Predicting Stroke Risk with Machine Learning Using Health and Lifestyle Data

Contents

[Abstract 2](#_Toc183283808)

[Acknowledgements 2](#_Toc183283809)

[Chapter 1: Introduction 2](#_Toc183283810)

[I. Overview 2](#_Toc183283811)

[II. Problem Statement 2](#_Toc183283812)

[III. Research Questions 2](#_Toc183283813)

[Chapter 2: Background 2](#_Toc183283814)

[I. Dataset 2](#_Toc183283815)

[II. Literature Review 2](#_Toc183283816)

[III. Research Gap 2](#_Toc183283817)

[IV. Algorithms 2](#_Toc183283818)

[a. TFIDF 2](#_Toc183283819)

[b. Count Vectorizer 2](#_Toc183283820)

[c. Logistic Regression 2](#_Toc183283821)

[d. SVM 2](#_Toc183283822)

[e. XGBoost 2](#_Toc183283823)

[f. Decision Tree 2](#_Toc183283824)

[g. Artificial Neural Network 2](#_Toc183283825)

[Chapter 3: Methodology 2](#_Toc183283826)

[I. Tools and Techniques 2](#_Toc183283827)

[II. EDA and Visualization 2](#_Toc183283828)

[Chapter 4: Results and Conclusion 2](#_Toc183283829)

[I. Critical Discussion 2](#_Toc183283830)

[II. Conclusion 2](#_Toc183283831)

[III. Future Work 2](#_Toc183283832)

[Chapter 5: Legal, Ethical and Professional Issues 2](#_Toc183283833)

[References 2](#_Toc183283834)

[Appendices 2](#_Toc183283835)

# Abstract

Stroke, therefore, is an important health problem for people of all ages, making the top five killers and disabling disorders globally. Proper and successful risk assessment for the stroke can greatly improve the patient condition because of early treatment. In this research, the performance of ML models that predict the vulnerability of a stroke patient is determined by five attributes; hypertension, heart disease, glucose level, BMI, and age. The research we use the “Stroke Prediction Dataset” available on Kaggle, the dataset provides information about more than 5000 patients with multiple characteristics, to build the prediction models and investigate the impact of the health state on the probability of a stroke.

Based on some machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, XGBoost and Artificial Neural Networks, the project builds strong predictor models. Basic data cleaning processes include cleaning for missing data, categorical data, converting them into encoding form, and normalizing the continuous features. Qualitative analysis such as the chi square tests and correlation analysis shows that stroke has a significant relationship with other medical conditions including Hypertensive heart disease. It is discovered that patients with hypertension and heart disease are 20% more likely to suffer a stoke than patients with either of the diseases at 13% or patients with no diseases at 3%.

General performance measurement tools are compared, including Accuracy, F1-score, and ROC AUC. Among all 6 methods applied, Logistic Regression and Random Forest get the highest accuracy of 95.09%, while Logistic Regression can be mentioned to have the best discriminative power, having ROC AUC as high as 0.807. The study underscores that hypertension and heart diseases are significant determinants of stroke and need to be included in such features.

This research also discusses legal, ethical, and professional concerns such as legal responsibility regarding prediction in clinical decisions, fair sharing of prediction methods in clinics, and the importance of model transparency. Future work shall then consider predicting larger databases, including other parameters into the model such as cholesterol levels and socioeconomic status, and using deep learning techniques to further improve the capabilities of the current model. It aims to establish valid, understandable and usable prognostic tools in the early stage of the stroke for diagnosis and improved clinical data for better patient management while adhering to high intra and inter-organizational, ethical and equitable principles of medical practice and health care delivery.

# Acknowledgments

# Chapter 1: Introduction

## Overview

The studies presented in this paper are related to the application of machine learning algorithms for the prediction of stroke risk for increasing the efficacy of early detection of the disease and for improving the patients’ outcomes (Holland, 2018). Stroke is still a leading cause of death and chronic impairment; therefore, proper and prompt prediction is essential. To assess the stroke risk factor, the project focuses on the patient’s characteristics such as hypertension, heart disease, age, smoking status, and body mass index. The applied dataset is based on more than 5000 patients and consists of categorical and numerical variables.

The research focuses especially on the synergetic impact of hypertension with heart disease on stroke risk and the critical importance of both variables for prediction. This two-stroke study performs the evaluation considering five machine learning algorithms, namely Logistic Regression, Decision Trees, Random Forests, XGBoost, and Neural Networks. The performance of each model is evaluated based on accuracy, F1-Score, and respectively, ROC-AUC, to select the model that will best be implemented in real-life scenarios.

The end vision of the research work is to erect an early warning system that clinicians, caregivers, and health practitioners can use to recognize those patients who are most likely to suffer from a stroke and then prevent it. In this context, this project will also have the objective of presenting how the different machine learning models can contribute to enhancing clinical decisions and, consequently, the patient’s treatment.

## Problem Statement

A review of the current research in the publication International Journal of Stroke indicates that there is a rising trend in the occurrence of stroke globally (*Stroke statistics*, no date), and this prompts a call for the development of reliable models for the prediction of the occurrence of the same. Stroke analysis as a diagnostic problem remains unsolved due to a multitude of health conditions contributing to the likelihood of a stroke including hypertension, heart disease, BMI, age, and smoking. Hypertension and heart disease are expected to be major precursors of strokes and they are well established but still, the synergistic impact of hypertension heart disease in stroke prediction is often overshadowed or not well investigated. However, it was found that ‘the probability of having a stroke in the presence of both conditions as compared with having only one of them or none of them.’ This lack of knowledge requires a critical examination of the relationship between hypertension and heart disease with other factors associated with stroke risk.

This research aims to answer the following questions: How much do hypertension and heart disease affect stroke risk? Does a history of hypertension plus heart disease increase one’s risk of stroke more than a history of either but not both? The first and foremost goals are, therefore, to evaluate these health attributes for the prediction of stroke using different classifiers such as Logistic Regression, Decision Tree, Random Forest, XGBoost, and Artificial Neural Networks. Through feature importance, ANOVA, and correlation analysis comparing means, the study aims to identify variables of great influence on stroke. Further, to contribute to the research question, the study will aim at creating an effective preprocessing step to process the data, and then produce predictive models that yield high accuracy which will be measured by accuracy measure, F1 score, precision, and recall measurements. In the long run, it aims at improving the ability to predict a stroke and possibly bringing further improvements to the field of health interventions.

## Research Questions

1. How significantly do hypertension and heart disease influence stroke prediction?
2. Do patients with both hypertension and heart disease have a higher likelihood of stroke compared to those with only one or neither condition?

## Objectives

* To assess the effect of health attributes (like hypertension and heart disease) in the prediction of stroke using the several learning models; Logistic regression, Decision tree, Random Forest, XGBoost and Artificial Neural Networks.
* To evaluate the degree of significance of each variable that has been selected, statistical models including feature importance, ANOVA, and correlation analysis will be used.
* The goal of this system would be to create a preprocessing pipeline and clean the dataset in addition to creating features and training models.
* To make predictions with high accuracy using metrics Such as, Accuracy, F1-Score, Precision & Recall.

# Chapter 2: Background

## Literature Review

The study by (Satapathy *et al.*, 2023) focused on evaluating important indicators for stroke risk determination like hypertension, heart diseases, glucose levels, and gender and age. A dataset with 5000 observations and 12 variables from the Kaggle site was used to build a machine-learning pipeline. This pipeline involved the preprocessing of datasets used, training of the models, and then their performance assessment. Logistic Regression, Random Forest, Decision Tree, and Support Vector Machines were used for the classification of risk as high or low. Of them, the Random Forest model had higher accuracy and gave more precise results among all the models.

The raw data was pre-processed effectively to enhance input parameters in a centralized dataset. Linear scaling reduced quantitative information to the same magnitude and rank value whereas correlation analysis established noteworthy co-linearity. Most of the graphical features; common mappings such as heatmaps, and hired referrals among the dataset attributes. The dataset was divided into training and testing subsets in a 70:30 split, to achieve a comprehensive assessment of the models’ quality. The datasets were pre-processed and the subsequent machine learning models were trained on these datasets using methods pertinent to classification problems.

Other features like exactness, specificity, recall, and precision measured the performance of the models. Out of all algorithms, Random Forest which uses multiple decision trees harvested an accuracy of 94.50% and a precision was .95. Indeed, Support Vector Machines and Logistic Regression yielded the same accuracy and precision, both of which at 94.50 percentage and 0.95 respectively. Decision Tree, with an accuracy of 0.9103 and precision of 0.93 performed worst among all. Random Forest outperformed other classifiers due to its potential to provide predictions from different decision trees; thus, it was the most accurate classifier in this work.

Coefficients from confusion matrices and Visual Assessment of the ROC curves supplemented the information concerning the discriminative abilities of algorithms. Again and again, Random Forest proves to have the best possible sensitivity as well as specificity, signifying real competency in producing the classification. On average, we observed similar performance of Logistic Regression and Support Vector Machines; however, Random Forest outperformed in terms of the ensemble. It was also observed that the Decision Tree had some weaknesses in dealing with a large number of features in the dataset as it had the least accuracy and precision as compared to other classifiers.

