Kylie Stephens, Kieran Purdue, Margaux Reynolds
Professor Johnson
DS 3021
9 May 2025

Final Paper

Abstract

Stroke is one of the leading causes of death and long-term disability worldwide. Early identification of individuals at higher risk for stroke could enable timely medical intervention and prevention. In this project, we explored whether basic demographic and health information can be used to predict stroke risk using machine learning techniques.

We used a publicly available dataset of 4,981 individuals containing 11 features, such as age, BMI, hypertension, glucose level, smoking status, and more, with a binary target variable indicating whether each person had experienced a stroke. A major challenge in this dataset was class imbalance: only about 5% of individuals had a recorded stroke.

After cleaning and encoding the data, we trained three classification models: (1) a decision tree, (2) a random forest using all individuals, and (3) a random forest trained only on individuals aged 35 and older. All models used `class_weight='balanced'` to compensate for class imbalance and were evaluated on a held-out test set.

The random forest trained on the full dataset achieved the highest recall (88%), meaning it correctly flagged most stroke cases. However, precision was low (12.7%), indicating many false positives. The decision tree offered slightly better precision (15.2%) but lower recall (76%). Filtering out individuals under 35 did not improve overall performance.

Our results suggest that while machine learning can identify general risk patterns, such as increased stroke risk with age, high BMI, and hypertension; the models struggled to make precise predictions due to limited and mostly categorical features. These models may be useful as preliminary screening tools but are not sufficient for clinical decision-making on their own. Future work could improve performance by incorporating more detailed medical histories, continuous biomarkers, or imaging data.

Introduction

Strokes are among the most serious and life-threatening medical conditions, ranking as the second leading cause of death worldwide according to the World Health Organization. Each year, strokes account for roughly 11% of all deaths globally, often striking without warning and leaving survivors with significant physical and cognitive disabilities. Early identification of individuals at high risk of stroke can drastically improve outcomes by enabling preventative interventions and timely care. Given the increasing availability of electronic health data, there is a growing interest in using machine learning to support clinical decision-making, including stroke risk prediction.

In this project, we explore the feasibility of using basic health and demographic data to predict stroke risk. Our analysis is based on the publicly available “Brain Stroke” dataset, which includes information on 4,981 individuals across 11 variables such as age, gender, BMI, hypertension, average glucose level, smoking status, and others. The target variable is binary, indicating whether or not a person has experienced a stroke. Our central research was centered around the questions: *What factors might predispose an individual to having a stroke/increase the risk of them having a stroke? What variables are the best predictors of if an individual will have a stroke? Can we build a predictive model that identifies individuals at higher risk of stroke based on these variables?*

We begin with a thorough exploratory data analysis (EDA) to better understand the distribution of each feature and its potential relationship with stroke. For example, we found that over 70% of stroke cases occurred in individuals over the age of 60, and that 45% of individuals who had a stroke had a BMI over 30, classifying them as clinically obese. These initial findings confirmed the importance of age and BMI as potential predictors and guided our modeling choices.

One major challenge in our analysis was the significant class imbalance in the dataset: only about 5% of individuals experienced a stroke. This poses difficulties for standard classification metrics like accuracy, which can be misleading when one class dominates. As a result, we prioritized more informative metrics like **recall** (how many actual stroke cases were correctly identified) and **precision** (how many predicted stroke cases were accurate).

To prepare the dataset for modeling, we performed several preprocessing steps. We dropped two categorical variables—ever_married and work_type—due to their low variation in stroke rates, and one-hot encoded others like gender and residence_type to make them compatible with machine learning algorithms. We also experimented with filtering out individuals under the age of 35 to assess whether removing extremely low-risk cases would improve model performance. Our initial modeling approach was to use weighted logistic regression, kNN, and tree classification and compare the performances of each model to each other; however, we ultimately transitioned to only using trees as kNN works best with continuous numeric data and several of the key clinical features were categorical variables. Additionally, while the class imbalance (low incidence of stroke in the dataset) was addressed through logistic regression, this approach often entails over-compensation by creating patterns and trends for the model to find that are not actually meaningful or reflective of reality. This is especially dangerous when creating models that could have real-world, clinical applications.

This led us to focus our modeling approach on tree-based classifiers due to their interpretability and ability to handle non-linear relationships. While our original plan involved testing a k-Nearest Neighbors (kNN) classifier, we found that the categorical nature of many features made kNN less suitable. Instead, we trained three models:

1. A **decision tree classifier** (max depth = 3) using the full dataset.

2. A **random forest classifier** (max depth = 3) using the full dataset.

3. A **random forest classifier** (max depth = 3) using only individuals aged 35 and older.

All models were trained using class_weight='balanced' to address the class imbalance. We split the data into an 80% training set and a 20% test set using stratified sampling to maintain the stroke/no-stroke ratio. The models were evaluated on the test set using accuracy, precision, and recall.

