Predicting Stroke Risk Based on Lifestyle Habits

1.0 Introduction

A stroke occurs when blood flow to part of the brain is disrupted, either due to a blocked artery (ischemic stroke) or a ruptured blood vessel (hemorrhagic stroke). This interruption deprives brain cells of oxygen and nutrients, leading to potential brain damage, long-term disability, or death.

The goal of this project is to predict the risk of stroke based on key lifestyle variables, making it possible to identify high-risk individuals early and help prevent strokes through timely lifestyle changes or medical interventions.

2.0 Exploratory Data Analysis (EDA)

Some interesting questions regarding the relationships between the variables were explored, including: a) How does BMI change with age? (b) How does smoking affect sleep? (c) What age groups smoke the most? (d) How does exercise affect sleep? (e) What is the relationship between smoking and BMI? (f) How does exercise frequency affect BMI? (g) How is exercise frequency related to age? (h) How is exercise frequency related to stroke risk? (i) How is age related to stroke risk? (j) How is BMI related to stroke risk? (k) How is sleep duration related to stroke risk? (l) How is smoking related to stroke risk?

Some of the questions were answered by analysis and visualization of the dataset variables, while others could not be directly inferred from the data.

Key insights from EDA:

1) Stroke risk was most prevalent in individuals who engaged in low levels of physical activity (Figure 1a). This can be explained by the fact that low exercise frequency is associated with poorer cardiovascular health, higher blood pressure, obesity, and reduced circulation, all of which increase the likelihood of developing a stroke.
2) The risk of stroke significantly increased after age 50 (Figure 1b). Aging is associated with physiological changes such as stiffening of blood vessels, higher blood pressure, and accumulation of cardiovascular risk factors, all of which increase the likelihood of stroke after age 50.
3) Average Body Mass Index (BMI) was higher in high-risk stroke individuals (Figure 1c). A higher BMI often reflects overweight or obesity, which is linked to hypertension, diabetes, and other metabolic conditions that increase the risk of stroke.

