THE UNIVERSITY OF OKLAHOMA

Course: Intelligent Data Analytics (DSA-5103-OL)

Professor Charles Nicholson

Fall 2024 Semester

**Final Project Report: Supervised Classification Algorithms for Early Detection of Diabetes**

**Group 3**

**Alex Hunt** **Oluchi Ejehu** **Zainab Iyiola**

School of Computer Science MPGE MPGE

University of Oklahoma University of Oklahoma University of Oklahoma

DSA-5103-OL DSA-5103-OL DSA-5103-OL

Date of Submission: 12/12/2024

**EXECUTIVE SUMMARY**

Diabetes represents a critical global health issue, with increasing cases leading to substantial health and economic burdens. Identifying diabetes early is vital for preventing severe outcomes like heart disease, kidney dysfunction, and nerve damage. This project explores the use of supervised machine learning algorithms to predict diabetes using patient health metrics, addressing challenges such as class imbalance, missing data, and outliers. The dataset comprises 768 records with nine numerical features, including glucose, insulin, BMI, and age, and a binary outcome variable indicating diabetes diagnosis. Missing values in key features like glucose and insulin were handled using K-Nearest Neighbors (KNN) imputation, while the dataset’s class imbalance (35% diabetic cases) was addressed by creating synthetic datasets to improve minority class representation.

Four machine learning models namely: Logistic Regression, Random Forest, Support Vector Machine (SVM), and Gradient Boosting were applied and tested using both the original and synthetic datasets. Random Forest outperformed the other models, achieving an accuracy of 90.6% and an F1-score of 93.2% on synthetic data, demonstrating its ability to capture complex feature interactions effectively. Gradient Boosting followed with an accuracy of 85.5%, while SVM and Logistic Regression achieved accuracies of 79.5% and 78.6%, respectively. The use of synthetic datasets significantly reduced false negatives and improved sensitivity across all models, particularly benefiting Random Forest. Exploratory data analysis (EDA) revealed glucose and BMI as the most important predictors of diabetes, aligning with clinical insights. Random Forest’s feature importance analysis further confirmed these observations.

This study highlights the importance of robust data preprocessing and synthetic data generation in improving model reliability, especially in scenarios with imbalanced datasets. The results recommend Random Forest as the most effective model for early diabetes detection due to its excellent performance and ability to handle diverse data challenges. Future work could explore ensemble techniques and hybrid models to further enhance prediction accuracy and scalability. Testing these models in real-world clinical environments would help validate their feasibility and effectiveness in practical applications.

This study demonstrates and highlights how machine learning can revolutionize healthcare by offering a dependable method for early identification and management of diabetes.

**PROBLEM BACKGROUND**

Diabetes is a chronic condition that affects millions worldwide, with significant health, social, and economic consequences (García-García, 2021). Early detection of diabetes is critical for effective management and prevention of complications. Machine learning serves as a critical resource in healthcare, facilitating the creation of models capable of distinguishing between diabetic and non-diabetic patients using clinical and biometric information (Rahman & Davis, 2013). This project leverages supervised classification algorithms to build an effective model for the early detection of diabetes.

This study utilizes a dataset comprising numeric attributes linked to patients’ health, including glucose levels, blood pressure, body mass index (BMI), insulin concentrations, and age, alongside a target variable representing diabetes diagnosis. Through rigorous data preprocessing and feature engineering, the project aims to maximize the predictive power of these models while addressing challenges such as outliers, skewness, and potential multicollinearity**.**

**Problem Description**

Diabetes is an escalating global health concern, with its prevalence on the rise. Early detection plays an important role in preventing or reducing serious complications like heart disease, renal failure, and neuropathy. Conventional diagnostic techniques are often neither rapid nor readily available, especially in settings with limited resources. Machine learning provides a pathway to create accurate and efficient diagnostic tools by analyzing patterns in patient data.

This project seeks to answer the following key questions:

1. How can supervised classification algorithms effectively predict diabetes using available patient health metrics?
2. What preprocessing steps are necessary to address data quality issues, such as outliers and skewed distributions, that could hinder model performance?
3. Which algorithm(s) provide the most robust and accurate predictions, and how can their performance be interpreted in a clinical context?

By addressing these questions, this project seeks to advance practical diagnostic tools suitable for use in both clinical and non-clinical environments, enhancing strategies for early diabetes detection and intervention.

