Heart Health Analysis and Stroke Risk Prediction

DATA 240
Under the guidance of Dr.Taehee Jeong

Presented By: Group 6
Sweekruthi Balivada
Sneha Karri

Agenda

- Introduction
- Motivation
- Methodology
- Dataset
- Data Mining and Analysis
- Preprocessing
- Modelling
- Key Findings and Insights
- Summary
- References

Introduction

- The project aims to create predictive models that can identify the likelihood of heart strokes in individuals.
- Our main goal is to use machine learning methods to effectively predict the occurrence of heart strokes by considering a range of contributing factors.
- We performed thorough data analysis and utilized data mining methods, such as feature engineering , to extract valuable insights from the dataset.
- We developed five additional functionalities that include new variables and innovative metrics to improve the predictive capacity of our models.

Motivation

Why is it important to predict Heart stroke?

- Stroke is the second leading cause of death and the third leading cause of disability worldwide.
- Early detection and intervention can significantly reduce the risk and severity of strokes.
- The timely detection of people at risk enables healthcare providers to focus resources and interventions on those who are most in need.
- Provides options for customized healthcare management based on personal risk factors.
- Contributes to initiatives in public health aimed at lessening the impact of stroke-related illness and death.

Methodology

- Our objective is to develop predictive models for identifying stroke likelihood in a person.
- To achieve our goal, we have followed a Machine Learning life cycle where,
    - We have collected the data related to Heart stroke from various sources and merged these datasets together to conduct extensive research on different factors effecting heart strokes.
    - We have applied data mining techniques and also analyzed the data further using exploratory data analysis to get better insights from the data and to apply a better approach.
    - There by proceeding to Predictive modelling and their evaluation we have achieved our goal.

Comprehensive Dataset

- For this project, we have merged various datasets to perform comprehensive analysis.

Features in our dataset:

```
Index(['HeartDisease', 'BMI', 'Smoking', 'AlcoholDrinking', 'Stroke',
       'PhysicalHealth', 'MentalHealth', 'DiffWalking', 'Sex', 'AgeCategory',
       'Race', 'Diabetic', 'PhysicalActivity', 'GenHealth', 'SleepTime',
       'Asthma', 'KidneyDisease', 'SkinCancer'],
      dtype='object')
```

- Target Feature is “Stroke”

