Enhancing Type 2 Diabetes Prediction: A Hybrid Approach with Rf & LightGBM

KARTHIKEYAN S[1], NALLASIVAM P V[2], KEERTHANAN G B I[3], A. RAIHANA, M. E.[4]

[1, 2, 3]*Student, Department of Information Technology, Sri Krishna College of Engineering & Technology, Coimbatore.*

[4]*Professor, Department of Information Technology, Sri Krishna College of Engineering & Technology, Coimbatore.*

Abstract- One major global health concern is diabetes mellitus, especially type 2 diabetes mellitus (T2DM). For T2DM to be effectively managed, early and precise prognosis is crucial. In order to improve accuracy and guarantee adaptability across a variety of datasets, this work presents a prediction model that combines the Random Forest (RF) and LightGBM algorithms. To maximize performance, the model uses extensive preprocessing methods, such as feature selection and data manipulation. The Pima Indians Diabetes Dataset and other publicly accessible datasets were used to assess the suggested methodology, which produced reliable and efficient prediction results. These results demonstrate how RF and LightGBM can be used to create dependable and flexible models for the analysis and prediction of diabetes.

Indexed Terms- Diabetes Prediction, Machine learning, Random Forest, LightGBM

I. INTRODUCTION

Millions of individuals worldwide suffer with diabetes mellitus, of which type 2 diabetes mellitus (T2DM) is the most common type. Until a diagnosis is made, people are frequently unaware of their risk factors because this disorder develops silently. Effective treatment and intervention of type 2 diabetes depend on early detection. In order to analyze complex medical data and find patterns that can help forecast disease, data mining and machine learning approaches have become extremely effective. In order to create a reliable predictive model for type 2 diabetes, we investigate the possibilities of the Random Forest (RF) and LightGBM algorithms in this work. The model is built to accommodate differences in data quality and structure between datasets by utilizing sophisticated preprocessing techniques. This flexibility guarantees that the model may be used with a variety of data sources, providing a workable solution for diabetes analysis and prediction.

1.1 DIABETES PREDICTION

Diabetes prediction is the use of computer methods to determine a person's risk of type 2 diabetes mellitus based on a number of health-related characteristics. Analyzing past data, determining possible risk factors, and creating models that can spot trends linked to the illness are all steps in this process. Diabetes prediction systems can process big datasets and identify trends that might not be immediately obvious through traditional analysis by using statistical techniques and machine learning algorithms. By helping to classify people into various risk groups, efficient prediction models can support early intervention tactics and better health management.

1.2 MACHINE LEARNING

The goal of the artificial intelligence field of machine learning is to create algorithms that can learn from data and make judgments without the need for explicit programming. These algorithms have the ability to examine intricate information, identify significant trends, and gradually enhance their functionality. Machine learning is essential to predictive modeling because it can handle both structured and unstructured data, optimize feature selection, and improve classification accuracy. To extract insights from datasets, a variety of methodologies can be used, including supervised, unsupervised, and ensemble learning methods. Machine learning makes ensuring that predictive systems are effective and flexible enough to handle a variety of data types by iteratively improving models over time.

1.3 RANDOM FOREST

An ensemble learning method called Random Forest builds several decision trees and combines their results to improve forecast stability and accuracy. To lessen bias and variation, a random subset of the dataset is used to build each tree in the model. Random Forest produces a more generalized model that is less likely to overfit by aggregating the predictions from several trees. Because it can effectively handle high-dimensional datasets, noisy data, and missing values, this approach is very useful for classification and regression problems. Furthermore, Random Forest offers feature importance ratings, which make it possible to determine which attributes in a dataset have the greatest influence. It is a popular method in many predictive analytics applications because of its capacity to handle intricate correlations found in data.

1.4 LIGHTGBM

A high-performance gradient boosting framework called LightGBM (Light Gradient Boosting Machine) was created to handle big datasets more effectively. LightGBM uses a leaf-wise growth strategy, in contrast to conventional boosting techniques, which allows for quicker training periods and improved handling of unbalanced datasets. It offers a number of optimization strategies that lower memory usage and computational expenses, including exclusive feature bundling and histogram-based learning. LightGBM ability to handle sparse data and carry out automatic feature selection makes it incredibly efficient in classification and regression tasks. It also provides optimized hyper parameter settings, which enhance customization and flexibility across various machine learning applications. It is a popular option for creating predictive models across a variety of disciplines because to its scalability and excellent accuracy.

