## A Mini- Project Report

**on**

## “Diabetes Prediction Tool”

Submitted to the

Pune Institute of Computer Technology, Pune In partial fulfillment for the award of the Degree of Bachelor of Engineering

in

#### Information Technology

by

|  |  |
| --- | --- |
| Vedant Puranik | 43152 |
| Sahil Naphade | 43141 |

Under the guidance of

## Prof. R. R. Chhajed



#### Department Of Information Technology

Pune Institute of Computer Technology College of Engineering

Sr. No 27, Pune-Satara Road, Dhankawadi, Pune - 411 043.

**2020-2021**

## CERTIFICATE

This is to certify that the project report entitled

**DIABETES PREDICTION TOOL**

**Submitted by**

|  |  |
| --- | --- |
| Vedant Puranik | 43152 |
| Sahil Naphade | 43141 |

is a bonafide work carried out by them under the supervision of Prof. R. R.Chhajed and it is approved

for the partial fulfillment of the requirement of **Computer Laboratory -X** for the award of the Degree of Bachelor of Engineering (Information Technology).

|  |  |
| --- | --- |
| **Prof. R. R. Chhajed** | **Dr. A. M. Bagade** |
| Lab Teacher | Head of Department |
| Department of Information Technology | Department of Information Technology |

**Place:** PICT, Pune

**Date:** 29/05/2021

II

## ACKNOWLEDGEMENT

We thank everyone who have helped and provided valuable suggestions for successfully creating a wonderful project.

We are very grateful to our guide, Prof. R. R. Chhajed, Head of Department Dr. A. M. Bagade and our principal Dr. P. T. Kulkarni. They have been very supportive and have ensured that all facilities remained available for smooth progress of the project.

We would like to thank our professor A. C. Karve and Prof. R. R. Chhajed for providing very valuable and timely suggestions and help.

Vedant Puranik

Sahil Naphade

III

## ABSTRACT

The recent covid-19 pandemic has taken its toll on the population across the world. Especially, the elderly and the people with comorbidities like Asthma and Diabetes are especially at risk. Having diabetes also increase the chances of contracting the Mucormycosis, a.k.a. the black fungus.

Having such comorbidities also delay the healing procedure of covid-19 infected patients and hence, should not go unknown and untreated.

Therefore, through this project, we wish to allow the general population to gain an insight into their diabetes situation to help them manage the disease better; and effectively help them lead a better lifestyle.

IV

## LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| Figure Number | Figure Title | Page Number |
| 1 | Flow of the system | 10 |
| 2 | Home page screenshot | 15 |
| 3 | Predictor page screenshot | 15 |

V

## LIST OF TABLES

|  |  |  |
| --- | --- | --- |
| Table Number | Table Title | Page Number |
| 1 | Empirical study of the algorithms tested | 16 |
| 2 | Confusion matrix | 16 |

**TOPICS**

1. **Introduction**

COVID-19 remains the biggest pandemic the world has ever witnessed. India is at the forefront with a huge number of cases and COVID-related deaths. As per reports, elderly people are highly susceptible to the Corona virus and those who have diabetes are even more prone. In general, people with diabetes are more likely to have more severe symptoms and complications when infected with any virus. Your risk of getting very sick from COVID-19 is likely to be lower if your diabetes is well-managed.

Therefore, finding out if a person has diabetes could be a very important step in understanding how severe would your COVID infection be, if you are ever affected by the virus. Lab methods generally take a while to deliver reports stating your diabetic status.

Hence, to provide a better intuition about the same problem, we propose a diabetes prediction tool. This tool takes 8 different inputs from the user and on the basis of that tells the user whether he or she is possibly diabetic. This tool makes use of Random Forest, which is a supervised machine learning technique for classifying whether the person has diabetes or not. As the prediction is purely based on numbers, this tool cannot be used to obtain a final verdict and in doubt the user is advised to contact a physician.

1. **Scope and Objective**

Scope of our project is to create a web application with considerable accuracy and speed to predict the condition of diabetes in a patient based on the outcomes of his results on certain parameters. The web application will also provide insights into the data gathered by the exploration of the dataset.

Objective of the project is to educate the populace and provide insights into the probability of one having diabetes and the related conditions. The insights into the data explored will also help in understanding the typical signs of one having diabetes and thus, enabling them to tackle the issue earlier.

