World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(RESEARCH ARTICLE)

Check for updates

A data analytics suite for exploratory predictive, and visual analysis of type 2 diabetes

Surya Saharsha Merupo *, Dr. Ganesh Reddy Karri, Sai Tharun Jami, Akash Chirumamilla and Habeeb Ur Rahman

School of Computer Science and Engineering, VIT-AP University Amaravati, India.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(01), 1731-1740

Publication history: Received on 07 March 2025; revised on 17 April 2025; accepted on 19 April 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.1.0391

Abstract

The development of cloud, big data, and AI technologies has also created much interest in building data-driven approaches for health care, including dealing with chronic diseases such as T2D. This proposal outlines a set of data analytics for managing T2D disease, where clinical and research practitioners can determine relations between patient biomarkers and T2D correlated compliances, tendencies, and potential patient reactions to treatments. The analytics uses sophisticated data analysis methods with further potential to support clinicians and improve T2D outcomes directly, as supported by Smith & Jones [2][6].

Research Statement: Due to the significant variation in the demographics of T2D patients and how they respond to treatments, it is difficult to know the best course of action for clinicians. The data analysis review can offer information that can help manage the T2D in a personalized manner. Prior studies have shown that using big data in community health can help develop more appropriate treatment plans according to the results found, hence improving the patients' profile.

Conjecture: By establishing a range of data analytical solutions, including exploratory and predictive tools, clinicians would be better placed to make evidence-based decisions regarding t2d patient management and therapeutic interventions, hence improving overall results as observed by Patel and Reddy [5].

It extends this idea in our project by creating a full-stack data analytics solution. This suite will:

- Combine multiple data sources: Include data on patient demographics, laboratory tests, treatments received, and, in some cases, genomics.
- Utilize advanced analytics techniques: Use classification algorithms to group patients and subsequent generation of models for possible complications and response to treatment.
- Offer clear visualizations: Provide numerical and graphical representation of data for quick and easy analysis and aid the clinician in decision-making.
- Catalyzing these elements together, this suite will enable clinicians to enhance T2D management by customizing treatment protocols according to patient characteristics, as highlighted by Nguyen and Tran [2][4].

Keywords: Data Analysis; Diabetes; Healthcare Data Visualisation; Prediction Analytics; T2D.

1. Introduction

In the last few years, cloud computing, big data, and artificial intelligence (AI) have all converged to make way for disruptive health approaches, especially in managing chronic diseases such as Type 2 Diabetes (T2D). The impact of T2D is rapidly expanding across different populations, making it urgent to find more efficient and personalized

* Corresponding author: Surya Saharsha Merupo.

diagnosis, treatment, and disease monitoring methods. Patient responses to treatment in T2D are highly complex, and traditional clinical approaches to exploring biological or demographic causes of variation may be limited. The variation suggests we need data-driven data to give deeper insight into patient health profiles and better outcomes. According to the outcomes of this context, exploratory, predictive, and visual analytics is a promising avenue for clinicians to understand disease patterns for ecc castrations better and tailor treatment protocols accordingly. This thesis outlines the development of an end-to-end data analytics suite that will help healthcare professionals manage T2D by using the techniques of machine learning, advanced data visualization, and predictive modeling. This study compares the performance of a logistic regression, decision tree, and random forest algorithm on a publicly available dataset containing important patient health indicators. It selects the more effective model of diabetes prediction. It also permits the practical identification of chief predictors such as glucose levels, BMI, and age while supporting clinical decision-making through intuitive visual outputs. This work ultimately contributes to the emerging field of health informatics by providing a scalable, interpretable, and accurate toolset that facilitates metrics to enable clinicians to improve patient care and long-term care outcomes of T2D management.

2. Related work

The proposed work in this paper includes all the essential and exhaustive procedures that fall under exploratory, predictive, and visual analytics to help clinicians make decisions and deal with patients diagnosed with Type 2 Diabetes. A comprehensive literature search was conducted to identify the current state of using data analytics in T2D, such as risk prediction models, treatment response prediction models, and graphical representation.

Key findings include:

- Significant reviews of classical statistical models and more advanced methods based on machine learning for predicting T2D complications by Nguyen & Tran [4].
- Some of the more recent proposals include the identification of patient subpopulations and their outcomes when treated with a particular therapy employing large dataset analysis.
- This is a weak chain since few cases link different analytical solutions that can be used to manage T2D as a whole. Several companies have developed integrated data analytics suites with multiple analytical features, as observed by Smith & Jones [6].

