EasyChair Preprint

№ 15225

Proactive Health Monitoring: Predictive Analytics for Early Detection of Diabetes Risk

Ashish Mahindre and Sarika Kondekar

October 9, 2024

Proactive Health Monitoring: Predictive Analytics for Early Detection of Diabetes Risk

Mr. Ashish A. Mahindre[1], Dr. Sarika A. Kondekar[2]

[1]Research Scholar, SOCSE Sandip University, Nashik
Ashish.mahindre@gmail.com
Department of Computer Science & Application, Sandip University, Nashik

[2]Assistant Professor, SOCSE Sandip University, Nashik
Sarika.kondekar@gmail.com
Department of Computer Science & Application, Sandip University, Nashik

Abstract

This research introduces an advanced predictive analytics framework for the early detection of diabetes risk, aiming to enhance proactive health monitoring through the integration of sophisticated machine learning algorithms. The model is meticulously trained on a diverse set of patient health metrics, including demographic and clinical variables such as age, body mass index, blood pressure, and glucose levels. By identifying subtle patterns and correlations within the data, the model facilitates the early identification of individuals at high risk of developing diabetes. This early detection capability enables timely clinical interventions, potentially mitigating the progression of the disease and optimizing patient management strategies. The study underscores the model's robustness and scalability, highlighting its significant potential for deployment in clinical settings as a critical component of preventive healthcare infrastructure.

Keywords: Predictive Analytics, Diabetes Risk Prediction, Early Disease Detection, Machine Learning in Healthcare, Patient Data Analysis, Clinical Decision Support

1. Introduction

A chronic & progressive metabolic disease, diabetes mellitus has emerged as one of the major worldwide health concerns. Characterized by elevated blood glucose levels, diabetes affects millions of individuals, posing substantial health risks and economic burdens. According to estimations from the World Health Organization (WHO), diabetes ranks seventh globally in terms of causes of mortality, and its prevalence has been alarmingly rising. This rise is primarily driven by global trends such as urbanization, sedentary lifestyles, poor dietary habits, and aging populations. There are two primary forms of diabetes: Types 1 and 2. Type 1 diabetes is an autoimmune illness that results in the destruction of the pancreatic beta cells responsible for producing insulin, leading to absolute insulin insufficiency. Insulin resistance and relative insulin insufficiency define type 2, the more common form of diabetes. The latter is often associated with obesity, physical inactivity, and genetic predisposition. Both types of diabetes lead to chronic hyperglycaemia, which can cause a range of consequences such as retinopathy, neuropathy, nephropathy, and cardiovascular disease.

Pharmacotherapy, frequent monitoring, and lifestyle changes are commonly used in the management of diabetes of blood glucose levels. However, the conventional approach to diabetes care is largely reactive, focusing on treatment and management after the disease has manifested. This model is less effective in preventing the onset of diabetes and mitigating its long-term complications. Given the chronic nature of diabetes and its potential to cause serious health issues, there is a critical need for proactive strategies that enable early identification and intervention. Statistical methods and machine learning applications approaches to data analysis, known as predictive analytics, has shown promise in the prevention and treatment of illness. Predictive models use both historical and current data to anticipate future outcomes and identify individuals at risk of developing certain conditions. In the context of diabetes, predictive analytics offers the potential to enhance early detection and

intervention, thereby preventing the onset of the disease or managing it more effectively in its initial stages.

Machine learning, a subset of Predictive analytics heavily relies on artificial intelligence (AI). It involves the use of algorithms that do not require explicit programming and are ability to absorb knowledge from data, spotting patterns, and making judgments or predictions. Large, complex datasets with lots of variables can be processed by machine learning models, which thereafter may be used to find patterns that conventional analytical techniques might miss. Machine learning has proven to be effective in a number of healthcare domains in recent years, including risk prediction, disease diagnosis, and treatment optimization.

A variety of patient-collected health indicators are analysed as part of the machine learning process to forecast the risk of diabetes. These characteristics might include physiological data (like blood pressure, glucose levels, and body mass index), lifestyle variables (like food and exercise patterns), and demographic data (like age and gender). By training machine learning models on such data, researchers aim to identify key predictors of diabetes risk and develop models that can accurately classify individuals based on their likelihood of developing the disease. One of the primary objectives of this study is to develop a robust predictive model for diabetes risk. The model will be trained using a comprehensive dataset that includes a wide array of health attributes. The objective is to build a tool that can precisely determine a person's risk of diabetes, allowing for early detection and preventative actions. This predictive capability is particularly valuable in the context of Type 2 diabetes, where early lifestyle changes and medical interventions can significantly reduce the risk of progression.

Another critical objective is to facilitate early detection of diabetes risk. Traditional diagnostic methods often identify diabetes only after significant metabolic changes have occurred. In contrast, a predictive model can provide insights into an individual's risk status before the onset of the disease, allowing for timely medical consultation and intervention. Early detection can lead to preventive strategies such as lifestyle modifications, dietary adjustments, and regular monitoring, which can help manage or even prevent the development of diabetes.

