Introduce the problem (Isaiah)

According to the World Health Organization (WHO), stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. A stroke occurs when the blood supply to part of the brain is interrupted or reduced, preventing brain tissue from getting oxygen and nutrients, which can lead to lasting brain damage, long-term disability, or death. This project will use classification techniques to determine whether a person is likely to have a stroke based on specific diagnostic and demographic factors, including age, hypertension status, heart disease, average glucose level, body mass index (BMI), and smoking status.

Introduce the data (Isaiah)

This dataset originates from a healthcare context and was posted on Kaggle by Fedesoriano. The healthcare-dataset-stroke-data.csv is 316.97 kB large and contains several features, including a unique id, gender, age, hypertension status, heart disease status, marital status, work type, residence type, average glucose level, body mass index (BMI), smoking status, and stroke outcome. For the target variable, 1 signifies that the patient had a stroke, and 0 signifies that they did not. The primary objective of this dataset is to predict whether a patient is likely to have a stroke based on their specific diagnostic and demographic information.

Preprocessing (Brandon)

The preprocessing step focused on preparing the dataset for classification so the model could learn effectively. The ID column was removed since it did not contribute to predicting a stroke. Missing values in the BMI column were filled using the median in order to avoid losing samples while keeping the data realistic. Categorical variables such as gender, marital status, and smoking status were converted into a numeric form using one-hot encoding so the model could understand them.Numerical features such as age, glucose level, and bmi, were scaled with

StandardScaler to put them all within a similar range. Lastly the data was split into a training and testing set using stratification to keep stroke vs. non-stroke proportions consistent between both sets.

Modeling (Jack)

To build our classification model for predicting stroke risk, we used the cleaned and preprocessed dataset to train a Random Forest Classifier. We chose this model because it performs well on tabular medical data, captures nonlinear relationships, and handles class imbalance effectively using the class_weight="balanced" setting. After splitting the data into training and testing sets, the model learned patterns based on features such as age, glucose level, BMI, hypertension status, and other demographic variables.

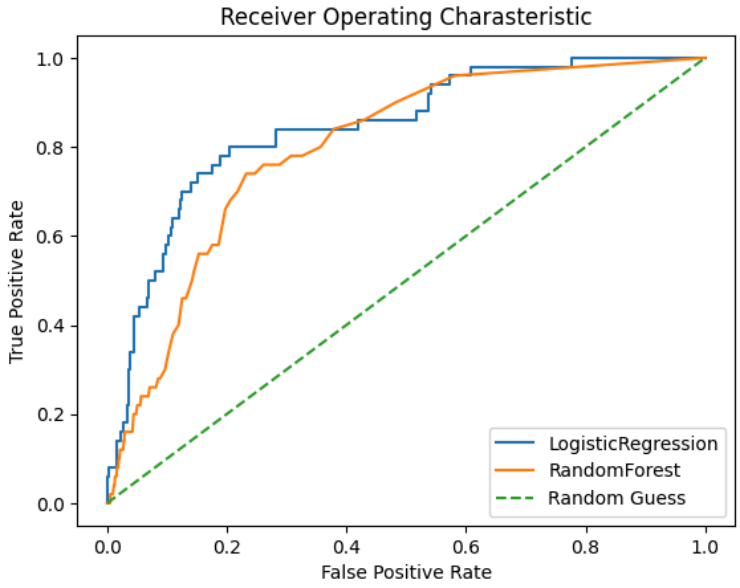
We evaluated the model using accuracy, precision, recall, F1-score, and the ROC-AUC score to measure how well it distinguishes between stroke and non-stroke cases. The confusion matrix and classification report helped us understand the types of errors the model made and how well it identified the minority class. Overall, this approach allowed us to build a reliable baseline model for predicting stroke risk from health and demographic information.

Evaluation (Gabriel)

