GESTATIONAL DIABETES MELLITUS PREDICTION USING MACHINE LEARNING ALGORITHM

A MINI PROJECT REPORT

Submitted by

PRIYAA. L. D (810020205064)

SARAVANADEVI. M (810020205075)

PODHUMANI. M (810020205313)

In partial fulfilment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

UNIVERSITY COLLEGE OF ENGINEERING, BIT CAMPUS
TIRUCHIRAPALLI

ANNA UNIVERSITY: CHENNAI 600 025

MAY 2023

ANNA UNIVERSITY: CHENNAI 600 025

BONAFIDE CERTIFICATE

Certified that this mini project report **"GESTATIONAL DIABETES MELLITUS PREDICTION USING MACHINE LEARNING ALGORITHM"** is a bonafide work of **"PRIYAA L D (810020205064), SARAVANA DEVI M (810020205075),** and **PODHUMANI M (810020205313)"**, who carried out the mini project work under my supervision, for the partial fulfilment of the requirements for the award of the degree of *Bachelor of Technology in Information Technology.*

SIGNATURE

Dr. G. ANNAPOORANI
HEAD OF THE DEPARTMENT
Assistant Professor (Sl. Gr)
Department of IT
University College of Engineering
BIT- Campus
Tiruchirapalli - 620 024

SIGNATURE

Dr. S. SATHIYA DEVI
SUPERVISOR
Assistant Professor
Department of IT
University College of Engineering
BIT- Campus
Tiruchirapalli - 620 024

Submitted for the University Practical Examination held on …………………

Internal Examiner **External Examiner**

DECLARATION

We hereby declare that the work entitled "**GESTATIONAL DIABETES MELLITUS PREDICTION USING MACHINE LEARNING ALGORITHM**" is submitted in partial fulfilment of the requirement for the award of the degree in B. Tech., (Information Technology) in University College of Engineering, BIT- Campus, Anna University, Tiruchirapalli. It is a record of our own work carried out by us during the academic year 2022 – 2023 under the supervision and guidance of Dr. S. Sathiya Devi, Assistant Professor, Department of Information Technology, University College of Engineering, BIT Campus, Anna University, Tiruchirapalli. The extend and source of information are derived from the existing literature and have been indicated through the dissertation of appropriate places. The matter embodied in this work is original and has not been submitted for the award of any other degree, either in this or any other University.

SIGNATURE OF CANDIDATES
PRIYAA L D (810020205064)
SARAVANADEVI M (810020205075)
PODHUMANI M (810020205313)

I certify that the declaration made above by the candidates is true.

SIGNATURE OF THE GUIDE
Dr. S. SATHIYA DEVI
Assistant Professor
Department of IT
University college of Engineering
BIT-Campus
Tiruchirapalli – 620 024.

ACKNOWLEDGEMENT

We would like to thank our honorable Dean **Dr. T. SENTHIL KUMAR**, Professor for having provided us with all required faculties to complete our project without hurdles.

We would also like to express our sincere thanks to **Dr. G. ANNAPOORANI**, Head of the Department of Computer Science and Engineering, for her valuable guidance, suggestions and constant encouragement paved the way for the successful completion of this project work.

We would like to thank and express our deep sense of gratitude to our project guide **Dr. S. SATHIYA DEVI**, Assistant Professor, Department of Information Technology, BIT-Campus, Anna University, Tiruchirapalli, for her valuable guidance throughout the project.

We also extend our thanks to all other teaching and non-teaching staff for their encouragement and support.

We thank our beloved parents and friends for their full support in the moral development of this project.

ABSTRACT

Diabetes Mellitus is one among the critical diseases and lots of people are suffering from this disease. Current practices followed in hospitals and clinics is to collect required information for diabetes diagnosis through various tests and appropriate treatment is provided based on diagnosis. Healthcare industries have large volume of databases. The project entitled "Gestational diabetes mellitus prediction using machine learning algorithm" has been implemented to predict diabetes as positive or negative. This project consists of five phases (i). Data collection (ii). Data pre-processing (iii). Model creation (iv). Performance evaluation (v). Prediction. Dataset is collected from National Institute of Digestive and Kidney Diseases and Diabetes with 768 records among which diabetes negative is 500 and diabetes positive is 268. The pre-processing of data is performed for missing values. The missing value is imputed by computing the mean of the corresponding attributes. Then the complete data set is given to Ada Boost classifier for prediction. The model produces high accuracy for training and testing as 84.6% and 100% respectively. The performance of this model is compared with other models such as (i). Random Forest, (ii). Decision Tree, (iii). K – Nearest Neighbour (K-NN) and (iv). Naïve Bayes. The experimental result shows that, Ada Boost yields better accuracy when compared with other models for both training and testing data. The Ada Boost produces a classification accuracy, precision, recall and F1 - score as 0.846, 0.742, 0.815 and 0.777 respectively for the above-mentioned dataset.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

GDM	Gestational Diabetes Mellitus
IGT	Impaired Glucose Tolerance
DH	District Hospital
MC	Medical College
GoI	Government of India
DESMLA	Diabetes Expert System using Machine Learning Analytics
PIDD	PIMA Indian Diabetes Dataset
DT	Decision Tree
RF	Random Forest
SVM	Support Vector Machine
GBC	Gradient Boost Classifier
K-NN	K-Nearest Neighbor
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative

CHAPTER 1
INTRODUCTION

Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high. Blood glucose is your main source of energy and comes from the food you eat. Insulin, a hormone made by the pancreas, helps glucose from food get into your cells to be used for energy. Sometimes your body doesn’t make enough—or any—insulin or doesn’t use insulin well. Glucose then stays in your blood and doesn’t reach your cells.

The most common types of diabetes are,

- Type 1 Diabetes
- Type 2 Diabetes
- Gestational Diabetes

Among the most common type of diabetes, Gestational diabetes are the most common diabetes.

