Addressing Critical Health Conditions through Data Science

Heart Disease: 16-18%

Cancer: 15-17%

Stroke: 10-12%

Chronic Lower Respiratory Diseases: 5-7%

Diarrheal Diseases: 3-5%

Surprising fact: 80% of strokes and Heart attacks are preventable

- Lifestyle Factors
- Medical Interventions
- Public Health Initiatives
- Regular Health Screenings

Data Sets

Heart Attack

Records: 8763

Features: 26

Age
Sex
Cholesterol
Blood Pressure
Heart Rate
Diabetes
Family History
Smoking
Obesity
Alcohol Consumption
Exercise Hours Per Week
Diet
Previous Heart Problems
Medication Use
Stress Level
Sedentary Hours Per Day
Income
BMI
Triglycerides
Physical Activity Days Per Week
Sleep Hours Per Day
Country
Continent
Hemisphere
Heart Attack Risk

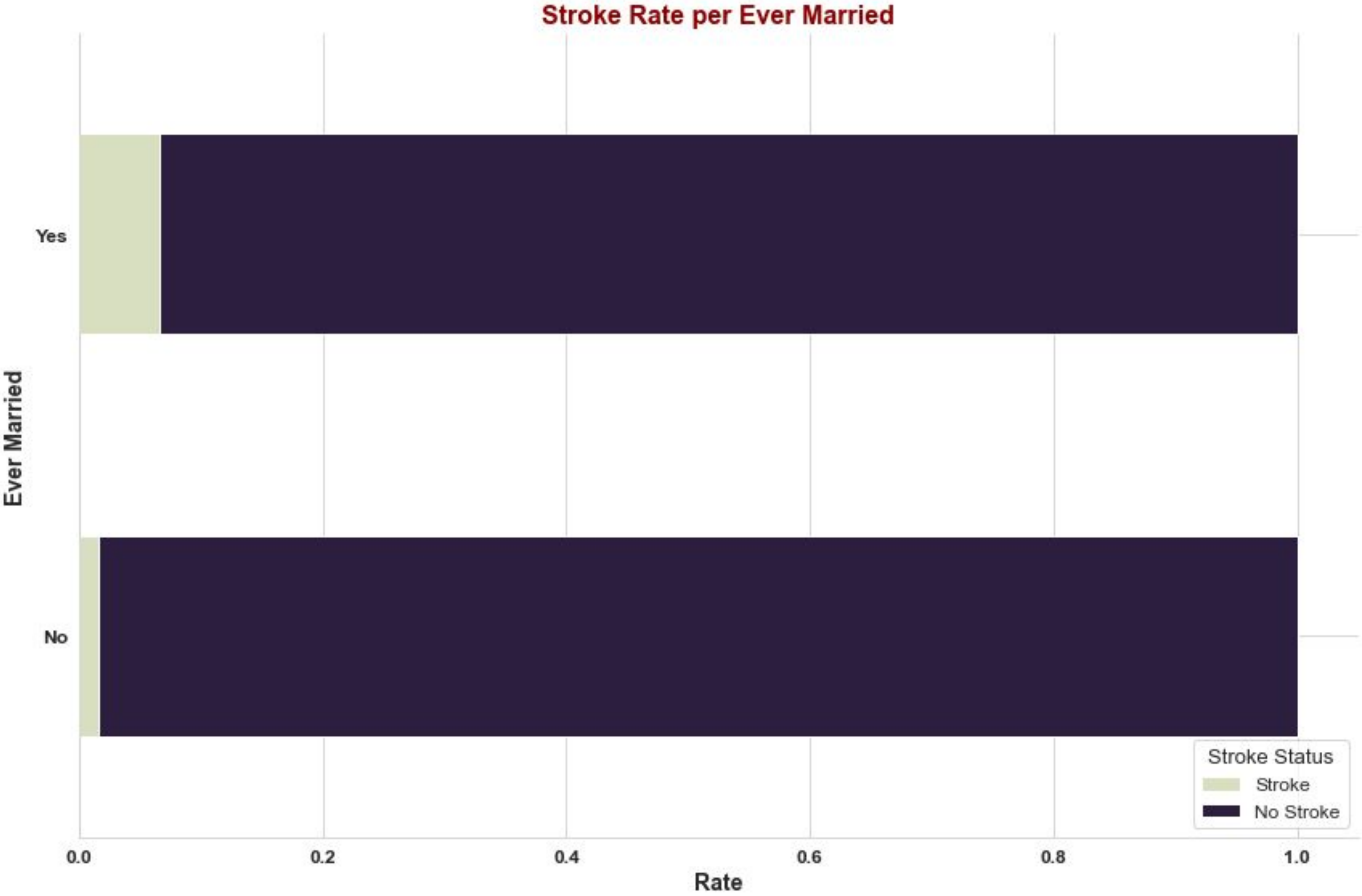
Stroke

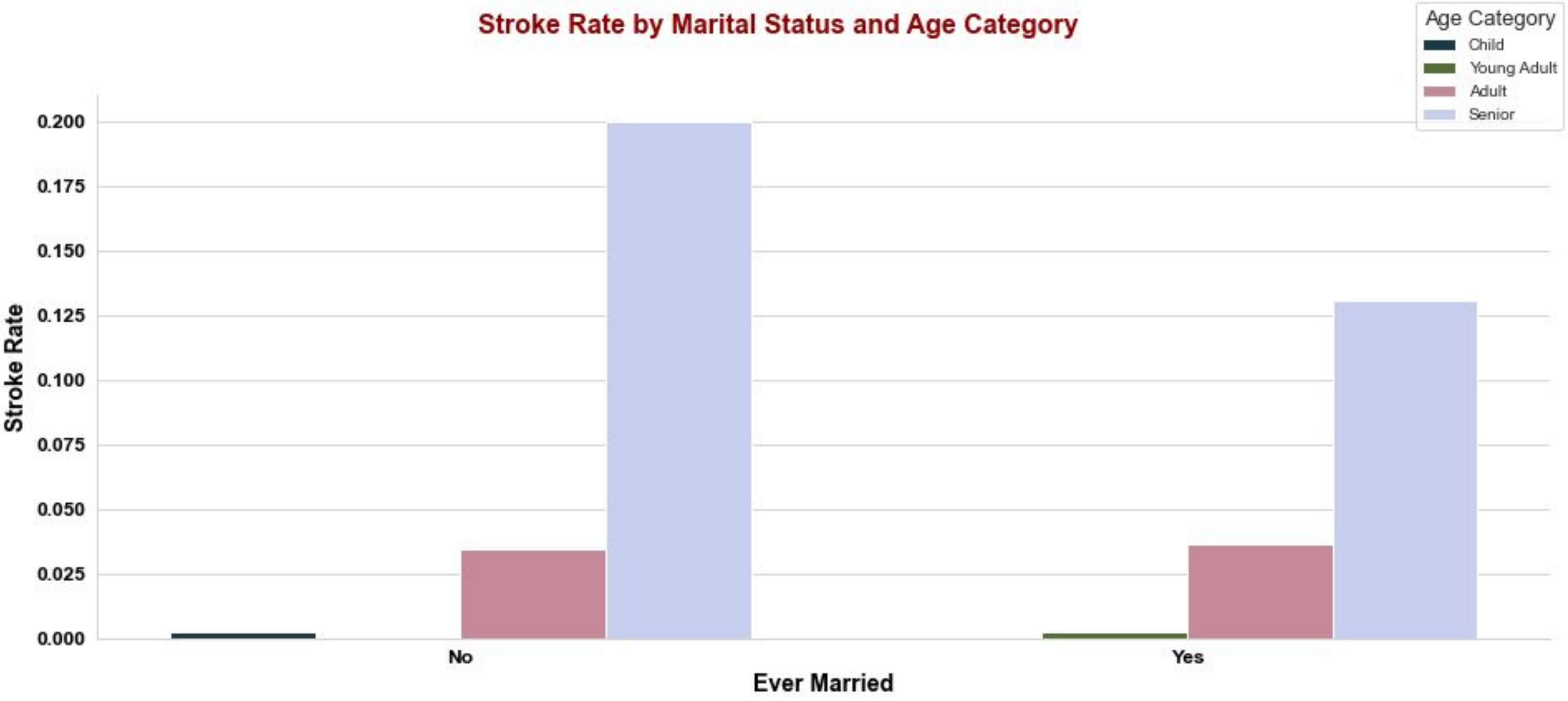
