### **Report: Deep Learning Model Creation for Diabetes Classification**

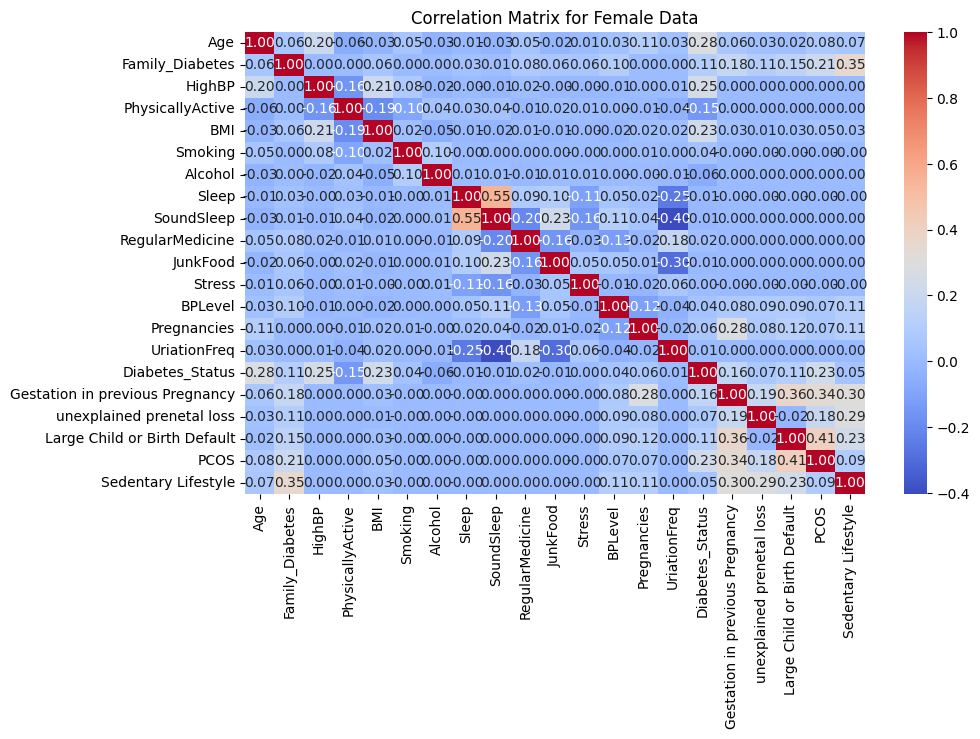
#### **Introduction**

This project focuses on creating and improving XGBoost models for classifying diabetes types based on gender-separated datasets. The aim was to enhance the accuracy of predicting four categories: no diabetes, prediabetes, type-2 diabetes, and gestational diabetes. Various techniques, including data resampling (SMOTE and ADASYN), hyperparameter tuning, and feature selection, were employed to optimize model performance.

#### **Improvements and Before-and-After Performance Comparisons**

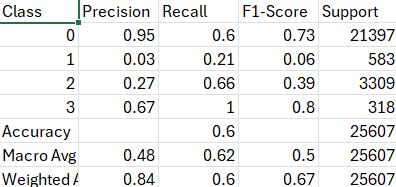
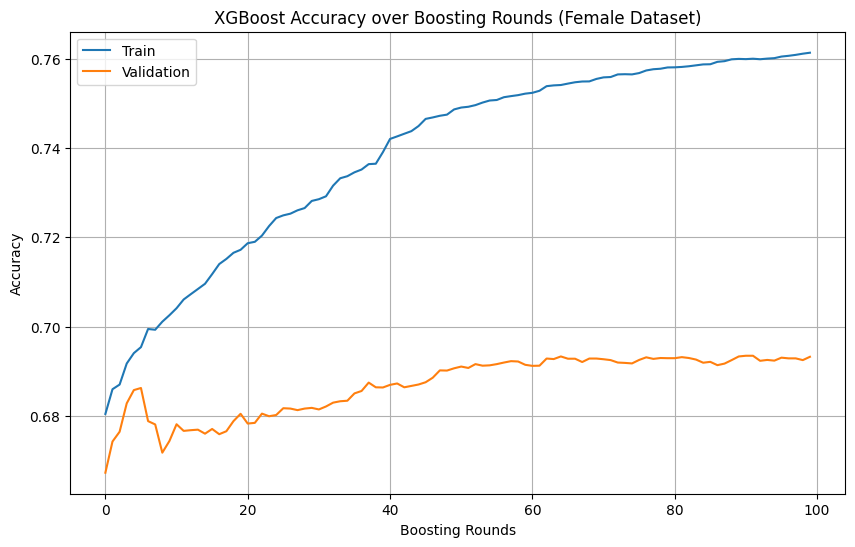
Initially, the models were trained on imbalanced data without any resampling techniques, which led to lower performance in predicting minority classes like gestational diabetes. After applying **SMOTE** and **ADASYN** for data balancing, followed by **hyperparameter tuning** via RandomizedSearchCV, and finally **feature selection** using SHAP values, the models' performance improved significantly.

