

Predictive Modeling of Diabetes, Cardiovascular Disease & Stroke

Using Public Health Survey Datasets

FINAL PROJECT
Presented by- Group 4

Devlin H, Ruman S
DATA 5100, Seattle University

The Challenge



Chronic diseases are leading causes of death and disability worldwide



Late diagnosis increases cost, complexity, and health risk



Early risk prediction can enable timely preventive intervention



Why This Matters: The Critical Need for Predictive Models in Healthcare

37M+ People Affected by Diabetes in the U.S.

Diabetes is a chronic condition impacting millions, leading to severe health complications if not managed early.

Cardiovascular Disease: Leading Global Cause of Death

Heart-related conditions remain the top cause of mortality worldwide, highlighting the urgency for early detection and intervention.

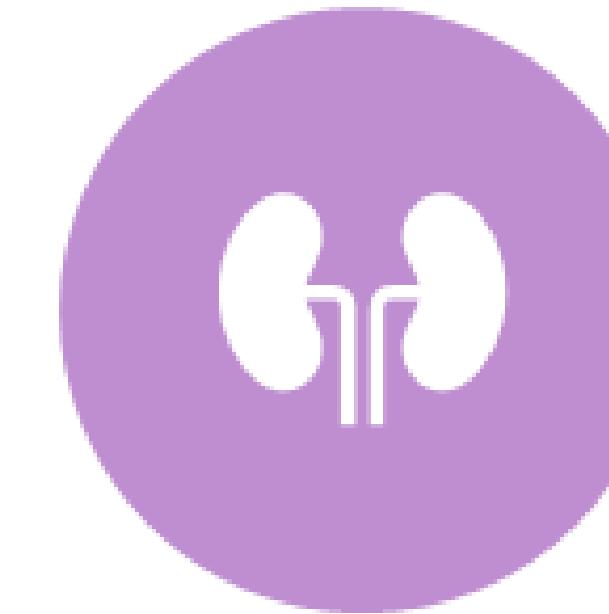
Stroke Often Leads to Long-Term Disability

A stroke can have devastating and lasting effects on a person's quality of life, emphasizing the importance of risk prediction.

Unlocking Insights: Key Research Questions



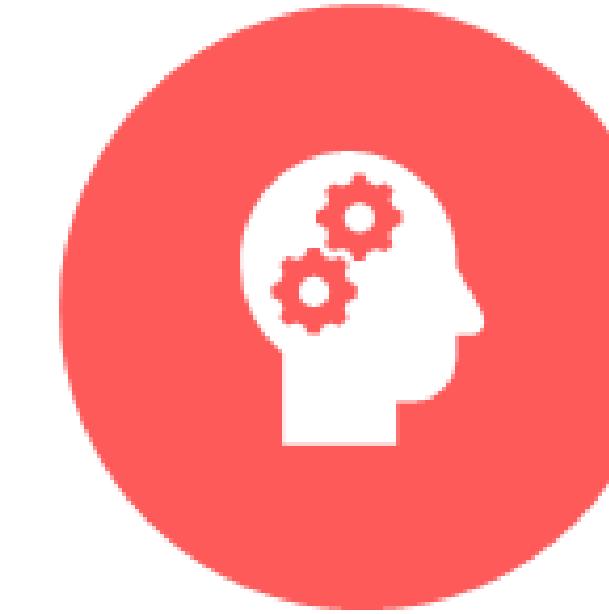
Our research aims to address critical gaps in understanding how machine learning can revolutionize disease prediction.



Can ML models effectively predict early disease risk for conditions like diabetes, cardiovascular disease, and stroke?



Which specific machine learning models demonstrate superior performance for each medical condition, and what are their strengths?



Beyond prediction, what are the most influential features or risk factors driving these predictions, offering clinical interpretability?

Dataset Overview: The Foundation of Our Analysis

Comprehensive Dataset Sizes

Our study leveraged substantial datasets to ensure robust model training and validation:

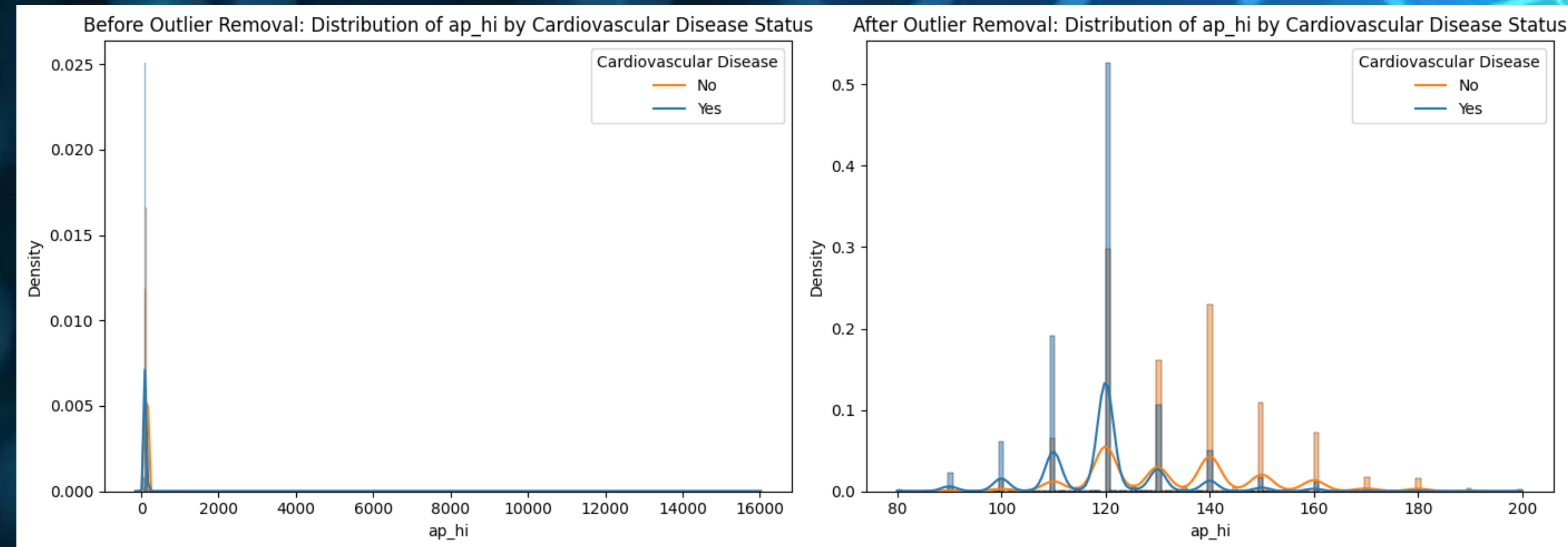
- **70,000 records** for Diabetes prediction
- **60,000 records** for Cardiovascular Disease risk
- **40,000 records** for Stroke prediction

