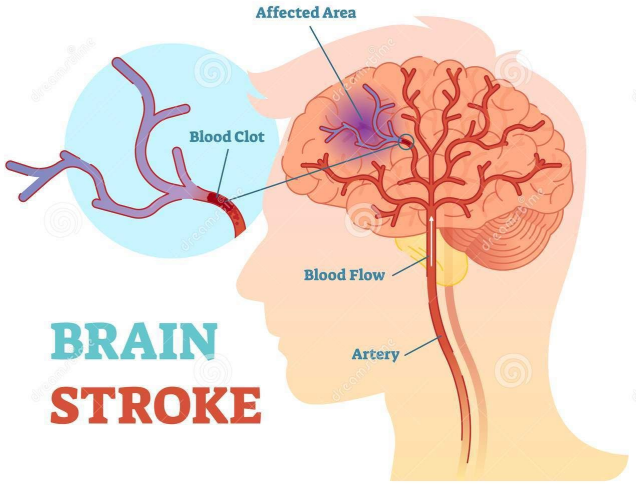
Stroke Prediction with Machine Learning and Neural Networks: A Comprehensive Predictive Intelligence Framework

By: Tanav Thanjavuru, Ashik Sathiya

Reasoning

- According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.
- Strokes occur when the blood supply to part of the brain is interrupted or reduced, depriving brain cells of the oxygen and nutrients they need to survive.
    - Strokes has a negative impact on society, and efforts have been made to improve stroke management and diagnosis.
        - Each year, about 795,000 people in the United States have strokes, and of these incidents, 137,000 of the people die.
    - The crossroad of Artificial Intelligence and the medical field could save countless lives and help medical professionals across the world



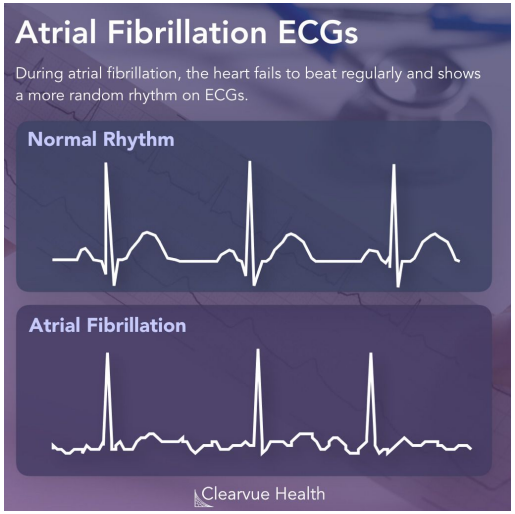
Literature Review (1)

- ‘A predictive analytics approach for stroke prediction using machine learning and neural networks’
    - Soumyabrata Dev, Hewei Wang, Chidozie Shamrock Nwosu, Nishtha Jain, Bharadwaj Veeravalli, Deepu John
- Principal Component Analysis
    - Did not improve scores

Way	Precision	Recall	F-score	Accuracy	Miss rate	Fall-out rate
DT	0.75	0.74	0.74	0.74	0.17	0.24
RF	0.74	0.73	0.73	0.74	0.18	0.25
NN	0.80	0.74	0.77	0.77	0.16	0.18
CNN	0.74	0.72	0.73	0.74	0.17	0.24
SVM	0.67	0.68	0.68	0.68	0.23	0.32
LASSO	0.78	0.72	0.75	0.76	0.19	0.20
ElasticNet	0.79	0.71	0.75	0.76	0.19	0.19

Literature Review (2)

- ‘A nationwide deep learning pipeline to predict stroke and COVID-19 death in atrial fibrillation’
    - Alex Handy, Angela Wood, Cathie Sudlow, Christopher Tomlinson, Frank Kee, Johan H Thygesen, Mohammad Mamouei, Reecha Sofat, Richard Dobson, Samantha Ip, Spiros Denaxas on behalf of the CVD-COVID-UK Consortium
- Machine learning pipeline to predict stroke and COVID-19 death in patients with atrial fibrillation in England
    - Atrial fibrillation: an irregular and often very rapid heart rhythm that can lead to blood clots in the heart
- There is a current scoring benchmark for stroke prediction called ‘CHA2DS2-VASc’ (Not ML based)
    - LSTM, Random Forest, and XGBoost
- The study showed that the machine learning pipeline improved first stroke prediction in atrial fibrillation by 17% compared to the current benchmark tool



