SC1015

REP2_Team 5

By : Kirtana Nair &Aadya Gupta

Brain Stroke Prediction

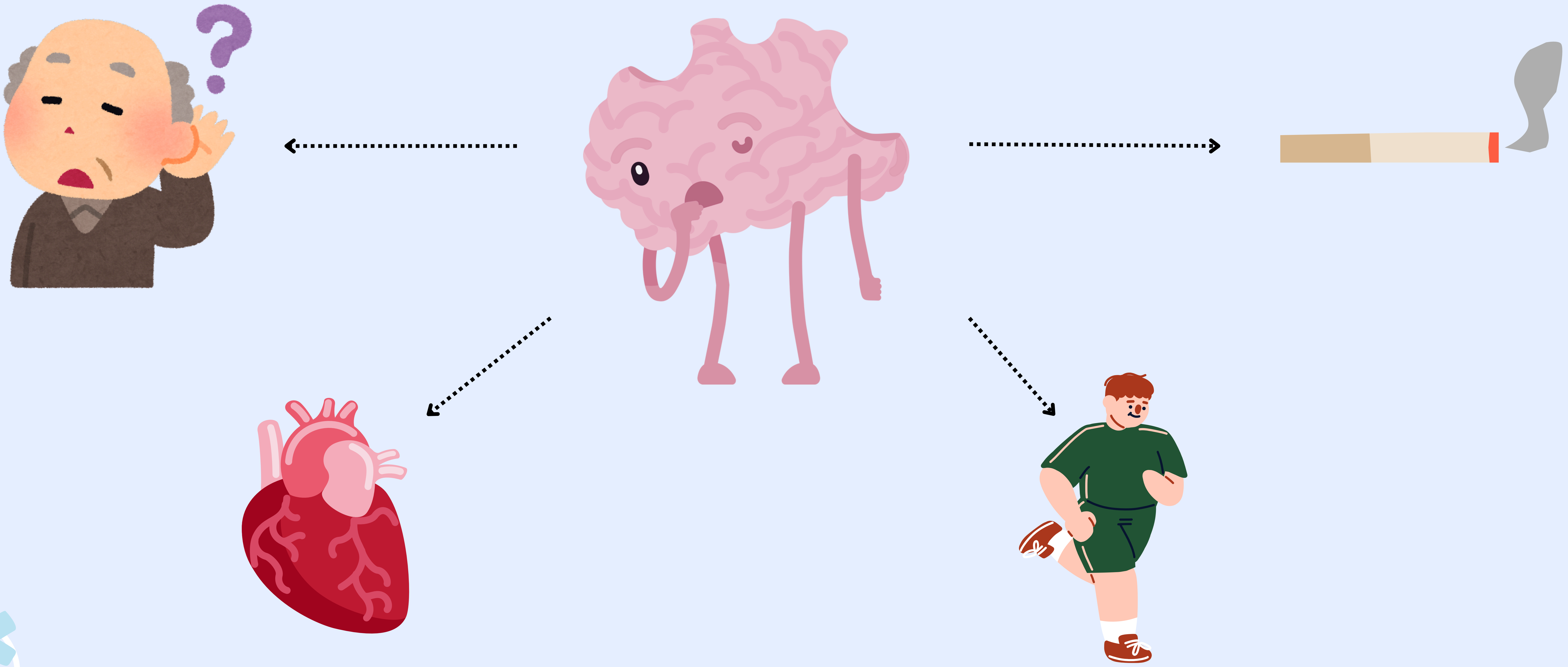
SC1015 – Mini Project

By- Kirtana Nair and Aadya Gupta

Brain stroke

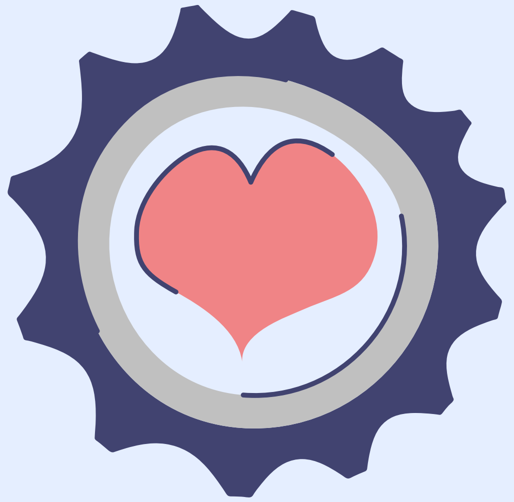
Kaggle - Brain Stroke Dataset

Brain Stroke Dataset Classification Prediction

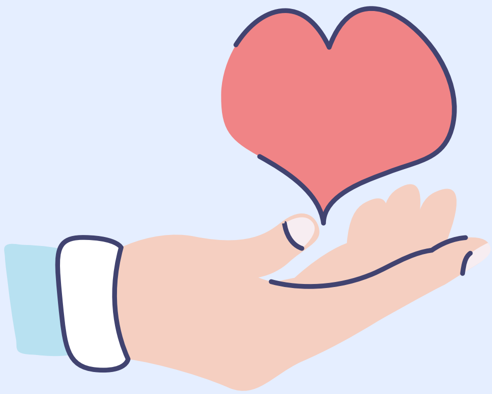


How do different variables affect the possibility of brain stroke in humans ?

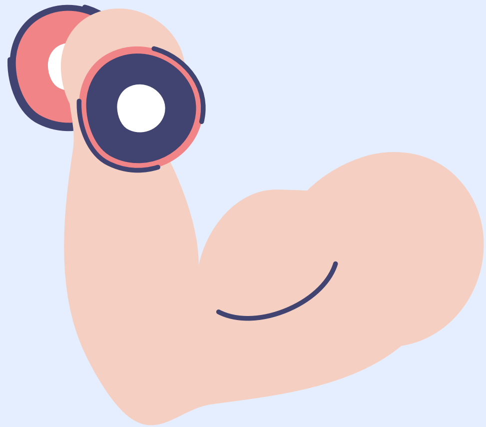
Why?



Early detection



Prevent delayed treatments



Mitigate long term effects of stroke

Removing Null sets

```
data.isnull().sum()
```

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

- data.isnull() returns a DataFrame of the same shape as original
- True if the value is missing
- False otherwise
- .sum() then adds up the number of missing values for each column.

Data cleanup

Variables

- Gender
- Age
- Hypertension
- Heart disease
- Ever married
- Work type
- Residence type
- Avg glucose level
- BMI
- Smoking status
- Stroke

Data shape

- 11 columns
- 4981 rows

Variables

- Gender
- Age
- Hypertension
- Heart disease
- Avg glucose level
- BMI
- Smoking status
- Gender encoded
- Smoking status encoded

