iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** III **Month of publication:** March 2024

DOI: https://doi.org/10.22214/ijraset.2024.59256

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 12 Issue III Mar 2024- Available at www.ijraset.com

Diabetes Prediction Using Machine Learning Algorithm

Miss. Vaishnavi Khalate[1], Prof. Borate Sukeshkumar[2]
Electronics Department, Pune University

Abstract: ***Diabetes, a chronic metabolic disorder affecting millions worldwide, requires early detection and management to mitigate its complications. Machine learning (ML) techniques have emerged as promising tools for predictive analytics in healthcare, offering the potential to improve diagnostic accuracy and patient outcomes. This paper presents a comprehensive review of ML algorithms applied to diabetes prediction, encompassing diverse methodologies and datasets.***

The study evaluates the performance of various ML algorithms, including but not limited to logistic regression, decision trees, support vector machines, random forests, and deep learning approaches, in predicting the onset or progression of diabetes. Additionally, feature selection techniques and data pre-processing methods are explored to enhance model robustness and interpretability.

Furthermore, this review highlights the significance of dataset characteristics such as size, imbalance, and feature diversity in influencing model performance. Challenges associated with model interpretability, scalability, and deployment in clinical settings are also discussed, alongside potential strategies to address these issues.

The findings suggest that ML algorithms demonstrate promising capabilities in diabetes prediction, with many studies reporting high accuracy, sensitivity, and specificity. However, there remains a need for standardized evaluation metrics and benchmark datasets to facilitate comparisons across studies. Moreover, efforts to enhance model interpretability and address data privacy concerns are crucial for promoting the adoption of ML-based predictive models in healthcare practice.

In conclusion, this review underscores the potential of ML techniques in diabetes prediction and emphasizes the importance of interdisciplinary collaboration between data scientists, clinicians, and healthcare stakeholders to leverage these advancements for improved patient care and disease management.

Keywords: Machine Learning, Supervised Machine Learning, Classification Algorithms, Logistic Regression, KNN, Naïve Bayes , Support Vector Machine, Random Forest

I. INTRODUCTION

Diabetes mellitus, a chronic metabolic disorder characterized by elevated blood glucose levels, poses a significant global health challenge. According to the International Diabetes Federation (IDF), approximately 463 million adults aged 20-79 years were living with diabetes in 2019, with projections indicating a rise to 700 million by 2045. The burden of diabetes extends beyond individual health, encompassing economic costs, reduced quality of life, and increased mortality rates due to its associated complications, including cardiovascular disease, kidney failure, and neuropathy.

Early detection and timely intervention are critical for managing diabetes and preventing its complications. Traditional diagnostic methods rely on clinical markers such as fasting blood glucose, oral glucose tolerance tests, and glycated hemoglobin (HbA1c) levels. While effective, these approaches may overlook subtle variations in individual risk factors and fail to account for the complex interplay of genetic, environmental, and lifestyle factors influencing diabetes onset and progression.

In recent years, machine learning (ML) techniques have gained traction as powerful tools for predictive analytics in healthcare. By leveraging large datasets comprising clinical, genetic, and lifestyle information, ML algorithms can uncover hidden patterns, identify high-risk individuals, and facilitate personalized interventions. This paradigm shift towards data-driven decision-making holds immense promise for improving diagnostic accuracy, optimizing treatment strategies, and ultimately enhancing patient outcomes in diabetes care.

This paper aims to explore the application of ML algorithms in predicting diabetes onset or progression. We conduct a comprehensive review of existing literature, examining the methodologies, datasets, and performance metrics employed across various studies. Furthermore, we discuss the challenges and opportunities associated with ML-based diabetes prediction, including model interpretability, data privacy concerns, and integration into clinical practice.

Through this review, we seek to shed light on the potential of ML techniques to transform diabetes management and pave the way for more proactive and personalized healthcare interventions. By bridging the gap between data science and clinical practice, we envision a future where predictive analytics empower healthcare providers to deliver timely and targeted interventions, thereby reducing the burden of diabetes on individuals and healthcare systems worldwide.