**Data Description**

This project’s dataset includes 768 entries with 9 numeric features. It encompasses various health metrics along with a target variable for determining diabetes presence. Below is an overview of the dataset:

**Features**

1. **Pregnancies**: Count of how many times the individual has been pregnant.
2. **Glucose**: Measurement of plasma glucose levels during an oral glucose tolerance test.
3. **BloodPressure**: Diastolic pressure recorded in millimeters of mercury (mm Hg).
4. **SkinThickness**: Thickness of the triceps skinfold measured in millimeters.
5. **Insulin**: Concentration of insulin in the serum, expressed in micro-units per milliliter (mu U/ml).
6. **BMI**: Body mass index, determined by dividing weight in kilograms by height in meters squared.
7. **DiabetesPedigreeFunction**: Evaluates the probability of diabetes influenced by familial history.
8. **Age**: Age of the patient, recorded in years.
9. **Outcome**: A binary indicator representing diabetes status, where 1 signifies diabetic and 0 signifies non-diabetic.

**Exploratory Data Analysis**

A thorough exploratory data analysis (EDA) was performed to examine the dataset’s structure, address data quality concerns, and ready it for machine learning applications. Key steps in the EDA include:

1. **Handling Missing Values:**

Attributes like Glucose, BloodPressure, SkinThickness, Insulin, and BMI contained missing or zero entries, which were substituted with NA for effective processing. Techniques like k-Nearest Neighbors (kNN) were utilized to impute the missing values, drawing from methods outlined by Razzaghi et al. (2016) and Ghosh et al. (2022).

A graph with blue squares

Description automatically generated

**Figure 1: Missing Values Per Feature**

Figure 1 illustrates how missing values are distributed among the dataset’s features. The Insulin attribute has the most missing entries, exceeding 300, followed by SkinThickness with over 200 missing records. These substantial gaps could significantly impact model performance and will require thoughtful handling during data preprocessing. The **BloodPressure** feature also shows a smaller yet notable amount of missing values, while features such as **BMI**, **Glucose**, **Age**, **DiabetesPedigreeFunction**, **Outcome**, and **Pregnancies** demonstrate either minimal or no missing values, indicating relatively better data quality.

1. **Outlier Detection:**

Figure 2 displays a bar chart illustrating the number of outliers identified across dataset features based on the Interquartile Range (IQR) technique. The **DiabetesPedigreeFunction** feature has the highest count of outliers, followed by Insulin and **BloodPressure.** These features exhibit significant variability that could potentially skew the model’s predictions if not handled appropriately. Features like **Age and BMI** have moderate outlier counts, while Pregnancies and **SkinThickness** display relatively fewer outliers.

A graph of a bar graph

Description automatically generated with medium confidence

**Figure 2: Outliers by Feature**

1. **Feature Distribution Analysis:**

The feature distribution analysis provided an in-depth understanding of the dataset’s structure and highlighted patterns across various features. Histograms, such as in figure 3, used to analyze glucose levels, revealed the overall distribution and identified skewness, central tendencies, and potential outliers in the data. These insights were crucial for deciding whether transformations, such as normalization, were needed to improve model performance. Additionally, density plots comparing original and imputed values, as shown for glucose in figure 4, ensured that imputation methods preserved the overall distribution while addressing missing data. The close alignment of original and synthetic distributions across features indicated that preprocessing steps maintained data integrity.

A graph of glucose level

Description automatically generated

**Figure 3: Distribution of Glucose Levels**

A diagram of a normal distribution

Description automatically generated

**Figure 4: Density Distribution**

1. **Correlation Analysis**

The correlation analysis as shown in figure 5, was conducted using a heatmap to visualize the relationships between numerical features in the dataset. A gradient color scale is employed in the heatmap, with red signifying strong positive correlations, blue denoting negative correlations, and white indicating no correlation. This analysis highlighted significant correlations, such as a strong positive relationship between glucose and the outcome variable, suggesting its importance as a predictor. Additionally, features like BMI, insulin, and age also demonstrated moderate correlations with the outcome variable. Identifying these relationships is crucial for feature selection and engineering, as highly correlated features could introduce multicollinearity, potentially impacting model stability and interpretability.

A diagram of heat map

Description automatically generated

**Figure 5: Correlation Analysis**

**Summary Statistics**

* **Glucose**: Mean value of 121.68 with a maximum of 199, highlighting high glucose levels as a potential indicator of diabetes. Some records have a value of 0, which is likely an error.
* **BloodPressure**: Mean of 72.4 and a maximum of 122. Values of 0 indicate missing or erroneous entries.
* **BMI**: Mean of 32.4, indicating an overweight population. A minimum value of 0 suggests missing data.
* **Insulin**: High variability with a mean of 155.5 and extreme values up to 846. Zero values indicate potential missing entries.