Rich Feature Types for Prediction

Each dataset included a variety of crucial demographic and clinical features:

- **Demographic:** Age, gender, and BMI
- **Biometric:** Blood pressure, glucose levels, HbA1c
- **Lifestyle:** Detailed smoking history

Data Preparation & Preprocessing: Building a Clean Foundation



Before model training, meticulous data preparation was essential to ensure data quality and model performance.

Raw Data Collection:

Initial aggregation of diverse patient health records.

Clean Data:

Handled missing values, removed outliers, and corrected inconsistencies.

Encode Categories:

Converted categorical variables into numerical formats for model compatibility.

Scale Features:

Normalized numerical features to prevent dominance by variables with larger ranges.

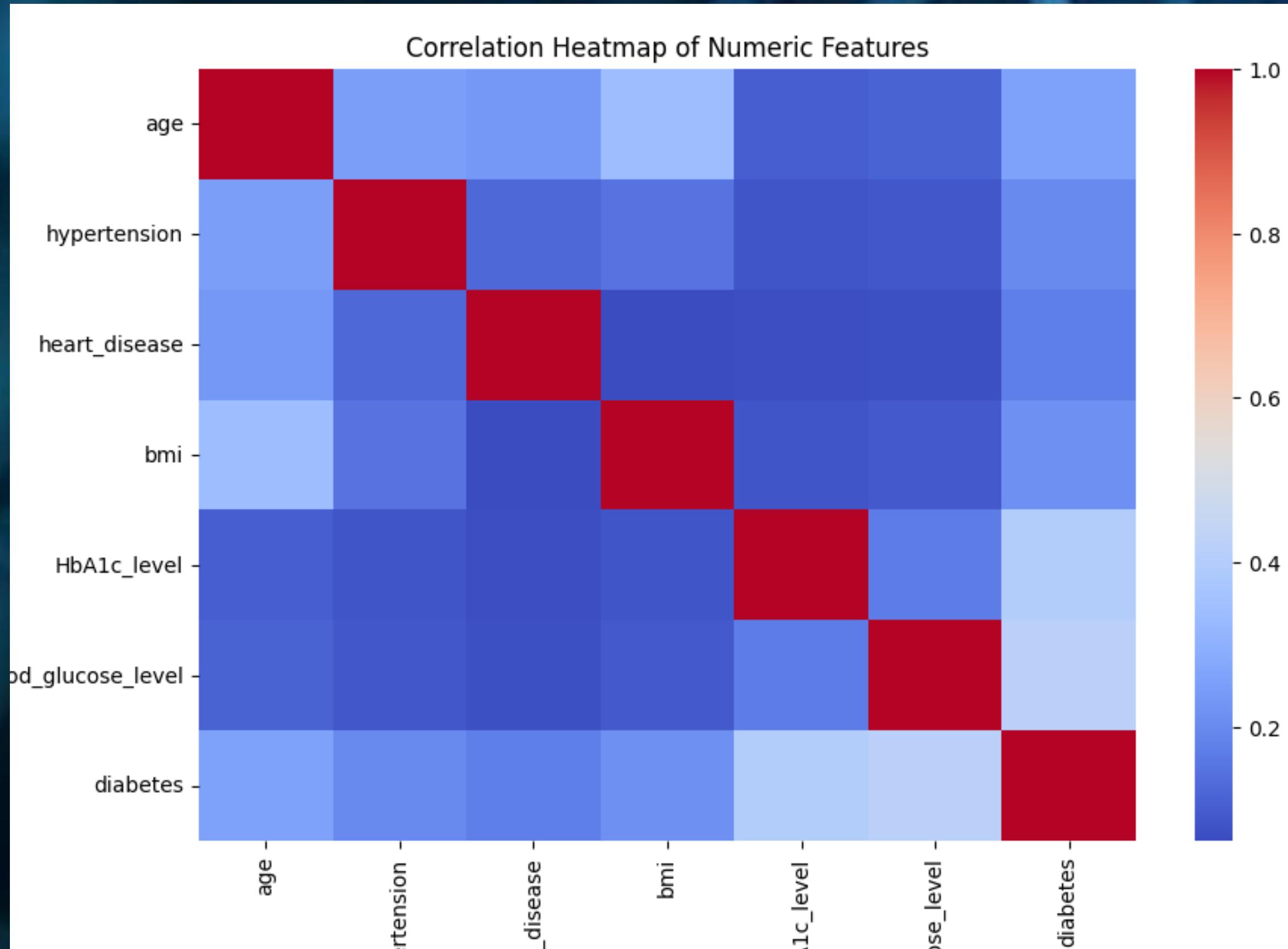
Feature Selection:

Reduced dimensionality by selecting the most relevant features and removing redundancy.

