DIABETES PREDICTION USING MACHINE LEARNING

Dr. Manasa P, Dr. Hithaishi P
Assistant Professor, Department of ECE,
Bangalore Institute of Technology, Bangalore.

Abstract: *Diabetes mellitus represents a global health challenge, its prevalence steadily rising due to contemporary lifestyles, unhealthy dietary habits, and rising obesity rates. To address this, numerous studies have explored the utility of predictive models utilizing both physical and chemical tests for diagnosing diabetes. Leveraging data science methodologies offers a promising approach to enhance understanding and prediction in this domain. In the proposed system, an efficient framework for diabetes detection employing machine learning and deep learning techniques is presented. Within the realm of machine learning, XGBoost and Random Forest algorithms are utilized. The experimental findings reveal the effectiveness of our approach, demonstrating high accuracy in predicting diabetes. By leveraging advanced computational techniques, presented work aims to contribute to the ongoing efforts in diabetes diagnosis and management, thereby potentially improving healthcare outcomes for affected individuals worldwide.*

Keywords: Diabetes, XGBoost, Random Forest algorithms, LightGBM, Decision Tree Classifier, AdaBoost.

1. INTRODUCTION

1.1 INTRODUCTION

Diabetes is a serious and chronic condition that currently has no cure. Once diagnosed, it can persist throughout a person's life. Elevated blood glucose levels associated with diabetes can lead to various health complications, including kidney disease, heart disease, stroke, eye issues, dental problems, foot issues, and nerve damage. Monitoring and managing diabetes effectively can help prevent these complications. As living standards improve, diabetes is becoming increasingly prevalent. Therefore, accurate and timely diagnosis is crucial. In medical practice, diabetes is diagnosed through measurements of fasting blood glucose, glucose tolerance and random blood glucose levels. Early diagnosis significantly improves management and control of the disease. Prediabetes is characterized by elevated blood glucose levels that are higher than normal but not yet high enough to be classified as diabetes. In type 1 diabetes, the body's insulin-producing cells in the pancreas are attacked, leading to the destruction of over 90% of these cells. This form of diabetes affects about 5 to 10 percent of people with diabetes. Diabetes can cause long-term damage and dysfunction in various organs, including the skin, liver, ears, and blood vessels [1-2].

Type 2 diabetes (T2D) is more common in middle-aged and older adults and is often associated with obesity, high blood pressure, dyslipidemia, and atherosclerosis. Unlike type 1 diabetes, which usually manifests in younger individuals, type 2 diabetes develops over time and is often linked to lifestyle factors. Symptoms may include increased appetite, frequent urination, and elevated blood pressure. While some cases of type 2 diabetes can be managed with oral medications, others may require insulin therapy. Preventive measures include maintaining a healthy weight, regular exercise, and a balanced diet. If blood sugar levels are not sufficiently controlled through lifestyle changes, metformin is commonly prescribed, and insulin injections may be necessary. Regular blood sugar monitoring is recommended for those on insulin, though it may not be required for those on oral medications. Bariatric surgery can also be beneficial for obese individuals with diabetes.

1.2 MOTIVATION

Machine learning techniques have been around us and have been compared and used for analysis for many kinds of data science applications. This work is carried out with the motivation to develop an appropriate computer-based system and decision support that can aid in the early detection of diabetes, in this work we have developed a model which classifies if the patient will have diabetes based on various features (i.e. potential risk factors that can cause diabetes) using random forest classifier. Hence, the early prognosis of diabetes can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine [3].

1.3 **PROBLEM STATEMENT**

The major challenge in predicting diabetes cases is its detection. There are instruments available that can predict diabetes but either they are expensive or are not efficient to calculate the chance of diabetes in humans. Early detection of diabetes can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more patience, time, and expertise. Since we have a good amount of data in today's world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data

2. LITERATURE SURVEY

Prabha [4] et al., discusses a new system for detecting diabetes using wrist PPG signals and physiological data. It highlights the rising global prevalence of diabetes and the need for accurate early detection methods. By employing machine learning algorithms like KNN and SVM, the system achieves promising classification accuracy. The study showcases the potential of non-invasive techniques for diabetes screening, offering a valuable contribution to healthcare technology. Sivaranjani [5] et al., Diabetes is a prevalent and life-threatening disease worldwide, affecting various age groups and attributed to factors like lifestyle, genetics, stress, and age. The presented work aims to enhance the accuracy of predicting diabetes-related diseases by employing machine learning algorithms, namely SVM and RF. The study utilizes the PIMA Indian diabetes dataset and employs techniques such as data pre-processing, feature selection, and dimensionality reduction to improve prediction outcomes.

Mahedy [6] et al., discusses the prevalence and challenges associated with diabetes, emphasizing the importance of early diagnosis. It highlights the chronic nature of diabetes and its significant impact on individuals' daily lives, requiring lifestyle adjustments. The World Health Organization's data indicates a substantial increase in the number of diabetes cases globally, making it a major health concern. The text emphasizes the need for accurate prediction models using machine learning techniques to classify diabetic patients early on..

Lyngdoh [7] et al., explores the prediction of diabetes disease using five supervised machine learning algorithms: K-Nearest Neighbors, Naïve Bayes, Decision Tree Classifier, Random Forest, and Support Vector Machine (SVM). By analyzing various risk factors and employing cross-validation, the study achieves stable accuracy rates, with KNN reaching the highest at 76%. The research aims to optimize accuracy and computational time for diabetes prediction, offering insights into potential advancements in disease detection using machine learning techniques.

