Diabetes Prediction App

ABSTRACT

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to the 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 increased 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 a 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. 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 three different supervised machine learning methods including: SVM, Logistic regression, ANN. This project also aims to propose an effective technique for earlier detection of the diabetes disease.

TABLE OF CONTENTS:

1. Introduction

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 center, 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 autoimmune logical 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.

Machine Learning

Machine learning is the scientific field dealing with the ways in which machines learn from experience. For many scientists, the term “machine learning” is identical to the term “artificial intelligence”, given that the possibility of learning is the main characteristic of an entity called intelligent in the broadest sense of the word. The purpose of machine learning is the construction of computer systems that can adapt and learn from their experience. A more detailed and formal definition of machine learning is given by Mitchel: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. With the rise of Machine Learning approaches we have the ability to find a solution to this issue, we have developed a system using data mining which has the ability to predict whether the patient has diabetes or not. Furthermore, predicting the disease early leads to treating the patients before it becomes critical. Data mining has the ability to extract hidden knowledge from a huge amount of diabetes-related data. Because of that, it has a significant role in diabetes research, now more than ever. The aim of this research is to develop a system which can predict the diabetic risk level of a patient with a higher accuracy. This research has focused on developing a system based on three classification methods namely, Support Vector Machine, Logistic regression and Artificial Neural Network algorithms.

Supervised Learning

In supervised learning, the system must “learn” inductively a function called target function, which is an expression of a model describing the data. The objective function is used to predict the value of a variable, called dependent variable or output variable, from a set of variables, called independent variables or input variables or characteristics or features. The set of possible input values of the function, i.e. its domain, are called instances. Each case is described by a set of characteristics (attributes or features). A subset of all cases, for which the output variable value is known, is called training data or examples. In order to infer the best target function, the learning system, given a training set, takes into consideration alternative functions, called hypothesis and denoted by h. In supervised learning, there are two kinds of learning tasks: classification and regression. Classification models try to predict distinct classes, such as e.g. blood groups, while regression models predict numerical values. Some of the most common techniques are Decision Trees (DT), Rule Learning, and Instance Based Learning (IBL), such as k-Nearest Neighbours (k-NN), Genetic Algorithms (GA), Artificial Neural Networks (ANN), and Support Vector Machines (SVM).

Unsupervised Learning

In unsupervised learning, the system tries to discover the hidden structure of data or associations between variables. In that case, training data consists of instances without any corresponding labels.Association Rule Mining appeared much later than machine learning and is subject to greater influence from the research area of databases. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

Reinforcement Learning

The term Reinforcement Learning is a general term given to a family of techniques, in which the system attempts to learn through direct interaction with the environment so as to maximize some notion of cumulative reward. It is important to mention that the system has no prior knowledge about the behaviour of the environment and the only way to find out is through trial and failure (trial and error). Reinforcement learning is mainly applied to autonomous systems, due to its independence in relation to its environment.

Flask

Flask is a web framework, it’s a Python module that lets you develop web applications easily. It has a small and easy-to-extend core: it’s a microframework that doesn’t include an ORM (Object Relational Manager) or such features.

It does have many cool features like url routing, template engine. It is a WSGI web app framework.

K Nearest Neighbour

K-Nearest Neighbor (KNN) is one of the most popular and simplest machine learning classification technique to predict disease using health data. KNN based model consists only of storing the training data set. To make a prediction for a new data point, the algorithm finds the closest data points in the training data set — its “nearest neighbors". However, in many cases the effectiveness of KNN depends on the K-values, such as K = 1, 2, 3, and so on. Such K-value is defined statically in the traditional KNN based prediction model. As a result, in many cases such static value of K, e.g., K = 1 may cause the lower prediction accuracy.

Thus, we present an optimal K-Nearest Neighbor (Opt-KNN) based disease risk prediction model based on relevant disease and patient’s habitual attributes.
This Opt-KNN based model dynamically determines the optimal number of neighbors for providing better prediction outcome. The effectiveness of this machine learning eHealth model is examined by conducting experiments on the realworld diabetes data collected from medical hospitals.

