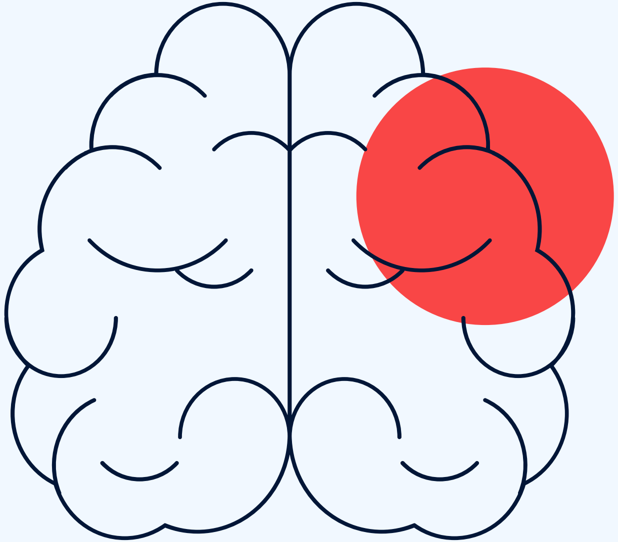
Stroke Prediction

By: Daphne O'Malley, Maxamed Jama, Zoe Cheng, Ashay Srivastava, Allen Liu, and Shemar Anglin

INST 452 | 13 May 2025



Research Question

Can we predict whether a patient will have a stroke (Yes or No) based on their clinical and lifestyle information?

Yes (1) → The patient is likely to have a stroke | No (0) → The patient is not likely to have a stroke

Background/Importance of the Research

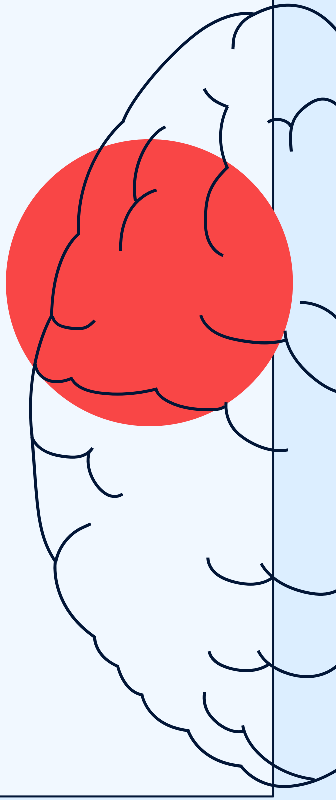
- **Why is this health issue important?**
  - Stroke is the second leading cause of death in the world.
  - It is the third leading cause of death and disability combined (calculated by disability-adjusted life-years lost).

    Source: [World Stroke Organization (2022)]

- **Who might benefit from this kind of prediction model.**
  - <u>Healthcare Providers</u>: WIth early detection methods, healthcare professionals can intervene quickly for patients who are high-risk.
  - <u>Patients</u>: Assessments of patients and their risk behaviors can help to make health and lifestyle decisions to reduce the risk of stroke.

- **Statistics of real-world use cases.**
  - In the United States:
    - <u>Every 40 seconds</u> someone experiences a stroke.
    - <u>Every 3 minutes and 11 seconds</u>, someone dies of a stroke.
    - <u>Annually, at least 795,000</u> people have a stroke.

    Source: [Center for Disease Control and Prevention]