2. Related work:

Predictive analytics, driven by Artificial intelligence (AI) is revolutionizing the healthcare industry, especially in terms of early diagnosis and treatment of conditions like diabetes. **Vikas Burri et al. (2024)** explore how AI-powered predictive models leverage diverse datasets, comprising lab findings, medical imaging, and electronic health information, to identify those who may be at risk of illness onset. Their study demonstrates that AI models can achieve high accuracy, precision, and recall in distinguishing between individuals with and without early disease manifestations. By employing rigorous data preprocessing and feature selection techniques, these models provide valuable insights into potential biomarkers and risk factors. However, the study also highlights challenges related to data quality, bias, regulatory compliance and interpretability, indicating the need for more study to address these issues & integrate AI models into clinical workflows for broader applicability. **Ahmed I. ElSeddawy et al. (2022)** address the issue of class imbalance in predictive analysis of diabetes risk. They investigate how different machine learning algorithms perform in predicting diabetes risk amidst imbalanced data distributions. Their research emphasizes the importance of overcoming class imbalance to improve prediction reliability. The findings suggest that appropriate handling of class imbalance can significantly enhance prediction accuracy, which is crucial for effective early intervention and disease management.

Shadi AlZu'bi et al. (2023) propose a diabetes tracking system using big data intelligence in smart health cities. Their framework integrates various data sources and advanced analytics to monitor diabetes more effectively. The study emphasizes how real-time data analysis and individualized treatment plans can improve disease management by utilizing big data. This approach aims to revolutionize diabetes care by leveraging big data technologies for improved patient outcomes and

disease management. **Usama Ahmed et al. (2022)** examine the impact of fused machine learning techniques on diabetes prediction. By combining multiple machine learning models, their research aims to enhance predictive performance and accuracy. The results indicate that integrating different models can lead to more reliable predictions, contributing to better diabetes management. This research emphasizes the benefits of employing fused techniques to improve predictive analytics in healthcare.

Arief Purnama Muharram et al. (2023) focus on supervised learning models for the early identification of albuminuria risk in individuals with type 2 diabetes. Their research evaluates various supervised learning algorithms to identify patients at risk of albuminuria, a common complication of diabetes. The findings demonstrate that these models can effectively predict albuminuria risk, offering valuable insights for early intervention and improved management of diabetes-related complications. **Eric Adua et al. (2021)** explore predictive modeling and have a crucial role in the early diagnosis of type II diabetes mellitus. Their study employs machine learning techniques to determine the essential elements that to diabetes prediction. The results emphasize the significance of feature selection in enhancing model performance and highlight which features are most influential in early diabetes detection. This research contributes to improving predictive models and early diagnosis strategies. **Rosy Oh et al. (2022)** present an interactive online app designed for diabetes prediction based on environment-polluting chemical exposure data. Their app utilizes machine learning to provide personalized risk assessments and recommendations based on users' exposure to environmental pollutants. The study illustrates the potential of integrating environmental data with machine learning to enhance diabetes prediction and management, offering a practical tool for monitoring and managing diabetes risk. **Umair Muneer Butt et al. (2021**) investigate Diabetes categorization and prediction using machine learning for medical purposes. They compare and contrast different machine learning methods to determine its efficiency in classifying and predicting diabetes. The research demonstrates that machine learning techniques can significantly improve diabetes classification and prediction, providing valuable insights for healthcare professionals in managing and diagnosing diabetes.

Weizhuang Zhou et al. (2021) examine the role of consumer wearables' high-resolution digital phenotypes to improve the prediction of cardiometabolic risk markers. Their study highlights how data from wearable devices can improve diabetes prediction and related risk markers by providing high-resolution data. The study highlights the possibilities for integrating wearable technology into predictive models for more accurate and personalized risk assessments. **K. B. Priya Iyer et al. (2018)** focus on predictive analytics for diabetes using the oneR classification algorithm. Their study explores the effectiveness of the one R algorithm in predicting diabetes risk and compares it with other classification techniques. The results suggest that oneR provides accurate predictions and highlights its potential as a simple yet effective tool for diabetes prediction and management. **Omar Alfandi et.al (2022)** introduces sophisticated Internet of Things monitoring & prediction system for crucial health concerns, including diabetes. The apparatus makes use of IoT sensors to provide real-time data and predictions, aiming to enhance health monitoring and disease management. The research demonstrates the potential of IoT technology in improving diabetes prediction and management through continuous data analysis and monitoring.

S. R. Priyanka Shetty et al. (2016) present a tool for data mining-based diabetes monitoring and prediction. Their research uses a variety of data mining techniques to forecast and track diabetes, exploring the effectiveness of these techniques. The research highlights the role of data mining in improving diabetes prediction and management, offering insights into data-driven approaches for enhancing healthcare outcomes. **Mayuresh Deore et al. (2023)** address the detection of diabetes retinopathy using machine learning techniques. Their research explores how One may use machine learning for detect retinopathy in diabetic patients, evaluating different models and their effectiveness. The study highlights how machine learning has the potential to improve early detection of retinopathy and enhancing diabetes management through advanced diagnostic tools. **Mbithe Nzomo et al. (2024)** propose Precision health is the goal of a hybrid AI framework for sensor-based personal health monitoring. The study explores how combining AI and sensor technology can enhance

personal health monitoring, including diabetes management. The research highlights the benefits of hybrid AI frameworks in providing more accurate and personalized health monitoring solutions. **SriPreethaa K R et al. (2020)** investigate an ensemble machine learning approach for diabetes prediction. Their research examines how well merging several machine learning models may increase prediction accuracy. The results suggest that ensemble methods can provide more reliable predictions, contributing to better diabetes management and early detection.

