Brunel University London

STUDENT ID: 2241023

CS5812-Distributed Data Analysis

Coursework Assessment for 2022/23

## Introduction:

Attrition in the HR department is a widespread issue today for many organizations. When employees depart, a business may lose significant talent, knowledge, and experience. Employer replacement can also be expensive for businesses to recruit and train new staff.
Therefore, it is critical for organizations to forecast staff attrition and identify the variables that contribute to it. The Kaggle HR Attrition dataset can be used in this situation. This dataset includes personnel information from multiple firm departments, including age, gender, education, job title, pay, and performance evaluation. Organizations can learn more about the causes of employee attrition and take action to mitigate it by analyzing this data. The research aims to anticipate employee attrition based on various variables, including workplace satisfaction, performance gains, travel time, work-life balance, and other pertinent variables. The attrition column, which indicates whether an employee has left the organization, would be the dependent variable in this study.

## Data Description: source: (www.kaggle.com, n.d.)

| Name | Description |
|---|---|
| AGE | Numerical Value |
| ATTRITION | Employee leaving the company (0=no, 1=yes) |
| BUSINESS TRAVEL | (1=No Travel, 2=Travel Frequently, 3=Tavel Rarely) |
| DAILY RATE | Numerical Value - Salary Level |
| DEPARTMENT | (1=HR, 2=R&D, 3=Sales) |
| DISTANCE FROM HOME | Numerical Value - THE DISTANCE FROM WORK TO HOME |
| EDUCATION | Numerical Value |
| EDUCATION FIELD | (1=HR, 2=LIFE SCIENCES, 3=MARKETING, 4=MEDICAL SCIENCES, 5=OTHERS, 6= TEHCNICAL) |
| EMPLOYEE COUNT | Numerical Value |
| EMPLOYEE NUMBER | Numerical Value - EMPLOYEE ID |
| ENVIROMENT SATISFACTION | Numerical Value - SATISFACTION WITH THE ENVIROMENT |
| GENDER | (1=FEMALE, 2=MALE) |
| HOURLY RATE | Numerical Value - HOURLY SALARY |
| JOB INVOLVEMENT | Numerical Value - JOB INVOLVEMENT |
| JOB LEVEL | Numerical Value - LEVEL OF JOB |
| JOB ROLE | (1=HC REP, 2=HR, 3=LAB TECHNICIAN, 4=MANAGER, 5= MANAGING DIRECTOR, 6= REASEARCH DIRECTOR, 7= RESEARCH SCIENTIST, 8=SALES EXECUTIEVE, 9= SALES REPRESENTATIVE) |
| JOB SATISFACTION | Numerical Value - SATISFACTION WITH THE JOB |
| MARITAL STATUS | (1=DIVORCED, 2=MARRIED, 3=SINGLE) |
| MONTHLY INCOME | Numerical Value - MONTHLY SALARY |

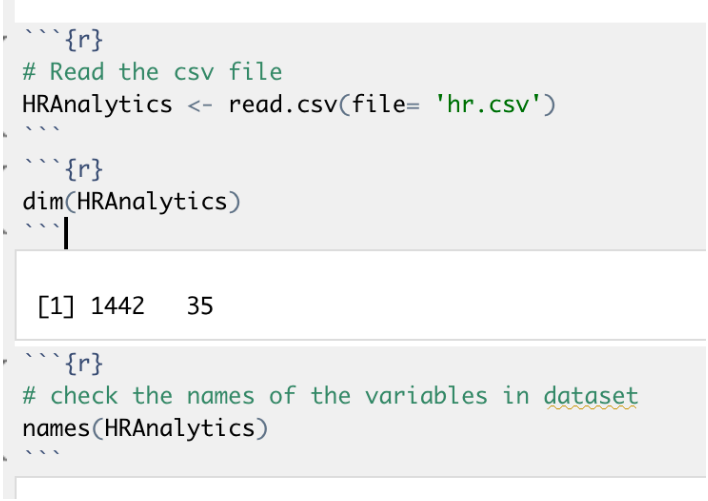
| | |
|---|---|
| MONTHY RATE | Numerical Value - MONTHY RATE |
| NUMCOMPANIES WORKED | Numerical Value - NO. OF COMPANIES WORKED AT |
| OVER 18 | (1=YES, 2=NO) |
| OVERTIME | (1=NO, 2=YES) |
| PERCENT SALARY HIKE | Numerical Value - PERCENTAGE INCREASE IN SALARY |
| PERFORMANCE RATING | Numerical Value - ERFORMANCE RATING |
| RELATIONS SATISFACTION | Numerical Value - RELATIONS SATISFACTION |
| STANDARD HOURS | Numerical Value - STANDARD HOURS |
| STOCK OPTIONS LEVEL | Numerical Value - STOCK OPTIONS |
| TOTAL WORKING YEARS | Numerical Value - TOTAL YEARS WORKED |
| TRAINING TIMES LAST YEAR | Numerical Value - HOURS SPENT TRAINING |
| WORK LIFE BALANCE | Numerical Value - TIME SPENT BEWTWEEN WORK AND OUTSIDE |
| YEARS AT COMPANY | Numerical Value - TOTAL NUMBER OF YEARS AT THE COMPNAY |
| YEARS IN CURRENT ROLE | Numerical Value -YEARS IN CURRENT ROLE |
| YEARS SINCE LAST PROMOTION | Numerical Value - LAST PROMOTION |
| YEARS WITH CURRENT MANAGER | Numerical Value - YEARS SPENT WITH CURRENT MANAGER |

