Diabetes Prediction Project

Overview

In this project, the objective is to test accuracy of Machine Learning algorithms to predict whether the person has Diabetes or not based on various features such as-

- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin Level
- BMI
- Age

Using various Machine Learning algorithms, we will train models and calculate accuracy of each algorithms.

The data set that has used in this project has taken from kaggle. "This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases.

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.

Learning Objective

- Data Gathering
- Descriptive Analysis
- Data Preprocessing
- Data Modelling
- Model Evaluation

Technologies Used

- Numpy
- Pandas
- Sklearn

•Jupyter Notebook

Importing Dataset

The Basic process of loading data from a CSV file into a pandas dataframe is achieved using the “READ_CSV” function in pandas.W

IMPORT PANDAS AS PD DATA=PD.READ_CSV("FILENAME.CSV")

Analysing Dataset

We need to analyse and remove the columns which are not necessary and use the ones which are necessary and handle the missing value if any by changing into mean value of dataset.

Splitting up of Data

Data is splitted into training and testing set based on size of dataset using

TRAIN_TEST_SPLIT FROM SKLEARN.MODEL_SELECTION

In our model, we have used 80% of dataset to train the model and rest 20% of dataset to test the model’s accuracy.

Applying ML algorithms

We will use classification predictive models to train and test our data.

Classification predictive modelling is the task of approximating a mapping function(F) from input variables(X) to discrete output variables(Y).

The following algorithms were used to train and test the dataset.

•**K Neighbour Classifier:-** Also known as K-Nearest Neighbour is a model that stores all the available data and classifies a new data point based on the similarity.

•**Decision Tree Classifier:-** In this algorithm, data is continuously split according to a certain parameter. The leaves are the decision or final outcomes and the decision nodes are where the data is split.

•**Logistic Regression:-** Logistic Regression is a predictive analysis algorithm and based on the concept of probability. logistic function normalizes everything to be between 0 and 1. Then we can interpret the number you get as a probability.

•**Random Forest:-** The Random Forest Classifier is a set of decision trees from randomly selected subset of training set. It aggregates the votes from different decision trees to decide the final class of the test object.

Evaluation of model

The following were the accuracy result of the algorithms used-

- K Neighbour Classifier -> 78.57142857142857**%**
- Decision Tree Classifier -> 79.22077922077922**%**
- Logistic Regression -> 82.46753246753246**%**
- Random Forest -> 83.11688311688312**%**