THE POTENTIAL OF EMPLOYEES LEAVING THE COMPANY (BECOME TERMINATED) MODEL

TITANS

Devila Bakrania
CWID : 10457590
Course Section: A
Department: CS
Level: Graduate

Syed Mahvish
CWID : 10456845
Course Section: A
Department: CS
Level: Graduate

Priya Gupta
CWID : 10457442
Course Section: A
Department: CS
Level: Graduate

PROBLEM STATEMENT

- To develop a classification model(s) to predict the potential of employees leaving the company (become terminated)

DATA SET EXPLANATION

- The dataset attrition_data.csv has 9613 rows and 27 columns
- The dataset contains data of Active and terminated employees of the company
- The Task is to predict the potential of employees leaving the company (become terminated)
- Target: Status - Active or Terminated
- Among them training data is: 6728 and testing data is: 2884

GOALS AND OBJECTIVES

- Major reason of termination is performance rating and Job satisfaction or not?
- Which year has high amount of termination?
- Which department has high termination rate?
- Termination pattern in terms of education
- Frequency of rehired employees being terminated
- Does Highly paid employees were terminated?
- Ratio of disable/vet vs normal in term of termination

ALGORITHMS

- K-Nearest Neighbor (KNN)
- Classification and Regression Trees (CART)
- C5.0 Methodology
- K-Mean Cluster
- Artificial Neural Networks (ANN)

K NEAREST NEIGHBOR (KNN)

CASE STUDY RESULTS ON BASIS OF SALARY

K value	Error rate	Accuracy
5	0.4472954	55.27046
10	0.4403606	55.96394
20	0.4226768	57.73232
50	0.4133148	58.66852
100	0.4223301	57.76699
150	0.4188627	58.11373
200	0.4147018	58.52982
300	0.417129	58.2871
500	0.4199029	58.00971
1000	0.4202497	57.97503

From above table it is clear that the accuracy lies between 55% to 58% for salary-based classification

K NEAREST NEIGHBOR (KNN)

CASE STUDY RESULTS BASED ON JOB-SATISFACTION

K value	Error rate	Accuracy
5	0.4282247	57.17753
10	0.4212899	57.87101
20	0.4088072	59.11928
50	0.4056865	59.43135
100	0.4095007	59.04993
150	0.3966713	60.33287
200	0.3987517	60.12483
300	0.398405	60.1595
500	0.398405	60.1595
1000	0.4053398	59.46602

From above table, it is clear that the accuracy lies between 57% to 60% for Performance and job-satisfaction based classification

K NEAREST NEIGHBOR (KNN)

CASE STUDY RESULTS BASED ON ALL DATA

K value	Error rate	Accuracy
5	0.4119279	58.80721
10	0.4084605	59.15395
20	0.4001387	59.98613
50	0.397018	60.2982
100	0.398405	60.1595
150	0.3987517	60.12483
200	0.3914702	60.85298
300	0.389043	61.0957
500	0.3883495	61.16505
1000	0.3942441	60.57559

K NEAREST NEIGHBOR (KNN) CONCLUSION

Following conclusion are drawn from KNN classification on given set of employment data:

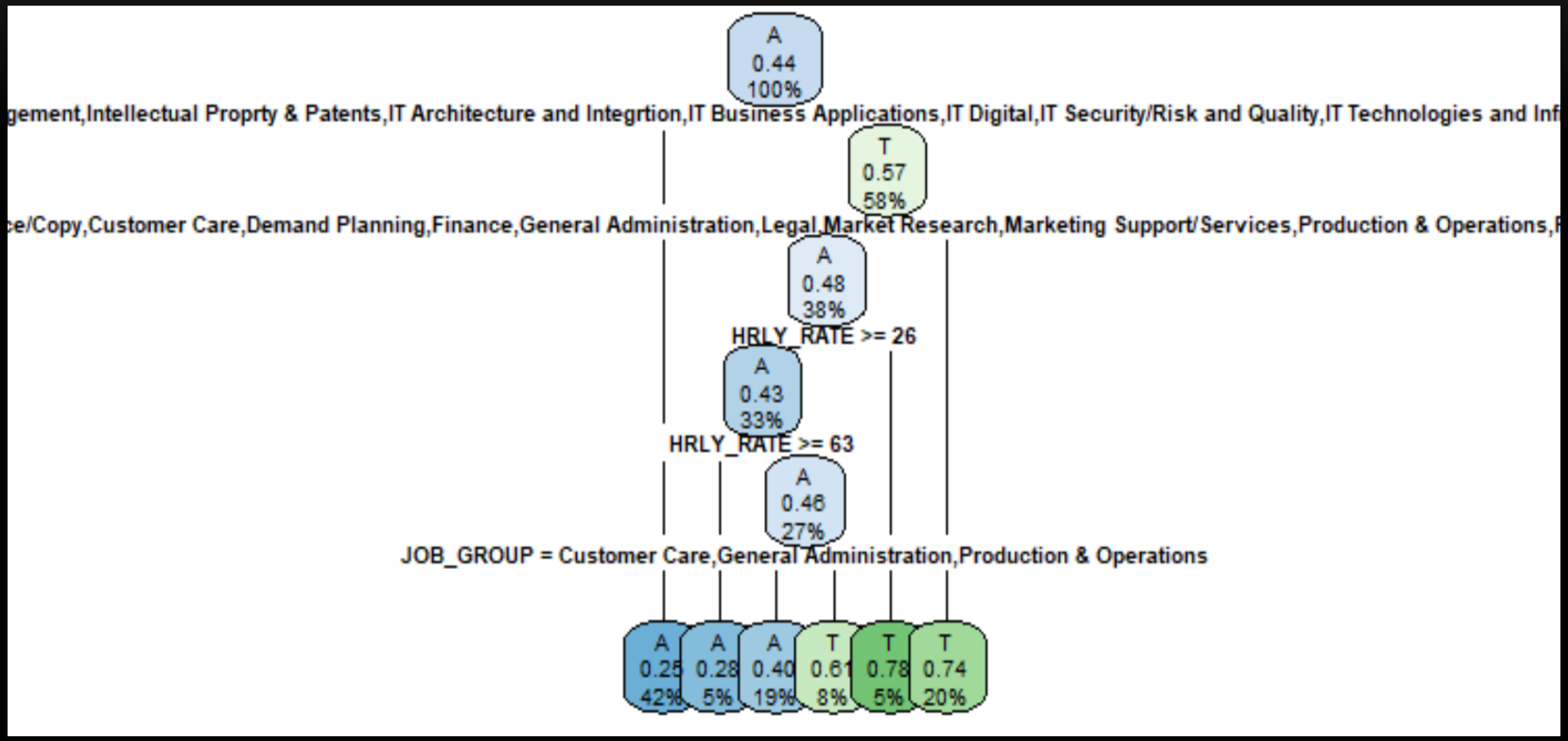
- 1- As value of K increases, accuracy increases in all three cases
- 2- The accuracy increases when multiple columns are considered
- As in case of salary, lowest accuracy is observed while in case of all data highest accuracy is observed

Hence from above observation, it can be concluded that termination of employee is based on all factor (majorly), and not limited to few factors such as salary or performance or job satisfication

CLASSIFICATION AND REGRESSION TREES (CART)

