DIABETES FORECASTING USING MACHINE LEARNING MODEL

Kapalavayi Ramesh Babu[1], Jonna Seshu[2], Kanala GopiChandra Sekhar Reddy[3], Pappula Dheeraj[4], Pallapu Tarun[5], Petla Srinivas[6]

[1] Assistant Professor, Dept of Electronics and Communication Engineering, Vasireddy Venkatadri Institute ofTechnology, Nambur, Andhra Pradesh, India
[2,3,4,5,6] UG Student, Dept of Electronics and Communication Engineering, Vasireddy Venkatadri Institute ofTechnology, Nambur, Andhra Pradesh, India

ABSTRACT

This project endeavors to delve into a diverse array of health-related variables and their intricate connections in order to develop an accurate diabetes forecasting model utilizing the random forest approach. Spanning factors such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level, the investigation aims to unravel the complex web of influences surrounding diabetes onset. By meticulously examining these variables, the study not only seeks to forecast diabetes risk with precision but also lays a robust foundation for subsequent research endeavors. Furthermore, the study's comprehensive analysis promises to yield valuable insights into the patterns and trends associated with diabetes susceptibility, paving the way for a deeper understanding of its occurrence and progression. Such insights are pivotal in informing healthcare practices aimed at improving patient care and outcomes in this increasingly critical domain. With a focus on elucidating the intricate interactions among these factors, the research sets the stage for future investigations to explore how these variables collectively shape the landscape of diabetes, thus offering invaluable knowledge for refining strategies in patient management and ultimately enhancing health outcomes.

Keyword:- *Diabetes Prediction, Random Forest Classifier, HbA1c_level, Hyperparamater Tuning, Accuracy Assessment*

1. INTRODUCTION

Diabetes has become increasingly prevalent in contemporary society, extending beyond its historical association with adulthood and old age to afflict individuals in their teenage years. A myriad of factors contribute to the development of this condition, including familial predisposition, age, dietary patterns, high blood pressure, and obesity. The condition manifests in two primary forms: type 1 and type 2 diabetes. Type 1 diabetes arises from insufficient insulin production, resulting from an autoimmune response that impairs pancreatic insulin-producing cells. Conversely, type 2 diabetes stems from insulin resistance within the body, often exacerbated by sedentary lifestyles, poor dietary choices, and obesity. Notably, both types of diabetes carry significant health risks if left undetected or uncontrolled. Recent research underscores concerning trends, with a surge in type-1 diabetes diagnoses among younger demographics and an increase in gestational diabetes cases linked to maternal obesity and unhealthy dietary habits. Additionally, conditions such as polycystic ovary syndrome further underscore the interconnectedness of lifestyle factors and health outcomes.

In the quest for effective management and prevention strategies, traditional approaches have proven insufficient. To address this gap, we propose a novel methodology leveraging a contemporary, digitally-driven medical dataset. This dataset captures nuanced insights into individuals' dietary patterns, exercise routines, and daily habits, diverging from static, fixed-value inputs common in traditional models. Through the integration of advanced machine learning algorithms such as Random Forest, Decision Trees, and Naïve Bayes, our approach aims to yield more accurate and efficient results. Our preliminary findings indicate that Random Forest, in particular, demonstrates promising performance when paired with our digital dataset, offering a step forward in diabetes research and predictive modeling.

