

**University of Mumbai**

**2020-2021**

**Third Year of Engineering**

In

**Electronics and Telecommunication Engineering**

Project on

**Diabetes classification using Machine Learning**

### Project by:

|  |  |
| --- | --- |
| 1. Dhananjay Kumbhar | (A-636) |
| 2. Tejas Kunder | (A-637) |
| 3. Bhumi Mota | (A-644) |
| 4. Veer Kumar | (A-646) |
|  |  |
|  |  |
|  |  |

Under Guidance of

**Prof.Poonam Sonar**



# CERTIFICATE

This is to certify that report entitled “**Diabetes classification using Machine Learning**” has been done by **Dhananjay Kumbhar ,Tejas Kunder , Bhumi Mota and Veer Kumar** under my guidance and supervision for the semester VI, 3rd year of Bachelor of Engineering in Electronics and Telecommunication during the academic year 2020-2021.

Prof. POONAM SONAR Prof. S.D Deshmukh HEAD OF DEPARTMENT

EXTC

DATE: 15/5/2021

PLACE: MUMBAI



## ACKNOWLEDGMENT

### We would like to express our deepest gratitude and thanks to our supervisor, Prof. Poonam Sonar, who suggested the idea of the diabetes classifier and who gave us the needed information to start working on the project. Also, we would like to thank her for being supportive and for her guidance through this semester and for giving us the necessary advices to be able to realize this project. We are really grateful to her contribution. Moreover, we would like to thank her for her supervising methodology that made our tasks easier and motivated us through this period. Finally, we would like to thank all our friends for their support. They have supported us during the difficult moments and encouraged us in carrying out this project work.



**TABLE OF CONTENTS**

1. Abstract
2. Introduction
   1. Problem: The Diabetes Crisis
   2. Diabetes Classifier
3. Concept used
   1. Machine learning
   2. Classification Algorithm(Random Forrest Classifier)
4. Application used
   1. Python
   2. Pycharm
   3. Streamlit

1. Algorithm
2. Program
3. Concept used in program
4. Output
5. Conclusion

10.References



**1.ABSTRACT**

Diabetes is taken into account together of the deadliest and chronic disease that causes a rise in glucose. Polygenic disease is that the kind wherever the exocrine gland doesn't manufacture hypoglycaemic agent in line with International polygenic disease Federation 382 million individuals live with polygenic disease across the world. By 2035, this will be doubled as 592 million. Diabetes mellitus or just sickness may be a disease caused due to the rise of blood glucose level. Many difficulties might occur if the diabetes remains untreated and unidentified by the doctor. The complications are excretory organ injury, typically resulting in chemical analysis, eye damage that may end in visual impairment, or associate degree enhanced risk for cardiopathy or stroke. The tedious identifying methodology ends up in visiting of a patient to a diagnostic center and consulting the doctor for more treatment. Rise in machine learning approaches solves this essential draw back. The objective of this mini project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm in machine learning technique. Random Forest algorithms are often used for each classification and regression tasks and also it is a type of ensemble learning method. The accuracy level is greater when compared to other algorithms. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, efficiently and most importantly, instantly.



## INTRODUCTION

#### The Diabetes Crisis

Over 30 million have now been diagnosed with diabetes in India. The CPR (Crude prevalence rate) in the urban areas of India is thought to be 9 per cent.In rural areas, the prevalence is approximately 3 per cent of the total population.The population of India is now more than 1000 million: this helps to give an idea of the scale of the problem.The estimate of the actual number of diabetics in India is around 40 million.This means that India actually has the highest number of diabetics of any one country in the entire world. IGT (Impaired Glucose Tolerance) is also a mounting problem in India.The prevalence of IGT is thought to be around 8.7 per cent in urban areas and 7.9 per cent in rural areas, although this estimate may be too high. It is thought that around 35 per cent of IGT sufferers go on to develop type 2 diabetes, so India is genuinely facing a healthcare crisis.

