iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.69351

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Identifying the Best Classifiers for Numerical Healthcare Data for Diabetes Prediction

Prof. Raghuvir Joshi, Dr. Dhaval S. Vyas
Department of Information Technology, Surendranagar University

Abstract: ***In the field of ML-Machine Learning, classification is one of the most widely used prediction tasks. In recent era, ML is being widely deployed in almost every field of real-world applications including heathcare. When we use ML for healthcare applications, it should be our main goal to achieve highest possible accuracy. Accuracy of any model is dependent on training dataset and algorithm being implemented. Different characteristics of training dataset contribute significantly to achieve highest possible accuracy. If we talk about general observations then the healthcare applications related data are mainly numerical like test reports showing numerical values. Classification is a categorical task that is easy to understand by patients like whether someone is having a particular disease or not. In this research work, we have evaluated and compared performances of various classifiers to decide which classifier works best when the training data is exclusively numerical. Based on our experiments, we have observed that Logistic Regression, Neural Network and Naive Bayes perform more accurately for exclusively numerical data to predict diabetes.***

Keywords: Machine Learning, Diabetes Prediction, Classification, Accuracy, Decision Tree

I. INTRODUCTION

ML-Machine Learning is becoming more and more an essential way of making predictions in many fields, including healthcare applications. ML is all about learn from the historic data to build a model that can make future predictions. One of the most widely used ML tasks is classification that predicts categorical label for the input data, for example whether a patient is having a particular disease or not, whether a patient should be discharged from hospital or not, whether a particular health equipment is effective or not etc. so in heathcare, classification task can be used to make predictions about diagnosis of diseases, effectiveness analysis of treatments etc. Heathcare applications are directly dependent on patients' lives so accuracy of models play utmost role in successful implementation of ML based systems. Accuracies of various ML algorithms are highly dependent on quality and type of dataset being used for training purposes. In general cases, healthcare data are mostly numerical like test reports with numerical values for various tests such as blood pressure level, glucose level, cholesterol level etc. classification is categorical prediction. So in such cases, it is important to select right ML algorithm that can handle numerical data to predict categorical output with desired accuracy [1,2,3].

Over the years, ML is evolved with introduction of large number of prediction algorithms. Each algorithm has its own advantages and limitations. We have worked on Orange tool where the available classification algorithms are: Logistic Regression, Neural Networks, Naive Bayes, Decision Trees, Support Vector Machines, and K-Nearest Neighbors. Each of these algorithms process numerical training data differently to build a prediction model in its own unique manner so accuracy may differ from algorithm to algorithm. Accuracy is the most important parameter to achieve in healthcare as inaccurate model can make incorrect diagnosis of disease that might be life threatening to patients in critical situations such as cancers, strokes etc. Rather than believing to any one classifier directly, it is advantageous to compare performances of more than one classifiers so we can have a classifier that best suits type of data we have. In our research work, we have done the same thing. Our dataset is numerical dataset related to diabetes predictions. Diabetes is a common and serious health condition that affects millions of people throughout the world. Many countries are becoming diabetic hubs with faster rates of growing patients of all ages. Diabetes predictions are being more and more popular for people of every age. Our work is to identify the most appropriate classifier for diabetes prediction when training data is numerical [1,2,3].

The paper is organized into various sections for effective understand. Section-II discusses literature review highlighting related contributions in recent time by various researchers. Section-III discusses how different classifiers used in this research work fundamentally differ from each other. This section also discusses the implementation details like dataset and flow in Orange tool. Section-IV discusses results and performance analysis of various classifiers. This paper ends with conclusions highlighting our observations.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

II. LITERATURE REVIEW

Several studies have explored the application of machine learning in healthcare, particularly for disease prediction. Researchers have examined various classification algorithms to determine their effectiveness with numerical datasets. Previous works have highlighted that Logistic Regression is widely used for binary classification tasks in medical diagnoses. Studies have also shown that Neural Networks perform well with complex patterns in health data, while Naive Bayes is effective for probabilistic classification. Comparative analyses of different classifiers have demonstrated varying results, emphasizing the importance of dataset characteristics. This literature review provides a foundation for our study, which focuses on classifier performance for diabetes prediction. Table-1 presents some of the recent findings.

