# Machine Learning Models for Stroke Prediction

Artificial Intelligence Master degree

Machine Learning

Academic Year 2024-2025

Author: Giacomo Lucato, VR522918

# Introduction

According to the World Health Organization, stroke is the second leading cause of death globally and the third leading cause of disability. Statistics indicate that one in four individuals will experience a stroke in their lifetime. Additionally, 70% of strokes occur in low- and middle-income countries, where both lifestyle and medical risk factors are less effectively managed. These risk factors include obesity, physical inactivity, smoking, excessive alcohol consumption, hypertension, high cholesterol, and diabetes.

Since most of these factors are modifiable, early detection of individuals at risk is crucial for timely intervention. Identifying high-risk individuals as soon as possible allows for preventive measures to be implemented early, potentially reducing the likelihood of a stroke and its severe consequences. [1]

# Machine Learning in Stroke Prevention

Stroke remains one of the leading causes of death and disability worldwide, with prevention being a key factor in reducing its impact. As healthcare costs continue to rise, early and non-invasive stroke risk assessment has become increasingly important. Traditional statistical methods, such as regression-based models, have long been used to identify stroke risk factors and predict high-risk individuals. However, these approaches often oversimplify complex relationships, limiting their predictive accuracy, particularly when dealing with multiple interacting risk factors.

Machine Learning (ML) is emerging as a powerful tool in healthcare, offering the ability to detect hidden patterns within vast datasets that may not be immediately apparent to human analysis or conventional statistical models. Unlike traditional methods, ML algorithms can capture complex, nonlinear relationships between risk factors and stroke occurrence, leading to more personalized and accurate risk stratification. Techniques such as Support Vector Machines (SVM), Random Forest (RF), and deep learning have shown promising results in stroke prediction, outperforming conventional models in several studies.

Despite these advancements, ML in stroke prediction is still in its initial stages. Many models face challenges in interpretability, generalizability, and regulatory acceptance. Currently, no ML-based predictive models for stroke have been approved by the U.S. Food and Drug Administration (FDA) or other major regulatory bodies. For widespread clinical adoption, ML models must demonstrate reliability and comparable to established risk calculators and meet rigorous validation standards.

Nevertheless, the potential of ML in stroke risk assessment is significant. By leveraging large datasets, ML can refine risk stratification, improve early detection, and support better preventive strategies. Ongoing research and advancements in explainability and validation will be crucial in integrating these models into routine clinical practice. [2]

# Objective

The project aims to develop and evaluate Machine Learning models for stroke prediction by analyzing key risk factors and comparing multiple ML algorithms. The goal is to identify the most effective approach for early stroke detection, with a focus on recognizing individuals at risk based on established medical and lifestyle factors. Additionally, the study seeks to assess the relative importance of these factors in determining stroke risk.

# Problem Definition

The project focuses on a classification problem in healthcare domain, where the aim is to predict whether an individual is at high-risk of stroke. The target variable is binary: ‘1’ indicates that an individual has experienced a stroke, ‘0’ indicates those who have not.

# Dataset Description

To address the presented problem of stroke prediction, a binary dataset of individuals who have or have not experienced a stroke was chosen. The target variable in the dataset takes the values '0' (indicating no stroke) and '1' (indicating a stroke occurred).

The dataset contains the following features:

- **Id:** Integer value, representing a unique identifier for each individual.
- **Gender:** Categorical value, indicating the gender of the individual ("Male" or "Female").
- **Age:** Integer value, indicating the age of the individual.
- **Hypertension:** Boolean value, indicating whether the patient suffers from hypertension (1 for Yes, 0 for No).
- **Heart_disease:** Boolean value, indicating whether the patient suffers from any heart disease (1 for Yes, 0 for No).
- **Ever_married:** Boolean value, indicating whether the patient has ever been married (1 for Yes, 0 for No).
- **Work_type:** Categorical value, representing the patient's work type ("children", "Gotv_jobv", "Never_worked", "Private" or "Self_employed").
- **Residence_type:** Categorical value, representing the patient's residence type ("Urban" or "Rural").
- **Avg_glucose_level:** Integer value, representing the average blood sugar level of the patient.
- **Bmi:** Integer value, representing body mass index of the patient.
- **Smoking_status:** Categorical value, indicating the smoking status of the patient ("formerly smoked", "smokes", "never smoked", "Unknown").
- **Stroke:** Boolean value, indicating whether the patient has experienced a stroke (1 for Yes, 0 for No).

The dataset consists of 5100 patients, of which 4861 did not experience a stroke, while only 219 patients did. This significant class imbalance is a common issue in healthcare datasets, where the occurrence of specific events such as strokes, is much less frequent. Such imbalances can pose challenges in model training, as the model may be biased towards the majority class, making it harder to detect the less frequent, but crucial, cases of stroke.

# Data Preprocessing

Before applying the dataset to the stroke prediction models, it is necessary to preprocess the data. The raw dataset contains irrelevant features, missing values, outliers, and categorical variables, all of which need to be addressed, to ensure that the machine learning algorithms work on clean and consistent data.

## Data Cleaning

The feature “id” serves solely as a unique identifier for each patient, and as such, it does not contribute any meaningful information regarding stroke risk. Since this feature does not hold predictive value for the models, it has been removed from the feature set to avoid unnecessary complexity and ensure the models only use relevant variables for prediction.

Additionally, the “Gender” variable initially contained a value of “Other” for one sample. Since this value is out of the expected categories (“Male” or “Female”) and only appears in one instance, it was decided to replace this value with “Male”. This approach ensures consistency in the dataset while minimizing any potential disruption caused by a rare, unrepresentative category.

## Missing Values

In the dataset there are several missing values in the “bmi” variable that need to be addressed. To handle these missing values, a common strategy is to impute them with a reasonable estimate based on existing data. In this case, the missing bmi values were replaced with the median bmi for each group of patients defined by specific categorical features, including gender, marital status, work type, and residence type.

The rationale behind this approach is that the bmi values might vary across different demographic and lifestyle categories, so imputing based on the median of these specific groups ensures that the imputed values reflect the distribution of bmi within each subgroup. By using the median, we avoid the influence of outliers that might skew the imputation.