A. *Types Of Diabetes*

Diabetes is classified into several types, each with distinct etiologies, clinical presentations, and management strategies. Understanding these different types is essential for accurate diagnosis and tailored treatment plans. The main types of diabetes include:

1) Type 1 Diabetes (T1D): Type 1 diabetes, previously known as juvenile diabetes or insulin-dependent diabetes, typically develops in childhood or adolescence, although it can occur at any age. In T1D, the body's immune system mistakenly attacks and destroys insulin-producing beta cells in the pancreas, leading to a deficiency in insulin production. Individuals with T1D require lifelong insulin therapy to regulate blood sugar levels and prevent complications.
2) Type 2 Diabetes (T2D): Type 2 diabetes is the most common form of diabetes, accounting for the majority of cases worldwide. It usually develops in adulthood, although there has been an alarming rise in its prevalence among children and adolescents in recent years. In T2D, the body either becomes resistant to the effects of insulin or fails to produce enough insulin to maintain normal blood sugar levels. factors for T2D include obesity, physical inactivity, unhealthy diet, genetics, and age. Management of T2D typically involves lifestyle modifications (e.g., diet, exercise) and may include oral medications or insulin therapy.
3) Gestational Diabetes Mellitus (GDM): Gestational diabetes occurs during pregnancy and is characterized by high blood sugar levels that develop or are first recognized during pregnancy. Although GDM usually resolves after childbirth, affected individuals and their offspring are at increased risk of developing T2D later in life. Management of GDM aims to control blood sugar levels to minimize the risk of complications for both the mother and the baby.

B. *Symptoms Of Diabetes*

1) Frequent Urination (Polyuria)
2) Increased Thirst (Polydipsia)
3) Extreme Hunger (Polyphagia)
4) Unexplained Weight Loss
5) Fatigue and Weakness
6) Blurred Vision
7) Slow Healing of Wounds
8) Frequent Infections
9) Tingling or Numbness in Hands and Feet
10) Dry Skin and Itching

C. *Causes Of Diabetes*

1) Type 1 Diabetes (T1D):
- Autoimmune
- Genetic Factors

2) Type 2 Diabetes (T2D):
- Insulin Resistance
- Obesity and Lifestyle Factors
- Genetic Predisposition

3) Gestational Diabetes Mellitus (GDM):
- Hormonal Changes during Pregnancy
- Obesity and Excess Weight Gain
- Previous History of Gestational Diabetes

II. LITERATURE REVIEW

Deeraj Shetty et al. [15] proposed diabetes disease prediction using data mining assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patient's database. In this system, they propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbour) to apply on diabetes patient's database and analyze them by taking various attributes of diabetes for prediction of diabetes disease. Muhammad Azeem Sarwar et al. [10] proposed study on prediction of diabetes using machine learning algorithms in healthcare they applied six different. The analysis of related work gives results on various healthcare datasets, where analysis and predictions were carried out using various methods and techniques. Various prediction models have been developed and implemented by various researchers using variants of data mining techniques, machine learning algorithms or also combination of these techniques. Joshi et al. [12] presented Diabetes Prediction Using Machine Learning Techniques aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, ANN. This project pro- poses an effective technique for earlier detection of the diabetes disease. Deeraj Shetty et al. [15] proposed diabetes disease prediction using data mining assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patient's database. In this system, they propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbour) to apply on diabetes patient's database and analyze them by taking various attributes of diabetes for prediction of diabetes disease. Muhammad Azeem Sarwar et al. [10] proposed study on prediction of diabetes using machine learning algorithms in healthcare they applied six different machine learning algorithms Performance and accuracy of the applied algorithms is discussed and compared. Comparison of the different machine learning techniques used in this study reveals which algorithm is best suited for prediction of diabetes. Diabetes Prediction is becoming the area of interest for researchers in order to train the program to identify the patient are diabetic or not by applying proper classifier on the dataset

III. PROPOSED SYSTEM ARCHITECTURE

SR.NO.	Atributes
1	Pregnancies
2	Glucose
3	BloodPressure
4	Skin Thickness
5	Insulin
6	BMI
7	DiabetesPedegreeFunction
8	Age

1) There are a total of 768 records and 9 features in the dataset.
2) Each feature can be either of integer or float datatype.
3) Some features like Glucose, Blood pressure , Insulin, BMI have zero values which represent missing data.
4) There are zero NaN values in the dataset.
5) In the outcome column, 1 represents diabetes positive and 0 represents diabetes negative.

A. Distribution of Diabetic Patient

We made a model to predict diabetes however the dataset was slightly imbalanced having around 500 classes labeled as 0 means negative means no diabetes and 268 labeled as 1 means positive means diabetic.Various techniques of Machine Learning can capable to do prediction, however it's tough to choose best technique. Thus for this purpose we apply popular classification and ensemble methods on dataset for prediction.

