Using Machine Learning Algorithms of Stroke Prediction

Daniela Slavkovska[1], Anita Petreska[2], Blagoj Ristevski[2], Saso Nikolovski[2], Nikola Rendevski [2]

[1] Children's Hospital Skopje
[2] Faculty of Information and Communication Technologies - Bitola, University "St. Kliment Ohridski"- Bitola, Republic of North Macedonia

dslavkovska@yahoo.com;blagoj.ristevski@uklo.edu.mk;etreska.anita@uklo.edu.mk; sasnik@gmail.com; nikola.rendevski@uklo.edu.mk

Abstract:

Stroke is a severe medical condition resulting from disrupted blood flow or ruptured blood vessels in the brain, often leading to life-threatening consequences. The World Health Organization (WHO) identifies stroke as a leading cause of death and disability worldwide. Although significant research has focused on heart-related diseases, stroke prediction has received comparatively less attention. To address this gap, this paper presents machine learning models developed to predict stroke likelihood, utilizing key physiological factors associated with stroke risk. Six algorithms: logistic regression, decision tree, random forest, KNN, SVM and Naïve Baye, were implemented to train and test prediction models. The primary objective was to determine the algorithm that provides the highest predictive accuracy.

Our findings reveal that the Naïve Bayes algorithm performed best, achieving an accuracy of approximately 82%. This is notable given Naïve Bayes' suitability for probabilistic data and its efficiency in handling complex variable interactions, suggesting its value for early stroke detection in clinical settings. The use of machine learning in stroke prediction highlights a promising approach for early intervention, potentially aiding in reducing stroke-related mortality and morbidity.

This paper contributes to expanding the application of machine learning in healthcare, emphasizing the need for focused stroke prediction research. Future work could enhance these models by integrating diverse datasets, testing additional machine learning techniques, and refining predictive algorithms to boost accuracy and reliability. By advancing stroke prediction, machine learning may play a key role in mitigating stroke's impact on global health.

Keywords:

Machine Learning, Logistic Regression, Decision Tree Classification, Random Forest Classification, KNN, SVM and Naïve Bayes

1. Introduction

About 11% of all deaths worldwide are due to stroke **Error! Reference source not found.**, according to the Centers for Disease Control and Prevention, strokes occur in the United States each year in about 795,000 people.

With the advancement of medical technology, machine learning can now be used to predict stroke. It is possible to make accurate predictions and analyses using machine learning algorithms. Strokes can be predicted using machine learning algorithms.

Six different machine learning algorithms were tested, with Naive Bayes achieving the highest accuracy.

Machine learning algorithms are useful for making accurate predictions and delivering accurate analytics. Research conducted on stroke has mainly focused on predicting heart attacks. The main elements of the methods used and the results achieved show that of the five classification algorithms tested, Naïve Bayes showed the highest performance by achieving a superior accuracy metric. The limitation of this model lies in its training on textual data instead of actual real-time brain images. This paper shows how six machine learning algorithms were put into practice. This paper has the potential to be extended to include the implementation of all existing machine learning algorithms.

A database from Kaggle, which contains a range of physiological traits as its attributes, was used for this paper.

These features are subsequently examined and used for final forecasting. The data is first cleaned and prepared for machine learning model understanding. Data preprocessing is the term used for this stage. To do this, the database is examined for all missing values and then filled. Then the string values are converted to integers using Label encoding, if necessary one-time encoding is performed.

After data preprocessing, the database is divided into training and testing data. Next, various classification algorithms are used to construct a model using the updated data. Accuracy is calculated for each of these algorithms and then compared to determine the most effective model for making predictions.

2. Machine learning algorithms for stroke prediction

Machine learning has been effectively applied to predict various diseases, such as diabetes, heart disease, and numerous efforts have been made to develop stroke prediction models using various classification techniques. Six classification algorithms were tested: Naïve Bayes, SVM, K-Nearest Neighbors (KNN) 0, Random Forest (RF), Decision Tree (DT) and Logistic Regression (LR) 000. The Naïve Bayes classifier achieved the best performance, with accuracy, precision, recall and F1-score of 82%, 79.2%, 85.7% and 82.3%. Also Naïve Bayes showed the highest area under the ROC curve (AUC) 0, reaching 82%.

