Data Science Project: Predicting Brain Strokes

By: Ananya Katyal, Bianca Sellemi, Victor Li

Abstract

Given a large dataset of possible factors leading to brain strokes and whether or not a brain stroke occurred, we hope to develop a method to predict whether or not a brain stroke will occur. We explored various forms of data visualization to display the correlation between variables and the occurrence of a brain stroke; this would give us a better grasp of identifying risk factors that we can use to better inform future users of our model. Next, we explored two types of machine learning models, a KNeighborsClassifier and a RandomForestClassifier, the former being better suited for smaller datasets and the latter for larger. The KNeighborsClassifier performed slightly better with a higher true positive rate.

Introduction

Brain strokes are a leading cause of disability and death worldwide, making early detection and prevention critical for improving patient outcomes. Predicting the likelihood of a stroke involves analyzing a variety of risk factors to identify individuals at higher risk. Machine learning algorithms offer a powerful approach to this problem by learning patterns from patient data and providing predictions that can guide clinical decision-making. In this work, we hope to develop a predictive model for brain strokes using patient data, aiming to create a tool that can assist healthcare providers in identifying at-risk individuals with greater accuracy and efficiency. The dataset includes 10 different potential factors along with whether or not a brain stroke occurred, the factors being: gender, age, hypertension, heart disease, ever married (marriage status), work type, residence type, average glucose level, BMI, and smoking status.

Methods

Victor:

Starting with data preprocessing, I converted all string variables into integers (e.g., "yes" and "no" became 1 and 0, respectively). A significant portion of the data had missing values for smoking status. To address this, I used a weighted random approach to fill these gaps proportionally, based on the distribution of known smoking statuses. This adjustment led to a notable improvement of 6–8% in the true positive rate. The dataset was then split into training and testing sets, with a test size of 40%, a proportion determined to yield the highest accuracy after rigorous testing. A KNeighborsClassifier was chosen for its simplicity and effectiveness in capturing local patterns in the data, particularly for datasets where relationships between features and outcomes may not be linear. After training, the model was evaluated based on its true positive rate (sensitivity) and true negative rate (specificity), alongside the counts of correct and incorrect predictions.

Ananya:

The dataset was loaded from a CSV file, inspected for structure, and preprocessed by converting categorical variables into numerical formats using one-hot encoding. Features and the target variable were separated, and the data was split into training and testing sets (80/20 split). To address class imbalance, Synthetic Minority Oversampling Technique (SMOTE) was applied to the training data to generate synthetic samples for the minority class ("stroke"). A Random Forest Classifier was then trained on the resampled data and evaluated using accuracy, AUC-ROC, and a confusion matrix. Hyperparameter tuning was performed using GridSearchCV to optimize parameters like the number of trees and maximum depth. The model was re-evaluated with the tuned parameters, with metrics such as recall, specificity, and F1-score calculated for detailed performance insights.

Bianca:

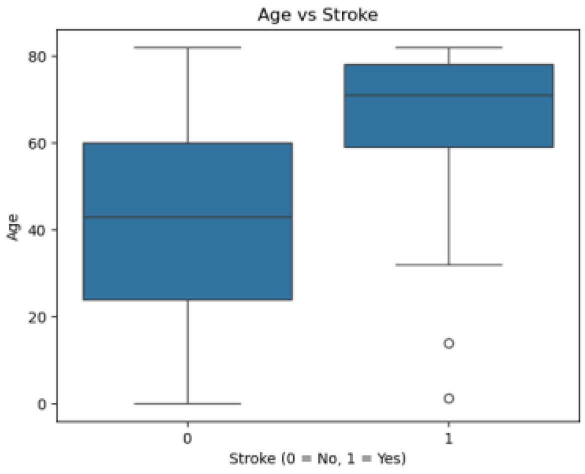
The CSV data file was first read and validated to ensure each column aligned correctly with its variable type. Missing or invalid rows were skipped, and string variables were converted into numerical formats for analysis ("Male" and "Female" were kept as categorical strings, while binary features like "Yes" and "No" were mapped to 1 and 0). To visualize the relationships between stroke and various health factors, categorical variables such as smoking status, hypertension, and heart disease were analyzed using bar charts, showing how different categories contributed to stroke occurrence. For continuous variables like age, BMI, and glucose levels, boxplots were employed to observe their distribution across stroke and no-stroke groups, highlighting variations in medians and ranges. The analysis aimed to identify patterns and correlations, such as higher glucose levels or older age being more common among stroke patients, providing actionable insights into stroke prediction and prevention.