Train/Test Split:

Divided data into training and testing sets to evaluate model generalization.

Exploratory Data Analysis: Uncovering Key Health Relationships



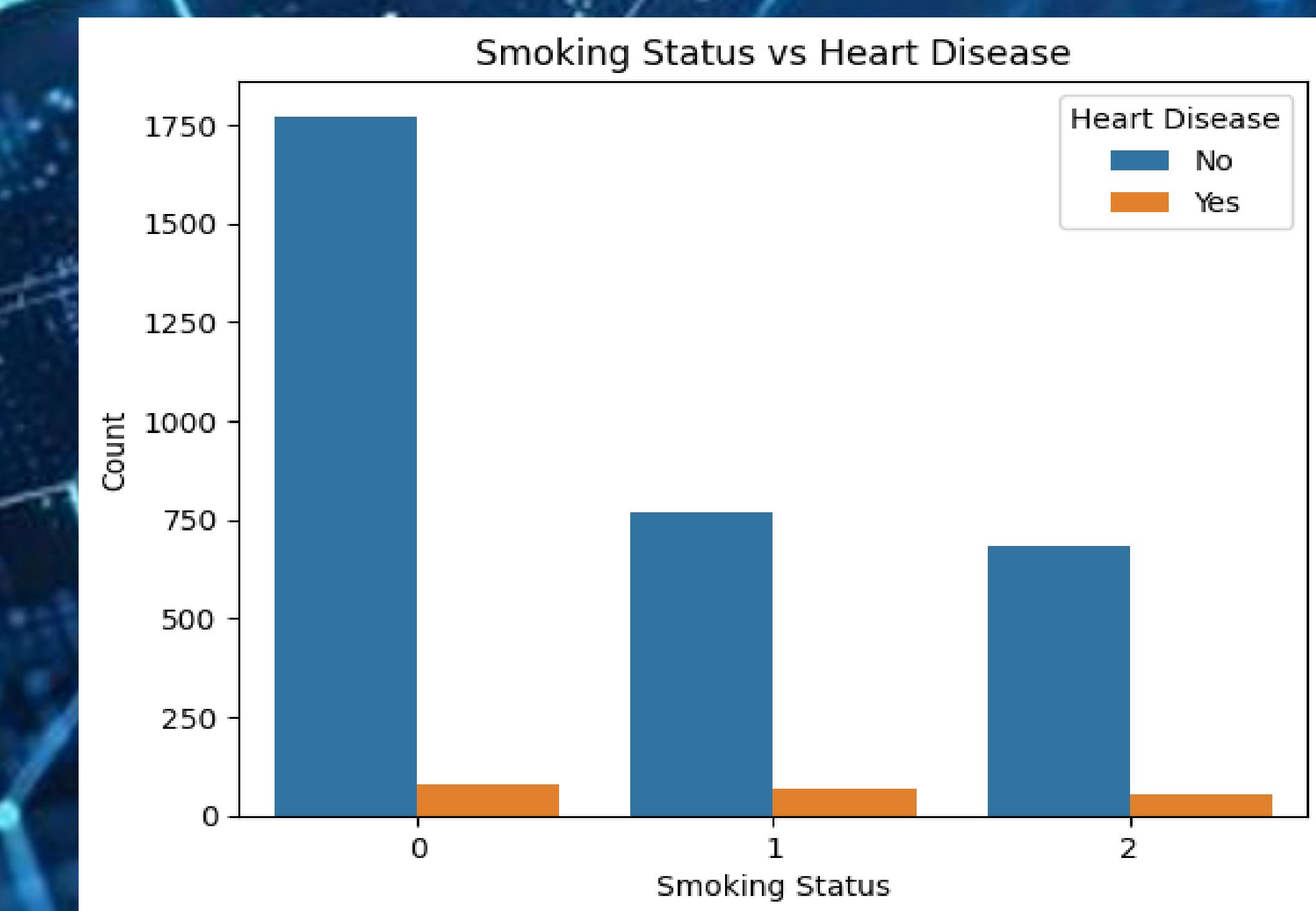
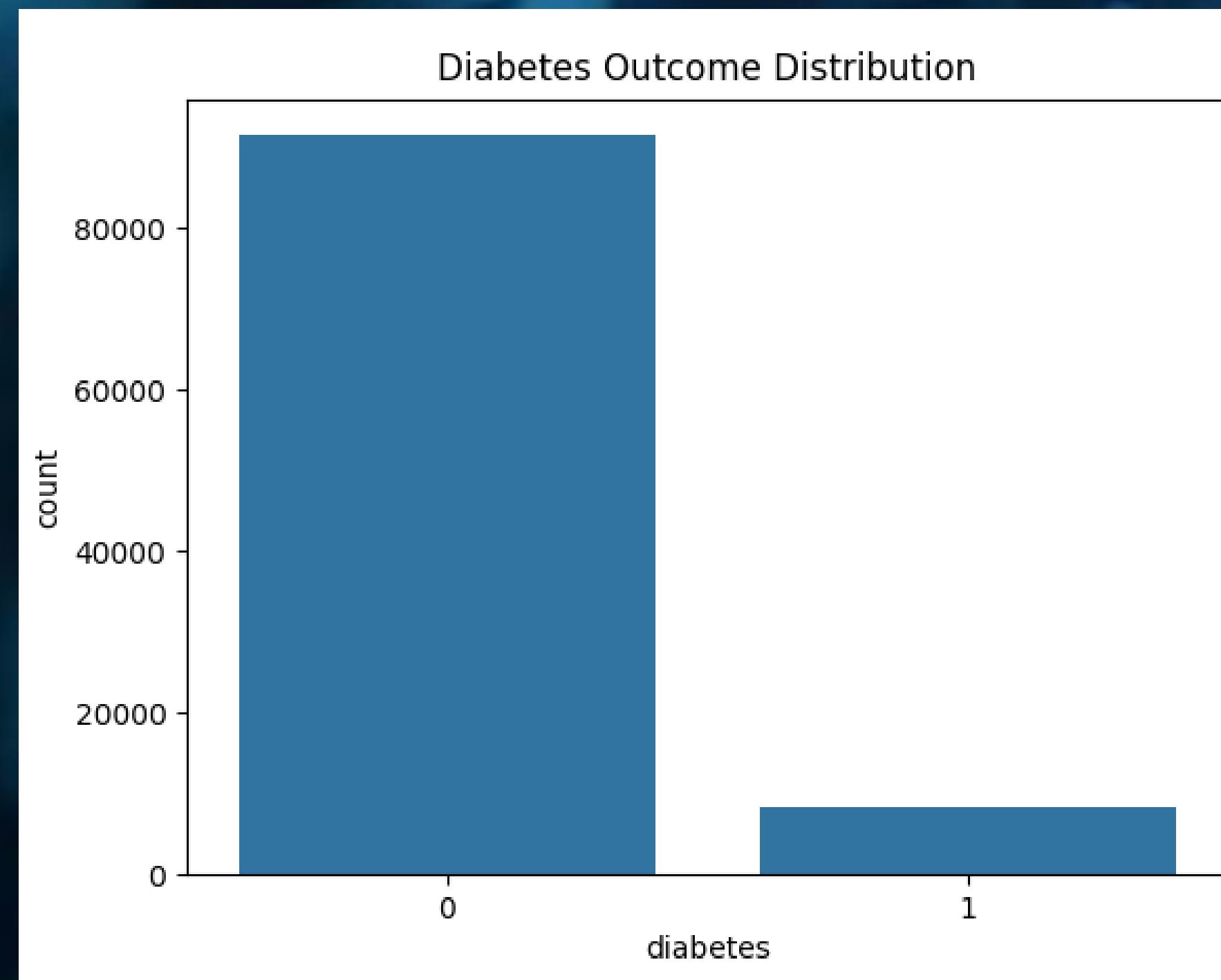
Correlation Heatmap

