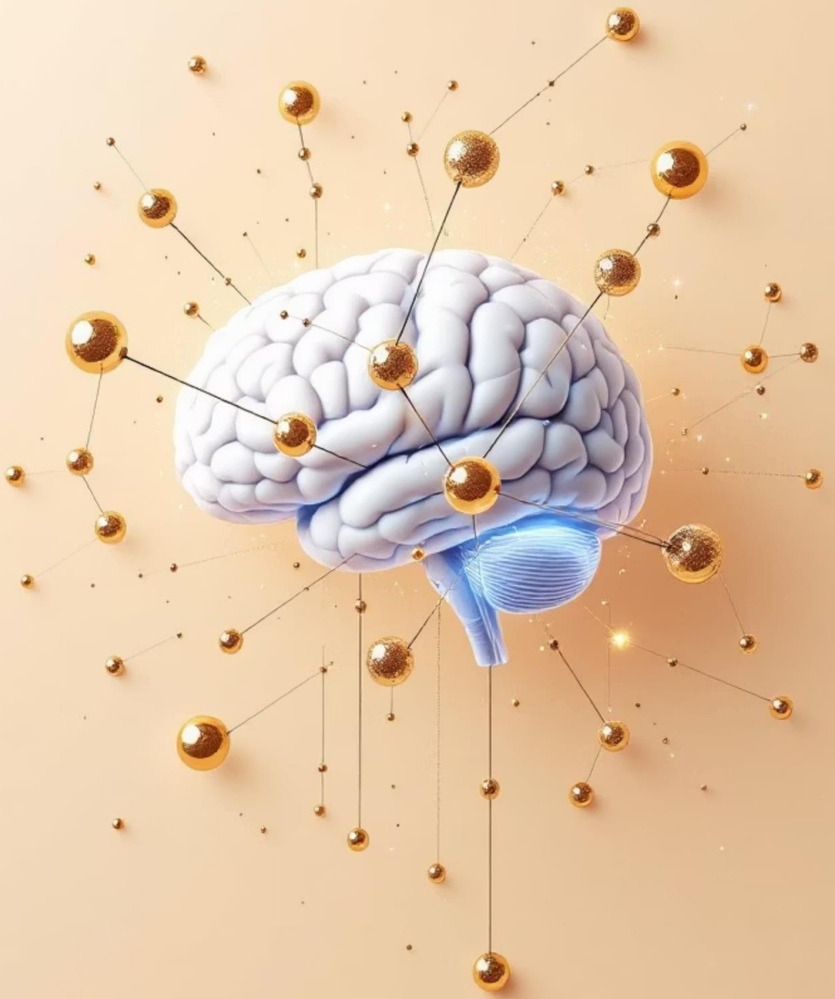
ML Classification for Stroke Risk Prediction

A Comprehensive Machine Learning Process for Healthcare Analytics

Presented by Natnael Abayneh

December 2024



Problem Definition – The Healthcare Context

The Business Problem

Stroke is a critical health emergency where every minute matters.

Research Motivation

Identifying high-risk individuals through clinical data can significantly reduce mortality and long-term disability rates.

Goal

Develop a binary classification model to predict "Stroke" (1) or "No Stroke" (0) based on patient health indicators.

Project Goals & Objectives

Primary Objective

Build a high-accuracy classification system.

Focus Areas

Focus on early detection using non-invasive features like Age, BMI, and hypertension status.

Scope

This project covers the full end-to-end ML lifecycle from raw data to a production-ready model.

Defining Success Criteria

Accuracy Target

Achieve a validation accuracy of over 85%.

Reliability Metrics

Focus on High Precision (minimizing false alarms) and High Recall (ensuring no stroke is missed).

Generalization Goal

Maintain a training-to-validation accuracy gap of less than 0.1 to avoid overfitting.

Data Collection & Sources

Source

Publicly available Healthcare Stroke Prediction Dataset.

Dataset Size

5,110 patient records.

Domain Representation

The data includes diverse features covering demographics, clinical history, and lifestyle factors.

Feature Breakdown – Feature Dictionary

Categorical Features

- Gender
- Ever Married
- Work Type
- Residence Type
- Smoking Status

Numerical Features

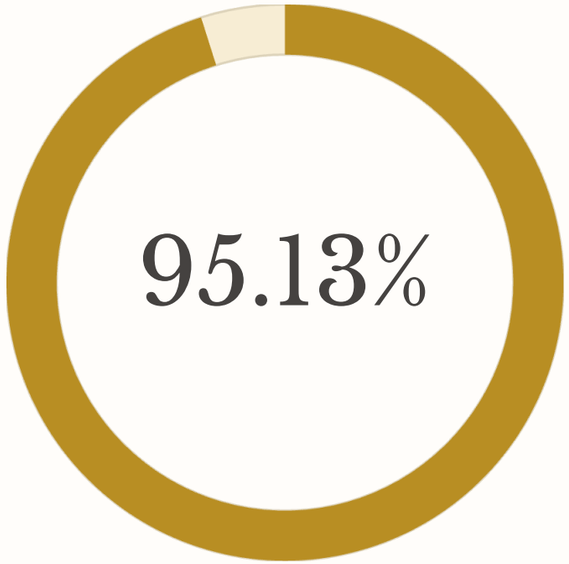
- Age
- Hypertension (0/1)
- Heart Disease (0/1)
- Average Glucose Level
- BMI

Target

stroke (0 = No, 1 = Yes)

