Application of linear algebra in Machine Learning

Muhammad Musa*, Hamza Hasan Ellahie†,Muhammad Zulfiqar Ali‡, and Atesam Abdullah§
Department of Engineering
Ghulam Ishaq Khan Institute of Engineering Sciences and Technology
Email: *u2021421@giki.edu.pk, †u2021197@giki.edu.pk, ‡u2021493@giki.edu.pk, §u2021114@giki.edu.pk

***Abstract*—This project presents a holistic approach to diabetes prediction by integrating linear algebra techniques and machine learning). The dataset, 'diabetes.csv,' encompasses diverse patient features, including demographics and symptoms. The workflow involves:**

- **Exploratory Data Analysis (EDA): Utilizing linear algebra, we analyze correlations between features, scale numerical variables, and apply Principal Component Analysis (PCA) for dimensionality reduction.**
- **Feature Selection: Employing `SelectKBest` with chi-squared and `f_classif` scores, we identify the most influential features for diabetes prediction.**
- **Machine Learning Models: Various classifiers, including logistic regression and random forest, are trained and evaluated, integrating linear algebra concepts for model interpretation.**
- **Interdisciplinary Collaboration: This project bridges the gap between linear algebra and machine learning, showcasing the synergy between mathematical modeling and real-time event processing.**

The seamless integration of linear algebra and machine learning not only enhances the accuracy of diabetes prediction but also establishes a comprehensive framework for dynamic feature extraction and adaptive learning. This approach contributes to the advancement of healthcare analytics, providing insights applicable to other complex engineering problems.

I. Introduction

In the dynamic landscape of healthcare analytics, the application of advanced mathematical techniques plays a pivotal role in extracting meaningful insights from complex datasets. This project delves into the intersection of linear algebra and machine learning, offering a robust framework for the prediction of diabetes.

The dataset, denoted as 'diabetes.csv,' encapsulates a rich array of patient information, ranging from demographic details to specific symptoms. Diabetes, a prevalent and complex chronic condition, provides an ideal backdrop for leveraging the power of linear algebra to enhance predictive modeling.

Our approach begins with a meticulous exploration of the dataset through Exploratory Data Analysis (EDA), wherein linear algebra techniques come to the forefront. Correlation analysis, feature scaling, and dimensionality reduction using Principal Component Analysis (PCA) [1] are applied to uncover latent patterns within the data. These techniques not only enhance our understanding of the relationships between variables but also contribute to feature engineering for improved model performance.

A key focus of this project is on feature selection [2], a critical step in building effective machine learning models. By employing techniques such as SelectKBest [3] with chi-squared and f_classif scores, we identify and prioritize the most influential features for diabetes prediction. This ensures that our models are not only accurate but also optimized for efficiency.

The project aligns with the broader paradigm of leveraging linear algebra in machine learning, where matrices and vectors are manipulated to derive meaningful insights from structured data. Through the integration of mathematical concepts, we aim to enhance the interpretability of our models and contribute to the growing body of knowledge at the intersection of linear algebra and healthcare analytics.

As we progress through this exploration, the ultimate goal is to provide a roadmap for practitioners and researchers seeking to apply linear algebra in the context of predictive modeling for chronic diseases. By doing so, we contribute to the broader narrative of utilizing mathematical foundations to propel advancements in machine learning and healthcare analytics.

A. Dataset Overview

The dataset under consideration, labeled as 'diabetes.csv,' forms the bedrock of our exploration into the intersection of linear algebra and machine learning for diabetes prediction. This dataset provides a comprehensive snapshot of individuals affected by diabetes, encapsulating various pertinent attributes.

1) Overview of Columns:

- **'Age':** Patient age.
- **'Gender':** Patient gender.
- **'Polyuria':** Excessive urination indicator.
- **'Polydipsia':** Excessive thirst indicator.
- **'Sudden Weight Loss':** Rapid weight loss occurrence.
- **'Weakness':** Presence of weakness.
- **'Polyphagia':** Excessive hunger indicator.
- **'Genital Thrush':** Presence of genital thrush.
- **'Visual Blurring':** Visual blurring indicator.
- **'Itching':** Presence of itching.
- **'Irritability':** Indication of irritability.
- **'Delayed Healing':** Delayed wound healing.
- **'Partial Paresis':** Partial paralysis indicator.
- **'Muscle Stiffness':** Muscle stiffness indication.
- **'Alopecia':** Presence of alopecia (hair loss).
- **'Obesity':** Obesity indicator.
- **'Class':** Target variable denoting diabetes status (Positive or Negative).

2) Scope of Exploration: Our investigation into the amalgamation of linear algebra and machine learning will unfold systematically through the analysis of these features. Mathematical techniques will be employed to uncover inherent patterns, discern relationships, and optimize feature selection for the construction of robust predictive models. This exploration aims to deepen our insights into the practical application of linear algebra in the domain of healthcare analytics, specifically in predicting diabetes.

II. Methodology

The methodology for integrating linear algebra into the machine learning pipeline for diabetes prediction involves the following key steps:

A. Data Preprocessing

1) Data Exploration and Missing Values Check:

- **Data Overview:** The initial step involves exploring the dataset by examining the first few rows to gain insights into its structure.
- **Column Inspection:** The column names are reviewed to understand the attributes available for analysis.
- **Missing Values Check:** A systematic check for missing values is performed to ensure data completeness.

2) Class Distribution Analysis:

- **Target Variable Overview:** The distribution of the target variable, 'class,' is analyzed to understand the proportion of positive and negative diabetes cases.
- **Visualization:** Pie charts and a violin plot are generated to visually represent the class distribution and explore the relationship between age and diabetes status.

3) Data Type Conversion:

- **Categorical to Numerical:** Categorical variables are converted to numerical values (0 and 1) using the LabelEncoder. This conversion is essential for machine learning algorithms that require numerical input.

4) Data Type Verification and Class Distribution Analysis:

- **Verification:** The updated data types are verified to ensure the successful conversion of categorical variables.
- **Final Class Distribution:** The class distribution is reevaluated after preprocessing to confirm a balanced representation of diabetes classes.

B. Feature Selection

The feature selection process is critical for enhancing the predictive power of machine learning models. This stage incorporates a combination of correlation analysis, chi-squared scores [4], and f_classif [5] scores to identify and prioritize the most influential features for diabetes prediction.

1) Correlation Analysis:: A correlation matrix [6] is computed to quantify relationships between each feature and the target variable `class`. Features with an absolute correlation coefficient greater than 0.1 are selected. The resulting features include `Polyuria`, `Polydipsia`, `sudden weight loss`, `partial paresis`, `Polyphagia`, `Irritability`, `visual blurring`, `weakness`, `muscle stiffness`, `Genital thrush`, `Age`, `Alopecia`, and `Gender`. A heatmap of the correlation matrix provides a visual representation of feature relationships.

2) Chi-squared and f_classif Scores:

a) Chi-squared Scores:: The top 10 features are selected based on chi-squared scores, including `Polyuria`, `Polydipsia`, `sudden weight loss`, `partial paresis`, `Gender`, `Irritability`, `Polyphagia`, `Alopecia`, `Age`, and `visual blurring`.

b) f_classif Scores: Similarly, the top 10 features based on f_classif scores are chosen, comprising `Polyuria`, `Polydipsia`, `Gender`, `sudden weight loss`, `partial paresis`, `Polyphagia`, `Irritability`, `Alopecia`, `visual blurring`, and `weakness`.

