DATA ATTRITION

TEAM MEMBERS

Nidhi Chovatiya
Course Section : A
Department : CS
Level : Graduate
CWID : 10457344

Janki Patel
Course Section : A
Department : CS
Level : Graduate
CWID : 10457365

Koral Kalathia
Course Section : A
Department : CS
Level : Graduate
CWID : 10457952

PROBLEM STATEMENT

- The **model** works by clustering/ classifying employee profiles based on various attributes such as age, sex, marital status, education level, work experience, group, etc.
- Attrition is a common issue that every company has to deal with. The goal of the data analytics project is to build a model that can help the company to predict whether or not a certain employee will leave as well as identify important factors of leave. The information can be vital in future recruitment and reduction in employee attrition.
- Through this kind of analysis, we can understand how many employees are likely to leave, while also determining which employees are at the highest risk and for what reasons.

DATASET DESCRIPTION

- This dataset contains all information about employees like ethnicity, annual and hourly rate, hiring month as well as termination year, also contain in which group they are working.
- Total number of Rows: 9612
- Total number of Columns : 27

```
> colnames(deleted_data)
 [1] "EMP_ID"                 "ANNUAL_RATE"
 [3] "HRLY_RATE"              "JOBCODE"
 [5] "ETHNICITY"              "SEX"
 [7] "MARITAL_STATUS"         "JOB_SATISFACTION"
 [9] "AGE"                    "NUMBER_OF_TEAM_CHANGED"
[11] "REFERRAL_SOURCE"        "HIRE_MONTH"
[13] "REHIRE"                 "TERMINATION_YEAR"
[15] "IS_FIRST_JOB"           "TRAVELLED_REQUIRED"
[17] "PERFORMANCE_RATING"     "DISABLED_EMP"
[19] "DISABLED_VET"           "EDUCATION_LEVEL"
[21] "STATUS"                 "JOB_GROUP"
[23] "PREVYR_1"               "PREVYR_2"
[25] "PREVYR_3"               "PREVYR_4"
[27] "PREVYR_5"
```

EXPLORATORY DATA ANALYSIS

- Total number or rows

  after omitting rows:

```
> deleted_data <- na.omit(file)
> nrow(deleted_data)
[1] 4218
```

- Summary of Status :

```
> summary(file$STATUS)
   A    T
5394 4218
```

EXPLORATORY DATA ANALYSIS

```
ANNUAL_RATE        HRLY_RATE         ETHNICITY    SEX       MARITAL_STATUS  JOB_SATISFACTION      AGE
 Min.   :0.00000   Min.   :0.00000   WHITE  :5820   F:5723   Divorced:1605   Min.   :0.0000  Min.   :0.0000
 1st Qu.:0.02761   1st Qu.:0.03030   ASIAN  :1389   M:3889   Married :4027   1st Qu.:0.2500  1st Qu.:0.2174
 Median :0.04653   Median :0.04882   BLACK  :1106            Single  :3980   Median :0.5000  Median :0.4565
 Mean   :0.05883   Mean   :0.06053   HISPA  :1067                            Mean   :0.4394  Mean   :0.4816
 3rd Qu.:0.07457   3rd Qu.:0.07576   TWO    : 176                            3rd Qu.:0.7500  3rd Qu.:0.7391
 Max.   :1.00000   Max.   :1.00000   PACIF  :  32                            Max.   :1.0000  Max.   :1.0000
                                     (Other):  22

REHIRE         TRAVELLED_REQUIRED PERFORMANCE_RATING  DISABLED_EMP   DISABLED_VET    EDUCATION_LEVEL  STATUS
 Mode :logical   Mode :logical      Min.   :0.0000    Mode :logical  Mode :logical   LEVEL 1:3398     A:5394
 FALSE:8726      FALSE:7781         1st Qu.:0.2500    FALSE:8665     FALSE:8682      LEVEL 2:1739     T:4218
 TRUE :886       TRUE :1831         Median :0.5000    TRUE :947      TRUE :930       LEVEL 3:2710
                                    Mean   :0.5005                                  LEVEL 4:1112
                                    3rd Qu.:0.7500                                 LEVEL 5: 653
                                    Max.   :1.0000

JOB_GROUP                    PREVYR_1         PREVYR_2         PREVYR_3         PREVYR_4          PREVYR_5
 Production & Operations:1714  Min.   :0.0000  Min.   :0.0000  Min.   :0.0000  Min.   :0.0000  Min.   :0.00000
 Marketing - Direct     : 849  1st Qu.:0.0000  1st Qu.:0.0000  1st Qu.:0.0000  1st Qu.:0.0000  1st Qu.:0.00000
 Physical Flows         : 816  Median :0.4000  Median :0.0000  Median :0.0000  Median :0.0000  Median :0.00000
 Finance                : 525  Mean   :0.2692  Mean   :0.2045  Mean   :0.1551  Mean   :0.1235  Mean   :0.09324
 Human Resources        : 396  3rd Qu.:0.6000  3rd Qu.:0.4000  3rd Qu.:0.4000  3rd Qu.:0.2000  3rd Qu.:0.00000
 Customer Care          : 355  Max.   :1.0000  Max.   :1.0000  Max.   :1.0000  Max.   :1.0000  Max.   :1.0000
 (Other)                :4957
```