Records: 5110

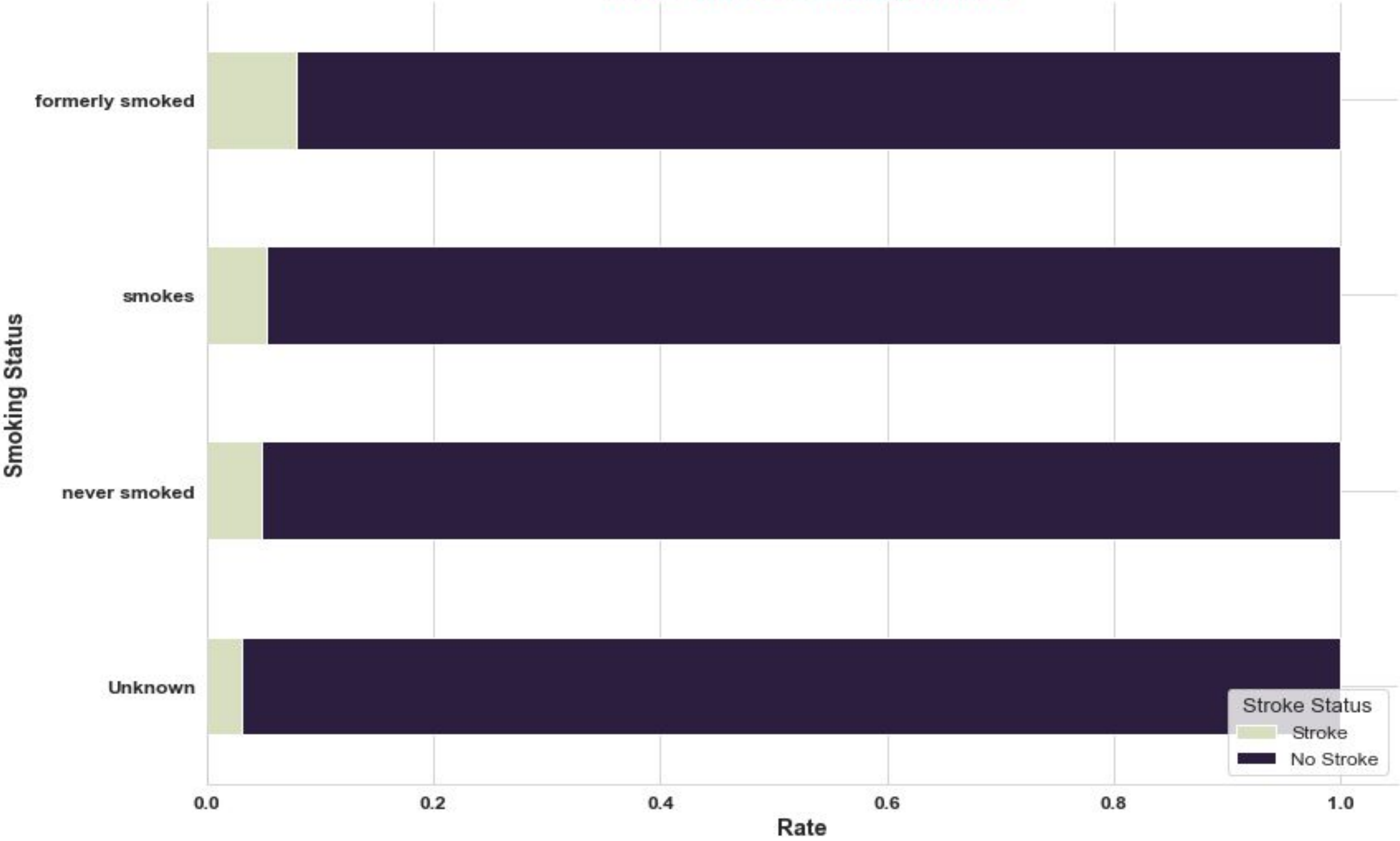
Features: 11

Gender
Age
Hypertension
Ever Married
Work Type
Residence Type
Glucose Level
BMI
Smoking Status
History of stroke
Heart Disease

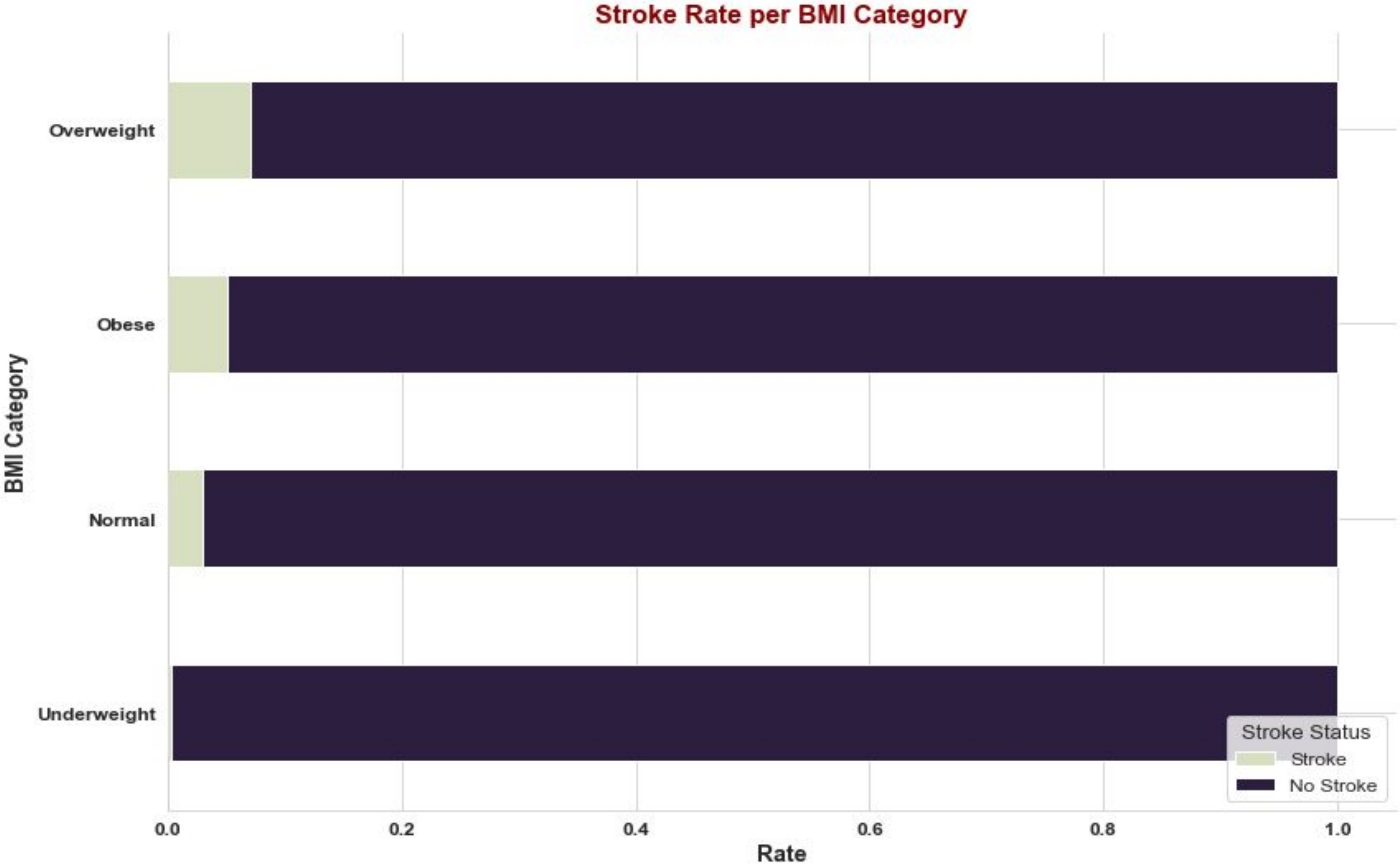
Challenges?