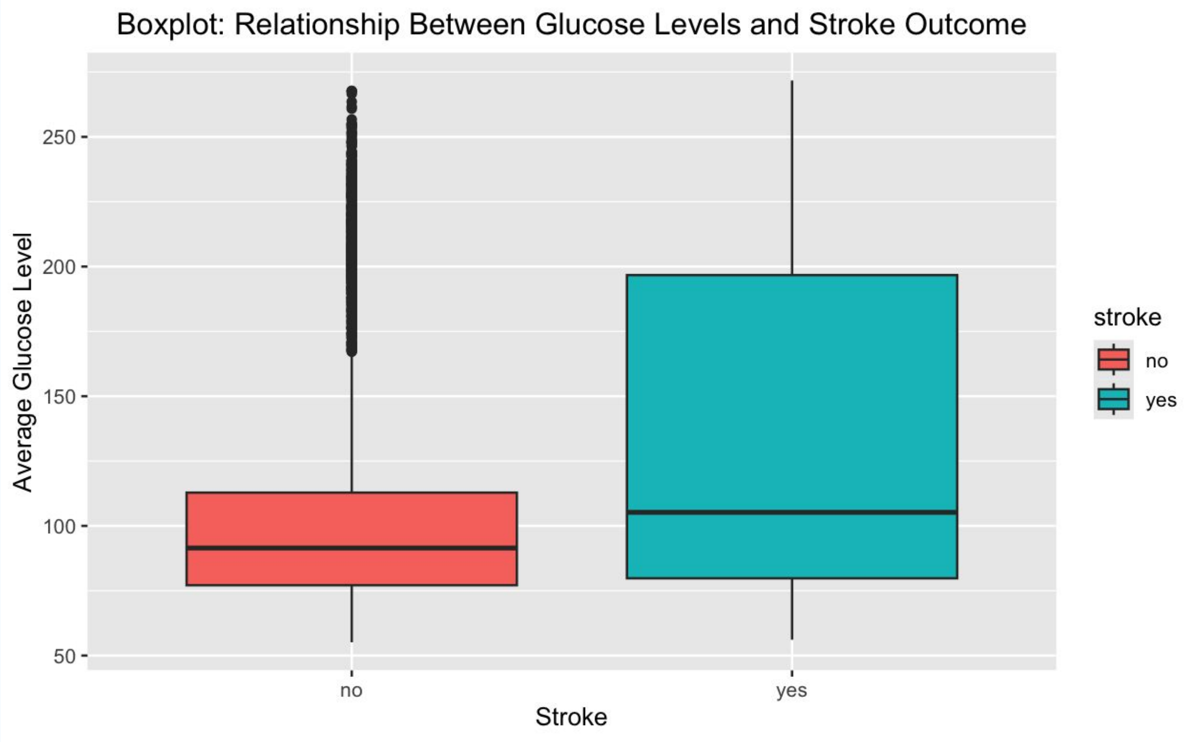
Data Cleaning & Preprocessing

Source: Stroke Prediction Dataset (strokepredictiondataset) - Kaggle
There are 5110 observations & 12 variables

Data preprocessing steps taken:

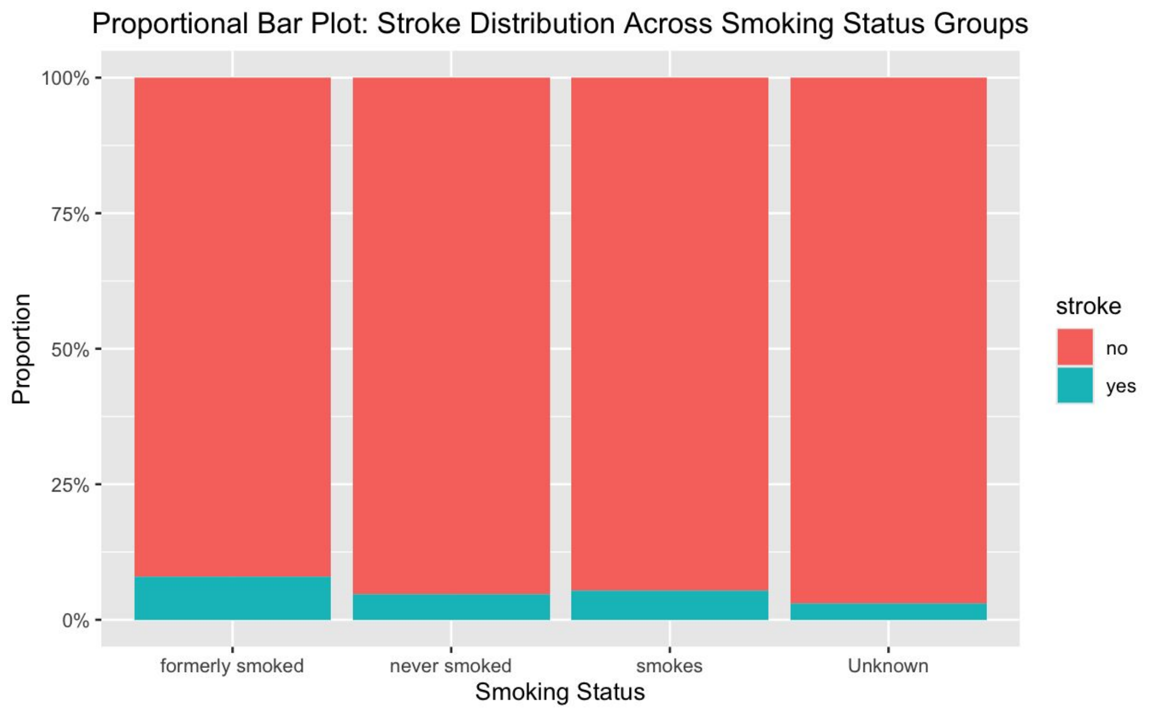
1. **Import & Format**
   - Loaded the stroke data set, using read.csv()
   - Converted the "stroke" variable into a factor, labeled: "yes" and "no"
2. **Addressed Missing Data**
   - Replaced "N/A" with NA; as recognized in R
   - Replace missing values in "bmi" column with the median
3. **Summary**
   - Ran "summary()" to analyze the variable distributions
4. **Encode Categorical Data**
   - One-hot encoded: gender, ever_married, work_type, residence_type, smoking_status
   - Used model.matrix() to perform a loop for converting each category to numeric
   - Dropped original columns
5. **Final Check**
   - Verify that all columns are ready for modeling
   - Confirmed no missing values in the encoded dataframe

Data Visualization 1 - Glucose vs. Stroke



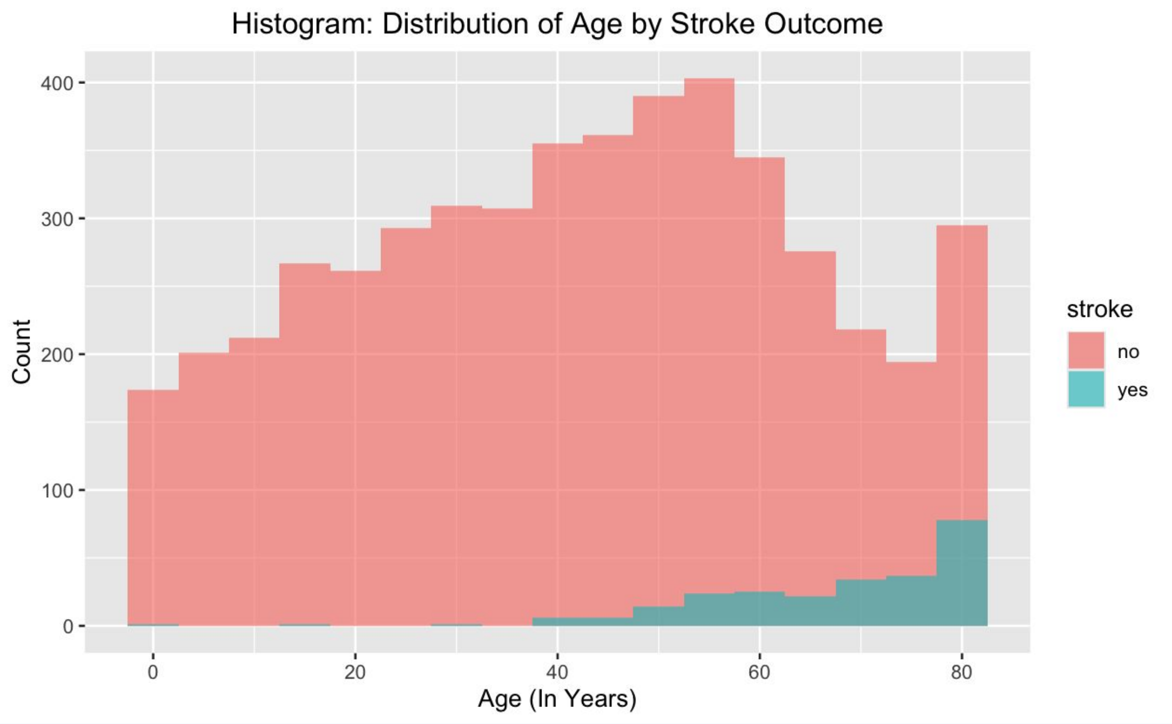
- Stroke patients have a wider and higher range of glucose levels.
- The outliers for non-stroke patients suggest that some patients till had very high glucose levels.
- This plot suggests that glucose levels are a strong clinical predictor for stroke modeling.

