Group Project
Diabetes Prediction

Introduction to Artificial Intelligence
(BDM 3014)

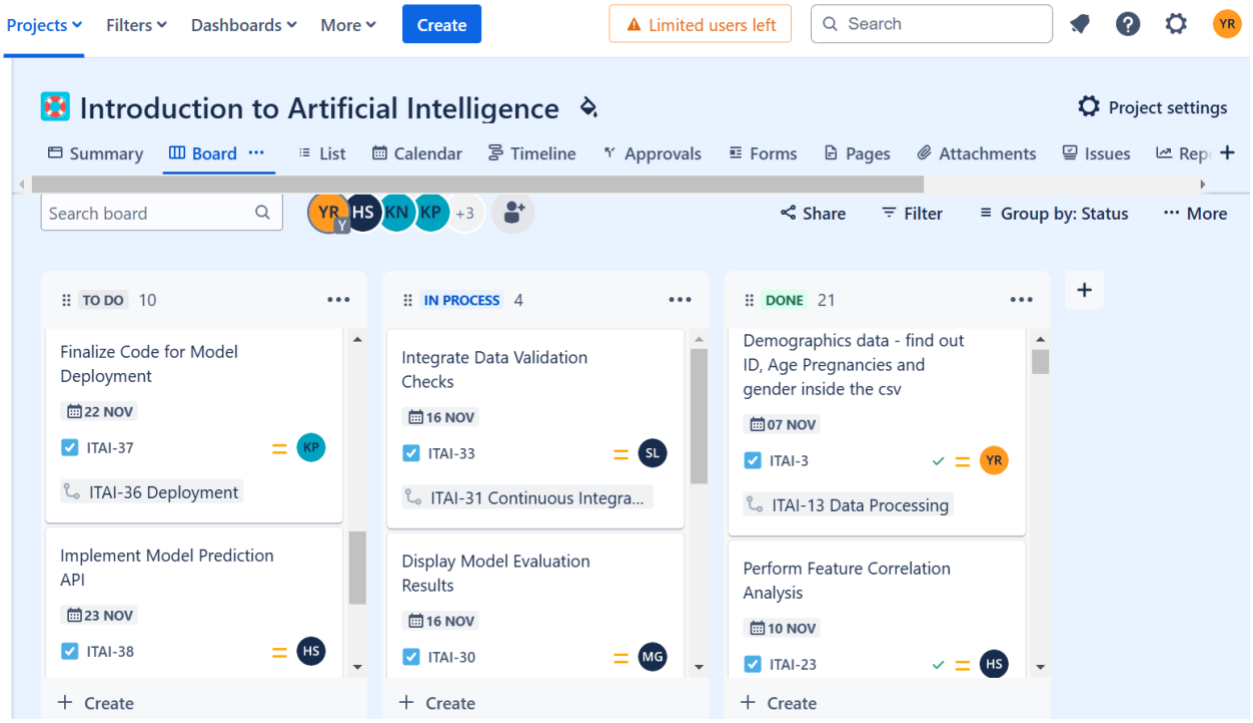
- Yorbis Daniel Alarcon (C0941168)
- Hazel Santons (C0915982)
- Komal Nandal (C0938518)
- Sangsun Lee (C0905412)
- Gustavo Vera Suarez (C0917164)
- Sneha Painuli (C0933116)
- Mahek Ghanchi (C0937254)
- Krunal Patel (C0936008)

Contents

Project Dashboard Jira

Board - Sprint 1

This is the main project board for the Introduction to Artificial Intelligence project in Jira. The board is organized into columns representing task statuses: To Do, In Process, and Done. This setup provides a clear view of the project's current progress and outstanding tasks, allowing team members to quickly understand what needs attention and what's completed.

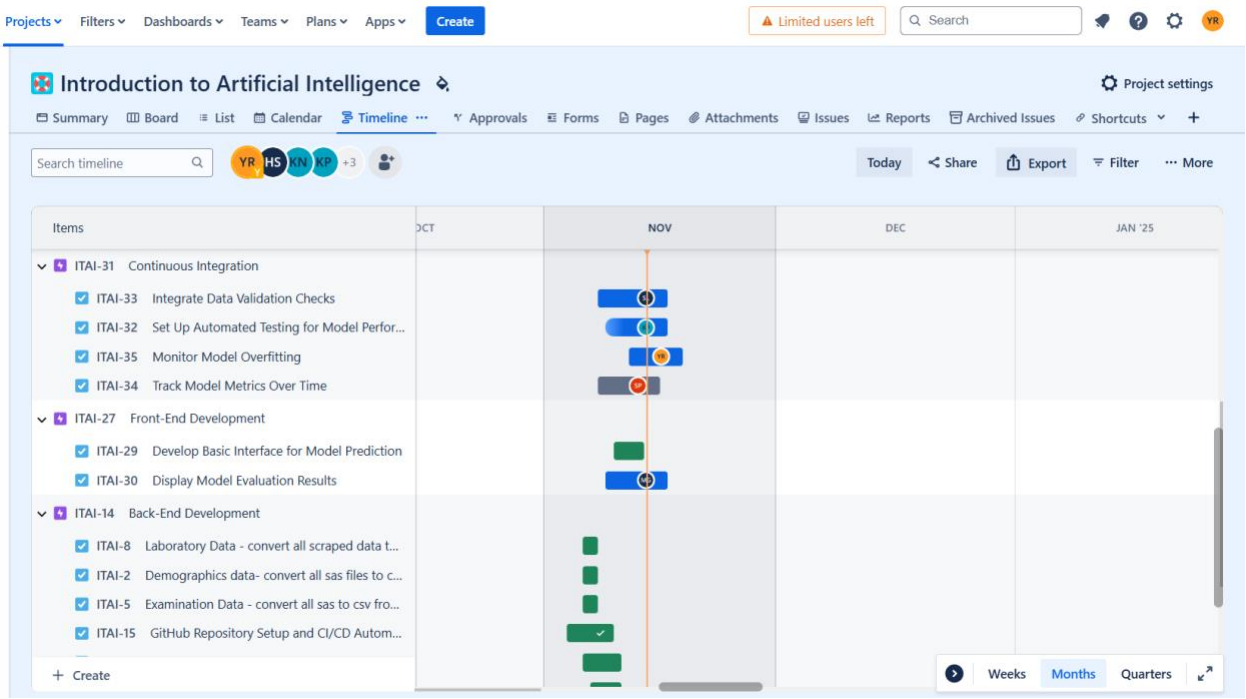


Timeline with Epics and Categories

This image displays the timeline view of the project in Jira, which is organized by epics and major categories, including:

1. Continuous Integration, Front-End Development, Back-End Development, ETL, Data Processing, and Deployment.
2. Task Status and Dates: Each task on the timeline shows its current status (such as completed or in progress) and due dates, enabling the team to visualize the project's timeline.

