DIABETES PREDICTION USING DATA SCIENCE

ABSTRACT:

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes mellitus or simply diabetes is a disease caused due to the increase level of blood glucose. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. However, early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors as diabetes affects human organs such as kidney, eye, heart, nerves, foot etc. Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. One such task is to help make predictions on medical data. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. This project aims to predict diabetes via different machine learning methods including: Logistic regression, SVM, Random forest, k-Nearest Neighbour ,Naïve Bayes Theorem. This project also aims to propose an effective technique for earlier detection of the diabetes disease.

DIABETES MELLITUS:

Diabetes is one of deadliest diseases in the world. It is not only a disease but also a creator of different kinds of diseases like heart attack, blindness, kidney diseases, etc. The normal identifying process is that patients need to visit a diagnostic centre, consult their doctor, and sit tight for a day or more to get their reports. Moreover, every time they want to get their diagnosis report, they have to waste their money in vain.Diabetes Mellitus (DM) is defined as a group of metabolic disorders mainly caused by abnormal insulin secretion and/or action. Insulin deficiency results in elevated blood glucose levels (hyperglycemia) and impaired metabolism of carbohydrates, fat and proteins. DM is one of the most common endocrine disorders, affecting more than 200 million people worldwide. The onset of diabetes is estimated to rise dramatically in the upcoming years. DM can be divided into several distinct types. However, there are two major clinical types, type 1 diabetes (T1D) and type 2 diabetes (T2D), according to the etiopathology of the disorder. T2D appears to be the most common form of diabetes (90% of all diabetic patients), mainly characterized by insulin resistance. The main causes of T2D include lifestyle, physical activity, dietary habits and heredity, whereas T1D is thought to be due to autoimmunological destruction of the Langerhans islets hosting pancreatic-β cells. T1D affects almost 10% of all diabetic patients worldwide, with 10% of them ultimately developing idiopathic diabetes. Other forms of DM, classified on the basis of insulin secretion profile and/or onset, include Gestational Diabetes, endocrinopathies, MODY (Maturity Onset Diabetes of the Young), neonatal, mitochondrial, and pregnancy diabetes. The symptoms of DM include polyuria, polydipsia, and significant weight loss among others. Diagnosis depends on blood glucose levels (fasting plasma glucose = 7.0 mmol/L

PROJECT DESCRIPTION:

we have collected dataset from online sources.

TECHNIQUES USED

- Data Cleaning
- Data Visualization
- Machine Learning Modelling

ALGORITHMS USED

1. Logistic Regression
2. Support Vector Machine
3. KNN
4. Random Forest Classifier
5. Naive Bayes

MODEL EVALUATION METHODS USED

1. Accuracy Score
2. ROC AUC Curve

DECSRIPTION OF CODE:

1. In the first line of code, we have imported several python libraries namely:

Pandas: It is mainly used for data analysis and also it allows various data manipulation operations such as merging, reshaping, selecting, as well as data cleaning, and data wrangling features.

NumPy: It is used to perform a wide variety of mathematical operations on arrays. It adds powerful data structures to Python that guarantee efficient calculations with arrays and matrices and it supplies an enormous library of high-level mathematical functions that operate on these arrays and matrices.

Seaborn: It is used for data visualization and exploratory data analysis. Seaborn works easily with data frames and the Pandas library. The graphs created can also be customized easily

Matplotlib: It is an amazing visualization library in Python for 2D plots of arrays. ... One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

2. In the second line of code ,we have displayed the fist five rows of the csv file using df.head().

Data Description of CSV file:

- **Pregnancies**: Number of times pregnant
- **Glucose:** Plasma Glucose concentration a 2 hours in an oral glucose tolerance test
- **Blood Pressure:** Diastolic blood pressure(mm Hg)
- **Skin Thickness:** Tricep skin fold thickness(mm)
- **Insulin:** 2-hour serum Insulin(mu U/ml)
- **BMI:** Body Mass Index(weight in kg/(height in m)^2)
- **Diabetes Pedigree Function:** Diabetes Pedigree Function
- **Age:** Age of a female
- **Outcome:** 1 if diabetes,0 if no diabetes

3. In the third line of code, we have used df.describe() for describing the dataset. It is also used to view some basic statistical details like percentile, mean, std,min,max etc.

4. In the fourth line of code ,we have used df.info() for displaying the information of dataset. This method prints information about a DataFrame including the index data type and column datatypes, non-null values and memory usage.

5. In the fifth line of code, we have created histogram for all attributes in the dataset. It is used to represent data provided in a form of some groups. It is accurate method for the graphical representation of numerical data distribution.

