THE GEORGE WASHINGTON UNIVERSITY

DATS6103 - Summary Report
Professor Divya Pandove Narula

STROKE PREDICTION

TEAM 6 - Aswin Balaji Thippa Ramesh, Barathkumar Anantharaj,
Gowri Sriram Lakshmanan, Rahul Arvind

INTRODUCTION

Stroke remains a leading cause of mortality and long-term disability worldwide, with its occurrence significantly influenced by a combination of medical, demographic, and lifestyle factors. According to the World Health Organization (WHO), approximately 15 million people suffer from strokes annually, with nearly 5 million fatalities and 5 million survivors facing permanent disabilities. Early identification of individuals at higher risk of stroke is critical for prevention and timely intervention.

This project analyzes stroke occurrence using the Stroke Prediction Dataset sourced from Kaggle, which comprises over 5,000 observations and various attributes like age, gender, medical history (hypertension, heart disease), and lifestyle factors (BMI, smoking status). While physical health factors like glucose levels and BMI are known to contribute to stroke risk, other variables such as work type, marital status, and residence type may also provide insights into identifying high-risk groups.

Through Exploratory Data Analysis and machine learning models like Logistic Regression and Random Forest, we aim to uncover the key predictors of stroke, assess their significance, and evaluate model accuracy. This study also explores trends in stroke occurrences across age groups and demographic segments, providing actionable insights to guide public health strategies and clinical practices.

INFORMATION ABOUT THE DATASET

We utilized the Stroke Prediction Dataset from Kaggle, which contains 5,110 observations and 11 key variables. These variables encompass multiple dimensions:

Demographic Factors: Age, gender, residence type, work type

Medical History: Hypertension, heart disease, BMI, average glucose level

Lifestyle Factors: Smoking status

Target Variable: Stroke occurrence (binary: 1 for stroke, 0 for no stroke)

The dataset presents a significant class imbalance, with stroke occurrences accounting for only about 5% of the total cases. This imbalance poses challenges for model training, as traditional algorithms may become biased toward the majority class. To address this, we plan to employ resampling techniques like SMOTE to create a balanced class representation, improve model performance, and ensure the minority class (stroke cases) is appropriately represented during training. By carefully preprocessing the data and addressing these challenges, we aim to build a robust predictive model that accurately identifies stroke risks

THE SMART QUESTIONS

With the intention of addressing critical aspects of stroke prediction using the Stroke Prediction Dataset, our project explores the following key questions:

1. What are the key factors most strongly associated with the occurrence of strokes?
2. How does each predictor's distribution differ between stroke and non-stroke cases?
3. Can we determine statistically significant correlations or patterns in stroke occurrence based on categorical variables?
4. Can we identify trends or variations in stroke risk based on time-related factors like age?
5. How accurate is it to build a machine learning model to predict stroke occurrence based on the dataset?

DATA PREPARATION

The data preparation stage involves essential steps to ensure the dataset is clean, well-structured, and ready for analysis or model training.

1. **Data Loading & Structure :**
   - We loaded the dataset using **Pandas** and confirmed successful loading by viewing the first few rows.
   - The dataset contained **5,110** observations and **11** columns, as verified using the shape function.

```
id                 int64
gender            object
age              float64
hypertension       int64
heart_disease      int64
ever_married      object
work_type         object
Residence_type    object
avg_glucose_level float64
bmi               float64
smoking_status    object
stroke             int64
dtype: object
```

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

```
✓ #Shape of the dataset ···

(Rows,columns): (5110, 12)
```

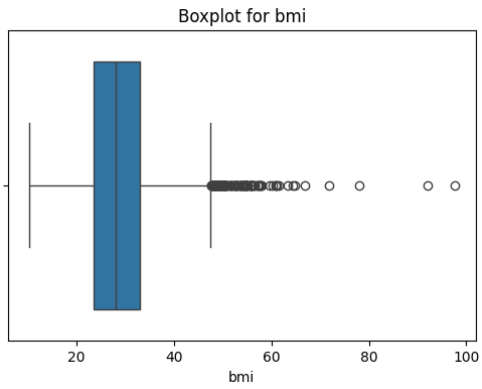
2. **Handling missing & unwanted values :**
   - We checked for missing values across all columns using the isnull().sum() method.
   - The **BMI** column had missing values, with approximately **4.5%** of its entries missing. To retain data integrity and avoid loss of observations, we performed **mean imputation** for the **BMI** column.
   - During the inspection of the **gender** column, we identified an **unwanted value** (e.g., "Other") that was inconsistent with the valid categories ("Male" and "Female"). This value was removed to ensure consistency in the data