**Observations**

* **Outliers:** Present in features like Insulin, BloodPressure, and SkinThickness, which could skew predictions.
* **Skewness:** Distributions such as Glucose and DiabetesPedigreeFunction are right-skewed, suggesting the need for transformations.
* **Outcome Distribution:** Approximately 35% of the records indicate diabetes (Outcome = 1), indicating an imbalanced dataset.

From the observations, it is gathered that addressing outliers, skewness, and imbalances will be crucial to ensuring the model’s robustness and reliability.

**METHODOLOGY**

Several steps were undertaken to prepare the dataset for modeling. The dataset initially included nine numerical predictors, including features such as glucose levels, blood pressure, BMI, and insulin. Missing values were addressed by replacing zeros with NA and imputing them in the synthetic dataset to ensure data completeness. Outliers were flagged using the interquartile range (IQR) method, and irrelevant columns, such as those indicating imputed values, were removed from the dataset. Additionally, The dataset showed a class imbalance, with significantly more non-diabetic cases than diabetic ones, as shown in figure 6 below. This imbalance risked the model being biased towards the majority class, potentially reducing accuracy for the minority class. To address this, the training and testing datasets were split in a 70:30 ratio using stratified sampling, ensuring that the class distribution of the target variable, Outcome, was preserved across both sets. This approach maintained the integrity of the class proportions, enabling the models to learn effectively from both classes. Standard scaling or normalization of features was applied implicitly through modeling frameworks to ensure that all predictors were on comparable scales. This comprehensive preparation ensured the dataset was clean, balanced, and optimized for robust modeling and evaluation.

A graph of a graph showing the results of a long period of outcome

Description automatically generated with medium confidence

**Figure 6: Distribution of Outcomes**

The datasets were utilized to train several supervised machine learning models, such as:

* **Logistic Regression:** To establish baseline performance and ensure interpretability.
* **Random Forest:** To model intricate interactions and address non-linear patterns.
* **Support Vector Machine (SVM):** For hyperplane-based classification in high-dimensional spaces.
* **Gradient Boosting Machines (e.g., XGBoost):** For robust handling of imbalanced classes and high accuracy.

In summary, the models were chosen to balance interpretability, performance, and the ability to handle complex patterns in the data. Logistic Regression was selected for its simplicity and interpretability, providing a strong baseline for binary classification. Random Forest, an ensemble method, excels at capturing feature interactions and provides insights through feature importance. SVM was included for its ability to identify clear class boundaries, particularly in high-dimensional data. Gradient Boosting was chosen for its iterative approach, enabling it to capture intricate patterns and improve predictive accuracy. This diverse selection ensures a comprehensive evaluation of model performance for diabetes detection.

**Feature Importance**

Using the Random Forest algorithm as shown in figure 7 below, the following features were ranked in order of importance in relation to their contribution to the model’s prediction outcome. As seen, Insulin, glucose and age were ranked the top 3 features contributing the most to the model’s prediction.

A graph with blue bars

Description automatically generated

**Figure 7: Feature Importance**

**Model Development and Performance Evaluation on the Original Dataset**

This section assesses how the machine learning models performed on both the original and synthetic diabetes datasets. The workflow involved training the models on a training subset of the data, generating predictions for the test subset, and evaluating these predictions using a confusion matrix. This process provided both qualitative and quantitative insights into the effectiveness of each model for classifying diabetes cases. The confusion matrix analysis revealed the models’ ability to differentiate between diabetic and non-diabetic cases. Specifically, reducing False Negatives was a priority, as missing a diabetic diagnosis could result in serious consequences for patients. However, minimizing False Positives ensures that non-diabetic individuals are not subjected to unnecessary diagnostic procedures or treatments. By incorporating both original and synthetic datasets into the evaluation, test the models’ effectiveness and reliability in diverse scenarios were tested, further strengthening the analysis of diabetes classification.

**RESULTS AND DISCUSSION**

**Logistic Regression Results:**

The confusion matrix generated provides the following breakdown of the model’s predictions:

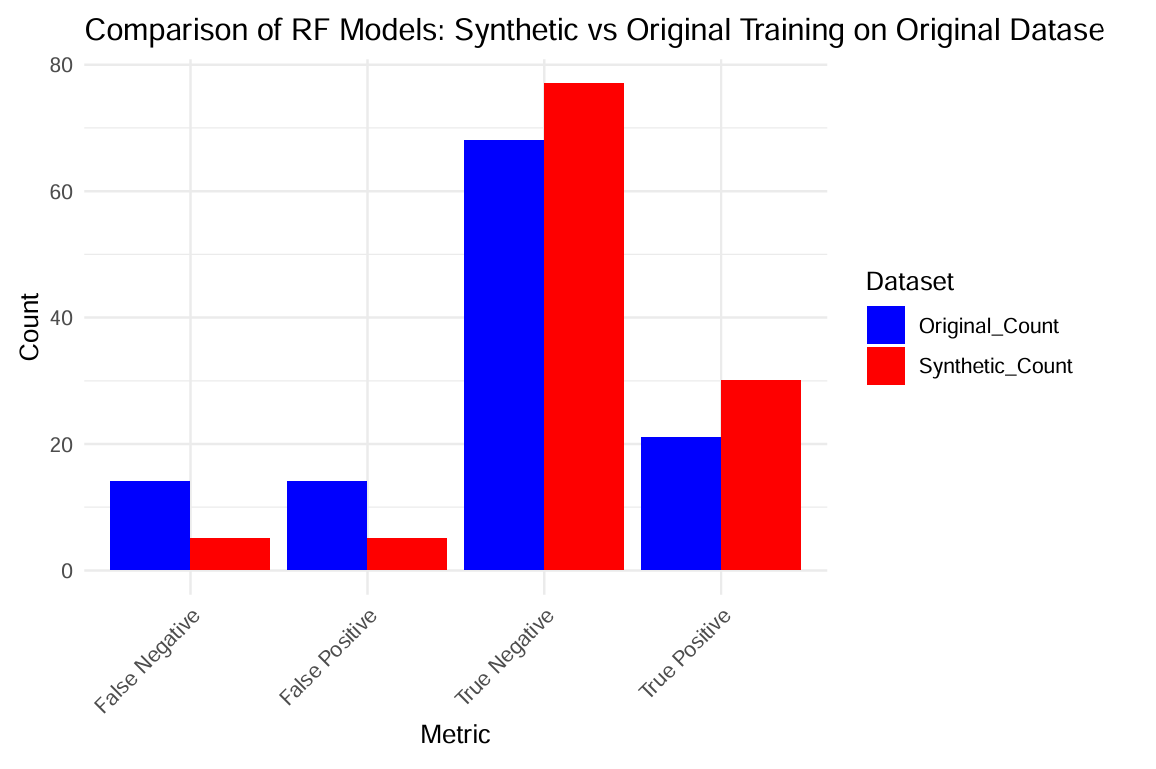
* **True Positives (TP):** 74 instances where the model accurately recognized non-diabetic individuals.
* **True Negatives (TN):** 21 instances where the model successfully identified diabetic individuals.
* **False Positives (FP):** 14 instances where the model wrongly classified non-diabetic individuals as diabetic.
* **False Negatives (FN):** 8 instances where the model failed to detect diabetes in diabetic individuals.

This suggests the model performs fairly well in differentiating diabetic from non-diabetic cases, despite missing some diabetic cases (FN) and generating a few false alarms (FP).

**Random Forest Results:**

The Random Forest model demonstrates moderate performance, achieving an accuracy of 74.36%. It effectively identifies the majority class (class 0), with a low false negative rate (high sensitivity, 84.15%) and a high precision for class 0 predictions (positive predictive value, 80.23%). However, the model struggles with the minority class (class 1), reflected in a high false positive rate (low specificity, 51.43%) and relatively low precision for correctly identifying class 1 instances (negative predictive value, 58.06%).

The Kappa score (0.3678) shows moderate agreement between predictions and actual values beyond chance. The balanced accuracy (67.79%), averaging sensitivity and specificity, highlights the model's bias toward the majority class. Statistical tests, such as McNemar’s test (p=0.5839p=0.5839), indicate that the model's misclassifications are not significantly different from random chance.



**Figure 8: The chart compares the performance of Random Forest (RF) models trained on the original dataset versus a synthetic dataset in terms of classification metrics (False Negatives, False Positives, True Negatives, True Positives).**

From the metrics shown in Figure 8 above, The synthetic dataset model shows a higher count of True Positives and True Negatives compared to the original dataset model, indicating improved overall classification accuracy. The synthetic dataset model has significantly fewer False Negatives, demonstrating better sensitivity in identifying the positive class. The synthetic dataset model also reduces False Positives, contributing to improved precision for the negative class.