This heatmap visually represents the interdependencies and strengths of relationships between various health metrics and disease outcomes.

Exploratory Data Analysis: Uncovering Key Health Relationships

Key Insights from EDA

- **Diabetes Indicators:** HbA1c, glucose levels, and BMI showed strong correlations with diabetes risk.
- **Cardiovascular Risk Factors:** Age and blood pressure were significantly correlated with an increased risk of cardiovascular disease.
- Outcome imbalance highlights the need for careful model evaluation and calibration.



Clear gradient: smokers show higher heart disease counts

Deeper Dive: Key Insights from EDA

Our exploratory data analysis revealed crucial characteristics of the datasets, guiding our modeling strategy.

- ✓ **Nonlinear Relationships:** Many health parameters exhibited complex, nonlinear relationships with disease incidence, suggesting the need for advanced ML techniques.
- ✓ **Feature Interactions:** We observed significant interactions between features, where the effect of one factor was modified by another, profoundly influencing overall risk prediction.
- ✓ **Complexity Confirmed:** These findings reinforced the necessity of employing machine learning models capable of capturing such intricate patterns, rather than relying on simpler statistical methods.



Modeling Approach: Selecting the Right Tools

To tackle the complexity of disease prediction, we employed a diverse set of machine learning models, each with unique strengths.

Models Tested:

Logistic Regression: A baseline model for binary classification, providing a linear perspective.

Decision Tree: A non-linear model that partitions data based on feature values, offering interpretability.

Random Forest: An ensemble method combining multiple decision trees to improve accuracy and reduce overfitting.

XGBoost: A gradient boosting framework known for its high performance and efficiency in complex datasets

Metrics Used:

Model performance was rigorously evaluated using a comprehensive suite of metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

Why Random Forest: The Optimal Choice for Healthcare Prediction

Our choice of Random Forest was driven by its unique combination of power and interpretability, making it ideal for the nuances of healthcare data.

1. Handles Nonlinearities

Effectively captures complex, non-linear relationships within the data, crucial for accurate disease prediction.