The Pima Indian dataset is an open-source dataset that is publicly available for machine learning algorithms, which has been used in this work along with a private dataset. It contains 768 patients' data, and 276 of them have established diabetes. The objective is to predict diabetes based on measures that show whether the patient is diabetic or not. Data set consists of 768 data points, with 9 features each. The feature we are going to predict is "Outcome", 0 means No diabetes, 1 means diabetes.

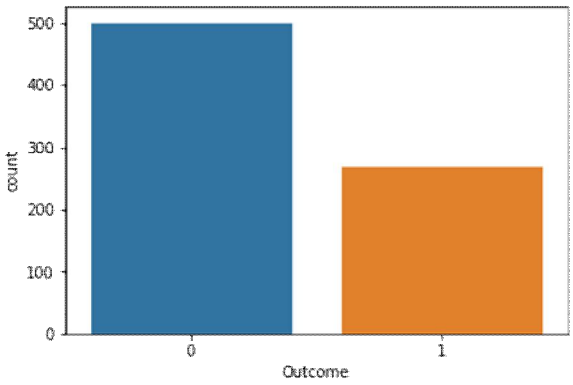


Fig1: Ratio Of Diabetes & Non-Diabetes Patients

1) *Data Collection and Preprocessing:* Data sources may include electronic health records (EHRs), patient demographics, medical history, laboratory test results (e.g., fasting blood glucose, HbA1c levels), lifestyle factors (e.g., diet, exercise), genetic information, and wearable sensor data (e.g., glucose monitors). Preprocessing involves cleaning the data, handling missing values, encoding categorical variables, scaling numerical features, and splitting the data into training and testing sets.
2) *Feature Engineering and Selection:* Feature engineering involves creating new features or transforming existing ones to improve model performance. Feature selection techniques such as correlation analysis, recursive feature elimination, and principal component analysis may be used to identify the most relevant features for prediction.
3) *Machine Learning Model Development:* Various machine learning algorithms can be employed for diabetes prediction, including logistic regression, decision trees, random forests,support vector machines, gradient boosting, and neural networks.

Multiple models may be trained and evaluated to determine the best-performing algorithm for the given dataset.

a) *Logistic Regression:* It is a supervised machine learning algorithm. Logistic regression is a statistical method used for binary classification tasks, meaning it predicts the probability of an instance belonging to one of two classes. Unlike linear regression, which predicts continuous values, logistic regression uses the logistic function (also called the sigmoid function) to squash the output into a range between 0 and 1. It predicts the probability between positive and negative class. With the sigmoid function, logistic regression calculates the probability that an instance belongs to the positive class. Typically, if the probability is above 0.5, the instance is classified as belonging to the positive class; otherwise, it belongs to the negative class. The decision boundary separates the classes in the feature space.
b) *Random Forest: Random Forest:* It is a popular machine learning algorithm used for both classification and regression tasks. It belongs to the ensemble learning methods, which combine multiple models to improve performance. Random Forest is an ensemble learning method that builds multiple decision trees, each trained on a random subset of data and features, and combines their predictions to produce a robust and accurate model for classification and regression tasks. It's known for its high performance, scalability, and resistance to overfitting.
c) *Decision Tree:* Decision tree is a supervised machine learning algorithm. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node. For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. To select attribute or feature it uses Attribute Selection measure (ASM) techniques.
d) *K- Nearest Neighbor:* K-Nearest Neighbors (KNN) is a simple and intuitive machine learning algorithm used for classification and regression tasks. KNN does not explicitly train a model. Instead, it stores the entire dataset in memory, effectively memorizing it. To predict the label (or target value) of a new instance, KNN calculates the distance between that instance and every other instance in the dataset. The distance measure used is typically Euclidean distance, but other measures like Manhattan distance or cosine similarity can also be used. KNN then selects the K nearest neighbors to the new instance based on the calculated distances. The value of K is a hyperparameter that needs to be specified beforehand. For classification tasks, KNN uses the class labels of the K nearest neighbors and employs majority voting to determine the predicted class of the new instance. For regression tasks, it calculates the average (or weighted average) of the target values of the K nearest neighbors to predict the target value for the new instance.

e) *Naïve Bayes:* Naive Bayes is a probabilistic machine learning algorithm based on Bayes' theorem. Naive Bayes learns the statistical relationship between the features and labels from the training data. It calculates the probabilities of each feature value occurring given each class label. Naive Bayes assumes that all features are conditionally independent given the class label. This means that the presence or absence of a particular feature is independent of the presence or absence of other features, given the class label. To classify a new instance, Naive Bayes calculates the probability of each class label given the features of the instance using Bayes' theorem. It multiplies the prior probability of each class (based on the frequency of each class in the training data) with the conditional probabilities of each feature given the class label. Naive Bayes assigns the class label with the highest probability as the predicted label for the new instance.