3. Methodology

A dataset from Kaggle was used, from all available datasets the appropriate one was selected to construct the model.

3.1. Methodology

Once the data is collected the next task is to organize it in order to improve its clarity and ensure that it is easily interpreted by machine learning algorithms. The process is known as data preprocessing.

The process includes handling missing values, managing unbalanced data, and implementing label coding that is unique to this data set.

With the pre-processed data, they are ready to construct the model. Data-driven and machine learning algorithms are essential for model creation. Logistic regression, decision tree, random forest, K-nearest neighbor, support vector machine and Naïve Bayes classifiers were used. After constructing six different models, they are evaluated using five metrics to determine their performance: accuracy score, precision score, recall score, F1 score, and receiver operating characteristic (ROC) curve.

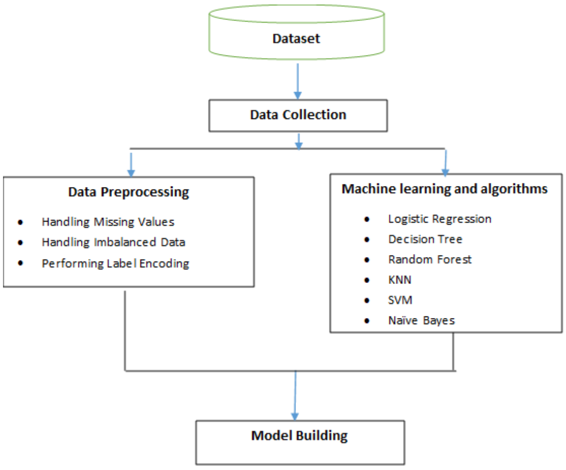

Figure 1: Methodology flow diagram

4. Implementation

The implementation of this paper is as follows:
- Data collection
- Data Preprocessing
- Label Encoding
- Handling Imbalanced Data

4.1. Data colection

The dataset used for stroke prediction was retrieved from Kaggle 0. It consists of 5110 rows and 12 columns, with the following attributes: "id", "sex", "age", "hypertension", "heart_disease", "ever_married", "job_type", "stay_type", "average_glucose level" , 'bmi', 'smoking_status' and 'stroke'. The 'stroke' column serves as an output variable, with values of '0' or '1'. A value of "0" indicates no detected risk of stroke, while "1" suggests a potential risk of stroke. This database is very unbalanced, as there are significantly more occurrences of "0" (4861 rows) compared to "1" (249 rows). To improve the accuracy, pre-processing of the data is undertaken to resolve the imbalance.

Table 1:
Stroke Dataset

Attribute name	Type (values)	Description
ID	Integer	A unique integer value for patients
Gender	Male, Female, Other	Tell the gender of the patient
Age	Integer	Age of patient
Hypertension	Integer (1 ,0)	Tall whether the patient has hypertension or not
Heart disease	Integer (1 ,0)	Tall whether the patient has heart disease or not
Ever married	String literal (Yes ,No)	Tall whether the patient is married or not
Work type	String literal (children, govt job, never worked, private, self- employed)	It gives different categories for work
Residence type	String literal (Urban ,Rural)	The patients residence type is stored
Avg glucose level	Floating point number	Gives the value of average glucose level in blood
Bmi	Floating point number	Gives the value of patients Body Mass Index
Smoking status	String literal (formerly, smoked, never smoked, smokes, unknown)	Gives the smoking status of the patient
Stroke	Integer (1 ,0)	A unique integer value for patients

4.2. Data Preprocessing

Data pre-processing is a crucial step before building a model, it helps to eliminate unwanted noise and outliers 0 in the data set, which can hinder model performance. This phase ensures that any factors that reduce the effectiveness of the model are addressed. Once the database is collected, the next step involves cleaning the data and preparing it for model development. The database used in this case contains 12 attributes, as shown in Table 1. Initially, the "id" column is discarded, as it does not contribute to the performance of the model. The data set is then examined for missing values. For example, the column 'bmi' contains null values, which are replaced by the mean value of the column. After addressing the missing data, the next step is to perform label coding.

