Springboard Final Capstone Project:

Predictive Modeling for Stroke Risk Assessment

Using Machine Learning

By **Ranjana Roka**

1. Problem Statement

Strokes are a leading cause of death and disability globally, affecting millions of people each year. Even with all the progress we've made in medicine, accurately predicting the risk of stroke remains a challenge to date. The traditional method mainly relies on a handful of well-known factors that lead to inaccurate risk assessments, potentially hindering preventive interventions and personalized healthcare strategies. This capstone project aims to develop a predictive model that accurately assesses the risk of stroke in individuals based on various health-related features.

Stroke prevention is crucial to reduce the disease burden and overall healthcare costs. Early identification of high-risk individuals can lead to timely interventions and minimize the disability by implementing effective preventative measures. Using machine learning in healthcare could pinpoint the crucial health factors linked to stroke risk.

2. Data Wrangling

In this project, data wrangling was a crucial step in cleaning and preparing the dataset for further analysis. I sourced the dataset from Kaggle https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?select=healthcare-dataset-stroke-data.csv.

Initially, the dataset was loaded using the pandas library from a CSV file and explored using various Python built-in features. we get an overview of the dataset, including the number of rows (5110), columns (12), data types. The dataset includes both numerical and categorical features such as age, gender, BMI, hypertension, heart disease, and lifestyle factors like smoking status.

After a detailed observation, I identified some gaps, particularly in the BMI values, with BMI having 4.1% missing data, a relatively small portion of the dataset. To address this, I calculated the mean of the BMI column and used it to fill in the missing values. After that, the final check was done to ensure that there were no remaining missing values and that the dataset was clean and ready for further analysis.

3. Exploratory Data Analysis

After the data had been collected and cleaned, the next step that I performed was data exploration and analysis. This helps me to understand the data's structure, visualize relationships between variables, and detect patterns that might contribute to stroke occurrences. I categorized the dataset into numerical and categorical variables.

a. Data distribution and visualization

I. **Numerical variables**:

`id:` Unique identifier for each individual.

`age`: Age of the individual.

`hypertension`: Presence of hypertension (1 = Yes, 0 = No).

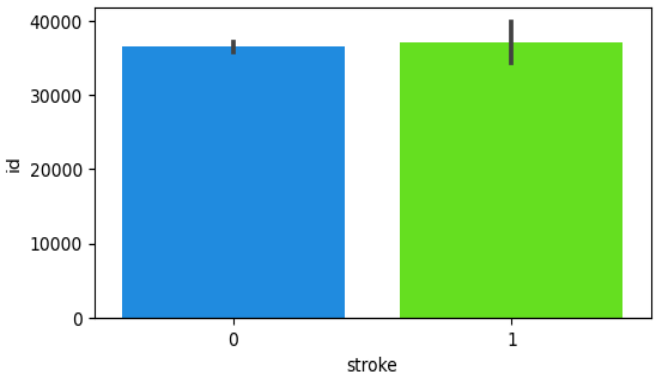
`heart_disease`: Presence of heart disease (1 = Yes, 0 = No).

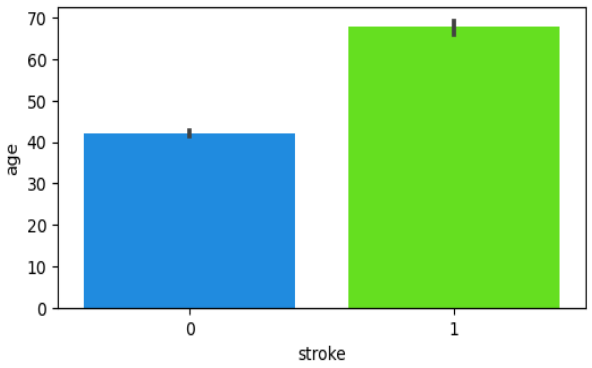
`avg_glucose_level`: Average glucose level in the blood.