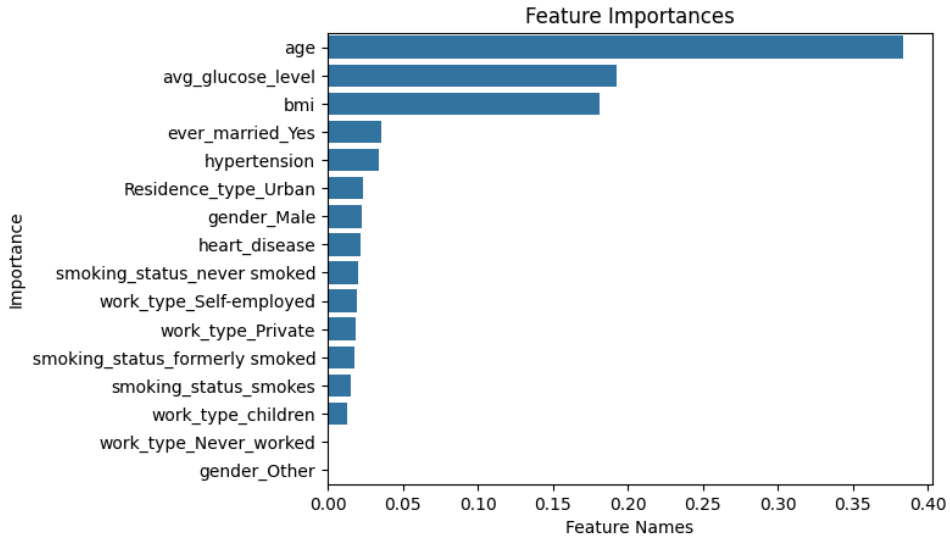
For this project, we employed several different metrics in order to evaluate the Machine Learning models. These metrics included the f1-score, precision, recall, and accuracy. In order to get a better sense of performance according to the data, we used a confusion matrix to visualize the actual predictions and ROC-AUC to assess the models ability to separate the classes. The dataset we chose for this project has a severe class imbalance as 95% of the data belongs to the negative class (no stroke). Additionally, with only 5% of the data belonging to the positive class (actual stroke) we could easily determine that most of the models would not perform well with

the raw data. Therefore, we had to experiment with the class weight ratios to obtain marginally better performance.

First, we assessed the performance of the Logistic Regression, this model initially achieved a 75% accuracy. However, after assessing the confusion matrix we gathered that it had completely misclassified all of the positive examples (actual stroke) as negative, a Type 2 error which in a real scenario most likely leads to a fatality depending on the stroke severity. Consequently, we experimented with several class weight ratios, including (1:10 (balancing point), 1:20, 1:30, 1:40) and concluded that after a (1:15) ratio, the False Positive rate (Type 1 error) was getting too high and no further improvements were realized in the False Negative rate (Type 2 error). The chosen class weight ratio provided a recall of 80% to the positive class, and 79% recall to the negative class. Lastly, by assessing the ROC-AUC curve below we can get a better sense of performance for this model. We can see that the Logistic Regression provides acceptable performance for this dataset, most importantly it can achieve over 80% True Positive Rate while maintaining less than 20% False Positive Rate.



In contrast to the Logistic Regression, the Random Forest Classifier achieved an exceptional accuracy score of 95%. However, this highlights the importance of assessing other metrics and was not indicative of good performance in this scenario since it incorrectly classified all of the positive examples as negatives (Type 2 error) which carries the most weight in stroke prediction. After assessing the confusion matrix we determined that the Random Forest is extremely sensitive to unbalanced datasets which was unexpected. Again, we experimented with different class weight ratios however, it did not improve the False Negatives in any regard. Therefore, Random Forest was not a good choice to model this dataset. While this model did not perform well it gave us a better insight and interpretation of the dataset. By assessing the feature importances we can gather that the main contributors to predicting stroke are age, average glucose level and, bmi. On another note, it seems that being married carries significant weight in determining stroke, likely due to the increased stress levels that come with this achievement.



Storytelling and Conclusion (Isaiah)

Our project's core insight was that with severe class imbalance with 95% of the data having no stroke, traditional accuracy is dangerously misleading, as high-scoring models initially

failed to identify any stroke patients. We successfully addressed this by using class weights in Logistic Regression, achieving an 80% recall for stroke cases to minimize fatal false negatives. We learned that model choice is critical, as random forest proved ineffective here despite its strong performance on balanced data. The feature analysis revealed expected factors like age and glucose, plus social insights like marital status. Ultimately, we created a viable screening tool but recognized its role must be supportive, not diagnostic, due to inherent precision-recall trade-offs. This underscored that in healthcare the cost of errors must define the model, not just its score. Future work requires advanced techniques to improve balance and rigorous bias auditing for equitable use. We started wanting to predict stroke and learned that the greater challenge is responsibly predicting *risk* while accounting for the profound imbalance in real-world medical data. The true outcome was not a single mode. This project demonstrated that using a model to predict the likely experience is not good enough and that it lacks the nuance of a real healthcare professional.

Impact (Brandon)

This project can have a meaningful impact, socially and ethically, because it deals with predicting stroke risk. While a well trained model could help identify high risk patients earlier and be used to support preventative care, there could also be risks. Misclassification could cause unnecessary anxiety or stress, or even worse it could overlook someone who needs the medical attention. Bias within a dataset could also lead to unequal predictions across demographic groups as well. Thus this highlights the importance of careful evaluation, transparency, and ensuring that the model is only used to support professional medical judgement, not replace.

References

Fedesoriano. (2021, January 26). *Stroke prediction dataset*. Kaggle. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset .

Code/Github

https://github.com/gabeLin300/Stroke-Prediction