International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Vol. 13, Issue 4, April-2024, Impact Factor: 8.285
Presented at "ICRETETM-2024", Organized by GSMCOE, Pune, on 22nd - 23rd April 2024

Predicting Diabetes with Machine Learning Methods

Aaditi Ranaware[1], Prof.Mukta Deshpande[2], Pratiksha Jadhav[3]

MCA (Masters of Computer Application), GenbaSopanrao Moze of Engineering Balewadi,Pune

ABSTRACT

High blood glucose levels are a chronic condition known as diabetes mellitus, which can lead to a number of consequences. With a surge in morbidity rates in recent years, projections indicate a staggering 642 million individuals worldwide afflicted by diabetes by 2040, indicating an impending health crisis demanding urgent attention. Using the recently emerging advances in machine learning, this work aims to be an example in the field of medical health, particularly in the field of diabetes, which has harmful effects on kidneys, eyes, hearts, nerves, and feet. As a new field of study within data science, machine learning presents opportunities to better understand how robots extract meaning from experience data. The goal of this study is to create a reliable system that can more accurately predict diabetes in the early stages.

Keywords: Diabetes Prediction, Supervised, SVM, XGBoost classifier, Random Forest, Decision tree.

INTRODUCTION

Diabetes, a prevalent chronic disease, poses a significant threat to human well-being by fostering the development of various other ailments such as heart attacks, blindness, and kidney diseases. The conventional diagnostic process entails patients visiting a diagnostic center, consulting with doctors, and awaiting reports, often causing delays in detection and prevention[1-2]. Early prevention and diagnosis are imperative to mitigate the risks associated with diabetes and prevent premature mortality. Machine learning, a subset of artificial intelligence, offers promising avenues for early disease prediction, thereby enabling timely intervention and treatment[3].

The Pima Indians Diabetes Database serves as a valuable resource for training and evaluating machine learning models aimed at predicting diabetes risk with heightened accuracy. This research endeavors to develop a predictive system leveraging four classification methods—Support Vector Machine, Random Forest, XGBoost, Decision Tree, and Artificial Neural Network algorithms. By harnessing machine learning techniques, this study contributes to the expanding knowledge base in the realm of diabetes prediction.

Diabetes mellitus, commonly known as diabetes, manifests as a chronic condition affecting blood sugar regulation. It manifests primarily in two forms:

Type 1 diabetes, characterized by the autoimmune destruction of insulin-producing beta cells in the pancreas, leading to insulin deficiency. On the other hand, Type 2 diabetes, which accounts for the majority of cases, is often associated with insulin resistance or insufficient insulin production. Factors contributing to each type vary:

Type 1 Diabetes Causes:

- Genetic predisposition increases susceptibility to type 1 diabetes.
- Environmental factors such as viral infections might trigger immune system attacks on insulin-producing cells.

Type 2 Diabetes Causes:

- Overweight and obesity contribute to insulin resistance.
- Physical inactivity exacerbates insulin sensitivity issues.
- Poor dietary habits, including high intake of processed foods and sugary beverages, exacerbate insulin resistance.

Through data mining and machine learning techniques, we aim to address the challenge of early diabetes detection, ultimately enhancing patient outcomes and mitigating the burden of this chronic disease[4].