f) *Support Vector Machine (SVM):* is a supervised machine learning algorithm used for classification and regression tasks, but primarily known for classification. SVM maps the input features into a higher-dimensional space using a kernel function. This transformation allows the algorithm to find a linear decision boundary that best separates the classes in the higher-dimensional space, even if the original data may not be linearly separable. SVM aims to find the hyperplane that maximizes the margin, which is the distance between the hyperplane and the nearest data points from each class, known as support vectors. Maximizing the margin helps improve the generalization ability of the model.

In cases where the data is not perfectly separable, SVM allows for a soft margin, meaning it tolerates some misclassification errors to find a better decision boundary. This is controlled by a regularization parameter (C) that balances between maximizing the margin and minimizing the classification error.

SVM utilizes a kernel function to efficiently compute the dot product in the higher-dimensional space without explicitly transforming the data. Common kernel functions include linear, polynomial, radial basis function (RBF), and sigmoid. To classify new instances, SVM evaluates which side of the decision boundary they fall on in the higher-dimensional space. If they fall on one side, they are classified as belonging to one class; if they fall on the other side, they are classified as belonging to the other class.

g) *Model Training and Evaluation:* K-Nearest Neighbor model is selected as per the evaluation result. It gives the better accuracy for a prediction. The training phase involves fitting the machine learning models to the training data to learn the underlying patterns and relationships. Model performance is evaluated using appropriate metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) on the test dataset. Cross-validation techniques may be employed to assess the robustness of the models and mitigate overfitting.

h) *Hyperparameter Tuning:* Hyperparameters of the machine learning algorithms are optimized to improve model performance. Techniques such as grid search, random search, and Bayesian optimization may be used to find the optimal hyperparameters.

i) *Model Deployment:* Once the best-performing model is identified, it is deployed into production for real-time predictions. The model may be integrated into existing healthcare systems or deployed as a standalone application accessible to healthcare providers or patients.

j) *Monitoring and Maintenance:* Deployed model is continuously monitored to ensure its performance remains optimal over time. Regular updates and retraining of the model may be necessary to adapt to changes in patient populations, clinical guidelines, or data distributions.

k) *Interpretability and Explainability:* Such as feature importance analysis, SHAP (SHapley Additive exPlanations) values, and model-agnostic interpretability methods are used to explain the predictions of the machine learning models.

Interpretable models such as decision trees or linear models may be preferred in clinical settings where transparency and explainability are crucial.

l) *Privacy and Security:* Measures are implemented to ensure the privacy and security of patient data throughout the entire process, including data encryption, access controls, and compliance with healthcare regulations such as HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation).

By following this proposed system architecture, healthcare organizations can develop and deploy effective machine learning-based models for diabetes prediction, thereby facilitating early intervention, personalized treatment, and improved patient outcomes.

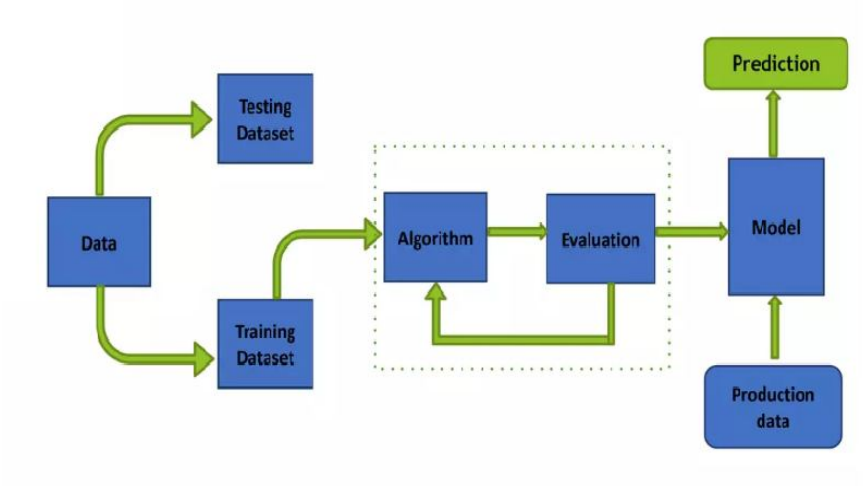


Fig2. Proposed Methodology Of Diabetes Prediction

Procedure of Proposed Methodology-

Step1: Import required libraries, Import diabetes dataset.