1.1 GESTATIONAL DIABETES

Gestational diabetes occurs when your body can’t make enough insulin during your pregnancy. Insulin is a hormone made by your pancreas that acts like a key to let blood sugar into the cells in your body for use as energy. During pregnancy, your body makes more hormones and goes through other changes, such as weight gain. These changes cause your body’s cells to use insulin less effectively, a condition called insulin resistance. Insulin resistance increases your body’s need for insulin. All pregnant women have some insulin resistance during late pregnancy. However, some women have insulin resistance even before they get pregnant. They start pregnancy with an increased need for insulin and are more likely to have gestational diabetes.

1.2 GESTATIONAL DIABETES DETECTION AND TREATMENT

Gestational Diabetes Mellitus (GDM) is defined as Impaired Glucose Tolerance (IGT) with onset or first recognition during pregnancy. Worldwide, one in 10 pregnancies is associated with diabetes, 90% of which are GDM. Undiagnosed or inadequately treated GDM can lead to significant maternal & fetal complications. Moreover, women with GDM and their off-springs are at increased risk of developing type 2 diabetes later in life. In India, one of the most populous Country globally, rates of GDM are estimated to be 10-14.3% which is much higher than the west.

As of 2021, there were an estimated 22 million women with diabetes between the ages of 20 and 39 & an additional 54 million women in this age group with impaired glucose tolerance (IGT) or pre-diabetes with the potential to develop GDM if they become pregnant. In a field study in Tamil Nadu performed under the Diabetes in Pregnancy – Awareness and Prevention project, of the 4151, 3960 and 3945 pregnant women screened in urban, semi urban and rural areas, respectively, the prevalence of GDM was 17.8% in the urban, 13.8% in the semi urban and 9.9% in the rural areas. The incidence of GDM is expected to increase to 20% i.e, one in every 5 pregnant women is likely to have GDM.

Despite a high prevalence of GDM in Indian women, currently screening of pregnant women for GDM is not being done universally as part of the essential antenatal package. The test is sporadically being done at DH and MC in some states as per direction of individual clinician except in the state of Tamil Nadu where every pregnant woman is being screened up to the level of PHC as a part of the government of Tamil Nadu initiative. Despite the fact that GDM is a sizeable public health problem with serious adverse effects on mother & child, we do not have a standard GoI guideline for diagnosis and management of GDM.

In today’s world, hospitals and healthcare organizations are under immense pressure to maximize resources – and predictive analytics makes that possible. Using predictive analytics, healthcare officials can improve financial and operational decision-making, optimize inventory and staffing levels, manage their supply chains more efficiently, and predict maintenance needs for medical equipment. Predictive analytics also makes it possible to improve clinical outcomes by detecting early signs of patient deterioration, identifying patients at risk for readmission, and improving the accuracy of patient diagnosis and treatment.

Over the past decade, computer detection has been widely used in various domains. Several methods have been changed from statistical methods to machine learning methods because it offers greater accuracy for tasks such as diabetes prediction. The technology can help computer scientists to develop tasks in various fields rapidly. It can automatically learn features from the given dataset.

CHAPTER 2

LITERATURE SURVEY

Aishwarya Mujumdar, Dr. V. Vaidehi (2019) [1], in this project a diabetes prediction model has been created by training and testing two different datasets with several different algorithms namely Ada Boost Algorithm, Decision Tree Classifier, K-Nearest Neighbor, Naïve Bayes, Logistic Regression, Support vector machine, Random Forest Classifier, etc. And they have trained the model with Pipelining algorithm as well. Then Compared the obtained results for better accuracy to build the predictive model. Ada Boost Classifier is the best model with more accuracy. It is clear that the model improves accuracy and precision of diabetes prediction with this dataset compared to existing dataset.

K M Jyoti Rani (2020) [2], In this study, systematic efforts are made in designing a system which results in the prediction of diabetes. During this work, five machine learning classification algorithms namely K-Nearest Neighbor, Logistic Regression, Decision Tree Classifier, Random Forest Classifier and Support Vector Machine are studied and evaluated on various measures. Experiments are performed on john Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 99% using Decision Tree algorithm.

S. Reshmi, Saroj Kr. Biswas, A.N. Bourah, D. M. Thounaojam, B. Purkayastha (2022) [3], DESMLA model uses Decision Tree (DT) and Random Forest (RF) as classifiers along with all the data pre-processing steps for diabetes prediction. DESMLA first treat the class imbalance problem by using SMOTE, Borderline SMOTE, ADASYN, K-Means and Gaussian smote and then by using DT and RF diabetes is predicted. The proposed procedure performs better for PIDD [6] but have not consider other crucial factors related to gestational diabetes, like family history, metabolic syndrome, the habit of smoking, some dietary patterns, lazy routines etc.

S. S. Bhat, V. Selvam, G. A. Ansari, M. D. Ansari, M. H. Rahman (2022) [4], prevalence of diabetes is showing an upward trend in Bandipora Kashmir. Using state of art algorithms for the early prediction can help in decreasing the upward trend of diabetes Six algorithms including RF, MLP, SVM, DT, GBC, and LR algorithms were utilized for this purpose amongst all algorithms we achieved RF has the highest accuracy of 98%. RF also has produced successful outcomes for several statistical metrics includes ROC Area, Recall, Precision, F-measure, and MCC. K-fold Machine Learning models such as cross-validation has been used to evaluate RF, MLP, SVM, DT, GBC, and LR. The framework utilized in this research will be applied to ensemble and hybridization Machine Learning in order to further recent research.

M. Soni, Dr. S. Varma (2020) [5], The main aim of the project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully. The proposed approach uses various classification and ensemble learning method in which SVM, KNN, Random Forest, Decision Tree, Logistic Regression and Gradient Boosting classifiers are used. And 77% classification accuracy has been achieved. The Experimental results ca assist health care to take early prediction and make early decision to cure diabetes and save humans life.

CHAPTER 3

GESTATIONAL DIABETES MELLITUS PREDICTION USING MACHINE LEARNING ALGORITHM

This project has been implemented to predict diabetes as positive or negative where positive states that the person is diabetic and negative states that the person is non-diabetic.