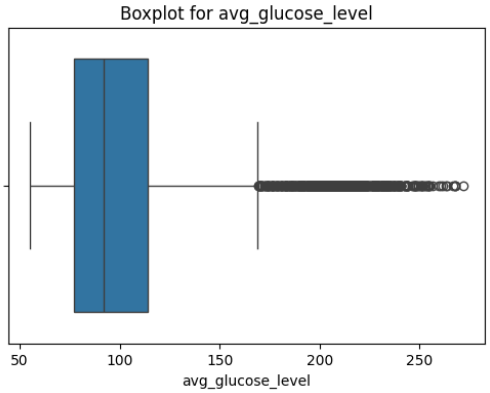
```
        gender
Female    2994
Male      2115
Other        1
Name: count, dtype: int64
```

```
id : 0
gender : 1
age : 0
hypertension : 0
heart_disease : 0
ever_married : 0
work_type : 0
Residence_type : 0
avg_glucose_level : 0
bmi : 201
smoking_status : 0
stroke : 0
```

```
df["bmi"]=df["bmi"].replace(np.nan,df["bmi"].mean())
df["gender"]=df["gender"].replace(np.nan,df["gender"].mode()[0])
```

3. **Outliers Check :**
    - Outliers in our dataset were identified using **boxplots**. These outliers were considered **useful** as they reflect real-world variations, such as high BMI or glucose levels, which are critical indicators of stroke risk.





4. **Encoding Variables:**
    - Categorical variables such as **gender**, **work_type**, **Residence_type**, and **smoking_status** were transformed using **Label Encoding** to prepare them for modeling. Label Encoding was chosen over one-hot encoding to keep the dataset compact and avoid increasing dimensionality.

```
categorical_columns = df.select_dtypes(include=['object']).columns

# Encode all categorical variables
label_encoders = {}
for col in categorical_columns:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])
    label_encoders[col] = le
```

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	1	67.0	0	1	1	2	1	228.69	36.600000	1	1
1	51676	0	61.0	0	0	1	3	0	202.21	28.893237	2	1
2	31112	1	80.0	0	1	1	2	0	105.92	32.500000	2	1
3	60182	0	49.0	0	0	1	2	1	171.23	34.400000	3	1
4	1665	0	79.0	1	0	1	3	0	174.12	24.000000	2	1

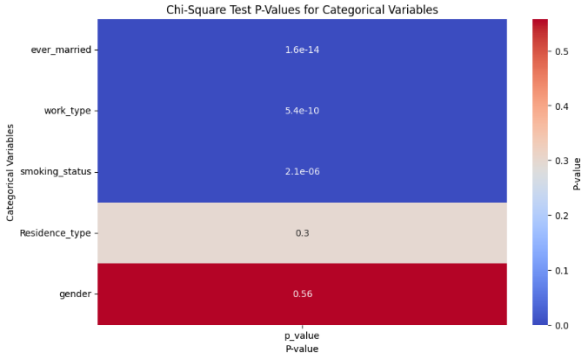
EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) was conducted to gain a deeper understanding of the dataset, uncover patterns, and answer the SMART questions related to stroke risk. We explored the distributions, relationships, and correlations between variables, with a focus on identifying factors that contribute to stroke occurrence.

1. **Can we determine statistically significant correlations or patterns in stroke occurrence based on categorical variables?**

    - **Objective**: We performed the **Chi-Square Test of Independence** to identify significant associations between categorical variables and the target variable (**stroke**).
    - **Significance Threshold**: Variables with p-values **< 0.05** were considered statistically significant.