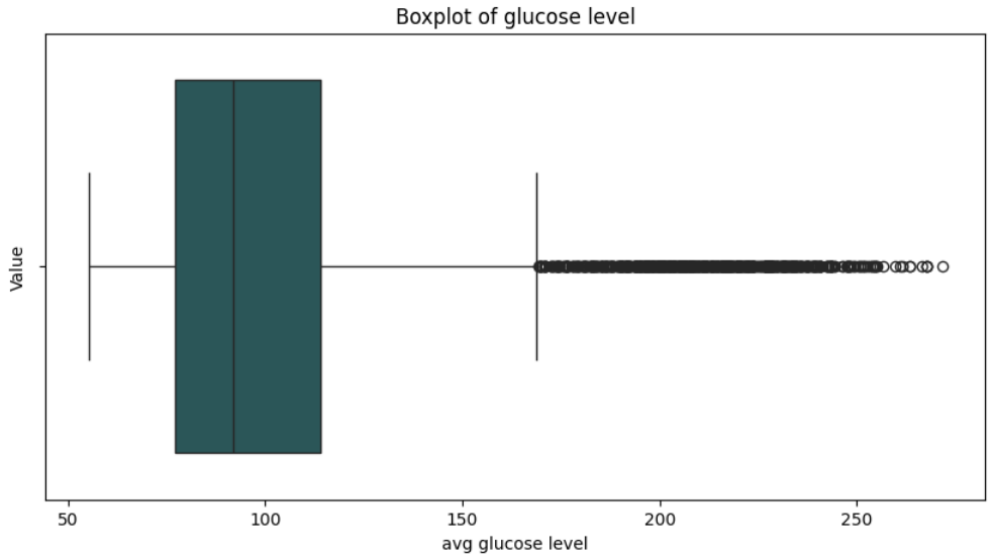
This method ensures that missing values are filled in a way that is contextually relevant to the individual’s characteristics.

## Outliers

Outliers are values significantly far from the center (mean) of their distribution. The presence of outliers can affect the performance of some classifiers, so they need to be handled carefully. Analyzing the distribution of the numerical features of the dataset emerged that both “avg_glucose_level” and “bmi” contain outliers.

### Outliers in avg_glucose_level

Glucose level refers to the concentration of glucose (a type of sugar) present in the blood. Below is reported the boxplot of the distribution of these values in the dataset:

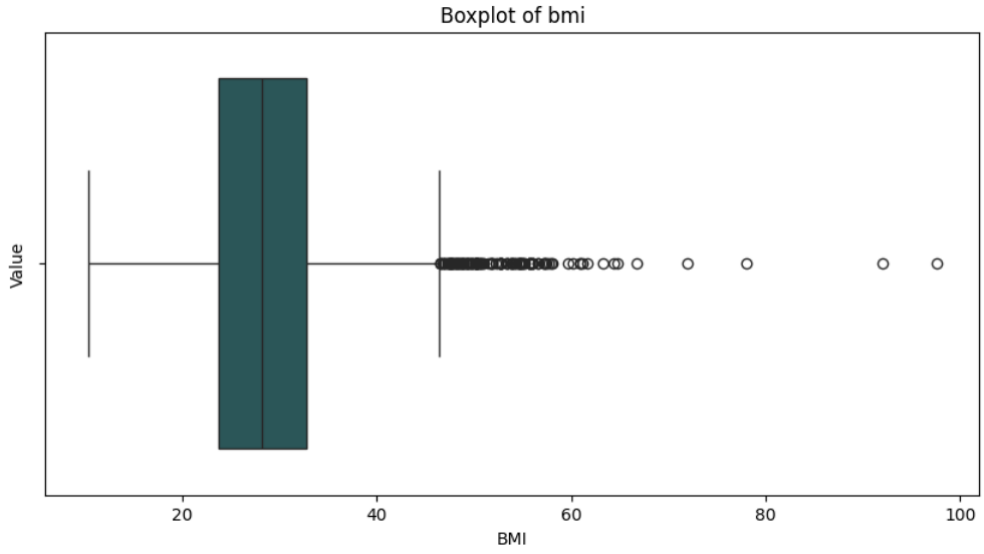



The plot shows that most of the values are concentrated around 70 mg/dL, but there are several outliers around 200 and 250 mg/dL.

According to data provided by the World Health Organization, values greater than 250 are common in people who suffer from diabetes. Since diabetes is commonly associated with stroke risk, these values represent valuable information for the models, and therefore they were not modified. [3]

## Outliers in BMI

BMI (body mass index) is a measure of the body fat present in the human body, calculated using a person's weight and height. Below is reported the boxplot of the distribution of the values of this feature in the dataset:



The plot reveals that most of the BMI values are concentrated between 25 and 30, with a notable presence of outliers above 50.

According to the World Health Organization guidelines, BMI values are classified as follows:

- **Below 18.5**: Underweight
- **18.5-24.9** : Normal Weight
- **25.0-29.9** : Pre-obesity
- **30.0-34.9** : Obesity Class I
- **35.0-39.0** : Obesity Class II
- **Above 40** : Obesity Class III

BMI values greater than 40 generally indicate severe obesity, and such values are expected to be rare in the general population. Based on WHO data, for a person with an average height of 75 inches (1.9 m) and an average weight of 431 pounds (195 kg), the maximum BMI would be approximately 54. [4]

Therefore, any BMI values exceeding 54 are considered unreliable and fall outside a reasonable range for this dataset.

To ensure data quality, all BMI values greater than 54 have been replaced by 54.

# Feature Correlation

Feature correlation was analyzed to identify variables strongly related to the target and to detect features that exhibit similar behavior. This analysis helps in eliminating redundant features during the feature selection phase.