2. Captures Feature Interactions

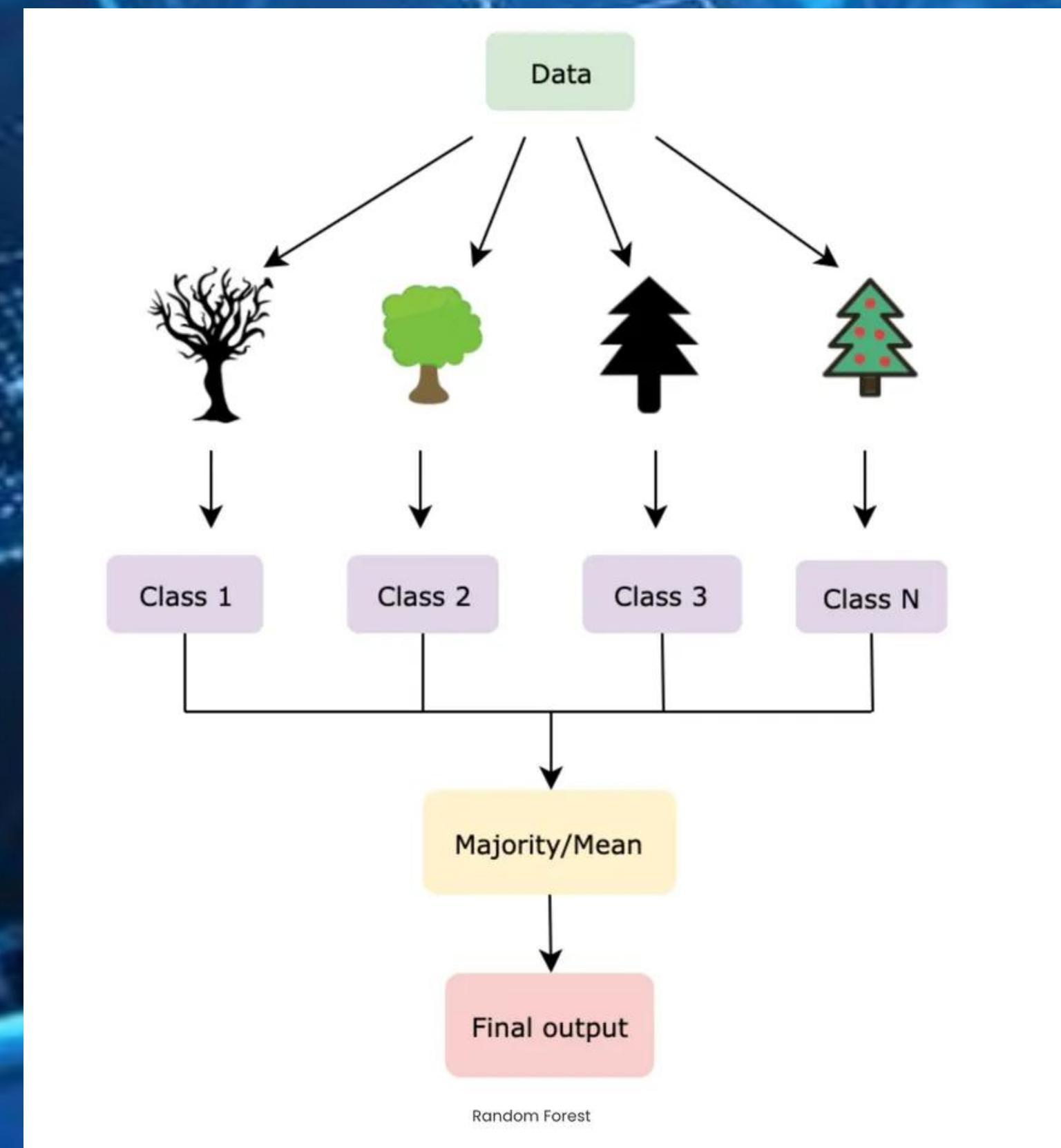
Excels at identifying how different risk factors combine to influence disease outcomes, providing a holistic view.

3. Low Overfitting

Its ensemble nature inherently reduces the risk of overfitting, leading to more generalizable and reliable predictions.

4. Interpretable Feature Importance

Provides clear insights into which features are most influential, aiding clinical understanding and decision-making.