1. **System architecture and flow**

We have considered using the Flask framework for the creation of the website. This mainly resorts to the fact that using the same environment of Python for data model and website ensures high cohesion. Flask is a light-weight micro web framework written in Python, which makes use of additional extensions if needed; to provide extra features like form validations, database abstraction layers, etc. It also contains its own light development server and debugger, support for Unit testing, RESTful service dispatching with support for client-side secure cookies.

We have created a website of two pages. First page of the site is an information page, which shows the insights gained after exploration of the data, and a link to the predictor on the second page. First page is available at decorator (/) which redirects to (/home). Second page contains a form which will accept the values input by the user under various heads, which are the features of our dataset, namely Skin Thickness, Blood pressure, BMI, Diabetes Pedigree function, age, etc. This form submits to the function predictor (decorator /predictor). In this function, the form data (the features input by the user) are extracted.

A pickle dump file (.pkl) was created for the trained model. This file basically contains the pretrained model. The values submitted through the form are scaled with one-hot encoding which is defined in the pickle dump file.

These features are then used to predict the value of the output, which outputs 1 for diabetic and 0 for non-diabetic values.

The final value is shown on the same page, just below the submit button.

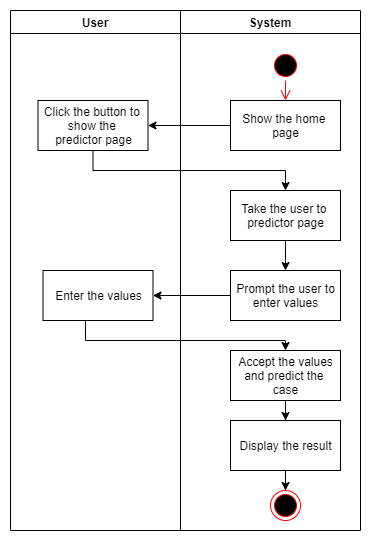


Figure : Flow of the system

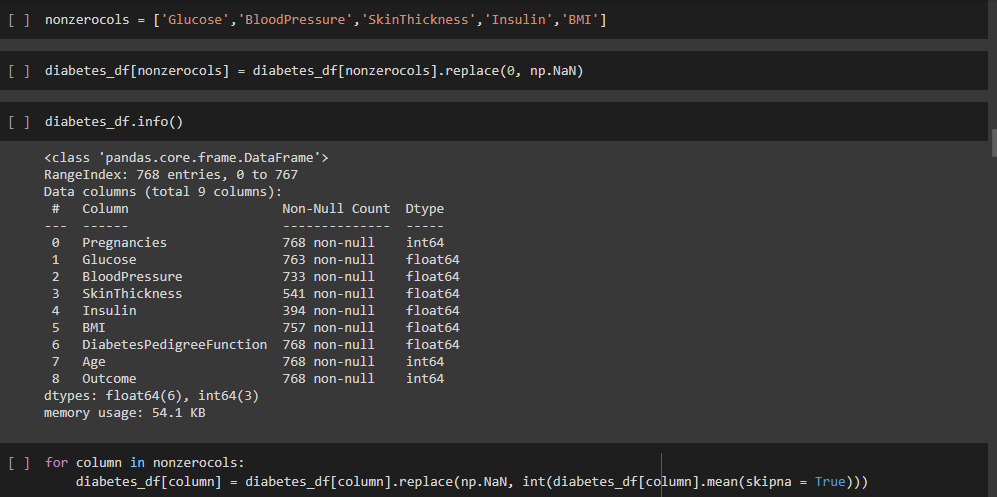
1. **Code and snapshots**

Code for data model:

