Diabetes Health Indicators

ID	Section.nb	Names
20201701066	Section 1	احمد ايمن احمد قاسم
20201700954	Section 32	هايدى حسن سعد عبدالله
20201700814	Section 23	مريم علاء الدين فضل علام
20201701153	Section 33	هناء احمد حامد محمد
20201700944	Section 32	نورهان محمد عبدالله محمد حماد
20201700934	Section 32	نوران سمير احمد حسب عوف
20201700308	Section 12	زياد شريف محمد عبدالفتاح مصطفى

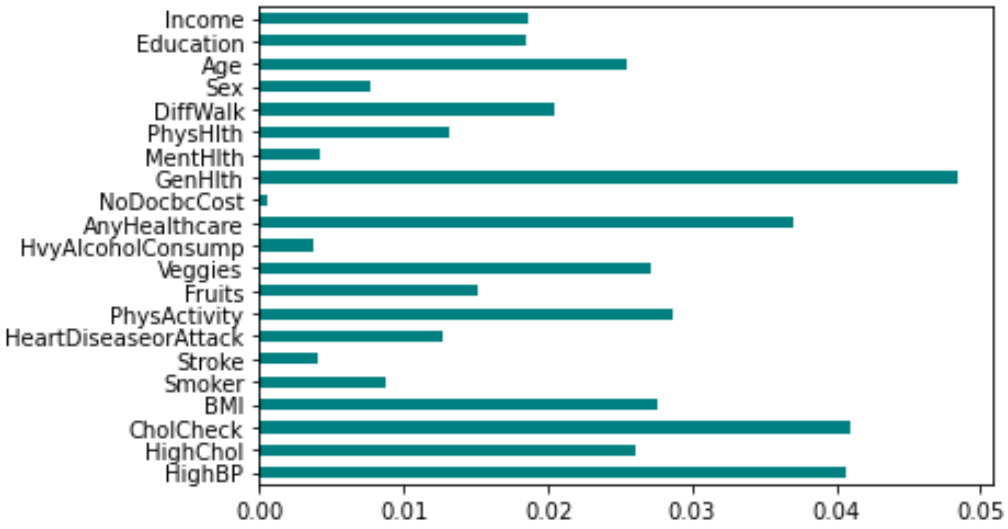
Introduction:

Diabetes mellitus refers to a group of diseases that affect the way your body uses blood sugar (glucose). Glucose is vital to your health because it is an important source of energy for the cells that make up your muscles and tissues. It is also the main source of energy for your brain.

Preprocessing:

Feature Selection:

- *We used mutual information method using information gain for feature selection and removed the lowest 4 columns which are Stroke, HvyAlcoholConsump, MentHlth, NoDocbcCost.*
- *information method graph shown below:*

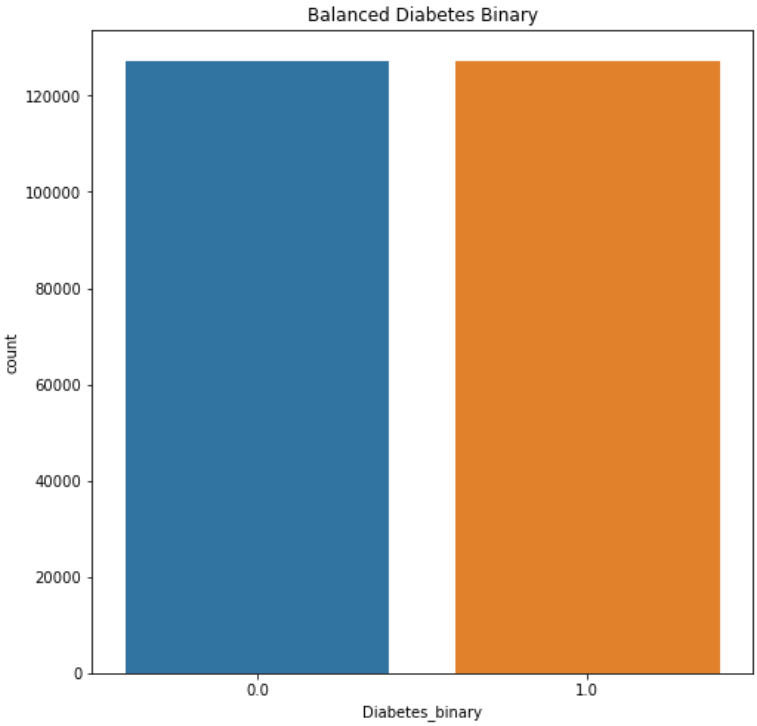


Data Cleansing:

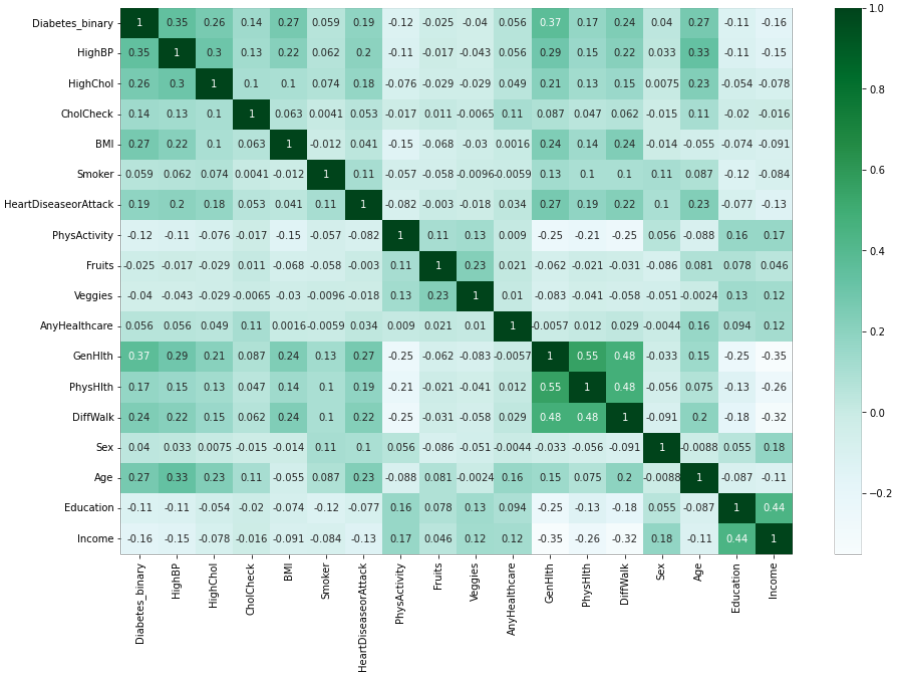
- *Removing all nulls from columns and rows using dropna function.*
- *Removing duplicates from rows using drop_duplicates function.*
- *Multiply the age by 5 so we have the actual age.*
- *After preprocessing number of rows change from **253680** to **216923**.*

Balancing dataset:

- *We used oversampling technique to balance the dataset using imblearn package with the sampling strategy set to minority and used the SMOTE strategy which solves the overfitting produced by the random oversampling strategy.*
- *Here is a count plot after balancing the dataset:*



Visualize Correlation:



Normalization:

- *Essential stage in machine learning process, the features distribution is not a bell ring so normalization is an important step here also the data scales are different so when we normalize it we get better accuracy.*
- *Normalize data using Pandas and sklearn library.*
- *min_max_scaling(df[col]) which subtracts the minimum value in the feature and then divides by the range (the difference between the original maximum and original minimum).*

Model Training:

- *Train test split is a function for splitting data arrays into two subsets: for training data and for testing data which will be set to 0.20.*

Logistic Regression:

- *Logistic Regression with max iterations 10000 used to perform linear and polynomial regression and make predictions accordingly, Then training data using fit() function.*

Model Evaluation:

- *Accuracy: 0.7078230403982974*
- *Precision: 0.32051282051282054*
- *F1 Score: 0.44749229964549314*
- *Confusion Matrix:*
  - *TP: 7700*
  - *TN: 38363*
  - *FP: 16324*
  - *FN: 2690*

SVM:

- *We used for the decision function shape One vs One strategy which is actually faster to train since we have a large dataset with a lot of features.*

Model Evaluation (kernel: linear):

- *Accuracy: 0.6937935061542481*
- *Precision: 0.31380720239031296*
- *F1 Score: 0.44479117327463713*
- *Confusion Matrix:*
  - *TP: 7982*
  - *TN: 37168*
  - *FP: 17454*
  - *FN: 2473*

Model Evaluation (kernel: rbf):

- *Accuracy: 0.7043348648524056*
- *Precision: 0.3202095440779242*
- *F1 Score: 0.4485081257703*
- *Confusion Matrix:*
    - *TP: 7824*
    - *TN: 38012*
    - *FP: 16610*
    - *FN: 2631*

Random Forecast:

- *A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.*
- *We used 400 for estimators which are number of trees in the forest.*

Model Evaluation:

- *Accuracy: 0.8191680624490988*
- *Precision: 0.4017398745008557*
- *F1 Score: 0.3237558901275715*
- *Confusion Matrix:*
    - *TP: 2817*
    - *TN: 50492*
    - *FP: 4195*
    - *FN: 7573*

Decision Tree:

- *A non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.*

Model Evaluation:

- *Accuracy: 0.7588395285584769*
- *Precision: 0.2768803634528016*
- *F1 Score: 0.2954749506195008*
- *Confusion Matrix:*
    - *TP: 3291*
    - *TN: 46092*
    - *FP: 8595*
    - *FN: 7099*

For Full Implementation Code:

link: https://colab.research.google.com/drive/1g2G4D3goFzXU6xyCnOOTWAzA84scohTn?usp=sharing