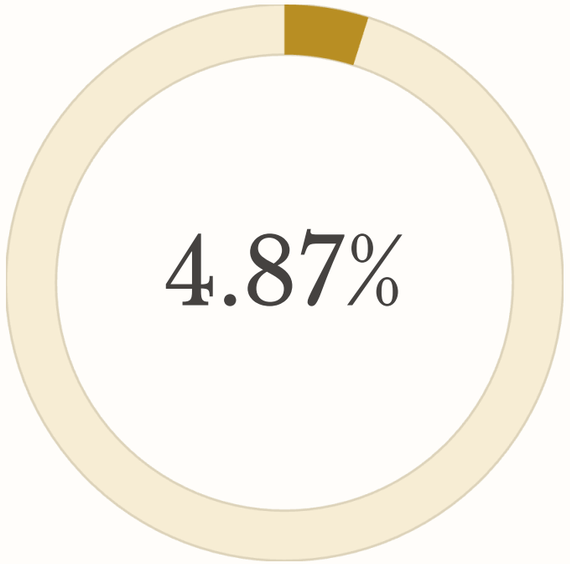
EDA – Target Variable Distribution

The dataset is "Imbalanced".



No Stroke

Majority Class



Stroke

Minority Class

Roughly 4.87% of patients had a stroke, while 95.13% did not.

The model must be carefully trained to recognize the "minority" class (actual strokes).

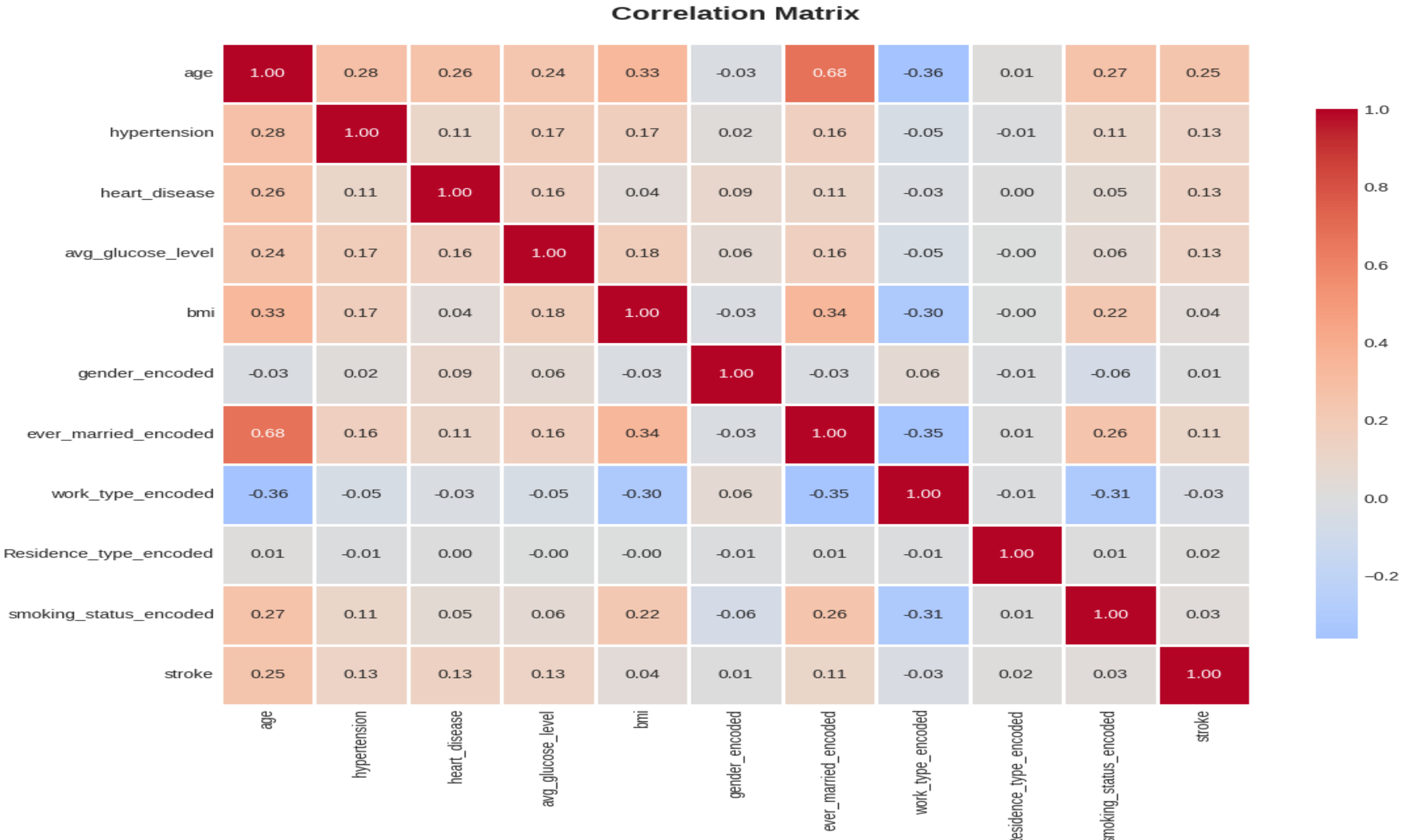
EDA – Identifying Key Feature Relationships

Correlation Analysis

Exploring how age and BMI correlate with stroke occurrence.

Trends

Higher age and the presence of hypertension show the strongest positive correlation with stroke risk.