Data Visualization 2 - Stroke by Smoking



- The stroke rate is highest for those who currently smoke and for those who used to smoke.
- Patients in “never smoked” have the lowest proportion of strokes (excluding ‘unknown’ group).
- This plot shows the relevance of smoking status/behavior as a risk factor.

Data Visualization 3 - Age by Stroke



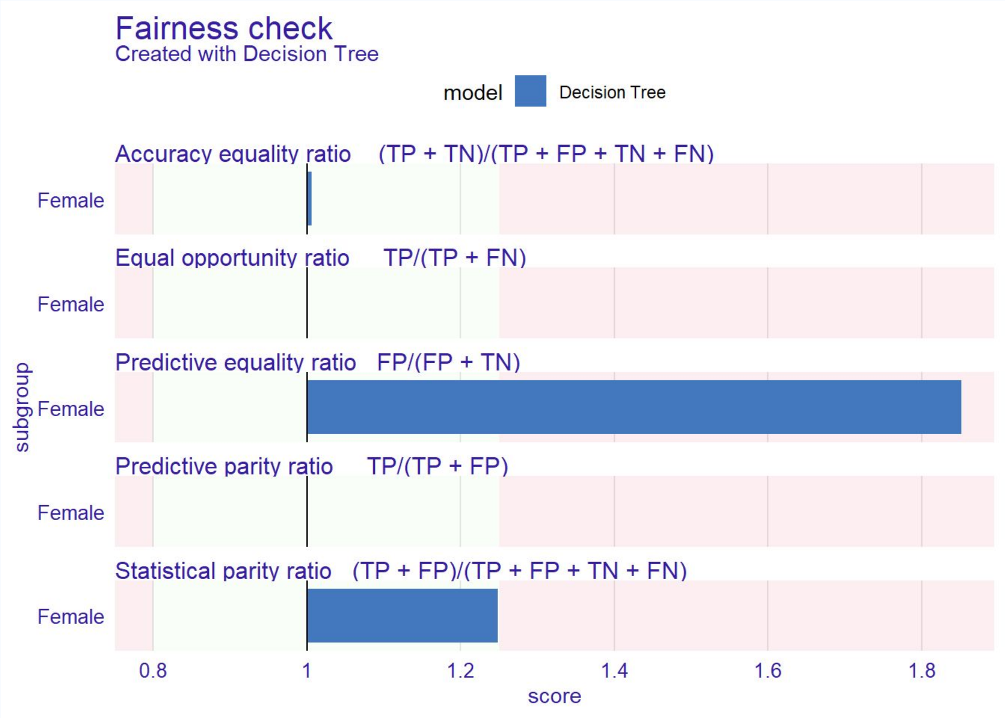
- There is a significant increase in stroke cases in patients aged 50+.
- Younger age groups show little to no stroke occurrences.
- This plot shows that age is a major contributor to predicting stroke events.