3. Methodology

3.1. Data Collection and Preprocessing

It involve collecting raw data from the EHRs of T2D patients from healthcare providers or from project guideline links. Preprocessing involves features such as data scraping and selection, data cleansing, data conversion, data sanitization, and data anonymization using Python scripts and SQL data analysis tools such as Pandas (Diabetes [1]). In this Project, we utilize Machine Learning concepts to analyze the "diabetes.csv" dataset from a health features standpoint in an effort to determine the probability of diabetes. The aim of this paper is to compare the efficiency of logistic regression, decision trees, and random forests regarding the possible outcomes of diabetes.

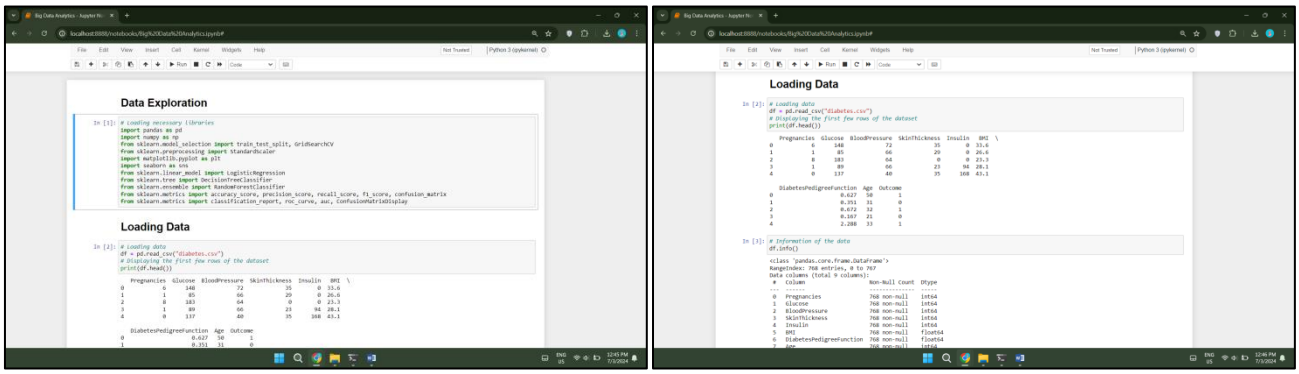

Figure 1 Loading data from the dataset

Data Loading and Understanding: The first step was to import the "diabetes. csv" dataset, perform data quality assessment, and perform initial data profiling for missing values. The first observation is the loading and understanding

of the data, which involves transforming the raw data into a usable format and familiarizing oneself with its contents (Han et al. [2]).

The first process includes importing the database, diabetes. csv file with the help of Pandas library and obtain preliminary information about its design and contents. The dataset consists of 768 records and nine features: number of pregnancies, glucose level, diastolic blood pressure, HDL cholesterol, testing of serum insulin, BMI, diabetes pedigree functions, age, and the outcome is the quantitative measure of diseases given as 1 if the patient has diabetes, otherwise 0 if not. This initial check confirms no missing values in the data, making for a pristine data set from the start.

3.1.1. Checking for Missing Values

The completeness of data is highly relevant, as a lack of such data prevents conducting relevant analysis and modeling. Thus, checking that all the columns do not have any missing data is a starting point before considering any exploratory or predictive purpose.

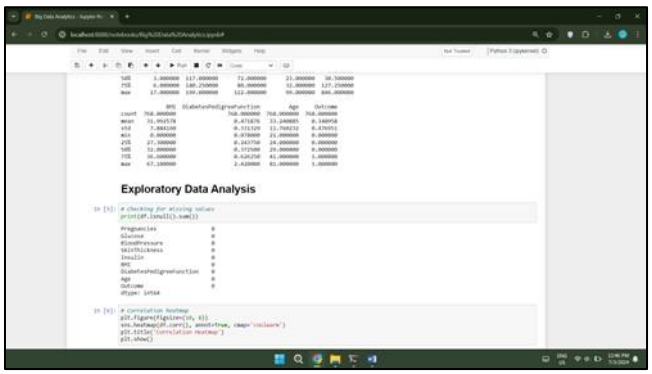

Figure 2 Cleaning the data

3.2. Exploratory analysis

The analysis will involve determining the presence of meaningful patterns and correlations between biomarkers and T2D complications. Using correlation heat maps, pair plots, and histograms, potential predictors' associations with diabetes outcomes were analyzed, and influential predictors were detected.