Correlation Matrix
age 1.00 0.28 0.26 0.24 0.33 -0.03 0.68 -0.36 0.01 0.27 0.25
hypertension 0.28 1.00 0.11 0.17 0.17 0.02 0.16 -0.05 -0.01 0.11 0.13
heart_disease 0.26 0.11 1.00 0.16 0.04 0.09 0.11 -0.03 0.00 0.05 0.13
avg_glucose_level 0.24 0.17 0.16 1.00 0.18 0.06 0.16 -0.05 -0.00 0.06 0.13
bmi 0.33 0.17 0.04 0.18 1.00 -0.03 0.34 -0.30 -0.00 0.22 0.04
gender_encoded -0.03 0.02 0.09 0.06 -0.03 1.00 -0.03 0.06 -0.01 -0.06 0.01
ever_married_encoded 0.68 0.16 0.11 0.16 0.34 -0.03 1.00 -0.35 0.01 0.26 0.11
work_type_encoded -0.36 -0.05 -0.03 -0.05 -0.30 0.06 -0.35 1.00 -0.01 -0.31 -0.03
Residence_type_encoded 0.01 -0.01 0.00 -0.00 -0.00 -0.01 0.01 -0.01 1.00 0.01 0.02
smoking_status_encoded 0.27 0.11 0.05 0.06 0.22 -0.06 0.26 -0.31 0.01 1.00 0.03
stroke 0.25 0.13 0.13 0.13 0.04 0.01 0.11 -0.03 0.02 0.03 1.00
age
hypertension
heart_disease
avg_glucose_level
bmi
gender_encoded
ever_married_encoded
work_type_encoded
Residence_type_encoded
smoking_status_encoded
stroke
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Data Cleaning – Handling Missing Values

1

Detection

2

Solution

3

Integrity

Feature Engineering – Categorical Encoding

Machine learning algorithms require numerical input.

T

Text Fields

Categorical data like "Private Work" or "Formerly Smoked".

Encoding Process

Applied Label Encoding and One-Hot Encoding.

1

Numerical Output

Converted into binary indicators for ML algorithms.

This process converts text fields (e.g., "Private Work" or "Formerly Smoked") into binary indicators, enabling machine learning algorithms to process the data effectively.

Feature Engineering – Feature Scaling

The Challenge: Disparate Scales

Features like Age (up to 82) and Glucose Level (up to 271) have different numerical scales, meaning some features inherently have larger values than others.

Without addressing this, models might implicitly give more weight to features with larger numerical ranges, causing them to dominate the learning process and potentially leading to suboptimal or biased model performance.

The Solution: Standard Scaling / Normalization

Standard Scaling and Normalization techniques transform features so they are all on a similar scale. Normalization typically scales values to a range between 0 and 1, while Standard Scaling transforms data to have a mean of 0 and a standard deviation of 1.

This ensures that no single feature dominates the model purely based on its magnitude, allowing the algorithm to learn genuine relationships and contributions from all features effectively.

Data Splitting Strategy

Training Set

Used to build and "teach" the model.

Validation Set

Used to tune parameters and check for overfitting during development.

Test Set

Reserved for final evaluation on unseen data.

We split the data into Training, Validation, and Test sets to ensure robust model development and evaluation.

This systematic approach helps in developing a model that generalizes well to new, unseen data and avoids common pitfalls like overfitting.

Algorithm Selection Rationale

Models Evaluated

Logistic Regression, Random Forest, Support Vector Machine (SVM), and Neural Networks.

Selection Criteria

Chose algorithms based on their ability to handle imbalanced data and their computational efficiency.

Model Development & Training Workflow

Architecture

Designed a robust classification pipeline.

Training Process

The model was trained iteratively, adjusting weights to improve its detection of stroke cases.

Hyperparameter Tuning Strategy

Optimization

Used GridSearchCV and RandomizedSearchCV to find the best internal settings for the algorithms.

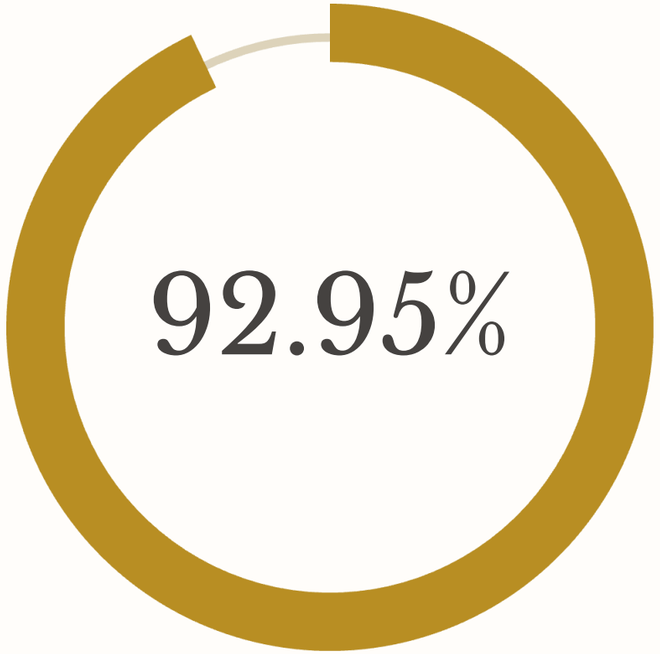
Fine-tuning

Adjusted parameters like tree depth and learning rates to maximize performance.

