# Stroke Risk Machine Learning Technical Report

Description of Process, Decisions, and Results

**Prepared For:** Prospective Employers - Data Scientist Position

**Author:** John Paul Medina

**Date:** May 3, 2025

**Github:** https://github.com/jmedinacs/stroke_risk_ML_addendum

**LinkedIn:** Https://www.linkedin.com/in/jpmedinacs

**Email:** jmedinacs715@gmail.com

# Table of Contents

# Project Summary

This project is an extension of an earlier Stroke Risk Data Analysis study. One of the key recommendations from that work was to apply machine learning to uncover deeper insights through multivariate analysis. Building on the same publicly available clinical dataset, several classification models were developed, compared, and evaluated. The best-performing model was selected and further optimized for recall to maximize the detection of high-risk individuals. To enhance transparency and trust, SHAP and PDP visualizations were used to explain individual predictions and highlight how specific features influenced stroke risk. The overarching goal was to simulate a high-impact clinical decision support tool that prioritizes early detection and prevention.

# Tools & Skills Demonstrated

## Languages & Libraries:

- Python (pandas, matplotlib, scikit-learn, XGBoost, imbalance-learn, SHAP, joblib)
- Google Sheets (Cleaning, EDA, Model, and Modularization logs)
- Git & GitHub for version control.

## Techniques & Tests:

- Feature Engineering and Data Imputation.
- Class Imbalance Handling with SMOTE
- Statistical Feature Selection:
    - Chi-Square Test (categorical vs binary target)
    - Point-Biserial Correlation (continuous vs binary target)
- Machine Learning Algorithms:
    - Logistic Regression
    - Random Forest
    - K-Nearest Neighbors (with normalization)
    - XGBoost (with hyperparameter tuning and threshold analysis)
- Model Evaluation: Confusion Matrix, Precision, Recall, F1, ROC AUC
- Model Explainability: SHAP Summary + Waterfall Plots, PDPs (Partial Dependence Plots)

# Problem Statement

Build a predictive model to estimate stroke risk using patient health and demographics data, enabling early detection and preventive care plan construction.

# Dataset

- Source: Kaggle - Stroke Prediction Dataset (fedosriano)
- Rows: 5,110 patients
- Target: stroke (0 = no, 1 = yes)
- Stroke class ratio 95:5 (highly imbalance no:yes)

# Dataset Preparation

- Identified and imputed missing BMI values using the median (avoid data skew effect on mean)
- Removed rare gender category (Other, with 1 instance)
- Standardized and trimmed all categorical text fields
- Applied one-hot encoding (drop_first = True) to nominal features
- Applied SMOTE to training data set due to severe stroke count imbalance (~5% of data)
- Statistical tests identified id, gender, and Residence_type as insignificant and were dropped from the model feature set.

# Model Training

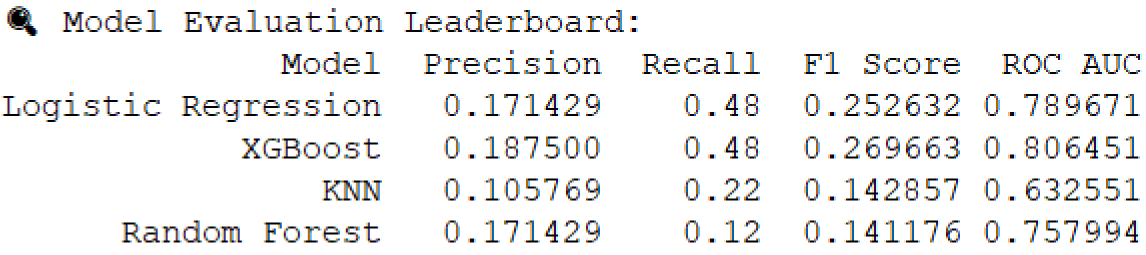
**Trained and evaluated four models:**

- Logistic Regression (baseline)
- Random Forest
- K-Nearest Neighbors (with normalization)
- XGBoost (baseline for comparison and then tuned)