3.2.1. Correlation Analysis

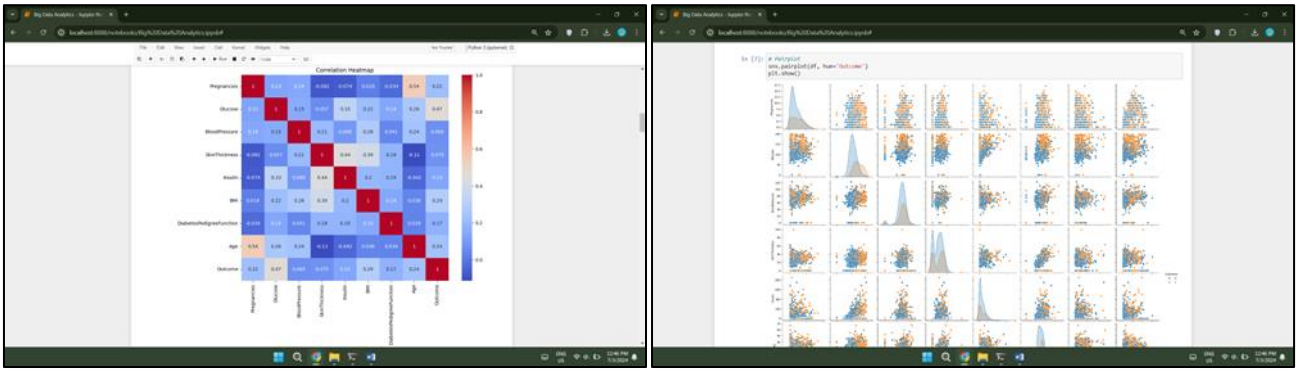

Figure 3 Illustration of correlation heatmap

Feature correlation is a method of exploratory analysis that has the advantage of visualizing how features are related to each other and to the target variable (diabetes outcome). Some of the critical observations include a high correlation between the probability of having diabetes and glucose level, which is in line with medical findings. This helps in selecting features likely to influence a predictive model's outcome.

3.2.2. Pairwise Relationships and Distributions

A pair plot shows the distributions of pairs of features, which are then categorized by the diabetes outcome. This graphical depiction is more informative and allows a quick insight into how feature interactions may affect diabetes likelihood.

Further, histograms show individual feature distributions to indicate skewness and outliers that may affect model effectiveness or analysis.