Data Mining and Analysis

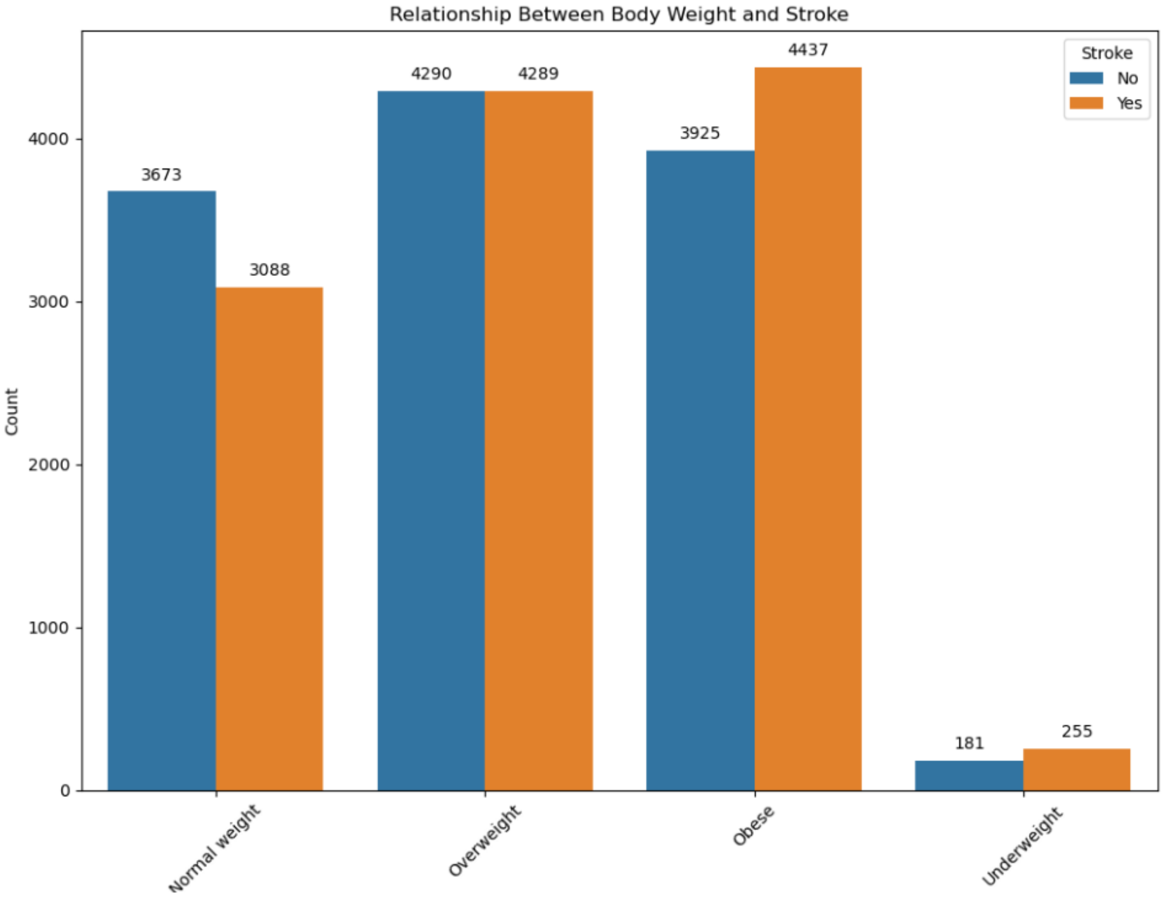
Relationship between BMI and Stroke

- Created a new feature - "Body_Weight" based on BMI (Body mass Index) feature.
- A person having BMI
    - <18.5 – Underweight
    - 18.5 – 25 – Normal weight
    - 25 – 30 – Overweight
    - >30 – Obese

3. Analyzed how Body weight effects Heart stroke in a person.

Observations:

- Count of obese people getting a stroke is more than the count of people not getting a stroke.
- Occurrence of Stroke in people having Normal weight is less.

