The Investigation of Prediction for Stroke Using Multiple Machine Learning Models

Jingrui Wen

Shenzhen OCAHS International Academy, Shenzhen, China

wenjingrui2024@hhu.edu.cn

Abstract. The primary objective of this research is to forecast stroke occurrence on an individual patient level. Through exploratory data analysis, the study has brought to light noteworthy disparities in the distribution of stroke and non-stroke cases, shedding light on the influence of diverse health and lifestyle factors on stroke susceptibility. This project underscores the immense potential of machine learning in the realm of medical prediction, serving to aid patients in risk assessment and aiding medical practitioners in devising treatment strategies. Concerning the predictive models employed, this research leveraged two distinct models, namely RandomForest and DecisionTree. Additionally, it utilized evaluation metrics such as the Confusion Matrix, Receiver Operator Characteristic (ROC) curve, and Precision-Recall curve, each of which provided comprehensive insights into the performance of the prediction models. One noteworthy aspect of this study is the presence of missing data within certain features, underscoring the challenges posed by data gaps in medical prediction and the imperative need for effective methods to handle missing data. The experimental outcomes unveiled an Area Under the Curve (AUC) of 95% for RandomForest and 92% for DecisionTree, indicating robust predictive capabilities. Future endeavors may concentrate on refining prediction models, achieving greater balance, and expanding the dataset to enhance prediction precision.

Keywords: Stroke, medical prediction, machine learning.

1. Introduction

According to the World Stroke Organization, stroke ranks as the second leading cause of death and the third leading cause of disability worldwide. Every year, approximately 3.3 million people lose their lives to stroke [1]. Stroke is a pervasive presence in our lives, impacting not only the affected individuals but also their families and the broader social environment. Moreover, stroke does not discriminate; it can affect anyone, regardless of age, gender, or physical condition.

A stroke is defined as an acute neurological condition that occurs when the blood supply to a part of the brain is interrupted [2-4], depriving brain cells of vital oxygen. There are two primary types of strokes: ischemic and hemorrhagic. Hemorrhagic strokes result from the rupture of a blood vessel, leading to bleeding within the brain. The more common type, ischemic stroke, is characterized by the blockage or narrowing of blood vessels, which leads to the cessation of blood flow to a specific brain area.

High blood pressure, obesity, physical inactivity, a poor diet, and smoking are some factors that raise the risk of stroke. Other risk factors depend on a person's environment, genetics, and way of life. Every age can experience a stroke; infants as young as one-year-old can even have a stroke. The risk of stroke doubles every ten years beyond the age of 55.

The onset of a stroke can be unpredictable, sometimes striking even as you awaken and adjust in bed. When a stroke happens, it can manifest through distinct symptoms, including: 1) Difficulty with speech and comprehension, making it challenging to understand or communicate with others. 2) Facial, arm, or leg paralysis or numbness. Vision issues affecting one or both eyes. 3) The onset of a sudden and intense headache, which may be accompanied by vomiting, dizziness, or changes in consciousness. 4) Difficulty walking, often leading to stumbling or loss of balance.

While certain individuals do make a recovery following a stroke, the extent of recuperation largely hinges on the stroke's severity. For the majority, persistent challenges persist, encompassing issues like 1) Memory, concentration, and attention deficits. 2) Speech difficulties and language

comprehension challenges. 3) Emotional struggles, including depression. 4) Impaired balance, walking capabilities, and the ability to move on one side of the body. 5) Trouble with swallowing food.

Although some patients recover after a stroke, depending on the severity of the stroke, the vast majority continue to experience problems such as memory, concentration, and attention problems, difficulty speaking or understanding language, emotional issues such as depression, loss of balance, ability to walk, loss of feeling on one side of the body and difficulty swallowing food.

In this case, machine learning algorithms can be considered for solving them to a certain extent due to their excellent performance [5-8]. This study will use the RandomForest [9] prediction model and DecisionTree [10] prediction model to predict the incidence of stroke. By comparing the accuracy of Randomforest and DecisionTree, it is concluded that the RandomForerst prediction model is more accurate. This study used a dataset from Kaggle, and the main problem was that part of the "BMI" feature was missing, emphasizing the necessity for effective missing data imputation methods in healthcare prediction tasks. The purpose is to evaluate RandomForest and DecisionTree. The experimental results show that the RandomForest method is good at large datasets, handles missing values, and reduces overfitting. And achieves high AUC, precision, and f1 score.