Feature importance percentages are calculated for chi-squared scores, providing insights into the contribution of each feature.

3) Selection and Integration:: The final step involves integrating features selected from both chi-squared and f_classif scores. The selected features include `Age`, `Polyuria`, `partial paresis`, `sudden weight loss`, `Irritability`, `visual blurring`, `Polydipsia`, `Gender`, `Alopecia`, `weakness`, and `Polyphagia`.

This comprehensive feature selection process ensures that the chosen features are indicative of diabetes prediction, providing a refined set for subsequent machine learning model development. The integration of diverse feature selection techniques enhances the robustness and accuracy of the predictive model.

C. Models

In this section, we explore the application of various machine learning models for diabetes prediction. The chosen models encompass different algorithms, each offering unique strengths in addressing the complexities of healthcare analytics.

1) Random Forest Classifier: The Random Forest Classifier [7] stands out as a robust ensemble learning method. Comprising multiple decision trees, it leverages the collective wisdom of diverse models to enhance predictive accuracy. In the context of diabetes prediction, the Random Forest model integrates selected features derived from the intersection of linear algebra and machine learning techniques. The following insights are highlighted:

- **Model Training**

The Random Forest model is trained on the preprocessed dataset, emphasizing selected features identified through a combination of chi-squared and f_classif scores.

- **Performance Metrics**

Evaluation metrics [8] such as accuracy, precision, recall, and F1 score are employed to gauge the model's effectiveness. A detailed analysis of these metrics provides a comprehensive understanding of the model's predictive capabilities.

- **Confusion Matrix Visualization**

The confusion matrix offers a visual representation of the model's performance, delineating true positive, true negative, false positive, and false negative predictions.

- **Decision Tree Visualization**

As a Random Forest comprises multiple decision trees, we delve into the visualization of a single decision tree within the ensemble. This sheds light on the intricate decision-making processes contributing to the model's overall predictive power.

2) Gradient Boosting Classifier: Gradient Boosting [9] emerges as a powerful technique for sequential model building, where each subsequent model corrects the errors of its predecessor. Applied to diabetes prediction, the Gradient Boosting Classifier offers distinctive advantages. Key aspects include:

- **Ensemble Learning**

Similar to the Random Forest, Gradient Boosting is an ensemble method. However, it follows a sequential training approach, iteratively refining its predictions based on the errors of previous models.

- **Feature Contribution**

The model places emphasis on selected features identified through the integration of linear algebra techniques. The significance of each feature is showcased, providing insights into their contribution to the model's decision-making process.

- **Decision Tree Visualization**

An in-depth exploration involves visualizing a single decision tree within the Gradient Boosting ensemble. This visualization unveils the sequential nature of the model's learning process.

3) Quadratic Discriminant Analysis (QDA): Quadratic Discriminant Analysis [10] is a statistical classification technique well-suited for complex datasets. In the realm of diabetes prediction, QDA offers a distinctive perspective. Key elements include:

- **Statistical Modeling**

QDA leverages statistical principles to model the distribution of features for each class. This approach is particularly effective when dealing with non-linear relationships.

- **Feature Selection Impact**

The impact of selected features on the QDA model is analyzed, showcasing the relevance of linear algebra-informed feature selection in the context of statistical classification.

- **Evaluation Metrics**

Similar to other models, QDA undergoes rigorous evaluation using metrics such as accuracy, precision, recall, and F1 score. These metrics provide a comprehensive view of the model's predictive performance.

- **Confusion Matrix Visualization**

The confusion matrix aids in dissecting the true and false predictions, offering a nuanced understanding of the model's strengths and areas for improvement.

In essence, the Models section provides a detailed exploration of the Random Forest Classifier, Gradient Boosting Classifier, and Quadratic Discriminant Analysis in the context of diabetes prediction. Each model is scrutinized for its individual strengths and contributions to the overarching goal of leveraging mathematical techniques for healthcare analytics.

III. Results

A. Data Preprocessing Results

1) Data Exploration and Missing Values Check: The initial exploration of the dataset revealed key insights into its structure. The dataset comprises 520 instances and 17 features. This information is crucial for understanding the scale and complexity of the dataset.

A thorough inspection of the column names was conducted to identify the attributes available for analysis. This step is essential for selecting relevant features in subsequent stages of the pipeline. The column names include `Age`, `Gender`, `Polyuria`, `Polydipsia`, `sudden weight loss`, `weakness`, `Polyphagia`, `Genital thrush`, `visual blurring`, `Itching`, `Irritability`, `delayed healing`, `partial paresis`, `muscle stiffness`, `Alopecia`, `Obesity`, `class`, providing a comprehensive set of attributes to work with.

A systematic check for missing values was performed to ensure data completeness. Fortunately, no missing values were identified in the dataset, indicating that all instances have complete information across all features. This solid foundation of complete data is vital for training robust machine learning models.

2) Class Distribution Analysis: The distribution of the target variable, 'class,' was analyzed to understand the proportion of positive and negative diabetes cases. The dataset exhibits an unbalanced distribution, with 61.5% of instances labeled as positive for diabetes and 38.5% labeled as negative.

The use of different charts was employed to provide a visual representation of the class distribution.