Literature Review (3)

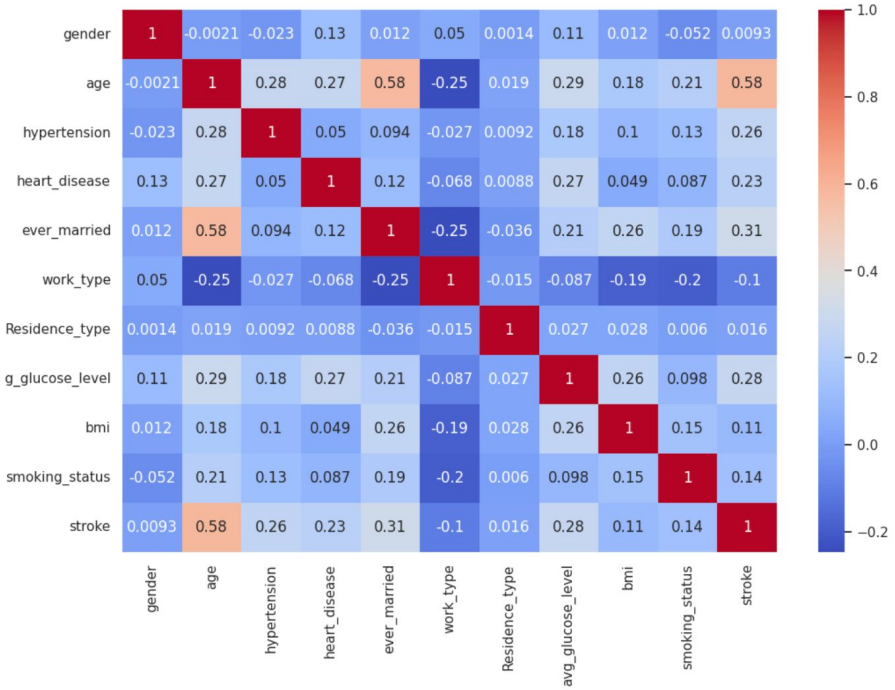
- ‘Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults’
    - Matthew Chun, Robert Clarke, Benjamin J Cairns, David Clifton, Derrick Bennett, Yiping Chen, Yu Guo, Pei Pei, Jun Lv, Canqing Yu,
- The study collected data on participants' lifestyle factors (such as smoking and alcohol habits), medical history, physical activity, and physical measurements.
    - 143 risk factor indicators
    - Focused on individuals with no prior history of stroke at baseline, and included incident cases of first stroke recorded for up to 9 years after the baseline survey
- Since GBT and COX have the highest scores, the researchers built an ensemble model out of it which had a 76% accuracy in men and 80% in women

Dataset

- Dataset of electronic health records
  - Released by Kaggle
  - 5110 total rows
  - **Features**
    - **id:** unique identifier
    - **gender:** "Male", "Female" or "Other"
    - **age:** age of the patient
    - **hypertension:** 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
    - **heart_disease:** 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
    - **ever_married:** "No" or "Yes"
    - **work_type:** "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
    - **Residence_type:** "Rural" or "Urban"
    - **avg_glucose_level:** average glucose level in blood
    - **bmi:** body mass index
    - **smoking_status:** "formerly smoked", "never smoked", "smokes" or "Unknown"*
    - **stroke:** 1 if the patient had a stroke or 0 if not