Data shape

- 9 columns
- 4981 rows

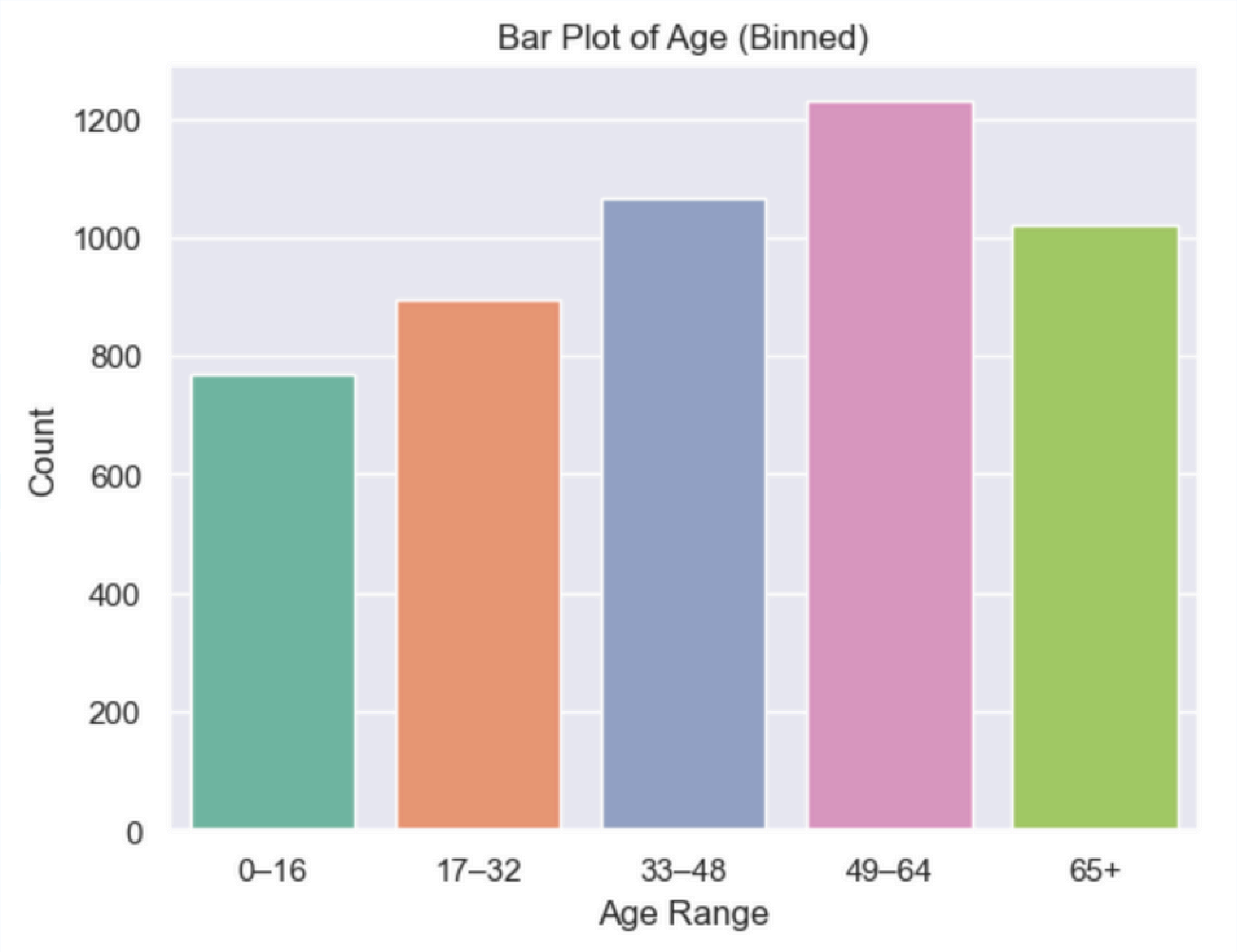
Encoding categorical variables

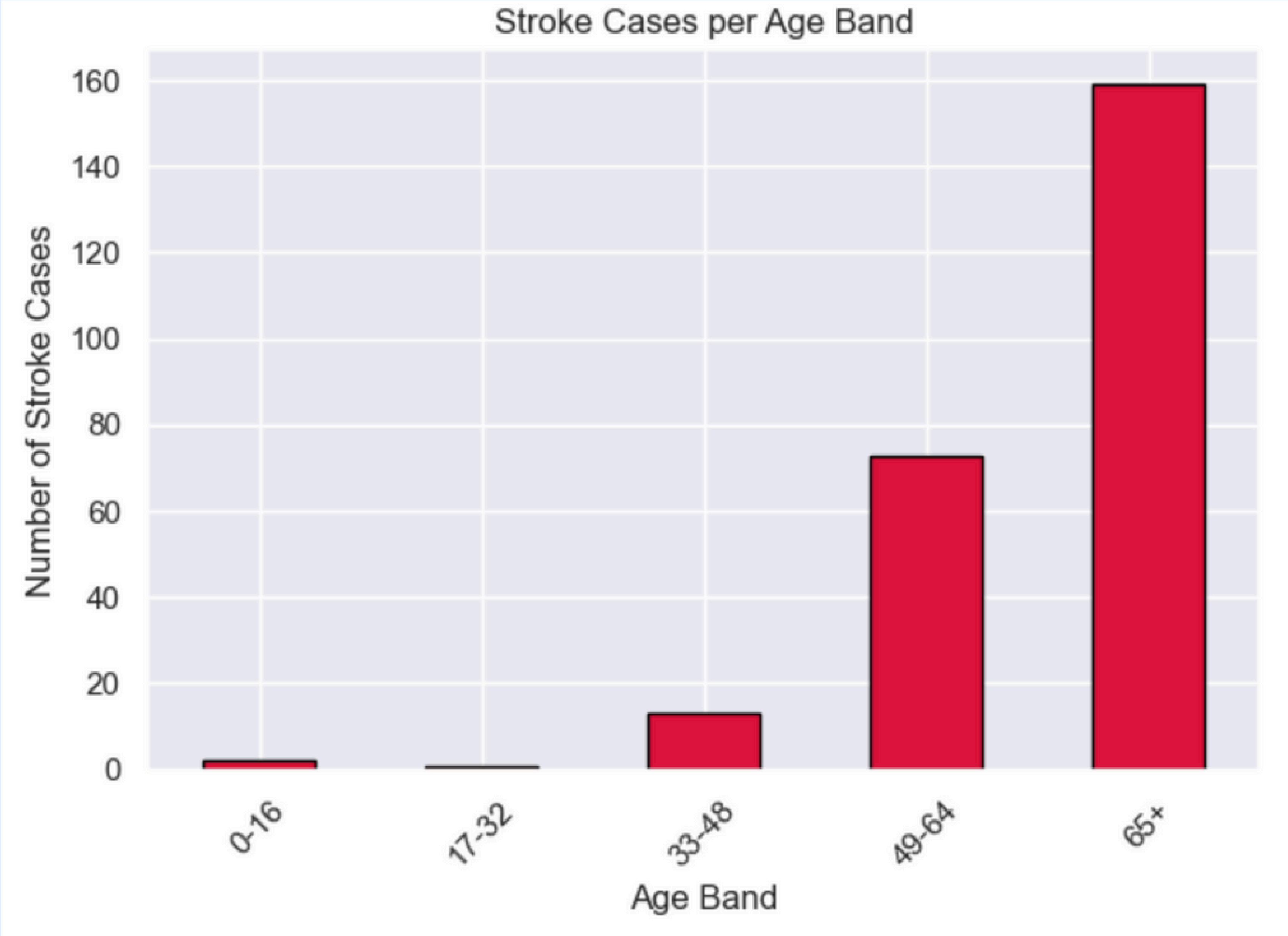
Encoding data is the process of converting categorical (non-numeric) values into a numeric format

```
for col in categorical_cols:
    data_cleaned[col + '_encoded'] = data_cleaned[col].astype('category').cat.codes
```

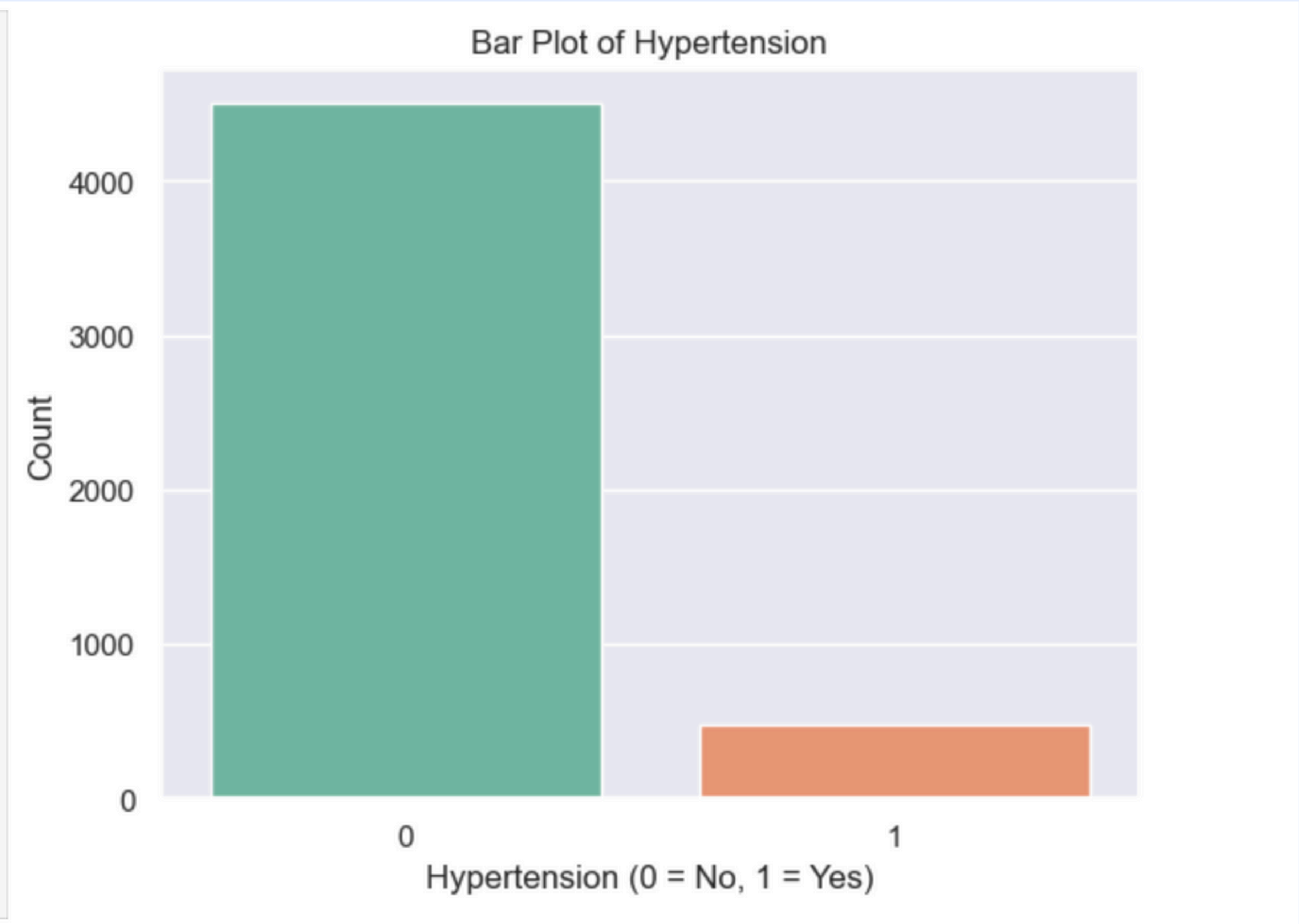
1. Gender
2. Hypertension
3. Heart disease
4. Smoking status