4.3. Label encoding

Label encoding 0 transforms categorical string values into numeric values to make the data set understandable to machine learning algorithms, which typically work with numeric data. In the dataset five columns contain string values. Label encoding is applied to convert these strings to integers with a database composed entirely of numeric values.

4.4. Handling Imbalanced Data

The data used to predict stroke is an unbalanced data set. Out of 5110 rows only 249 indicate the occurrence of a stroke while 4861 indicate no stroke. This imbalance is illustrated in Figure 2. Training a model with such skewed data may yield high accuracy, but may result in poor precision. To solve this problem and improve the performance of the model, the data is balanced 0 using the undersampling technique. The sample adjusts the dataset by reducing the majority class to match the size of the minority class. In this case rows with a value of "0" are undersampled to equal 249 rows with a value of "1".

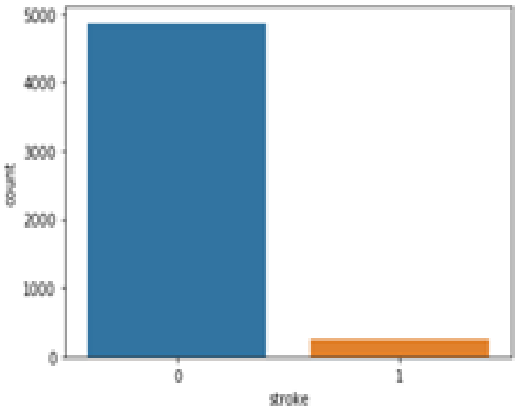

Figure 2: Before Undersampling

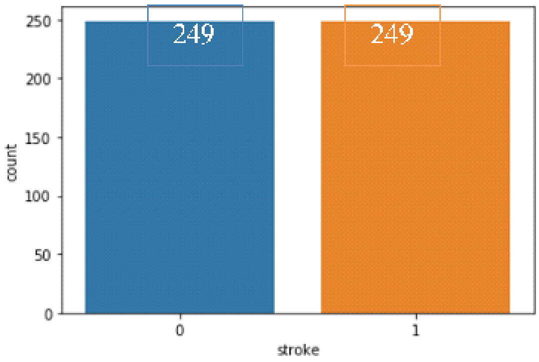

Figure 3: After Undersampling

4.5. Matrix correlation

The matrix correlation highlights the influence of different characteristics on an attribute. Figure 4 illustrates the relationship between other attributes and the stroke attribute. The graph shows that no single parameter has a dominant effect on the occurrence of stroke. The key factors that significantly affect the risk of stroke are: gender, age, hypertension, heart disease, average blood sugar levels, body mass index and smoking status. Attributes with the least impact are type of work, type of residence and marital status.

