Predicting Potential Strokes with Machine Learning Algorithms

By Jacob Kuhn

GitHub Repo: https://github.com/Jacob-Kuhn/SupervisedLearningFinal/tree/main

Problem

Can machine learning techniques help doctors predict the likelihood of an individual having a stroke?

Motivation

- Is the US, 795,000 people have strokes every year.[2]
- Every year, 610,000 new people have first time strokes in the US.[2]
- According to the WHO, strokes are the second leading cause of death in the world. [3]

The Data

Collected from Kaggle author Kedesoriano titled Stroke Prediction Dataset.[1]

Features

5110 Records - 12 Features total:

- Id
- Gender
- Age
- Hypertension
- Heart_Disease
- Ever_Married
- Work_Type
- Residence_Type
- Avg_Glucose_Level
- BMI
- Smoking_Status
- Stroke – outcome variable

Outcome Variable - Categorical

Stroke:

249 Positive Cases: (stroke = “1”)

4861 Negative Cases: (stroke = “0”)

Baseline Accuracy:

95.2% -- Therefore, sensitivity is key.

Machine Learning Approach & Methods

K-Nearest-Neighbors

Strategies used

- K value hyperparameter tuning with 5-fold cross validation
- Synthetic Minority Oversampling Technique (SMOTE)
- Train/Test stratified split

Started with K=5 and no oversampling.

Iterated to include oversampling and Cross-Validation to increase sensitivity.

Random Forest

Strategies used

- Hyperparameter tuning – n_estimators, max_depth, min_simples_split, min_samples_leaf, max_features.
- Synthetic Minority Oversampling Technique (SMOTE)
- GridSearch with 5-fold cross-validation and scorer of sensitivity
- Train/Test stratified split