Diabetes Model Results: High Efficacy in Early Prediction

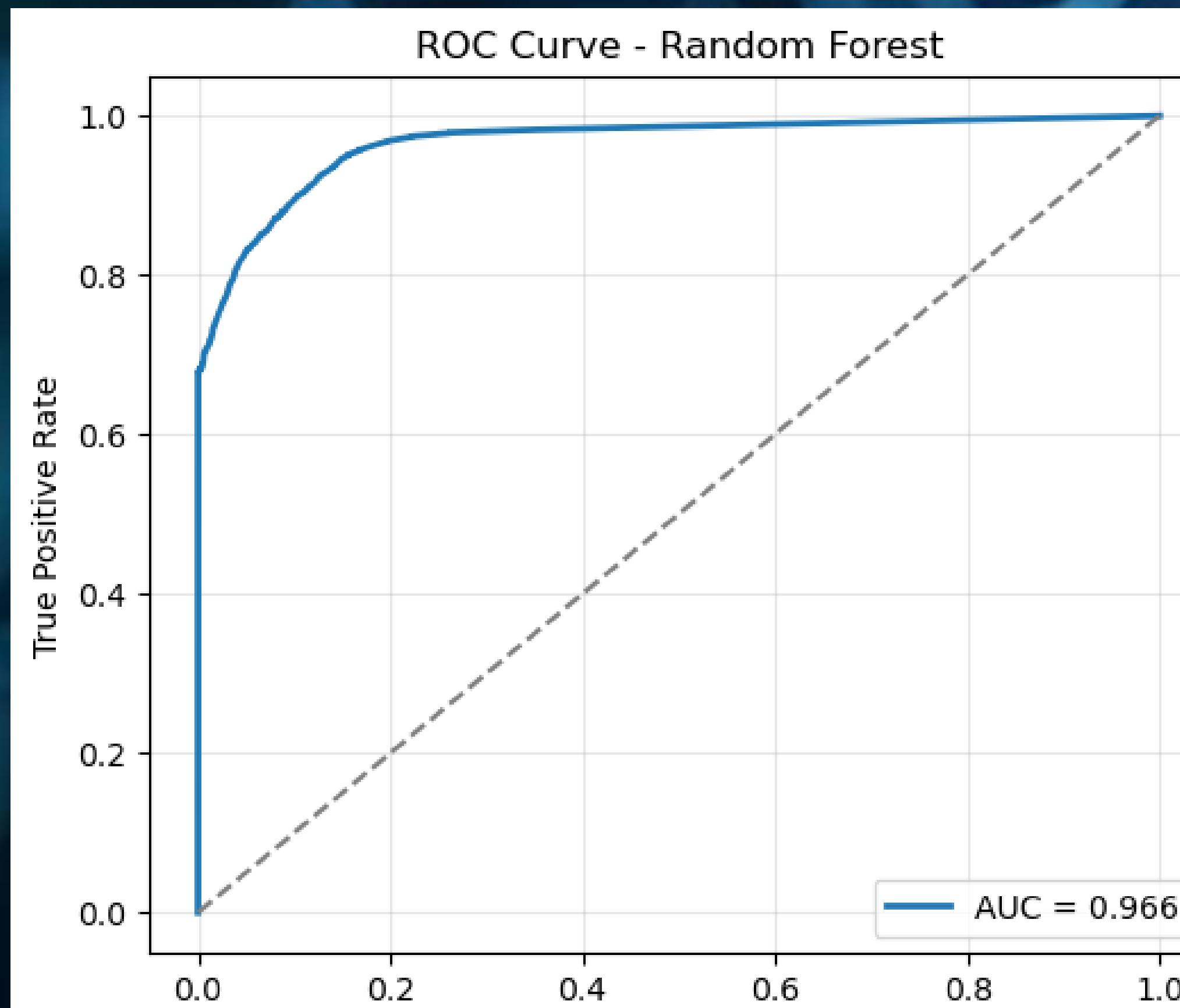


Figure. ROC Curve (Diabetes)

ROC Curve for Diabetes Prediction

The high AUC score indicates excellent discrimination ability, meaning the model is very good at distinguishing between patients who will and will not develop diabetes.

