IJARST
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

Supervised Learning Algorithm for Prediction of Type 2 Diabetes Mellitus Disease

Babu Rao[1], A. Vyshnavi[2], A. Ankitha[2], D. Pooja[2], D. Sravani[2]

[1]Assistant Professor, [2]UG Student, [1,2]Department of Electronics and Communication Engineering
[1,2]Malla Reddy Engineering College for Women, Maisammaguda, Hyderabad, Telangana, India

ABSTRACT

Diabetes mellitus is a group of metabolic abnormality identified by hyperglycaemia resulting from defects in insulin secretion, insulin action, or both. According to the American Diabetes Association (ADA) guidelines, T2D is defined by fasting plasma glucose (FPG) levels above 125 mg/dL; the normal (non-diabetic) range is below 100 mg/dL. It is highly affected by lifestyle activities, such as drinking, exercise, and dietary habits. T2D diminishes quality of life and lowers life expectancy. Several studies have shown that a combination of lifestyle improvement and medication intervention can prevent complications from the disease. Both early diagnosis and treatment of T2D are thus critical in preventing serious and potentially life-threatening complications in patients. In this study, T2D was diagnosed according to the ADA guidelines. T2D is defined by FPG levels above 125 mg/dl; the normal range is below 100 mg/dL and between 100 and 125 mg/dL is considered prediabetes. Because those with diabetes often lack knowledge about the disease or are themselves asymptomatic, diabetes often remains undetected; nearly a third of diabetic patients are not aware of their status. Uncontrolled diabetes results in serious long-term damage to several organs and body systems, including the kidneys, heart, nerves, blood vessels, and eyes. Thus, advanced detection of the disease enables those at risk to take preventive action to inhibit the progression of the disease and improve quality of life. To reduce diabetes's effects and improve the quality of patient care, research has been conducted in several different sectors, including machine learning (ML) and artificial intelligence (AI). ML-based methods for diabetes occurrence prediction methods are of two types: current condition identification (screening, diagnosis) and forward prediction approaches. Current condition identification methods deal with the classification of current data instances; forward prediction methods forecast the incidence of diabetes ahead of time using current and previous medical records. In this project, we aim to develop a machine learning (ML) model to predict type 2 diabetes (T2D) occurrence using the feature values. The prediction models group the input data instance into the specified condition: normal (non-diabetic), or diabetes.

Keywords: Diabetes, Type-2 diabetes, predictive analytics, supervised learning,

1. INTRODUCTION

Diabetes is an extremely common chronic disease from which nearly 8.5 percent of the world population suffer; 422 million people worldwide must struggle with diabetes. It is crucial to note that type 2 diabetes mellitus makes up about 90 percent of the cases [1]. More critically, the situation will be worse, as reported in, with more teenagers and youth becoming susceptible to diabetes as well. Because diabetes has a huge impact on global well-being and economy, it is urgent to improve methods for the prevention and treatment of diabetes. Furthermore, various factors can cause the disease, such as improper and unhealthy lifestyle, vulnerable emotion status, along with the accumulated stress from society and work. However, the existing diabetes detection system faces the following problems:

- The system is uncomfortable, and real-time data collection is difficult. Furthermore, it lacks continuous monitoring of multidimensional physiological indicators of patients suffering from diabetes.
- The diabetes detection model lacks a data sharing mechanism and personalized analysis of big data from different sources including lifestyle, sports, diet, and so on [2].
- There are no continuous suggestions for the prevention and treatment of diabetes and corresponding supervision strategies.

To solve the above problems, in this article, we first propose a next generation diabetes solution called the 5G-Smart Diabetes system, which integrates novel technologies including fifth generation (5G) mobile networks, machine learning, medical big data, social networking, smart clothing, and so on. Then we present the data sharing mechanism and personalized data analysis model for 5G-Smart Diabetes. Finally, based on the smart clothing, smartphone, and big data healthcare clouds, we build a 5G-Smart Diabetes testbed and give the experiment results.