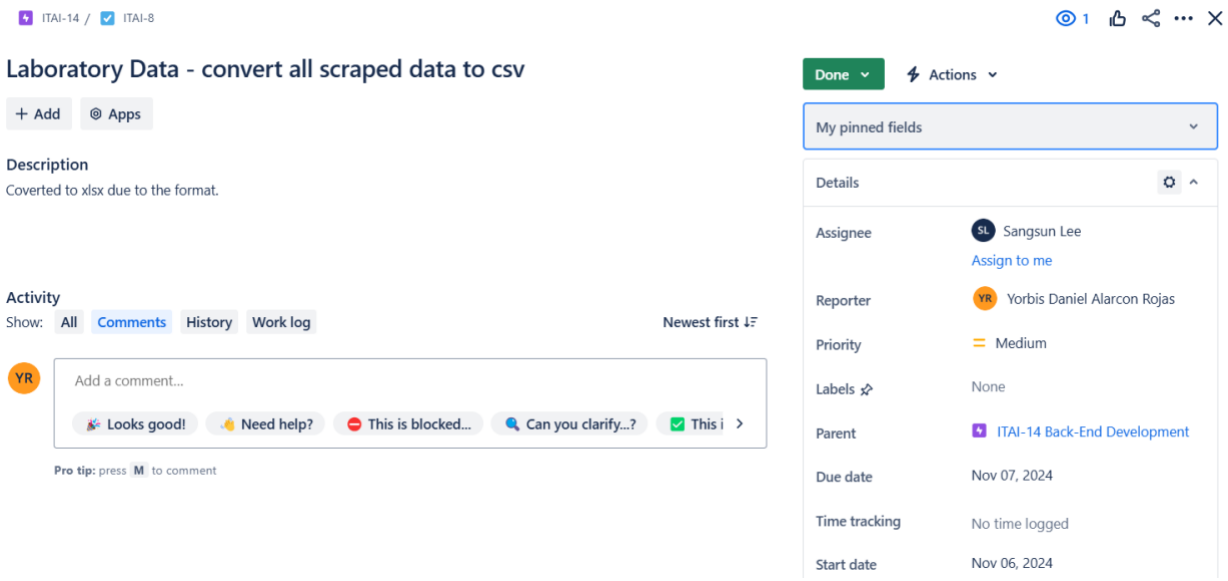
This timeline helps the team track progress across various project components, maintain alignment on schedules, and identify dependencies or potential bottlenecks in specific areas. Segmenting tasks into these categories allows the team to focus on different project parts while staying aware of the overall timeline, ensuring timely delivery of each phase and facilitating more effective sprint planning.



Task Details Example

This image shows a detailed view of a specific task within Jira, titled "Laboratory Data - convert all scraped data to csv". This task is part of the Back-End Development epic and is assigned to Sangsun Lee. The task includes:

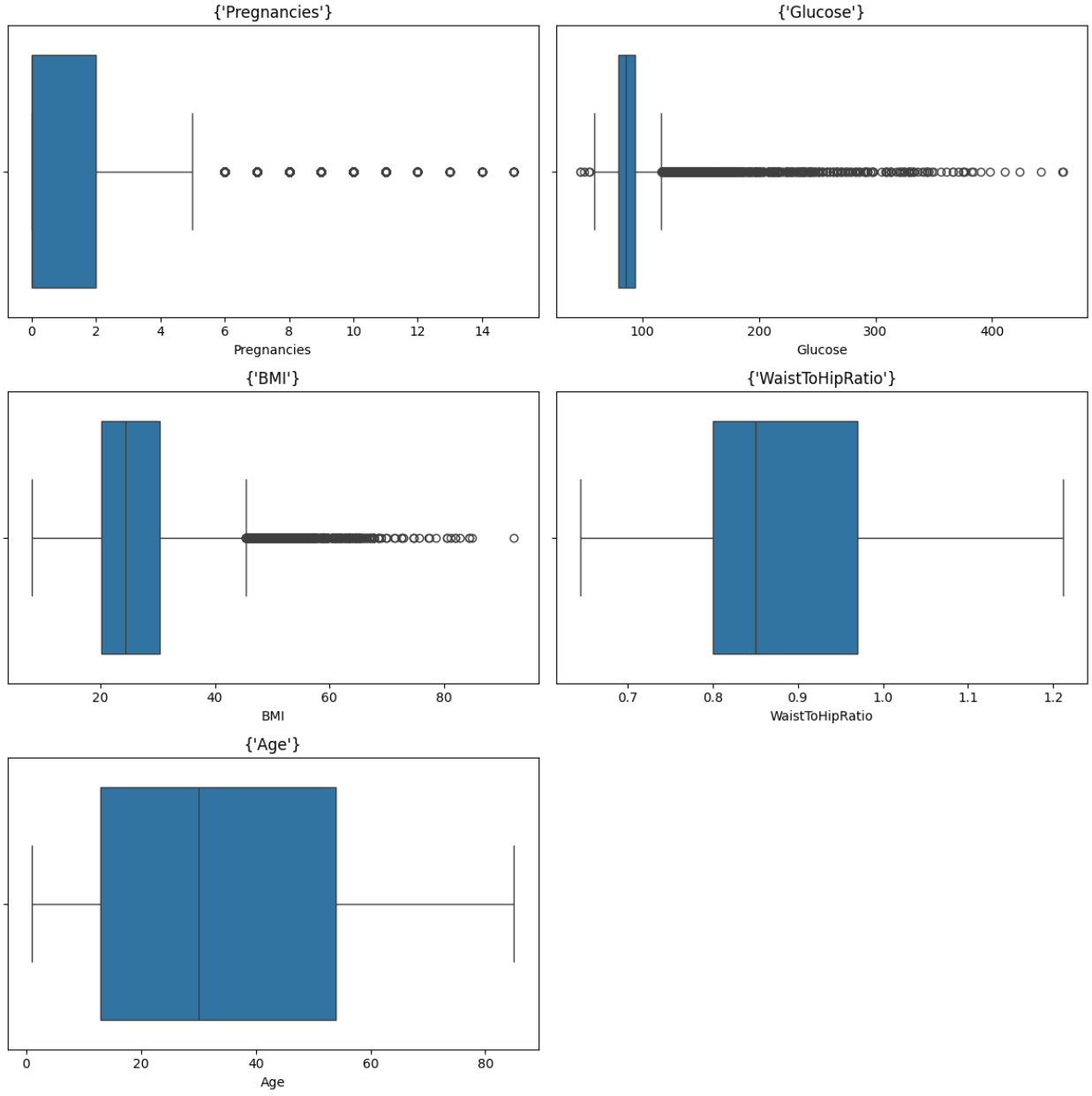
1. Status: Marked as "Done", indicating that the task has been completed.
2. Details: Includes a brief description, "Converted to xlsx due to the format", to clarify the actions taken during the task.
3. Activity and Comments: Provides a space for team members to leave comments, clarifications, or request further help, enhancing communication and ensuring alignment on task expectations.


