International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6594
***Article Received**: 22 February 2023 **Revised**: 18 March 2023 **Accepted**: 18 April 2023*

Application and Analysis of Machine Learning Algorithms on Pima and Early Diabetes Datasets for Diabetes Prediction

Lakshmi H N*1, Atluri Vani Vathsala2, Bhavik K Upadhyay3, A Narasimha Rao4
*1Professor, Computer Science and Engineering (AIML, CS and DS), CVR College of Engineering, Hyderabad.
E-mail: hn.lakshmi@cvr.ac.in
2Professor, Department of Computer Science and Engineering, CVR College of Engineering, Hyderabad, India.
Email: vani_vathsala@cvr.ac.in
3IV Year B. Tech, CSIT, Department of Computer Science and Information Technology, CVR College of Engineering, Hyderabad, India.
Email: bhavikupadhyay08@gmail.com
4Scrum Master, Tech Mahindra, Hyderabad, India.
Mail Id: atlurinr@yahoo.com

Abstract-Diabetes is a chronic condition that strike how your body burns food for energy. Much of the food you consume is converted by your body into sugar (glucose), which is then released into your bloodstream. Your pancreas releases insulin when your blood sugar levels rise. Over the years, several scholars have sought to create reliable diabetes prediction models. Due to a lack of adequate data sets and prediction techniques, this discipline still faces many unsolved research issues, which forces researchers to apply big data analytics and ML-based methodology. Four distinct machine learning algorithms are used in the study to analyze healthcare prediction analytics and solve the issues. In this investigation, the Pima and Early detection datasets were employed. We applied the Decision Tree, MLP, Naive Bayes, and Random Forest algorithms to these datasets and evaluated the accuracy and F-Measure. The goal of this research is to develop a system that could more precisely predict a patient's risk of developing diabetes.

Keywords: — Random Forest, Deep Neural Networks, Type 2 Diabetes, , Naive Bayes, and Supervised Learning.

I. INTRODUCTION

Diabetes is a chronic disease that has an mark on how your body uses food as fuel. Your body converts the bulk of the food you eat into sugar (glucose), which is subsequently released into your circulation. When your blood sugar levels increase, your pancreas releases insulin. Diabetes in its majority has no recognized definite cause. In every circumstance, blood sugar levels rise. This is a result of the pancreas' insufficient insulin production. Both types of diabetes may be catalyzed on by a combination of genetic and environmental factors. However, many diabetics claim that the most common symptoms of the disease are increased thirst, increased urination, exhaustion, and weight loss.

When the pancreas is unable to produce enough hypoglycemic agent, type 1 diabetes develops. "Juvenile diabetes" or "insulin-dependent polygenic disease mellitus" were terms used to describe this type (IDDM). The type 1 polygenic disorder that disturbs children under 20 is unknown as to its cause. People with type 1 diabetes who rely on insulin injections go through life with pain. The doctors' advised exercise and fitness regimens must frequently be followed by diabetes patients. Ineffective systemic insulin utilization or insufficient pancreatic insulin synthesis are the two main contributing factors to diabetes, a chronic disease. Blood sugar is regulated by the hormone insulin. Hyperglycemia, or increased blood sugar [1], is a common complexity of uncontrolled diabetes that, over time, has a negative impact on a number of physiological systems, including the blood vessels and neurons., heart attacks strokes, blindness, Kidney disease, and lower limb amputation are all significantly more common in those with diabetes. Type 2 diabetes is mostly brought on by hypoglycemic agent resistance, which happens when cells do not react to hypoglycemic drugs as efficiently as they should. The sickness arises because there is no additional built-in hypoglycemia agent. This sort was referred to as "non-insulin-dependent polygenic illness mellitus." One of the main causes of type 2 diabetes is obesity. Type 3 Gestational diabetes is identified when a gravid woman manifests elevated blood glucose concentrations devoid of any antecedent diabetic medical records.

