National College of Ireland

Comparative Modeling of Stroke Prediction Using Advanced Machine Learning Techniques

MSc Research Project
Data Analytics

Livia Anthony Pereira
Student ID: X22158294

School of Computing
National College of Ireland

Supervisor: Arjun Chikkankod

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Livia Anthony Pereira
Student ID:	X22158294
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Arjun Chikkankod
Submission Due Date:	14/12/2023
Project Title:	Comparative Modeling of Stroke Prediction Using Advanced Machine Learning Techniques
Word Count:	7500
Page Count:	30

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Livia Anthony Pereira
Date:	14th December 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Comparative Modeling of Stroke Prediction Using Advanced Machine Learning Techniques

Livia Anthony Pereira
X22158294

Abstract

A stroke transpires when there is a sudden cessation of blood flow to a specific region of the brain. This abrupt interruption leads to the gradual demise of brain cells, resulting in disability contingent on the affected brain area. Timely identification of symptoms serves as a crucial factor in providing valuable insights for stroke prediction and facilitating a health-conscious lifestyle. The current research employs machine learning (ML) methodologies to craft and assess multiple models, aiming to construct a resilient framework for the enduring prognostication of stroke occurrences. This research scrabbles into the critical dominion of stroke prediction, employing a comprehensive approach encompassing detailed data preprocessing, innovative feature engineering, and strategic model training. The implementation involves a systematic process, from initial data collection and preprocessing to the application of two distinct feature engineering approaches – correlation analysis and feature importance evaluation. The ultimate model training unfolds with the evaluation of four prominent algorithms: K Nearest Neighbor, Balanced Random Forest, Catboost, and XGBoost. Notably, XGBoost emerges as the most surpassing algorithm, attaining an outstanding accuracy of 98.11%, F1-Score of 0.9811, and Roc_Auc score of 0.9977, showcasing its unparalleled efficacy in stroke prediction. The research underscores the importance of diverse feature engineering approaches influencing stroke risk as Feature selection using one-vs-all feature selection surpasses the feature selection by correlation analysis. The findings contribute to the burgeoning field of healthcare analytics and signify a step towards more accurate and targeted stroke prevention strategies.

1 Introduction

Stroke has been identified as a critical health condition in which blood vessels are ruptured in the central nervous system (CNS), i.e. brain, thus causing a damaging impact. As stated by the World Health Organization, stroke is identified to be one of the potential causes of increased mortality rates and disability (paralyzed) worldwide Tazin et al. (2021). With an understanding of the increased prevalence of death due to cerebral stroke, studies have provided necessary information on the mechanism action of the disease. As explained by Tazin et al. (2021), stroke occurs due to irregular blood flow throughout the central nervous system (brain) due to disruption and damaging impact of the vessel. It is because of this condition, that brain cells present in the disrupted part do not receive adequate nutrients or oxygen, thus leading to cell death (apoptosis). Since

cerebral attack is a significant health concern, the WHO, medical professionals, and research experts have been investigating subsequent measures to detect and appropriately manage stroke in the early stage of its impact. According to the statistical estimation of the WHO, nearly 15 million individuals globally are affected by cerebral strokes every year, while 1 out of 4 people die within 4-5 minutes of the attack Tazin et al. (2021). On the other hand, "The Centre for Disease Control and Prevention" (CDC) has informed that cerebral stroke is one of the leading death prevalent conditions in the USA, which although non-communicable, can affect a maximum number of people Tazin et al. (2021). Thus, it has become integral to investigate the matter with high alertness on disease prediction and control.

Numerous research studies have been developed over the years to forecast medical data on cerebral stroke by using different methods. Evidence-based understanding has represented that text mining (TM) and machine learning (ML) are two standard approaches in classifying cerebral health problems by utilizing extensive datasets Tazin et al. (2021). There are various well-recognized ML methods have been trained and tested for stroke prediction including "artificial neural networks" (ANN). As per the evidence in the literature, "stochastic gradient descent" (SGD) is one of the suitable algorithms, which provides a significant value estimated to be 95%. In recent times, stroke has been predicted to be a "global threat" and an epidemiological condition, which is associated with premature death as well as increased financial burden. Thus, researchers and medical professionals have been exploring efficient diagnosis and treatment regimes by predicting the condition at an early stage of its occurrence. Artificial intelligence and machine learning have been appropriate tools in this exploration for various reasons. Evidence-based understanding suggests that machine learning can predict the problem by coping with numerous class imbalance dataset problems. One of the significant reasons for these imbalances is that the cerebral patients' class is considerably smaller compared to the healthy class. Apart from using class imbalance datasets for stroke prediction, machine learning can interpret the overall outcome to acknowledge the decision-making. Thus, it is imperative to explain that repetitive research studies have prioritized machine learning over other methods.

As already discussed, stroke is a significant reason of death and disability in people globally. It affects individuals regardless of their age and gender although individuals above 40 years are at a higher risk Kokkotis et al. (2022). The impact of this disorder typically reduces the "quality of life" and increases the burden on the healthcare system. As implored in research studies, a high prevalence of the impact of cerebral stroke on society has forced the "scientific community" to emphasize the exploration of different stroke prediction models at an early stage. Therefore, a contemporary approach has been marked with the assistance of artificial intelligence, which ensures a reliable approach to preventing several diseases. As per the recent review of the literature, many studies have prioritized machine models to diagnose stroke and predict treatment responses with assured patient outcomes. In this approach, as implicated by Kokkotis et al. (2022), the main objective is to form a personalized approach toward rehabilitation protocols. The "data mining" approach introduced by experts for the prediction of ischemic stroke typically includes optimized data from 80 subjects and 112 normal or healthy subjects. It is a remarkable process that has considered extensive features from the class imbalance data and enabled the identification of risk factors related to ischemic stroke. The

contemporary analysis typically considers class imbalance datasets for proposed models to detect cerebral stroke. Although the data is imbalanced due to insignificant classes compared to healthy datasets, the singular or hybrid machine learning models are used for the prediction of cerebral stroke using this data. Moreover, the prediction scores of these models are compared among each other based on various parameters such as accuracy, false positive-negative alarm rate, and area under concentration (AUC). The overall approach in stroke prediction is based on a computerized approach, where algorithms play a vital role and have been prioritized in recent years to reduce the risk prevalence of this disorder and enhance the quality of life as well as survival rate.

1.1 Research Objectives

As we navigate the complexities of predictive analytics in healthcare, our investigation is guided by a series of targeted research objectives. These goals are crafted to dissect the efficacy of machine learning models in the context of stroke prediction. The research objectives are as follows:

- To Evaluate the Predictive Performance of Different Machine Learning Models on Cerebral Stroke Prediction Dataset.
- To Determine the Impact of Feature Selection Methods on Model Performance.
- To Identify the Most Optimal Model for Stroke Prediction.

1.2 Research Question

To advance the accurate prediction of Cerebral Stroke through machine learning approaches, our research is centered around addressing the following pivotal research question.

- What is the most effective machine learning model for predicting stroke occurrences when comparing kNN, balanced random forest, CatBoost, and XGBoost, and how do different feature selection methods, such as correlation analysis and Random Forest feature importance ranking, influence the predictive accuracy of these models?

2 Related Work

The second chapter has specified a comparative evaluation of different methods in the prediction of cerebral stroke based on class imbalance datasets. The medical dataset used for the prediction has been studied based on the literature review and critically explored the reliability of each method in the prediction process. With this approach, further exploration of the convenience of machine learning models is also recognized by exploring the facts and information provided by existing researchers. Since this study is based on stroke prediction, the applicability of algorithms in the medical domain is significantly associated with ethical obligations and challenges. Therefore, this study has predisposed knowledge on these specific matters and further delineates the scope for future research.

2.1 Empirical Reviews on Stroke Prediction

The condition of stroke, which is also popularized as the "cerebrovascular accident" has become a potential area of research in recent times since there is an increased prevalence of cerebral attacks worldwide which impacted people's living conditions and quality of life. A comprehensive approach therefore is marked with an automated detection process through forecasting of the extracted medical data. As explained by Tiwari (2021), the medical dataset used in the prediction process contains 12 features, which also include highly imbalanced target columns. Ischemic stroke is a blood-deprived condition of brain cells, which therefore leads to cell damage and death. According to the information provided by Liu et al. (2019), medical datasets typically contain patient symptoms as well as health conditions. In recent research advances, these datasets have been prioritized to extract useful features to train and test classification and detection models, which serve the potential prognostic and diagnostic measures. While understanding the relevance of this medical dataset, many studies have provided information on the accuracy, effectiveness, and efficiency of machine learning models in clinical assessments for stroke prediction. Therefore, this chapter has reviewed the potential of machine learning and other models to put forward a comparison of the reliability rate using class imbalance datasets.

2.2 Machine Learning Algorithms in Cerebral Stroke Prediction

The application of machine learning models is widespread in contemporary research regarding classification, prediction, and automation processes. The data feasibility to extract features for the classification and prediction process is a significant step to ensure that the model is efficiently working and provides a comprehensive score based on different parameters. The contribution of research knowledge by Kokkotis et al. (2022) explains that utilizing combined clinical data, "trans-thoracic echocardiography" and "Computed Tomography Angiography" (CTA Imaging, various machine learning models have achieved a higher accuracy rate and AUC score of 96.4%. Moreover, the review of information by Uttam (2022) has stated that 4 machine learning (ML) classifiers among which "random forest" (RF) has achieved a higher accuracy, which is estimated to be 96% compared to other classifiers. Uttam (2022) in their study further specified that using under-sampling and over-sampling strategies, contemporary approaches to stroke prediction have become more convenient since these strategies have improved the data balancing mechanism. The above-specified study has provided extensive knowledge through the prediction process where the proposed ML model has used balanced datasets and achieved a performance accuracy of 98%. As per the experimental observation, it has been further revealed that the over-sampling strategy has improved the class imbalanced condition of datasets, thus yielding a better result by 46%.

Understandably, class-imbalanced medical datasets have provided extensive features to enhance the prediction process by utilizing the same to train machine learning models and forecast stroke risk factors to promote early clinical diagnosis. As explained by Wang et al. (2021), early prediction of strokes and increased risk factors like heart attack has become a promising approach in recent times, where data analysis has been integrated to reduce the mortality rate. In this regard, evidence shows that the processing of imbalanced data has been enhanced with the leverage of random sampling (under and over)

and clustering methods, which are specified as "undersampling-clustering-oversampling algorithms" (UCO) Wang et al. (2021). As further explained by Wang et al. (2021), the above algorithm has generated a significant balanced dataset to be used by machine learning classifiers to extract features. In the above-specified study, a total of 5 ML classifiers were used to precisely predict cerebral stroke and heart attack. As per the experimental outcomes, one of the renowned models, the random forest has achieved better performance with an estimated accuracy level of nearly 70.3% and a precision rate of 70% respectively. Thus, it can be stated that random forest along with the UCO(120) algorithm has performed a suitable prediction of stroke occurrence among patients and the severity with increased chances of heart attack.