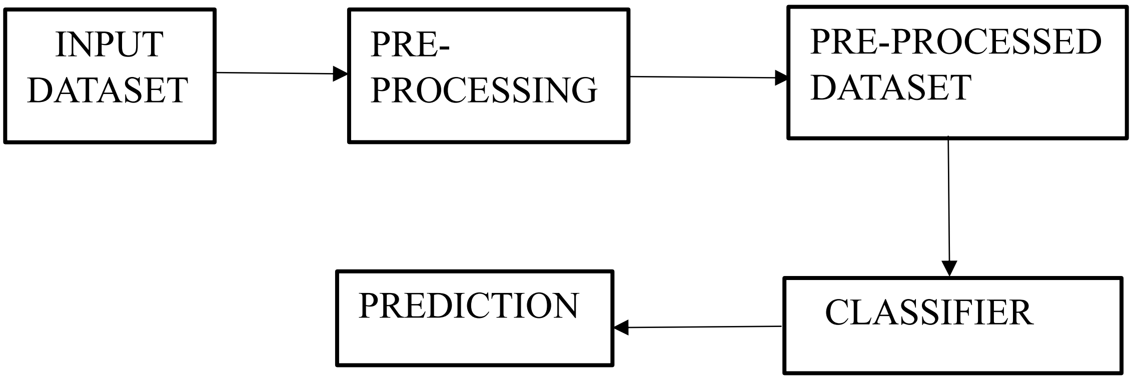


Fig. 1 FLOW DIAGRAM OF DIABETES PREDICTION

The whole process consists of five phases as shown in Fig. 1 and are (i). Dataset collection (ii). Dataset Pre-processing (iii). Model creation (iv). Performance Evaluation (v). Prediction. Initially the dataset is collected from National Institute of Digestive and Kidney Disease and Diabetes with total of 768 records. Data Pre-processing is performed on this dataset to increase the size of the dataset. The pre-processed dataset is trained with the machine learning algorithms namely Ada Boost Algorithm, Decision Tree Classifier, K-Nearest Neighbor Algorithm, Naïve Bayes Algorithm, and Random Forest Classifier. Classification Accuracies obtained are compared and better algorithm is chosen among them to build prediction model.

3.1 DATA COLLECTION

Dataset is collected from National Institute of Digestive and Kidney Disease and Diabetes entitled “PIMA Indian Diabetes Dataset (PIDD)” [6] with total of 768 records. The dataset consists of diabetes prediction and it is classified into two, (i). Negative (500) and (ii). Positive (268). The number of records in each class is shown in Table. 1. The Dataset is categorized into training and testing data as shown in Table. 2.

Table. 1 Class Label Description

S.no	Class	Samples
1	Positive	268
2	Negative	500

Table. 2 Dataset description: Training and Testing Data

Dataset	No. of records
Training Data	614
Testing Data	154

3.2 DATA PRE-PROCESSING

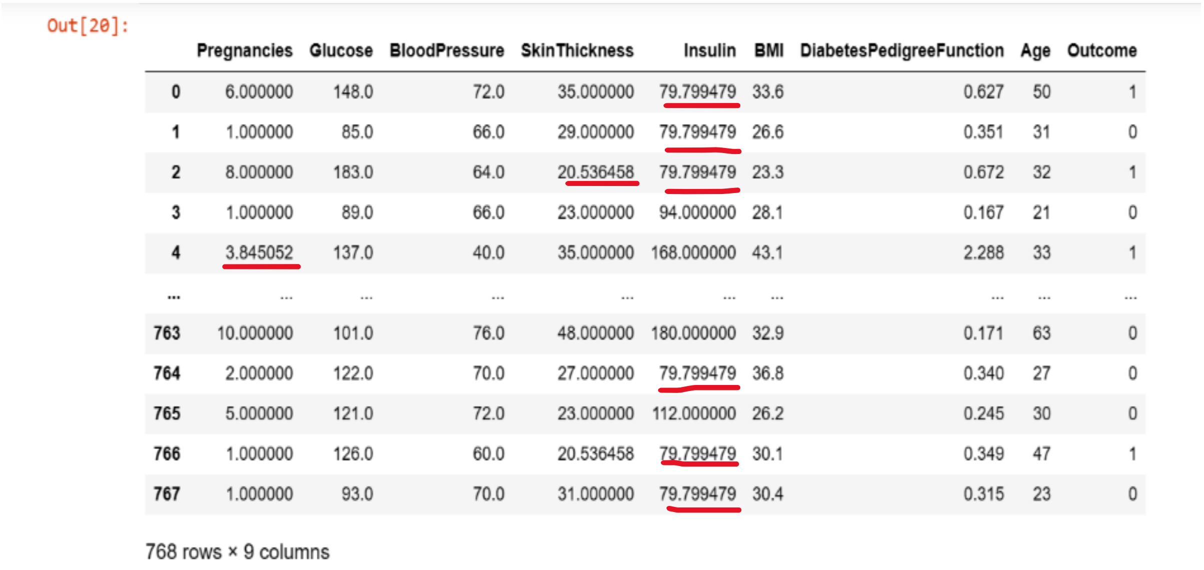
The PIMA Indian Diabetes Dataset (PIDD) [6] collected from National Institute of Digestive and Kidney Diseases and Diabetes contains a total of 768 records with some amount missing-values. Due to this missing-values the prediction made is inaccurate. To overcome this, the dataset is pre-processed by replacing the missing values with the mean value of data of that particular attribute.

This is depicted by the equation 1,

$$1.\text{Mean} = \frac{sum\ of\ all\ values\ of\ the\ attribute}{Total\ no\ of\ records}$$

Table. 3 Obtained mean values

	Pregnancies	**Glucose**	**Blood Pressure**	**Skin Thickness**	**Insulin**	**BMI**	**Diabetes Pedigree Function**	**Age**
Mean	3.845	120.894	69.105	20.536	79.799	31.992	0.471	33.240



Out[20]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6.000000	148.0	72.0	35.000000	79.799479	33.6	0.627	50	1
1	1.000000	85.0	66.0	29.000000	79.799479	26.6	0.351	31	0
2	8.000000	183.0	64.0	20.536458	79.799479	23.3	0.672	32	1
3	1.000000	89.0	66.0	23.000000	94.000000	28.1	0.167	21	0
4	3.845052	137.0	40.0	35.000000	168.000000	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10.000000	101.0	76.0	48.000000	180.000000	32.9	0.171	63	0
764	2.000000	122.0	70.0	27.000000	79.799479	36.8	0.340	27	0
765	5.000000	121.0	72.0	23.000000	112.000000	26.2	0.245	30	0
766	1.000000	126.0	60.0	20.536458	79.799479	30.1	0.349	47	1
767	1.000000	93.0	70.0	31.000000	79.799479	30.4	0.315	23	0

768 rows × 9 columns

Fig. 2 Dataset After Pre-processing

3.3 MODEL CREATION

The diabetes prediction model is built with Ada Boost Algorithm with 84.6% accuracy. The Ada Boost Algorithm has provided a better accuracy when compared with classification accuracies of other four algorithms.