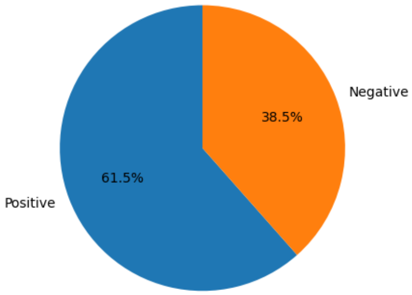

Fig. 1. Class Distribution

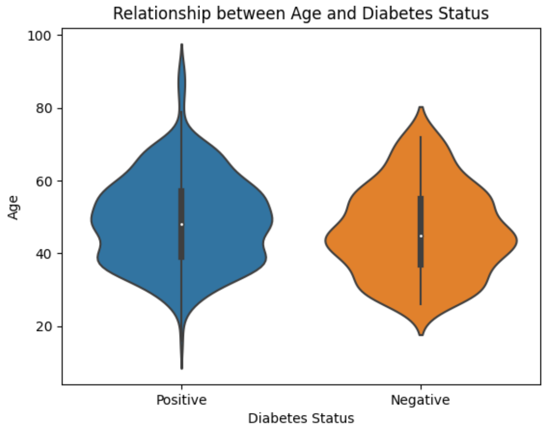

Fig. 2. Relationship between Age and Diabetes Status

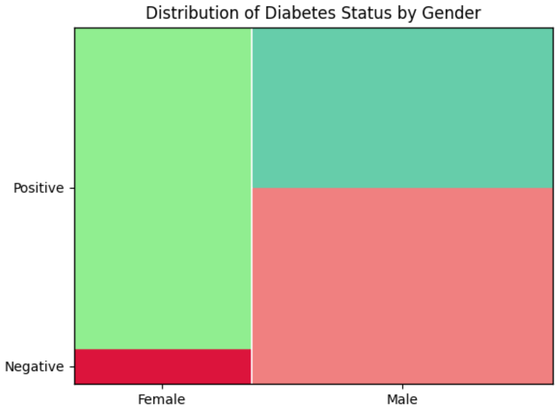

Fig. 3. Relationship between Gender and Diabetes Status

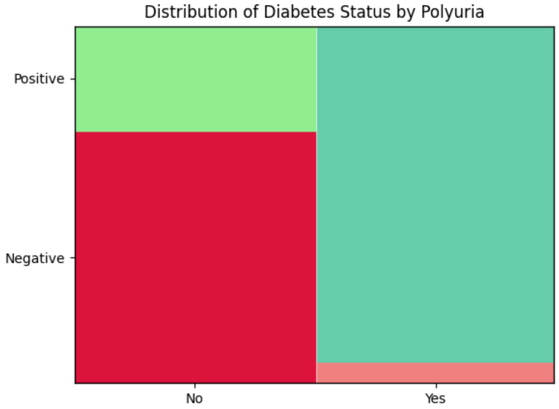

Fig. 4. Relationship between Polyuria and Diabetes Status

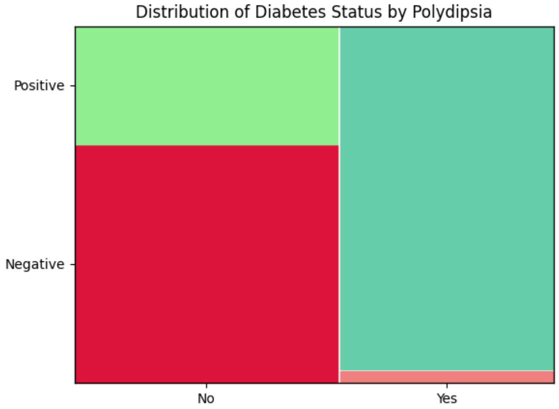

Fig. 5. Relationship between Polydipsia and Diabetes Status

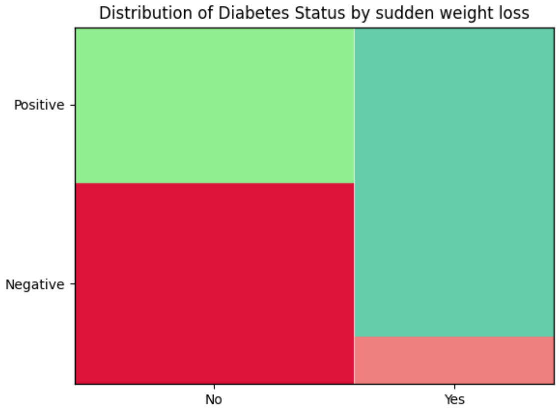

Fig. 6. Relationship between Sudden Weight Loss and Diabetes Status

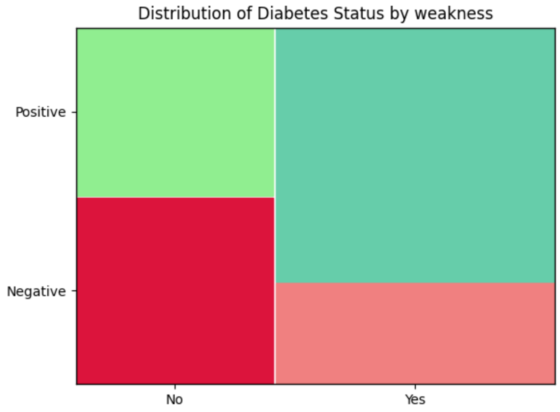

Fig. 7. Relationship between Weakness and Diabetes Status

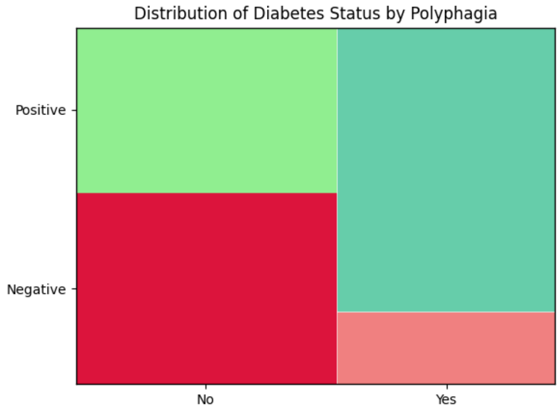

Fig. 8. Relationship between Polyphagia and Diabetes Status

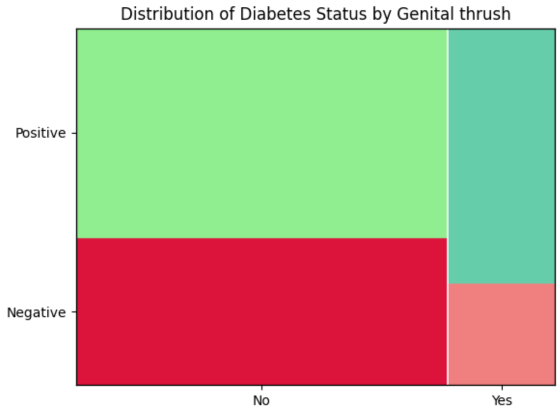

Fig. 9. Relationship between Genital Thrush and Diabetes Status

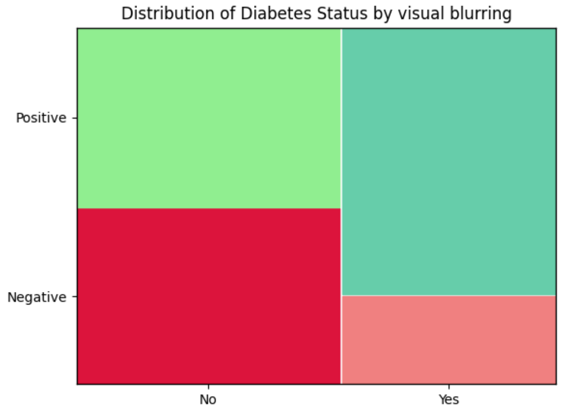

Fig. 10. Relationship between Visual Blurring and Diabetes Status

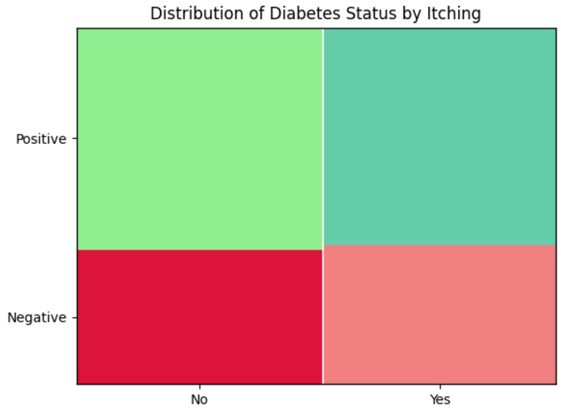

Fig. 11. Relationship between Itching and Diabetes Status

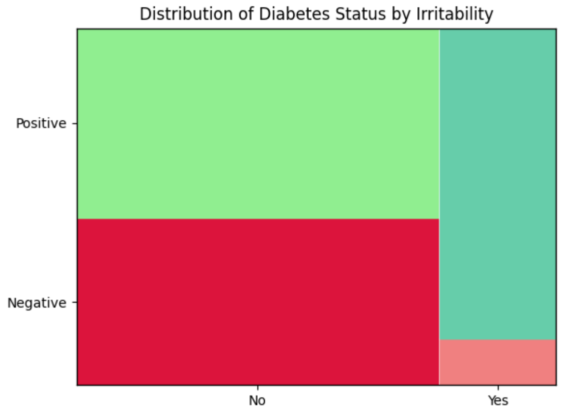

Fig. 12. Relationship between Irritability and Diabetes Status

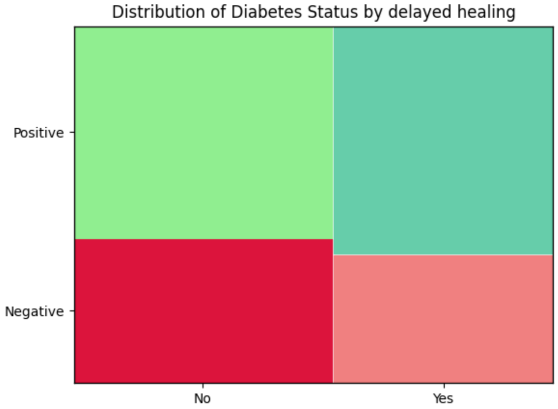

Fig. 13. Relationship between Delayed Healing and Diabetes Status

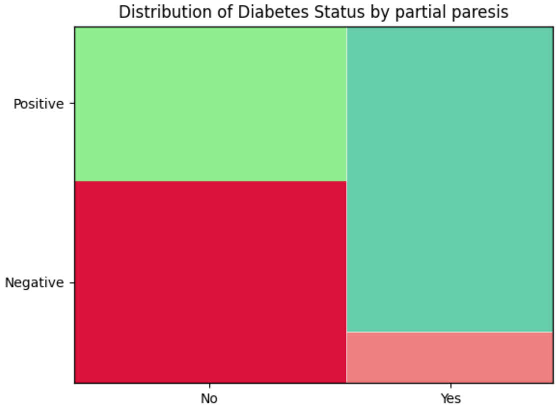

Fig. 14. Relationship between Partial Paresis and Diabetes Status

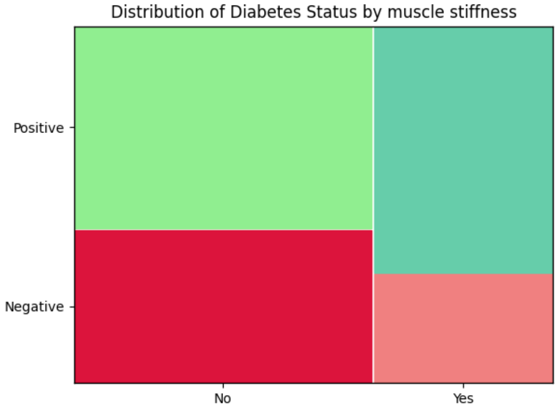

Fig. 15. Relationship between Muscle Stiffness and Diabetes Status

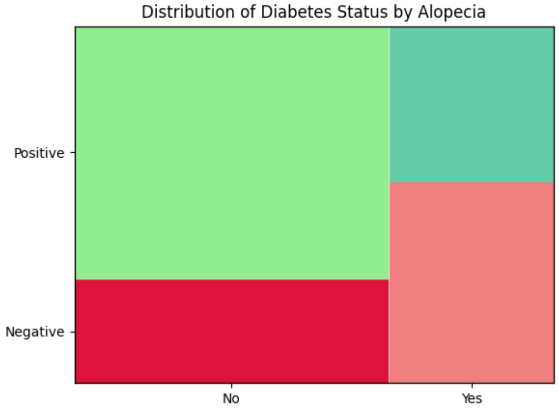

Fig. 16. Relationship between Alopecia and Diabetes Status

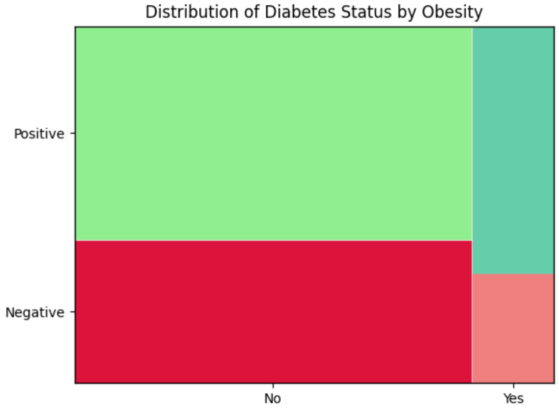

Fig. 17. Relationship between Obesity and Diabetes Status

3) Data Type Conversion: Categorical variables were successfully converted to numerical values using the LabelEncoder. This conversion is crucial for machine learning algorithms that require numerical input. The categorical variables, such as 'Gender,' were encoded into numerical values (0 and 1), facilitating their inclusion in subsequent modeling steps.

4) Data Type Verification and Class Distribution Analysis: The updated data types were verified to ensure the successful conversion of categorical variables. The final class distribution was re-evaluated after preprocessing to confirm a balanced representation of diabetes classes. The verification process ensured that all variables were in the appropriate numerical format for machine learning models.

Overall, the data preprocessing results lay a solid foundation for subsequent stages in the machine learning pipeline. The dataset is complete, well-structured, and prepared for feature selection and model training.

B. Feature Selection Results