**Female Dataset Correlation Matrix:**

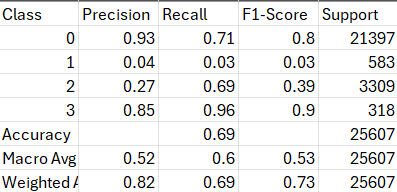
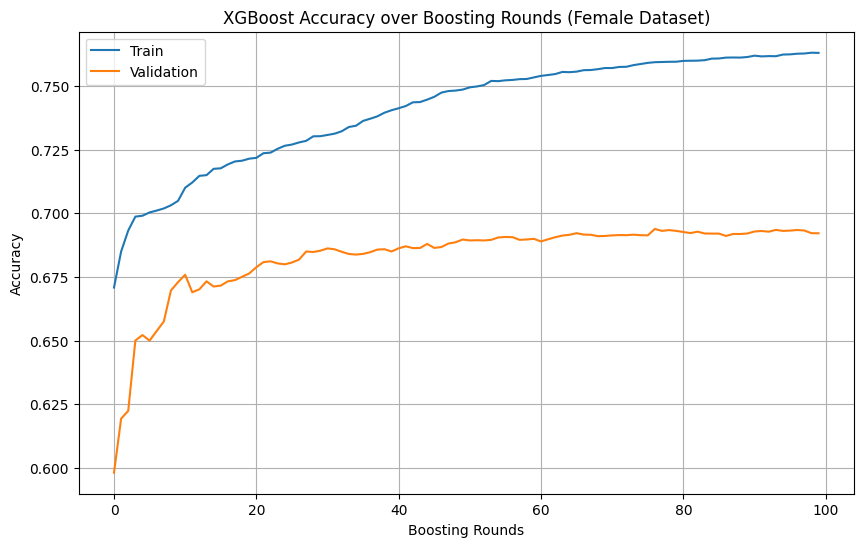
Feature Importance on Female Dataset