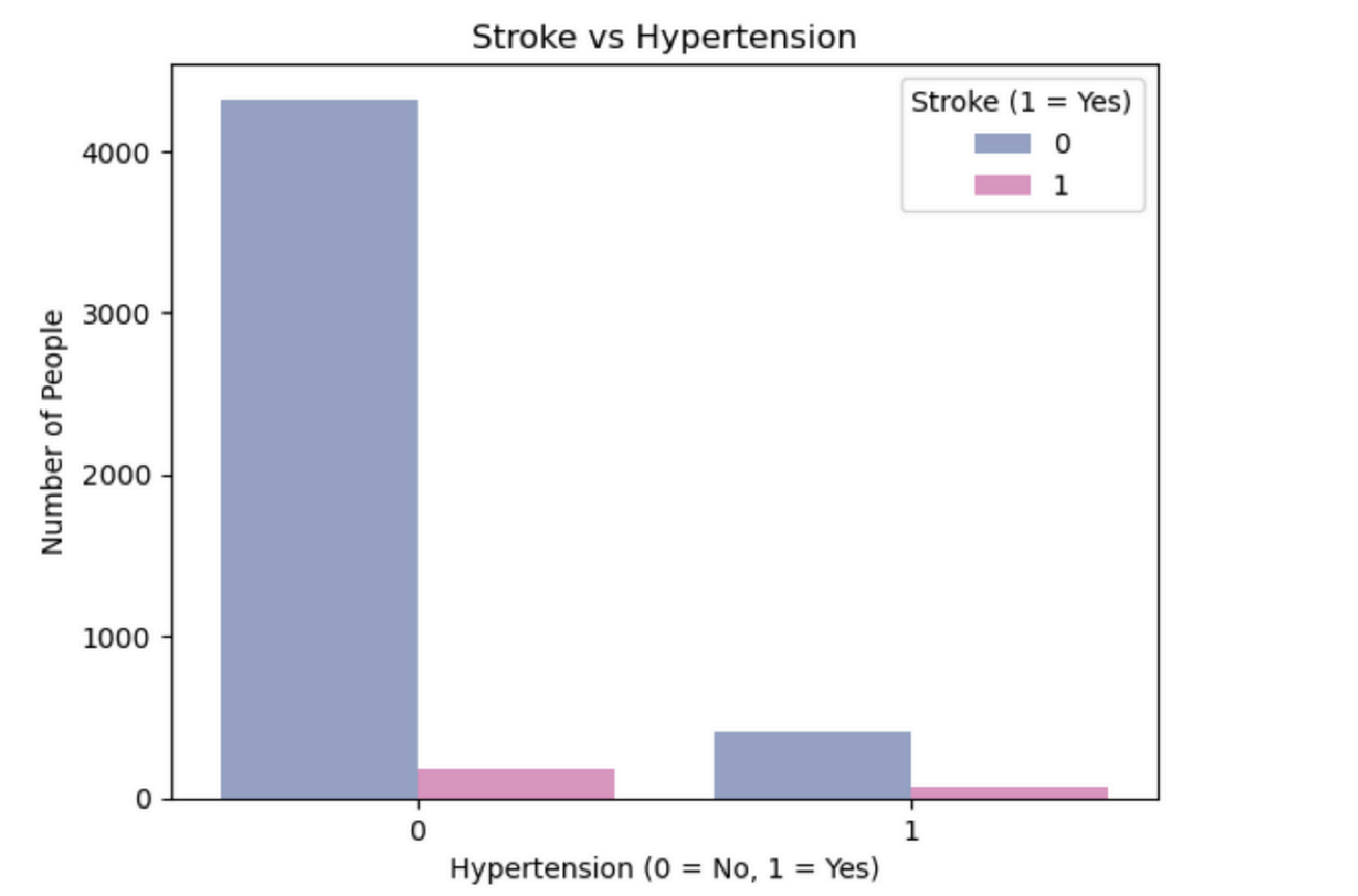
Exploratory data analysis





Exploratory data analysis

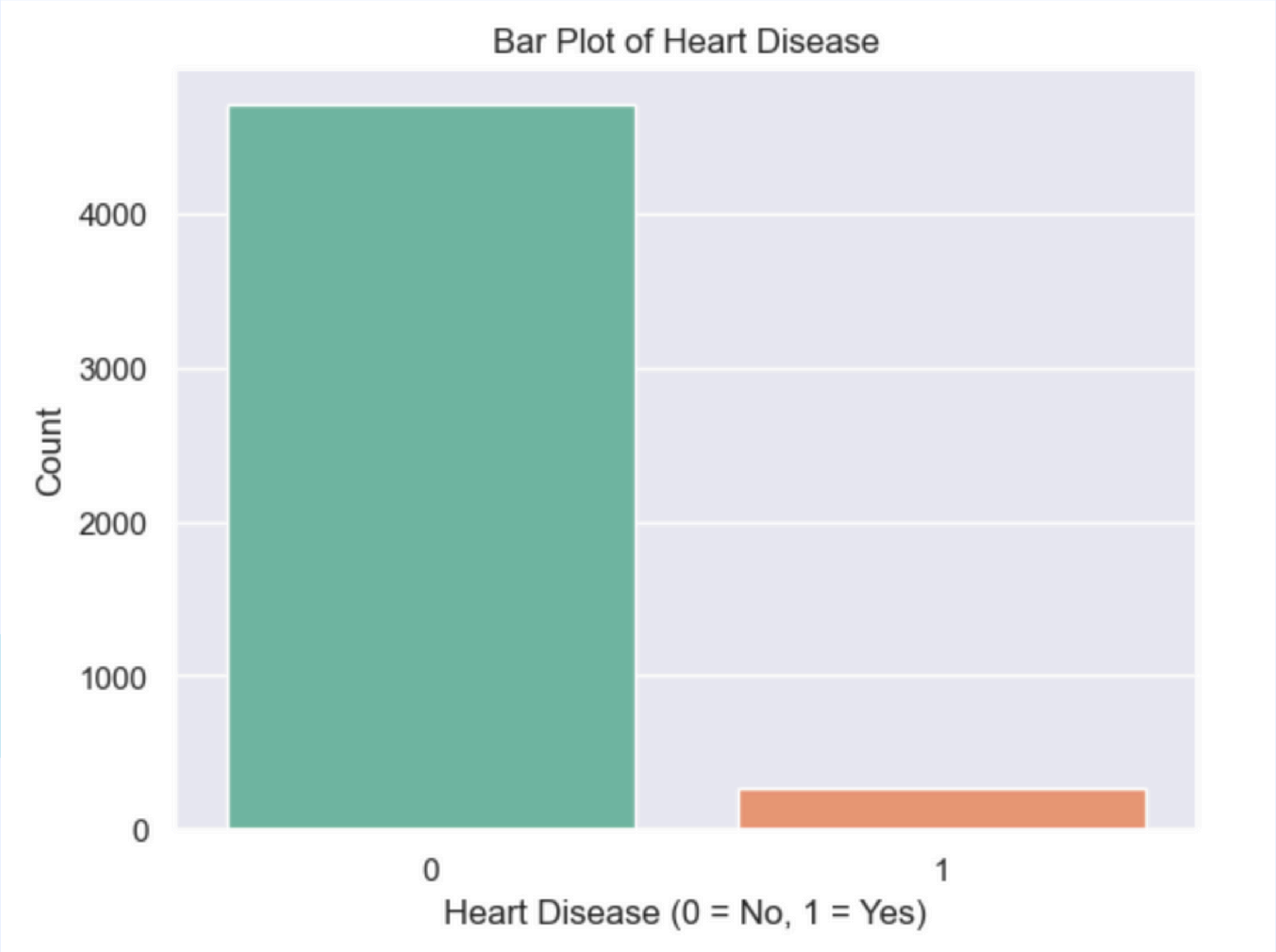


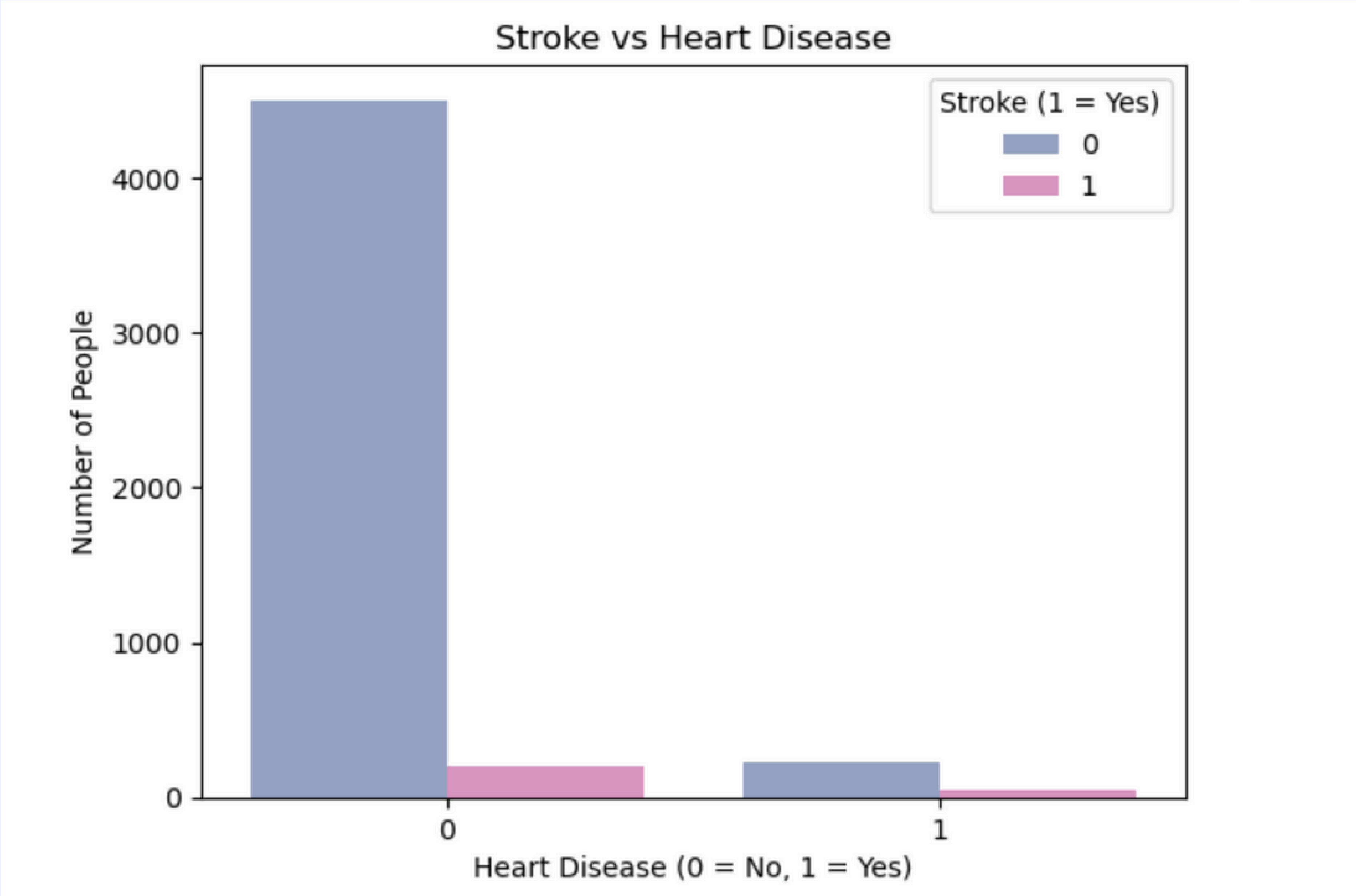


No Hypertension:
Stroke rate = 4.04%

With Hypertension:
Stroke rate = 13.78%

Exploratory data analysis