ITAI-14 / ITAI-8
1
Laboratory Data - convert all scraped data to csv
+ Add
Apps
Description
Coverted to xlsx due to the format.
Activity
Show: All Comments History Work log
Newest first
YR
Add a comment...
Looks good!
Need help?
This is blocked...
Can you clarify...?
This i
Pro tip: press M to comment
Done
Actions
My pinned fields
Details
Assignee
SL
Sangsun Lee
Assign to me
Reporter
YR
Yorbis Daniel Alarcon Rojas
Priority
Medium
Labels
None
Parent
ITAI-14 Back-End Development
Due date
Nov 07, 2024
Time tracking
No time logged
Start date
Nov 06, 2024

Data Preprocessing and Feature Engg.

Data Preprocessing and Feature Engineering Pipeline

Outlier Treatment

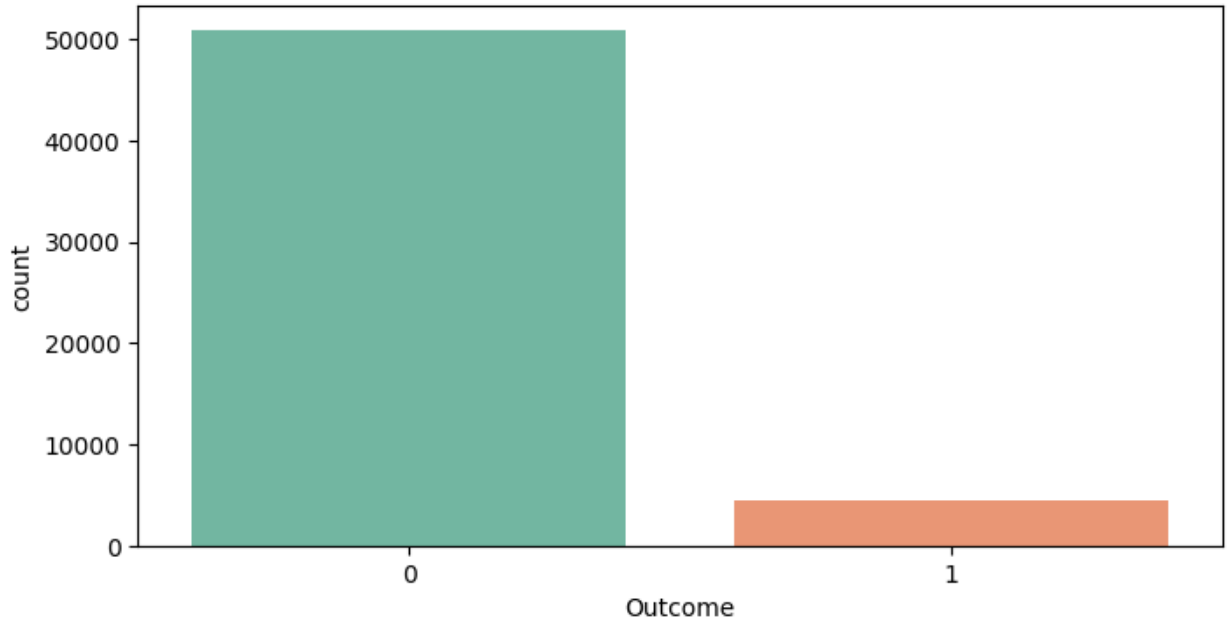


The distribution analysis of key features, visualized through boxplots, reveals several outliers that may have significant implications for diabetes risk. Most values for Pregnancies cluster between 0 and 4, though some higher outliers exceed 10, potentially representing unique cases that could influence diabetes risk. Glucose levels are generally concentrated in the middle range, with outliers above 250, which aligns with the expectation that elevated glucose levels are a primary indicator

of diabetes. For BMI, most values fall within a typical range, but outliers above 45 suggest cases of obesity, a known risk factor for diabetes.

In addition, WaistToHipRatio values are largely contained within a reasonable range, although a few outliers exceed 1.0, indicating central obesity—a condition associated with increased metabolic risks, including diabetes. Age, however, shows a broad distribution without extreme outliers, allowing for an exploration of diabetes risk across different age groups. Overall, the presence of outliers in features like BMI, WaistToHipRatio, Pregnancies, and Glucose underscores the importance of these variables in understanding and predicting diabetes, as they align with known health risks related to obesity and glucose regulation.

