Project Report

- **Project Objective-** The purpose of this project is to predict future medical expenses of individuals that help medical insurance to make decisions on charging the premium.

- **Dataset Source-** **Insurance Premium Prediction | Kaggle**

- **Dataset Introduction-** The dataset contains 1338 observations (rows) and 7 features (columns). The dataset contains 4 numerical features (age, bmi, children and expenses) and 3 categorical features (sex, smoker and region) that were converted into factors with numerical value designated for each level. We have to look into different features to observe the relation between target and independent features as well as autocorrelation among independent features, and plot a regression model based on several features of individuals such as age, physical condition, number of children and location against their existing medical expense.

- **Data Cleaning Techniques-**
  Removed duplicate rows
  Detected only 0.02% outliers so decided not to remove that
  Null values identification technique used , dataset is absence of null values

- **Data Analysis-**
  Analysis of data by exploring it.
  Checked for numerical and character values, identification of unique values in all attributes, 6 point summary
  Univariate analysis of discrete data and numerical data

- **Regression Models building-**
  Applied following models on given dataset :
  - OLS statistics model
  - Linear Regressor
  - Linear Regression with polynomial features
  - Decision Tree Regressor
  - Random Forest Regressor
  - Support Vector Regressor
  - XGBoost [eXtreme Gradient Boosting] Regressor

- **Steps taken to improve accuracy of models-**
  Following steps are taken to improve upon accuracy of models:
  - Hyperparameter tuning in Decision trees
  - Hyperparameter tuning in Random forests
  - Hyper parameter tuning in SVR
  - Hyper parameter tuning XGBoost

- **Conclusion-**Following points are concluded from the project:
  - Random Forest done good on dataset, have achieved similar RMSE on train dataset and test dataset
  - Train data Root Mean Squared Error: 0.06664856232505958
  - Test data Root Mean Squared Error: 0.07543950366865448
  - Achieved ~86% accuracy for 10-Fold CV Random Forest Regression Model: 0.858888230155341