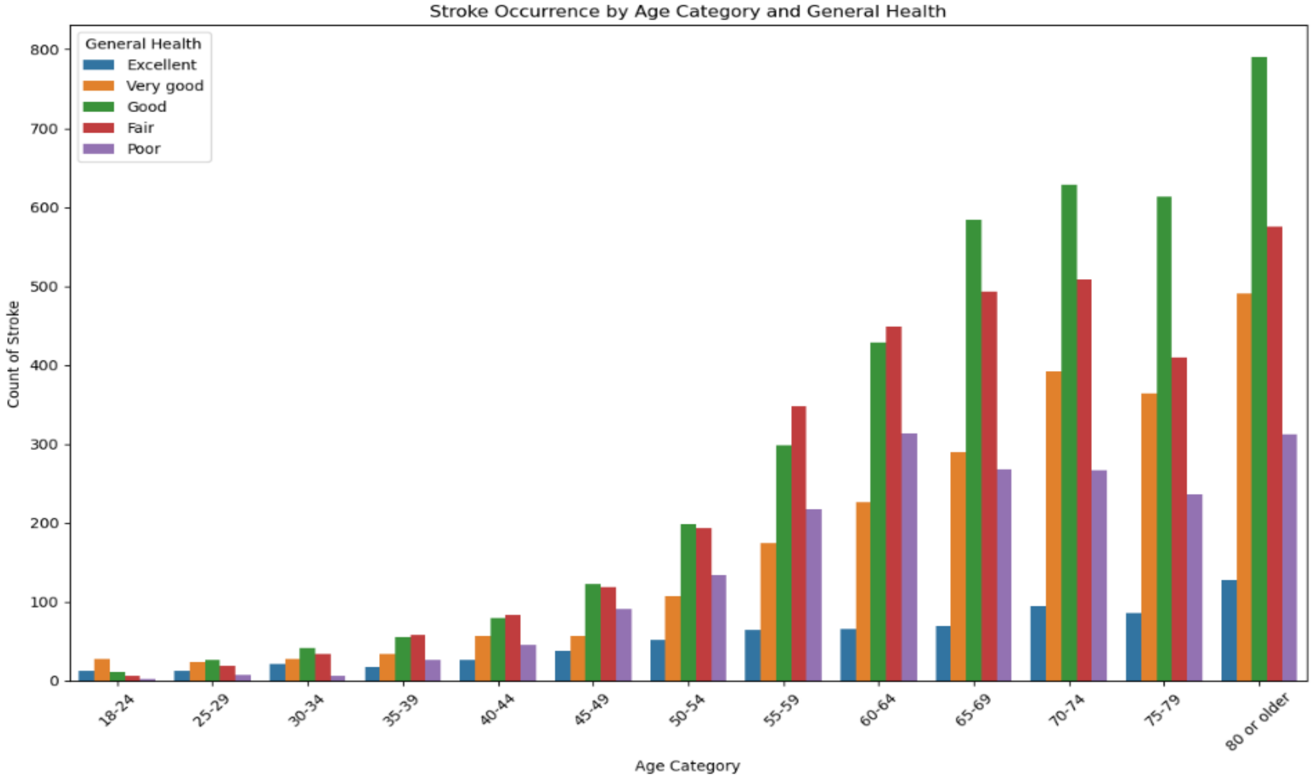


Relationship of Age and General Health with Stroke risk

- Analyzed relationship between Age, General Health and Likelihood of stroke occurrence in a person

Observations:

- A person having “Good” general health also is prone to heart strokes
- Ages between 50-64 are most vulnerable.



When Stroke= “Yes”, count of patients in each category

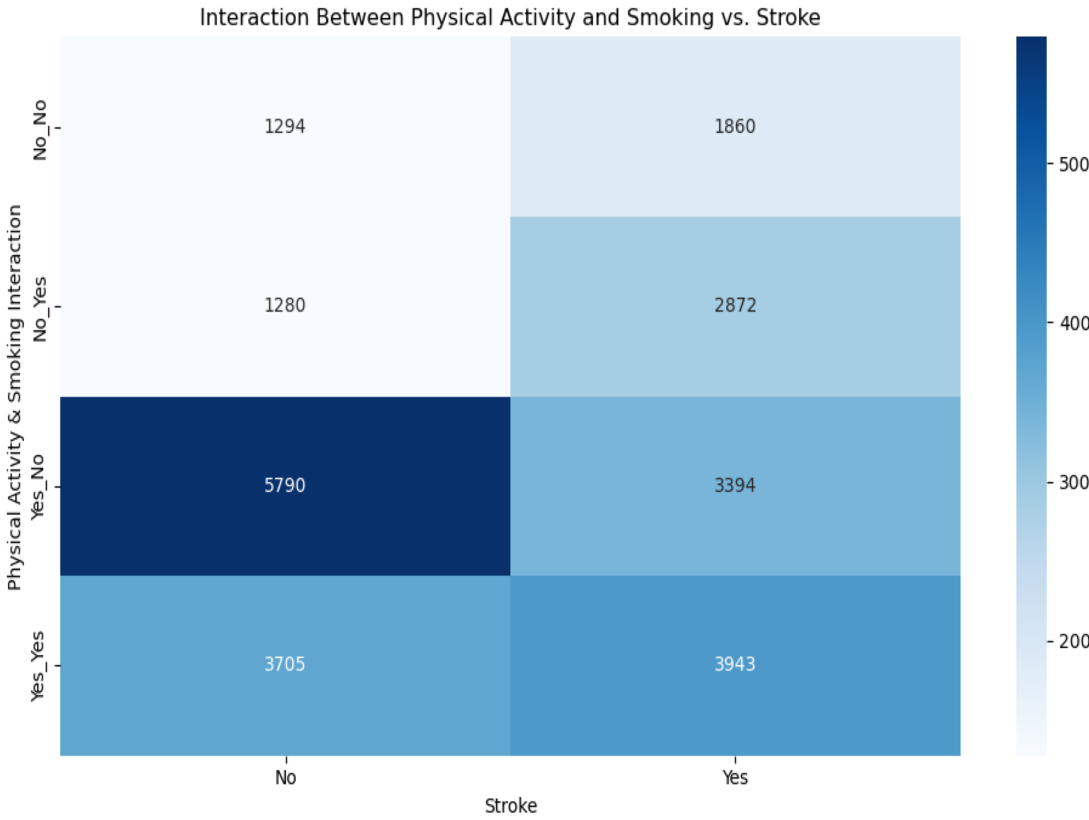
Good	3879
Fair	3298
Very good	2272
Poor	1929
Excellent	691

How having Physical Activity and Smoking habit will effect stroke occurrence?