Support Vector Machine

The Support Vector Machine (SVM) was first proposed by Vapnik, and SVM is a set of related supervised learning method always used in medical diagnosis for classification and regression. SVM simultaneously minimize the empirical classification error and maximize the geometric margin. So SVM is called Maximum Margin Classifiers. SVM is a general algorithm based on guaranteed risk bounds of statistical learning theory, so called structural risk minimization principle. SVMs can efficiently perform nonlinear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. The kernel trick allows constructing the classifier without explicitly knowing the feature space. Recently, SVM has attracted a high degree of interest in the machine learning research community. Several recent studies have reported that the SVM (support vector machines) generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms. SVM is a technique suitable for binary classification tasks, so we choose SVM to predict diabetes. The reason is SVM is well known for its discriminative power for classification, especially in the cases where a large number of features are involved.

Naive Bayes Classifier

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. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Naive Bayes is a machine learning classifier which employs the Bayes Theorem. Using Bayes theorem posterior probability P(C|X) can be calculated from P(C), P(X) and P(X|C). Therefore, P(C|X) = (P(X|C) P(C))/P(X)
Where, P(C|X) = target class's posterior probability.
P(X|C) = predictor class's probability.
P(C) = class C's probability being true.
P(X) = predictor's prior probability

2. Proposed method

The whole project will be completed in 2 complex steps

a. *Creating a model using machine learning*

b. *Creating a web app using flask and connecting it with model*

3. Implementation

Creating a model using machine learning

a. Install all required packages

all the packages can be installed using pip from cmd(terminal)

```
pip install pandas,numpy,matplotlib,scikit-learn,seaborn
```

b. Install jupyter notebook

```
pip install jupyter
```

```
OR
```

```
Install Anaconda that comes with all necessary packages in built
```

c. Open jupyter notebook

```
python -m notebook
```

Open a new notebook in jupyter, follow the below steps along

a. Import the necessary libraries
b. Load the dataset
c. EDA on dataset
d. Modeling and training
e. Improve accuracy using Hyperparameter tuning
f. Evaluate the model
g. Save and load the model

Creating a web app using flask and connecting it with model

. Create a web app and write its code in app.py

. Create two html files i.e, detect.html and result.html for taking input data and displaying results

