**Project Summary :**

--------------------------------------------------------------------------------------------------------------------------- The Employee Performance prediction is a Supervised Machine Learning problem with Multiclass classification , having 1200 recods with 22 features.

* **The following tasks are expected from this project.**
  + Department wise performances
  + Top 3 Important Factors effecting employee performance
  + A trained model which can predict the employee performance based on factors as inputs. This will be used to hire employees
  + Recommendations to improve the employee performance based on insights from analysis
* Primary libraries used for data loading and visualization :
  + **Matplotlib**
  + **Seaborn**
  + **Pandas**
  + **Numpy**
* Predictor and Target variables:
  + Target Variable : 'PerformanceRating'
  + Predictor Variables:

['Gender', 'EducationBackground', 'MaritalStatus', 'EmpDepartment',

'EmpJobRole', 'BusinessTravelFrequency', 'DistanceFromHome',

'EmpEducationLevel', 'EmpEnvironmentSatisfaction', 'EmpHourlyRate',

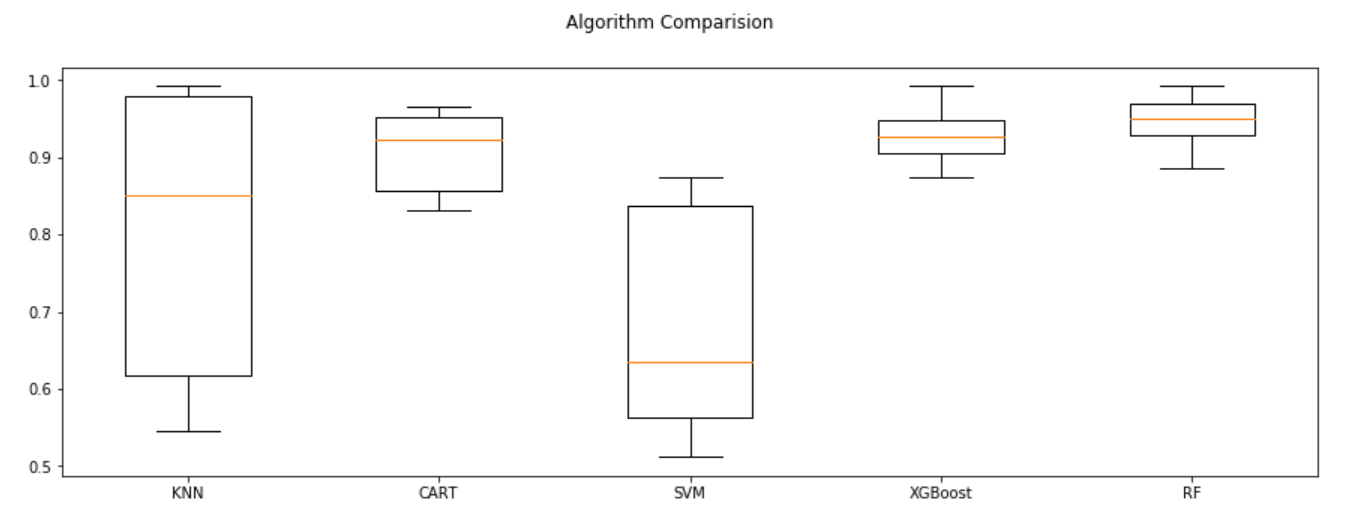
'EmpJobInvolvement', 'EmpJobLevel', 'EmpJobSatisfaction',

'NumCompaniesWorked', 'OverTime', 'EmpLastSalaryHikePercent',

'EmpRelationshipSatisfaction', 'TrainingTimesLastYear',

'EmpWorkLifeBalance', 'YearsSinceLastPromotion', 'Attrition']

* Data Preprocessing :
  + Data looks pretty clean ,no null values.
  + But few continous features having outliers . Outliers handled by IQR(Inter Quartile range) i.e imputing outliers with respective upper and lower whiskers values.
  + Few columns convey same information so dropping one of the column which shows duplication of data.
* Techniques used to determine Correlation:
  + Correlation matrix
  + Variable Inflation Factor(VIF) : Multicollinearity
* Feature Selction Techniques used :
  + SelectKBest
* The following Classification algorithms are considered to predict the Employee performance:



|  |  |
| --- | --- |
| * + **Algorithms** | * + **Accuracy** |
| * + KNN | * + 80% |
| * + SVM | * + 67% |
| * + XGBoost | * + 93% |
| * + RF | * + 94% |

* Since there was a difference between training and test accuracy ,To overcome Overfitting Problem **Kfold Cross Validation** is implemented.
* Since the data is unbalanced **SMOTETomek Oversampling technique** is used to achieve the better accuracy.
* Model evaluation techniques : accuracy\_score , classification\_matrix, classification\_report
* Finally with tuning the model with hyperparams ,cross validation and oversampling achieved accuracy of 99% with XGBoost and 98% with RandomForest.