- Analyzed how having physical activity and Smoking together will effect the likelihood of stroke occurrence in a person
- For this we have used feature engineering, which is a fundamental technique in Data mining.
- Combined two relevant features to capture a potentially meaningful interaction that might influence the likelihood of experiencing a stroke

Observations:

- People who are having physical activity and doesn't smoke are tend to live more and more likely to not get a Stroke.
- Physical activity helps in lowering the risk of strokes in people who smoke and don't smoke.
    - 69.2% - No Physical activity but are smokers
    - 51.5% - Yes Physical activity and are smokers
    - 58.9% - No physical activity and Non-smokers
    - 36.9% - Yes physical activity and Non-smokers

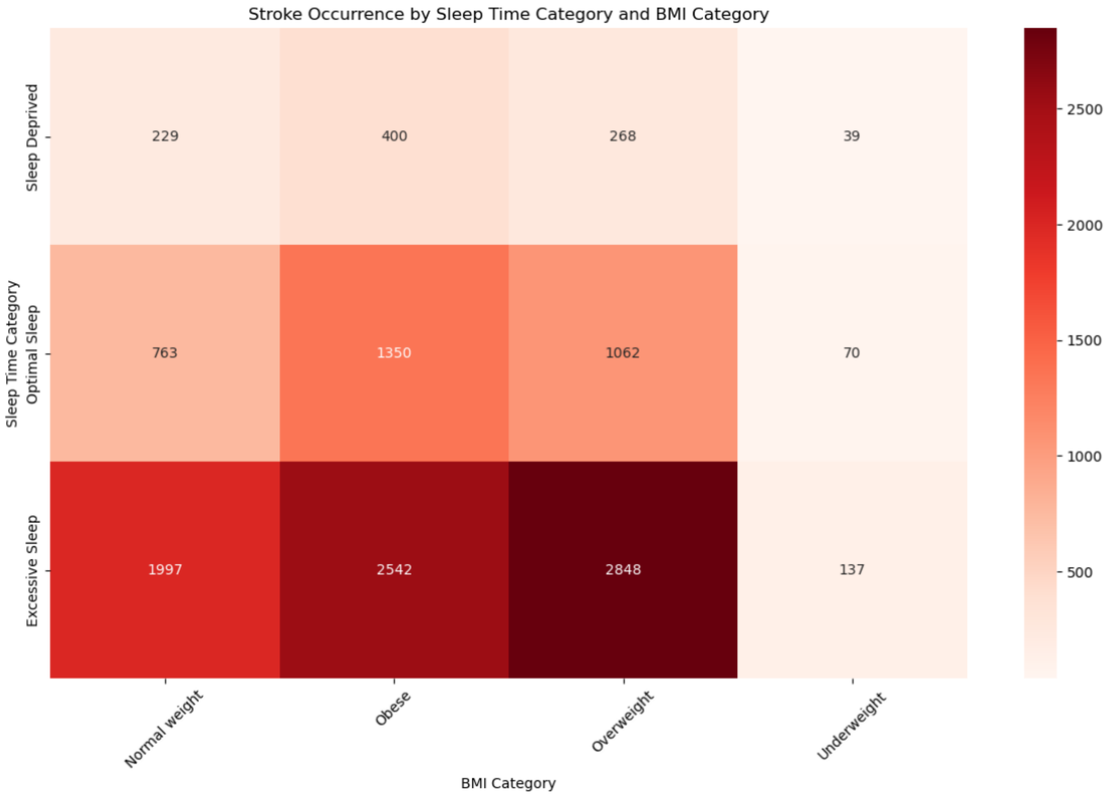


How Sleep duration effect Body weight as well as Stroke occurrence

- Created a new feature - “Sleep Category” based on Sleep duration feature.
- A person who sleeps for
    - < 5hrs – Sleep Deprived
    - 6 – 7hrs – Optimal Sleep
    - > 8hrs – Excessive Sleep

Observations:

- Higher chances of stroke occurrence in People who are Overweight and Obese, and tend to sleep more than 8hrs
- Lower counts are observed in people who Have optimal sleep schedules and have normal Weight.