Exploratory Data Analysis

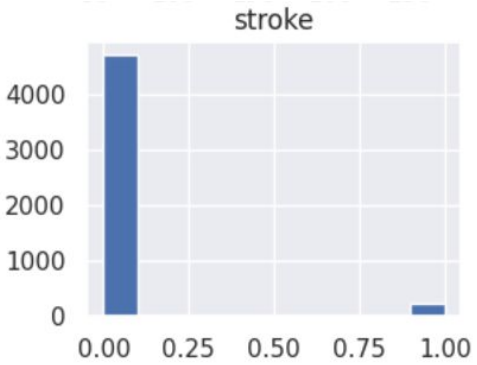
- **Heatmap**
  - Graphical representation of data where values in a matrix are represented as colors
  - Heatmaps can help identify patterns and trends in the data
- **Findings**
  - 0.58 Correlation - Age & Stroke
  - 0.31 Correlation - Marriage & Stroke
  - 0.28 Correlation - Avg. Glucose Level & Stroke
  - 0.26 Correlation - Hypertension & Stroke
  - 0.23 Correlation - Heart Disease & Stroke

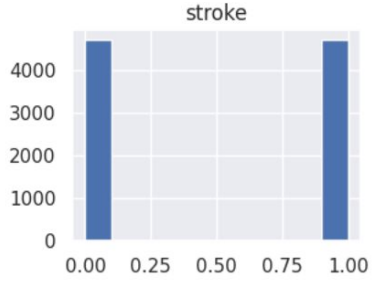


Exploratory Data Analysis/Data Preprocessing

- **Imbalanced Dataset**
- **RandomOverSampler**
  - Randomly replicates minority class samples until the number of samples in each class is roughly equal
  - This helps to address the issue of class imbalance and improve the performance of machine learning models on the minority class.

`





Data Preprocessing

- **Drop Nulls**
- **Label Encoder**
  - Work Type
  - Smoking Status
  - Gender
  - Ever Married
  - Residence Type
- **Train-test Split**
  - 0.7 Training
  - 0.3 Testing

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
5105	18234	Female	80.0	1	0	Yes	Private	Urban	83.75	NaN	never smoked	0
5106	44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
5107	19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5108	37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
5109	44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

5110 rows × 12 columns

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	1	67.0	0	1	1	2	1	228.69	36.6	1	1
2	31112	1	80.0	0	1	1	2	0	105.92	32.5	2	1
3	60182	0	49.0	0	0	1	2	1	171.23	34.4	3	1
4	1665	0	79.0	1	0	1	3	0	174.12	24.0	2	1
5	56669	1	81.0	0	0	1	2	1	186.21	29.0	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...
5104	14180	0	13.0	0	0	0	4	0	103.08	18.6	0	0
5106	44873	0	81.0	0	0	1	3	1	125.20	40.0	2	0
5107	19723	0	35.0	0	0	1	3	0	82.99	30.6	2	0
5108	37544	1	51.0	0	0	1	2	0	166.29	25.6	1	0
5109	44679	0	44.0	0	0	1	0	1	85.28	26.2	0	0

4909 rows × 12 columns

Machine Learning Models - Neural Network

- **Sequential**
  - Layers
    - Relu - 12
    - Relu - 8
    - Sigmoid - 1
    - 100 Epochs
- **Results**
  - Mean Squared Error- 0.1489
  - Accuracy - 79.67%

```
from keras.models import Sequential
from keras.layers import Dense

# define the model
model = Sequential()
model.add(Dense(12, input_dim=X_train.shape[1], activation='relu'))
model.add(Dense(8, activation='relu'))
model.add(Dense(1, activation='sigmoid'))

# compile the model
model.compile(loss='mse', optimizer='adam', metrics=['accuracy'])

# fit the model
model.fit(X_train, y_train, epochs=100, batch_size=64)
```

```
Epoch 88/100
103/103 [==============================] - 0s 2ms/step - loss: 0.1498 - accuracy: 0.7942
Epoch 89/100
103/103 [==============================] - 0s 2ms/step - loss: 0.1499 - accuracy: 0.7941
Epoch 90/100
103/103 [==============================] - 0s 2ms/step - loss: 0.1508 - accuracy: 0.7910
Epoch 91/100
103/103 [==============================] - 0s 2ms/step - loss: 0.1501 - accuracy: 0.7926
Epoch 92/100
103/103 [==============================] - 0s 2ms/step - loss: 0.1492 - accuracy: 0.7927
Epoch 93/100
103/103 [==============================] - 0s 2ms/step - loss: 0.1499 - accuracy: 0.7929
Epoch 94/100
103/103 [==============================] - 0s 2ms/step - loss: 0.1487 - accuracy: 0.7994
Epoch 95/100
103/103 [==============================] - 0s 2ms/step - loss: 0.1488 - accuracy: 0.7960
Epoch 96/100
103/103 [==============================] - 0s 2ms/step - loss: 0.1499 - accuracy: 0.7910
Epoch 97/100
103/103 [==============================] - 0s 2ms/step - loss: 0.1520 - accuracy: 0.7892
Epoch 98/100
103/103 [==============================] - 0s 2ms/step - loss: 0.1529 - accuracy: 0.7836
Epoch 99/100
103/103 [==============================] - 0s 2ms/step - loss: 0.1489 - accuracy: 0.7967
Epoch 100/100
103/103 [==============================] - 0s 2ms/step - loss: 0.1488 - accuracy: 0.7933
```