II. LITERATURE REVIEW

Riccardo B, Blaz Z, et al. have suggested Predictive data mining is quickly becoming a vital tool for medical researchers and clinicians. Deploying these techniques and disseminating the results requires an understanding of the key concerns behind them as well as the implementation of established and standardized procedures. The process of choosing, examining, and modeling vast volumes of data in order to find unidentified patterns or relationships that give the data analyst a clear and practical outcome is known as data mining. The phrase "data mining," which was first used in the middle of the 1990s, is now synonymous with "knowledge discovery in databases," emphasizing the process of analyzing data rather than the use of particular analysis techniques. A variety of methods from computer science, such as multi-dimensional databases, machine learning, soft computing, and data visualization, and statistics, such as hypothesis testing, clustering, classification, and regression techniques, are frequently used to solve data mining problems. Generally speaking, data mining jobs fall into two categories: description tasks and prediction tasks. After taking into account the data as a whole and building a model, prediction tries to forecast some response of interest, whereas description looks for patterns and relationships that are understandable to humans. Decision trees are commended for their openness, which enables the decision-maker to study and comprehend the decision model and how it functions. Every path in the decision tree can also be thought of as a decision rule [1].

Given the rise in the number of cases reported globally, diabetes is today one of the biggest hazards to human life, according to this study by Mechelle Gittens, Reco King, et al. Since type 2 diabetes accounts for the bulk of cases identified, this abrupt increase has been attributed to changes in human lifestyle. To help patients live healthier lives, mobile health (m-health) technologies are being used in every sector of the healthcare sector. With one of the highest rates of diabetes and amputation in the world, the society we chose for our study is in crisis and has a population that is primarily of African heritage. A mobile phone, a health data server, and a data acquisition module (DAM) make up the suggested system. Using a variety of sensors, the DAM collects patient data, which it then transmits over Bluetooth to the mobile device. The readings are transmitted to a distant health data center via an I network, such as the Internet, after they arrive at the mobile device. After viewing the readings, the medical practitioners can respond accordingly. According to the authors, the system's round-the-clock patient monitoring might take the place of in-person consultations between patients and physicians. As a result, patients can get the care they require in the convenience of their own homes. This solution can be used without requiring consumers to own a smartphone, in contrast to the majority of the research that has been investigated. An alternative to using pricey smartphones is provided by this research, as the majority of people with chronic illnesses are typically elderly individuals who may find them confusing [2].

Among others, Marcano-Cede-no Alexis These days, diabetes affects people of all ages and in all populations. Various artificial intelligence techniques have been used to address the issue of diabetes. The artificial met plasticity on multilayer perceptron (AMMLP) was proposed in this study as a diabetes prediction model. Diabetes leads to heart disease and raises the risk of renal illness, blindness, nerve damage, blood vessel damage, and other conditions. According to the globe Health Organization, there were around 170 million diabetics in the globe in 2000. By 2030, that number is expected to more than double to 366 million. Type 1, or insulin-dependent diabetes, and type 2, or non-insulin-dependent diabetes, are the two main types of diabetes. The hallmark of type 1 diabetes is a complete lack of insulin secretion. Genetic markers and serological evidence of an autoimmune pathologic process in the pancreatic islets are frequently used to identify people who are more likely to acquire this kind of diabetes. Type 2 diabetes is the most prevalent type. Applying artificial meta plasticity to multilayer perceptrons (AMMLP) as a prediction model for diabetes prediction was the primary goal of the current study. The suggested model AMMLP was tested using the Pima Indian diabetes data set. The outcomes of AMMLPT were contrasted with those of other algorithms that were applied to the same database and recently proposed by other academics. It is challenging for doctors to diagnose diabetes because there are numerous factors to consider [3].

According to Veena Vijayan V et al.'s studies, diabetes mellitus is brought on by an elevated blood sugar level. Numerous consequences, including kidney failure, stroke, cancer, heart disease, and blindness, may result from this. These problems can be recognized and prevented with the aid of early diagnosis and identification. Several computerized information systems for diabetes diagnosis and prediction were created utilizing various classifiers. It is obvious that using the right categorization algorithms improves the system's accuracy and effectiveness. This study's primary goal is to examine the advantages of various preprocessing methods for diabetes prediction decision support systems that rely on support vector machines (SVM), naive bayes classifiers, and decision trees. Computational approaches, statistical techniques, clustering, classification, pattern recognition, and transformation are all part of data mining. Medical data mining opens up a vast array of medical data analysis sources by identifying hidden patterns in vast amounts of heterogeneous data. Health and medical organizations must work together at a higher level in biomedical and healthcare systems. Maintaining consistency in a coordinated background is one of the biggest challenges facing biomedical practitioners. Diabetes is a severe medical condition in which the body is unable to control its sugar intake [4].