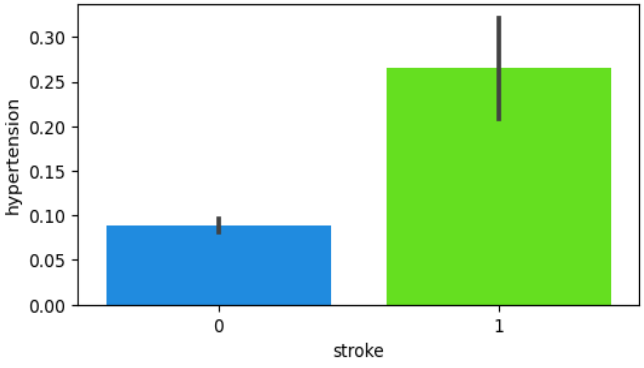
`bmi`: Body Mass Index (BMI).

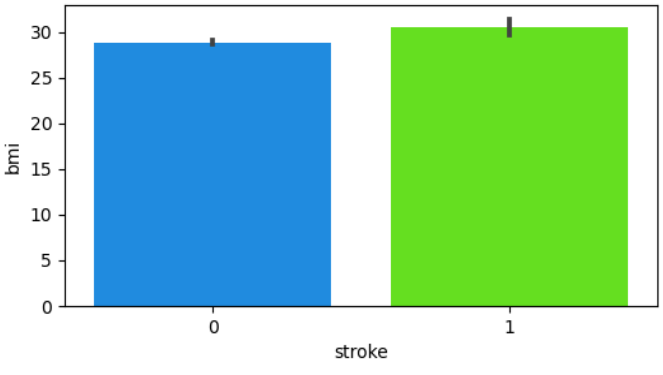
`stroke`: Indicates whether the individual had a stroke (1 = Yes, 0 = No).

I used the bar plots to see the relationship between each numerical variable and the target variable, stroke.









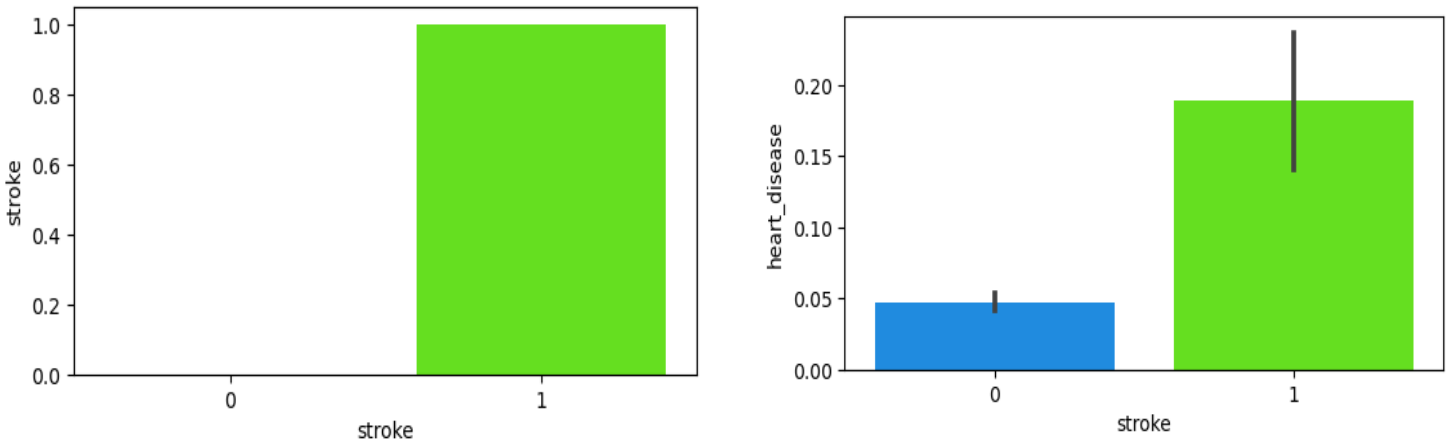


Figure 1. Bar plot of Numerical Features

II. **Categorical Variables**: I used the count plots to observe their distribution and relationship with stroke occurrences.