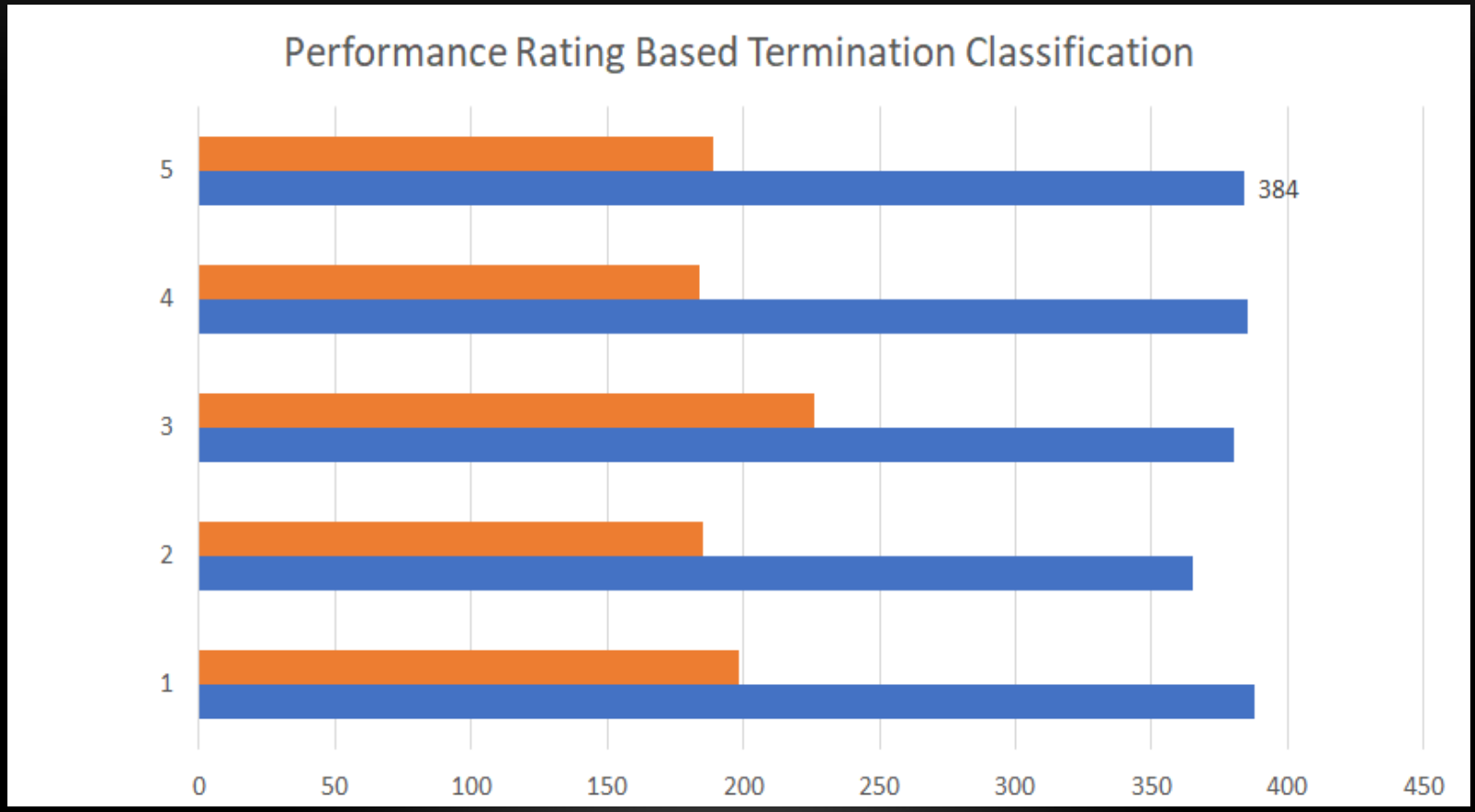
- The decision tree method is a powerful and popular predictive machine learning technique that is used for both classification and regression. So, it is also known as Classification and Regression Trees (CART)
- The algorithm of decision tree models works by repeatedly partitioning the data into multiple sub-spaces, so that the outcomes in each final sub-space is as homogeneous as possible. This approach is technically called recursive partitioning