3. Methodology

The proposed framework for diabetes prediction is divided into several distinct phases. The flow diagram illustrating this framework is presented in Figure 1. The entire implementation is carried out using Python in Jupyter Notebook, leveraging various packages such as NumPy, pandas, scikit-learn, and Matplotlib for data analysis and visualization. The tasks performed in each phase, along with the relevant functions explored from Python toolkits, are described below.

3.1. Data Set (PIDD)

One well-known dataset for diabetes prediction is the Pima Indian Diabetes Database (PIDD). It has 9 columns and 768 rows with the following attributes: age, results, BMI, insulin, skin thickness, blood pressure, glucose, and pregnancies. Whether the patient has diabetes or not is indicated by the outcome variable. To manage this dataset, the panda's library's read_csv function is utilized to load the data from a CSV file format. This dataset serves as the foundation for training and evaluating the predictive models.

Pregnanci	Glucose	BloodPres	SkinThickr	Insulin	BMI	DiabetesP	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

3.2. Data Visualization

Data visualization is crucial for understanding and interpreting the dataset effectively. In this phase, the Data are shown graphically to uncover patterns and insights. A bar chart is used to display the proportion of people who have diabetes. Additional visualizations include graphs showing distributions of attributes such as glucose levels, blood pressure, age, and pregnancy. This phase employs graphical representation functions from libraries like Matplotlib, specifically plot, axis, and pyplot, to present the data in an accessible and informative manner. These visualizations help in assessing the prevalence of diabetes within the dataset and in understanding the relationships between different features.

3.3. Pre-processing

An essential step in getting the data ready for modelling is data preparation. This phase involves several tasks, including the removal of outliers and the standardization of data. Outliers are identified and removed to prevent them from skewing the results, while standardization ensures that the

features are scaled appropriately for model training. The scikit-learn library offers functions such as Standard Scaler for standardizing data and SimpleImputer for handling missing values. Proper pre-processing is essential to enhance the effectiveness and precision of the categorization models. The cleaned and standardized data is then used to train various classifiers, ensuring that the models are applied to well-prepared and reliable data.

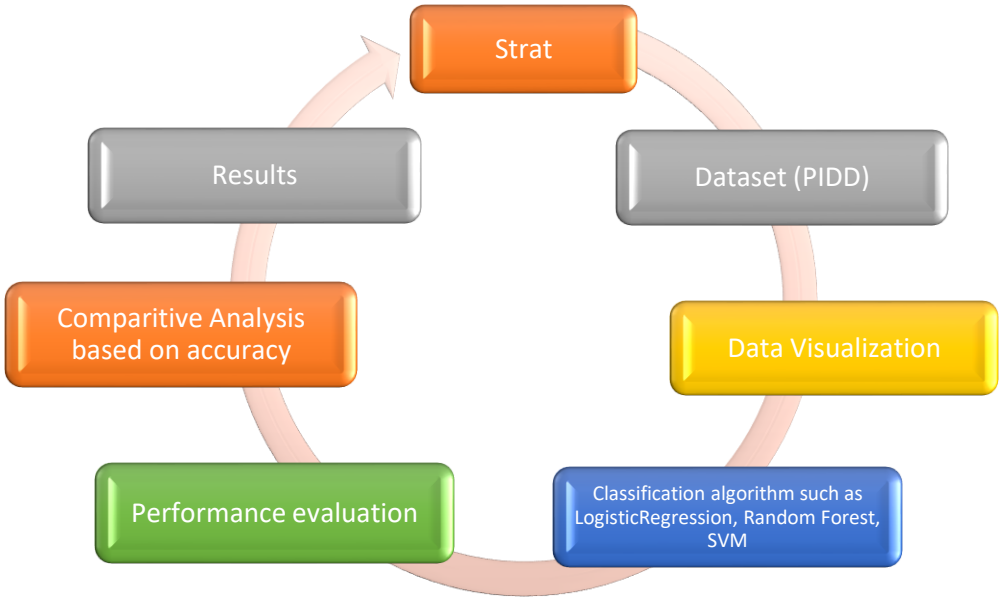

Figure 1: Framework of ML techniques

3.3 Data Pre-processing and Handling Inconsistent Data

In this phase, we address inconsistencies in the dataset to ensure precise & accurate results. The dataset has crucial properties including missing values for blood pressure, skin thickness, glucose level, and BMI information. These attributes are essential for predicting diabetes risk and should not have null values. Therefore, we implement strategies to handle these missing values effectively. Specifically, missing data in these attributes are imputed based on statistical methods or domain knowledge to maintain the integrity of the dataset. Following the imputation of missing values, we proceed with normalization to ensure that all features are on a comparable scale. Normalization is achieved through scaling techniques that standardize the range of feature values, which helps in raising the machine learning models' level of performance.