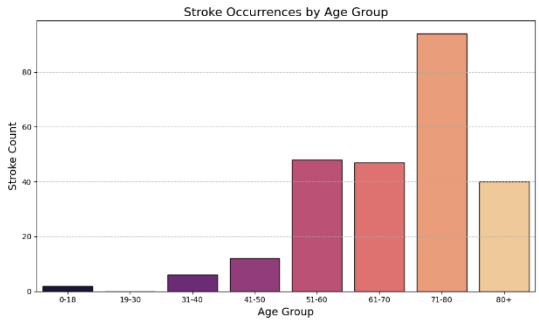
```
Statistically significant categorical correlations with the target variable:
{'ever_married': 1.6389021142314745e-14, 'work_type': 5.397707801896119e-10, 'smoking_status': 2.0853997025008455e-06}
```

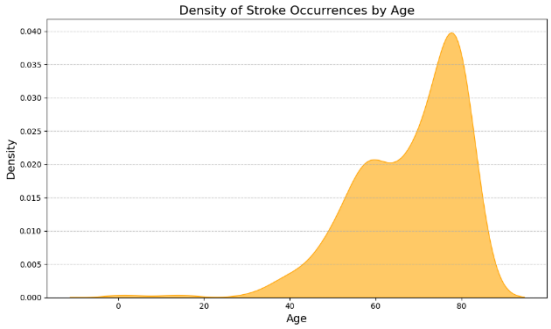


From the graph, variables like **ever_married**, **work_type**, and **smoking_status (dark blue)** are important for predicting stroke, while **Residence_type (light shade)** and **gender (red)** are less relevant. This highlights where the focus should be during feature selection for modeling.

2. Can we identify trends or variations in stroke risk based on time-related factors like age?

- We grouped ages into predefined bins (e.g., 0-18, 19-30, 31-40, etc.) to better analyze stroke occurrences by age range.
- A **Density Plot** was created to show the distribution of stroke occurrences across age.
- A **Bar Chart** was plotted to display stroke counts for each age group, making trends across age ranges easier to interpret.

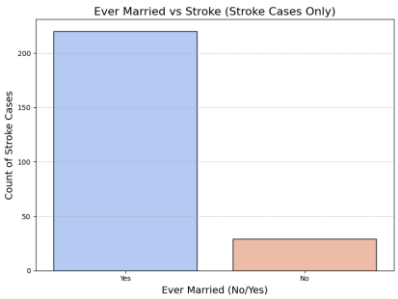


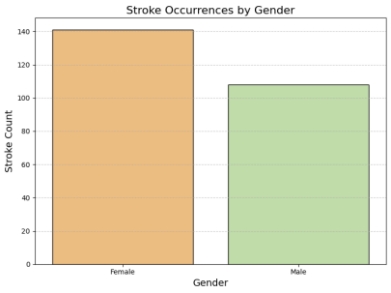


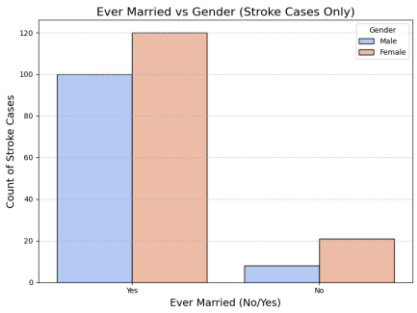
From both visualizations, it is evident that stroke occurrences are heavily concentrated among older individuals, particularly those aged **71–80 years**, with a notable increase starting from age **50**.

3. How does each predictor's distribution differ between different stroke cases ?

- These insights provide a foundation for identifying significant variables and understanding patterns that contribute to stroke risk.
- Examined the proportions of stroke occurrences across categories (e.g., work type, smoking status, marital status) using bar charts and stacked plots.
- Observed clear differences in distributions, such as older individuals, higher BMI, and glucose levels being more common among stroke cases.