## Research Question:

The goal of the study is to identify parameters that can be utilized to forecast employee turnover, including workplace satisfaction, performance increases, travel time, work-life balance, and other pertinent variables. The attrition column, which indicates whether an employee has left the organization, would be the dependent variable in this study.

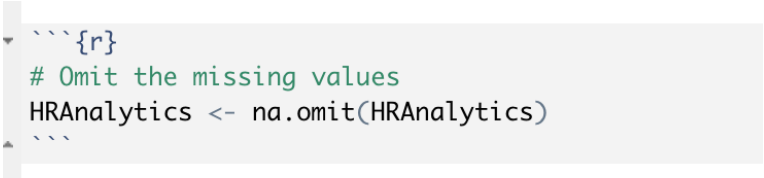
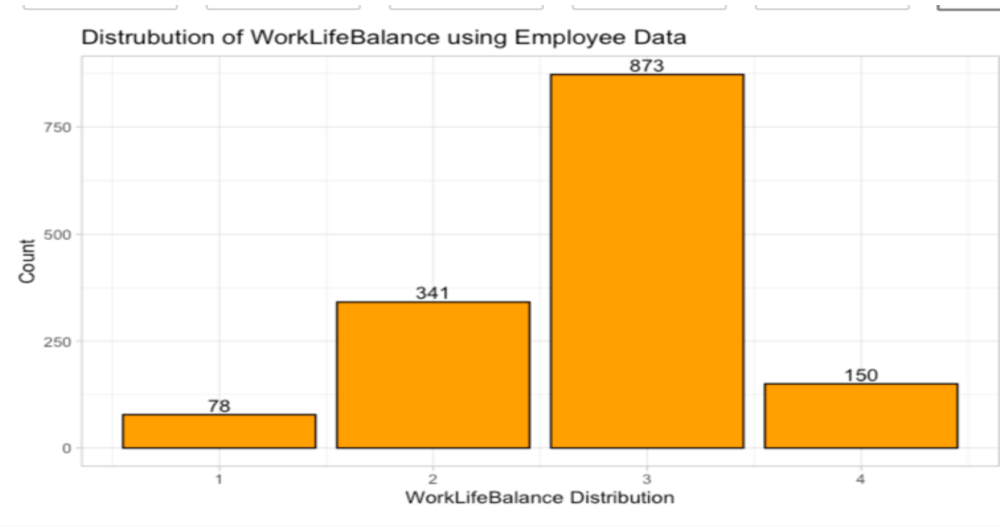
## Data Preparation and Cleaning:

Data preparation is the process of analysis of the data to analyse the issues associated with the dataset.
First step is to read the input file and analyse the structure of the data frame using read.csv, names(), str(), and summary() method.



~~~
```{r}
# Read the csv file
HRAnalytics <- read.csv(file= 'hr.csv')
~~~
```{r}
dim(HRAnalytics)
```

[1] 1442   35

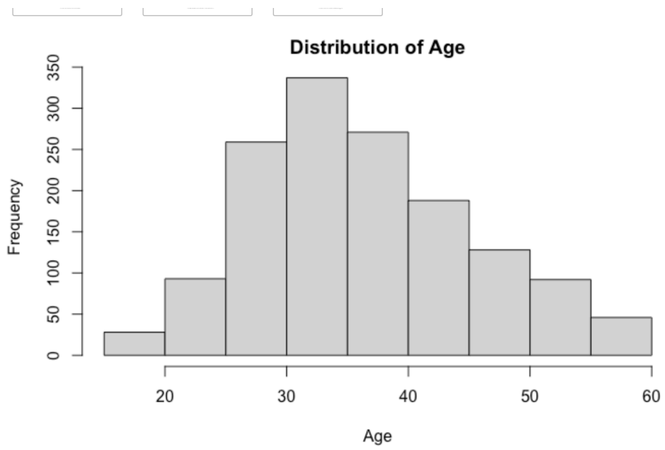
```{r}
# check the names of the variables in dataset
names(HRAnalytics)
```
```

The dataset consists of 1442 rows and 35 columns, and our dependent column is “Attrition”. From str () we can see the datatype of columns are char and int.