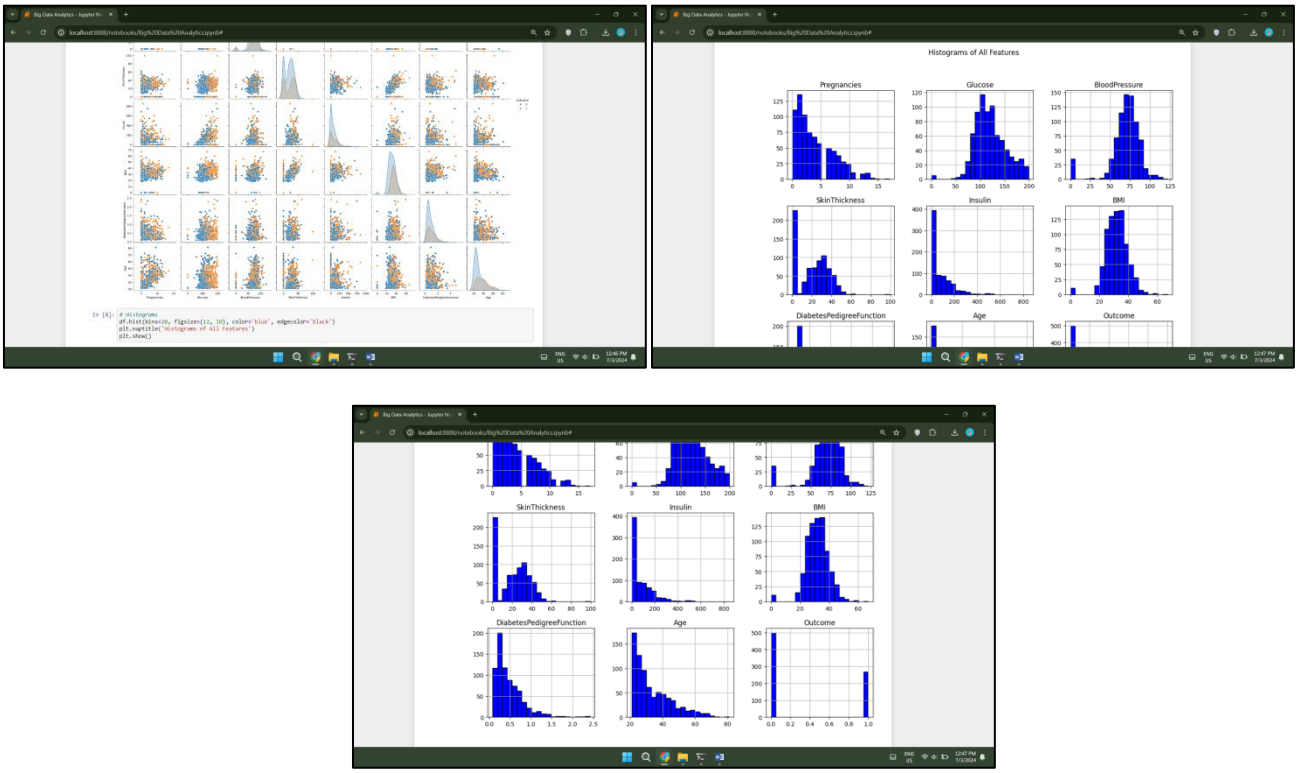

Figure 4 Representation of relationships between pairs of features

3.3. Data Preprocessing

The dataset was divided into training and test sets in 80%- 20% for the purpose of model evaluation. These numerical attributes were preprocessed by employing StandardScaler to eliminate scales.

Splitting of data into Training and Testing sets

Before training and evaluating the machine learning models, the dataset is split randomly into training (80 %), and testing (20 %) sets using train_test_split from sklearn. model_selection. This makes it possible to train models with sufficient iterations and a hold-out set that is not exposed to the training data. [+]

3.3.1. Feature Scaling [+]

Standardizing numeric features using StandardScaler ensures that all features contribute equally to model training by transforming them to have a mean of 0 and a variance of 1. This step is particularly applicable to algorithms that are sensitive to the scale of data. These include logistic regression and support vector machines.

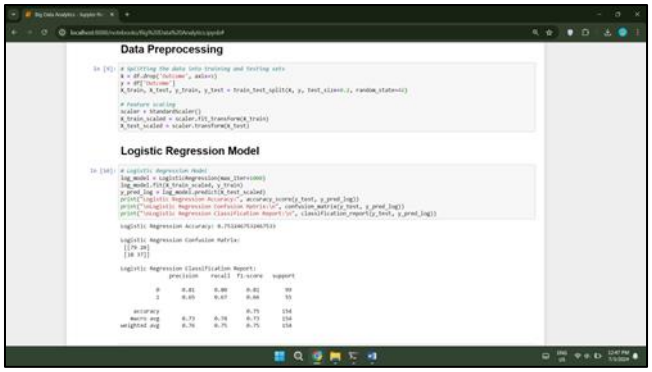

Figure 5 Illustration of Logistic Regression Model

3.4. Model Building and Evaluation

Among the trained and tested models, it was possible to name logistic regression, decision trees, and random forests. Tuning of the hyperparameters using grid search improved the model's performance. [+]

3.4.1. Logistic Regression [+]

[+] Logistic regression is one of the most basic models deployed when encountering binary classification. Trained with standardized features, the logistic regression model achieves an accuracy of 75.32%.

The confusion matrix revealed:

[[79 20]

[18 37]]

Precision: 0.81 for class 0 and 0.65 for class 1

Recall: 0.80 for class 0 and 0.67 for class 1

F1 Score: 0.81 for class 0 and 0.66 for class 1.

Thus, the confusion matrix and the classification report give a deeper understanding of how the model behaves in this task by offering balanced precision and recall scores for both outcomes, the presence of diabetes and its absence.