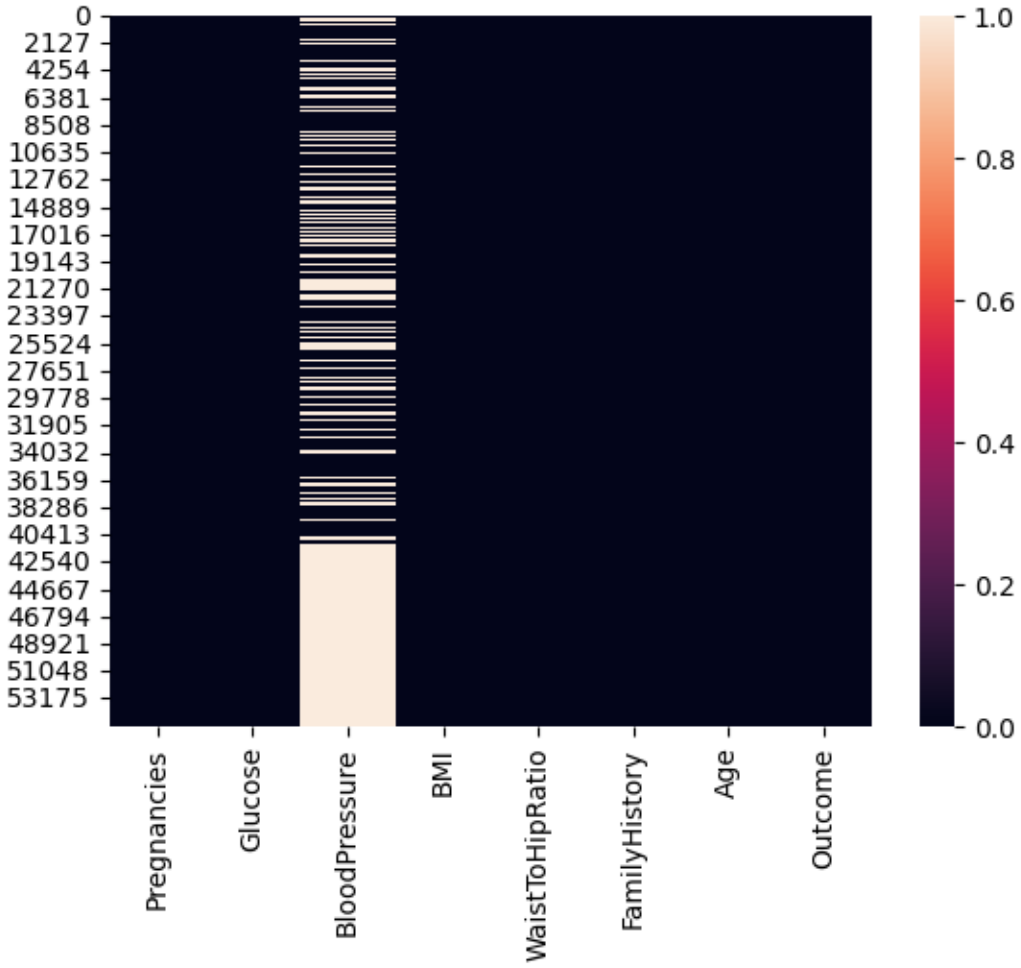
Class Distribution and Under-sampling Strategy



The initial class distribution shows a significant imbalance between the majority class (non-diabetic) and the minority class (diabetic), with the non-diabetic class being far more prevalent. This imbalance can lead to biased model predictions, as the model may favor the majority class to minimize overall error, which would result in poor predictive accuracy for the minority (diabetic) class.

To address this issue, we applied an undersampling technique to reduce the number of non-diabetic samples, creating a balanced dataset with equal representation from both classes. After undersampling, the dataset shows an even distribution between diabetic and non-diabetic cases, allowing the model to learn patterns from both classes effectively. This balanced dataset is expected to enhance the model's ability to accurately predict outcomes for the minority class, ultimately improving the overall performance and robustness of the model.

Missing Data Analysis and Imputation Strategy



The above heatmap visualizes missing data across the different features in the diabetes dataset. Each row represents an observation, while each column corresponds to a feature. White lines indicate missing values, with the color intensity showing the frequency of missing data for each feature.

Key Observations:

1. **Blood Pressure**:
    - This feature contains a substantial number of missing values, as indicated by the high concentration of white lines in the column. These missing entries represent a significant portion of the dataset, requiring attention during preprocessing.
2. **Other Features**:
    - The remaining features, including Pregnancies, Glucose, BMI, WaistToHipRatio, FamilyHistory, Age, and Outcome, display minimal or no missing values, as shown by the solid dark color (near 0 on the scale). This consistency across features implies less preprocessing needed for handling missing values in these columns.

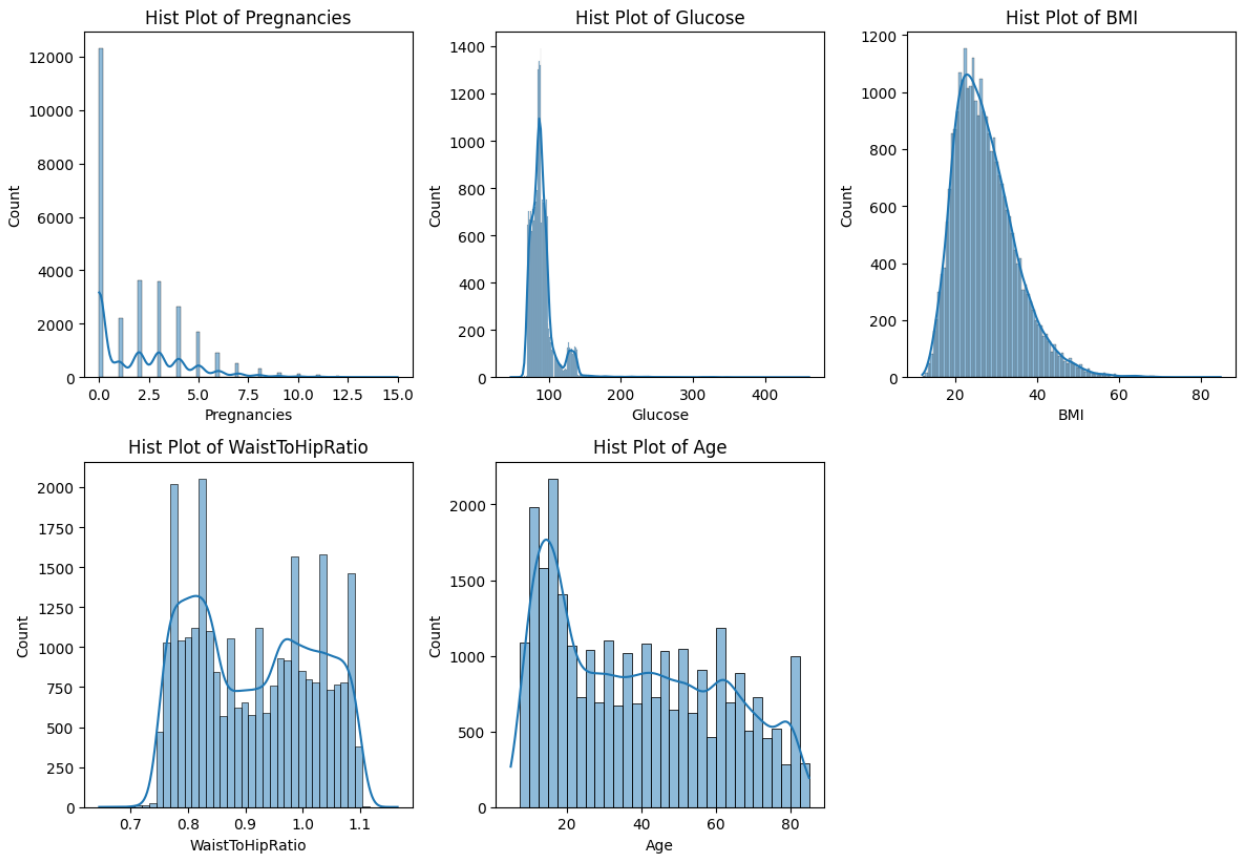
3. **Implications for Data Preprocessing**:
    - Given the presence of missing data primarily in BloodPressure, appropriate imputation strategies are necessary. This may involve median or mean imputation or potentially more advanced techniques if the missing values exhibit a specific pattern or correlation with other features.