Using summary () method we can see that the dataset is having 11 missing values. Removal of NA values is important because as NA values can affect the corelation and covariance between the variables and some machine learning algorithms cannot handle null values. To remove NA values from the data frame, we have used na.omit().



```
```{r}
# Omit the missing values
HRAnalytics <- na.omit(HRAnalytics)
```
```

Once the NA values are removed next step is to analyse the categorical features using table () and converting the values of the categorical column to factors. For the visualisation, we have used histogram to check the distribution of the numerical columns and for categorical columns we have used bar plot.


```

## Exploratory Data Analysis:

To better understand the information included in a dataset, exploratory data analysis (EDA) is employed. Discovering patterns, correlations, and outliers in the data requires visualizing and summarizing the data using statistical and graphical tools.

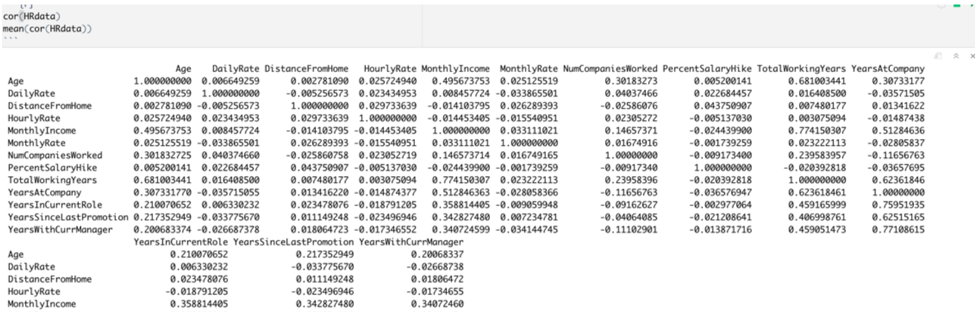
### Univariate Analysis

1. To check the normal distribution among numerical variable we have used histogram.



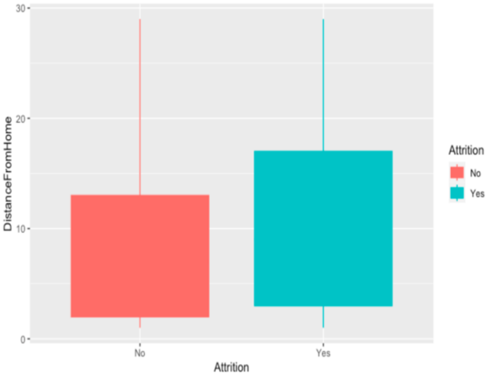
### Corelation

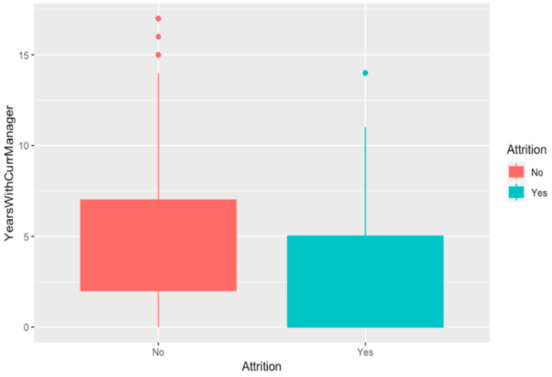
Next is to analyse the relationship between the variables. Therefore, to analyse the relation between numerical columns we have used cor test.