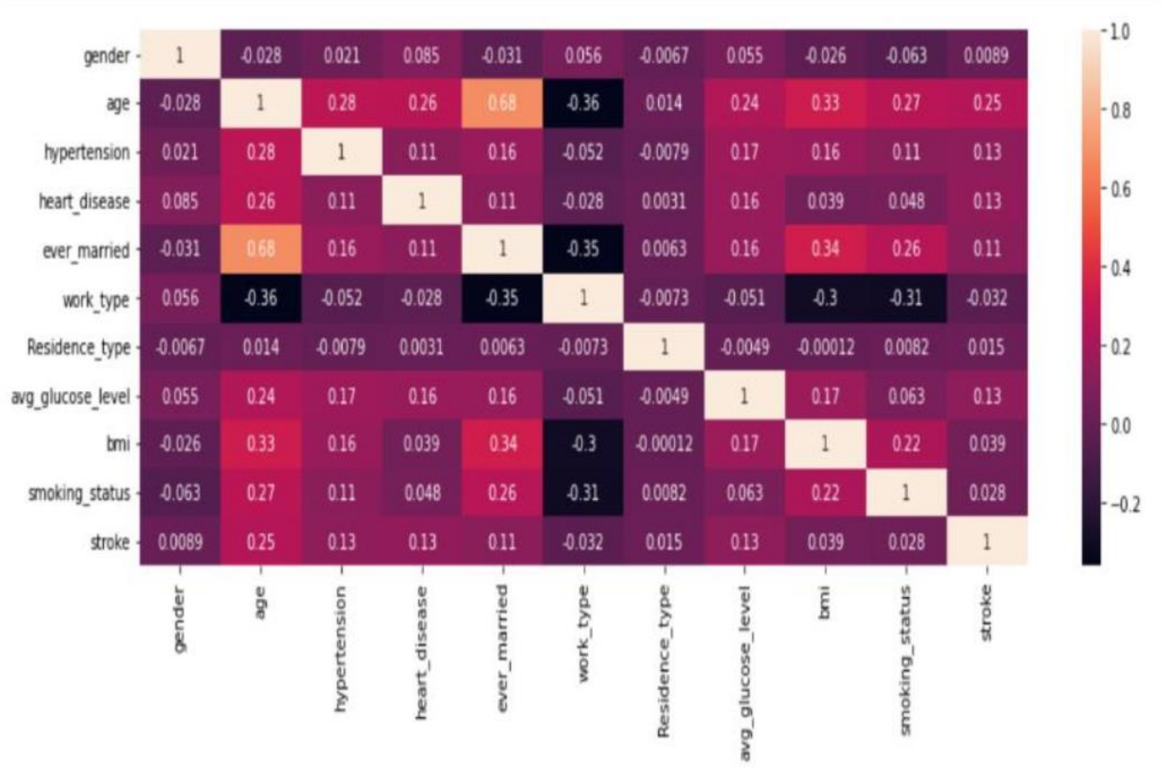

Figure 4: Matrix correlations between sociodemographics, lifestyle, and disease

5. Model building

5.1. Splitting the Data

After the data preprocessing is completed and the imbalanced database is resolved, the next step is model building. The data is split into a training dataset and a testing dataset to ensure better accuracy and efficiency. The data is split 80:20, with 80% used for training and 20% for testing. After splitting the data, several classification algorithms are applied to train the model. Algorithms used include logistic regression, decision tree classification, random forest classification, K-nearest neighbors (KNN), support vector machine (SVM), and Naïve Bayes classification.

5.2. Classification Algorithms

5.2.1. Logistic regression

Logistic regression 0 is a supervised learning algorithm commonly used to predict the probability of a binary output variable (0 or 1). Since the output attribute in this dataset is binary, logistic regression is an ideal choice. After applying this algorithm, the model achieved an accuracy of 78%. The performance of the algorithm will be further evaluated using additional metrics such as precision and recall, both yielded a result of 77.6%. The F1 score is also 77.6%. A receiver operating characteristic (ROC) curve for logistic regression 0 has a performance of 78%, as illustrated in Figure 5.

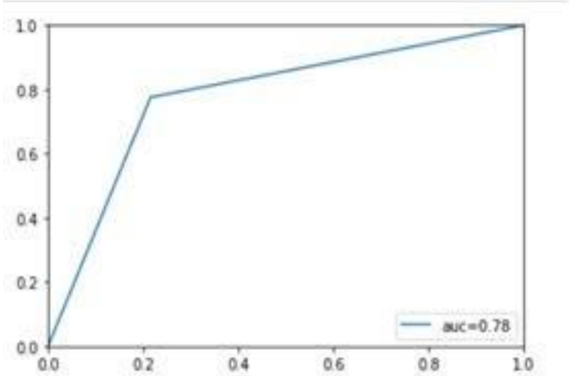

Figure 5: ROC curve for logistic regression

5.2.2. Decision tree classification

Decision tree 0 classification is a supervised learning algorithm used for both regression and classification tasks. It works by partitioning data based on specific parameters into a tree-like structure consisting of decision nodes (where the data is partitioned) and leaf nodes (which provide the result). In this stroke prediction model the Decision Tree algorithm achieved an accuracy of 66%, which is lower than the accuracy obtained with logistic regression. Like the logistic regression, the precision and recall scores are equal to 77.6%. The F1 score also matches this value with 77.6%. The ROC curve for decision tree classification 0 showed an accuracy of 66%, as shown in Figure 6.

