PROJECT REPORT

AI – Based Diabetes Prediction Systems

Team Leader : Pandeeswaran C.K. (pandeeswaranckit21@jkkmct.edu.in)

Team Members: Dinesh.M (dineshmit21@jkkmct.edu.in)

Krishnan.S (krishnansit21@jkkmct.edu.in)

Parthiban.M (parthibanmit21@jkkmct.edu.in)

Naveen.S (naveensit21@jkkmct.edu.in)

ABSTRACT

Diabetes Prediction:

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 imply 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 three different supervised machine learning methods including: SVM, Logistic regression, KNN. This project also aims to propose an effective technique for earlier detection of the diabetes disease using Machine learning algorithms and end to end deployment using flask.

CONTENTS

1) INTRODUCTION

Diabetes is a chronic metabolic disorder characterized by elevated blood sugar levels, which can have severe and life-altering consequences if left unmanaged. Early detection of diabetes risk is crucial, as it allows for timely interventions that can prevent or delay the onset of the disease and its complications. In recent years, the integration of artificial intelligence (AI) and machine learning techniques into healthcare has opened up new possibilities for predicting diabetes risk with higher accuracy and efficiency.

Traditional diabetes risk assessment relies on well-established clinical risk factors such as family history, body mass index (BMI), age, and glucose levels. While these factors are valuable, they may not provide a comprehensive and personalized assessment of an individual's risk. This is where AI-based diabetes prediction systems come into play. These systems leverage large datasets and advanced algorithms to identify hidden patterns and correlations in data that are beyond the capacity of human clinicians. By analyzing a broader range of features, including genetic information, lifestyle habits, and historical health records, AI models can provide more accurate and individualized predictions of diabetes risk.

This introduction sets the stage for the exploration of AI-based diabetes prediction systems, which have the potential to revolutionize diabetes risk assessment and contribute significantly to the global efforts to combat this pervasive and challenging health condition. In the following sections, we will delve into the methodology, advantages, challenges, and real-world applications of these innovative AI systems.

Causes Of Diabetes

Genetic factors are the main cause of diabetes. It is caused by at least two mutant genes in the chromosome 6, the chromosome that affects the response of the body to various antigens. Viral infection may also influence the occurrence of type 1 and type 2 diabetes .Studies have shown that infection with viruses such as rubella, Coxsackie virus, mumps ,hepatitis B virus, and cytomegalovirus increase the risk of developing diabetes.

Types of Diabetes

Type 1

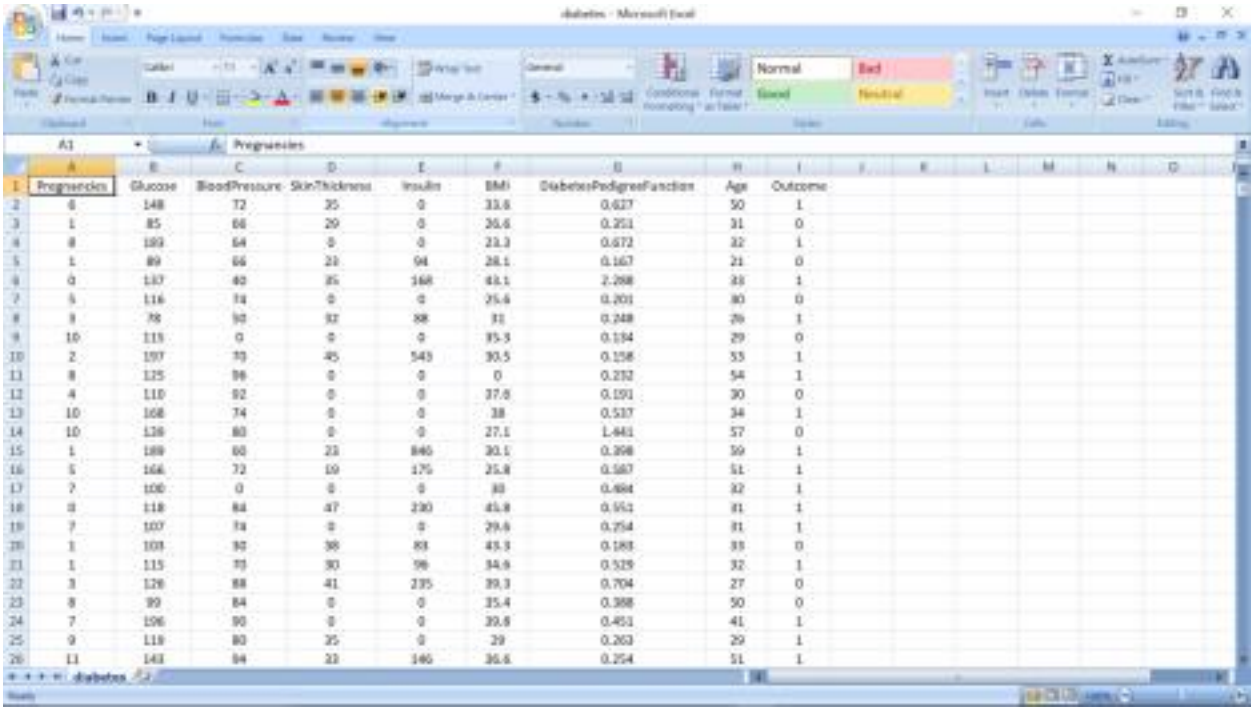
Type 1 diabetes means that the immune system is compromised and the cells fail to produce insulin in sufficient amounts. There are no eloquent studies that prove the causes of type 1 diabetes and there are currently no known methods of prevention.