Table-1 Literature Review Summary

Sr.No.	Research Paper	Observation
1	Identifying Ethical Considerations for Machine Learning Healthcare Applications [4]	This paper outlines a way for identifying ML-HCA (Machine Learning – Health Care Applications) ethical concerns. The framework aims to facilitate ethical evaluations of ML-HCAs throughout their development and implementation stages.
2	Ethical Machine Learning in Healthcare [5]	This paper outlines different ethical concerns like amplification of existing health inequities by models. It is also discussed that how ML can be used in healthcare by focusing on fairness, equality and other such requirements.
3	Secure and robust machine learning for healthcare: A survey [6]	This paper outlines various application areas in healthcare and focuses on security and privacy aspects and associated challenges. In addition, potential methods to ensure secure and privacy-preserving ML for healthcare applications are also discussed.
4	Automated machine learning: Review of the state-of-the-art and opportunities for healthcare [7]	This paper outlines how Automated Machine Learning (AutoML) can be helpful and effective for healthcare application seven when people have limited data science expertise. Different opportunities and hurdles are also discussed.
5	Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine [8]	This paper outlines how AI and machine learning solutions help in advance healthcare discovery. Authors discuss how precision medicine improves healthcare by tailoring treatments based on patient data, enabled by technological advances. Integrating electronic health records and AI can optimize real-time decision support.
6	Synthetic data in machine learning for medicine and healthcare [9]	This paper outlines the proliferation of synthetic data in artificial intelligence for medicine and healthcare. Authors raise concerns about the vulnerabilities of the software and the challenges of current policies.
7	How to develop machine learning models for healthcare [10]	This paper outlines development of machine learning models for healthcare applications by considering various points for developing, validating and implementing these models. The main concern is improving clinical decision support and eventually improving patient care.
8	Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers [11]	This paper is focused on building a framework with a set of machine learning based methods with weighted ensembling concept.
9	Diabetes Prediction using Machine Learning Algorithms [12]	This work is based on analyzing external factors along with regular factors for classification purposed.
10	A review on current advances in machine learning based diabetes prediction [13]	This work summarizes the progress of diabetes prediction so far and associated challenges.
11	A comparison of machine learning algorithms for diabetes prediction [14]	This work discusses how different algorithms provide accurate results for a given dataset
12	Diabetes prediction using machine learning techniques [15]	This work discusses how random forest technique performs better as compared to other techniques

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

III.IMPLEMENTATIONS

A. Classification Algorithms

Classification is task to predict categorical outcome for example, student will pass or fail, a patient has a particular disease or not. It is mainly a part of supervised learning where historical data is processed to build a model that can classify future data for which output values will be unknown. The process remains the same though the way historical data (also called training data) is processed to build a model differ across various algorithms. A brief idea about most widely used classification algorithms is given below.

- Neural Network: A neural network is a computer model that works like a human brain. It has many small parts called "neurons" that connect to each other. These neurons take numbers as input, process them, and give an output. Neural networks are good at finding patterns and are used in things like speech recognition and image recognition.
- SVM (Support Vector Machine): SVM separates data into different groups. It works like deriving a line or a complex shaper between different groups of data. Here the goal is to find best possible line to separate different groups as far as possible.
- Logistic Regression: Logistic regression works like linear regression but using sigmoid function to find probability of each of the categorical output. The output that is having highest probability will be considered as output.
- kNN (k-Nearest Neighbors): kNN finds the distance between various data elements to identify which are similar and which are dissimilar. Based on such distance based calculations, it identifies the category of new data by considering new data is more closed to which type of data.
- Decision Tree: A decision tree derives a flowchart with yes/no type questions. Each question splits the data into smaller groups until a final decision is made. It is easy to understand and is used for things like diagnosing diseases or deciding loan approvals.
- Naïve Bayes: Naïve Bayes is a simple way to guess which category something belongs to. It assumes that different features of the data do not affect each other, even if they actually do. This makes it very fast and good for things like spam detection and text classification.

B. DataSet

Our research work is for the prediction of diabetes where we will classify a patient having diabetes vs a patient who is not having diabetes. We have used dataset available on Kaggle for this research work [16]. The dataset has records of 768 patients (268 positive records and 500 negative records). Various features (predictors) are given below. All these features (predictors) are numerical only.

The target is either 0 (patient has no diabetes) or 1 (patient has diabetes).

Pregnancies	Glucose	BloodPressure
SkinThickness	Insulin	BMI
DiabetesPedigreeFunction	Age	

C. Orange Tool for Implementation

Orange is a free tool for machine learning and data analysis. It helps people understand data using visual blocks instead of coding. You can drag and drop different tools to clean data, make charts, and train machine learning models. It is useful for beginners because it is easy to use and does not need programming skills. Scientists, students, and business people use Orange to find patterns in data and make smart decisions.

Figure-1 shows the general flow of Orange tool to evaluate performances of various classifier algorithms. The steps are listed below. The process is divided into four steps.