Cross-Validation Model Performance

```
Models: DecisionTree, SVM, kNN
Number of resamples: 5

ROC
                  Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
DecisionTree 0.5000000 0.6803841 0.7091925 0.6886105 0.7374377 0.8160383    0
SVM          0.5691647 0.6641086 0.6641827 0.6491640 0.6643836 0.6839803    0
kNN          0.5939087 0.6206381 0.6322468 0.6392175 0.6540985 0.6951952    0

Sens
                  Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
DecisionTree 0.9821674 0.9863014 0.9903978 0.9909503 0.9958848 1.0000000    0
SVM          0.9972603 0.9986283 1.0000000 0.9991777 1.0000000 1.0000000    0
kNN          0.9931413 0.9945130 0.9945130 0.9950629 0.9958904 0.9972565    0

Spec
             Min.    1st Qu.     Median       Mean    3rd Qu.       Max. NA's
DecisionTree    0 0.02631579 0.02702703 0.03726885 0.05405405 0.07894737    0
SVM             0 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000    0
kNN             0 0.00000000 0.02631579 0.02147937 0.02702703 0.05405405    0
```

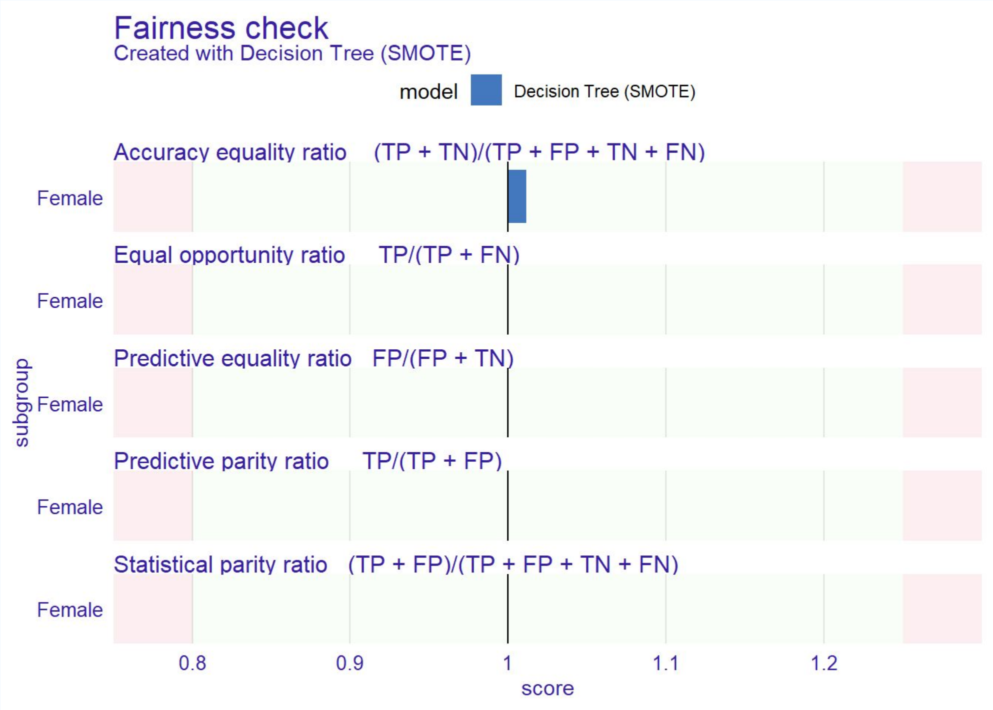
- The SVM and k-NN models have very high sensitivity and very low specificity indicating a high case of false alarms.
- The Decision Tree model has the highest ROC score (0.69) as well as a good balance between sensitivity and specificity.
- The Decision Tree is the best model.