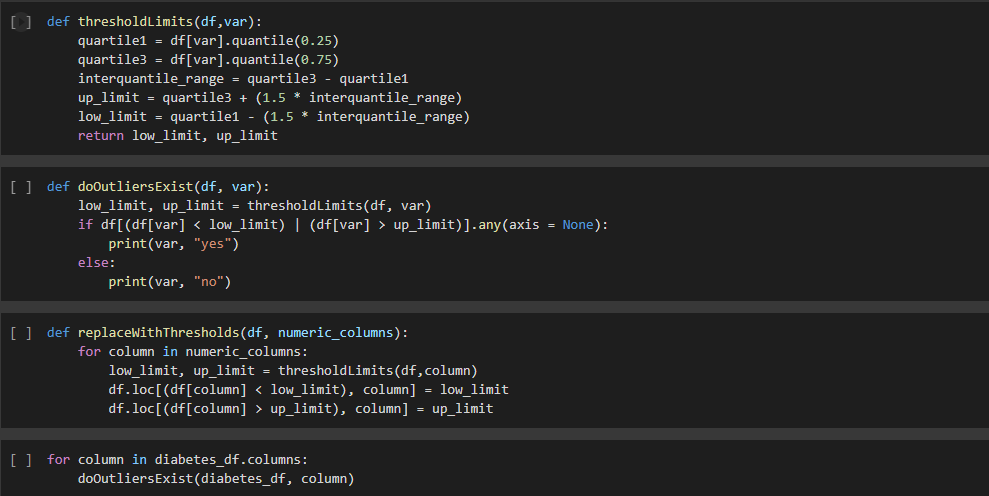
* 1. **Reading data**

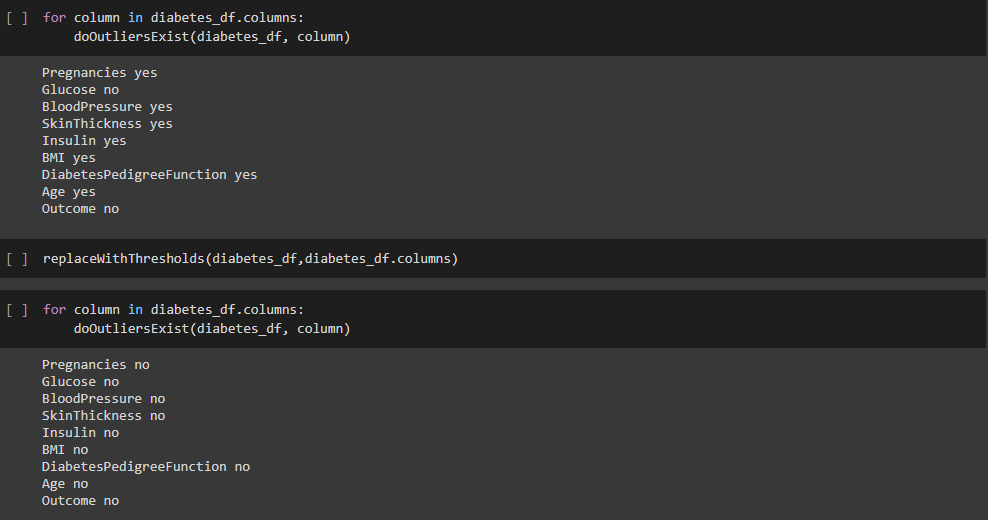


* 1. **Preprocessing**
     1. **Replacing 0’s in columns with average value**

****

* + 1. **Detecting and removing outliers**

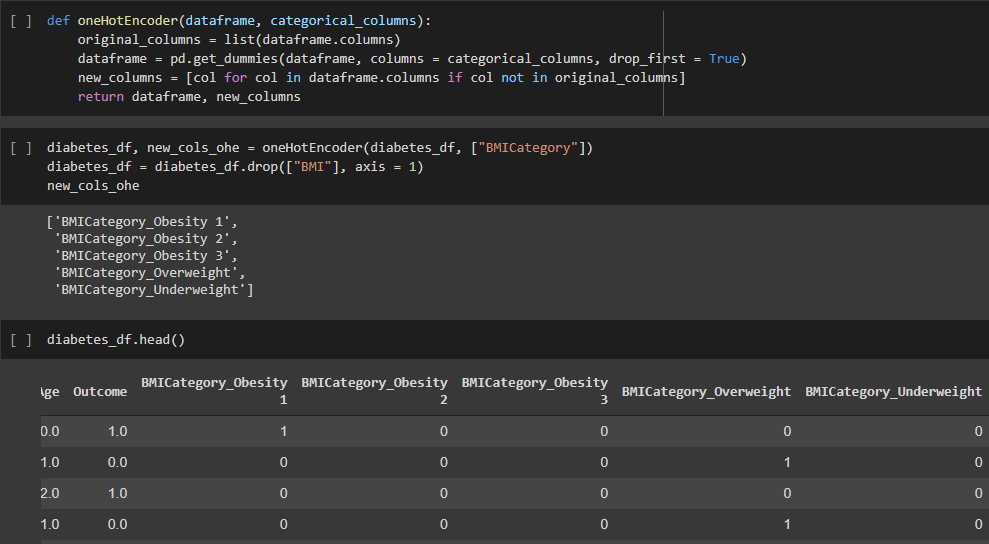




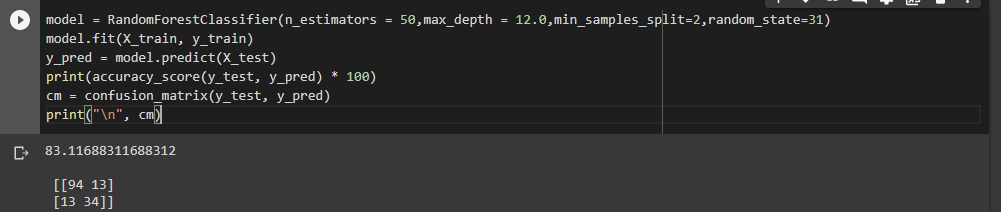
* + 1. **Converting BMI into a categorical column**



* + 1. **One-hot encoding BMI categories**

****

* 1. **Hyperparameter-tuned Model Training and Result**

****

**BACKEND**

* 1. **Code for extracting the values through the form:**

Note: This is the code inside the predictor function.

if request.method == "POST" and form.validate():

pregnancies = form.pregnancies.data