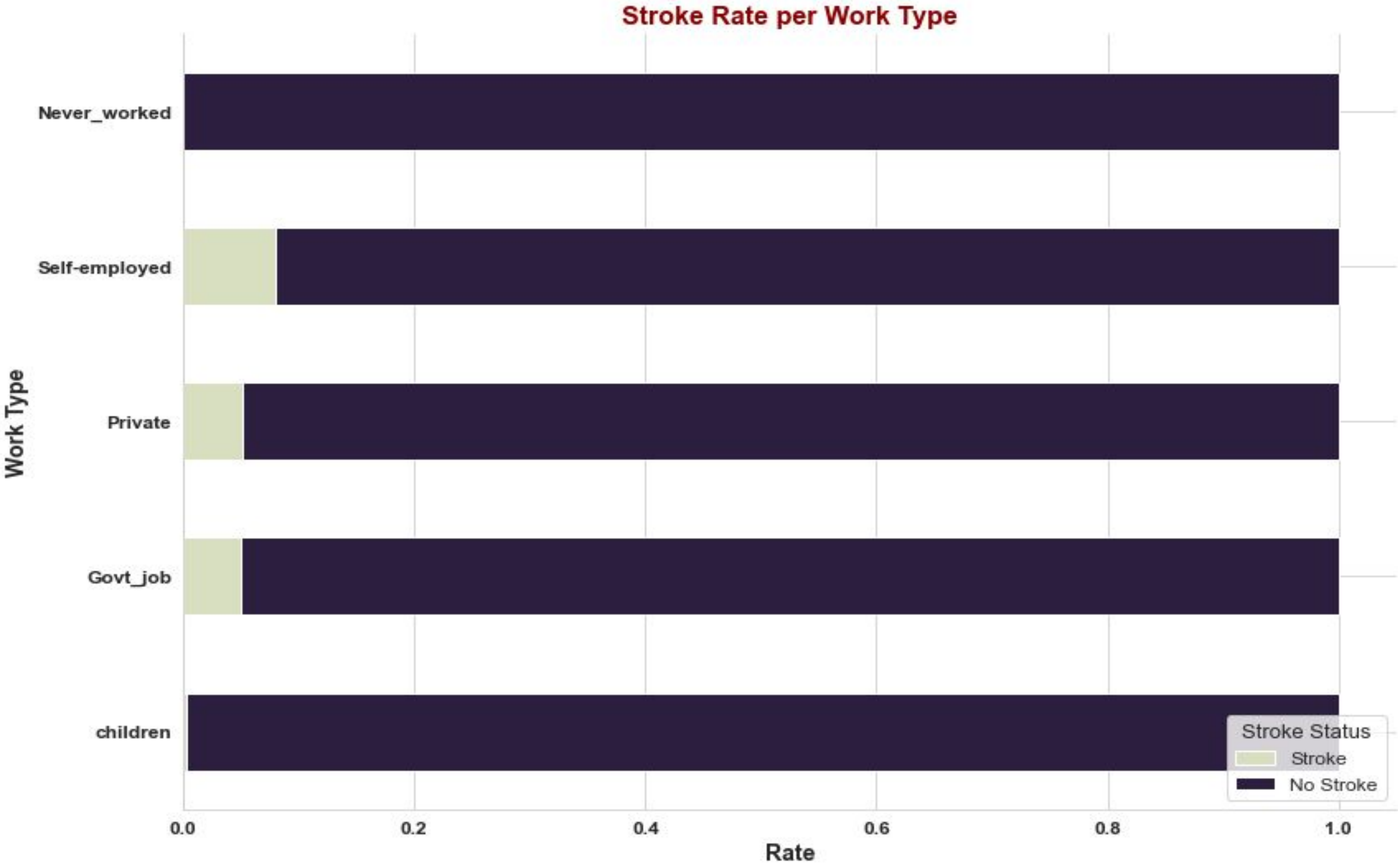

Stroke Rate per Ever Married
Ever Married
Yes
No
0.0
0.2
0.4
0.6
0.8
1.0
Rate
Stroke Status
Stroke
No Stroke


Stroke Rate by Marital Status and Age Category
Age Category
Child
Young Adult
Adult
Senior
Stroke Rate
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
No
Yes
Ever Married



Stroke Rate per Smoking Status


Stroke Rate per BMI Category
BMI Category
Overweight
Obese
Normal
Underweight
0.0
0.2
0.4
0.6
0.8
1.0
Rate
Stroke Status
Stroke
No Stroke


Stroke Rate per Work Type
Work Type
Never_worked
Self-employed
Private
Govt_job
children
Rate
0.0
0.2
0.4
0.6
0.8
1.0
Stroke Status
Stroke
No Stroke

Key Challenges

Data Quality and Reliability: Ensuring the datasets used are accurate and reliable.

Model Accuracy and Validation: Building models that are not only accurate but also clinically valid.

Handling Imbalanced Data: Dealing with the common issue of imbalanced datasets in medical data.

Ethical Considerations: Addressing data privacy and ethical concerns in predictive healthcare analytics.

Logistic Regression Performance

Train Accuracy: 81.74%

Test Accuracy: 77.69%

Recall: 50.0%

Precision: 11.0%

F1 Score: 18.0%

Brief note: Balanced recall but low precision, indicating a tendency to over-predict strokes.

K-Nearest Neighbors (KNN) Performance