1) Import Dataset
2) Select Predictors and Target.
3) Train a Model
4) Test Model and Calculate Various Measures
5) Represent Model Performance in the form of Confusion Matrix

©IJRASET: All Rights are Reserved

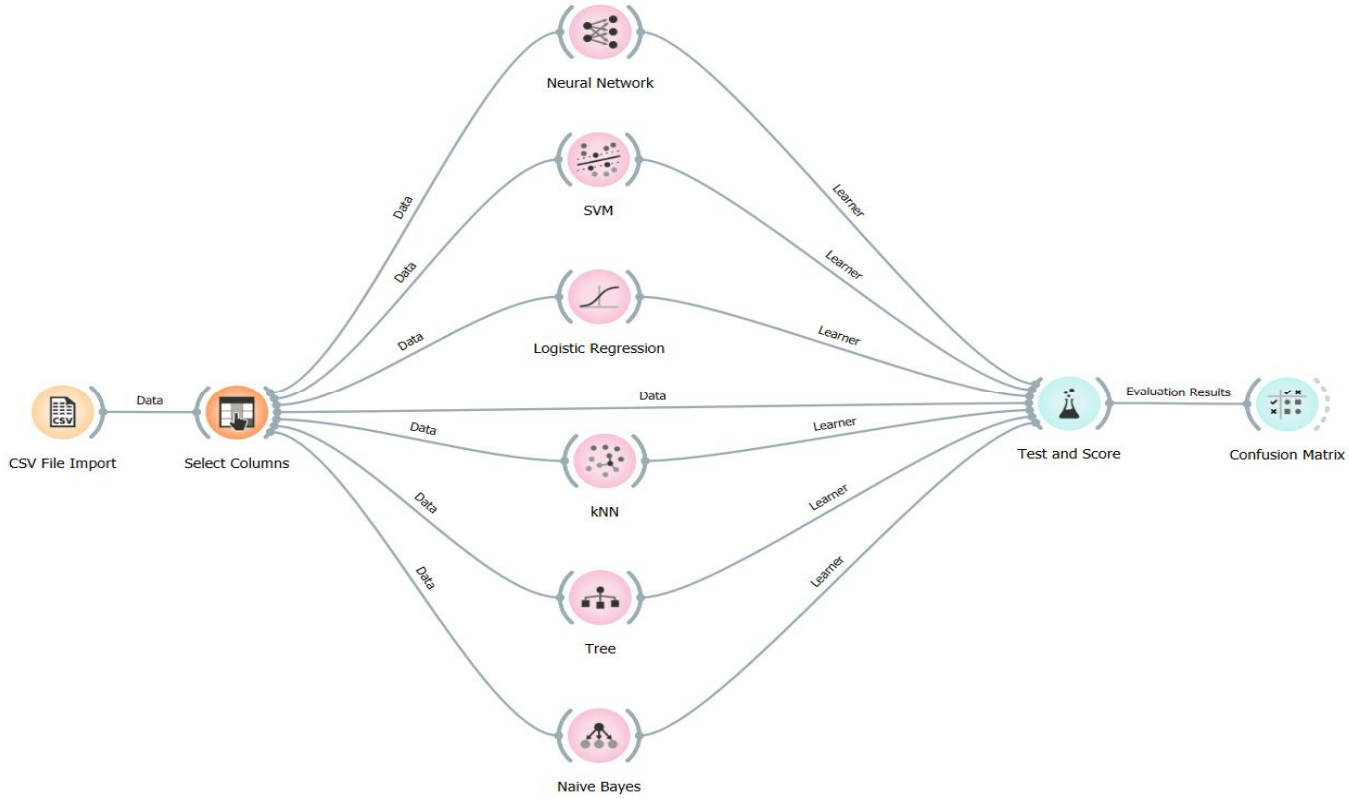


Figure – 1 Implementation Flow in Orange Tool

IV. RESULTS AND PERFORMANCE ANALYSIS

Cross-validation is a way to test how well a machine learning model works. Instead of using all the data to train the model and then testing it on a small part, cross-validation splits the data into multiple parts. The model is trained on some parts and tested on the remaining part, and this process is repeated several times. This helps to check if the model is learning well and will work on new data, not just memorizing what it has seen before. It also helps find the best model by comparing different ones and making sure it is not overfitting or underfitting the data. The subsequent content shows comparative results of various classifiers using different cross validations.