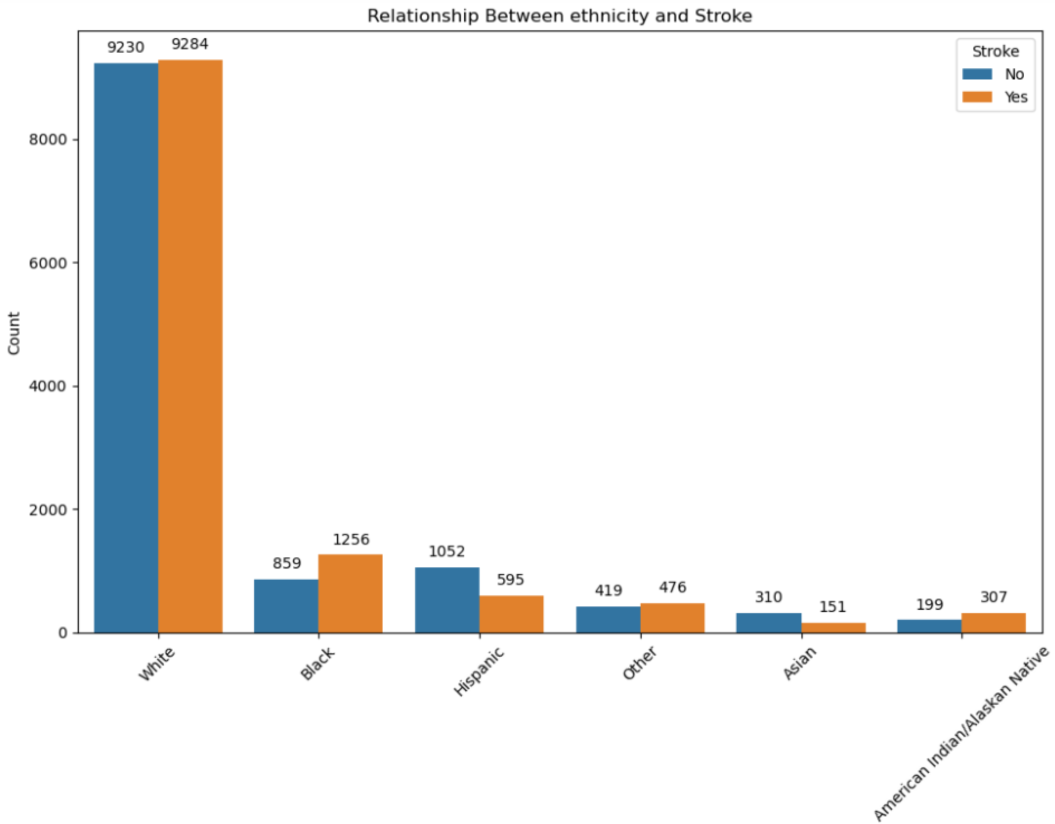


Relationship between Ethincity and Stroke

- Analyzed the relationship between people belonging to different races and Stroke occurrence in them.

Observations:

- Significantly less percentage of Asian and Hispanic people have strokes.
- Where as Blacks and American Indians show a significant raise in the count of stroke occurrences.