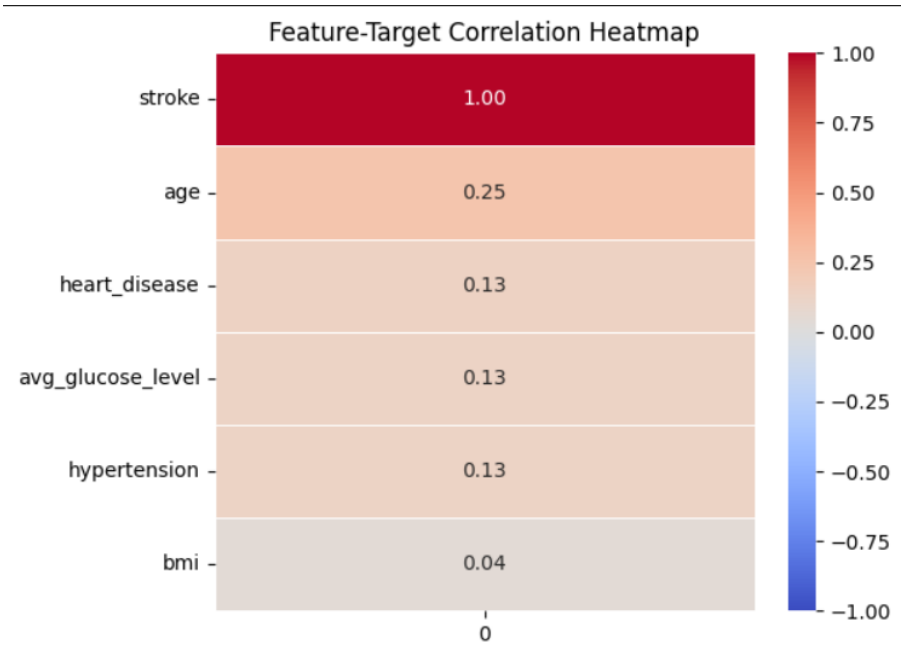
## Numerical Features and Target Variable

### Correlation Matrix

To understand the relationship between numerical features and stroke occurrence, a correlation matrix was computed. Correlation values range from –1 to 1:

- **Positive correlation** (close to 1) indicates that as one variable increases, the other tends to increase.
- **Negative correlation** (close to –1) indicates an inverse relationship.
- **A correlation close to 0** suggests little to no linear relationship between the variables.

The correlation matrix helps identify features that are most associated with stroke risk. However, it is important to consider that the dataset is highly imbalanced, with significantly fewer stroke cases than non-stroke cases. This imbalance may affect correlation values, potentially underrepresenting the true impact of certain features on stroke occurrence. Therefore, while correlation analysis provides useful insights, it should not be the sole criterion to determine feature relevance.




The correlation matrix indicates that there is little to no strong linear correlation between stroke risk and the other numerical features. Among them, age exhibits the highest correlation with stroke occurrence, with a coefficient of approximately 0.25, suggesting that older individuals are more likely to experience a stroke. No other numerical variable appears to have a particularly strong correlation with the target variable.

## Visualization of Feature Relationships with Stroke Risk

While the correlation matrix provides a quantitative measure of the linear relationship between numerical features and stroke occurrence, it does not capture nonlinear patterns or distribution differences between stroke and non-stroke cases. To gain deeper insights, violin plots were generated for each numerical feature against the target variable.

Violin plots combine aspects of boxplots and density plots, allowing for a more detailed view of the distribution of values within each category.

### *Age vs Stroke*

The following violin plot illustrates the distribution of age for both stroke and non-stroke cases:


The age distribution varies notably between individuals who experienced a stroke and those who did not. While non-stroke cases are spread across all age groups, with a slight concentration around 40 years, stroke cases are primarily clustered among older individuals, peaking between 70 and 80 years. This pattern indicates that age plays a significant role in stroke risk, with older adults being more susceptible.

This aligns to what emerged from the matrix correlation: as age increases, stroke risk increases.

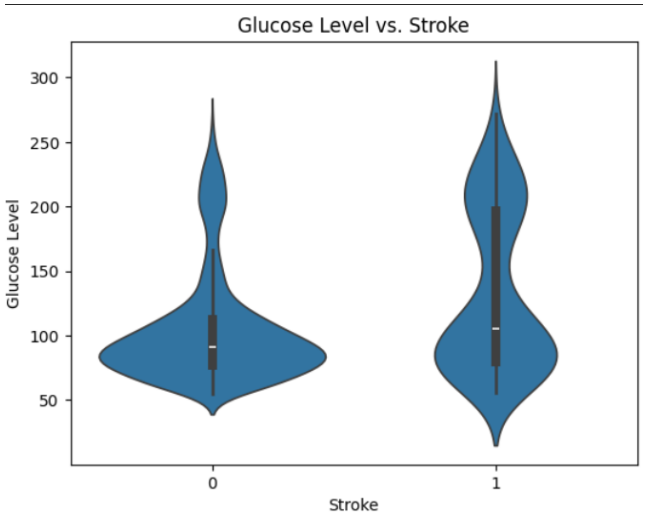
### *BMI and Stroke*

The following violin plot illustrates the distribution of bmi for both stroke and non-stroke cases:


The plot reveals that BMI values are similarly distributed across both groups, with a peak around 30 in both stroke and non-stroke cases. This aligns with the findings from the correlation matrix, indicating no strong relationship between BMI and stroke risk. However, this observation is specific to this dataset and is likely influenced by its imbalance.

### *Average Glucose Level and Stroke*

The following violin plot illustrates the distribution of average glucose level for both stroke and non-stroke cases:

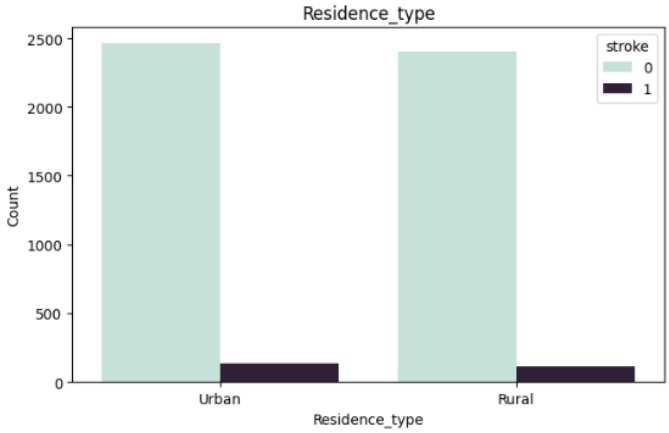



Similarly to bmi, values of average glucose level are similarly distributed across both groups, suggesting no strong relationship between target variable and glucose level.