#### Diabeates Classifier?

Classification strategies are broadly used in the medical field for classifying data into different classes according tosome constrains comparatively an individual classifier. Diabetes is an illness which affects the ability of the body inproducing the hormone insulin, which in turn makes the metabolism of carbohydrate abnormal and raise the levels of glucose in the blood. In Diabetes a person generally suffers from high blood sugar. Intensify thirst, Intensify hungerand Frequent urination are some of the symptoms caused due to high blood sugar. Many complications occur if dia-betes remains untreated. Experiments for diagnosing the diseases using various classification algorithmsof machine learning approaches like J48, SVM, Naive Bayes, Decision Tree, Decision Table etc. as researchers have proved that machine-learning algorithms works better in diagnosing different diseases. Data Mining and Machine learning algorithms gain its strength due to the capability of managing a large amount of data to combine data from several different sources and integrating the background information in the mini project.We will be using Random Forrest Classifier while developing our Machine learning classification model.



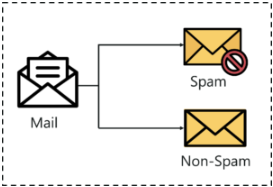
## 3. CONCEPT USED

**3.1 MACHINE LEARNING**

**Machine learning** (**ML**) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a field of study within machine learning, and focuses on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

**3.2 Classification Algorithm:**

Classification is a process of categorizing a given set of data into classes, It can be performed on both structured or unstructured data. The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories.The classification predictive modeling is the task of approximating the mapping function from input variables to discrete output variables. The main goal is to identify which class/category the new data will fall into.



Let us try to understand this with a simple example.

Diabetes disease detection can be identified as a classification problem, this is a binary classification since there can be only two classes i.e has diabetes disease or does not have diabetes disease. The classifier, in this case, needs training data to understand how the given input variables are related to the class. And once the classifier is trained accurately, it can be used to detect whether heart disease is there or not for a particular patient.



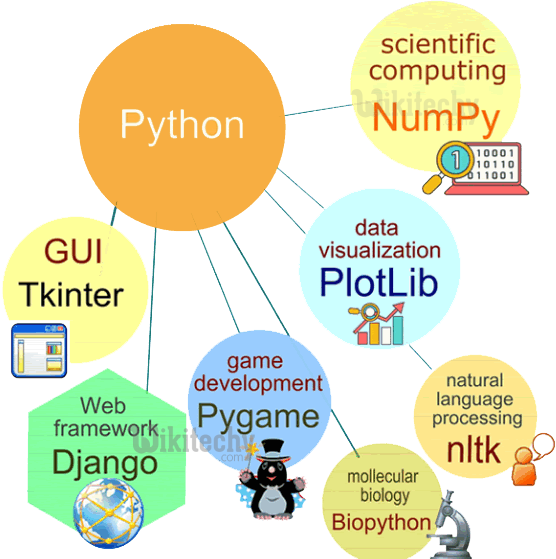
## APPLICATION USED

#### 4.1 WHAT IS PYTHON ?

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together.

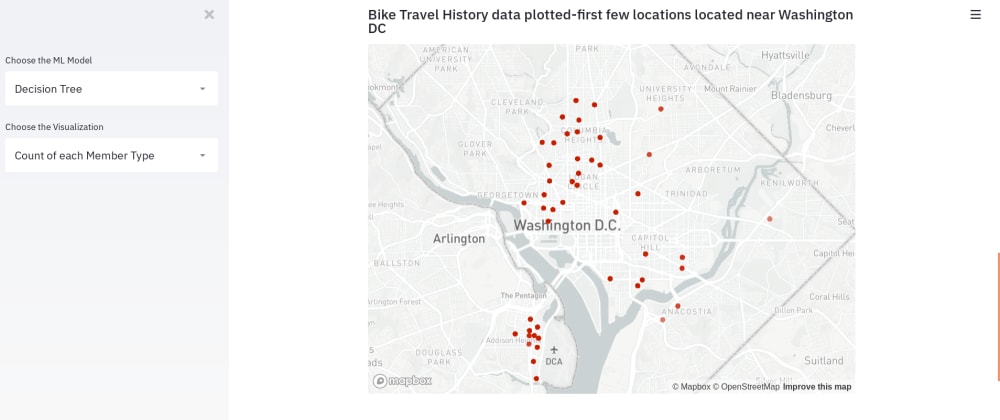
Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance.

Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed.



#### 4.2WHAT IS STEAMLIT

[Streamlit](https://streamlit.io/) is an open-source Python library that makes it easy to create and share beautiful, custom web apps for machine learning and data science. In just a few minutes you can build and deploy powerful data apps

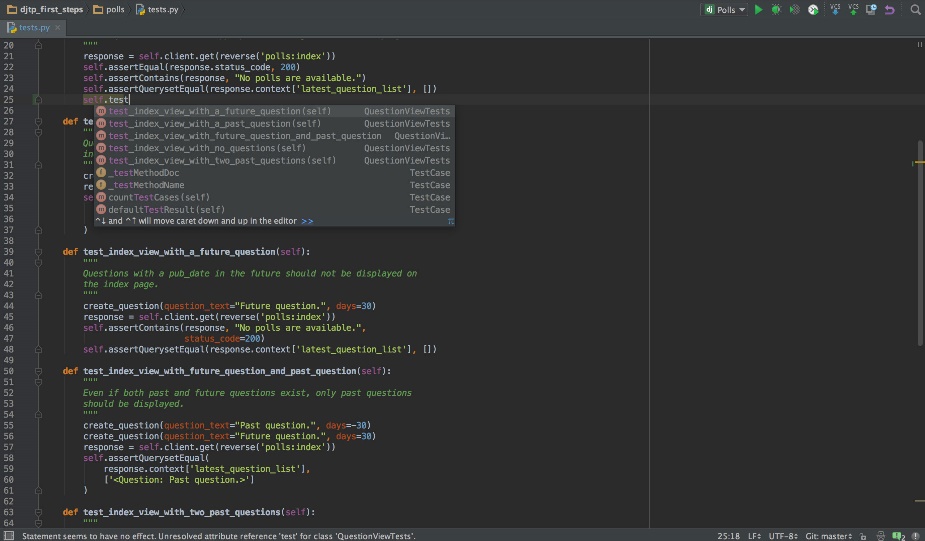




**4.3 WHAT IS PYCHARM?**

**PyCharm** is an [integrated development environment](https://en.wikipedia.org/wiki/Integrated_development_environment) (IDE) used in [computer programming](https://en.wikipedia.org/wiki/Computer_programming), specifically for the [Python](https://en.wikipedia.org/wiki/Python_(programming_language)) language. It is developed by the [Czech](https://en.wikipedia.org/wiki/Czech_Republic) company [JetBrains](https://en.wikipedia.org/wiki/JetBrains) (formerly known as IntelliJ).[[5]](https://en.wikipedia.org/wiki/PyCharm#cite_note-5) It provides code analysis, a graphical debugger, an integrated unit tester, integration with [version control systems](https://en.wikipedia.org/wiki/Revision_control) (VCSes), and supports web development with [Django](https://en.wikipedia.org/wiki/Django_(web_framework)) as well as [data science](https://en.wikipedia.org/wiki/Data_science) with [Anaconda](https://en.wikipedia.org/wiki/Anaconda_(Python_distribution)).[[6]](https://en.wikipedia.org/wiki/PyCharm#cite_note-6)

PyCharm is [cross-platform](https://en.wikipedia.org/wiki/Cross-platform), with [Windows](https://en.wikipedia.org/wiki/Windows), [macOS](https://en.wikipedia.org/wiki/MacOS) and [Linux](https://en.wikipedia.org/wiki/Linux) versions. The Community Edition is released under the [Apache License](https://en.wikipedia.org/wiki/Apache_License),[[7]](https://en.wikipedia.org/wiki/PyCharm#cite_note-community-7) and there is also Professional Edition with extra features – released under a [proprietary license](https://en.wikipedia.org/wiki/Proprietary_software).

