Stroke Risk Prediction using Machine Learning and Logic-Based Safety Layers

Technical Report – Mini Research Paper

Author: Karan Shah
Date: November 2025

1. Introduction

Stroke is one of the leading causes of death and long-term disability worldwide, yet many strokes are potentially preventable through earlier risk awareness and lifestyle or clinical interventions. However, individual stroke risk is influenced by a combination of demographic, clinical, and lifestyle factors, and these patterns can be difficult to interpret without data-driven tools.

This project explores how a **machine learning model**, combined with **medical domain logic and interpretability tools**, can be used to provide **stroke risk awareness** for the general public. The goal is not to provide a clinical diagnosis, but to:

- Estimate a **personalised probability of stroke** based on common health attributes
- Highlight which features contribute most to the prediction
- Demonstrate **responsible AI design** through safety overrides and transparency

The model is deployed as a **FastAPI backend** with a **Streamlit web interface**, making predictions interactively and visualising outputs in a user-friendly way.

2. Dataset

The model is trained on the public **Kaggle Stroke Prediction Dataset**, which contains anonymised records of individuals with the following attributes (among others):

- Age
- Gender
- Residence type (Urban/Rural)
- Ever married (Yes/No)

- Work type
- Smoking status
- Average glucose level
- Body Mass Index (BMI)
- Hypertension (0/1)
- Heart disease (0/1)
- Target variable: stroke (0 = no stroke, 1 = stroke)

2.1 Data Challenges

Working with this dataset presents several realistic machine learning challenges:

- **Class imbalance:** Stroke events (label = 1) are relatively rare compared to non-stroke cases (label = 0). A naive model that predicts “no stroke” for everyone can have high accuracy but be useless for detecting actual risk.
- **Missing values:** Some entries, especially BMI and smoking status, are missing or marked as “Unknown”.
- **Noisy or non-clinical values:** Public data may contain unrealistic combinations or extreme values (e.g., very high BMI or glucose) that do not correspond to realistic real-world recordings.
- **Limited feature set:** Important clinical variables like cholesterol, blood pressure trends, or family history are not included, limiting the model’s potential performance in a clinical setting.

Because of these challenges, the project focuses not only on building a classifier, but also on **feature engineering, class imbalance handling, and safety-conscious post-processing**.

---

3. Methodology

The model is implemented as a **pipeline**, combining preprocessing, feature engineering, class balancing during training, and a final classifier.

3.1 Preprocessing

Key preprocessing steps include:

- **Imputation** of missing numerical values (e.g., BMI)
- **One-hot encoding** of categorical variables (gender, residence type, smoking status, work type, etc.)
- **Scaling** of numerical features (such as age, glucose, BMI) to stabilise the training process

These steps are implemented using `scikit-learn` transformers and combined into a pipeline so that training and inference share identical transformations.

3.2 Feature Engineering

To better capture medical relationships, several engineered features are created from the raw inputs, for example:

- **BMI category** (e.g., underweight, normal, overweight, obese)
- **Age groups** and a **senior flag**
- **High BMI / high glucose flags**
- **Ratios and interactions**, such as:
  - BMI ÷ Age
  - Glucose ÷ BMI
  - Age × BMI
  - Age × smoking status
- Binary flags like `smoker_flag`, `cardio_flag`, etc.

These engineered variables help the model learn non-linear relationships that are more aligned with medical intuition, such as the compounding effect of high BMI and high glucose.

3.3 Handling Class Imbalance

Stroke cases are rare relative to non-stroke cases. To address this, **SMOTE (Synthetic Minority Oversampling Technique)** is applied during training (but not at inference time). SMOTE synthesises new minority-class examples to give the model more signal about positive stroke cases and prevent it from simply learning “always predict zero”.

3.4 Model Training (XGBoost)

The final classifier is an **XGBoost** model, chosen for its:

- Strong performance on tabular data
- Ability to handle non-linear relationships
- Support for class weighting and calibration

The training objective is tuned towards **rare-event performance** rather than just overall accuracy. In particular, hyperparameters and thresholds are evaluated using **Precision–Recall AUC (PR-AUC)** and **F1 score** to better reflect behaviour on stroke cases.

---

4. Model Performance

The model was evaluated on a validation set separated from the training data. During early development, the model achieved approximately **80–90% overall accuracy** on hold-out testing, showing strong signal despite the limitations of a public dataset.

On a more detailed validation with the tuned threshold around **0.34 (≈30/100 as shown in the app UI)**, the following metrics were observed:

- **Precision:** ~57.4%
- **Recall:** ~51.4%
- **F1 score:** ~54.2%
- **PR-AUC:** ~0.54
- **Overall accuracy:** ~95.8% (driven by the large number of non-stroke cases)

These values come from the confusion matrix:

- **True Negatives (TN):** 4,766
- **False Positives (FP):** 95
- **False Negatives (FN):** 121
- **True Positives (TP):** 128

This means:

- The model correctly identifies most non-stroke cases (high TN, low FP).

- Around half of the true stroke-risk cases are detected (recall ~51%), which is reasonable for a rare-event problem on public data.

- More than half of the high-risk warnings are correct (precision ~57%), indicating the model is not “shouting stroke” too often.

The overall accuracy (~95.8%) is high, but in the context of imbalanced data, **accuracy alone is not the main success metric**. PR-AUC, precision, and recall provide a clearer picture of how the model behaves on the minority stroke class.

> **Interpretation:** The model is **conservative but meaningful** in flagging risk — it does not raise too many false alarms, and when it does flag high risk, it is often correct.