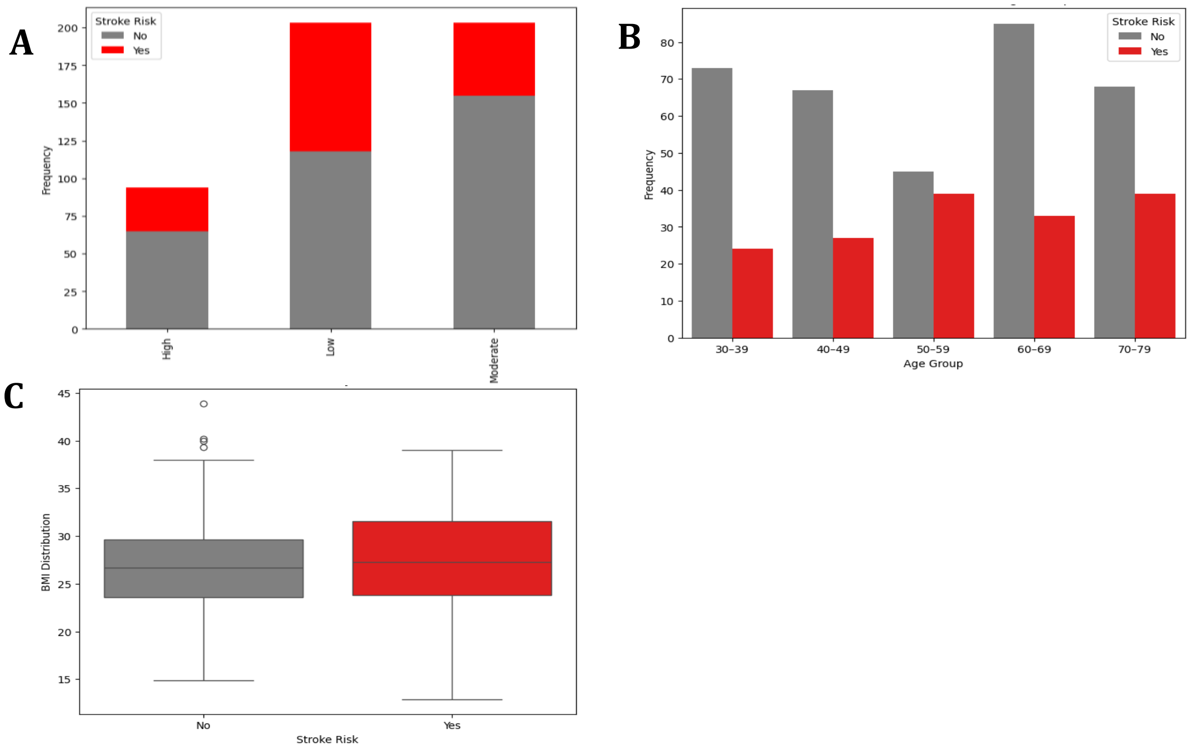

Figure 1: Exploratory data analysis. A) Stroke risk distribution with exercise frequency. The Low exercise frequency class contains more high-risk stroke subjects. B) Stroke risk distribution across age groups. Risk increases significantly after age 50 C) BMI distribution across the two stroke risk classes. Slightly higher average BMI in the high stroke risk class

3.0 Feature Engineering (FE)

S/N	FE Step	Method	Justification
1	Handing outliers in the bmi and sleep_hours columns	Winsorization	Winsorization preserves the overall sample size and distribution, while reducing the influence of extreme values on model performance.
2	Encoding the exercise_frequency and stroke_risk columns	Ordinal Encoding	Ordinal encoding preserves the inherent order of the categories
3	Encoding the smoking_status column	One-hot Encoding	The relationship between smoking categories and stroke risk was not inherently ordered, and one-hot encoding allows the model to treat each category independently, which improved predictive performance compared to using ordinal encoding
4	Adding a variable to indicate if the subject is below or above 50 (age_above_50)		Data visualization revealed a significant increase in stroke risk after age 50
5	Adding a discrete variable (exercise_frequency x age)		Low levels of physical activity may pose a higher risk in old age
6	Adding a continuous variable (exercise_frequency x bmi)		Low levels of physical activity may pose a higher risk in overweight or obese subjects

7	Adding a continuous variable (exercise_frequency x bmi x age_above_50)		Capture the interaction between exercise frequency, age, and whether the subject is above 50 or not on stroke risk
8	Adding a continuous variable (sleep hours x exercise_frequency)		Capture the effect of the combination of sleep hours and frequency of exercise on stroke risk
9	Adding a discrete variable (age^2)		Increase the weight given to age by the model
10	Adding a variable to Classify BMI based on WHO metrics (bmi_class)		Capturing clinically meaningful distinctions, rather than treating BMI as a continuous variable, may help the model learn non-linear relationships with stroke risk.
11	Dropping rows with BMI values in class 5 (35-40)		EDA revealed that most of the subjects in this class are not at risk of stroke. This contradicts the pattern of high BMI = high risk in the rest of the data and confuses the model. Dropping these values increased model performance by 16.25%
12	Adding a continuous variable (bmi x bmi_class)		The variability in BMI scores is inherently low, and weighting with the appropriate class increases the variability, thereby providing more information for the model.
13	Rounding the sleep_hours column to whole numbers		Reduced noise and improved model performance
14	Dropping columns ("stroke_risk", "age_group", "age", "age_above_50", "bmi", "sleep_hours")		High correlations between dropped columns and other columns contradicts the logistic regression assumption that variables are independent
15	Handling class imbalance	SMOTE and RandomUnderSampler	Upsampling to equalize produced worse results than the baseline model. SMOTE upsampling to 200 plus downsampling gave the best results.

4.0 Model Training and Evaluation

For predicting stroke risk, a logistic regression model was trained using the preprocessed dataset. Standard scaling was used to prevent features with larger scales from disproportionately influencing the model. Standard scaling gave a better performance than Min-max scaling.

A grid search was performed to optimize hyperparameters, including regularization strength (C), penalty, and solver. The best-performing configuration was: C = 1, Penalty = L1 (L1 regularization), Solver = saga, Max iterations = 5000. Different train-test splits were also evaluated, and a split of 80% training and 20% testing yielded the best results (Figure 2).