CART PLOT

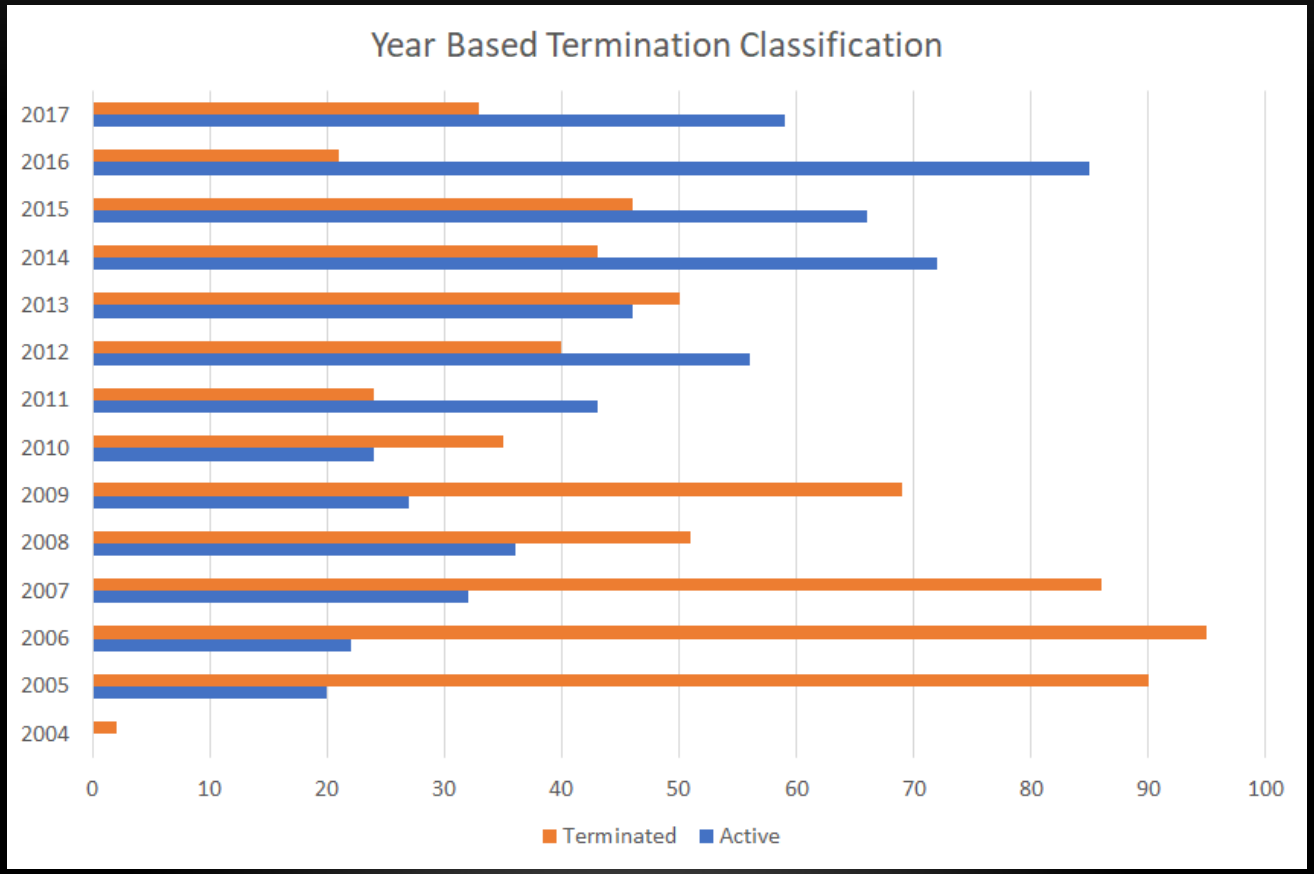


Data Set classified on the basis of Hourly Rate >= 26 and Hourly Rate >= 63

PERFORMANCE RATING BASED CLASSIFICATION (CART)