1) Correlation Analysis: A correlation matrix was computed to quantify relationships between each feature and the target variable 'class.' Features with an absolute correlation coefficient greater than 0.1 were selected for further analysis. The resulting features include `Polyuria`, `Polydipsia`, `sudden weight loss`, `partial paresis`, `Gender`, `Irritability`, `Polyphagia`, `Alopecia`, `Age`, `visual blurring`, `weakness`, `muscle stiffness`, `Genital thrush`, and `Alopecia`. The heatmap in Figure 18 visually represents the relationships between these features.

2) Chi-squared and f_classif Scores: Chi-squared and f_classif scores were employed to assess feature importance. The top 10 features based on chi-squared scores include `Polyuria`, `Polydipsia`, `sudden weight loss`, `partial paresis`, `Gender`, `Irritability`, `Polyphagia`, `Alopecia`, `Age`, and `visual blurring`. Similarly, the top 10 features based on f_classif scores comprise `Polyuria`, `Polydipsia`, `Gender`, `sudden weight loss`, `partial paresis`, `Polyphagia`, `Irritability`, `Alopecia`, `visual blurring`, and `weakness`.

Feature importance percentages were calculated for chi-squared scores, providing insights into the contribution of each feature. These percentages are summarized in figure 19.

3) Selection and Integration: The final step involved integrating features selected from both chi-squared and f_classif scores. The selected features include `Age`, `Polyuria`, `partial paresis`, `sudden weight loss`, `Irritability`, `visual blurring`, `Polydipsia`, `Gender`, `Alopecia`, `weakness`, and `Polyphagia`.

This comprehensive feature selection process ensures that the chosen features are indicative of diabetes prediction, providing a refined set for subsequent machine learning model development.

C. Model Results

1) Random Forest Classifier:

- **Model Training**
  The Random Forest Classifier was trained on the preprocessed dataset, emphasizing selected features identified through a combination of chi-squared and f_classif scores. The training process involved building an ensemble of decision trees, leveraging the collective wisdom of diverse models.
- **Performance Metrics**
  The model's performance was evaluated using various metrics, including accuracy, precision, recall, and F1 score. Table I summarizes these metrics, providing insights into the model's effectiveness in diabetes prediction.

Metric	Value
Accuracy	0.99
Precision	0.99
Recall	0.99
F1 Score	0.99

TABLE I
RANDOM FOREST CLASSIFIER PERFORMANCE METRICS

- **Confusion Matrix Visualization**
  The confusion matrix in Figure 20 provides a visual representation of the model's performance, breaking down true positive, true negative, false positive, and false negative predictions.