3.3.1 ADABOOST ALGORITHM

Ada Boost, short for Adaptive Boosting, is a machine learning algorithm formulated by Yoav Freund and Robert Schapire. AdaBoost technique follows a decision tree model with a depth equal to one. AdaBoost is nothing but the forest of stumps rather than trees. AdaBoost works by putting more weight on difficult to classify instances and less on those already handled well. AdaBoost algorithm is developed to solve both classification and regression problem.

Idea behind AdaBoost:

- Stumps (one node and two leaves) are not great in making accurate classification so it is nothing but a week classifier/ weak learner. Combination of many weak classifier makes a strong classifier and this is the principle behind the AdaBoost algorithm.
- Some stumps get more performance or classify better than others.
- Consecutive stump is made by taking the previous stumps mistakes into account.

3.3.1.1 STEPS INVOLVED:

a. Assign equal weight to all the observations.
b. Classify random samples using stumps.
c. Calculate Total Error.
d. Calculate performance of the stump.
e. Update weights.
f. Update weights in iteration.
g. Final prediction.

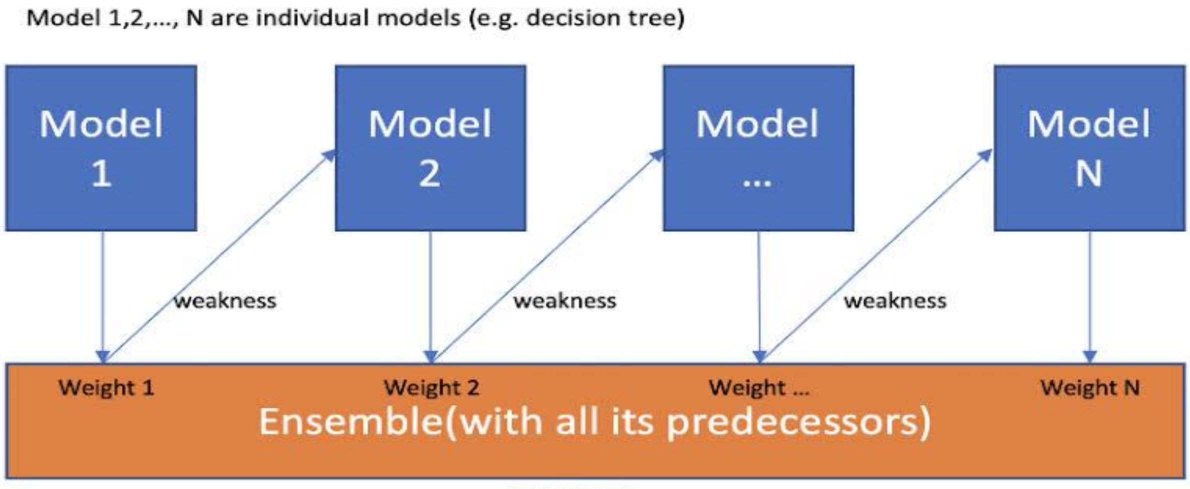

Model 1,2,..., N are individual models (e.g. decision tree)
Model 1
Model 2
Model ...
Model N
weakness
weakness
weakness
Weight 1
Weight 2
Weight ...
Weight N
Ensemble(with all its predecessors)

Fig. 3 ADABOOST ALGORITHM

3.3.1.2 PARAMETERS INVOLVED:

1.**Base estimator**: The base estimator from which the boosted ensemble is built. Support for sample weighting is required, as well as proper classes_ and n_classes_ attributes. If None, then the base estimator is DecisionTreeClassifier initialized with max_depth=1.

2.**n_estimator**: The maximum number of estimators at which boosting is terminated. In case of perfect fit, the learning procedure is stopped early. Values must be in the range [1, inf).

3.**Learning_rate**: Weight applied to each classifier at each boosting iteration. A higher learning rate increases the contribution of each classifier. There is a trade-off between the learning_rate and n_estimators parameters. Values must be in the range [0.0, inf).

4.**Random_state**: Controls the random seed given at each estimator at each boosting iteration. Thus, it is only used when estimator exposes a random_state. Pass an int for reproducible output across multiple function calls.

3.3.2 DECISION TREE CLASSIFIER(DTC)

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

3.3.2.1 STEPS INVOLVED**:**

a. Determine the Root of the tree.
b. Calculate the Entropy after spilt for the classes.
c. Calculate the Entropy after spilt for each attribute.
d. Calculate information gain for each spilt.
e. Perform the spilt.
f. Perform further splits.
g. Complete the Decision tree.