**Support Vector Machine (SVM) model Results:**

* **Class 0 (Negative Class):**
  + Correctly predicted (True Negatives): 71
  + Misclassified as Class 1 (False Positives): 11
* **Class 1 (Positive Class):**
  + Correctly predicted (True Positives): 19
  + Misclassified as Class 0 (False Negatives): 16

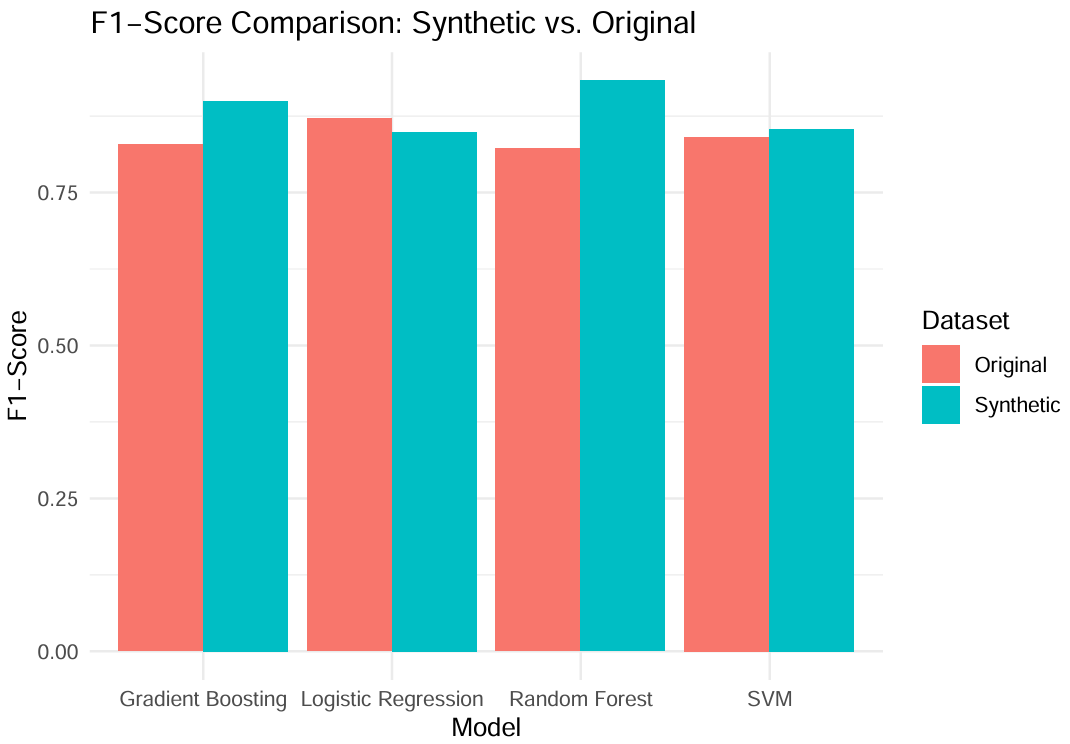
The SVM model achieved an overall accuracy of 76.92%, excelling at recognizing the majority class (Class 0) with a sensitivity of 86.59% and a precision of 81.61%. However, it struggles with the minority class (Class 1), achieving a moderate specificity of 54.29% and a lower precision (negative predictive value) of 63.33%.

**Gradient Boosting Results:**

* **Class 0 (Negative Class):**
  + Correctly predicted (True Negatives): 70
  + Misclassified as Class 1 (False Positives): 12
* **Class 1 (Positive Class):**
  + Correctly predicted (True Positives): 18
  + Misclassified as Class 0 (False Negatives): 17

The model achieves a moderate accuracy (75.21%) and performs well in detecting the majority class (Class 0), with high sensitivity (85.37%) and good precision (80.46%). We can deduce that the model is effective at identifying non-diabetic cases, making it reliable for initial screening to rule out individuals unlikely to have diabetes.

Bar charts are highly effective for providing clarity and insight when comparing data analysis results. The comparison of model performance involved evaluating key metrics, including F1-score and accuracy, as depicted in figure 9.