The prediction process of cerebral stroke has combined text mining techniques and machine-learning models which portrays a magnificent approach to tracking areas in a medical domain such as health surveillance, data management, and medicine based on which features can be extracted to train the model to execute the classification and detection process. According to Govindarajan et al. (2020), the data mining approach critically applied in the area provides a significant review of tracked information which can develop semantic and syntactic insights. According to the review of the study by Govindarajan et al. (2020), it has been determined that the processed medical data were used or fed in various ML models such as ANN, SVM, boosting & bagging, and RF. The experimental result obtained from the study has retained a specific configuration of the outcome which stated that ANN trained with SGD algorithm has outperformed other models with an accuracy level estimated as 95% with SD 14.69. Previously, it was discussed that the "stochastic gradient descent" (SGD) algorithm provides a more significant outcome in stroke prediction than singular machine learning algorithms Liu et al. (2019). Thus, it is imperative to consider that experts have integrated innovative applications of algorithms to distinctly predict the outcome related to stroke occurrence. Another study developed by Phankokkruad and Wacharawichanant (2022) has presented information on four tested and trained machine learning models, - XGBoost, RF, AdaBoost, and K-Nearest Neighbour (kNN) based on their accuracy as well as performance convenience. In this approach, some risk factors are considered from the class imbalance medical dataset of patients, which include age, blood pressure (hypertension), cardiovascular disease, BMI rate, and glucose level. According to Phankokkruad and Wacharawichanant (2022), these factors are determined to be critical attributes in the prediction of cerebral strokes. Aligned further information on the study outcome, it has been identified that each of the classifiers has an AUC score of 0.85 (Random Forest), 0.86 (XGBoost), 0.67 (AdaBoost), and 0.85 (KNN) respectively. These show that each ML model has shown equal confidence values although XGBoost shows a significantly higher performance compared to other models in stroke prediction.

Over the years, research has progressed significantly in the field of predicting cerebral strokes. According to statistical estimation, stroke is one of the leading mortality aspects throughout the world and disrupts normal living conditions as well as the quality of life of individuals Biswas et al. (2022). In the study conducted by Biswas et al. (2022), the author has marked an approach to the early prediction of cerebral stroke to prevent the damaging and unprecedented impact on affected patients. In the above study, approximately 11 machine learning classifiers are selected and trained with preprocessed and processed imbalanced datasets. A "random over-sampling" (ROS) strategy is applied for

accurate prediction based on slightly balanced data Biswas et al. (2022). As the evaluation of the result nearly 10 classifiers have shown an accuracy rate above 90%prior with imbalanced data and 4 classifiers have shown an accuracy level of 96% and above with balanced datasets Biswas et al. (2022). This information implies that the performance level of these classifiers varies significantly based on the class of datasets. Moreover, the "hyperparameter tuning and cross-validation", which have been performed on the model, have further enhanced the experimental result. It has been observed that the "Support Vector Machine" (SVM) shows the highest accuracy with an estimation of 99.9%, recall rate estimated to be 99.9%, precision rate estimated to be 99.9%, and F1-score of 99.9% respectively. Thus, the review of this study has implied that SVM is significant in providing high-performance accuracy in stroke prediction with imbalanced, balanced, and fine-tuned datasets.

2.3 Comparison of Models in Significant Stroke Prediction

In the previous section, a review of studies has determined the performance accuracy of various machine learning models in cerebral stroke prediction. According to the information, it can be stated that research has empowered a distinguished path to successfully retain the effectiveness and efficiency of ML classifiers in the prediction process. However, as the result from the study by Biswas et al. (2022) implied, some model has provided lower performance accuracy with imbalanced datasets. This can be a challenging fact since medical data is mostly classified as class imbalanced. Nevertheless, machine learning models have gained significant attention as well as interest in cerebral stroke prediction at an early stage, which thereby aided in advancing clinical diagnosis and prognosis measures. According to the study review presented by Singh and Choudhary (2017), it has been identified that different methods are compared based on the "Cardiovascular Health Study" (CHS) datasets. In this approach, an ML model - "decision trees" (DT) is used to extract features, "principle component analysis" (PCA) is used to reduce the dimension and back propagated neural network model for the construction of a predicted classifier Singh and Choudhary (2017). As per the understanding of the optimum result obtained from the prediction, it can be stated that the classification model has obtained a higher accuracy of 97.7% in the disease detection process.

Over the years, extensive research has been performed to develop suitable prediction techniques that can provide reliable outcomes for early intervention to the clinical diagnosis of cerebral strokes. In this approach, automated or computerized approaches have provided clinical relevance in identifying risk factors and utilizing medical data to forecast the stroke prediction process. According to the information presented by Paliwal et al. (2023), a contemporary experimental determination of the relevance of machine learning models is compared to identify the clinically relevant one with higher performance accuracy. In this regard, Paliwal et al. (2023) have selected algorithms such as KNN, SVM, DT, LR, RF and GNB, BNB, and VC. A specific emphasis has been given to the over-sampling strategy which typically shows the effectiveness of the models using imbalanced datasets based on different parameters. Depending on the experimental outcomes showing the comparison level among the models, it has been implicated that k-Nearest Neighbour (KNN) shows a more accurate performance than other classifiers and achieves a comparative ROC score (0.97). Based on this information, it can be stated that KNN in contemporary stroke prediction provides a reliable outcome with class imbalance datasets.

Another study highlighted by Sailasya and Kumari (2021) has further revealed a comparative evaluation of different machine learning classifiers, which indicates the effectiveness and efficiency in stroke prediction. With significant concern regarding increased ischemic stroke occurrence worldwide, health administrators and experts are showing interest in different classification models among which specific ML classifiers have gained unique considerations. According to Sailasya and Kumari (2021), the most specified algorithms in the literature are kNN, SVM, DT, RF, NB, and LR. Based on suggestive experimental outcomes, among all the algorithms, Naive Baiyes (NB) provides an accuracy of 82%, which is outlined as the highest performance accuracy.

Depending on the above information, it is integral to state that a significant comparison between the singular approach to machine learning models in stroke prediction is achieved in the literature. However, evidence on other methods including hybrid ML classifiers also achieved to some extent regarding their reliability in the prediction process. As stated by Santos et al. (2022), ML models in most instances promote an early prediction and effective cost approach to clinical diagnosis for stroke sequels. However, there is a challenging impact always prevails with these models when trained with imbalanced datasets. Therefore, Santos et al. (2022) have introduced an ensemble ML model based on "decision trees" (DTs) and "artificial immune system" (AIS), which were induced through "Genetic Programming" (GP). This approach has increased interpretability even using highly imbalanced datasets. As per the resultant observation, it can be stated that a higher sensitivity (70%) and specificity (78%) were achieved thus, showing a competitive benefit with comparable results of this model with existing ML models. Another study presented by Lyashevska et al. (2021) has introduced an enhanced algorithm, the "Gradient Boosting" (GB) algorithm which shows greater confidence and relevance with highly imbalanced datasets. Since the prediction accuracy is impacted by the class imbalance condition of the dataset, contemporary research has revealed the necessity to introduce improved models that can be fitted with the processing of this data without hampering the performance.

Understandably, machine learning has provided higher relevance in the classification and prediction of diseases and other aspects of the healthcare system. The model is more promising to achieve reliable clinical outcomes for further diagnosis and medical management than human interpretation. However, evidence provided a concerning fact regarding this model to be used in the medical domain Michelson et al. (2022). Despite the potential, the model has accurate medical management; biasedness in datasets with continuous replication of the feature has induced a "question of trust". Moreover, lack of trust and mistrust of the technology is a potential ethical challenge, which has been raised in recent years, thus, influencing patients', families', doctors', and healthcare administrators' decisions to rely on machine learning models.

The overall assessment of the review has predisposed comprehensive knowledge of the effectiveness of machine learning models in stroke prediction. It has been understood that most of the ML models are reliable and can provide high-performance accuracies. However, model incompetency most of the time in handling imbalance datasets (common medical data) and ethical obligations have directed a future research scope to explore integrated models such as ensemble ML models or other hybrid approaches that can produce reliable outcomes in stroke prediction with class imbalance datasets.

3 Methodology

According to the World Health Organization (WHO), stroke stands as a predominant cause of global mortality and disability. While extensive research has focused on predicting heart strokes, limited attention has been devoted to forecasting the risk of brain strokes. Recognizing stroke as a pernicious ailment predominantly affecting individuals, the prediction of stroke poses a labor-intensive and time-consuming task for medical practitioners. Thus, the primary objective of this work lies in forecasting the probability of stroke occurrence through the utilization of cutting-edge machine learning methodologies. A pre-structured process is enabled in this research to begin with the data collection, extensive and detailed data preprocessing followed by the data analysis and its visualization. Further two-pronged feature engineering is executed followed by model training and evaluation using various metrics on each feature engineering prong. A detailed description of each step performed in each process is provided in subsequent subsections. The Methodology flow diagram of our research is shown in Figure 1.

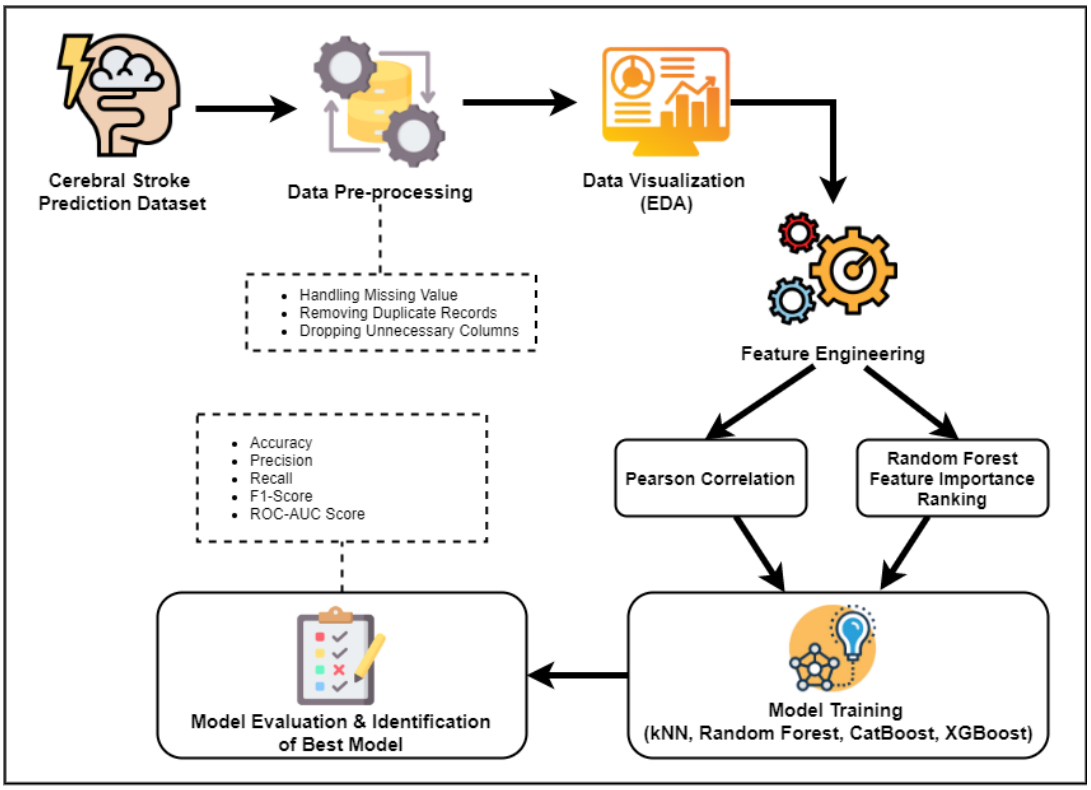

Figure 1: Methodology for Stroke Prediction Using Machine Learning Methods

3.1 Dataset Description

In this study, the dataset for stroke prediction was sourced from Kaggle *Cerebral Stroke Prediction-Imbalanced Dataset — kaggle.com* (n.d.). The specific dataset utilized for this research is centered around stroke detection. Upon obtaining the dataset, the initial step involved a comprehensive exploration to gain insights into its structure and characteristics. Visualizations of both the top and bottom rows were performed, offering a preliminary understanding of the dataset's composition. The dataset comprises 43,400

instances with 12 distinct columns, including features such as gender, age, hypertension, heart disease, marital status, work type, residence type, average glucose level, body mass index (BMI), smoking status, and the occurrence of a stroke. The dimensions of the dataset, as represented by its shape (43400, 12), illustrate its substantial size, totaling 26 megabytes. The chosen dataset not only provides a comprehensive set of attributes but also aligns with the research objective of stroke prediction using machine learning algorithms.