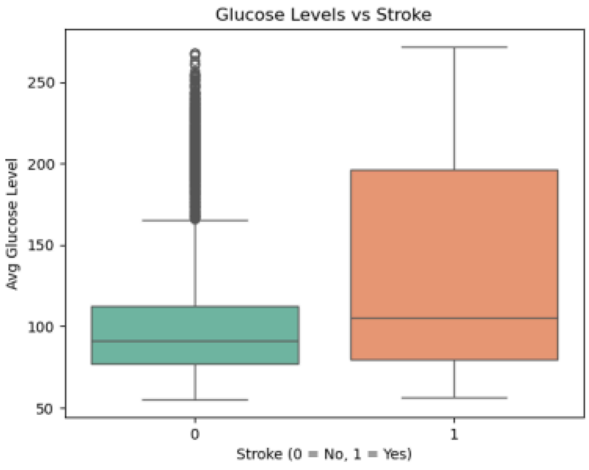
Results

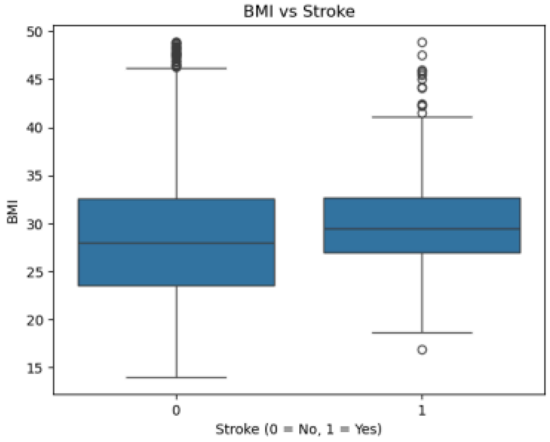
After 10 trials, our first model, the KNeighborsClassifier, achieved an average true positive rate of 15.07% and an average true negative rate of 95.49%. The highest recorded true positive rate was 20.99%, while the lowest was 13.04%. True negative rates ranged from 94.7% to 96.1%. This disparity in performance can be attributed to the imbalanced dataset, which had significantly more non-stroke cases than stroke cases. The limited number of stroke cases provided insufficient data for the model to learn patterns effectively for positive predictions.

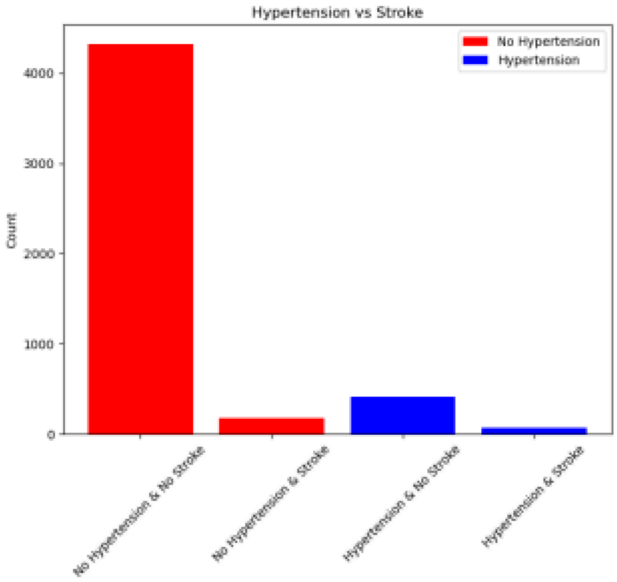
Our second model, the Random Forest Classifier, achieved a true positive rate (recall) of 14% and a true negative rate (specificity) of 97%, correctly identifying 7 stroke cases and 914 non-stroke cases. However, the model produced 33 false positives and 43 false negatives, highlighting difficulty in detecting stroke cases due to dataset imbalance. While specificity was high, the low recall indicates that the model struggled with the minority class, even after applying SMOTE and hyperparameter tuning.

