Stroke Prediction

Jason A., Kimberly, and Oliver

Why Stroke Prediction?

Stroke remains the second leading cause of death in the world. Being able to have early prediction of the likelihood of a person havng a stroke would allow to provide preventative care.

Data Set

Features

Demograchics

- Age
- Gender

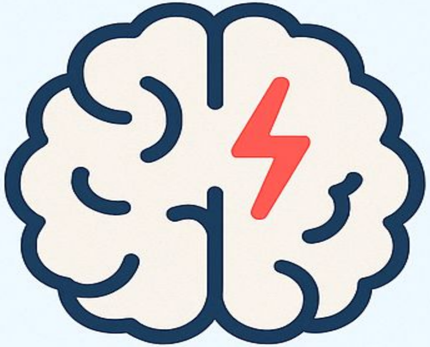
Medical History

- Avg glucose
- Hypertension
- Heart disease

Lifestyle

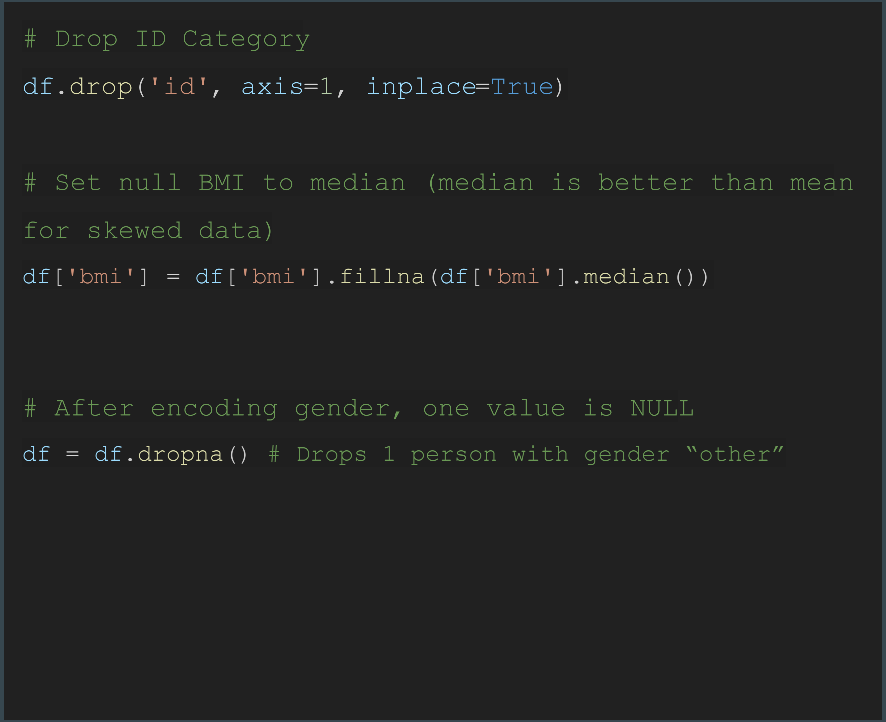
• Work type • BMI • Smoking status Residence

Target Variable

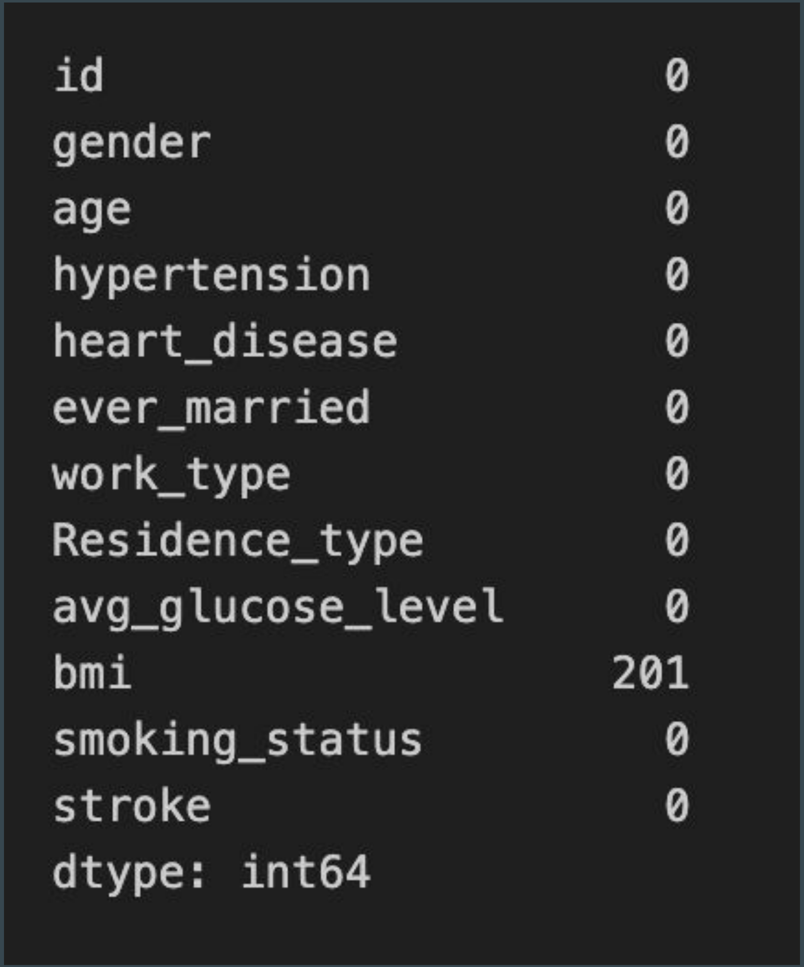


Stroke

Data Cleaning



```
# Drop ID Category
df.drop('id', axis=1, inplace=True)

# Set null BMI to median (median is better than mean
for skewed data)
df['bmi'] = df['bmi'].fillna(df['bmi'].median())


# After encoding gender, one value is NULL
df = df.dropna() # Drops 1 person with gender “other”
```