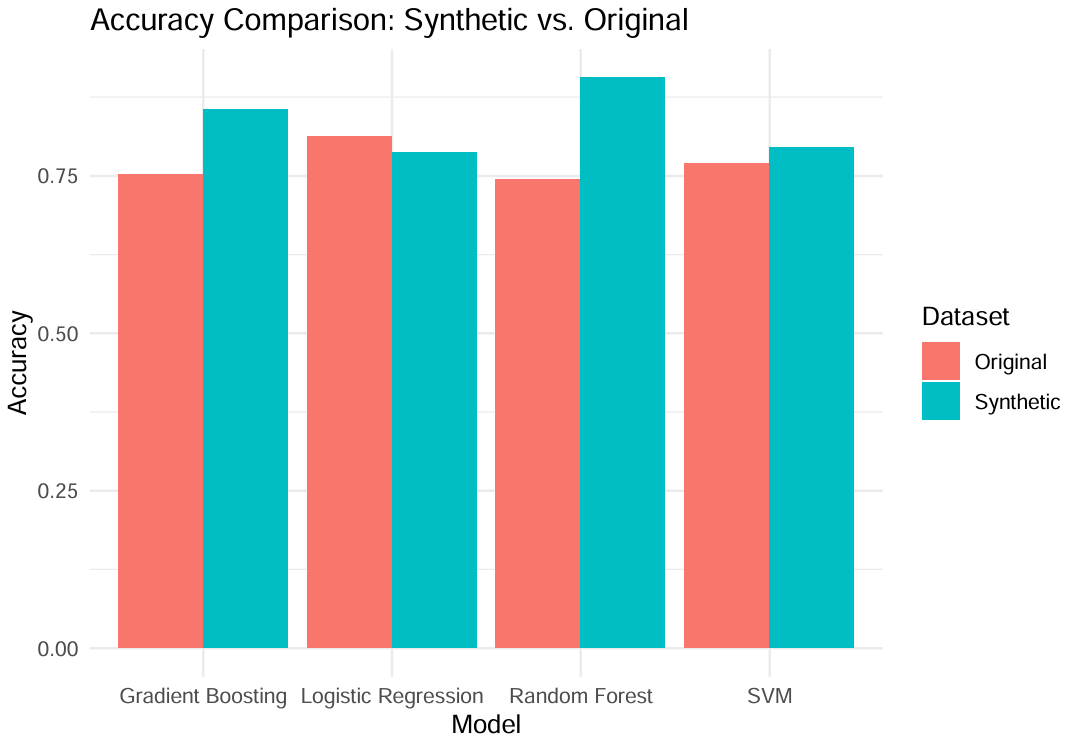


**Figure 9: F1-Score Comparison of Models Trained on Original vs. Synthetic Datasets**

Random Forest demonstrates the most significant improvement, suggesting it benefits the most from the synthetic data, potentially due to enhanced representation of minority class instances, consistent with findings by Hennebelle et al. (2023).

Gradient Boosting and Logistic Regression show moderate improvements, indicating that while the synthetic data enhances their balance of precision and recall, the benefit is less pronounced compared to Random Forest.

For SVM, while its improvement is less pronounced than Random Forest, it still achieves higher F1-scores with synthetic data, indicating better generalization across classes.



**Figure 10: Accuracy Comparison of Models Trained on Synthetic vs. Original Datasets**

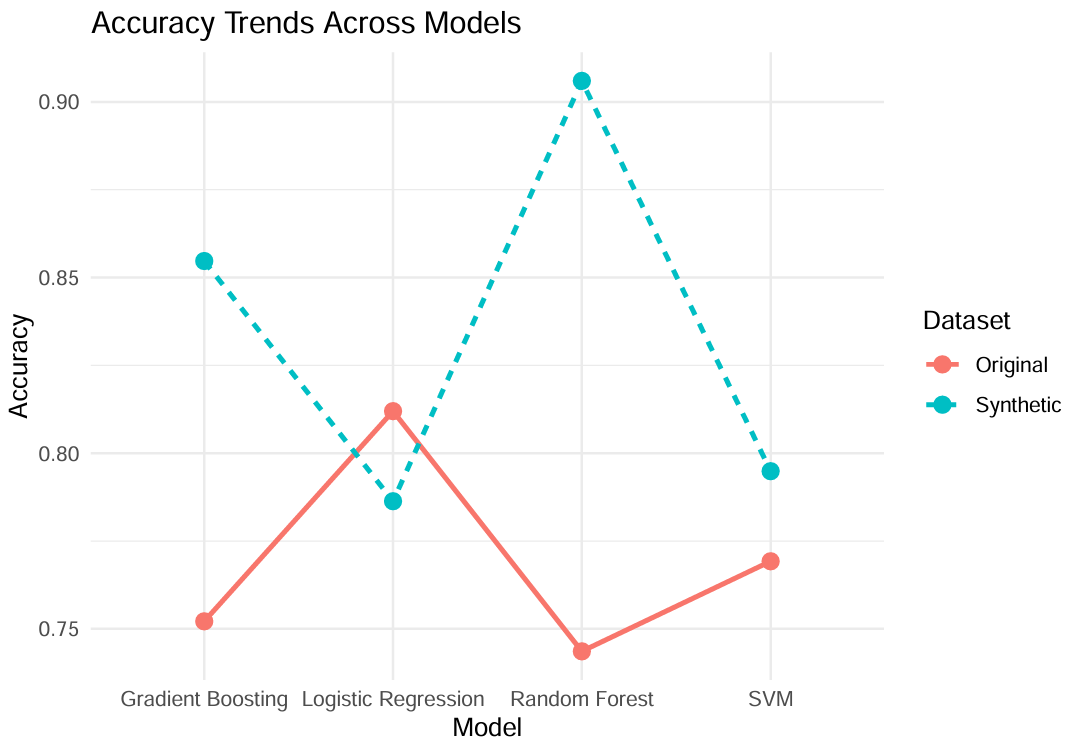
As depicted in figure 10, **Random Forest** shows the largest improvement in accuracy when trained on synthetic data, emphasizing its effectiveness in leveraging the added information from synthetic examples