```
cor(HRdata)
mean(cor(HRdata))
```
```

                                Age    DailyRate DistanceFromHome   HourlyRate MonthlyIncome  MonthlyRate NumCompaniesWorked PercentSalaryHike TotalWorkingYears YearsAtCompany
Age                     1.000000000  0.006649259      0.002781090  0.025724940   0.495673753  0.025125519         0.30183273       0.005200141       0.681003441     0.30733177
DailyRate               0.006649259  1.000000000     -0.005256573  0.023434953   0.008457724 -0.033865501         0.04037466       0.022684457       0.016408500    -0.03571505
DistanceFromHome        0.002781090 -0.005256573      1.000000000  0.029733639  -0.014103795  0.026289393        -0.02586076       0.043750907       0.007480177     0.01341622
HourlyRate              0.025724940  0.023434953      0.029733639  1.000000000  -0.014453405 -0.015540951         0.02305272      -0.005137030       0.003075094    -0.01487438
MonthlyIncome           0.495673753  0.008457724     -0.014103795 -0.014453405   1.000000000  0.033111021         0.14657371      -0.024439900       0.774150307     0.51284636
MonthlyRate             0.025125519 -0.033865501      0.026289393 -0.015540951   0.033111021  1.000000000         0.01674916      -0.001739259       0.023222113    -0.02805837
NumCompaniesWorked      0.301832725  0.040374660     -0.025860758  0.023052719   0.146573714  0.016749165         1.00000000      -0.009173400       0.239583957    -0.11656763
PercentSalaryHike       0.005200141  0.022684457      0.043750907 -0.005137030  -0.024439900 -0.001739259        -0.00917340       1.000000000      -0.020392818    -0.03657695
TotalWorkingYears       0.681003441  0.016408500      0.007480177  0.003075094   0.774150307  0.023222113         0.23958396      -0.020392818       1.000000000     0.62361846
YearsAtCompany          0.307331770 -0.035715055      0.013416220 -0.014874377   0.512846363 -0.028058366        -0.11656763      -0.036576947       0.623618461     1.00000000
YearsInCurrentRole      0.210070652  0.006330232      0.023478076 -0.018791205   0.358814405 -0.009059948        -0.09162627      -0.002977064       0.459165999     0.75951935
YearsSinceLastPromotion 0.217352949 -0.033775670      0.011149248 -0.023496946   0.342827480  0.007234781        -0.04064085      -0.021208641       0.406998761     0.62515165
YearsWithCurrManager    0.200683374 -0.026687378      0.018064723 -0.017346552   0.340724599 -0.034144745        -0.11102901      -0.013871716       0.459051473     0.77108615
                        YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
Age                            0.210070652             0.217352949           0.20068337
DailyRate                      0.006330232            -0.033775670          -0.02668738
DistanceFromHome               0.023478076             0.011149248           0.01806472
HourlyRate                    -0.018791205            -0.023496946          -0.01734655
MonthlyIncome                  0.358814405             0.342827480           0.34072460
```

To analyse the relationship between the categorical and numerical column we have used boxplot. The below box plot represents that the minmum distance from same is equal, but the maximum percenatge of the people who are leaving the company has the highest distance from home, Next with the comparison of attrition with the years with current manager , is high in the people who stays in the company. It means that who stays long with the current manager are not likely to leave.

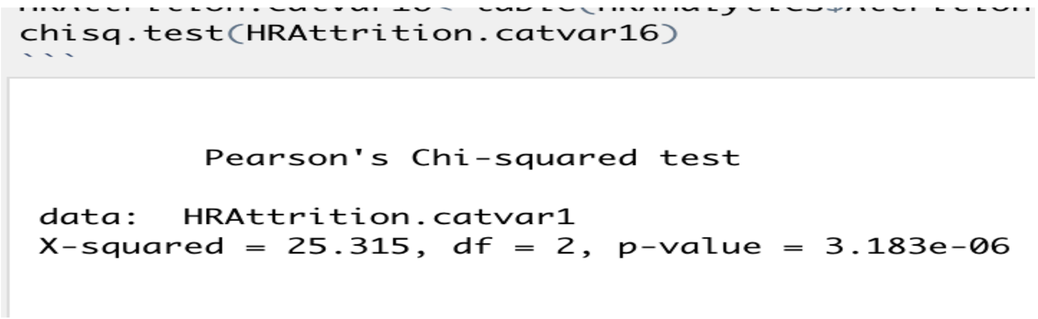




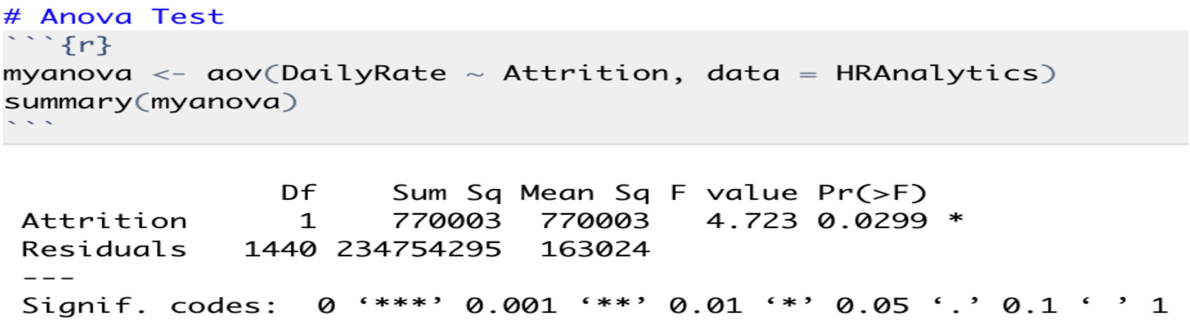
- The median employee age is 38, according to the boxplot. The interquartile range (IQR) is 10 years, therefore 50% of employees are between 28 and 48 years old. One 65-year-old employee stands out.
- The median YearsAtCompany boxplot is 5 years. The IQR is 4 years, therefore the middle 50% of employees have been with the company for 1–9 years. One exception is a 20-year corporate veteran.

To compare between categorical columns, we have used chi-square test.

H0: The null hypothesis (H0) states that no meaningful correlation exists between any two category variables.