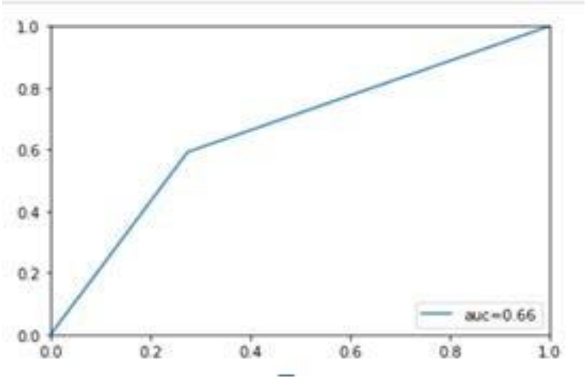

Figure 6: ROC curve for Decision Tree

5.2.3. Random Forest Classification

The next classification algorithm used is Random Forest Classification 0. A random forest is composed of multiple decision trees, each trained independently on random subsets of the data. During training, these trees are generated and each one gives a result. For the final prediction, a process called 'voting' is used, where each tree votes for an output class ('stroke' or 'no stroke'). The class with the most votes is chosen as the final prediction. The model achieved an accuracy of 73% using this algorithm. The precision and recall scores are 72% and 73.5%, the F1 score is 72.7%. A receiver operating characteristic (ROC) curve for random forest classification0 showed an accuracy of 73%, as shown in Figure 7.

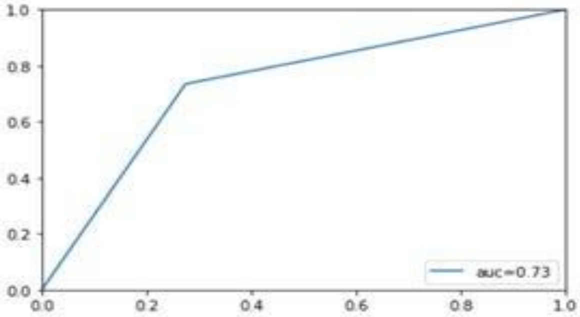

Figure 7: ROC curve for Random Forest

5.2.4. K-Nearest Neighbours Classification

Another classification algorithm applied is K-Nearest Neighbors (KNN) 0, a supervised learning technique. KNN is a lazy algorithm, meaning that it does not perform training immediately after obtaining the dataset. Instead, it stores the database and acts on it only during classification. The algorithm works by finding similarities between new data and existing data, and then assigning the new data to the category most similar to the existing categories. This algorithm achieved an accuracy of 80%. Precision and recall scores were 77.4% and 83.7%, respectively. The F1 score was 80.4%. The ROC curve for KNN 0 showed an accuracy of 80%, as illustrated in Figure 8.

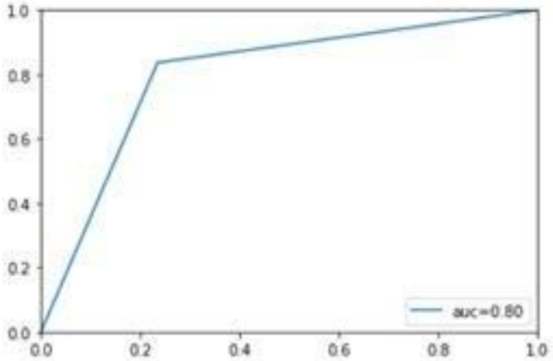

Figure 8: ROC curve for KNN

5.2.5. Supper Vector Machine (SVM)

Support Vector Machine 0 is a supervised learning algorithm used for both classification and regression tasks. SVM is particularly effective for high-dimensional data. In this case, the algorithm achieved an accuracy of 80%. The precision and recall scores were 78.6% and 83.8%, respectively, resulting in an F1 score of 81.1%. A receiver operating characteristic (ROC) curve for SVM 0 indicates a performance of 80%, as shown in Figure 9.

