**HR Analytics**

**Chapter 1: introduction**

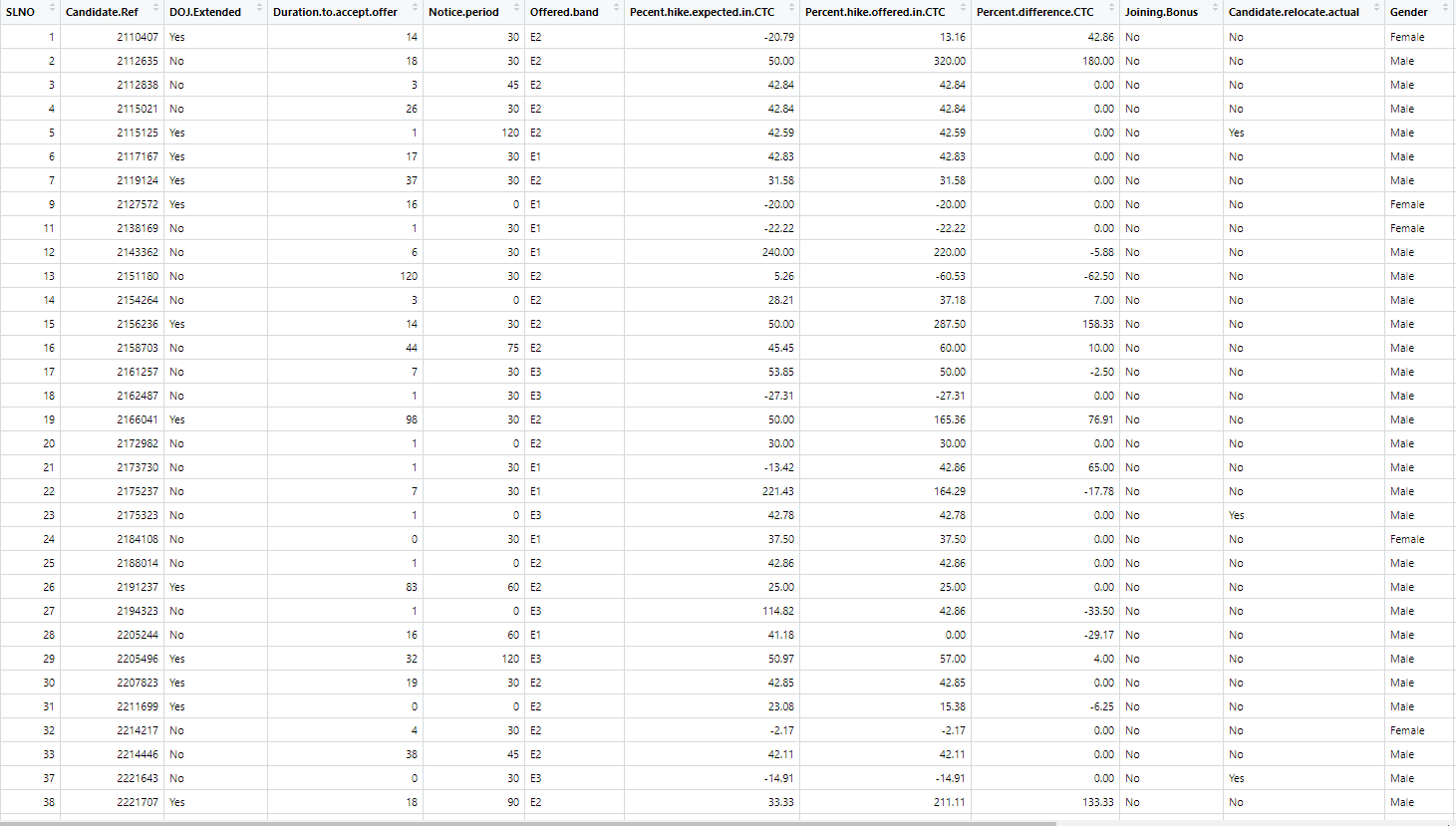
**Scope of Analysis:**

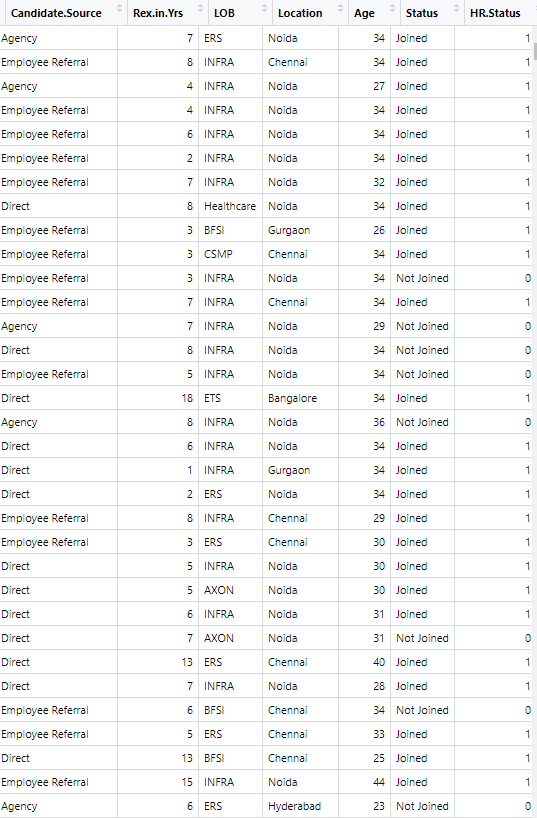
The dataset HR.Analytics.csv contains sample of candidates that we part of a recruitment process of particular client of Scalene Works. supports several information technology(IT) companies in India with their talent acquisition. One of the challenge they face is about "30% of the candidates who accept the jobs offers, do not join the company."