To identify the degree to which ML can help advance the field of healthcare, the study used a rigorous method of data preprocessing and feature extraction and performed a thorough model assessment. The cross-sectional observations that emerged provided pointers as well to glucose levels and older age, even among females, as important determinants of increased stroke risk. It also emerged that Random Forest is an ensemble method that performed well when dealing with the various attributes, and thus is useful in real-world applications. In this context, the integration of the received user-input data with the results of medical tests demonstrated the applicability of the highlighted framework.

This research established that point-of-care machine learning has a critical role in early detection of stroke. The Random Forest algorithm stands out as the most accurate showing its feasibility as a decision-aid tool for medical practitioners. Its accuracy rate was established at 94.50% which confirmed its predictive power on stroke risk. Suggestions for further research are the use of greater sample size and the application of more sophisticated feature selection techniques to enhance the predictive mode. Higher advancement may increase objectivity and widen uses in healthcare management; bringing more value to patients and facilities.

Cerebrovascular accident or simply stroke is a leading global health crisis, and in fact, ranks second as the leading cause of death globally. Identifying reaches early is essential in achieving good results and decreasing mortality rates. This research (Prasad *et al.*, 2024) aims to point out how current methods in stroke prediction can be supplemented through improved machine learning methodologies. With the help of models like eXtreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LGBM), this study exhibits how improved these models’ accuracy and robustness are. For instance, while comparing the capability of XGBoost to other algorithms, it reaches an accuracy of 98% which makes it pertinent for clinical utility. It was evident from the findings that machine learning innovation has taken a new dimension towards the diagnosis of stroke to indicate the need to embrace it and integrate it into the usual practice of health care.

stroke means a blood supply interruption means that the brain is starved of oxygen and nutrients. This condition can impact cognitive, emotional, and physical health so much. Where intervention is required, early detection is critical, but standard testing processes are usually slow and imprecise. This research seeks to fill this gap by applying machine learning algorithms that are capable of analysing rich medical data patterns. Taking early detection as the primary direction, the study presents hybrid feature selection methods, as well as advanced algorithms such as XGBoost, and LGBM to redesign the process of stroke prediction. These developments not only provide proven accuracy but also allow healthcare providers to take anticipatory action that can overall boost patient care greatly.

The dataset embraces 5110 patients’ records including 11 features; age, hypertension, glucose level, smoking, etc., and body mass index. Preprocessing of data is the first step of the analysis in this research work as it helps in maintaining the quality of the data by handling missing values, normalizing attributes and variables, encoding nominal data, and dealing with outliers. To overcome bias in the data, the data is made equally composed and then partitioned into 80-20 training and test data sets. This careful method protects the generation of accurate and precise predicting models.

XGBoost stands out in this research as an algorithm that leverages a strong gradient-boosting technique. It elaborates the predictions progressively by constructing the decision trees that minimize the errors made in previous steps hence reducing bias. Feature selection, hyperparameter tuning, measures for feature importance, and interrogation of the internal model through cross-validation again improve on this. Apart from XGBoost, LGBM is used because of its fast-learning capabilities in terms of working with big data. LGBM also offers improvement to the performance of the model by turning its attention to the important set of features and lower use of memory. These characteristics make both algorithms very appropriate to be used in predicting a stroke.

Employment of experimental assessments indicates that those models yield far superior performances. XGBoost delivers a perfect accuracy of 98% and precision, recall, and f-score as well are above 96%. LGBM is not far behind with an accuracy of 96% confirming the efficiency of the model. These findings therefore highlight the significant importance of machine learning for new benchmarking standards in stroke prediction pointing to the increased precision, possibility, prevalence, and rapidity of these enhanced algorithms in practice.

To that end, feature importance analysis has established glucose levels, hypertension, and age as predictors of stroke in the research. This insight helps direct the aims and goals of clinical assessments towards risky attributes, thus increasing diagnostic acuity and timeliness of interventions. The effectiveness of XGBoost and LGBM models in this study implies the potential of these models for broader application to large-scale problems within the healthcare sector to reduce mortality through timely diagnosis of various diseases.

Following this research, this paper identifies the following areas for future research: Consequently, the application of the model to more extensive and different databases would increase the versatility and stability of the results. The studies of deep learning approaches could be fragmented to advance the predictive prowess in order to identify more complex relationships. Further, proposed solutions of boosting model interpretability and incorporating uncertainty estimation would further aid in the adoption of its predictions in clinical settings. Such enhancements could help bridge the gap between the application of machine learning models and actual efficient use cases in healthcare, guaranteed reliability.

The findings of the present research are not only applicable to the realm of stroke risk prediction. In this regard, the study offers an excellent opportunity to implement new approaches to solving the most vital and urgent problems of human health using the possibilities of machine learning, as well as modern knowledge of medicine. Combining such algorithms as XGBoost and LGBM into healthcare practices enables early intervention and precise diagnosis of developmental diseases which can improve patients’ quality of life tremendously. Besides contributing to enhanced stroke detection, this research also underlines the importance of the application of machine learning in the development of several healthcare technologies, ongoing improvement of treatment methods, and saving the lives of millions of patients worldwide through accuracy and productivity.

A stroke occurs when blood flow to the brain is interrupted, leading to a lack of oxygen and nutrients, which can cause irreversible damage to nerve cells. This damage may result in death, long-term disability, or permanent brain injury. Timely detection of stroke is critical to minimizing its impact, and early intervention can significantly reduce mortality rates. The challenge lies in accurately predicting strokes at an early stage to enable quick medical intervention. This study leverages machine learning (ML) algorithms to develop a prediction model capable of identifying individuals at risk of suffering a stroke, thereby facilitating early diagnosis and treatment.

The dataset used in this research (Srivastav, Guleria, and Sharma, 2023) was sourced from Kaggle, consisting of 4,982 patient records with 11 features, including age, hypertension, glucose levels, smoking status, and body mass index. The dataset includes 249 cases of stroke, labelled as "Yes," and 4,734 non-stroke cases, labelled as "No." The data was pre-processed to handle missing values, eliminate duplicates, and standardize the features for better prediction accuracy. This pre-processed dataset was then used to train and test various machine learning algorithms, including Naïve Bayes (NB), Decision Tree (DT), K-Star, and Logistic Regression (LR).

The primary goal of the research was to assess the performance of these algorithms in predicting the likelihood of a stroke. The model's effectiveness was evaluated using several metrics, including accuracy, precision, recall, F1-score, and root mean square error (RMSE). These metrics provide a comprehensive view of how well the model identifies true positives, minimizes false negatives, and balances precision and recall. The logistic regression algorithm emerged as the top performer, achieving the highest accuracy of 95.02%, followed by decision tree (94.80%), K-Star (93.88%), and Naïve Bayes (88.40%).

Logistic regression, a binary classification algorithm, was particularly effective due to its simplicity and ability to handle linearly separable data. Its ability to calculate probabilities made it an ideal choice for predicting stroke occurrence. In comparison, other algorithms like decision trees and Naïve Bayes performed well but did not surpass the logistic regression model in terms of accuracy. While decision trees are known for their interpretability and versatility, they are prone to overfitting, especially with deep trees. K-Star, an instance-based learning method, relies on distance-based metrics to classify instances, but its performance was also less accurate than logistic regression.

The performance evaluation also included a comparison of precision, recall, and F1-score, where logistic regression maintained a strong lead. Precision, which measures the proportion of true positives out of all predicted positives, was highest for logistic regression (0.928), closely followed by the decision tree (0.912). Recall, that the metric that gauges the ability of the model to identify all relevant instances, was also highest for logistic regression (0.95). The F1-score, which is the harmonic mean of precision and recall, reinforced logistic regression’s superior performance. Additionally, the RMSE, which quantifies the error between predicted and actual values, was lowest for logistic regression, indicating its ability to provide highly accurate predictions.

A comparison with existing stroke prediction models revealed that the proposed logistic regression model outperformed other techniques in terms of accuracy. For instance, models using algorithms like K-Nearest Neighbours (KNN) and Support Vector Machine (SVM) reported accuracies of 85.02% and 90%, respectively, while the proposed logistic regression model achieved an impressive accuracy of 95.02%. This underscores the effectiveness of logistic regression in predicting stroke risk based on the given features.

In conclusion, this study demonstrates the significant potential of machine learning in stroke prediction. The logistic regression model, in particular, provided the highest accuracy and performance metrics, offering a promising tool for early stroke detection. The results suggest that further research and expansion of the dataset could further improve the model’s predictive power. Additionally, the integration of advanced machine learning and deep learning techniques in the future could lead to even more robust models, capable of handling larger and more complex datasets. Ultimately, the application of such predictive models in healthcare could help reduce the incidence of strokes and improve patient outcomes through early detection and intervention.