To solve the missing values issue in the Blood Pressure (BP), we chose using linear regression rather than using mean for imputing, because it is beneficial in a several ways:

- Provides individualized estimates that reflect each patient's characteristics
- Leverages correlations to improve the accuracy of estimates
- Preserves the original data distribution, reducing the risk of distortion
- Minimizes bias by avoiding oversimplified imputations
- Ultimately contributes to improved model performance by providing a realistic and representative dataset

Feature Selection, Combination Strategies, and Target Predictability / Class Separability

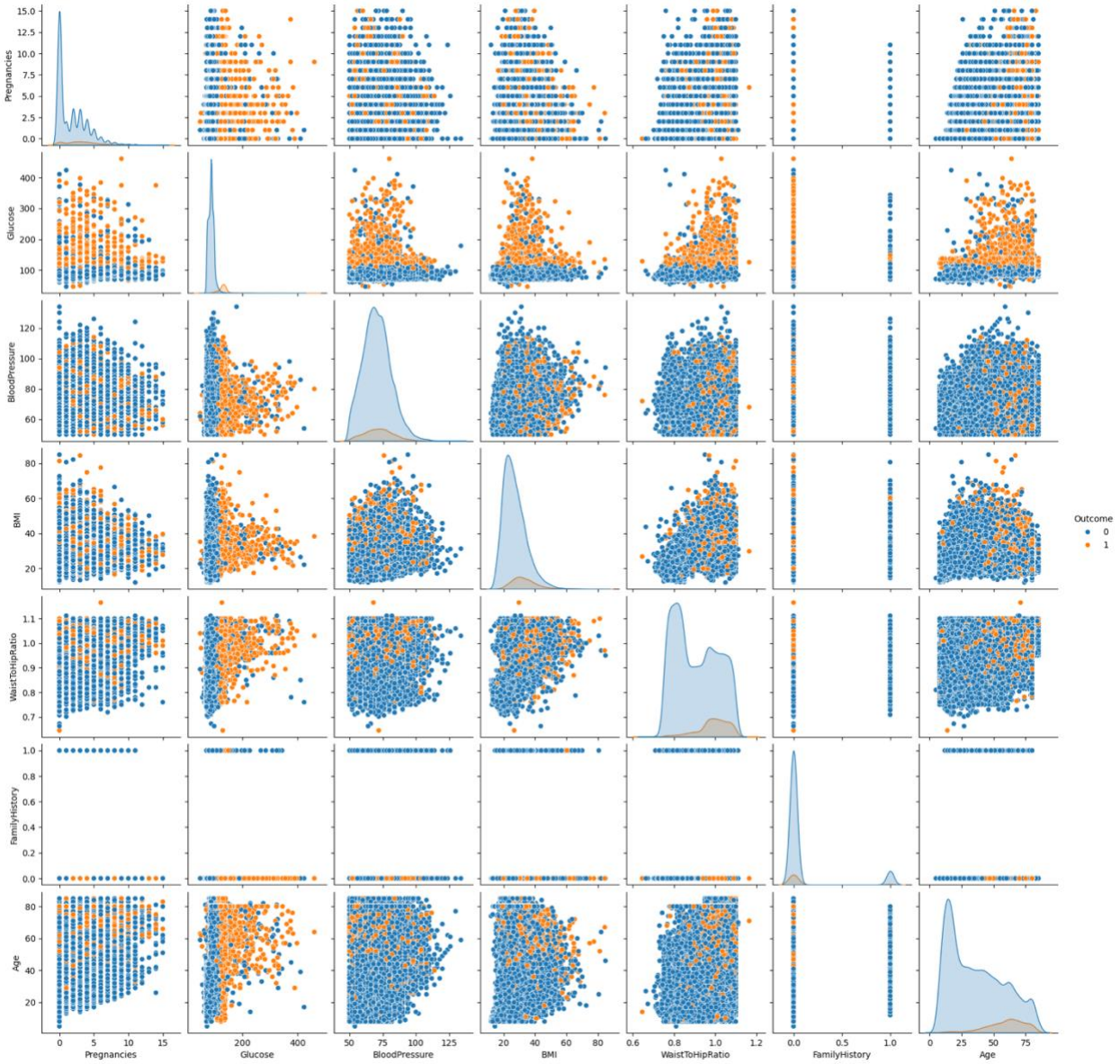
Feature Distributions and Predictability



The histograms with Kernel Density Estimate (KDE) overlays provide a detailed view of the distribution for key features in the diabetes dataset, revealing important patterns and potential outliers. The Pregnancies feature shows a right-skewed distribution, with most values concentrated between 0 and 3, and a long tail extending to higher values, indicating some outliers. Glucose levels are similarly right-skewed, with the majority of values clustering around 100–130, while a few high values suggest cases with elevated glucose levels, potentially signaling diabetes risk.

BMI displays a relatively normal distribution, peaking around 30, with a slight right skew due to high-BMI outliers, which may correspond to individuals with obesity—a known risk factor for diabetes. The Waist-To-Hip Ratio has a nearly symmetric distribution, peaking between 0.85 and 0.9; higher values indicate central obesity, which is associated with metabolic health risks. The Age distribution spans a broad range, allowing for the analysis of diabetes risk across different age groups, with a slight concentration of younger to middle-aged individuals around 20–30 years.

Overall, these histograms reveal insights into the general distribution and range of each feature. The skewness and outliers observed in certain features, such as Pregnancies and Glucose, suggest that data transformations or outlier handling strategies may be necessary to optimize model performance.