Type 2

Type 2 diabetes means that the cells produce a low quantity of insulin or the body can’t use the insulin correctly. This is the most common type of diabetes, thus affecting 90%of persons diagnosed with diabetes. It is caused by both genetic factors and the manner of living.

2.Dataset

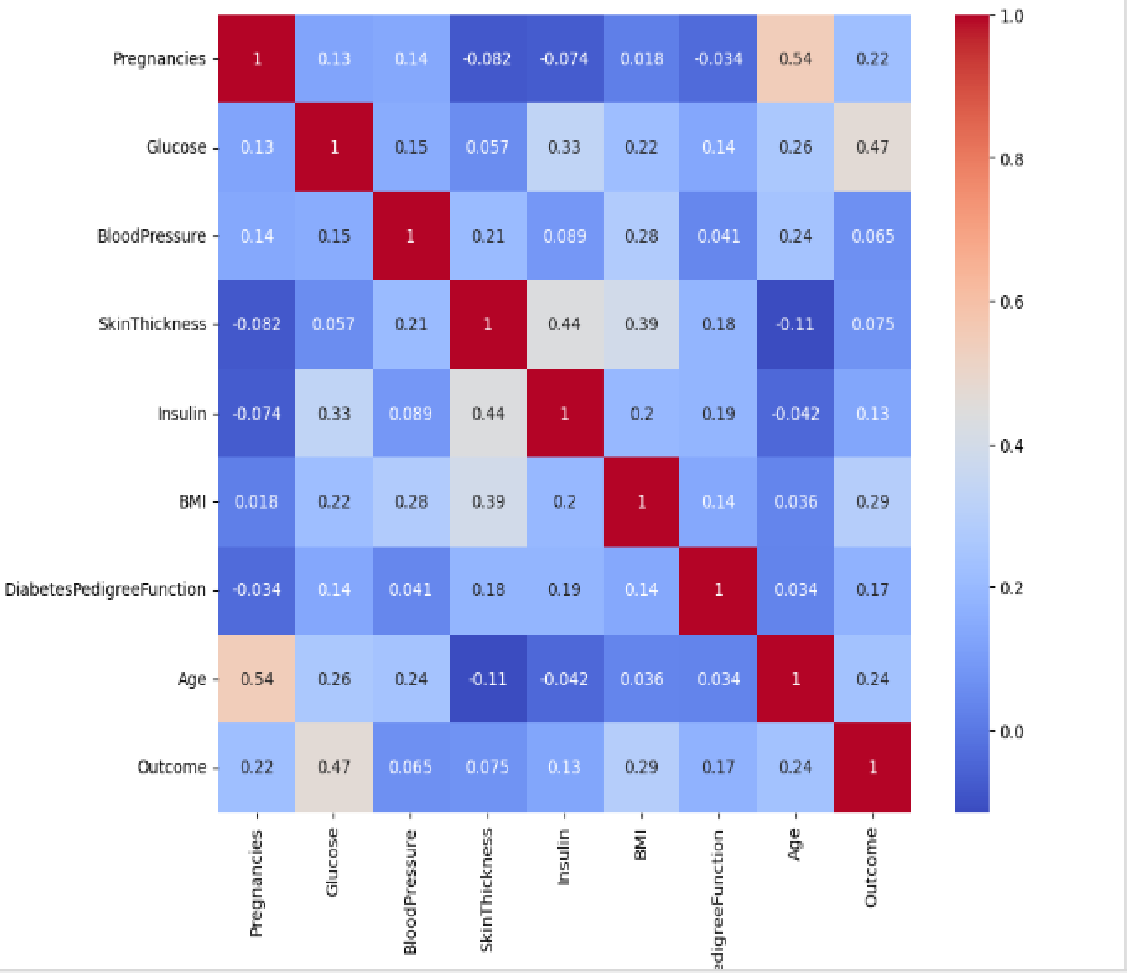
The dataset collected is originally from the given dataset. It consists of several medical analyst variables and one target variable .The objective of the dataset is to predict whether the patient has diabetes or not. The dataset consists of several independent variables and one dependent variable, i.e., the outcome .Independent variables include the number of pregnancies the patient has had their BMI ,insulin level, age, and so on.