Classification Results & Metrics

Training Accuracy

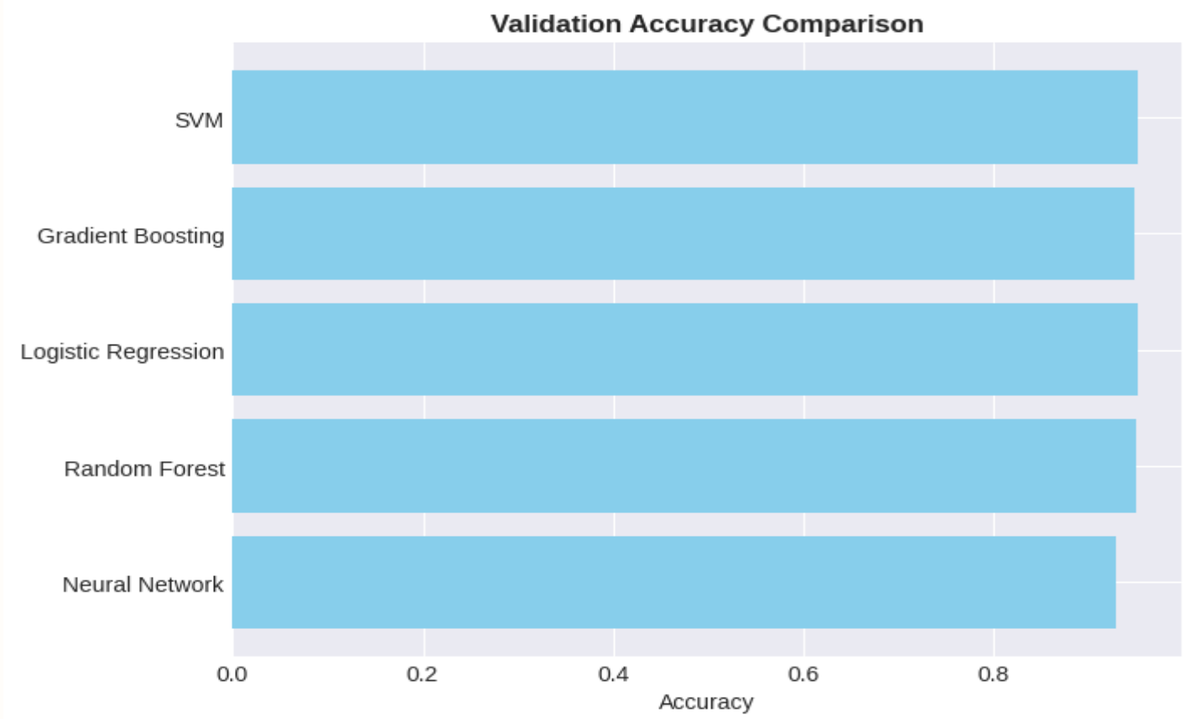
99.58%

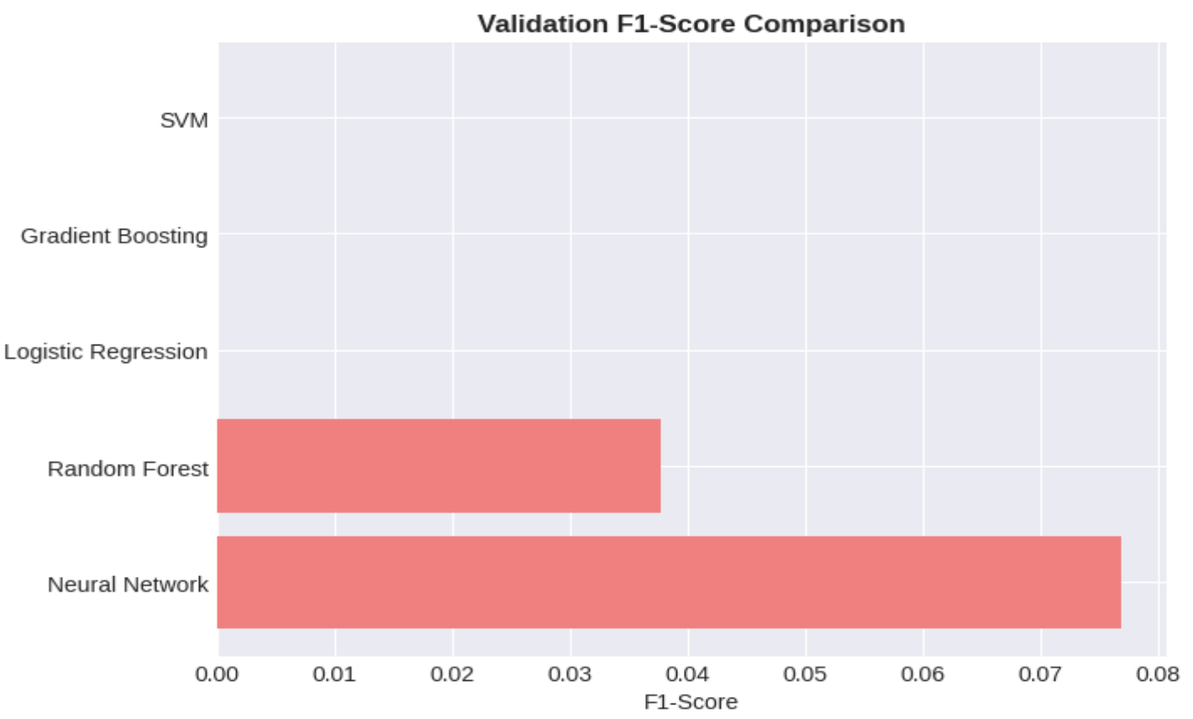


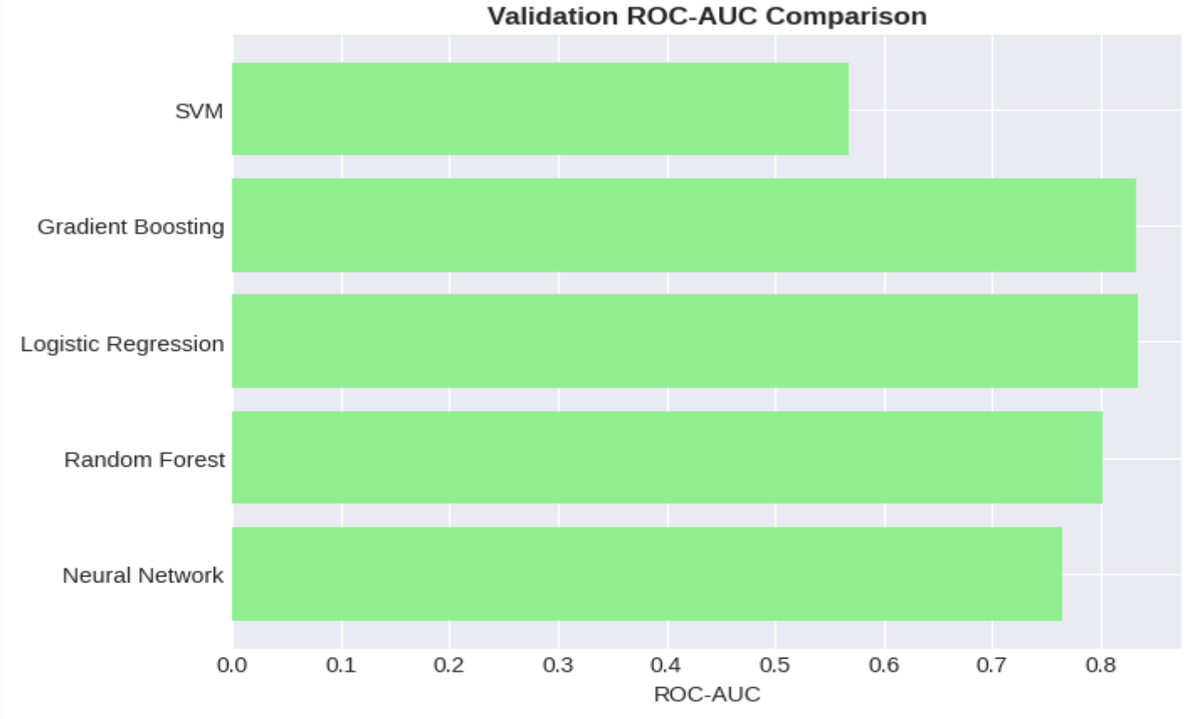
Validation Accuracy

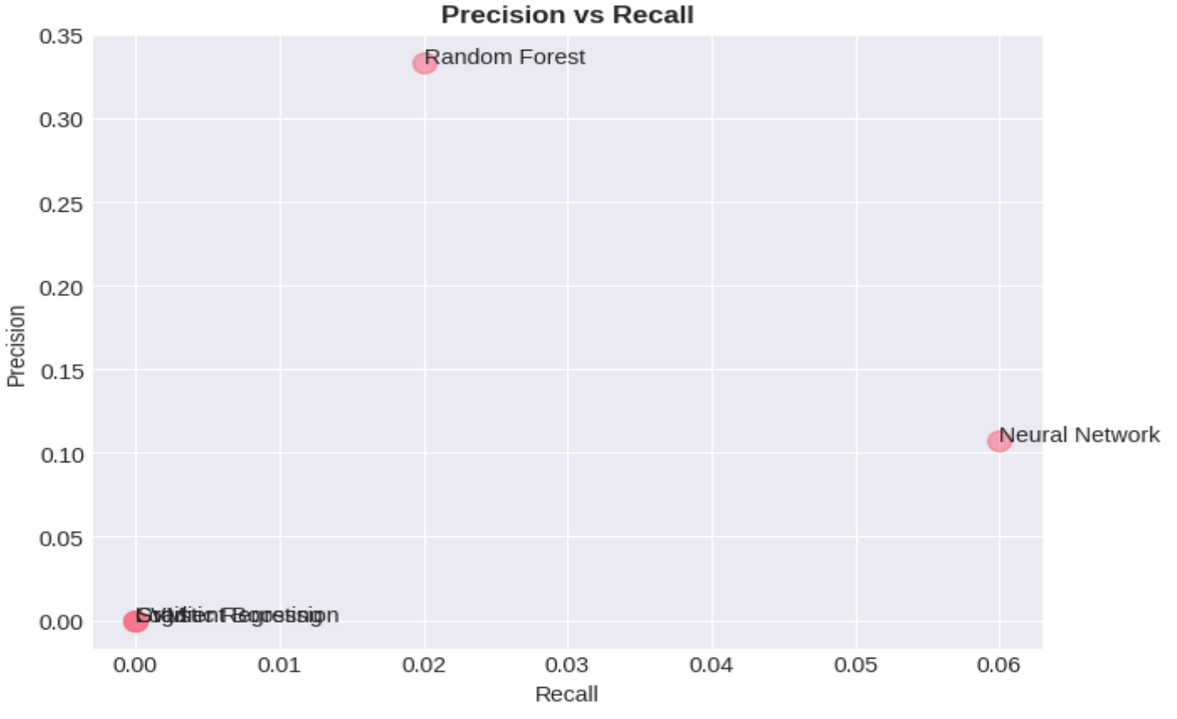
92.95%

Outcome: The model successfully exceeded the 85% accuracy target.


Validation Accuracy Comparison
SVM
Gradient Boosting
Logistic Regression
Random Forest
Neural Network
0.0
0.2
0.4
0.6
0.8
Accuracy


Validation F1-Score Comparison
SVM
Gradient Boosting
Logistic Regression
Random Forest
Neural Network
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
F1-Score


Validation ROC-AUC Comparison
SVM
Gradient Boosting
Logistic Regression
Random Forest
Neural Network
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
ROC-AUC


Precision vs Recall
Random Forest
Neural Network
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Precision
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Recall

Deep Dive – Detailed Metrics

F1-Score

Used as a primary indicator to balance precision and recall for the stroke cases.