Our results showed that the **random forest classifier trained on all individuals** achieved the highest recall (88%) but had low precision (12.7%), indicating that while it successfully identified most stroke cases, many of its positive predictions were incorrect. The **decision tree** offered better precision (15.2%) with slightly lower recall (76%). Filtering out individuals under 35 did not improve performance, suggesting that even low-risk individuals can provide useful training signals.

These findings underscore a key insight: predicting stroke from a limited number of basic clinical features is possible, but difficult. The models were able to detect broad patterns, such as increased risk with age, hypertension, and obesity but struggled to make highly precise predictions, likely due to the absence of richer medical history or diagnostic information.

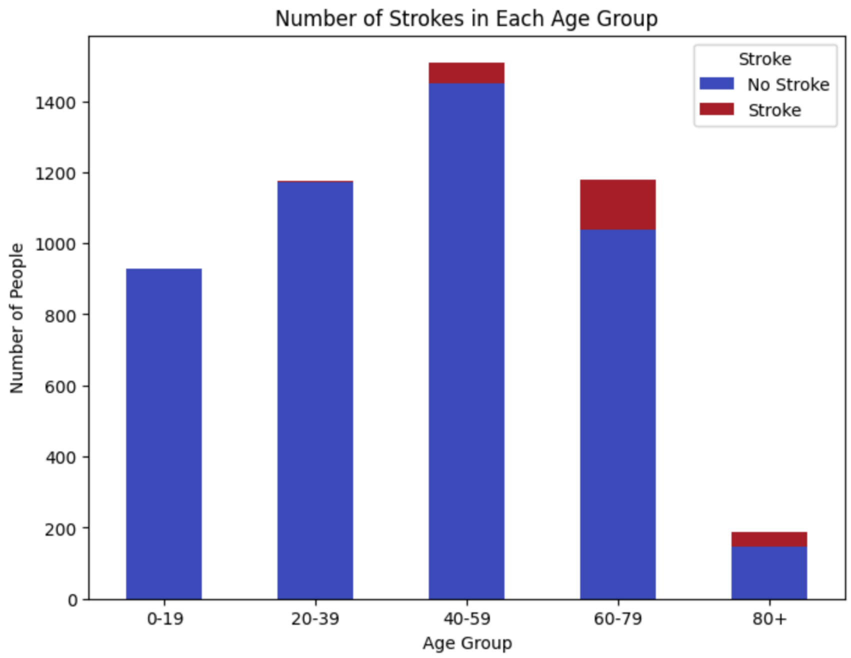
In real-world applications, such a model could serve as an initial screening tool, highlighting high-risk individuals who may benefit from further testing. However, the limitations of the dataset and the modest precision of the models mean they should be used in conjunction with, not as a replacement for, professional clinical evaluation.

In the pages that follow, we detail our data preprocessing steps, modeling decisions, evaluation metrics, and results. We also reflect on the limitations of our work and propose directions for future research, including the need for more informative features and alternative modeling strategies to better handle rare medical outcomes.