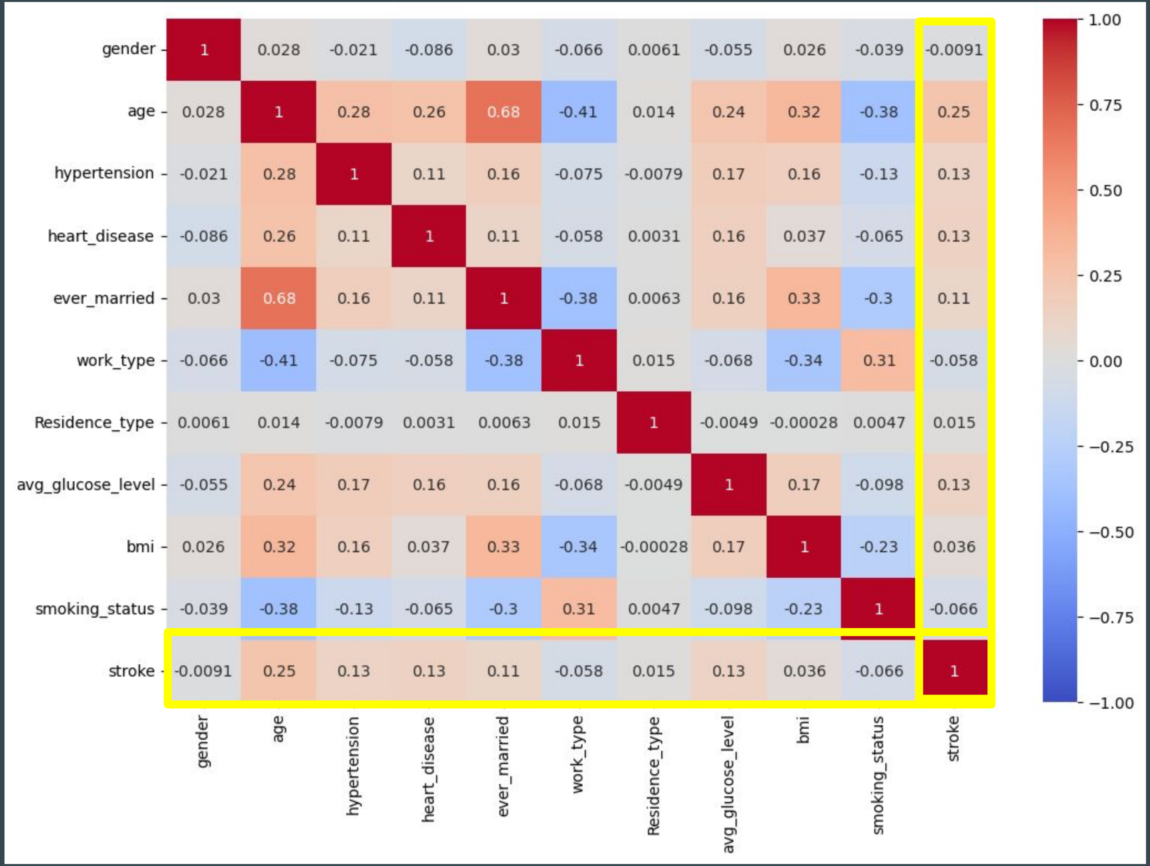


```
id                     0
gender                 0
age                    0
hypertension           0
heart_disease          0
ever_married           0
work_type              0
Residence_type         0
avg_glucose_level      0
bmi                  201
smoking_status         0
stroke                 0
dtype: int64
```

Data Encoding

```
df['gender'] = df['gender'].map({'Male': 0, 'Female':
1})

# Married?
df['ever_married'] = df['ever_married'].map({'Yes':
1, 'No': 0})

# Work Type
work_types = {
    'Private': 0,
    'Self-employed': 1,
    'Govt_job': 2,
    'children': 3,
    'Never_worked': 4
}
df['work_type'] = df['work_type'].map(work_types)
```

```
# Residence Type: Urban = 1, Rural = 0
df['Residence_type'] =
df['Residence_type'].map({'Urban': 1, 'Rural': 0})

# Smoking_status
df['smoking_status'] = df['smoking_status'].map({
    'formerly smoked': 0,
    'never smoked': 1,
    'smokes': 2,
    'Unknown': 3
})
```



	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
gender	1	0.028	-0.021	-0.086	0.03	-0.066	0.0061	-0.055	0.026	-0.039	-0.0091
age	0.028	1	0.28	0.26	0.68	-0.41	0.014	0.24	0.32	-0.38	0.25
hypertension	-0.021	0.28	1	0.11	0.16	-0.075	-0.0079	0.17	0.16	-0.13	0.13
heart_disease	-0.086	0.26	0.11	1	0.11	-0.058	0.0031	0.16	0.037	-0.065	0.13
ever_married	0.03	0.68	0.16	0.11	1	-0.38	0.0063	0.16	0.33	-0.3	0.11
work_type	-0.066	-0.41	-0.075	-0.058	-0.38	1	0.015	-0.068	-0.34	0.31	-0.058
Residence_type	0.0061	0.014	-0.0079	0.0031	0.0063	0.015	1	-0.0049	-0.00028	0.0047	0.015
avg_glucose_level	-0.055	0.24	0.17	0.16	0.16	-0.068	-0.0049	1	0.17	-0.098	0.13
bmi	0.026	0.32	0.16	0.037	0.33	-0.34	-0.00028	0.17	1	-0.23	0.036
smoking_status	-0.039	-0.38	-0.13	-0.065	-0.3	0.31	0.0047	-0.098	-0.23	1	-0.066
stroke	-0.0091	0.25	0.13	0.13	0.11	-0.058	0.015	0.13	0.036	-0.066	1

