BITS Pilani
Pilani | Dubai | Goa | Hyderabad
Work Integrated Learning Programmes

Pima Indians Diabetes Assignment

271- R A Bharat, Kulkarni Nishant Mohanrao ,S Sumalatha

Introduction to Data Science

M.Tech Data Science and Engineering

Overview

- Objective: Diabetes analysis - **To determine whether a person can have diabetes based on various features which will contribute towards this outcome**
- Methodology: **Various analytic techniques like Descriptive, Diagnostic and Predictive are used to derive the diabetes outcome**

Dataset

- How many features: **8 features**
- Size of the dataset: **768 records**
- Multiple files: **1 file**
- What kind of data – numerical or character: **All are Numerical**
- Balanced or imbalanced – what is the distribution: **Imbalanced on diabetes outcome for 0**
- Distribution of Training set, validation set, testing set: **70/30**
- Missing data and Preprocessing challenges: **No missing dataset, however, mean and min-max normalization are used for data processing**

Methodology

Feature 1: **Outliers rows for Pregnancies, Insulin and BMI are removed from the dataset**
Feature 2: **Outliers for Pregnancies, Insulin and BMI are converted to mean values in the data set and new features are created in their place Pregnancies_univariate, Insulin_univariateand and BMI_univariate respectively**

- The 3 classifiers used
- Ensemble pipeline: **RandomForestClassifier is used for both feature 1 and feature 2 dataset**
- Other models considered: **DecisionTreeClassifier and GradientBoostingClassifier are built for both Feature 1 and Feature 2**
- Hyper-parameter tuning: **KNN is used for K values between 1 and 15 to find the right value of K with best fit model and accuracy. K = 6,7 and 8 are giving best fit with Micro accuracy of 94% and weighted accuracy of 92%**

Feature Engineering Techniques

- Features removed: **Pregnancies,Insulin and BMI as these had more number of outlier records**
- Feature creation: **Pregnancies_univariate, Insulin_univariateand and BMI_univariate features outliers were changed to mean values of the dataset**
- Feature ranking: **Glucose, BMI, Age, Blood Pressure, Pregnancies, Skin Thickness, Insulin**
- Class imbalance treatment: **DiabetesPedigreeFunction is preprocessed to hold min max normalized value**

BITS Pilani
Pilani | Dubai | Goa | Hyderabad
Work Integrated Learning Programmes

Pima Indians Diabetes Assignment

271- R A Bharat, Kulkarni Nishant Mohanrao ,S Sumalatha

Introduction to Data Science

M.Tech Data Science and Engineering

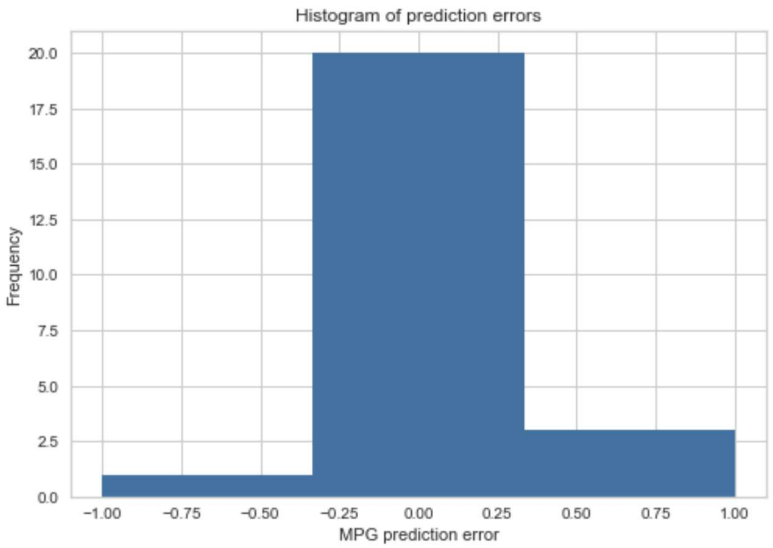
Results

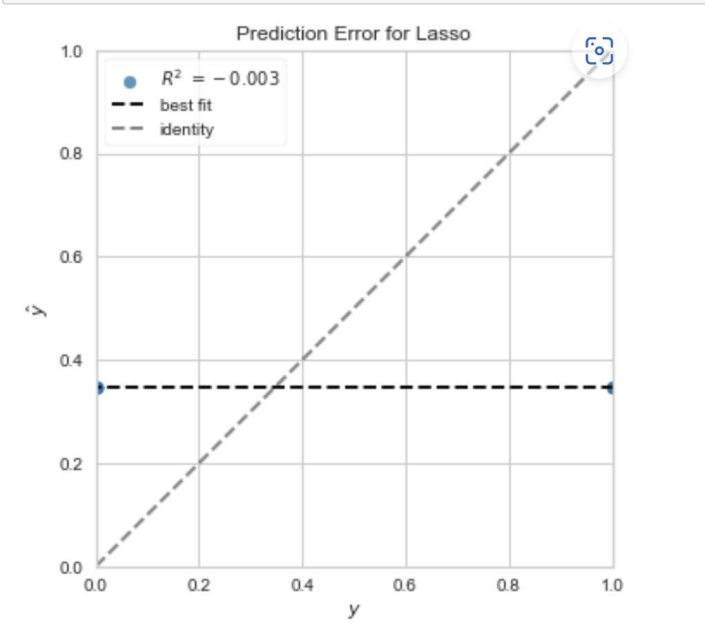
- Table for the evaluation metric for each ML technique used:

Methods	Accuracy
Feature 1 : ML1 - Decision Tree classifier	77.27
Feature 2 : ML1 - Decision Tree classifier	83.33
Feature 1 : ML2 - Random Forest Ensembler	75.65
Feature 2 : ML2 - Random Forest Ensembler	75.3
Feature 2 : ML3 - GradientBoostingClassifier	66.67
Feature 1 : ML1 - Pipeline DTC	72.73
Feature 1 : Hyper Parameter tunning KNN:6,7,8	94.0

- Plot of the curves:

```
Root Squared Mean Error for Decision tree is : 0.408248290463863
```





- Conclusion: **Diabetes outcome to determine if a person is diabetes or not is best predicted using the KNN model with K=6,7 and 8 as the error of prediction is less and outcome is as close to the dataset provided. This model is followed by Decision tree classifier with accuracy of diabetes prediction 83% for the feature set in which outliers are converted to mean values.**