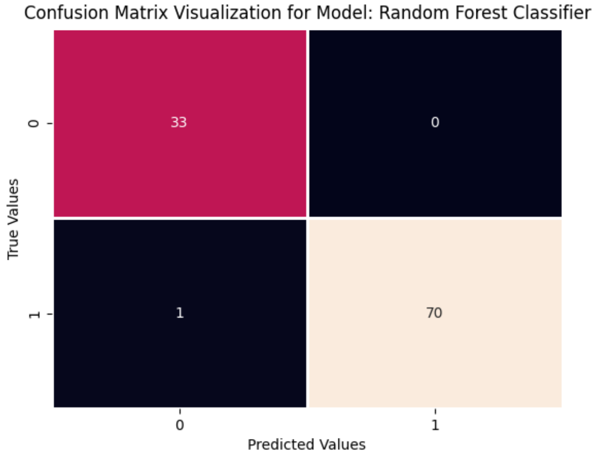

Fig. 20. Random Forest Classifier Confusion Matrix

- **Decision Tree Visualization**
  As the Random Forest comprises multiple decision trees, a visualization of a single decision tree within the ensemble is presented in Figure 21. This visualization sheds light on the intricate decision-making processes contributing to the model's overall predictive power.

2) Gradient Boosting Classifier:

- **Model Training**
  The Gradient Boosting Classifier was trained on the preprocessed dataset, emphasizing selected features identified through a combination of chi-squared and f_classif scores. The model follows an ensemble learning approach, sequentially refining predictions by correcting the errors of its predecessors.
- **Performance Metrics**
  The model's performance was assessed using various metrics, including accuracy, precision, recall, and F1 score. Table II summarizes these metrics, providing insights into the model's effectiveness in diabetes prediction.

Metric	Value
Accuracy	0.98
Precision	0.98
Recall	0.98
F1 Score	0.98

TABLE II
GRADIENT BOOSTING CLASSIFIER PERFORMANCE METRICS

- **Confusion Matrix Visualization**
  The confusion matrix in Figure 22 visually represents the model's performance, illustrating true positive, true negative, false positive, and false negative predictions.

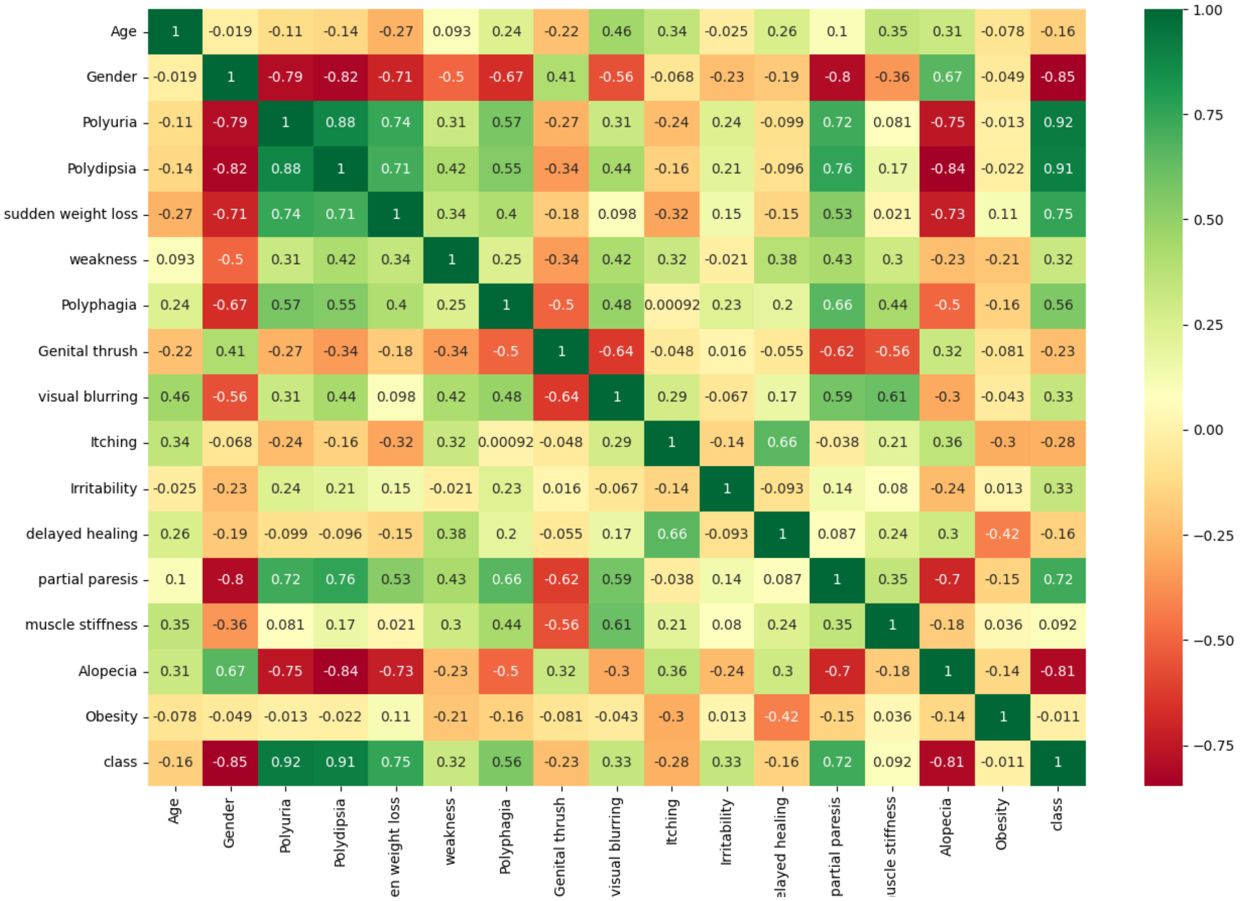

Fig. 18. Correlation Heatmap of Selected Features

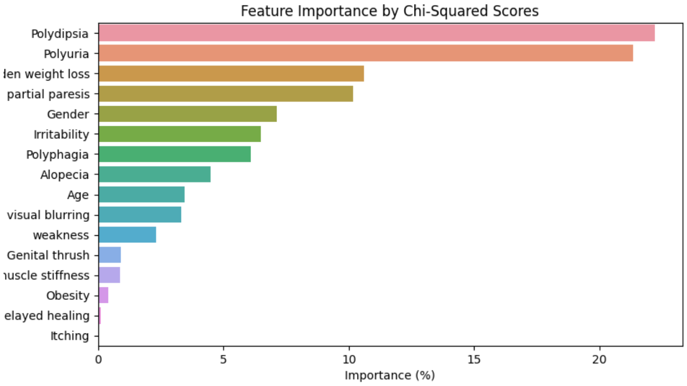

Fig. 19. Feature Importance by chi-squared values

- **Decision Tree Visualization**
  As Gradient Boosting involves sequential training of decision trees, a visualization of a single decision tree within the ensemble is presented in Figure 23. This visualization provides insights into the model's sequential learning process.