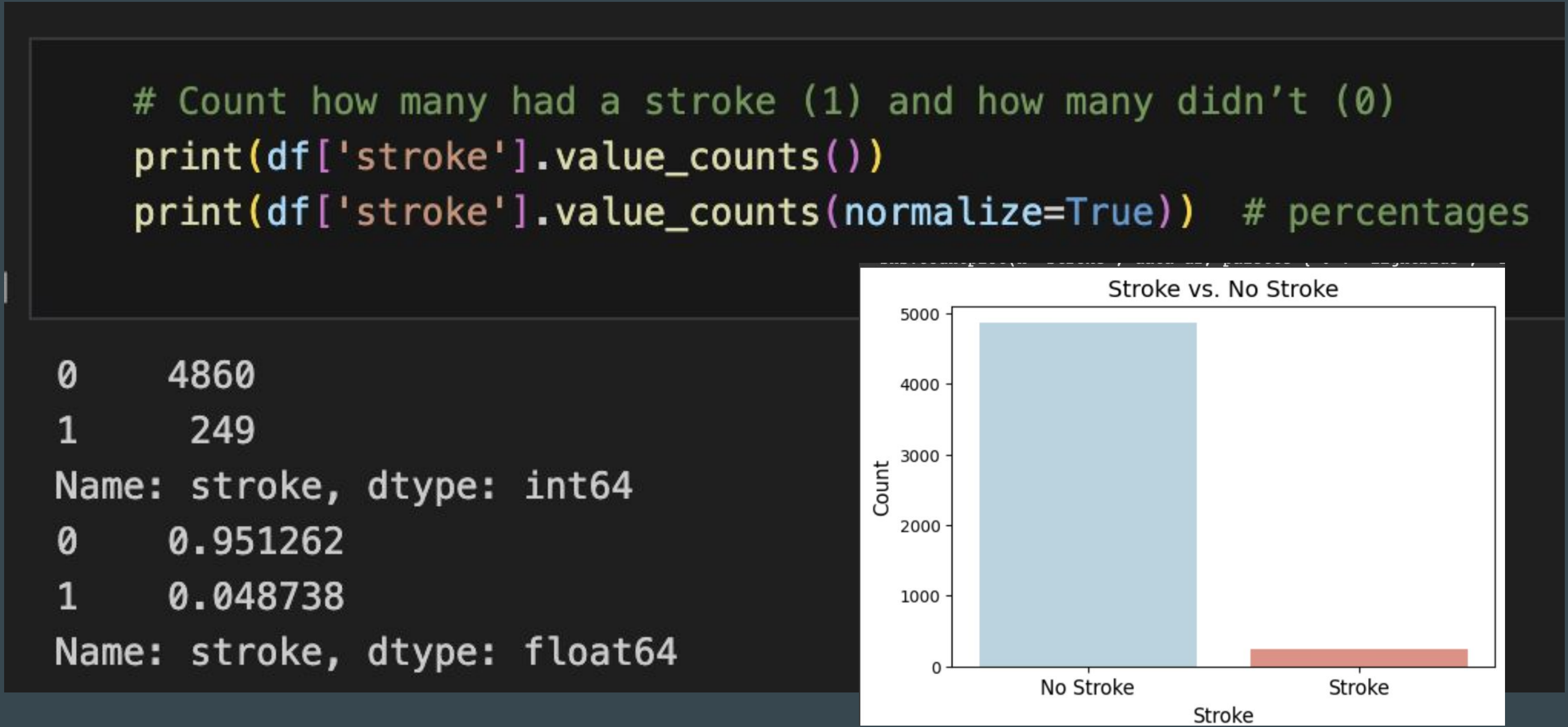
Top correlations with stroke:

- Age + .25
- Hypertension + .13
- Heart disease + .13
- Average glucose level + .13

Negative correlations:

- Smoking status - 0.066
- Work type - 0.058
- Gender - 0.0091

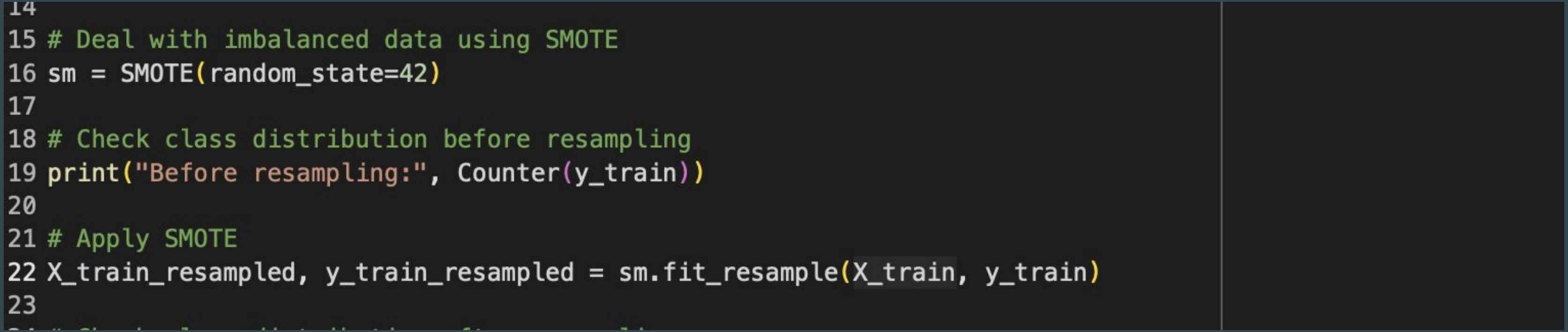
Dealing with Imbalanced Data



```
# Count how many had a stroke (1) and how many didn’t (0)
print(df['stroke'].value_counts())
print(df['stroke'].value_counts(normalize=True))  # percentages
```

```
0    4860
1     249
Name: stroke, dtype: int64
0    0.951262
1    0.048738
Name: stroke, dtype: float64
```

