Stroke Risk Prediction Tool: A Machine Learning Approach

Author: John Medina Date: 5/17/2025

Why This Tool Matters

According to the Centers for Disease Control and Prevention (CDC), in the United States, someone experiences a stroke **every 40 seconds**, and someone dies from a stroke **every 3 minutes and 11 seconds**. Understanding the risk factors that contribute to stroke is critical for prevention and early intervention.

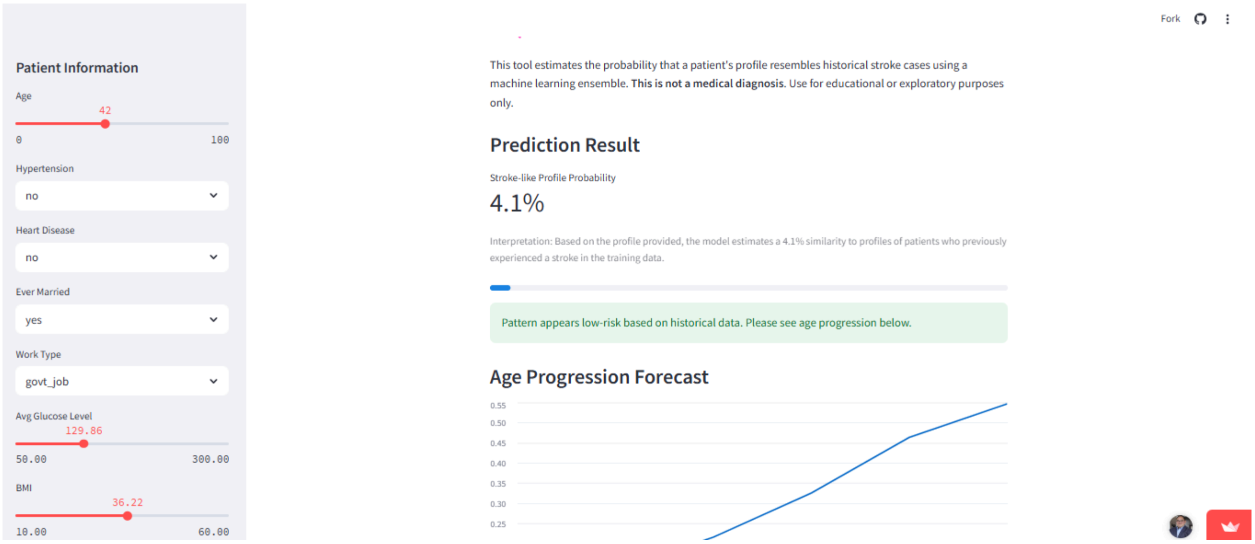
Why this Project

This project developed a machine learning-powered risk estimation tool that allows users to enter basic patient information and receive a stroke-like profile probability using a deployed Streamlit app. While not intended for diagnosis, this tool supports **early identification** and encourages further screening or consultation when patterns align with high-risk historical profiles.

The Data

The dataset used in this project was taken from Kaggle, which was uploaded by Fedesoriano. It consists of **5110 de-identified patient records**, with features including age, BMI, average glucose level, heart disease, work type, and other demographic and health indicators. The model was trained to detect stroke-like patterns based on these features using a soft-voting ensemble of logistic regression and random forest classifiers.

The Streamlit App



Patient Information

Age: 42 (0–100)

Hypertension: no

Heart Disease: no

Ever Married: yes

Work Type: govt_job

Avg Glucose Level: 129.86 (50.00–300.00)

BMI: 36.22 (10.00–60.00)

This tool estimates the probability that a patient's profile resembles historical stroke cases using a machine learning ensemble. **This is not a medical diagnosis.** Use for educational or exploratory purposes only.

Prediction Result

Stroke-like Profile Probability

4.1%

Interpretation: Based on the profile provided, the model estimates a 4.1% similarity to profiles of patients who previously experienced a stroke in the training data.

Pattern appears low-risk based on historical data. Please see age progression below.

Age Progression Forecast

The end-product app is deployed in **Streamlit Cloud** and accessible to users (here: aistrokerisktool.streamlit.app). After conducting statistical analysis, features that did not significantly contribute to stroke prediction were removed to simplify the model and improve interpretability.

The app allows the user to enter their **health profile**, including age, heart disease, hypertension, marriage history, work type, average glucose level, BMI, and smoking status.

