**Objective: To develop a KNN-classification model(s) to predict the potential of employees leaving the company (become terminated).**

**KNN definition:**

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.

The distance function used is :

A screenshot of a cell phone

Description automatically generated

**The working of KNN can be explained through below diagram :**

A close up of a device

Description automatically generated

**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 colums are considered :

1-ANNUAL\_RATE ,

2-HRLY\_RAT

3-Status (target value)

The result are as follow:

|  |  |  |
| --- | --- | --- |
| 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 you can see the accuracy lies between 55% to 58% for salary based classification.

2- Classification is done based on Performance and job satisfication :

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 satisfication.

While performing classification on basis of Performance and job satisfication following colums are considered :

1. JOB\_SATISFACTION
2. PERFORMANCE\_RATING
3. STATUS
4. PREVYR\_1
5. PREVYR\_2
6. PREVYR\_3
7. PREVYR\_4

8. PREVYR\_5

The result are as follow:

|  |  |  |
| --- | --- | --- |
| 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 you can see the accuracy lies between 57% to 60% for Performance and job satisfication based classification.

3- Classification is done based on all data:

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

While performing classification on basis of all data following colums are considered :

1. ANNUAL\_RATE
2. HRLY\_RATE
3. JOBCODE
4. ETHNICITY
5. SEX
6. MARITAL\_STATUS
7. JOB\_SATISFACTION
8. AGE
9. NUMBER\_OF\_TEAM\_CHANGED
10. REFERRAL\_SOURCE
11. HIRE\_MONTH
12. REHIRE
13. IS\_FIRST\_JOB
14. TRAVELLED\_REQUIRED
15. PERFORMANCE\_RATING
16. DISABLED\_EMP
17. DISABLED\_VET
18. EDUCATION\_LEVEL
19. STATUS
20. JOB\_GROUP
21. PREVYR\_1
22. PREVYR\_2
23. PREVYR\_3
24. PREVYR\_4
25. PREVYR\_5

The result are as follow:

|  |  |  |
| --- | --- | --- |
| 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 |

From above table you can see the accuracy lies between 58% to 61% for all data.

**Conclusion :**

The above algorithm is performed on data of size : 9k

Among them training data is : 6728 and testing data is : 2884

Following conclusion are drawn from KNN classification on given set of employement 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.**