**Gradient Boosting and Logistic Regression** also benefit significantly from synthetic data, showing noticeable accuracy gains, aligning with comparative results from Kowsher et al. (2023).

**SVM** demonstrates moderate improvement, indicating that while synthetic data enhances its performance, the impact is less pronounced compared to Random Forest.

The consistent improvements across all models highlight the utility of synthetic data in boosting overall classification performance.

The results suggest that synthetic data effectively augments training datasets, leading to better model generalization and reliability.



**Figure 11: Accuracy Trends Across Models Trained on Original vs. Synthetic Datasets**

**Gradient Boosting**: Shows moderate accuracy improvement with synthetic data compared to the original dataset.

**Logistic Regression**: Achieves higher accuracy than other models when trained on the synthetic dataset, indicating its effectiveness in leveraging synthetic data.

**Random Forest**: Experiences the most significant improvement in accuracy with synthetic data, highlighting its ability to utilize enhanced training data.

**SVM**: Demonstrates an improvement with synthetic data but lags slightly behind Random Forest and Logistic Regression in terms of overall accuracy.

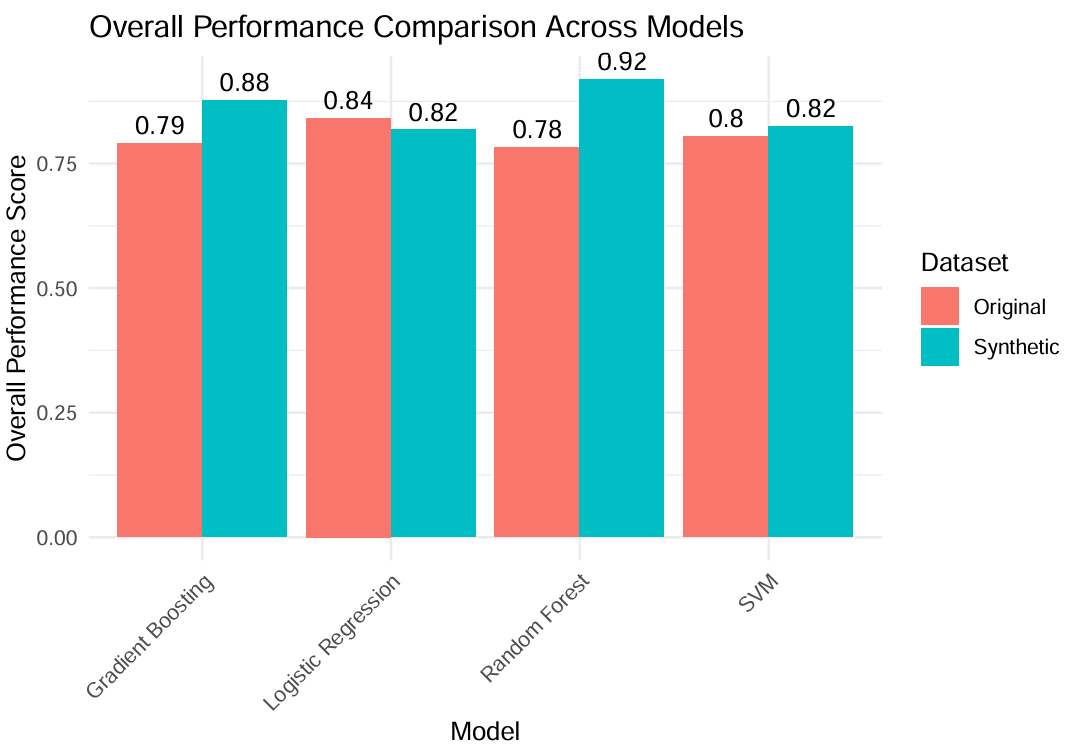
Synthetic data provides a clear performance boost across all models, with Random Forest and Logistic Regression benefiting the most. The graph highlights that synthetic data helps overcome limitations in the original dataset, such as class imbalance or insufficient representation of minority classes.