YEARLY CLASSIFICATION USING CART



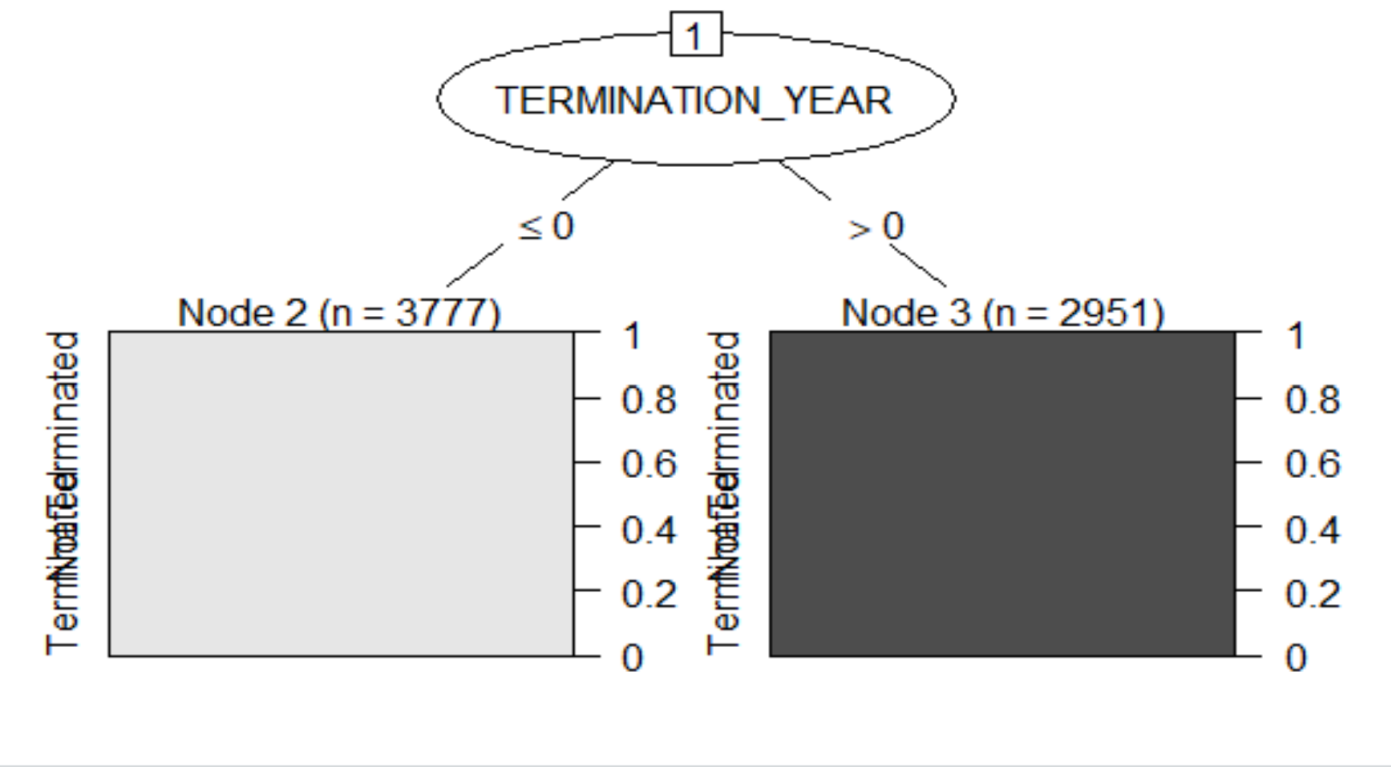
CART CONCLUSION

- **Classification is done basis of :**
  - -Performance Rating
  - -Department
  - -Education
  - -Year
- **Error Rate: .3068655**
- **Accuracy: 70%**

C5.0

- The C5.0 algorithm has become the industry standard for producing decision trees, because it does well for most types of problems directly out of the box. Compared to more advanced and sophisticated machine learning models (e.g. Neural Networks and Support Vector Machines), the decision trees under the C5.0 algorithm generally perform nearly as well but are much easier to understand and deploy

C5.0 PLOT



Data Set classified on the basis of Termination Year

STATUS BASED C5.0 CLASSIFICATION

```
Decision tree:

TERMINATION_YEAR <= 0: Not Terminated (3777)
TERMINATION_YEAR > 0: Terminated (2951)


Evaluation on training data (6728 cases):

           Decision Tree
         ----------------
         Size      Errors

            2    0( 0.0%)   <<


          (a)   (b)    <-classified as
         ----  ----
         3777          (a): class Not Terminated
               2951    (b): class Terminated


        Attribute usage:

        100.00% TERMINATION_YEAR


Time: 0.0 secs
```

C5.0 CONCLUSION

- After taking 12 columns into consideration, algorithm C5.0 is giving Decision Tree of size 2 and it's Accuracy is 100%

K-MEAN CLUSTER
FOR REHIRE

- Clustering is done for employee rehired or not. That creates two clusters. Cluster of rehired and not rehired are tabulate against Status column i.e. terminated or not terminated

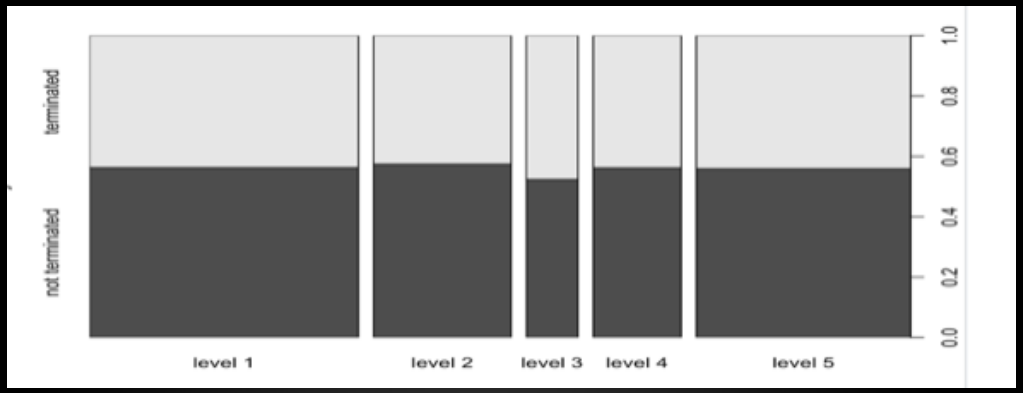
	not terminated	terminated
not Rehire	4886	3840
Rehire	508	378

