Personalized Health Plan for Stroke Risk Prediction

Introduction

Strokes are among the leading causes of death worldwide. There are numerous risk factors that vary across individuals and it is vital that individuals get screened to determine their likelihood of a stroke. Early identification of stroke risk can enable timely interventions and save lives, as it would reduce both the incidence of and severity of strokes. Our project aims to create a personalized health plan based on an individual's stroke risk profile which could help detect strokes early on and could even prevent them from having a stroke altogether. An individual's stroke risk profile is derived from a dataset containing key health metrics. By using advanced AI techniques and an intuitive user interface, this system seeks to make stroke risk assessment more accessible and quick for users to access. By integrating a Random Forest classifier with an interface built on Streamlit, we want to provide users with personalized insights into their stroke risk, and offer individualized recommendations to improve their health outcomes.

This report will cover our successes, challenges, and the overall results of our project. It will cover our progress from data preprocessing to model optimization and user interface development, with our objective being the creation of a model that could help predict stroke risk and is accessible to users seeking to evaluate their stroke risk.

Background

The dataset utilized for this project comprises 5,110 entries with 12 risk factors, each related to factors influencing stroke prediction. Key features include BMI, average glucose levels, age, hypertension, heart disease, and smoking status. These variables are critical in understanding the multifaceted nature of stroke risk, which is influenced by both physiological and lifestyle factors.

Understanding the relationship between these features is essential for accurate prediction. Machine learning provides powerful tools for identifying patterns and relationships within the data that may not be immediately apparent. Random Forests have the ability to handle complex, non-linear interactions between variables. Additionally, understanding that the risk of a stroke is relatively rare is crucial for developing a reliable model. Techniques such as SMOTE (Synthetic Minority Oversampling Technique), class weighting, and resampling are instrumental in mitigating this imbalance.

The user interface is designed with a user-centric approach, prioritizing ease of use and clarity. By providing an intuitive platform for users to input their data and receive personalized feedback, we aim to empower individuals to take proactive steps in managing their health.

Methodology

The methodology of our project encompasses several key stages: data preprocessing, model development, and user interface design.

In the data preprocessing stage, we addressed challenges such as missing values and feature selection. The BMI column contained missing values, which we imputed using the mean BMI of the available data to maintain dataset integrity. The smoking status feature included "unknown" values. Initially, these were retained to preserve as much data as possible, but further analysis revealed that smoking status and being a former smoker showed no significant correlation with stroke risk. Due to their lack of influence on the prediction factor, we removed these features to streamline the model and eliminate excess data.

Feature engineering involved transforming categorical variables into numerical formats suitable for machine learning algorithms. We converted variables such as gender, hypertension, heart disease, marital status, work type, and then put their likelihood of having a stroke into categorical numerical representations (e.g., 0s and 1s). This process is vital for algorithms that require binary input and ensures that the categorical nature of the data is preserved in a form the model can interpret.

We conducted a correlation analysis to identify the most influential features. The correlation matrix indicated strong relationships between stroke risk and factors like glucose levels, BMI, hypertension, and heart disease. These insights guided our feature selection and informed the focus of our predictive modeling efforts.

In developing the predictive model, we explored several machine learning techniques. We initially attempted to use decision trees due to their simplicity and interpretability, but decision trees resulted in low accuracy rates and did not handle the class imbalance effectively. The decision tree's lack of performance led us to adopt Random Forests. Random Forests are an ensemble method that combines multiple decision trees to enhance predictive performance. Random Forests provided improved accuracy.

Addressing the class imbalance was a critical aspect of our modeling process. It is very important to our model that it does not predict a false negative as strokes are incredibly serious and can lead to death. The rarity of stroke cases in the dataset posed a challenge for accurate prediction. We employed several techniques to prevent our model from mispredicting a stroke. Class weighting adjusted the importance assigned to the stroke class during model training, while resampling methods involved oversampling the minority class to achieve a more balanced dataset. SMOTE was particularly effective, generating synthetic samples of the minority class to improve model learning without duplicating existing records.

Hyperparameter tuning was conducted to optimize the Random Forest model's performance. We experimented with various parameters, such as the number of trees in the forest, maximum depth of the trees, and the minimum number of samples required to split an internal node. Through cross-validation, we identified the optimal settings that maximized the model's recall—the ability to correctly identify true positives—while maintaining an acceptable level of precision.