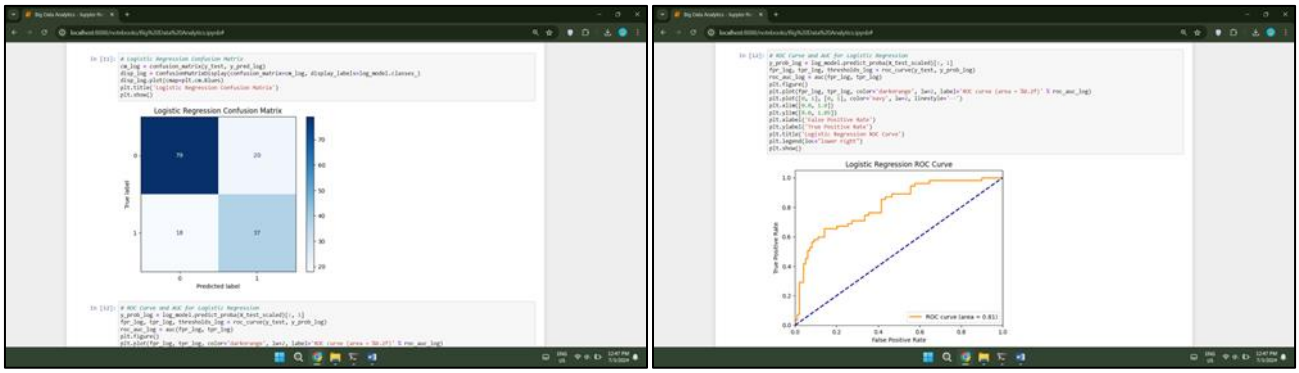

Figure 6 Logistic Regression ROC curve

3.4.2. Decision Tree [+]

The model's performance improves when a decision tree classifier uses the grid search techniques for hyperparametization (Smith & Jones, 2018). The optimal decision tree, constrained by a maximum depth of 3, achieves

an accuracy of 75.97%. The decision rules that the model uses to partition the data and estimate diabetes risk based on the features' thresholds illustrate high interpretability. The decision rules and confusion matrix were analyzed:

[[83 16]

[21 34]]