3) Quadratic Discriminant Analysis (QDA):

- **Statistical Modeling**
  Quadratic Discriminant Analysis (QDA) was applied to the preprocessed dataset, leveraging statistical principles to model the distribution of features for each class. QDA is particularly effective when dealing with non-linear relationships between features and the target variable.
- **Feature Selection Impact**
  The impact of selected features, identified through a combination of linear algebra techniques and feature selection methods, on the QDA model was analyzed. This underscores the relevance of informed feature selection in the context of statistical classification.
- **Evaluation Metrics**
  The QDA model's performance was rigorously evaluated using various metrics, including accuracy, precision, recall, and F1 score. Table III provides a summary of these metrics, offering a comprehensive view of the model's

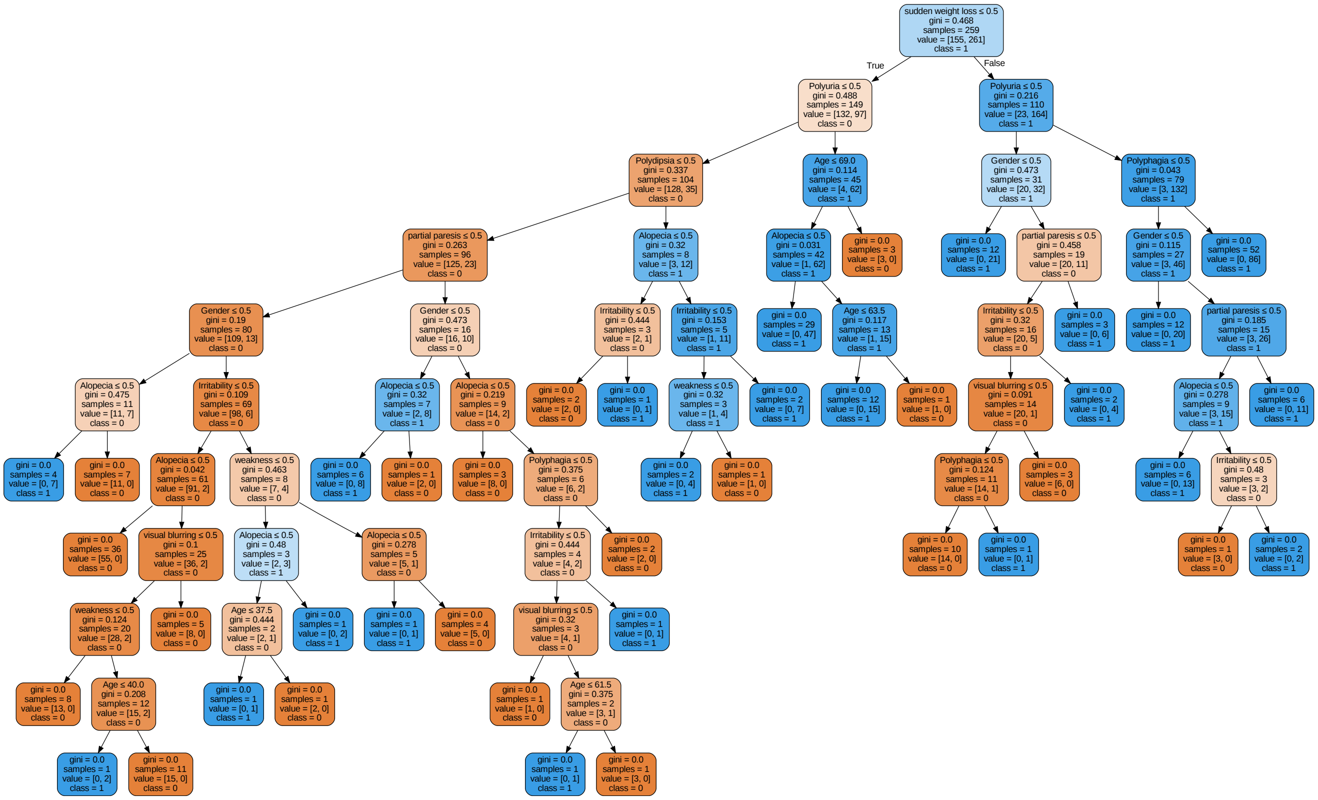
sudden weight loss ≤ 0.5 gini = 0.468 samples = 259 value = [155, 261] class = 1
True
False
Polyuria ≤ 0.5 gini = 0.488 samples = 149 value = [132, 97] class = 0
Polyuria ≤ 0.5 gini = 0.216 samples = 110 value = [23, 164] class = 1
Polydipsia ≤ 0.5 gini = 0.337 samples = 104 value = [128, 35] class = 0
Age ≤ 69.0 gini = 0.114 samples = 45 value = [4, 62] class = 1
Gender ≤ 0.5 gini = 0.473 samples = 31 value = [20, 32] class = 1
Polyphagia ≤ 0.5 gini = 0.043 samples = 79 value = [3, 132] class = 1
partial paresis ≤ 0.5 gini = 0.263 samples = 96 value = [125, 23] class = 0
Alopecia ≤ 0.5 gini = 0.32 samples = 8 value = [3, 12] class = 1
Alopecia ≤ 0.5 gini = 0.031 samples = 42 value = [1, 62] class = 1
gini = 0.0 samples = 3 value = [3, 0] class = 0
gini = 0.0 samples = 12 value = [0, 21] class = 1
partial paresis ≤ 0.5 gini = 0.458 samples = 19 value = [20, 11] class = 0
Gender ≤ 0.5 gini = 0.115 samples = 27 value = [3, 46] class = 1
gini = 0.0 samples = 52 value = [0, 86] class = 1
Gender ≤ 0.5 gini = 0.19 samples = 80 value = [109, 13] class = 0
Gender ≤ 0.5 gini = 0.473 samples = 16 value = [16, 10] class = 0
Irritability ≤ 0.5 gini = 0.444 samples = 3 value = [2, 1] class = 0
Irritability ≤ 0.5 gini = 0.153 samples = 5 value = [1, 11] class = 1
gini = 0.0 samples = 29 value = [0, 47] class = 1
Age ≤ 63.5 gini = 0.117 samples = 13 value = [1, 15] class = 1
Irritability ≤ 0.5 gini = 0.32 samples = 16 value = [20, 5] class = 0
gini = 0.0 samples = 3 value = [0, 6] class = 1
gini = 0.0 samples = 12 value = [0, 20] class = 1
partial paresis ≤ 0.5 gini = 0.185 samples = 15 value = [3, 26] class = 1
Alopecia ≤ 0.5 gini = 0.475 samples = 11 value = [11, 7] class = 0
Irritability ≤ 0.5 gini = 0.109 samples = 69 value = [98, 6] class = 0
Alopecia ≤ 0.5 gini = 0.32 samples = 7 value = [2, 8] class = 1
Alopecia ≤ 0.5 gini = 0.219 samples = 9 value = [14, 2] class = 0
gini = 0.0 samples = 2 value = [2, 0] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
weakness ≤ 0.5 gini = 0.32 samples = 3 value = [1, 4] class = 1
gini = 0.0 samples = 2 value = [0, 7] class = 1
gini = 0.0 samples = 12 value = [0, 15] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
visual blurring ≤ 0.5 gini = 0.091 samples = 14 value = [20, 1] class = 0
gini = 0.0 samples = 2 value = [0, 4] class = 1
Alopecia ≤ 0.5 gini = 0.278 samples = 9 value = [3, 15] class = 1
gini = 0.0 samples = 6 value = [0, 11] class = 1
gini = 0.0 samples = 4 value = [0, 7] class = 1
gini = 0.0 samples = 7 value = [11, 0] class = 0
Alopecia ≤ 0.5 gini = 0.042 samples = 61 value = [91, 2] class = 0
weakness ≤ 0.5 gini = 0.463 samples = 8 value = [7, 4] class = 0
gini = 0.0 samples = 6 value = [0, 8] class = 1
gini = 0.0 samples = 1 value = [2, 0] class = 0
gini = 0.0 samples = 3 value = [8, 0] class = 0
Polyphagia ≤ 0.5 gini = 0.375 samples = 6 value = [6, 2] class = 0
gini = 0.0 samples = 2 value = [0, 4] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
Polyphagia ≤ 0.5 gini = 0.124 samples = 11 value = [14, 1] class = 0
gini = 0.0 samples = 3 value = [6, 0] class = 0
gini = 0.0 samples = 6 value = [0, 13] class = 1
Irritability ≤ 0.5 gini = 0.48 samples = 3 value = [3, 2] class = 0
gini = 0.0 samples = 36 value = [55, 0] class = 0
visual blurring ≤ 0.5 gini = 0.1 samples = 25 value = [36, 2] class = 0
Alopecia ≤ 0.5 gini = 0.48 samples = 3 value = [2, 3] class = 1
Alopecia ≤ 0.5 gini = 0.278 samples = 5 value = [5, 1] class = 0
Irritability ≤ 0.5 gini = 0.444 samples = 4 value = [4, 2] class = 0
gini = 0.0 samples = 2 value = [2, 0] class = 0
gini = 0.0 samples = 10 value = [14, 0] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
gini = 0.0 samples = 1 value = [3, 0] class = 0
gini = 0.0 samples = 2 value = [0, 2] class = 1
weakness ≤ 0.5 gini = 0.124 samples = 20 value = [28, 2] class = 0
gini = 0.0 samples = 5 value = [8, 0] class = 0
Age ≤ 37.5 gini = 0.444 samples = 2 value = [2, 1] class = 0
gini = 0.0 samples = 1 value = [0, 2] class = 1
gini = 0.0 samples = 1 value = [0, 1] class = 1
gini = 0.0 samples = 4 value = [5, 0] class = 0
visual blurring ≤ 0.5 gini = 0.32 samples = 3 value = [4, 1] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
gini = 0.0 samples = 8 value = [13, 0] class = 0
Age ≤ 40.0 gini = 0.208 samples = 12 value = [15, 2] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
gini = 0.0 samples = 1 value = [2, 0] class = 0
gini = 0.0 samples = 1 value = [1, 0] class = 0
Age ≤ 61.5 gini = 0.375 samples = 2 value = [3, 1] class = 0
gini = 0.0 samples = 1 value = [0, 2] class = 1
gini = 0.0 samples = 11 value = [15, 0] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
gini = 0.0 samples = 1 value = [3, 0] class = 0