Islam [8] et al., explores the rising issue of Diabetes Mellitus (DM), emphasizing the need for early detection to mitigate its impact. It employs machine learning (ML) algorithms—AdaBoost, Bagging, and Random Forest—to predict DM accurately. Using real-time data from 464 instances with 22 risk factors, the study achieves high accuracies: 97.84% for AdaBoost, 98.28% for Bagging, and 99.35% for Random Forest. This research underscores ML's potential in improving DM prediction and advocates for further exploration with larger datasets and alternative algorithms

Alanazi [9] et al., presents a model using machine learning algorithms, namely Support Vector Machine (SVM) and Random Forest (RF), to predict diabetes. Achieving 98% accuracy and 99% ROC, the model utilizes real data from a primary health care center. Results favor RF over SVM. The study underscores the potential of computational intelligence in expediting diabetes diagnosis, offering valuable insights for healthcare technology.

The reviewed papers on diabetes prediction using machine learning (ML) techniques collectively aim to improve the accuracy and efficiency of disease detection. These studies leverage various ML algorithms, including Support Vector Machine (SVM), Random Forest (RF), XGBoost, LightGBM, Decision Tree Classifier, and ensemble methods like AdaBoost and Bagging. By utilizing different datasets such as the PIMA Indian diabetes dataset and real time data from primary healthcare centers, researchers strive to develop robust prediction models capable of accurately identifying diabetes-related diseases. A notable focus across the papers is the adoption of methodological approaches to enhance prediction outcomes. Techniques such as data pre-processing, feature selection, and dimensionality reduction are employed to optimize computational efficiency and improve the overall performance of ML models. Moreover, the exploration of non-invasive detection methods, such as wrist photo plethysmography (PPG) signals and physiological data, underscores the potential for early screening and diagnosis of diabetes, offering valuable contributions to healthcare technology.

3.MACHINE LEARNING AND ALGORITHMS

Machine learning, a branch of artificial intelligence, focuses on developing algorithms that enable computers to learn from data and past experiences autonomously. By utilizing historical data, known as training data, machine learning algorithms create mathematical models that aid in making predictions or decisions without direct programming. This field merges computer science with statistics to build predictive models, with performance improving as more data is provided. The basic operation of a machine learning model involves learning from historical data to build prediction models. When new data is introduced, the model uses these predictions to provide output. The accuracy of these predictions typically increases with the volume of data, as larger datasets enable the creation of more precise models. For complex problems requiring predictions, rather than manually coding solutions, data is fed into generic algorithms, which then develop logic and make predictions based on the input data. Machine learning has transformed our approach to problem-solving by automating the learning process [10].

The Growing Need for Machine Learning: The demand for machine learning continues to rise due to its ability to handle tasks that are too complex for direct human implementation. Humans face limitations in manually processing vast amounts of data, making machine learning essential for automating data analysis, model construction, and prediction. The efficiency of machine learning algorithms is influenced by the amount of data available and is assessed using cost functions. Machine learning not only saves time and resources but also has practical applications across various industries. For instance, it is used

in self-driving cars, cyber fraud detection, facial recognition, and social media features like Facebook's friend suggestions. Leading companies such as Netflix and Amazon leverage machine learning to analyze user preferences and provide personalized recommendations. Machine learning can be broadly categorized into three types:Supervised Learning, Unsupervised Learning and Reinforcement Learning. Supervised Learning is a machine learning approach where models are trained using labeled data, meaning that each input is paired with the correct output. The labeled data serves as a guide or "supervisor" to help the machine learn how to predict outcomes accurately. This method is akin to a student learning under the guidance of a teacher. In supervised learning, the goal is to develop a function that maps input variables (x) to output variables (y) based on the training data provided. Applications of supervised learning in real-world scenarios include risk assessment, image classification, fraud detection, and spam filtering.

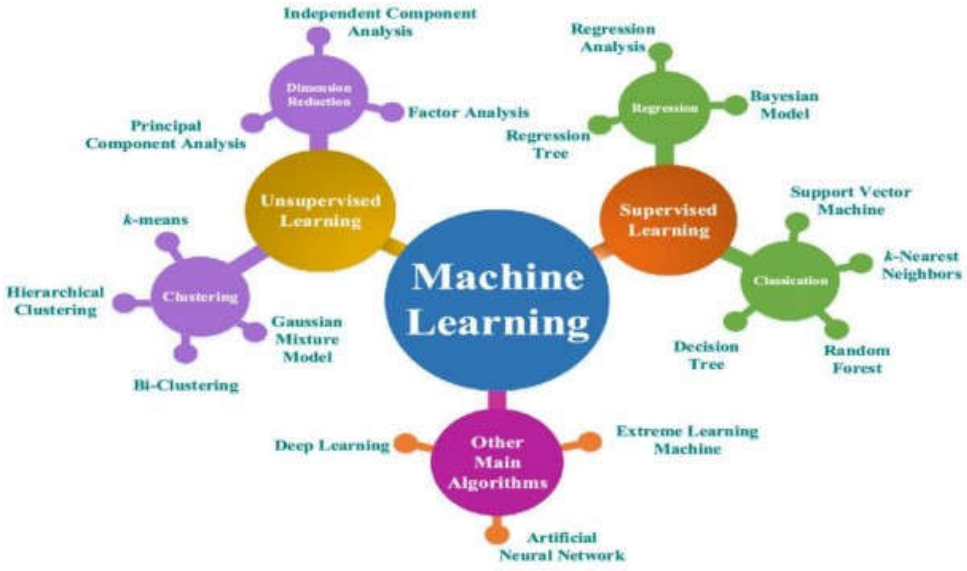


Figure 1: Types of Machine Learning

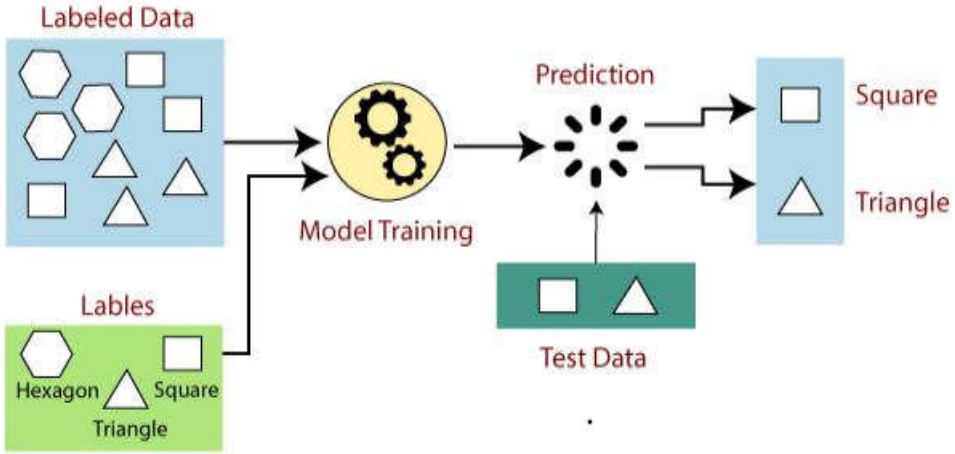
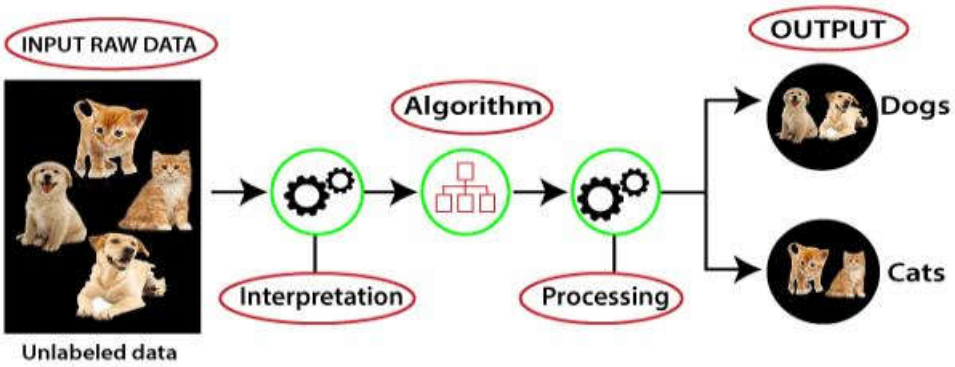


Figure 2: Supervised Learning



iii)

Figure 3: Unsupervised Learning

Reinforcement Learning: Reinforcement Learning is a feedback-based Machine learning technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions. For each good action, the agent gets positive feedback, and for each bad action, the agent gets negative feedback or penalty. In Reinforcement Learning, the agent learns automatically using feedbacks without any labelled data, unlike supervised learning. Since there is no labelled data, so the agent is bound to learn by its experience only. RL solves a specific type of problem where decision making is sequential, and the goal is long-term, such as game-playing, robotics, etc. The agent interacts with the environment and explores it by itself. The primary goal of an agent in reinforcement learning is to improve the performance by getting the maximum positive rewards.

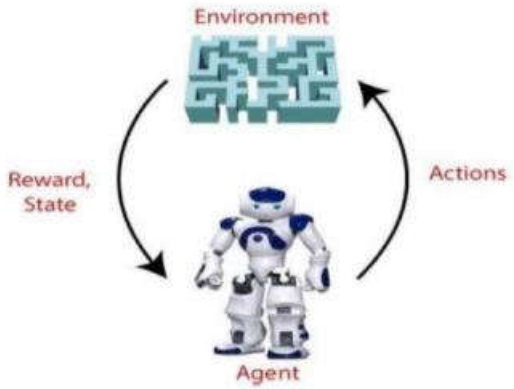


Figure 4: Unsupervised Learning

4. IMPLEMENTATION OF PROPOSED METHODOLOGY

Figure 5 shows the block diagram of diabetes prediction methodology. The following are the steps involved in implementing the above proposed methodology: Step 1: Data Collection: Obtain a dataset containing relevant information for diabetes prediction. Step 2: Data Preprocessing: Clean and organize the data for analysis. Step 3: Feature Engineering: Extract meaningful features from the data. Step 4: One Hot Encoding: Convert categorical variables into a format suitable for machine learning algorithms. Step 5: Model Development: Train machine learning models on the preprocessed data. Step 6: Model Tuning: Optimize model performance by adjusting the model parameters. Step 7: Testing and Evaluation: Test and evaluate the models to assess their predictive accuracy

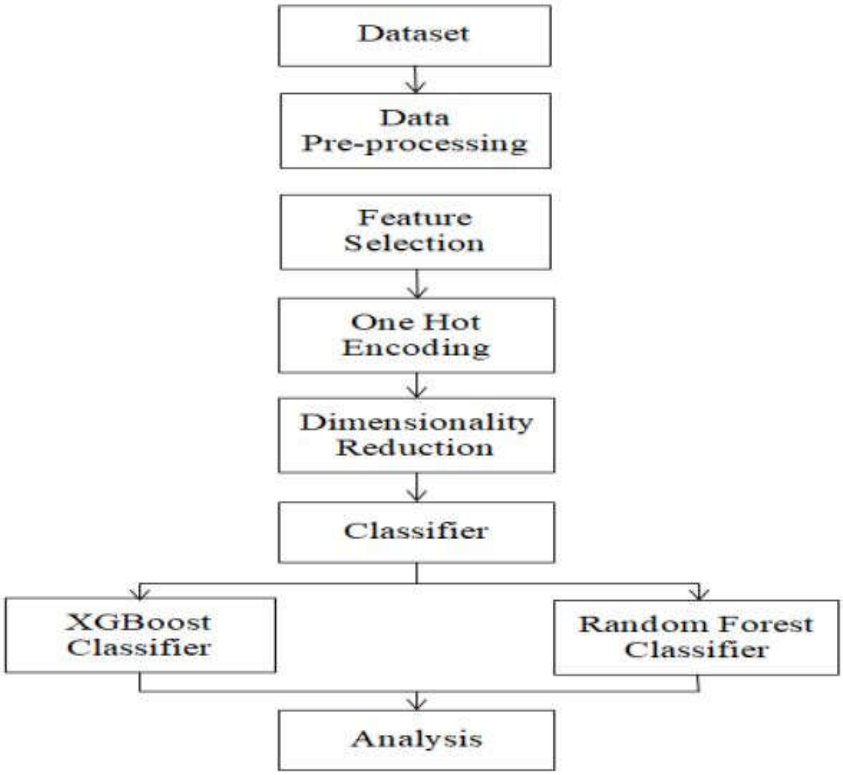


Figure 5 : Block Diagram of Diabetes Prediction Methodology

Step 1: Data Collection Initially, the journey commences with the crucial task of data collection. We rely on the esteemed Pima Indian dataset, which is renowned for its rich repository of patient data encompassing various demographic, clinical, and lifestyle attributes. This dataset serves as the cornerstone of our analysis, providing the necessary foundation upon which our predictive models are built. Step 2: Data Preprocessing Subsequently, the collected data undergoes meticulous pre-processing to ensure its integrity and reliability. This involves a thorough examination of the dataset to address any anomalies, such as missing values, outliers, or inconsistencies. By cleansing and organizing the data, we endeavor to create a robust and standardized dataset conducive to accurate model training.

The Local Outlier Factor (LOF) method is a part of the preprocessing pipeline which is employed to identify and handle outliers within the dataset and could potentially skew the results and compromise the integrity of our predictive models. Outliers are either removed from the dataset or subjected to further scrutiny, depending on their impact on the analysis and predictive modeling. By mitigating the influence of outliers, the LOF method helps to improve the robustness and reliability of our predictive models, ensuring that they are trained on highquality data free from undue influence from aberrant observations Feature Engineering Feature engineering emerges as a pivotal step in enhancing the predictive capabilities of our model. Here, we extract meaningful features from the dataset, including glucose levels, BMI, age, family history, and other relevant parameters. These features provide valuable insights into the factors influencing diabetes diagnosis