This lead to huge loss of revenue and time as the companies initiate the recruitment process again to fill the workforce demand. Scalene Works want to find out if a model can be build to predict the likelihood of a candidate joining the company. If the likelihood is high, then the company can go ahead and offer the jobs to the candidates.

**Chapter 2: Data understanding**

**Data Dictionary :**





**Candidate reference number :**

Unique number to identify the candidate

**DOJ extended :**

Binary variable identifying whether candidate asked for date of joining extension (Yes/No)

**Duration to accept the offer :**

Number of days taken by the candidate to accept the offer (continuous variable)

**Notice period :**

Notice period to be served in the parting company before candidate can join this company (continuous variable)

**Offered band :**

Band offered to the candidate based on experience and performance in interview rounds (categorical variable labelled C0/C1/C2/C3/C4/C5/C6)

**Percentage hike (CTC) expected :**

Percentage hike expected by the candidate (continuous variable)

**Percentage hike offered (CTC) :**

Percentage hike offered by the company (continuous variable)

**Joining bonus :**

Binary variable indicating if joining bonus was given or not (Yes/No)

**Gender :**

Gender of the candidate (Male/Female)

**Candidate source :**

Source from which resume of the candidate was obtained (categorical variables with categories: Employee referral/Agency/Direct)

**REX (in years) :**

Relevant years of experience of the candidate for the position offered (continuous variable)

**LOB :**

Line of business for which offer was rolled out (categorical variable)

**DOB :**

Date of birth of the candidate

**Joining location :**

Company location for which offer was rolled out for candidate to join (categorical variable)

**Candidate relocation status :**

Binary variable indicating whether candidate has to relocate from one city to another city for joining (Yes/No)

**HR status :**

Final joining status of candidate (Joined/Not-Joined)

**Install library package :**

The install.packages() can be used to install packages at the R console. The library() function loads packages that have been installed so that you may access the functionality in the package.

library(tidyverse)

library(ggplot2)

library(e1071)

library(caret)

library(rpart)

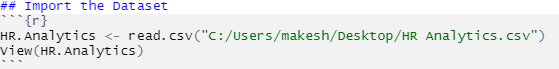
library(rpart.plot)

library(tune)

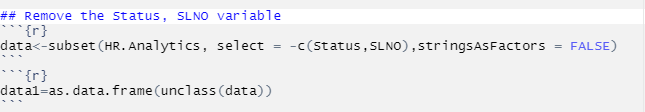
library(tree)

**Import the Dataset :**

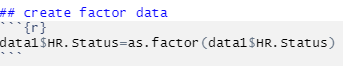
Importing data into R is a necessary step that, at times, can become time intensive. To ease this task, RStudio includes new features to import data from: csv, xls, xlsx, sav, dta, por, sas and stata files.



**Remove the Status, SLNO variable :**

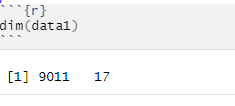


**Create factor data :**



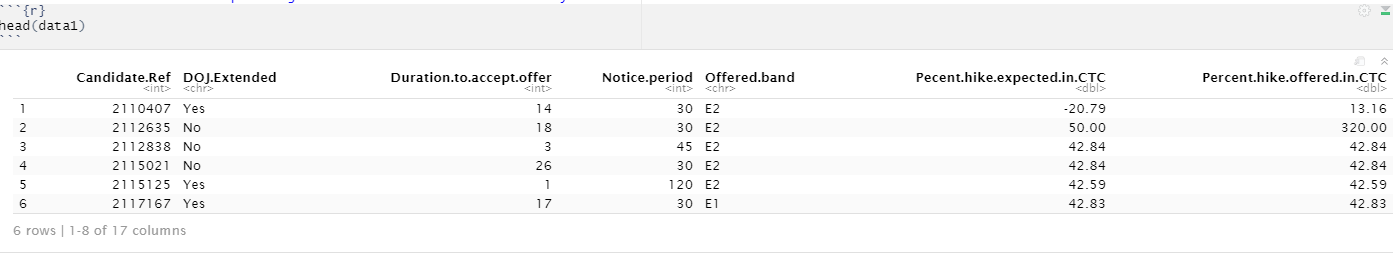
**Count of variable and observation :**

In the HR.Analytics data set, we can see immediately that there are only 9011 rows and 17 columns. This function is useful, because it tells us whether it would be okay to print the entire data frame to the console.