Neural Network

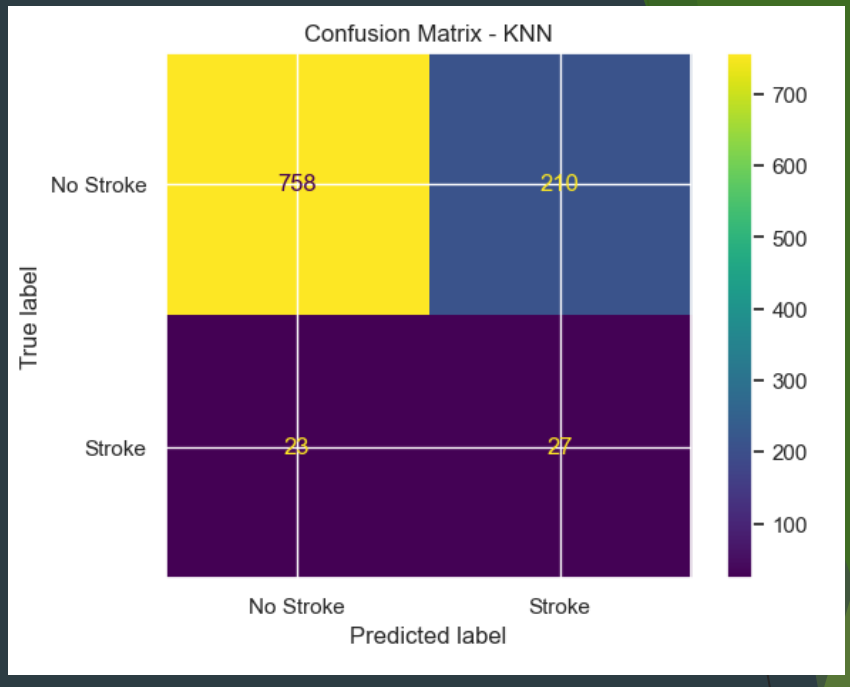
Strategies used

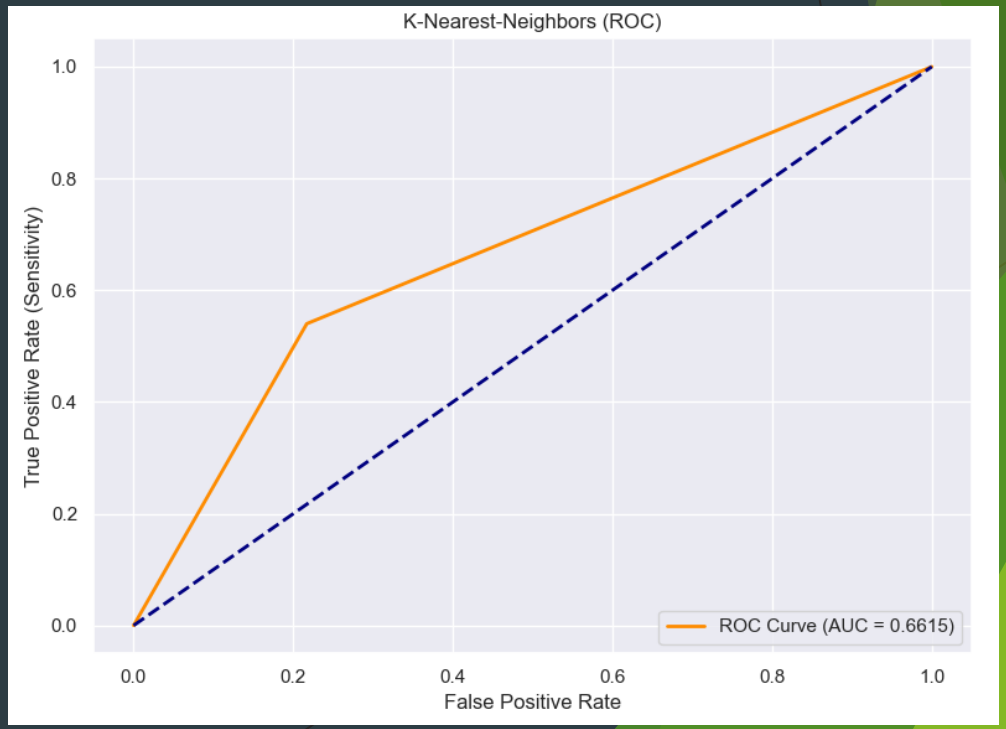
- Multiple layers: 64 nodes, 32 nodes, 1 outcome node
- Dropout layers to keep from overfitting
- Early stopping to end epochs if improvements flatten
- Train/Test stratified split

Results

K-Nearest-Neighbors

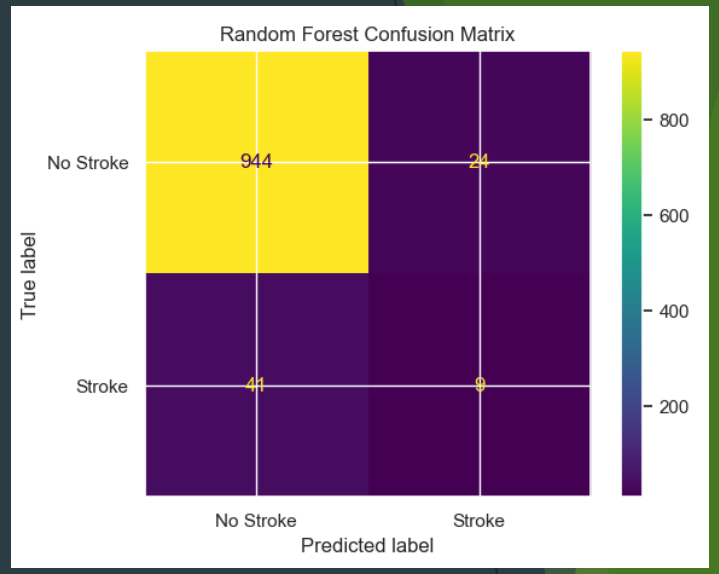
- Sensitivity 54%
- No better than a random guess