GitHub Repository: github.com/jpmedinacs/ai_stroke_risk_tool

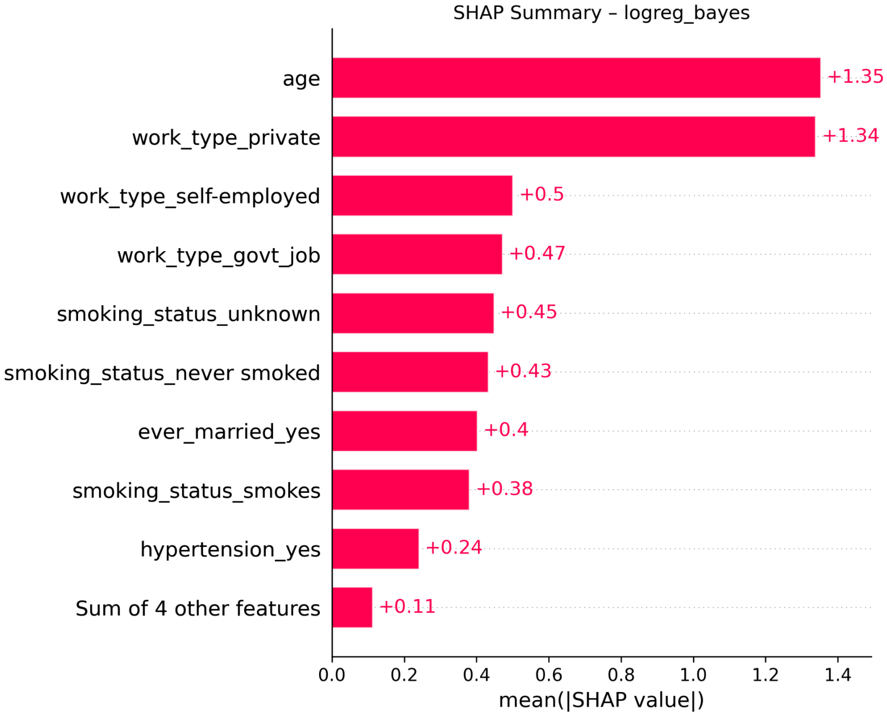
Based on the user's information, the model returns a **stroke-like profile probability** indicating how closely the input resembles historical cases of stroke. It also displays a line graph showing how risk is projected to change with age, assuming all other health factors remain constant.

The Predictive Models

Four machine learning models — Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), and XGBoost were selected, trained, and evaluated. Each model brings unique strengths and limitations to the task. To optimize performance, all models were tuned using **BayesSearchCV**, which automatically explored different hyperparameter combinations to find the best-performing configuration for each algorithm.

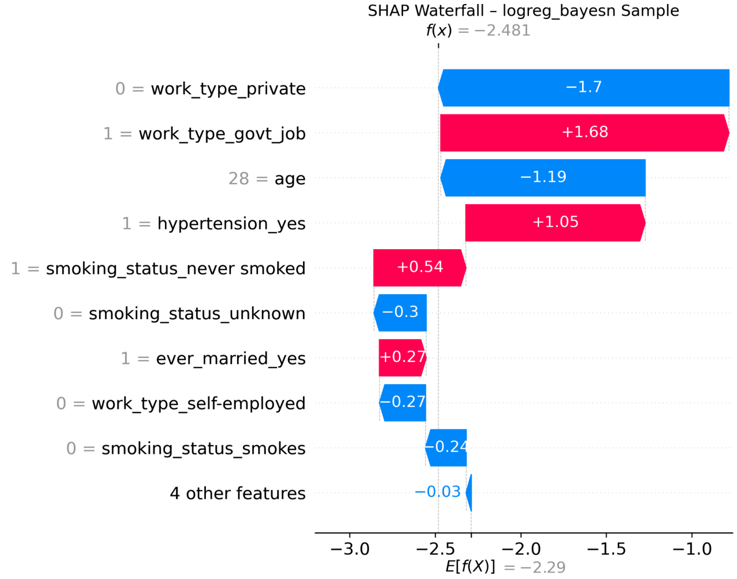
Ultimately, two models were deemed viable and were combined into a **soft-voting ensemble** to produce the final predictive model.

Logistic Regression Model



The Logistic Regression (LR) model identified both **age** and **employment type** (particularly private and self-employed work) as the strongest contributors to stroke risk predictions. While **age** is a well-established medical risk factor, the model's emphasis on **work type** suggests potential underlying patterns worth further exploration. Among all individual models, LR achieved the **highest recall at 68%**, making it the most effective at identifying stroke cases in the dataset.