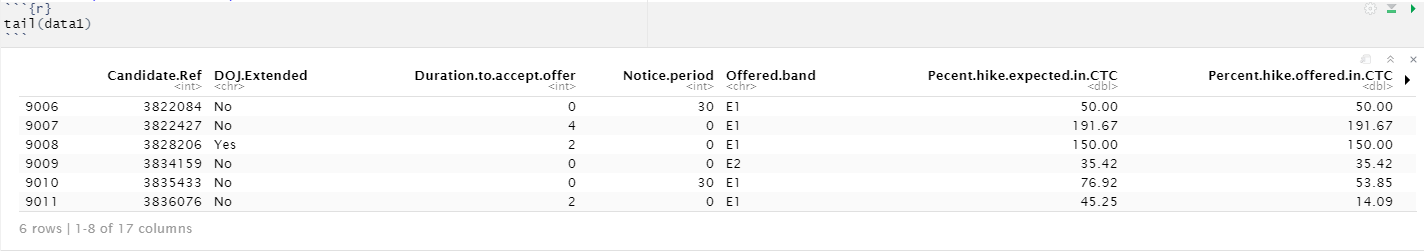
**Show the Head :**

This function defaults to printing the first 6 rows. In the HR.Analytic data set.



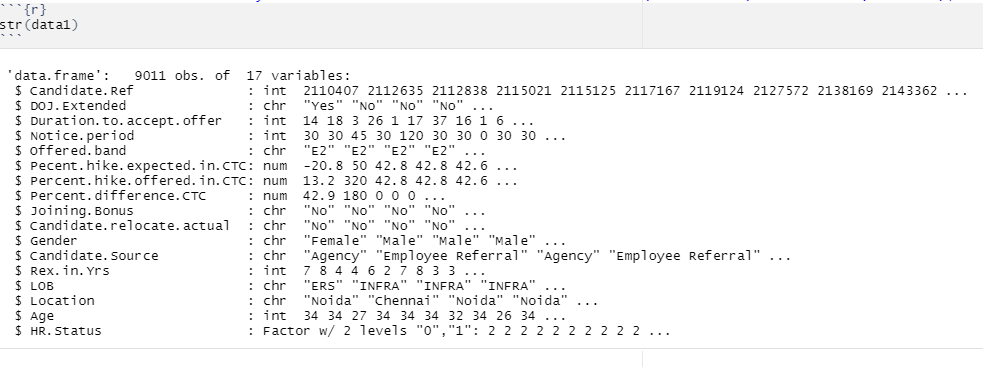
**Show the Tail :**

The except this function prints the end of the data frame. In this case, we’ve called the last 5 observations.



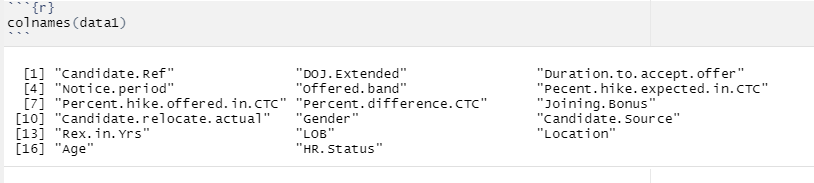
**Show the String :**

The structure of the HR.Analytics data set also tells us the number of rows (observations) and columns (variables), but it provides even more information.



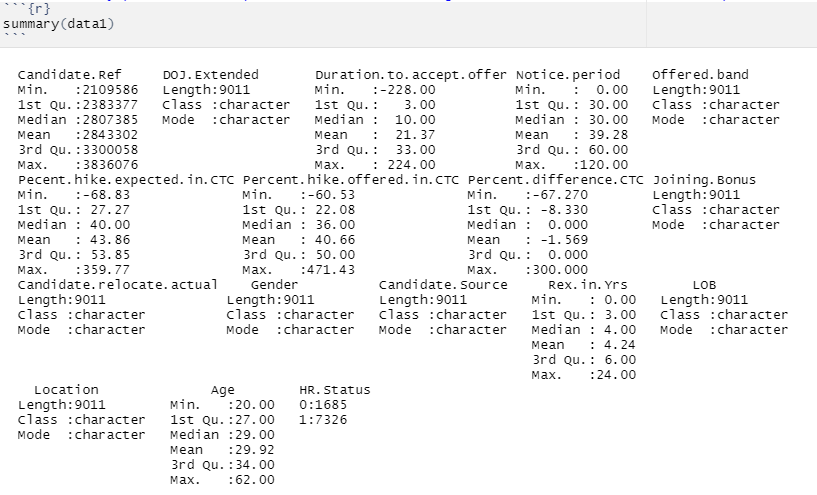
**Show the Column Name :**

This function prints a vector of the column names, which can be useful if you’re trying to reference a particular column. For the HR.Analytics data set.