## Research Gap

This is a significant research gap in my project because apart from an accurate machine learning model, more focus is given to the feature engineering stage and a strict examination of several machine learning models. Although earlier research focuses on a small number of models like Logistic Regression, Decision Trees, and Random Forest, my experiment uses more variables, for instance, XGBoost, and Neural Networks. They also quantify model performance in other measures such as accuracy, F1 measure, and ROC AUC which gives a broader view of each of the model’s predictive abilities.

In addition, my project presents a new aspect of the relationship between the features – hypertension and heart disease and the risk of stroke with the help of Chi-squared testing. This goes further than feature selection by investigating how these conditions and features are dependent on one another to affect the possibility of a stroke occurring which is not often deeply discussed in present literature.

Additionally, my project pays special attention to a stricter data preprocessing procedure, such as one-hot encoding, Min-Max scaling, and missing value handling, to provide the best models that are trained with the prepared data. This step for preprocessing, together with the use of several assessment metrics and models, helps to avoid a critical drawback of existing stroke prediction studies, which is the high variability of the predictive accuracy but the lack of comparable interpretability of the achieved results.

## Algorithms

### Logistic Regression

Logistic Regression is a statistical technique used in data analysis that results in nominal dependent variables. This happens by finding the likelihood of a target variable belonging to a given class by the sum of independent variables. Logistic Regression in its simplest form uses the same sigmoid function to map the inputs to an output between zero and one, perfect for classification purposes.

The algorithm estimates the odds of an obtained dependent variable belonging to a specific class and establishes a linear model. The values obtained are then passed through the sigmoid function so that we get probabilities (Anshul, 2021). For instance, in stroke prediction, Logistic Regression gives probability in the event the patient gets a stroke (1) or not (0) under considerations of age, hypertension, and glucose levels. An operating point is defined by cross entropy and a decision threshold is usually taken as 0.5.

Logistic Regression’s linear relationship supposition between features and the log odds of the outcome makes it easy and comprehensible. It employs the route of maximum likelihood estimation to set optimum the coefficients of the model. The data should be debugged and the use of the L1 (LASSO) and L2 (Ridge) (Jain, 2016) models for regularization can help to reduce overfitting and increase the accuracy of the model. Still, Logistic Regression is a powerful tool for datasets where the relations between variables and outcomes are linear as closely as possible.

### Decision Tree

Decision Trees are tree-structured procedures used in classification and regression analyses. They use the work by dividing data into subsets based on the values of input features. Each internal node corresponds to an attribute, whereas each branch corresponds to a decision and each terminating node corresponds to the class label or the regression value (“What is a decision tree?,” 2024).

The splitting process is facilitated by measures such as the Gini index, entropy (in the case of information gain), or even variance. For instance, in stroke prediction, a Decision Tree might split values into data with Age or Glucose level thresholds and the branches that denote the patients as Risky or Non-Risky to contract a stroke.

The algorithm proceeds in a way that evenly splits the data set to minimize the number of misclassifications at the nodes where the data is split the algorithm splits the dataset into increasingly finer splits until it reaches the nodes of the tree that are all pure or are meets some predetermined terminal conditions. Advantages of Decision Trees: The Decision Trees can handle both numerical and categorical data and are easy to interpret but it has the disadvantage that it is liable to overfitting if the tree is too complex (Ravindran, 2023). This is done to enhance the model generality and other techniques like pruning where some inputs or branches that do not add much to the predictor’s ability to make a diagnosis are cut off. Decision Trees are highly useful when used on datasets that have some sort of relation from top to bottom.

### Random Forest

Random Forest is the technique based on a Decision Tree that increases the classification or re-grading by using several Decision Trees randomly. It does this by establishing a ‘forest’ of trees, each of which is trained on a different random sample (bootstrap sample) of the overall dataset and which operates independently. The last prediction is computed by combining the output produced: in the case of classification, using a simple voting system, while in the case of regression, taking an average of the results (Baladram, 2024).

Random Forest also includes randomness in addition to node selection in the tree, so each tree reaches for a distinct set of data. This minimizes overfitting compared to the building of a single Decision Tree and improves the model’s generalizability (Nouman, 2021).

Finally, in strokes, Random Forest assesses features such as hypertension, heart disease, and age in several trees after which it sums up the results to classify the degree of risk of strokes. From the algorithm, the feature importance measures are obtained that illustrate which variables are more important in making a decision. Random Forest works perfectly when there are interactions between the variables that do not have a linear relationship. But it may take a long time to calculate or when the data size or number of trees increases it may be very time-consuming.

### XGBoost

Gradient boosting is an extension of boosting that allows making the number of weak learners in the tree model very large; XGBoost, or eXtreme Gradient Boosting, is a robust implementation of gradient boosting. It creates a set of weak learners (mostly decision trees) where every next learner tries to minimize mistakes made by the earlier one (Sonawane, 2023). The reason is that the structural variation of this iterative procedure reduces a loss function, for example, log-loss in classification tasks, by refining model predictions.

XGBoost uses L1 (Lasso) & L2 (Ridge) norm-based standards to control over-fitting and is thus more robust to overfitting than other boosting algorithms. It also employs shrinkage in which the output of each tree is called by a learning rate to avoid large shifts.

I find using the algorithm relatively easy, especially when dealing with missing values and categorical variables which makes it appropriate for any dataset. In the prediction of stroke, risk factors such as age and glucose levels that may interact with each other can be accurately captured by XGBoost. Hyperparameters such as the number of trees, tree depth, and learning rate mean that optimization is crucial for achieving its best result. The advantage of speed, scalability as well as accuracy make XGBoost one of the most popular algorithms in structured data applications.

### Neural Network

Neural Networks are computational models using multiple layers of interconnected points called nodes (or neurons). Each neuron takes input data together with weight and bias, applies them, and then runs the result through some activation function, like ReLU or sigmoid, to introduce the nonlinearity into this simple structure (Aytan, 2021).

Typically, Neural Networks have an input layer, an output layer, and one or more hidden layers. The hidden layers transform the raw data (age, glucose levels) to your hidden layer results by weighted connections on the input layer. Later, the final prediction like the likelihood of a stroke is provided in the output layer.

Backpropagation is a process by which the network learns by adjusting the weights and biases with a process of calculating the gradient of the loss function (i.e., mean squared error or cross-entropy loss) with respect to each parameter. This is the loss to be optimized, for example, using stochastic gradient descent (SGD), Adam, and so on, to improve the model accuracy (Kalirane, 2023).

Capturing complex, but nonlinear, relationships in data is what Neural Networks do best, but also need large datasets and are computationally intensive. Although these are less interpretable than other models, they can be very accurate (as with stroke prediction given sufficient training data). Dropout is a regularization technique designed to prevent overfitting and improve generalization.

# Chapter 3: Methodology

## Dataset

The data in the dataset covers the patient's features including the patient’s age, gender, the diseases that have ever had, hypertension, heart disease, smoking, the type of work, etc. This resulted in 12 columns as well as 5110 records (“Stroke Prediction Dataset,” 2021). The dependent variable is ‘stroke,’ where 1 represents a patient diagnosed as having a stroke and 0 otherwise.

## Proposed Work

The work outlined for this research project is as follows to achieve a detailed examination of the influence of hypertension and heart diseases on stroke prediction. In these sections, the authors pay much attention to the steps that include data preprocessing, feature engineering, modeling/evaluation, and statistical analysis. The planned approach will be as follows:

1. **Dataset Collection and Preprocessing**

In this section, the data set is obtained from the Kaggle website and imported into the environment. The dataset consists of several parameters which include medical parameters like Age, BMI, etc, diseases like hypertension, heart disease, and so on, and lifestyle parameters like smoking and the like. After loading the data, the following preprocessing steps will be performed:

* Handling Missing Values: Such records missing BMI value will be assumed to have the column mean to ensure that the formula yields an actual value.
* Feature Encoding: Other categorical variables like; gender, marital status, work type, residence type, and smoking status will be encoded as numbers using one hot method.
* Feature Scaling: The values like age, hypertension, heart disease, average glucose levels, and BMI will be normalized, and will be converted into the (0, 1) scale by Min-Max scaler to make all features contribute to both models equally.

2. **Exploratory Data Analysis (EDA)**

Here various analyses such as graphical and statistical shall be done to identify correlations between various features and the target class (stroke). Key activities include:

* Histograms and Box Plots: Age, BMI, and risk of stroke: Exploratory analyses of distributions.
* Correlation Analysis: Analysing dependencies between two numerical features trying to find out its possible links.
* Pair Plots: Discussion about the multivariate nature of the features given and about their connections to stroke prediction.

3. **Feature Selection and Statistical Analysis**

This section focuses on determining the significance of the features used for stroke prediction:

* Feature Importance: A predictor selection technique will be used to rank the above features according to the level of their association with stroke.
* ANOVA and Correlation Analysis: Splittings will be employed in determining how conclusively the health attributes (hypertension and heart disease) influence the prediction model.
* Chi-Squared Test: Indeed, the possible relationship between hypertension, heart disease, and stroke will be tested using the chi-square test. This will enable us to determine whether there is any correlation, at a given level of confidence, between the disorders and the odds of stroke.