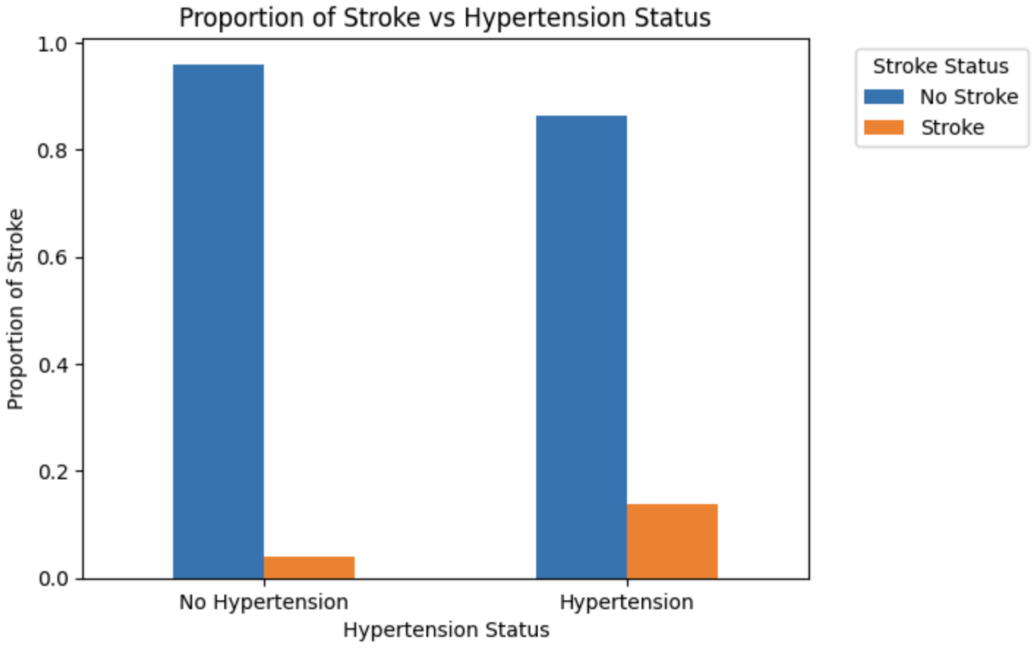
Data

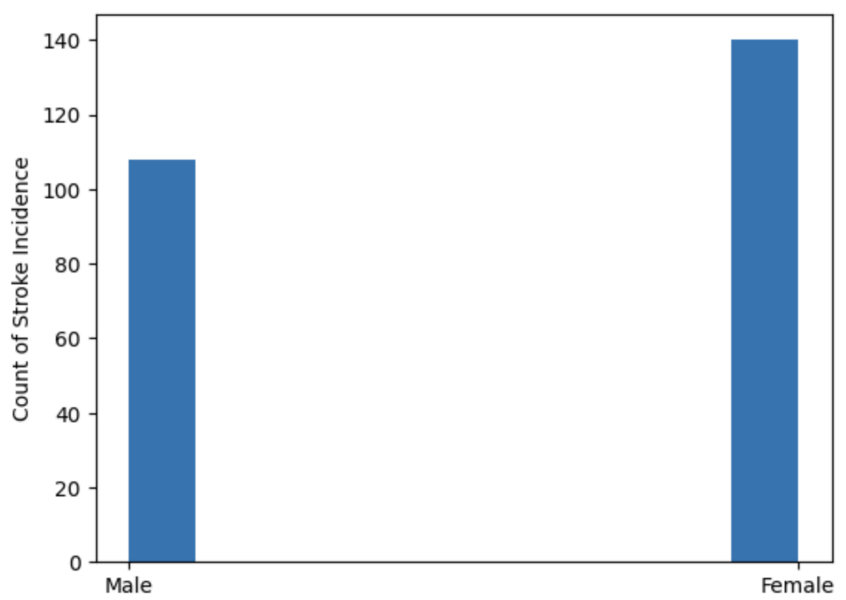
This project uses data from the “Brain Stroke” dataset, which includes demographic and health information for 4,981 individuals. Each observation represents a single person, described by attributes such as age, gender, BMI, smoking status, residence type, and whether they experienced a stroke. The target variable, `stroke`, is binary (1 = stroke, 0 = no stroke).

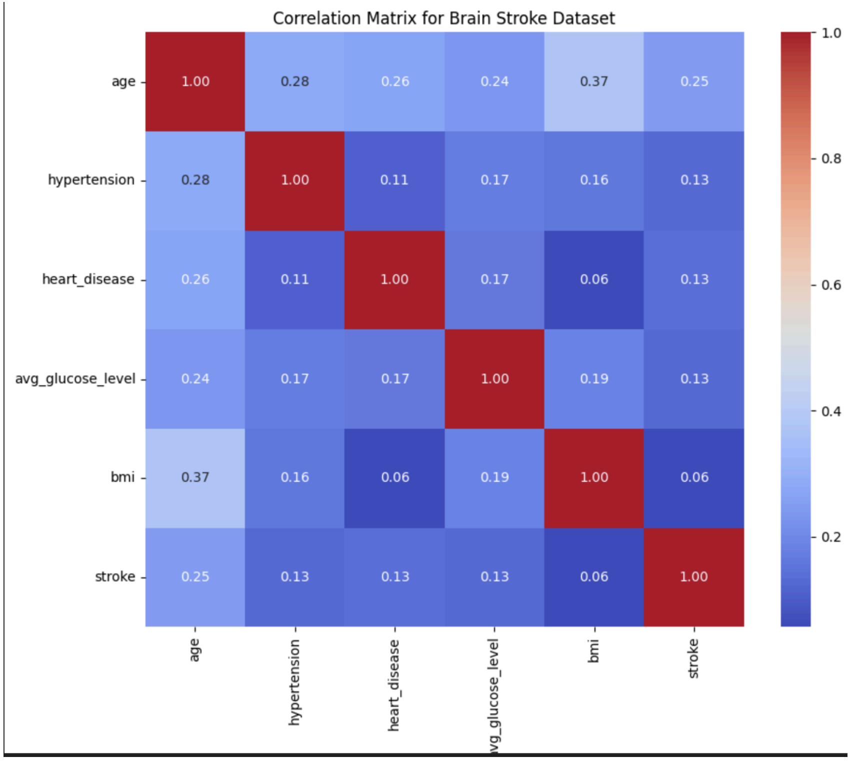
As part of the exploratory data analysis section of this project, we created some basic distributions and plots between variables. We also calculated the populations of people that had higher proportions of people that experienced a stroke.