Utilizing a variety of machine learning algorithms, data mining techniques, and statistical tools, predictive analysis is a technique for learning new information and predicting what will happen in the future [2]. By applying predictive analysis to healthcare data, significant conclusions and forecasts can be drawn. Predictive analytics has the potential to utilize machine learning and regression methodologies. By enhancing patient care, maximising resource usage, and providing the most accurate disease diagnosis possible, predictive analytics aims to enhance healthcare outcomes. Therefore, for everyone who is at risk for getting the condition, diabetes prediction and early identification are crucial. Currently, artificial intelligence (AI) techniques can be used to diagnose a variety of illnesses, with the most accurate classification results [3]. Supervised learning aims to generate a precise model of the distribution of class labels in relation to predictor attributes. The ensuing classifier is employed to assign class labels to testing instances, given that the values of predictor features are known, but the class label values are undisclosed. This paper covers a number of supervised machine learning categorization methods.

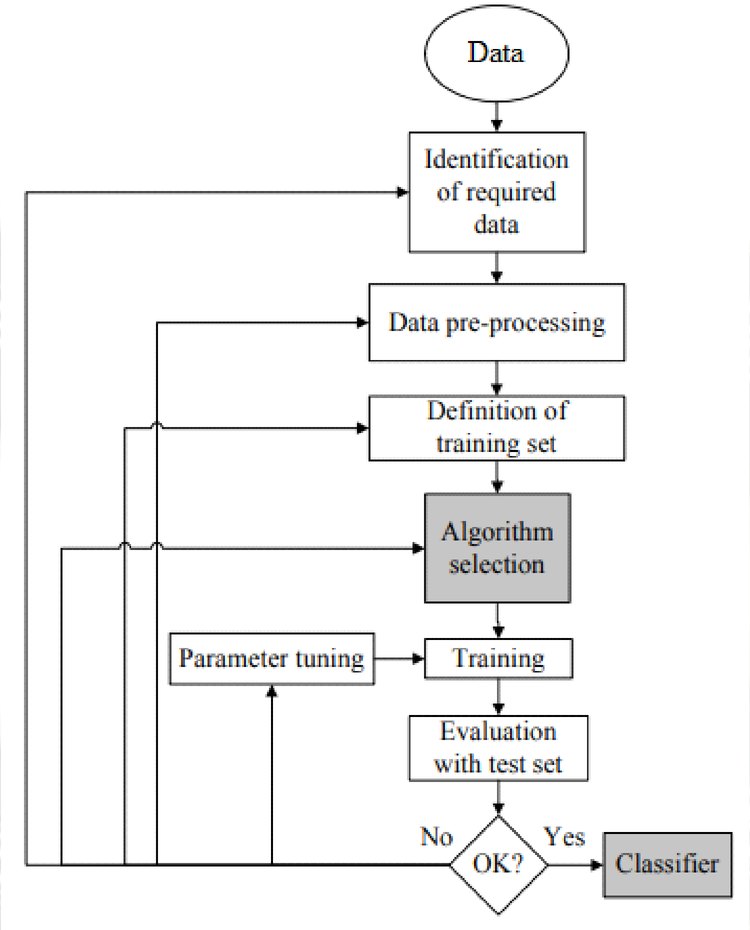


Fig. 1. A synopsis of algorithms used in supervised learning

II. BACKGROUND STUDY

Henock M. Deberneh and Intaek Kim [1] presented that "If someone is aware of when they will develop type 2 diabetes (T2D), they can take steps to prevent the start or progression of the condition. In this study, utilising features from the current year, We developed a machine learning (ML) model with the purpose of predicting the incidence of T2D in the subsequent year (Y + 1).(Y). The study's dataset, which was made up of electronic health records, was acquired from a private medical facility between 2013 and 2018. The key attributes for the prediction model were initially selected using recursive feature reduction, chi-square testing, and ANOVA tests. The residual attributes comprised age, uric acid levels, gender, alcohol consumption, smoking habits, physical activity, and family medical history.

Gamma-GTP, BMI, lipids, HbA1c, and fasting plasma glucose were also included (FPG). The outcome was then predicted using ensemble machine learning algorithms as normal (non-diabetic), prediabetes, or diabetes using these parameters. Among the techniques utilised were support vector machines, ensemble machine learning, logistic regression, random forest, and XGBoost. The empirical findings evinced that the prognostic model exhibited commendable performance in estimating the prevalence of T2D in the Korean community. Clinicians and patients may find the model's accurate predictions regarding the likelihood of developing T2D useful. The ensemble models outperformed the solo models, according to the cross validation (CV) results. More medical data from the dataset was added to the prediction models for CVs to enhance their performance".