SMOTE for Imbalanced Data Sets

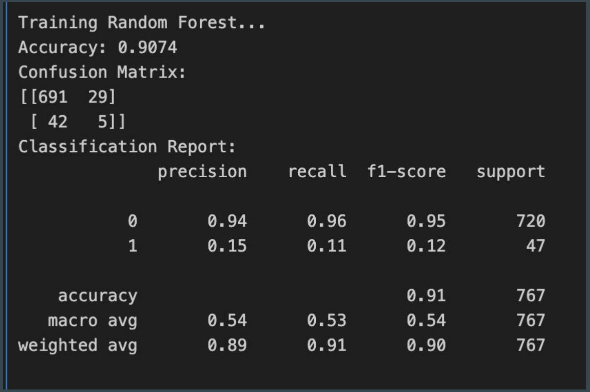


```python

# Deal with imbalanced data using SMOTE
sm = SMOTE(random_state=42)

# Check class distribution before resampling
print("Before resampling:", Counter(y_train))

# Apply SMOTE
X_train_resampled, y_train_resampled = sm.fit_resample(X_train, y_train)

```

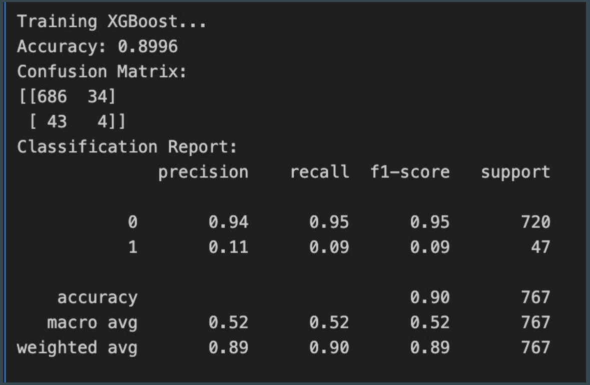
SMOTE = Synthetic Minority Over-sampling Technique

It creates new, synthetic examples of the minority class (e.g., strokes = 1) by interpolating between existing cases, helping the model better recognize these rare events.

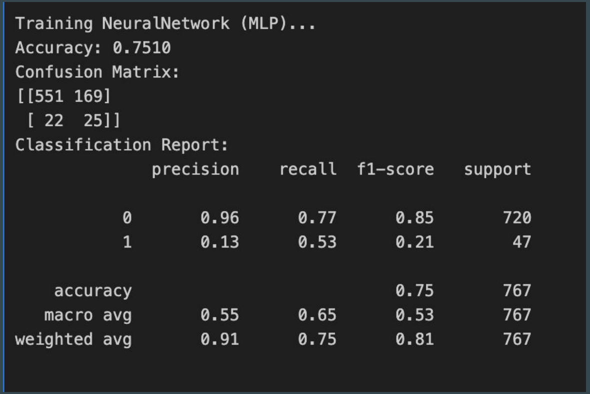
Training and Comparing Different Models



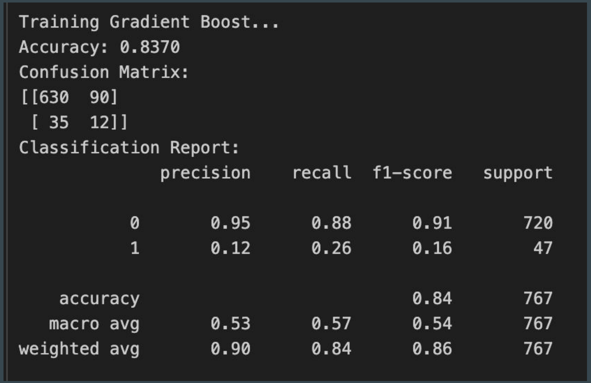
```
Training Random Forest...
Accuracy: 0.9074
Confusion Matrix:
[[691  29]
 [ 42   5]]
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.96      0.95       720
           1       0.15      0.11      0.12        47

    accuracy                           0.91       767
   macro avg       0.54      0.53      0.54       767
weighted avg       0.89      0.91      0.90       767
```