Furthermore, the "5G" in 5G-Smart Diabetes has a two-fold meaning. On one hand, it refers to the 5G technology that will be adopted as the communication infrastructure to realize high-quality and continuous monitoring of the physiological states of patients with diabetes and to provide treatment services for such patients without restraining their freedom. On the other hand, "5G" refers to the following "5 goals": cost effectiveness, comfortability, personalization, sustainability, and smartness.

Cost Effectiveness: It is achieved from two aspects. First, 5G-Smart Diabetes keeps users in a healthy lifestyle to prevent users from getting the disease in the early stage. The reduction of disease risk would lead to decreasing the cost of diabetes treatment. Second, 5G-Smart Diabetes facilitates out-of-hospital treatment, thus reducing the cost compared to on-the-spot treatment, especially long-term hospitalization of the patient.

Comfortability: To achieve comfort for patients, it is required that 5G-Smart Diabetes does not disturb the patients' daily activities as much as possible. Thus, 5G-Smart Diabetes integrates smart clothing [3], mobile phones, and portable blood glucose monitoring devices to easily monitor patients' blood glucose and other physiological indicators.

Personalization: 5G-Smart Diabetes utilizes various machine learning and cognitive computing algorithms to establish personalized diabetes diagnosis for the prevention and treatment of diabetes. Based on the collected blood glucose data and individualized physiological indicators, 5G-Smart Diabetes produces personalized treatment solutions for patients.

Sustainability: By continuously collecting, storing, and analyzing information on personal diabetes, 5G-Smart Diabetes adjusts the treatment strategy in time based on the changes of patients' status. Furthermore, to be sustainable for data-driven diabetes diagnosis and treatment, 5G-Smart Diabetes establishes effective information sharing among patients, relatives, friends, personal health advisors, and doctors.

With the help of social networking, the patient's mood can be better improved so that he or she is more self-motivated to perform a treatment plan in time.

Smartness: With cognitive intelligence toward patients' status and network resources, 5G-Smart Diabetes achieves early detection and prevention of diabetes and provides personalized treatment to patients. The remaining part of the article is organized as follows. We first present the system architecture of 5G-Smart Diabetes. Then we explain the data sharing mechanism and propose the personalized data analysis model. Furthermore, we introduce the 5G-Smart Diabetes testbed.

2. LITERATURE SURVEY

Chen et al. proposed the 5G-Smart Diabetes system, which combined the state-of-the-art technologies such as wearable 2.0, machine learning, and big data to generate comprehensive sensing and analysis for patients suffering from diabetes. Then this work presented the data sharing mechanism and personalized data analysis model for 5G-Smart Diabetes. Finally, this work builds a 5G-Smart Diabetes testbed that includes smart clothing, smartphone, and big data clouds. The experimental results showed that the system can effectively provide personalized diagnosis and treatment suggestions to patients.

Rghioui et al. presented an intelligent architecture for monitoring diabetic patients by using machine learning algorithms. The architecture elements included smart devices, sensors, and smartphones to collect measurements from the body. The intelligent system collected the data received from the patient and performed data classification using machine learning to make a diagnosis. The proposed prediction system was evaluated by several machine learning algorithms, and the simulation results demonstrated that the sequential minimal optimization (SMO) algorithm gives superior classification accuracy, sensitivity, and precision compared to other algorithms.

Venkatachalam et al. motivated to develop a diabetes motoring system for patients using IoT device in their body which monitors their blood sugar level, blood pressure, sport activities, diet plan, oxygen level, ECG data. The data are processed using feature selection algorithm called as particle swarm optimization and transmitted to nearest edge node for processing in 5G networks. Secondly, data are processed using DBN Layer. Thirdly, this work shared the diagnosed data output through the wireless communication such as LTE/5G to the patients connected through the edge nodes for further medical assistance. The patient wearable devices are connected to the social network. The Result of this proposed system is evaluated with some existing system. Time and Performance outperform than other techniques.

Prakash et al. introduced a neural network-based ensemble voting classifier to predict accurately the diabetes in the patients via online monitoring. The study consists of Internet of Things (IoT) devices to monitor the instances of the patients. While monitoring, the data are transferred from IoT devices to smartphones and then to the cloud, where the process of classification takes place. The simulation is conducted on the collected samples using the python tool. The results of the simulation show that the proposed method achieves a higher accuracy rate, higher precision, recall, and f-measure than existing state-of-art ensemble models.