glucose = form.glucose.data

bp = form.bp.data

st = form.st.data

insulin = form.insulin.data

bmi = form.bmi.data

pedigreeFunc = form.diabetesPedigree.data

age = form.bmi.data

* 1. **Code for loading the pickle file and predict the output:**

scalerLoad = pickle.load(open("scaler.pkl", "rb"))

modelLoad = pickle.load(open("randomForestModel.sav", "rb"))

featureList = np.array([pregnancies, glucose, bp, st, insulin, pedigreeFunc, age]).reshape(1, -1)

featureList = scalerLoad.transform(featureList)

finalFeatures = np.concatenate((featureList, BMIVar), axis = 1)

value = modelLoad.predict(finalFeatures)

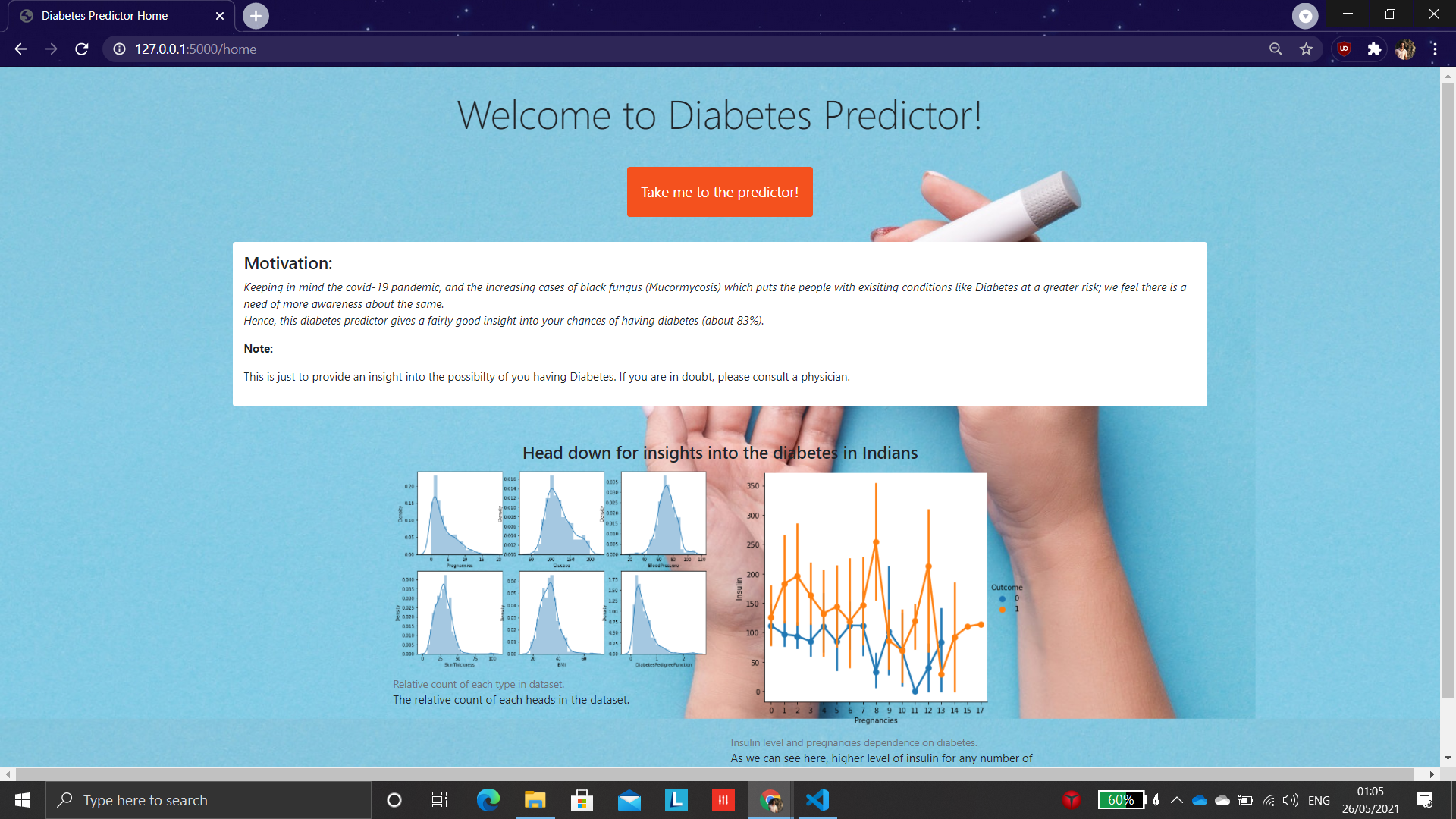
print("Value = {}".format(value))