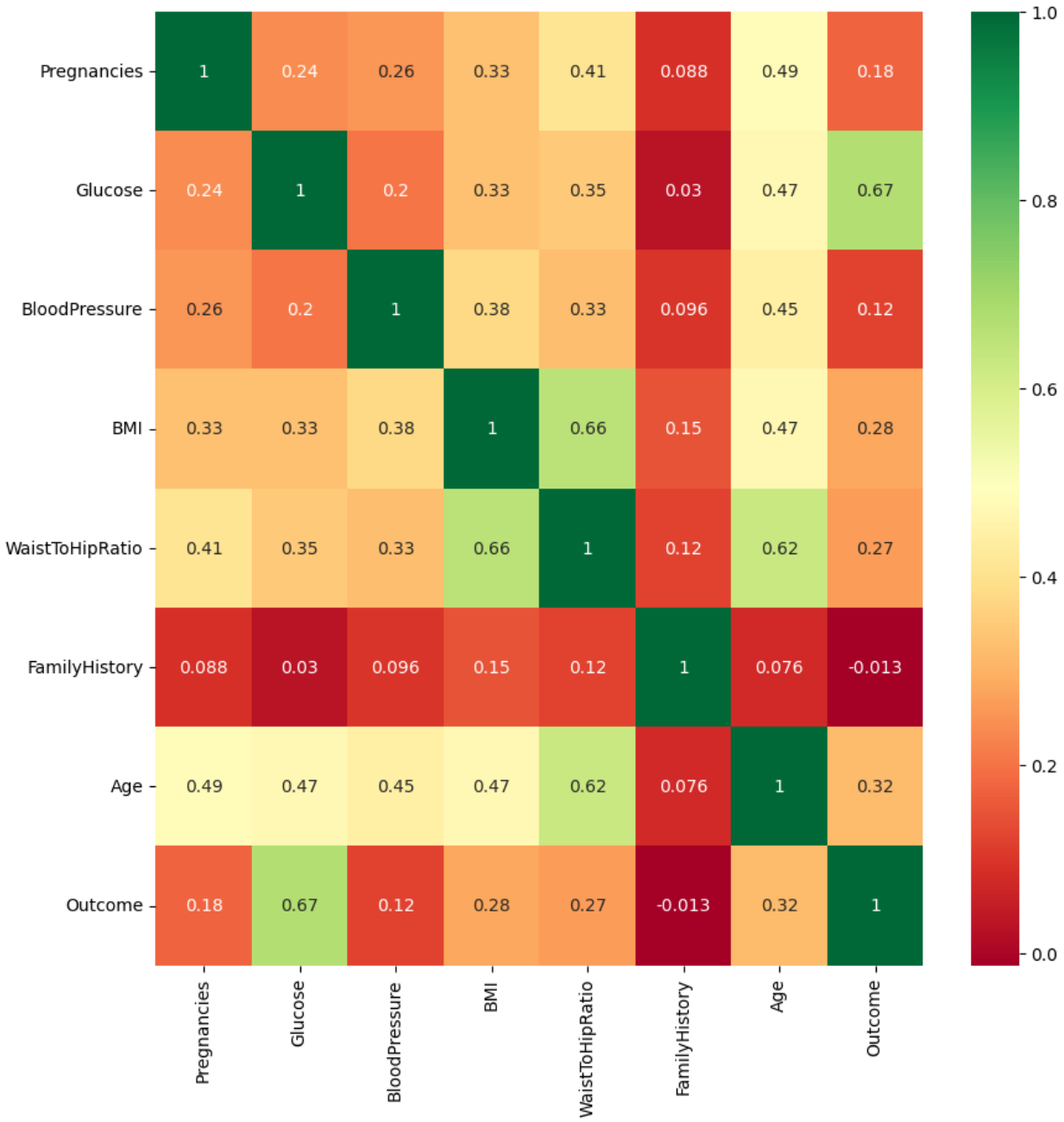


The pair plot visualization reveals significant insights into the relationships and distributions of features in the diabetes dataset, particularly highlighting Glucose, BMI, and WaistToHipRatio as strong indicators of diabetes risk. Diabetic cases (labeled as 1) are generally associated with higher values for these features, which suggests that these attributes should be prioritized in feature selection. The right-skewed distribution of Pregnancies and Glucose also indicates the presence of outliers, especially at higher values, potentially pointing to unique risk profiles among individuals with higher pregnancies or elevated glucose levels.

In terms of feature relationships, Glucose and BMI show a positive correlation, with higher values for both often corresponding to diabetic cases. Similarly, BMI and WaistToHipRatio are moderately correlated, reflecting their shared link to body composition and obesity, both of which are known diabetes risk factors. Although features like BloodPressure and FamilyHistory show less separation between classes, their inclusion may still add value due to their health-related context, particularly as FamilyHistory indicates a genetic predisposition to diabetes.

This analysis suggests that Glucose, BMI, and WaistToHipRatio are essential features for model development due to their strong class separability and predictive power. Age and FamilyHistory also provide moderate predictive value, particularly for older individuals or those with a family history of diabetes. Overall, this plot informs feature selection and combination strategies by identifying the features that are most indicative of diabetes, thus helping to enhance the model's accuracy and interpretability.

Correlation Heatmap For Feature Selection



The correlation heatmap serves as a valuable tool for feature selection, providing insight into the relationships between each feature and the target variable, Outcome. Notably, features such as Glucose and BMI exhibit strong positive correlations with Outcome, indicating their potential predictive power for identifying diabetes risk. This suggests that these features should be prioritized in the final feature set due to their relevance to the target. Additionally, the heatmap reveals notable inter-feature correlations, such as between BMI and Waist-To-Hip Ratio, which

may indicate redundancy. Recognizing these relationships is essential for informed feature selection, as highly correlated features may contribute similar information to the model, increasing the risk of multicollinearity in certain algorithms. Overall, the heatmap helps refine the feature set by highlighting both the predictive strength of individual features for the target variable and potential redundancies that could affect model performance.

Model Building and Evaluation

Model Selection and Building Strategy

To effectively classify diabetic versus non-diabetic cases, we employed six machine learning models to assess their suitability based on performance, interpretability, and robustness against overfitting and underfitting. This approach allowed us to leverage a variety of algorithms, each with unique strengths, to determine the model best suited to this medical prediction task. We carefully chose models with distinct characteristics to ensure a comprehensive comparison across different machine learning paradigms, including linear models, distance-based classifiers, probabilistic models, margin-based methods, and tree-based models.