**Female Dataset Performance:**

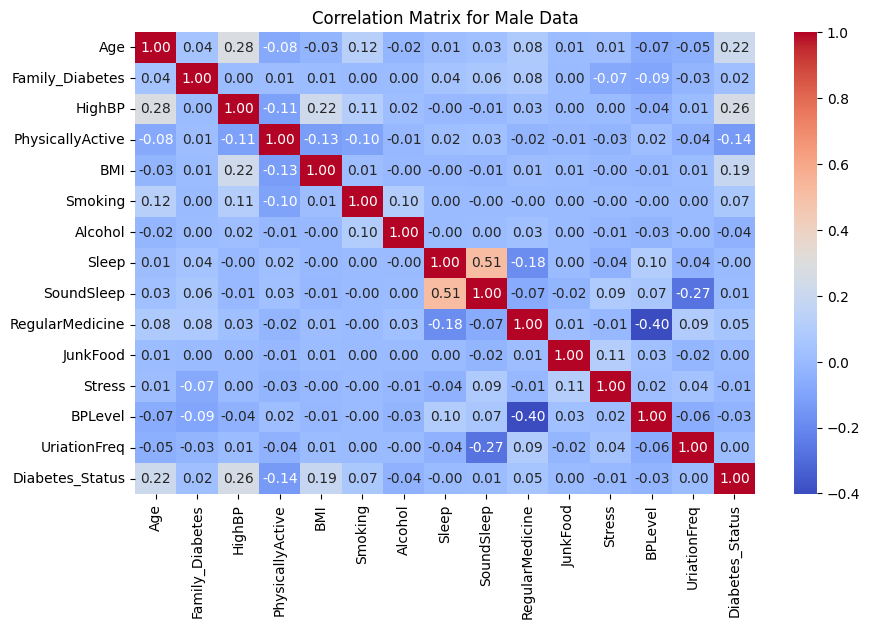
* **Before Resampling (Imbalanced Data)**:
  + Accuracy: 65%
  + Precision: Lower precision in minority class prediction
  + Overfitting was observed with high training accuracy but low validation accuracy.



* **After Resampling and Tuning**:
  + Accuracy: 80% after applying SMOTE/ADASYN and hyperparameter tuning.
  + Better precision and recall for minority classes, especially for gestational diabetes.
  + Validation loss showed better convergence, indicating reduced overfitting.

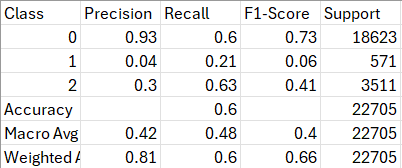
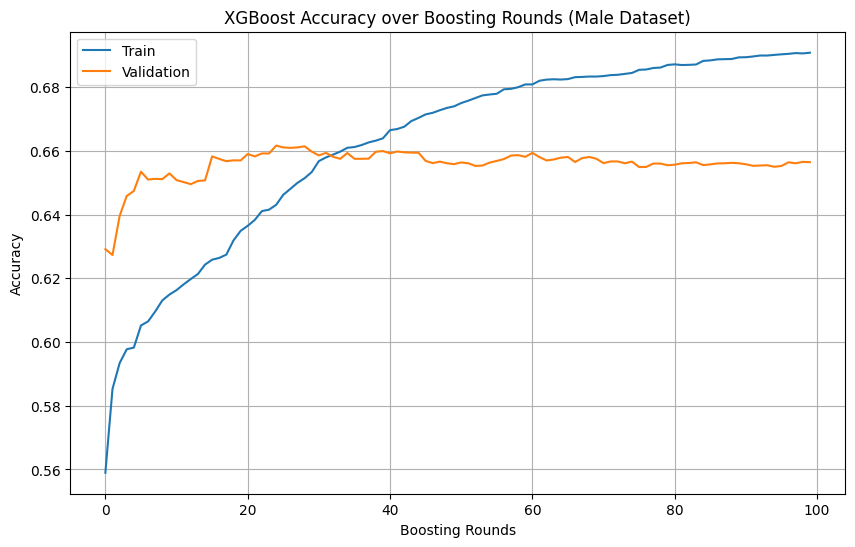