1.1 Domain Knowledge:

- **Age:** The influence of age on diabetes risk is well-established within the medical community. As individuals advance in age, their susceptibility to diabetes escalates, attributed in part to lifestyle changes and physiological alterations. Reduced physical activity, hormonal fluctuations, and a higher prevalence of comorbid health conditions collectively contribute to this heightened risk among older adults.
- **Gender:** Gender dynamics also play a role in diabetes susceptibility, albeit with nuanced variations. Notably, women with a history of gestational diabetes exhibit an elevated risk of type 2 diabetes later in life, underscoring the enduring impact of pregnancy-related metabolic changes. Conversely, while some studies suggest a marginally higher diabetes risk among men, the interplay of gender-specific factors in diabetes etiology warrants further investigation.
- **Body Mass Index (BMI):** Body Mass Index (BMI) serves as a fundamental indicator of adiposity and is instrumental in predicting diabetes risk. Elevated BMI levels, indicative of excess body fat, are closely associated with insulin resistance and impaired glucose regulation.
- **Hypertension:** Hypertension, a prevalent comorbidity in diabetes, shares a complex interrelationship with the condition. Mutual exacerbation ensues, as hypertension contributes to the pathogenesis of type 2 diabetes, while diabetes-related metabolic abnormalities exacerbate hypertension. Consequently, the coexistence of these conditions amplifies cardiovascular risks, necessitating comprehensive management strategies.
- **Heart Disease:** Bidirectional links between heart disease and diabetes underscore the shared pathophysiological mechanisms and risk factors. Obesity, hypertension, and dyslipidemia constitute common denominators, driving the concomitant development of both conditions. As such, individuals afflicted with one ailment face an augmented risk of succumbing to the other, accentuating the imperative for holistic cardiovascular risk management.
- **Smoking History:** Cigarette smoking emerges as a modifiable risk factor intricately intertwined with diabetes etiology. Chronic smoking exerts deleterious effects on insulin sensitivity and glucose metabolism, thereby heightening the propensity for type 2 diabetes development. Encouragingly, cessation interventions offer tangible benefits in mitigating diabetes risk and ameliorating long-term metabolic health.
- **HbA1c Level:** HbA1c levels furnish invaluable insights into glycemic control and serve as a pivotal marker in diabetes management. Elevated HbA1c levels signify suboptimal blood sugar regulation, portending an augmented risk of diabetes and its complications. Consequently, vigilant monitoring of HbA1c levels is imperative in guiding therapeutic interventions and averting adverse health outcomes.
- **Blood Glucose Level:** The measurement of blood glucose levels constitutes a cornerstone in diabetes diagnosis and management. Elevated blood glucose levels, particularly in fasting states or postprandially, signal aberrant glucose metabolism and heighten diabetes risk. Regular monitoring of blood glucose levels empowers clinicians to initiate timely interventions, thereby forestalling disease progression and optimizing patient outcomes.

2. LITERATURE SURVEY:

Hasan et al. concentrate on the challenge of precisely predicting diabetes when faced with outliers or missing values in the dataset, as well as limited labeled data. The suggested framework makes use of several machine learning classifiers, multilayer perceptron, data standardization, feature selection, outlier rejection, filling in missing values, and K-fold cross-validation. Assembling various classifiers with weights calculated from the AUC metric is advised to increase prediction accuracy. On the Pima Indian Diabetes Dataset, the effectiveness of the suggested framework is assessed, and it is discovered to outperform existing approaches, attaining an AUC of 0.950, a 2% improvement over the results of the most recent studies. The source code for predicting diabetes is accessible to anyone interested.

Sarwar et al. investigate the implementation of machine learning algorithms for anticipatory analytics in healthcare. The study examines six distinct machine-learning methods to anticipate diabetes using the medical records of patients. The algorithms are contrasted based on performance and accuracy to determine which algorithm is best for predicting diabetes. The major goal is to use machine learning approaches to enable early diabetes prediction, ultimately helping practitioners and healthcare providers. The effects of diabetes on a global scale as well as the difficulties in making an early prognosis because of the many relationships it has with several other factors. Despite the existence of conventional methods for diabetes diagnosis, data science techniques, particularly machine learning, may help with early prediction with improved accuracy. A system that combines the results of three different supervised machine learning approaches—namely, Support Vector Machine, logistic regression, and Artificial Neural Network to predict diabetes and provide a useful strategy for early illness diagnosis is critical.

Amani Yahyaoui et al. focus on predicting and detecting diabetes by evaluating the effectiveness of medical Decision Support Systems (DSS) in supporting healthcare professionals in making clinical decisions. The suggested decision support system (DSS) employs Machine Learning (ML) techniques and contrasts conventional machine learning methodologies with deep learning approaches. To achieve this, the research employs a Support Vector Machine (SVM) and Random Forest (RF), which are the two most commonly used classifiers in conventional machine learning, and a fully convolutional neural network (CNN) in deep learning. When tested on the Pima Indians Diabetes database, RF outperformed deep learning and ML methods in predicting diabetes. The study described investigates the escalating incidence of diabetes, which is associated with high blood sugar and obesity. The primary objective is to identify the critical factors that contribute to diabetes and underscore the attributes that are essential for predicting an individual's likelihood of developing the disease. The authors also note that in fields where enormous datasets are available, variable and feature selection represent vital areas of research.