Step2: Pre-process data to remove missing data.

Step3: Perform percentage split of 80% to divide dataset as Training set and 20% to Test set.

Step4: Select the machine learning algorithm i.e. K- Nearest Neighbor, Support Vector Machine, Decision Tree, Logistic regression, Random Forest and Naïve Bayes.

Step5: Build the classifier model for the mentioned machine learning algorithm based on training set.

Step6: Test the Classifier model for the mentioned machine learning algorithm based on test set.

Step7: Perform Comparison Evaluation of the experimental performance results obtained for each classifier.

Step8: After analyzing based on various measures conclude the best performing algorithm.

IV. EXPERIMENTAL RESULTS

In this work different steps were taken. The proposed approach uses different classification and ensemble methods and implemented using python. These methods are standard Machine Learning methods used to obtain the best accuracy from data. In this work we see KNN classifier achieves better compared to others. Overall we have used best Machine Learning techniques for prediction and to achieve high performance accuracy. Figure shows the result of these Machine Learning methods.

Logistic Regression: 71.42857142857143

K Nearest neighbors: 78.57142857142857

Support Vector Classifier: 73.37662337662337

Naive Bayes: 71.42857142857143

Decision tree: 68.18181818181817

Random Forest: 75.97402597402598

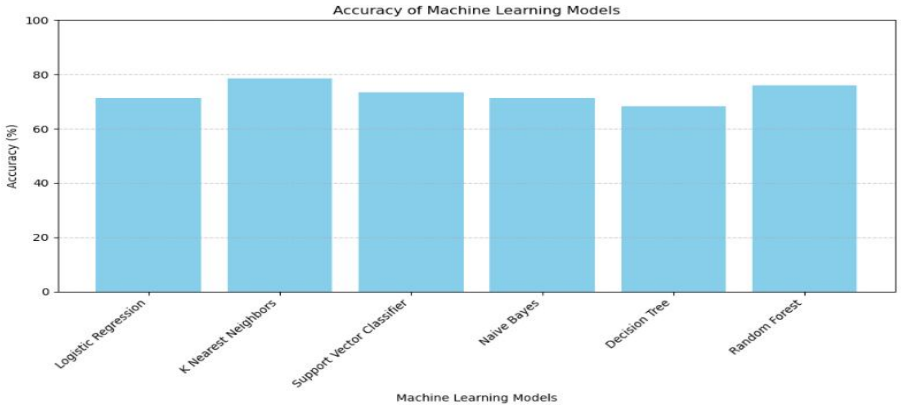


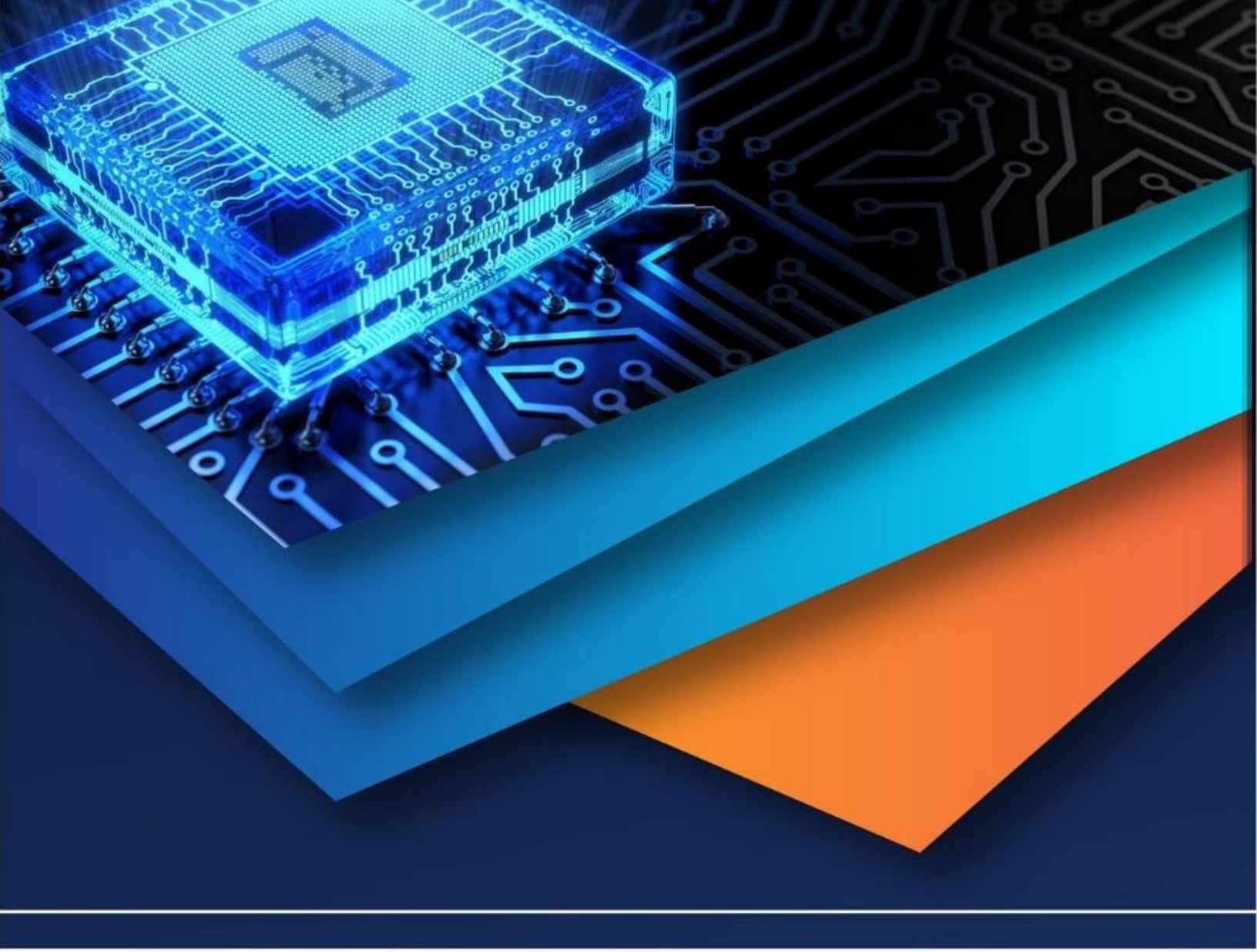
Fig3: Graph Between Models And Their Corresponding Accuracy

V. CONCLUSIONS

The main aim of this 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 are used. And 78.5% classification accuracy has been achieved by KNN algorithm. The Experimental results can be asst health care to take early prediction and make early decision to cure diabetes and save humans life.

REFERENCES

[1] Debarred Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analysing Feature Importance's for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.
[2] Salliah Shafi, Prof. Gufran Ahmad Ansari, "Early Prediction of Diabetes Disease & Classification of Algorithms Using Machine Learning Approach", International Conference on Smart Data Intelligence, 2021.