Percentage of People that had a stroke that were over the age of 60: 72.98387096774194 %
Percentage of People that had a stroke that were under the age of 20: 0.8064516129032258 %







This correlation matrix demonstrates a lack of strong correlation between most variables and stroke. Age and stroke have the highest correlation of .25. This suggests the importance of interaction between variables.

One major challenge in the dataset is class imbalance. Only around 5% of individuals experienced a stroke, while the remaining 95% did not. This imbalance has important implications for model evaluation, as metrics like accuracy can be misleading when one class dominates the data.

Before modeling, we made several changes to improve the structure and predictive power of the dataset. Specifically, a few variables were removed due to limited predictive power. We dropped `ever_married` and `work_type` after exploratory data analysis showed little variation in stroke rate across these categories. Categorical variables such as `gender` and `residence_type` were one-hot encoded to make them usable in scikit-learn models. `Gender` was encoded with male = 1, female = 0, and `residence_type` with urban = 1, and rural = 0.

In some of our analysis, we also filtered out individuals under age 35. This subgroup had a very low stroke rate (less than 0.1%), and we wanted to explore whether excluding them might improve model performance. However, we did not remove them entirely from the dataset. Instead, we compared models trained on all individuals versus models trained only on individuals aged 35 or older to assess the effect of age-based filtering.

To better understand the relationships between variables and stroke risk, we performed exploratory data analysis. One key insight was that stroke risk increases with age. Individuals over 60 showed a higher rate of stroke, consistent with real-world medical knowledge. We also found that men had a higher stroke rate across all age groups. Notably, 45% of individuals who experienced a stroke had a BMI over 30, the clinical threshold for obesity, suggesting that BMI may be an important risk factor.

In summary, our cleaned dataset consists of individuals with complete information and a focus on interpretable health and demographic predictors. We accounted for class imbalance and evaluated models on both the full dataset and a filtered older population to better understand the dynamics of stroke prediction. Additionally, we found that a combination variables was best at predicting stroke incidence as a complex, often unforeseen medical event is not easily predicted by one clinical feature, especially the 11 clinical features in this data set that did not divulge medical history or other factors that might typically be included when evaluating a patient for stroke risk.

Methods

Each observation in this study represents an individual patient or respondent, described by features such as gender, age, hypertension status, heart disease status, average glucose level, body mass index (BMI), smoking status, and whether or not the individual has experienced a stroke. The goal of the project was to build a predictive model that could estimate stroke risk based on these demographic and health-related characteristics.

This is a supervised classification problem; we aim to predict a binary outcome (stroke = 1 or stroke = 0) using a set of known features. Because of the medical context, our prediction is framed less as a binary “yes/no” and more as a risk estimation problem. In practice, we interpret the model’s output as helping to identify higher-risk individuals who may benefit from further screening or intervention.

While our original plan was to use a k-Nearest Neighbors (kNN) classifier to categorize individuals into low, moderate, and high stroke risk, we ultimately shifted our focus to decision trees and random forests.

This shift was motivated by the presence of categorical variables, the importance of model interpretability, and the non-linear relationships between predictors.

We trained three classification models:

1. A decision tree (max depth = 3) using the full dataset
2. A random forest (max depth = 3) using the full dataset
3. A random forest (max depth = 3) using only individuals aged 35 and older

This setup allowed us to compare simple models and to explore how excluding younger, lower-risk individuals affected prediction quality.

One anticipated challenge was the large number of categorical variables, which can complicate modeling if not encoded properly. We addressed this through one-hot encoding and variable selection. Another issue was the extremely imbalanced target variable, which made evaluation metrics like accuracy slightly unreliable and required rethinking which performance metrics we used.

To account for class imbalance (only ~5% of individuals had a stroke), we applied `class_weight=’balanced’` in all models, ensuring the minority class (stroke cases) received proportionally more weight during training.