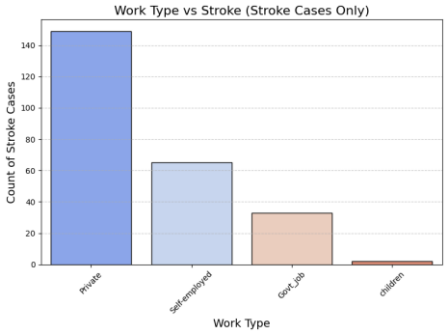


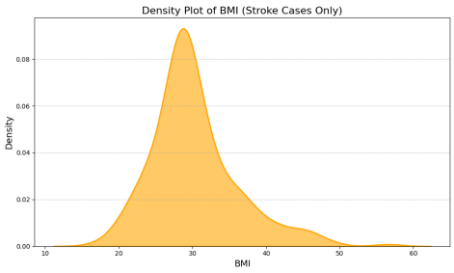


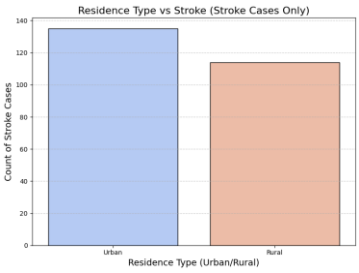


- **Marital Status**: Individuals who were **ever married** experienced significantly more stroke cases compared to those who were not, indicating marital status may be linked to lifestyle or age-related factors contributing to stroke risk.
- **Gender**: Stroke cases were **higher among females** compared to males overall, suggesting gender could play a role in stroke susceptibility.

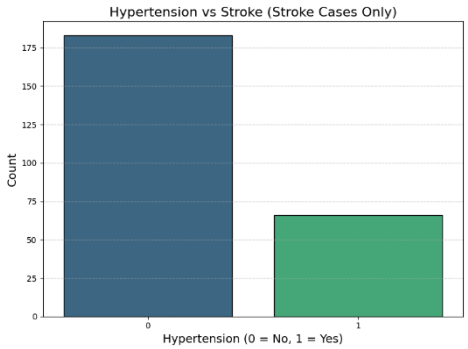
Among stroke cases, females who were **ever married** represented the largest group, indicating that marital status and gender together may amplify stroke risk.



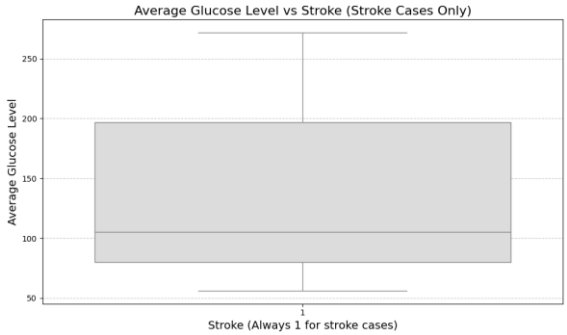


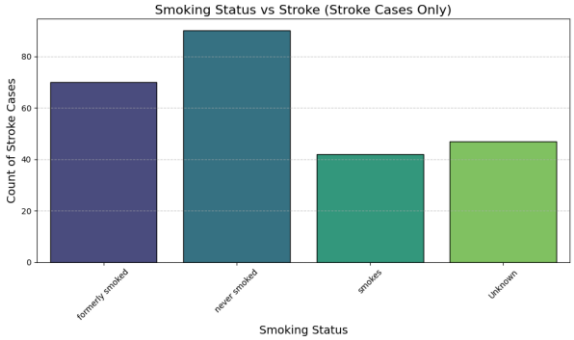


- **Work Type:** Stroke cases are most common among individuals in the private sector, followed by the self-employed, while occurrences are much lower for government workers and children.
- **BMI:** The BMI distribution shows that stroke cases are concentrated in the overweight to moderately obese range (BMI 25–35)
- **Residence Type:** Stroke cases are almost uniformly distributed between urban and rural areas, indicating that residence type has minimal impact on stroke occurrences.



The analysis reveals that while hypertension is a known risk factor for stroke, the majority of stroke cases occur in individuals **without hypertension**.





- **Average Glucose Level**: Stroke cases are associated with a **wide range of glucose levels**, with many individuals having elevated glucose levels above **100 mg/dL**, indicating that hyperglycemia could be a contributing factor.
- **Smoking Status**: Stroke cases are more prevalent among individuals who **never smoked** and those who were **formerly smokers**, followed by lower counts in current smokers and those with **unknown smoking status**.