```
HRAttrition.catvar16<- table(HRAnalytics$Attrition
chisq.test(HRAttrition.catvar16)
```
```

```
	Pearson's Chi-squared test

data:  HRAttrition.catvar1
X-squared = 25.315, df = 2, p-value = 3.183e-06
```

To analyse the relationship between the categorical and numerical column we have used Anova test.

H0: ANOVA's null hypothesis (H0) states that there is no difference between the means of any two or more groups.



```
# Anova Test
```{r}
myanova <- aov(DailyRate ~ Attrition, data = HRAnalytics)
summary(myanova)
```

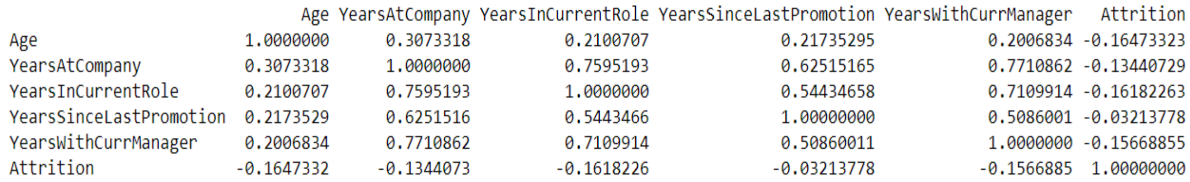
              Df    Sum Sq Mean Sq F value Pr(>F)  
Attrition      1    770003  770003   4.723 0.0299 *
Residuals   1440 234754295  163024                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```
```

The key insights from the EDA, from the corelation table we can see the columns which are interrelated to each other as they have cor() value <0.05. therefore, all the input variables are not mutually independent to each other, and same for chi squared.

- Attrition increases with age. Older workers are slightly more likely to quit than younger workers.
- YearsAtCompany: Attrition is moderately positively correlated with years in the company. This suggests that longer-tenured employees are slightly more likely to depart the organization.
- YearsInCurrentRole: Attrition is weakly positively correlated with years in the current role. This suggests that longer-tenured employees are slightly more likely to depart the organization.

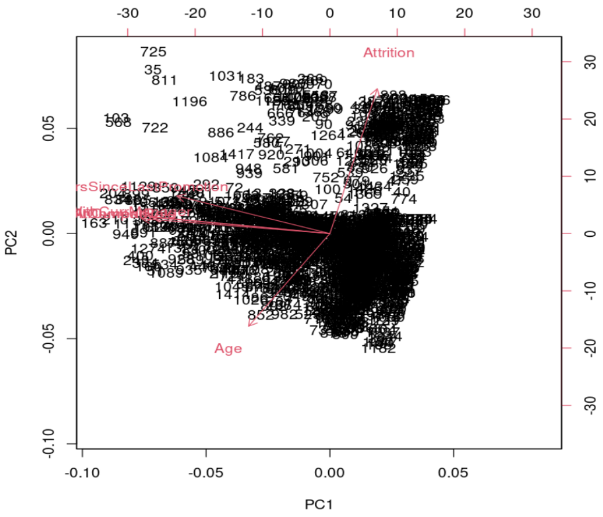
## **Principal Component Analysis:**

The next step is to perform the Principal Component Analysis, or PCA, is a statistical technique used to find the most important variables or features in a dataset, thereby reducing the dimensionality of the dataset. A high-dimensional dataset's key information must be extracted to display it in a more manageable and compact manner without substantially losing any of the original data.



```
                              Age YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager   Attrition
Age                     1.0000000      0.3073318          0.2100707              0.21735295            0.2006834 -0.16473323
YearsAtCompany          0.3073318      1.0000000          0.7595193              0.62515165            0.7710862 -0.13440729
YearsInCurrentRole      0.2100707      0.7595193          1.0000000              0.54434658            0.7109914 -0.16182263
YearsSinceLastPromotion 0.2173529      0.6251516          0.5443466              1.00000000            0.5086001 -0.03213778
YearsWithCurrManager    0.2006834      0.7710862          0.7109914              0.50860011            1.0000000 -0.15668855
Attrition              -0.1647332     -0.1344073         -0.1618226             -0.03213778           -0.1566885  1.00000000
```

- PC1: The first principal component accounts for 44.8% of data variance. PC1 is inversely connected with age and years at the company and favourably correlated with years in the current role, since the previous promotion, with the current management, and attrition. PC1 may indicate job satisfaction. Dissatisfied workers are older, longer-tenured, and lower-paid. Satisfied employees are more likely to be in higher-paying occupations, have been in their present function for a longer period, have been promoted recently, and have been with their current management for a longer time.