**Figure 12: F1-Score Trends Across Models Trained on Original vs. Synthetic Datasets**

Random Forest achieves the highest F1-score with synthetic data (~0.93), showing the largest improvement compared to its performance on the original dataset.

The SVM model exhibits minimal improvement (~2%), indicating it gains less advantage from synthetic data relative to other models. F1-scores are consistently higher across all models when synthetic data is used for training compared to the original dataset. This demonstrates that synthetic data improves the models’ ability to balance precision and recall, especially in handling imbalanced datasets.



**Figure 13: Overall Performance Score Comparison Across Models Trained on Original vs. Synthetic Datasets.**

From figure 13 above, it can be deduced that **Random Forest** achieves the highest performance score with synthetic data (**0.92**), showing the most significant improvement. The consistent improvement across all models highlights the effectiveness of synthetic data in enhancing model performance.

The greatest relative improvement is observed in Random Forest, while SVM benefits the least from synthetic data. Synthetic data significantly improves overall performance scores across all models, with Random Forest seeing the largest benefit. This emphasizes the value of synthetic datasets in addressing limitations such as class imbalance or insufficient data representation.

**CONCLUSION**

This study showcases the capability of supervised classification algorithms for early diabetes detection, emphasizing the significance of machine learning models in healthcare. The problem of diabetes, with its rising prevalence and significant health implications, underscores the need for innovative diagnostic tools. Through rigorous preprocessing, feature engineering, and the use of synthetic data, this study addressed critical challenges such as class imbalance and missing data, achieving improved model performance.

Random Forest emerged as the top-performing algorithm among those assessed, delivering the best accuracy, F1-score, and overall results when using synthetic data. This reflects its strong ability to leverage enhanced data representation for robust predictions. Gradient Boosting and Logistic Regression displayed significant enhancements, whereas SVM achieved moderate progress. These results highlight the necessity of choosing suitable algorithms based on the specific problem and data attributes.

Key insights from this study include the utility of synthetic datasets in enhancing machine learning model performance. By reducing false negatives and improving sensitivity, synthetic data strengthens the reliability of diagnostic tools, especially in imbalanced and incomplete datasets. However, critical assumptions include the generalizability of synthetic data to real-world scenarios and the need for robust validation to ensure model scalability.

The results have important implications for healthcare applications. With Random Forest achieving a near-perfect F1-score, this model demonstrates its feasibility for real-world clinical use. However, limitations such as the need for consistent data quality and handling outliers must be addressed before large-scale implementation. These findings suggest that adopting synthetic data and rigorous preprocessing techniques can significantly enhance machine learning models' reliability and diagnostic accuracy in clinical applications.

Future work could explore ensemble techniques, advanced sampling methods, and additional algorithms for improving minority class detection. Moreover, deploying these models in real-world clinical environments will help validate their scalability and robustness. By bridging the gap between theoretical research and practical applications, this study establishes a foundation for applying machine learning to enhance healthcare outcomes, enabling precise and timely diabetes diagnoses with reduced errors.

**REFERENCES**

[1] García-García, J. (2021). Machine learning and deep learning predictive models for type 2 diabetes: A systematic review. *Diabetology & Metabolic Syndrome, 13*(148). <https://doi.org/10.1186/s13098-021-00767-9>

[2] Hennebelle, A., Materwala, H., & Ismail, L. (2023). HealthEdge: A machine learning-based smart healthcare framework for prediction of type 2 diabetes in an integrated IoT, edge, and cloud computing system. *arXiv preprint arXiv:2301.10450*. <https://arxiv.org/abs/2301.10450>

[3] Kowsher, M., Turaba, M. Y., Sajed, T., & Rahman, M. M. M. (2023). Prognosis and treatment prediction of type-2 diabetes using deep neural network and machine learning classifiers. *arXiv preprint arXiv:2301.03093*. <https://arxiv.org/abs/2301.03093>

[4] Razzaghi, T., Roderick, O., Safro, I., & Marko, N. (2016). Multilevel weighted support vector machine for classification on healthcare data with missing values. *arXiv preprint arXiv:1604.02123*. <https://arxiv.org/abs/1604.02123>

[5] Rahman, M. M., & Davis, D. N. (2013). Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing, 3*(2), 224–228. <https://doi.org/10.7763/IJMLC.2013.V3.307>

[6] Ghosh, S., Baranowski, E. S., Biehl, M., Arlt, W., Tino, P., & Bunte, K. (2022). Interpretable models capable of handling systematic missingness in imbalanced classes and heterogeneous datasets. *arXiv preprint arXiv:2206.02056*. <https://arxiv.org/abs/2206.02056>