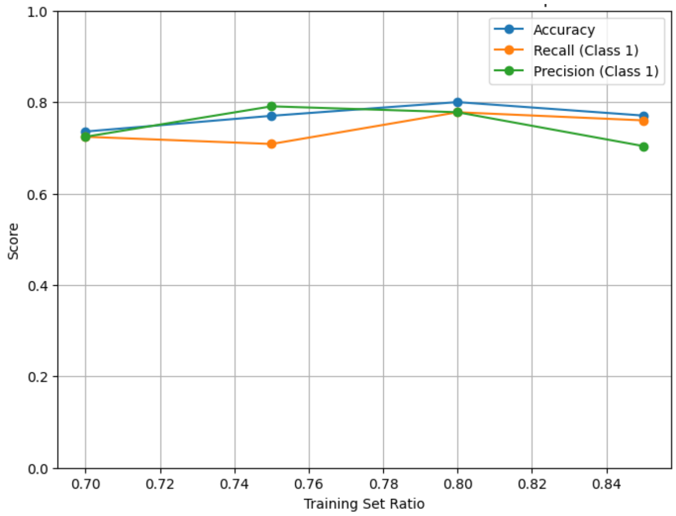

Figure 2: Model performance across different train-test splits. Split of 0.8 gave the best performance.

The model achieved an **accuracy** of **82.50%**, **recall** of **0.83**, and **precision** of **0.79** (Figure 3).
In the context of medical prediction, recall is a particularly important metric. Recall measures the proportion of true high-risk individuals correctly identified by the model. Missing a high-risk patient (false negative) could lead to severe consequences, whereas a false positive (lower precision) is less critical, and may only result in additional monitoring or preventive measures. Therefore, a high recall ensures that the model minimizes the risk of overlooking patients who may benefit from early interventions, which is crucial in healthcare applications.