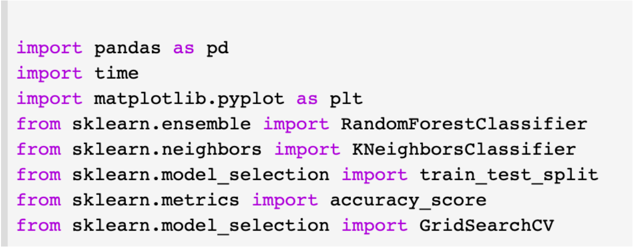


- PC2: The second main component accounts for 15.7% of data variance. Age, years in the role, years since promotion, and years with the manager negatively affect PC2. PC2 may indicate work stability. Older workers, those who have been promoted recently, and those who have worked with their manager for a long time are more likely to have stable jobs.
- Importantly, the primary components are not the original variables. Principal components are linear combinations of original variables. The primary components can represent the original variables, but they do not carry all their information.

- The first principal component (PC1) is responsible for the most variation in the data, whereas PC2 is responsible for the next most variation. According to this, PC1 and PC2 are the two most crucial PCs.

Employees who are more likely to leave the company can be pinpointed with the help of the principal component analysis (PCA), but it's vital to look at other aspects like job happiness, job security, career growth opportunities, and satisfaction with one's manager as well.

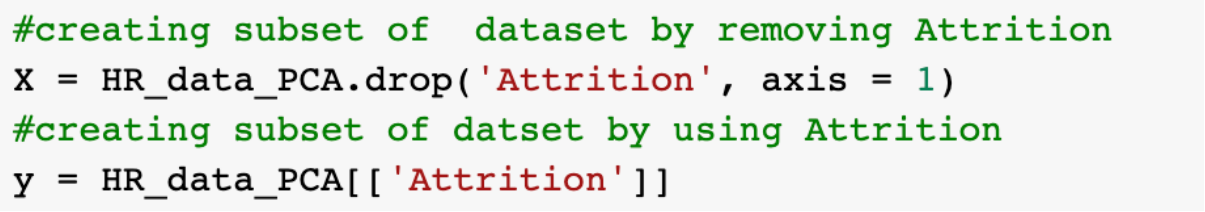
## Machine Learning Prediction



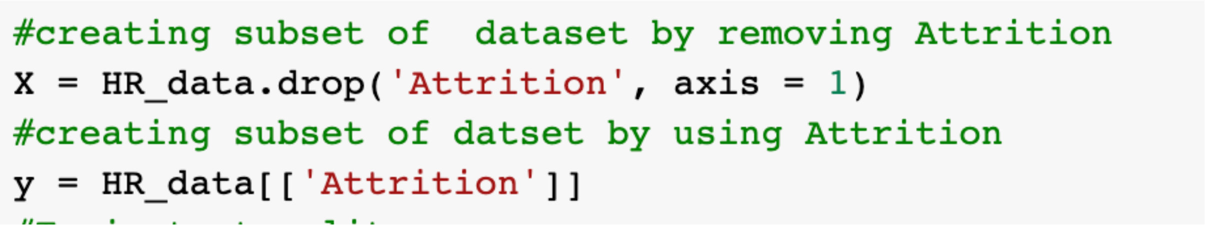
```
import pandas as pd
import time
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.model_selection import GridSearchCV
```

Above libraries have been imported for the implementation of our model.
First, we import the data from a CSV file and divide both the dataset into two subsets, one without the target variable (X) and the other with just the target variable (y) present. We then divided the both data using the Scikit-learn module into training and testing sets.

two datasets, one on which we performed PCA and taken the most important features. Other dataset we are using on which PCA is not performed so that we can check by reducing the dimension how it effect the accuracy.

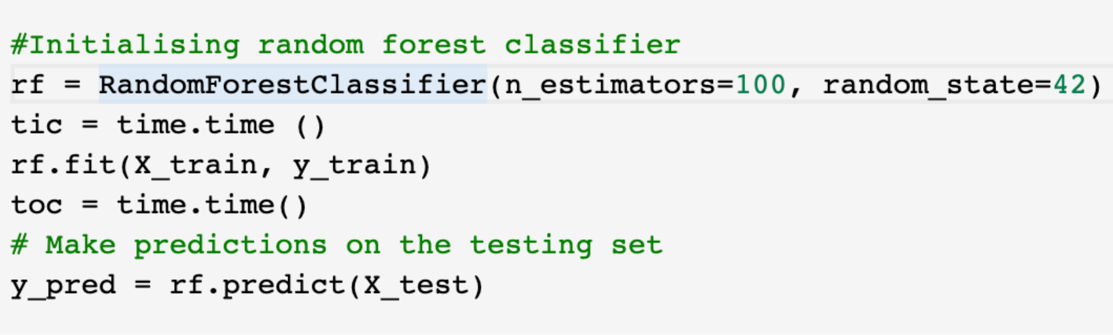


```
#creating subset of  dataset by removing Attrition
X = HR_data_PCA.drop('Attrition', axis = 1)
#creating subset of datset by using Attrition
y = HR_data_PCA[['Attrition']]
```