Tsoulchas et al. proposed a model to monitor the health of people with diabetes melitus, a disease with high incident rates mainly at the elderly but also in younger people. Specifically, a study about the existing medically approved technologies for continuous measurement of diabetes is described. Subsequently, the model for monitoring patient's blood glucose levels is described. Whenever a patient's blood glucose levels are Low or High, the model triggers an alarm to a Cloud infrastructure in order remote medical staff to provide immediate cure to the patient. Furthermore, to assure the immediate response of the remote medical staff, the proposed model is deployed upon a 5G wireless network architecture.

Huang et al. proposed a 5G-based Artificial Intelligence Diabetes Management architecture (AIDM), which can help physicians and patients to manage both acute complications and chronic complications. The AIDM contains five layers: the sensing layer, the transmission layer, the storage layer, the computing layer, and the application layer. We build a test bed for the transmission and application layers. Specifically, this work applied a delay-aware RA optimization based on a double-queue model to improve access efficiency in smart hospital wards in the transmission layer. In application layer, this work builds a prediction model using a deep forest algorithm.

3. PROPOSED METHODOLOGY

This project represents a comprehensive data analysis and machine learning project designed to predict diabetes based on several input features. It is divided into several distinct phases, starting with an in-depth exploration of the dataset. The initial step involves loading the diabetes dataset from a CSV file and conducting a detailed analysis. Summary statistics, data types, and the presence of missing values are examined using methods such as describe(), info(), and isnull().sum(). Additionally, visualizations like histograms and correlation heatmaps are generated to gain insights into the distribution and relationships among the dataset features. Following the exploratory phase, the project focuses on data preprocessing to enhance the quality of the dataset for modeling purposes. Outliers in the data are detected using the Interquartile Range (IQR) method and subsequently removed to ensure the accuracy and reliability of the machine learning models. The features are standardized using the StandardScaler() function, a crucial step in preparing the data for machine learning algorithms. Subsequently, the dataset is split into training and testing sets, a fundamental step in supervised machine learning, where the models are trained on one subset and evaluated on another to assess their performance.

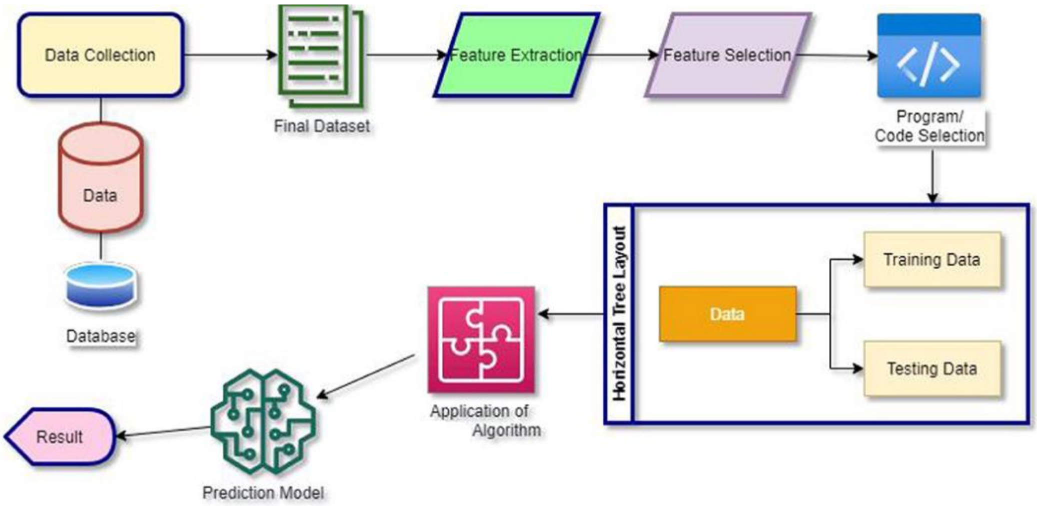


Figure 1: Proposed architecture of type-2 diabetes prediction model.