Priyanka Sonar et al, focus on how dangerous diabetes is and how it can cause serious illnesses like heart failure, renal problems, and blindness. In addition to developing machine learning techniques to forecast the risk of diabetes in patients, routine checks are essential. The goal is to construct a system that employs Naive Bayes, Decision Tree, Artificial Neural Network, and SVM algorithms to properly forecast a patient's probability of getting diabetes. 85% accuracy is achieved by the Decision Tree model, according to the results, while 77% and 77.3% accuracy are attained by the Naive Bayes and SVM models, respectively. The results imply that machine learning techniques can accurately forecast a patient's probability of developing diabetes.

Quan Zou et al. utilized physical examination data from a Chinese hospital located in Luzhou and established three distinct machine learning models: decision tree, random forest, and neural network. The researchers enhanced the models' performance by employing dimensionality reduction methods like principal component analysis and minimum redundancy maximum relevance. When all characteristics were taken into account, the random forest model generated the most precise outcome of 0.8084 accuracy.

3. EXISTING MODELS:

Based on the data mining techniques, the most widely used machine learning techniques are blood glucose anomalies detection, blood glucose dynamics and decision making models. Information science provides a huge number of order calculations, for example, Logistic Regression, Support Vector Machine, Decision Trees and Naive Bayes classifier. Be that as it may, the Random Forest (RF) algorithm is one such algorithm that remains very close to the highest point of the classifier progressive system. The proposed supervised machine learning model has been developed with the following assumptions: The bias and variance error components are mostly present in every model. Bias and variance are inversely related to each other and hence while trying to reduce one component, the other component of the model will increase. The true art lies in creating a good fit by balancing both, wherein the ideal model will have both low bias and low variance. Errors from the bias component come from erroneous assumptions made in the underlying learning algorithm but being a nonparametric model, Random Forest (RF) can handle skewed and multimodal data. It is robust to outliers and works well for nonlinear data by having a low risk of overfitting. Moreover, it runs efficiently on large datasets. Since decision trees form a tree like structure and remain as the building blocks of a Random Forest (RF) algorithm, the factors that comes under decision tree principle has been emphasized initially.

4. PROPOSED MODEL:

In our analysis, we have opted for the Random Forest Classifier as our model of choice. The Random Forest algorithm is classified as an ensemble learning technique, renowned for its efficacy in constructing numerous decision trees during training and deriving predictions based on the mode of classes for classification tasks or mean prediction of individual trees for regression.

Several compelling factors guided our selection of the Random Forest approach for this particular task:

- **Accommodating Large Datasets:** Random Forest exhibits remarkable efficiency in handling sizable datasets characterized by high dimensionality. Given our dataset's substantial volume comprising numerous rows and diverse features, this attribute aligns seamlessly with our analytical requirements.
- **Mitigation of Overfitting:** Overfitting poses a prevalent concern in decision tree-based models. However, Random Forest mitigates this risk by aggregating predictions from multiple decision trees, thereby averting the tendency towards excessive model complexity often associated with individual decision trees.
- **Versatility in Data Types:** Our dataset encompasses a blend of numerical and categorical features. Random Forest excels in accommodating such diverse data types, rendering it a pragmatic choice for our analytical endeavor.
- **Facilitation of Feature Importance Analysis:** An invaluable characteristic of Random Forest is its ability to elucidate feature importance comprehensively. Given our objective to explore the influence of various factors on diabetes risk, this feature empowers us to discern the relative significance of each predictor variable.
- **Capacity to Capture Non-linear Relationships:** Medical data frequently exhibits intricate, non-linear associations between variables. Leveraging its inherent non-linear modeling capabilities, Random Forest excels in capturing and discerning these complex relationships, thereby enhancing the interpretability and predictive accuracy of our analyses.

4.1 Data collection and pre-processing:

Collecting a dataset for diabetes prediction involves gathering diverse health-related information from sources like electronic health records and clinical databases. This includes variables such as age, gender, BMI, blood pressure, cholesterol levels, family history, lifestyle factors, and medication history. Ensuring data quality and adhering to ethical guidelines are essential aspects of the process. Careful attention to these factors lays the groundwork for developing accurate and reliable predictive models to inform healthcare decisions.

Three important measures to be taken before preprocessing:

- Handling Duplicates
- Uniqueness
- Missing Values

Preprocessing is an essential step in machine learning workflows as it prepares the data for model training, ensuring that it is in a format that the model can effectively interpret and learn from. When dealing with numerical features, standardization is often employed to transform the data so that it has a mean of 0 and a standard deviation of 1. This process, typically achieved using techniques like the StandardScaler in scikit-learn, is beneficial for models that assume a Gaussian distribution of the features. However, even for models that do not strictly require standardized input, standardization can still improve convergence rates and overall performance by making the optimization process more efficient and less sensitive to the scale of the features.