As discussed above, we prepared the dataset by first dropping two categorical variables (`ever_married` and `work_type`), one-hot encoded binary categorical variables (`gender` and `Residence_type`), created a subset of the data filtering out individuals under 35, and verified that no missing values were present.

Because of the class imbalance, we did not rely on accuracy as our main evaluation metric but still included it in our final results. A model could achieve over 90% accuracy simply by predicting that no one had a stroke. We also used the following metrics: recall/sensitivity (proportion of actual stroke cases that were correctly identified) and precision (proportion of predicted stroke cases that were correct).

We split the data into 80% training and 20% testing using stratified sampling to maintain the proportion of stroke cases. Performance metrics were calculated on the test set only.

Our goal was not just to build a model that works well, but to understand the tradeoffs between false positives and false negatives in a health screening context, where missing a stroke case can have serious consequences, but over-predicting can lead to unnecessary worry and cost.

The major weakness in this dataset lies in the fact that there are only 248 stroke incidents in the dataset containing almost 5,000 individuals. Also, many distributions for variables compared between those who had a stroke and did not do not differ greatly or point to significant correlations. Low incidence might have caused issues with the precision and recall of the model. It did not have a lot to learn from. Most of the clinical features are categorical. They need to be continuous for kNN. Using a dataset with more informative clinical features (not mainly categorical) would be better for future model training. Additionally, accuracy will most likely be negatively impacted when weighted logistic regression is used because it focuses more on the minority event. This effectively will increase recall, but will also result in more false predictions of stroke. However, in a medical setting with a condition that is life or death (ie stroke), it is probably better to have false positives than false negatives.

Results

Our central prediction question was: Which individuals are at highest risk of experiencing a stroke based on demographic and health information? The models we developed could hypothetically be used in a hospital setting as a preliminary screening tool to flag high-risk individuals for further evaluation.

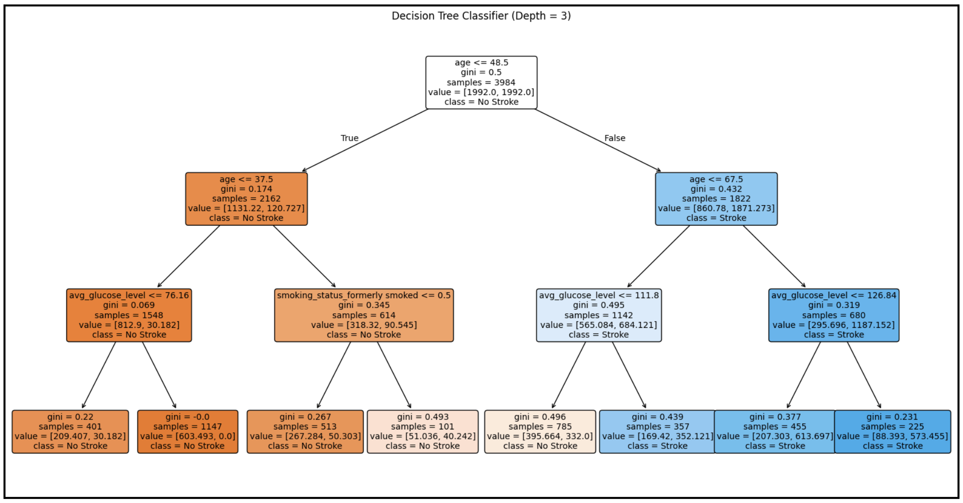
To answer this question, we trained and evaluated three classification models:

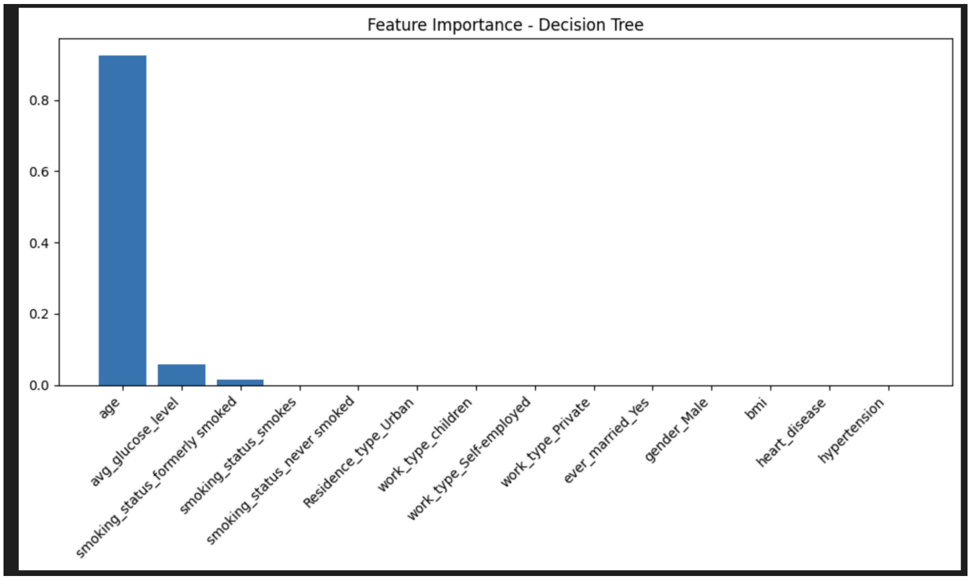
1. A decision tree trained on all individuals
2. A random forest trained on all individuals
3. A random forest trained only on individuals aged 35 and older