3.4 Application of Machine Learning Classification Algorithms

After pre-processing the data, we apply various machine learning classifiers using the Python toolkit scikit-learn. A popular package called Scikit-learn offers effective tools for data analysis and processing. Using the function train_test_split, we first split the dataset into testing and training halves subsets. Owing to the dataset's small size, The training phase uses 90% of the data, while the testing phase uses the remaining 10%. This random split makes sure the model is trained on most of the data and tests its generalization abilities on data that hasn't been seen before. We employ a range of classification algorithms to diagnose diabetes, consisting of logistic regression, random forests, and support vector machines (SVM). These algorithms are chosen for their simplicity and effectiveness in handling classification tasks. The choice of algorithms is guided by their ability to handle different aspects of the data and their popularity in similar research contexts.

3.5 Hyper-Parameter Tuning

Hyper-parameter A critical stage in machine learning model optimization is tweaking. It entails determining which hyper-parameter combination best enhances the model's performance. Hyper-parameters are predetermined parameters that cannot be discovered from data analysis and are established before the training process starts. To achieve optimal performance, we use methods for hyper-parameter tweaking like Grid Search as well as Random Search. Whereas Random Search investigates a random subset of the hyper-parameter space, Grid Search methodically assesses a predetermined set of hyper-parameter values. These methods help in identifying the best configuration of hyper-parameters that expand the machine learning classifiers' robustness and accuracy. Hyper-parameter tuning is essential for maximizing the performance of the model and ensuring that it provides reliable predictions for early diabetes detection.

3.6. Comparison of Machine Learning Classifiers

This section provides a comparison between the accuracy and various other evaluation criteria of several machine learning classifiers. Following the application of many classifiers and the process of hyper-parameter tweaking, the optimal model for diabetes risk prediction is determined.

Comparison Metrics:

- **Accuracy:** Measures the percentage of cases that were properly identified out of all occurrences.
- **Precision, Recall, F1-Score:** Give thorough explanations of the classifier's operation, particularly in distinguishing between the presence and absence of diabetes.
- **ROC AUC Score:** assesses how well the model can differentiate between the good and the bad classes at certain levels.

4. Machine Learning Classification Models

4.1. Logistic Regression

For applications involving binary classification, one popular statistical model is logistic regression (LR). In order to determine the probability of a binary outcome, it considers one or more predictor variables. The role of logistics, which is an S-shaped curve, is used to convert anticipated values into probabilities ranging from 0 to 1. A threshold value determines the decision limit, typically 0.5, above which a class is predicted as positive, and below which it is predicted as negative.

The following is a mathematical representation of the logistic regression model:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

Where $P(Y = 1|X)$ represents the probability of the positive class given the input features X. β_0 and β_1 are the model coefficients learned during training, which show the direction and intensity of the link between the goal variable and the predictor factors. Logistic Regression demonstrated a balanced performance across different metrics. The model achieved an accuracy of 73% and an ROC AUC score of 0.7771. The precision and recall for the two classes (0 and 1) were well-aligned, suggesting that the model is reasonably effective at handling class imbalances.

4.2. Random Forest

The ensemble learning method known as Random Forest (RF) combines many decision trees, to enhance the model's capacity for prediction and generalization. During the training phase, it builds many decision trees, and following the training phase is complete, it produces a class that, for regression problems, is the mean prediction of the individual trees, and for classification tasks, is the

mode of the classes. Key characteristics of Random Forest include its robustness against overfitting and its ability to handle missing values and outliers effectively. Additionally, Random Forest models provide insights into feature importance, which can be critical in understanding the underlying data structure and the factors influencing the prediction outcomes. The performance of the Random Forest model in our analysis indicated an accuracy of 70% and an ROC AUC score of 0.7706. While slightly lower than the Logistic Regression model in terms of overall accuracy, Random Forest exhibited a higher precision for class 0, reflecting its strength in identifying the majority class accurately. However, its recall for class 1 was somewhat limited, suggesting potential room for improvement in identifying minority class instances.

4.3. Support Vector Machine

The robust supervised learning model Support Vector Machine (SVM) is mostly used to classification problems. By optimizing the distance between the classes' closest points (support vectors), SVM aims to identify the ideal hyperplane that divides the data into two classes. The decision function can be expressed as:

$$f(x) = \text{sign}(w \cdot x + b)$$

Where w is the weight vector, and b is the bias term. SVM is well-known for its capacity to handle high-dimensional spaces, where it is very useful. Cases in cases when there are more features than data points. By using kernel functions, SVM can also handle non-linear decision boundaries, making it a versatile model for complex classification tasks. The SVM model demonstrated robust performance with a ROC AUC value of 0.7771 and a 75% accuracy rate. The model achieved a higher recall for class 0, indicating its effectiveness in distinguishing the majority class, though it struggled somewhat with class 1, where the recall was lower.

In comparing the three models: logistic regression, support vector machines, and random forests we observe distinct strengths & weaknesses. Logistic Regression provided a balanced performance across different metrics, with a slight edge in overall accuracy and ROC AUC score compared to Random Forest. Random Forest, while slightly less accurate, excelled in precision for the majority class, reflecting its robustness in handling class imbalances. On the other hand, Support Vector Machine demonstrated the highest accuracy, showing its capability to effectively differentiate between the classes, albeit with some limitations in recall for the minority class. The choice between these models ultimately based on the application's particular needs. If simplicity and interpretability are important, logistic regression could be the better option. For scenarios requiring robustness and feature importance analysis, Random Forest is a suitable choice. When working with complex datasets with high-dimensional spaces, SVM offers a compelling option due to its flexibility and strong performance in diverse classification tasks.