## Categorical Features and Target Variable

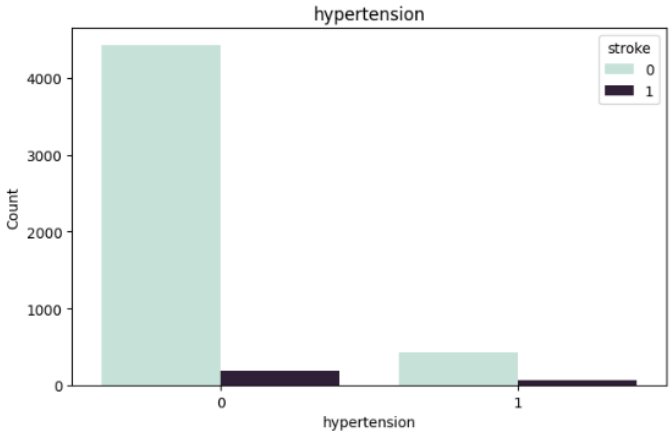
The dataset includes several categorical variables (smoking status, marital status, residence type, work type, gender), but no clear or strong relationship with stroke risk is immediately evident from their distribution. While some differences may exist, they do not suggest a significant pattern without further analysis.

As an example, the plot below illustrates the distribution of Residence_type, showing no significant trend between its categories and stroke occurrence. Similar results were observed for the other categorical features, suggesting that these variables may not be as strongly linked to stroke risk in this dataset.



## Binary Features and Target Variable

Finally, the distributions of the binary variables, heart_disease and hypertension, were analyzed. Although the correlation matrix did not reveal a strong relationship between these variables and stroke risk, the distributions suggest something more significant. Specifically, very few individuals without heart disease and/or hypertension experienced a stroke. This finding is crucial as it indicates that the absence of these conditions can serve as a strong indicator of lower stroke risk. Therefore, the presence of heart disease or hypertension can be considered valuable information for predicting stroke risk.




# Handling Categorical Variables

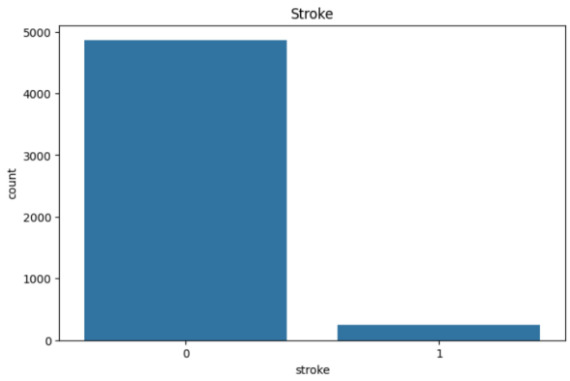
Before applying any machine learning algorithms, categorical variables must be converted into numerical format. To achieve this, one-hot encoding was employed, transforming each categorical feature into a set of binary indicators, with a separate binary column representing each possible category.

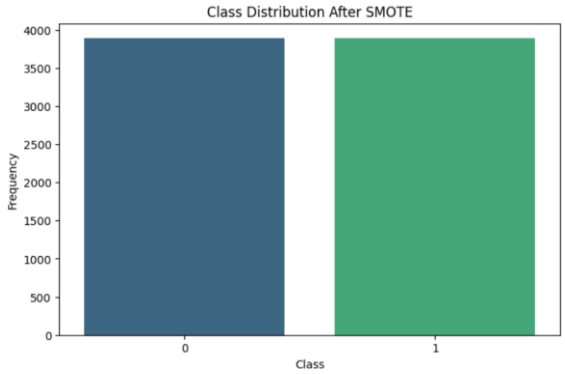
# Handling Imbalance

Class imbalance can pose significant challenges for certain machine learning models. To address this, the training set was resampled to ensure an equal representation of both classes: stroke risk and non-stroke.

SMOTE (Synthetic Minority Over-sampling Technique) was used to mitigate this issue. SMOTE is a widely used technique that generates synthetic samples for the minority class. It does so by selecting a sample from the minority class, identifying its k-nearest neighbors, and then creating new samples through interpolation between the original sample and one of its neighbors.

The following visualizations display the class distributions before and after the application of SMOTE:




# Model Selection

## Evaluation Metrics

In classification tasks, accuracy is commonly used as an evaluation metric. However, in cases where the dataset is imbalanced, accuracy can be misleading. This is because even if all instances of the minority class (e.g., stroke risk patients) are misclassified, the accuracy can still be high, as the majority class has more influence on the final result.

To address this issue, recall and precision are more reliable evaluation metrics. Recall measures the True Positive Rate, indicating how well the model detects positive instances (patient at risk of stroke). Precision, on the other hand, measures how many of the predicted positive instances are actually true positives, providing insight into the accuracy of positive predictions.

Additionally, f1-score, the harmonic mean of recall and precision, was used to balance these two metrics. Finally, confusion matrix was also employed to further analyze the performance of the model, providing a more comprehensive view on how predictions are distributed across all classes.

## Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. It is particularly effective for handling datasets with complex relationships between features, making it a strong candidate for stroke risk prediction.

In this project, Random Forest was applied to classify individuals based on their likelihood of experiencing a stroke. The algorithm was chosen for its ability to handle both numerical and categorical variables, its robustness to noise, and its capability to model non-linear relationships between risk factors and stroke occurrence.