[3] KM Jyoti Rani, "Diabetes Prediction Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN: 2456- 3307, Volume 6 Issue 4, pp. 294-305, July-August 2020.
[4] A.K., Dewangan, and P., Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques," International Journal of Engineering and Applied Sciences, vol. 2, 2015.
[5] Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.
[6] B. G. Lee. Y. Chung, "A Smartphone based Driver safety Monitoring System Using Data Fusion", Sensors, 12, 2012, pp. 17536-17552.
[7] Choubey, D.K., Paul, S., Kumar, S., Kumar, S., 2017. Classification of Pima Indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection, in: Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016), pp. 451– 455.
[8] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC,978-1-5090-3243-3,2017.
[9] Ayush Anand and Divya Shakti," Prediction of Diabetes Based on Personal Lifestyle Indicators", 1st International Conference on Next Generation Computing Technologies, 978-1-4673-6809-4, September 2015.
[10] B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7,2017.
[11] Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing,2015.
[12] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly," Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.
[13] P. Suresh Kumar and S. Pranavi "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics", International Conference on Infocom Technologies and Unmanned Systems, 978-1-5386-0514-1, Dec. 18-20, 2017.
[14] Mani Butwall and Shraddha Kumar," A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier", International Journal of Computer Applications, Volume 120 - Number 8,2015.
[15] K. Rajesh and V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.
[16] Humar Kahramanli and Novruz Allahverdi,"Design of a Hybrid System for the Diabetes and Heart Disease", Expert Systems with Applications: An International Journal, Volume 35 Issue 1-2, July, 2008.
[17] B.M. Patil, R.C. Joshi and Durga Toshniwal,"Association Rule for Classification of Type-2 Diabetic Patients", ICMLC '10 Proceedings of the 2010 Second International Conference on Machine Learning and Computing, February 09 - 11, 2010.
[18] Dost Muhammad Khan1, Nawaz Mohamudally2, "An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm ", Journal Of Computing, Volume 3, Issue 12, December 2011.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY