e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:07/July-2024 Impact Factor- 7.868 www.irjmets.com

DETECTION OF DIABETES MELLITUS USING ARTIFICIAL INTELLIGENCE

Gudapati Sri Sri Ram Karthikeya*1, Kodi Nandan*2, Mullapudi Charan*3, Sirigireddy Bharath Chandra Reddy*4, Basaba Bikram*5

*1,2,3,4,5Department Of Computer Science Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India.

DOI : https://www.doi.org/10.56726/IRJMETS59890

ABSTRACT

Diabetes mellitus, a chronic metabolic disorder characterized by elevated blood glucose levels, poses a significant global health challenge. Predictive modeling using machine learning techniques has emerged as a promising approach for early diagnosis and intervention. This study investigates the performance of various machine learning models in predicting diabetes based on comprehensive datasets. The dataset comprises a diverse set of features, including demographic information, clinical markers, and lifestyle factors. We employ logistic regression, decision trees, random forests, support vector machines, k-nearest neighbors, neural networks, and gradient boosting models (XGBoost, LightGBM) to assess their efficacy in diabetes prediction. The dataset undergoes rigorous preprocessing, including handling missing values and feature scaling, to ensure the robustness of the models. Comparative analysis reveals the strengths and limitations of each model, considering factors such as interpretability, accuracy, and computational efficiency. The results highlight the significance of feature selection and model tuning in optimizing predictive performance. Additionally, we discuss the implications of these findings for early diabetes detection and personalized healthcare. Our study contributes to the evolving landscape of diabetes prediction methodologies, offering insights into the suitability of different machine learning models for diverse datasets and laying the groundwork for future advancements in this critical field of medical research.

Keywords: Random Forest Classifier, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Decision Tree Classifier, Ada Boost Classifier, Gradient Boosting Classifier, XG Boost Classifier, Light GBM Classifier.

I. INTRODUCTION

Diabetes is a prevalent metabolic disorder characterized by hyperglycemia, poses a substantial global health burden with profound implications for individuals and healthcare systems. Early detection and effective management are paramount in mitigating the complications associated with diabetes. The advent of machine learning (ML) techniques has ushered in a new era of predictive modeling, offering unprecedented opportunities for accurate and timely diabetes prediction. This study seeks to unravel the intricacies of diabetes prediction by conducting a comprehensive analysis of various ML models, examining their performance on a diverse dataset encompassing demographic, clinical, and lifestyle factors. The urgency of addressing diabetes stems from its escalating prevalence and the myriad complications it engenders, ranging from cardiovascular diseases to neuropathy. ML, with its capacity to discern intricate patterns within vast datasets, has become instrumental in the pursuit of proactive healthcare solutions. As such, this research navigates the landscape of ML models, evaluating their applicability and efficacy in predicting diabetes. We delve into established models like logistic regression and decision trees, exploring their interpretability and simplicity, as well as advanced techniques such as neural networks and gradient boosting models, which excel in capturing complex relationships within the data. The dataset under consideration is carefully curated to reflect the multifaceted nature of factors influencing diabetes, ensuring a holistic analysis. Preprocessing steps, including handling missing values and normalizing features, are rigorously implemented to enhance the robustness of the models. Through a systematic comparative analysis, this study sheds light on the relative strengths and limitations of each model, offering valuable insights into their performance metrics and computational efficiency. The implications of this research extend beyond the realm of predictive modeling. By discerning which ML models are best suited for diabetes prediction, we pave the way for optimized clinical decision-making and personalized healthcare interventions. This study not only contributes to the evolving field of diabetes research

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:07/July-2024 Impact Factor- 7.868 www.irjmets.com

but also provides a foundation for future endeavors aiming to harness the power of ML for predictive healthcare analytics.

II. LITERATURE WORK

Rani's comprehensive study [1] serves as a foundational exploration into the application of machine learning techniques for diabetes prediction. The emphasis on predictive modeling techniques underscores the potential of these technologies in healthcare. Varun Jaiswal and colleagues [2] extended the discourse through a meticulous review of current advances in machine learning-based diabetes prediction. Their work provides invaluable insights into the dynamic landscape of predictive modeling within the healthcare domain, identifying emerging trends and advancements. Isfafuzzaman Tasin et al. [3] and Sandip Kumar Singh Modak [4] brought a crucial dimension to the literature by focusing on the incorporation of explainable AI techniques in diabetes prediction models. The transparency and interpretability introduced by these studies are pivotal for building trust in the application of machine learning in medical contexts. Jobeda Jamal Khanam, Simon Y. Foo [5], Yifan Qin et al. [6], and Abhinav Juneja et al. [7] delved into the comparative analysis of various machine learning algorithms for diabetes prediction. This comparative approach sheds light on the nuanced strengths and weaknesses of different models, aiding researchers and practitioners in making informed choices tailored to specific scenarios. Shimoo Firdous et al. [8], Mitushi Soni, Sunita Varma [9], and Quan Zou et al. [10] conducted surveys and studies on diabetes risk prediction using machine learning approaches. These surveys offer comprehensive overviews of existing methodologies, elucidating trends and challenges within the field, and serving as valuable resources for researchers navigating this dynamic landscape. B. Shamreen Ahamed et al.'s exploration [11] into predictive modeling for diabetes mellitus disease prediction and type classification showcases the diverse application of various machine learning techniques and classifiers in this domain. The literature survey further extends its purview into related areas such as diabetic foot ulcers [12], Hallux Valgus classification [13], and ensemble learning for disease prediction [14], illustrating the interdisciplinary nature of diabetes prediction and its far-reaching applications in healthcare. In addition, the contributions of P. Mahajan et al. [14] and Mowafaq Salem Alzboon et al. [15] exploring the early diagnosis of diabetes underscore the timeliness and effectiveness of machine learning methods in predicting and detecting diabetes at its nascent stages. The collective insights from these studies underscore the evolving landscape of diabetes prediction, integrating advancements in machine learning and AI technologies to enhance healthcare outcomes.

III. METHODOLOGY

Dataset:

In this examined research, a text-based dataset is utilized to investigate the predictive elements associated with diabetes, employing a variety of parameters. The dataset incorporates pivotal variables, including the number of pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, body mass index (BMI), diabetes pedigree function, and age. Together, these factors serve as integral components for the evaluation and prognosis of diabetes. The inclusion of pregnancy history is imperative, as it may influence the likelihood of developing diabetes. Glucose levels provide crucial insights into an individual's blood sugar levels, serving as a key diagnostic indicator for diabetes. Factors such as blood pressure and skin thickness contribute valuable information regarding cardiovascular health and potential complications linked to diabetes. The consideration of insulin levels, a hormone impacting blood sugar regulation, further enriches the dataset. The significance of body mass index (BMI) cannot be overstated, as obesity stands out as a recognized risk factor for diabetes. The diabetes pedigree function offers a familial perspective, acknowledging the hereditary dimension of the disease. Additionally, age emerges as a fundamental parameter, given the observed correlation between advancing age and increased diabetes prevalence. Through a comprehensive analysis of these parameters, the research endeavors to refine the accuracy of diabetes prediction models, providing valuable insights into the intricate interplay of factors contributing to the onset of this medical condition.

Data Analysis

In the initial phase of our data analysis, we embarked on a comprehensive data pre-processing journey aimed at enhancing the quality of our dataset. Our primary focus was on addressing missing values, a common challenge in real-world datasets. Through a meticulous approach, we strategically filled in the gaps, ensuring

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:07/July-2024 Impact Factor- 7.868 www.irjmets.com

that the dataset was robust and ready for further analysis. This involved employing techniques such as imputation and interpolation, guided by the nature of the missing data and the context of the variables under consideration. Following the successful completion of data pre-processing, we delved into the realm of Exploratory Data Analysis (EDA) to unravel the intricate relationships and patterns within the dataset. EDA served as a crucial step in understanding the factors that play a pivotal role in predicting diabetes. This multifaceted process involved employing statistical methods, graphical representations, and summary statistics to gain insights into the distribution, central tendencies, and variabilities of the variables. Visualization techniques, such as histograms, box plots, and correlation matrices, were instrumental in highlighting trends and uncovering potential associations between different factors. As we navigated through the EDA process, each variable's contribution to the overall predictive framework for diabetes became apparent. We meticulously examined factors like age, body mass index, blood pressure, and glucose levels, among others, to discern their impact on the outcome variable. The synergy of statistical rigor and visual exploration empowered us to identify outliers, understand data distributions, and discern potential patterns that could inform subsequent modeling efforts. The holistic approach to data pre-processing and EDA not only fortified the dataset against inconsistencies but also laid the foundation for informed feature selection and model building in our pursuit of accurate diabetes prediction.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1

Figure 1: Data Of Different Diabetic Patients Used in Research

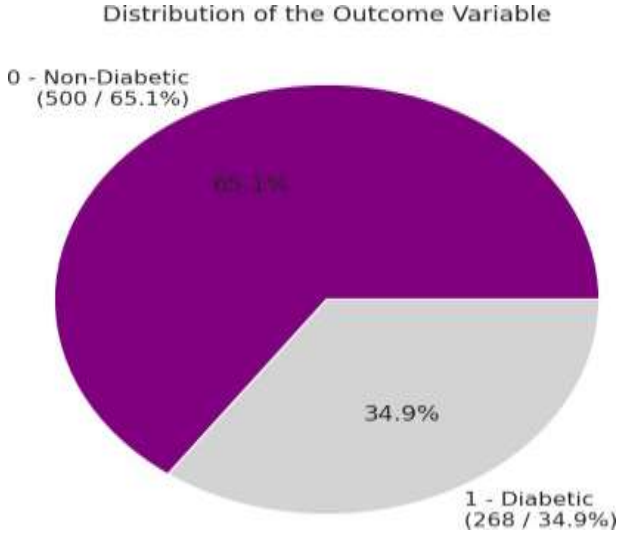

Figure 2: Distribution Of Diabetic and Non-Diabetic Patients

AI Algorithms

In the domain of text-based datasets, the selection of an appropriate machine learning algorithm is pivotal for extracting meaningful insights and ensuring accurate predictions. Various algorithms, each characterized by

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:07/July-2024 Impact Factor- 7.868 www.irjmets.com

distinct strengths, can be deployed to navigate the intricacies of textual data. Prominent options include the Random Forest Classifier, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Decision Tree Classifier, AdaBoost Classifier, Gradient Boosting Classifier, XGBoost Classifier, and Light Classifier. The Random Forest Classifier stands out for its proficiency in handling text data, leveraging its ensemble nature to construct multiple decision trees during training and aggregate their outputs. This approach proves effective in capturing complex relationships within textual features. Logistic Regression, a linear model, emerges as a versatile choice for text classification tasks, offering a probabilistic interpretation of the relationships between input features and binary outcomes-Nearest Neighbors (KNN), a non- parametric algorithm, finds application in text datasets by measuring the similarity between instances. Operating on the principle that similar texts belong to the same class, KNN is valuable for tasks such as text clustering or document categorization. The Support Vector Classifier (SVC), a robust algorithm for binary and multi- class classification, proves effective in text analysis by mapping textual features into high-dimensional spaces and identifying optimal hyperplanes for separation. Ensemble methods, including Decision Tree Classifier, AdaBoost Classifier, Gradient Boosting Classifier, XGBoost Classifier, and LightGBM Classifier, demonstrate excellence in handling text-based datasets. Their ability to amalgamate weak learners into a robust predictive model makes them well-suited for capturing nuanced relationships within textual features. Ultimately, the selection of the most suitable algorithm hinges on the specific characteristics of the text dataset and the objectives of the analysis. Rigorous experimentation with diverse models, accompanied by a careful consideration of their strengths and weaknesses in the context of the textual data, is imperative for achieving optimal results in text-based machine learning tasks.

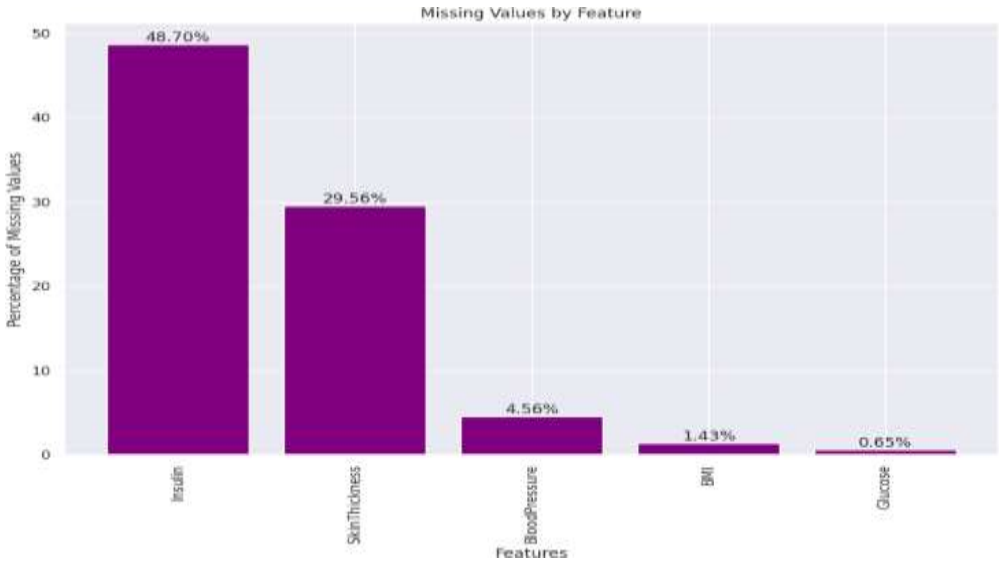

Figure 3: Missing Value of Each Feature

IV. RESULTS AND ANALYSIS

The Random Forest Classifier is a robust machine learning algorithm widely employed for classification tasks. Operating on an ensemble of decision trees, it leverages a technique called bagging to enhance predictive accuracy and mitigate overfitting. By training on random subsets of the data and aggregating predictions, the Random Forest yields a more resilient and accurate model. It excels in handling large datasets, offers insights into feature importance, and is known for its versatility across various domains. With minimal hyperparameter tuning required, it stands as a reliable and efficient choice for practitioners seeking a powerful solution for classification in diverse and complex datasets. The model was trained with 1000 entries and underwent metric evaluations, including accuracy, recall, precision, and F1-score, achieving 77%, 70%, 59%, and 64%, respectively. The process begins with inputting data containing features such as age, BMI, blood pressure, glucose levels, and diabetes labels. Multiple random subsets of the dataset are created through bootstrapping. For each subset, decision trees are constructed, considering features like age, BMI, blood pressure, and glucose levels for splitting. At each split, a random subset of features is considered, introducing diversity among the

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:07/July-2024 Impact Factor- 7.868 www.irjmets.com

trees and preventing overfitting. Predictions from all decision trees are aggregated, often involving majority voting for classification. The performance of the Random Forest is evaluated using metrics such as accuracy, precision, recall, and F1-score, using a separate validation set or cross-validation. The importance of each feature is determined based on how often they are used for splitting across all trees. Finally, the Random Forest model is ready to predict diabetes in new data based on learned patterns.

Logistic Regression is a foundational and widely applied statistical technique for binary and multiclass classification. Despite its name, it is primarily utilized for classification tasks rather than regression analysis. Functioning on the logistic function, the model probabilistically estimates the likelihood of an instance belonging to a specific class. Through the estimation of coefficients for each feature, Logistic Regression maps input variables to a linear combination, which is then transformed into a probability score. Renowned for its simplicity, interpretability, and computational efficiency, Logistic Regression finds utility in scenarios where understanding the individual features' impact on the outcome holds significance. The model was trained with 1000 entries and underwent metric evaluations, including accuracy, recall, precision, and F1-score, achieving 78%, 76%, 56%, and 65%, respectively. The process begins with inputting data containing features such as age, BMI, blood pressure, glucose levels, and diabetes labels. Data preprocessing involves handling missing values, scaling or normalizing features, and splitting the dataset into training and testing sets. Features are standardized or normalized to ensure equal contribution to the model. The logistic regression model is trained by initializing coefficients and intercept, followed by iteratively optimizing parameters by minimizing the cost function, typically using gradient descent. Performance is assessed using metrics like accuracy, precision, recall, and F1-score. An appropriate probability threshold is chosen to convert continuous probability scores into binary predictions. Predictions on new data are made using the trained logistic regression model, and the model's predictions are evaluated and analyzed.

K-Nearest Neighbors (KNN) is a non-parametric and versatile machine learning algorithm used for classification and regression tasks. Operating on the principle that similar instances are likely to share the same class or outcome, KNN classifies a new data point based on the majority class of its k nearest neighbors. It adapts well to various datasets and is effective for both numerical and categorical data. However, its performance can be sensitive to the choice of the distance metric and the value of k. Despite these considerations, KNN remains a straightforward and powerful algorithm for pattern recognition and predictive modeling tasks. The model was trained with 1000 entries and underwent metric evaluations, including accuracy, recall, precision, and F1-score, achieving 76%, 67%, 62%, and 64%, respectively. The process begins with inputting data containing features such as age, BMI, blood pressure, glucose levels, and diabetes labels. Data preprocessing involves handling missing values, scaling or normalizing features, and splitting the dataset into training and testing sets. Features are standardized or normalized to ensure equal contribution to distance calculations. The value of 'k,' the number of neighbors to consider during classification, is determined. KNN does not involve an explicit training phase as it is an instance-based algorithm, simply memorizing the training data. For each new data point, distances to all points in the training set are calculated, and the 'k' nearest neighbors are identified. The class label is assigned based on the majority class among the 'k' neighbors. The KNN model then predicts whether a new data point belongs to the diabetic or non-diabetic class based on the majority vote of its nearest neighbors.

Support Vector Machine (SVM) is a powerful supervised learning algorithm that excels in binary and multiclass classification tasks. Operating on the principle of finding an optimal hyperplane that separates different classes while maximizing the margin between them, SVM is effective even in high-dimensional spaces. It transforms input data into a higher-dimensional space using a kernel trick, allowing for complex decision boundaries. With a focus on support vectors, the data points nearest to the hyperplane, SVM ensures robustness to outliers. Widely used in various domains, SVM's versatility and ability to handle both linear and non-linear classification make it a prominent choice in machine learning applications. The model was trained with 1000 entries and underwent metric evaluations, including accuracy, recall, precision, and F1-score, achieving 74%, 68%, 50%, and 58%, respectively. The process begins with inputting data containing features such as age, BMI, blood pressure, glucose levels, and diabetes labels. Data preprocessing involves handling missing values, scaling or normalizing features, and splitting the dataset into training and testing sets. Features are standardized or

normalized to ensure equal contribution to decision boundaries. An appropriate kernel (linear, polynomial, radial basis function, etc.) is chosen based on the characteristics of the dataset. The SVM is trained using the training set, optimizing parameters such as C (regularization parameter) and kernel-specific parameters. The optimal hyperplane that separates different classes is determined, maximizing the margin. New data points are classified based on their position relative to the decision boundary. Performance is assessed using metrics such as accuracy, precision, recall, and F1-score. The trained SVM is then ready for predicting diabetes based on the learned decision boundaries.

The Decision Tree Classifier is a versatile and interpretable machine learning algorithm employed for both classification and regression tasks. Its functionality involves iteratively partitioning the dataset based on the most informative features, constructing a tree-like structure. At each node, the algorithm strategically selects the feature that maximizes information gain or minimizes impurity, forming effective decision boundaries. Renowned for its ease of comprehension and visualization, decision trees prove valuable in elucidating the decision-making process. However, they are susceptible to overfitting, prompting the application of techniques like pruning to enhance generalization. Despite this challenge, Decision Tree Classifiers find widespread use owing to their simplicity and adaptability across diverse domains. The model was trained with 1000 entries and underwent metric evaluations, including accuracy, recall, precision, and F1-score, achieving 72%, 61%, 58%, and 59%, respectively. The process begins with inputting data containing features such as age, BMI, blood pressure, glucose levels, and diabetes labels. Data preprocessing involves handling missing values, encoding categorical variables, and splitting the dataset into training and testing sets. Features are chosen based on their relevance and importance for predicting diabetes. The decision tree is constructed by selecting the best feature at each node based on criteria like Gini impurity or information gain. Conditions for splitting nodes are determined, considering feature values that optimize the chosen impurity measure. Optionally, tree pruning is performed to prevent overfitting by removing branches that do not contribute significantly to predictive accuracy. New data points are classified by traversing the decision tree based on their feature values. Performance is assessed using metrics such as accuracy, precision, recall, and F1-score. The trained Decision Tree Classifier is then ready for predicting diabetes based on learned decision rules.

AdaBoost (Adaptive Boosting) Classifier is an ensemble learning algorithm known for enhancing the performance of weak learners in binary classification. It iteratively assigns weights to misclassified instances, guiding subsequent weak learners to focus on these errors. The algorithm combines the predictions of multiple weak classifiers, giving more weight to those with higher accuracy. AdaBoost adjusts the weights at each iteration, emphasizing challenging instances and creating a strong classifier. Robust and versatile, AdaBoost is less prone to overfitting and excels in improving classification accuracy, making it valuable in various applications, from face detection to bioinformatics, where accurate predictions are paramount. The model was trained with 1000 entries and underwent metric evaluations, including accuracy, recall, precision, and F1-score, achieving 75%, 67%, 56%, and 61%, respectively. The process begins with inputting data containing features such as age, BMI, blood pressure, glucose levels, and diabetes labels. Equal weights are initially assigned to all data points. For each iteration, a weak classifier (e.g., Decision Stump) is trained on the weighted dataset, calculating the classifier's error and emphasizing misclassified instances. A weight is assigned to the weak classifier based on its accuracy, giving more weight to accurate classifiers. Weights of misclassified instances are increased, making them more influential in subsequent iterations. Predictions of all weak classifiers are combined based on their weights. Performance is assessed using metrics like accuracy, precision, recall, and F1-score. The trained AdaBoost Classifier is then ready for predicting diabetes, leveraging the collective strength of multiple weak learners.

Gradient Boosting Classifier is a powerful ensemble learning technique widely used for both classification and regression tasks. It builds a strong predictive model by combining multiple weak learners, typically decision trees, sequentially. Each tree corrects the errors of its predecessor, optimizing the overall model performance. The algorithm minimizes a loss function by adding new models that predict the residuals of previous models. Gradient Boosting is highly flexible, allowing customization of loss functions and tree structures. Despite its susceptibility to overfitting, careful tuning of hyperparameters like learning rate and tree depth ensures robust and accurate predictions. The model was trained with 1000 entries and underwent metric evaluations,

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:07/July-2024 Impact Factor- 7.868 www.irjmets.com

including accuracy, recall, precision, and F1-score, achieving 78%, 72%, 60%, and 65%, respectively. The process begins with inputting data containing features such as age, BMI, blood pressure, glucose levels, and diabetes labels. Data preprocessing involves handling missing values, encoding categorical variables, and splitting the dataset into training and testing sets. Features are chosen based on their relevance and importance for predicting diabetes. The gradient boosting algorithm initializes with a simple model, often a constant value. Iteratively, new decision trees are added, each trained to predict the residuals (errors) of the combined ensemble model from the previous iteration. Trees are constructed by optimizing a loss function, typically mean squared error for regression or log loss for classification. The final model combines the predictions of all individual trees, weighted by their respective contributions. Performance is assessed using metrics like accuracy, precision, recall, and F1-score. The trained Gradient Boosting Classifier is then ready for predicting diabetes based on the learned patterns and corrections of previous models.

Table 1: Comparison of Accuracy of different models.

MODEL	ACCURACY
Random Forest Classifier	77%
Logistic Regression	78%
K-NearestNeighbors	76%
Support Vector Classifier	74%
Decision TreeClassifier	72%
AdaBoostClassifier	75%
GradientBoostingClassifier	76%
XGBoost Classifier	76%
LightGBM Classifier	80%

V. FUTURE SCOPE

The future scope for the use of LightGBM in diabetes prediction is promising. Further research can explore the algorithm's performance across diverse datasets and its adaptability to different healthcare scenarios. Fine-tuning hyperparameters and optimizing the model for specific characteristics of diabetes datasets could enhance its predictive capabilities. Additionally, integrating advanced feature engineering techniques and exploring interpretability aspects of the model could provide valuable insights for medical practitioners. Collaborative efforts between machine learning experts and healthcare professionals can lead to the development of more accurate and reliable predictive models for diabetes diagnosis and management

VI. CONCLUSION

In conclusion, the LightGBM Classifier stands out as a powerful and efficient tool for predicting diabetes. Its unique features, such as histogram-based learning and leaf- wise growth, contribute to its speed, scalability, and low memory usage. Leveraging these characteristics, LightGBM can effectively handle large datasets with high-dimensional features, making it particularly well-suited for predictive modeling tasks in healthcare, such as diabetes prediction. The model's accuracy and efficiency make it an asset in applications where rapid and accurate predictions are crucial.

VII. REFERENCES

[1] Rani, KM Jyoti. "Diabetes prediction using machine learning." International Journal of Scientific Research in Computer Science Engineering and Information Technology 6 (2020): 294-305.

[2] Jaiswal, Varun, Anjli Negi, and Tarun Pal. "A review on current advances in machine learning based diabetesprediction." Primary Care Diabetes 15.3 (2021): 435-443

[3] Tasin, Isfafuzzaman, et al. "Diabetes prediction using machine learning and explainable AI techniques." Healthcare Technology Letters 10.1-2 (2023): 1-10.

[4] Modak, Sandip Kumar Singh, and Vijay Kumar Jha. "Diabetes prediction model using machine learning

techniques." Multimedia Tools and Applications (2023): 1-27

[5] Khanam, Jobeda Jamal, and Simon Y. Foo. "A comparison of machine learning algorithms for diabetes prediction." Ict Express 7.4 (2021): 432- 439

[6] Qin, Yifan, et al. "Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type." International Journal of Environmental Research and Public Health 19.22 (2022): 15027

[7] Juneja, Abhinav, et al. "Predicting diabetes mellitus with machine learning techniques using multi-criteria decision making." International Journal of Information Retrieval Research (IJIRR) 11.2 (2021): 38-52

[8] Firdous, Shimoo, Gowher A. Wagai, and Kalpana Sharma. "A survey on diabetes risk prediction using machine learning approaches." Journal of Family Medicine and Primary Care 11.11 (2022): 6929.

[9] Soni, Mitushi, and Sunita Varma. "Diabetes prediction using machine learning techniques." International Journal of Engineering Research & Technology (Ijert) Volume 9 (2020).

[10] Zou, Quan, et al. "Predicting diabetes mellitus with machine learning techniques." Frontiers in genetics 9 (2018): 515.

[11] Ahamed, B. Shamreen, et al. "Diabetes Mellitus Disease Prediction and Type Classification Involving Predictive Modeling Using Machine Learning Techniques and Classifiers." Applied Computational Intelligence and Soft Computing 2022 (2022).

[12] Cassidy, Bill, et al. "Artificial intelligence for automated detection of diabetic foot ulcers: A real- world proof-of-conceptclinical evaluation." Diabetes Research and Clinical Practice 205 (2023): 110951.

[13] Hida, Mitsumasa, et al. "Development of Hallux Valgus Classification Using Digital Foot Images with Machine Learning." Life 13.5 (2023): 1146.

[14] Mahajan, P., et al. "Ensemble Learning for Disease Prediction: A Review. Healthcare 2023, 11, 1808." (2023)

[15] Alzboon, Mowafaq Salem, et al. "Early Diagnosis of Diabetes: A Comparison of Machine Learning Methods." International Journal of Online & Biomedical Engineering 19.15 (2023).