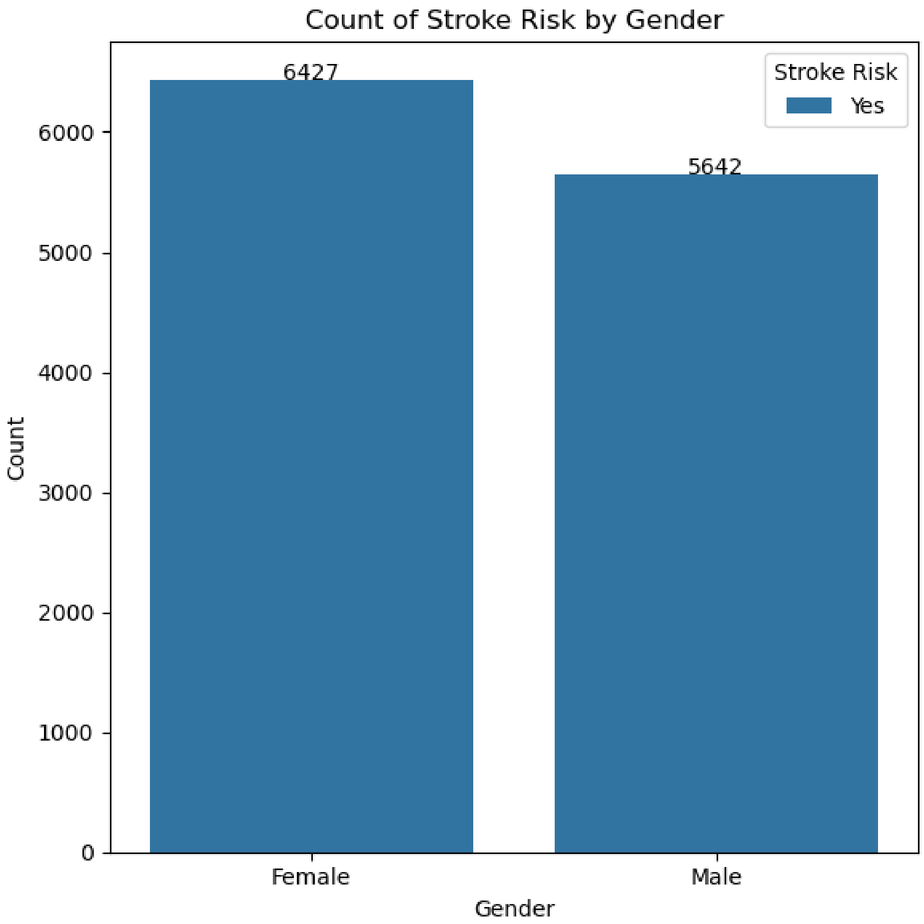


Relationship between Gender and Stroke

- Analyzed the relationship between Gender and Stroke occurrence

Observations:

- Significantly less is observed in Males.
- Females have higher chances of getting strokes in their lifetime



New Feature- Comorbidity Score

What is Comorbidity Score?

- A comorbidity score is a numerical measure in healthcare that assesses the severity of multiple chronic conditions a patient has, providing a standardized evaluation of their overall health.

Comorbidity	Score
Prior myocardial infarction	1
Congestive heart failure	1
Peripheral vascular disease	1
Cerebrovascular disease	1
Dementia	1
Chronic pulmonary disease	1
Rheumatologic disease	1
Peptic ulcer disease	1
Mild liver disease	1
Diabetes	1
Cerebrovascular (hemiplegia) event	2
Moderate-to-severe renal disease	2
Diabetes with chronic complications	2
Cancer without metastases	2
Leukemia	2
Lymphoma	2
Moderate or severe liver disease	3
Metastatic solid tumor	6
Acquired immuno-deficiency syndrome (AIDS)	6

```python
# Scores for each health condition
health_scores = {
    'HeartDisease': 3,
    'Diabetic': 1,
    'Asthma': 2,
    'KidneyDisease': 2,
    'SkinCancer': 2
}
```

	HeartDisease	Diabetic	Asthma	KidneyDisease	SkinCancer
265589	0	0.0	1	0	0
74795	0	1.0	0	1	0
303183	0	0.0	0	0	0
209960	0	0.0	0	0	0
177613	0	1.0	0	0	0