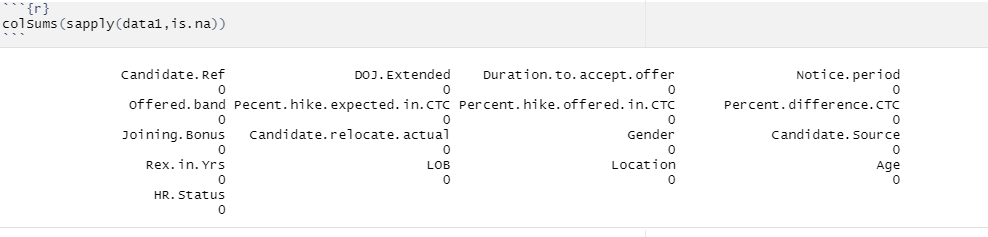


**Summary of dataset :**

The summary provides descriptive statistics including the min, max, mean, median, and quartiles of each column.

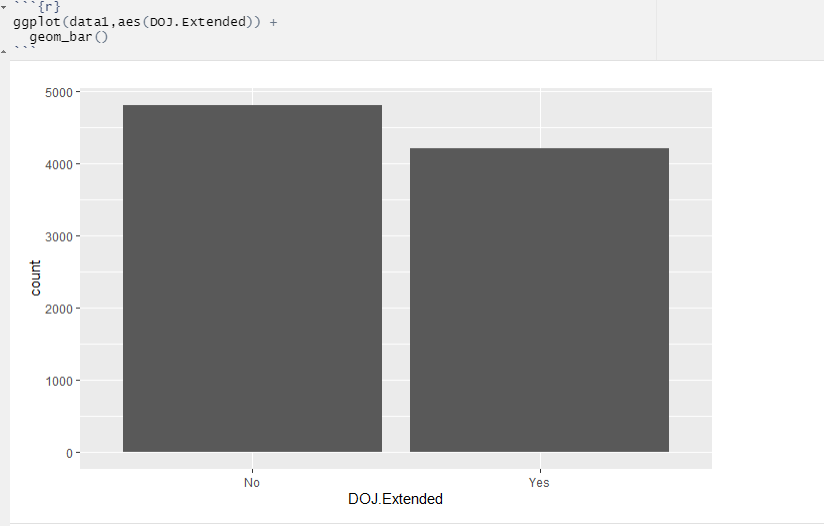


**check the clean data or not :**

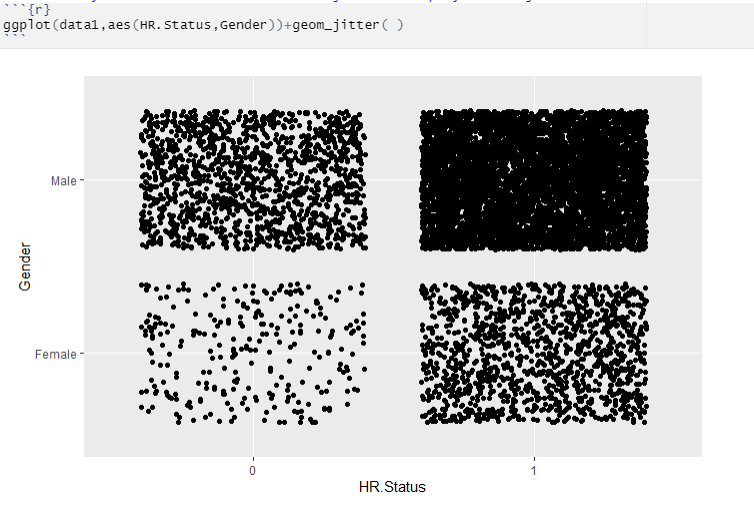


**Chapter 3: EDA using R**

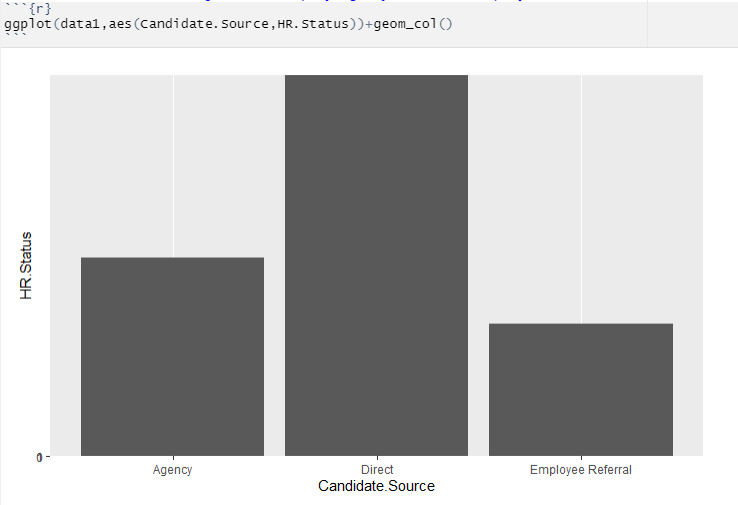
**3.1 How many people candidate asked for date of joining extension YES/NO**