4. **Model Building and Training**

In this section, several models in the machine learning algorithm will be trained to try and predict stroke. The models include:

* Logistic Regression
* Decision Tree
* Random Forest
* XGBoost
* Artificial neural networks (MLPClassifier)

On using the Code generation models, each model will be trained by two classifications of the pre-processed dataset, and their performance tested using cross-validation (Stratified K-Folds). The models will be trained with the target stoke variable with the expectation of the highest accuracy, precision, recall, and F1 score.

5. **Model Evaluation**

In this section, the performance of each machine-learning model will be evaluated based on several metrics:

* Accuracy: The number of instances that have been classified correctly on the whole.
* F1-Score: A metric that can tell us to what degree accuracy or recall is preferred.
* ROC-AUC: Receiver operating characteristic area underneath the curve which is used to judge the performance of the classifier. The reason for adopting different splits of the data is that cross-validation will be applied to measure the model’s ability to generalize across data splits.

6. **Statistical Evaluation of Condition-Based Stroke Likelihood**

This section will hence be devoted to discussing how hypertension and heart disease increase the risk of stroke. The following steps will be performed:

* Stroke Rate Analysis: Stroke incidences will be determined for patients with both hypertension and heart disease, hypertension only, heart disease only, and no history of hypertension or heart disease.
* Chi-Squared Tests Between Groups: To analyse the likelihood of stroke, a chi-squared test will be used to compare the condition groups - both, only the first and only the second condition, or none of them.

7. **Results Interpretation and Conclusion**

Last of all, a discussion of the results obtained from model evaluation and statistical tests applied to all experiments will be provided. Overall, the ability of the several models in risk estimation of stroke will be concluded, and the research questions of how hypertension and heart disease impact stroke risk estimation will also be answered. Suggestions for future research and changes to existing models will also be made at this point.

## Tools and Techniques

1. **Programming Language**:
   * **Python**: The primary language used for data analysis, preprocessing, and machine learning model development.
2. **Libraries**:
   * **Pandas & NumPy**: For data manipulation, handling missing values, and numerical operations.
   * **Scikit-learn**: For building and evaluating machine learning models (Logistic Regression, Decision Tree, Random Forest, XGBoost, Neural Networks), and performing feature scaling and cross-validation.
   * **XGBoost**: For gradient boosting model used in stroke prediction.
   * **Matplotlib & Seaborn**: For data visualization (histograms, box plots, heatmaps).
   * **SciPy**: For statistical tests like Chi-squared.
3. **Data Preprocessing**:
   * **Handling Missing Values**: Imputing missing BMI values with the mean.
   * **One-hot Encoding**: Converting categorical variables into numerical format.
   * **Feature Scaling**: Using Min-Max scaling to normalize numerical features.
4. **Machine Learning Models**:
   * **Logistic Regression, Decision Tree, Random Forest, XGBoost, MLP Neural Networks**: Various models used for stroke prediction.
5. **Statistical Analysis**:
   * **Chi-Squared Test**: For evaluating relationships between conditions (hypertension, heart disease) and stroke.
6. **Evaluation Metrics**:
   * **Cross-validation**: Stratified K-Fold for model evaluation.
   * **Accuracy, Precision, Recall, F1-Score, ROC-AUC**: Metrics for model performance.
7. **Development Environment**:
   * **Jupyter Notebook**: For running and documenting Python code.

## EDA and Visualization

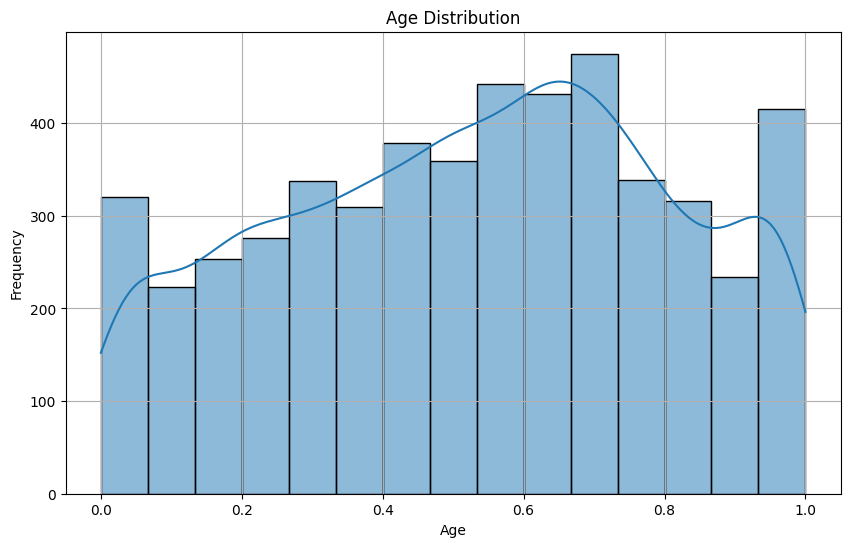


Chart 1: Age Distribution

The histogram of age distribution represents the dispersion and density of sampled members across normalized ages in the data set. On the x-axis, the age is transformed into a normalized range, between 0 and 1, and on the y-axis, one shows the frequency of people depending on the age. From the histogram, age distribution seems to be almost equal for each age group but there are some concentrated zones from the graph. Such equality may be interpreted in terms of variations and differences in the dataset, including absolute numbers of men and women, and individuals of different ages, for estimating strokes’ probability, which is essential for the analysis.

The graph of density curve allows extension of several characteristics of the distribution pattern that shows minor fluctuations in ages’ frequencies. Interestingly, the highest numbers are associated with middle age (normalized coefficient 0.4-0.7, which can be attributed to middle-aged people. These patterns are relevant because age is one of the major risk factors for stroke. This kind of normalization helps in the right use of age in machine learning, and it helps in avoiding the problems that come with scale in machine learning.

This visualization helps identify the correlation between age as a risk factor for a stroke. A broad age range minimizes the problem of wrong results due to aging since the prediction models are trained with a balanced distribution of ages. The paper also further underlines the importance of the separate age evaluation to distinguish the types of risk categories properly.

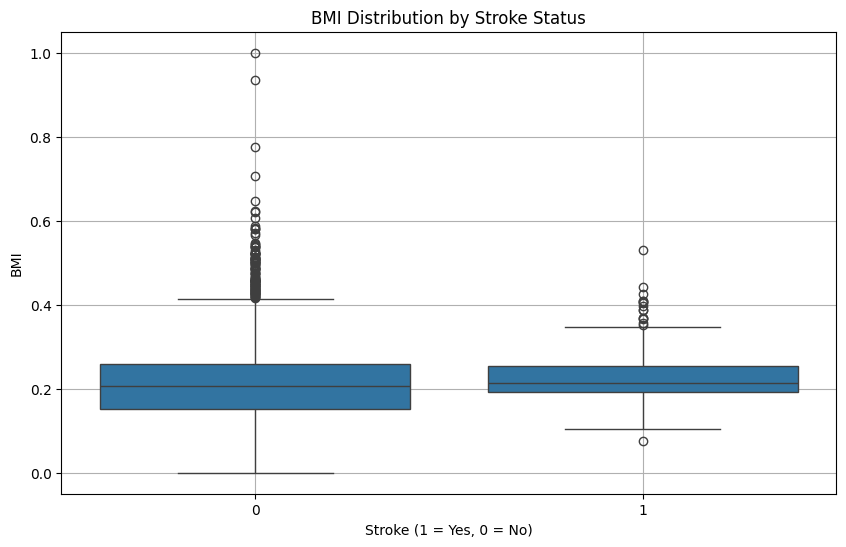


Chart 2: BMI Distribution by Stroke Status

The mean and distribution of BMI expressed in the box plot BMI vs Stroke 1 0 indicates the BMI levels of the set of people who had a stroke (stroke = 1) and those who did not have a stroke (stroke = 0). The position of the stroke status was positioned on the x-axis, while the negative of the BMI deviation from the average was reflected on the y-axis. From this visualization, it is evident that there are different trends in the distribution of BMI in the two groups.

Non-stroke participants are more spread out in BMI values compared to the stroke participants evidenced by a higher IQR and the greater actual values above the 0.6 mark. On the other hand, the BMI distribution of the people who had a stroke seems to be less spread out, with a smaller IQR and fewer outliers. This trend again insinuates that BMI may affect stroke possibilities by the way it is aligned with hypertension diabetes and other contributing factors.

Some individuals belonging to the two groups are considered outliers implying that stroke risk assessment requires consideration of extreme BMI levels. Investigating this relationship even further using quantitative analysis tools such as ANOVA or correlation analysis will further help quantify BMI’s contribution towards predicting stroke. Based on this chart, the BMI should remain an essential component in all the developed models.