Summary of Dataset

MODELS WE USED

- Naïve Bayes
- KNN
- SVM
- Neural Network
- Dtree

NAÏVE BAYES

- **Supervised Machine** Learning algorithm
- Based on the **Bayes Theorem** that is used to solve classification problems by following a **probabilistic approach**.
- Based on the idea that the **predictor variables** in a Machine Learning model are independent of each other.
- Outcome of a model depends on a set of independent variables that have nothing to do with each other.

```
> print(paste("Total bad Predictions:" , wrongprediction))
[1] "Total bad Predictions: 937"
> print(paste("Error rate :" , wrongpredictionrate))
[1] "Error rate : 0.390091590341382"
> print(paste("Accuracy :" , 100-(wrongpredictionrate*100)))
[1] "Accuracy : 60.9908409658618"
```

NAÏVE BAYES ACCURACY AND ERROR RATE

KNN

- KNN is a **Supervised Learning** algorithm.
- Uses **labelled** input data set to predict the output of the **data points**.
- **Most simple** Machine learning algorithm
- **Easily implemented** for a varied set of problems.
- Based on **feature similarity**.
- Checks **similarity** of a data point between its **neighbour**
- Classifies data point into most **similar classes**.

KNN ACCURACY & ERROR RATE

K=3

```
> error<-sum(newk3 != train_data$STATUS) #error
> error_rate <- sum(newk3 != train_data$STATUS)/length(newk3 != train_data$ST
ATUS)
> error_rate
[1] 0.4418863
> accuracy<-100-(error_rate*100)
> accuracy
[1] 55.81137
```

K=5

```
> error<-sum(newk5 != train_data$STATUS) #error
> error_rate <- sum(newk5 != train_data$STATUS)/length(newk5 != train_data$ST
ATUS)
> error_rate
[1] 0.4335645
> accuracy<-100-(error_rate*100)
> accuracy
[1] 56.64355
```

K=10

```
> error<-sum(newk10 != train_data$STATUS) #error
> error_rate <- sum(newk10 != train_data$STATUS)/length(newk10 != train_data
$STATUS)
> error_rate
[1] 0.4171983
> accuracy<-100-(error_rate*100)
> accuracy
[1] 58.28017
```

SVM

- SVM (Support Vector Machine) is a supervised machine learning algorithm.
- Used to classify data into different classes
- Used to generate multiple separating hyperplanes.
- Data is divided into segments
- Each segment contains only one kind of data.

```
> SVM_wrong<- (test_data$STATUS!=svm.pred)
> error_rate<-sum(SVM_wrong)/length(SVM_wrong)
> error_rate
[1] 0.3134888
> accuracy<-100-(error_rate*100)
> accuracy
[1] 68.65112
```

SVM ACCURACY & ERROR RATE

NEURAL NETWORK

Neural Networks and Data Mining.