Fairness and Bias Evaluation



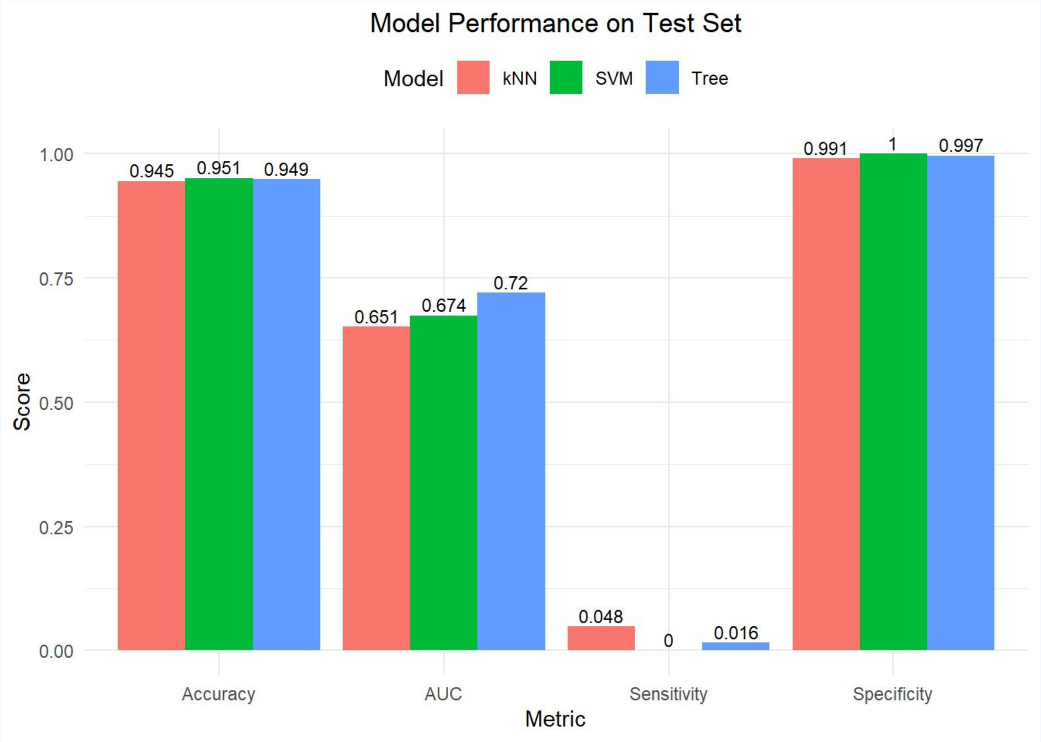
- Bias was detected for the female group. With high disparity for the Predictive Equality Ratio and the Statistical Parity Ratio
- This means that the model had a high chance of giving false diagnosis among females, which is very riky from a clinical standpoint.

Fairness and Bias Evaluation



- SMOTE was applies to balance the stroke classes within the training data.
- After SMOTE was used, the ratios of fairness came closer to 2, which showed a significant reduce in bias for gender.
- SMOTE was the best option because of its low sensitivity AND class imbalance, as described in the course.

Test Set Evaluation



- Each model shows high accuracy, which means that predicted stroke outcomes are overall reliable.
- Decision Tree had the highest ROC, meaning that this model was best at making distinctions between true positive and false positive rates for distinguishing stroke vs. no stroke patients.
- Sensitivity was extremely low, which meant that the models had trouble identifying actual stroke cases for patients.

Metric Selection

Chosen Metric: Sensitivity

- Sensitivity measures the models' ability to identify true stroke cases for patients.

Why Sensitivity Matters?

- Clinically speaking, if healthcare providers were to miss true stroke cases (false negatives), this can lead to a lack of proper intervention for patients who are at risk. This can lead to severe disability or even death.

Real-World Implications

- Medical professionals are concerned more with mitigating stroke, rather than dealing with it after the the event occurs.
- Sensitivity ensures that they are able to catch as many high-risk patients that they can, regardless of them being false alarms.
- For stroke care, its better for patients and medical professionals to over identify the risk of a stroke than missing a potential stroke.

Recommendation to Clinicians

Recommendation: Due to the low sensitivity of each model, none of them should be used clinically.

Reasoning: Models that failed to identify stroke cases correctly would be dangerous to use in real-world clinical environments. This risks not alerting the appropriate people when urgent care regarding stroke is needed.

Possible Improvements:

- Improve models for better detection of stroke cases.
- Try alternative models to improve performance beyond kNN and/or SVM models.
- Trying different inputs for parameters like k for kNN or cost in SVM.
- Reassess any data preprocessing steps.