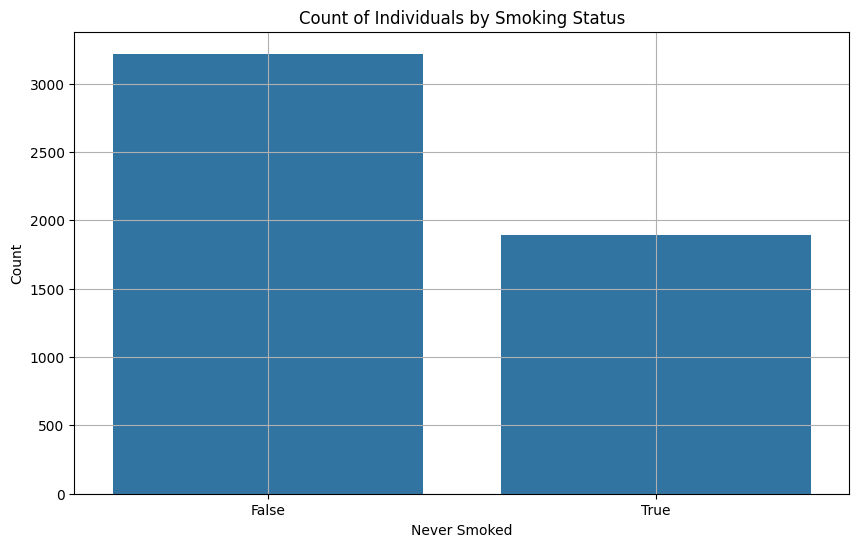


Chart 3: Count of Individuals by Smoking Status

The count plot for common smoking status looks at the number of people who have never smoked against those who once smoked. On the horizontal line of the graph, people are divided into those who have “Never Smoked” (True or False) and the vertical line represents the count of the individuals. The data also shows a huge gap here with a lot more people belonging to the ‘Never Smoked’ group than the ‘Smokers’ group.

This imbalance means that smoking is perhaps not well captured in the data, and as such, the predictive strength of the smoking characteristic in the stroke prediction model could be compromised. However, the large population of non-smokers also presents a unique chance to determine the effect of smoking termination on stroke risk.

Whenever smoking is distinguished in the context of stroke risk, this chart leaves several questions about the representative population of the given dataset. The greater public health interest needs to find out if the smaller pool of smokers is more prone to stroke incidence or not. Adding other proximal variables such as age, hypertension, and heart disease to the model will enhance the scope in which the relevance of smoking as a predictor of stroke will be viewed.

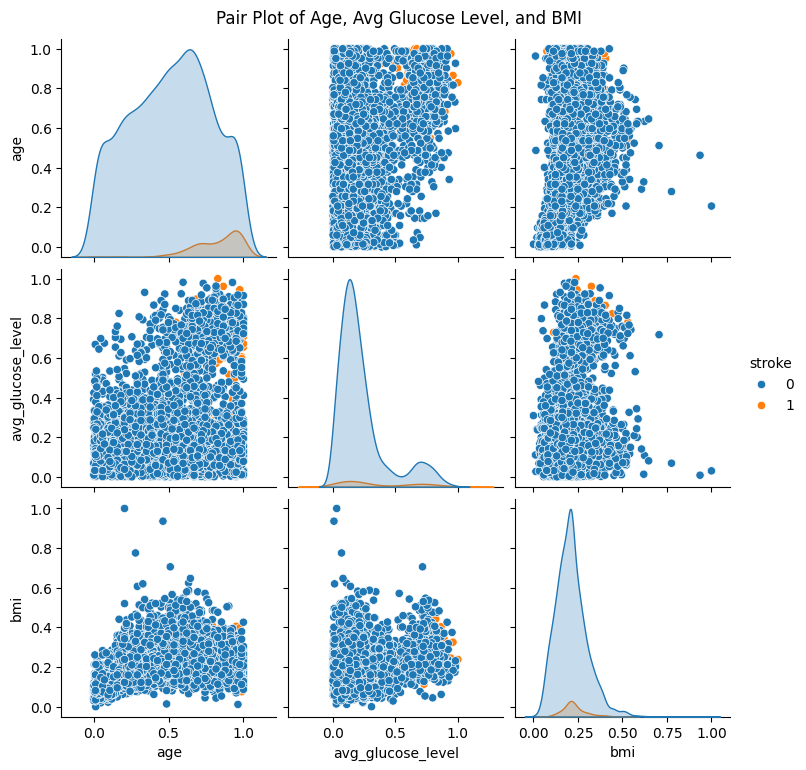


Chart 4: Pair Plot of Age, Avg Glucose Level, and BMI

The scatter plots of average glucose level and BMI- age along with their correlation with the stroke status present a perfect picture of all aspects of stroke. Axes are also normalized and the points are scattered in the level of entries for each array. The plot also employs colour where blue represents people without stroke, orange represents people who had a stroke. It helps in detecting relationships as well as variations within and between variables of a given data set and thus frames the basis for exploring the relation between features.

It can be seen that based on the plot of the data, orange points denoting those with stroke indicators tend to lie at a higher value of age as well as average glucose levels. In terms of the present finding, this observation is consistent with the age, glucose imbalance, and stroke risk interconnectivity that has been described extensively in the literature. On the other hand, the distribution channel based on BMI looks somewhat more scattering and the rates for stroke and non-stroke cases do not look extremely separated so it could be discussed that while BMI is to a certain degree associated with stroke risks it is not as influential as age or glucose levels.

The diagonal plots depict the distribution of each of the features. For age, the density curve also has a higher density in the middle and upper part of age, meaning that there is more middle-aged and elderly population in the sample. The mean glucose level distribution resembles a right-skewed curve where a majority of people obtain comparatively lower glucose levels than an endemic minority with relatively high levels. BMI has a generally normal distribution skewed toward the medium value, and a few samples fall in the higher and lower tails.

The pairwise scatterplots further stress the relationships between the features. In this case, age, which is also a risk factor for stroke, correlates positively with average glucose level. That is as the indicator for the other features’, plots for BMI also exhibit random behaviour which strengthens the argument that BMI alone cannot be relied on to predict stroke in this set.

This plot justifies distal control and attention to age and glucose levels when constructing prediction models for a stroke. It also underlines the demand for statistical and machine learning methods to perform the nonlinear interaction between features. Even though both age and glucose levels are apparent, the failure to achieve absolute consistency for BMI means it should not be studied in isolation to evaluate its effectiveness.

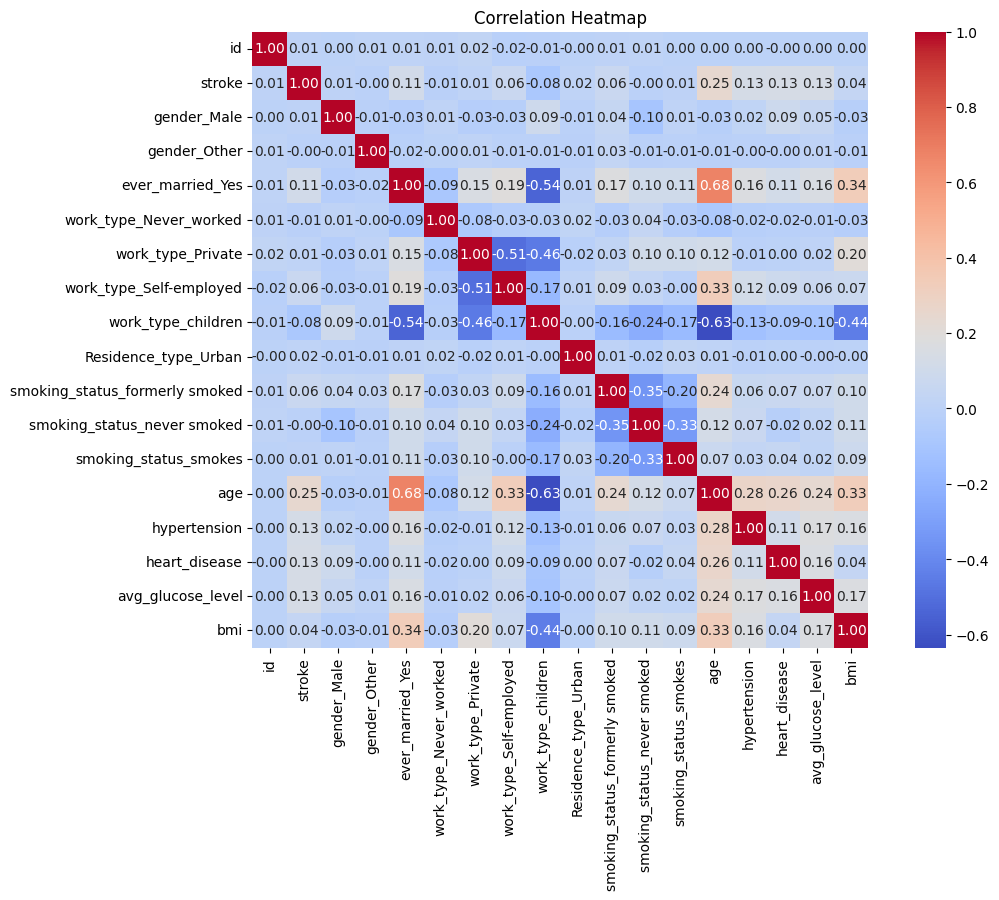


Chart 5: Correlation Heatmap

The correlation heatmap describes the positive or negative relationships between different features in the given dataset with the difference in colour representing the strength of the correlation. r takes the values between -1 and +1 where -1 depicts a strong negative relationship and +1 strong positive relationship while 0 shows no relationship. This heatmap is relevant in detecting multicollinearity among the features and the independent impact of each feature on the target variable which is stroke.