. Connect the web app to the classifier model

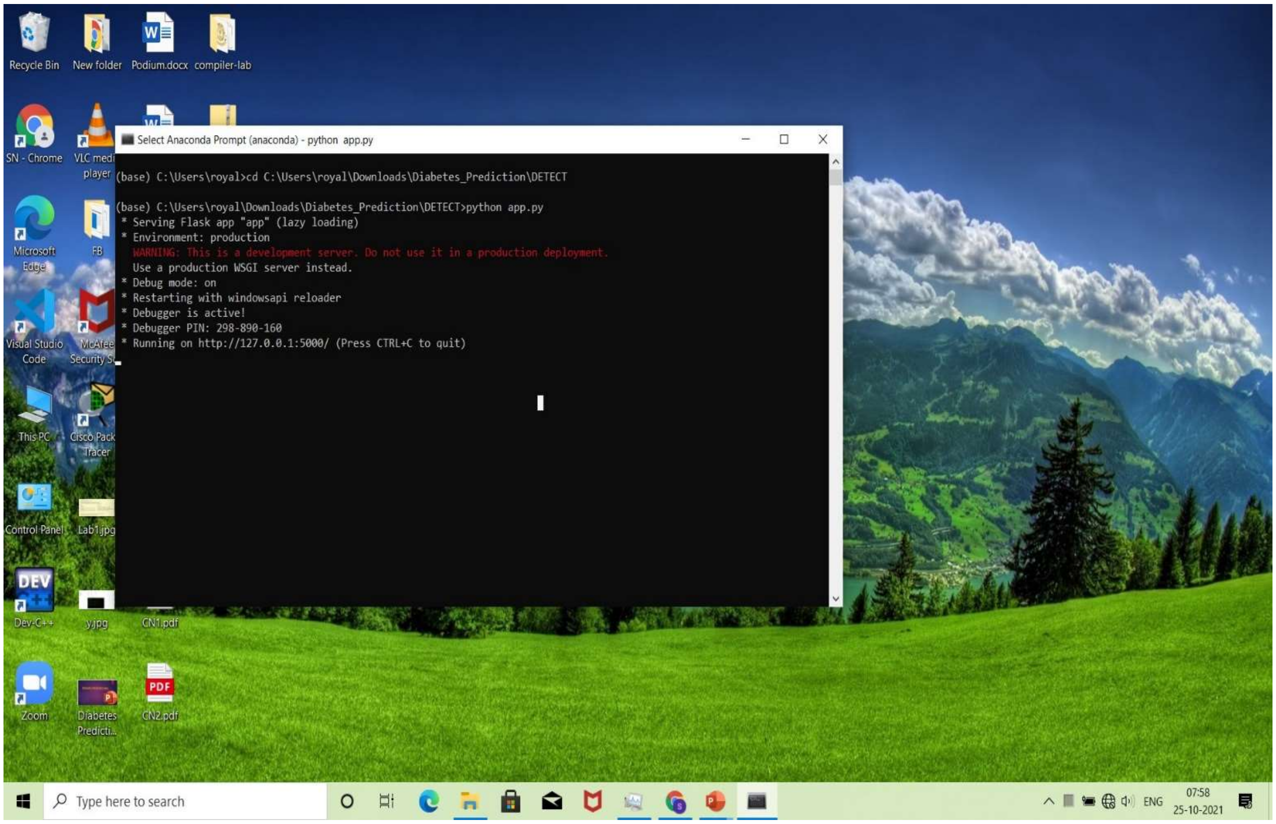
EXECUTION

a.Open cmd navigate to project>code>flask

b. Run the command

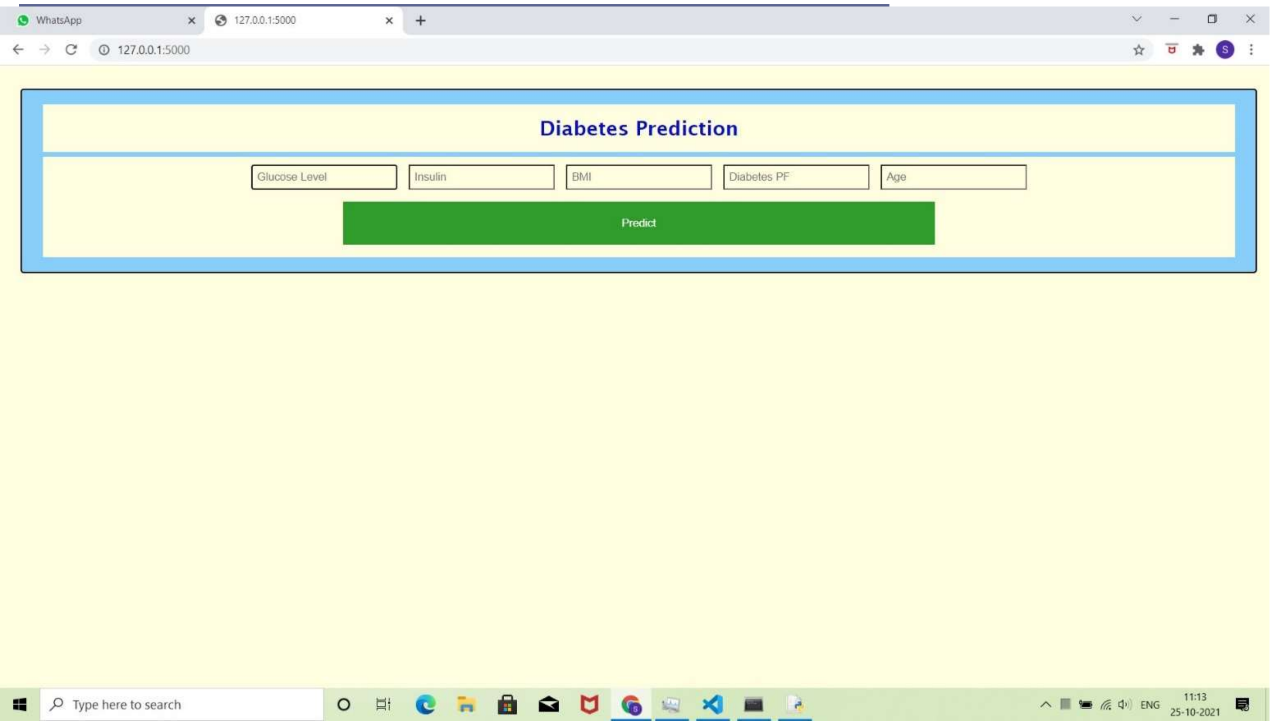
python app.py

Screenshots Of Execution -

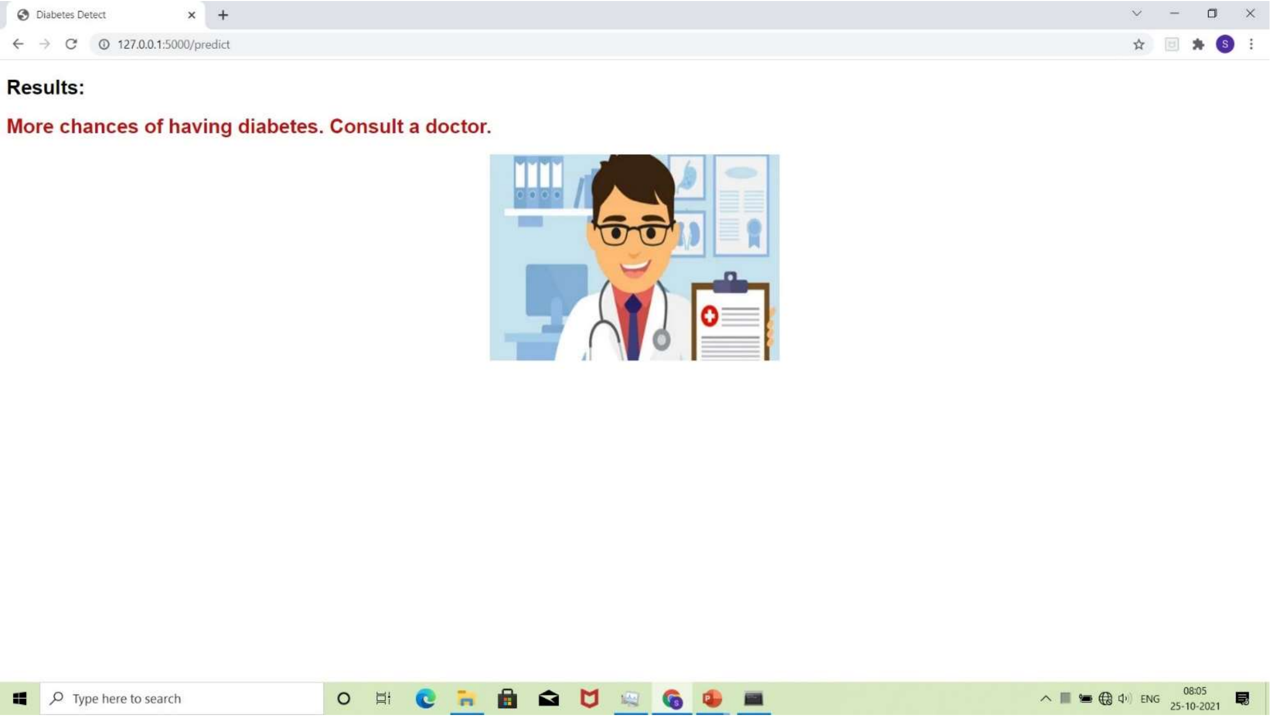


```
(base) C:\Users\royal>cd C:\Users\royal\Downloads\Diabetes_Prediction\DETECT

(base) C:\Users\royal\Downloads\Diabetes_Prediction\DETECT>python app.py
 * Serving Flask app "app" (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: on
 * Restarting with windowsapi reloader
 * Debugger is active!
 * Debugger PIN: 298-890-160
 * Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
```



Diabetes Prediction



Results:

More chances of having diabetes. Consult a doctor.

4. Conclusion

Machine learning has the great ability to revolutionize the diabetes risk prediction with the help of advanced computational methods and availability of large amount of epidemiological and genetic diabetes risk dataset. Detection of diabetes in its early stages is the key for treatment. This work has described a machine learning approach to predicting diabetes levels. The technique may also help researchers to develop an accurate and effective tool that will reach at the table of clinicians to help them make better decision about the disease status.