On the other hand, categorical features require a different preprocessing approach. Since machine learning models operate on numerical data, categorical variables need to be transformed into a numerical format. One common technique for achieving this is one-hot encoding, where each categorical variable is converted into a binary vector representation, with each category represented as a binary feature. This ensures that the model can properly interpret and learn from the categorical data, without assigning any ordinal relationship between categories. By performing one-hot encoding, categorical variables are effectively expanded into a series of binary features, each representing the presence or absence of a specific category, thereby preserving the information encapsulated within the categorical data.

4.2 Dataset Analysis:

4.2.1 Bivariate Analysis:

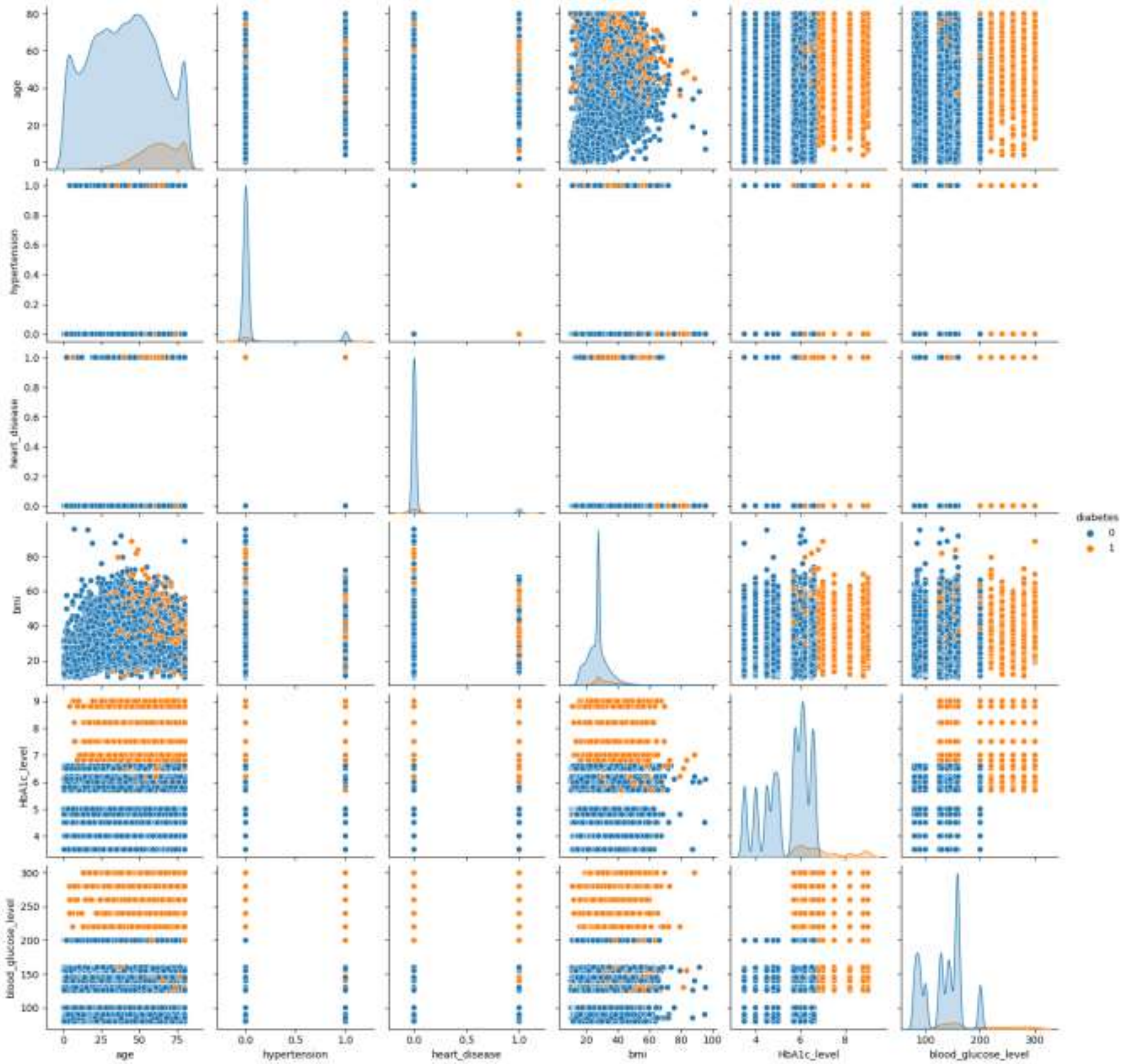

Fig -1: Pair Plot for numeric features

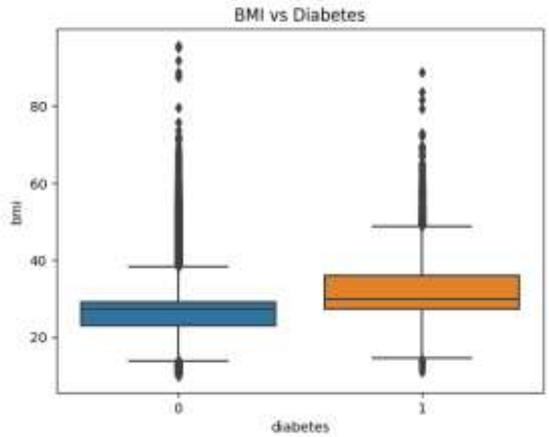

Fig -2 : Boxplot BMI vs Diabetes classification

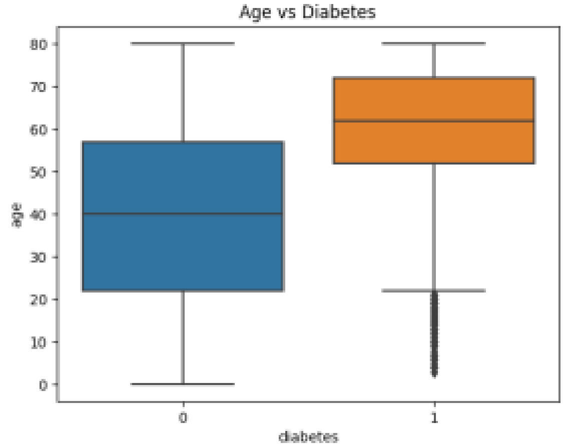

Fig -3 : Boxplot Age vs Diabetes classification

4.2.2 Multivariate Analysis:

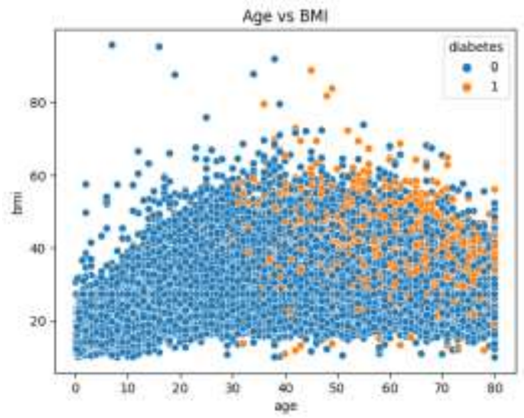