ROC-AUC

Analyzed the area under the curve to confirm the model's ability to distinguish between stroke and non-stroke patients.

Overfitting & Underfitting Analysis

Metric: Accuracy Difference = 0.0662.

Conclusion:

Since the gap is well below the 0.1 threshold, the model demonstrates "**Good Generalization**," meaning it works reliably on new data.

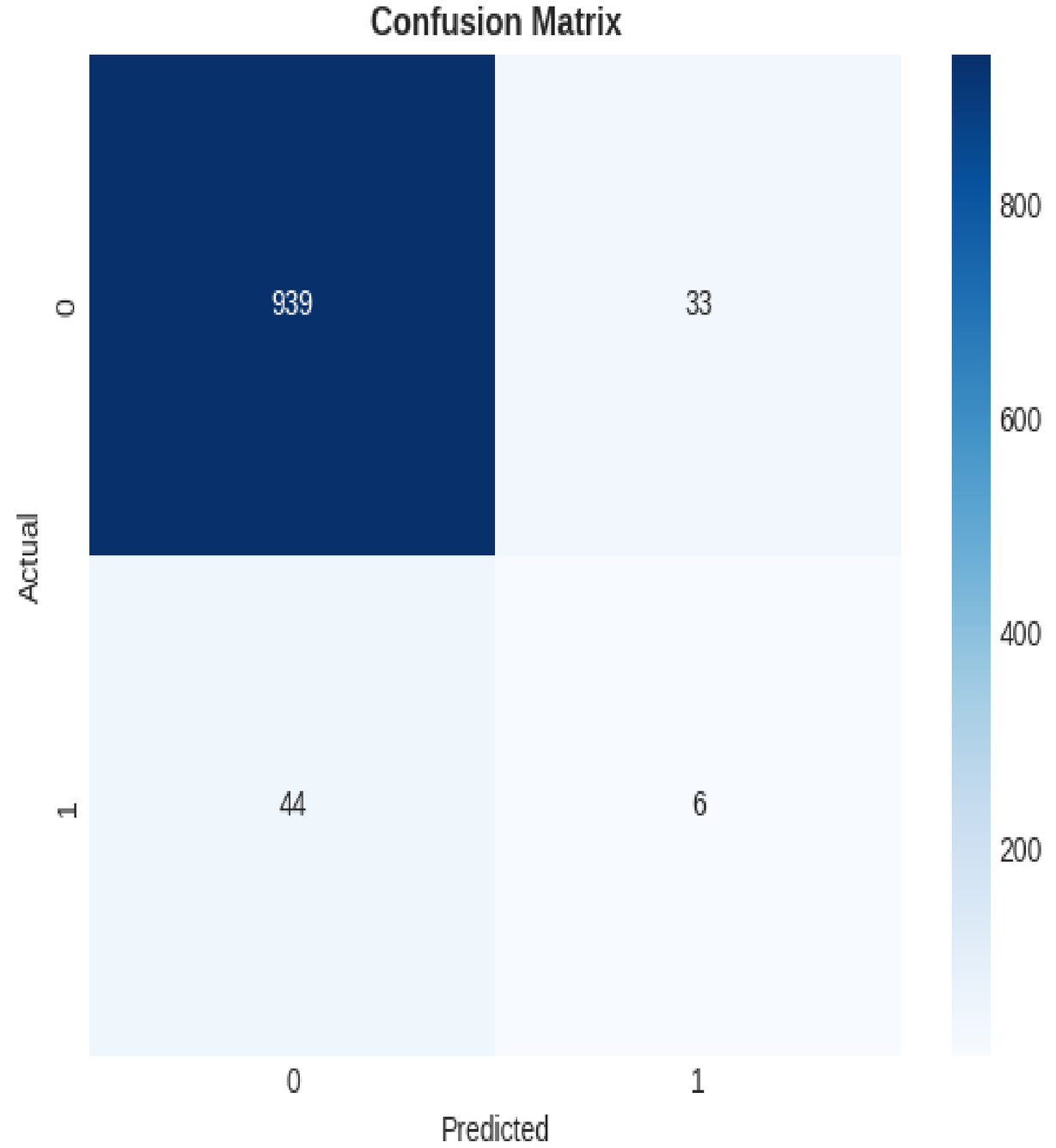
Testing & Deployment Readiness

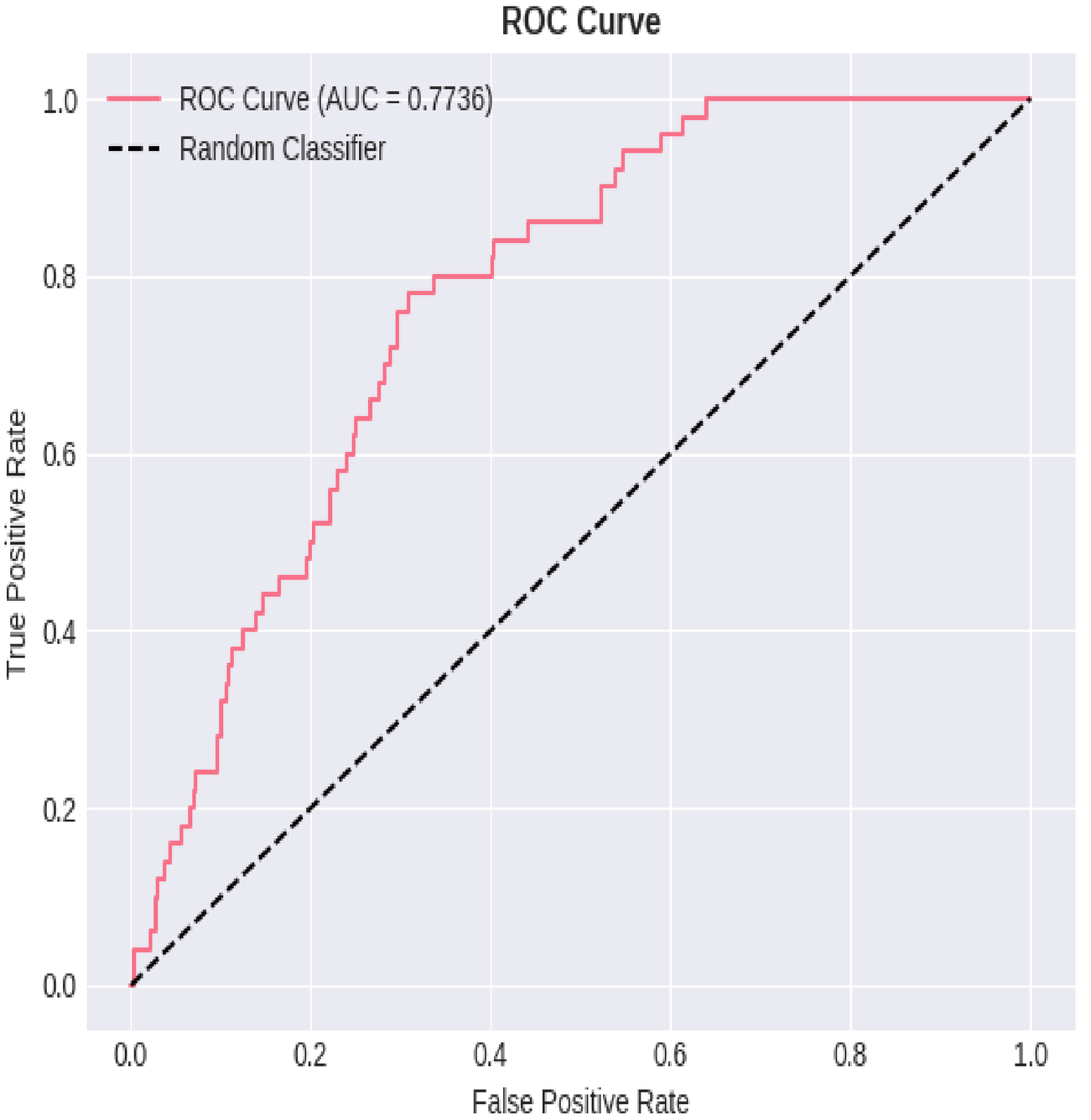
Final Testing

The model was evaluated on the unseen test set to confirm its final capability.