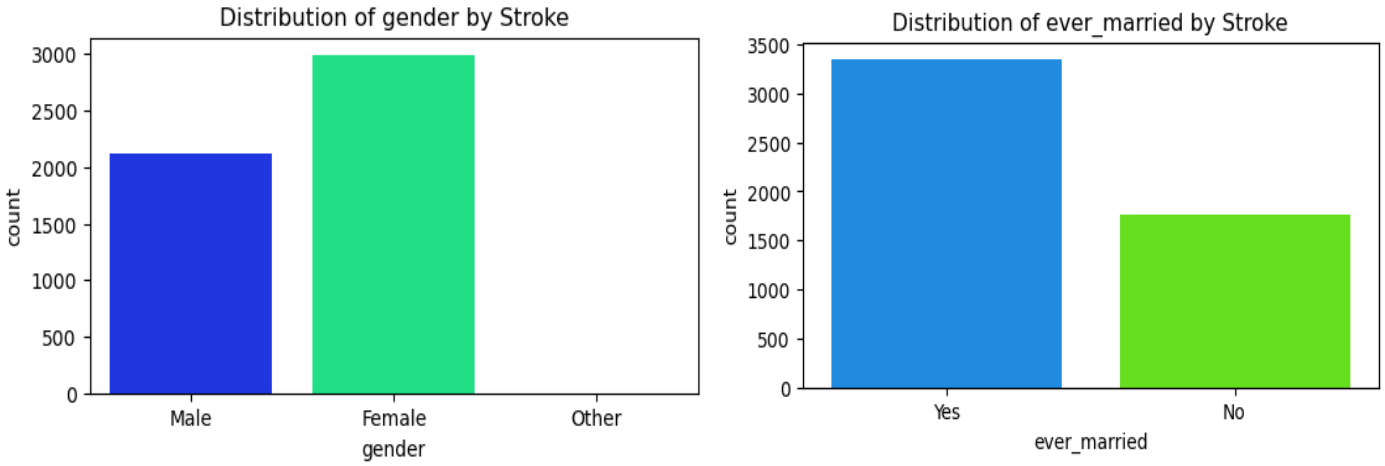
`gender`: Gender of the individual (Male, Female, Other).

`ever_married`: Whether the individual was ever married (Yes, No).

`work_type`: Type of work (Private, Self-employed, Govt_job, children, Never_worked).

`Residence_type`: Type of residence (Urban, Rural).

`smoking_status`: Smoking habits (formerly smoked, never smoked, smokes, Unknown)