The modelLoad is the loaded model from the pickle file; whereas scalerLoad variable is the scalar defined and trained for the dataset using sklearn.

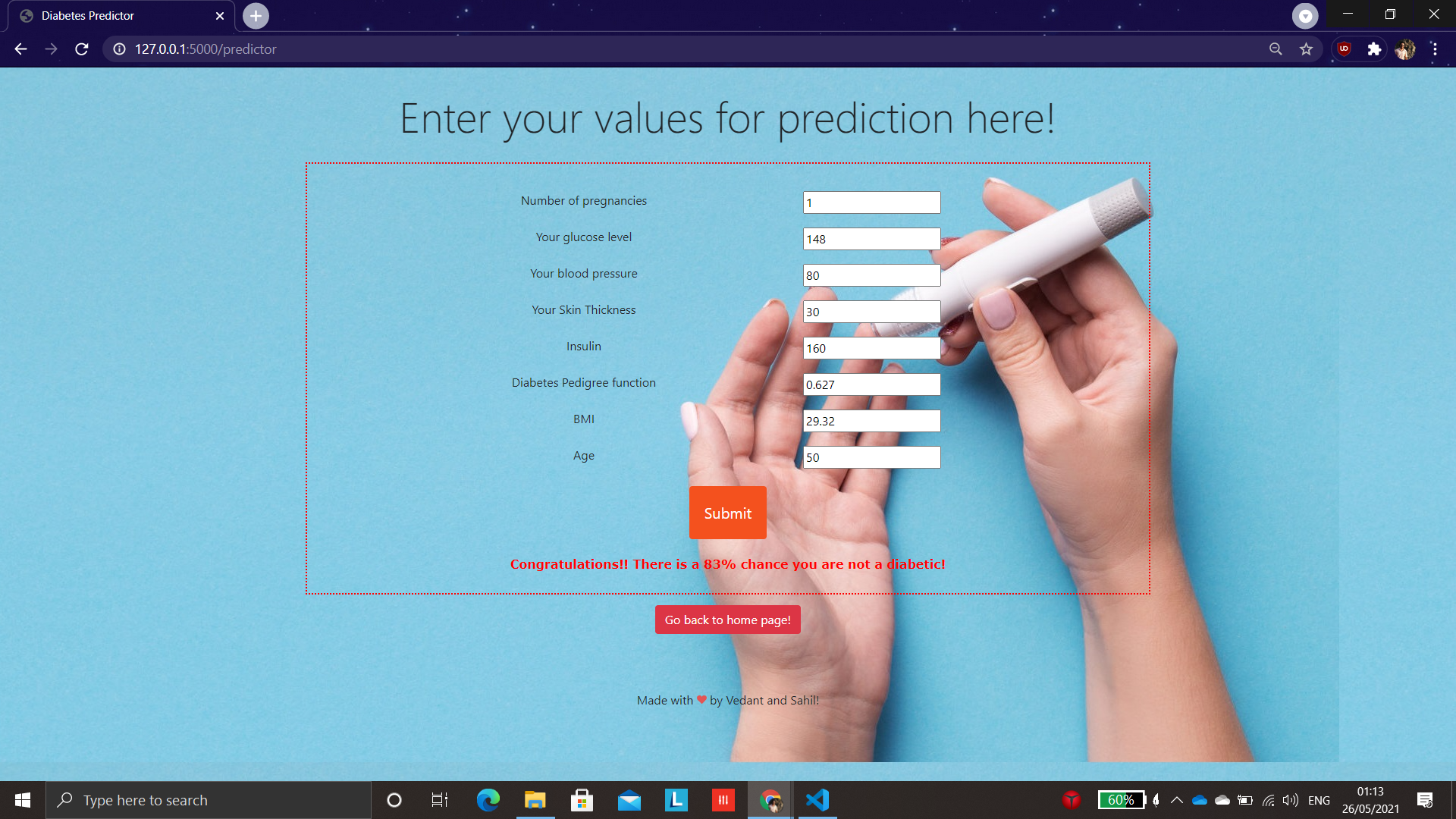
Finally, the numpy array of features; finalFeatures; is used to predict the outcome from the given data.