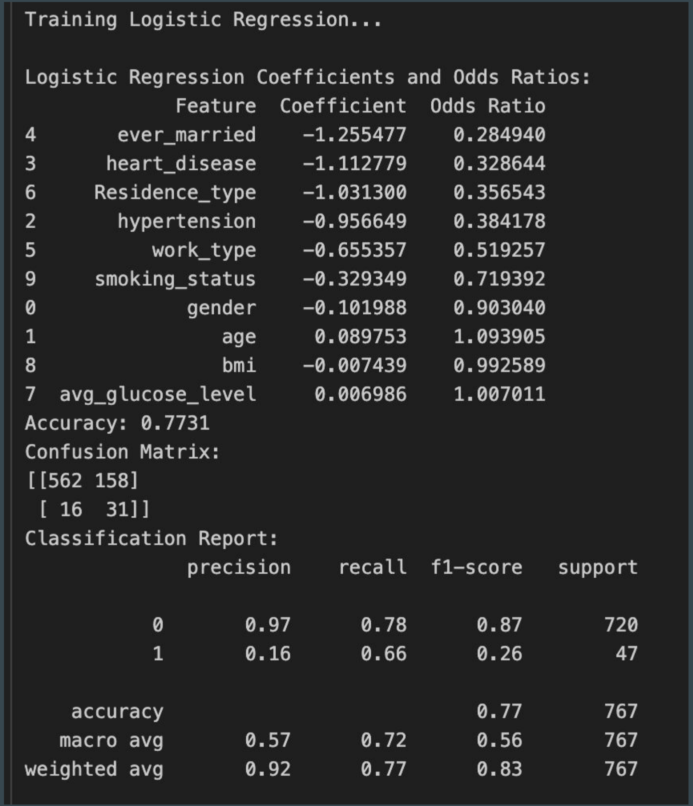
```
Training XGBoost...
Accuracy: 0.8996
Confusion Matrix:
[[686  34]
 [ 43   4]]
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.95      0.95       720
           1       0.11      0.09      0.09        47

    accuracy                           0.90       767
   macro avg       0.52      0.52      0.52       767
weighted avg       0.89      0.90      0.89       767
```



```
Training NeuralNetwork (MLP)...
Accuracy: 0.7510
Confusion Matrix:
[[551 169]
 [ 22  25]]
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.77      0.85       720
           1       0.13      0.53      0.21        47

    accuracy                           0.75       767
   macro avg       0.55      0.65      0.53       767
weighted avg       0.91      0.75      0.81       767
```



```
Training Gradient Boost...
Accuracy: 0.8370
Confusion Matrix:
[[630  90]
 [ 35  12]]
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.88      0.91       720
           1       0.12      0.26      0.16        47

    accuracy                           0.84       767
   macro avg       0.53      0.57      0.54       767
weighted avg       0.90      0.84      0.86       767
```



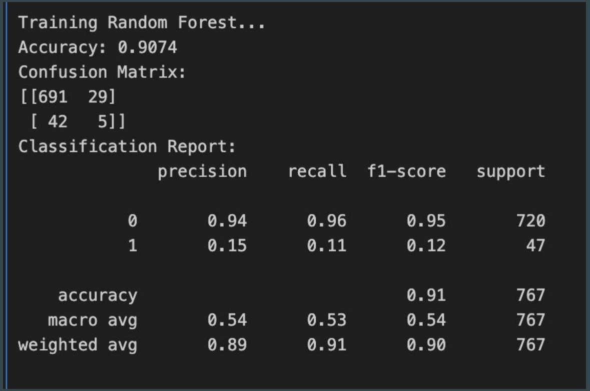
```
Training Logistic Regression...

Logistic Regression Coefficients and Odds Ratios:
            Feature  Coefficient  Odds Ratio
4       ever_married    -1.255477    0.284940
3      heart_disease    -1.112779    0.328644
6     Residence_type    -1.031300    0.356543
2       hypertension    -0.956649    0.384178
5          work_type    -0.655357    0.519257
9     smoking_status    -0.329349    0.719392
0             gender    -0.101988    0.903040
1                age     0.089753    1.093905
8                bmi    -0.007439    0.992589
7  avg_glucose_level     0.006986    1.007011
Accuracy: 0.7731
Confusion Matrix:
[[562 158]
 [ 16  31]]
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.78      0.87       720
           1       0.16      0.66      0.26        47

    accuracy                           0.77       767
   macro avg       0.57      0.72      0.56       767
weighted avg       0.92      0.77      0.83       767
```

Deciding What’s Valuable

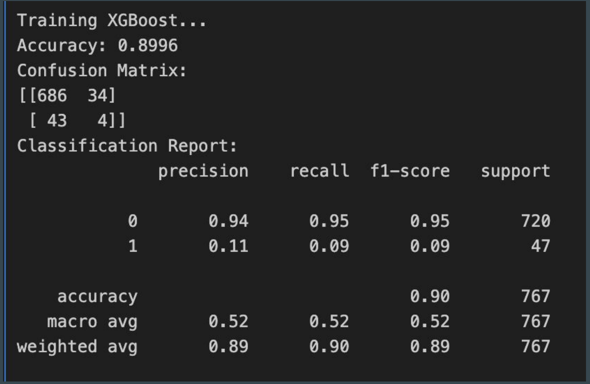
- Picking “No” every time gives 95% accuracy.
- What’s the important metric ethically?

- **Recall** (true positive rate) measures how well the model accurately predicts true positives.
- Out of all the positive cases in the data set, how many did the model predict correctly?