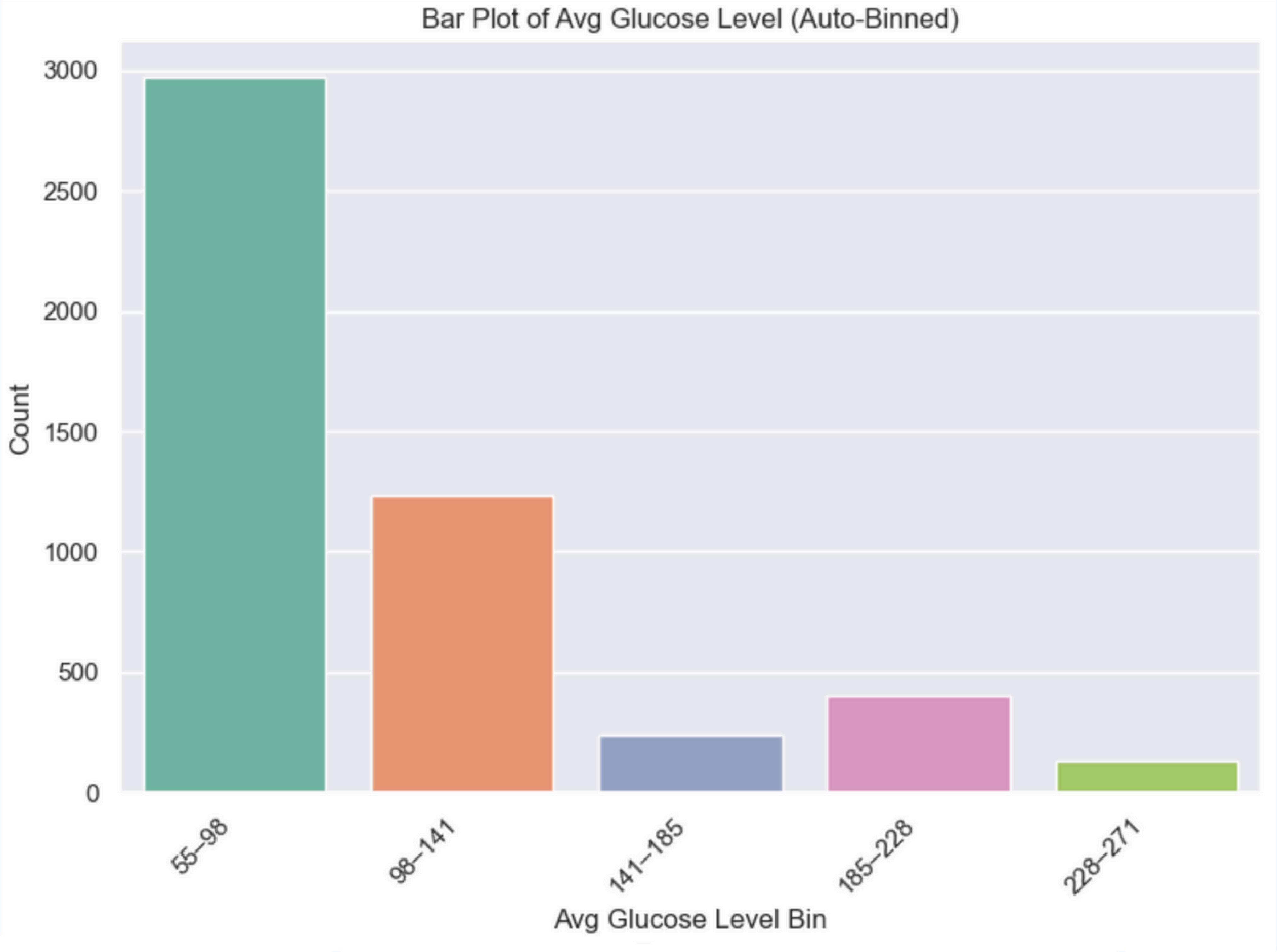




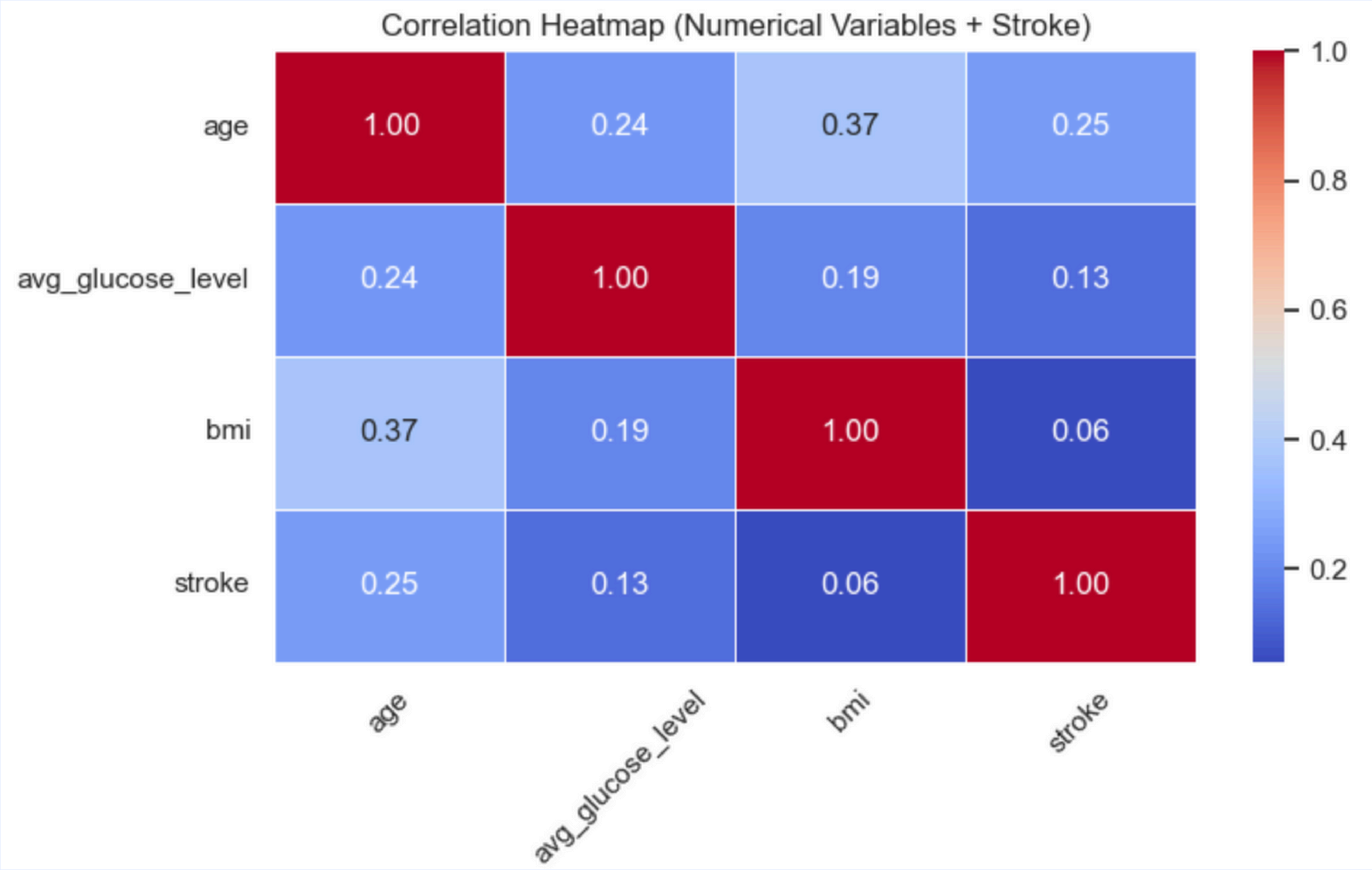
No heart disease:
Stroke rate = 4.27%

With heart disease :
Stroke rate = 17.09%

Exploratory data analysis



Correlation heatmap



Prediction models

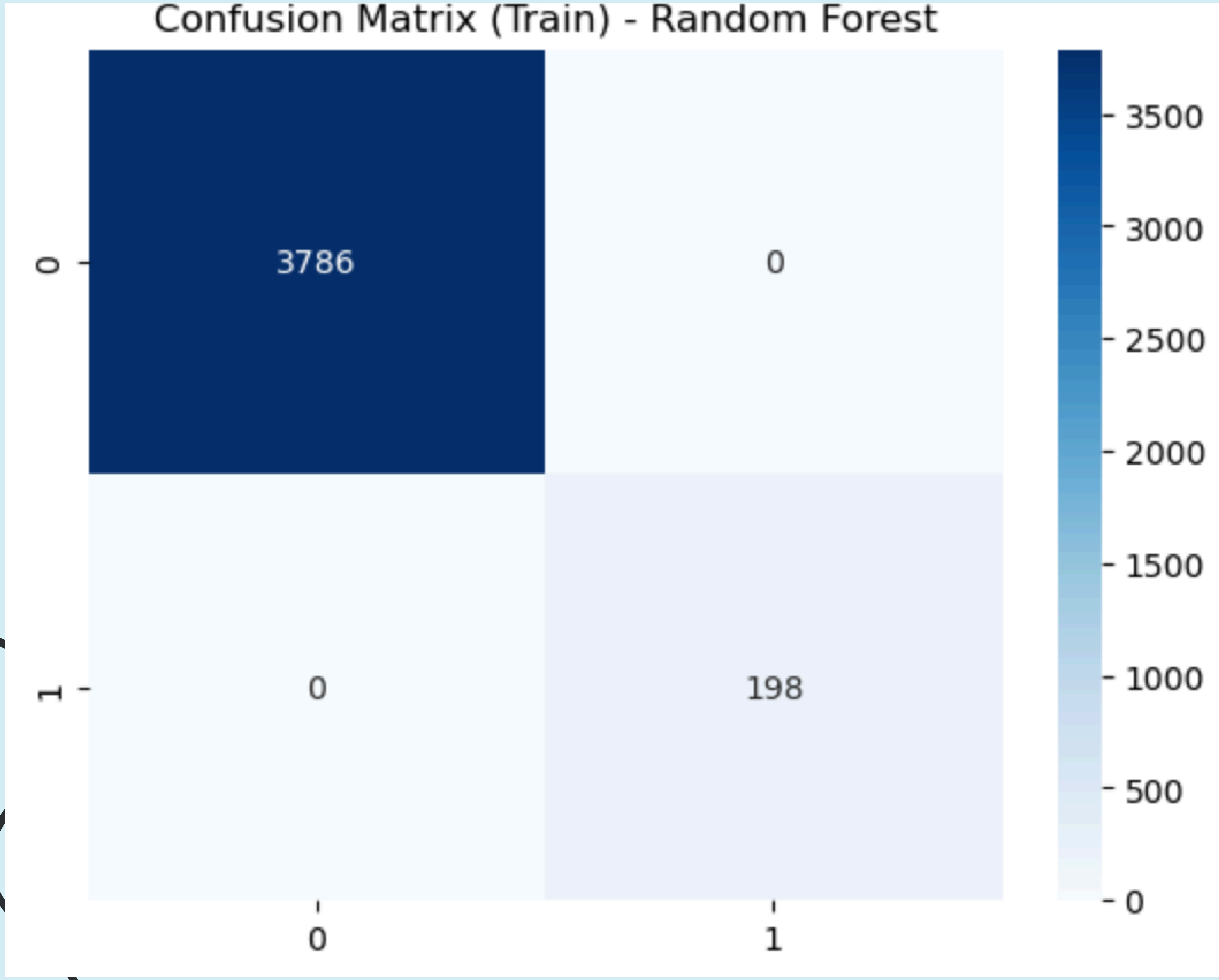