**All models were trained using an 80/20 stratified split and evaluated on:**

- Recall (stroke = 1)
- Precision (stroke = 1)
- F1 Score
- ROC AUC
- Confusion Matrix

**Base Model Performance Leaderboard**



```
🔍 Model Evaluation Leaderboard:
                 Model  Precision  Recall  F1 Score   ROC AUC
   Logistic Regression   0.171429    0.48  0.252632  0.789671
               XGBoost   0.187500    0.48  0.269663  0.806451
                   KNN   0.105769    0.22  0.142857  0.632551
         Random Forest   0.171429    0.12  0.141176  0.757994
```

- XGBoost and Logistic Regression (LR) had the highest recall of 48%

- XGBoost outperformed LR in F1, precision, and ROC AUC
- Random Forest and KNN underperformed in roughly all categories.

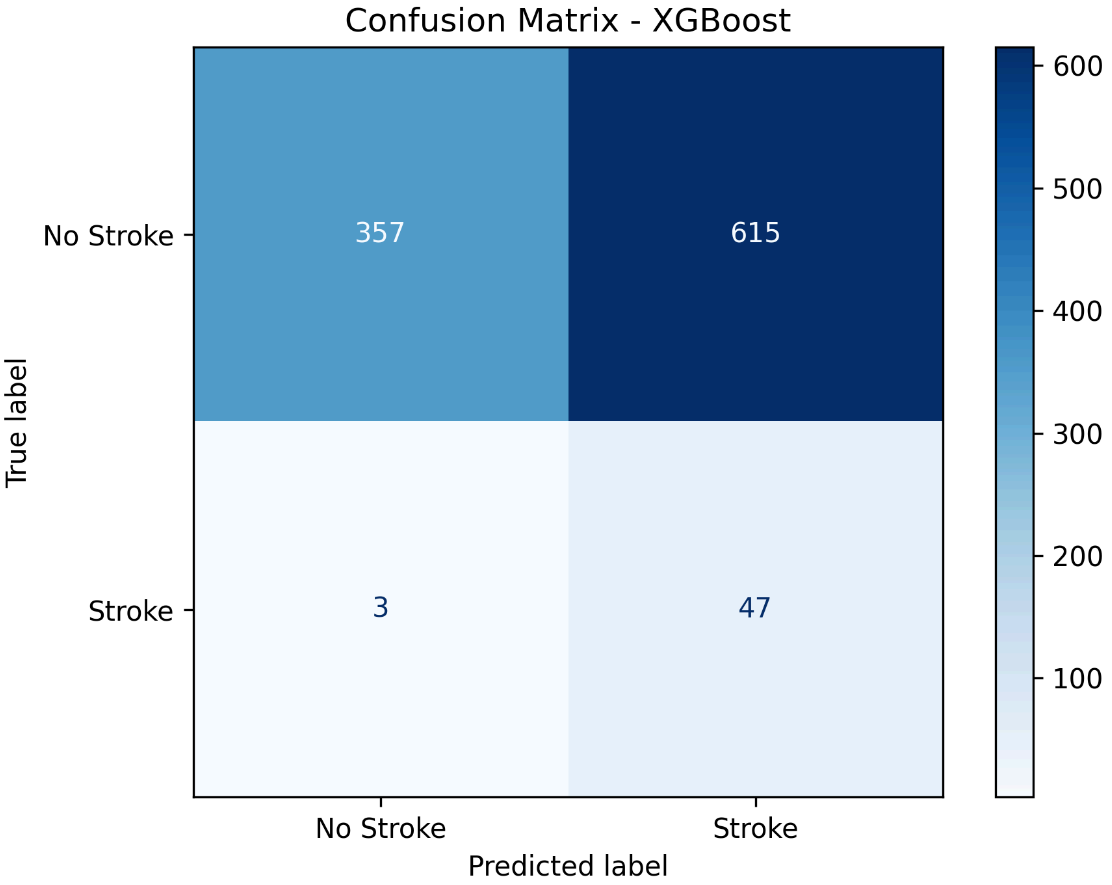
# Final Champion Model: Tuned XGBoost

This version was optimized for maximum recall (catching most stroke cases):

- Recall: 94%
- Precision: 8%
- F1 Score: 14.8%
- ROC AUC: 80.7%

Detected 47 out of 50 stroke cases in the test set.