Diabetes Model Results: High Efficacy in Early Prediction

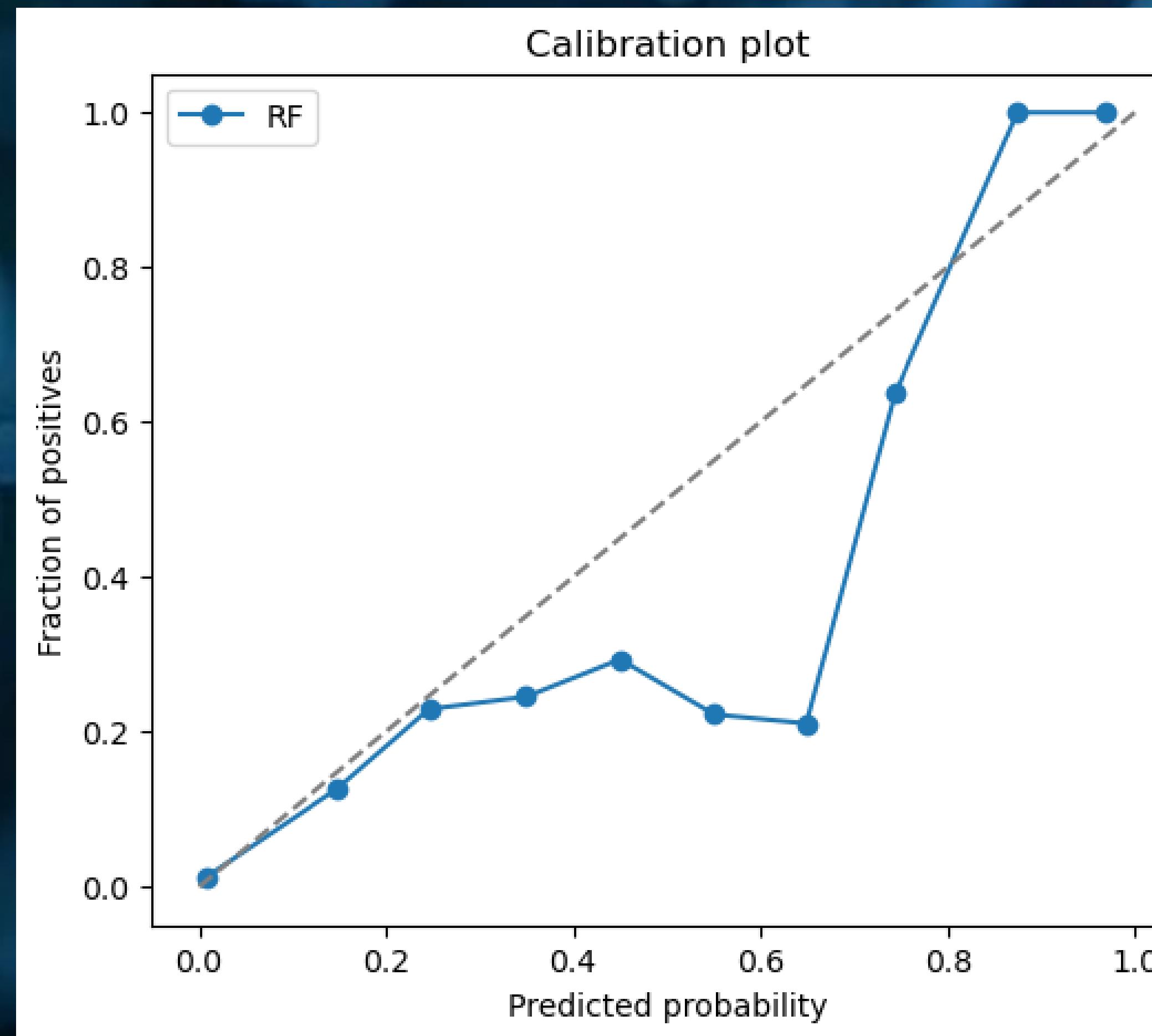
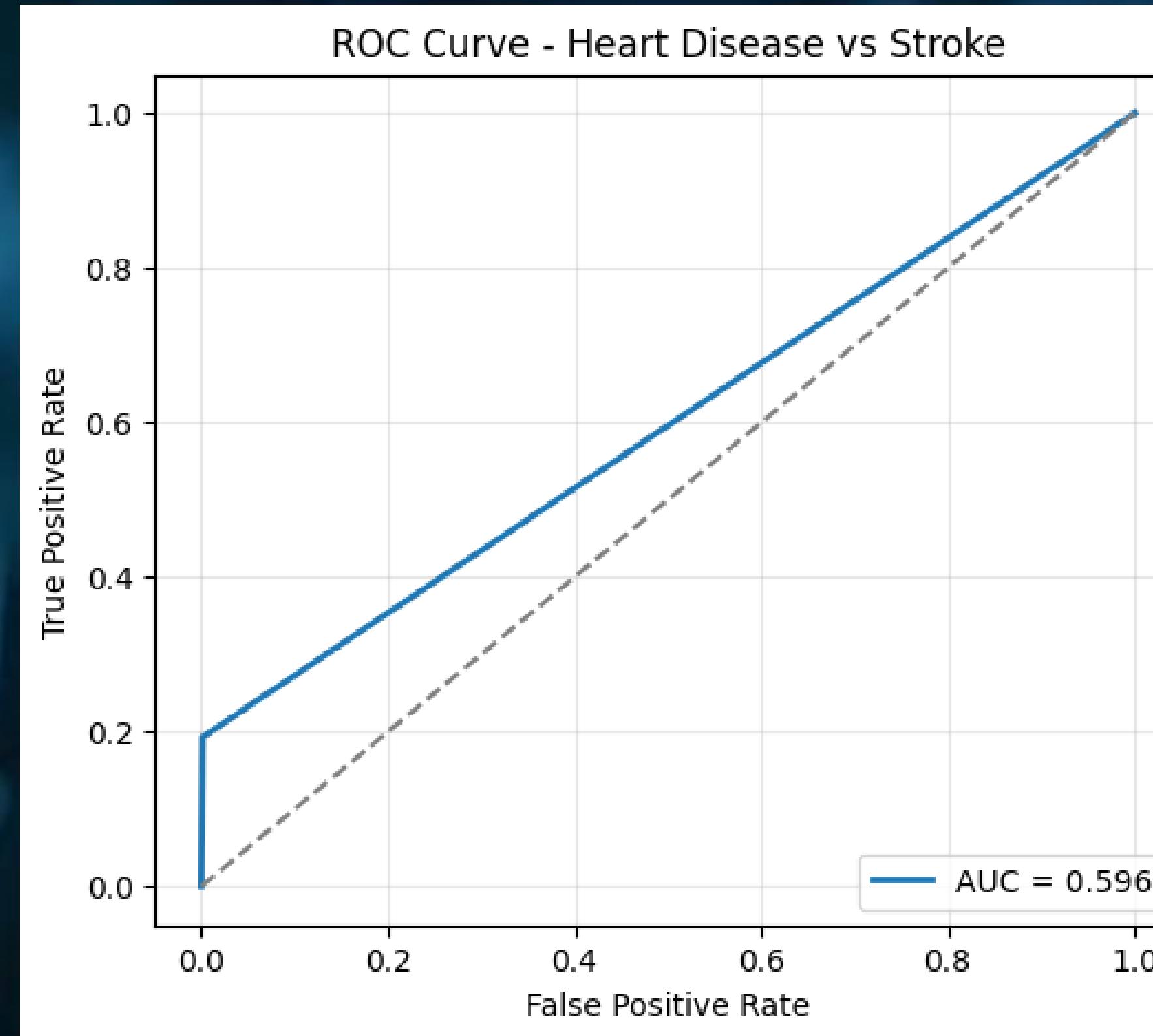


Figure. Calibration Plot - Random Forest (Diabetes)

Key Performance Indicators

1. **ROC-AUC ≈ 0.96 :** Demonstrates outstanding predictive accuracy and discrimination for diabetes risk.
2. **Top Predictors:** HbA1c, Glucose levels, and Body Mass Index (BMI) were identified as the most significant factors, aligning with medical understanding.