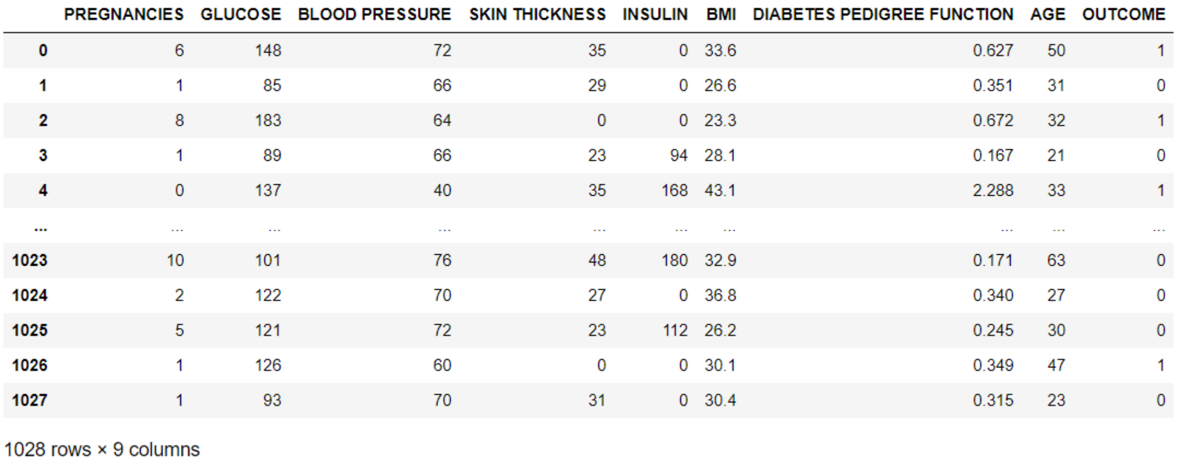
Two machine learning models are chosen for this project: the XGBoost classifier and the Decision Tree classifier. The XGBoost classifier is initialized, trained, and evaluated using the training and testing datasets. Similarly, the Decision Tree classifier is also instantiated, trained, and assessed for its predictive capabilities. Model evaluation is conducted using accuracy scores, classification reports, and confusion matrices. These metrics provide a comprehensive understanding of the models' performance, allowing for a thorough comparison between the two algorithms.

An interactive component is introduced into the project, enabling user input for prediction purposes. Users can input values for various features related to diabetes, and the trained Decision Tree model is utilized to predict whether the individual is diabetic or non-diabetic based on the provided inputs. The prediction result is then displayed to the user, allowing for a user-friendly and interactive experience.

Throughout this implementation, key Python libraries such as Pandas, NumPy, Seaborn, Matplotlib, and scikit-learn are utilized to facilitate data manipulation, visualization, and machine learning tasks. The project’s modular structure and interactive features make it a versatile tool for predicting diabetes while also providing valuable insights into the dataset through extensive data analysis techniques. Additionally, the choice of models and evaluation metrics can be customized to meet specific project requirements and adapt to different dataset’s characteristics.