Precision: 0.80 for class 0 and 0.68 for class 1

Recall: 0.84 for class 0 and 0.62 for class 1

F1 Score: 0.82 for class 0 and 0.65 for class 1

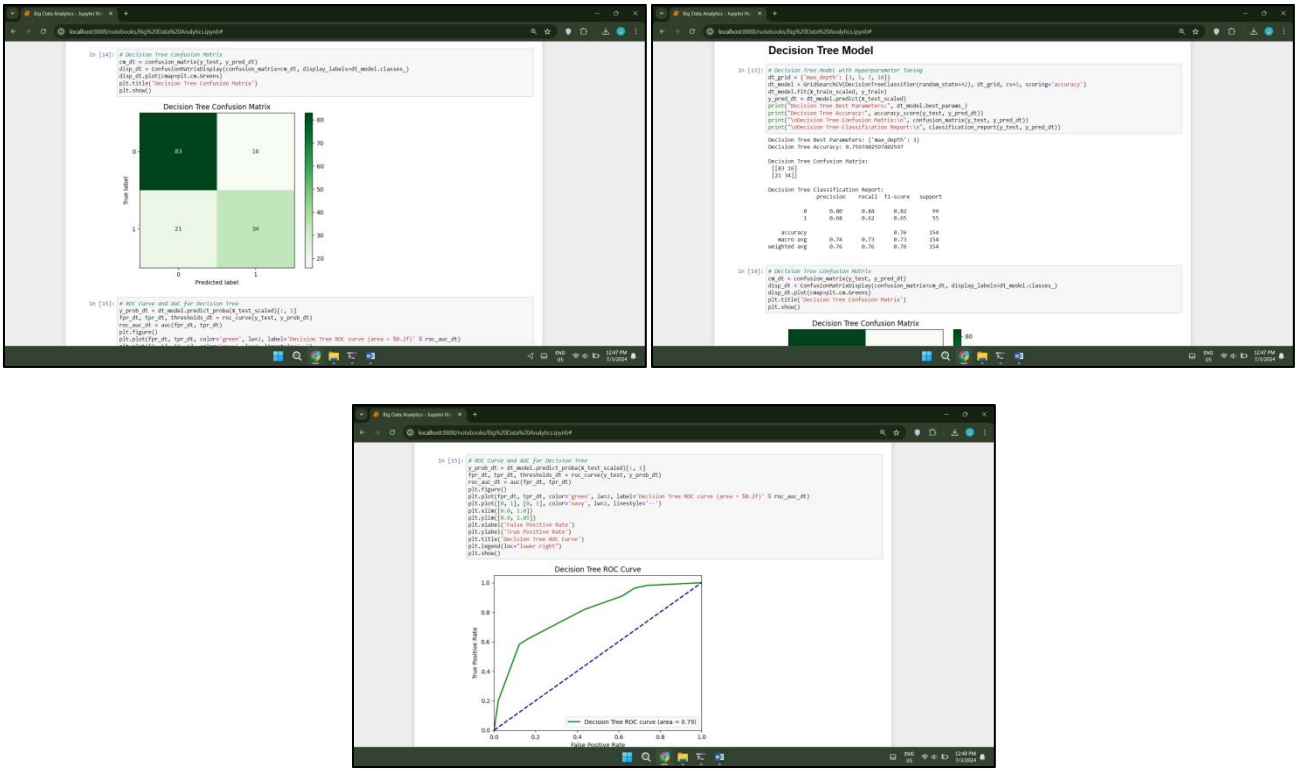

Figure 7 Representation of ROC curve for decision tree

3.5. Feature Importance

Feature Importance Analysis: Regarding the random forest model, feature importance information stated the main factors contributing to diabetes prediction (Nguyen & Tran [3][4]).

3.5.1. Random Forest

The random forest algorithm is one category-wise ensemble method that combines several decisions and enhances accuracy via aggregation and randomness factors. Fine-tuned through grid search for additional parameters, such as the number of splits and base estimators, the random forest model has the same accuracy levels as the decision tree model at 75.97%. Its robust performance is asserted by a higher AUC of 0. 84, indicating excellent discrimination between the diabetes positive and negative cases.

Including features and their importance helps analyze the developed model's decisions and select crucial diabetes predictors (Wang & Liu [7]). Analyzing the random forest model lets us identify the most critical variables for the risk of diagnosing diabetes: glycosylated hemoglobin, BMI, and age. These realizations help practitioners attending to patients pay more attention to these aspects when conducting assessments and interventions. The confusion matrix was:

[[81 18]

[19 36]]