2. Dataset

This dataset shown in Figure 1 contains 5110 samples and 11 factors. The data set to be used includes the patient's various health and lifestyle features. Each row represents a patient's age, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI, smoking status, and gender. The dataset also contains a target variable, stroke, which represents the occurrence of a stroke.

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	1	88	0	1	1	2	1	3850	240	0	1
1	0	82	0	0	1	3	0	3588	162	1	1
2	1	101	0	1	1	2	0	2483	199	1	1
3	0	70	0	0	1	2	1	3385	218	2	1
4	0	100	1	0	1	3	0	3394	113	1	1
...	...	...	...	...	...	...	...	...	...	...	...
5105	0	101	1	0	1	2	1	1360	162	1	0
5106	0	102	0	0	1	3	1	3030	274	1	0
5107	0	56	0	0	1	3	0	1314	180	1	0
5108	1	72	0	0	1	2	0	3363	129	0	0
5109	0	65	0	0	1	0	1	1454	135	1	0

5110 rows × 11 columns

Figure 1. The samples in the collected dataset (Photo/Picture credit: Original).

Age: This is the patient's age. Age is a crucial factor in stroke prediction because the risk of stroke increases with age. According to the World Health Organization, the risk of stroke doubles every ten years after age 55.

High blood pressure is a significant risk factor for stroke because it can damage blood vessels and make them more susceptible to blockages or ruptures.

Heart disease: This binary characteristic indicates whether the patient has a heart disease. Patients with heart disease are at higher risk of stroke because these diseases can lead to the formation of blood clots in the heart that can travel to the brain.

There are six steps in the research to create a model to predict. Data loadin, which means load and preprocess the data for analysis and modeling. Exploratory Data Analysis (EDA), is to perform exploratory data analysis to gain insights into the dataset, understand the distributions of features, and explore potential relationships between the elements and the stroke outcome. Data Cleansing is perform imputing missing values and data transformation to improve the model's performance. Model Training and Validation, train the model using a train-test split strategy and make predictions on the test set. Model Evaluation, evaluate the performance of the trained model using appropriate evaluation metrics such as the confusion matrix, ROC curve, and Precision-Recall curve, and assess

the model's ability to generalize to unseen data using the test set. Prediction: Use the trained model to predict new, unseen data. If applicable, deploy the model for practical use or further analysis.

For the exploratory data analysis (EDA), this study will proceed with the following steps:

Univariate Analysis: This paper will inspect each variable individually to understand its distribution and potential outliers. Variate Analysis: This paper will explore the relationship between each variable and the target variable stroke. Multivariate Analysis: This paper will study the interactions between different variables and how they collectively relate to the target variable stroke. This study will start with the target variable stroke and then move on to the other variables. Since stroke is a binary variable, this study used a bar chart to visualize its distribution.

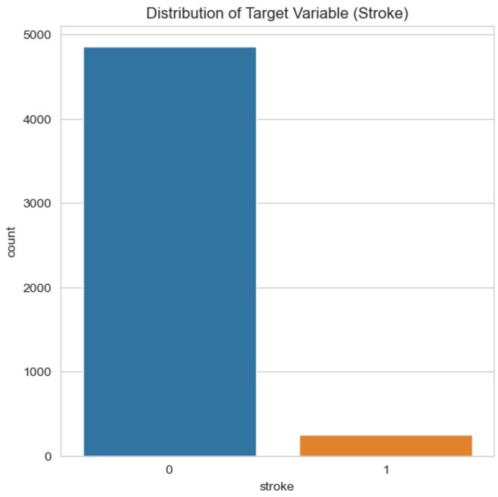

Figure 2. The distribution of Target Variable (Photo/Picture credit: Original)

The target variable “stroke” is highly imbalanced shown in Figure 2 because there are many more instances of class 0 (no stroke) than class 1 (stroke). This observation is important because it impacts the choice of machine learning model and evaluation metrics.