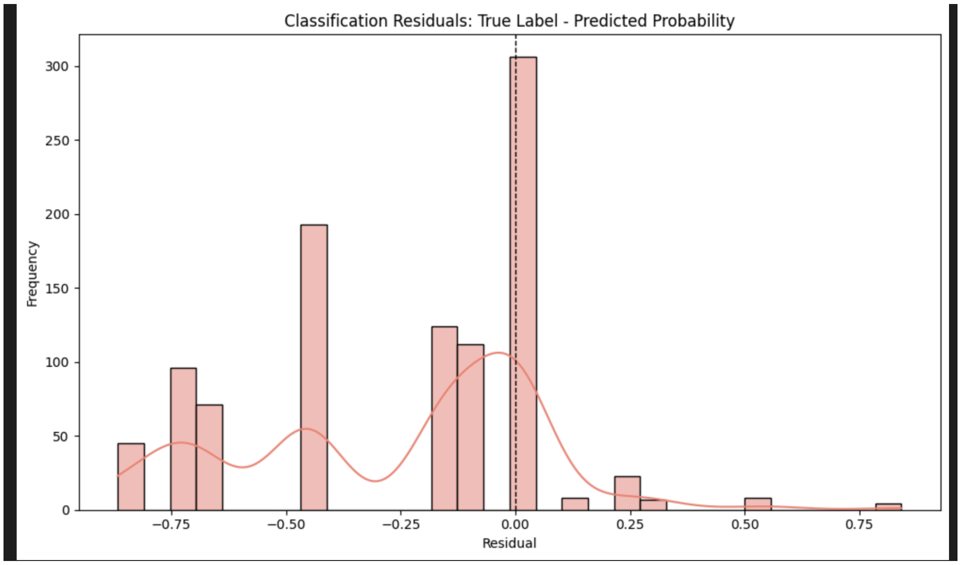
Decision Tree (All Individuals)

Metric	Value
Accuracy	77.5%
Precision	15.2%
Recall	76.0%

The decision tree model provides a clear and interpretable baseline. While it achieved high accuracy, this is largely due to correctly classifying the majority “no stroke” class. The more meaningful result is its 76% recall, indicating that it correctly flagged over three-quarters of stroke cases. Its precision was 15.2%, reflecting a significant number of false positives. Still, this tradeoff may be acceptable in a screening context where catching as many true stroke cases as possible is more important than avoiding false alarms.