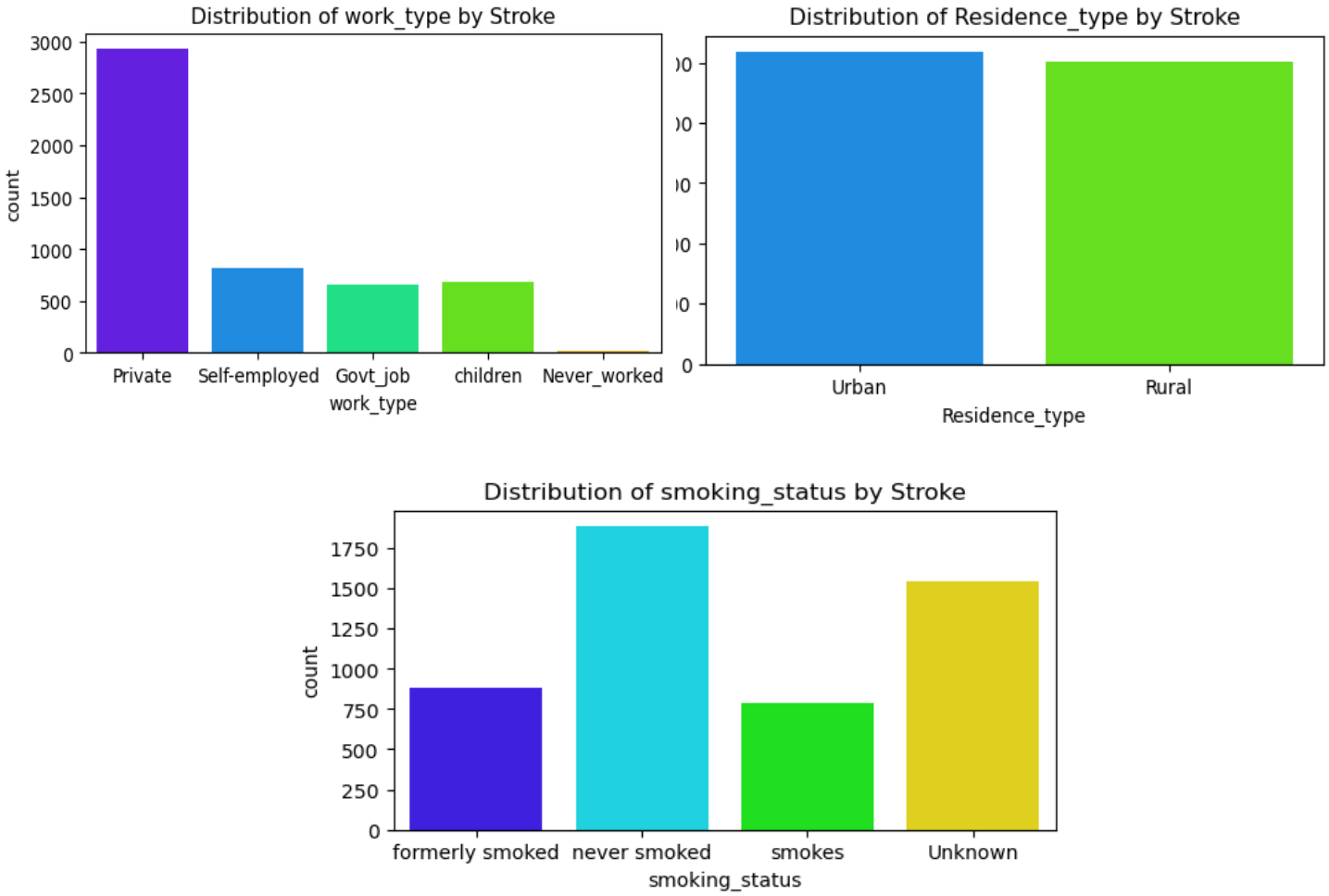


Figure 2. Count plot of Categorical Features

I visualized the relationship between numerical and categorical features with stroke occurrences with the above plots. Key features showing a relationship with stroke include age, where older individuals are at higher risk, along with hypertension, heart disease, and elevated glucose levels.