Leila Ismail et al. [2] presented that "Diabetes, which can be brought on by a variety of lifestyle factors, including genetic, psychological, and physiological risk factors, is one of the top 10 killers in the world. Type 2 diabetes prediction is crucial for allied health professionals to forecast or diagnose type 2 diabetes and assist in the creation of an effective and efficient preventative strategy. Machine learning techniques for predicting type 2 diabetes have been proposed in a number of research. It is challenging to compare the findings, though, because each study evaluates algorithms using a different dataset and evaluation criteria. To conduct a fair and impartial evaluation, we present a classification system of risk factors associated with diabetes and assess the efficacy of 35 distinct machine learning algorithms, with and without feature selection techniques for the prediction of type 2 diabetes. For the evaluation, we employ 9 feature selection techniques and 3 real-world diabetes datasets. We assess the model construction and validation precision, F-measure, and execution time of the under-consideration approaches on diabetic and non-diabetic patients".

Huaping Zhou et al. proposed that "Because of some complications, diabetes is most prominent prevalent, fatal, and Chronic illnesses are amongst the most pervasive conditions globally.Prompt diagnosis is pivotal to the efficacious management of diabetes, as it can impede the advancement of the ailment.The suggested approach can be used to identify the ailment a person has and forecast if they will eventually develop diabetes. This approach will aid in giving the patient the right medication because type 1 diabetes and type 2 diabetes are managed in very different ways. Our approach converts the task into a classification issue and employs dropout regularisation to avoid overfitting. Usually, a deep neural network's hidden layers are used to develop it. Numerous studies have focused on the diabetes and diabetic type databases in the Pima Indian population.

The results of the studies demonstrate the advantages of By leveraging specific personalized parameters and the binary cross-entropy loss function, we successfully designed a deep neural network predictive model with exceptional precision. The trial outcomes affirm the potency and practicality of the DLPD (Deep Learning for Predicting Diabetes) model. Notably, the Pima Indians and diabetes types data sets exhibited the highest training accuracy, registering at 94% and 99.4%, respectively. Numerous studies have focused on the diabetes and diabetic type databases in the Pima Indian population. The results of the studies demonstrate the advantages of our suggested strategy over cutting-edge techniques".

Raja Krishnamoorthi et al. [4] implemented that "Healthcare prediction c. The main objective of the study was to examine the potential applications of big data analytics and machine learning in the management of diabetes. According on the data analysis, the suggested ML-based framework can score an 86. Health care providers and other interested parties are striving to develop categorization models to help with diabetes prediction and the creation of prevention interventions. The authors conducted a comprehensive review of machine learning literature and developed an intelligent framework for predicting diabetes based on their discoveries. Besides critically evaluating machine learning techniques, the authors showcase and evaluate an intelligent machine learning-based architecture for diabetes prediction. The decision tree (DT)-based random forest (RF) and support vector machine (SVM) learning models for diabetes prediction, currently the most popular approaches in the literature, were developed and evaluated by the authors of this work. This study argues for the development of a state-of-the-art machine learning system for anticipating diabetes mellitus (IDMPF). The framework states that it was created after carefully analysing the prior Prediction models documented in the literatureand determining if they were appropriate for diabetes. Using the framework, the authors provide training techniques, model evaluation techniques, problems with diabetes prediction, and potential solutions".

III. MACHINE LEARNING ALGORITHMS AND DATASETS USED FOR DIABETES PREDICTION

A. Machine Learning Algorithms

We start by outlining the several machine learning methods that were employed in this study to predict diabetes. The chosen techniques integrate statistical, perceptron-based, and supervised learning algorithms. The effectiveness of these strategies was evaluated after they were applied to the two discussed well-known datasets [4].