	ComorbidityScore
265589	1.0
74795	2.0
303183	0.0
209960	0.0
177613	1.0

New Feature- Vulnerability

- Based on a person’s age and Comorbidity severity score, this feature is calculated.
- This feature has 4 categories:
  - Highly Vulnerable
  - Vulnerable
  - Less Vulnerable
  - Not Vulnerable
- We can help the healthcare professionals identify the vulnerable population and provide preventative care

```
          HeartDisease    BMI Smoking AlcoholDrinking Stroke  PhysicalHealth  \
319588               0  30.56      No              No    Yes              21
319619               1  39.31      No              No    Yes               3
319620               1  27.64      No              No    Yes               1
319740               0  26.07      No              No    Yes               0
319765               1  38.45      No              No    Yes              30

          MentalHealth DiffWalking     Sex AgeCategory  ... GenHealth  \
319588               2         Yes    Male       50-54  ...      Good
319619               0         Yes  Female       65-69  ...      Fair
319620               0         Yes    Male       50-54  ...      Good
319740               0          No  Female       60-64  ...      Good
319765              15         Yes  Female       55-59  ...      Poor

          SleepTime Asthma KidneyDisease  SkinCancer  Body_Weight  \
319588            8      0             0           0        Obese
319619            4      1             1           0        Obese
319620            6      0             1           0   Overweight
319740            6      0             0           0   Overweight
319765            6      1             0           0        Obese

          PhysicalActivity_Smoking_Interaction  Sleep_Category  \
319588                                  Yes_No Excessive Sleep
319619                                   No_No  Sleep Deprived
319620                                  Yes_No   Optimal Sleep
319740                                   No_No   Optimal Sleep
319765                                  Yes_No   Optimal Sleep

         ComorbidityScore        Vulnerablity
319588                1.0      Not vulernable
319619                3.0   Highly vulnerable
319620                3.0     Less vulenrable
319740                0.0      Not vulernable
319765                3.0     Less vulenrable
```

Preprocessing

- In the Preprocessing stage we have balanced the data as the target variable was not balanced, using under sampling technique.
- Checked for missing values.
- Encoded the categorical values using Label Encoder

```
df.isnull().sum()
```

```
HeartDisease        0
BMI                 0
Smoking             0
AlcoholDrinking     0
Stroke              0
PhysicalHealth      0
MentalHealth        0
DiffWalking         0
Sex                 0
AgeCategory         0
Race                0
Diabetic            0
PhysicalActivity    0
GenHealth           0
SleepTime           0
Asthma              0
KidneyDisease       0
SkinCancer          0
dtype: int64
```

As the target variable is unbalanced lets apply resampling technique - Undersampling

```
majority_class = df[df['Stroke'] == 'No']
minority_class = df[df['Stroke'] == 'Yes']

# Undersampling the majority class
majority_undersampled = resample(majority_class,
                                 replace = False,
                                 n_samples = len(minority_class),
                                 random_state = 68)

balanced_df = pd.concat([majority_undersampled, minority_class])

print(balanced_df['Stroke'].value_counts())
```

```
No     12069
Yes    12069
Name: Stroke, dtype: int64
```

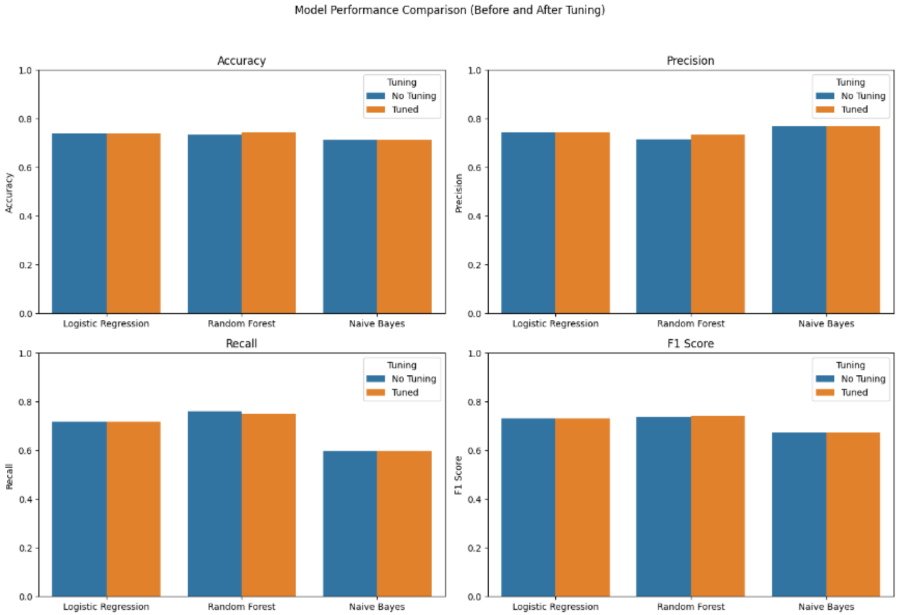
```
label_encoder = LabelEncoder()

for col in categorical_cols:
    balanced_df[col] = label_encoder.fit_transform(balanced_df[col])
print(balanced_df.head())
```

```
        HeartDisease    BMI  Smoking  AlcoholDrinking  Stroke  PhysicalHealth
265589             0  24.78        0                0       0               0
74795              0  26.63        0                0       0               8
303183             0  23.41        0                0       0               0
209960             0  23.67        0                0       0               0
177613             0  49.13        0                0       0               0

        MentalHealth  DiffWalking  Sex  AgeCategory  ...  GenHealth  \
265589             0            0    1            8  ...          0
74795              0            1    0           10  ...          2
303183             0            0    0            3  ...          0
209960             5            0    1            3  ...          4
177613             0            1    0            9  ...          2

        SleepTime  Asthma  KidneyDisease  SkinCancer  Body_Weight  \
265589          7       1              0           0            0
74795           9       0              1           0            2
303183          6       0              0           0            0
209960          5       0              0           0            0
177613          7       0              0           0            1

        PhysicalActivity_Smoking_Interaction  Sleep_Category  \
265589                                     2               0
74795                                      2               0
303183                                     0               1
209960                                     2               1
177613                                     0               0

        ComorbidityScore  Vulnerablity
265589               1.0             2
74795                2.0             2
303183               0.0             2
209960               0.0             2
177613               1.0             2

[5 rows x 23 columns]
```