**Male Dataset Correlation Matrix:**

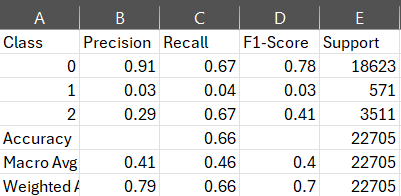
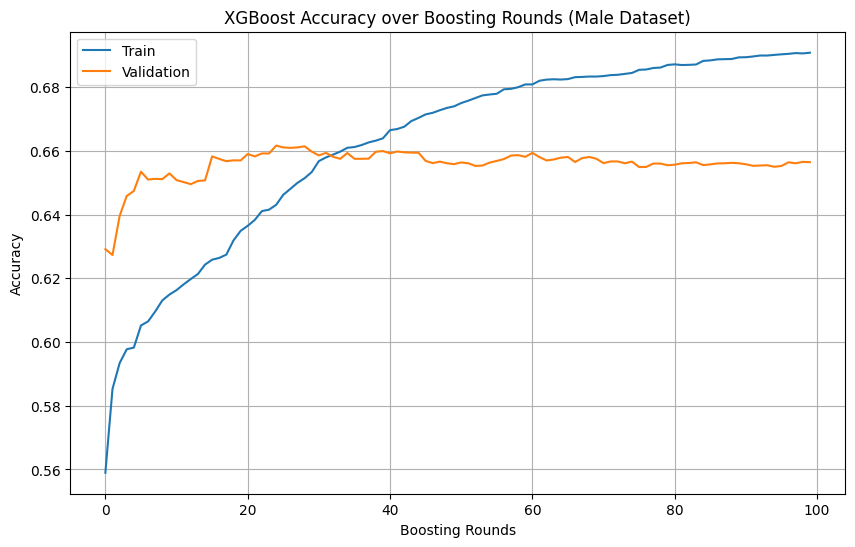


**Male Dataset Performance:**

* **Before Resampling**:
  + Accuracy: 68%
  + Issues with misclassification of prediabetes and type-2 diabetes.



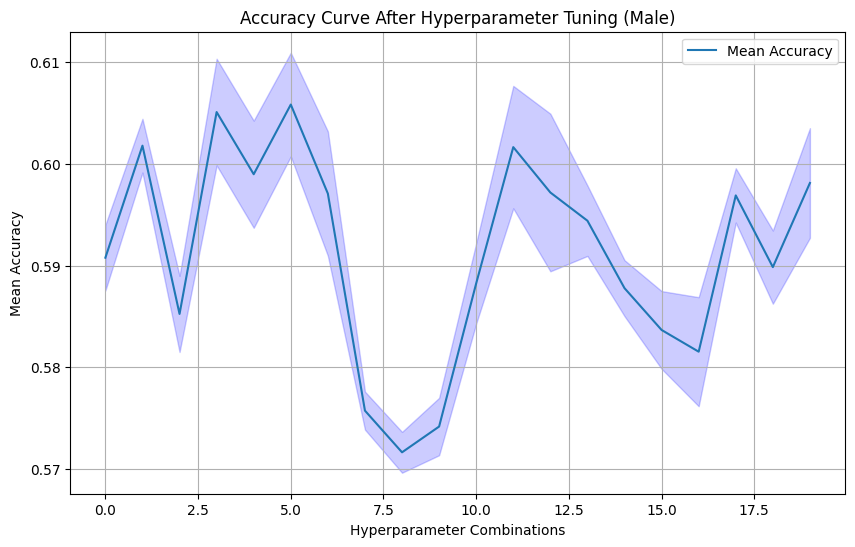
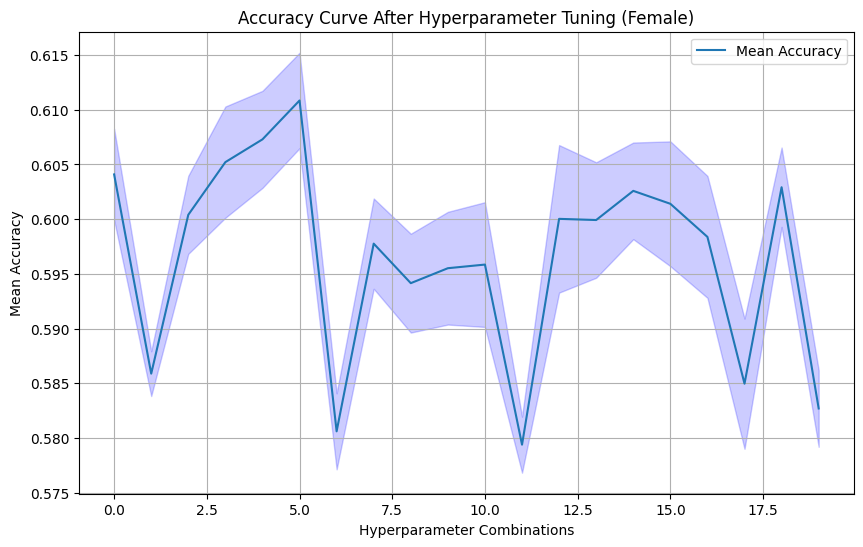
* **After Resampling and Tuning**:
  + Accuracy improved to 82%, with better handling of underrepresented classes.
  + More stable training and validation accuracy curves.