Visualizing the Correlation Matrix with Heatmap

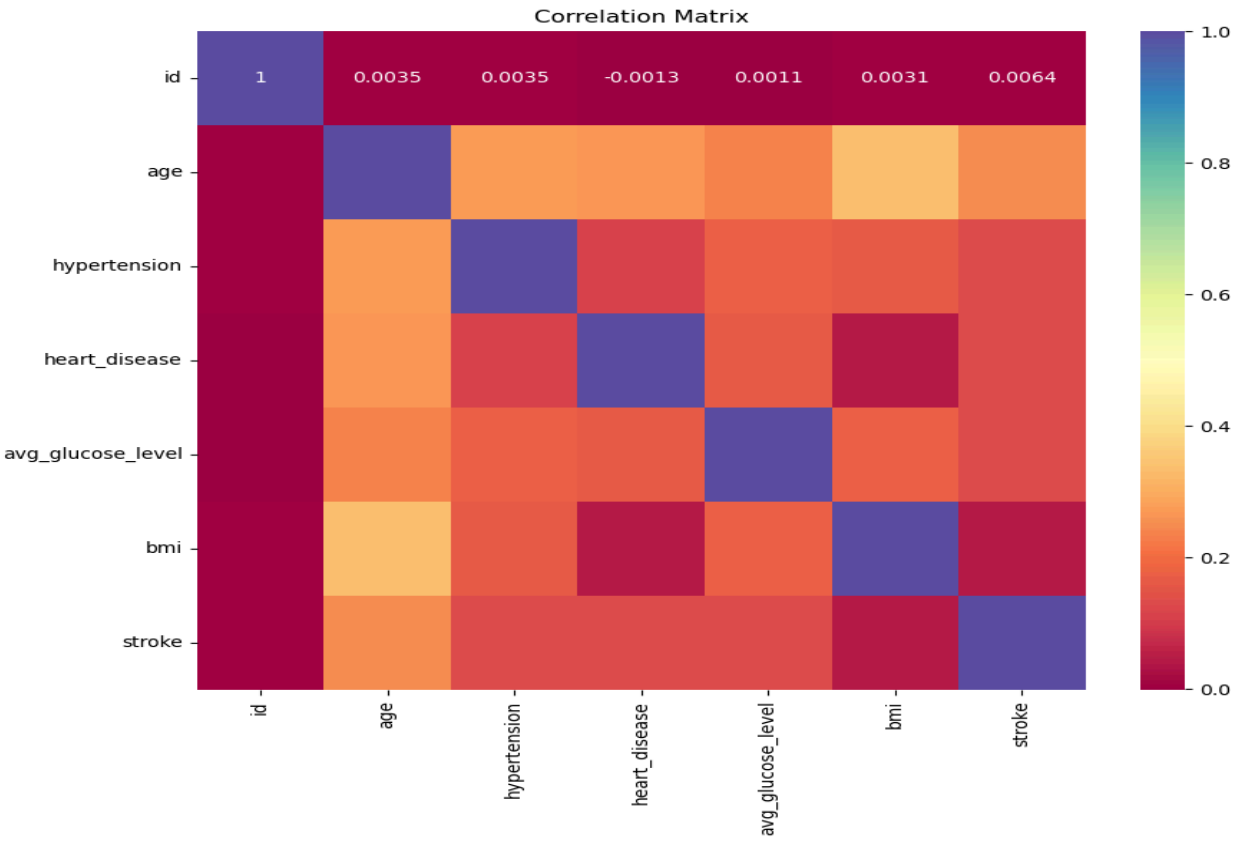


Figure 3. Correlation Matrix

A heatmap visually represents the correlation matrix, making it easier to identify strong and weak correlations at a glance. The diagonal from the top left to the bottom right is all 1. This is because each variable is perfectly correlated with itself. There's a moderate positive correlation between age and stroke. This suggests that as age increases, the likelihood of having a stroke also increases. The correlation seems low to moderate between hypertension and stroke. Hypertension does have a positive relationship with stroke, but it is not very strong. Again, the correlation is moderate between heart_disease and stroke. Heart disease is positively correlated with stroke occurrence, but the relationship is not as strong as it might be with age. The correlation appears to be low between avg_glucose_level and stroke, indicating that average glucose levels may not have a strong linear

relationship with stroke risk in this dataset. The correlation is also low between bmi and stroke, suggesting that BMI might not have a strong direct correlation with stroke in this dataset. Seeing the matrix, age appears to have the most significant positive correlation with stroke (target variable).

4. Machine Learning

I used four different machine learning models—Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting to determine the most effective approach to stroke prediction regarding accuracy score. I created a classification report and confusion matrix for each model and compared them against each other.

Output for Confusion Matrix:

Model	True Negatives (TN)	False Positives (FP)	False Negatives (FN)	True Positives (TP)
Logistic Regression	1452	0	81	0
Decision Tree	1452	0	81	0
Random Forest	1452	0	81	0
Gradient Boosting	1452	0	80	1

All models have 100% accuracy in predicting non-stroke cases (True Negatives = 1452, False Positives = 0). Logistic Regression, Decision Tree, and Random Forest failed to identify stroke cases (True Positives = 0, False Negatives = 81).

Gradient Boosting predicted 1 stroke case correctly but still missed 80 stroke cases. While there is a slight improvement, the performance is still poor when it comes to detecting stroke cases due to the imbalance in the data (many more non-stroke cases).