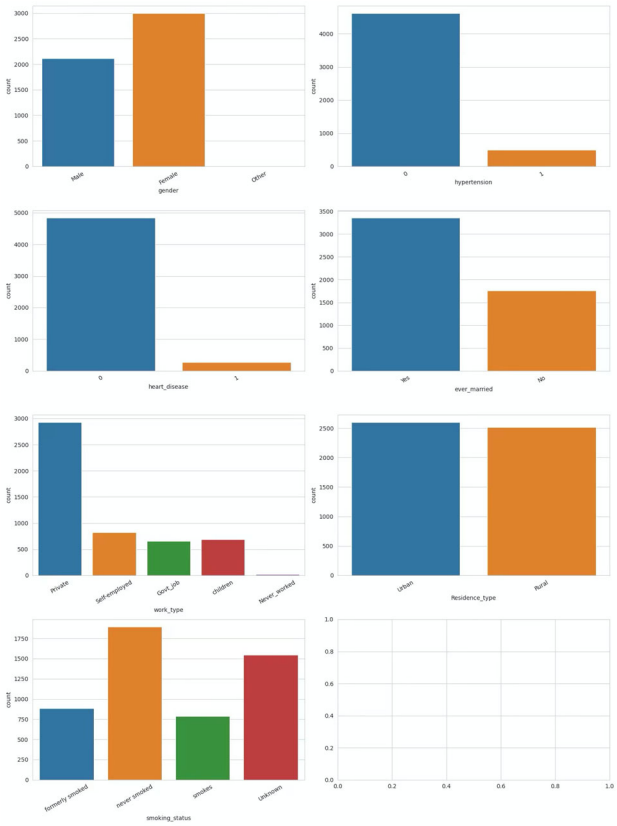

Figure 3. The distribution of Category Variables (Photo/Picture credit: Original)

Here are the observations from the distribution of categorical variables shown in Figure 3.

Gender: There are more female patients than male patients, and some identify as “other.”

High blood pressure: Most patients do not have high blood pressure.

Heart disease: Most patients do not have high blood pressure heart disease.

Ever_married: Most patients have been married at least once.

Work_type: Most patients fall into the private work category. There are also a significant number of self-employed people and children. Govt_job and Never_worked types have fewer patients.

Residence_type: The number of patients living in urban and rural areas is almost the same.

Smoking_status: Most patients have never smoked. The early smokers and smokers categories have fewer patients. There is a significant proportion of patients with unknown smoking status.

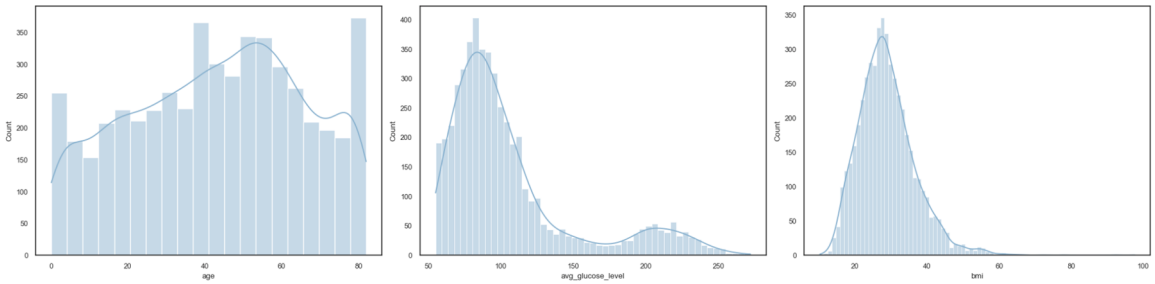

Figure 4. The distribution of Continuous Variables (Photo/Picture credit: Original)

Here are the observations from the distribution of continuous variables shown in Figure 4: Patients' ages vary from young to old, with most patients ranging from 40 to 80 years old. avg_glucose_level: Most patients have an average glucose level in the range of 50-125, although there are many patients with higher values. The distribution is skewed to the right.BMI: Most patients have a BMI range of 20-40, which is considered normal to overweight. There are some outliers with highly high BMI values. Figure 5 shows the distribution of stroke in each property.