3.2 Data Preprocessing

Data preprocessing serves as a crucial phase in the research process, ensuring the dataset is refined and conducive to accurate model development for stroke prediction using machine learning algorithms. The initial step involved the removal of the 'id' column, deemed unnecessary as it merely represented the index. This process was executed to streamline the dataset, resulting in a reduced dimensionality with a shape of (43400, 11). To gain deeper insights into the dataset, statistical summaries were generated using the 'describe()' function, offering a comprehensive overview of the data distribution. Additionally, the 'info()' function was employed to elucidate the data types of each column, fostering a better understanding of the attribute characteristics. Addressing data quality, duplicate records were identified and subsequently eliminated, ensuring the integrity and uniqueness of the dataset. The examination of null values revealed that the 'BMI' and 'smoking_status' columns exhibited varying degrees of missing data as depicted in Figure **??**. A strategic imputation process ensued, leveraging the SimpleImputer with a 'most_frequent' strategy for categorical variables, specifically the 'smoking_status' column. The 'BMI' column, a numerical variable, underwent imputation as well, ensuring a comprehensive and reliable dataset for subsequent analyses. To provide a visual representation of the null percentages before and after imputation, plots were generated as shown in Figure **??**, highlighting the efficacy of the imputation strategy in preserving data completeness.

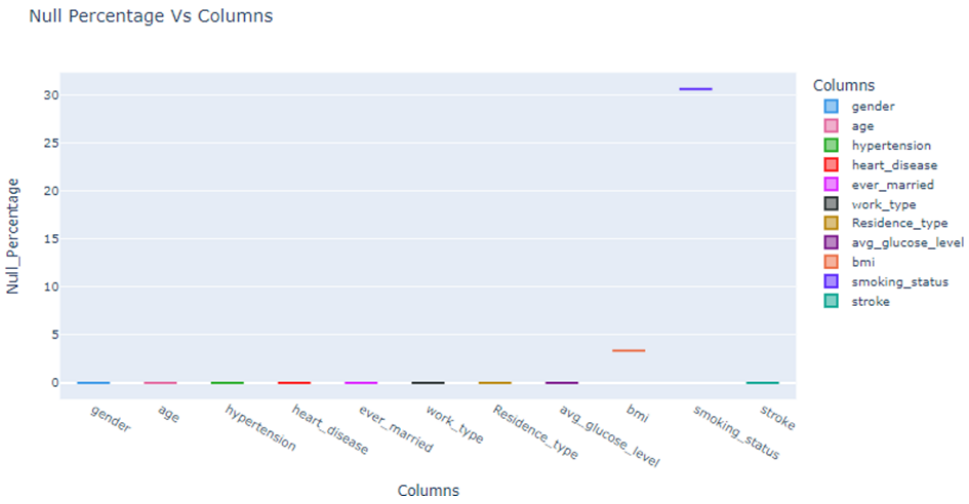

Figure 2: Null Percentage in Each Column

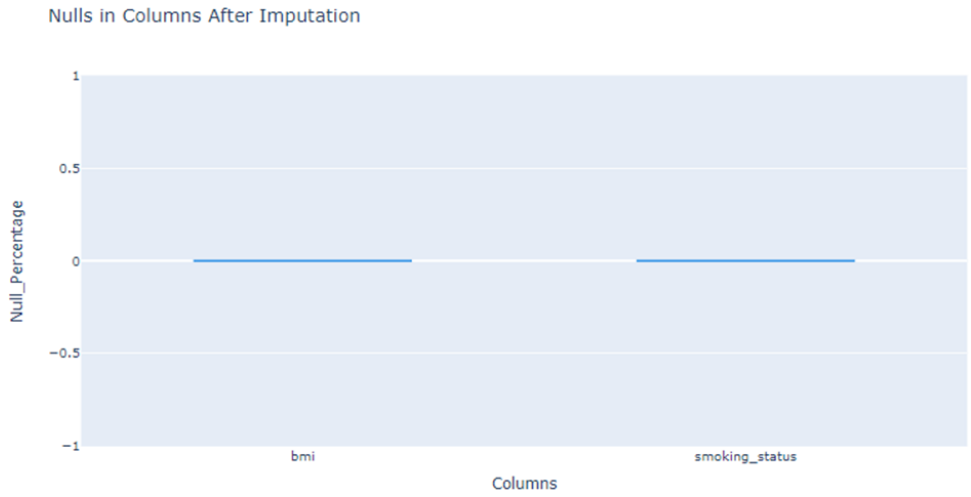

Figure 3: No Null percentage After Simple Imputation

3.3 Exploratory Data Analysis

Exploratory data analysis plays a pivotal role in unraveling patterns and insights crucial for the stroke prediction model. Beginning with the 'age' variable, a box plot was employed to understand the distribution of the stroke occurrence as shown in Figure 4. Outliers were identified and subsequently removed, revealing a refined representation of age's impact on stroke likelihood displayed in Figure 5.

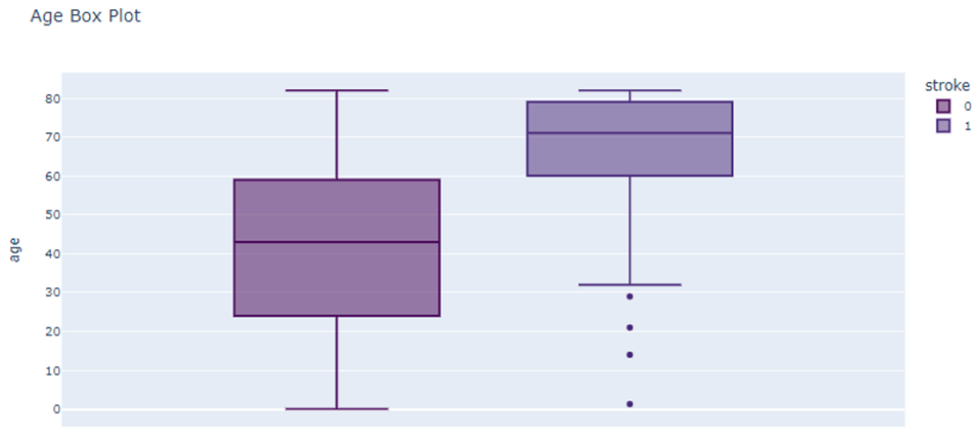

Figure 4: Age Box Plot with Outliers

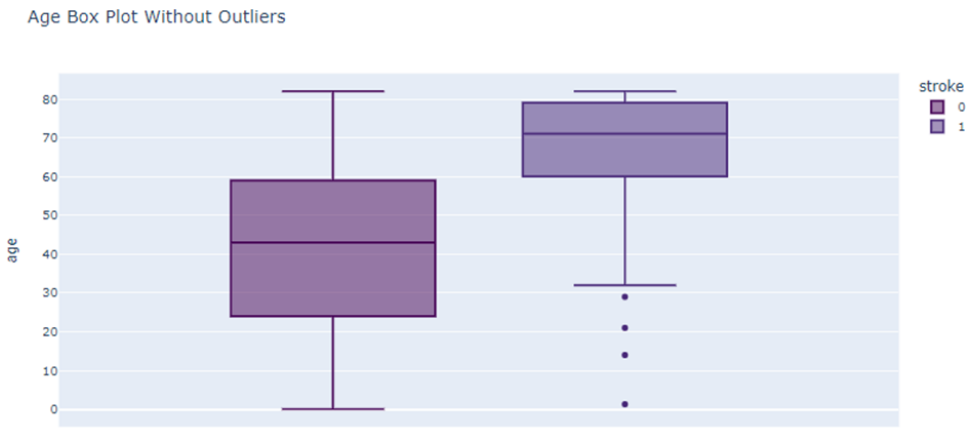

Figure 5: Age Box Plot with no Outliers

Moving to the 'BMI' variable, a similar approach was adopted, employing box plots to visualize the distribution and impact on stroke occurrence as depicted in Figure 6. By addressing potential outliers through a Quantile removal process, the subsequent box plot presented a more focused perspective on the relationship between BMI and the likelihood of stroke as shown in Figure 7.

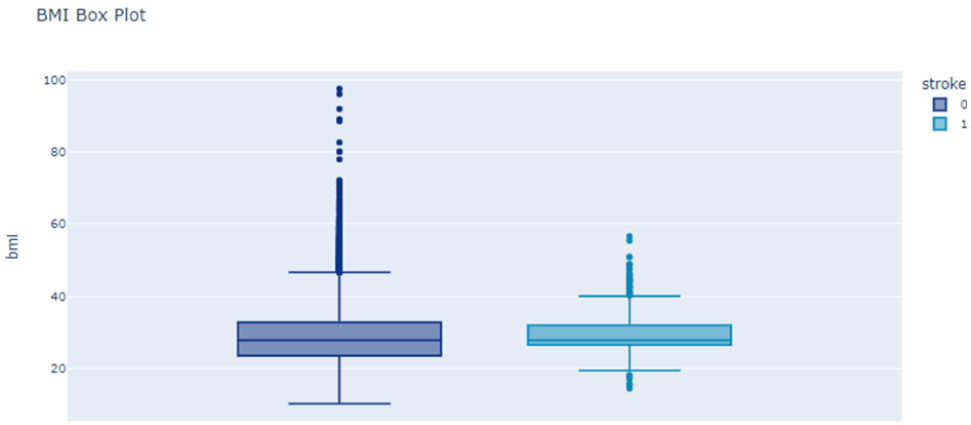

Figure 6: BMI Box Plot with Outliers

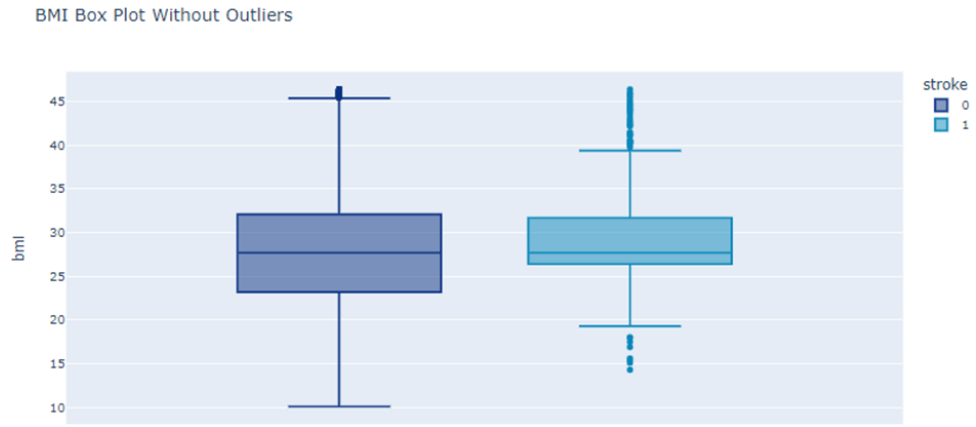

Figure 7: BMI Box Plot with no Outliers

The examination extended to the 'avg_glucose_level' variable, where box plots were employed to illustrate its distribution concerning stroke status. Outliers were methodically identified and removed as depicted in Figure 8 and Figure 9 respectively.