Logistical regression

Random forest

Naive Bayes

Random forest

Train

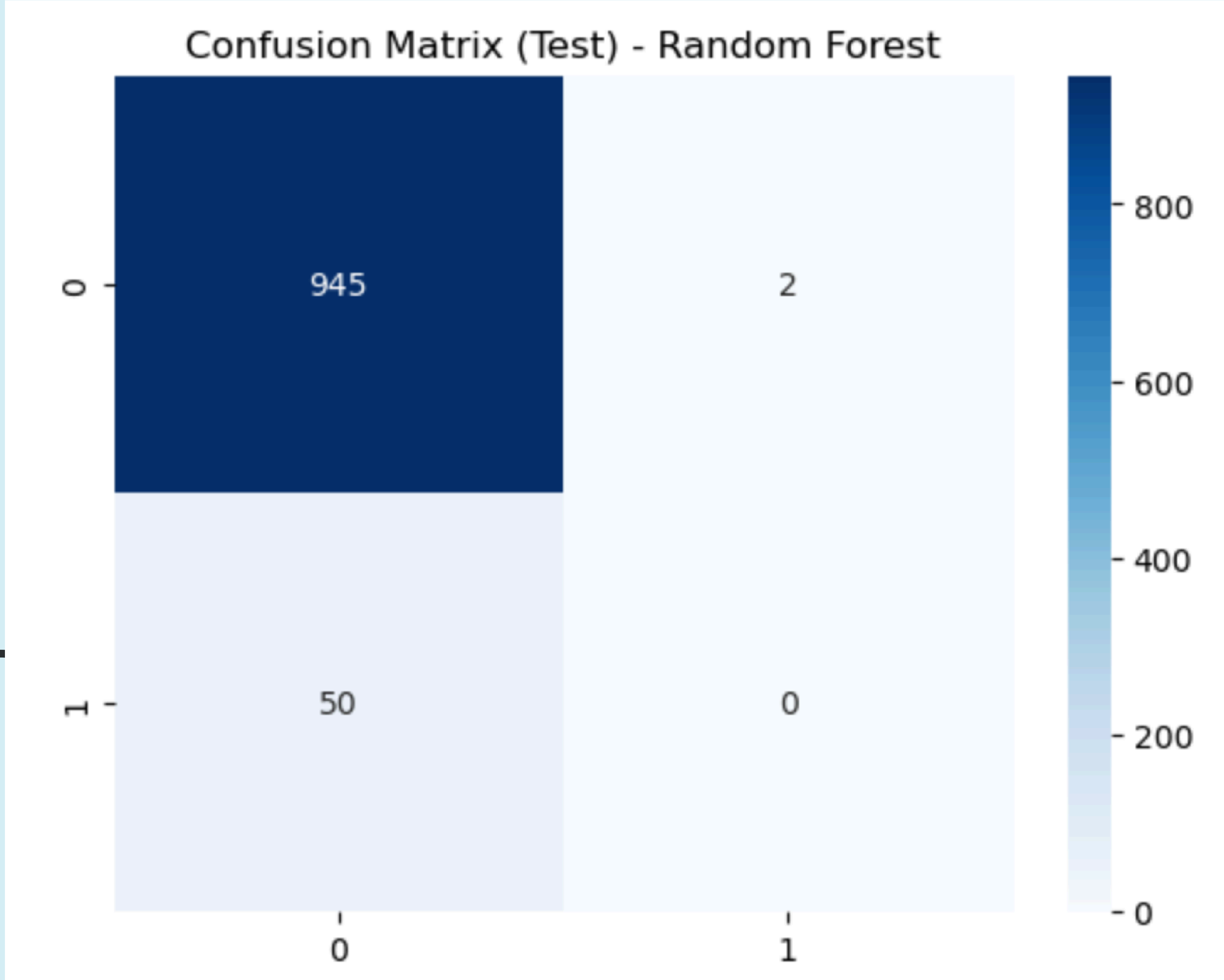


- An ensemble of decision trees that votes on the final prediction.
- It captures complex, non-linear patterns in the multivariate data but tends to overfit without balancing.

```