Cross Validation: 2 Fold

Model	AUC	CA	Prec	Recall
Logistic Regression	0.821	0.768	0.763	0.768
Tree	0.660	0.697	0.689	0.697
Naive Bayes	0.810	0.738	0.747	0.738
kNN	0.733	0.725	0.716	0.725
Neural Network	0.826	0.760	0.756	0.760
SVM	0.808	0.738	0.731	0.738

Cross Validation: 5 Fold

Model	AUC	CA	Prec	Recall
Logistic Regression	0.828	0.771	0.765	0.771
Tree	0.640	0.693	0.688	0.693
Naive Bayes	0.820	0.747	0.755	0.747
kNN	0.733	0.698	0.690	0.698
Neural Network	0.825	0.768	0.763	0.768
SVM	0.730	0.667	0.679	0.667

Cross Validation: 10 Fold

Cross Validation: 20 Fold

Model	AUC	CA	Prec	Recall
Logistic Regression	0.829	0.776	0.771	0.776
Tree	0.653	0.707	0.703	0.707
Naive Bayes	0.818	0.736	0.745	0.736
kNN	0.737	0.711	0.703	0.711
Neural Network	0.826	0.760	0.756	0.760
SVM	0.713	0.665	0.681	0.665

Model	AUC	CA	Prec	Recall
Logistic Regression	0.829	0.777	0.772	0.777
Tree	0.647	0.716	0.711	0.716
Naive Bayes	0.818	0.737	0.746	0.737
kNN	0.740	0.717	0.710	0.717
Neural Network	0.826	0.770	0.765	0.770
SVM	0.695	0.659	0.676	0.659