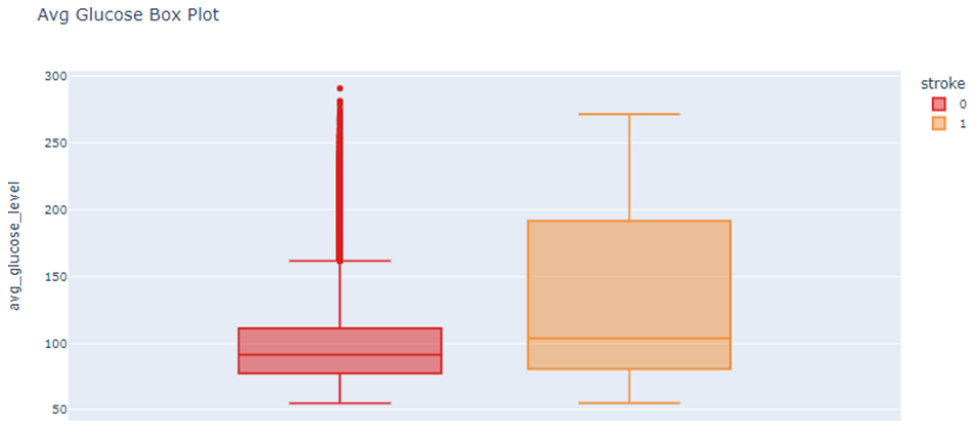

Figure 8: Avg Glucose Box Plot

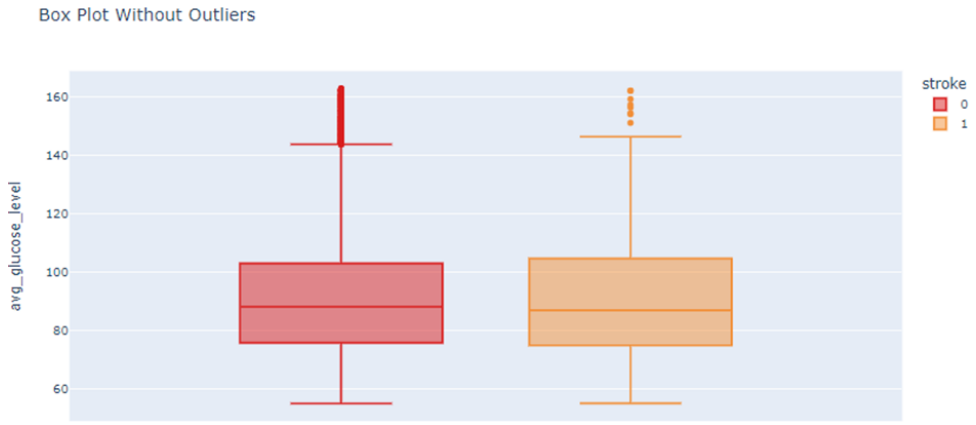

Figure 9: Avg Glucose Box Plot with no Outliers

The gender distribution within the dataset was explored through a pie chart, offering insights into the relative proportions of male and female as 40.4% and 59.6% respectively subjects displayed in Figure 10.

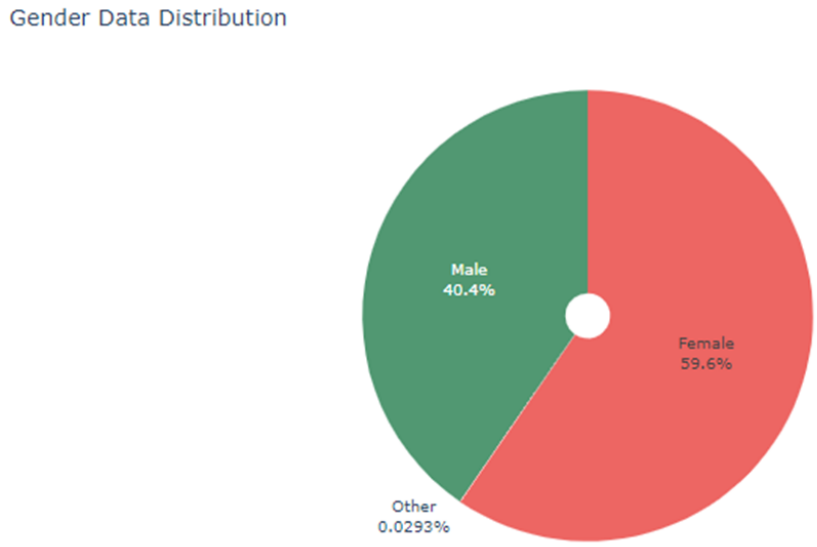

Figure 10: Distribution of Gender

Further, a bar plot as depicted in Figure 11 representing gender with stroke occurrence provided a comprehensive overview of the gender-specific impact on stroke likelihood, contributing to a nuanced understanding of potential gender-related patterns.

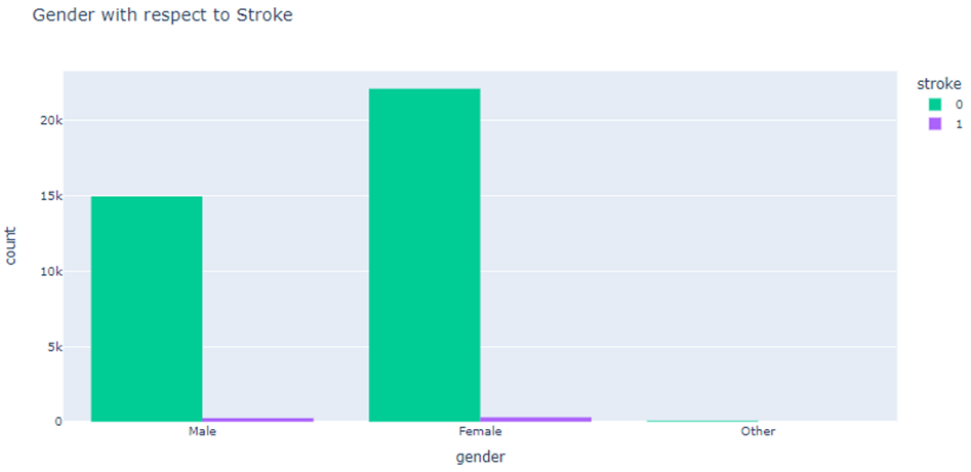

Figure 11: Distribution of Gender vs Stroke

Delving into the 'smoking_status' variable, a detailed analysis was conducted, illustrating the distribution of smoking statuses among individuals with and without strokes as depicted in Figure 12.

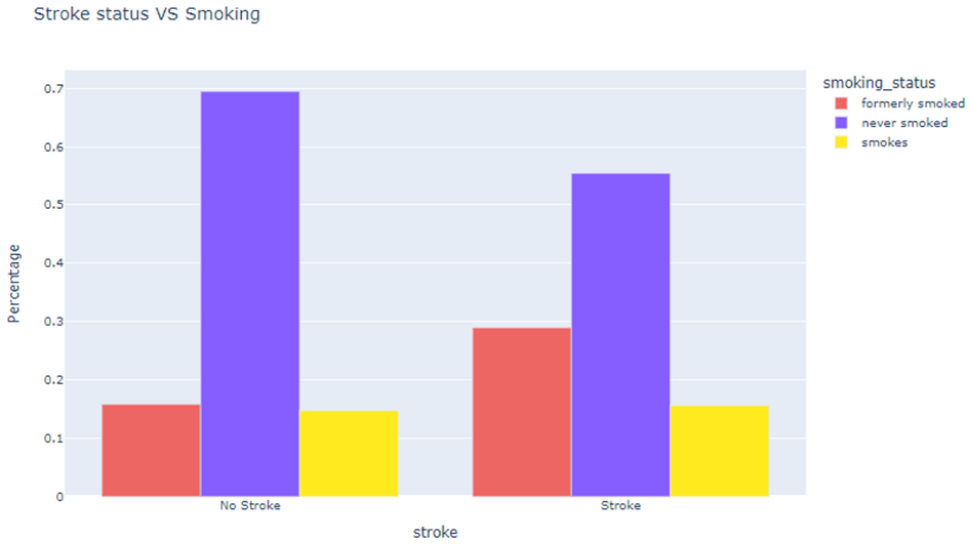

Figure 12: Distribution of Smoking Status vs Stroke

The influence of hypertension on stroke likelihood was explored through a bar plot, presenting the distribution of subjects with and without hypertension concerning stroke occurrence. This analysis illuminates the significance of hypertension as a potential contributing factor to stroke risk as shown in Figure 13.

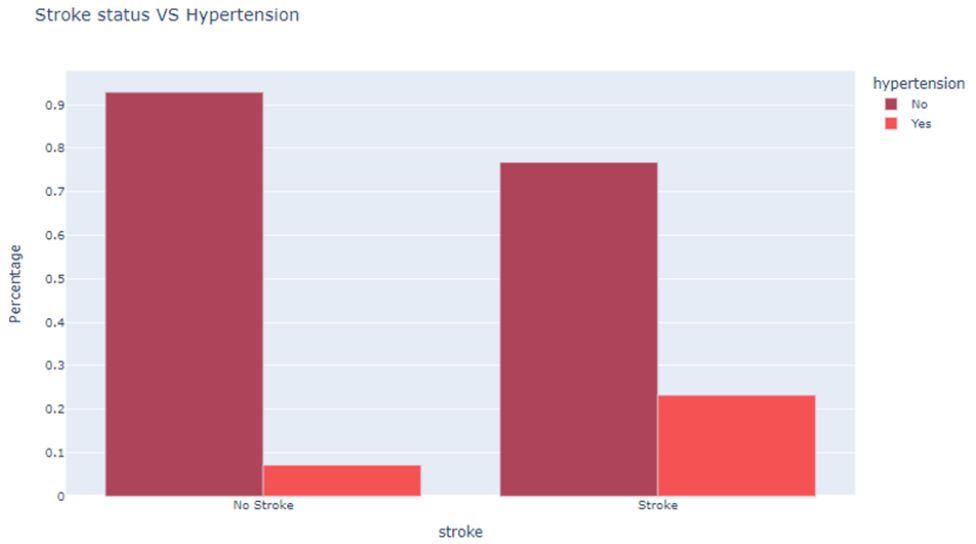

Figure 13: Distribution of Hypertension vs Stroke

A parallel exploration was conducted for the 'heart_disease' variable, visually representing its impact on stroke likelihood as shown in Figure 14. The resulting pie chart provides insights into the distribution of subjects with and without heart disease concerning stroke occurrence.

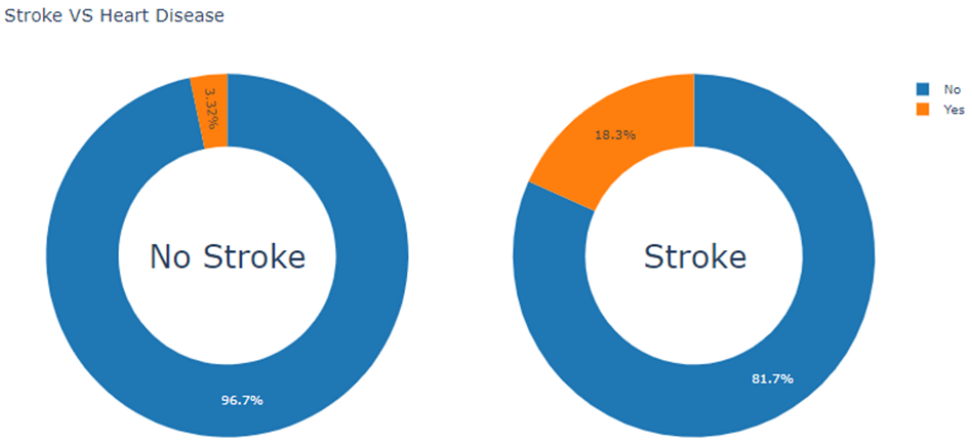

Figure 14: Distribution of Heart Disease vs Stroke

The work type and residence type variables were jointly examined through a dual-axis bar plot as shown in Figures 15 and Figure 16. This visualization method facilitates a comparative understanding of how work type and residence type may correlate with stroke occurrence, offering nuanced insights into potential occupational and residential influences.