=== Random Forest - Train ===
Accuracy: 1.0000
TPR (Recall): 1.0000
FNR: 0.0000
TNR: 1.0000
FPR: 0.0000
```

Random forest

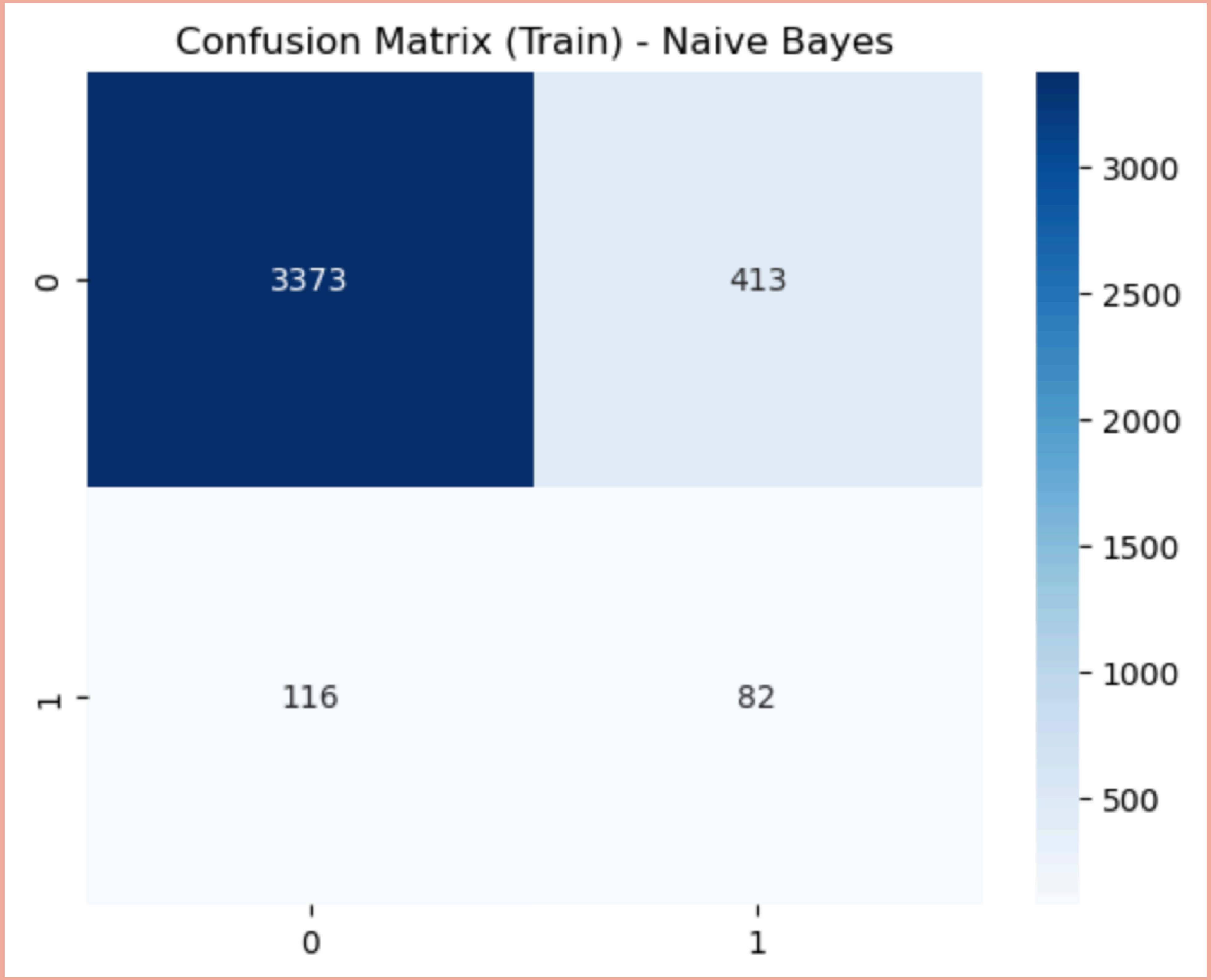
Test data




```
=== Random Forest - Test ===
Accuracy: 0.9478
TPR (Recall): 0.0000
FNR: 1.0000
TNR: 0.9979
FPR: 0.0021
```