Artificial Neural Network, often just called a neural network.

Mathematical model inspired by biological neural networks.

Consists of an interconnected group of artificial neurons, and it

Processes information using a connectionist approach to computation.

```
> error_rate
[1] 0.2872606
> accuracy<-100-(error_rate*100)
> accuracy
[1] 71.27394
```

NEURAL NETWORK ACCURACY & ERROR RATE

DTREE

Decision **tree** is a graph to represent choices and their results in form of a **tree**.

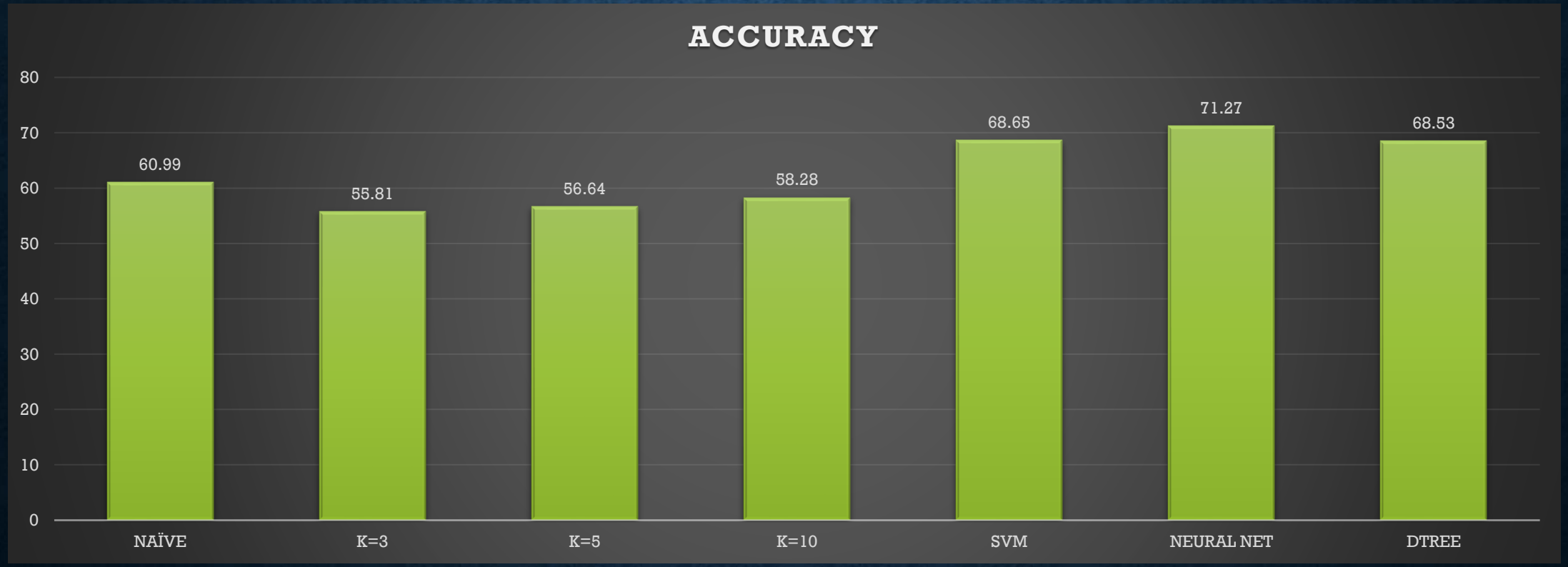
The decision tree classifier is a supervised learning algorithm which can use for both the classification and regression tasks.