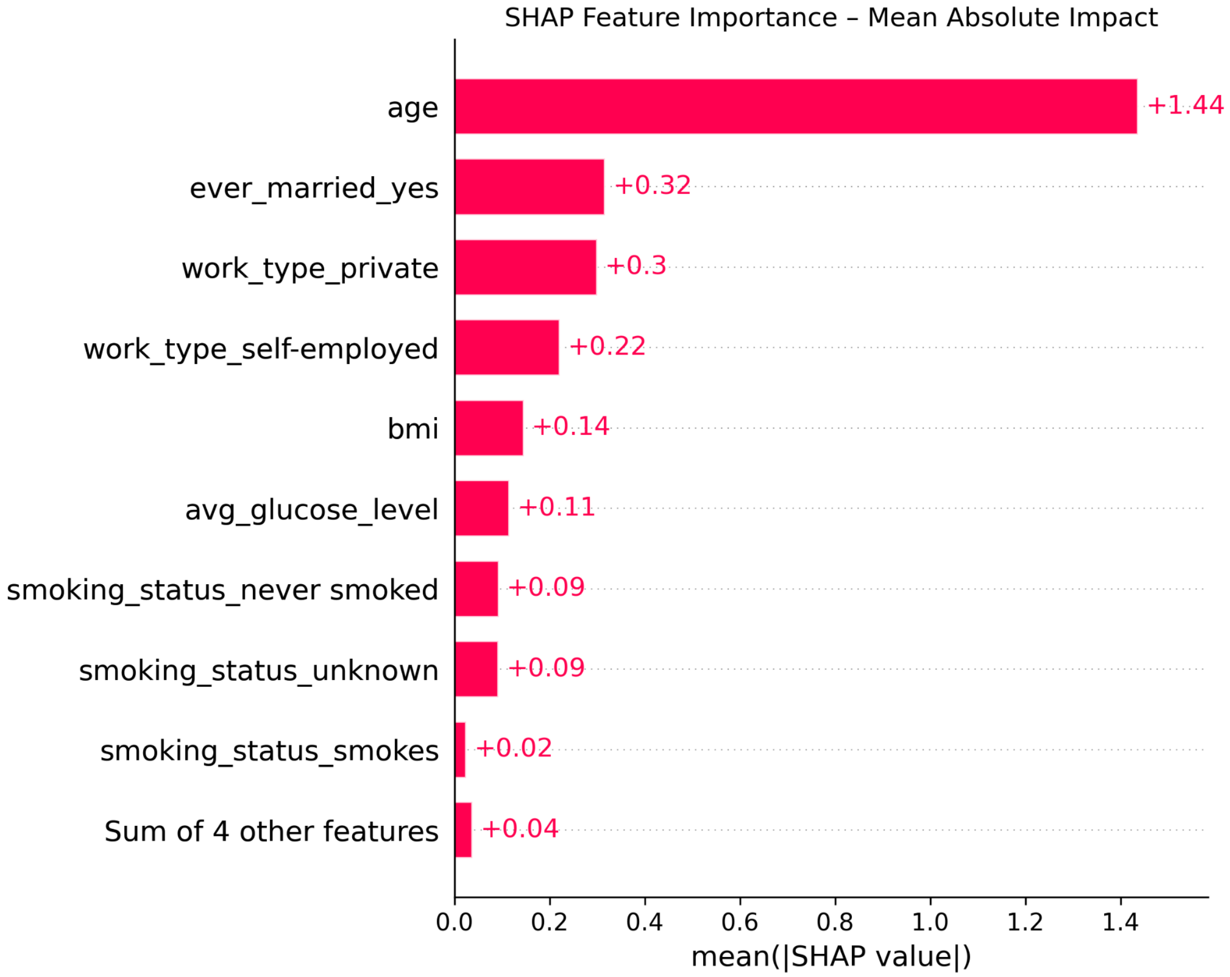
False positives were expected due to the severe imbalance in the dataset (~5%, test set not balanced with SMOTE for realism).  The model yields a high level of false positives; however, with a recall of 94%, it was deemed acceptable as misidentifying an actual stroke is far more costly than a false positive.