This research paper delves into the creation of early detection systems for diabetes and their potential advantages. Diabetes often progresses silently, especially in its early stages, making early detection crucial for preventing complications. By utilizing prediction models, healthcare providers can design personalized treatment plans tailored to each person's risk factors and disease progression. This aids in better managing blood sugar levels and reducing the likelihood of severe long-term complications such as heart disease, stroke, blindness, and kidney failure. Early detection also allows healthcare systems to allocate resources more efficiently by prioritizing individuals at high risk who may require more frequent monitoring and care. This can lead to cost savings by reducing healthcare expenses associated with managing complications from undiagnosed or poorly controlled diabetes. On a larger scale, predicting diabetes within entire communities helps identify high-risk areas. This enables targeted initiatives to promote healthy lifestyles and increase awareness about diabetes. Early intervention at the community level has the potential to slow the spread of diabetes and alleviate its burden on healthcare systems[5]. Furthermore, the use of prediction models aids researchers in analyzing vast amounts of data to uncover new risk factors or patterns related to diabetes. This knowledge can drive the development of improved diagnostic tools, prevention strategies, and treatments for diabetes. Nevertheless, there are limitations to consider, such as the potential lack of generalizability of models trained on specific datasets to different population groups. Additionally, the accuracy of prediction models hinges on the quality of the data used for training, and there may be instances where insufficient information hampers the ability to provide specific preventive recommendations [6-8]. Despite these challenges, diabetes prediction systems hold immense promise for enhancing diabetes management and reducing its impact on individuals and healthcare systems alike [9].

II LITERATURE REVIEW

This section provides Literature review for prediction of diabetes using different approaches. Jaiswal et al. [1] provide a review of current advancements in machine learning for diabetes prediction. They highlight the use of various algorithms, including Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), Naive Bayes, Partial Least Squares – Discriminant Analysis (PLS-DA), and deep learning techniques. The review emphasizes the need for further improvement and validation of existing methods on diverse datasets. Tafa et al. [2] propose a hybrid classification approach combining Naive Bayes and Support Vector Machines (SVM) for enhanced diabetes diagnosis. This method achieves a high accuracy of 97.6%, suggesting its potential for improved diabetes prediction compared to single algorithms. It's valuable to explore further details of the study, such as the specific features used and how the hybrid approach leverages the strengths of both Naive Bayes and SVM. Sisodia & Sisodia [3] investigated the use of various classification algorithms for diabetes pred (SVMs). Interestingly, their findings showed Naive Bayes achieving the highest accuracy of 76.30%. Hussain and Naaz [4] take a theoretical approach in their review, comparing machine learning algorithms for diabetes diagnosis through mathematical methods. They compared the performance of Naive Bayes, decision trees, and Support Vector Machines analysis. They likely focus on Random Forest, Naive Bayes, and Neural network. Olaniyi & Adnan [5] explore the application of Artificial Neural Networks (ANNs), a type of deep learning technology, for diagnosing diabetes mellitus. Their model achieved an overall accuracy of 84.64%. Kandhasamy and Balamurali [6] investigated the use of various machine learning algorithms for predicting diabetes mellitus. Their study focused on Decision Tree (J48), K-Nearest Neighbors (KNN), Random Forest, and Support Vector Machines (SVM). Interestingly, their findings showed the J48 Decision Tree achieving the highest accuracy of 73.82% among the tested algorithms. Birjais et al. [7] explored the use of machine learning for predicting future diabetes risk. They compared three algorithms: Gradient Boosting, Logistic Regression, and Naive Bayes. Their findings suggest Gradient Boosting achieved the highest accuracy, reaching 86%. This indicates its potential for effectively identifying individuals at risk of developing diabetes. Swapna et al. [8] investigate the potential of deep learning for diabetes detection. Their study explores three approaches: Support Vector Machines (SVM), Convolutional Neural Networks (CNNs), and a hybrid network combining CNNs and LSTMs with SVM. The most impressive results came from the CNN LSTM-SVM network, achieving a maximum accuracy of 95.7%. Ooiting kee &harmizaharun [9] explores the potential of machine learning for incorporating cardiovascular complications into a diabetes prediction model. They compared three machine learning algorithms – Neural Networks, Decision Trees, and Support Vector Machines (SVM) to identify which performs best in predicting diabetes while considering the risk of cardiovascular complications. The Neural Network model achieved the highest accuracy, reaching 91%. This suggests it may be most effective in this specific context of predicting diabetes while also considering cardiovascular risk. Detail literature review is provided in Table 2 .