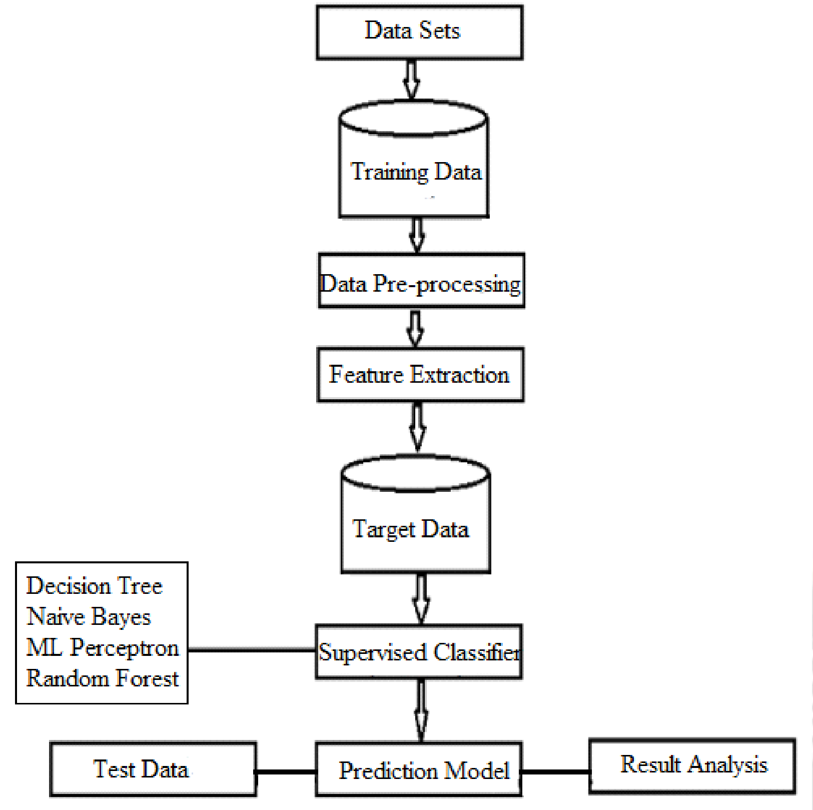


Fig. 2. Proposed Methodology

1) *Decision Tree Algorithm:*

Classification and regression the challenges can be addressed through implementation of the supervised learning methodology commonly referred to as decision tree. however this approach is frequently preferred. A decision tree is a classifier with a tree-like structure, where the leaves denote the classification results, and internal nodes correspond to dataset features.

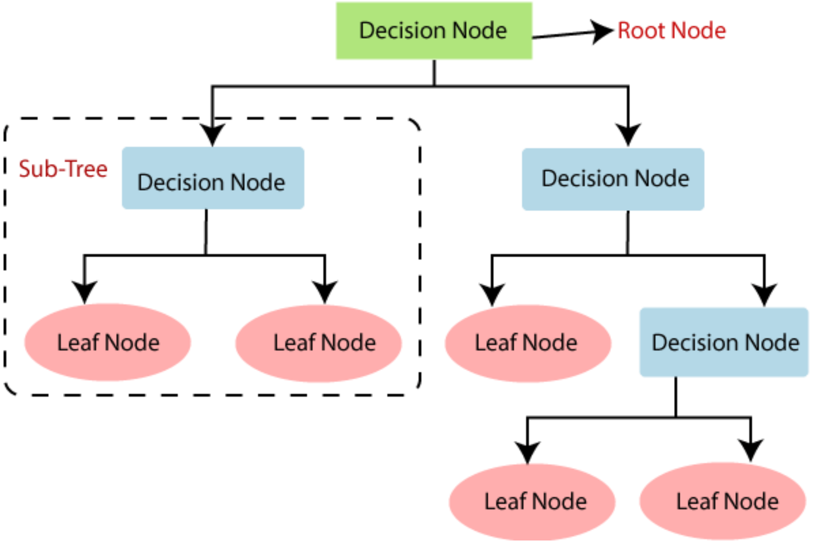


Fig. 3. Decision Tree Algorithm

The basic technique builds a tree [5] outlining the possible options and outcomes, which is then utilised for prediction.

Below equation 1 specifies that the method selects the branch at each tree node with the greatest information gain.

$$IG_R = H_D - H_{D|R} \tag{1}$$

DOI: https://doi.org/10.17762/ijritcc.v11i5s.6594
***Article Received**: 22 February 2023 **Revised**: 18 March 2023 **Accepted**: 18 April 2023*

Where H_D stands for the basic entropy and $H_{D|R}$ for the conditional entropy, and R is the risk factor.