Selected Models

1. **Logistic Regression**
    - Logistic Regression, a linear model, was selected for its interpretability and its capacity to provide probabilities, which can be particularly valuable in a medical context for understanding the likelihood of diabetes. Logistic Regression is well-suited for binary classification tasks, and after scaling the data, we trained the model to identify linear relationships between the features and the target. This model fits a decision boundary that best separates diabetic from non-diabetic cases in the feature space.
2. **K-Nearest Neighbors (KNN)**
    - KNN is a non-parametric model that classifies instances based on their proximity to other points in the feature space, making it an effective choice for classification tasks when data is properly scaled. The model assigns a class to each point based on the majority class of its nearest neighbors, a process that requires scaled data to ensure equal contribution of all features to the distance calculations. This allows KNN to make more accurate neighbor-based predictions.
3. **Naive Bayes**
    - As a probabilistic model, Naive Bayes provides a quick, efficient, and straightforward baseline model. Despite its assumption of feature independence, scaling can still improve Naive Bayes by ensuring that all feature ranges are comparable. Naive Bayes is computationally efficient, making it suitable for initial comparisons to assess the baseline performance of other more complex models.
4. **Support Vector Machine (SVM)**

- SVM is particularly effective for complex classification tasks, especially in high-dimensional spaces. It works by maximizing the margin between classes, making it a robust choice for datasets where separation boundaries are complex. Since SVM relies on a margin-based approach, scaling is essential to prevent certain features from disproportionately influencing the margin, which could lead to suboptimal decision boundaries. This model is advantageous when there are clear separations in the data but can struggle with large-scale datasets if not carefully tuned.

5. **Decision Tree**
    - Decision Trees offer a highly interpretable model structure, as they split data based on feature values, making it easy to visualize and explain. However, they are prone to overfitting as they capture noise and intricacies of the training data. Decision Trees are non-parametric and do not rely on feature distances; thus, they do not require scaling. This model splits the data based on criteria like Gini Impurity or Entropy to maximize information gain at each node, making it especially useful when interpretability is a priority.
6. **Random Forest**
    - Random Forest, an ensemble of Decision Trees, mitigates the overfitting tendencies of individual trees by combining predictions from multiple trees, effectively reducing variance. This model typically achieves higher accuracy and generalization by averaging the predictions of a diverse set of trees. Like Decision Trees, Random Forests are also non-parametric and do not rely on feature scaling, which makes them flexible across various data types and distributions.

Data Scaling and Leakage Prevention

To prevent data leakage and ensure a fair evaluation, we performed the train-test split before applying any scaling. This practice ensures that the scaler learns parameters (mean, standard deviation) from the training set alone, preserving the independence of the test set. By doing so, we avoid contaminating the training process with information from the test set, allowing for a realistic assessment of model performance on genuinely unseen data.

- **Scaling Strategy**:
    - For models sensitive to feature scaling, such as Logistic Regression, KNN, and SVM, we initialized and fitted the scaler on the training set only, then applied the same transformation to the test set. This ensured that scaling parameters were independent of the test data, preventing leakage.
    - For models that do not require scaling, such as Decision Tree and Random Forest, we used the unscaled data to preserve their interpretability and natural robustness to feature ranges.

This multi-model approach, combined with careful preprocessing and scaling practices, demonstrates a comprehensive application of machine learning concepts to determine the best model for predicting diabetes. By comparing accuracy, robustness, and generalization capacity, we aim to select the most effective model that can provide interpretable, accurate predictions in a clinical context, ensuring it performs well on new, unseen data. This structured methodology aligns

with best practices in model building and evaluation, fostering a reliable and interpretable solution for diabetes classification.

Evaluation

To evaluate the predictive models and identify the best-performing model for diabetes classification, we compared six machine learning algorithms: Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, Support Vector Machine (SVM), Decision Tree, and Random Forest. By examining the training and test accuracy for each model, as well as observing their handling of overfitting and generalization capabilities, we aimed to select the model best suited for this medical application.

Logistic Regression

Logistic Regression demonstrated high training accuracy (95.3%) and test accuracy (95.07%), indicating that it effectively learned the linear relationships in the dataset without significant overfitting. Logistic Regression's interpretability and ability to provide probability scores make it suitable for medical predictions where understanding likelihoods is essential. However, as a linear model, it may struggle with complex, nonlinear patterns.

K-Nearest Neighbors (KNN)

KNN yielded a training accuracy of 95.63% and a slightly lower test accuracy of 94.00%. The model's performance highlights its sensitivity to scaling, as it relies on feature distances to determine classifications. The difference between training and test accuracy suggests that KNN could be prone to mild overfitting, especially as it considers the nearest neighbors in the training data. Nonetheless, KNN remains a valuable model due to its simplicity and effectiveness in well-scaled datasets.

Naive Bayes

Naive Bayes achieved training and test accuracies of approximately 87.91% and 87.73%, respectively, making it a suitable baseline model. Although Naive Bayes assumes feature independence, which may limit its accuracy for this application, it provides quick and efficient predictions. The lower accuracy compared to other models suggests that Naive Bayes may not capture the full complexity of diabetes-related features, but it is valuable as a simple and interpretable model.

Support Vector Machine (SVM)

SVM achieved a strong training accuracy of 95.92% and test accuracy of 95.18%, indicating good generalization and a strong performance in high-dimensional feature space. SVM's approach of maximizing the margin between classes is effective in this application, as it accurately separates diabetic from non-diabetic cases. Scaling was essential for SVM, as unscaled features would have distorted the margin, leading to suboptimal boundaries. SVM's slight overfitting tendency is balanced by its robustness and margin-based separation, making it a reliable choice.

Decision Tree

Using unscaled data, the Decision Tree model achieved high training accuracy but demonstrated a tendency to overfit, with a gap between training and test accuracy. Decision Trees split data based on feature values rather than distances, making scaling unnecessary. However, the Decision Tree model's flexibility and propensity to learn data intricacies contribute to overfitting, limiting its ability to generalize well on unseen data. While interpretable, its overfitting issue reduces its applicability as the primary model for this application.

Random Forest

The Random Forest model achieved perfect training accuracy (100%) and a high test accuracy of 95.74%, indicating potential overfitting. However, Random Forest's ensemble approach, which combines predictions from multiple decision trees, helps reduce variance and improves generalization compared to a single Decision Tree. This model offers high accuracy, robustness against overfitting, and strong feature importance insights, making it an ideal choice despite the slight overfitting. Random Forest's test accuracy is competitive, and its stability in predictions makes it suitable for real-world applications.