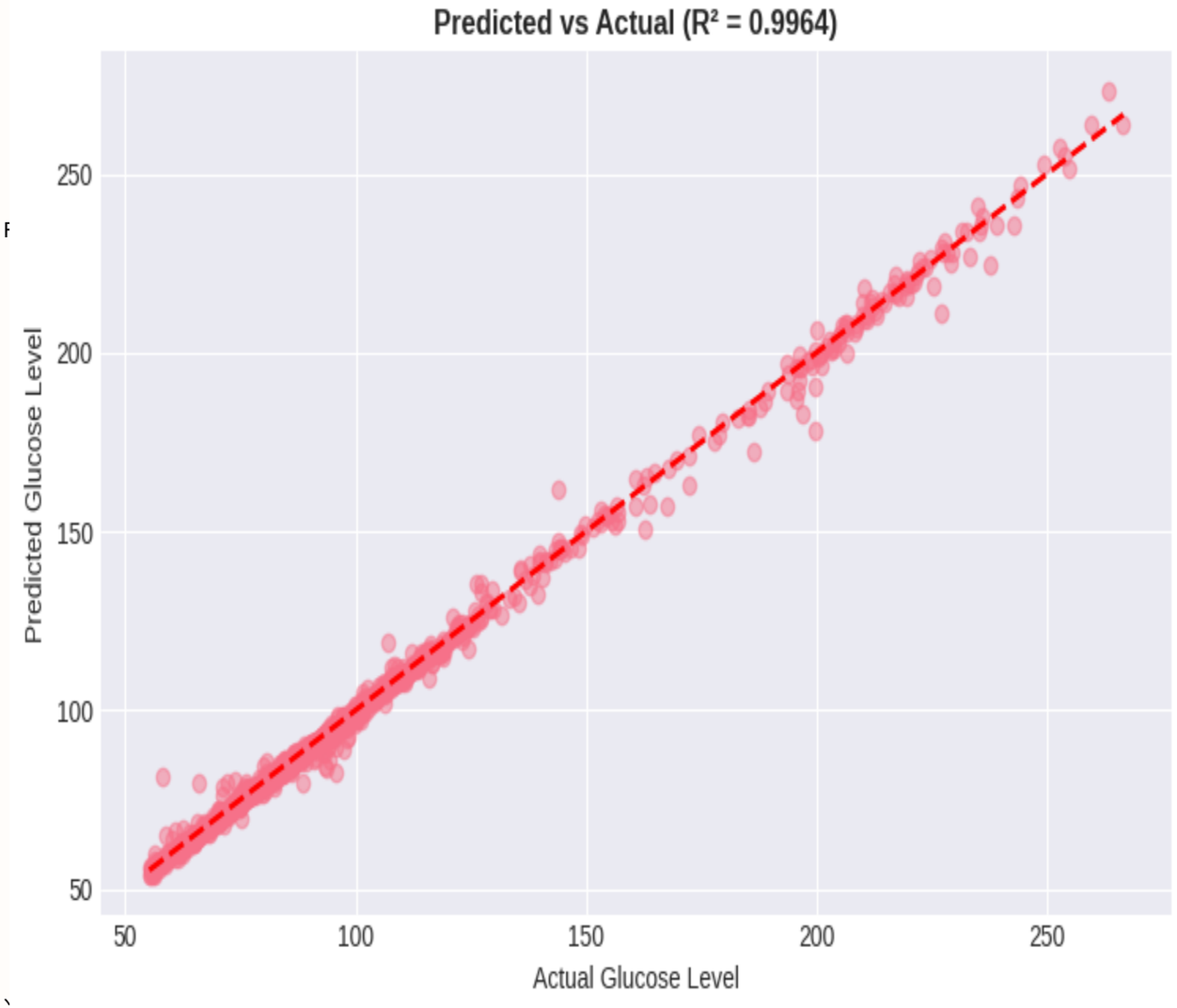
Tools

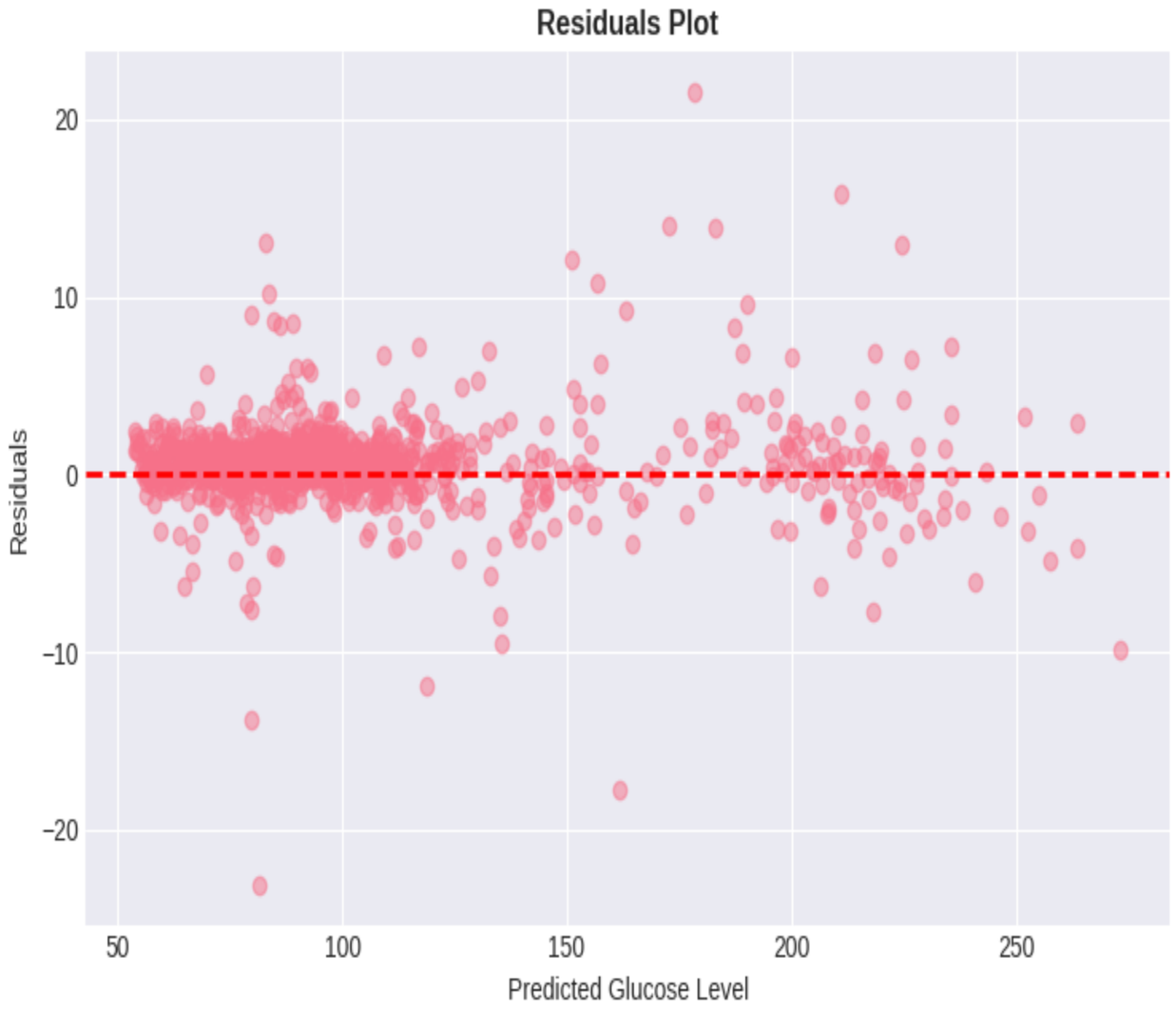
Saved the final model using joblib or pickle for integration into web apps like Streamlit or Flask.(On progress)


Confusion Matrix
939
33
44
6
0
1
Actual
Predicted
800
600
400
200


ROC Curve
ROC Curve (AUC = 0.7736)
Random Classifier
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Regression Model - Test Set Results:
RMSE: 2.6226
MAE: 1.5287
R^2 Score: 0.9964





Monitoring & Maintenance

Data Drift Monitoring

In a real-world setting, the model would be monitored for "Data Drift" (changes in patient health trends over time).

Continuous Feedback Loop

Continuous data collection would allow for future retraining to improve recall further.

Thank You

Questions & Discussion

Presenter: Natnael Abayneh