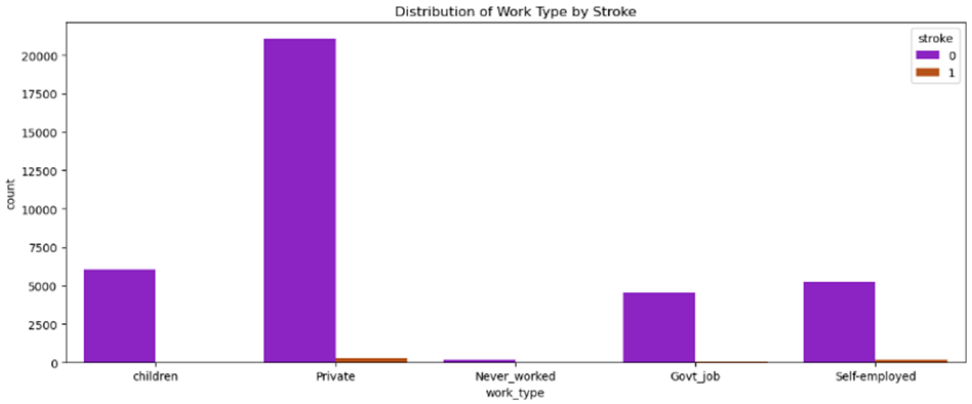

Figure 15: Distribution of WorkType vs Stroke

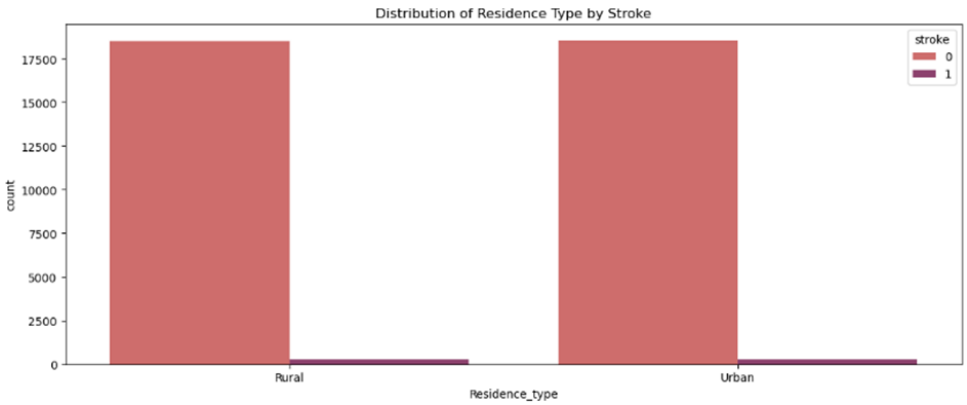

Figure 16: Distribution of Residence Type vs Stroke

Lastly, the 'ever_married' Column was analyzed in conjunction with stroke occurrence, providing a compelling overview of the distribution of married and unmarried individuals in the context of stroke likelihood as shown in Figure 17.

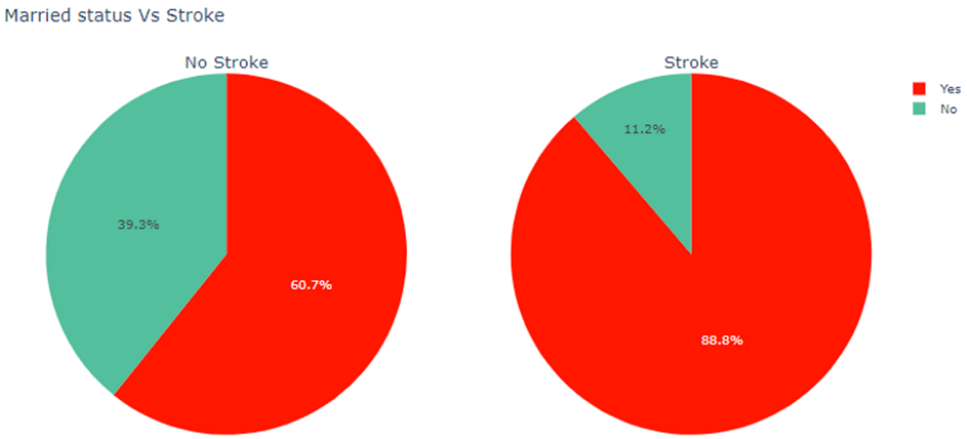

Figure 17: Distribution of Married Status vs Stroke

These comprehensive analyses collectively contribute to a nuanced understanding of the dataset's variables, laying the groundwork for informed feature selection and model development in the pursuit of accurate stroke prediction.

3.4 Feature Engineering

Feature engineering plays a crucial role in enhancing the predictive capabilities of machine learning models. Post-data analysis and visualization, First, all the categorical variables are converted into numerical ones using a label encoder followed by the separation of the target variable from the features data. Since this is the variable on which prediction will be made. From the data analysis, it is evident that the data is highly imbalanced as depicted in Figure 18 therefore SMOTE oversampling strategy is deployed to remove the imbalance in the dataset which is a potential threat to non-biased predictions. After applying the SMOTE Oversampling method data is found to be balanced as depicted in Figure 19. Feature selection plays an crucial role for predicting the model performance therefore in this research two feature selection methods has been deployed to understand how different feature selection methods affect the machine learning performance. Two Feature selection methods has been used in this research which will be discussed in further subsection.

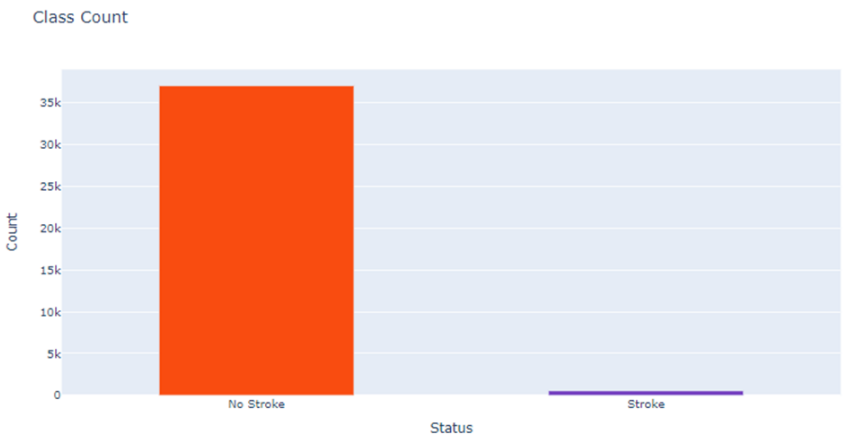

Figure 18: Imbalanced Classes in Stroke Variable

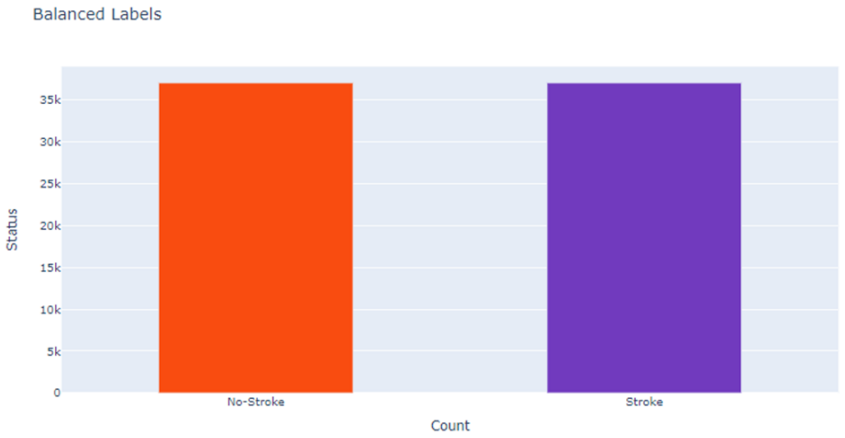

Figure 19: Balanced Classes in Stroke Variable After Oversampling

3.4.1 First Method (Correlation Analysis)

In the first method, correlation analysis, and subsequent feature selection were employed to identify relevant attributes for stroke prediction. By setting a correlation threshold of 0.02 and evaluating the Pearson correlation coefficients, a subset of features, including 'age,' 'hypertension,' 'heart disease,' 'ever_married,' and 'smoking_status,' were selected. The significance lies in isolating features that exhibit a meaningful relationship with the target variable, thereby focusing the model on key predictors and potentially improving its accuracy. The correlation analysis helps filter out attributes with weak associations, focusing on those most likely to impact stroke prediction. The correlation heatmap is shown in Figure 20

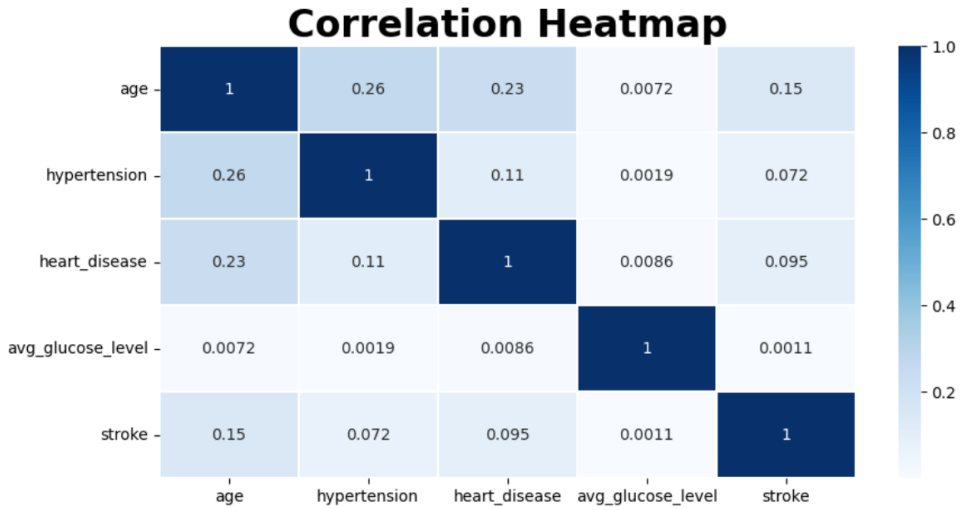

Figure 20: Correlation Matrix to Identify Feature Importance

3.4.2 Second Method (Feature Importance Ranking from Random Forest Classifier)

The second method of feature engineering involved leveraging the importance of features derived from a One-vs-Rest classifier, specifically employing a Random Forest classifier. Through this approach, features were ranked based on their importance scores, with 'age,' 'avg_glucose_level,' 'BMI,' 'smoking_status,' and 'work_type' emerging as the most influential predictors as shown in Figure 21. Significantly, this method emphasizes the intrinsic contribution of each feature to the model's predictive power, allowing for the identification and prioritization of impactful variables. The importance scores provide a quantitative measure of each feature's contribution, aiding in the selection of the most relevant attributes for stroke prediction.