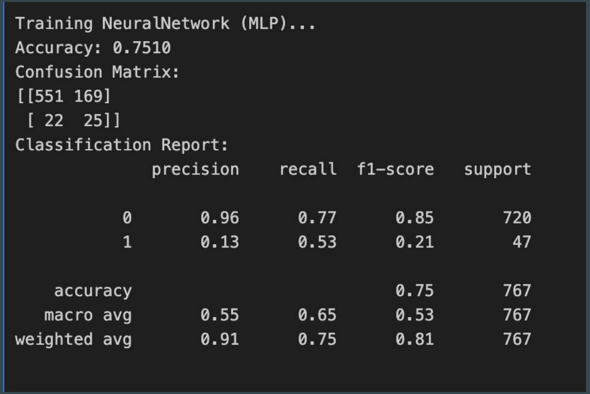
Training and Comparing Different Models



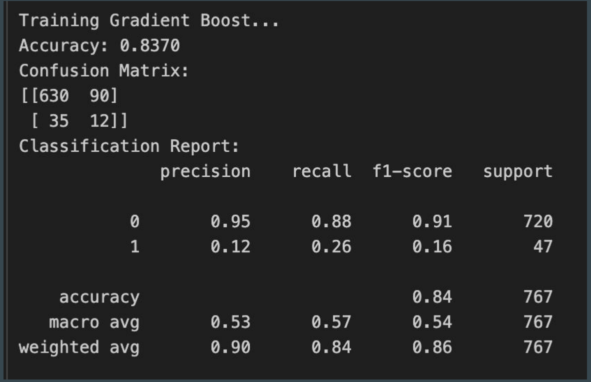
```
Training Random Forest...
Accuracy: 0.9074
Confusion Matrix:
[[691  29]
 [ 42   5]]
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.96      0.95       720
           1       0.15      0.11      0.12        47

    accuracy                           0.91       767
   macro avg       0.54      0.53      0.54       767
weighted avg       0.89      0.91      0.90       767
```



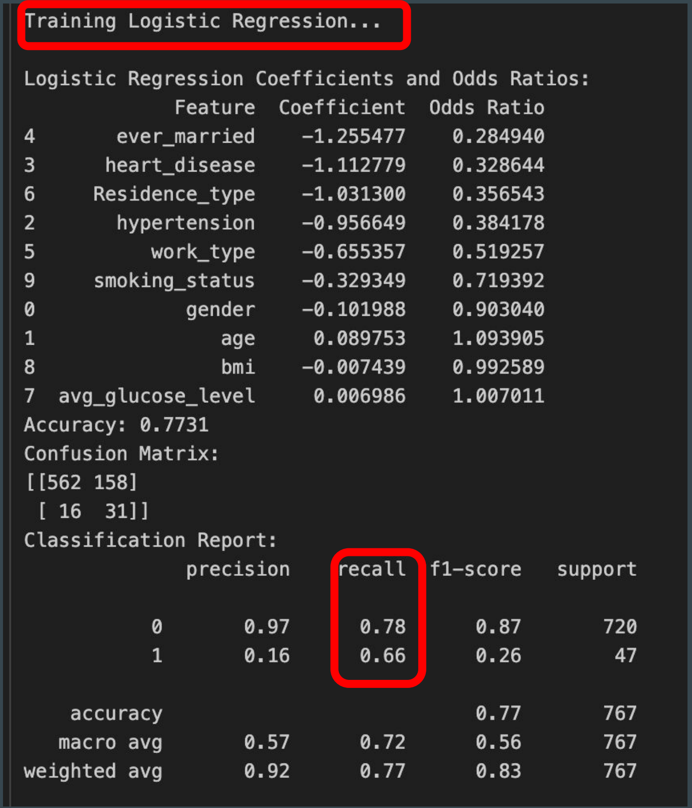
```
Training XGBoost...
Accuracy: 0.8996
Confusion Matrix:
[[686  34]
 [ 43   4]]
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.95      0.95       720
           1       0.11      0.09      0.09        47

    accuracy                           0.90       767
   macro avg       0.52      0.52      0.52       767
weighted avg       0.89      0.90      0.89       767
```



```
Training NeuralNetwork (MLP)...
Accuracy: 0.7510
Confusion Matrix:
[[551 169]
 [ 22  25]]
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.77      0.85       720
           1       0.13      0.53      0.21        47

    accuracy                           0.75       767
   macro avg       0.55      0.65      0.53       767
weighted avg       0.91      0.75      0.81       767
```



```
Training Gradient Boost...
Accuracy: 0.8370
Confusion Matrix:
[[630  90]
 [ 35  12]]
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.88      0.91       720
           1       0.12      0.26      0.16        47

    accuracy                           0.84       767
   macro avg       0.53      0.57      0.54       767
weighted avg       0.90      0.84      0.86       767
```



```
Training Logistic Regression...

Logistic Regression Coefficients and Odds Ratios:
             Feature  Coefficient  Odds Ratio
4       ever_married    -1.255477    0.284940
3      heart_disease    -1.112779    0.328644
6     Residence_type    -1.031300    0.356543
2       hypertension    -0.956649    0.384178
5          work_type    -0.655357    0.519257
9     smoking_status    -0.329349    0.719392
0             gender    -0.101988    0.903040
1                age     0.089753    1.093905
8                bmi    -0.007439    0.992589
7  avg_glucose_level     0.006986    1.007011
Accuracy: 0.7731
Confusion Matrix:
[[562 158]
 [ 16  31]]
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.78      0.87       720
           1       0.16      0.66      0.26        47

    accuracy                           0.77       767
   macro avg       0.57      0.72      0.56       767
weighted avg       0.92      0.77      0.83       767
```