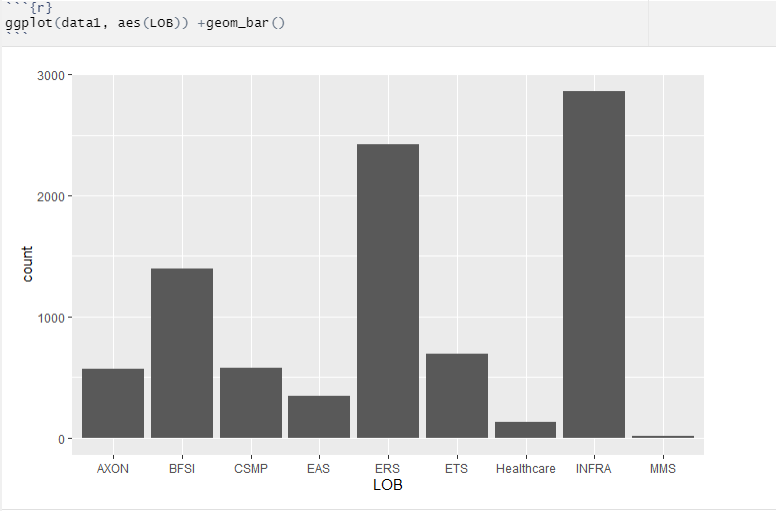
**3.2 How many candidate male or female to join the company or not join**



**3.3 The candidate how to join the company agency,direct or Employee referreal**



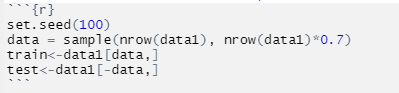
**3.4 which candidate work the previous company for which offer was rolled out**



**Chapter 4 : Model building**

**Train and test dataset :**

The splits 70% of the data selected randomly into training set and the remaining 30% sample into test data set.train=6307 obs. test=2704 obs, both 17 variable



**SVM Algorithm :**

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.

To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

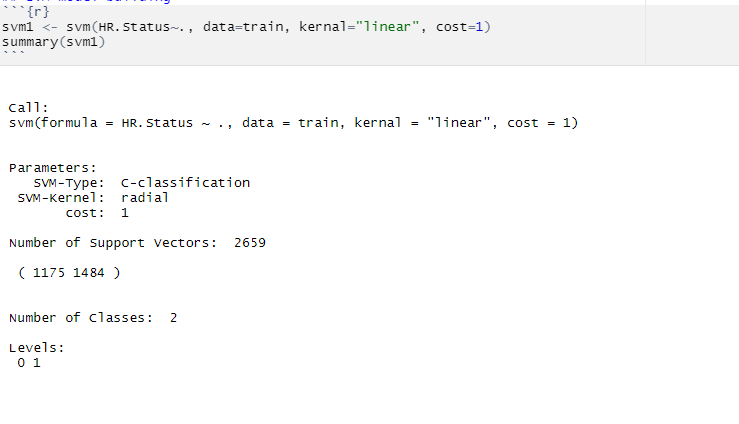
**SVM** is a supervised machine learning **algorithm** which can be **used** for classification or regression problems. It **uses** a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs.

# Advantages:

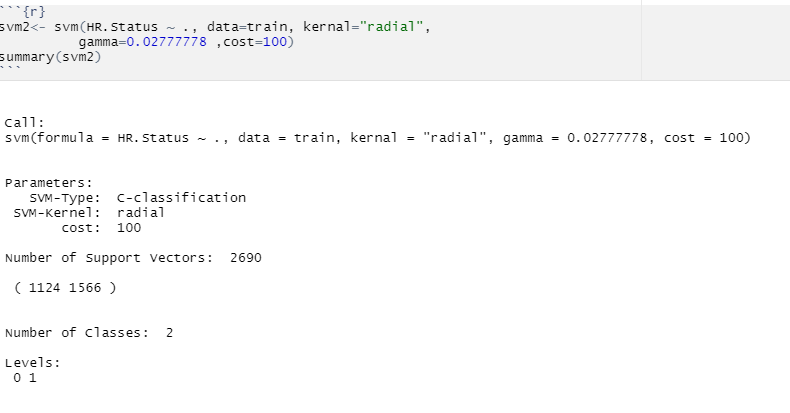
1. SVM works relatively well when there is a clear margin of separation between classes.
2. SVM is more effective in high dimensional spaces.
3. SVM is effective in cases where the number of dimensions is greater than the number of samples.
4. SVM is relatively memory efficient

**SVM model building**

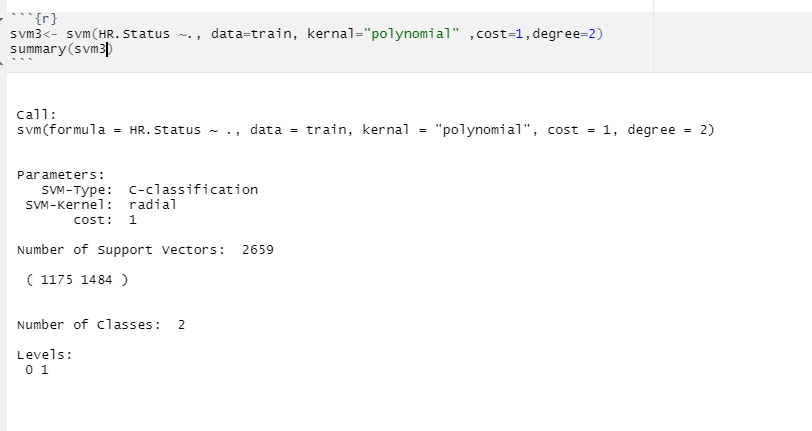
**Model 1 :**



**Model 2 :**



**Model 3 :**



**Decision Tree Algorithm :**

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving **regression and classification problems** too.