The user interface was developed using Streamlit, a Python library that enables the creation of interactive web applications. Initially, we used Live Server through Visual Studio Code to host our website, but since that was only a temporary host, we had to move to Streamlit. The interface comprises several components designed to guide the user through the assessment process seamlessly. The Welcome Page introduces the system's purpose and encourages users to proceed with the assessment. The Input Form Page allows users to enter related health data, such as age, BMI, average glucose level, hypertension status, and heart disease history. The form is designed for simplicity, using dropdown menus and binary options to facilitate ease of use. We also have interactive buttons that animate when hovered over that make the user-interface visually appealing and interactive. Everything about the UI was meticulously chosen in order to enhance the user experience from the color scheme to animated buttons to the simplicity of a dropdown menu.

Upon submitting their information, users are directed to the Risk Analysis Page, where their stroke risk is displayed as a probability percentage. A color-coded risk bar visually represents their risk level, with green indicating low risk, yellow for moderate risk, and red for high risk. The page also lists the key contributing factors affecting their risk, providing transparency and helping users understand the rationale behind the assessment.

The final thing included in the UI is the Recommendation Page which offers personalized advice based on the user's specific risk factors. Recommendations encompass lifestyle changes, medical advice, and dietary modifications. For example, users with high BMI may receive suggestions to incorporate regular exercise into their routine, while those with elevated glucose levels might be advised to monitor their sugar intake.

Results

The exploratory data analysis provided valuable insights into the factors most strongly associated with stroke risk. Elevated average glucose levels and high BMI were consistently identified as top contributors. Hypertension and heart disease also emerged as significant predictors. These findings corroborate existing medical knowledge about stroke risk factors and validate the relevance of the features used in the model.

User testing of the interface indicated that the system is accessible and user-friendly. Test users reported that the process of entering data and receiving feedback was straightforward. The visual representations of risk levels and recommendations were particularly appreciated, as they enhanced understanding and made the information more accessible and visually appealing.

Discussion

The project demonstrates the potential of integrating machine learning with user-centric design to create effective preventive healthcare tools. The emphasis on recall in the model aligns with the ethical imperative to minimize harm by ensuring that individuals at risk are not overlooked. While this approach increases the number of false positives, the benefits of early detection and intervention justify this outcome.

Addressing the class imbalance was a significant challenge. The use of SMOTE proved instrumental in enhancing the model's ability to learn from the minority class without introducing bias or overfitting. Feature engineering and careful selection of variables contributed to the model's improved accuracy and generalizability.

Our decision to prioritize recall over precision was deliberate. In the context of stroke prediction, false negatives (failing to identify someone at risk) have more severe consequences than false positives. By maximizing recall, we aim to minimize the likelihood of overlooking individuals who are at risk of stroke.

The project also highlighted areas for future improvement. Incorporating additional features, such as genetic factors or real-time health monitoring data, could further enhance the model's predictive capabilities. Additionally, integrating the system with healthcare providers could facilitate more comprehensive care and follow-up.

Conclusion

Our personalized health plan system successfully combines advanced machine learning techniques with an intuitive user interface to provide individuals with valuable insights into their stroke risk. By prioritizing recall and employing effective data balancing methods, we have developed a model that accurately identifies those at risk.

The project underscores the transformative potential of AI in preventive healthcare. By empowering individuals with knowledge and personalized recommendations, we contribute to the broader goal of reducing the incidence and impact of strokes. The project represents a significant step toward harnessing technology to address critical health challenges. We created a tool that not only predicts risk, but also fosters positive health behaviors, ultimately contributing to improved health outcomes.

Contributions

A brief section (1 sentence per group member) describing each member’s contributions.

Diya: I was responsible for the entire model-building process (handling missing values, feature engineering (categorical transformations, feature selection, and removal of low-correlation variables), balancing the dataset using techniques like SMOTE and class weights, hyperparameter tuning, and implementing the Random Forest model to achieve high recall and precision. I also conducted data preprocessing/cleaning as well as EDA in regards to the correlation matrices and findings. I was also a significant contributor to the slides and the (now optional) final paper, as well as the proposal and the check-in. In addition, Priyanshi and I took on more of a “team lead” position within the project group, and helped keep the group on track over the quarter and delegated tasks as we saw fit.

