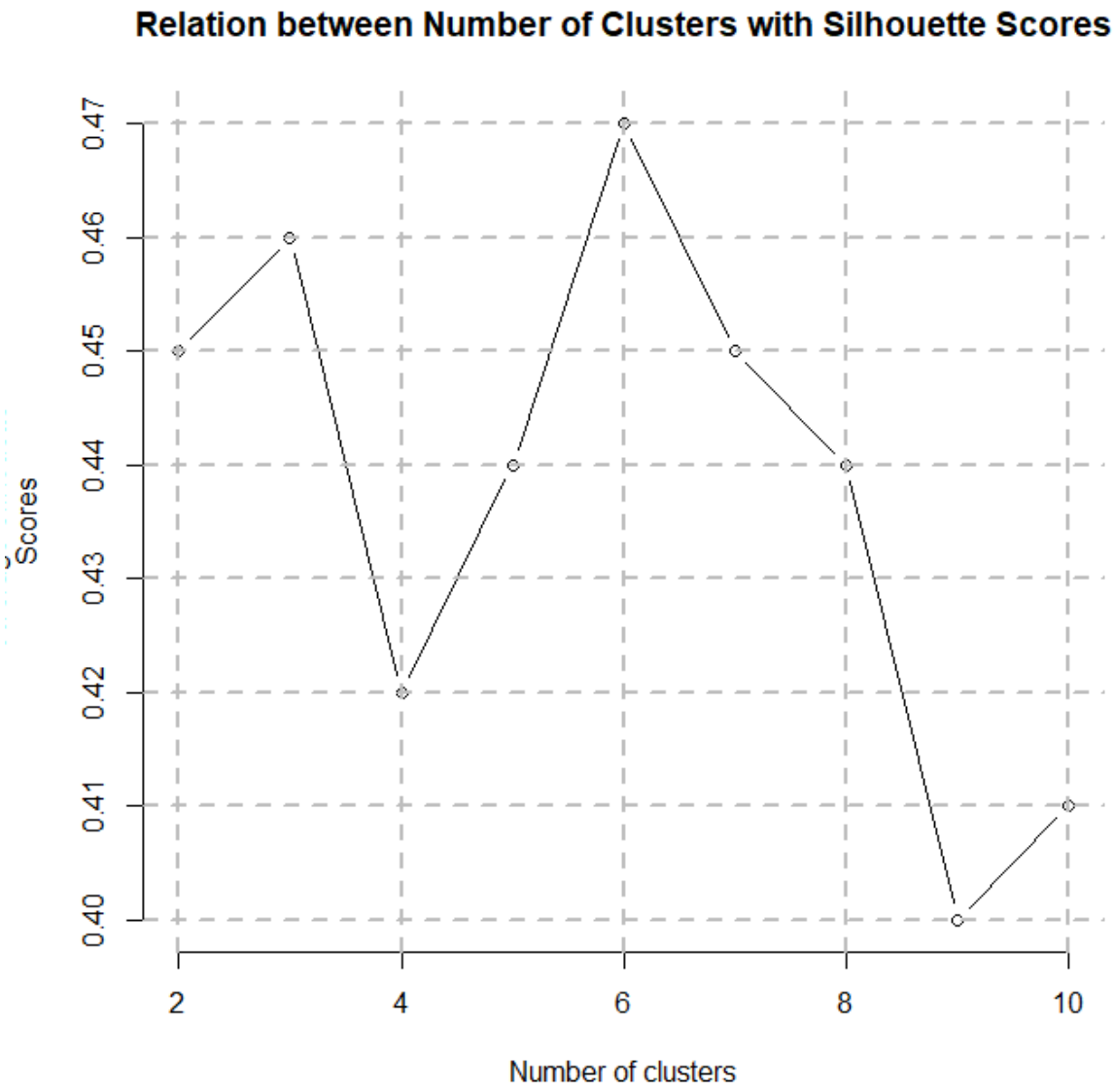
Appendix05 - Result of Neural Network on Imbalanced Dataset



Appendix06 - Result of Neural Network on Feature Selection



Appendix06 - Feature Selection verification with help of R Part



Appendix07 – Relation between Number of Cluster With Silhouette Scores