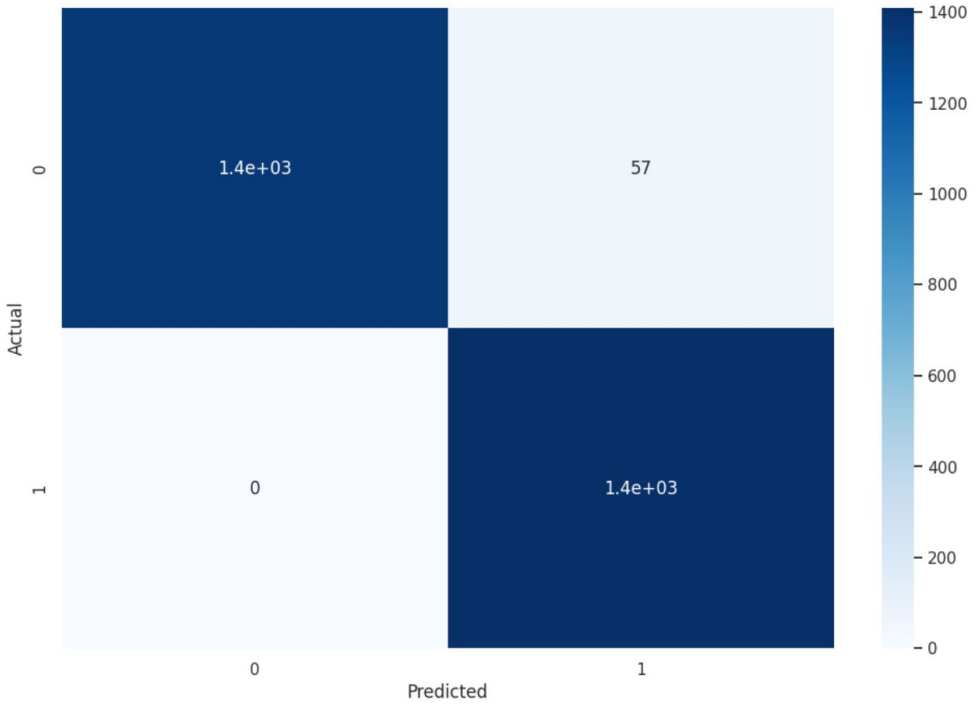
Machine Learning Models - Ensemble Learning

- **XGBoostClassifier**
  - GridSearchCV
    - Learning Rate - 0.3
    - Max_Depth - 9
    - Alpha - 0.1
- **Results**
  - Mean Squared Error - 0.0195
  - Accuracy - 98.05%

```
Best Hyperparameters:  {'alpha': 0.1, 'learning_rate': 0.3, 'max_depth': 9}
Best Score:  0.9758358662613983
Mean Squared Error: 0.01950354609929078
Accuracy: 98.05%
```

Model Validation

- **Confusion Matrix**
  - Compares the predicted values of the model with the actual values in the test dataset, and displays the number of true positives, true negatives, false positives, and false negatives.
- **Results**
  - 1400 TP
  - 1400 TN
  - 0 FN
  - 57 FP