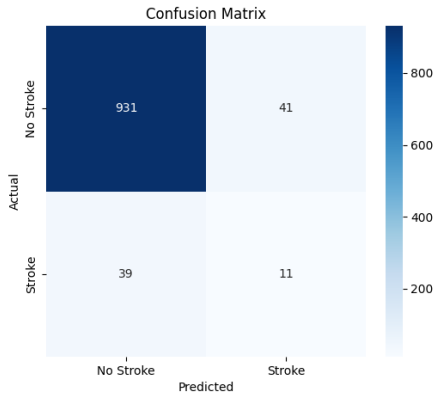
## Random Forest without SMOTE

Initially, Random Forest was trained on the original, non-resampled dataset to assess whether the available data contained enough information for effective stroke risk prediction.

When evaluated on the test set (a separate portion of the dataset reserved for validation), the model produced the following results:

- **Accuracy:** 0.92
- **Recall:** 0.22
- **Precision:** 0.21
- **F1_score:** 0.21

Below is the confusion matrix for the model's predictions, illustrating the distribution of True Negatives, True Positives, False Negatives, and False Positives:




These results suggest that the model struggles to accurately classify samples from the minority class, with most stroke cases being misclassified as non-stroke. This poor performance is likely due to class imbalance, which makes it difficult for the model to learn distinct patterns associated with stroke cases.

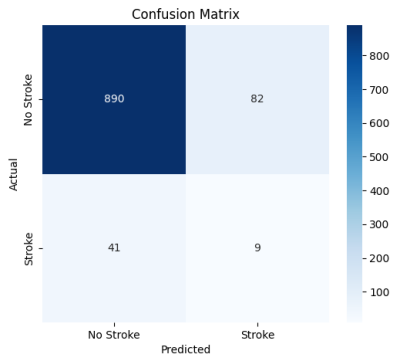
## Random Forest with SMOTE

To address the dataset imbalance, Random Forest was also trained on the resampled version of the training set, and evaluated on the original, non-resampled version of the test set. The following results were obtained:

- **Accuracy:** 0.87
- **Recall:** 0.18
- **Precision:** 0.09

- **F1_score:** 0.12

Below is the confusion matrix for predictions:




These results indicate that, despite resampling, Random Forest still struggles to accurately detect stroke cases, leading to poor performance and unreliable predictions. This issue is likely due to overfitting, a common consequence of working with imbalanced or synthetically resampled datasets.

## Random Forest Hyperparameters Tuning

To mitigate the effects of class imbalance and improve model performance, hyperparameter tuning was conducted to identify the optimal settings for Random Forest, aiming to enhance its ability to correctly classify stroke cases.

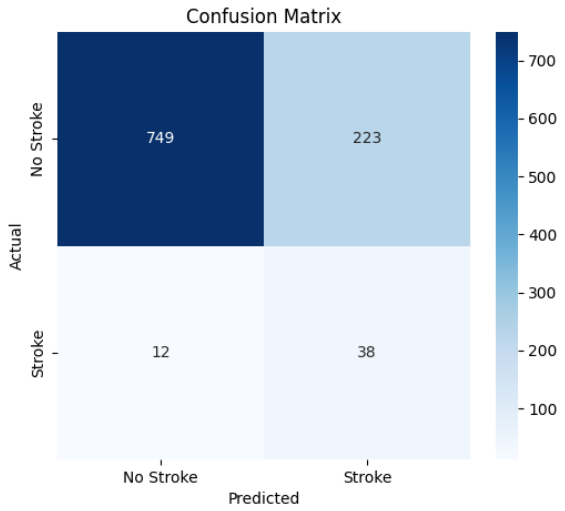
Hyperparameter tuning was performed using k-fold cross validation on the original training dataset, with the f1_score as primary evaluation metric to optimize the model's ability to classify stroke cases. The best hyperparameters found were:

- **Max_depth:** 6
- **N_estimators:** 50

When tested on the test set, these optimized parameters provided the following results:

- **Accuracy:** 0.77
- **Recall:** 0.76
- **Precision:** 0.14
- **F1_score:** 0.24

Below is reported confusion matrix for predictions:




These results indicate that the model successfully addresses the dataset's imbalance when training using a limited depth and a limited number of estimators, achieving a reasonable ability to predict samples from the minority class. While a notable number of non-stroke cases were misclassified as stroke-cases, this is less concerning in healthcare context, where prioritizing the detection of True Positives is often more critical than minimizing False Positives. Ensuring that at-risk individuals are identified is crucial, even if it means some non-stroke cases are mistakenly classified.

## K-Nearest Neighbors (KNN)

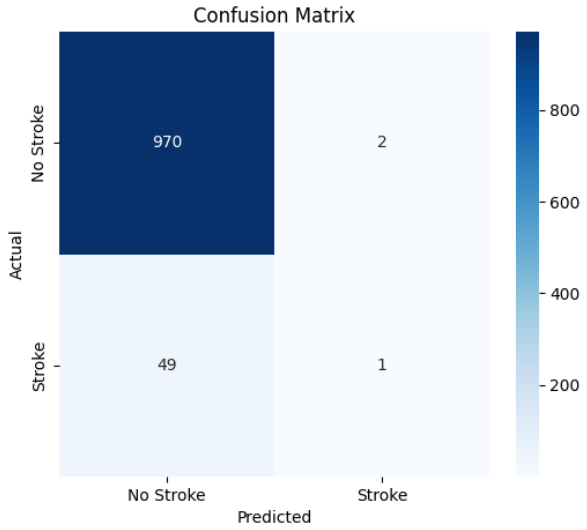
Although KNN is not typically the most suitable classifier for imbalanced datasets with categorical variables, it was still employed as part of the model evaluation process. Despite its potential limitations in handling class imbalance and categorical features, KNN was included to assess its performance on this specific dataset.