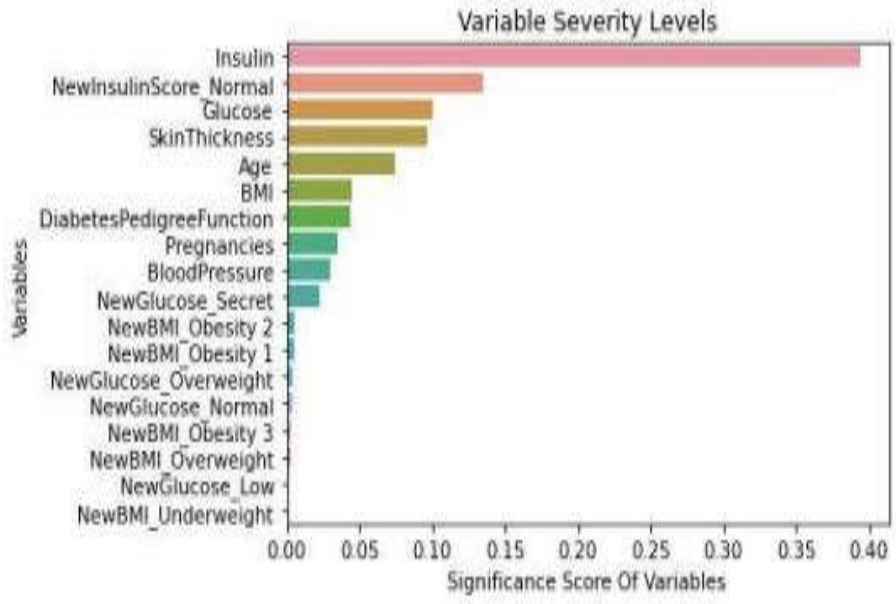


Figure 6: Random Forest Feature Importance

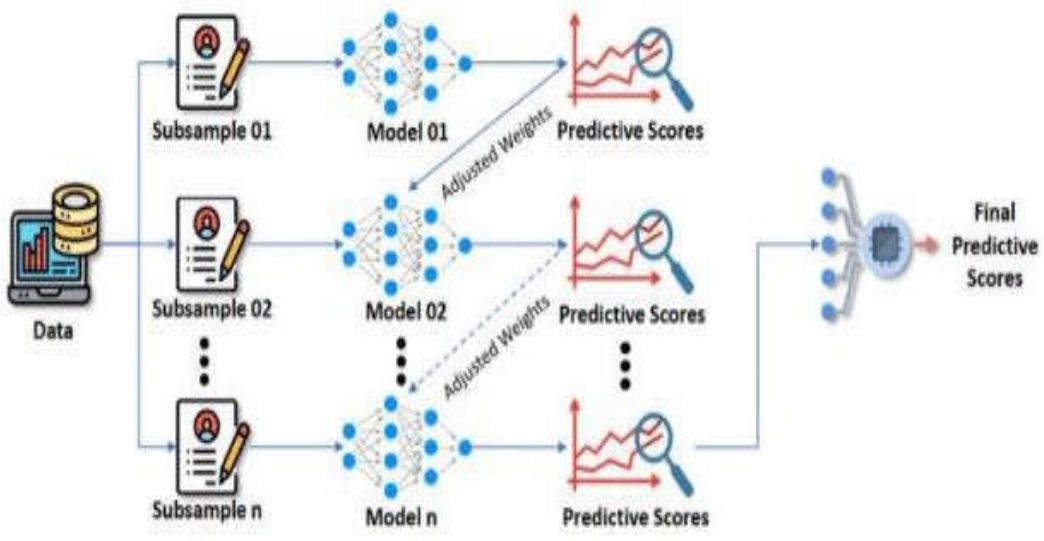


Figure 7: XGBoost Classifier

Step 4: One Hot Encoding One-hot encoding is employed to handle categorical variables within the dataset, converting them into a numerical format compatible with machine learning algorithms. This process ensures that categorical attributes, such as patient ethnicity or medical history, can be effectively incorporated into our predictive models without introducing bias or ambiguity. For example, if there's a feature like "BMI" with categories "Normal", "Underweight" and "Overweight", it would be converted into binary features: "BMI_Normal" ,"BMI_Underweight" and "BMI_Overweight".

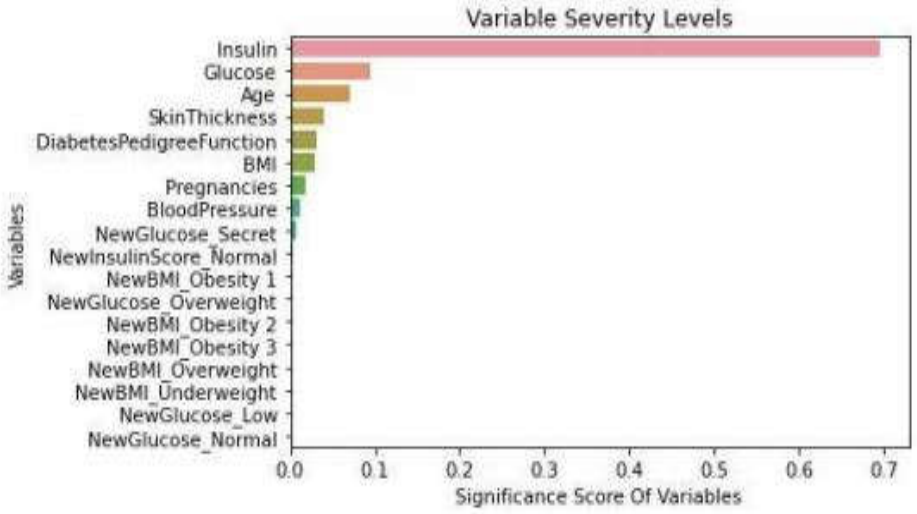


Figure 8: XGBoost Feature Importance

Step 5: Model Development The XGBoost algorithm is then employed for training the diabetes prediction model. XGBoost is chosen for its effectiveness in handling complex datasets, feature interactions, and imbalanced classes. During the training phase, Hyperparameter tuning is performed to optimize model performance and generalization. Cross-validation techniques are used to assess the model's robustness and prevent overfitting. Step 6: Model Tuning Once the model is trained, it is evaluated using evaluation metrics such as accuracy, precision, recall, and F1-score. The model's performance is assessed on both training and validation datasets to ensure it generalizes well to unseen data. Model interpretability techniques may be applied to gain insights into the factors influencing diabetes risk predictions. Model tuning is a critical phase where we optimize the performance of our machine learning models. By adjusting model Hyperparameters and fine-tuning the algorithms, we strive to achieve the best possible predictive accuracy and generalization.

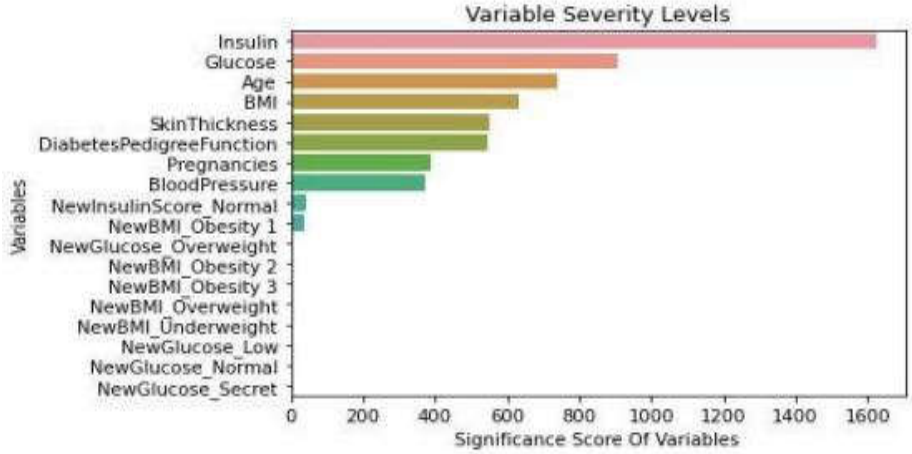


Figure 9: LightGBM Feature Importance

Step 7: Testing and Evaluation Finally, the trained XGBoost model is deployed in a production environment, where it can be used to predict diabetes risk for new individuals. Integration with existing healthcare systems or mobile applications allows for seamless access to the predictive model, enabling early intervention and personalized healthcare recommendations. Ongoing monitoring and model maintenance ensure that the predictive model remains accurate and up-to-date with evolving healthcare trends and patient data.