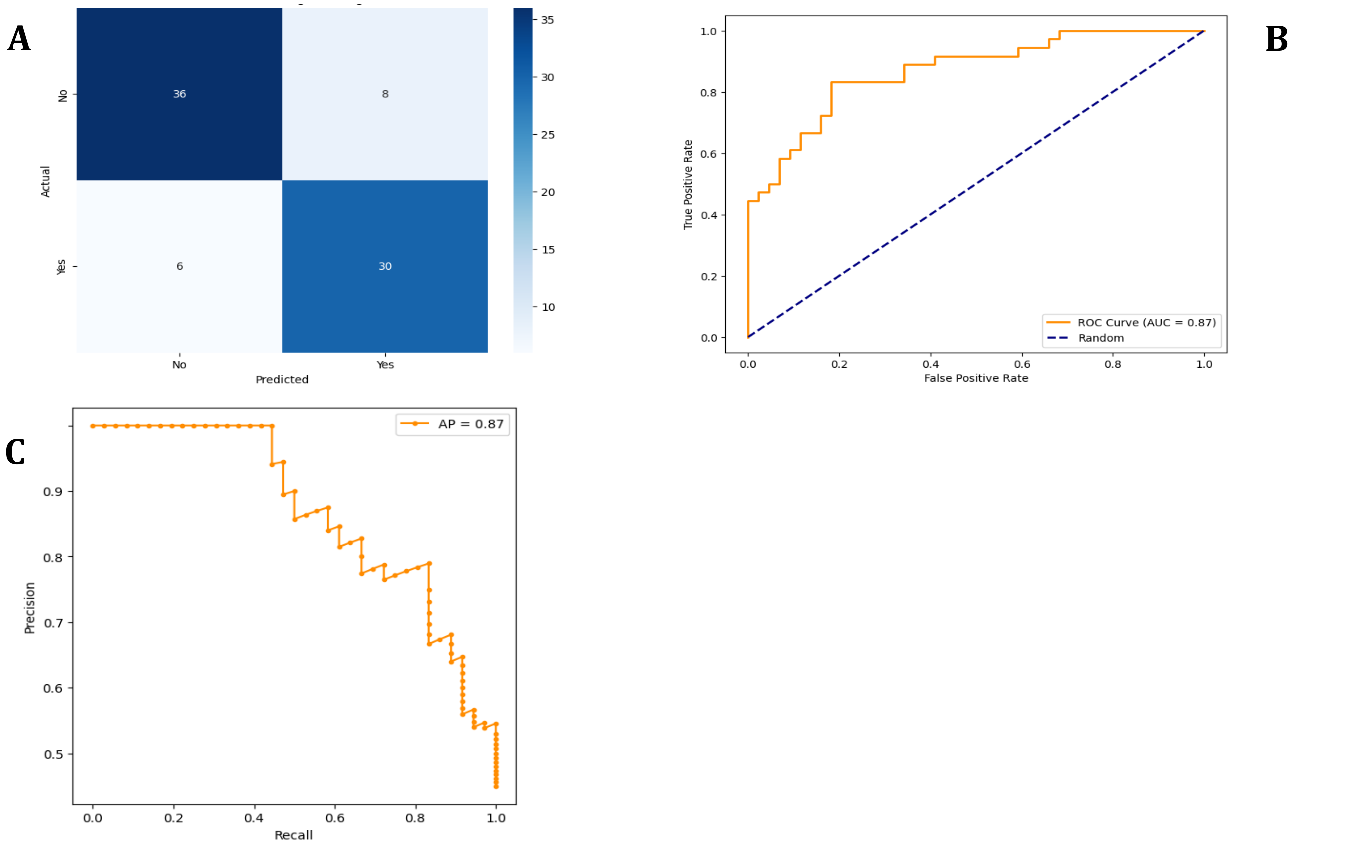

Figure 3: Model metrics. A) Confusion matrix showing recall=0.83 and precision=0.79 for the positive class. B) Receiver Operating Characteristic (ROC) curve showing Area Under the Curve (AUC) of 0.87. C) Precision-Recall curve showing the trade-off between precision and recall for stroke risk prediction. The model achieved an average precision (AP) of 0.87.

5.0 Interpretability

In the logistic regression model, a positive coefficient means that an increase in a feature corresponds to an increase in stroke risk, while a negative coefficient suggests a protective effect that lowers stroke risk. From the results of the model coefficients (Figure 4), the strongest positive contribution was from the interaction between exercise and age (exercise_x_age), indicating that age combined with low exercise habits significantly increases stroke risk. Similarly, belonging to a higher BMI category (bmi_class) and having a higher age-squared value both raised stroke risk. This means that both weight status and advancing age strongly influence stroke outcomes. On the other hand, being a former or non-smoker put subjects at a lower risk for stroke. In real-world applications, these findings can guide targeted prevention strategies, for example, focusing on promoting physical activity in older adults and weight management across all age groups could help reduce stroke risk.

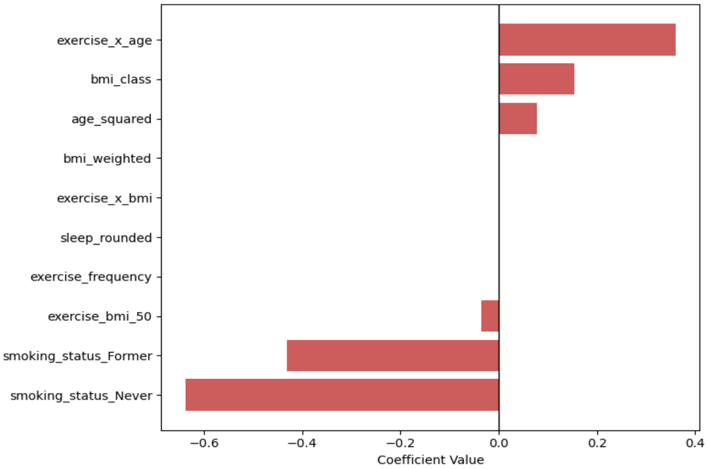

Figure 4: Logistic regression model coefficients showing positive coefficients for exercise_x_age, bmi_class, and age_squared, and negative coefficients for exercise_bmi_50, smoking_status_Former, and smoking_status_Never

6.0 Future Work

Future improvements for this project would include exploring other feature engineering techniques and machine learning algorithms (such as Random Forest and Gradient Boosting) to improve the recall and accuracy scores. Additionally, the model can be deployed on the web and integrated with a large language model to request appropriate data from users via text or medical report uploads.