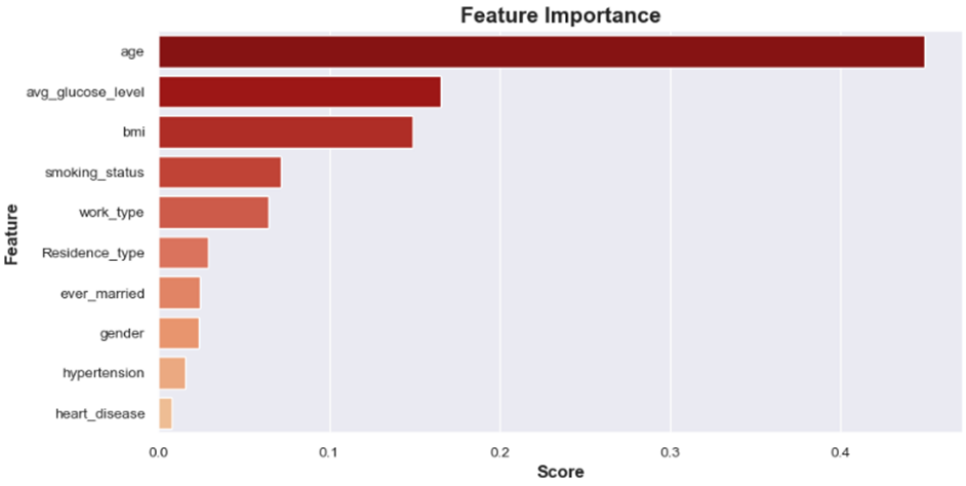

Figure 21: Feature Importance By One-vs-All Classifier

3.5 Model Training

Model training is a pivotal phase in the development of machine learning models, serving as a decisive step towards achieving accurate predictions. Following feature engineering, the dataset is divided into training and testing sets to facilitate the evaluation of model performance. In this study, the data is split using the train_test_split function, with 80% allocated for training and 20% for testing. This division ensures a robust assessment of the models on unseen data, a critical aspect in determining their real-world applicability. Two distinct feature engineering approaches were employed in tandem with four machine learning algorithms: K Nearest Neighbors (KNN), Balanced Random Forest, CatBoost, and XGBoost. Therefore each algorithm is trained on training data twice independently, the first training set extract is based on the first feature selection method and similarly second training set extract is based on the second feature selection method. The training of each algorithm involved fitting the model to the training data, allowing it to learn from the provided features and labels. Post-training, the models were evaluated using the testing set to gauge their generalization performance.

3.6 Model Evaluation

The evaluation of model performance in a binary classification task is a critical aspect of assessing the predictive capabilities of machine learning algorithms. Four key metrics, namely Accuracy, Precision, Recall, F1-Score, and ROC_AUC Score, were employed for a comprehensive evaluation. After training algorithms through each distinct method of feature selection, the algorithms are evaluated on test data produced by each distinct feature selection process separately to compare which method of feature selection is better and which algorithm is viable and accurate among all other implemented algorithms. The comprehensive use of these metrics contributes to a thorough evaluation, ensuring a well-informed decision-making process in the context of stroke prediction.

4 Design Specification

This research task is a classification task consisting of four machine learning algorithms: K-Nearest Neighbour, Balanced Random Forest, Catboost, and Xgboost. A crisp detail about these algorithms is present in the below subsections.

4.1 K-Nearest Neighbour Model

The K-Nearest Neighbor (KNN) algorithm, a versatile and intuitive approach in machine learning, serves both classification and regression purposes. In KNN, predictions for a new data point are determined by the majority class or mean of its k-nearest neighbors in the feature space. This non-parametric, lazy learning method excels in scenarios where local patterns are critical. KNN's simplicity and effectiveness make it a valuable tool, especially when dealing with datasets where the underlying distribution is not explicitly known or may vary across the data space. Its reliance on local information and ease of implementation contribute to its widespread applicability. Classification method of KNN is also shown with the help of graph in Figure 22.

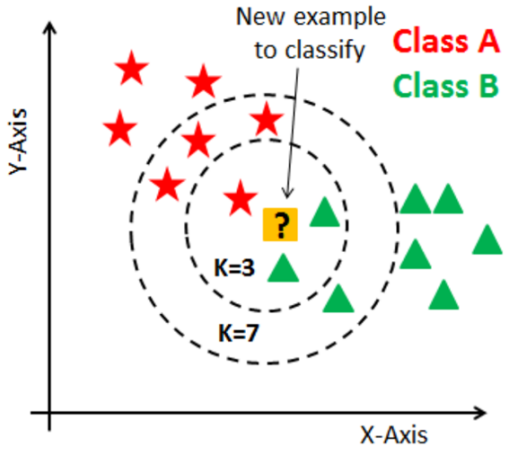

Figure 22: K-Nearest Neighbor Model Ye et al. (2013)

4.2 Balanced Random Forest Model

The Balanced Random Forest Regressor (BRF) is an ensemble learning technique that combines the strengths of randomization and bagging. Constructing multiple decision trees while balancing the class distribution within each tree, BRF excels in handling imbalanced datasets, a common challenge in real-world applications. The aggregation of predictions from diverse trees enhances predictive accuracy and robustness. BRF's ability to adapt to complex relationships within the data and its ensemble approach make it a valuable asset in regression tasks where mitigating bias and variance is crucial. Architecture of Random forest model is shown in Figure **??**.

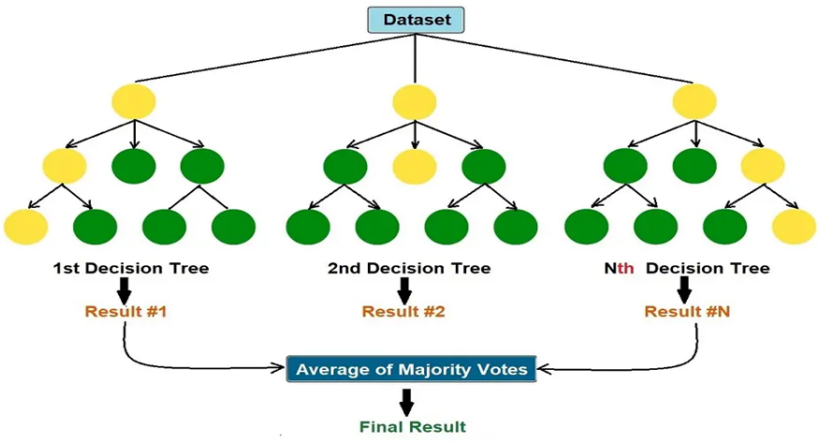

Figure 23: Balanced Random Forest Model *Spotfire — Demystifying the Random Forest Algorithm for Accurate Predictions — spotfire.com* (n.d.)

4.3 Catboost Model

Catboost, a specialized gradient-boosting algorithm, addresses the challenges associated with categorical feature handling. Its efficient gradient computation, particularly well-suited for large datasets, sets it apart. Catboost's unique strength lies in its ability

to handle categorical variables directly, sparing the need for extensive preprocessing. By incorporating a robust mechanism to prevent overfitting, Catboost strikes a balance between model complexity and generalization. Its efficiency, especially in scenarios with categorical features, positions Catboost as a go-to algorithm for predictive tasks where interpretability and ease of use are paramount. Working mechanism of Catboost algorithm is shown in Figure 24.

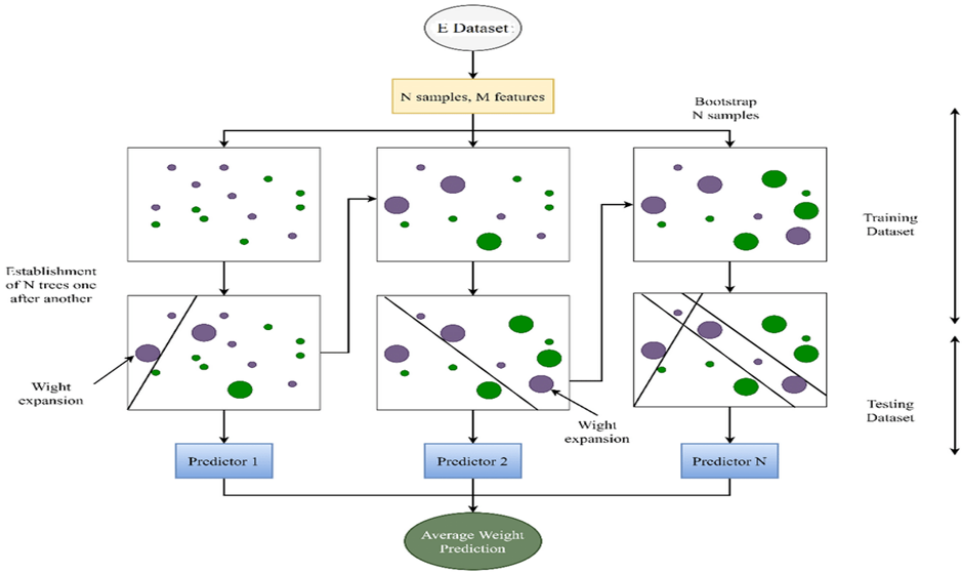

Figure 24: Catboost Model Shahani et al. (2022)

4.4 Xgboost model

Xgboost, or Extreme Gradient Boosting, stands out as a powerful ensemble learning algorithm renowned for its speed and performance. It employs a boosting framework to combine weak learners, typically decision trees, into a robust and accurate predictive model. Xgboost's strength lies in its ability to handle complex relationships and patterns within data, providing flexibility and customization options. Regularization techniques are applied to control overfitting, ensuring the model's adaptability across various datasets. Widely embraced in machine learning competitions and industry applications, Xgboost's efficiency and versatility contribute to its status as a leading algorithm in predictive modeling. Architecture diagram of XGBoost is shown in Figure 25.

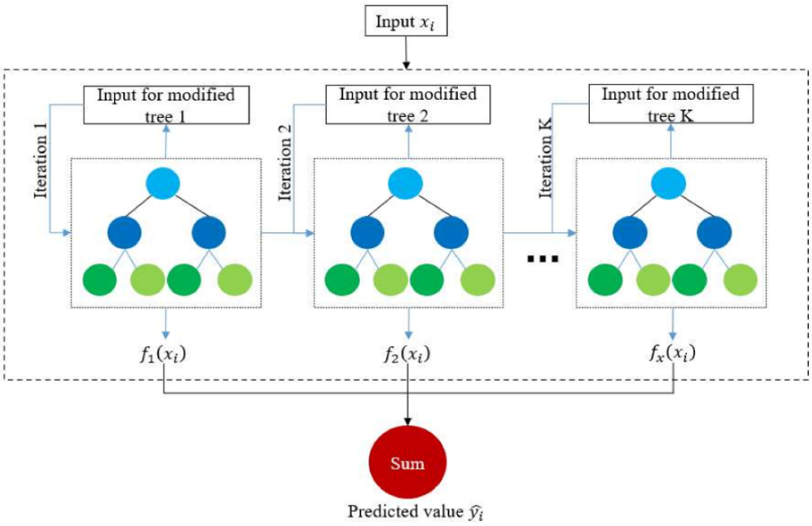

Figure 25: Xgboost Model Wang et al. (2020)

5 Implementation

This research work, conducted using the Python programming language and the Anaconda framework on a Windows operating system with 16GB RAM, focuses on the critical task of stroke prediction using machine learning algorithms. The necessity of such predictive models in the real world lies in their potential to assist healthcare professionals in identifying individuals at risk of stroke, allowing for timely interventions and preventive measures. The dataset for this study was collected from Kaggle, a reliable repository of diverse datasets, particularly from the medical domain. The task unfolds in two distinct feature engineering approaches. Feature selection based on correlation analysis and using One-Vs-Rest Classifier. Subsequently, the selected features were scaled down for compatibility with machine learning algorithms. Following feature engineering, the dataset was split into training and testing sets using the train_test_split function from Scikit-learn, ensuring a robust evaluation of model generalization. Four machine learning algorithms—K Nearest Neighbors (KNN), Balanced Random Forest, CatBoost, and XGBoost—were implemented for each feature engineering approach. Model training involved fitting each algorithm to the training data, allowing them to learn from the refined features. The model evaluation was conducted using regression metrics—Accuracy, Precision, Recall, F1-Score, and ROC_AUC Score—to comprehensively assess predictive performance. The significance of employing two feature engineering approaches lies in the nuanced understanding of attribute importance, enabling models to learn from refined features that contribute meaningfully to stroke prediction. The use of diverse machine learning algorithms ensures a comprehensive exploration of the dataset's predictive potential. This project, executed in Python and leveraging various libraries such as Pandas, NumPy, Scikit-learn, CatBoost, and XGBoost, not only contributes to the field of medical prediction but also underscores the practical applicability of machine learning in healthcare decision support systems.

