ECEN 360 - Final Report

Stroke Risk Prediction Model

Affiliation: Texas A&M University

Names: Ashrith Thallam - Ashrith Thallam
Kevin Eldho - Kevin Eldho
Aryan Chaubey - Aryan Chaubey

Executive Summary

Strokes are and have been a leading cause of death or long-term disability worldwide. This can place heavy burdens on families, patients, and healthcare systems. The goal of this project was to develop a reliable and accurate machine learning model that could be used as a tool for the early detection of individuals with a high risk of developing a stroke. By using a related dataset, we aimed to deliver a predictive model that balances accuracy and sensitivity.

The main goals were to first do the risk prediction by creating machine learning classifiers that were capable of detecting an individual's probability of developing a stroke based on some routinely collected health variables from the data. The next step was to identify these variables, factor identification. Determine whether features such as age, hypertension status, heart disease and medical history, smoking status, etc, relate to or influence stroke risk. Then we assess the model by evaluating and comparing multiple algorithms to find the best approach and focus on performance metrics that could have real-world utility.

Introduction

Stroke is a significant global health concern, ranked as the second leading cause of death worldwide by the World Health Organization (WHO). The risk of stroke is influenced by various factors, including but not limited to age, gender, smoking status, and pre-existing medical conditions. This project aims to develop a machine learning model capable of predicting an individual's likelihood of stroke. Being able to identify these risk factors and accounting for them could help in reducing stroke-related mortality rates. Our main goal for our model is to aim for 70% in recall for both stroke and non-stroke cases and 75% for model accuracy.

Developing a predictive model for stroke risk offers substantial benefits. By enabling early detection, the model would allow healthcare professionals to intervene proactively, reducing the likelihood of severe health consequences. Additionally, it would help hospitals by minimizing emergency stroke cases. Preventing strokes through early risk assessment can lead to significant cost savings in medical treatment and rehabilitation for both individuals and insurance providers. Moreover, by offering insights into stroke risk, the model could encourage individuals to adopt healthier lifestyles, ultimately reducing their susceptibility to strokes and improving overall public health.

Methodology

We started by loading our dataset (stroke-data.csv) into a pandas DataFrame. Since some BMI values were missing, we grouped the patients into age ranges of ten years each, then filled in the missing BMIs with the median BMI from their age group. After that, we converted categorical features like gender, smoking status, work type, and residence type into numerical form using one-hot encoding. For the binary variable "ever married," we simply converted it to 0 or 1. We haven't scaled any of our continuous variables (age, average glucose level, and BMI) yet, since we planned to handle them later in our pipelines.

Next, we did some exploratory data analysis on our data. We created violin plots to see how the distributions of continuous variables (age, average glucose level, and BMI) differ between people who had strokes and those who didn't. We also used a seaborn pairplot to visualize these variables together and identify any strong relationships or potential issues with collinearity. To explore our categorical data, we made grouped bar charts to visualize stroke incidence across different categories, such as gender, smoking status, work type, residence type, and marital status.

After preparing the data, we built three basic machine learning models to serve as a baseline: logistic regression, random forest, and histogram gradient boosting. Each model included steps to encode categorical variables, scale the continuous variables, and perform classification. We trained these models on 80% of the dataset using default hyperparameters and evaluated their performance using cross-validation. This step helped us understand how each model performed initially, before tuning it further.

Since our dataset had a severe imbalance, with less than 5% of cases experiencing a stroke, we used a technique called SMOTE to balance the classes by oversampling stroke cases only within the training data. We then tuned hyperparameters of the random forest and histogram gradient boosting classifiers using RandomizedSearchCV with five-fold cross-validation, focusing specifically on optimizing recall for stroke cases. However, each tuned model didn't meet our goals, so we wanted to get the best of the three models into one model.

Thus, we built a stacking classifier model. This stacked model combined the predictions from our tuned logistic regression, random forest, and histogram gradient boosting models, using logistic regression as a final step to blend their predictions. After confirming that this stacking model outperformed our models, we calibrated its decision threshold by testing different probability cutoffs on a small validation set, selecting the lowest threshold that still achieved our target recall of at least 75%. We finalized this pipeline, including SMOTE, preprocessing, the stacking classifier, and the calibrated threshold, for predicting stroke risk in new patients.