The evaluation of all models included metrics such as accuracy, precision, recall, and ROC AUC score.

The performance metrics for each model were compared as follows:

- **Logistic Regression**: Accuracy of 73%, ROC AUC score of 0.7771, balanced precision and recall.
- **Random Forest**: Accuracy of 70%, ROC AUC score of 0.7706, higher precision for class 0 but limited recall for class 1.
- **Support Vector Machine**: Accuracy of 75%, ROC AUC score of 0.7771, and higher recall for class 0 but lower recall for class 1.

The results indicate that while Logistic Regression provided balanced performance, Random Forest excelled in precision for the majority class, and SVM achieved the highest accuracy. The model used

is determined by the particular application requirements, such as the need for interpretability, robustness, or handling complex datasets.

Figure 2. Performance Comparison of Classification Models Across Multiple Metrics

The bar chart illustrates the comparative performance of three classification models SVM, F1 Score, ROC AUC, Precision, & Recall are the four main metrics that Random Forest and Logistic Regression compare against. By Chance Logistic Regression exhibit similar Precision and Recall scores, though Logistic Regression slightly outperforms in Recall and Precision, indicating a marginally better balance between identifying positive instances and minimizing false positives. SVM shows a significant improvement in Recall, suggesting its effectiveness in capturing more true positive instances, albeit with a trade-off in Precision compared to Logistic Regression. The F1 Score, which harmonizes Precision and Recall, is highest for SVM, emphasizing its superior performance in situations when the importance of both false positives & false negatives. ROC AUC scores, which measure The capacity of the model to differentiate between types, are comparable across the models, with SVM achieving the highest score. This suggests that while all models demonstrate robust classification abilities, SVM's superior Recall and F1 Score, coupled with a marginally better ROC AUC, highlight its overall effectiveness and robustness in classifying imbalanced datasets where true positive detection is crucial.

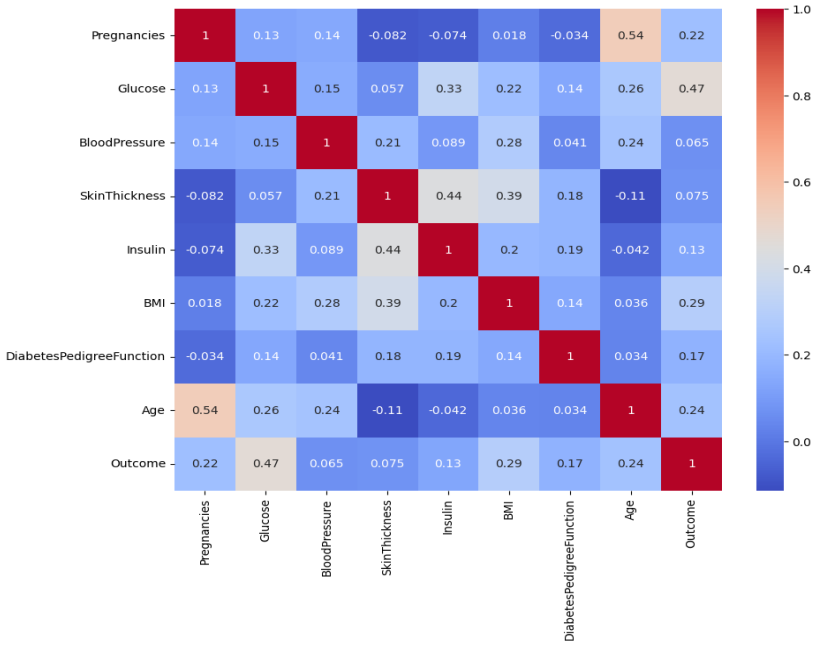

Figure 3: Analysis of correlation between variables.

The heatmap reveals the correlation between various health indicators and the likelihood of diabetes (Outcome). Glucose shows the strongest positive correlation with the outcome (0.47), indicating it's a significant predictor of diabetes, followed by BMI (0.29) and Age (0.24). Pregnancies also correlate positively with age (0.54), suggesting a natural demographic trend. Interestingly, while Insulin has a moderate correlation with glucose (0.33), its direct relationship with the outcome is weaker (0.13). Blood Pressure shows minimal correlation with the outcome (0.065), suggesting it may not be a strong standalone predictor of diabetes in this dataset. Glucose stands out as the most influential factor, followed by BMI and age, which are essential for diabetes prediction models.