Table 2 Literature Review for Diabetic prediction

Sr. no	Paper	Name	Algorithms	Result
1	A review on current advances in Machine Learning based diabetes prediction	Varun Jaiswal,Anjl inegi,Tarun Pal	ANN, SVM, NaiveBayes, PLS DA and deeplearning	Existing Methods need to be improvised tested on different datasets.
2	Enhanced Diabetes Diagnosis Using a Hybrid Classification Approach	Tafa et al.	Combination of Naive Bayes and SVM	Achieving97.6%accuracy
3	Prediction Of Diabetes Using Classification algorithms	Sisodia & Sisodia	Applied Naïve Bayes, decision trees, and SVM	Naive Bayes achieving thehighestaccuracy76. 30%
4	Machine Learning in Medical Diagnosis: A Review of SL Algorithms for Diabetes Diagnosis	Hussain and Naaz	Random Forest, Naïve Bayes, Neural Networks	Compared using math
5	Applications of deep Learning Neural network technology for diagnosis of diabetes mellitus	Olaniyi & Adnan	Artificial Neural Network	Overall accuracy of the modelwas84.64%
6	Performance Analysis of Classifier Models to Predict Diabetes Mellitus	Kandhasamy and Bamurali	DecisionTree ,KNN,Random Forest, SVM	Decision TreeJ48 Classifier achieves high accuracy73.82%
7	Prediction & diagnosis of future diabetes risk Using ML	Birjais et.al.[8]	Gradientboosting, logistic regression, and naiveBayes classifiers	Gradient boosting having high accuracy of86%
8	Diabetes detection using deep learning	SwapnaG, Vinayakumar R,SomanK.P.	SVM, CNN, Hybrid network	The maximum accuracy value of95.7%wasobtainedforCN N 5-LSTMwithSVMnetwork
9	Cardiovascular Complications In a diabetes prediction model Using ML	Ooi ting kee,harmiz aharun	Neural network model, Decision tree, SVM	Neural network model with0.91%accuracy

International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Vol. 13, Issue 4, April-2024, Impact Factor: 8.285
Presented at "ICRETETM-2024", Organized by GSMCOE, Pune, on 22nd - 23rd April 2024