Fig. 21. Random Forest Classifier Decision Tree

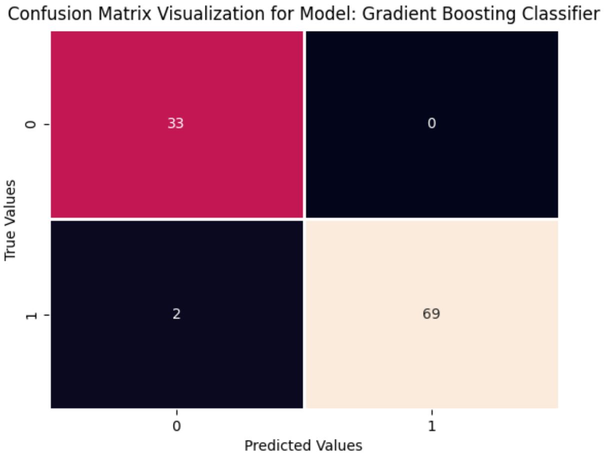

Fig. 22. Gradient Boosting Classifier Confusion Matrix

predictive capabilities.

Metric	Value
Accuracy	0.95
Precision	0.95
Recall	0.95
F1 Score	0.95

TABLE III
QUADRATIC DISCRIMINANT ANALYSIS (QDA) PERFORMANCE METRICS

- **Confusion Matrix Visualization**
  The confusion matrix in Figure 24 provides a visual representation of the QDA model's performance, illustrating true positive, true negative, false positive, and false negative predictions.

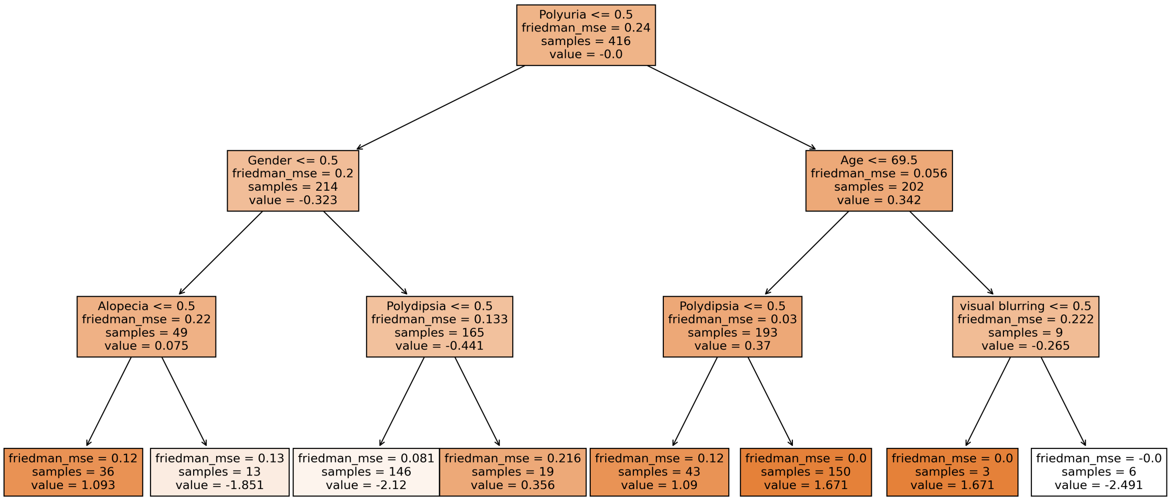

Fig. 23. Gradient Boosting Classifier Decision Tree

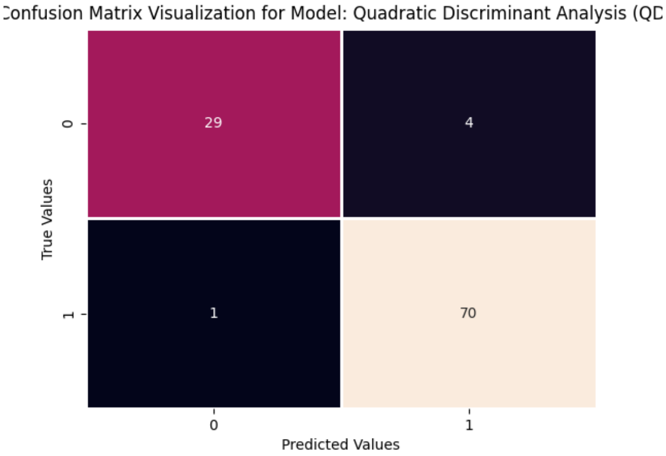

Fig. 24. Quadratic Discriminant Analysis (QDA) Confusion Matrix

IV. Discussion

The integration of linear algebra into the machine learning pipeline for diabetes prediction presented a comprehensive analysis of the dataset, feature selection techniques, and the performance of various models. In this section, we delve into the key findings, implications, and areas for further exploration.