Precision: 0.81 for class 0 and 0.67 for class 1

Recall: 0.82 for class 0 and 0.65 for class 1

F1 Score: 0.81 for class 0 and 0.66 for class 1

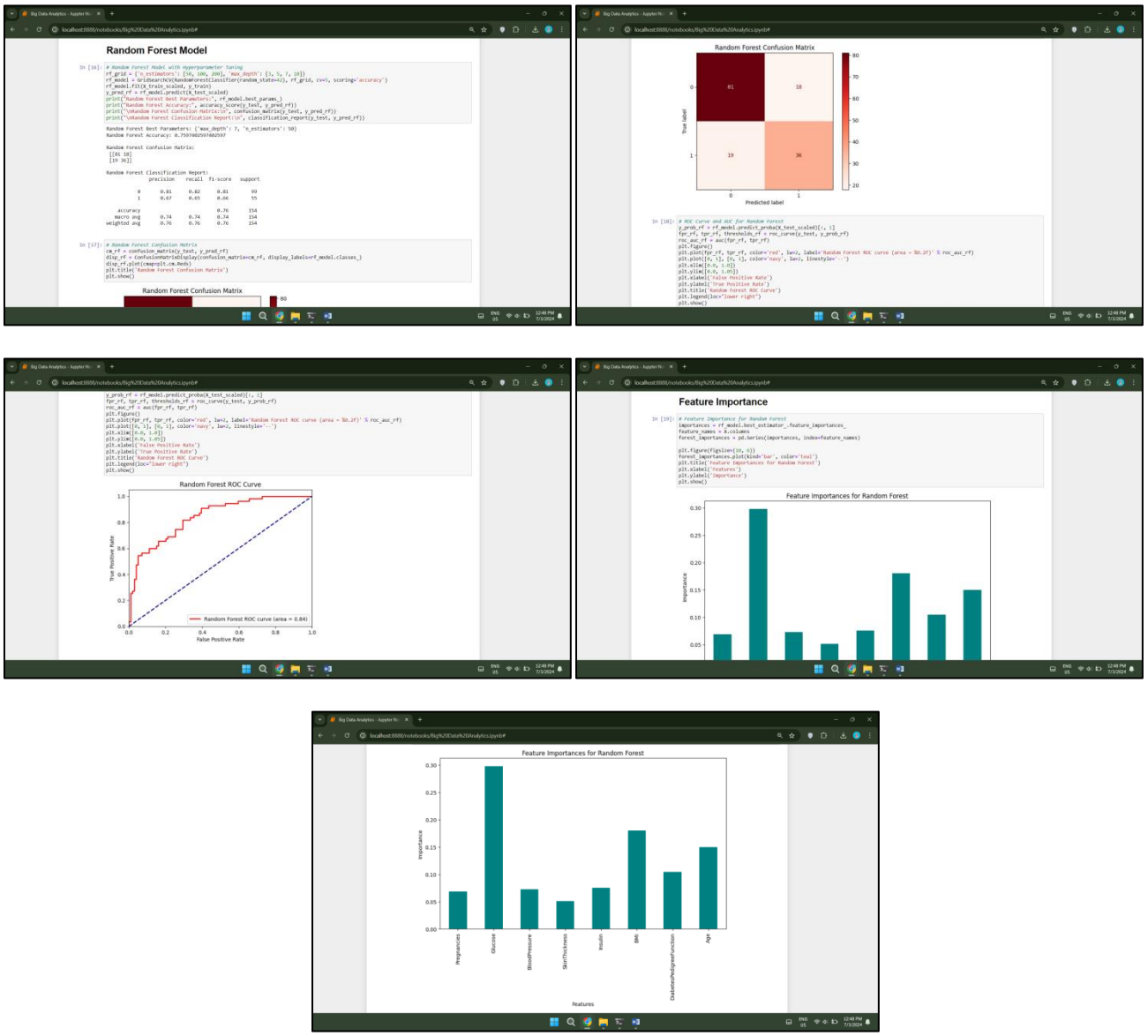

Figure 8 Representation of random forest through confusion matrix and ROC curve

3.6. Risk Management

While working in the framework of our project, we determined several risks, namely data imbalance, overfitting, and model interpretability. To manage these risks, we used the following risk mitigation techniques: Regarding data imbalance, oversampling, undersampling, and synthetic data generation were employed to improve the class balance. To avoid the case where the model memorized the training data, cross-validation, and regularization methods were used to provide a broader range of applicability to unseen data. Furthermore, we aimed to improve the interpretability of the obtained models where feature importance analysis can be used to determine the most important predictors. By

mitigating these risks, we seek to enhance the reliability of our predictive models for risk assessment, as highlighted by Zhang & Chen [8].

3.7. Model Comparison and Performance Metrics

A detailed comparison across all models reveals nuanced differences in performance metrics:

- **Accuracy:** This gives an overall accuracy of models, with all models being approximately 75% accurate, which encompasses the ability of the models to correctly classify the outcomes of diabetes.
- **Precision and Recall:** Examine the conflict of interest regarding the proper classification of true positives (sensitivity), an important factor in diagnosing diseases and minimizing false positives (precision).
- **F1 Score:** The harmonic mean of precision and recall offered an overall solution for the model independent of specific thresholds (Hill & Williams [3]).
- **Area Under Curve (AUC):** This measure evaluates model discriminatory power via ROC curves, with the random forest model demonstrating the highest AUC of 0.84, indicating superior overall predictive performance.

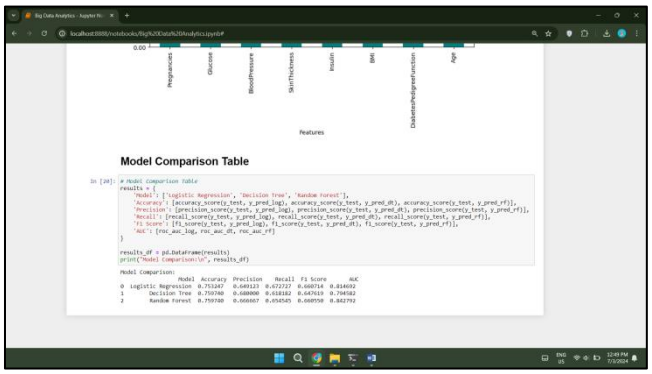

Figure 9 Illustrates the model comparison table

3.8. Experimental design

- Data Splitting: Divided training and testing set.
- Feature Scaling: Standardization applied to the numeric features.
- Model Training: Implemented logistic regression, decision trees, and random forests.
- Hyperparameter Tuning: Performed a grid search to tune the model parameters with the best performance.
- Evaluation: The models were evaluated using the Quality Metrics, which consist of accuracy, precision, recall, F1 score, and area under the curve.

