REPORT: DATA ATTRITION

- **Data Attrition:**

  - The problem statement is mainly based on data attrition. The objective of this project is to predict the attrition rate for the each employee, to find out who's more likely to leave the organization.
  - It will help organization to find ways to prevent attrition or to plan in advance the hiring of new candidate.
  - Where it required to model the probability of attrition using logistic regression. And the result thus obtained will be used by the company to understand what changes they should make to their workplace.

- **Dataset Analysis:**

  - In attrition dataset total no. of employees are 9612. Where there are 27 types of attributes include employee id, age, job group, etc. This dataset contains all information about employees.
  - After expunge data we get status of employees and the summary status is 5394 active employees and 4218 terminated employees.
  - Exploratory dataset analysis refers to the critical process of performing intimal investigations on data so as to discover data in kind of pattern manner. So here we get summary of dataset like sex: 5723 -female, 3889-male, disabled employee: 947, and so on.

- **Models:**

  This models is used to find accuracy and error rate here and data mining using the given dataset, which help company to understand knowledge discovery process and provide road map to follow and understand data attrition.

  - **Naïve Bayes**: Based on Bayes' theorem with an assumption of independence predictor. After applying Naive Bayes model in dataset we get accuracy up to 60, error rate 0.39 and bad prediction around 937.

- **KNN:** This method used to categorize a new observation based on past observation. It classifies new observation by identifying its closest k neighbor's value. We get highest error rate 0.44 and less accuracy 55.81 by taking k=3, where get less error rate 0.41 and better accuracy 58.28 by taking k=10.
- **SVM:** Support vector machine model divide data into segments and creates a line or hyperplane which separates the data into classes and it gives 68.65% accuracy.
- **Neural Network:** This model used for classification, clustering, feature mining, prediction and pattern recognition, which gives 71.27% accuracy.

- **Graphs:**

  - **Comparison Graph:**
    We get highest accuracy in Neural Network model, which shows it is the best model for this analysis, where lowest accuracy found in KNN model. According to this greatest error rate 0.441 in k=3 KNN model and minimum error rate 0.287 by Neural Network among all other model which we used.
  - **Status Of Employee:**
    We can see in graph that almost 1100 employee difference between active and terminated.
  - **No. Of Active Employee:**
    In dataset as compared to male, female is more so according to that in this graph we can see there are more active female then male.
  - **Termination Year Vs. No Of Employee:**
    In all the year there are more female then male and highest termination happened in 2007 and lowest termination happened in 2010, 2005, 2006, 2014 and 2017 almost same rate of termination happened in these years.
  - **Scatter Plot:**
    This represent actual values in dataset against the values predicted by the model. So, this graph assigning probability and setting a manual cut off to p=0.5.
  - **Job Group Vs. Job Satisfaction:**
    As we can see in graph more employees are satisfied in production group than any other groups.
    This are the graphs to analysis different aspect.