III. RESEARCH METHODOLOGY

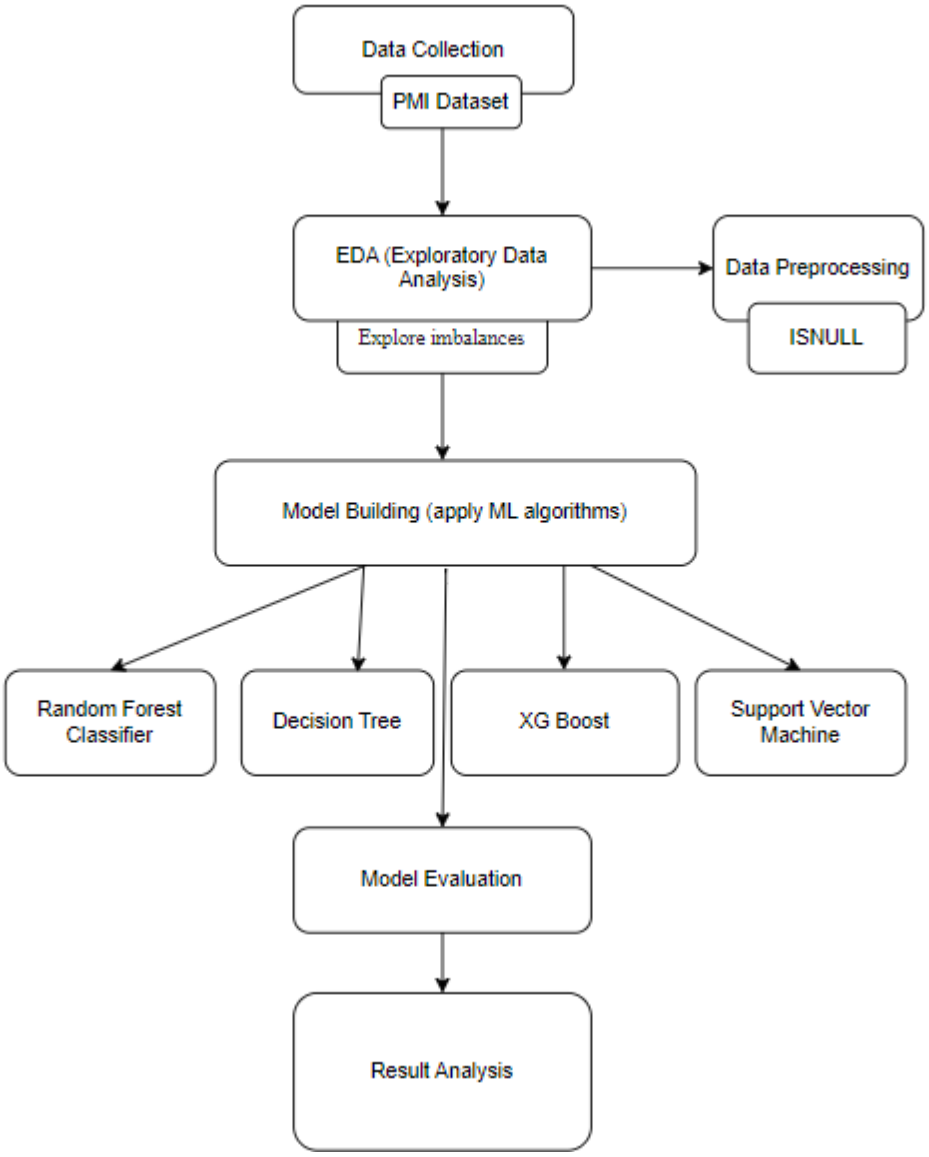


Figure.1 Research Methodology for Diabetes Prediction

Step 1 Data Collection:

Our research hinges on a dataset containing medical information about individuals, encompassing variables such as blood sugar levels, weight, family medical history, and age. Our primary aim is to predict the likelihood of diabetes occurrence in patients by leveraging specific features in our machine learning model. To achieve this objective, we will employ the widely recognized Pima Indians Diabetes Database, which provides valuable insights conducive to diabetes prediction using machine learning techniques. This dataset comprises 768 entries, each including 9 features per individual, such as Pregnancies, blood pressure, Glucose, BMI, Skin thickness, Insulin, BMI, Pedigree, and age, as illustrated in Figure 2. An essential binary outcome variable signifies the presence or absence of diabetes.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Pedigree	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure.2 Sample Data Set

Step 2 Exploratory Data Analysis (EDA)

In our EDA for diabetic prediction, we will thoroughly examine key parameters such as pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree function, age, and the outcome variable indicating diabetes presence. Through a combination of visualizations and statistical analyses, our objective is to uncover any discernible patterns or correlations between these parameters and the occurrence of diabetes. Our focus will include assessing the variation of each parameter among individuals with and without diabetes, with the aim of identifying potential risk factors and insights that can guide our predictive modeling efforts. This comprehensive exploration forms the foundation for the development of an effective diabetic prediction model.

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

Figure.3 EDA Analysis for diabetic prediction

Step 3 Data pre-processing

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

The raw data might have inconsistencies or missing values. This stage ensures the data is clean and formatted appropriately for the machine learning algorithms.