Table 1: System Configuration and Resource Information

Component	**Specification**
CPU	Intel Core i5 (8 Cores)
RAM	16GB DDR4
Storage	1TB HDD
Operating System	Windows 11
Python Version	3.11
Machine Learning Libraries	Pandas, Numpy, Scikit-Learn, XGBoost, Plotly, Matplotlib, Imblearn
Additional Software	Jupyter Notebook

6 Evaluation

This research encompasses the utilization of four distinct machine learning algorithms: K-Nearest Neighbors, Balanced Random Forest, Catboost, and Xgboost. Given the binary classification nature of the task at hand, five key classification metrics—namely, accuracy, precision, recall, F1-score, and Roc_AUC Score—are employed to ensure a comprehensive evaluation of each algorithm. The subsequent subsections delve into a detailed discussion of the outcomes derived from the assessment of each algorithm based on these metrics.

6.1 Evaluation Based on Accuracy

Accuracy is a measure of how often a machine learning model makes correct predictions. It's calculated as the ratio of correctly predicted observations to the total observations. An optimal model should have a high accuracy score.

In the first experimentation, The evaluation of each model based on the Accuracy metric reveals insightful perspectives on their performance. In the first approach of feature engineering (Correlation Analysis), the K Nearest Neighbors algorithm achieved an accuracy of 91.67%, showcasing its effectiveness in predicting stroke outcomes. The Balanced Random Forest and CatBoost models exhibited high accuracies of 93.53% and 92.97%, respectively. Notably, XGBoost outperformed others with an impressive accuracy of 93.86%. This suggests that the initial feature engineering approach, which incorporated correlation analysis, effectively facilitated model learning and prediction accuracy. In contradistinction, the second approach of feature engineering demonstrated (Feature ranking with random forest) its efficacy in enhancing model accuracy further. The K Nearest Neighbors model achieved an accuracy of 92.74%, showcasing the positive impact of the second feature engineering method. The Balanced Random Forest and CatBoost models displayed remarkable accuracy levels of 97.52% and 97.64%, respectively. XGBoost emerged as the top-performing algorithm, attaining an accuracy of 98.11%. The superior accuracy of XGBoost in the second approach underscores the significance of employing feature importance through a One-vs-Rest classifier. The algorithm that exhibited the best performance overall is XGBoost in the second feature engineering approach, achieving an accuracy of 98.11%. This approach, utilizing feature importance derived from a One-vs-Rest classifier, proved instrumental in highlighting critical features, enabling XGBoost to make more informed predictions. Figure 26 represents the comparative analysis of models based on Accuracy.

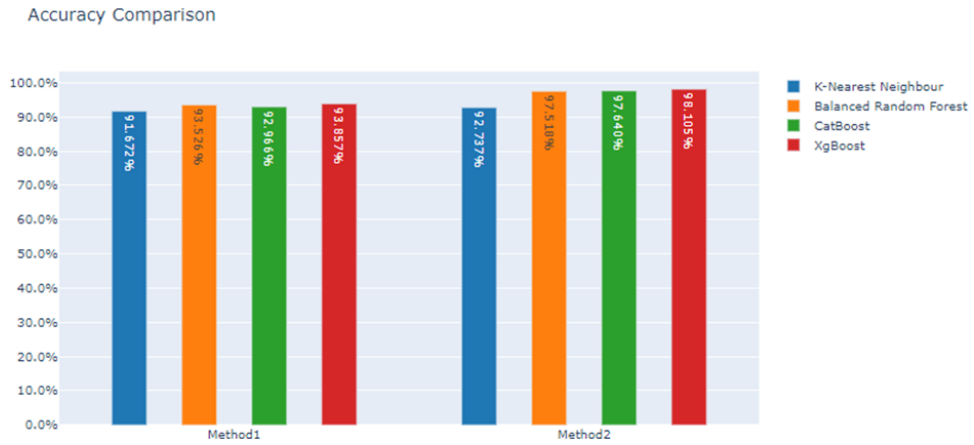

Figure 26: Accuracy-based Comparative Analysis of Models

6.2 Evaluation Based on Precision

In the second experiment, the Precision metric provides a detailed insight into the models' ability to correctly identify positive instances. In the first approach of feature engineering, the K Nearest Neighbors demonstrated a commendable precision of 0.9172. The Balanced Random Forest and CatBoost models exhibited high precision values of 0.9353 and 0.9298, respectively. XGBoost showcased the highest precision among all models, achieving an impressive 0.9386. Moving to the second feature engineering approach, K Nearest Neighbors maintained a solid precision of 0.9396, showcasing the efficacy of the alternative feature engineering method. The Balanced Random Forest and CatBoost models exhibited exceptional precision values of 0.9753 and 0.9764, respectively. XGBoost outperformed others with the highest precision of 0.9811, affirming the positive impact of the second feature engineering approach. The algorithm that emerged as the top performer overall is XGBoost in the second feature engineering approach. The nuanced understanding of feature importance provided by the second approach likely facilitated XGBoost's superior precision, emphasizing the importance of strategic feature selection and refinement in the model training process. Comparative analysis of models based on precision is depicted in Figure 27.

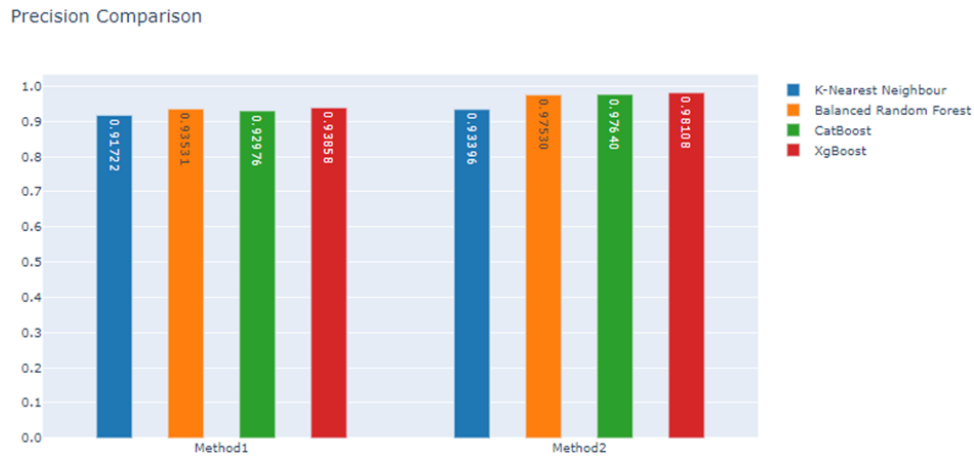

Figure 27: Precision-based Comparative Analysis of Models

6.3 Evaluation Based on Recall

In the third experiment, the evaluation based on the Recall metric provides insights into the models' ability to capture positive instances comprehensively. In the context of the first feature engineering approach, K Nearest Neighbors exhibited a solid recall of 91.67%, showcasing its effectiveness in identifying true positive cases. Balanced Random Forest, CatBoost, and XGBoost demonstrated high recall values of 0.9353, 0.9297, and 0.9386, respectively. Transitioning to the second feature engineering approach, K Nearest Neighbors maintained a strong recall of 0.9274, emphasizing the efficacy of the alternative feature engineering method. The Balanced Random Forest and CatBoost models showcased exceptional recall values of 0.9752 and 0.9764, respectively. XGBoost surpassed others, achieving the highest recall of 0.9811, affirming the positive impact of the second feature engineering approach. The incorporation of feature importance from a One-vs-Rest classifier likely played a crucial role in emphasizing critical features, enabling XGBoost to achieve superior recall. Recall comparison of models is implemented in Figure 28.

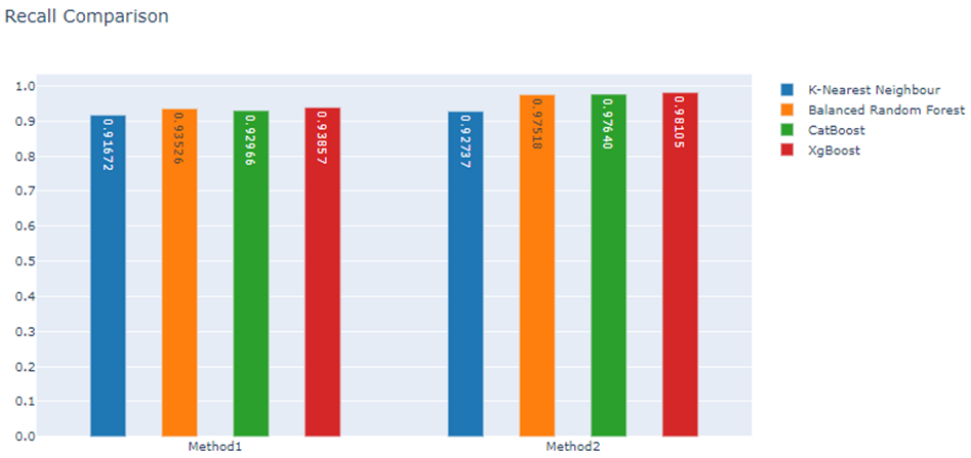

Figure 28: Recall-based Comparative Analysis of Models

6.4 Evaluation Based on F1-Score

In the fourth Experiment, the F1-Score metric delves into the balance between precision and recall, providing insights into a model's ability to achieve both high positive predictive value and sensitivity. In the context of the first feature engineering approach, all algorithms demonstrated commendable F1 scores. K Nearest Neighbors, Balanced Random Forest, CatBoost, and XGBoost achieved a solid F1-Score of 0.9167, 0.9353, 0.9297, and 0.9386, respectively while in the second feature engineering approach, K Nearest Neighbors maintained a high F1-Score of 0.9272, emphasizing the robustness of the alternative feature engineering method. Balanced Random Forest and CatBoost excelled further, achieving impressive F1 scores of 0.9752 and 0.9764, respectively. XGBoost outperformed others with the highest F1-Score of 0.9811, highlighting the substantial impact of the second feature engineering approach. The algorithm that emerged as the top performer overall is XGBoost in the second feature engineering approach. The result obtained by models through the second feature engineering method is more accurate than the First one which indicates in this task second method is more compatible with

real-world deployment. Figure 29 represents the comparative analysis of models based on the F1-score.

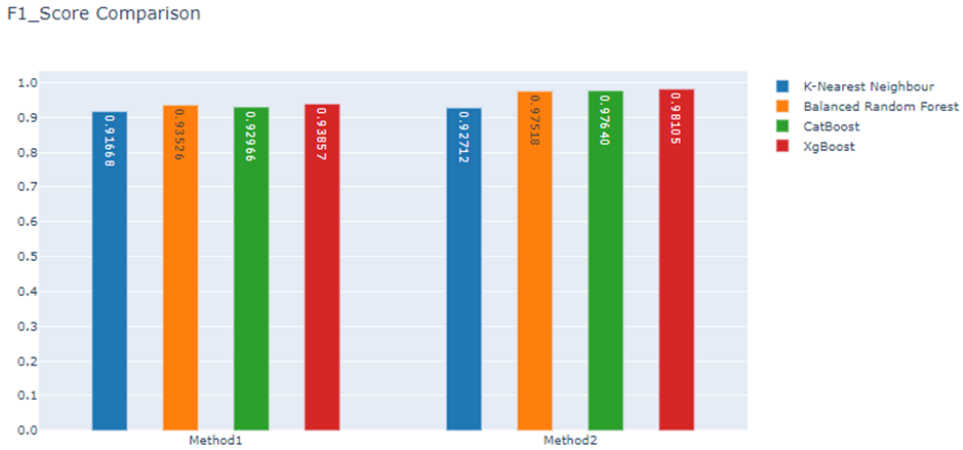

Figure 29: F1-Score-based Comparative Analysis of Models

6.5 Evaluation Based on Roc_AUC Score

The last experiment indulges the ROC_AUC Score metric that provides a comprehensive assessment of the model's ability to discriminate between positive and negative instances across various classification thresholds. In the first feature engineering approach, K Nearest Neighbors achieved a commendable ROC_AUC Score of 0.9633, emphasizing its discriminative power. Balanced Random Forest, CatBoost, and XGBoost also exhibited strong ROC_AUC Scores of 0.9840, 0.9828, and 0.9868, respectively. Concerning the second feature engineering approach, K Nearest Neighbors maintained a high ROC_AUC Score of 0.9683, underscoring the efficacy of the alternative feature engineering method. Balanced Random Forest and CatBoost excelled, achieving impressive ROC_AUC Scores of 0.9974 and 0.9828, respectively. XGBoost outperformed others with the highest ROC_AUC Score of 0.99.77 highlighting the substantial impact of the second feature engineering approach. The algorithm that emerged as the top performer overall is XGBoost in the second feature engineering approach. It is highly evident from the results obtained by models utilizing the second approach, that can surpass the first method of feature Engineering. The comparative analysis of models is depicted in Figure 30.