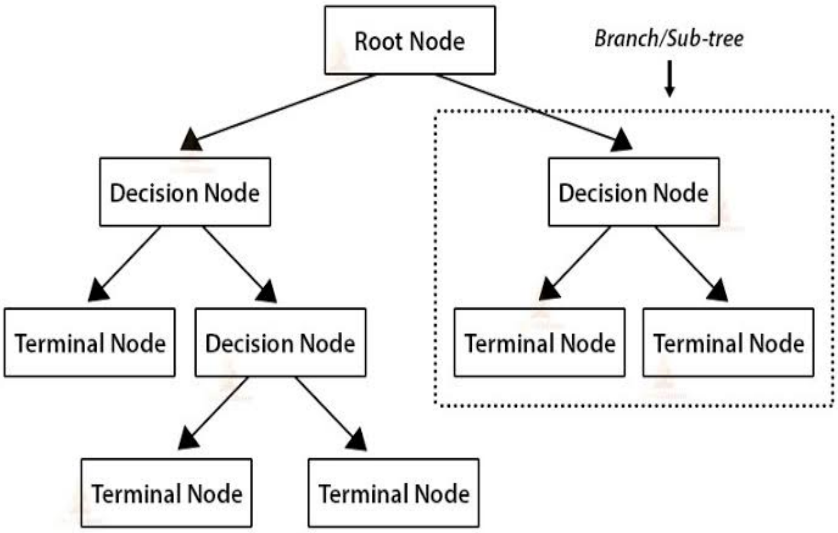


Fig. 4 DECISION TREE CLASSIFIER

3.3.2.2 PARAMETERS USED:

1.**Criterion**: The function to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “log_loss” and “entropy” both for the Shannon information gain.

2.**Max_depth**: The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.

3.**Random_state**: Controls the random seed given at each estimator at each boosting iteration. Thus, it is only used when estimator exposes a random_state. Pass an int for reproducible output across multiple function calls.

3.3.3 K-NEAREST NEIGHBOUR

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

3.3.3.1 STEPS INVOLVED:

a. Select the number K of the neighbors.
b. Calculate the Euclidean distance of K number of neighbors.
c. Take the K nearest neighbors as per the calculated distance.
d. Among these k neighbors, count the number of the data points in each category.
e. Assign the new data points to that category for which of the neighbour is maximum.
f. Our model is ready.

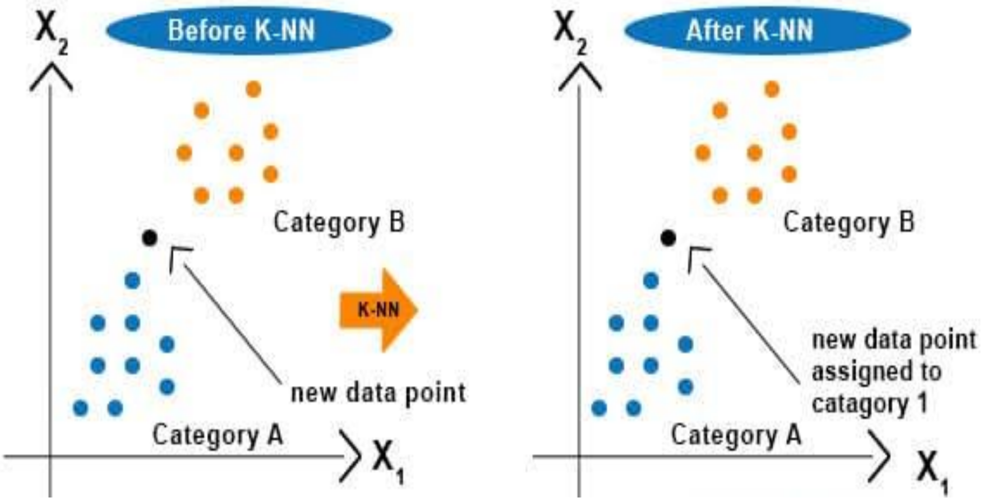


Fig. 5 KNN ALGORITHM

3.3.3.2 PARAMETER USED:

1.**n_neighbors**: Number of neighbors to use by default for **k neighbors** queries.

2.**Random_state**: Controls the random seed given at each estimator at each boosting iteration. Thus, it is only used when estimator exposes a random_state. Pass an int for reproducible output across multiple function calls.

3.3.4 NAÏVE BAYES ALGORITHM

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

3.3.4.1 STEPS INVOLVED:

a. Calculate prior probability for given class labels.
b. Calculate conditional probability with each attribute for each class.
c. Multiply same class conditional probability.
d. Multiply prior probability with step3 probability.
e. See which class has higher probability. Higher probability class belongs to given input set step.

3.3.4.2 PARAMETERS USED:

1. **priors**: Prior probabilities of the classes. If specified, the priors are not adjusted according to the data.

3.3.5 RANDOM FOREST CLASSIFIER

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning**,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

3.3.5.1 STEPS INVOLVED:

a. Start with the selection of random samples from a given dataset.
b. This algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
c. In this step, voting will be performed for every predicted result.
d. At last, select the most voted prediction result as the final prediction result.

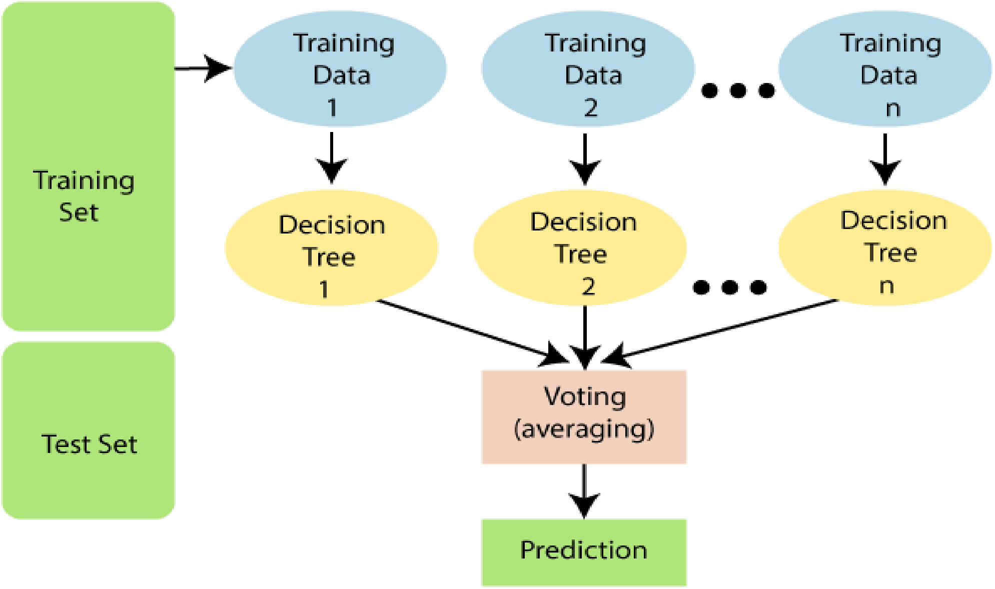


Fig. 6 RANDOM FOREST CLASSIFIER

3.3.5.2 PARAMETERS USED:

1. **Criterion**: The function to measure the quality of a split. Supported criteria are "gini" for the Gini impurity and "log_loss" and "entropy" both for the Shannon information gain.