# Model Interpretability

## SHAP Summary Plot

Shows the global impact of each feature in the stroke prediction

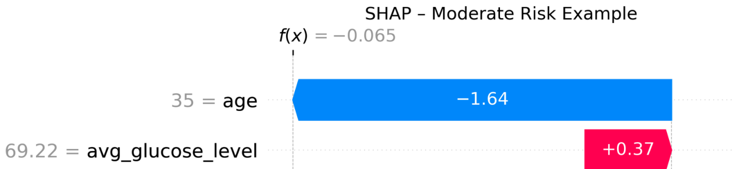


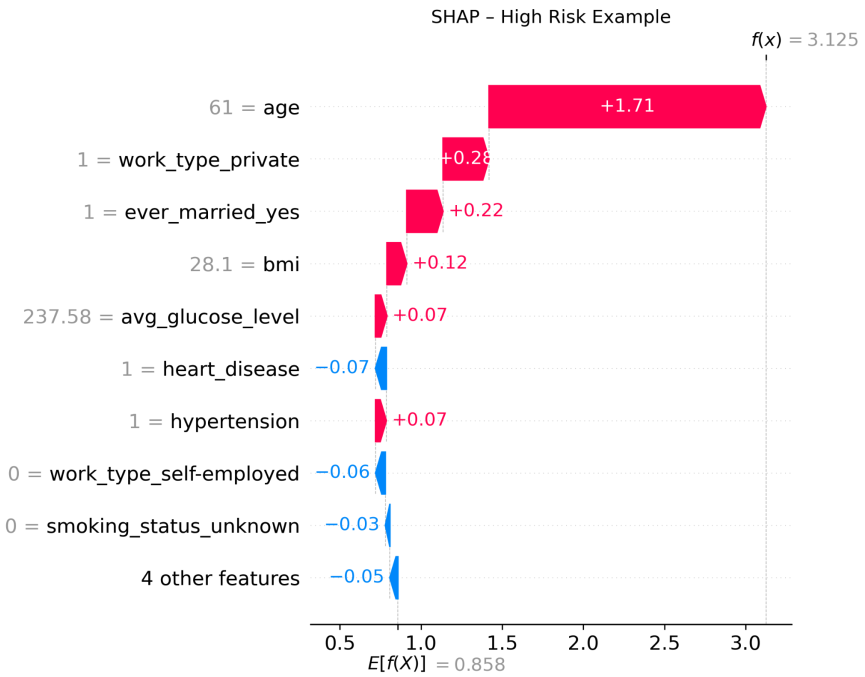
**Top Features:**

- Age (dominant risk predictor)
- Ever Married (yes)
- Work Type: Private

## SHAP Waterfall Plots

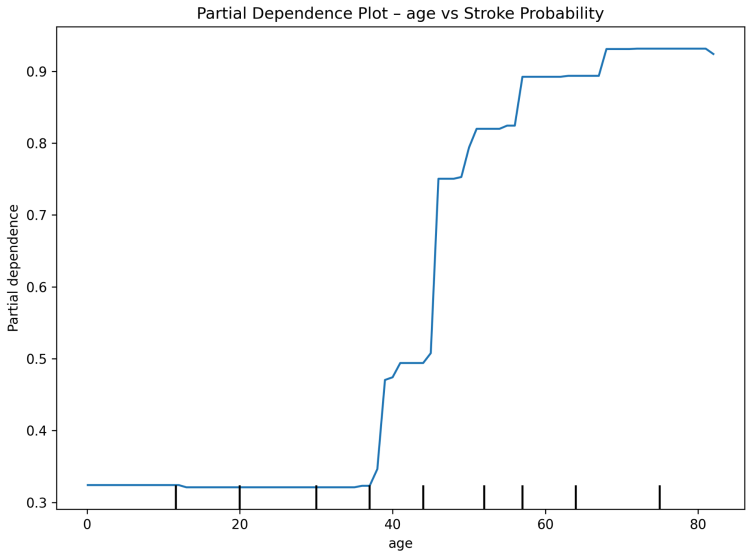
Visual samples and explanation of individual predictions. Each bar represents the feature's contribution (+ or -) to the patient's stroke risk.