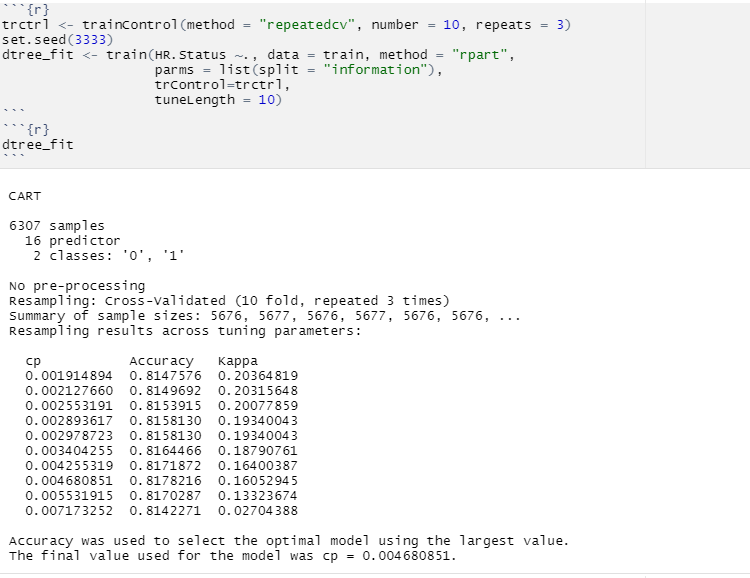
The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by **learning simple decision rules** inferred from prior data(training data).

In Decision Trees, for predicting a class label for a record we start from the **root** of the tree. We compare the values of the root attribute with the record’s attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

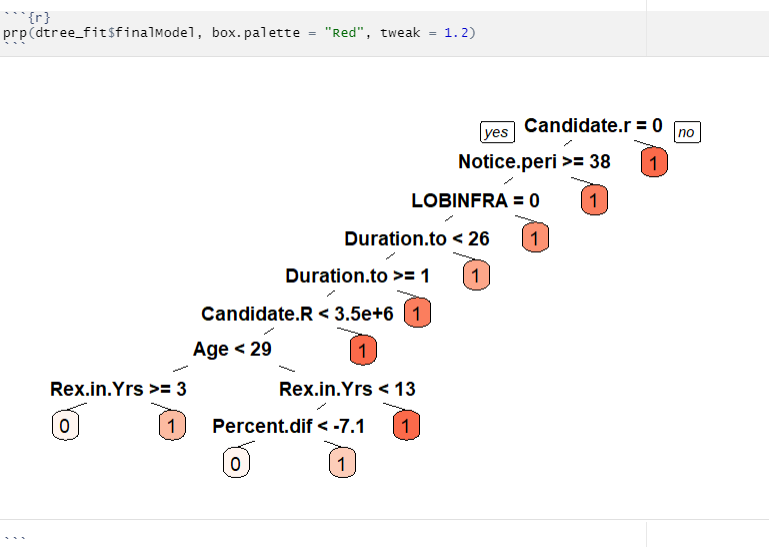
# ****Advantages:****

1. Compared to other algorithms decision trees requires less effort for data preparation during pre-processing.
2. A decision tree does not require normalization of data.
3. A decision tree does not require scaling of data as well.
4. Missing values in the data also do NOT affect the process of building a decision tree to any considerable extent.
5. A Decision tree model is very intuitive and easy to explain to technical teams as well as stakeholders.

**Decision Tree Model :**



**Decision Tree Plot :**



# Random Forests Algorithm :

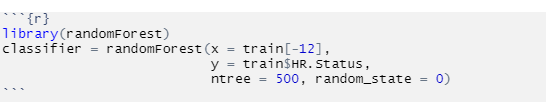
# Random forest is a tree-based algorithm which involves building several trees (decision trees), then combining their output to improve generalization ability of the model. ... Random Forest can be used to solve regression and classification problems. In regression problems, the dependent variable is continuous.

# But as stated, a random forest is a collection of decision trees. ... With that said, random forests are a strong modeling technique and much more robust than a single decision tree. They aggregate many decision trees to limit overfitting as well as error due to bias and therefore yield useful results.

**Advantages :**

1. Random Forest is based on the **bagging** algorithm and uses **Ensemble Learning** technique. It creates as many trees on the subset of the data and combines the output of all the trees. In this way it **reduces overfitting** problem in decision trees and also**reduces the variance**and therefore **improves the accuracy**.
2. Random Forest can be used to **solve both classification as well as regression problems**.
3. Random Forest works well with both **categorical and continuous variables**.
4. Random Forest can automatically **handle missing values**.
5. **No feature scaling required:** No feature scaling (standardization and normalization) required in case of Random Forest as it uses rule based approach instead of distance calculation.
6. **Handles non-linear parameters efficiently:** Non linear parameters don't affect the performance of a Random Forest unlike curve based algorithms. So, if there is high non-linearity between the independent variables, Random Forest may outperform as compared to other curve based algorithms.
7. Random Forest can automatically **handle missing values**.
8. Random Forest is usually **robust to outliers** and can handle them automatically.
9. Random Forest algorithm is very **stable**. Even if a new data point is introduced in the dataset, the overall algorithm is not affected much since the new data may impact one tree, but it is very hard for it to impact all the trees.
10. Random Forest is comparatively **less impacted by noise**.