Results

Our project had two secondary goals: first, figure out which features are most important for predicting stroke risk, and second, see how well different classifiers perform on this prediction task. In the tuned Random Forest model, we found the most important features were average glucose level (0.2808), age (0.2357), and BMI (0.2327). Demographic features like gender, residence type, and work type all had importance scores below 0.03, showing that health indicators were way more critical for predicting strokes. Exploratory plots backed this up, violin plots showed stroke patients generally had higher and more variable glucose levels, while age-versus-glucose plots made it clear that older patients with higher glucose levels were more likely to have strokes. Boxplots of BMI showed that while BMI alone didn't separate stroke from non-stroke patients, it was still useful when combined with other factors.

Next, we tried out baseline models and then tuned them using SMOTE with RandomizedSearchCV, which revealed some interesting trade-offs. The tuned Logistic Regression model had great recall (0.8400), meaning it caught most stroke cases, but had low precision (0.1273), resulting in lots of false positives. Its overall accuracy was 71.04% with a solid ROC-AUC of 0.8342. The tuned Random Forest, on the other hand, gave a balanced result: high precision (0.9736) and recall (0.8354) for non-stroke cases, but lower precision (0.1489) and recall (0.5600) for stroke cases. This led to better overall accuracy (82.19%) but slightly lower ROC-AUC (0.7916). Logistic Regression might be best if missing a stroke would be harmful, whereas Random Forest gives better overall accuracy but slightly fewer stroke detections.

Finally, we created a stacked ensemble model by combining tuned Logistic Regression, Random Forest, and Histogram Gradient Boosting models, using Logistic Regression as the meta-model. This turned out to be our best model. With a probability threshold of 0.07, it had solid recall (0.7600) for stroke cases and great recall (0.7479) for non-stroke cases, plus high precision (0.9838) for non-stroke cases. Overall accuracy was 74.85%, and ROC-AUC reached 0.8281. This stacked model performed better than any of our tested models at detecting stroke cases while maintaining strong overall performance.

Discussion

Our project highlighted some key insights into stroke prediction, notably that average glucose level, age, and BMI were consistently the most influential factors in our models. Clinical indicators turned out to be way more important than demographic factors like gender or work type, reinforcing the idea that managing glucose, body weight, and age-related health interventions can significantly impact stroke prevention efforts. These findings suggest that health professionals should prioritize clinical assessments and targeted health interventions rather than focusing solely on demographic characteristics.

In terms of model performance, we successfully achieved the main goals of our analysis. Logistic Regression was great at identifying potential stroke cases (high recall), which would be useful in situations where missing an actual stroke could be dangerous. On the other hand, the Random Forest provided higher overall accuracy and balanced the rate of false positives and false negatives more effectively. Ultimately, our stacking ensemble model combined these strengths to deliver solid recall for strokes (76%) while maintaining high overall accuracy, making it the best option out of all of our tested models.

However, there are some limitations to our study. First, our dataset had a big class imbalance, with fewer stroke cases compared to non-stroke cases, which probably impacted our models' precision and recall, even after applying techniques like SMOTE. Second, we didn’t have access to some important medical information, like full medical histories, medications, or genetic factors, which could have improved our predictions. Finally, although our models performed well internally, we didn't test them on external datasets, meaning we can't yet confirm how well they'd generalize to new patient groups.

Future research should include more comprehensive clinical data and longitudinal studies tracking patient health over time, as well as validating these models with external datasets to improve reliability across diverse populations. Exploring advanced ensemble methods or hybrid approaches might further enhance performance by capturing both the sensitivity of Logistic Regression and the accuracy of Random Forest models. Additionally, incorporating explainable AI techniques could help healthcare providers better understand and trust these models, encouraging their practical use in clinical decisions.

Conclusion

This project showed that machine learning can be used to effectively predict the likelihood of a person developing a stroke to a degree by using proper datasets. Through the data processing and model evaluation, we saw that age, heart disease, and glucose level were the most influential predictors of a stroke. The random forest model and the linear regression model were both successful versions. For future reference and real-world applications, the model should get some external validation to enhance its predictive performance.