Ken: I helped with the feature engineering, and helped Bill with the UI integration on streamlit.

Chris: I helped with the project check in, feature engineering, and presentation.

Bill: I was responsible for conducting the initial literature review, which helped us understand the background of our project. I also worked on portions of the data preprocessing to ensure it was ready for analysis and compatible with our model. Additionally, I focused on the user interface integration with our trained model on streamlit to allow users to interact with our model and visualize the results.

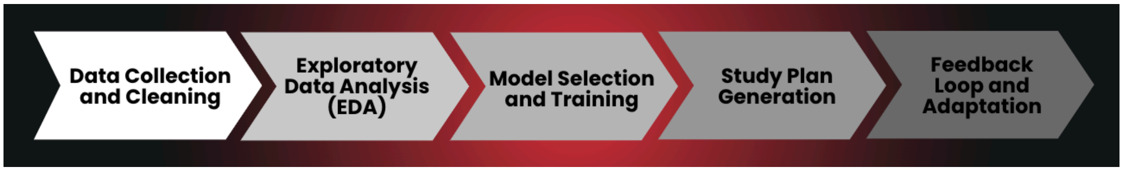
Priyanshi: Diya and I look over more of a project manager role throughout the quarter and helped ensure that tasks we being completed and done. I also was responsible for our project proposal (including a literature review and formatting things into markdown), was a significant contributor to the project check-in, was an active contributor to finding ideas and project ideas for our team to pivot towards, and helped with the presentation.

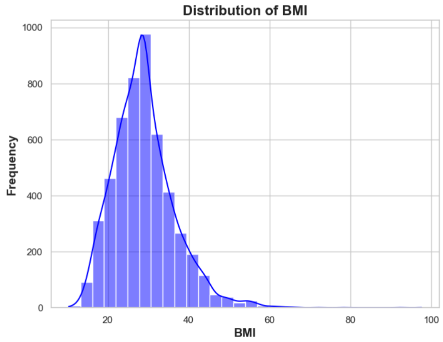
Maithreyi: I was primarily responsible for data preprocessing and conducting exploratory data analysis (EDA) on the dataset. This involved analyzing skewness through histograms, pair plots, correlation matrices, and Q-Q plots, as well as creating multiple graphs to visualize key insights. Additionally, I contributed to the planning of the project's implementation, the proposal, and the project check-ins. I also assisted in preparing the slides and creating other necessary materials for the presentation.

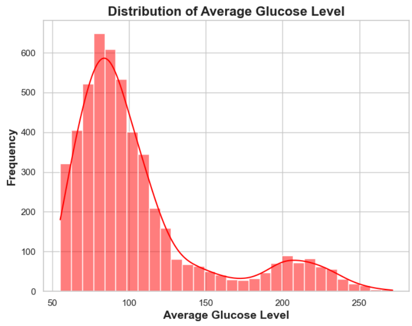
Nissi: Worked primarily on data preparation, identifying, evaluating, and preprocessing datasets to ensure optimal training for our model. I helped with the project proposal, check-in, and project planning stages, and also conducted a literature review to strengthen our methodology.

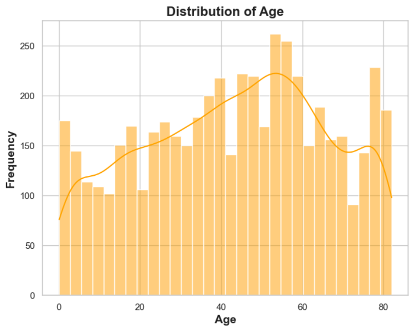
Sheda: I assisted with the UI, the final paper, the presentation, the proposal, and the check-in. I also made the Heatmap. I assisted in the original development of the project idea, and helped come up with new ideas when our original one proved too challenging.