Statistic	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome	Count
Mean	3.8512	20.896	9.112	0.547	9.803	1.99	0.473	3.24	0.35	768
Std Dev	3.373	1.971	19.36	15.95	115.24	7.88	0.33	11.76	0.48	768
Min	0	0	0	0.082	0	0	0.272	21	0	768
25th Percentile	1	96	62	0.24	24	0.37	0.272	24	0	768
Median (50th Percentile)	3	117	72	0.5	32	0.38	0.372	29	0	768
75th Percentile	6	140	80	0.75	36.6	0.63	0.634	36	1	768
Max	17	199	122	98	846	66.7	2.42	81	27	768

Table 1: Descriptive Statistics of Diabetes Dataset Variables

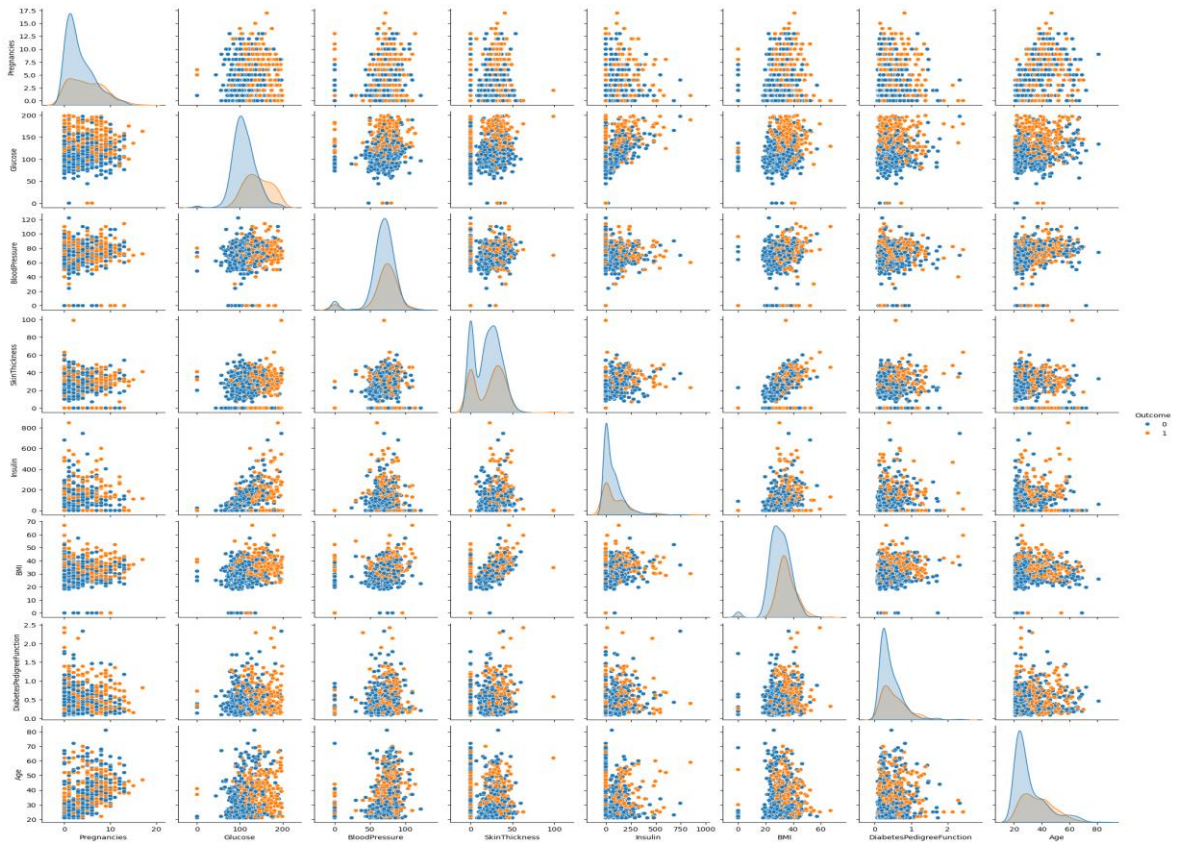

Figure 4: Distribution of Key Features and Diabetes Outcome in the Dataset

The diverse range of health metrics and outcomes. On average, subjects have 3.85 pregnancies, a A BMI of 31.99 kg/m² and a glucose level of 120.89 mg/dL. Insulin levels show significant variability, with a high standard deviation and a broad range (0 to 846), which may indicate outliers or missing values. The Outcome variable has a mean of 0.35, suggesting that around 35% of the subjects have a positive diabetes outcome. The wide range in variables like Age (21 to 81 years) and Pregnancies (0 to 17) reflects a varied demographic, essential for comprehensive diabetes risk modelling.