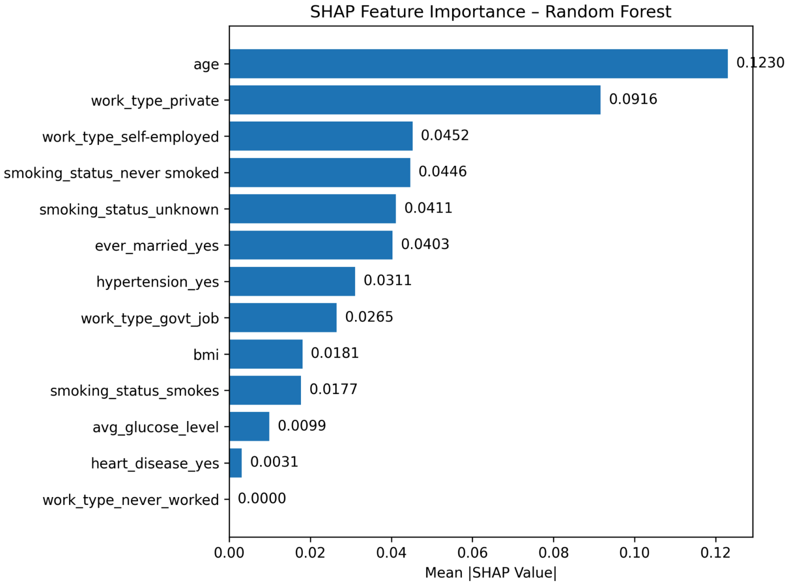
A Peek Under the Hood



The image above shows a **SHAP waterfall plot** from the Logistic Regression model, breaking down how it arrives at a patient's **stroke-like risk score**. The model assigns a score to each of the patient’s **health factors,** such as age, work type, and smoking status, based on how much they increase or decrease stroke risk and then **adds them together** to get the final prediction.

In this example, work type strongly lowers the predicted risk (blue bars), while age, hypertension, and smoking history increase it (pink bars). Even though one factor may seem dominant, the model balances all features, showing how multiple risk contributors interact to influence the outcome.

Random Forest Model



The Random Forest (RF) model also identified **age** as the most influential predictor of stroke, followed by **private-sector employment**, **self-employment**, and **marriage history**. The slight variation in feature rankings and contribution intensity, compared to other models, is expected, as different algorithms evaluate patterns in unique ways. Random Forest, for instance, relies on decision trees that split the data based on the most informative variables at each step, which can result in a different prioritization of features even when the overall trends remain consistent.

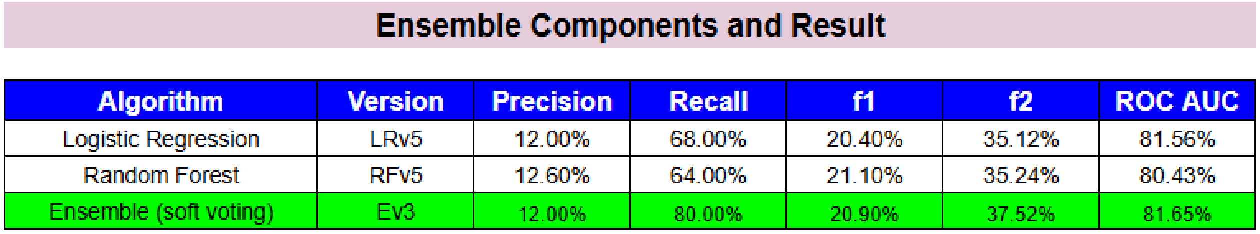
XGBoost Model

The XGBoost model identified **age** as the most influential contributing feature, with SHAP values showing it contributed over **30% more** than the next strongest factor, **work type**. After tuning, the best-performing XGBoost model required a very low classification threshold of **0.09** to maximize recall. In that setting, it correctly identified **46%** of true stroke cases and achieved the **highest precision (13.5%)** among all individual models. However, its lower sensitivity made it less suitable for inclusion in the final ensemble.

K-Nearest-Neighbor

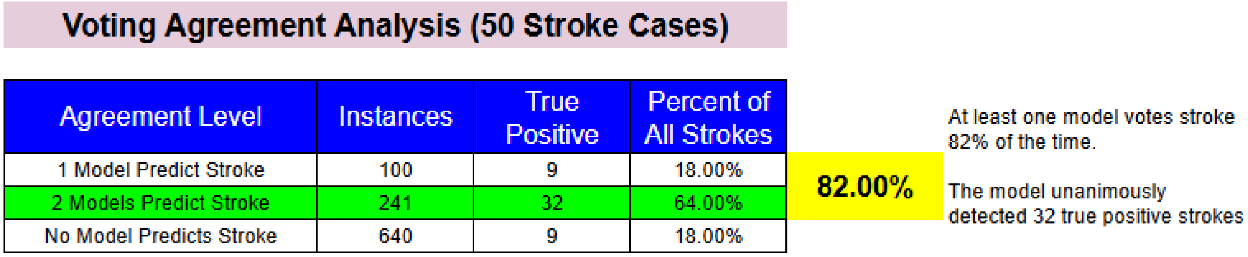
The K-Nearest Neighbors (KNN) model was excluded from the final ensemble due to its comparatively **lower performance and limited interpretability**. While KNN can be effective in well-structured problems, it struggled to generalize in this dataset and offered minimal insight into feature influence. These limitations made it a poor fit for a tool that prioritizes both prediction accuracy and explainability.