A1 fx Pregnancies

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0
8	99	84	0	0	35.4	0.388	50	0
7	196	90	0	0	39.8	0.451	41	1
9	119	80	35	0	29	0.263	29	1
11	143	94	33	146	36.6	0.254	51	1

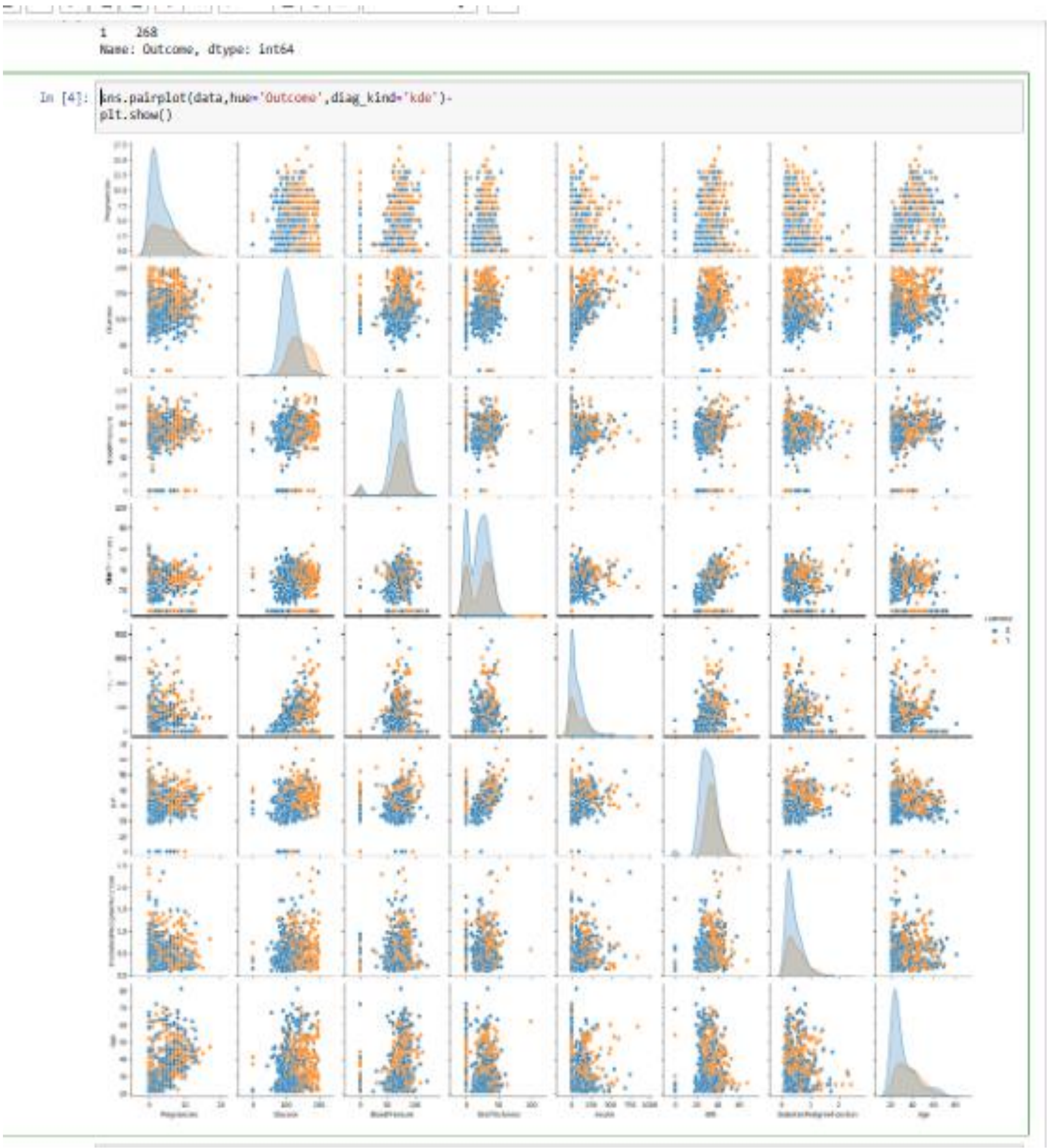
Correlation matrix :-



Correlation matrix

It is easy to see that there is no single feature that has a very high correlation with our outcome value. Some of the features have a negative correlation with the outcome value and some have positive.

Bar Plot for Outcome Class



```
1    268
Name: Outcome, dtype: int64
```

```
In [4]: sns.pairplot(data,hue='Outcome',diag_kind='kde')-
        plt.show()
```

Proposed Methods

Dataset collection

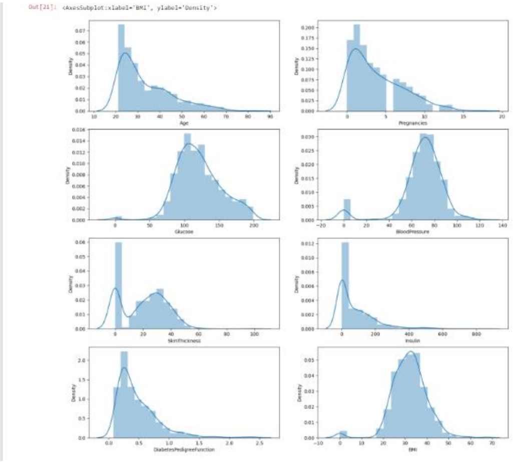
It includes data collection and understanding the data to study the hidden patterns and trends which helps to predict and evaluating the results. Dataset carries1405 rows i.e., total number of data and 10 columns i.e., total number of features. Features include Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age .

Data Pre-processing:

This phase of model handles inconsistent data in order to get more accurate and precise results like in this dataset Id is inconsistent so we dropped the feature .This dataset doesn’t contain missing values. So, we imputed missing values for few selected attributes like Glucose level, Blood Pressure, Skin Thickness, BMI and Age because these attributes cannot have values zero. Then data was scaled using Standard Scalar. Since there were a smaller number of features and important for prediction so no feature selection was done.

