DATA SCIENCE PROJECT REPORT
DIABETES PREDICTION

1. Project Description and Scope

Diabetes is a type of chronic disease which is more common among the people of all age groups. Predicting this disease at an early stage can help a person to take the necessary precautions and change his/her lifestyle accordingly to either prevent the occurrence of this disease or control the disease (For people who already have the disease).

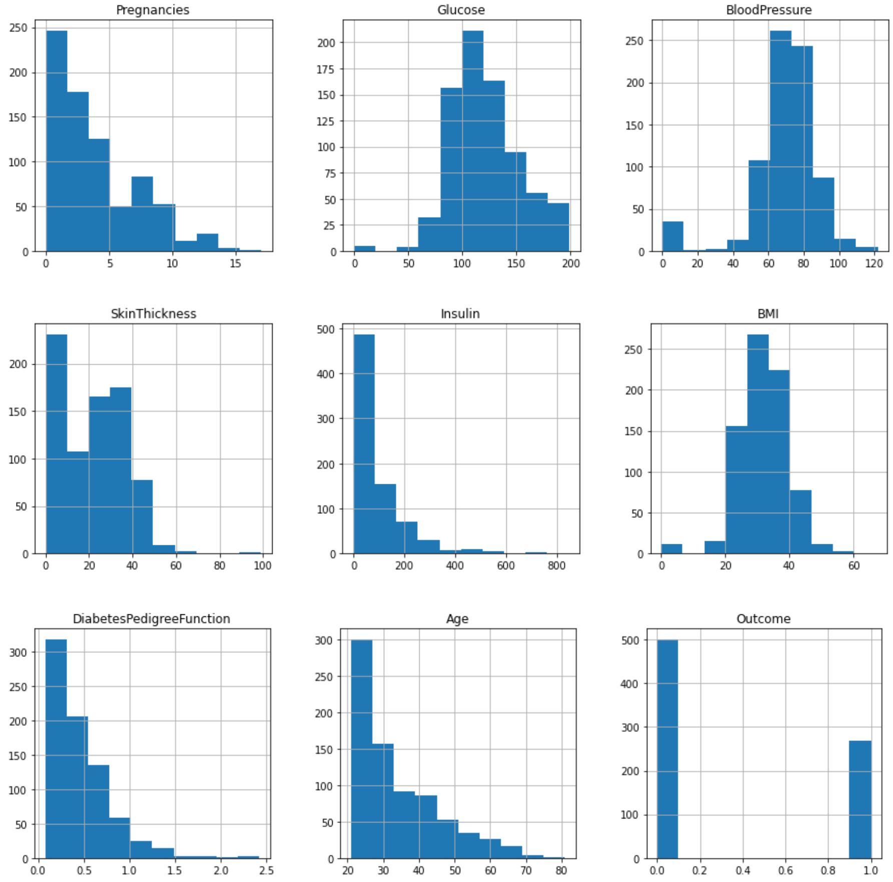
The focus of this project is developing machine learning models that can accurately predict person has diabetes or not based on its parameters. We implement and evaluate various learning methods on a dataset.

2. Dataset

For this project we are using Pima Indians Diabetes Database available on Kaggle [References - 1]. The features available in this dataset are Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Outcome.