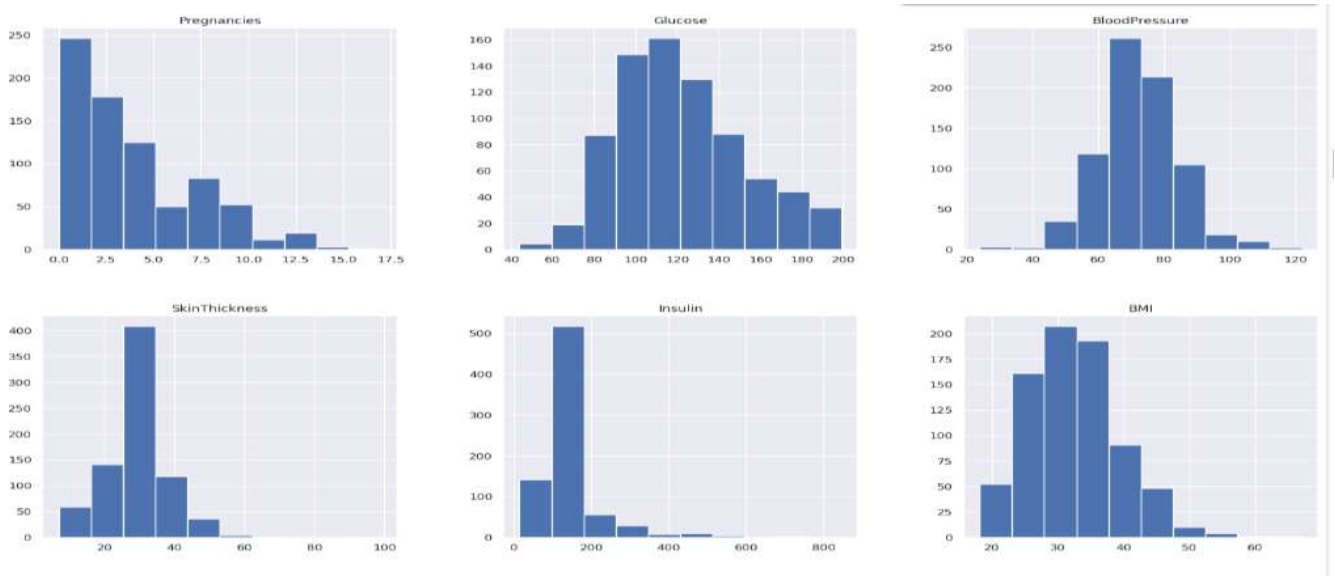
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False
5	False	False	False	False	False	False	False	False	False
6	False	False	False	False	False	False	False	False	False
7	False	False	False	False	False	False	False	False	False
8	False	False	False	False	False	False	False	False	False
9	False	False	False	False	False	False	False	False	False

Figure.4 Data Preprocessing

Step 4 Classification Model Training: Here, a machine learning model is trained on the prepared data. Common algorithms for thistask include Logistic Regression, Random Forest, and Support Vector Machines. The model learns to identify patterns in the data that differentiate diabetic from non- diabetic individuals. **Step 5 Analysis and Evaluation of trained models for Prediction:** Once trained, the model can then analyze new, unseen data. Based on the patterns learned, it predicts whether a new individual is likely to be diabetic or not.

IV EXPERIMENTAL RESULT

As figure.5 shows the occurrence of every kind of value in the graphical structure which in turn lets us know the range of the data**.**