2) Multi Layered Perceptron (MLP)

The Multi-Layer Perceptron is used in addition to the feed forward neural network (MLP). It has input, output, and hidden layers, which are three different types of layers. The input layer is responsible for receiving the signal that requires processing, while the output layer accomplishes the task at hand, such as classification or prediction. The input and output layers are positioned between an arbitrary quantity of hidden layers that comprise the actual computational engine of the Multilayer Perceptron (MLP) [6]. simply changing the weights in the hidden layers to increase the subsequent iteration's prognostication accuracy.

$$a = \emptyset(Wi + b) \qquad (2)$$

i is the input vector of the risk factors, W is the weight matrix for each risk factor, b is the bias vector, (.) is the sigmoid activation function, and an is the vector output consisting of diabetes and non-diabetes class labels.

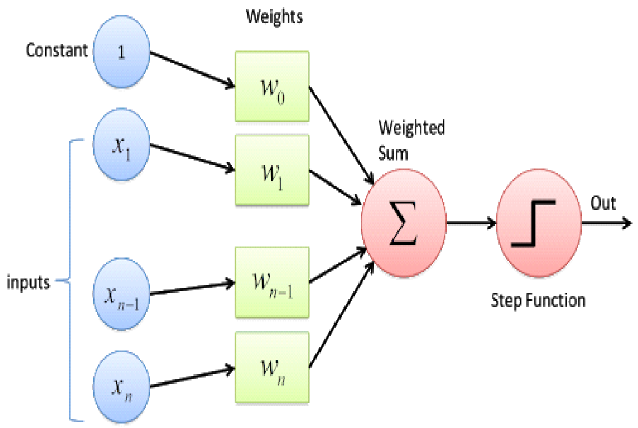


Fig. 4. MLP Algorithm

2) Naive Bayes (NB)

Based on the Bayes theorem, the Nave Bayes algorithm is a supervised learning technique for classification problems. It provides predictions based on the likelihood that an object will occur because it is a probabilistic classifier.

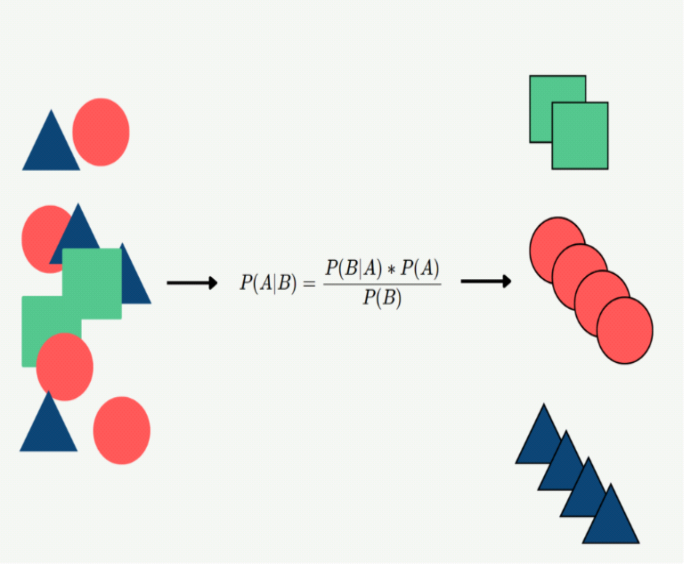


Fig. 5. NB Algorithm

$$P(C|R) = \frac{P(R|C).P(C)}{P(R)} \qquad (3)$$

where P(R|C) is the Likelihood, P(C) is the Class Prior Probability, and P(R) is the Predictor Prior Probability, and P(C|R) is the Posterior Probability.

3) Random Forest (RF)

Random Forest can be used for ML problems involving both regression and classification. It is based on the concept of ensemble learning, which is a technique for combining several classifiers to handle challenging problems and improve model performance.