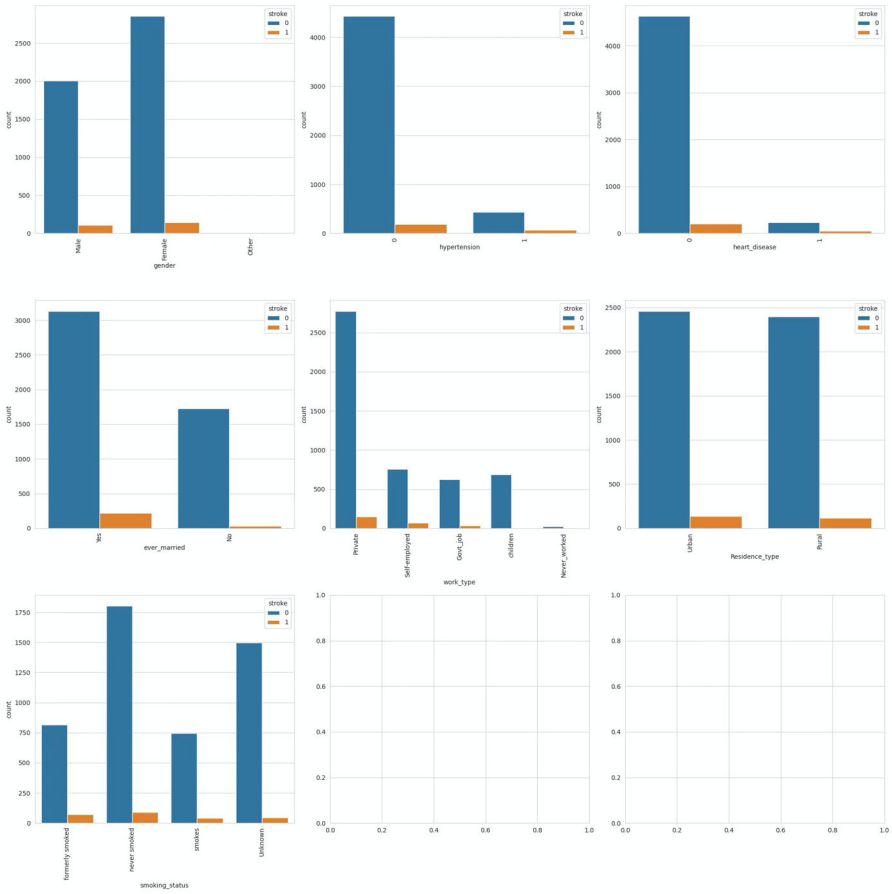

Figure 5. The distribution of stroke in each property (Photo/Picture credit: Original).

Gender: Both men and women have a similar proportion of stroke cases, with men having slightly more. There are no stroke cases in the "Other" category, but this could be due to the small sample size of this category.

High blood pressure: More stroke cases occur in patients with high blood pressure than in those without high blood pressure.

Heart disease: People with heart disease have more strokes than people without heart disease.

Ever_married: Married people have more stroke cases than unmarried people.

Work_type: Patients who are self-employed or work in private jobs have a higher proportion of stroke cases than other occupational types.

Residence_type: The ratio of stroke cases is almost the same between urban and rural residents.

Smoking_status: The proportion of stroke cases is higher in patients who have previously or currently smoked than in patients who have never smoked. The ratio of strokes for the "Unknown" category is lower.

The first heatmap shown in Figure 6 visualizes the correlation between all pairs of features in the dataset. The color of each cell represents the correlation coefficient between the pair of variables: blue for positive correlation and red for negative correlation. The darker the color, the stronger the correlation.