NAIVE BAYES

Train data

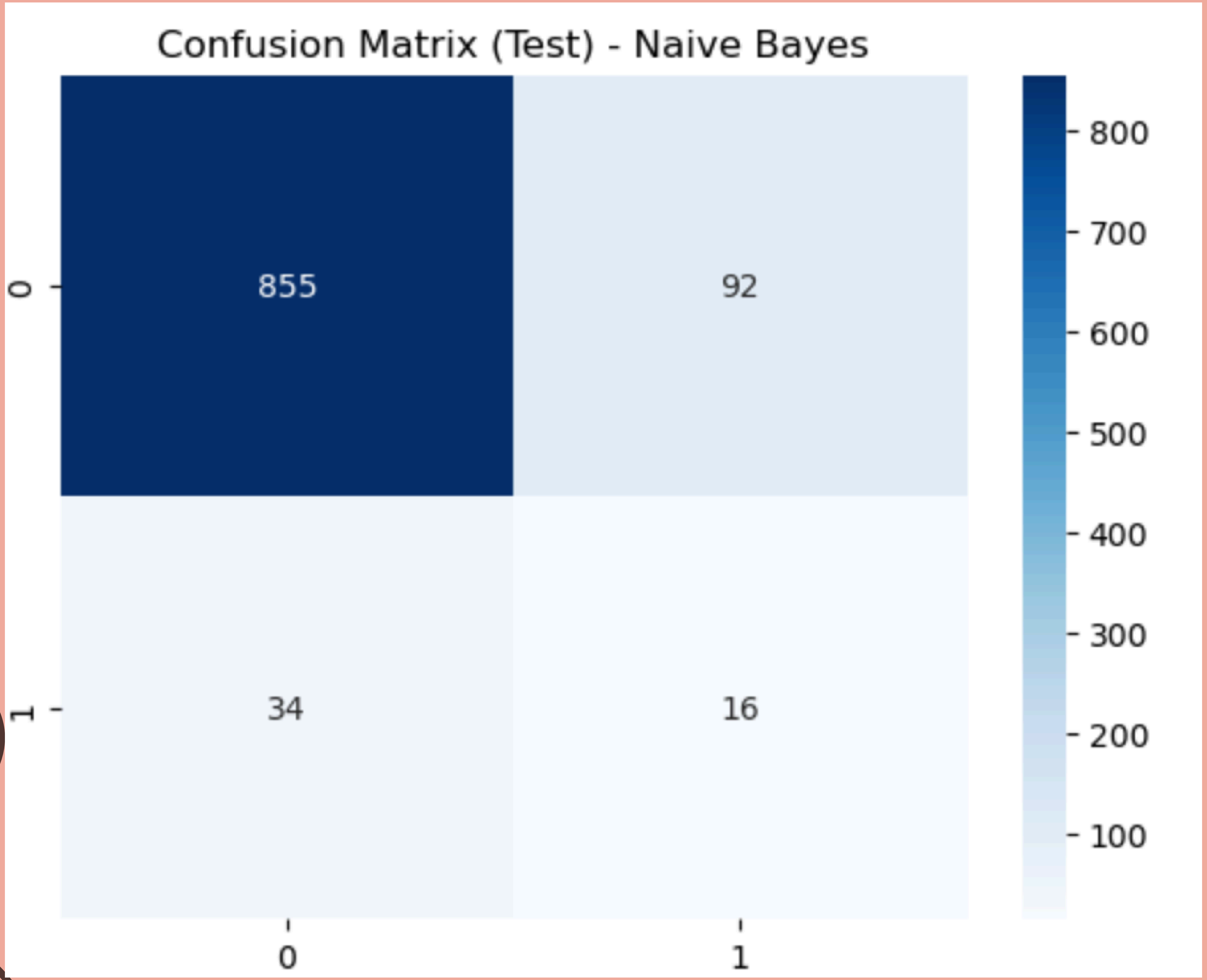


- A probabilistic model based on Bayes' Theorem that assumes feature independence.
- It's simple, fast, and worked well with categorical and numerical data in the dataset.

```
=== Naive Bayes - Train ===
Accuracy: 0.8672
TPR (Recall): 0.4141
FNR: 0.5859
TNR: 0.8909
FPR: 0.1091
```

NAIVE BAYES

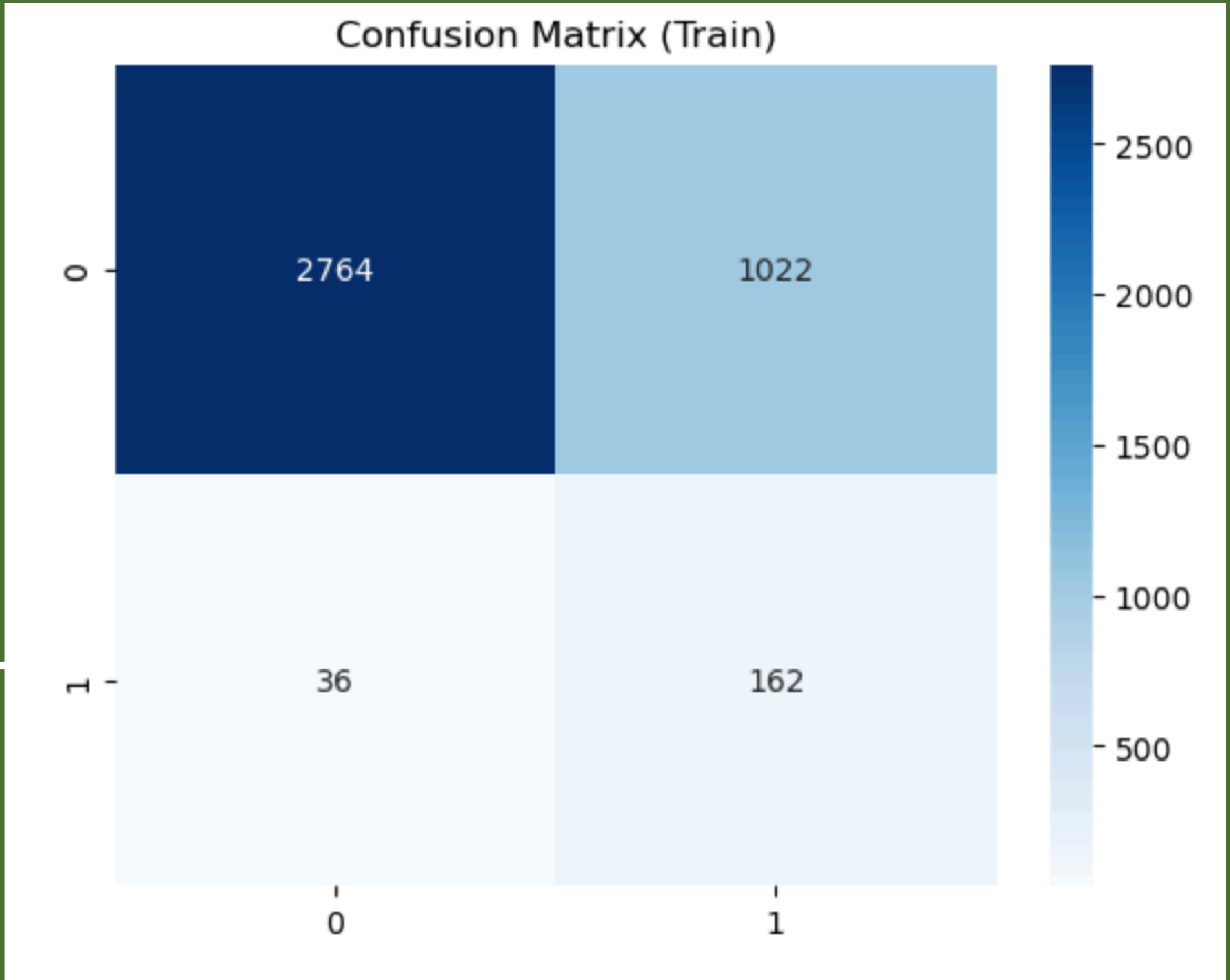
Test data




```
=== Naive Bayes - Test ===
Accuracy: 0.8736
TPR (Recall): 0.3200
FNR: 0.6800
TNR: 0.9029
FPR: 0.0971
```