Finally, the results of our visualization code demonstrate that the strongest correlation between stroke and factor *x* is with age, glucose levels, hypertension, and BMI levels. The charts and graphs below represent our findings:

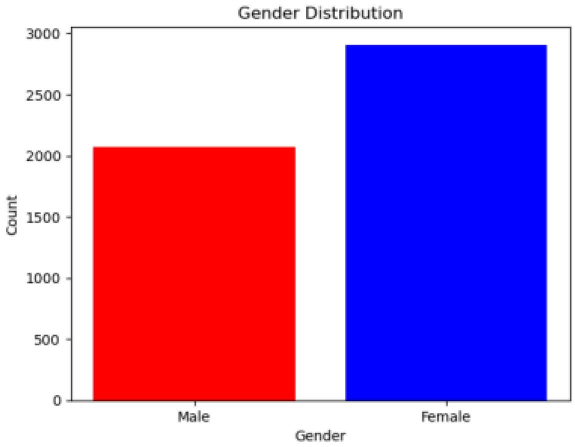


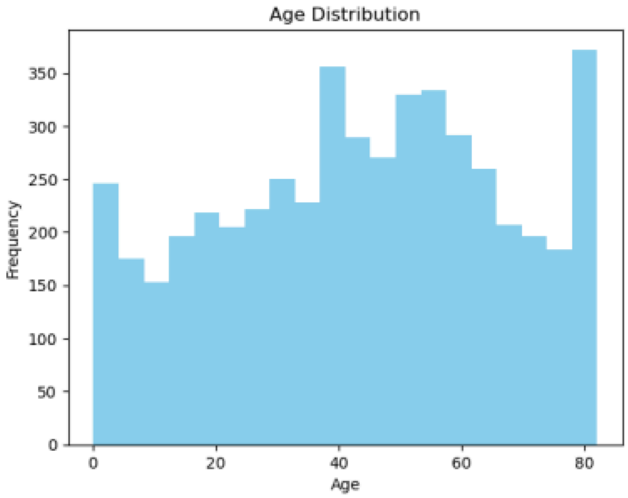


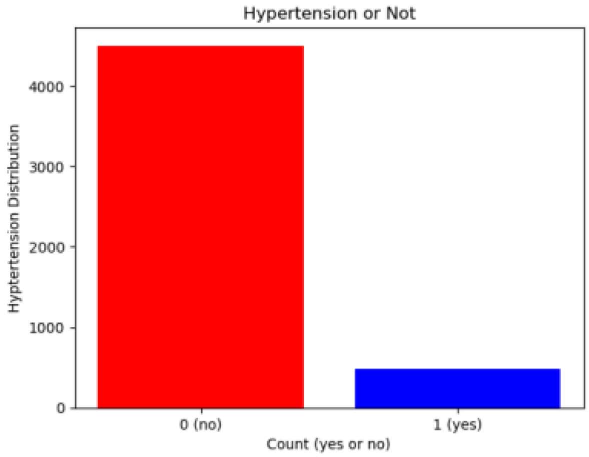