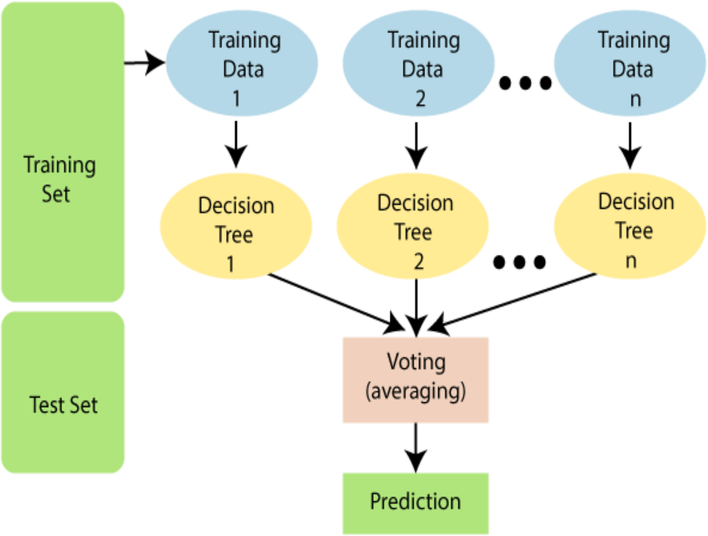


Fig. 6. RF Algorithm

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6594
***Article Received**: 22 February 2023 **Revised**: 18 March 2023 **Accepted**: 18 April 2023*

B. Datasets for Analysis

We give a brief description of the two datasets that will be used to assess the performance of the diabetes prediction algorithms described in the section above.

1) Pima Dataset Description

The dataset was initially created by the National Institute of Diabetes and Digestive and Kidney Diseases with the intention of estimating a patient's risk of getting diabetes based on clinical diagnostic variables. Female patients over the age of 21 are included in the dataset, along with a few medical predictor variables and one outcome variable called "Outcome." Some of the factors considered are age, the number of pregnancies, blood pressure, triceps skinfold thickness, insulin, body mass index (BMI), and diabetes pedigree function (DPF). There are 768 items in the dataset being used, 268 of which have diabetes, and 500 of which do not. All these traits are listed and described in Table I.

Feature	**Description**
Pregnancies	Number of pregnancies a patient has had earlier
Glucose	Glucose level present in the patient
Blood Pressure	Recorded blood pressure level at that particular time
Skin Thickness	Skin thickness level of the patient
Insulin	Amount of Insulin present in the patient's body
BMI	Body Mass Index of the patient
Diabetes Pedigree Function	The patient's family history of diabetes
Age	The age of the patient

2) Early Diabetes Classification Dataset

The dataset's 16 attributes are used to estimate the probability of developing diabetes. Patients at the Sylhet Diabetes Hospital in Sylhet, Bangladesh, provided 520 observations with 17 features using direct surveys and diagnosis results. The information about the following characteristic is shown:

a) Age	: 20-65
b) Sex	: Male/Female
c) Polyuria	: Yes/No
d) Polydipsia	: Yes/No
e) Sudden weight loss	: Yes/No
f) Weakness	: Yes/No
g) Polyphagia	: Yes/No
h) Genital thrush	: Yes/No
i) Visual blurring	: Yes/No
j) Itching	: Yes/No
k) Irritability	: Yes/No
l) Delayed healing	: Yes/No
m) Partial paresis	: Yes/No
n) Muscle stiffness	: Yes/No
o) Alopecia	: Yes/No
p) Obesity	: Yes/No
q) Class	: Positive/Negative

IV. RESULTS AND DISCUSSION

The execution times, F-measure, and accuracy of the various algorithms are contrasted and compared in this section. We assessed the machine learning methods reported in Section II-A using the datasets described in Section II-B. The average of the outcomes from each algorithm's five runs was used to enhance the reliability and validity of the experimental findings. Equations 4 and 5 are used to determine the accuracy and F-measure, respectively [7].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (4)$$

$$F - Measure = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \qquad (5)$$

$$Precision = \frac{TP}{TP + FP} \qquad (6)$$

$$Recall = \frac{TP}{TP + FN} \qquad (7)$$

A. Performance Analysis

The performance of the Decision Trees, Naive Bayes, Multi-layered Perceptron, and Radial Basis algorithms on the mixed datasets is examined in this section using several carefully selected features.

