**Problem Statement** : To develop a classification model(s) to predict the potential of employees leaving the company (become terminated)

**Following Algorithms are used:**

KNN K-Nearest Neighbor (KNN): K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions).

A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

Slide no 6:

The KNN is used to develop a classification model to predict the potential of employee leaving the company(become terminated).

The case study is performed in three parts :

1- Classification is done based on salary : Because in most of cases, salary plays important role for leaving company or being terminated. Mostly in case of layoffs, its being notice employee with high salaries are laid off.While performing classification on basis of Salary following columns are considered : ANNUAL\_RATE , HRLY\_RAT , Status (target value). The accuracy lies between 55% to 58% for salary-based classification

Slide no 7 :

Classification is done based on Performance and job satisfaction :

The performance of employee plays a vital role while terminating from company. Most termination is based on low performance.

Also low performance of employee may be directly or indirectly related to job satisfaction.

While performing classification on basis of Performance and job satisfaction following columns are considered :JOB\_SATISFACTION ,PERFORMANCE\_RATING ,STATUS ,PREVYR\_1 ,PREVYR\_2 ,PREVYR\_3,PREVYR\_4 , PREVYR\_5. The accuracy lies between 57% to 60% for Performance and job-satisfaction based classification

Slide no 8:

Classification is done based on all data:

Now all data is considered to perform classification of employee termination.

The Accuracy lies between 58% to 61% for all data

Slide no 9:

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 satisfaction

Slide no 10:

 Classification and Regression Trees (CART):

Classification has been done multiple columns like:  Performance Rating ,Department  ,Education ,year

Slide no 12:

As can see below, on basis of performance rating, Active and terminated employee are classified.

|  |  |  |
| --- | --- | --- |
| Performance  Rating | Active | Terminated |
| 1 | 388 | 198 |
| 2 | 365 | 185 |
| 3 | 380 | 226 |
| 4 | 385 | 184 |
| 5 | 382 | 189 |

Slide no 14 :

Conclusion: In attrition \_data after filtering out and applying the CART algorithm, analysis is done on basis of all columns. The error rate is .3068655  and Accuracy: 70%

 Slide no 15:

Applying algorithms on columns: RLY\_RATE, JOB\_SATISFACTION, REHIRE, TERMINATION YEAR, IS\_FIRST\_JOB, TRAVELLED\_REQUIRED, PERFORMANCE\_RATING, DISABLED\_EMP, DISABLED\_VET, EDUCATION\_LEVL, STATUS and JOB\_GROUP

Slide no 16:

split is happening up to 2 level.

Total Active employees are 3777 and terminated employees are 2951

Slide no 18:

Conclusion:

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

Slide no 19:

K-Mean Cluster : The K-mean cluster is used to develop a cluster model to predict the potential of employee leaving the company(become terminated).

The case study is performed in four parts :

1-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. Columns considered are :

1- Rehired 2-Status

Side no 20:

2- 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.Columns considered are :

1- Education level 2- Status

Slide no 21:

3- 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.Columns considered are :

1- Disable Employee 2- Disable vet 3- Status

Slide no 22:

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.Columns considered are :

1- Year of termination 2- Status

Slide no : 23

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

Slide no 24:

Following conclusion are drawn from K-mean cluster on given set of employment data:

1- Get a meaningful intuition of the structure of the data we’re dealing with.

2- Cluster-then-predict where different models will be built for different subgroups if we believe there is a wide variation in the behaviours of different subgroups.

Slide no 25:

Artificial Neural Networks (ANN):

* 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

Slide no 27:

Output:

ANN classified data as Active : 5394 and terminated : 4218. It provide 100% accuracy while classifying data.

Slide no 28:

Comparison of Classification Algorithms:

On basis of given graph. It can be concluded that while classifying data on basis of all factor for terminated and active employee :

KNN provides 60% accuracy.

C5.0 provides 100% accuracy.

CART provides 70% accuracy.

ANN provides 100% accuracy.

While perform clustering, using K mean it provide 100% accuracy.

Slide no 29:

Conclusion:

Hence, 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

Slide no 30:

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)