2. **n_estimators**: The maximum number of estimators at which boosting is terminated. In case of perfect fit, the learning procedure is stopped early. Values must be in the range [1, inf).

3. **max_depth**: The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.

4.**Random_state**: Controls the random seed given at each estimator at each boosting iteration. Thus, it is only used when estimator exposes a random_state. Pass an int for reproducible output across multiple function calls.

3.4 PERFORMANCE EVALUATION

The following metrics are used to evaluate the performance of the classification.

3.4.1 CONFUSION MATRIX

A confusion matrix is a tabular representation of prediction outcomes of any binary classifier, which is used to describe the performance of the classification model on a set of test data when true values are known.

		ACTUAL VALUES	
		POSITIVE (1)	NEGATIVE (0)
PREDICTED VALUES	POSITIVE (1)	TP	FP
	NEGATIVE (0)	FN	TN

Fig. 7 CONFUSION MATRIX

1.True positive (TP): Outcome where the model correctly predicts the positive class.

2.True Negative (TN): Outcome where the model correctly predicts the negative class.

3. False Positive (FP): It is also called as type 1 error. Outcome where the model incorrectly predicts the positive class when it is actually negative.

4. False Negative (FN): It is also called as type 2 error. Outcome where the model incorrectly predicts the negative class when it is actually positive.

3.4.2 ACCURACY

The accuracy metric is one of the simplest Classification metrics to implement, and it can be determined as the number of correct predictions to the total number of predictions.

It can be depicted by equation 2,

2. Accuracy = $\frac{TP+TN}{TP+FN+TN+FP}$

3.4.3 PRECISION

The precision metric is used to overcome the limitation of Accuracy. The precision determines the proportion of positive prediction that was actually correct. It can be calculated as the True Positive or predictions that are actually true to the total positive predictions (True Positive and False Positive).

It can be depicted by equation 3,

3. Precision=$\frac{TP}{TP+FP}$

3.4.4 RECALL

It is also similar to the Precision metric; however, it aims to calculate the proportion of actual positive that was identified incorrectly. It can be calculated as True Positive or predictions that are actually true to the total number of positives, either correctly predicted as positive or incorrectly predicted as negative (true Positive and false negative).

It can be depicted by equation 4,

4. Recall=$\frac{TP}{TP+FN}$

3.4.5 F1-SCORE

F-score or F1 Score is a metric to evaluate a binary classification model on the basis of predictions that are made for the positive class. It is calculated with the help of Precision and Recall. It is a type of single score that represents both Precision and Recall. So, the F1 Score can be calculated as the harmonic mean of both precision and Recall, assigning equal weight to each of them.

It can be depicted by equation 5,

5. F1-Score=$\frac{\boldsymbol{2\times precision\times recall}}{\boldsymbol{precision+recall}}$

CHAPTER 4

EXPERIMENT AND RESULTS

4.1 SOFTWARE REQUIREMENTS

The software requirements are as follows:

- Framework: streamlit
- Tool : Jupyter Notebook, IDLE python
- Language : python3

4.2 HARDWARE REQUIREMENTS

The hardware requirements are as follows:

- Desktop or PC
- Processor: i3 or any equal
- RAM: minimum 4GB
- Storage: 128GB and more

4.3 PACKAGES USED

This project entitled, “Gestational Diabetes Mellitus Prediction using Machine Learning Algorithm” is implemented using python3 programming language in a Jupyter Notebook. The experiments were conducted by making use of following python packages.

4.3.1 Sklearn

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modelling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

4.3.2 Streamlit

Streamlit lets you turn data scripts into shareable web apps in minutes, not weeks.

4.3.3 Numpy

Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data.

4.3.4 Pandas

Pandas is a Python package that provides fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python.

4.4 DATASET DESCRIPTION

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

- Size of the dataset: - 9 KB
- Total number of records: - 768
- Total number of attributes: - 9
- **List of Attributes**: -
  - Input variables:

Pregnancy, BMI, Glucose, BP, Skin Thickness, Age, Insulin, DiabetesPedigreeFunction.

- Output variable:

  Outcome.

4.5 PARAMETERS TUNNED

Table. 4 PARAMETERS TUNNED

S.NO	PARAMETER	VALUE
1	base_estimator	DecisionTreeClassifier
2	n_estimator	100
3	learning_rate	1
4	n_neighbors	5
5	random_state	0
6	criterion	entropy
7	max_depth	5

4.6 METRICS CALCULATED

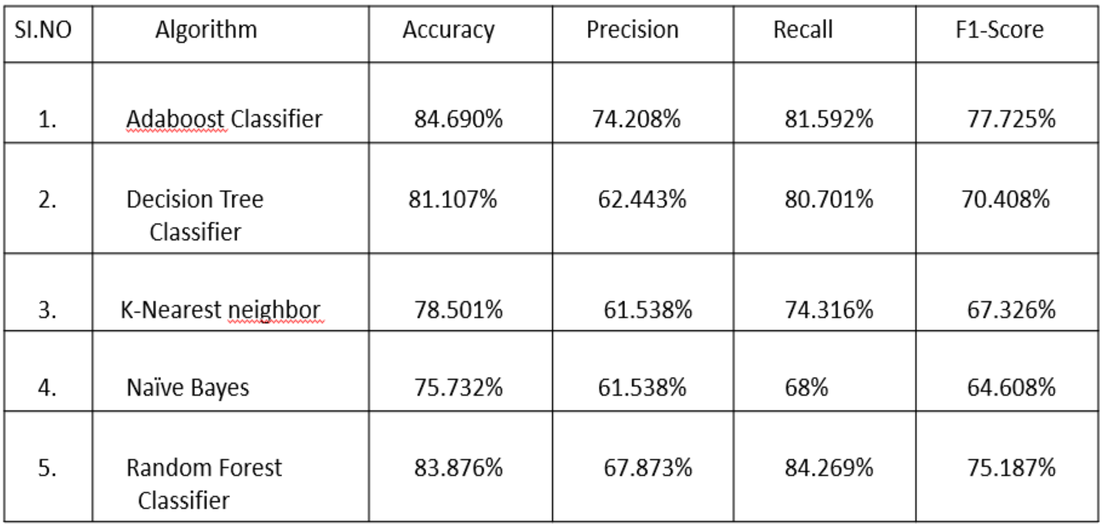
- Confusion Matrix
- Accuracy
- Precision
- Recall
- F1-score