6. In the sixth line of code, we have used sns.heatmap(df.corr()) for finding the correlation. Correlation means association - more precisely it is a measure of the extent to which two variables are related. A positive correlation is a relationship between two variables in which both variables move in the same direction. From this, we can analyse skin thickness, insulin, pregnencies and age are full independent to each other. And also it is observed that the age and pregencies has negative correlation.

7. In the seventh line of code, we have counted the total outcome in each target 0 or 1 using sns.countplot(y=df['Outcome'],palette='Set1').Here 0 indicates no diabetes and 1 indicates patient has diabetes.

8. In the eighth line of code we have done two scatter plots considering Age versus Glucose and Age versus BloodPressure.

9. In the ninth line of code, we have extracted the features for X and y.

10. In the tenth line of code, we have divided 80% of data for training the model and 20% of data for testing the model.

11. In the eleventh line of code,we have displayed the shape of the test and train data.

12. In the 12,13,14 ,15 line of code, we have displayed first five rows of the train_x , test_X, train_y, test_y using the method head().

BUILDING THE MODEL:

1. LOGISTIC REGRESSION:

Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. Based on historical data about earlier outcomes involving the same input criteria, it then scores new cases on their probability of falling into a particular outcome category.

The accuracy of Logistic Regression is: 0.775
The ROC_AOC curve is: 0.7254774305555556

2.SUPPORT VECTOR MACHINE (SVM):

Support vector machines (SVMs) are particular linear classifiers which are based on the margin maximization principle. They perform structural risk minimization, which improves the complexity of the classifier with the aim of achieving excellent generalization performance.

The Accuracy of SVM is : 0.785
The ROC_AOC curve is : 0.7424045138888888

3. RANDOM FOREST:

A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees.

The Accuracy of Random Forest is : 0.9925
The ROC_AOC curve is : 0.9895833333333333

4. K-NEAREST NEIGHBORS:

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows.

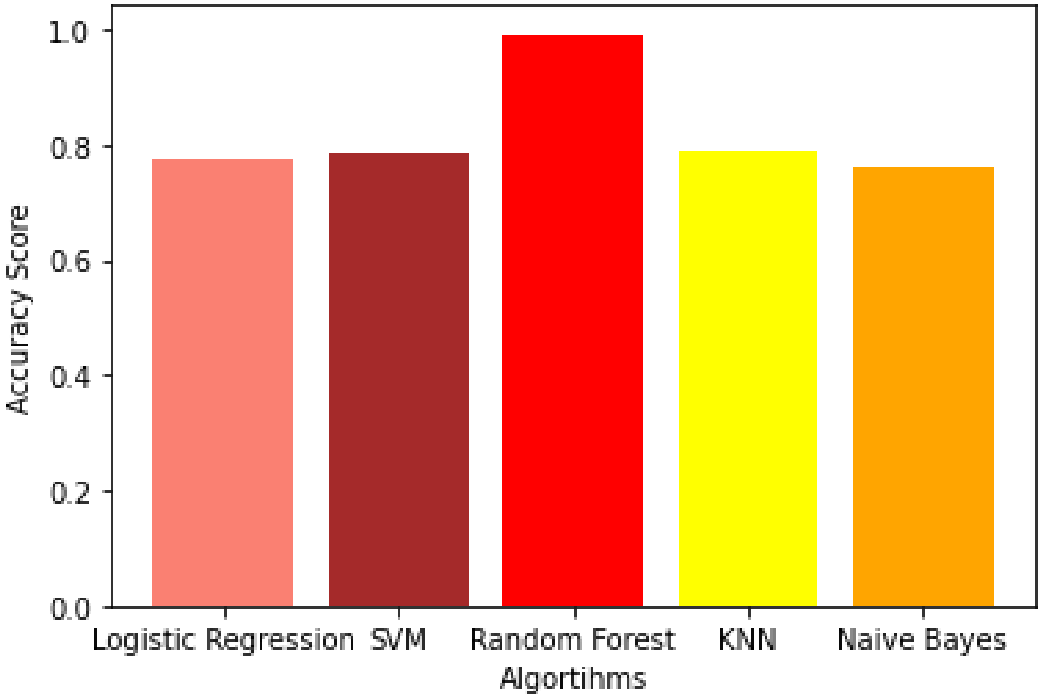
The Accuracy of KNN is : 0.7875
The ROC_AOC curve is : 0.7641059027777778

5. NAIVE BAYES THEOREM:

It is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

The Accuracy of Naive Bayes is : 0.76
The ROC_AOC curve is : 0.716796875

PLOTTING THE GRAPH:



From the above graph we can conclude that Random forest has highest accuracy 95%+-

THANK YOU