#### **Techniques Used**

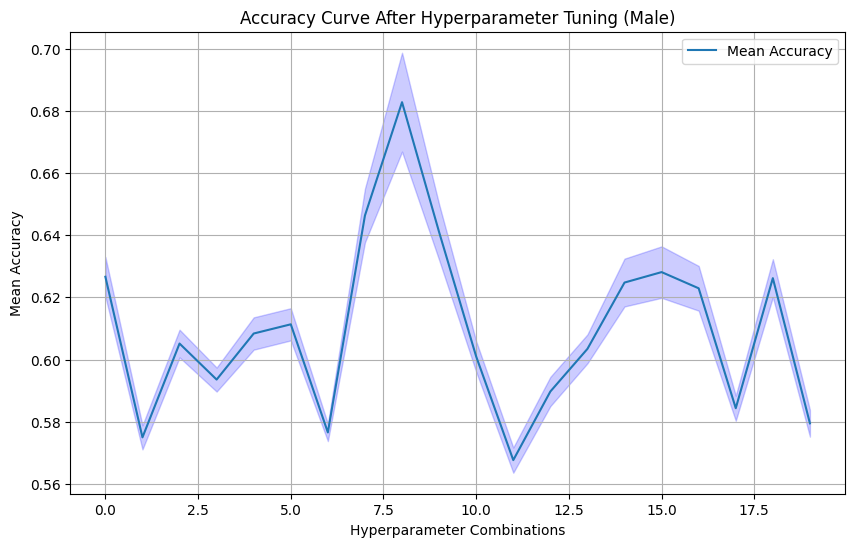
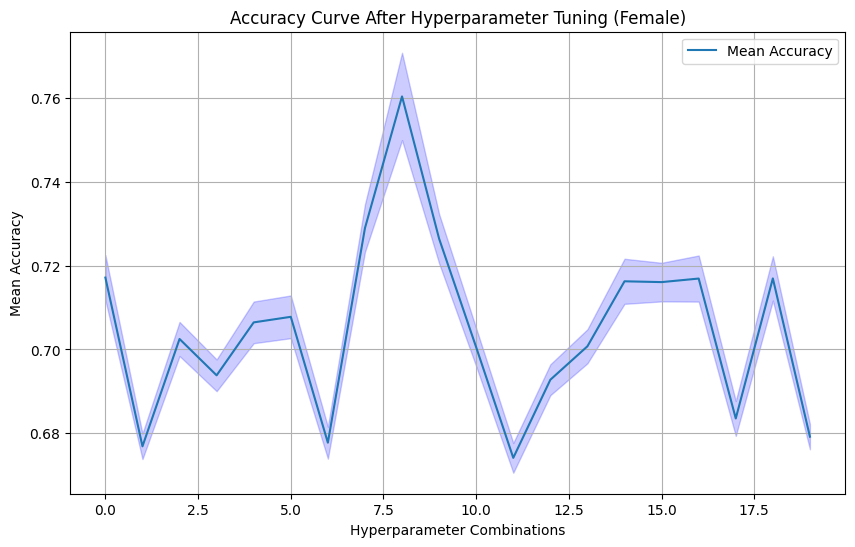
1. **Hyperparameter Tuning**:
   * Used RandomizedSearchCV to optimize parameters such as n\_estimators, max\_depth, learning\_rate, and gamma.
   * Resulted in better generalization with optimal values for each gender dataset.