Why Logistic Regression is the Best Model for Stroke Prediction

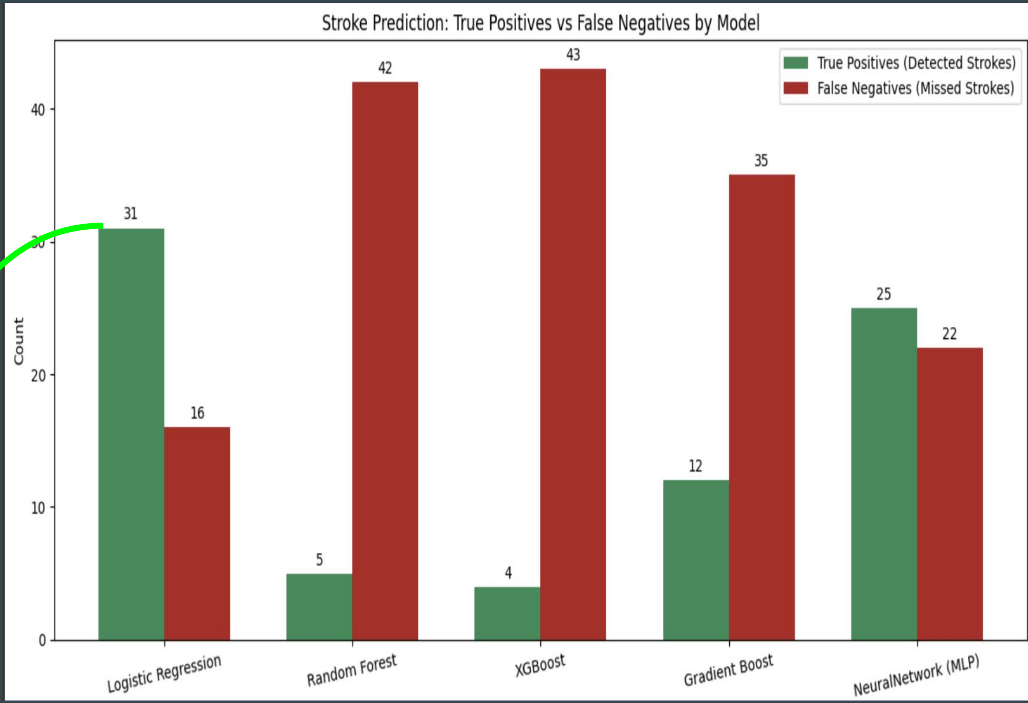
Model Chosen: Logistic Regression

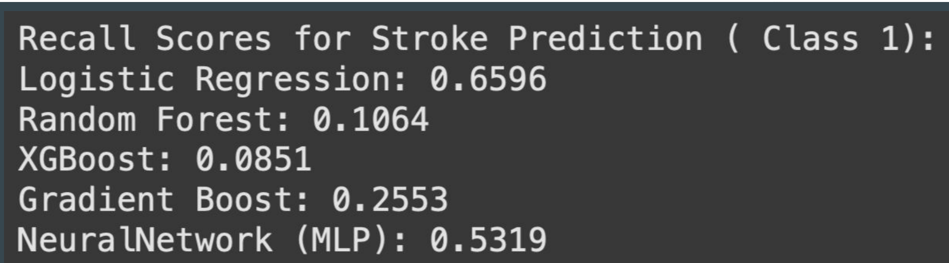
Why?

- Missing a stroke case is worse than a false alarm.
- Prioritize Recall, which measures how well we catch actual stroke cases.
- NIH states: “This metric [Recall] is also regarded as being among the most important for medical studies, since it is desired to **miss as few positive** instances as possible, which translates to a high recall.”
- Logistic Regression had the highest recall and the most true positives.

Goal:

- Maximize identification of real stroke cases to save lives.

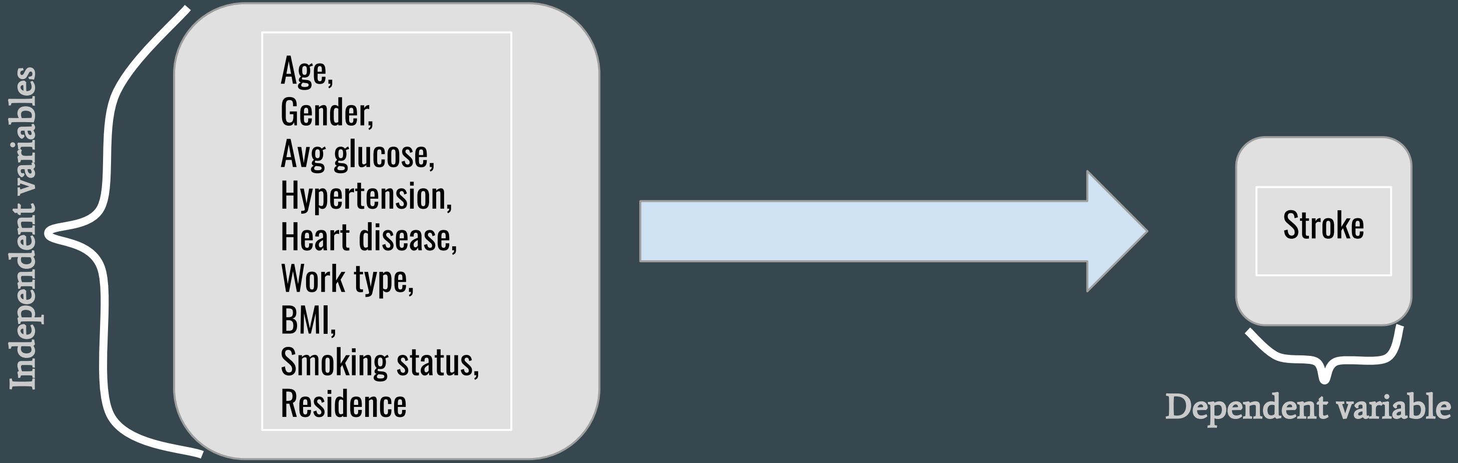




```
Recall Scores for Stroke Prediction ( Class 1):
Logistic Regression: 0.6596
Random Forest: 0.1064
XGBoost: 0.0851
Gradient Boost: 0.2553
NeuralNetwork (MLP): 0.5319
```

An Interpretable Model

- Logistic Regression allows us to interpret how each feature impacts the prediction.
- Doctors can ask: “Why did the model flag this patient?” and get a clear answer.
- In contrast: Black-box models like neural networks offer less explainability.