4. What are the key factors most strongly associated with the occurrence of strokes?

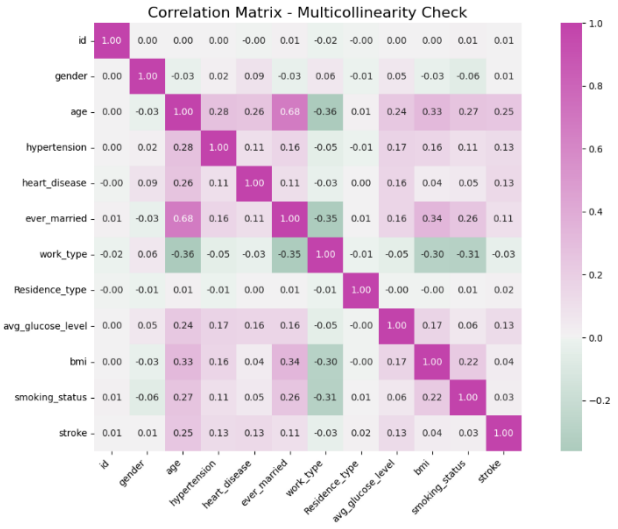
- We applied **Recursive Feature Elimination (RFE)** using a **linear model (lm)** on the imbalanced dataset to identify and rank the most important features associated with stroke occurrence. This method iteratively eliminated less important features to highlight the strongest predictors.

Feature	Rank
age	1
avg_glucose_level	2
bmi	3
work_type	4
smoking_status	5
gender	6
Residence_type	7
ever_married	8
hypertension	9
heart_disease	10)

The top predictors include age, average glucose level, BMI, work type, and smoking status, with age being the most significant. These findings emphasize that age-related factors, metabolic indicators (like glucose and BMI), and lifestyle attributes are crucial in understanding and predicting stroke risk.

MULTICOLLINEARITY CHECK

To detect and address multicollinearity among predictors, we created a **correlation heatmap** to visualize and identify highly correlated numerical predictors and calculated **Variance Inflation Factor (VIF)** values, using a threshold of **5** to detect and remove predictors with high multicollinearity.



VIF values for features:

	feature	VIF
1	gender	1.017599
2	age	2.306963
3	hypertension	1.115878
4	heart_disease	1.113590
5	ever_married	1.969231
6	work_type	1.287976
7	Residence_type	1.000680
8	avg_glucose_level	1.106752
9	bmi	1.226924
10	smoking_status	1.160771
11	stroke	1.088418

The multicollinearity check, conducted using the **correlation matrix** and **Variance Inflation Factor (VIF)**, confirms that there is no significant multicollinearity among the predictors:

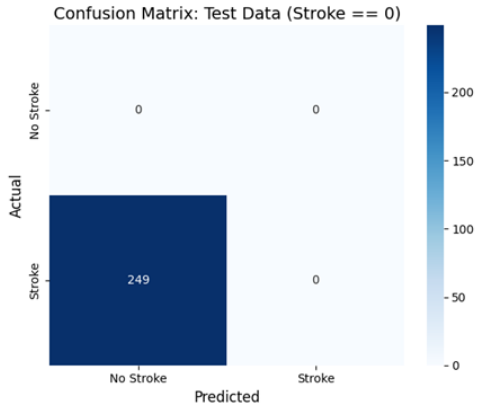
- The **correlation heatmap** shows no strong correlations between features, with correlation coefficients remaining well below the critical threshold of **0.8**.
- The **VIF values** for all predictors are below the threshold of **5**, with the highest being **2.3** for the age variable. This indicates that none of the predictors are highly correlated with others, ensuring the dataset does not suffer from multicollinearity.

The results confirm that the predictors are independent and can be safely included in the modeling process.

MODEL BUILDING

We implemented various **machine learning models** on both the **imbalanced** and **balanced datasets** to evaluate their performance. The balanced dataset was created using techniques like **SMOTE** to address class imbalance. The analysis yielded **surprising results**, highlighting differences in model performance based on data balance and the significance of certain predictors in stroke prediction.

1. Logistic Regression on Imbalanced data:



- The model achieved 0% accuracy for stroke predictions (True Positives = 0).
- It achieved 100% accuracy for non-stroke predictions, as it classified all instances as "No Stroke" due to the dominance of the majority class.

Balancing the data:

To address the severe class imbalance in the dataset (4861 non-stroke cases vs. 249 stroke cases), we applied **SMOTE** (Synthetic Minority Oversampling Technique) to generate synthetic samples for the minority class.