Cardiovascular Disease & Stroke: Prediction Outcomes



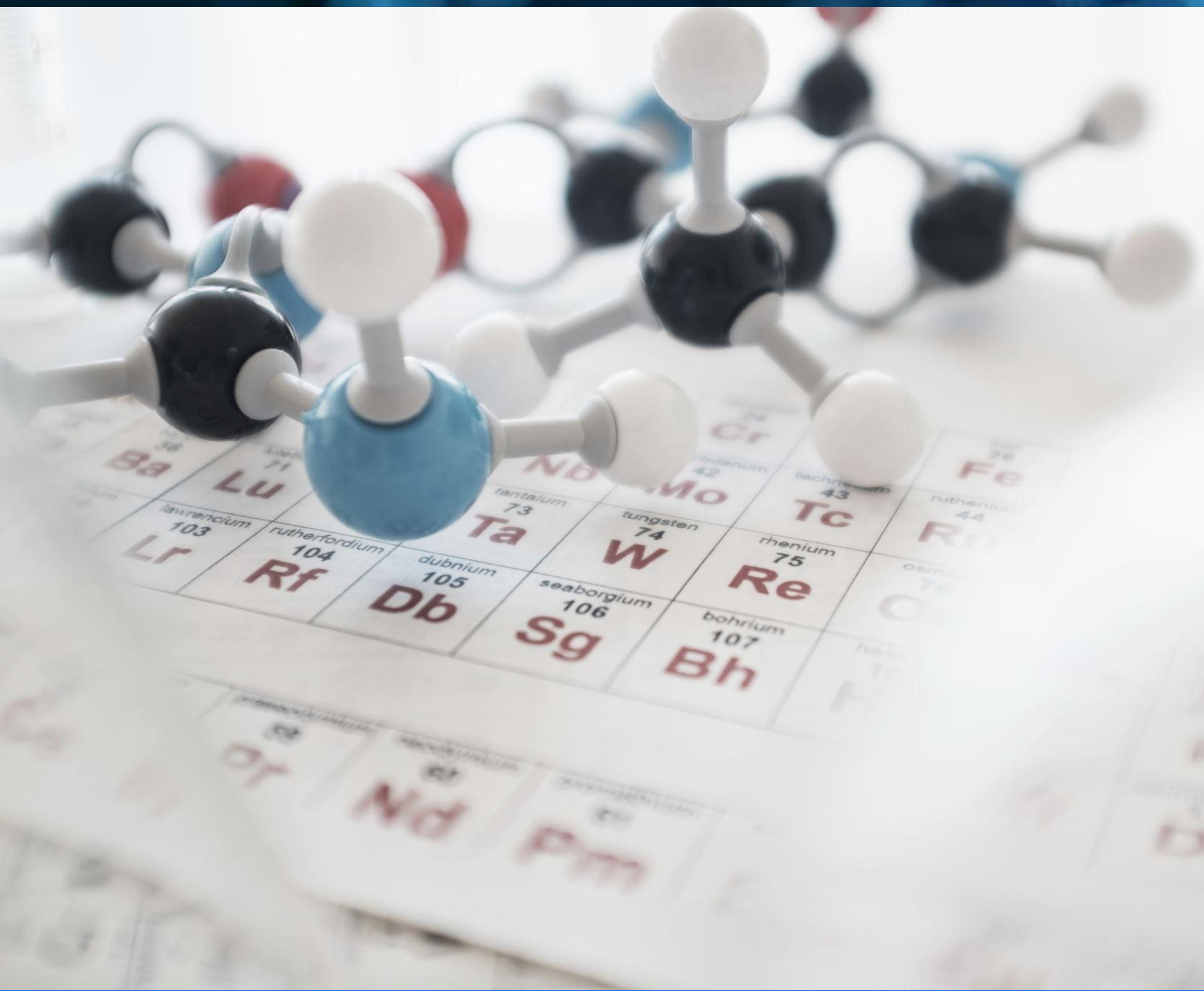
Cardiovascular Disease Results

- **Strong Performance:** Achieved high scores across accuracy, precision, and recall, indicating robust model efficacy.
- **Top Predictors:** Age, smoking history, blood pressure, and BMI were consistently identified as critical risk factors.

Stroke Results

- **Meaningful Performance:** While moderate, the model showed significant predictive power, offering valuable insights for early intervention.
- **Important Predictors:** Age, hypertension (high blood pressure), and smoking emerged as key factors influencing stroke risk.

Figure. Cardiovascular Vs Stroke (ROC Curve)



Feature Importance & Interpretation: Aligning with Medical Science

This chart illustrates the relative importance of each feature in our predictive models, with higher values indicating greater influence on disease risk.

- ✓ **Age:** Consistently the most influential factor across all disease models, underscoring its profound impact on health risk.
- ✓ **Metabolic Factors:** BMI, blood pressure, and glucose levels emerged as strong contributors to risk prediction, aligning with established medical understanding.
- ✓ **Medical Alignment:** Our findings strongly support known medical science, providing robust evidence for the utility of ML in preventive healthcare.

Ethical Considerations & Responsible AI in Healthcare



Dataset Imbalance

Acknowledging and mitigating potential biases from imbalanced datasets to ensure fair predictions across all patient groups.

Potential Demographic Bias

Carefully assessing and addressing any inherent biases related to demographic factors to prevent discriminatory outcomes.

Support, Not Replace, Clinicians

Emphasizing that ML predictions are assistive tools designed to augment, not supersede, expert medical judgment and patient care.

Limitations & Future Work: Expanding the Horizon



Current Limitations

- **Class Imbalance:** The unequal distribution of disease cases in datasets requires advanced handling techniques.
- **Limited Demographic Variety:** The generalizability of models could be enhanced with broader and more diverse demographic data.
- **Static Survey Data:** Reliance on historical, static data might not fully capture dynamic health changes over time.

Future Directions

- **Hyperparameter Tuning:** Further optimization of model parameters to achieve peak performance.
- **Additional Clinical Variables:** Incorporating more real-time and longitudinal clinical data for enhanced prediction accuracy.
- **Interactive Demo Development:** Building a small, interactive demonstration tool for clinicians to explore model insights firsthand.

Conclusion

Machine learning holds immense promise in supporting the early prediction of chronic disease risk. Our models demonstrated strong predictive performance and identified meaningful, medically relevant risk factors, paving the way for proactive healthcare interventions.



GitHub Link: https://github.com/simplyyweirdd3/data5100_project

Any Questions?