Scaling and Normalization:

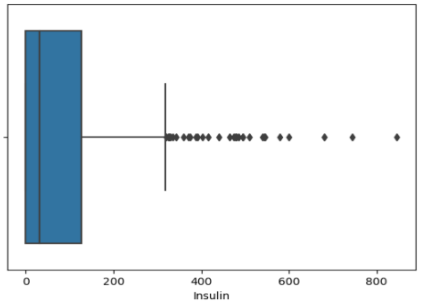
We performed feature scaling by normalizing the data from 0 to 1 range, which boosted the algorithm’s calculation speed .scaling means that you're transforming your data so that it fits within a specific scale, like 0-100 or 0-1. You want to scale data when you're using methods based on measures of how far a part data points are, like support vector machines (SVM) or k-nearest neighbours (KNN).With these algorithms, a change of "1" in any numeric feature is given the same importance.



```
Out[21]: <AxesSubplot:xlabel='BMI', ylabel='Density'>
```

Splitting of data:

After data cleaning and pre-processing, the dataset becomes ready to train and test. In the train/split method, we split the dataset randomly into the training and testing set. For Training we took 1600 sample and for testing we took 400 sample.



Machine learning classifier:

We have developed a model using Machine learning Technique. Used different classifier and ensemble techniques to predict diabetes dataset. We have applied SVM, LR, DT and RF Machine learning classifier to analyse the performance by finding accuracy of each classifier All the classifiers are implemented using scikit learn libraries in python. The implemented classification algorithms are described in next section.