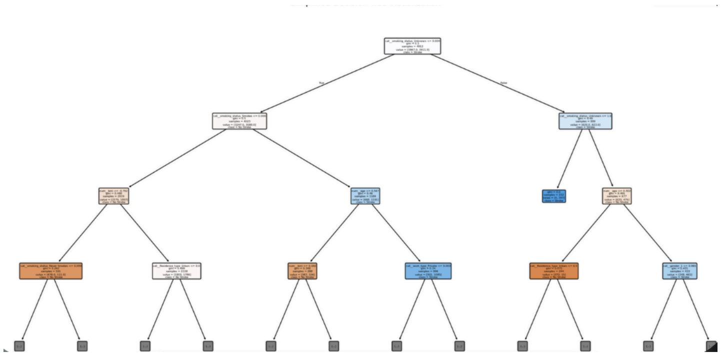
Relevant Images:

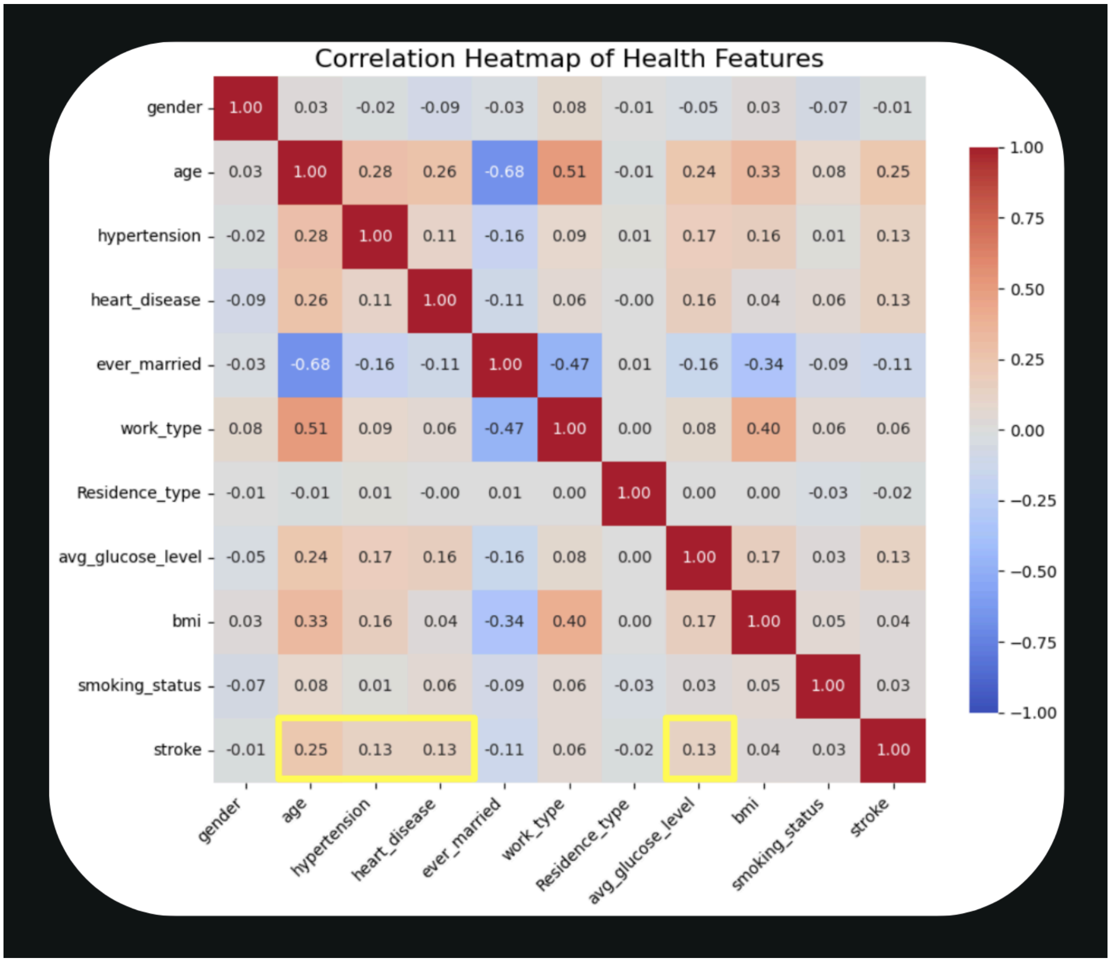

Data Collection and Cleaning
Exploratory Data Analysis (EDA)
Model Selection and Training
Study Plan Generation
Feedback Loop and Adaptation










Correlation Heatmap of Health Features