The nodes in the graph represent an event or choice and the edges of the graph represent the decision rules or conditions.

```
> error<-sum(preds != test_data$STATUS) #error
> error_rate <- sum(preds != test_data$STATUS)/length(preds != test_data$STATUS)
> error_rate
[1] 0.3146067
> accuracy<-100-(error_rate*100)
> accuracy
[1] 68.53933
```

DTREE ERROR RATE AND ACCURACY

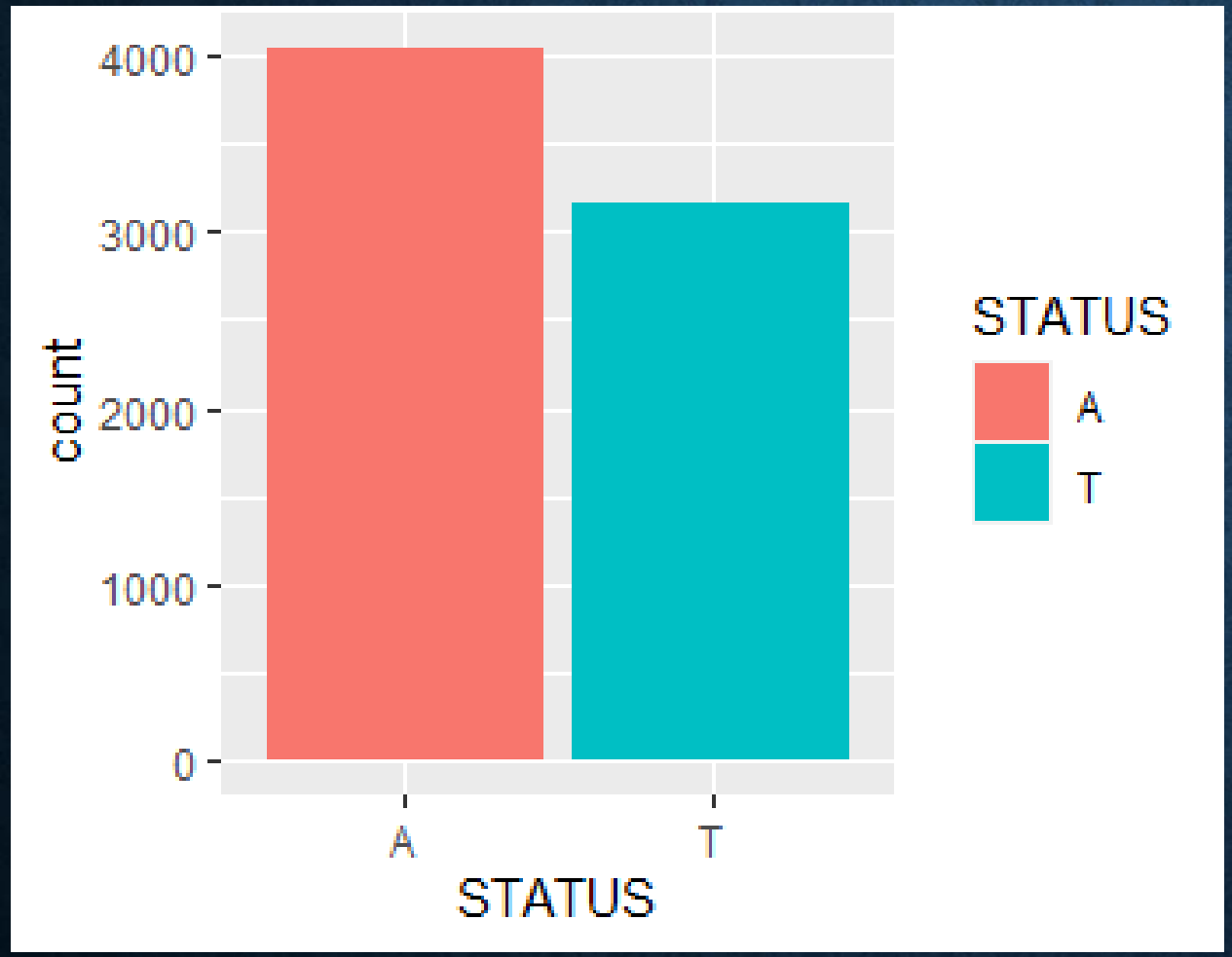
COMPARISON


ACCURACY
80
70
60
50
40
30
20
10
0
60.99
55.81
56.64
58.28
68.65
71.27
68.53
NAÏVE
K=3
K=5
K=10
SVM
NEURAL NET
DTREE
KNN

COMPARISON

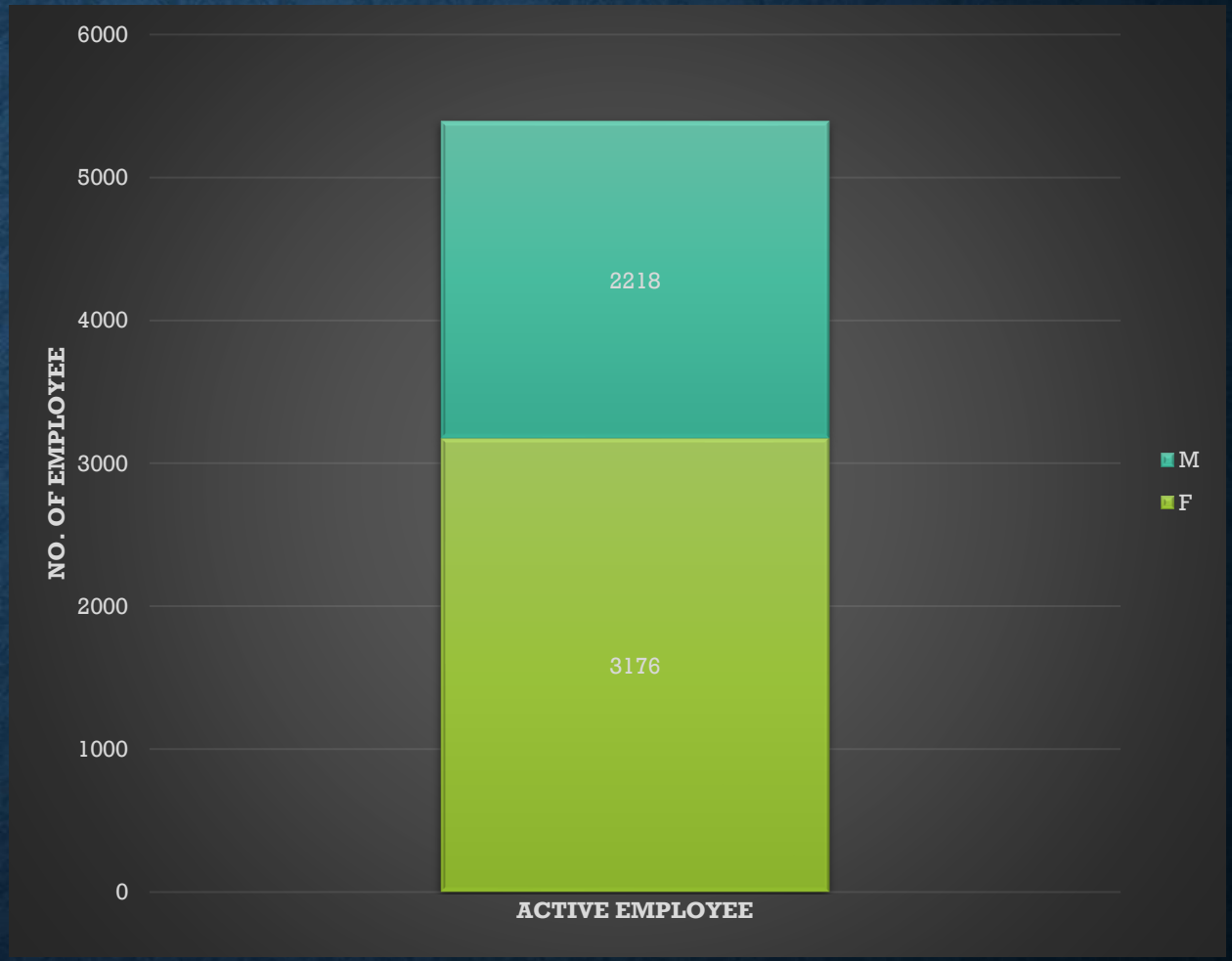

ERROR RATE
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.39
0.441
0.433
0.417
0.313
0.287
0.314
NAÏVE
K=3
K=5
K=10
SVM
NEURAL NET
DTREE
KNN