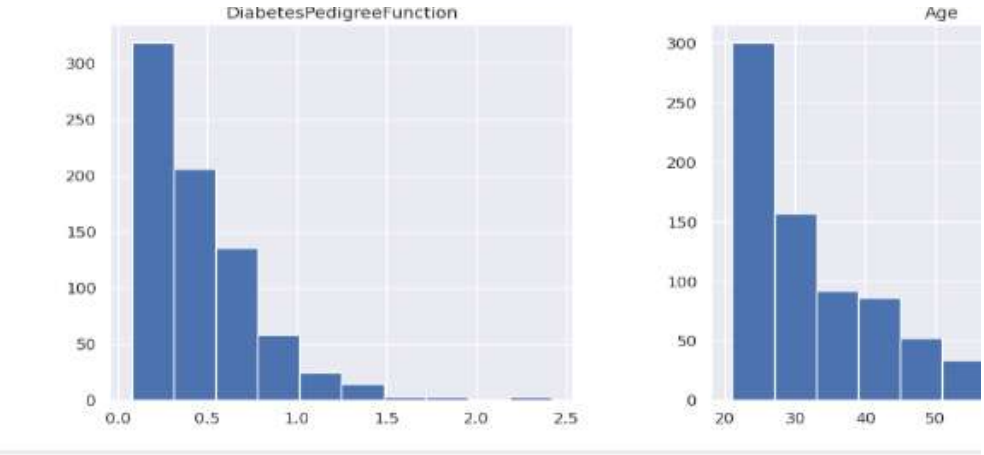


Figure.5 Graphical Representation of Data

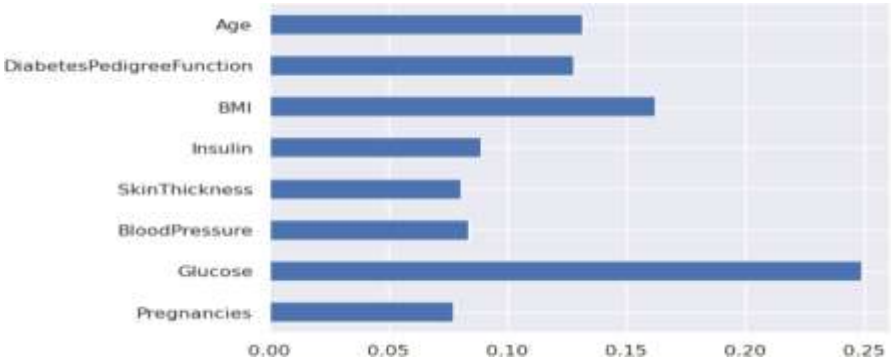


Figure.6 Important features from dataset

Here from the above Figure 6, it is clearly visible that Glucose as a feature is the most important in this dataset. Figure .7 depicts a correlation between all features and simplified view of a machine learning model for diabetes prediction. By analyzing patterns in the data, the model aims to make accurate predictions about the risk of diabetes.