Train Accuracy: 84.07%

Test Accuracy: 84.93%

Recall: 38.0%

Precision: 13.4%

F1 Score: 19.8%

Brief note: Higher accuracy and precision, but lower recall compared to Logistic Regression.

Decision Tree Performance

Train Accuracy: 83.40%

Test Accuracy: 77.50%

Recall: 50.0%

Precision: 10.9%

F1 Score: 17.9%

Brief note: Similar to Logistic Regression in recall and F1 Score, but slightly lower in precision.

Best ML model?

🎲 **KNN Accuracy Insights:** Exhibits superior test accuracy, yet caution is advised as accuracy metrics can be misleading in datasets balanced by SMOTE.

🎯**Recall and Precision Dynamics:** Logistic Regression and Decision Tree demonstrate enhanced recall, effectively identifying stroke instances but with a higher rate of false positives. KNN, conversely, shows improved precision but at the cost of lower recall.

⚖️ **F1 Score Analysis:** Across all models, the F1 scores are moderate, indicating ongoing challenges in achieving an optimal balance between precision and recall in balanced datasets.

🚀 **Model Selection Strategy:** The choice of model hinges on specific application requirements: prioritize Logistic Regression or Decision Tree for higher sensitivity to stroke detection, or opt for KNN for greater precision and fewer false positives.