K-MEAN CLUSTER

FOR EDUCATION LEVEL

- Clustering is done for employee education level. That creates five clusters. Cluster of education level are tabulate against Status column i.e. terminated or not terminated

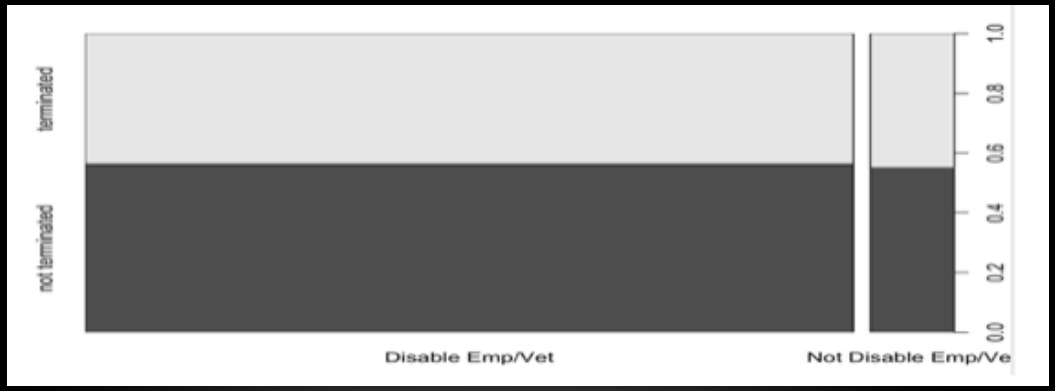
	not terminated	terminated
level 1	1913	1485
level 2	999	740
level 3	342	311
level 4	625	487
level 5	1515	1195



K-MEAN CLUSTER
FOR DISABLE EMPLOYEE/VETERAN

- Clustering is done for disable employee and vet. That creates two clusters. Cluster of disable employee and vet are tabulate against Status column i.e. terminated or not terminated

	not terminated	terminated
Disable Emp/Vet	4874	3791
Not Disable Emp/Vet	520	427

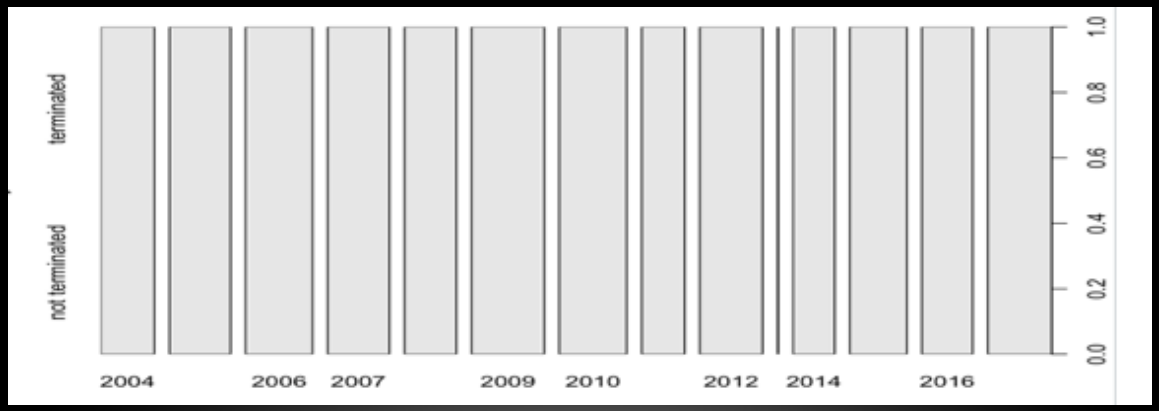


K-MEAN CLUSTER

FOR YEAR OF TERMINATION

- Clustering is done for Year of termination. That creates 14 clusters. Cluster of Year of termination are tabulate against Status column i.e. terminated or not terminated

	not terminated	terminated
2004	0	299
2005	0	341
2006	0	374
2007	0	346
2008	0	290
2009	0	406
2010	0	377
2011	0	240
2012	0	350
2013	0	3
2014	0	232
2015	0	319
2016	0	284
2017	0	357



K-MEAN CLUSTER

FOR CLUSTER OF ALL DATA

- Clustering is done for all factors. We creates 10 clusters. Cluster are tabulate against Status column i.e. terminated or not terminated

	not terminated	terminated
1	945	795
2	665	560
3	9	4
4	1004	717
5	573	244
6	351	98
7	479	208
8	102	26
9	644	816
10	622	750