4.1 Precision–Recall Behaviour

InsertFigure1:Precision–RecallcurvehereifyouhaveitInsert Figure 1: Precision–Recall curve here if you have itInsertFigure1:Precision–Recallcurvehereifyouhaveit

The Precision–Recall curve illustrates how different thresholds shift the balance between catching more stroke cases (higher recall) and avoiding unnecessary warnings (higher precision). Threshold ≈0.34 is chosen as a compromise between the two, suitable for an awareness-focused tool.

4.2 Calibration

InsertFigure2:CalibrationcurvehereifyouhaveitInsert Figure 2: Calibration curve here if you have itInsertFigure2:Calibrationcurvehereifyouhaveit

The calibration curve shows how well predicted probabilities match actual observed frequencies. Reasonable calibration supports the interpretation of outputs as “out of 100 people like you, around X might experience a stroke,” which is used in the app’s UI text.

5. Explainability (SHAP)

To avoid a “black box” model, the project uses **SHAP (SHapley Additive exPlanations)** to understand which features drive predictions.

5.1 Global Feature Importance

InsertFigure3:SHAPsummaryplothere(shapsummary.png)Insert Figure 3: SHAP summary plot here (shap_summary.png)InsertFigure3:SHAPsummaryplothere(shapsummary.png)

The SHAP summary plot shows which features most strongly influence stroke risk across the entire dataset. Consistent with medical expectations, features like:

- **Age**
- **BMI**
- **Average glucose level**
- **Smoking status**
- Relevant interaction features (e.g., glucose × BMI)

appear as dominant contributors. High age, high BMI, high glucose, and active smoking are associated with higher SHAP values (pushing risk upwards).

5.2 Individual Prediction Explanations

InsertFigure4:SHAPforceplothere(shapforceexample0.png)Insert Figure 4: SHAP force plot here (shap_force_example_0.png)InsertFigure4:SHAPforceplothere(shapforceexample0.png)

For individual users, SHAP force plots show how each feature pushes the predicted risk upward or downward from the baseline. For example:

- Age and high BMI might push the prediction strongly toward higher risk.
- “Never smoked” or “normal BMI” may push in the opposite direction, reducing risk.

This makes the model’s reasoning more transparent and educational, allowing users and reviewers to see that the model is relying on medically meaningful patterns rather than arbitrary correlations.

---

6. App Deployment

The model is deployed as a functioning web application with:

- A **FastAPI** backend:
    - Exposes a `/predict` endpoint
    - Receives JSON input containing user features
    - Applies model and logic rules
    - Returns probability + risk category + threshold

- A **Streamlit** frontend:
    - Collects inputs: age, gender, residence, smoking status, marital status, hypertension, heart disease, height, weight (to compute BMI), and average glucose
    - Displays:
        - Estimated stroke risk (as a percentage)
        - Risk category (e.g., Low / High)
        - A short natural-language interpretation
        - In some versions, model performance and SHAP visuals
    - Tracks prediction latency (with a target of **<500 ms** per prediction)

The frontend is designed to be clean and understandable, with text emphasising that the result is **educational and not diagnostic**.

---

7. Safety and Ethics

Because this project operates in the health domain, safety and ethics are central to the design.

7.1 Not a Medical Device

The application displays clear messaging:

- It is **not a diagnostic tool**
- It does **not replace professional medical advice**
- It should be used for **awareness and education** only

Users are encouraged to speak to a licensed doctor, especially if they receive a high-risk output.

7.2 Logic-Based Safety Overrides

In addition to the statistical model, several **logic-based rules** are applied after prediction to enforce medical sanity:

- **Smoking status:** If the model appears to underweight smoking, additional risk may be added for active smokers.
- **Extreme BMI:** Extremely high or low BMI values trigger adjusted risk levels.
- **Very high glucose:** Very high average glucose values (e.g., >300 mg/dL) raise risk.
- **Unknown values:** “Unknown” inputs (e.g., unknown smoking status) can add a small conservative penalty.

These overrides are applied **without retraining the model**, allowing the system to correct counterintuitive or unsafe outputs from the learned model based on domain common sense.

7.3 Data Limitations and Bias

The report explicitly acknowledges that:

- The dataset is not a substitute for real clinical data.
- Certain demographic or medical groups may be underrepresented.
- The model’s performance is constrained by the available features and labels.

Therefore, results must be treated as **approximate and educational**, not clinical evidence.

---

8. Conclusion and Future Work

This project demonstrates a full-stack approach to stroke risk prediction:

- **Data:** Public Kaggle dataset with real-world challenges.
- **Model:** XGBoost classifier with SMOTE balancing and engineered features.
- **Metrics:** Rare-event-focused evaluation using precision, recall, F1, PR-AUC, and confusion matrix.
- **Explainability:** SHAP-based insights at both global and individual levels.
- **Deployment:** FastAPI + Streamlit application, designed for accessible public use.
- **Safety:** Logic-based overrides and clear disclaimers to avoid misuse.

While early development achieved **~80–90% accuracy** on hold-out validation and extended testing shows consistent, cautious high-risk detection, this model is **not** a replacement for professional medical judgement or real clinical risk scoring systems.

8.1 Limitations

- Limited feature set compared to clinical practice (e.g., no cholesterol, blood pressure history, or detailed family history).
- Dependence on a single public dataset.
- Potential dataset bias and unmeasured confounders.

8.2 Proposed Improvements

Future work could include:

- Training or fine-tuning on **real clinical datasets** (with appropriate approvals and ethics).
- Further optimisation to **increase recall** on stroke cases while keeping false positives at a manageable level.
- Improved calibration across different subgroups.
- Enhanced UI explanations, including personalised preventive suggestions co-designed with medical professionals.

Overall, this project is best seen as a demonstration of how a high-school student can build a responsible, interpretable, deployed machine learning system in the health domain, while recognising and communicating its limitations clearly.