The Voting Ensemble



Ensemble Components and Result

Algorithm	Version	Precision	Recall	f1	f2	ROC AUC
Logistic Regression	LRv5	12.00%	68.00%	20.40%	35.12%	81.56%
Random Forest	RFv5	12.60%	64.00%	21.10%	35.24%	80.43%
Ensemble (soft voting)	Ev3	12.00%	80.00%	20.90%	37.52%	81.65%



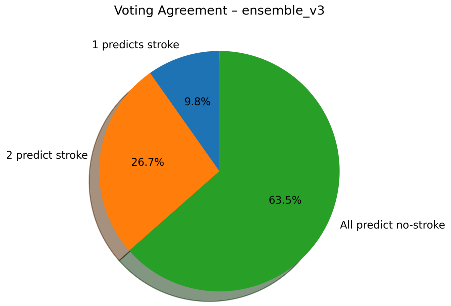
Voting Agreement Analysis (50 Stroke Cases)

Agreement Level	Instances	True Positive	Percent of All Strokes
1 Model Predict Stroke	100	9	18.00%
2 Models Predict Stroke	241	32	64.00%
No Model Predicts Stroke	640	9	18.00%

82.00%

At least one model votes stroke 82% of the time.

The model unanimously detected 32 true positive strokes

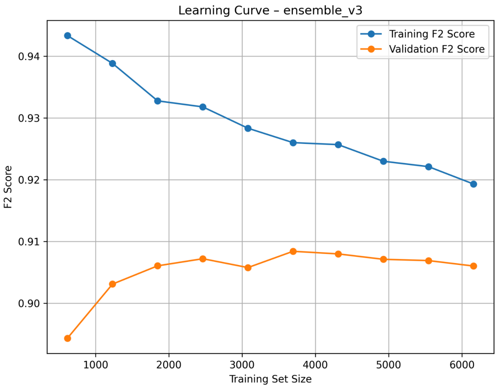


The final ensemble model (Ev3) outperformed both Logistic Regression and Random Forest in terms of **recall (80%)** and **F2-score (37.52%)**, while maintaining similar precision and a strong ROC AUC (81.65%). The voting agreement analysis further supports the model's effectiveness: **82% of stroke cases were flagged by at least one model**, and **64% were detected unanimously by both**. This confirms the ensemble's value as a high-recall screening tool, even if some predictions are cautious in nature.

The Voting Ensemble Interpretation

These results show that the ensemble model successfully identifies **8 out of 10 actual stroke cases** and correctly identifies a true stroke case in roughly **1 out of every 8** high-risk predictions. This version was selected based on the project's goal of **maximizing recall** and prioritizing the detection of true stroke risks while maintaining a reasonable level of false positives. In this context, **missing an actual stroke poses greater risk** than recommending a patient for further noninvasive evaluation.

The Ensemble's Learning Curve



F2 score is a way to measure how well a model catches positive cases (having a stroke), while still trying not to make too many false alarms. It puts **extra weight on recall**, which is essential when **missing a real case is worse than raising a false one**.

In the chart above, we compare how the model performs on training data (blue line) vs. new, unseen data (orange line) as we give it more examples. The gap between the two lines shows that the model is learning well: it starts very confident on small datasets, then becomes more balanced and realistic as more data is added.

The orange line, representing how well it does on new data, **gets better up to around 4,000 records** and then levels off. That means the model is no longer significantly improving with more data and has likely learned most of what it can from the current features. At this stage, the model is stable and consistent, but if more data is introduced in the future, **recalibration or additional feature engineering** would be needed to unlock further improvements.

Conclusion and Next Steps

This project developed a recall-optimized stroke risk tool using a soft-voting ensemble that balances sensitivity and precision. The final model identifies **80% of stroke cases** while maintaining a manageable false positive rate, making it well-suited for early screening and outreach.

Because it relies only on basic health inputs, the tool is easy to deploy in clinical forms or digital health platforms for **low-friction, noninvasive risk assessment**. Future work can focus on expanding the feature set, integrating clinical timelines, and adding interpretability features to support real-world decision-making.

This project demonstrates how machine learning can support preventative care by translating historical data into **accessible, action-oriented insights**.

The dataset is openly available and intended for educational and research use (Fedesoriano).