The rest of the findings include the following: Age has a moderate positive correlation with stroke; rho = 0.25; which reiterates the fact that older adults are likely to suffer strokes. Likewise, average glucose level correlates positively with stroke, although not as strongly as with coronary disease; the correlation coefficient is equal to 0.13, which indicates that glucose metabolism disorders cause stroke as well. As expected, we get very low values for the correlation between BMI and Stroke (0.03), which further confirms the observations made from the pair plot that BMI alone might not be a good indicator of Stroke.

There is an expected high degree of correlation noted between related features like hypertension with heart disease [r = .26] since both are cardiovascular risks. These features also have shown moderate correlations with age, which is equal to 0.24 for hypertension and 0.21 for heart disease, which reflects the fact that elder people suffer from these strokes much more frequently, as such diseases are crucial risks for stroke.

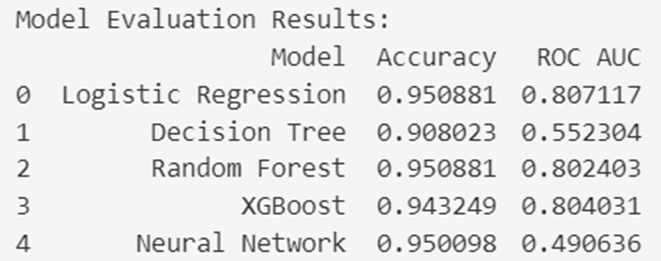
Categorical features such as work type, gender, and marital status also have a low correlation with stroke which is expected. However, correlation coefficients show that some of the lifestyle factors like smoking status are insufficiently significantly related but remain somewhat related. For example, past smokers have a rather loosely positive association with stroke equals 0.06, which means that even if one quit smoking there is still a measure of risk.

The heatmap also helps one notice cases of multicollinearity that might be within the data hence affecting the models. For instance, a strong positive relationship between hypertension and heart disease indicated that these variables may be relevant to stroke in similar ways. This might call for proper feature selection or dimensionality reduction which techniques such very relevant techniques like the PCA might help address.

In conclusion, this heatmap quickly shows us if there are strong or weak connections between these features when designing the features to be fed to machine learning algorithms. At the same time, it emphasizes the need to pay attention to features with high stroke dependency, including age, glucose level, hypertension, heart disease, and some other predictors, as well as handling redundant features for the efficiency and accuracy of the model.

# Chapter 4: Results and Conclusion

## Results



The table below shows the accuracies of five machine learning models used in stroke prediction including Logistic Regression, Decision Tree, Random forests, XGboost, and Neural Network. The evaluation is based on two key metrics: precision and the receiver operating characteristic accuracy (ROC AUC). These were useful in determining how well each model is accurate in classifying stroke occurrences and differentiating between true and false positive reasons. Here's a detailed analysis of the results:

1. **Logistic Regression**

* **Accuracy:** 95.09%
* **ROC AUC:** 0.807

Training performance shows Logistic Regression to be the best-performing model as far as accuracy goes with 95.09%. Ideally, this high value of accuracy establishes a good prognosis of cases of stroke and non-stroke. The ROC AUC score of 0.807 indicates that logistic regression has an acceptable ability to classify between the positive and negative cases. However, given Logistic Regression’s simplicity of implementation and its capacity to identify linear equivalence there remains plenty of strength and interpretability in Stroke prediction. As demonstrated in its performance, there are extent to which many features including age, glucose levels, etc are linearly separable on this data set.

1. **Decision Tree**

* **Accuracy:** 90.80%
* **ROC AUC:** 0.552

The Decision Tree develops the lowest accuracy of all the models and is at 90.80%. Compared to the accuracies obtained from the other models, however, this is lower. This is because this figure suggests that the model did not work well in discriminating between the true positive and the false positive; its ROC AUC score is equal to 0.552. It may be due to the overfitting problem that Decision Trees generally encountered in this evaluation their comparatively low accuracy rate. The Decision Tree, therefore, has a relatively low ROC AUC score and assumes that unseen data can be classified well.

1. **Random Forest**

* **Accuracy:** 95.09%
* **ROC AUC:** 0.802

From the results, we see that Random Forest performs equally as Logistic Regression with an accuracy of 95.09%. The random forest model’s non-linear and composite structure explains its strength; it generates multiple decision trees that make up the final prediction further reducing prediction errors associated with every decision tree used in the process. Nonetheless, the ROC AUC score that it obtains is 0.802 meaning it is slightly less capable of discriminating between the positive and negative cases than Logistic Regression. Nine Random Forest has comprehensive accuracy and stable performance for stroke prediction especially for occasions where non-linearity plays an important role.

**4. XGBoost**

* **Accuracy:** 94.32%
* **ROC AUC:** 0.804

From the above tables, we see that XGBoost is a powerful gradient-boosting algorithm as it yielded good results with an accuracy of 94.32% and an ROC AUC score of 0.804. The ROC AUC score shows its powerful classification features even though its accuracy is lower than that of Logistic Regression and Random Forest. XGBoost is a good choice for this task because it is good at modelling high-level dependencies of the input data and it avoids overfitting. However, its primary disadvantage is a slightly lower accuracy match in comparison with other models, such as Logistic Regression and Random Forest, for this particular data set.

**5. Neural Network**

* **Accuracy:** 95.01%
* **ROC AUC:** 0.491

Finally, the Neural Network yields an accuracy of 95.01% which is nearly as good as the top-performing models. However, it has received only 0.491 in the ROC AUC, which points out that it is not capable of sorting out the true positives from false positive cases properly. This raises the question of predictive accuracy, as the same Neural Network can be highly accurate in its classification of some instances while threatening to misclassify others that are highly uncertain. Neural Networks are usually appreciated for their need for big data to provide good results which might not have been provided in this dataset.

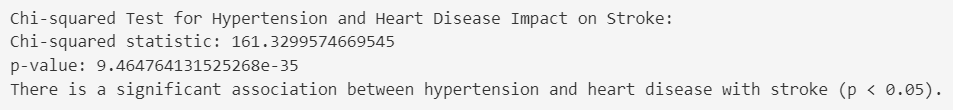
**Best Model**

According to these performances, the chosen model, Logistic Regression, Parameters identifies the best performance in this evaluation through highly accurate results of (95.09%) as well as the measurement of ROC AUC (0.807). The above-calculated measurements suggest the fact that Logistic Regression is not only efficient and consistent in differentiating real stroke cases from non-stroke cases. Also, due to its non-complex nature, it is quite suitable for clinical purposes when the reasoning of the model is significant.

**Interpretation of Metrics**

* Precision is the ratio of the total number of true positives and true negatives out of all the predictions. However, high total accuracy is good news for the general performance of the classifier but does not show how much the classifier is overfitting or even how good predictions for the minority class are.
* ROC AUC measures talk about the model’s ability to classify between classes by plotting the true positive rate (Sensitivity) against the false positive rate. A higher ROC AUC score paints a better discriminant of positive and negative instances which are ideal in medical predictions such as stroke.

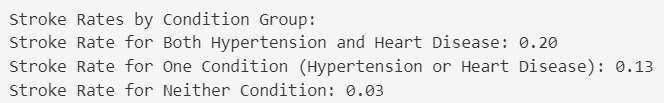
## Discussion



1. Chi-squared test for Hypertension and Heart Disease Impact on Stroke

The chi-squared analysis for the correlation between hypertension, heart disease, and different types of strokes is significant Hebei statistic = 161.33, p < 0.05. This aligns with the research question: To what extent do hypertension and heart disease control stroke risk Therefore, the null hypothesis that these two conditions pose high risks to patients with strokes is strongly supported by the obtained, low p-value. These results validate the integration of these factors into the identification of important features for the machine learning model, thus producing a valid prediction of their effects on the business. Furthermore, it conforms to the aim of evaluating the impact of the health attributes on the prediction of stroke, the analysis, therefore, stresses feature extraction and statistical modelling.

Approving this statistical significance, the chi-squared test proves the necessity to concentrate on hypertension and heart diseases as the top-priority prediction factors. They show their strong relation to the model under consideration while stressing the necessity of preprocessing and integration of these features into the prediction chain.