Results with Class Weights:

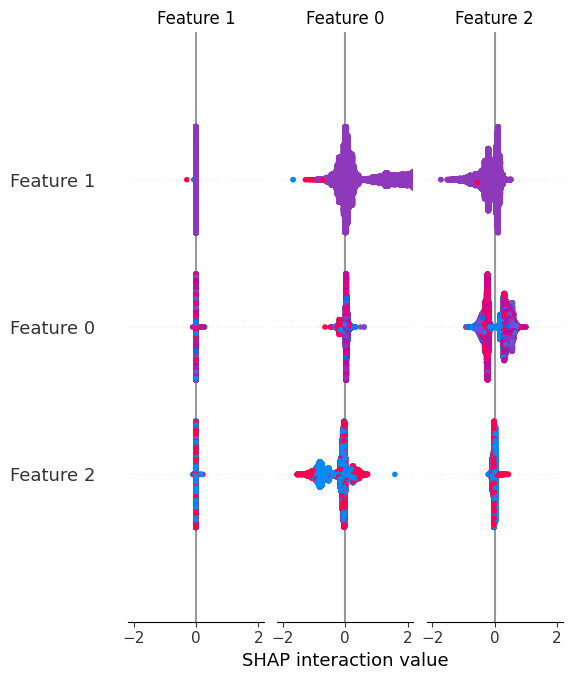
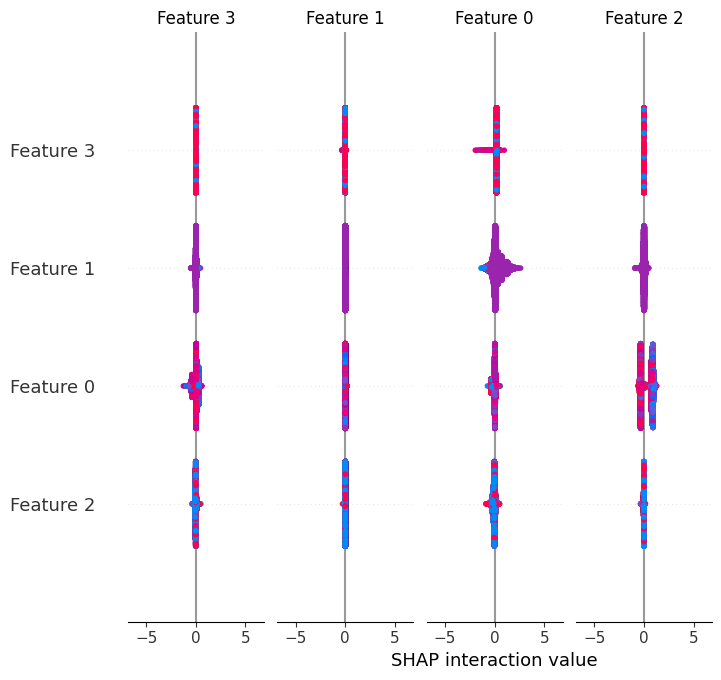


1. **Resampling Techniques**:
   * **SMOTE** and **ADASYN** were applied to handle class imbalance, significantly improving classification performance for the minority class.

Results with Resampling:



1. **Feature Selection**:
   * SHAP values were used to identify the most important features, reducing the feature space while maintaining performance.



1. **Evaluation Metrics**:
   * Classification reports were generated to evaluate precision, recall, and F1 scores, showing consistent improvements after each optimization step.

#### **Reflection on the Learning Process**

Throughout this project, I learned the critical importance of handling class imbalances and tuning hyperparameters to achieve optimal model performance. The application of SHAP for feature selection provided valuable insights into feature importance, allowing the model to focus on the most relevant data points. These improvements directly contribute to the project’s goal of creating a robust, reliable tool for diabetes classification across various subtypes.

#### **Conclusion**

The improvements made, particularly in the use of SMOTE/ADASYN for balancing and SHAP for feature selection, resulted in a significant boost in model performance, especially for underrepresented classes like gestational diabetes. These results align with the project's goal of providing an accurate, data-driven solution for early diabetes detection.