**** 



## Algorithm

## Code the program using python in pycharm.

## Import necessary libraries including sklearn,pandas and streamlit

## Read the ‘diabetes.csv’ file into the script

## Split the the train and test data,we use a 25/85 (train/test)split over here.

## Get user input regarding the various parameteres like pregnanacies,glucose,blood pressure,insulin,skin thickness, BMI and age,this user data will be used to make predictions by our classiier model

## Now,we use the RandomForrestClassifier,which trains and tests our model,following which the model will

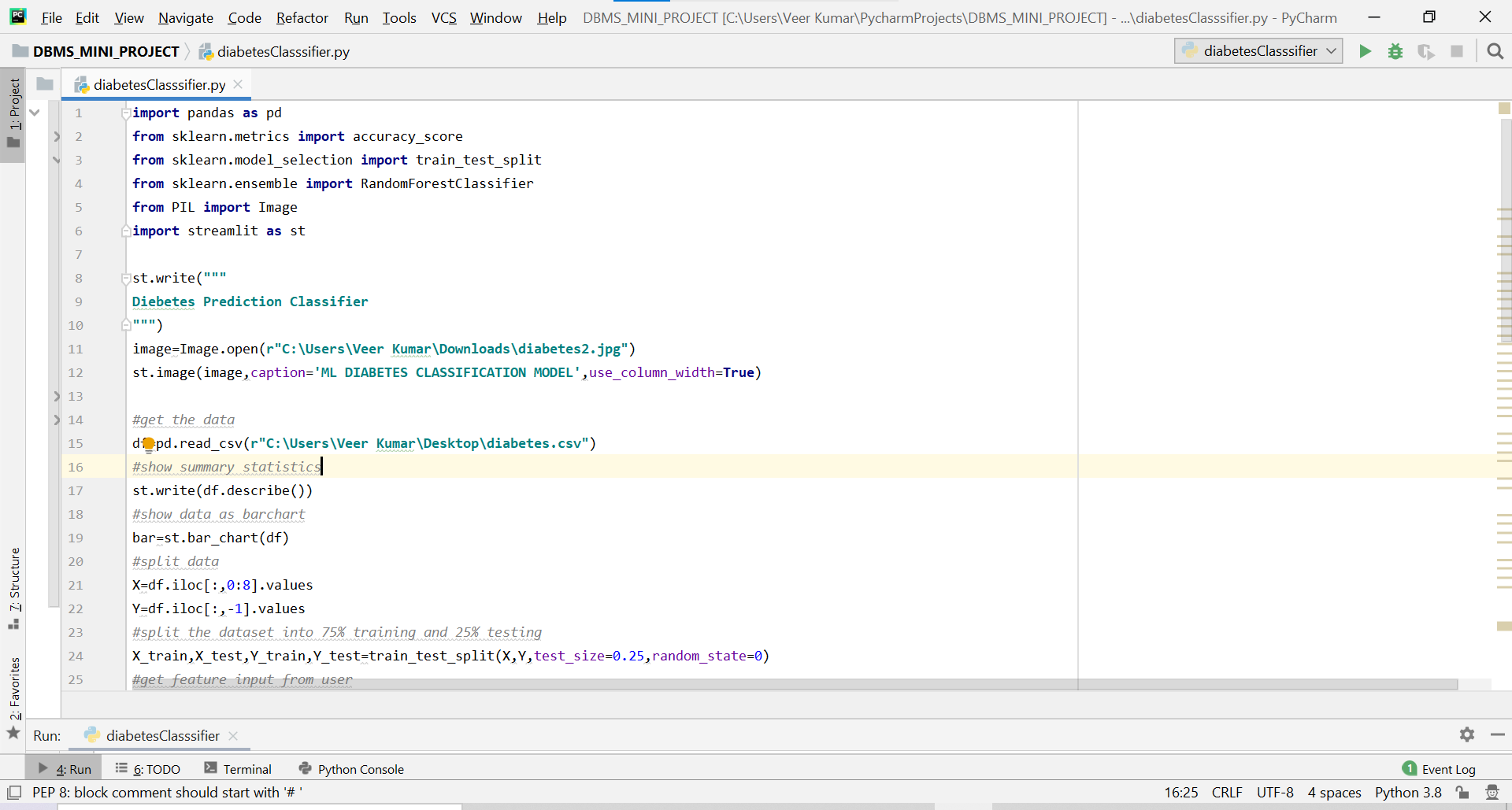
## Make the predictions and give us a model accuracy score as well.