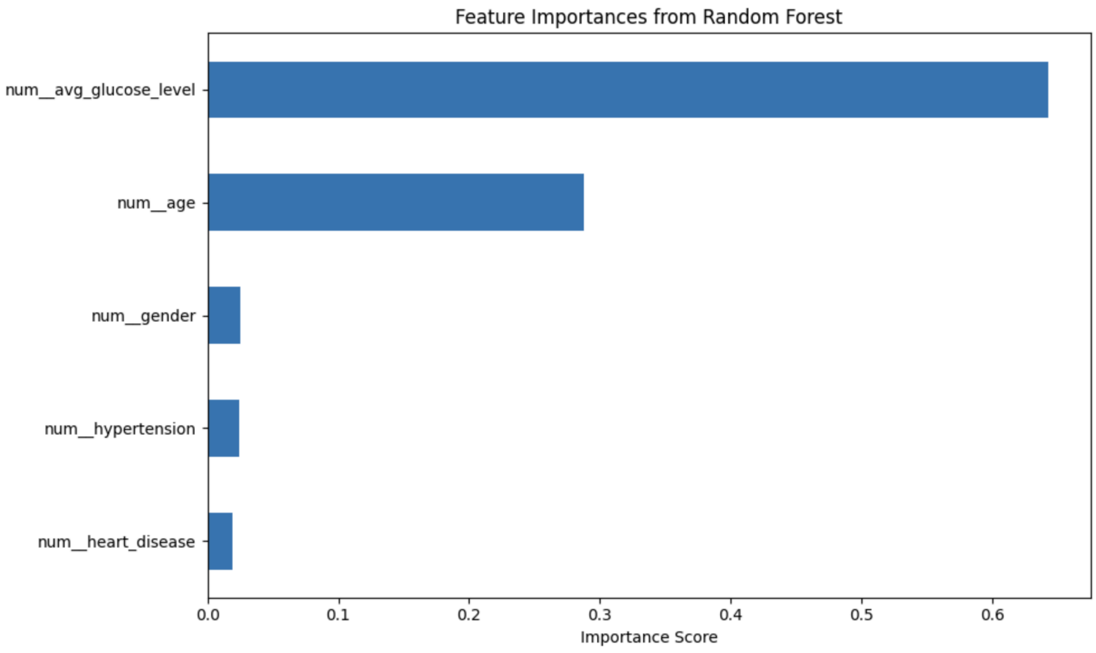




Random Forest (All Individuals)

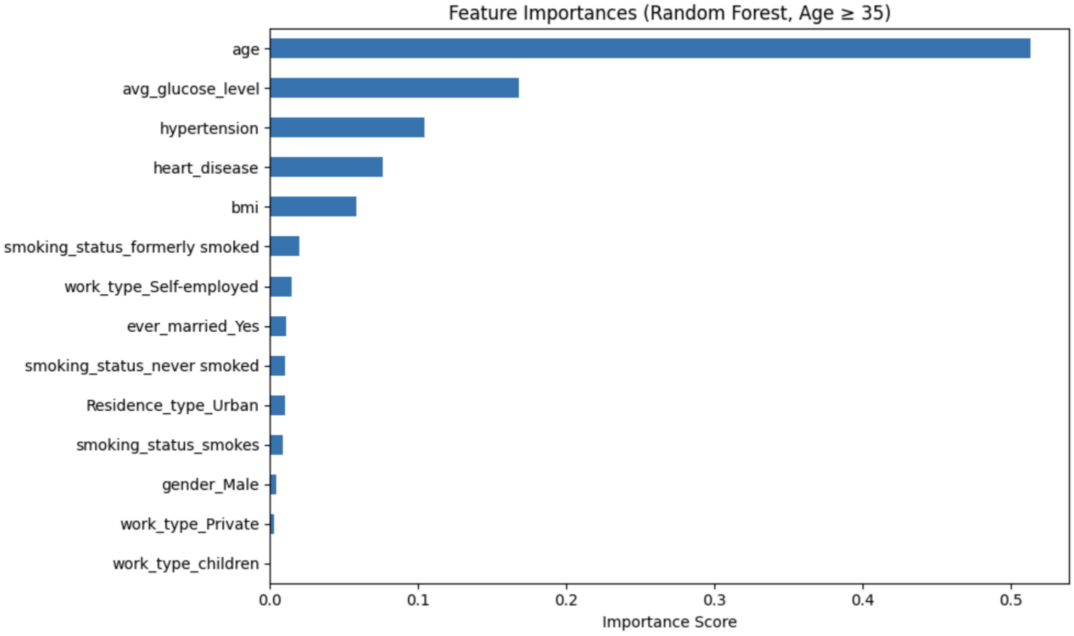
Metric	Value
Accuracy	69.0%
Precision	12.7%
Recall	88.0%

The random forest trained on the full dataset achieved the highest recall of all models at 88%, correctly identifying nearly 9 out of 10 stroke cases. However, precision fell to 12.7%, meaning that most of the stroke predictions were incorrect. In contexts like public health screening, this tradeoff may not be horrible, since it is better to over-warn than to miss high-risk signs. The drop in accuracy is expected when prioritizing recall.



Random Forest (Ages 35 and Older Only)

Metric	Value
Accuracy	66.5%
Precision	15.5%
Recall	75.5%



When we filtered out individuals under 35, the model's recall decreased and precision increased slightly, suggesting that while stroke is rare in younger people, including them may help the model identify important patterns. Overall, this version of the model underperformed the random forest trained on all ages, indicating that age-based filtering did not lead to stronger predictions.

Among the three models, the random forest using all individuals had the highest recall, making it best for flagging at-risk patients. The decision tree performed nearly as well, while offering clearer interpretability. The age-filtered model did not improve results and actually missed more stroke cases.

These results show that tree-based models, even with limited input data, can reveal useful patterns in stroke risk. However, the low precision across all models highlights the limitations of working with imbalanced datasets with limited features.