4.7 RESULT

The below mentioned Table. 5 and Table. 6 shows the performance metrics evaluated for training and testing data respectively. The Fig. 8 and Fig. 9 demonstrates the prediction of diabetes for two different class labels.

4.7.1 PERFORMANCE EVALUATED FOR TRAINING DATA

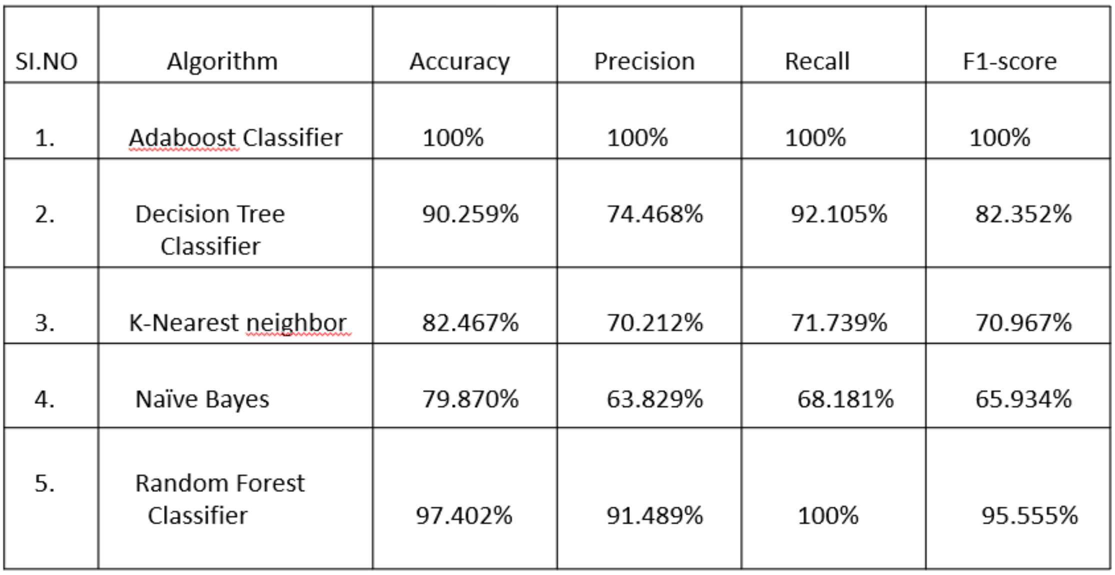
Table. 5 PERFORMANCE OF TRAINING DATA



SI.NO	Algorithm	Accuracy	Precision	Recall	F1-Score
1.	Adaboost Classifier	84.690%	74.208%	81.592%	77.725%
2.	Decision Tree Classifier	81.107%	62.443%	80.701%	70.408%
3.	K-Nearest neighbor	78.501%	61.538%	74.316%	67.326%
4.	Naïve Bayes	75.732%	61.538%	68%	64.608%
5.	Random Forest Classifier	83.876%	67.873%	84.269%	75.187%

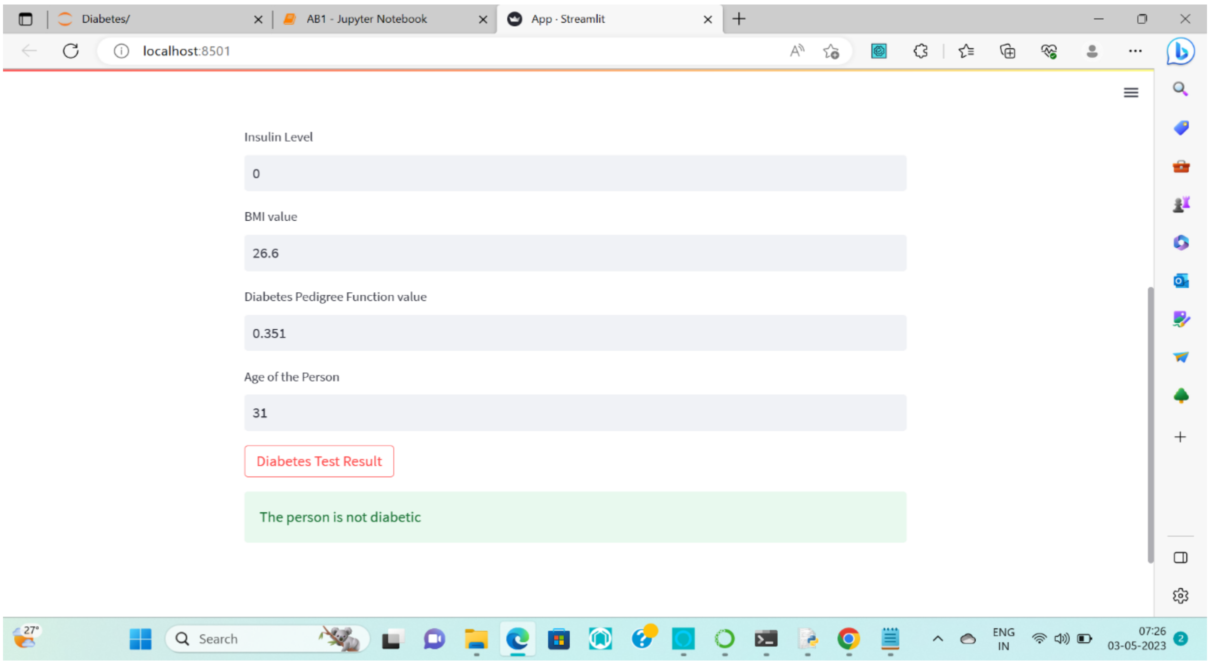
4.7.2 PERFORMANCE EVALUATED FOR TESTING DATA