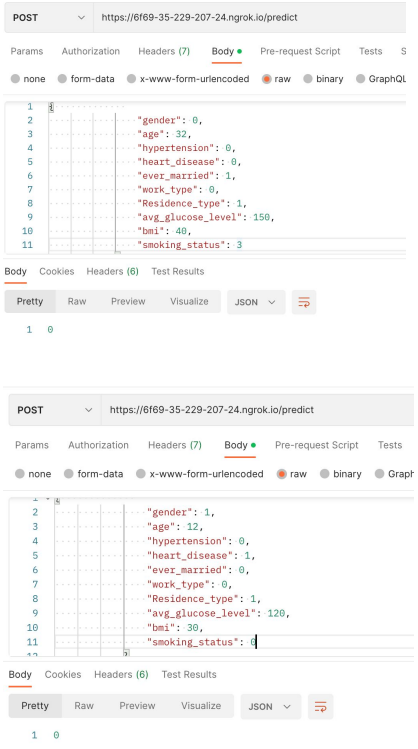


Endpoint Testing

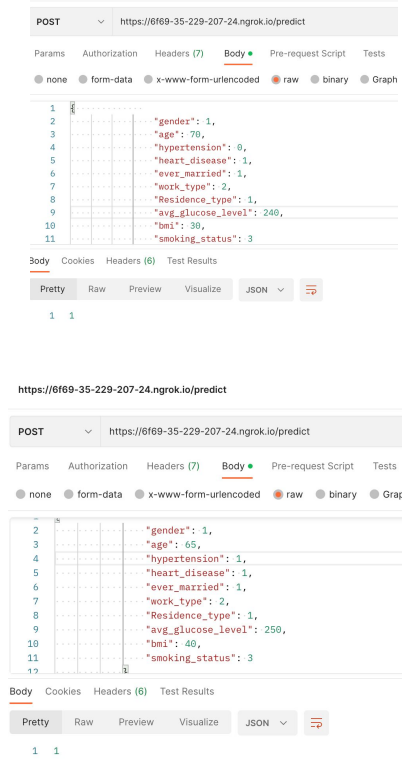
- **Endpoint Deployment**
  - ColabCode
  - FastAPI
  - Pydantic
    - BaseModel
  - Postman API testing

- **Test Machine Learning model with different values**
  - Connect Endpoint to Postman API tester

Negative Stroke



Positive Stroke



Conclusion

- **Deep Learning and Neural Networks**
  - Using both of these forms of Artificial Intelligence we were able to predict a stroke with over 75% accuracy based on risk factors
- **Future**
  - With results like these, medical professionals can successfully predict strokes using certain risk factors
  - This could help medical professionals affirm their own diagnoses and could help improve patient outcomes
  - Since these models were able to be successful with stroke predictions it could also work for other diseases which could include:
    - Heart Disease Prediction
    - Diabetes Prediction
    - Cancer Prediction

References

A predictive analytics approach for stroke prediction using machine learning and neural networks. Soumyabrata Dev, Hewei Wang, Chidozie Shamrock Nwosu, Nishtha Jain, Bharadwaj Veeravalli, Deepu John, Healthcare Analytics, Volume 2, 2022, 100032, ISSN 2772-4425, https://doi.org/10.1016/j.health.2022.100032.

A nationwide deep learning pipeline to predict stroke and COVID-19 death in atrial fibrillation. Alex Handy, Angela Wood, Cathie Sudlow, Christopher Tomlinson, Frank Kee, Johan H Thygesen, Mohammad Mamouei, Reecha Sofat, Richard Dobson, Samantha Ip, Spiros Denaxas, medRxiv 2021.12.20.21268113; doi: https://doi.org/10.1101/2021.12.20.21268113

Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults. Matthew Chun, Robert Clarke, Benjamin J Cairns, David Clifton, Derrick Bennett, Yiping Chen, Yu Guo, Pei Pei, Jun Lv, Canqing Yu, Ling Yang, Liming Li, Zhengming Chen, Tingting Zhu, the China Kadoorie Biobank Collaborative Group, Journal of the American Medical Informatics Association, Volume 28, Issue 8, August 2021, Pages 1719–1727, https://doi.org/10.1093/jamia/ocab068