Independent variables

Age,
Gender,
Avg glucose,
Hypertension,
Heart disease,
Work type,
BMI,
Smoking status,
Residence

Stroke

Dependent variable

Factor Contribution

Feature	Odds Ratio	Interpretation
ever_married (Yes = 1)	0.285	Being married significantly reduces the odds of stroke (~72% lower than unmarried).
heart_disease (Yes = 1)	0.329	Having heart disease decreases stroke odds by ~67% counterintuitive. This could point to a data issue, label imbalance, or confounding features.
Residence_type (Urban = 1)	0.357	Living in an urban area reduces stroke odds by ~64% compared to rural. Possibly reflects better access to healthcare.
hypertension (Yes = 1)	0.384	Hypertension reduces stroke odds by ~62% another counterintuitive result. Clinically, this should increase stroke risk.

work_type (Private = 0 → Never_worked = 4)	0.519	As the work type shifts toward less conventional employment (e.g., never worked, children), stroke risk decreases. But it's a multi-category ordinal, so this interpretation needs caution.
smoking_status (formerly smoked = 0 → Unknown = 3)	0.719	Higher smoking associated with **lower stroke risk** , which is the opposite of expected. Suggests potential encoding or sampling bias.
gender (Male = 0, Female = 1)	0.903	Being female **slightly reduces** stroke risk compared to male (~10% less).
age (numeric)	1.094	Every additional year **increases** stroke odds by ~9%.
bmi (numeric)	0.993	Small effect — higher BMI **very slightly reduces** stroke odds (~0.7% per unit).
avg_glucose_level (numeric)	1.007	Slight **increase** in stroke risk with higher glucose (~0.7% per unit).

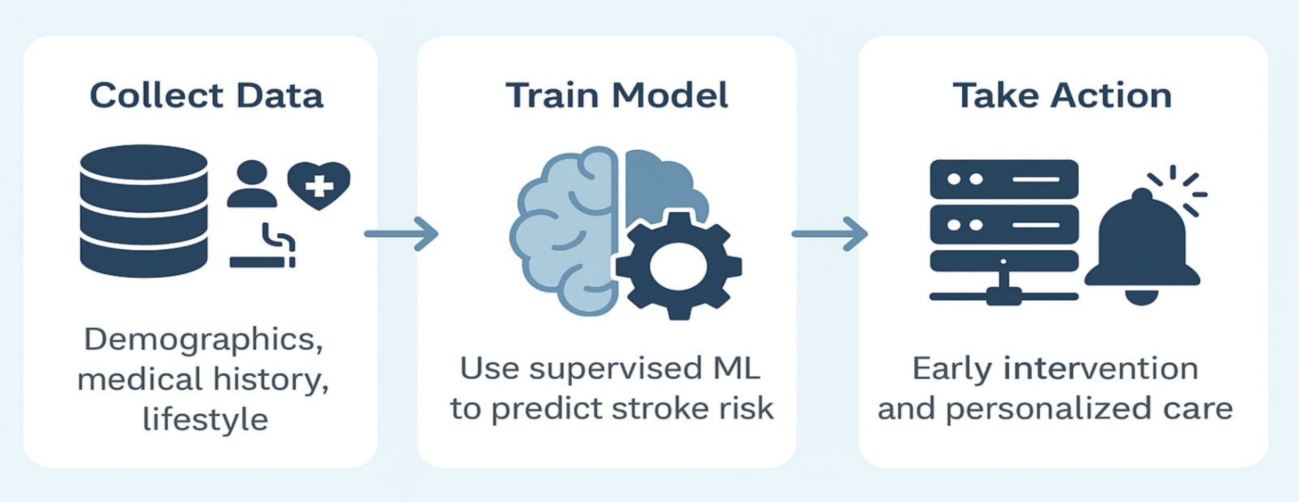
Learning

- Recall vs. Precision
- Critical thinking:
  - What metric matters most for this problem?
- Interesting factors:
  - Work type
  - Ever married

Potential Implementations

Preventative Care (Online Prediction)

- Doctors input your data to these models at an annual checkup
- If you are predicted to have a stroke, you can consult with your doctor to take preventative measures before hand.



Collect Data

Demographics, medical history, lifestyle

Train Model

Use supervised ML to predict stroke risk

Take Action

Early intervention and personalized care

Sources

Patni, Ayush. “How to Choose the Right Evaluation Metrics for Your ML Model ?” *Medium*, Medium, 27 Nov. 2023, ayushdpatni.medium.com/how-to-choose-the-right-evaluation-metrics-for-your-ml-model-ad1f448ae3a5.

Hicks, Steven A, et al. “On Evaluation Metrics for Medical Applications of Artificial Intelligence.” *Scientific Reports*, U.S. National Library of Medicine, 8 Apr. 2022, pmc.ncbi.nlm.nih.gov/articles/PMC8993826/.

Feigin VL;Brainin M;Norrving B;Martins SO;Pandian J;Lindsay P;F Grupper M;Rautalin I; “World Stroke Organization: Global Stroke Fact Sheet 2025.” *International Journal of Stroke : Official Journal of the International Stroke Society*, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/39635884/. Accessed 23 Apr. 2025.

Kaggle DataSet: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/code

Kaggle Dataset SMOTE reference:
https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/discussion?sort=undefined