Conclusion

In this project, we set out to explore stroke prediction using a dataset of nearly 5,000 individuals, each described by a set of demographic and clinical features. Our goal was to determine whether it is possible to flag high-risk individuals using readily available health data. Given the real-world importance of early stroke detection, this project blends both statistical modeling and public health relevance.

Our primary finding is that while basic machine learning models like decision trees and random forests can achieve relatively high recall, they struggle with precision. For instance, the best-performing model, a random forest trained on the full dataset, correctly identified 88% of actual stroke cases, but only 12.7% of its positive predictions were accurate. In contrast, the decision tree offered slightly better precision (15.2%) with marginally lower recall (76%). These results reflect a consistent challenge: imbalanced data and limited predictive information.

Despite these limitations, our results still suggest that demographic and basic clinical features can offer value in preliminary screening. High-recall models like our random forest could be used to identify individuals who merit further diagnostic evaluation, especially in resource-constrained settings. That said, these tools should never serve as a substitute for clinical judgment or more thorough medical tests.

We also looked into whether removing individuals under age 35 (population with a very low stroke rate)could improve model performance. Surprisingly, excluding these individuals led to decreased recall and only marginal gains in precision, suggesting that younger participants may still provide useful variation that helps the model learn more generalizable patterns.

Below are several limitations of our dataset and modeling process.

- *Limited Feature Set*: The dataset lacked more granular clinical variables such as blood pressure readings, cholesterol levels, family medical history, and medication use. Including richer clinical data could greatly enhance predictive power. A major factor to include would be stress level, which is known to be somewhat indicative of stroke incidence for a patient.

- *Imbalanced Classes*: With only approximately 5% of individuals experiencing a stroke, our models were trained on highly imbalanced data (unequal classes). While we used techniques like class weighting to mitigate this, future work could incorporate resampling methods or anomaly detection frameworks specifically designed for rare event modeling. Weighted logistic regression modeling was not the best solution to address this problem as it essentially fabricates data for the sake of pattern/trend recognition.

- *Categorical Dominance*: Many features were categorical or binary, limiting the utility of distance-based models like k-Nearest Neighbors and reducing the resolution of certain relationships. A dataset with more continuous variables would enable more sophisticated modeling approaches. This would entail more blood tests or variables that can be quantified (potentially, even ranking severity of heart disease or hypertension on a scale could deepen insights found/made by the model).

- *Context-Specific Modeling*: Our analysis assumed all individuals are equally at risk without considering geographic, cultural, or healthcare-access differences. A natural next step would be to train models on region-specific subsets or to include social determinants of health as additional features. This once again represents the shallowness of the clinical features within this dataset.

- *Precision vs. Recall Tradeoff*: While we favored recall in a medical context (to avoid missing true stroke cases), the cost of false positives must be addressed. Future models could include cost-sensitive learning or explore threshold-tuning techniques that optimize a balance based on practical constraints (e.g., hospital screening capacity).

- *Model Interpretability*: Although we leaned toward interpretable models like decision trees and shallow forests, future work might explore more complex models like gradient-boosted trees or neural networks, which could improve performance while retaining transparency. These complex models were not suitable for our relatively small dataset, though.

The limitations of our project and model highlighted the importance of having enough data to train the model on, the effect of class imbalance on model prediction, and the strengths and weaknesses of the various model building approaches we have learned this semester. The scarcity of positive cases emphasized the importance of having a sufficiently large and representative dataset, especially for rare but critical outcomes. Second, the severe class imbalance skewed the performance metrics of our models, particularly accuracy, which appeared deceptively high due to the dominance of the majority class. This forced us to think more critically about which metrics (such as recall and precision) were most appropriate for evaluating model performance in an imbalanced classification problem. Finally, this project allowed us to compare different model-building strategies we encountered during the semester, such as decision trees, random forests, and k-nearest neighbors, and better understand the tradeoffs each method brings.

Ultimately, this project highlights the potential promise and the current limits of using basic demographic and clinical data for stroke prediction. While our models performed reasonably well in recall, more robust, in-depth datasets and advanced modeling techniques are needed to reduce false positives and improve legitimate, actionable insights. In the future with further improvement and fine tuning, tools like this could become part of a larger digital health infrastructure aimed at early detection and prevention of strokes and or other medical conditions/events.

References/Bibliography

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

Hassan, A., Ahmad, S. G., Munir, E. U., Khan, I. A., & Ramzan, N. (2024). Predictive modelling and identification of key risk factors for stroke using machine learning. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-61665-4