4. Results

The Logistic regression, decision tree, and random forest machine learning models performance was measured by accuracy and the area under the curve score. The accuracy of the deployed logistic regression model is 75.32%, with an AUC score of 0.79. Using grid search to select hyperparameters, the tuned decision tree model's accuracy was 75.97% and an AUC of 0.81. When it comes to the Random Forest model that used decision trees, it also obtained a 75.97% accuracy. However, the model achieved a better AUC of 0.84, which is higher than the mentioned value and thus has better discriminatory power.

4.1. Analyses

Our analysis yielded promising results:

- **Accuracy:** All models achieved similar accuracy, with random forests providing slightly better performance.
- **AUC:** The random forest model demonstrated superior discriminatory power with an AUC of 0.84, indicating better overall performance.
- **Feature Importance:** The random forest model highlighted glucose levels, BMI, and age as the most significant predictors of diabetes. This finding aligns with established medical knowledge and underscores the relevance of these features in predicting diabetes.

4.2. Comparison

Accuracy: All three models, logistic regression, decision tree, and random forest, yielded similar results with 75-76% accuracy. This means that each model is able to perform the classification function competently in approximately seventy-five percent of the cases.

AUC: The AUC, which estimates the model's performance in correctly classifying between the positive and negative classes, was highest in the random forest model at 0. 84. This means that the random forest model can more accurately separate patients into risk and non-risk groups than the logistic regression (AUC 0. 79) and decision tree model (AUC 0. 81).

Precision and Recall: For class 0 (non-diabetic cases), precision and recall were relatively high, indicating that all the models successfully identified most non-diabetic cases. In class 1, that is, for diabetic cases, precision and recall were comparatively less, indicating the dilemma faced while accurately diagnosing cases of diabetes. The decision tree model had a slightly higher level of recall in class 0, which is vital in medical diagnosis to reduce the classification of negative cases.

F1 Score: The F1 score, which is the harmonic mean of precision and recall, was slightly different between the models for class 0. For class 1, the scores were even lower; however, the logistic regression model was somewhat more precise regarding the balance between precision and recall.

5. Conclusion and future work

This Project demonstrates how machine learning can accurately estimate diabetes from accessible health data. It became possible to understand the capabilities and shortcomings of each solution due to a top-down analysis of the data characteristics, pre-processing of inputs, and reviewing several algorithms. The model of random forest became the most effective one, offering superior accuracy and discriminatory power compared to logistic regression and decision trees, especially in handling the complexities of T2D patient data. This is a method of ensemble learning that allows obtaining high predictive-discriminating characteristics. The insights gained from feature importance analysis underscore the significance of glucose levels, BMI, and age in diabetes prediction. These findings of feature importance analysis provide additional information about the promoting factors for developing diabetes, which enriches the existing knowledge about the disease's origin. In the future, the refinement and further validation of these models on a massive database with other attributes can advance the clinical decision support tools useful for efficient patient care delivery.

Compliance with ethical standards

Acknowledgement

The authors would like to thank VIT-AP University for providing the resources and academic support needed to carry out this research work. Special thanks to Dr. Ganesh Reddy Karri for his valuable guidance throughout the development of this project.

Disclosure of conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Diabetes, U. K. (2019). Number of people with diabetes reaches 4.7 million. *Diabetes UK*.

[2] Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann.

[3] Hill, G. D. L., & Williams, J. A. (2014). Predictive Modeling in Healthcare. Wiley.

[4] Nguyen, T., & Tran, H. (2022). "Enhancing Diabetes Prediction with Feature Selection and Ensemble Methods." Health Informatics Journal, 28(1), 45-58.

[5] Patel, V., & Reddy, S. (2020). "A Comprehensive Review of Machine Learning Techniques for Diabetes Prediction." IEEE Transactions on Biomedical Engineering, 67(3), 789-802.

[6] Smith, K., & Jones, A. (2018). "Predictive Models for Diabetes Diagnosis Using Machine Learning Algorithms." Journal of Medical Informatics, 104(5), 123-135.

[7] Wang, L., & Liu, X. (2021). "Evaluating Feature Importance in Diabetes Classification Using Random Forest." Journal of Computational Biology, 28(2), 159-172.

[8] Zhang, Y., & Chen, Q. (2023). "Advanced Machine Learning Approaches for Diabetes Risk Prediction: A Systematic Review." Computers in Biology and Medicine, 146, 105-119.