Table 6 PERFORMANCE OF TESTING DATA



SI.NO	Algorithm	Accuracy	Precision	Recall	F1-score
1.	Adaboost Classifier	100%	100%	100%	100%
2.	Decision Tree Classifier	90.259%	74.468%	92.105%	82.352%
3.	K-Nearest neighbor	82.467%	70.212%	71.739%	70.967%
4.	Naïve Bayes	79.870%	63.829%	68.181%	65.934%
5.	Random Forest Classifier	97.402%	91.489%	100%	95.555%

4.7.3 OUTPUT



Insulin Level

0

BMI value

26.6

Diabetes Pedigree Function value

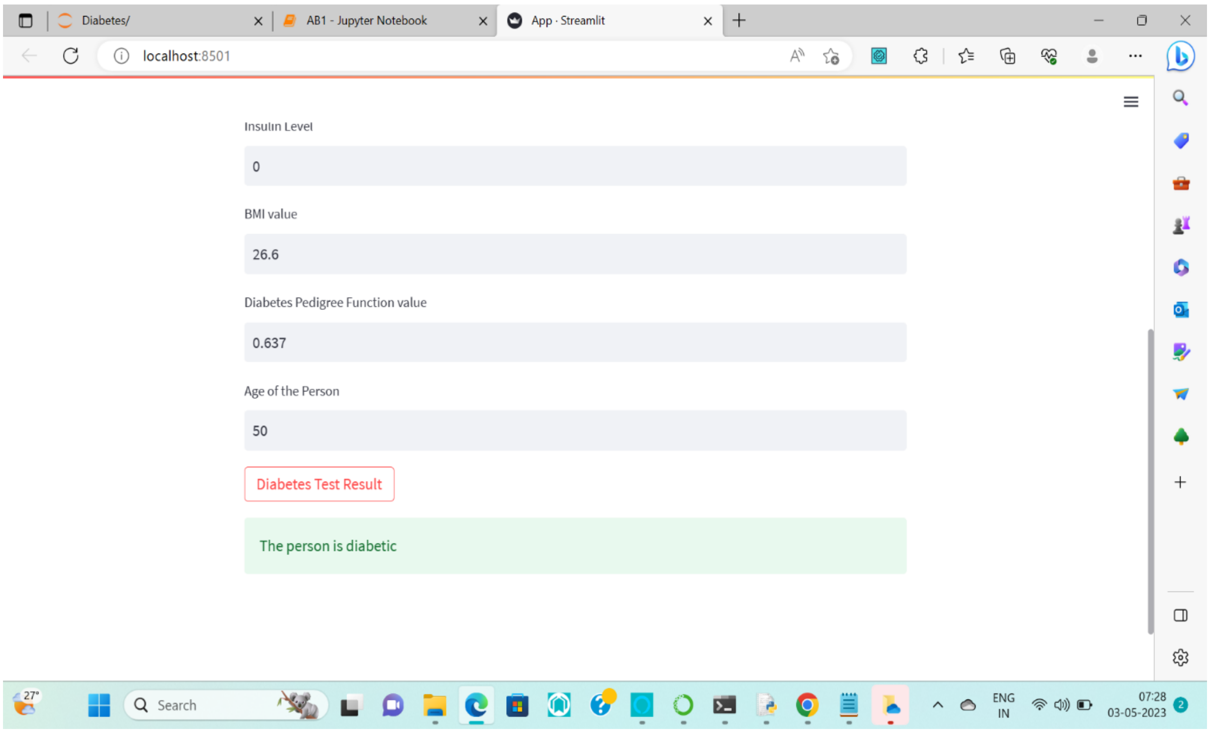
0.351

Age of the Person

31

The person is not diabetic

Fig. 8 DEMONSTRATION I



Insulin Level

0

BMI value

26.6

Diabetes Pedigree Function value

0.637

Age of the Person

50

The person is diabetic

Fig. 9 DEMONSTRATION II

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this study, various machine learning algorithms are applied on the dataset and the classification has been done using various algorithms of which Ada Boost gives highest accuracy of 84.6%. We have seen comparison of machine learning algorithm accuracies with the pre-processed dataset. It is clear that the model improves accuracy and precision of diabetes prediction with this dataset compared to existing dataset. Further this work can be extended to find how likely non diabetic people can have diabetes in next few years.

In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.

REFERENCES

[1] Aishwarya Mujumdar, Dr.V. Vaidehi,” Diabetes Prediction using Machine Learning Algorithms”, International Conference on Recent Trends in Advanced Computing (ICRTAC)2019, Procedia Computer Science, volume 165, pp.292-299, 2019.

[2]KM Jyoti Rani,” Diabetes Prediction using Machine Learning”, International Journal of Scientific Research in Computer Science Engineering and Information Technology (IJSRCSE-IT), Volume 6, Issue 4, July-August, pp.294-305, 2020.

[3] S. Reshmi, Saroj Kr. Biswas, Arpita Nath Boruah, Dalton Meitei Thounaojam, Biswajit Purkayastha, “Diabetes Prediction using Machine Learning Analytics” International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, pp.1-6, 2022.

[4] Salliah Shafi Bhat, Venkatesan Selvam, Gufran Ahmad Ansari, Mohd Dilshad Ansari and Md Habibur Rahman, “Prelevance and Early Prediction of Diabetes using Machine Learning in North Kashmir: A case-study of District Bandipora”, Hindawi, Computational Intelligence and Neuroscience, volume 2022, pp.1-12, 4 October 2022.

[5] Mitushi Soni, Dr, Sunita Varma, “Diabetes Prediction using Machine Learning Techniques”, International Journal of Engineering Research and Technology (IJERT), Volume 9, Issue 9, PP.922-925, September 2020.

[6] Dataset: PIMA Indian Diabetes Dataset.

(https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database)