## After model metrics,finally the result will be in the binary classification format,it will either a “1” which means positive for diabetes or “0” which indicates negative for diabetes.



1. **PROGRAM**

**import** pandas **as** pd  
**from** sklearn.metrics **import** accuracy\_score  
**from** sklearn.model\_selection **import** train\_test\_split  
**from** sklearn.ensemble **import** RandomForestClassifier  
**from** PIL **import** Image  
**import** streamlit **as** st  
  
st.write(**"""  
Diebetes Prediction Classifier  
"""**)  
image=Image.open(**r"C:\Users\Veer Kumar\Downloads\diabetes2.jpg"**)  
st.image(image,caption=**'ML DIABETES CLASSIFICATION MODEL'**,use\_column\_width=**True**)  
  
*#get the data*df=pd.read\_csv(**r"C:\Users\Veer Kumar\Desktop\diabetes.csv"**)  
*#show summary statistics*st.write(df.describe())  
*#show data as barchart*bar=st.bar\_chart(df)  
*#split data*X=df.iloc[:,0:8].values  
Y=df.iloc[:,-1].values  
*#split the dataset into 75% training and 25% testing*X\_train,X\_test,Y\_train,Y\_test=train\_test\_split(X,Y,test\_size=0.25,random\_state=0)  
*#get feature input from user***def** get\_user\_input():  
 pregnancies=st.sidebar.slider(**'pregnancies'**,0,17,3)  
 glucose=st.sidebar.slider(**'glucose'**,0,199,117)  
 blood\_pressure=st.sidebar.slider(**'blood\_pressure'**,0,122,72)  
 skin\_thickness=st.sidebar.slider(**'skin\_thickness'**,0,99,23)  
 insulin=st.sidebar.slider(**'insulin'**,0.0,846.0,30.0)  
 BMI=st.sidebar.slider(**'BMI'**,0.0,67.1,32.0)  
 DPF=st.sidebar.slider(**'DPF'**,0.078,2.42,0.3725)  
 age=st.sidebar.slider(**'age'**,21,81,29)  
  
 *#store a dictionary into a variable* user\_data={**' pregnancies'**: pregnancies,**'glucose'**:glucose,**'blood\_pressure'**:blood\_pressure,**'skin\_thickness'**: skin\_thickness,**'insulin'**:insulin,**'BMI'**:BMI,**'DPF'**:DPF,**'age'**:age}  
 *#Transform the data into dataframe* features=pd.DataFrame(user\_data,index=[0])  
 **return** features  
*#store user input into variable*user\_input=get\_user\_input()  
st.subheader(**'User input:'**)  
st.write(user\_input)  
*#create and train the model*RandomForestClassifier=RandomForestClassifier()  
RandomForestClassifier.fit(X\_train,Y\_train)  
*#show model metrics*st.subheader(**'Model Test Accuracy Score:'**)  
st.write(str(accuracy\_score(Y\_test,RandomForestClassifier.predict(X\_test))\*100)+**'%'**)  
*#store model predictions in a variable*prediction=RandomForestClassifier.predict(user\_input)  
*#set a subheader and display classification:*st.subheader(**'Classification:'**)  
st.write(prediction)

****

****

****



**7.CONCEPTS USED IN PROGRAM**

#### OBJECT ORIENTED PROGRAMMING

1. Python is a multi-paradigm programming language. Meaning, it supports different programming approach.
2. One of the popular approach to solve a programming problem is by creating objects. This is known as Object-Oriented Programming (OOP)

#### INHERITANCE

Inheritance is the capability of one class to derive or inherit the properties from some another class. The benefits of inheritance are:

* + 1. It represents real-world relationships well.
    2. It provides **reusability** of a code. We don’t have to write the same code again and again. Also, it allows us to add more features to a class without modifying it.
    3. It is transitive in nature, which means that if class B inherits from another class A, then all the subclasses of B would automatically inherit from class A.