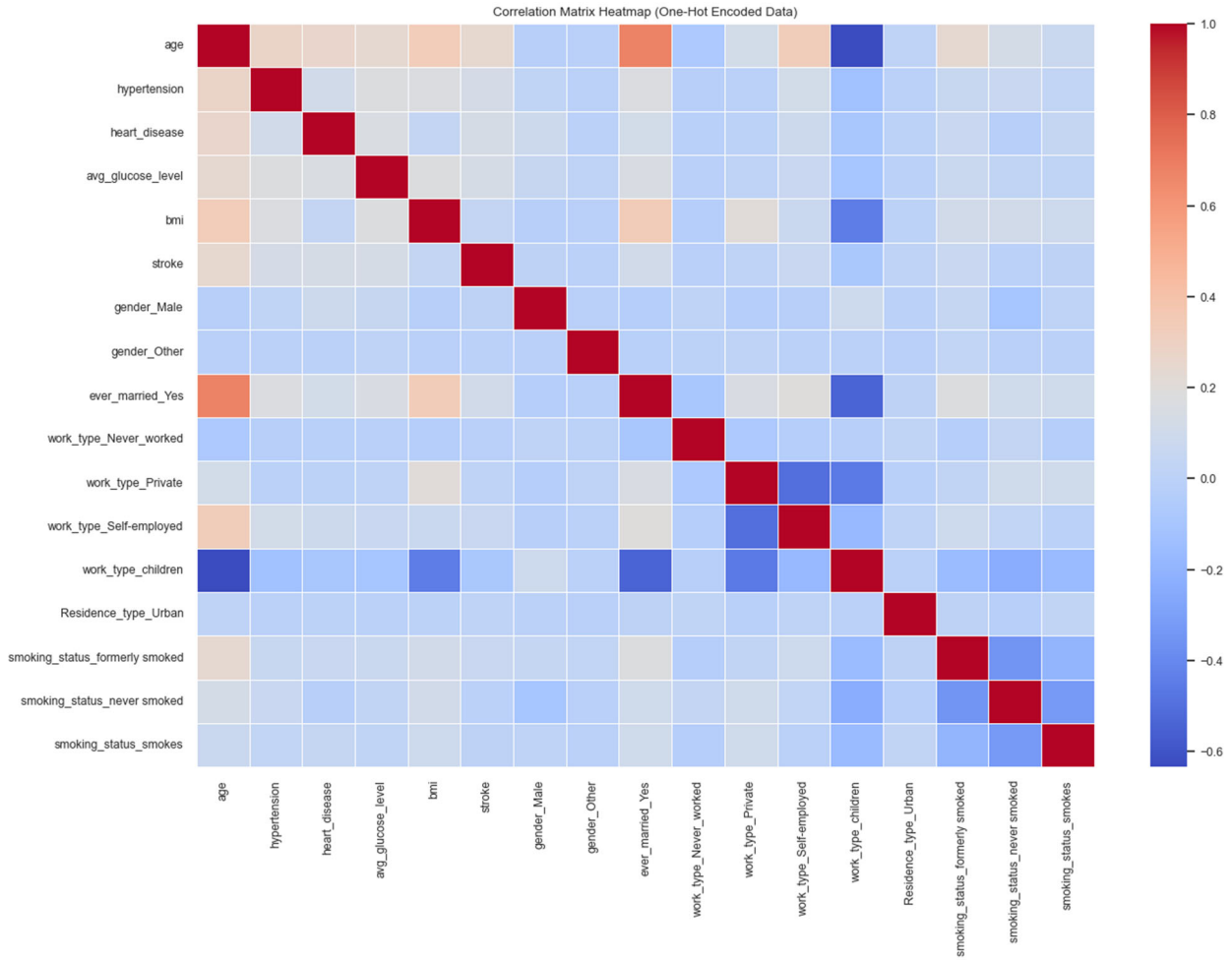

Figure 6. The visualization of heat map (Photo/Picture credit: Original).

The second heatmap shown in Figure 7 ranks the features based on their correlation with the target variable stroke. The age feature shows the highest positive correlation with the stroke target variable. This suggests that older people might be at a higher risk of stroke, aligning with known medical observations.