**Random Forests Model :**



**Chapter 5: Performance Evaluation**

**Predicting on test data set :**

So let us understand in details how to build a Predictive model and know the most important algorithms to be learned in Predictive Analytics.

Predictive Analytics is a branch of advanced data analytics that involves the use of various techniques such as machine learning, statistical algorithms and other data mining techniques to forecast future events based on historical data.

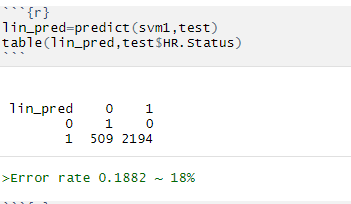
The model is then applied to current data to predict what would be the next course of action or suggestion for the outcome.

There are various algorithms available in the categories of data mining, machine learning and statistics when you assemble your predictive analysis model. As you explore the data it becomes easier to take further decision.

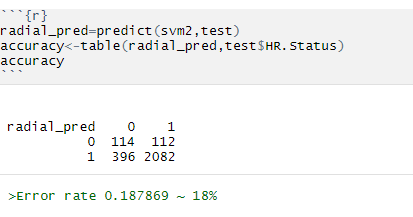
**Confusion matrix SVM:**

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing

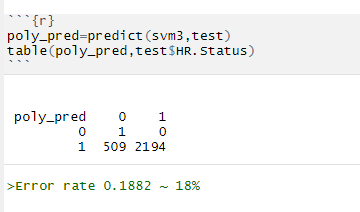
**Linear Error rate :**



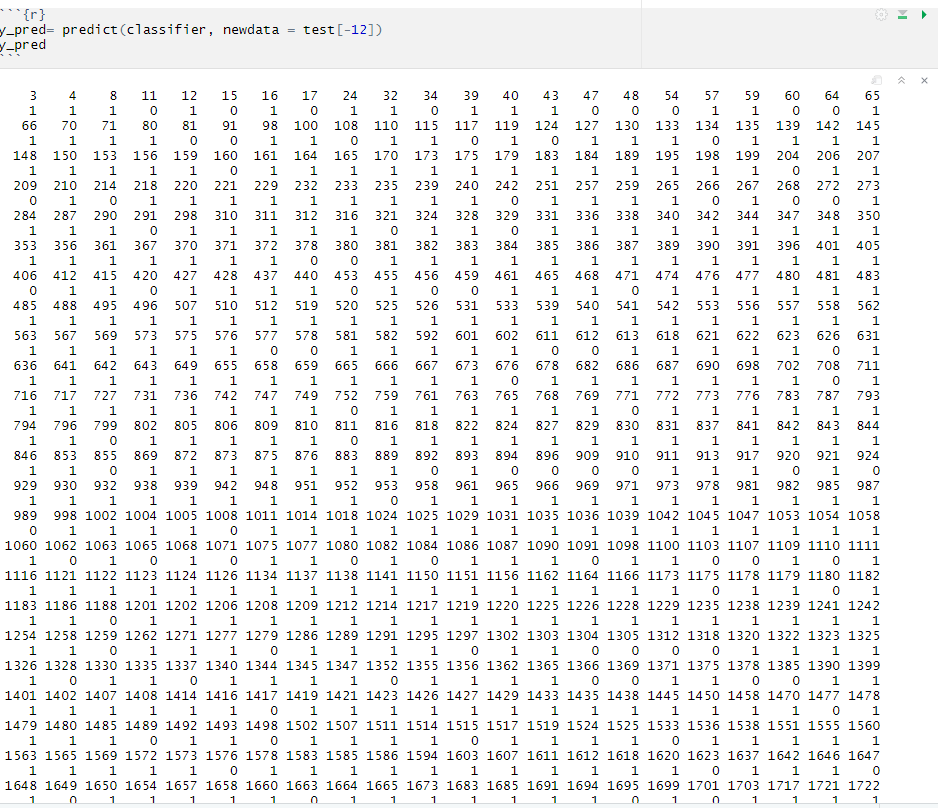