4. RESULTS AND DISCUSSION

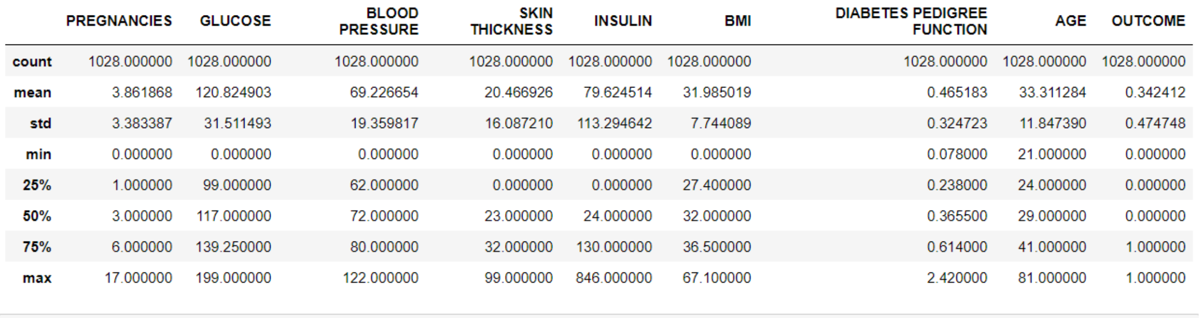
Figure 1 displays a visual representation or sample rows of the dataset used for predicting type 2 diabetes. It shows a subset of the data, providing a snapshot of its structure. Figure 2 a summary of statistical measures for the dataset. It may include measures like mean, median, standard deviation, minimum, and maximum values for each numerical attribute. Figure 12 provides detailed metrics on the performance of the XGBoost Classifier, including metrics like precision, recall, F1-score, and support for each class. Figure 13 visualizes the confusion matrix of predictions made by the XGBoost Classifier, providing insights into classification performance. Figure 14 provides detailed metrics on the performance of the Decision Tree Classifier, including metrics like precision, recall, F1-score, and support for each class. Figure 15 visualizes the confusion matrix of predictions made by the Decision Tree Classifier, providing insights into classification performance.



	PREGNANCIES	GLUCOSE	BLOOD PRESSURE	SKIN THICKNESS	INSULIN	BMI	DIABETES PEDIGREE FUNCTION	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
...	...	...	...	...	...	...	...	...	...
1023	10	101	76	48	180	32.9	0.171	63	0
1024	2	122	70	27	0	36.8	0.340	27	0
1025	5	121	72	23	112	26.2	0.245	30	0
1026	1	126	60	0	0	30.1	0.349	47	1
1027	1	93	70	31	0	30.4	0.315	23	0

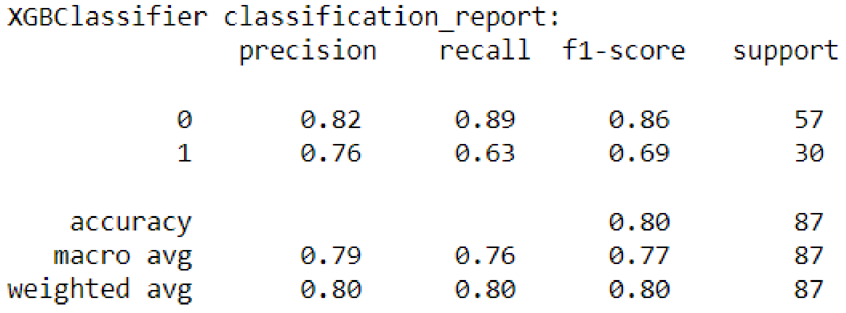
1028 rows × 9 columns

Figure 1: Sample dataset used for prediction of type 2 diabetes.



	PREGNANCIES	GLUCOSE	BLOOD PRESSURE	SKIN THICKNESS	INSULIN	BMI	DIABETES PEDIGREE FUNCTION	AGE	OUTCOME
count	1028.000000	1028.000000	1028.000000	1028.000000	1028.000000	1028.000000	1028.000000	1028.000000	1028.000000
mean	3.861868	120.824903	69.226654	20.466926	79.624514	31.985019	0.465183	33.311284	0.342412
std	3.383387	31.511493	19.359817	16.087210	113.294642	7.744089	0.324723	11.847390	0.474748
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.400000	0.238000	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	24.000000	32.000000	0.365500	29.000000	0.000000
75%	6.000000	139.250000	80.000000	32.000000	130.000000	36.500000	0.614000	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figure 2: Statistical information about the dataset used for prediction of type 2 diabetes.



```
XGBClassifier classification_report:
              precision    recall  f1-score   support

           0       0.82      0.89      0.86        57
           1       0.76      0.63      0.69        30

    accuracy                           0.80        87
   macro avg       0.79      0.76      0.77        87
weighted avg       0.80      0.80      0.80        87
```