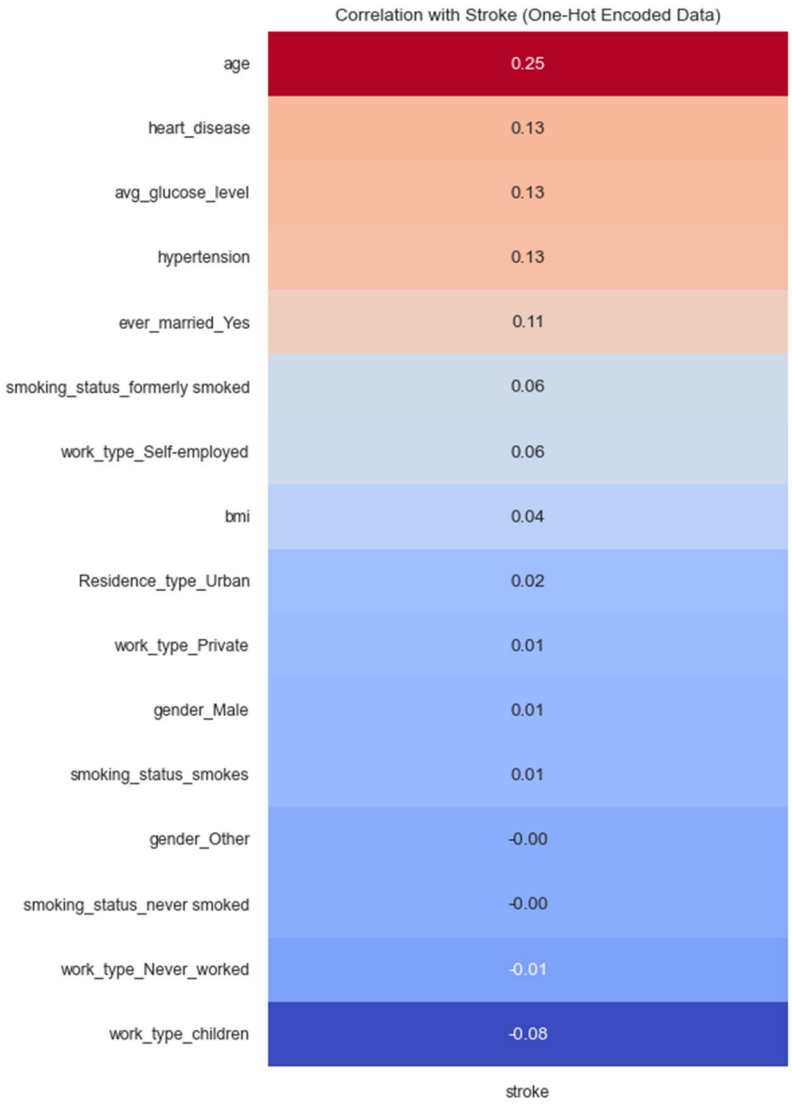

Figure 7. The visualization of correlation with stroke (Photo/Picture credit: Original).

The pair plot gives us a visual overview of the pairwise relationships and distributions of age, avg_glucose_level, and BMI, separated by stroke status. Here are some observations:

The diagonal plots show the distribution of each variable for stroke and non-stroke patients. They confirm our findings from the univariate analysis: stroke patients tend to be older and have higher glucose levels. At the same time, their BMI distribution is similar to that of non-stroke patients.

3. Machine learning models

There are 201 missing values in the bmi column, must filled them up.

For the 'avg_glucose_level' and 'bmi' columns, apply a log transformation, as it is a standard method to reduce the influence of outliers in skewed data. Before doing that, ensure no zero or negative values are in these columns, as log transformation is not defined for those.

Because the dataset is small, removing the outliers might lead to a significant loss of information. Therefore, transform the data. No 0 or -, use log transformation. The log transformation has reduced the outliers in the avg_glucose_level and bmi columns. The average_glucose_level column now has 380 outliers (down from 627), and the BMI column has 48 outliers (down from 110).

3.1. Decision tree

A decision tree serves as a supervised learning tool in the realm of classification and regression modeling. Within the domain of predictive modeling, regression plays a vital role. Decision trees are employed for the purpose of data categorization and forecasting future outcomes.

These decision trees take on the appearance of flowcharts, initiating from a root node that poses a specific data inquiry and branching into potential answers. These branches subsequently lead to internal decision nodes, posing further questions and yielding additional outcomes. This iterative process continues until the data reaches an endpoint known as an "end node" or "leaf node."

In the field of machine learning, there exist four primary methodologies for training algorithms: supervised, unsupervised, reinforcement learning, and semi-supervised knowledge. Decision trees serve as a valuable tool for visually comprehending how supervised learning algorithms generate specific outcomes.

The basic idea of a decision tree is to use a set of predictor variables to build a tree that uses decision rules to predict the value of a response variable.

1/ Fit a decision tree classifier with NO set limits on maximum depth features and leaves.

2/ Determine how many nodes are present and what the depth of this (very large) tree is.

3/ Using this tree to measure the predication error in the train and test dataset.

3.2. Random forest

Random Forest stands as a widely employed machine learning technique that amalgamates the outputs of multiple decision trees to yield a unified result. Its user-friendly nature and adaptability have led to its widespread adoption, enabling it to effectively address both classification and regression challenges.

The core of the Random Forest algorithm lies in the creation of numerous decision trees, which are then combined to enhance prediction accuracy and stability.

Two pivotal hyperparameters play a significant role in shaping the Random Forest's behavior. Firstly, there's "n_estimators," denoting the quantity of trees the algorithm constructs before determining the ultimate consensus through majority voting or averaging of predictions. Generally, a higher number of trees enhances performance by rendering predictions more robust, albeit at the cost of increased computational demands. Secondly, "max_features" serves as another crucial hyperparameter, defining the maximum number of features considered by a random forest when splitting a node.

To employ Random Forest effectively, one should start by preparing the dataset, ensuring alignment between the input features (X) and the target variable (y). Subsequently, the data is divided into a training set and a test set. Following this, a Random Forest regression model is constructed and trained using the training dataset.