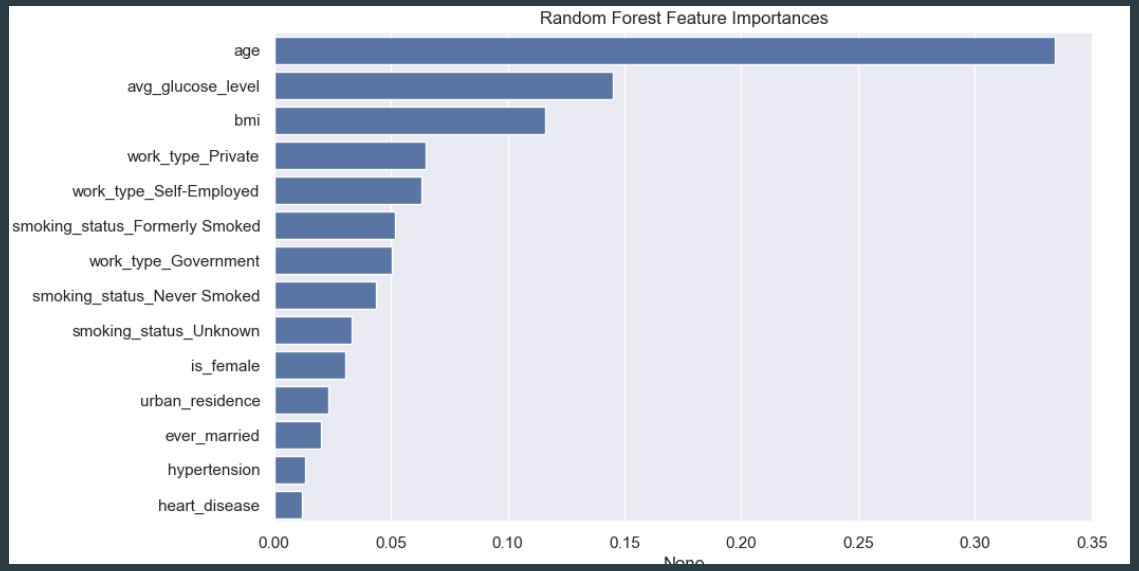


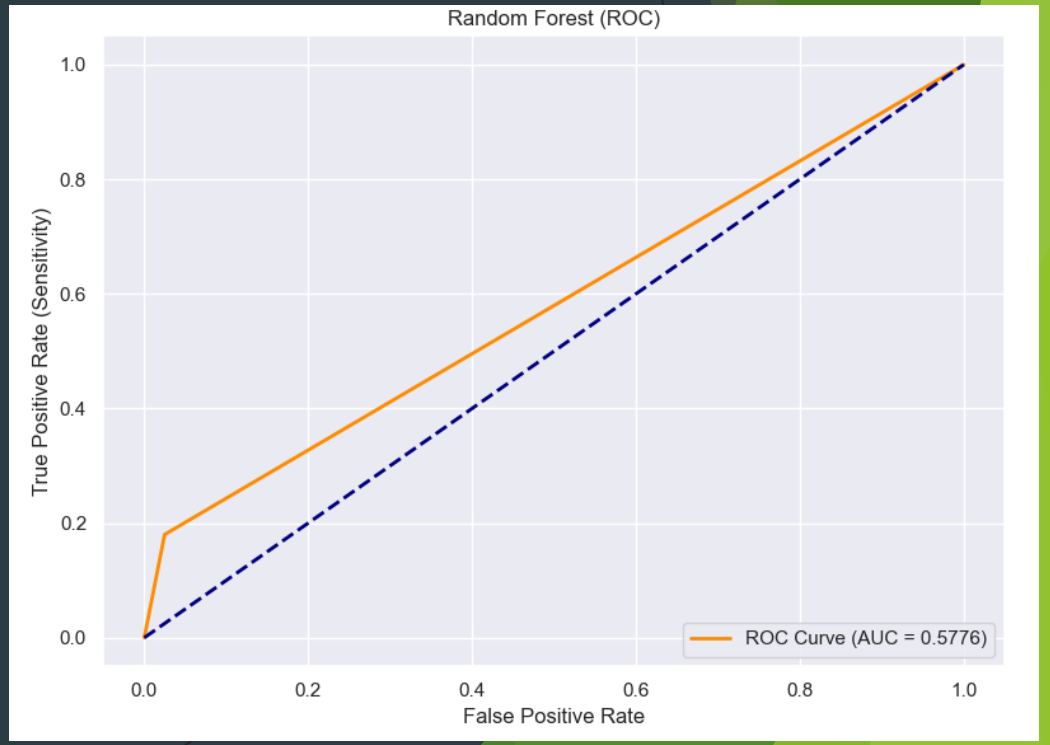


Random Forest

- Sensitivity 18%
- No better than a random guess

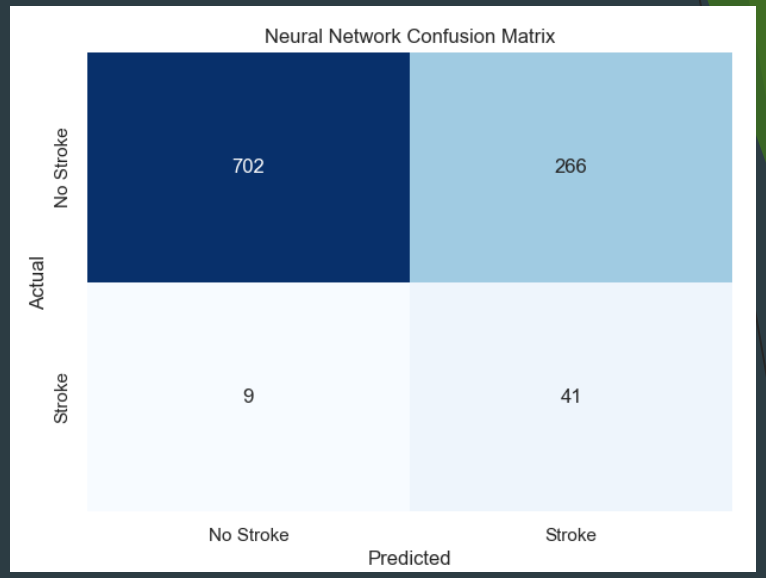


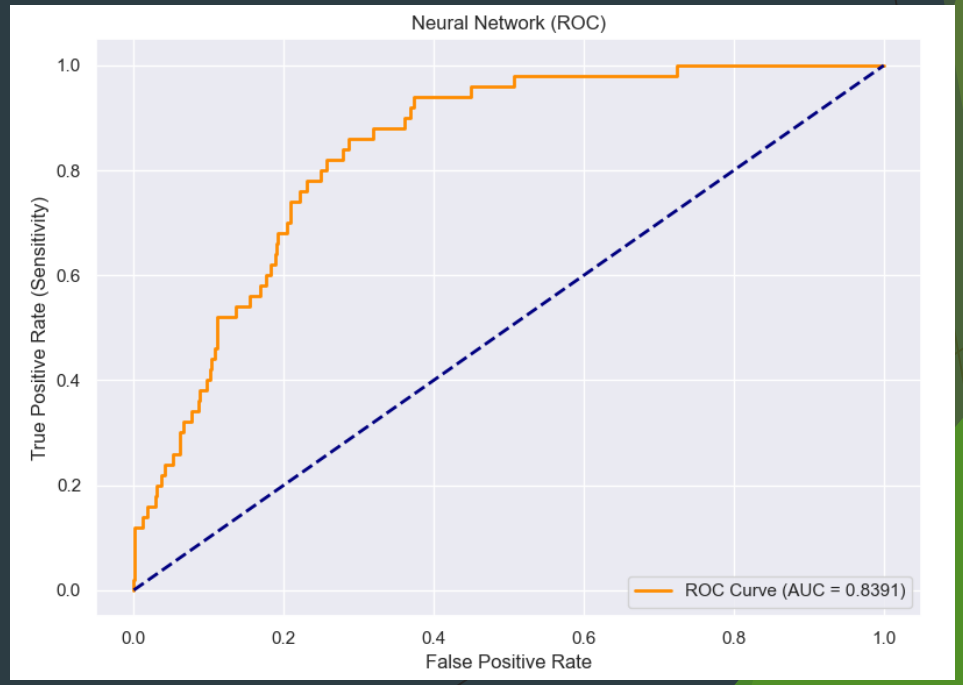




Neural Network

- Sensitivity 82%
- Good AUC score. Good predictor.





Conclusion

Can doctors use these machine learning models to predict stroke likelihood?

With this model…
no.

Baseline accuracy - 95.2 %

Best Sensitivity achieved - 82%

Limitations

- Baseline accuracy is too high.
- Many records had unknown smoking status.
- Missing BMI data.
- Small number of records.
- Many of the features seemed uncorrelated.

References:

- [1] Fedesoriano. (2021, January 26). Stroke prediction dataset. Kaggle. Retrieved May 6, 2025, from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

- [2] Centers for Disease Control and Prevention. (2022, October 14). Stroke facts. Centers for Disease Control and Prevention. Retrieved May 6, 2025, from https://www.cdc.gov/stroke/facts.htm

- [3] Singh, P. K. (2021, October 28). World stroke day. World Health Organization. Retrieved May 6, 2025, from https://www.who.int/southeastasia/news/detail/28-10-2021-world-stroke-day