### KNN without SMOTE

KNN was first trained on original non-resampled dataset. The results are displayed below:

- **Accuracy:** 0.95
- **Recall:** 0.02
- **Precision:** 0.33
- **F1_score:** 0.03

Below is reported confusion matrix for predictions:



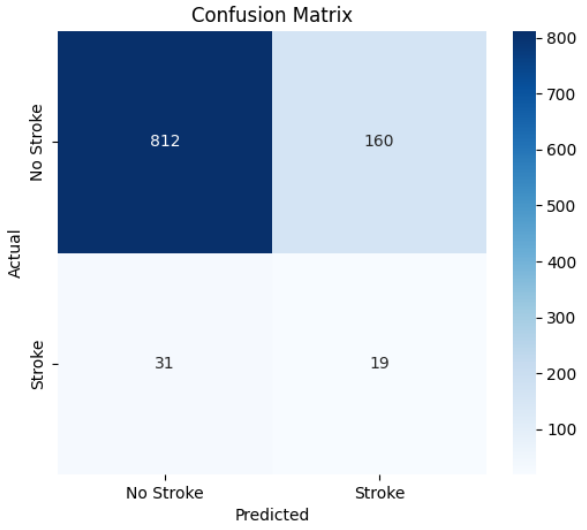

The results demonstrate that the model fails to predict any stroke cases, classifying all samples as non-stroke. This poor performance is likely due to the imbalance of the dataset and to the presence of many categorical features.

## KNN with SMOTE

To moderate the heavy impact of class imbalance on KNN model's performance, it was trained on the resampled version of the training set and then evaluated on the original version of the test set. Results are listed below:

- **Accuracy:** 0.81
- **Recall:** 0.38
- **Precision:** 0.10
- **F1_score:** 0.16

Below is presented confusion matrix for predictions:



Thanks to training set resampling, the model's performance on the test set appears slightly improved. However, it is still not good enough, since the model continues to struggle to detect stroke cases.

## KNN Hyperparameter Tuning

To improve KNN model's performance, hyperparameter tuning through cross validation was adopted to try to find a hyperparameter combination suitable for predicting stroke risk.

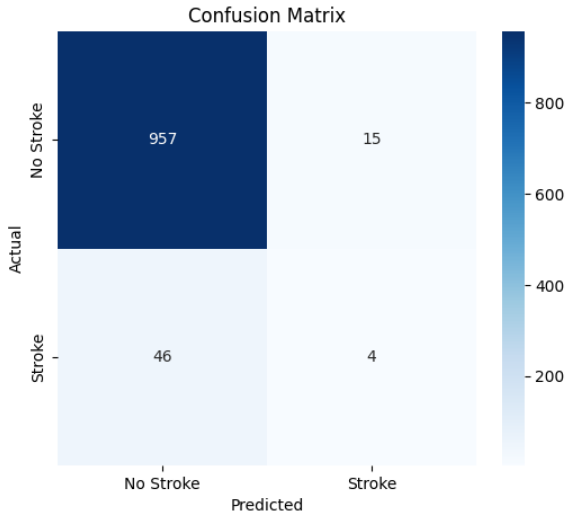
The best hyperparameter found were:

- **Metric:** manhattan
- **N_neighbors:** 3

The following results were obtained:

- **Accuracy:** 0.94
- **Recall:** 0.08
- **Precision:** 0.21
- **F1_score:** 0.11

The confusion matrix is reported below:




Despite hyperparameter tuning, the model still struggles to recognize stroke cases, failing to predict most of them.

The poor performance of the KNN model to predict stroke cases can largely be attributed to two factors. First, KNN is sensitive to class imbalance, which is prominent in this dataset, where the number of non-stroke cases greatly overweighs the stroke cases. This imbalance makes it challenging for KNN to effectively identify the minority class, resulting in the model predominantly predicting the majority class. Second, KNN struggles with categorical variables, especially when they are not properly encoded or when the relationships between the variables are complex. In this dataset, the presence of multiple

categorical variables, combined with the class imbalance, likely contributed to KNN's inability to correctly classify stroke cases. Consequently, despite its simplicity, KNN proved less effective compared to other models in this particular context.

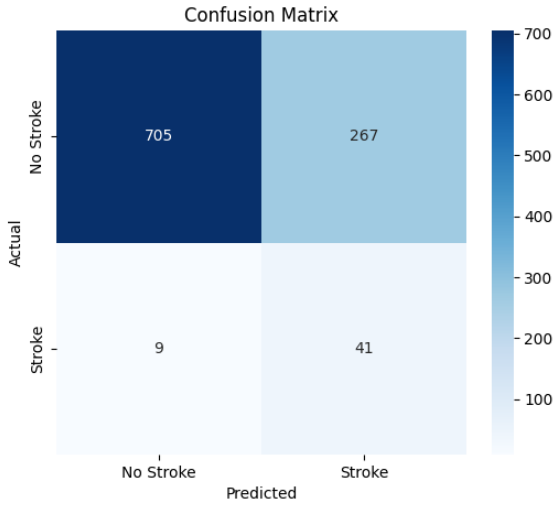
## Support Vector Machines (SVM)

### SVM without SMOTE

Training an SVM's model on the original non-resampled training set the following results were obtained:

- **Accuracy:** 0.72
- **Recall:** 0.82
- **Precision:** 0.13
- **F1_score:** 0.22

Below is presented the confusion matrix of the predictions:



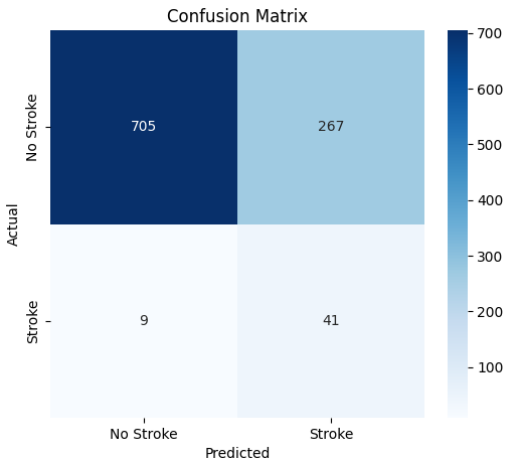

The results indicate that the SVM model was successful in detecting stroke cases, although some non-stroke cases were misclassified, similar to the behavior observed with the Random Forest model. Notably, these results were achieved without the need for resampling the training set or hyperparameter tuning. This suggests that SVM may be a promising model for stroke prediction. To confirm this, further analysis was conducted.