Fig -4: Scatter plot Age vs BMI colored by diabetes classification

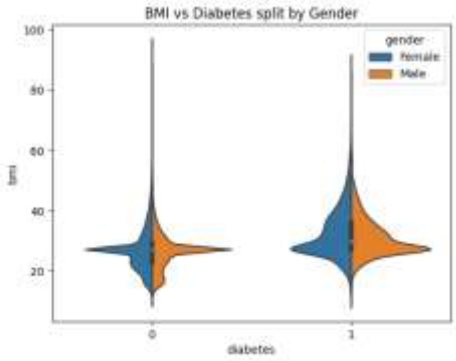

Fig -5: Violin plot of BMI vs Diabetes split by gender

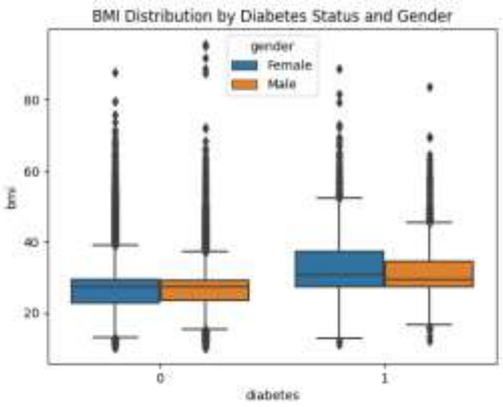

Fig -6: Interaction between gender, BMI, and diabetes

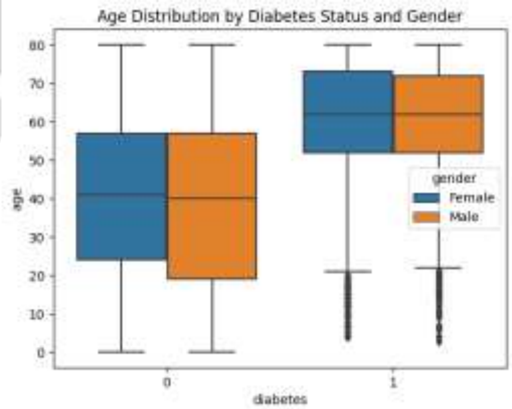

Fig -7: Interaction between gender, age, and diabetes

4.2.3 Correlation Matrix

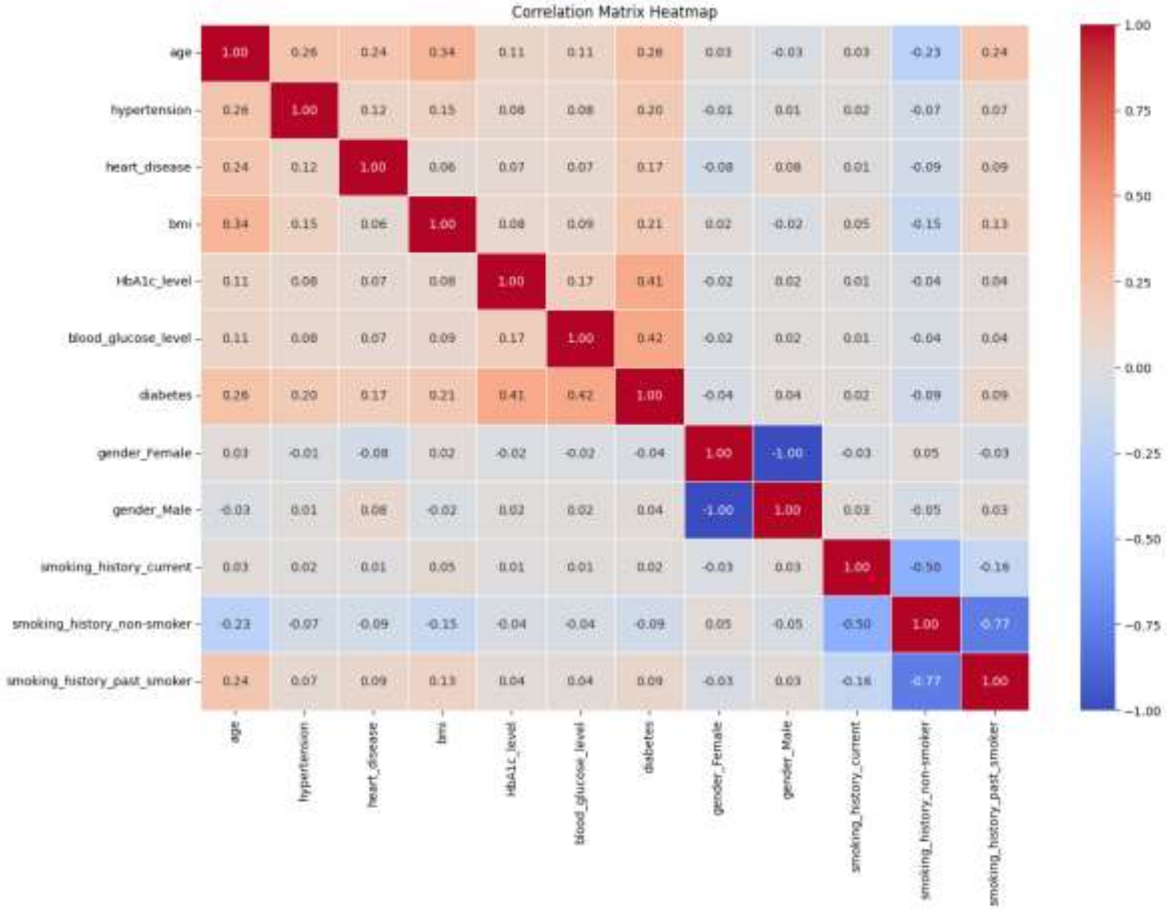

Fig -8: Correlation Matrix heatmap

4.3 Model Training and Testing:

To train a Random Forest model for predicting diabetes, the following steps are typically followed:

- **Data Preparation:** Prepare a dataset containing relevant features (such as age, BMI, blood pressure, etc.) and the corresponding target variable indicating the presence or absence of diabetes. Ensure the dataset is properly formatted and free from missing values.
- **Splitting Data:** Divide the dataset into training and testing sets. The training set is used to train the model, while the testing set is used to evaluate its performance.
- **Feature Selection:** Identify the most informative features for predicting diabetes. This may involve conducting exploratory data analysis and feature engineering to select relevant predictors.
- **Model Training:** Train the Random Forest model using the training data. The model will learn patterns and relationships between the features and the target variable. Adjust hyperparameters such as the number of trees, tree depth, and minimum samples per leaf to optimize model performance.
- **Model Evaluation:** Evaluate the trained model using the testing set. Calculate metrics such as accuracy, precision, recall, and F1-score to assess its performance in predicting diabetes.
- **Tuning and Validation:** Fine-tune the model parameters using techniques like cross-validation to ensure robustness and prevent overfitting. Validate the model on additional datasets if available to confirm its generalizability.
- **Deployment:** Once the model demonstrates satisfactory performance, it can be deployed for predicting diabetes in new data. This may involve integrating the model into a healthcare system or application for real-time predictions.

- **Monitoring and Updating:** Continuously monitor the model's performance and update it as necessary with new data or changes in healthcare practices to ensure ongoing accuracy and relevance.

By following these steps, we can effectively train a Random Forest model for predicting diabetes and deploy it for practical use in healthcare settings.

In addition to these steps, it's essential to document the process thoroughly and maintain transparency in model development. This includes documenting any data preprocessing steps, such as handling missing values or scaling features, to ensure reproducibility. Moreover, considering the ethical implications of deploying a predictive model in healthcare settings is crucial. This involves addressing issues related to data privacy, patient consent, and potential biases in the dataset. Furthermore, ongoing monitoring and evaluation of the deployed model are necessary to assess its performance in real-world scenarios and identify areas for improvement. Incorporating these considerations ensures responsible and effective deployment of the Random Forest model for predicting diabetes in healthcare settings.

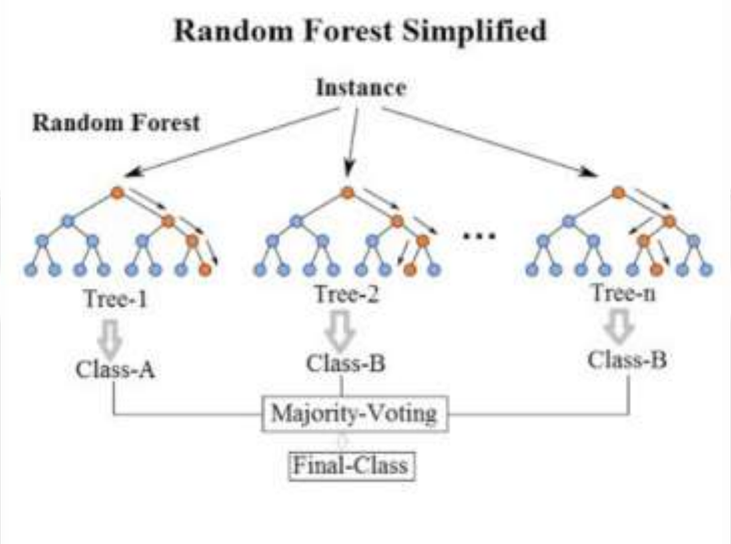

Fig -9: Random Forest Simplified

5. RESULTS:

After training, our Random Forest Model exhibited an impressive accuracy rate of approximately 95%, indicating its ability to correctly classify the majority of cases in the test set.

Further examination of the classification metrics unveils nuanced performance insights for each class (0 and 1) within the dataset:

A | Class 0 (Non-diabetes):

For class 0, the model demonstrates a commendable precision of 0.98. This signifies that among instances predicted as non-diabetes, 98% were accurately classified.

The recall for class 0 is also noteworthy at 0.96, indicating the model's capability to correctly identify 96% of actual non-diabetes cases.

B | Class 1 (Diabetes):

In contrast, the precision for class 1 stands at a relatively lower level of approximately 0.65, suggesting that around 65% of predictions for diabetes were correct.