**Snapshots of the project:**

* 1. **UI**
     1. **Homepage**



* + 1. **Predictor page**



1. **Results**

To get an estimate of the highest accuracy achievable on our dataset, we implemented a number of machine learning algorithms for the same. Out of all the algorithms tried, Random Forest performed the best with a test accuracy of 83%. The open-source dataset we used is called the PIMA Indians Diabetes Dataset, which is a standard dataset used for diabetes related problem statements.

|  |  |
| --- | --- |
| **Model** | **Test Accuracy** |
| Naive Bayes | 33.11% |
| Decision Tree | 70.12% |
| K-Nearest Neighbors | 79.22% |
| XGBoost | 79.8% |
| SVM | 80.5% |
| Logistic Regression | 81.16% |
| Random Forest | 83.11% |

Following is an empirical analysis of several supervised machine learning algorithms we tried:

As Random Forest performed the best as compared to others, we used other metrics to get a better idea about the model’s classification ability, for which we used confusion matrix.

**Confusion Matrix**:

|  |  |  |
| --- | --- | --- |
|  | Predicted True | Predicted False |
| Actually True | 94 (TP) | 13 (FN) |
| Actually False | 13 (FP) | 34 (TN) |

Precision = TP/(TP+FP) = 94/(94+13) = 0.878

Recall = TP/(TP+FN) = 94/(94+13) = 0.878

F1-score = 2\*precision\*recall/(precision + recall) = 0.878

1. **Conclusion and Future Scope**

Hence, as COVID-19 remains the biggest health problem faced by the world today, knowing whether or not a person has diabetes can help him assess his health and be prepared in advance for steps to be taken, as diabetics face more complications than non-diabetics. To get a numerical prediction, we employed several supervised machine learning algorithms, out of which Random Forest proved to deliver the highest accuracy of 83%. This tool uses traditional machine learning. If more data is available, using neural networks for classification would be a better solution in the future. Additionally, also making an integrated app for obtaining preliminary predictions of several diseases related to kidney, liver, heart, etc. would prove to provide a one-stop solution to a number of problems.

**GitHub link for project:** <https://github.com/ItsNaSa/Diabetes-Predictor>

1. **References**
2. <https://machinelearningmastery.com/>
3. <https://scikit-learn.org/stable/>
4. <https://www.datacamp.com/community/tutorials/pickle-python-tutorial>
5. <https://www.kaggle.com/uciml/pima-indians-diabetes-database>
6. <https://flask.palletsprojects.com/_/downloads/en/1.1.x/pdf/>
7. [Deploy a machine learning model using flask](https://towardsdatascience.com/deploy-a-machine-learning-model-using-flask-da580f84e60c#:~:text=Pickle%20is%20used%20to%20serializing,be%20used%20by%20the%20server.)