MODELING AND ANALYSIS:

Logistic Regression:

Logistic regression is a machine learning technique used when dependent variables are able to categorize. The outputs obtained by using the logistic regression is based on the available features. Here sigmoidal function is used to categorize the output.

K-Nearest Neighbors:

K-nearest neighbors (KNN) algorithm uses ‘feature similarity’ to predict the values of new data points which further means that the new data point will be assigned a value based on how closely it matches the points in the training set. Predictions are made for a new instance (x) by searching through the entire training set for the K most similar instances (the neighbors) and summarizing the output variable for those Kinstances.

SVM

SVM is supervised learning algorithm used for classification. In SVM we have to identify the right hyper plane to classify the data correctly. In this we have to set correct parameters values. To find the right hyper plane we have to find right margin for this we have choose the gamma value as 0.0001 and rbf kernel. If we select the hyper plane with low margin leads to miss classification.

Decision Tree:

Decision tree is non parametric classifier in supervised learning. In this method all the details are represented in the form of tree, where leaves are corresponds to the class labels and attributes are corresponds to internal node of the tree. We have used Gini Index for splitting the nodes.

Random Forest:

Random forest is an ensemble learning method for classification. This algorithm consists of trees and the number of tree structures present in the data is used to predict the accuracy .Where leaves are corresponds to the class labels and attributes are corresponds to internal node of the tree. Here number of trees in forest used is 100 in number and Gini index is used for splitting the nodes.

Measurements

To find the efficient classifier for diabetes prediction we have applied a performance matrices are confusion matrix and accuracy are discussed as follows:

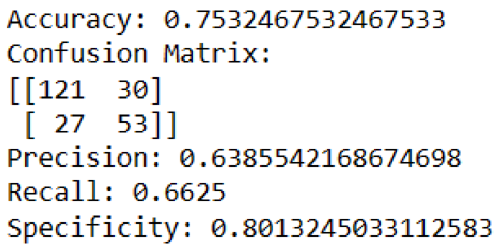
Confusion matrix: - which provides output matrix with complete description performance of the model .Here,

Tp: True positive

FP: False positive

TN: True negative

FN: False negative

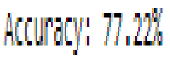


```
Accuracy: 0.7532467532467533
Confusion Matrix:
[[121  30]
 [ 27  53]]
Precision: 0.6385542168674698
Recall: 0.6625
Specificity: 0.8013245033112583
```

Accuracy

We have chooses accuracy matrix to measure the performance of all the models .The ratio of number of correct predictions to the total number of predictions Made.