DOI: https://doi.org/10.17762/ijritcc.v11i5s.6594
***Article Received**: 22 February 2023 **Revised**: 18 March 2023 **Accepted**: 18 April 2023*

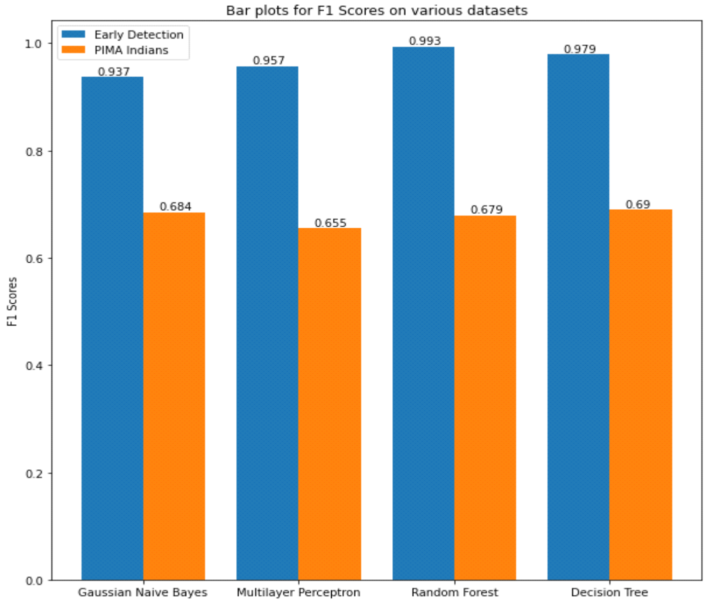


Fig. 7. F-Measure of Algorithms

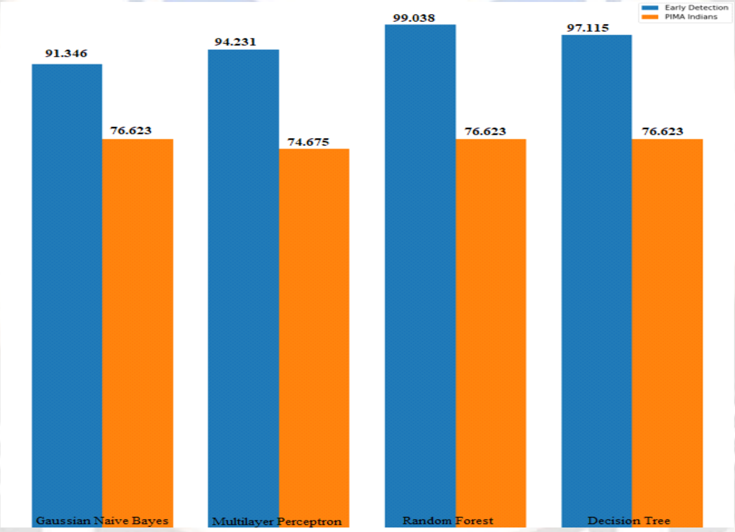


Fig. 8. Accuracy of Algorithms

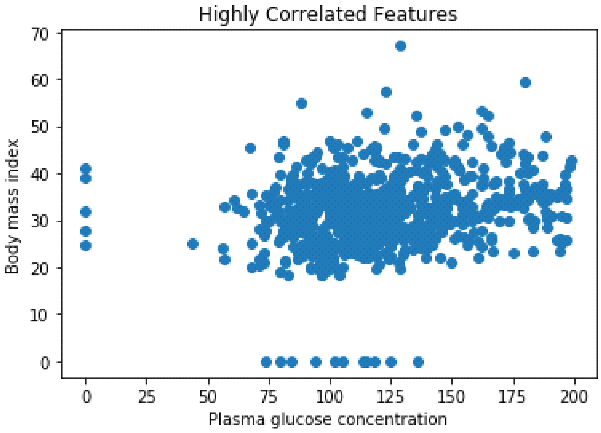
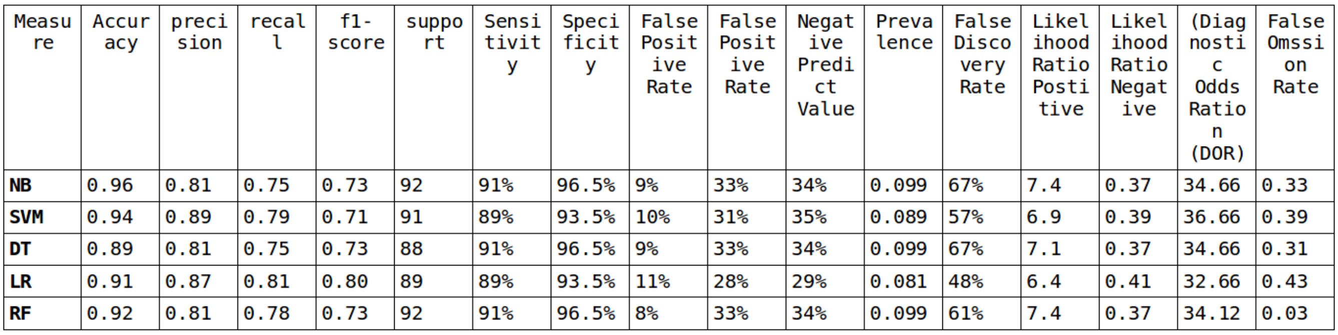