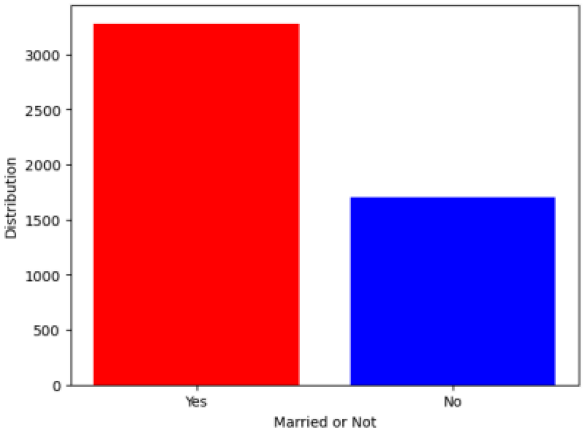


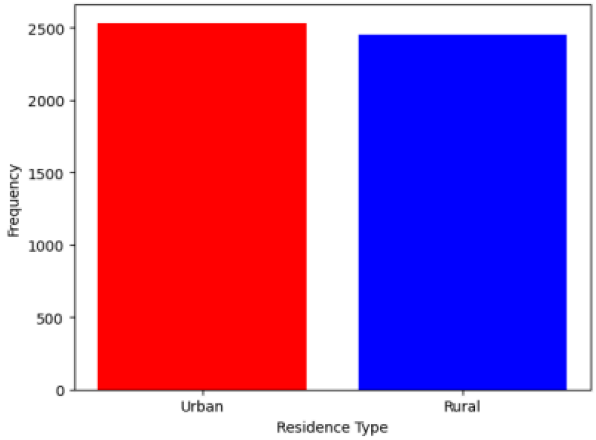
The remaining graphs below visualize each respective variable from the CSV file:

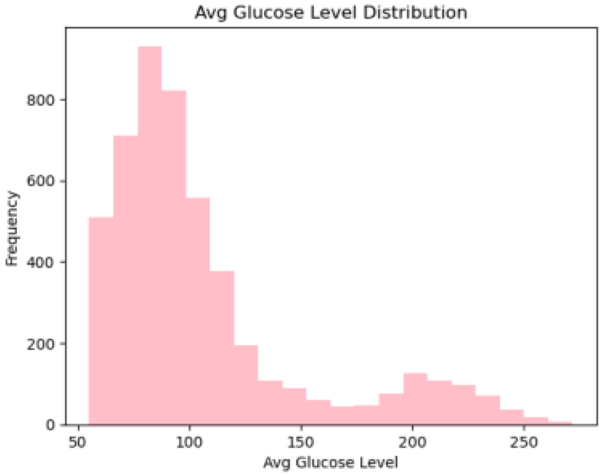


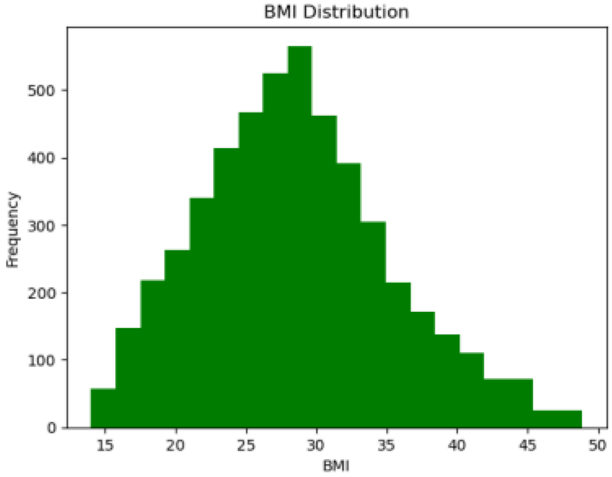


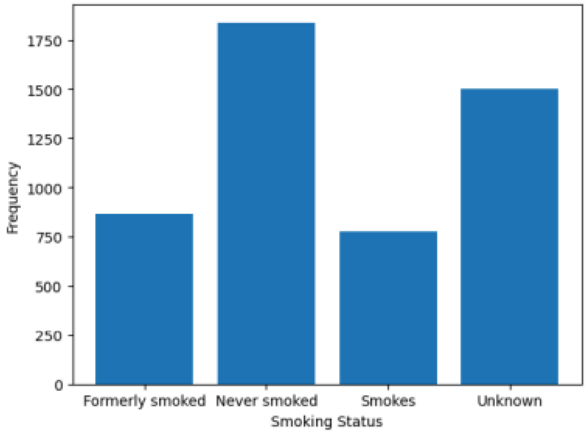

Hypertension or Not
Hyptertension Distribution
4000
3000
2000
1000
0
0 (no)
1 (yes)
Count (yes or no)