VISUALIZATION

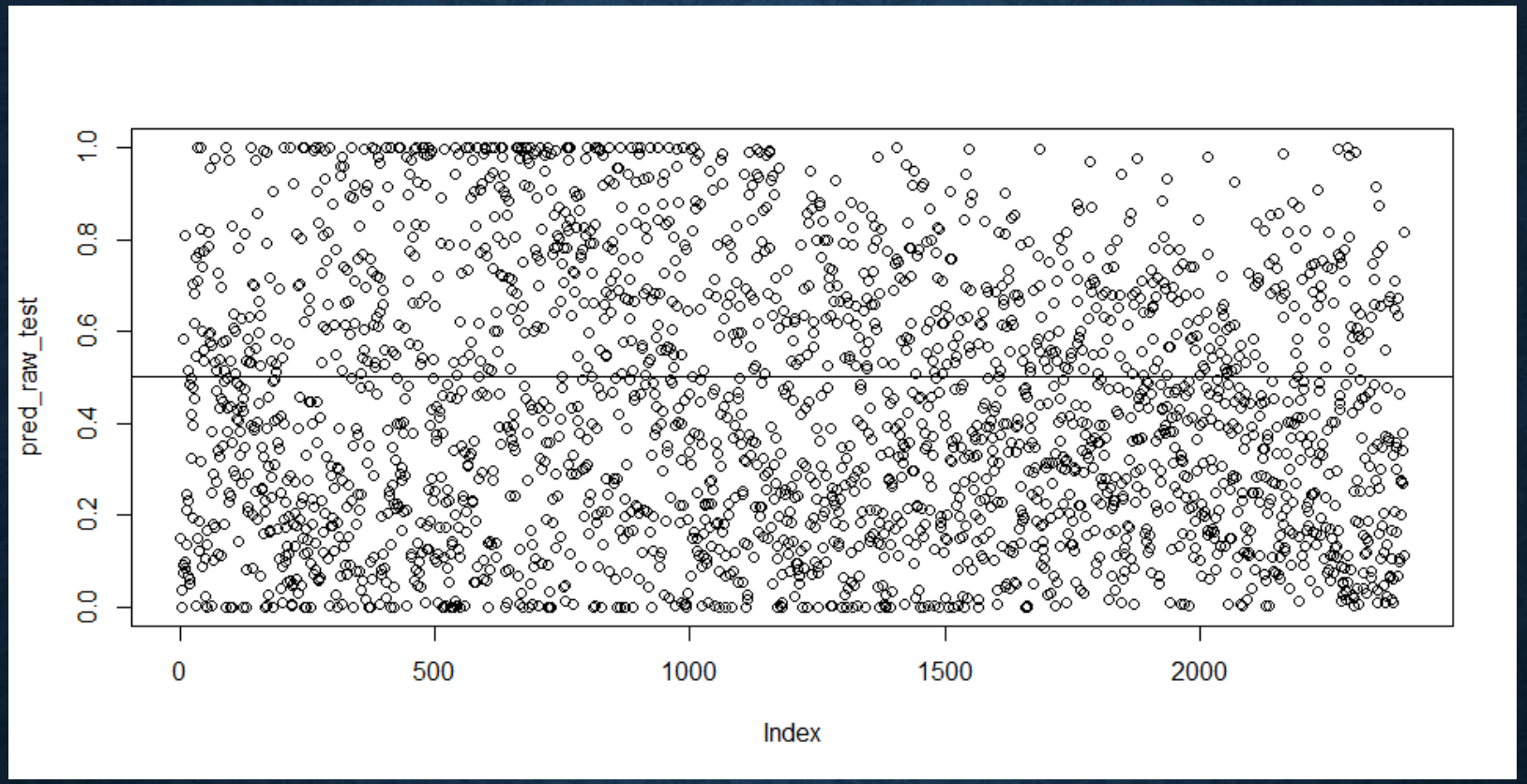


STATUS OF EMPLOYEE

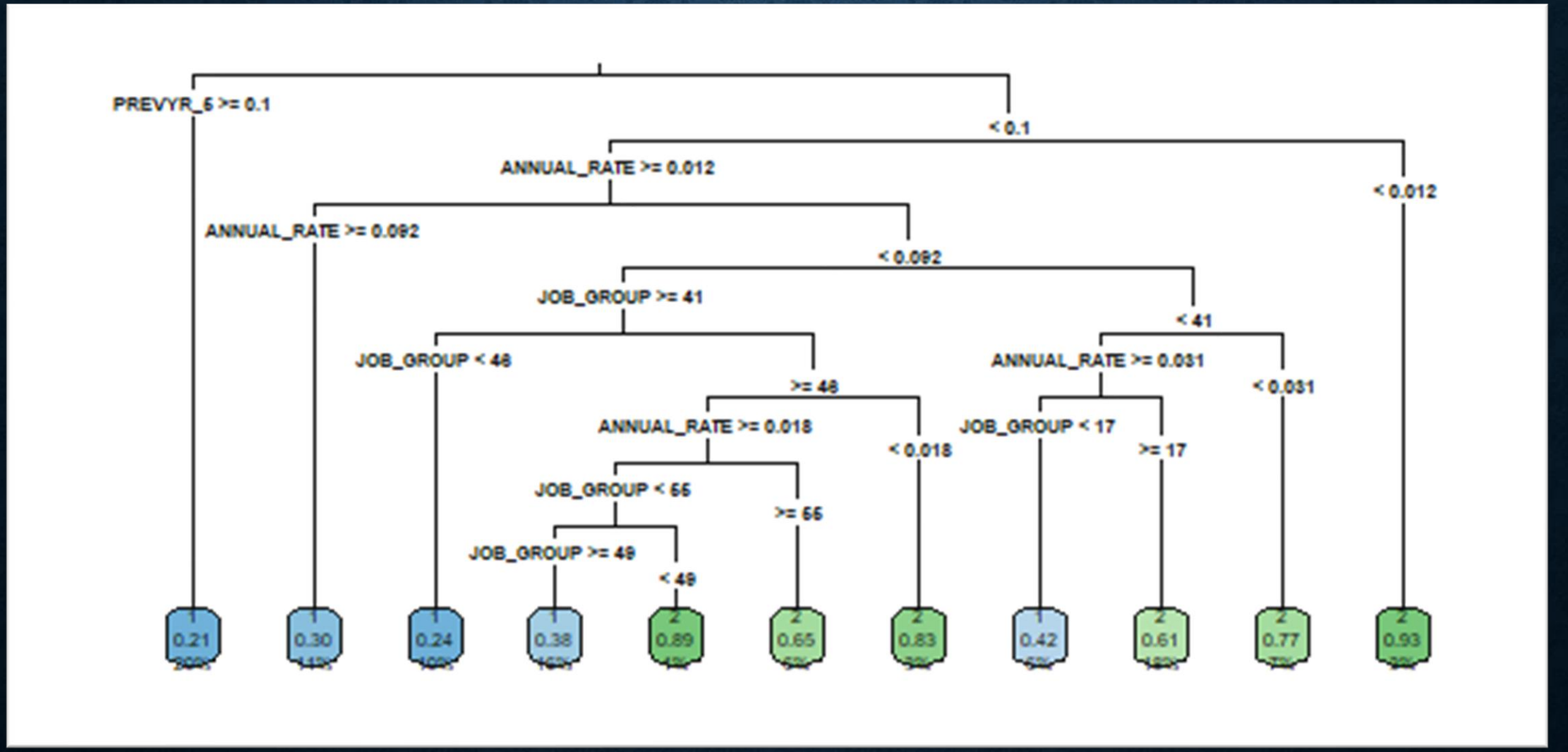
NO. OF ACTIVE EMPLOYEE



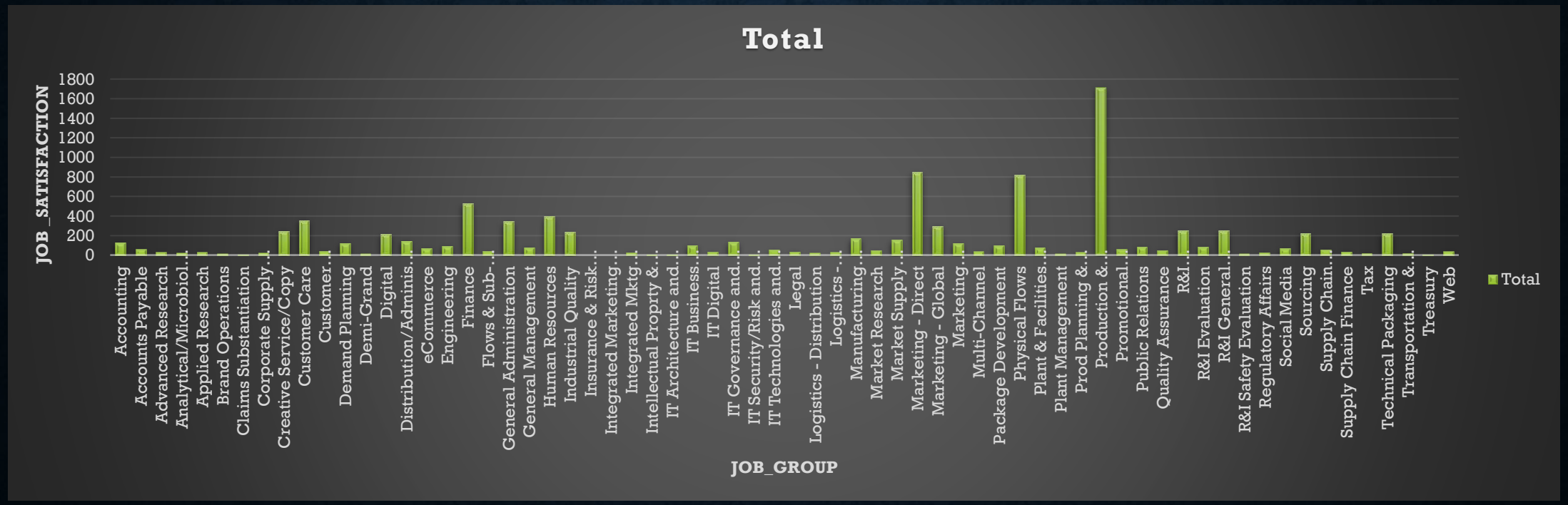
TERMINATION YEAR VS NO. OF EMPLOYEE



ASSIGNING PROBABILITY AND SETTING A MANUAL CUT-OFF OF P=0.5


PREVYR_6 >= 0.1
< 0.1
ANNUAL_RATE >= 0.012
< 0.012
ANNUAL_RATE >= 0.092
< 0.092
JOB_GROUP >= 41
< 41
JOB_GROUP < 48
ANNUAL_RATE >= 0.031
>= 48
< 0.031
ANNUAL_RATE >= 0.018
JOB_GROUP < 17
< 0.018
>= 17
JOB_GROUP < 55
>= 55
JOB_GROUP >= 49
< 49
0.21
0.30
0.24
0.38
2
0.89
2
0.65
2
0.83
0.42
2
0.61
2
0.77
2
0.93

JOB_GROUP VS JOB_SATISFACTION



FUTURE SCOPE

This analytics helps human resources to interpret data, find out the trends & help take required steps to keep the organization running smoothly & profitably.

Helps in future recruitment

Helps the companies to understand which group need more focus and employee

CONCLUSION

- Neural Net gives best accuracy rate of 71.27%
- Help in making decisions to the company
- Help to know number of active and terminated employee
- Help to predict the hiring rate
- Task management.
- Helps in future recruitment

THANK YOU