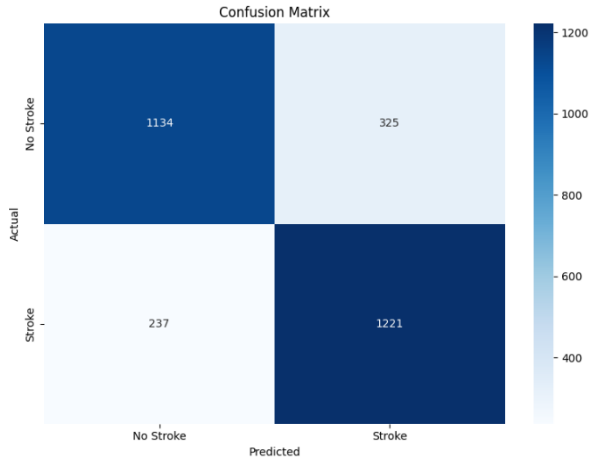
stroke	
0	4861
1	249

stroke	
1	4861
0	4861

This resulted in a balanced dataset with **4861 stroke cases** and **4861 non-stroke cases**, ensuring equal class representation. Balancing the data improves model performance by enabling the model to learn patterns for both classes effectively, reducing bias toward the majority class.

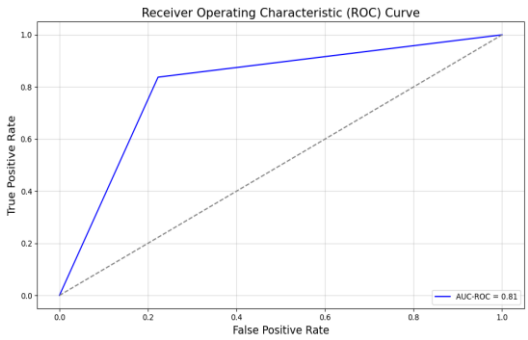
2. Logistic Regression on balanced data:

We built a classification model on the **balanced dataset** to predict stroke occurrences. The model's performance was evaluated using the following metrics:



Classification Report:

	precision	recall	f1-score	support
0	0.83	0.78	0.80	1459
1	0.79	0.84	0.81	1458
accuracy			0.81	2917
macro avg	0.81	0.81	0.81	2917
weighted avg	0.81	0.81	0.81	2917

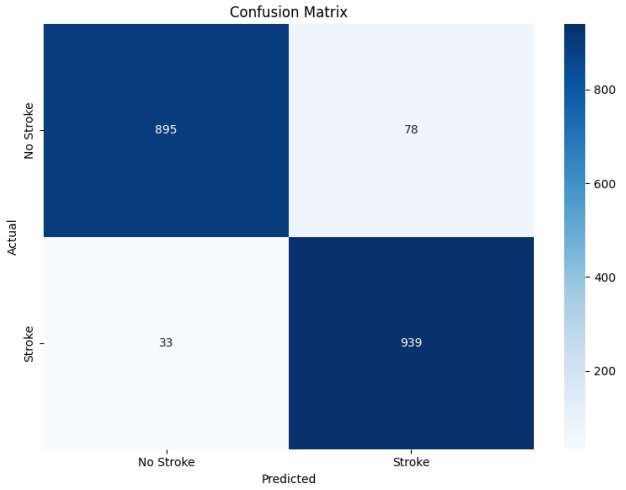


- **ROC-AUC Score**: The model achieved an **AUC of 0.81**, indicating good predictive performance.
- **Confusion Matrix**: Correctly predicted **1134 No-Stroke** cases and **1221 Stroke** cases. Misclassified **325 No-Stroke** cases and **237 Stroke** cases.
- **Classification Report:** Precision, recall, and F1-score were balanced for both classes, with an overall **accuracy of 81%**.

The balanced dataset significantly improved the model's ability to predict both stroke and non-stroke cases. With an **AUC of 0.81** and strong precision-recall scores, the model demonstrates reliable performance, successfully overcoming the limitations observed with the imbalanced data.

3. Random Forest Model on balanced data:

We implemented a Random Forest model on the balanced dataset to predict stroke occurrences. To evaluate its performance, we assessed key metrics that we used previously. This approach allowed us to measure the model's ability to classify stroke and non-stroke cases effectively.

