Course Title : Diabetes Prediction using Machine Learning Algorithms

Team Members :
Madhavendra Srivastava (RA1911003010901)
Prinitha V (RA1911003010902)
Mohd Azaan (RA1911003010900)
Aayush Mehta (RA1911003010905)

**Project Link : https://github.com/Madhavendra21/Ai-Project**

Problem Statement :

Diabetes has proven to be the most common disease in aged people since decades irrespective of ethnicity, gender or physique. Accurate predictive measures are crucially essential to avoid further serious immunity consequences. We plan to solve this problem using non-traditional methods involving various machine learning algorithms providing the maximum accuracy over the manual testing methodologies hence eliminating the factor of human error.

Methodology :
1. **Data preprocessing:** Cleaning and removal of inconsistent data.
2. **Data analysis and viz:** EDA is one of the most important steps in the data science project life cycle and making inferences from the visualizations and data analysis.
3. **Model development:** Here we will be developing various machine learning models optimizing the accuracy metrics over the data.
4. **Model testing :** Validating the metrics over the test data concluding the best performing model.
5. **Model deployment:** Saving the best model using pickle to make the prediction from real data

Approach :
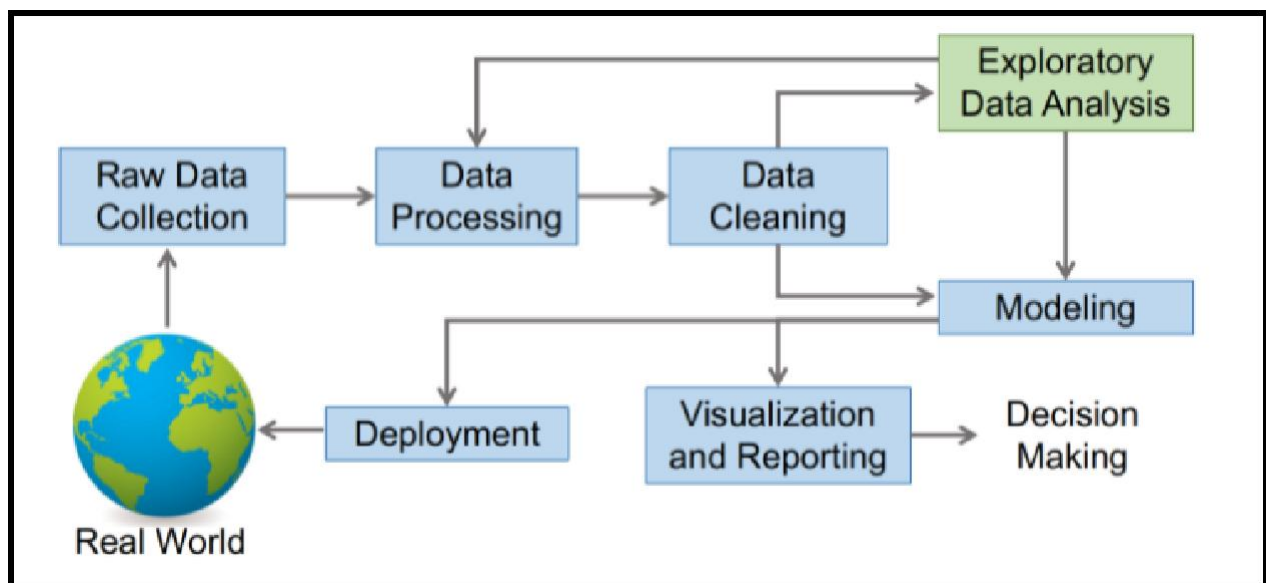
Tech Stack :

Numpy, Pandas, Matplotlib, Seaborn, Pickle, Sklearn, Xgboost, Random Forest Classifier, Support Vector Machines, KNeighborsClassifier, Google Colab, Jupyter Notebook
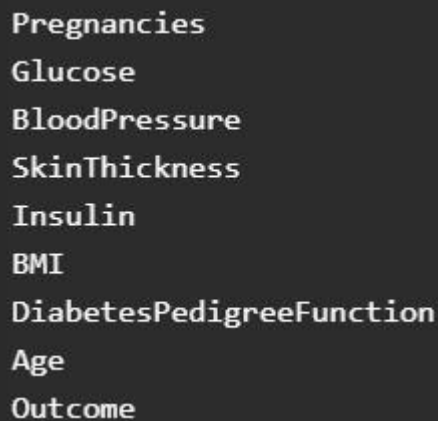
Data :

Dataset Usage : Pima Indians Diabetes Dataset

Description :

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

The datasets consist of several medical predictor variables and one target variable, Outcome. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

Column Labels

```
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Outcome
```

Purpose of the problem statement:

After testing the patient records, we are able to develop a machine learning model to accurately predict if the patients in the dataset have diabetes also plotting some insights from the data via data analysis and visualization techniques, finally concluding the random forest as the best classifier algorithmic model with an accuracy_score of 0.76.

Inferences :

https://towardsdatascience.com/understanding-random-forest-58381e0602d2

https://towardsdatascience.com/exploratory-data-analysis-eda-a-practical-guide-and-template-for-structured-data-abfbf3ee3bd9

https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/