Output for Classification Report:

Model	Accuracy	Precision (1)	Recall (1)	F1-score (1)
Logistic Regression	0.947162	0.00	0.00	0.00
Decision Tree	0.947162	0.00	0.00	0.00
Random Forest	0.947162	0.00	0.00	0.00
Gradient Boosting	0.947162	0.50	0.01	0.02

In this analysis, all models achieved high accuracy (94.7%); since there are many more non-stroke cases than stroke cases in the data, the models are primarily good at predicting people who don't have strokes. Logistic Regression, Decision Tree, and Random Forest failed to correctly identify stroke cases, resulting in 0 precision, recall, and F1-scores for stroke prediction.

Gradient Boosting did a bit better, correctly identifying 1% of stroke cases with a precision of 50%, but overall, it still missed the majority of actual stroke cases. Out of 4 models, Gradient Boosting performed the best overall in predicting strokes, even though the imbalanced data made it challenging.

Handling Class Imbalance using SMOTE:

I used SMOTE technique to oversample the minority class (stroke cases) to balance the dataset and improve model training.

Four machine learning models (Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting) were trained on the resampled data. I evaluated the model performance based on these three factors:

Accuracy and ROC AUC (Area Under the ROC Curve)

Confusion matrices

Classification reports

Model Performance Summary:

The output of the performance metrics includes both **accuracy** and **ROC AUC**, which help in comparing the models' effectiveness.

Model	Accuracy	ROC AUC
Logistic Regression	0.737769	0.821583
Decision Tree	0.662753	0.793818
Random Forest	0.720157	0.792436
Gradient Boosting	0.875408	0.777901

Classification Report After SMOTE:

Model	Accuracy	Precision (1)	Recall (1)	F1-score (1)
Logistic Regression	0.737769	0.143979	0.679012	0.237581
Decision Tree	0.662753	0.122010	0.629630	0.204409
Random Forest	0.720157	0.128713	0.641975	0.214433
Gradient Boosting	0.875408	0.159030	0.728395	0.261062

Confusion Matrix table after applying SMOTE:

Model	True Negatives (TN)	False Positives (FP)	False Negatives (FN)	True Positives (TP)
Logistic Regression	1125	327	26	55
Decision Tree	1085	367	30	51
Random Forest	1100	352	29	52
Gradient Boosting	1140	312	22	59

CONCLUSION

All models achieved high accuracy scores, around 94.7%, but more than accuracy alone was needed due to the imbalance in the dataset. SMOTE helped to oversample the minority class (stroke cases), creating a more balanced dataset. After applying SMOTE, I focused more on ROC-AUC and classification metrics like precision, recall, and F1-score, which provided a clearer picture of how well the models performed on stroke cases. While Gradient Boosting has the highest accuracy (87.54%), Logistic Regression has a better ROC AUC score (0.82), which is more critical when working with imbalanced datasets. Gradient Boosting is the best model if I prioritize accuracy, but it performs best if I prioritize ROC AUC: Logistic Regression.

RECOMMENDATION

1. **Improve early stroke risk detection:**

   The findings of this project showed the gradient-boosting model performed the best in the early detection of stroke risk among patients. Using this model in their health monitoring system, the client can detect the stroke risk early even though the dataset is imbalanced. This can lead to timely interventions for those who are at high risk of having a stroke.

2. **Data-driven decision making:**

After applying the SMOTE technique, the dataset became more balanced than it was before. My analysis using SMOTE demonstrated that balanced datasets improve stroke prediction accuracy. The clinical support systems or insurance risk assessment can benefit from this approach, where missing high-risk patients due to an imbalanced dataset can be a problem. Therefore, balancing the dataset is critical to ensure that such overlooked high-risk cases get attention on time.

3. **Optimizing Resources in Healthcare:**

   By applying the Gradient Boosting model to focus on the minority class (stroke cases), the client can optimize healthcare resources by prioritizing attention to high-risk patients. This helps healthcare settings allocate resources by identifying the cases that need prompt attention, leading to better patient outcomes and reduced healthcare costs.