According to this study by Ms. K. Sowjanya et al., diabetes mellitus (DM) is becoming potentially epidemic in India. Diabetes and its possible complications cause a great deal of disease and devastation, which has led to a substantial health care burden on households and society as a whole. The worrying aspect is that diabetes is increasingly being shown to be associated with several problems and to be occurring in the nation at a relatively younger age. Diabetes Mellitus (DM), or simply diabetes, is a condition brought on by either the pancreas' inability to generate insulin or the body's cells' inability to respond to it. Insulin, one of the hormones that serves as a key to enable glucose from the blood to enter our cells, powers our body's cells. A person is diagnosed with diabetes if the insulin-producing beta cells in the pancreas are depressed. This results in inadequate regulation of blood glucose, which causes the blood glucose level to rise suddenly. There are basically four forms of diabetes on that point. A glucose level that is higher than usual but not yet high enough to be classified as diabetes is called prediabetes [5].

III. RELATED WORK

Among the world's most well-known non-transmittable diseases, diabetes stands out in the current system. According to estimates, it ranks as the sixth most common cause of mortality. Medical experts and researchers have made early diabetes detection a top priority. Collaborative studies have demonstrated that by utilizing computer abilities and algorithms (such data mining), effective, economical, and quick methods for diagnosing diabetes can be developed thanks to the extensive technical advancement in computer science. The experiment's results demonstrate that the Adaboost machine learning ensemble technique performs better than both a J48 decision tree and bagging. The primary goal of a diabetes prediction system is to forecast whether a candidate will get diabetes at a given age. By using decision trees, the suggested system is created using the machine learning approach. The results were good since the system that was created was able to forecast the incidence of diabetes at a specific age with greater accuracy by employing decision trees.

IV. METHODOLOGY

The objective of the suggested system is to use machine learning methods, particularly the Random Forest (RF) and LightGBM algorithms, to create a reliable model for predicting type 2 diabetic mellitus (T2DM). To guarantee that the model is flexible and able to manage numerous datasets with different architectures, the system makes use of extensive data preprocessing procedures, such as feature selection and transformation. The system is intended to forecast T2DM with high accuracy and flexibility across various data sources by utilizing the advantages of both RF and LightGBM. The design of the system guarantees that it can spot significant trends and connections in the data, offering insightful information for accurate diabetes prediction. Furthermore, the suggested method is designed to be dependable and effective, guaranteeing steady performance across a range of diabetes-related datasets.

A. DATA COLLECTION

Importing the datasets needed for the prediction model is the responsibility of this module. It guarantees that information is gathered and structured consistently from a variety of sources. The data is verified for completeness and may originate from databases or structured files. Making sure that data is properly integrated reduces discrepancies and improves the quality of subsequent processing. Prior to proceeding, this stage additionally confirms data integrity by looking for duplicate entries or missing values. Handling the data correctly at this point guarantees that the next processes can go forward without mistakes or misunderstandings.

B. DATA PRE-PROCESSING

This module improves the quality of the dataset by cleaning and transforming it. Categorical variables are encoded to make them appropriate for model training, numerical features are scaled for uniformity, and missing values are managed to avoid bias in the model. To increase efficiency, redundant or unnecessary qualities are eliminated. To keep extreme results from compromising the accuracy of the model, outlier detection techniques are used. Effective data preparation guarantees that the data is optimized for use in Random Forest and LightGBM, improving predictions. To ensure consistency in the dataset and minimize noise, the preprocessing phase is crucial.

C. FEATURE EXTRACTION

In order to increase prediction accuracy, this module focuses on choosing the most pertinent attributes from the dataset. The objective is to remove less helpful information while keeping characteristics that significantly influence the prediction of type 2 diabetes mellitus. Random Forest and LightGBM's feature importance strategies aid in determining which variables have the greatest influence on the classification process. Key feature extraction boosts model performance and increases computational efficiency. The training process is kept from becoming needlessly complex by making sure that only significant attributes are employed.

D. TRAINING AND TESTING

To assess model performance, the dataset is divided into training and testing subsets during this session. Patterns in the training data are discovered using the Random Forest and LightGBM algorithms. By modifying parameters like the Random Forest tree count and the LightGBM learning rate, hyperparameter tuning is done to maximize model performance. In order to ensure that it generalizes successfully to new data, the model learns the correlations between features and target labels. The testing phase helps to further optimize the method by evaluating how well the trained model works on unknown data. Overfitting is avoided and the model's ability to produce accurate predictions is guaranteed by appropriate training and testing protocols.

E. MODEL EVALUATION

This module employs a number of evaluation criteria to gauge the trained model's efficacy. The quality of predic`tions produced by Random Forest and LightGBM is evaluated using accuracy, precision, recall, F1-score, and AUC-ROC. The model's ability to differentiate between cases with and without diabetes is measured by these measures. Confusion matrices aid in further model refinement and offer insights on misclassifications. To increase accuracy, changes like hyperparameter tuning and data reprocessing are made as needed. A thorough review guarantees that the finished model is reliable and appropriate for predictive analysis.