Logistic Regression

Actual \ Predicted	0	1	Σ
0	442	58	500
1	113	155	268
Σ	555	213	768

Tree

Actual \ Predicted	0	1	Σ
0	404	96	500
1	122	146	268
Σ	526	242	768

Naïve Bayes

Actual \ Predicted	0	1	Σ
0	383	117	500
1	85	183	268
Σ	468	300	768

kNN

Actual \ Predicted	0	1	Σ
0	409	91	500
1	126	142	268
Σ	535	233	768

Neural Network

Actual \ Predicted	0	1	Σ
0	425	75	500
1	102	166	268
Σ	527	241	768

SVM

Actual \ Predicted	0	1	Σ
0	343	157	500
1	105	163	268
Σ	448	320	768

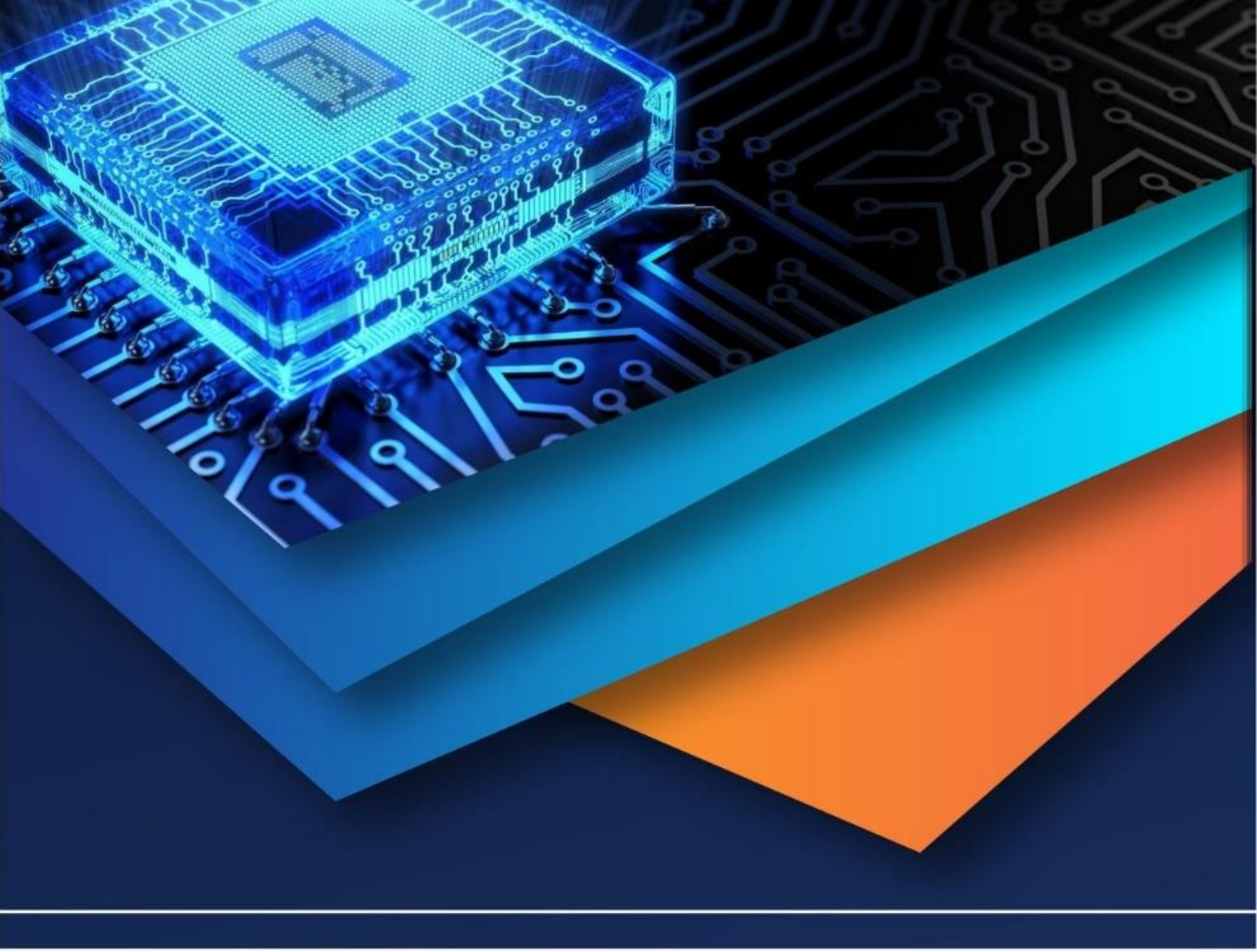
V. CONCLUSIONS

This research work aimed to identify the best classification algorithms to process exclusively numerical data. We used dataset to predict diabetes. As a part of our research work, 6 most widely used classification algorithms are evaluated using Orange tool. As a part of testing, various cross validations are made using different folds. We have observed that while processing numerical data, Logistic Regression, Naive Bayes and Neural Network algorithms perform the best as compared to Tree, kNN and SVM. This observation helps us to select which method to use for what type of dataset for to achieve higher accuracy. This work can be further extended to be evaluated with different datasets.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

REFERENCES

[1] Education, Pearson. Machine Learning, 1e. Pearson Education India., 2018
[2] Rebala, Gopinath, Ajay Ravi, and Sanjay Churiwala. An introduction to machine learning. Springer, 2019.
[3] Pereira, F. C., and S. S. Borysov. "Machine Learning Fundamentals Mobility Patterns, Big Data and Transport Analytics." (2019): Elsevier 9-29.
[4] Char, Danton S., Michael D. Abràmoff, and Chris Feudtner. "Identifying ethical considerations for machine learning healthcare applications." The American Journal of Bioethics 20.11 (2020): 7-17.
[5] Chen, Irene Y., et al. "Ethical machine learning in healthcare." Annual review of biomedical data science 4 (2021): 123-144.
[6] Qayyum, Adnan, et al. "Secure and robust machine learning for healthcare: A survey." IEEE Reviews in Biomedical Engineering 14 (2020): 156-180.
[7] Waring, Jonathan, Charlotta Lindvall, and Renato Umeton. "Automated machine learning: Review of the state-of-the-art and opportunities for healthcare." Artificial intelligence in medicine 104 (2020): 101822.
[8] Ahmed, Zeeshan, et al. "Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine." Database 2020 (2020): baaa010.
[9] Chen, Richard J., et al. "Synthetic data in machine learning for medicine and healthcare." Nature Biomedical Engineering 5.6 (2021): 493-497.
[10] Chen, Po-Hsuan Cameron, Yun Liu, and Lily Peng. "How to develop machine learning models for healthcare." Nature materials 18.5 (2019): 410-414.
[11] Hasan, Md Kamrul, et al. "Diabetes prediction using ensembling of different machine learning classifiers." IEEE Access 8 (2020): 76516-76531.
[12] Mujumdar, Aishwarya, and V. Vaidehi. "Diabetes prediction using machine learning algorithms." Procedia Computer Science 165 (2019): 292-299.
[13] Jaiswal, Varun, Anjli Negi, and Tarun Pal. "A review on current advances in machine learning based diabetes prediction." Primary Care Diabetes 15.3 (2021): 435-443.
[14] Khanam, Jobeda Jamal, and Simon Y. Foo. "A comparison of machine learning algorithms for diabetes prediction." Ict Express 7.4 (2021): 432-439.
[15] Soni, Mitushi, and Sunita Varma. "Diabetes prediction using machine learning techniques." International Journal of Engineering Research & Technology (Ijert) Volume 9 (2020).
[16] Diabetes dataset https://www.kaggle.com/datasets/mathchi/diabetes-data-set



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)