Stroke Prediction Interface for Doctor

Team 5

1. Project Goals

The primary goal of our project is to develop an interactive web application that aids physicians in predicting the likelihood of a patient experiencing a stroke. By inputting patient data, such as age, gender, medical history, and lifestyle factors, the application will provide a prediction based on a pre-trained machine learning (ML) model. Additionally, the application will offer insights into the data and the model's performance, along with explanatory plots to help physicians understand the underlying factors contributing to the prediction.

2. Intended Target Audience

Our target audience for this application is doctors working in hospitals. These professionals are often pressed for time and need quick, accurate tools to aid clinical decision-making. The application aims to support them by providing a reliable stroke risk assessment tool that is easy to use and integrates into their workflow. The predictions and explanations provided by the application can help physicians make informed decisions about patient care and preventive measures. It will enable physicians to offer targeted advice to patients who are currently not experiencing a stroke but are at potential risk. This guidance can help patients take proactive steps to reduce their stroke risk, leading to better long-term health outcomes.

3. Dataset and Documentation

The dataset used for training our model is sourced from Kaggle, specifically the Stroke Prediction Dataset (https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data). This dataset includes various attributes that are relevant to predicting stroke risk. The attributes are as follows:

No.	Feature	Description
1	id	Unique identifier for each patient
2	gender	The gender of the patient, categorized as "Male," "Female," or "Other."
3	age	The age of the patient.

4	hypertension	Indicates whether the patient has hypertension (1) or not (0)
5	heart disease	Indicates whether the patient has a heart disease (1) or not (0)
6	ever married	Indicates if the patient has ever been married ("Yes" or "No")
7	work type	The type of work the patient is engaged in, categorized as "children," "Govt_job," "Never_worked," "Private," or "Self-employed."
8	Residence type	The patient's residence type, either "Rural" or "Urban."
9	avg glucose level	The average glucose level in the patient's blood
10	bmi	The patient's body mass index
11	smoking status	The patient's smoking status, categorized as "formerly smoked," "never smoked," "smokes," or "Unknown."
12	stroke	The target variable indicating if the patient had a stroke (1) or not (0).

This dataset provides a comprehensive set of features pertinent to assessing stroke risk, covering demographic, medical, and lifestyle factors.

4. Key Design Decisions

Several key design decisions were made during the development of this application:

4.1. User Interface (UI) Design

The UI was designed to be intuitive and straightforward using Streamlit, allowing doctors to input patient data easily. We prioritized a clean layout with clear labels and input fields to minimize the learning curve and reduce the risk of errors. The design ensures that doctors can quickly enter the required information and obtain predictions without unnecessary complexity.

4.2. Model Selection

After experimenting with various machine learning models, we selected logistic regression classifier for its accuracy and robustness. The model was trained and validated using the provided dataset, ensuring it performs well on unseen data.

4.3. Prediction and Interaction with Doctor

Upon entering patient data, the application provides a prediction of whether the patient is at risk of having a stroke. Additionally, the application explains which features significantly influenced the prediction. For example, if a patient has hypertension and a high average glucose level, these factors will be highlighted as major contributors to the stroke risk prediction. The application was designed with direct input from physicians to ensure it meets their needs.

4.4. Explainability

To ensure the model's predictions are interpretable, we integrated SHAP (SHapley Additive exPlanations) values. This allows the application to provide visual explanations of the factors contributing to each prediction, helping doctors understand the model's reasoning. For instance, a SHAP plot might show that age and hypertension are the top contributors to a particular patient's stroke risk.

4.5. Performance Metrics

The application displays the model's key performance metrics, such as accuracy, precision, recall, and F1 score. These metrics help users assess the reliability of the model. Detailed confusion matrices and ROC curves are also provided to give a deeper understanding of the model's performance across different thresholds.

5. Problems Faced

During the development process, we encountered several challenges:

5.1. Data Imbalance

The dataset was imbalanced, with a smaller number of stroke cases compared to non-stroke cases. This imbalance could potentially bias the model towards predicting the majority class. To address this, we employed techniques such as SMOTE (Synthetic Minority Over-sampling Technique) to balance the training data.

5.2. Missing Values

Some attributes had missing values. For the attribute **bmi**, which also had missing values, we decided to remove those entries from the dataset to maintain data integrity and quality. The **Smoke_status** attribute contained a significant number of unknown values, which posed a challenge for accurate analysis and modeling. To address this, we consolidated similar categories by replacing 'formerly smoked' with 'smokes'. Additionally, we removed entries with 'Unknown' smoking status to maintain data integrity.

5.3. Standardization of Age, BMI, and Avg_Glucose Level

We had to standardize the attributes age, BMI, and avg_glucose level to ensure they are on a similar scale, which is crucial for the performance of many machine learning algorithms.

5.4. Adjusting Classification Threshold

We needed to set an appropriate threshold for classification. Using a threshold of 0.35 as an example, we converted the probability predictions y_proby_proby_prob to binary outcomes based on this threshold. This adjustment was necessary to balance the precision and recall of our model.

5.5. Model Explainability

Ensuring the model's predictions were interpretable was a significant challenge. We integrated SHAP values to provide explanations, but visualizing and presenting these explanations in a user-friendly manner required careful design and iteration. (explanation or manuals)

5.6. User Testing

Gathering feedback from physicians to refine the application was essential but logistically challenging. Coordinating with busy healthcare professionals and incorporating their feedback into the design required effective communication and project management.

6. Reflections on the Development Process

The development process for this application was both challenging and rewarding. Key reflections include:

6.1. Iterative Development

Adopting an iterative development approach allowed us to refine the application based on user feedback and testing. This approach ensured we could quickly identify and address issues, improving the application's usability and functionality.

6.2. Focus on Explainability

Emphasizing the explainability of the model's predictions was a critical decision. It not only helps build trust with users but also provides valuable insights that can inform clinical decision-making.

6.2. Balancing Complexity and Usability

Striking the right balance between providing comprehensive functionality and maintaining a simple, intuitive interface was challenging. Prioritizing usability without sacrificing essential features required careful consideration and user testing.

6.3. User-Friendly Design

Considering that this application is intended for users without a data science background, it was essential to ensure the interface and explanations were user-friendly. We avoided overly technical language and complex data visualizations to ensure that physicians could easily understand and interpret the predictions and explanations. This focus on simplicity and clarity was critical in making the application accessible and practical for everyday clinical use.

6.4. Continuous Learning

The development process was a continuous learning experience. Staying updated with the latest advancements in machine learning and user experience design was essential to build a robust and user-friendly application.