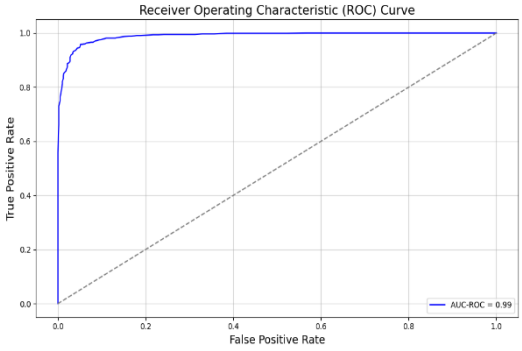



```
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.92      0.94       973
           1       0.92      0.97      0.94       972

    accuracy                           0.94      1945
   macro avg       0.94      0.94      0.94      1945
weighted avg       0.94      0.94      0.94      1945

AUC-ROC: 0.99
```



- **ROC-AUC Score**: The Random Forest model achieved an impressive **AUC of 0.99**, significantly improving from the previous model's **AUC of 0.81**, demonstrating stronger discriminative ability.
- **Confusion Matrix**: The Random Forest model correctly predicted **895 No-Stroke** and **939 Stroke** cases, reducing misclassifications (**78 No-Stroke** and **33 Stroke**) compared to the previous model's higher misclassification rates (**325 No-Stroke** and **237 Stroke**).
- **Classification Report:** Precision, recall, and F1-score improved to **94%** for both classes, up from **80-81%** in the previous model. Overall **accuracy** increased to **94%**, compared to the previous model's **81%**.

The **Random Forest model** outperformed the previous model across all metrics, achieving higher accuracy, precision, recall, and AUC scores. This significant improvement highlights the Random Forest's ability to better capture patterns in the balanced dataset, making it a superior choice for stroke prediction.

CONCLUSION

We evaluated three models for stroke prediction: **Logistic Regression**, **a prior balanced model**, and **Random Forest**. Logistic Regression on the imbalanced dataset failed to predict stroke cases, achieving **0% accuracy** for strokes due to class bias. The second model on balanced data improved performance significantly, achieving an **AUC of 0.81** and an accuracy of **81%**. Finally, the **Random Forest model** on the balanced dataset outperformed all, with an **AUC of 0.99** and **94% accuracy**, demonstrating exceptional precision, recall, and minimal misclassifications. The results confirm that addressing class imbalance and using ensemble methods like Random Forest significantly enhance model performance for stroke prediction.

LIMITATIONS

- The dataset suffers from **class imbalance** (only 5% stroke cases), leading to bias toward the majority class. Additionally, **population bias** exists, with females showing higher stroke occurrences, which may reflect sampling issues rather than actual trends, limiting generalizability across different demographics.
- Based on general stroke case study, critical medical predictors like **cholesterol levels, blood pressure trends**, and **family medical history** are missing, while contextual factors such as **treatment history, environmental influences**, and socioeconomic status are not included. This reduces the dataset's depth and predictive power.
- The data is **cross-sectional** and does not track risk progression over time. Measurement errors in self-reported values (e.g., BMI, glucose) may reduce reliability, and synthetic balancing methods like **SMOTE** can introduce overfitting risks, making cautious interpretation necessary.

REFERENCES

A predictive analytics approach for stroke prediction using machine learning and neural networks
https://www.sciencedirect.com/science/article/pii/S2772442522000090

Predicting stroke occurrences: a stacked machine learning approach with feature selection and data preprocessing
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-024-05866-8

Stroke Risk Factors, Genetics, and Prevention
https://pmc.ncbi.nlm.nih.gov/articles/PMC5321635/#:~:text=The%20Framingham%20Stroke%20Risk%20Profile,the%20presence%20of%20cardiovascular%20disease%20(

The most efficient machine learning algorithms in stroke prediction: A systematic review
https://pmc.ncbi.nlm.nih.gov/articles/PMC11443322/

Stroke Prediction and Contributing Factors Using Machine Learning
https://www.researchgate.net/publication/380424935_Stroke_Prediction_and_Contributing_Factors_Using_Machine_Learning