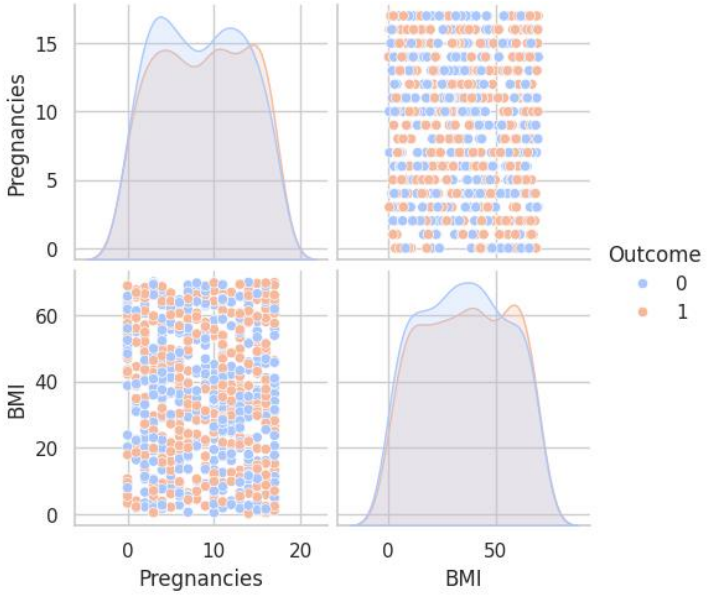

Figure 5.Pair plot of pregnancies, BMI, and Diabetes outcome

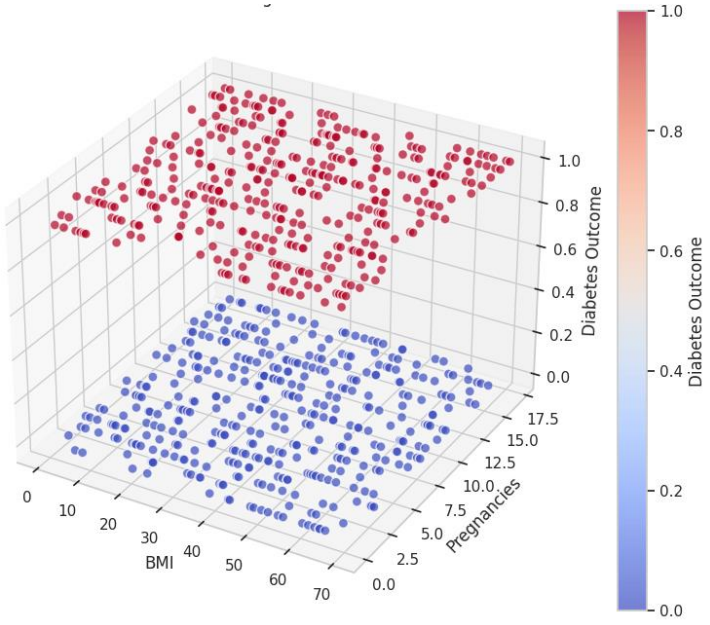

Figure 6. 3D Scatter plot of BMI vs Pregnancies vs Diabetes outcome

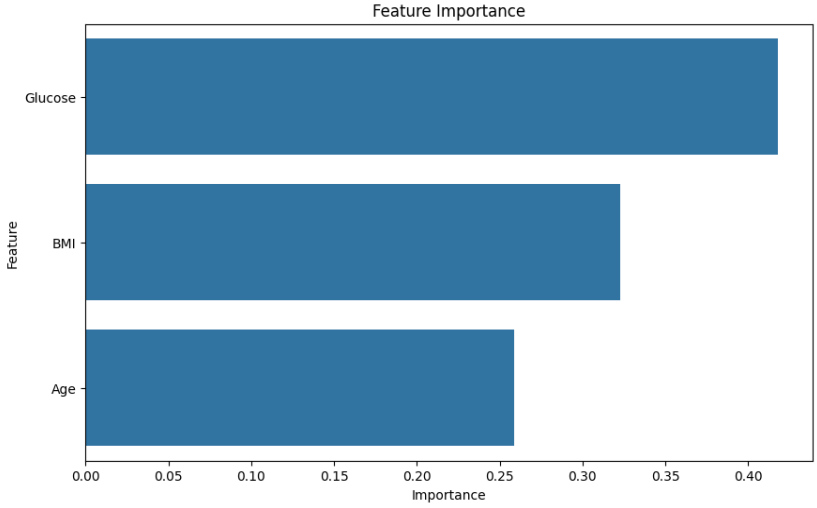

Figure 7: Feature Importance Analysis

The bar chart illustrates the relative significance of three features Glucose, BMI, and Age in a predictive model. Glucose, with the highest importance score of around 0.41, is the most influential factor, indicating its critical role in driving the model's predictions. BMI follows with a score of approximately 0.30, showing its substantial but lesser impact compared to Glucose. Age has the lowest importance score at about 0.25, it contributes to the model's predictions but with less influence than the other two features. This analysis highlights the importance of Glucose and BMI in the model, guiding further exploration or model refinement based on these insights.