#### FUNCTIONS

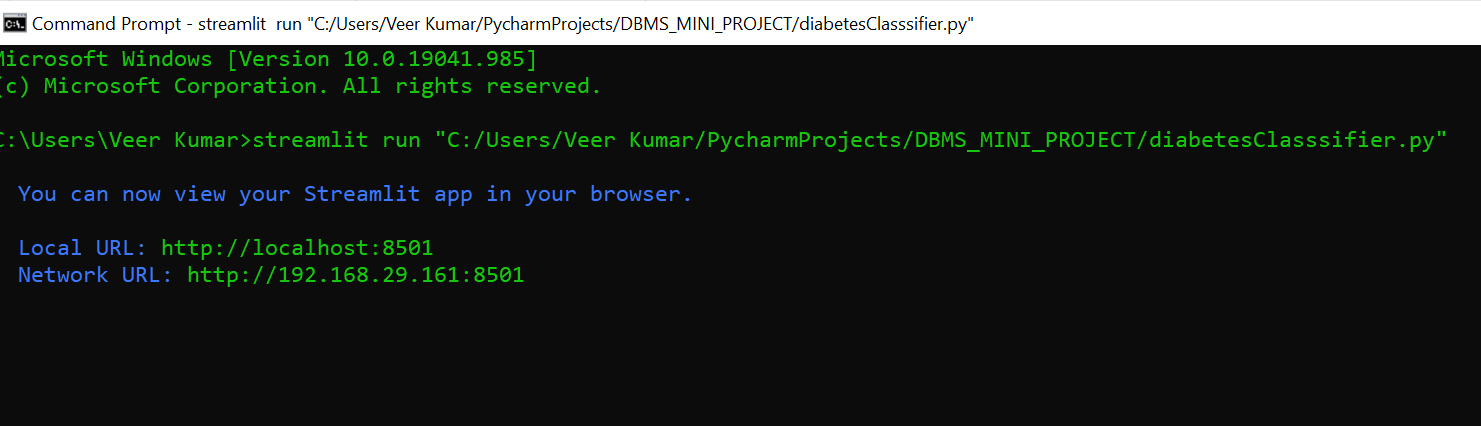
A function is a block of code which only runs when it is called.You can pass data, known as parameters, into a function.A function can return data as a result. Arguments are specified after the function name, inside the parentheses. You can add as many arguments as you want, just separate them with a comma.