Modelling

- Experimented with 3 different models: Logistic regression, Random Forest, and Gaussian Navie Bayes
- As per the evaluation metrics, Logistic Regression is our best model.

	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.740679	0.744904	0.718539	0.731484
1	Random Forest	0.736813	0.717060	0.767416	0.741384
2	Naive Bayes	0.715824	0.766880	0.606180	0.677126



Key Findings and Insights

- People who are over weight or obese are at a higher risk of suffering from stroke.
- As the age increases there are more chances of getting a stroke even with good health.
- Physical activity reduces the risk of getting a stroke greatly in non-smokers when compared to smokers.
- People who are Obese or overweight have a higher stroke occurrences. Being underweight doesn’t have much of an impact on stroke occurrences.
- Some ethnicities have a higher likelihood of getting a stroke compared to others.
- Females are prone to more strokes when compared with males
- Comorbidity score assess the previous health condition in a person that might have an impact on their likelihood of getting a stroke.
- Vulnerability is a customizable metric, used to a measure how vulnerable a person is.

Summary

To summarize, we have performed comprehensive data analysis as well as extensive data mining techniques such as feature engineering to analyze the factors which lead to Heart Strokes. We have also created 5 new features out of which 3 are derived from existing features(Body weight, Sleep Category, Physical Activity Smoking Interaction) and 2 new features(Comorbidity Score and Vulnerability) are formulized. To predict Heart stroke occurrences precisely we have experimented with 3 different Machine learning models out of which Logistic regression was our best model with a precision score of about 74%. Our models aim to predict the heart stroke occurrences precisely in a given population based on different features.

References

[1] Chen, A., Chen, D.O. “Simulation of a machine learning enabled learning health system for risk prediction using synthetic patient data”. Sci Rep 12, 17917 (2022). https://doi.org/10.1038/s41598-022-23011-4

[2] Sarah Friedrich, Stefan Groß, Inke R König, Sandy Engelhardt, Martin Bahls, Judith Heinz, Cynthia Huber, Lars Kaderali, Marcus Kelm, Andreas Leha. “Applications of artificial intelligence/machine learning approaches in cardiovascular medicine: a systematic review with recommendations”. European Heart Journal - Digital Health, Volume 2, Issue 3, September 2021, Pages 424–436. https://doi.org/10.1093/ehjdh/ztab054

[3] Elias Dritsas and Maria Trigka. “Stroke Risk Prediction with Machine Learning Techniques”. Department of Computer Engineering and Informatics, University of Patras, 26504 Patras, Greece,2022,22(13),4670. https://doi.org/10.3390/s22134670

Thank You