Accuracy = Number of correct Prediction **/** Total numbers of predictions made



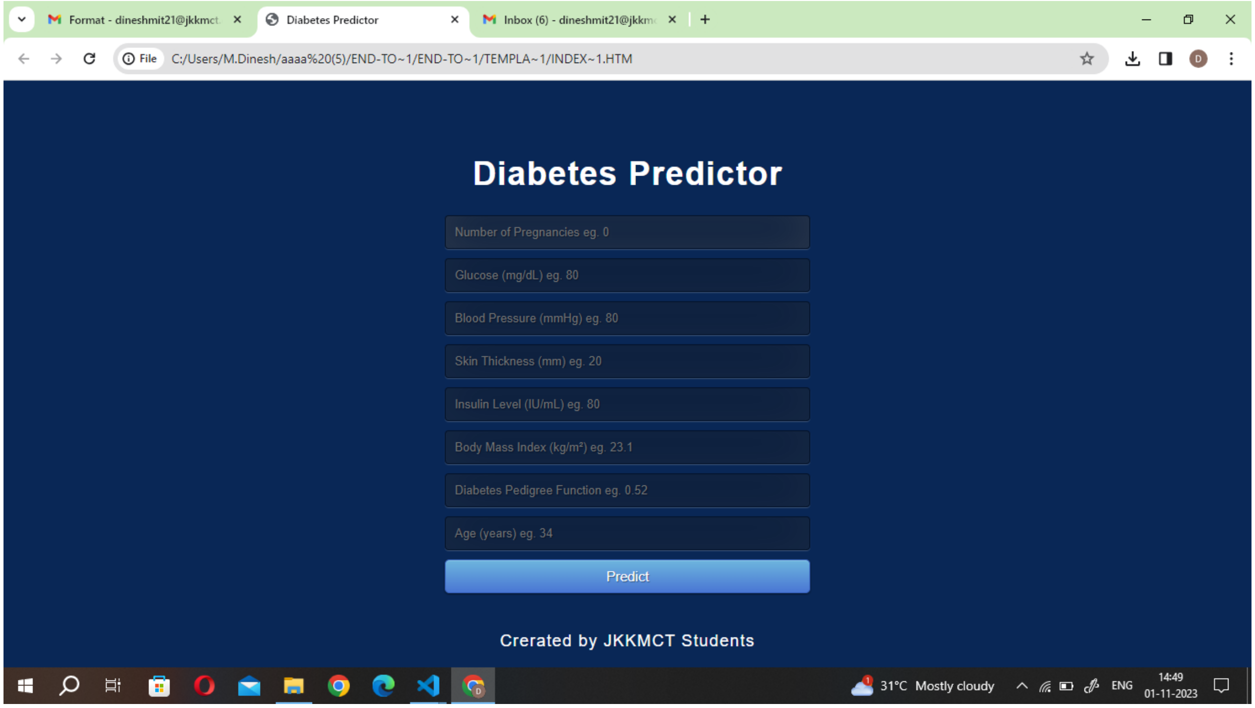
```
Accuracy: 77.22%
```

RESULTS

Machine learning classification algorithms developed for prediction of diabetes in earlier stage. We used 70% of data for training and 30% of data for testing. In this ratio of data splitting Here we found that Random Forest Classifier predicted with 99% of accuracy as highest accuracy for the dataset.

Creating a User Interface for Accessibility:

The last part of the project is the creation of a user interface for the model. This user interface is used to enter unseen data for the model to read and then make a prediction. The user interface is created using “Flask” Web app, Hyper Text Markup Language, and Cascading Style Sheets.


Format - dineshmit21@jkkmct
Diabetes Predictor
Inbox (6) - dineshmit21@jkkm
File
C:/Users/M.Dinesh/aaaa%20(5)/END-TO~1/END-TO~1/TEMPLA~1/INDEX~1.HTM
Diabetes Predictor
Number of Pregnancies eg. 0
Glucose (mg/dL) eg. 80
Blood Pressure (mmHg) eg. 80
Skin Thickness (mm) eg. 20
Insulin Level (IU/mL) eg. 80
Body Mass Index (kg/m²) eg. 23.1
Diabetes Pedigree Function eg. 0.52
Age (years) eg. 34
Predict
Crerated by JKKMCT Students
31°C Mostly cloudy
ENG
14:49
01-11-2023

Results and Analysis

The project predicts the onset of diabetes in a person based on the relevant medical details collected. When the person enters all the relevant medical data required in the online Web portal, this data is then passed on to the trained model for it to make predictions whether the person is diabetic or non-diabetic the model then makes the prediction with an accuracy of98%, which is fairly good and reliable. Following figure shows the basic UI form which requires the user to enter the specific medical data fields. These parameters help determine if the person is prone to develop diabetes Our research has the added benefit of an associated Web app, which makes the model more user friendly and easily understandable for a novice On submission of this form, data the model gives the result in the form of Text.

Conclusion

The objective of the project was to develop a model which could identify patients with diabetes who are at high risk of hospital admission. Prediction of risk of hospital admission is a fairly complex task. Many factors influence this process and the outcome. There is presently a serious need for methods that can increase healthcare institution’s understanding of what is important in predicting the hospital admission risk. This project is a small contribution to the present existing methods of diabetes detection by proposing a system that can be used as an assistive tool in identifying the patients at greater risk of being diabetic .This project achieves this by analyzing many key factors like the patient’s blood glucose level, body mass index, etc., using various machine learning models and through retrospective analysis of patients’ medical records. The project predicts the onset of diabetes in a person based on the relevant medical details that are collected using a Web application .When the user enters all the relevant medical data required in the online Web application, this data is then passed on to the trained model for it to make predictions whether the person is diabetic or non diabetic. The model is developed using artificial neural network consists of total of six dense layers. Each of these layers is responsible for the efficient working of the model. The model makes the prediction with an accuracy of 77.22%,which is fairly good and reliable.