Logistical regression

Train data

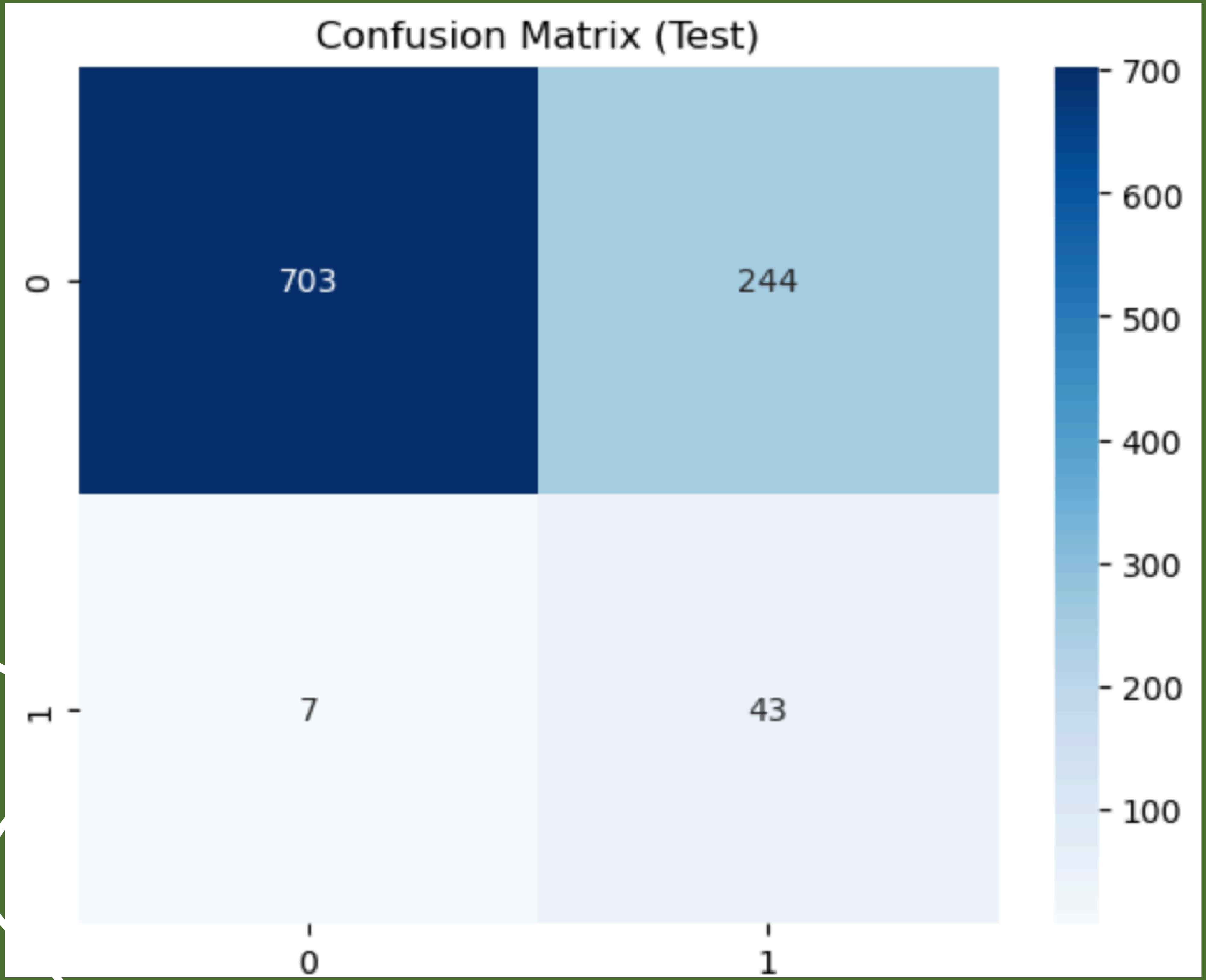


- A linear model used to predict binary outcomes.
- Uses class_weight='balanced' option to address class imbalance and improve sensitivity to the minority class (stroke cases).

```
=== Logistic Regression - Train ===
Accuracy: 0.7344
TPR (Recall): 0.8182
FNR: 0.1818
TNR: 0.7301
FPR: 0.2699
```

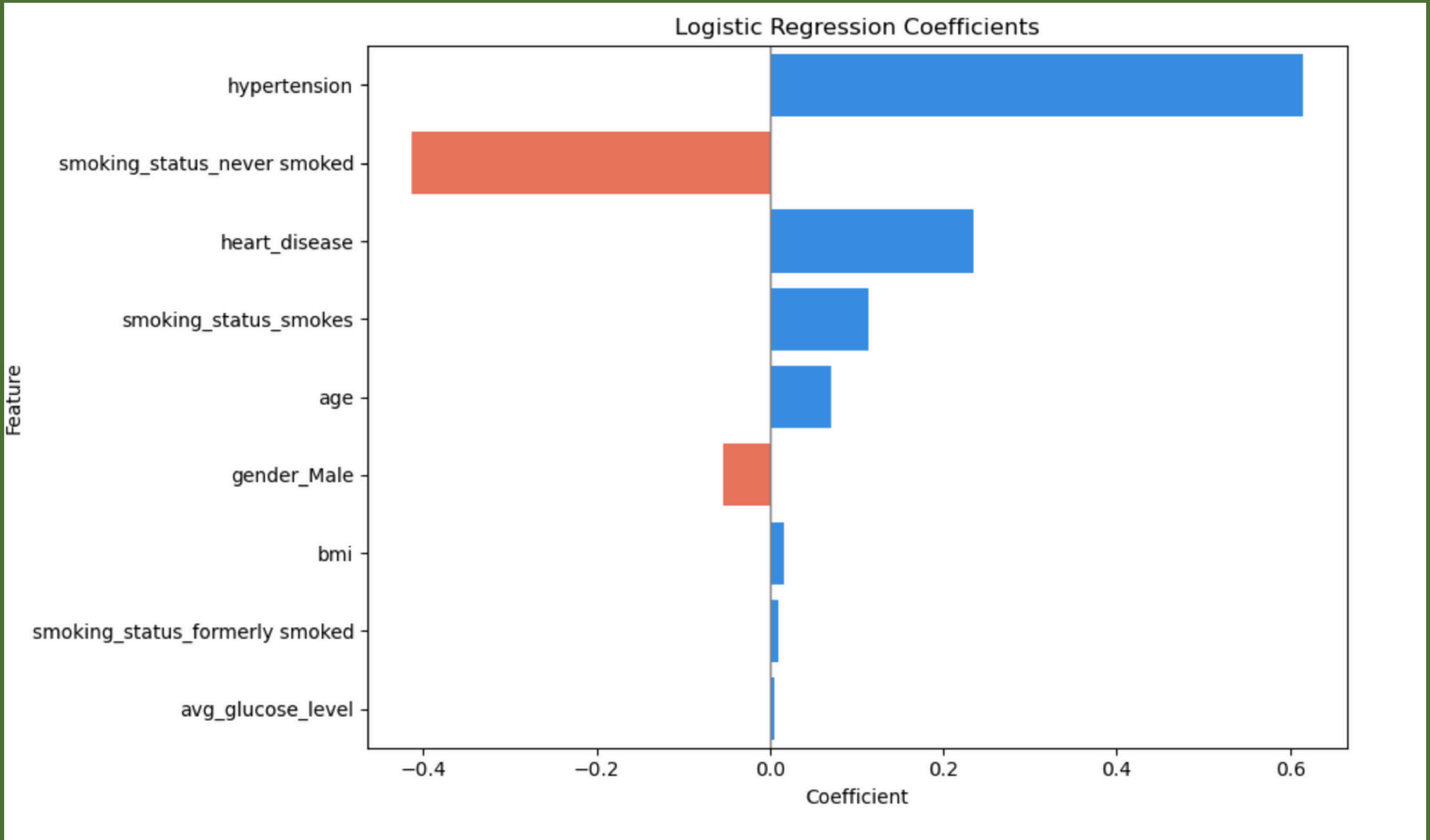
Logistical regression

Test data




```
=== Logistic Regression - Test ===
Accuracy: 0.7482
TPR (Recall): 0.8600
FNR: 0.1400
TNR: 0.7423
FPR: 0.2577
```

Logistical regression coefficients



Comparing models

Random forest

- Handles non-linearity
- High accuracy

- Can overfit
- Slower
- Less interpretable

Naive Baye's

- Simple
- Good with categorical features

- Assumes feature independence
- Lower accuracy

Logistical regression

- Works well with balanced data
- Fast training

- Struggles with non-linear patterns
- Performance depends on feature scaling

Insight

Reccomendation

Stroke rates are much higher among people with heart disease and hypertension

Prioritize screening and early intervention for individuals with cardiovascular conditions.

Higher average glucose levels are linked to increased stroke risk.

Promote lifestyle interventions for those with pre-diabetes or diabetes to prevent strokes.

Insight

Reccomendation

Smoking status plays a role in stroke likelihood

Even people who have quit smoking need to be monitored. Health campaigns should reinforce this fact.

Stroke risk increases steadily with age, but not all older individuals have equal risk.

Use age plus other variables in screening tools instead. Prevents overgeneralization and optimizes care.

Conclusion

- Developed problem statement for focus
- Cleaned and prepared data for neat analysis
- Analysed and interpreted given and derived data for insights
- Used machine learning and new learnings to analyse differnt models for prediciton
- Formulated our recommnedations for the best model that can be used through interepretation and analysis

Thank You