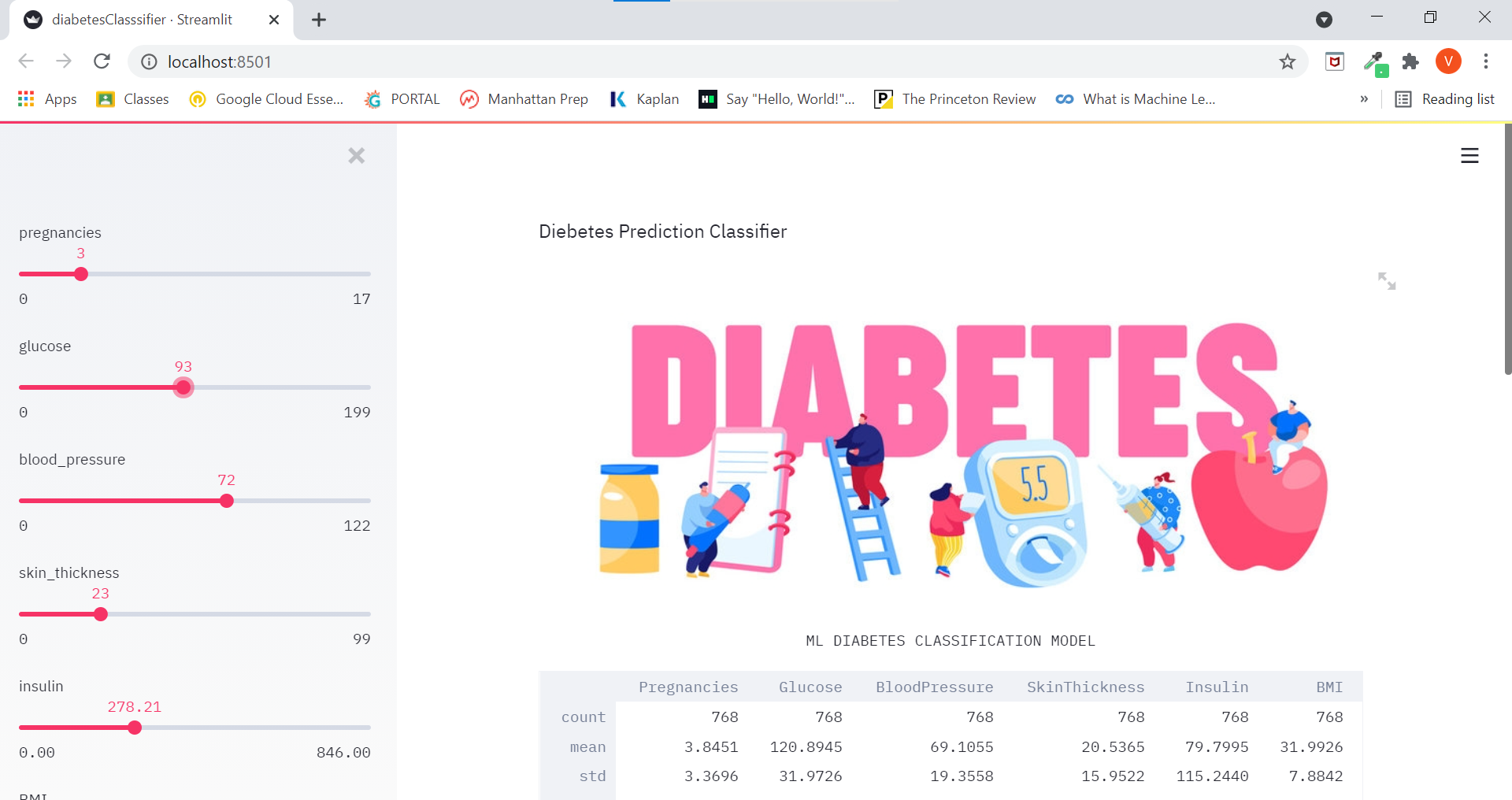
#### SK LEARN ML LIBRARY:

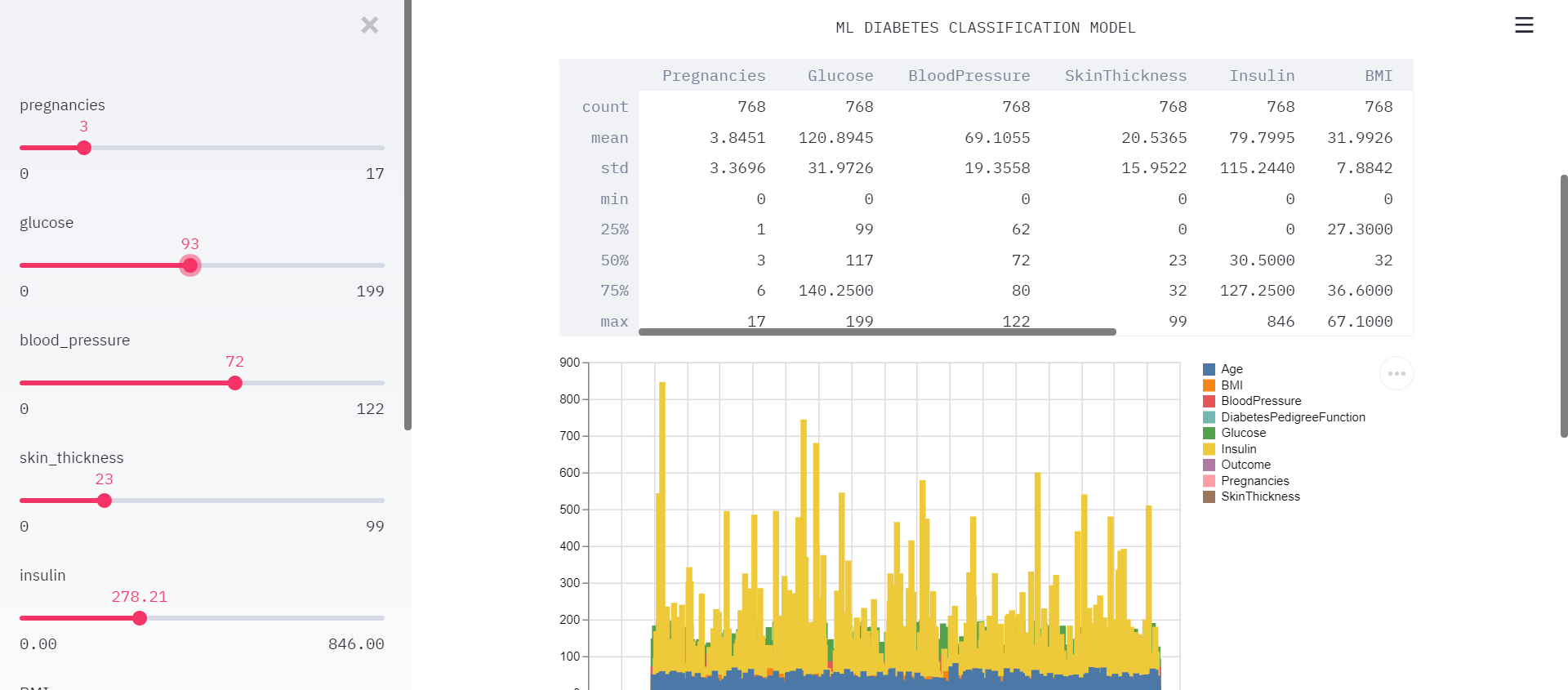
#### Scikit-learn (formerly scikits.learn and also known as sklearn) is a [free software](https://en.wikipedia.org/wiki/Free_software) [machine learning](https://en.wikipedia.org/wiki/Machine_learning) [library](https://en.wikipedia.org/wiki/Library_(computing)) for the [Python](https://en.wikipedia.org/wiki/Python_(programming_language)) [programming language](https://en.wikipedia.org/wiki/Programming_language). It features various [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and [clustering](https://en.wikipedia.org/wiki/Cluster_analysis) algorithms including [support vector machines](https://en.wikipedia.org/wiki/Support_vector_machine), [random forests](https://en.wikipedia.org/wiki/Random_forests), [gradient boosting](https://en.wikipedia.org/wiki/Gradient_boosting), [*k*-means](https://en.wikipedia.org/wiki/K-means_clustering) and [DBSCAN](https://en.wikipedia.org/wiki/DBSCAN), and is designed to interoperate with the Python numerical and scientific libraries [NumPy](https://en.wikipedia.org/wiki/NumPy) and [SciPy](https://en.wikipedia.org/wiki/SciPy).



**8.OUTPUT**

****









## 9.CONCLUSION

* An accurate and efficient diabetes classifer has been developed which achieves highly satisfactory metrics
* This project uses recent techniques in the field of Machine learning classification algorithms,and also employs data analytics tools like pandas and numpy in order to derive insights from csv file(excel sheet used)
* This can be used as an effective model to classify the presence or absence of diabetes before making an informed medical decision.A mdeical professional/doctors can make use of this model in order to decide the further line of treatment to be administered to the patient by modifying various feature parameters and simulating different outcomes,
* An important scope would be to train the classifer model for to achieve a higher accuracy.We know very well that a false negative is the most undesirable result,since it puts a vulnerable patient under the impression that he/she is free from disease,consequently depriving them of the required medical treatment and associated procedures,thus minimizing false negative results would be the primary scope of improvement for this mini project.



**10.REFERENCES**

* + - 1. <https://en.wikipedia.org/wiki/Python_%28programming_language%29>
      2. <https://en.wikipedia.org/wiki/PyCharm>
      3. https://www.analyticsvidhya.com/blog/2015/01/scikit-learn-python-machine-learning-tool/
      4. <https://www.researchgate.net/publication/259235118_Random_Forests_and_Decision_Trees>
      5. https://towardsdatascience.com/building-a-machine-learning-classifier-model-for-diabetes-4fca624daed0