Preventing Data Leakage: Scaling After Train-Test Split

To ensure fair evaluation and prevent data leakage, we performed the train-test split before scaling. This approach ensures that the scaler parameters (e.g., mean, standard deviation) are learned from the training data only, preserving the independence of the test set. Scaling after splitting ensures that no information from the test set influences the training process, thus enabling an unbiased assessment of model performance on genuinely unseen data.

Model-Specific Scaling Application

- For models sensitive to scaling, such as Logistic Regression, KNN, and SVM, we applied scaling only to the training data and then transformed the test data based on these parameters. This process preserves the integrity of the test set.
- For models that do not require scaling, including Decision Tree and Random Forest, we used unscaled data to maintain interpretability and allow these models to leverage their inherent ability to handle various feature ranges.

Classification Report and Confusion Matrix Analysis

The classification report provided precision, recall, and F1-scores for each class (diabetic and non-diabetic), with an overall accuracy of 96%. The confusion matrix highlighted the model's strengths and weaknesses in correctly identifying diabetic versus non-diabetic cases. Both precision and recall scores are high, which is critical in a medical setting, as the cost of misclassifying a diabetic case can be significant.

SVM as an Alternative Model

While Random Forest emerged as the preferred model due to its accuracy, robustness, and interpretability, Support Vector Machine (SVM) was also a strong contender in the analysis. SVM offers several advantages, making it suitable for scenarios that prioritize balanced performance and precision. Here's why SVM was considered a viable alternative:

- **High Accuracy and Balanced Metrics**: SVM achieved high accuracy and balanced precision-recall scores, indicating its effectiveness in managing both false positives and false negatives. This makes it a strong choice for medical applications where the consequences of misclassification can be significant.
- **Sensitivity to Hyperparameters**: SVM's performance is highly dependent on selecting the right kernel function, regularization parameter, and other hyperparameters. With careful tuning, SVM can achieve optimal performance, though this tuning process can be computationally intensive and time-consuming. SVM's flexibility to adapt through parameter tuning provides an opportunity for further refinement, albeit at a cost of additional computational resources.
- **Scalability Constraints**: SVM can become computationally demanding, especially with large datasets or complex kernels. For scenarios that require real-time predictions or have limited computational resources, this model might be less practical. Random Forest, by comparison, offers quicker training and inference, making it more scalable for applications where speed is critical.

Final Model Selection

After systematically evaluating each model's accuracy, generalization capability, and sensitivity to scaling, we identified **Random Forest** as the optimal candidate for diabetes classification in this application. Random Forest demonstrated high test accuracy (95.74%) and provided stable, accurate predictions, even though it exhibited potential overfitting with a perfect training accuracy. Its ensemble nature, combining predictions from multiple decision trees, reduces variance and enhances robustness, making it less prone to overfitting than a single Decision Tree. Additionally, Random Forest's interpretability through feature importance insights offers valuable information for understanding the impact of each health metric on diabetes risk, which is particularly beneficial in medical settings.

Support Vector Machine (SVM) emerged as a strong alternative model. SVM achieved comparable test accuracy with balanced precision and recall, indicating effective performance in distinguishing diabetic from non-diabetic cases. Its margin-based classification approach ensures clear class separation, making it suitable for high-dimensional spaces and scenarios where balanced metrics are essential. However, SVM's reliance on hyperparameter tuning (such as kernel choice and regularization) can make it computationally intensive, especially for large datasets, which limits its scalability for real-time applications.

Conclusion

While SVM offers high accuracy and balanced metrics, the ease of use, scalability, and robustness of Random Forest make it the more practical and effective choice for this project. Random Forest's

ability to generalize well and its faster inference times make it ideal for clinical applications where consistent, reliable predictions are essential. However, in scenarios where higher precision is needed and computational resources are abundant, SVM could be considered as a secondary choice with potential for further performance improvement through hyperparameter tuning.

In summary, Random Forest is the recommended model for deployment due to its accuracy, robustness, and interpretability, while SVM remains a viable alternative if specific requirements favor it. This strategic selection ensures the application of a model that provides both reliability and practicality in predicting diabetes risk.

Insights Report

Insight 1: Glucose Levels as a Primary Predictor of Diabetes

Insight: Elevated fasting glucose levels are strongly linked with diabetes, indicating that individuals with higher glucose levels are at a significantly increased risk.

Statistical Technique: Correlation analysis and clustering in data visualizations reveal a robust relationship between glucose levels and diabetes incidence.

Organizational Use: Healthcare providers can implement regular glucose screening to identify high-risk individuals, even before a formal diagnosis. Setting up alerts for elevated glucose levels can prompt timely interventions, leading to better health outcomes.

Impact: Early detection through glucose monitoring helps prevent complications, improves patient health, and reduces long-term healthcare costs. Adding glucose-level tracking to standard health assessments could accelerate diagnosis and enhance preventive measures.

Insight 2: BMI and Waist-to-Hip Ratio as Indicators of Diabetes Risk

Insight: Elevated BMI and waist-to-hip ratios are key indicators of diabetes risk, with a clear positive correlation between higher values in these metrics and the presence of diabetes.

Statistical Technique: Regression analysis and association studies confirm a strong link between BMI, waist-to-hip ratio, and diabetes prevalence.

Organizational Use: Healthcare providers and wellness programs can include BMI and waist-to-hip ratio checks in routine health evaluations. Identifying individuals with elevated metrics allows for targeted interventions, like customized nutrition and fitness plans, to reduce risk.

Impact: Addressing obesity and central fat as modifiable risk factors can help lower diabetes rates, decrease healthcare costs, and improve quality of life by delaying or preventing the disease onset.

Insight 3: Age as a Secondary Risk Factor for Diabetes

Insight: Older individuals, especially those over 40, show a higher incidence of diabetes, though age is not as strong a predictor as glucose or BMI.

Statistical Technique: Age-related analysis and stratification highlight a moderate correlation between age and diabetes, underscoring its role as a secondary risk factor.

Organizational Use: Health organizations and insurers can prioritize age-based risk assessments, particularly for older adults with additional risk factors such as high glucose or BMI. Tailored screenings and interventions for these individuals can be implemented to mitigate risk.

Impact: Age-targeted preventive measures allow healthcare providers to allocate resources efficiently, reduce undiagnosed diabetes cases, and improve health outcomes among aging populations, leading to substantial cost savings in diabetes care.