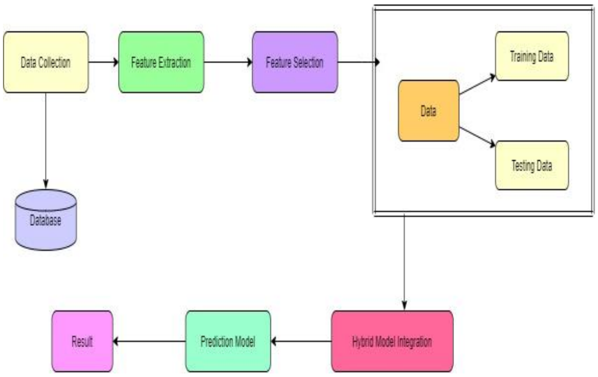

Figure 1. Block diagram

V. RESULT ANALYSIS

Analyzing the results entails assessing how well the Random Forest and LightGBM models predict type 2 diabetes mellitus. To evaluate the model's efficacy, a number of performance indicators are measured, such as accuracy, precision, recall, and F1-score. The findings demonstrate how feature extraction and preprocessing enhance classification performance. The models' comparison highlights each algorithm's advantages in managing distinct patterns seen in the dataset. The assessment demonstrates that the chosen methods yield accurate forecasts, guaranteeing a well-rounded strategy for determining diabetes risk. The model's overall prediction effectiveness and adaptability to various datasets can both be improved with more tuning.

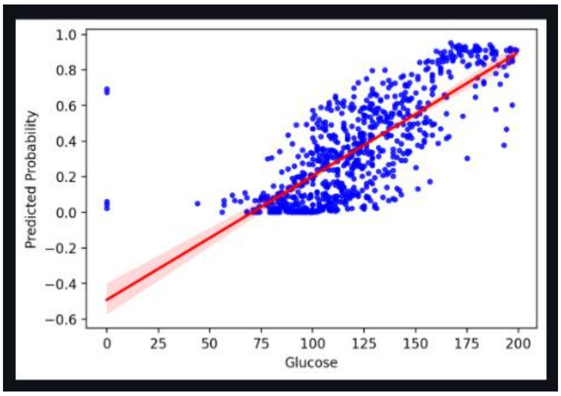

Figure 2: Regression plot

CONCLUSION

To improve accuracy and efficiency, Random Forest and LightGBM are included into the type 2 diabetes mellitus prediction model. The system efficiently detects patterns linked to diabetes risk using structured data processing, which includes preprocessing, feature extraction, and model evaluation. By ensuring that only pertinent features are used in decision-making, the method increases prediction reliability. The evaluation results show how well the chosen algorithms handle a variety of datasets while retaining performance consistency. Better decision support in health-related data analysis is facilitated by the model's organized implementation and streamlined workflow. Future improvements can expand the model's applicability to a wider range of datasets and further improve its adaptability.

REFERENCES

[1] Ricardo B, Blaz Z. Current concerns and recommendations regarding predictive data mining in clinical medicine. 77:81–97; Int J Med Inf 2008.

[2] Adrian Als, Curtis Gittens, Mecelle Gittens, and Reco King IEEE 16th International Conference on e-Health Networking, Applications, and Services, 2014; Post-diagnosis Management of Diabetes via a Mobile Health Consultation Application (Healthcom).

[3] Andina Diego, Torres Joaquín, and Marcano-Cede~no Alexis. An artificial metaplasticity-based diabetes prediction model. LNCS 6687, Part II, IWINAC 2011, pp. 418–25.

[4] A evaluation of decision support systems for diabetes mellitus prediction by Veena Vijayan V. and Anjali C. Global Conference on Communication Technologies (GCCT 2015) Proceedings.

[5] Ms. K. Sowjanya, MobDBTest: A mobile device-based machine learning system for diabetes risk prediction. International Advance Computing Conference, IEEE, 2015 (IACC).

[6] Design and Implementation of a Diabetes Risk Assessment Model Based on Mobile Things by Gang Shi, Shanshan Liu, and Ding Ye, 7th International Conference on Information Technology in Medicine and Education, 2015.

[7] Xiaolong Su and Junao Wang, 2011 IEEE 3rd International Conference on Communication Software and Networks (ICCSN): An enhanced K-Means clustering technique.

[8] Enhanced k-means clustering for longitudinal data using Efros distance, Yanhui Sun, Liying Fang, and Pu Wang, Chinese Control and Decision Conference (CCDC), 2016.

[9] Shunye Wang, 2013 3rd International Conference on Computer Science and Network Technology (ICCSNT), Enhanced K-means clustering technique based on the optimized initial centroids.

[10] "Improving Type 2 Diabetes Mellitus Risk Prediction Using Classification," by Phattharat Songthung and Kunwadee Sripadkulchai, 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2016.