Fig. 9. Highly correlated features.



Measure	Accuracy	precision	recall	f1-score	support	Sensitivity	Specificity	False Positive Rate	False Positive Rate	Negative Predict Value	Prevalence	False Discovery Rate	Likelihood Ratio Positive	Likelihood Ratio Negative	(Diagnostic Odds Ration (DOR)	False Omssion Rate
NB	0.96	0.81	0.75	0.73	92	91%	96.5%	9%	33%	34%	0.099	67%	7.4	0.37	34.66	0.33
SVM	0.94	0.89	0.79	0.71	91	89%	93.5%	10%	31%	35%	0.089	57%	6.9	0.39	36.66	0.39
DT	0.89	0.81	0.75	0.73	88	91%	96.5%	9%	33%	34%	0.099	67%	7.1	0.37	34.66	0.31
LR	0.91	0.87	0.81	0.80	89	89%	93.5%	11%	28%	29%	0.081	48%	6.4	0.41	32.66	0.43
RF	0.92	0.81	0.78	0.73	92	91%	96.5%	8%	33%	34%	0.099	61%	7.4	0.37	34.12	0.03

Table-2. F1-score values before ensemble.

V. CONCLUSION

Type 2 diabetes is spreading globally because of urbanization, novel meals, and altering lifestyles. We classify the risk factors for type 2 diabetes to establish which categories are most important for predicting the condition. Numerous studies have been conducted utilizing various algorithms to precise estimation type 2 diabetes. Due to the fact that these algorithms were assessed using various datasets and evaluation metrics, it was challenging to compare how well they performed. In this study, we combine two diabetes datasets into a single configuration and analyse the precision, F-measure, and processing times of several methods. Our test findings demonstrate that, when compared to the other three models, the Random Forest technique is the most accurate.

References

[1] Deberneh, H.M.; Kim, I. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. Int. J. Environ. Res. Public Health 2021, 18, 3317.

[2] Leila Ismail, “Type 2 Diabetes with Artifcial Intelligence Machine Learning: Methods and Evaluation”, Archives of Computational Methods in Engineering (2022) 29:313–333 .Akib, M.G.A., Ahmed, N., Shefat, S.N. and Nandi, D., 2022. A Comparative Analysis among Online and On-Campus Students Using Decision Tree.

[3] Huaping Zhou, Diabetes prediction model based on an enhanced deep neural network”, EURASIP Journal on Wireless Communications and Networking (2020) 2020:148.

[4] Raja Krishnamoorthi, “A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques”, Journal of Healthcare Engineering Volume 2022, Article ID 1684017.

[5] Ravindra Changala, “Development of Predictive Model for Medical Domains to Predict Chronic Diseases (Diabetes) Using Machine Learning Algorithms and Classification Techniques”, ARPN Journal of Engineering and Applied Sciences, VOL. 14, NO. 6, March 2019, ISSN 1819-6608.

[6] Reddy, A. Srinivasa. "Effective CNN-MSO method for brain tumor detection and segmentation." Materials Today: Proceedings 57 (2022): 1969-1974.

[7] WHO.Diabetes. Available online: https://www.who.int/newsroom/fact-sheets/detail/diabetes (accessed on 20 Aug 2022).