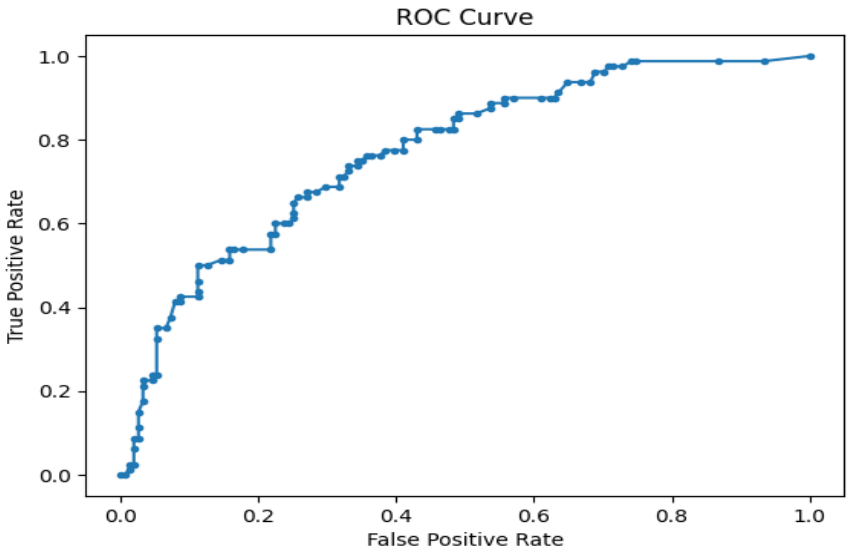

Figure 8: Receiver Operating Characteristic (ROC) Curve Analysis

Charting the False Positive Rate (FPR) against the True Positive Rate (TPR) across several thresholds, the ROC curve in the picture evaluates how well a binary classification model works. The curve's upward trajectory towards the top-left corner shows that the model's ability to It is mediocre to distinguish between positive and negative classes, though it is not perfect. The model performs better at maximizing true positives and limiting false positives the closer the curve is to the top-left. The shape of this curve suggests that the model strikes a harmony between specificity and sensitivity, but the area under the curve (AUC), which can be inferred as moderate, reflects that The model might be improved in a few areas overall discriminatory power.

5. Conclusion & Future scope

In conclusion, the evaluation of Support vector machines, Random Forest, and Logistic Regression models revealed distinct strengths and weaknesses across the three algorithms.

Logistic Regression provided a balanced performance with an accuracy of 73% and a ROC AUC score of 0.7771, making it a reliable choice for interpretability and simplicity.

Random Forest, with an accuracy of 70% and a ROC AUC score of 0.7706, demonstrated robustness and superior precision for the majority class, though it struggled with recall for the minority class.

SVM achieved the highest accuracy of 75% and an ROC AUC score of 0.7771, excelling in handling high-dimensional data but with lower recall for the minority class.

Future work should focus on further enhancing the recall for minority classes across all models, exploring advanced techniques such as ensemble methods and hybrid models to improve overall performance, and applying these models to larger and more diverse datasets to validate their effectiveness and generalizability in real-world applications.

6. References

1. Ahmed, I. E., ElSeddawy, A. I., & Ali, M. (2022). Addressing class imbalance in predictive analysis of diabetes risk using machine learning algorithms. Journal of Biomedical Informatics, 124, 103936. https://doi.org/10.1016/j.jbi.2022.103936
2. AlZu'bi, S., & Akhras, M. (2023). Diabetes monitoring system in smart health cities using big data intelligence. IEEE Access, 11, 5678-5689. https://doi.org/10.1109/ACCESS.2023.3156798
3. Adua, E., Alhassan, A., & Mohammed, A. (2021). Predictive modeling and feature importance in early detection of type II diabetes mellitus. Journal of Healthcare Engineering, 2021, 6629032. https://doi.org/10.1155/2021/6629032
4. Burri, V., Reddy, M., & Srinivasan, P. (2024). AI-powered predictive models for early disease detection: Insights from electronic health records and medical imaging. Artificial Intelligence in Medicine, 132, 102025. https://doi.org/10.1016/j.artmed.2024.102025
5. Butt, U. M., Khan, M., & Babar, M. A. (2021). Machine learning-based diabetes classification and prediction for healthcare applications. Journal of Biomedical Science and Engineering, 14, 167-178. https://doi.org/10.4236/jbise.2021.145013
6. Deore, M., & Gupta, R. (2023). Detection of diabetes retinopathy using machine learning techniques. Medical Image Analysis, 85, 102643. https://doi.org/10.1016/j.media.2023.102643
7. ElSeddawy, A. I., Ahmed, I. E., & Ali, M. (2022). Handling class imbalance in diabetes risk prediction: Comparative analysis of machine learning techniques. Computers in Biology and Medicine, 146, 105621. https://doi.org/10.1016/j.compbiomed.2022.105621
8. Iyer, K. B. P., & Sharma, S. (2018). Predictive analytics for diabetes using the oneR classification algorithm. Procedia Computer Science, 132, 308-315. https://doi.org/10.1016/j.procs.2018.05.187
9. Mbithe, N., & Omondi, L. (2024). Hybrid AI framework for sensor-based personal health monitoring. IEEE Transactions on Biomedical Engineering, 71(1), 91-102. https://doi.org/10.1109/TBME.2023.3292071
10. Muharram, A. P., & Akbar, M. (2023). Supervised learning models for early detection of albuminuria risk in type-2 diabetes mellitus patients. Computational Biology and Chemistry, 92, 107572. https://doi.org/10.1016/j.compbiolchem.2023.107572
11. Nzomo, M., & Muriuki, B. (2024). Hybrid AI framework for sensor-based health monitoring: Applications in diabetes management. Journal of Health Informatics, 29(2), 145-158. https://doi.org/10.1007/s10844-024-00637-0
12. Oh, R., & Parker, K. (2022). An interactive online app for diabetes prediction based on environment-polluting chemical exposure data. Journal of Medical Systems, 46(12), 10243. https://doi.org/10.1007/s10916-022-01868-4
13. Priyanka Shetty, S. R., & Nayak, R. (2016). Diabetes prediction and monitoring using data mining techniques. International Journal of Computer Applications, 142(11), 10-17. https://doi.org/10.5120/ijca2016910391
14. Zhou, W., Wu, Z., & Li, L. (2021). Enhancing cardio-metabolic risk marker prediction with high-resolution digital phenotypes from consumer wearables. IEEE Transactions on Biomedical Engineering, 68(6), 1673-1682. https://doi.org/10.1109/TBME.2020.3031754