### SVM with SMOTE

Training the SVM model on the resampled training set leaded to the following results:

- **Accuracy:** 0.72
- **Recall:** 0.82

- **Precision:** 0.13
- **F1_score:** 0.22




The results obtained with the SVM model were consistent with those from the non-resampled dataset, indicating that class imbalance did not significantly impact the model's performance. This suggests that SVM is inherently more robust to imbalance compared to other models, making it a strong candidate for stroke prediction in this case.

## SVM Hyperparameter Tuning

After applying hyperparameter tuning to the SVM model, the results remained consistent with those observed both with and without resampling. This indicates that the model's performance was not significantly influenced by the choice of resampling or the fine-tuning of hyperparameters. It suggests that the SVM model is already well optimized for handling class imbalance in this specific dataset, and the selected hyperparameters were appropriate for maintaining class balance.

The chosen hyperparameters were:

- **C=1:** the parameter C controls the trade-off between achieving a low error on the training data and ensuring that the model generalizes well to unseen data (avoiding overfitting). A value of C=1 indicates a balanced approach, where the model tries to correctly classify the training data but also allows for some misclassifications to avoid overfitting. In cases like this where there is class imbalance or noise in the data, a moderate value of C helps the model focus on finding general patterns without being overly influenced by noise.
- **Kernel=linear:** the choice of a linear kernel suggests that the data might not require a highly complex, non-linear decision boundary to separate the stroke and non-stroke classes effectively. Even though the dataset is imbalanced, the linear kernel allows the SVM to find the best hyperplane for separating the two classes. A linear kernel is often preferred in cases where the decision boundary can be relatively simple and does not require intricate transformations of the data.

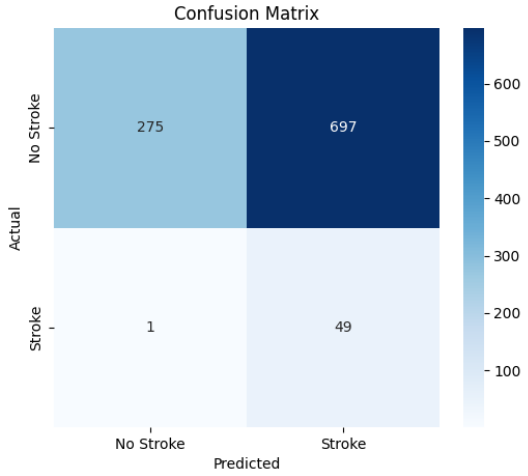
# Naive Bayes Classifier

## Naive Bayes without SMOTE

Testing Naive Bayes classifier on the original non-resampled dataset were obtained the following results:

- **Accuracy:** 0.30
- **Recall:** 0.99
- **Precision:** 0.06
- **F1_score:** 0.12

Below is reported the confusion matrix of the predictions:



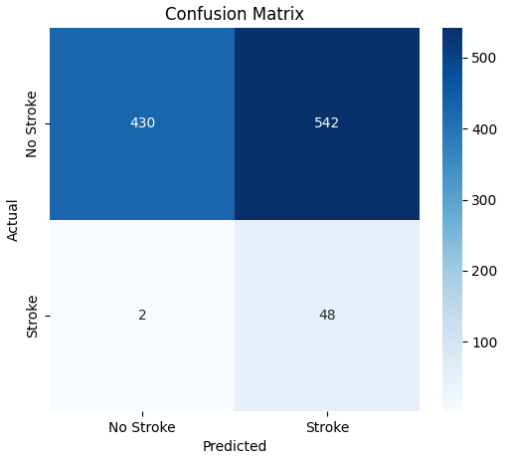

The Naive Bayes model demonstrated exceptional performance in identifying stroke cases, with a recall score of 0.99, making it the most effective classifier for stroke risk prediction so far. However, the model also misclassified a significant number of non-stroke cases as stroke cases. While in healthcare contexts accurately identifying True Positives is often prioritized over minimizing False Positives, the high rate of False Positives in this case is concerning. Such a high rate could result in an excessive number of unnecessary diagnoses and additional follow-up procedures for patients, which may lead to increased healthcare costs and patient anxiety.

### Naive Bayes with SMOTE

To lower the rate of False Positives, Naive Bayes model was trained on the resampled training set. The following results were obtained:

- **Accuracy:** 0.46
- **Recall:** 0.96
- **Precision:** 0.08
- **F1_score:** 0.15

Below is reported the confusion matrix of the predictions:




Thanks to resampling, the number of False Positives has decreased, but it remains concerning high. This suggests that Naive Bayes alone cannot be used as model for stroke prediction, but it could be a powerful tool combined with other classifiers.

# Models Comparison

The following table reports the results obtained by each classifier, helping to compare them and to understand which one is more suitable for stroke prediction:

| | Model | Precision | Recall | F1-Score |
|---|---|---|---|---|
| 0 | Random Forest | 0.14 | 0.76 | 0.24 |
| 1 | Random Forest (SMOTE) | 0.11 | 0.44 | 0.18 |
| 2 | KNN | 0.21 | 0.08 | 0.11 |
| 3 | KNN (SMOTE) | 0.08 | 0.20 | 0.12 |
| 4 | SVM | 0.13 | 0.82 | 0.22 |
| 5 | SVM (SMOTE) | 0.11 | 0.80 | 0.20 |
| 6 | Naive Bayes | 0.08 | 0.98 | 0.12 |
| 7 | Naive Bayes (SMOTE) | 0.08 | 0.96 | 0.15 |

The results clearly indicate that KNN is not suitable for stroke prediction tasks, as it is overly sensitive to overfitting and fails to accurately detect stroke risk cases. In contrast, The Naive Bayes model was able to identify stroke cases with excellent precision (high recall), but it struggled significantly with misclassifying non-stroke cases, leading to an unacceptably high rate of False Positives. Although resampling the training set showed some improvements, it was not sufficient to make Naive Bayes a reliable standalone model for stroke prediction.

On the other hand, the SVM and Random Forest models performed much better, managing to correctly identify stroke cases while keeping False Positives under control. These models demonstrated good balance in terms of recall and precision, offering a more reliable approach for stroke prediction. The combination of SVM's ability to handle imbalanced datasets without the need for resampling and Random Forest's robustness to overfitting positions these two models as the most suitable for this type of healthcare problem.