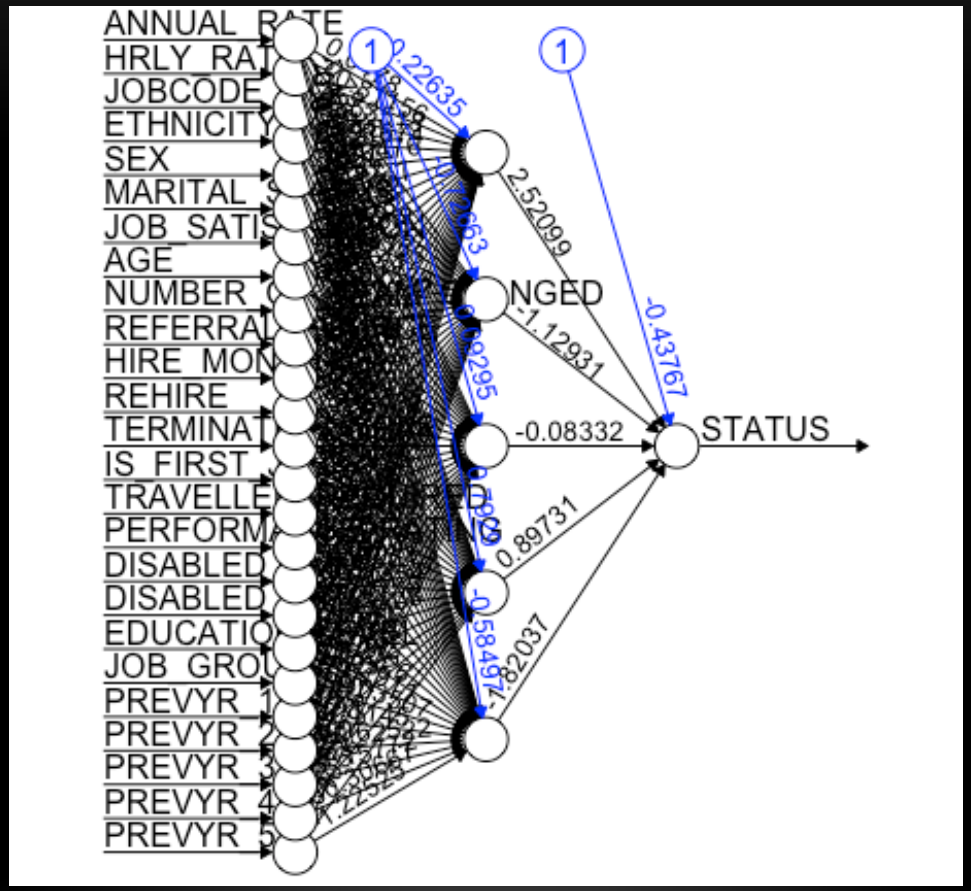
K-MEAN CLUSTER CONCLUSION

- Following conclusion are drawn from K-mean cluster on given set of employment data:
  - Get a meaningful intuition of the structure of the data we're dealing with
  - Cluster-then-predict where different models will be built for different subgroups if we believe there is a wide variation in the behaviors of different subgroups

ARTIFICIAL NEURAL NETWORKS(ANN)

- a system of neurodes (*nodes*) and *weighted connections* (synapses) inside the memory of a computer
- nodes are data storage locations (like variables in a program, cells in a spreadsheet)
- nodes are arranged in *layers* with weighted connections running between layers
- *balls* represent nodes and *lines* represent connection weights
- *input* layer nodes (Considered all columns) receive the data
- *output* layer nodes (Target Node : Status) relay the response of the neural network out of the net
- *hidden* layer nodes (5 columns considered hidden from the outside world) conduct the internal processing
- data are fed into the net through the input nodes
- data are processed internally by hidden nodes, based on the inter-node connection weights
- result are passed on to the outside world by output nodes
- "learning" takes place through adjusting connection weights
- a "learned" neural network has adjusted its weights properly