The problem of stroke prevention is closely related to these discoveries as early detection can significantly improve outcomes. Healthcare professionals can move from reactive treatment to proactive management by identifying patients whose risk factors match those high-importance characteristics. Furthermore, the ensemble technique provides additional robustness in more data-rich contexts, while the interpretability of the linear model makes it ideal for settings that require transparency.

Future Work

There are several ways we could improve our stroke-prediction model in future work. First, adding more detailed patient data, such as lab test results, medication histories, or genetic risk factors, could help identify new predictors of stroke risk. Another critical issue is the model's low precision, which is around 13%, meaning many predictions might be false alarms. To improve this, we could experiment with different resampling methods, cost-sensitive algorithms, or calibration techniques to better balance sensitivity and specificity. Without addressing this precision issue, the model probably wouldn't be practical for clinical use.

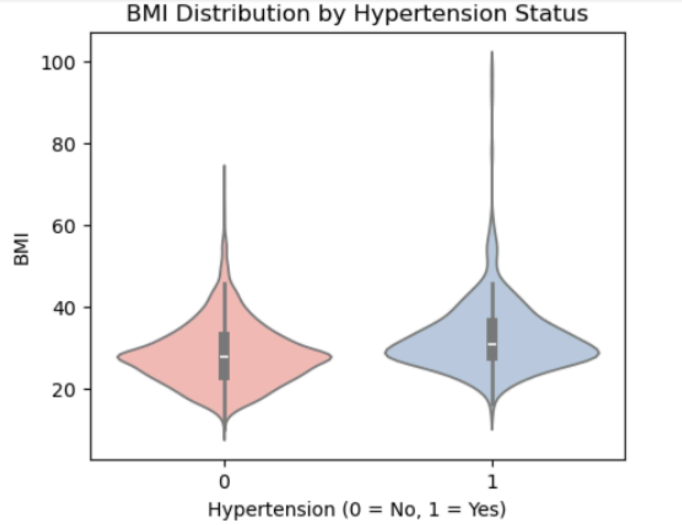
Additionally, we should test the model using new data from different hospitals or real-world clinical environments. This type of external validation would reveal how generalizable our model is and whether adjustments are needed to improve performance across diverse patient populations. It would also be important to check how fairly the model performs across different groups based on age, gender, or ethnicity, ensuring it provides accurate predictions for everyone.

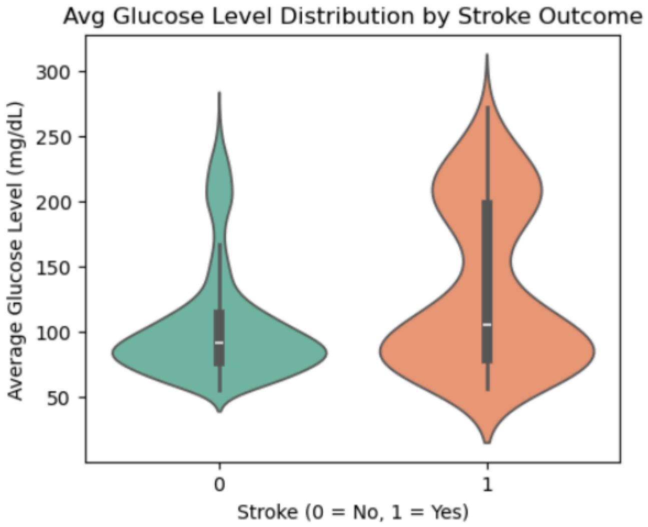
Finally, before integrating this model into clinical practice, we need to consider how it would work in real healthcare settings. Using interpretability tools like SHAP could help clinicians understand and trust the predictions. Pilot studies could be conducted within electronic health record systems to evaluate issues such as alert fatigue, decide how frequently the model needs updating, and measure whether it positively impacts patient outcomes.