**Radial Error rate :**

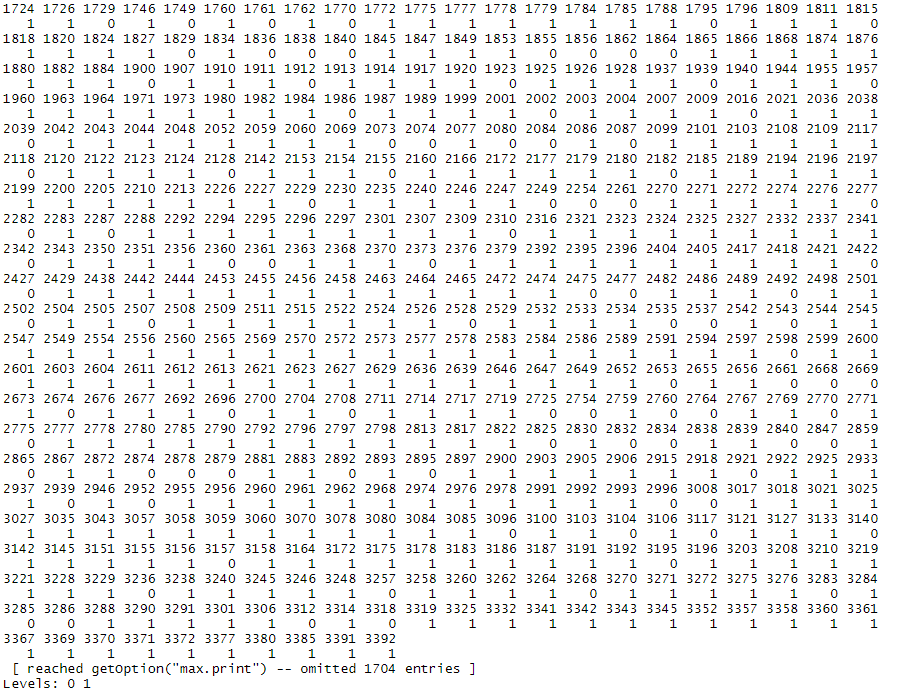


**Polynomial Error rate :**

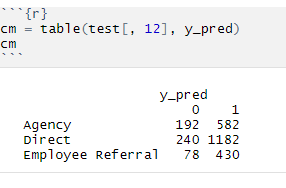


**Predict the model Random Forests :**



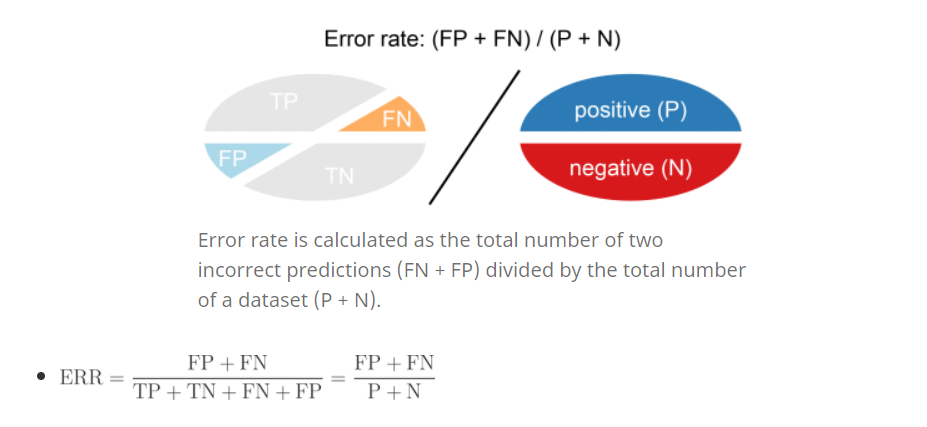


**Confusion matrix Random Forests:**



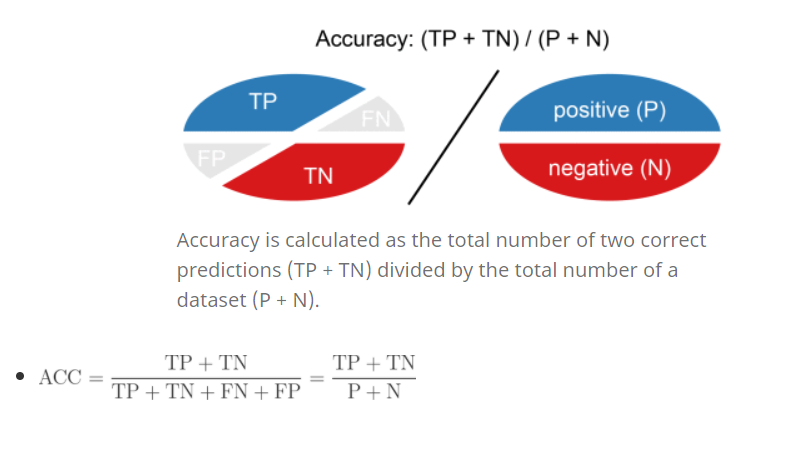
**Error rate :**

Error rate (ERR) is calculated as the number of all incorrect predictions divided by the total number of the dataset. The best error rate is 0.0, whereas the worst is 1.0.



#### **Accuracy :**

Accuracy (ACC) is calculated as the number of all correct predictions divided by the total number of the dataset. The best accuracy is 1.0, whereas the worst is 0.0. It can also be calculated by 1 – ERR.



**Chapter 6 : Conclusion**

## **Conclusion:**

The top 5 key drivers that influence the candidate's joining of a company are:

* Notice Period
* Candidate Source
* Line of Business
* Offered Band
* Age

The Support Vector Machine Model was built to predict the candidates probability of acceptance of an offer. The Code and analysis of the Model are available