4. Results and discussion

4.1. Decision tree

	precision	recall	f1-score	support
0	0.95	1.00	0.97	1457
1	0.33	0.01	0.03	76
accuracy			0.95	1533
macro avg	0.64	0.51	0.50	1533
weighted avg	0.92	0.95	0.93	1533

Figure 8. The performance of the decision tree (Photo/Picture credit: Original).

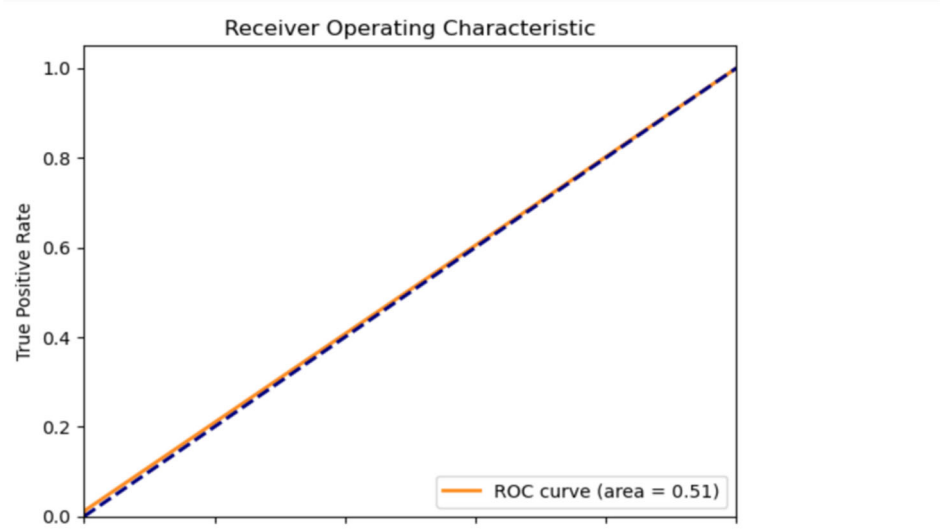

Figure 9. The ROC curve of the decision tree (Photo/Picture credit: Original).

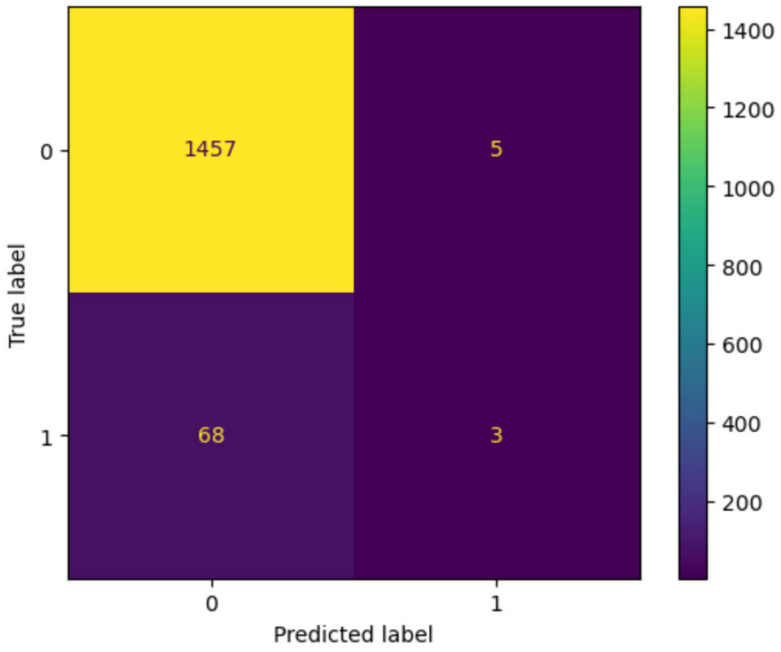

Figure 10. The confusion matrix of the decision tree (Photo/Picture credit: Original).

4.2. Random forest

	precision	recall	f1-score	support
0	0.96	0.96	0.96	1457
1	0.17	0.17	0.17	76
accuracy			0.92	1533
macro avg	0.57	0.56	0.56	1533
weighted avg	0.92	0.92	0.92	1533

Figure 11. The performance of the random forest (Photo/Picture credit: Original).