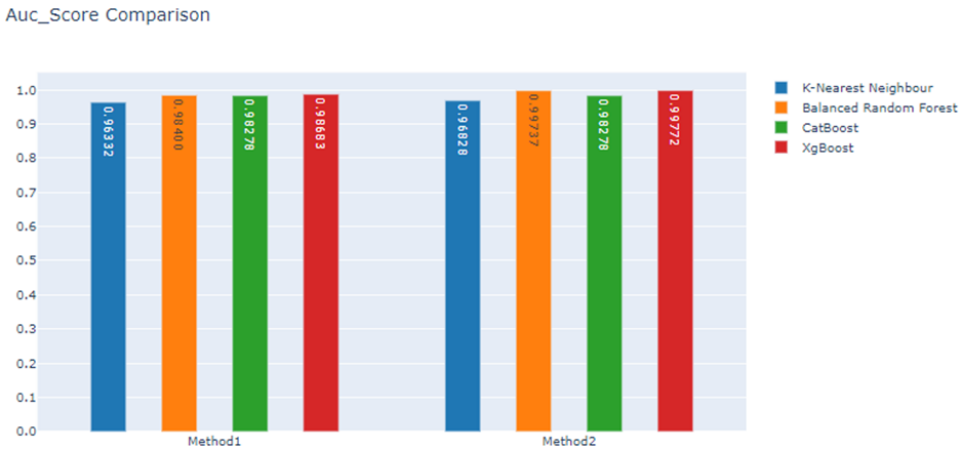

Figure 30: Roc_AUC Score-based Comparative Analysis of Models

6.6 Discussion

The discussion revolves around the comprehensive evaluation of machine learning algorithms for stroke prediction in the context of the employed dataset and task. Among the algorithms investigated, XGBoost emerged as the top-performing model, showcasing superior performance across various evaluation metrics. Notably, XGBoost achieved an outstanding F1-Score of 0.9811, signifying its exceptional balance between precision and recall. This algorithm also excelled in other metrics, including accuracy 98.11%, precision 0.9811, recall (0.9811, and ROC_AUC score of 0.9977. The confusion matrix obtained from the XG-Boost model for the Test Set is shown in Figure 31. From the confusion matrix, it can seen that the number of True positive and true negative are quite high, and false positive and false negative values are quite low which represent the most optimal outcomes achieved by the XG-Boost model. On analysing the AUC score of XG-Boost model an score of 0.9977 has been obtained which is very close 1. The closer the AUC score is to 1, the better the model is at distinguishing between the positive class (people who have had a stroke) and the negative class (people who have not had a stroke) as shown in Figure 32.

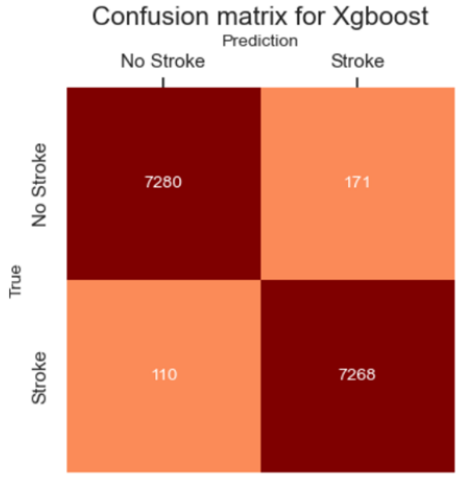

Figure 31: Confusion Matrix Analysis of XG-Boost Model

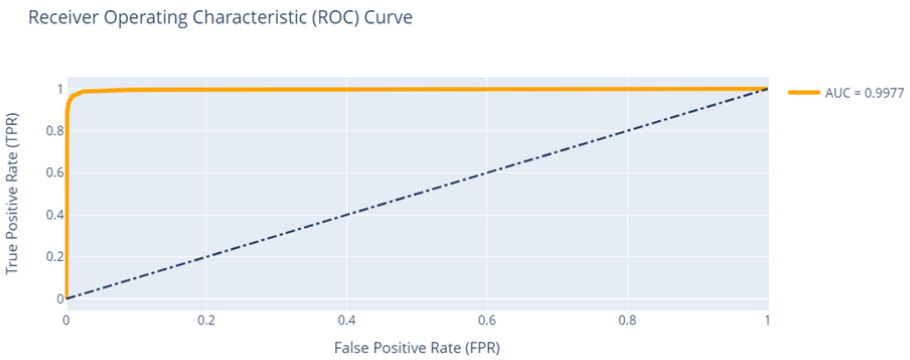

Figure 32: Roc_AUC Score-Graph of XG-Boost Model

While XGBoost demonstrated remarkable performance, it is crucial to acknowledge the commendable results achieved by other algorithms, such as K Nearest Neighbors, Balanced Random Forest, and CatBoost, underlining the overall efficacy of the models. The results of each model, when trained and evaluated on each metric using the features selected by the second method, surpass the results obtained by each model when trained and evaluated utilizing the features selected with the first method. The success of XGBoost can be attributed to its inherent capacity for capturing complex relationships within the dataset, as well as the nuanced feature engineering approach employed in the second method, leveraging feature importance. This strategic combination facilitated superior discrimination between positive and negative instances, contributing to XGBoost's exceptional predictive accuracy and making it the algorithm of choice for stroke prediction in this specific context.

7 Conclusion and Future Work

A stroke poses a significant threat to human life, necessitating preventive measures and timely intervention to avert unforeseen complications. In the contemporary era marked by the swift evolution of machine learning, one can leverage established models to discern pertinent features, commonly referred to as risk factors, associated with stroke occurrence. Furthermore, these models facilitate the assessment of the respective probability or risk of experiencing a stroke. Our Research includes a detailed comparison of four different machine learning models (kNN, Balanced Random Forest, Cat-Boost, and XG-Boost). This extensive comparison across multiple models is unique in providing a broader understanding of which algorithms perform best in stroke prediction under varied conditions. The implemented research embraced a noble approach, leveraging meticulous data preprocessing, insightful feature engineering, and systematic model training to predict stroke occurrences. Among the ensemble of machine learning algorithms explored, XGBoost emerged as the most proficient model, exhibiting superior predictive capabilities with an Accuracy of 98.11%, F1-Score of 0.9811, and ROC_AUC Score of 0.9977. The judicious application of feature engineering, particularly through the identification of crucial features using correlation analysis and feature importance, contributed significantly to the success of XGBoost in discriminating between positive and negative instances. Moving forward, the future scope of this research involves the exploration of additional advanced machine learning techniques, integration of more diverse datasets, and consideration of

evolving medical knowledge to enhance predictive accuracy. Furthermore, the potential incorporation of real-time health monitoring systems and continuous model refinement could elevate the practical applicability of the predictive models developed in this thesis, thereby contributing to more effective and personalized stroke prevention strategies.

References

Biswas, N., Uddin, K. M. M., Rikta, S. T. and Dey, S. K. (2022). A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach, *Healthcare Analytics* **2**(100116): 100116.

Cerebral Stroke Prediction-Imbalanced Dataset — kaggle.com (n.d.). `https://www.kaggle.com/datasets/shashwatwork/cerebral-stroke-predictionimbalaced-dataset/`.

Govindarajan, P., Soundarapandian, R. K., Gandomi, A. H., Patan, R., Jayaraman, P. and Manikandan, R. (2020). Classification of stroke disease using machine learning algorithms, *Neural Comput. Appl.* **32**(3): 817–828.

Kokkotis, C., Giarmatzis, G., Giannakou, E., Moustakidis, S., Tsatalas, T., Tsiptsios, D., Vadikolias, K. and Aggelousis, N. (2022). An explainable machine learning pipeline for stroke prediction on imbalanced data, *Diagnostics (Basel)* **12**(10): 2392.

Liu, T., Fan, W. and Wu, C. (2019). A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset, *Artif. Intell. Med.* **101**(101723): 101723.

Lyashevska, O., Malone, F., MacCarthy, E., Fiehler, J., Buhk, J.-H. and Morris, L. (2021). Class imbalance in gradient boosting classification algorithms: Application to experimental stroke data, *Stat. Methods Med. Res.* **30**(3): 916–925.

Michelson, K. N., Klugman, C. M., Kho, A. N. and Gerke, S. (2022). Ethical considerations related to using machine learning-based prediction of mortality in the pediatric intensive care unit, *J. Pediatr.* **247**: 125–128.

Paliwal, S., Parveen, S., Alam, M. A. and Ahmed, J. (2023). Improving brain stroke prediction through oversampling techniques: A comparative evaluation of machine learning algorithms.

Phankokkruad, M. and Wacharawichanant, S. (2022). Performance analysis and comparison of cerebral stroke prediction models on imbalanced datasets, *2022 IEEE/ACIS 7th International Conference on Big Data, Cloud Computing, and Data Science (BCD)*, IEEE.

Sailasya, G. and Kumari, G. L. A. (2021). Analyzing the performance of stroke prediction using ML classification algorithms, *Int. J. Adv. Comput. Sci. Appl.* **12**(6).

Santos, L. I., Camargos, M. O., D'Angelo, M. F. S. V., Mendes, J. B., Medeiros, E. E. C. d., Guimarães, A. L. S. and Palhares, R. M. (2022). Decision tree and artificial immune systems for stroke prediction in imbalanced data, *Expert Syst. Appl.* **191**(116221): 116221.

Shahani, N., Zheng, X., Guo, X. and Wei, X. (2022). Machine learning-based intelligent prediction of elastic modulus of rocks at thar coalfield, *Sustainability* **14**: 3689.

Singh, M. S. and Choudhary, P. (2017). Stroke prediction using artificial intelligence, *2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)*, IEEE.

Spotfire — Demystifying the Random Forest Algorithm for Accurate Predictions — spotfire.com (n.d.). `https://www.spotfire.com/glossary/what-is-a-random-forest`.

Tazin, T., Alam, M. N., Dola, N. N., Bari, M. S., Bourouis, S. and Monirujjaman Khan, M. (2021). Stroke disease detection and prediction using robust learning approaches, *J. Healthc. Eng.* **2021**: 7633381.

Uttam, A. K. (2022). Analysis of uneven stroke prediction dataset using machine learning, *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE.

Wang, M., Yao, X. and Chen, Y. (2021). An imbalanced-data processing algorithm for the prediction of heart attack in stroke patients, *IEEE Access* **9**: 25394–25404.

Wang, W., Chakraborty, G. and Chakraborty, B. (2020). Predicting the risk of chronic kidney disease (ckd) using machine learning algorithm, *Applied Sciences* **11**: 202.

Ye, R., Zhang, L. and Suganthan, P. (2013). K-nearest neighbor based bagging svm pruning, pp. 25–30.