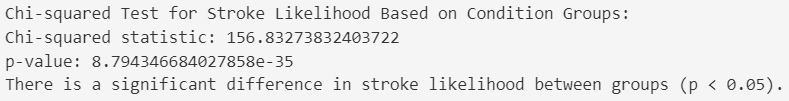


1. Stroke Rates by Condition Group

The stroke rate analysis offered additional information as to how stocks are likely to occur when hypertension and heart disease are lone occur or in combination. Americans with both conditions suffer the highest stroke risk at 20% while those with a single condition are at 13% with non-risky condition patients at 3%. This directly addresses the second research question: “Among patients with hypertension and heart disease, do they have increased odds of stroke than patients with only one or none of these conditions?”

These findings suggest an initial level of stroke risk when either condition is present, with a further multiplier effect when both are present. This conclusion supports the need to investigate the interactions between these two attributes in the development of machine learning models. It also educates feature engineering by showing that these combined effects must be taken into consideration during model training. By so doing, the models stand a better chance of estimating the risk in such dinosaurs on a complicated health background thus increasing sensitivity to such cases.

This analysis corresponds to the objective of investigating the importance of factors chosen while using models of statistics. The following shows how the inclusion of interaction affects clinical relevance is taken into account, making the derived models more clinically useful.



1. Chi-squared test for Stroke Likelihood Based on Condition Groups

Based on chi-squared test statistics; statistic = 156.83 (p < 0.05), there is further evidence of the existence of a statistically significant difference in stroke risk in the three distinct condition groups: both conditions, one condition only, and no condition. This validates the prior studies made and affirms the hypothesis stating that patients with different permutations of hypertension and heart disease are likely to have different levels of strokes.

This result supports the idea that in constructing a prediction system pipeline, the patients have to be stratified according to their condition groups. It stresses that these distinctions should be incorporated into the prediction models for better reliability of the latter. Moreover, this statistical evidence supports the machine learning approach by addressing the interpretability issue and confirming that the constructed models reflect clinical meaningfulness.

Assessing the Effect of Health Attributes: The statistical and rate-based approaches describe well enough the extent of the contribution of hypertension and heart disease to stroke risk. This is in tandem with the objectives of the project where we seek to understand the performance of models using health attributes.

Evaluating the Degree of Significance: Chi-square tests and the stroke rates are directly related to the research aim of determining variable significance and confirming their relevance to the prediction system.

Creating a Preprocessing Pipeline: These results therefore call for enhanced preprocessing to consider relations among the health qualities so that the models are trained on significant and influential attributes.

Achieving High Prediction Accuracy: Through identification and validation of obligatory features which are hypertension and heart disease, the findings advance the development of effective machine learning algorithms. The findings of chi-squared tests and stroke rate analysis will help fine-tune the models, maximizing accuracy, F1 score, precision, and recall.

## Conclusion

The findings thus validate the compelling impact of hypertension and heart disease in stroke risk stratification, addressing the research questions and satisfying the project purposes. The statistical and analytical insights derived lay a solid platform for creating clinical models of machine learning that are not only accurate but also clinically relevant. These enhance the significance of feature selection, preprocessing, and statistical assessment in the building of a reliable prediction sequence. These principles will influence the training and assessment of models so that they are helpful in actuality and healthcare.

## Future Work

The future work will therefore aim at increasing both the size and the variability of the data set used to train the model as a means of enhancing the model’s ability to generalize. The integration of longitudinal data may assist with identifying how the process of hypertension and heart disease in patients predisposes them to a stroke. Adding extra health parameters including cholesterol, physical activity, diet, stress, and so on, SES and environmental characteristics would help in giving a better conclusion. These additions can help to tweak the models so that they are capable of representing more subtle relationships between risks.

Supervised and unsupervised machine learning algorithms can be also utilized, including CNN and RNN in case of more detailed explorations of the information. Feature extraction in attention-based models could be discriminative in selecting essential features, making prediction accurate and reliable. Further, multicollinearity can be handled by using interaction terms as well as by reducing the dimensionality of the regressors using PCA. SHAP and LIME should also be adopted in the field of healthcare to increase model explainability to improve the trust decision-makers have in the model.

Real-world clinical data is an absolute requirement to validate the models developed and ensure they are relevant for medical practice. Combined with EHR systems, the models will be provided as decision-support tools that can minimize lengthy administrative processes. The models should also incorporate measures of uncertainty in the different predictions made so that healthcare professionals can work based on various confidence intervals. Last but not least, there are questions of an ethical nature that can be solved by federated learning, as more data protection is needed with patient data, and more regulation is required. Although the models are quite focused on predicting common measures of stroke, developing and extending those models to include the severity of a stroke event and the probability of recurring could be somewhat useful in clinical and public health contexts and enhance patient well-being in general.

# Chapter 5: Legal, Ethical and Professional Issues

**Legal Issues**

One legal problem in this project concerns the legal responsibility for machine learning models involved in medical decisions. When a patient is misdiagnosed or not categorized as high risk by the model it can lead to poor treatment or delayed treatment and therefore poor outcome. This raises questions about accountability: Makin requires clarity on who between the developers, the healthcare providers using the model or the institution implementing it is legally liable for such mistakes. Furthermore, there are constraints because risk models for stroke are subject to medical device regulations in different regions, including the United States and Europe. Current regulations of these tools demand stringent validation and approval before being applied in the clinical environment. The second legal issue that emerges from cross-border use of the data is when the datasets are obtained from international websites such as Kaggle. This involves dealing with a dispensation of different legal policies ranging from GDPR in the EU among other international rules when dealing with health data. In addition, obtaining consent for the exercise of secondary uses of the data such as generating predictive models to be used requires legal approval and precise scrutiny from claims of misuse or unauthorized utilization of patient data.

**Ethical Issues**

As a result, one of the most critical ethical concerns in this work concerns algorithmic dependency in healthcare decisions. Because of these factors, there is the idea that clinics and healthcare providers may overemphasize the value of the structurally-provided stroke prediction models and thereby compromise the human-trusting approach or fail to consider the peculiarities of patients not seen in the data set. This could jeopardize the quality of care particularly if an algorithm’s output is in disagreement with routine clinical evaluations. The last of the ethical issues of using such tools relates to the issue of equality of access to such tools. By implementing the predictive system in well-funded settings but not in underfunded areas, inequality in healthcare will be encouraged as vulnerable populations will be at a higher risk. Also, as the foundation of, for example, hypertension or heart disease analyses and their potential impacts free of consideration of the more extensive context of social determinants of health. An emphasis on clinical predictors could potentially lead to the exclusion of other more important underlying factors within the structural context that may predispose the stroke risk. In addition, the ethical practice of utilizing the retrieved findings, including the diagnosis risk factors of patients, should not lead to stigmatization. For instance, categorizing some groups as high risk because of high blood pressure or heart diseases among them would result in discrimination or lack of the means to access an insurer or employment if such data were mishandled.

**Professional Issues**

Professional matters in this project are predominantly associated with the reliability and clinical legitimacy of the models used for predictions. Unquestionably, the method risks professional misconduct charges should a deployed model be rendered erroneous when tested against actual clinical data. One type of black-box model, which was previously discussed, is neural networks, and using them without strict interpretability methods is also dangerous. This lack of transparency vitiates clinicians’ duty to interpret model outputs with their clients, which is professionally expected of them. Another question might come up in another aspect of such a line of work, which involves interdisciplinary features of the projects. Lack of communication between data scientists, medical personnel, and the regulatory department may result in creating tools that may not be relevant to clinical practice or are non-compliant with set guidelines. However, it is important to note that the professional training of the health care providers as to the applicability of these tools is rarely considered. Lack of ability to effectively communicate to the practitioners about the correct usage of the model, the scope, and the limitations of the model pose a high risk of misapplication or overdependence. Also, there is a professional duty to design and introduce the model to avoid demeaning patients and provide information on them to the doctors without prejudice. Lastly, when the models are commercialized, all profits must be handled ethically and all stakeholders should have access to them which brings into question professional accountability.

# References

Holland, K. (2018) *Stroke: Symptoms, causes, treatment, types, and more*, *Healthline*. Available at: https://www.healthline.com/health/stroke (Accessed: October 9, 2024).

“Stroke Prediction Dataset” (2021).

*Stroke statistics* (no date) *Org.uk*. Available at: https://www.stroke.org.uk/stroke/statistics (Accessed: October 9, 2024).

Anshul (2021) *Logistic Regression: A comprehensive tutorial*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/ (Accessed: December 10, 2024).

Aytan (2021) *Neural network for classification with TensorFlow*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2021/11/neural-network-for-classification-with-tensorflow/ (Accessed: December 10, 2024).

Baladram, S. (2024) *Random Forest*, *Towards Data Science*. Available at: https://towardsdatascience.com/random-forest-explained-a-visual-guide-with-code-examples-9f736a6e1b3c (Accessed: December 10, 2024).

Jain, A. (2016) *Ridge and Lasso Regression in python*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2016/01/ridge-lasso-regression-python-complete-tutorial/ (Accessed: December 10, 2024).