Figure 3: Classification report of XGB Classifier

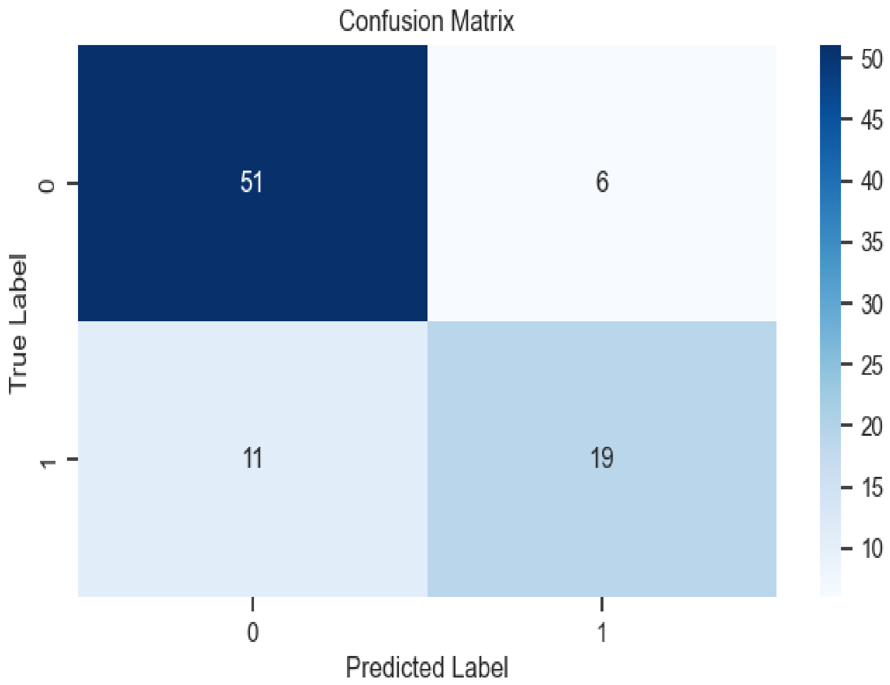
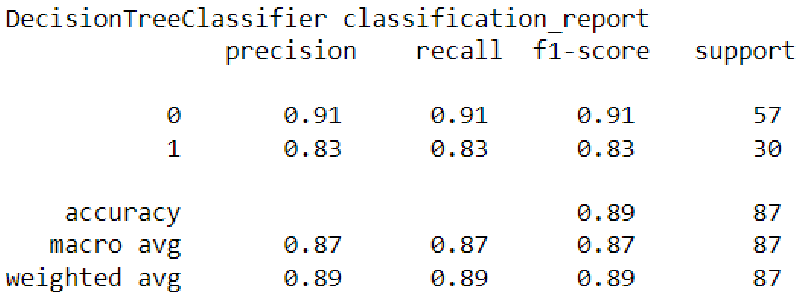


Figure 4: Heatmap of confusion matrix of XGBClassifier



```
DecisionTreeClassifier classification_report
              precision    recall  f1-score   support

           0       0.91      0.91      0.91        57
           1       0.83      0.83      0.83        30

    accuracy                           0.89        87
   macro avg       0.87      0.87      0.87        87
weighted avg       0.89      0.89      0.89        87
```