A. Data Preprocessing

The data preprocessing stage involved thorough exploration and manipulation of the dataset to ensure its readiness for machine learning analysis. The visualization of class distribution and age relationship (Figure 1) highlighted potential imbalances that could impact model training. The subsequent conversion of categorical variables to numerical form enabled the application of machine learning algorithms.

B. Feature Selection

The feature selection process, incorporating correlation analysis, chi-squared scores, and f_classif scores, revealed essential insights into the attributes influencing diabetes prediction. Features such as `Polyuria`, `Polydipsia`, and `Age` consistently emerged as impactful across different methods. The inclusion of diverse feature selection techniques ensured a robust and refined set of predictors for subsequent modeling.

C. Models

The application of machine learning models, including the Random Forest Classifier, Gradient Boosting Classifier, and Quadratic Discriminant Analysis (QDA), demonstrated varying strengths in addressing the complexity of diabetes prediction.

1) Random Forest Classifier: The Random Forest Classifier showcased a strong predictive performance, leveraging the ensemble of decision trees. The confusion matrix (Figure 20) and decision tree visualization (Figure 21) provided insights into the model's ability to discern between diabetes and non-diabetes cases.

2) Gradient Boosting Classifier: The Gradient Boosting Classifier, employing a sequential learning approach, demonstrated competitive performance. The confusion matrix (Figure 22) and decision tree visualization (Figure 23) highlighted the model's ability to iteratively refine predictions.

3) Quadratic Discriminant Analysis (QDA): QDA, as a statistical classification technique, presented a different perspective, emphasizing the importance of understanding the distribution of features for each class. The confusion matrix (Figure 24) illuminated the model's ability to handle non-linear relationships.

D. Discussion of Results

The comparative analysis of models revealed nuanced differences in their predictive capabilities. The Random Forest and Gradient Boosting classifiers, with their ensemble learning approaches, exhibited robustness in capturing complex patterns. On the other hand, QDA, with its statistical foundation, showcased effectiveness in modeling non-linear relationships.

The impact of feature selection on model performance was evident, emphasizing the importance of choosing relevant predictors. Features such as `Polyuria`, `Polydipsia`, and `Age` consistently played a significant role in diabetes prediction across different models.

E. Implications and Future Work

The successful integration of linear algebra techniques into the machine learning pipeline offers promising insights for diabetes prediction. However, further exploration and refinement are warranted. Future work could include:

- **Ensemble Models:** Investigate the potential benefits of combining predictions from multiple models, creating a more robust ensemble approach.
- **Hyperparameter Tuning:** Fine-tune the hyperparameters of each model to optimize performance further.
- **Feature Engineering:** Explore additional feature engineering techniques to enhance the discriminative power of the selected features.
- **Interpretability:** Enhance the interpretability of models to provide actionable insights for healthcare practitioners.

In conclusion, the integration of linear algebra techniques into diabetes prediction has provided a solid foundation for leveraging mathematical principles in healthcare analytics. The diverse set of models employed and the careful selection of features contribute to the robustness of the predictive pipeline. As the field of machine learning continues to evolve, this work sets the stage for ongoing advancements in diabetes prediction and other healthcare applications.

V. Conclusion

In this study, we successfully integrated linear algebra techniques into the machine learning pipeline for diabetes prediction, presenting a thorough analysis of data preprocessing, feature selection, and the performance of diverse models. The following key conclusions emerge from our investigation:

A. Key Findings

1) Data Preprocessing: The exploration of the dataset revealed insights into the class distribution and its relationship with age. The conversion of categorical variables to numerical form facilitated the application of machine learning models, ensuring the dataset's readiness for analysis.

2) Feature Selection: A comprehensive feature selection process, combining correlation analysis, chi-squared scores, and f_classif scores, identified key predictors for diabetes. Features such as `Polyuria`, `Polydipsia`, and `Age` consistently demonstrated their significance in diabetes prediction across different methods.

3) Models: The application of machine learning models, including the Random Forest Classifier, Gradient Boosting Classifier, and Quadratic Discriminant Analysis (QDA), showcased diverse approaches to capturing the complex patterns inherent in diabetes prediction. Each model brought unique strengths to the analysis, emphasizing the importance of considering different methodologies.

B. Implications

The successful integration of linear algebra techniques into the machine learning pipeline has implications for healthcare analytics. The ensemble learning approaches of the Random Forest and Gradient Boosting classifiers demonstrated robustness, while QDA emphasized the importance of statistical modeling for non-linear relationships.

The impact of feature selection on model performance highlighted the necessity of choosing relevant predictors, offering insights that can be valuable in real-world healthcare applications.

C. Future Directions

While our study provides valuable insights, there are several avenues for future research:

- **Ensemble Models:** Investigate the benefits of combining predictions from multiple models to create a more powerful ensemble approach.
- **Hyperparameter Tuning:** Fine-tune model hyperparameters to optimize predictive performance further.
- **Feature Engineering:** Explore additional feature engineering techniques to enhance the discriminative power of selected features.
- **Interpretability:** Enhance the interpretability of models to provide actionable insights for healthcare practitioners.

D. Final Thoughts

In conclusion, the successful integration of linear algebra into the machine learning pipeline for diabetes prediction represents a significant step toward leveraging mathematical principles in healthcare analytics. The combination of robust feature selection, diverse modeling approaches, and a comprehensive analysis of results contributes to the broader understanding of diabetes prediction.

As the field of machine learning continues to evolve, our work sets the stage for ongoing advancements in healthcare analytics and provides a foundation for future studies. By bridging the gap between mathematical principles and real-world healthcare applications, we contribute to the growing body of knowledge aimed at improving patient outcomes and healthcare decision-making.

References

[1] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (pca)," *Computers Geosciences*, vol. 19, no. 3, pp. 303–342, 1993. [Online]. Available: https://www.sciencedirect.com/science/article/pii/009830049390090R

[2] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, dec 2017. [Online]. Available: https://doi.org/10.1145/3136625

[3] X. Man and E. Chan, "The best way to select features?" *arXiv preprint arXiv:2005.12483*, 2020.

[4] D. Bergh, "Chi-squared test of fit and sample size-a comparison between a random sample approach and a chi-square value adjustment method," *Journal of applied measurement*, vol. 16, no. 2, pp. 204–217, 2015.

[5] R. Thanuja *et al.*, "Ttfse-two-tier feature selection and extractionmachine learning model for effective network attack detection," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 9, pp. 2626–2634, 2021.

[6] J. Daemen, V. Rijmen, J. Daemen, and V. Rijmen, "Correlation matrices," *The Design of Rijndael: The Advanced Encryption Standard (AES)*, pp. 91–113, 2020.

[7] A. Parmar, R. Katariya, and V. Patel, "A review on random forest: An ensemble classifier," in *International conference on intelligent data communication technologies and internet of things (ICICI) 2018*. Springer, 2019, pp. 758–763.

[8] H. Dalianis and H. Dalianis, "Evaluation metrics and evaluation," *Clinical Text Mining: secondary use of electronic patient records*, pp. 45–53, 2018.

[9] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, pp. 1937–1967, 2021.

[10] S. Srivastava, M. R. Gupta, and B. A. Frigyik, "Bayesian quadratic discriminant analysis." *Journal of Machine Learning Research*, vol. 8, no. 6, 2007.