5. RESULTS AND OBSERVATION

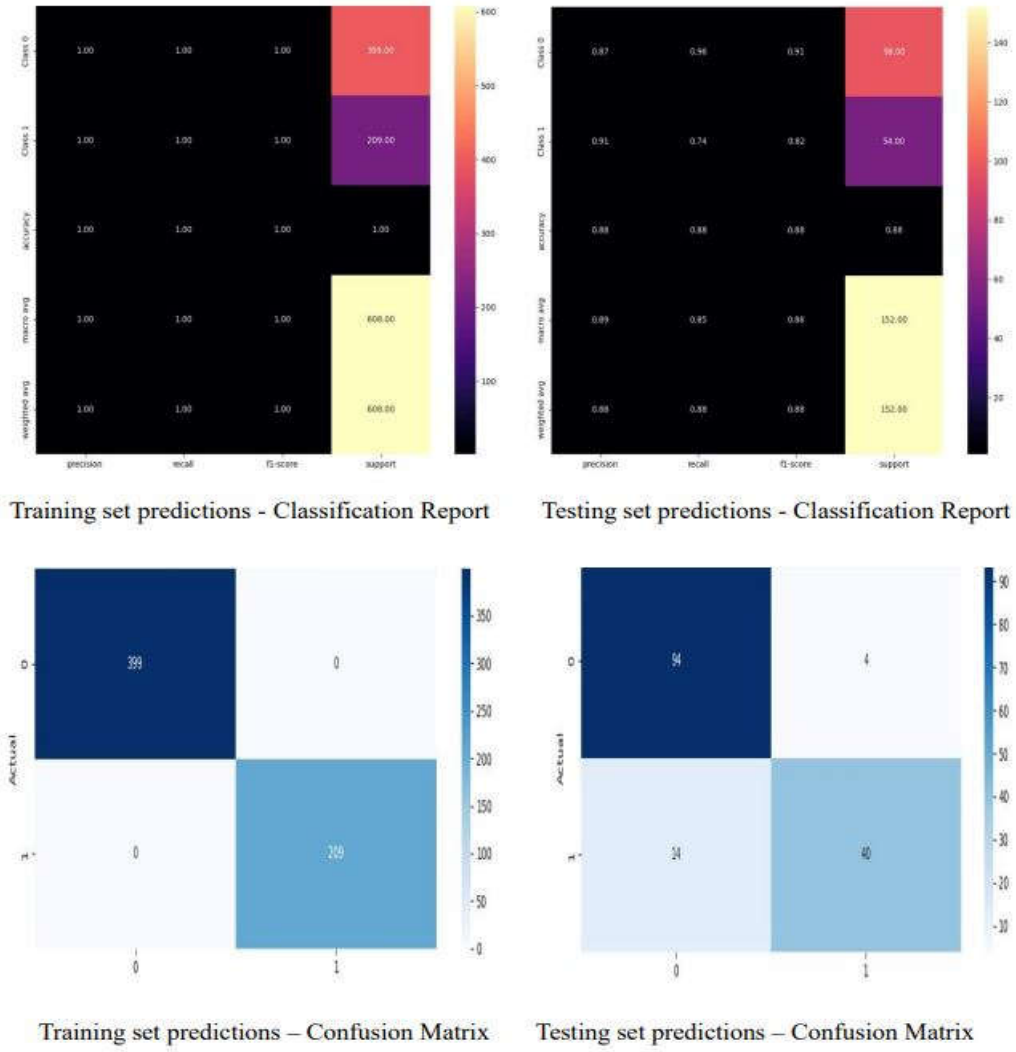


Figure 10: Classification Report and Confusion Matrix of XGBoost Model on training and testing sets.

Figure 10 shows the classification Report and Confusion Matrix of XGBoost Model on training and testing sets. On the training set, we can see that we have an excellent classification report and confusion matrix. In the case of the Testing set, 89 percent of the retrieved instances are relevant, implying that the precision is 89 percent when averaged across both positive and negative classes. 89 percent of all relevant instances are retrieved, implying that the recall is 89 percent when both positive and negative classes are averaged. We can see from the Confusion Matrix that there are 4 False Positive values, which means that our XGBoost classifier misclassified 4 values of Non-Diabetic patients instances as having Diabetes.With 14 false negative values , our XGBoost Classifier misclassified 14 values of Diabetic patients instances as not having Diabetes. Our goal is to reduce the False Negative rate to as low as possible, ideally close to zero, because we need to accurately classify a person with diabetes in order to assist patients in taking necessary precautions and limiting the further escalation and we might get a False Positive rate, which we accept as a cost of building the best model – a trade off

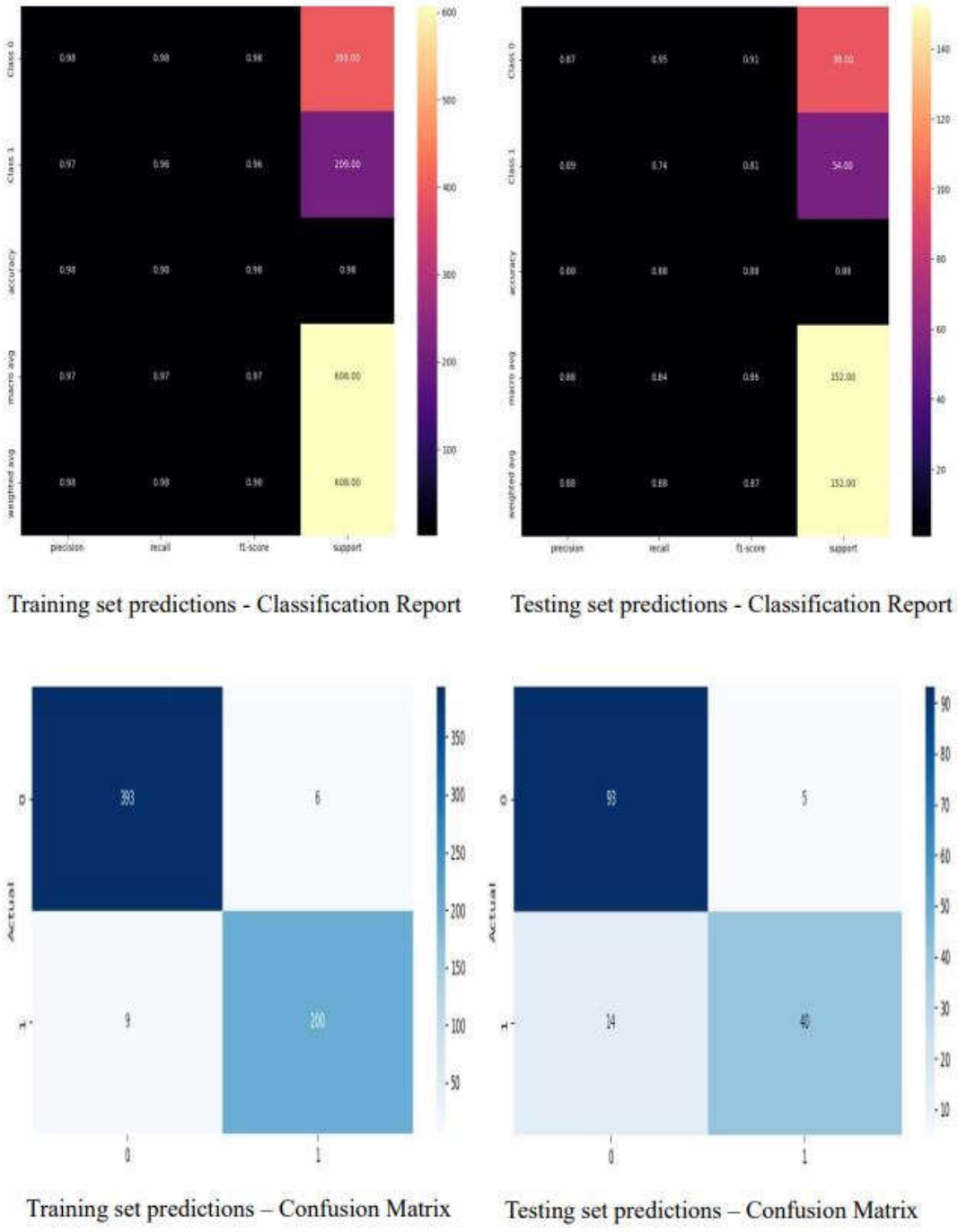


Figure 11: Classification Report and Confusion Matrix of LightGBM Model on training and testing sets.

Figure 11 shows the classification Report and Confusion Matrix of LightGBM Model on training and testing sets. On the training set, we can see that we have a good classification report and confusion matrix. In the case of the Testing set, 88 percent of the retrieved instances are relevant, implying that the precision is 88 percent when averaged across both positive and negative classes. 88 percent of all relevant instances are retrieved, implying that the recall is 88 percent when both positive and negative classes are averaged. We can see from the Confusion Matrix that there are 5 False Positive values, which means that our LightGBM classifier misclassified 5 values of Non-Diabetic patients instances as having Diabetes. And, with 14 false negative values , our LightGBM Classifier misclassified 14 values of Diabetic patients instances as not having Diabetes.

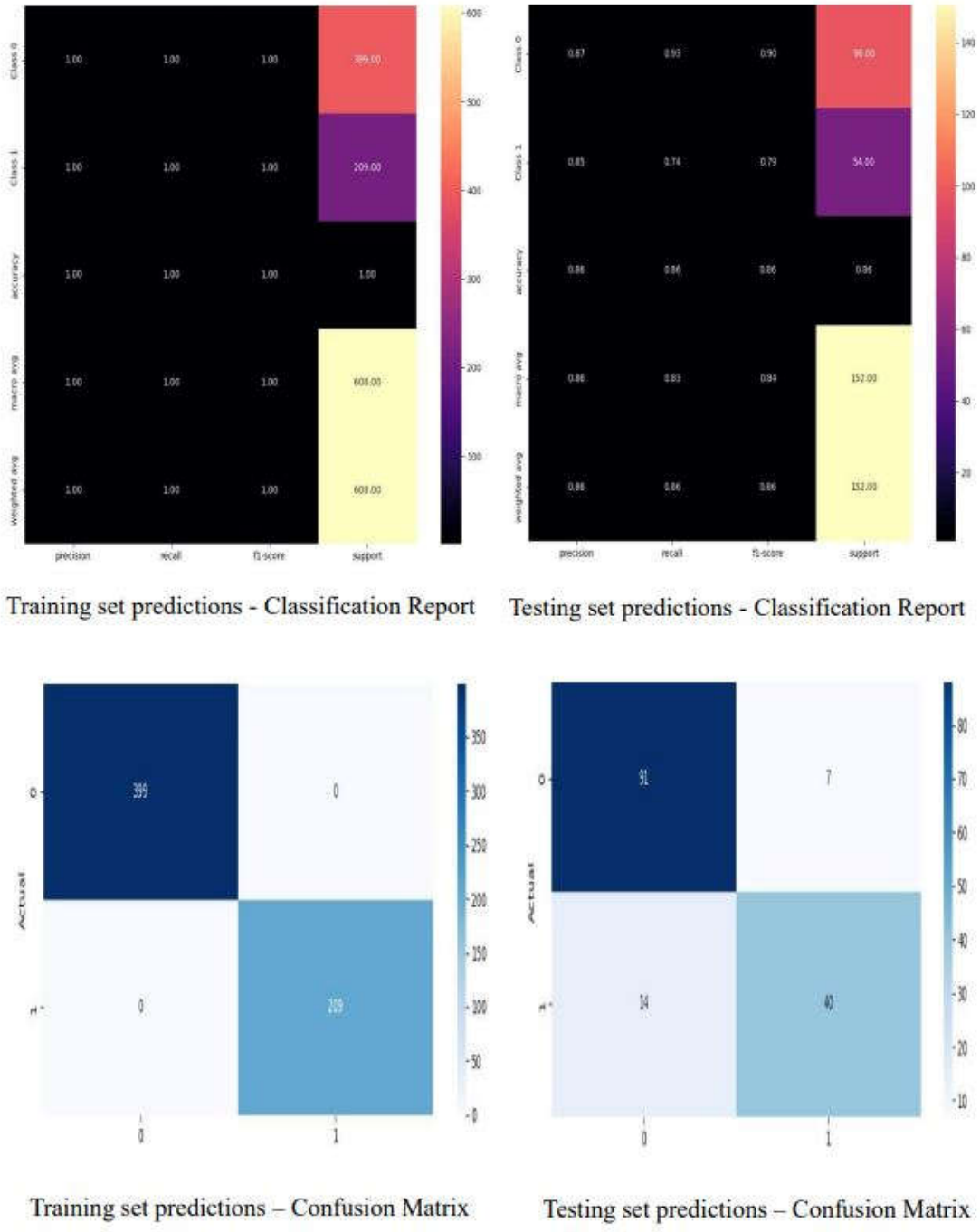


Figure 12: Classification Report and Confusion Matrix of Random Forest Model on training and testing sets.

Figure 12 shows the Classification Report and Confusion Matrix of Random Forest Model on training and testing sets. On the training set, we can see that we have a good classification report and confusion matrix. In the case of the Testing set, 86 percent of the retrieved instances are relevant, implying that the precision is 86 percent when averaged across both positive and negative classes. 86 percent of all relevant instances are retrieved, implying that the recall is 86 percent when both positive and negative classes are averaged. We can see from the Confusion Matrix that there are 7 False Positive values, which means that our Random Forest classifier misclassified 7 values of Non-Diabetic patients instances as having Diabetes. with 14 false negative values , our

Random Forest Classifier misclassified 14 values of Diabetic patients instances as not having Diabetes.

CONCLUSION

The application of machine learning in diabetes detection stands at the forefront of a technological revolution in healthcare where machine learning algorithms are used to analyze diverse and extensive datasets to offer a promising avenue for more accurate, timely, and personalized diagnosis of diabetes. Through the identification of intricate patterns and correlations, machine learning contributes to early risk assessment, enabling healthcare providers to implement proactive interventions and personalized treatment plans. This technological integration not only enhances the efficiency of diabetes detection but also holds the potential to alleviate the burden on healthcare systems by facilitating preventive measures so as we witness the ongoing advancements in machine learning techniques, it becomes increasingly clear that these innovations have the capacity to revolutionize the way we approach and manage diabetes.

REFERENCES

1. Nimmagadda, Satyanarayana Murthy, et al. "A Comprehensive Survey on Diabetes Type-2 (T2D) Forecast Using Machine Learning." Archives of Computational Methods in Engineering (2024): 1-19.
2. Hasan, Md Kamrul, et al. "Diabetes prediction using ensembling of different machine learning classifiers." IEEE Access 8 (2020): 76516-76531.
3. Sonar, Priyanka, and K. JayaMalini. "Diabetes prediction using different machine learning approaches." 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2019.
4. Prabha, Anju, et al. "Non-invasive diabetes mellitus detection system using machine learning techniques." 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2021.
5. Sivaranjani, S., et al. "Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction." 2021 7th international conference on advanced computing and communication systems (ICACCS). Vol. 1. IEEE, 2021.
6. Hasan, SM Mahedy, et al. "A machine learning-based model for early stage detection of diabetes." 2020 23rd International Conference on Computer and Information Technology (ICCIT). IEEE, 2020.
7. Lyngdoh, Arwatki Chen, Nurul Amin Choudhury, and Soumen Moulik. "Diabetes disease prediction using machine learning algorithms." 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES). IEEE, 2021.
8. Islam, Md Tanvir, et al. "Diabetes mellitus prediction using different ensemble machine learning approaches." 2020 11th international conference on computing, communication and networking technologies (ICCCNT). IEEE, 2020.
9. Alanazi, Aeshah Saad, and Mohd A. Mezher. "Using machine learning algorithms for prediction of diabetes mellitus." 2020 international conference on computing and information technology (ICCIT-1441). IEEE, 2020.
10. Acheampong, Emmanuel, et al. "Predictive modelling of metabolic syndrome in Ghanaian diabetic patients: an ensemble machine learning approach." Journal of Diabetes & Metabolic Disorders (2024): 1-17.