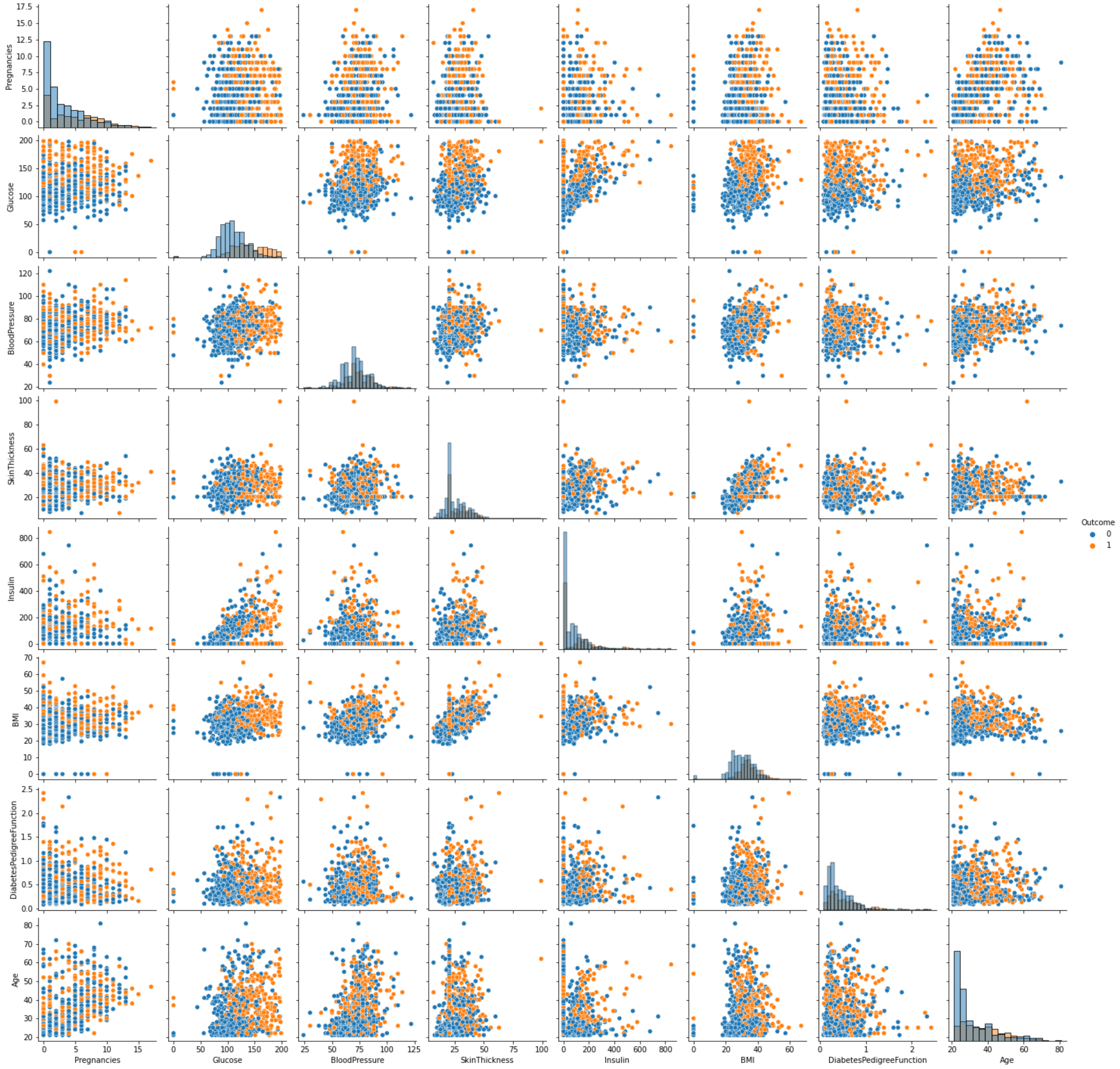
3. Pre-processing and Visualization

In order to better understanding of dataset histograms of data are plotted. We can see that many values like Glucose, Blood pressure, insulin should not be ‘0’ as it does not fit normal human condition. This is because of unavailability of data they could be give 0. As a solution we can replace that value with mean.



We will not drop any feature as none is causing problem in training and every parameters are playing important role in detecting decease.

Given below graph is to visualize how outcome is distributed according to features. For example In most of the Glucose graphs outcome is not mixed and easy to analyze whereas in BloodPressure vs [Skin thickness, insulin, BMI] graphs outcome is mixed and not clear.



4. Handling Unbalanced data

In outcome column of dataset which are labels if you can observe, number of outputs as ‘0’ is almost double than number of outputs as ‘1’. Which is practically true, most of the people who diagnose will not be diabolic. But while training a model it causes problem, as it gives more priority to ‘0’. Then probability of prediction will be ‘0’ will increase. This happens because of unbalanced dataset.

As a solution to this problem, we need to do sampling of dataset. As in this dataset there is around 700-800 amount of data. Upsampling should be batter choice for sampling. It will create new data points around low number of datapoints.

5. Methodology

We utilized several classic and state-of-the-art methods, including ensemble learning techniques, with a 80% - 20% split for the training and test data. Logistic Regression, Random Forest and KNN are baseline methods. For most of the model implementations, the open-source Scikit-Learn package [References - 2] was used.

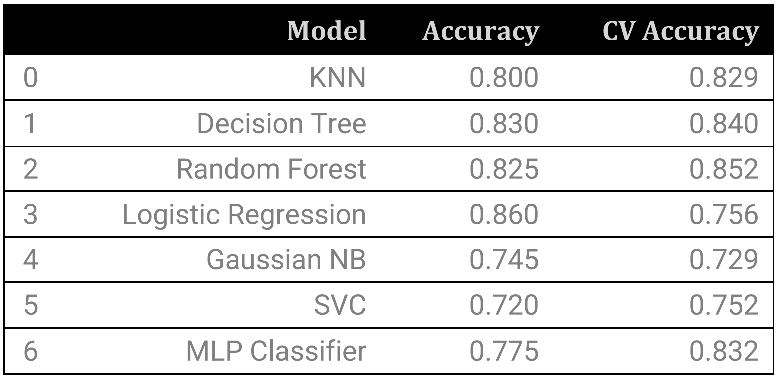
As we want best accuracy and test data could be uncertain, we will cross validate with k-fold cross validation and get mean of all accuracy for each model.

Used methods are,

1. K-nearest neighbor
2. Decision tree
3. Random forest
4. Logistic Regression
5. Naïve bayes
6. Support Vector Machine
7. Neural Network (MLP)

6. Results

The results of our tests were quantified in terms of the accuracy score of our predictions.



	Model	Accuracy	CV Accuracy
0	KNN	0.800	0.829
1	Decision Tree	0.830	0.840
2	Random Forest	0.825	0.852
3	Logistic Regression	0.860	0.756
4	Gaussian NB	0.745	0.729
5	SVC	0.720	0.752
6	MLP Classifier	0.775	0.832

Basing our selection criteria on the accuracy score, the best model for this project is Random Forest and Decision Tree which gives an Accuracy Score of 85% and 84%. There are different ways of determining the best model like using Standard deviation etc. For this project, I decided to use the accuracy score as the main metric in choosing the best model.

7. References

1. *https://www.kaggle.com/uciml/pima-indians-diabetes-database*
2. *https://scikit-learn.org/stable/modules/classes.html : Sklearn in python API Reference*