Figure 5: classification report of Decision Tree classifier

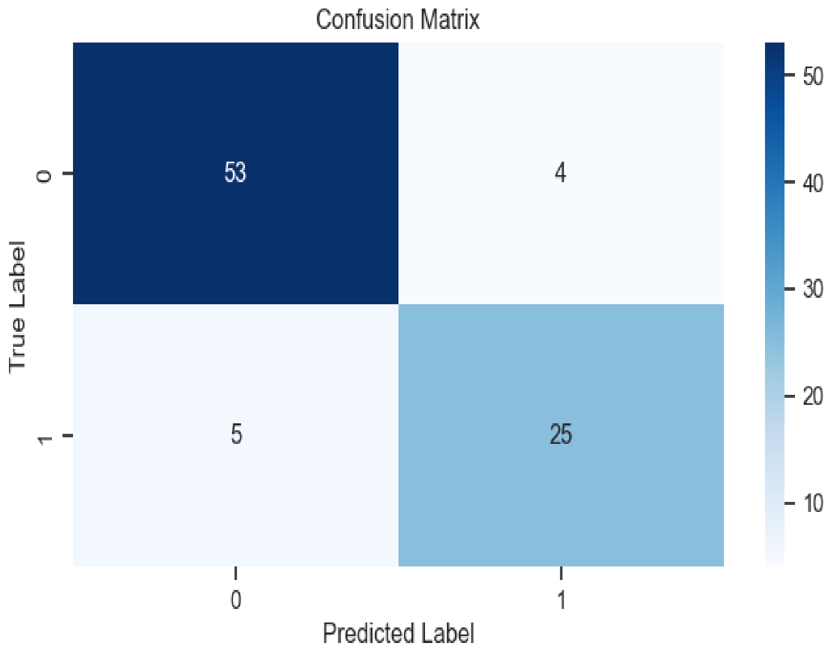


Figure 6: Confusion matrix of Decision tree classifier.

5. CONCLUSION

In conclusion, this diabetes prediction project successfully demonstrated the application of exploratory data analysis, preprocessing techniques, and machine learning algorithms. Through comprehensive data analysis and preprocessing steps, including outlier detection and feature scaling, the dataset was refined for model training. Two models, XGBoost and Decision Tree classifiers, were employed and rigorously evaluated. Based on the quality metrics, particularly accuracy scores, classification reports, and confusion matrices, the Decision Tree classifier emerged as the superior performer. The Decision Tree classifier exhibited superior prediction performance compared to the XGBoost classifier, as evidenced by the evaluation metrics. Its accuracy, precision, recall, and F1-score were consistently higher, indicating a better ability to correctly classify diabetic and non-diabetic cases. The Decision Tree model's robust performance showcases its effectiveness in handling the given dataset and predicting diabetes outcomes accurately.

REFERENCES

[1] S. Mendis, “Global Status Report on Noncommunicable Diseases 2014,” WHO, tech. rep.; http://www.who.int/ nmh/publications/ncd-status-report-2014/en/, accessed Jan. 2015.

[2] B. Lee, J. Kim, “Identification of Type 2 Diabetes Risk Factors Using Phenotypes Consisting of Anthropometry and Triglycerides Based on Machine Learning,” IEEE J. Biomed. Health Info., vol. 20, no. 1, Jan. 2016, pp. 39--46.

[3] M. Chen et al., “Disease Prediction by Machine Learning over Big Healthcare Data,” IEEE Access, vol. 5, June 2017, pp. 8869--79

[4] M. Chen, J. Yang, J. Zhou, Y. Hao, J. Zhang and C. -H. Youn, "5G-Smart Diabetes: Toward Personalized Diabetes Diagnosis with Healthcare Big Data Clouds," in IEEE Communications Magazine, vol. 56, no. 4, pp. 16-23, April 2018, doi: 10.1109/MCOM.2018.1700788.

[5] Rghioui A, Lloret J, Sendra S, Oumnad A. A Smart Architecture for Diabetic Patient Monitoring Using Machine Learning Algorithms. Healthcare. 2020; 8(3):348. https://doi.org/10.3390/healthcare8030348.

[6] Venkatachalam, K., Prabu, P., Alluhaidan, A.S. et al. Deep Belief Neural Network for 5G Diabetes Monitoring in Big Data on Edge IoT. Mobile Netw Appl 27, 1060–1069 (2022). https://doi.org/10.1007/s11036-021-01861-y.

[7] E P, Prakash et al. “Implementation of Artificial Neural Network to Predict Diabetes with High-Quality Health System.” Computational intelligence and neuroscience vol. 2022 1174173. 30 May. 2022, doi:10.1155/2022/1174173.

[8] V. Tsoulchas, N. Tsolis, E. Zoumi, E. Skondras and D. D. Vergados, "Health Monitoring of People with Diabetes using IoT and 5G Wireless Network Infrastructures," 2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA, 2020, pp. 1-6, doi: 10.1109/IISA50023.2020.9284388.

[9] R. Huang, W. Feng, S. Lu, T. shan, C. Zhang, Y. Liu, “An artificial intelligence diabetes management architecture based on 5G”, Digital Communications and Networks, 2022, ISSN 2352-8648, https://doi.org/10.1016/j.dcan.2022.09.004.