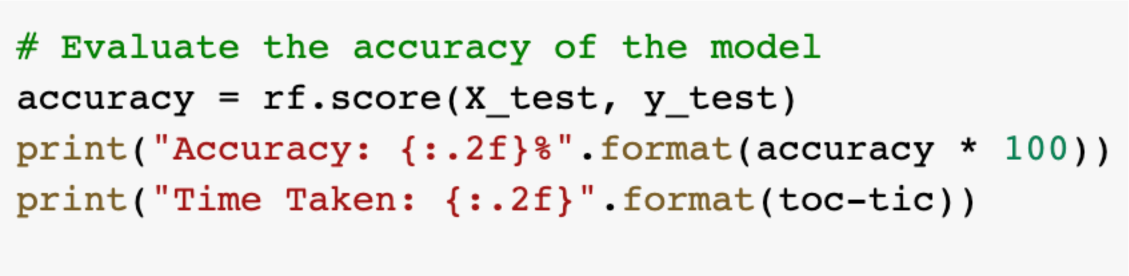
```
#creating subset of  dataset by removing Attrition
X = HR_data.drop('Attrition', axis = 1)
#creating subset of datset by using Attrition
y = HR_data[['Attrition']]
```

Next, one first dataset-initialise a Random Forest Classifier with the quantity of estimations (estimators) and random state (random state) in the forest. Using the fit () and predict () functions, we fit the Random Forest model to the training set of data and generate predictions for the testing set.



```
#Initialising random forest classifier
rf = RandomForestClassifier(n_estimators=100, random_state=42)
tic = time.time ()
rf.fit(X_train, y_train)
toc = time.time()
# Make predictions on the testing set
y_pred = rf.predict(X_test)
```

Using the score () method, we assess the model's accuracy and print the accuracy as well as the length of time it took to train and generate predictions.
Same process is applied on second dataset.

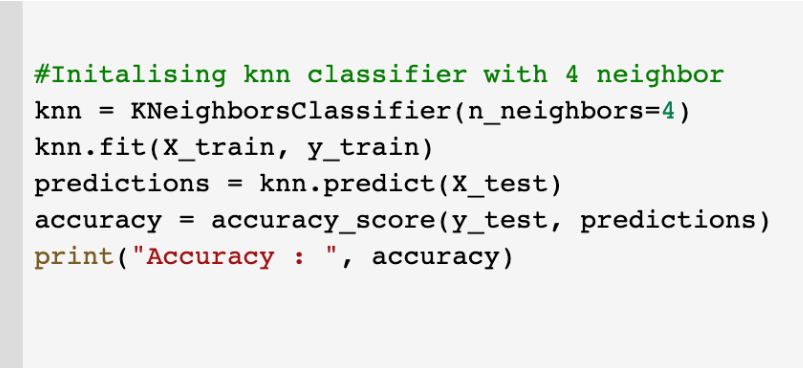


```
# Evaluate the accuracy of the model
accuracy = rf.score(X_test, y_test)
print("Accuracy: {:.2f}%".format(accuracy * 100))
print("Time Taken: {:.2f}".format(toc-tic))
```

Accuracy of random forest model is 84.3%

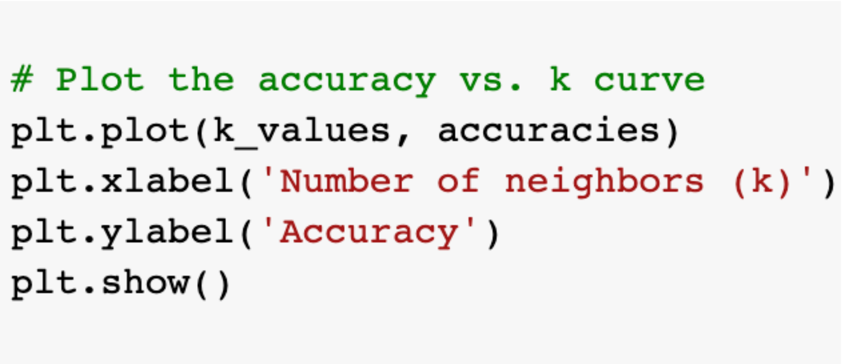
## K Nearest Neighbour:

KNN is most popular model for classification and regression issues in machine learning. The most common class (in classification) or average value (in regression) among those K neighbours is assigned as the predicted value in KNN.
This is done by locating the K number of training observations that are closest to the new observation, as determined by a distance metric (such as Euclidean distance).
we can use the Scikit-learn library's KNeighborsClassifier() function to conduct k-nearest neighbour classification. Using a variety of k values, we fit the model, generate predictions, and assess the model's efficacy using the accuracy_score() function. Using the Matplotlib tool, we visualise the correlation between the accuracy of the KNN classifier and the number of neighbours (k). With k=4, we fit the KNN classifier, generate predictions, and use the accuracy_score() method to assess the model's efficacy. We output the model's final accuracy.

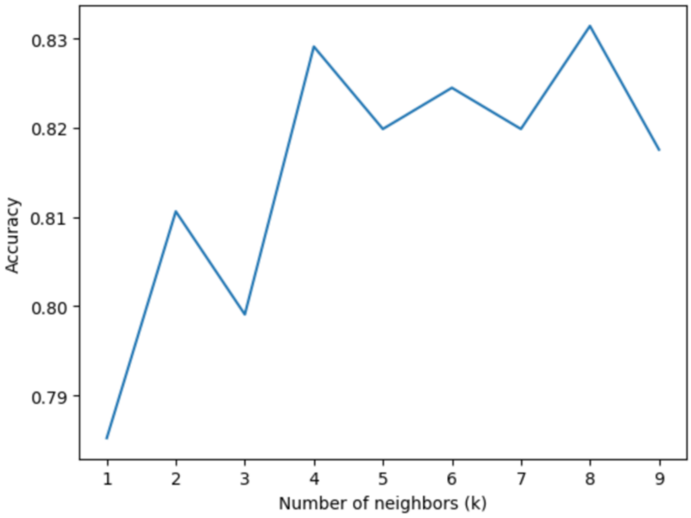


```
#Initalising knn classifier with 4 neighbor
knn = KNeighborsClassifier(n_neighbors=4)
knn.fit(X_train, y_train)
predictions = knn.predict(X_test)
accuracy = accuracy_score(y_test, predictions)
print("Accuracy : ", accuracy)
```

We can visualise the accuracy to the number of neighbours.



```
# Plot the accuracy vs. k curve
plt.plot(k_values, accuracies)
plt.xlabel('Number of neighbors (k)')
plt.ylabel('Accuracy')
plt.show()
```



```
Accuracy :  0.8290993071593533
/usr/local/lib/python3.9/dist-packages/sklearn/neighbors/_classificat
  return self._fit(X, y)
```

Accuracy of K nearest neighbour is 82.

## Comparision between the models:

The HR dataset's complicated correlations between the input characteristics and the target variable were captured by the Random Forest Classifier without PCA, yielding a high accuracy score on the test set. This might be because Random Forest is a strong and adaptable method that handles high-dimensional datasets and unbalanced class distributions by mixing several decision trees to produce predictions with higher accuracy. However, because K-Nearest Neighbours (KNN) depends on the closeness of data points to produce predictions and has trouble with high-dimensional datasets and class imbalance, it might not perform well in this situation. Therefore, the Random Forest Classifier without PCA is the best model for this situation with 84% of accuracy.

## Performance Evaluation:

The models used by teammates are Machine Learning Model

Random Forest- 84%

KNN (K-Nearest Neighbours)-82%

Neural Network (R)-95%

Neural Network (Pyspark)-82%

Decision Tree-82%

Logistic Regression-85%

Suppert Vector Machine-78%

According to the accuracies provided the best model is Neural network. which is best suitable for this data. In this instance, the neural network model may have learnt a more intricate representation of the input features, enabling it to identify the underlying patterns more accurately in the data. A variety of optimisation methods, including backpropagation, can also be used to train neural networks, which can enhance their performance.

## HPC implementation:

The model without pyspark runs for 0.31 seconds and the model with the pyspark runs for 3.99 seconds.

## Author Contribution Statement:

Together, the five of us—including myself—worked to ensure the success of our endeavour. We divided the tasks according to our individual abilities and areas of expertise. Together, Krupa Kakadiya, Diksha Jadhav, and Hasitha Kutala selected the data and used RStudio to clean the data. The item was approved by the group following a thorough debate. Me and Diksha prepared the data, and they also worked on outlier correction and principal component analysis (PCA) of the data. The exploratory data analysis (EDA) charts were made in R by Rushikesh and Rohit Shivhare and featured the correlation and Chi-Square test in addition to other EDA activities. Overall, this collaborative approach enabled us to effectively finish our project and achieve our objectives.