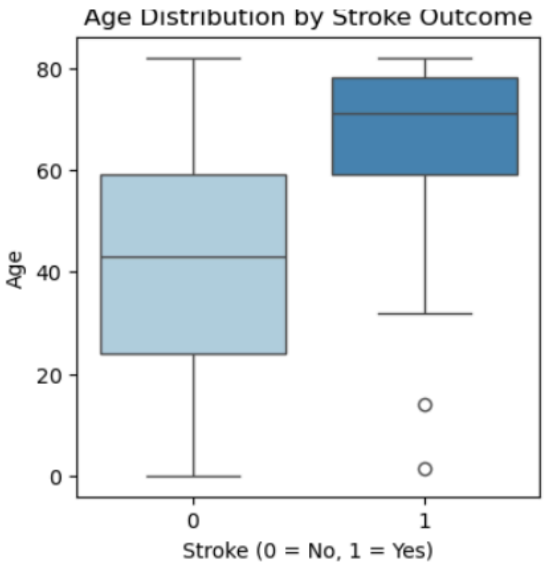
References

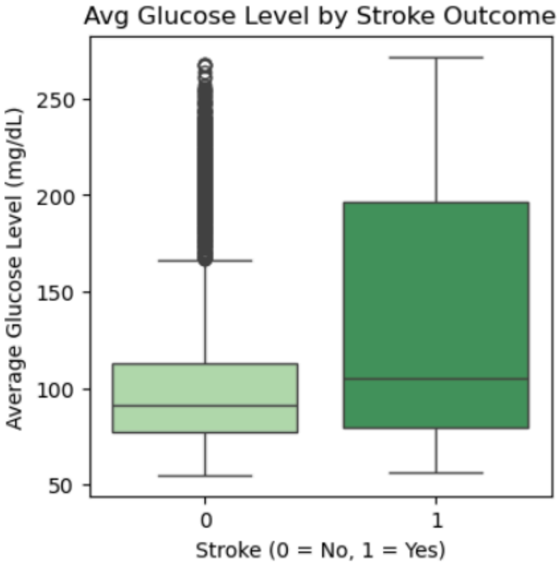
Kaggle. (n.d.). *Stroke Prediction Dataset*. Retrieved January 2025, from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

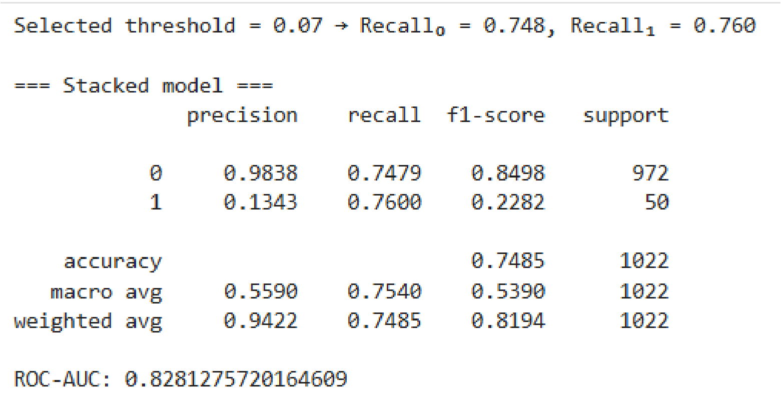
Appendices:


BMI Distribution by Hypertension Status
BMI
100
80
60
40
20
0
1
Hypertension (0 = No, 1 = Yes)


Avg Glucose Level Distribution by Stroke Outcome
Average Glucose Level (mg/dL)
300
250
200
150
100
50
0
1
Stroke (0 = No, 1 = Yes)







```
Selected threshold = 0.07 → Recall₀ = 0.748, Recall₁ = 0.760

=== Stacked model ===
              precision    recall  f1-score   support

           0     0.9838    0.7479    0.8498       972
           1     0.1343    0.7600    0.2282        50

    accuracy                         0.7485      1022
   macro avg     0.5590    0.7540    0.5390      1022
weighted avg     0.9422    0.7485    0.8194      1022

ROC-AUC: 0.8281275720164609
```

Signature Page

Aryan Chaubey

Kevin Eldho

Ashrith Thallam