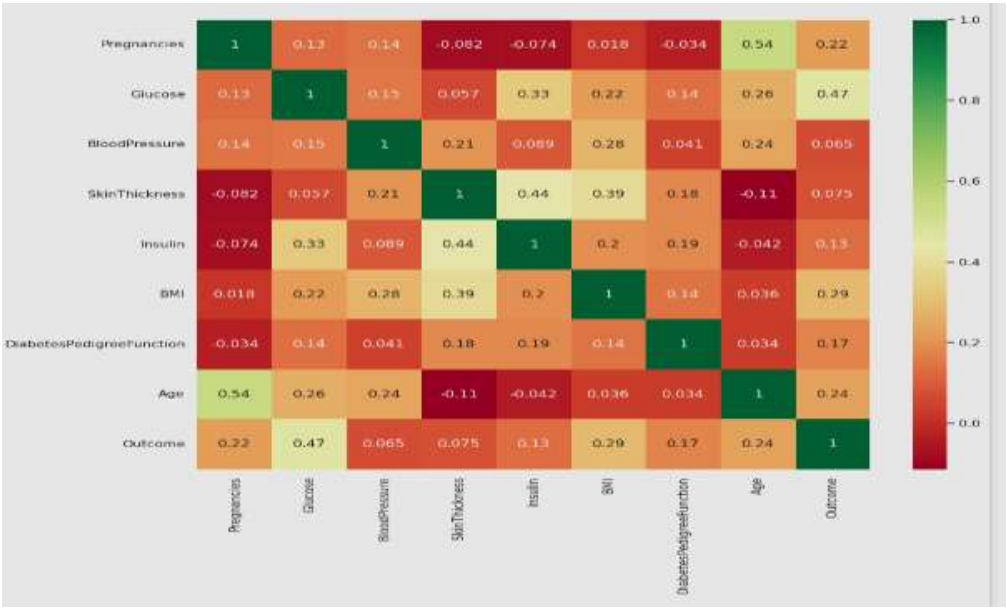


Figure.7 Correlation Matrix Between Features

Table 3 Obtained results with Classification models

Classification	Training Accuracy	Testing Accuracy
Decision Tree	0.7	0.73
SVM	0.5	0.74
Random Forest	1.0	0.76
XGboost Classifier	0.6	0.74

Table 3 provides information about results obtained with classification models. The Decision Tree model shows decent performance with a testing accuracy slightly higher than the training accuracy. This indicates that the model generalizes reasonably well to unseen data.

The SVM model demonstrates better testing accuracy compared to training accuracy. However, the low training accuracy suggests that the model may not have learned the underlying patterns effectively during training. The Random Forest model exhibits perfect training accuracy, suggesting potential overfitting to the training data. However, the testing accuracy is reasonably high, indicating that the model still performs well on unseen data. The XGBoost Classifier shows relatively balanced performance between training and testing accuracies. While the testing accuracy is decent, it could potentially be improved with further tuning or feature engineering.

Overall, the Random Forest model stands out with the highest testing accuracy, followed closely by the Decision Tree and XGBoost Classifier. The SVM model, despite having a higher testing accuracy, exhibits lower training accuracy, suggesting potential limitations in its learning capacity.

CONCLUSION

In conclusion, this research has investigated the application of machine learning techniques for predicting diabetes using the Pima Indians Diabetes Dataset. Through comprehensive exploration and analysis, several noteworthy findings have been uncovered. Four prominent machine learning algorithms were scrutinized: Random Forest, Decision Tree, XGBoost Classifier, and Support Vector Machine (SVM). Among these, the Random Forest model emerged as the most effective, attaining an impressive accuracy score of 0.76 on the testing dataset. Notably, feature importance analysis revealed "Glucose" as the most influential predictor in the model, emphasizing its significance in diabetes prediction. Leveraging its superior performance, the Random Forest model was saved using pickle for potential deployment in real-world applications, offering efficient and accurate diabetes predictions. Overall, this study underscores the promise of machine learning in diabetes prediction and lays the foundation for future research endeavors aimed at advancing diabetes management through sophisticated machine learning methodologies.

REFERENCES

[1]. Jaiswal, V., Negi, A., & Pal, T. (2021). A review on current advances in machine learning based diabetes prediction. *Primary Care Diabetes*, *15*(3), 435-443.
[2]. Sisodia, D., Singh, L., Sisodia, S., 2014. Fast and Accurate Face Recognition.
[3]. 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 (ICCCS2016), pp. 451 455.
[4]. Muhammed, M. A., Al-Janabi, S. S., & Khamit, N. F. (2023). Diabetes detection based on machine learning and deep learning approaches. Multimedia Tools and Applications, 1-22
[5]. https://www.kaggle.com/johndasilva/diabetes
[6]. Agrawal, P.,Dewangan, A.:A brief survey on the technique used for the diagnosis diabetes-mellitus. Int.Res.J. Eng.Technol. (IRJET).02(03)(2015). e-ISSN:2395-0056;p-ISSN:2395-0072
[7]. Diabetes, World Health Organization (WHO):30Oct2018.

[8]. Devi, M.Renuka, and J.Maria Shyla. "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus." International Journal of Applied Engineering Research 11.1(2016):727-730.
[9]. Witten, Ian H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.
[10]. Ramachandran A, Snehalatha C, Salini J, Vijay V. Use of glimepiride and insulin sensitizers in the treatment of type2 diabetes- a study in Indians. *JAssoc Physicians India.* 2004;52:459–63.[PubMed][Google Scholar]