However, the recall for class 1 remains reasonably high at approximately 0.80, signifying the model's ability to capture around 80% of all actual diabetes cases.

The F1 score, which combines precision and recall, measures around 0.97 for class 0 and approximately 0.72 for class 1. The weighted average F1 score aligns closely with the overall accuracy, hovering around 0.94.

This observed performance disparity between classes likely stems from the inherent dataset imbalance. Class 0 (Non-diabetes) represents the majority class, offering more instances for the model to learn from and potentially leading to greater predictive accuracy in this category.

	Precision	Recall	F1-score	Support
0	0.98	0.96	0.97	17525
1	0.68	0.80	0.74	1701
Accuracy			0.95	19226
Micro avg	0.83	0.88	0.85	19226
Weighted avg	0.95	0.95	0.95	19226

Table 1 : Classification report

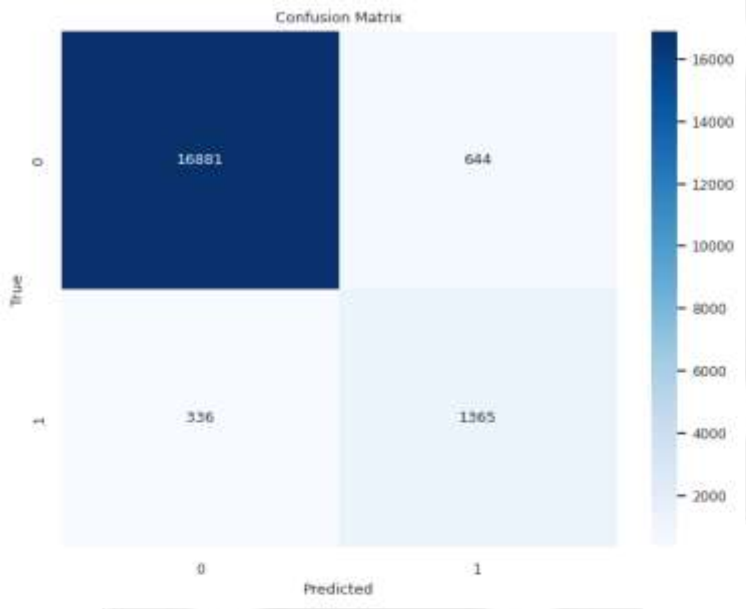

Fig -10: Confusion Matrix

6. CONCLUSIONS:

Utilizing a Random Forest classifier, the analysis aimed to predict diabetes based on a range of health indicators and lifestyle factors. The model underwent training and evaluation on a dataset comprising 100,000 records, with Hyperparameter tuning conducted to optimize performance. Achieving an accuracy of approximately 95%, the model demonstrated precision rates of 0.98 for non-diabetic (class 0) and 0.69 for diabetic (class 1) instances. Notably, it also exhibited a recall of 96% for non-diabetic cases and 81% for diabetic cases, showcasing a well-tuned and balanced performance across both classes. Feature importance analysis identified HbA1c_level and blood glucose level as critical predictors of diabetes, followed by age and BMI. Conversely, factors like smoking history and gender had minimal influence on model predictions.

7. REFERENCES

[1]. Chou, Chun-Yang, Ding-Yang Hsu, and Chun-Hung Chou."Predicting the Onset of Diabetes with Machine Learning Methods." Journal of Personalized Medicine 13, no. 3 (2023): 406.

[2]. A. Misra, H. Gopalan, R. Jayawardena, A. P. Hills, M. Soares, A. A. Reza-Albarrán, and K. L. Ramaiya, “Diabetes in developing countries," J. Diabetes, vol. 11, no. 7, pp. 522_539, Mar. 2019.

[3]. Febrian, Muhammad Exell, Fransiskus Xaverius Ferdinan,Gustian Paul Sendani, Kristien Margi Suryanigrum, and Rezki Yunanda. "Diabetes prediction using supervised machine learning." Procedia Computer Science 216 (2023): 21-30.

[4]. Sarwar, Muhammad Azeem, Nasir Kamal, Wajeeha Hamid,and Munam Ali Shah. "Prediction of diabetes using machine learning algorithms in healthcare." In 2018 24th international conference on automation and computing (ICAC), pp. 1-6. IEEE, 2018.

[5]. Dutta, Debadri, Debpriyo Paul, and ParthajeetGhosh."Analysing feature importances for diabetes prediction using machine learning." In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 924- 928. IEEE, 2018.