# Combination of Models

Given the Naive Bayes classifier's excellent ability to detect stroke cases and the strength of Random Forest and SVM in minimizing False Positives, these methods were combined to leverage their complementary strengths and explore their combined predictive capabilities.

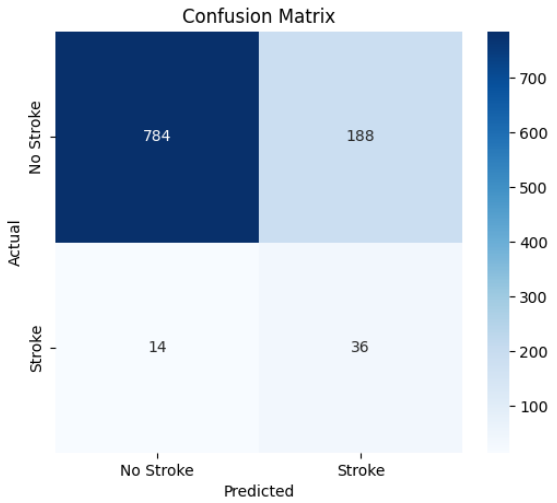
## Bagging Ensemble

Bagging is a powerful method that combines multiple base models to produce a more robust and accurate prediction. In this project, Bagging was used to combine the strengths of the models that had shown promising performance – Naive Bayes, Random Forest, and SVM – in hopes of further improving their performance by reducing variance and increasing generalization.

The following results were achieved:

- **Accuracy:** 0.80
- **Recall:** 0.72
- **Precision:** 0.16
- **F1_score:** 0.26

Below is reported the confusion matrix of the predictions:




Results show that Bagging does not achieve a better performance compared to SVM alone and Random Forest.

## Stacking Ensemble

In order to improve the performance of the model by leveraging the strengths pf different classifiers, a stacking ensemble approach was employed. Stacking involves training multiple base models and combining their predictions though a meta-model. The goal of this approach is to enhance predictive performance by utilizing a diverse set of classifiers that may each capture different aspects of the data.

For this analysis, three distinct base models were used: Naive Bayes (trained on SMOTE data), Random Forest (tuned) and Support Vector Machine. The predictions of these three base models were combined through a meta-model, which was another Naive Bayes classifier. The meta-model is trained on the outputs (predictions) of the base models, learning how to best combine them for final classification.

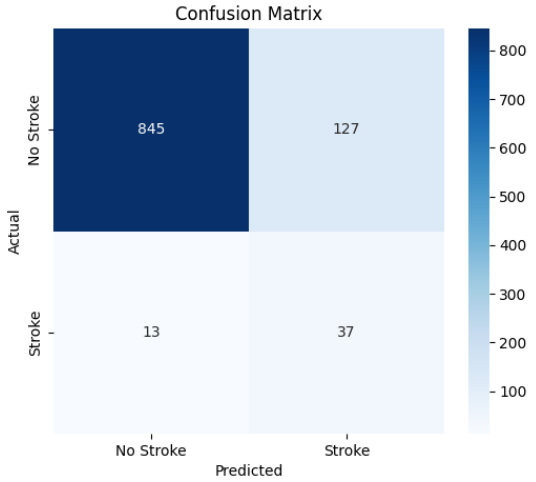
Cross-validation with 5 folds was used during the training process to ensure the model's generalizability and to prevent overfitting.

After training the stacking model, it was evaluated to the test set to assess its predictive performance. The results are reported below:

- **Accuracy:** 0.86
- **Recall:** 0.74
- **Precision:** 0.22

- **F1_score:** 0.34

Below is shown the confusion matrix for the predictions:




The results demonstrate that combining Naive Bayes, SVM, and Random Forest offers an effective trade-off between False Negatives and False Positives. This ensemble approach preserves the model's ability to accurately identify stroke cases when minimizing the number of False Positives. The f1_score achieved with this combination is the highest observed so far, highlighting the balance between precision and recall.

These promising results suggest that further experimentation with different combinations of machine learning models could lead to even more accurate predictions, enhancing the reliability and effectiveness of stroke risk detection.

# Conclusion

This project explored various machine learning models to predict stroke risk based on demographic, medical and lifestyle-related features. Given the significant class imbalance in the dataset, multiple techniques were employed to address the challenge it posed, including resampling methods and alternative evaluation metrics like precision, recall, and f1_score.

Among the individual models tested, Support Vector Machines and Random Forest demonstrated the best balance between detecting stroke cases and minimizing false positives. Naive Bayes, while achieving an exceptionally high recall, suffered from an excessive number of false positives, limiting its practical usability. K-Nearest Neighbors, on the other hand, failed to correctly classify stroke cases, confirming its unsuitability for this task.

To further enhance performance, ensemble learning techniques were applied. A stacking approach, combining Naive Bayes, SVM, and Random Forest, yielded the most promising results, achieving the best f1_score by effectively balancing sensitivity and specificity. This suggests that integrating multiple models can improve stroke prediction accuracy, making it a valuable direction for future research.

While the results are encouraging, there remain challenges to address. The dataset's inherent imbalance continues to influence model performance, and further refinement of feature selection, hyperparameter tuning, and additional ensemble methods could lead to further improvements.

Ultimately, this project highlights the potential of machine learning in stroke risk prediction and underscores the importance of model selection, data preprocessing, and evaluation metrics in handling imbalanced datasets.

# Bibliography


- [1] National Library of Medicine. (2017). Stroke Risk Factors, Genetics and Prevention. Amelia K Boheme, Charles Esenwa, Mitchell S V Elkind
  URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC5321635/
- [2] National Library of Medicine. (2023). Machine Learning and the Conundrum of Stroke Risk Prediction. Yaacoub Chahine, Matthew J Magoon, Bahetihazi Maidu, Juan C del
  URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC10326666/
- [3] World Health Organization. Mean fasting blood glucose
  URL: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/2380
- [4] World Health Organization. A healthy lifestyle – WHO recommendations