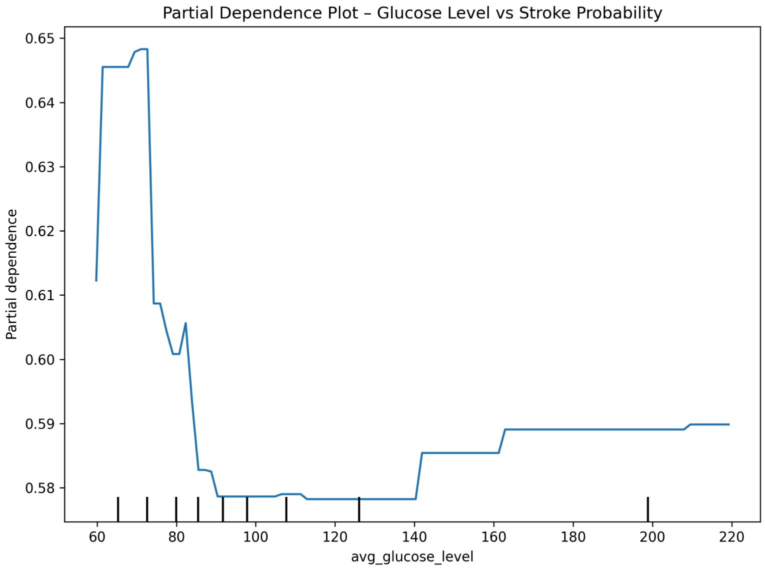




# Partial Dependence Plot (PDP)

Reveals the relationship between continuous features and predicted stroke probability.





The PDP shows the behavior of each feature and how it contributes to the increase or decrease of stroke risk as the continuous feature increases in value.

**Findings:**

- Stroke risk increases sharply around age 45+
- Glucose: elevated risk below 80, flattens after 100
- BMI: increased risk above 23, plateau around 30–40

# Conclusion

This addendum presents findings that may initially appear to contradict those of the original stroke risk data analysis. However, these differences stem from the distinct contexts in which each analysis was conducted. The earlier data analysis emphasized elderly individuals, heart disease, and hypertension as the leading stroke risk factors—findings drawn from bivariate comparisons between individual features and stroke occurrence. In contrast, the machine learning models developed here evaluate all features simultaneously, uncovering age, work type (private), and marital status as top predictors.

These results are not conflicting but complementary. The original analysis spotlighted the high stroke risk within the elderly population—a critical observation—but did not explicitly account for how lower age can significantly reduce risk. Machine learning's multivariate approach, by contrast, captures both high-risk and protective effects across the full range of patient data.

Ultimately, the strength of machine learning lies in its ability to analyze complex, interacting variables at scale—surfacing patterns that are impractical to identify manually. Still, no model is better than the data it learns from. Whether through traditional EDA or automated algorithms, data quality, accuracy, and imbalance remain pivotal in shaping the insights and conclusions we derive.

# Recommendations & Future Work

- Consider a two-model ensemble: use a high-recall model (XGBoost) followed by a precision-tuned classifier to reduce false positives.
- Explore undersampling or cost-sensitive learning
- Expand with additional features (e.g., medication use, family history)
- Deploy model via web app or dashboard for clinical simulation

# Author & Credits

Prepared by **John Medina**, aspiring data scientist with 16 years of experience teaching mathematics and computer science.
GitHub: github.com/jmedinacs
LinkedIn: linkedin.com/in/jpmedinacs

Data by: Fedsoriano, Stroke Prediction Dataset. Kaggle
From: Kaggle.
Date: 2021
Available at: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
License: Open Data Commons Public Domain Dedication and License (PDDL)