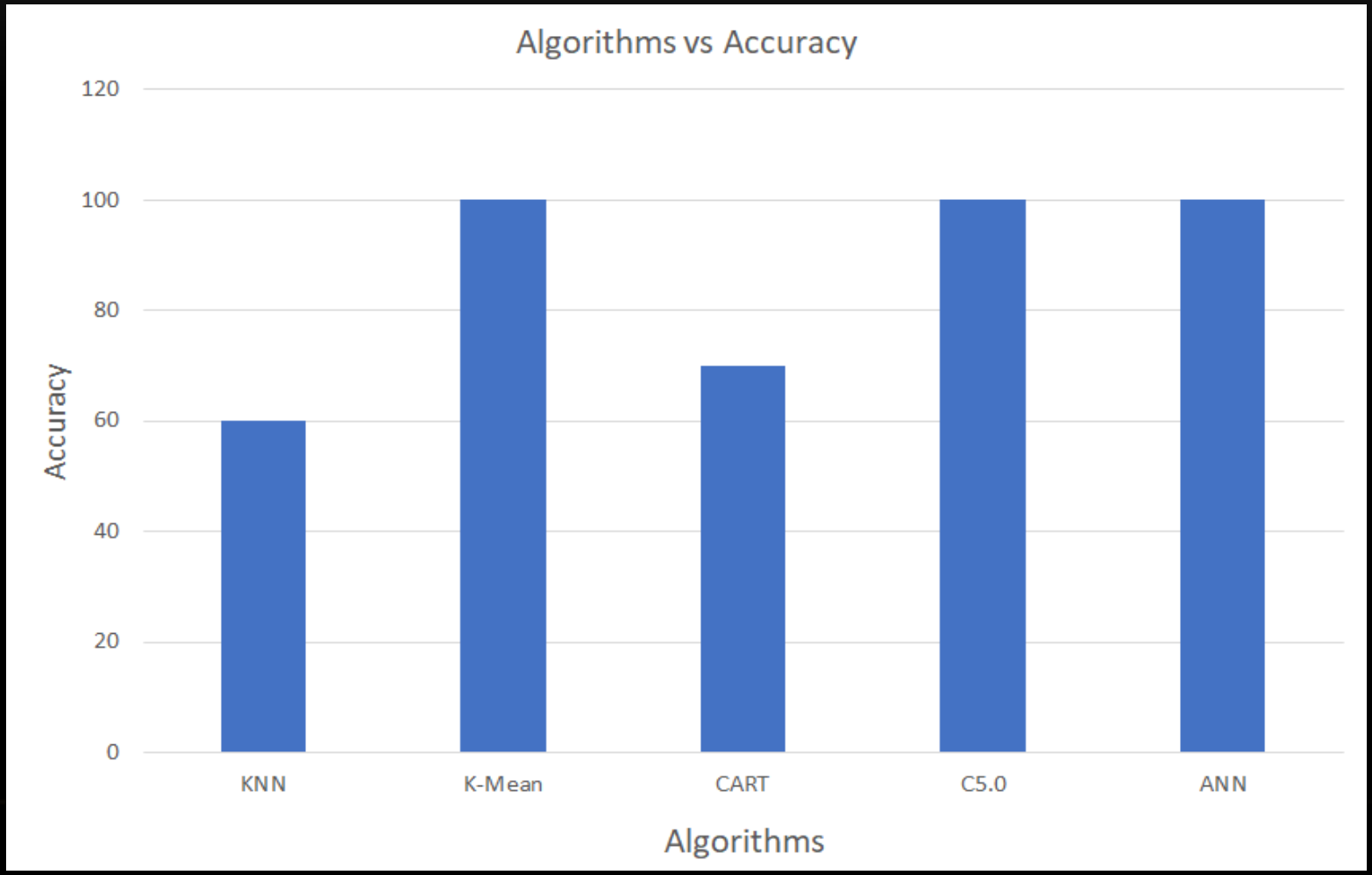
APPLICATION OF ARTIFICIAL NEURAL NETWORKS



ARTIFICIAL NEURAL NETWORKS STATUS OUTPUT

A	T
5394	4218

COMPARISON OF CLASSIFICATION ALGORITHM



CONCLUSION

- We have explored different prediction models. By measuring the performance of the models using real data, we have seen interesting results on the potential of employees leaving the company (become terminated)
- As per comparison between algorithms, ANN has 0 error rate and the Maximum Accuracy of 100% for most of the attributes in predicting the potential of employees leaving the company(become terminated) data

FUTURE SCOPE

- Employee Efficiency and Transparent appraisal decision related data can lead to again more accuracy to predicate the potential of employees leaving the company (become terminated)

THANK YOU!