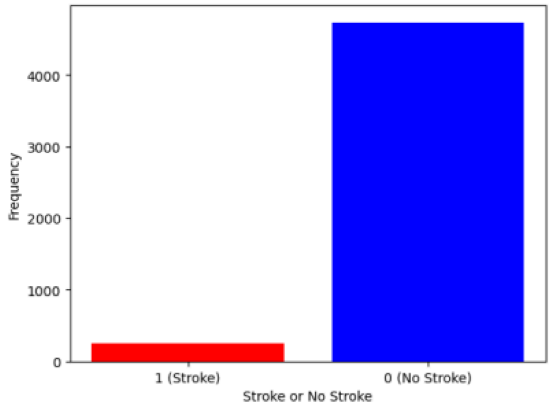

Distribution
3000
2500
2000
1500
1000
500
0
Yes
No
Married or Not


Frequency
2500
2000
1500
1000
500
0
Urban
Rural
Residence Type


Avg Glucose Level Distribution
Frequency
800
600
400
200
0
50
100
150
200
250
Avg Glucose Level


BMI Distribution
Frequency
500
400
300
200
100
0
15
20
25
30
35
40
45
50
BMI


Frequency
1750
1500
1250
1000
750
500
250
0
Formerly smoked
Never smoked
Smokes
Unknown
Smoking Status



We experienced imbalance in our data because out of the nearly 5K data logs we have, a very small proportion actually experienced a stroke. As a result, there is likely a skewed error with our results which we plan to take into account when we enhance the machine learning model.

Conclusion:

Based on the visualizations provided above, we have identified key variables as risk factors that could enhance an individual's chance of undergoing a brain stroke. Age is the leading risk factor and it is very well known that a significantly higher proportion of individuals who experienced a stroke tended to be older on average compared to those who did not. However, because of how broad the chart is, the data demonstrates that younger individuals are not completely exempt and may very well still be victims. Additionally, hypertension is a common risk factor because of its impact on blood vessel integrity and blood pressure regulation; our results suggest that monitoring and controlling blood pressure in hypertensive individuals is necessary to reduce their chances of stroke. Our third most powerful risk factor is glucose levels, which are often associated with diabetes, often a catalyst for strokes as well due to its effects on blood vessels and the nervous system. It's recommended that blood sugar levels are monitored and managed to play a role in prevention. Our other variables ranked in the data from most to least significant risk factors include: heart disease, BMI, and smoking status while gender, marital status, and residence type offer minimal predictive conclusions.

For the future, our predictive brain stroke model can make a crucial impact by continuing to leverage advanced machine learning and AI techniques to enhance accuracy and scalability. By using algorithms like Random Forests and K-Neighbors Classifiers, the model can identify critical stroke risk factors with high precision, enabling early interventions and personalized healthcare. Expanding the dataset with more diverse and comprehensive health metrics—such as cholesterol levels, physical activity patterns, and real-time monitoring data from wearable devices—can further refine predictions. Additionally, integrating explainable AI (XAI) methods will help medical professionals understand the rationale behind predictions, fostering trust and adoption in clinical settings. As the model evolves, we believe it has the potential to be implemented in real-time health monitoring systems, empowering individuals with timely alerts and reducing stroke-related fatalities through proactive care.

References

Tech, J. S. (2022). *Brain Stroke Dataset*. Kaggle.com. https://www.kaggle.com/datasets/jillanisofttech/brain-stroke-dataset

parrt. (2023, July 13). *GitHub - parrt/dtreeviz: A python library for decision tree visualization and model interpretation.* GitHub. https://github.com/parrt/dtreeviz

1.11. Ensembles: Gradient boosting, random forests, bagging, voting, stacking. (2022). Scikit-Learn. https://scikit-learn.org/stable/modules/ensemble.html#random-forests

2. Over-sampling — Version 0.12.4. (2014). Imbalanced-Learn.org. https://imbalanced-learn.org/stable/over_sampling.html#smote

3.2. Tuning the hyper-parameters of an estimator. (2016). Scikit-Learn. https://scikit-learn.org/stable/modules/grid_search.html

3.4. Metrics and scoring: quantifying the quality of predictions. (2015). Scikit-Learn. https://scikit-learn.org/stable/modules/model_evaluation.html#confusion-matrix