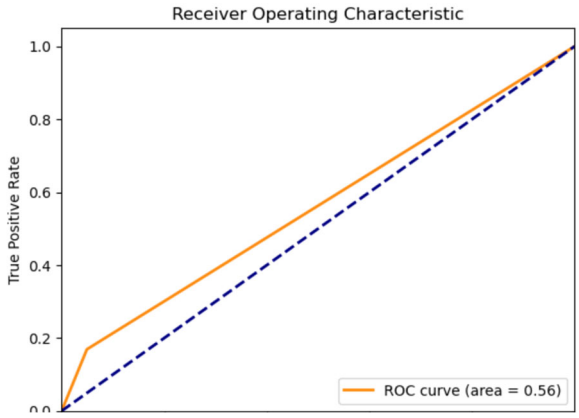

Figure 12. The ROC curve of the random forest (Photo/Picture credit: Original).

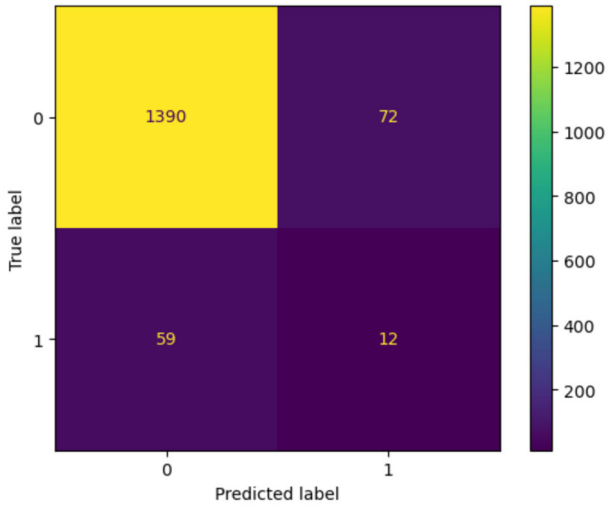

Figure 13. The confusion matrix of the random forest (Photo/Picture credit: Original).

The performance of the models is shown in Figure 8, Figure 9, Figure 10, Figure 11, Figure 12 and Figure 13. The accuracy improved from 92% to 95% when RandomForest was used instead of DecisionTree and RandomForest is good at handling large datasets, missing values, and reducing overfitting. Although a random forest model is a collection of decision trees, there are some differences. Decision trees can be affected by overfitting. In most cases, random forests can prevent this by creating random subsets of features and using those subsets to build smaller trees. Subtrees are then combined. It is important to note that this only sometimes works and may also slow down the calculation depending on how many trees the random forest forms.

5. Conclusion

This project highlights the significant potential of machine learning in medical prediction, benefiting both patients in risk assessment and healthcare practitioners in treatment planning. It employed two distinct predictive models, RandomForest and DecisionTree, and assessed their performance using evaluation metrics such as the Confusion Matrix, Receiver Operator Characteristic (ROC) curve, and Precision-Recall curve. The study also underscored the challenge of missing data in certain features, emphasizing the need for effective strategies to address this issue in medical prediction. Notably, the experimental results yielded a robust predictive capability, with an Area Under the Curve (AUC) of 95% for RandomForest and 92% for DecisionTree. Future research can focus on refining prediction models, achieving better balance, and expanding the dataset to enhance prediction accuracy.

References

[1] World stroke Organization, 2023, https://www.world-stroke.org/assets/downloads/WSO_Global_Stroke_Fact_Sheet.pdf

[2] Wolfe C D A. The impact of stroke. British medical bulletin, 2000, 56(2): 275-286.

[3] Weimar C, Roth M P, Zillessen G, et al. Complications following acute ischemic stroke. European neurology, 2002, 48(3): 133-140.

[4] Johnson W, Onuma O, Owolabi M, et al. Stroke: a global response is needed. Bulletin of the World Health Organization, 2016, 94(9): 634.

[5] Heaton J B, Polson N G, Witte J H. Deep learning for finance: deep portfolios. Applied Stochastic Models in Business and Industry, 2017, 33(1): 3-12.

[6] Culkin R, Das S R. Machine learning in finance: the case of deep learning for option pricing. Journal of Investment Management, 2017, 15(4): 92-100.

[7] Qiu Y, Wang J, Jin Z, et al. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control, 2022, 72: 103323.

[8] Monil P, Darshan P, Jecky R, et al. Customer segmentation using machine learning. International Journal for Research in Applied Science and Engineering Technology (IJRASET), 2020, 8(6): 2104-2108.

[9] Rigatti S J. Random forest. Journal of Insurance Medicine, 2017, 47(1): 31-39.

[10] Myles A J, Feudale R N, Liu Y, et al. An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society, 2004, 18(6): 275-285.