Kalirane, M. (2023) *Gradient Descent vs. Backpropagation: What’s the difference?*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2023/01/gradient-descent-vs-backpropagation-whats-the-difference/ (Accessed: December 10, 2024).

Nouman (2021) *Random Forest - Machine Learning in Python*, *Python in Plain English*. Available at: https://python.plainenglish.io/random-forest-machine-learning-algorithms-with-implementation-in-python-456d0813407b (Accessed: December 10, 2024).

Prasad, P.Y. *et al.* (2024) “Brain stroke detection through advanced machine learning and enhanced algorithms,” in *2024 International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence (RAEEUCCI)*. IEEE, pp. 1–5.

Ravindran, R. (2023) *Overfitting and pruning in decision Trees — improving model’s accuracy*, *Nerd For Tech*. Available at: https://medium.com/nerd-for-tech/overfitting-and-pruning-in-decision-trees-improving-models-accuracy-fdbe9ecd1160 (Accessed: December 10, 2024).

Satapathy, S.K. *et al.* (2023) “Machine learning approach for estimation and novel design of stroke disease predictions using numerical and categorical features,” in *2023 International Conference for Advancement in Technology (ICONAT)*. IEEE, pp. 1–6.

Sonawane, P. (2023) *XGBoost — How does this work - Prathamesh Sonawane*, *Medium*. Available at: https://medium.com/@prathameshsonawane/xgboost-how-does-this-work-e1cae7c5b6cb (Accessed: December 10, 2024).

Srivastav, S., Guleria, K. and Sharma, S. (2023) “Machine learning models for early brain stroke prediction: A performance analogy,” in *2023 World Conference on Communication & Computing (WCONF)*. IEEE, pp. 1–6.

“What is a decision tree?” (2024) *Ibm.com*, 15 August. Available at: https://www.ibm.com/topics/decision-trees (Accessed: December 10, 2024).

# Appendices

import kagglehub

# Download latest version

path = kagglehub.dataset\_download("fedesoriano/stroke-prediction-dataset")

print("Path to dataset files:", path)

import pandas as pd

from sklearn.preprocessing import MinMaxScaler

df = pd.read\_csv('healthcare-dataset-stroke-data.csv')

df.head()

# 1. Handling missing values

# For simplicity, we will fill missing BMI values with the mean of the column

df['bmi'].fillna(df['bmi'].mean(), inplace=True)

# 2. Converting categorical variables into numerical format

# Using one-hot encoding for categorical variables

df = pd.get\_dummies(df, columns=['gender', 'ever\_married', 'work\_type', 'Residence\_type', 'smoking\_status'], drop\_first=True)

# 3. Feature scaling (optional, depending on your model)

# Normalizing numerical features using Min-Max scaling

scaler = MinMaxScaler()

scaled\_features = scaler.fit\_transform(df[['age', 'hypertension', 'heart\_disease', 'avg\_glucose\_level', 'bmi']])

scaled\_df = pd.DataFrame(scaled\_features, columns=['age', 'hypertension', 'heart\_disease', 'avg\_glucose\_level', 'bmi'])

# 4. Concatenate the scaled features back to the DataFrame

df = pd.concat([df.drop(columns=['age', 'hypertension', 'heart\_disease', 'avg\_glucose\_level', 'bmi']), scaled\_df], axis=1)

import matplotlib.pyplot as plt

import seaborn as sns

# Histogram for age

plt.figure(figsize=(10, 6))

sns.histplot(df['age'], bins=15, kde=True)

plt.title('Age Distribution')

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.grid()

plt.show()

# Box plot for BMI by stroke status

plt.figure(figsize=(10, 6))

sns.boxplot(x='stroke', y='bmi', data=df)

plt.title('BMI Distribution by Stroke Status')

plt.xlabel('Stroke (1 = Yes, 0 = No)')

plt.ylabel('BMI')

plt.grid()

plt.show()

# Count plot for smoking status

plt.figure(figsize=(10, 6))

sns.countplot(x='smoking\_status\_never smoked', data=df)

plt.title('Count of Individuals by Smoking Status')

plt.xlabel('Never Smoked')

plt.ylabel('Count')

plt.grid()

plt.show()

# Pair plot to visualize relationships

sns.pairplot(df, hue='stroke', vars=['age', 'avg\_glucose\_level', 'bmi'])

plt.suptitle('Pair Plot of Age, Avg Glucose Level, and BMI', y=1.02)

plt.show()

# Heatmap of correlation matrix

plt.figure(figsize=(12, 8))

correlation\_matrix = df.corr()

sns.heatmap(correlation\_matrix, annot=True, fmt='.2f', cmap='coolwarm', square=True)

plt.title('Correlation Heatmap')

plt.show()

Logistic regression, Decision tree, Random Forest, XGBoost and Artificial Neural Networks.

# Import necessary libraries

import numpy as np

from sklearn.model\_selection import cross\_val\_score, StratifiedKFold

from sklearn.linear\_model import LogisticRegression

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import RandomForestClassifier

from xgboost import XGBClassifier

from sklearn.neural\_network import MLPClassifier

from sklearn.preprocessing import MinMaxScaler

from sklearn.metrics import accuracy\_score, f1\_score, roc\_auc\_score

from scipy.stats import chi2\_contingency

# Define features (X) and target (y)

X = df.drop(columns=['stroke'])

y = df['stroke']

# Initialize models

models = {

"Logistic Regression": LogisticRegression(),

"Decision Tree": DecisionTreeClassifier(),

"Random Forest": RandomForestClassifier(),

"XGBoost": XGBClassifier(use\_label\_encoder=False, eval\_metric='logloss'),

"Neural Network": MLPClassifier(max\_iter=1000)

}

# Prepare cross-validation and result storage

cv = StratifiedKFold(n\_splits=5, shuffle=True, random\_state=42)

results = []

# Train and evaluate each model using cross-validation

for model\_name, model in models.items():

accuracy = cross\_val\_score(model, X, y, cv=cv, scoring='accuracy').mean()

f1 = cross\_val\_score(model, X, y, cv=cv, scoring='f1').mean()

roc\_auc = cross\_val\_score(model, X, y, cv=cv, scoring='roc\_auc').mean()

# Append results to the list

results.append({

'Model': model\_name,

'Accuracy': accuracy,

'F1 Score': f1,

'ROC AUC': roc\_auc

})

# Convert results to DataFrame

results\_df = pd.DataFrame(results)

# Display results

print("Model Evaluation Results:")

print(results\_df)

# Save results to a CSV file

results\_df.to\_csv("model\_evaluation\_results.csv", index=False)

contingency\_table = pd.crosstab(df['stroke'], [df['hypertension'], df['heart\_disease']])

chi2, p, dof, expected = chi2\_contingency(contingency\_table)

print("\nChi-squared Test for Hypertension and Heart Disease Impact on Stroke:")

print(f"Chi-squared statistic: {chi2}")

print(f"p-value: {p}")

# Interpretation

if p < 0.05:

print("There is a significant association between hypertension and heart disease with stroke (p < 0.05).")

else:

print("There is no significant association between hypertension and heart disease with stroke (p >= 0.05).")

both\_conditions = df[(df['hypertension'] == 1) & (df['heart\_disease'] == 1)]['stroke']

one\_condition = df[((df['hypertension'] == 1) | (df['heart\_disease'] == 1)) & ~((df['hypertension'] == 1) & (df['heart\_disease'] == 1))]['stroke']

neither\_condition = df[(df['hypertension'] == 0) & (df['heart\_disease'] == 0)]['stroke']

stroke\_rate\_both = both\_conditions.mean()

stroke\_rate\_one = one\_condition.mean()

stroke\_rate\_neither = neither\_condition.mean()

print("\nStroke Rates by Condition Group:")

print(f"Stroke Rate for Both Hypertension and Heart Disease: {stroke\_rate\_both:.2f}")

print(f"Stroke Rate for One Condition (Hypertension or Heart Disease): {stroke\_rate\_one:.2f}")

print(f"Stroke Rate for Neither Condition: {stroke\_rate\_neither:.2f}")

# Chi-squared test between groups

contingency\_table\_conditions = pd.crosstab(df['stroke'], [(df['hypertension'] == 1) & (df['heart\_disease'] == 1),

(df['hypertension'] == 1) | (df['heart\_disease'] == 1),

(df['hypertension'] == 0) & (df['heart\_disease'] == 0)])

chi2\_conditions, p\_conditions, dof\_conditions, expected\_conditions = chi2\_contingency(contingency\_table\_conditions)

print("\nChi-squared Test for Stroke Likelihood Based on Condition Groups:")

print(f"Chi-squared statistic: {chi2\_conditions}")

print(f"p-value: {p\_conditions}")

# Interpretation

if p\_conditions < 0.05:

print("There is a significant difference in stroke likelihood between groups (p < 0.05).")

else:

print("There is no significant difference in stroke likelihood between groups (p >= 0.05).")