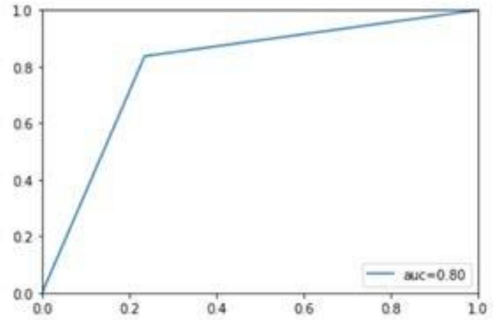

Figure 9: ROC curve for SVM.

5.2.6. Naïve Bayes classification

Naïve Bayes 0 is a supervised learning technique that works on the assumption that each feature is independent of the others, based on Bayes' theorem. The Naïve Bayes classifier assumes that the presence of any feature in a class is unrelated to the presence of other features. This algorithm achieved an accuracy of 82%. Precision and recall scores were 79.2% and 85.7%, respectively, with an F1 score of 82.3%. The ROC curve for Naïve Bayes 0 classification showed a performance of 82%, as shown in Figure 10.

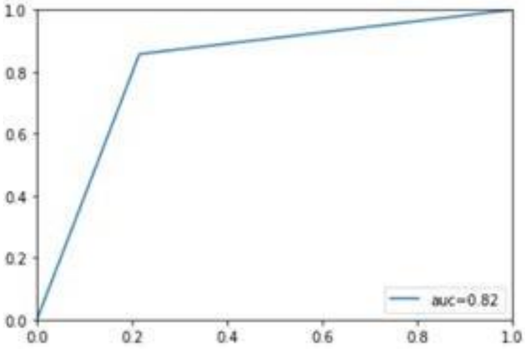

Figure 10: ROC curve for Naïve Bayes classification

After building the models, it is concluded that Naïve Bayes outperformed the other algorithms.

6. Conclusion

A stroke is a serious medical condition that requires prompt intervention to prevent further complications. The development of a machine learning model could aid in the early detection of stroke, helping to mitigate its long-term effects. This paper evaluates the performance of several machine learning algorithms in stroke prediction based on various physiological attributes. Among the algorithms tested, Naïve Bayes Classification performed best, achieving an accuracy of 82%. The comparison of accuracy of different algorithms is illustrated in Figure 11. In terms of precision, recall and F1 scores, Naive Bayes also outperformed the others. These comparisons are presented in Figure 12, Figure 13 and Figure 14.

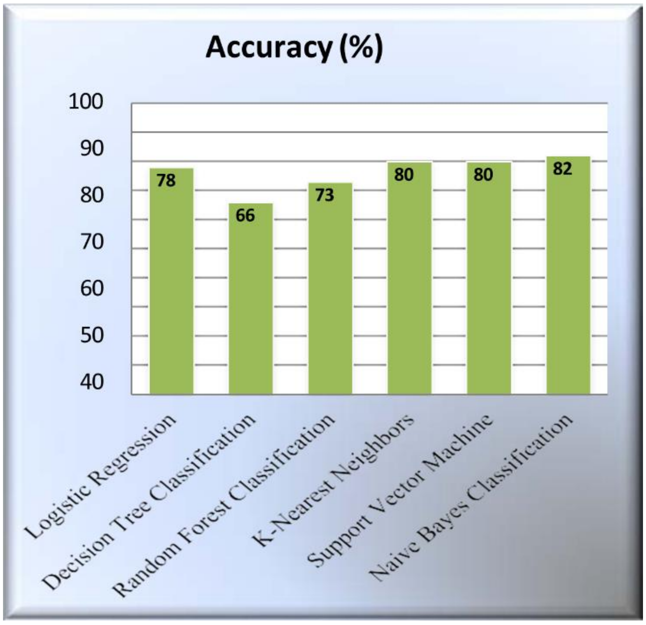

Figure 11: Comparing the Accuracies of ML Algorithms

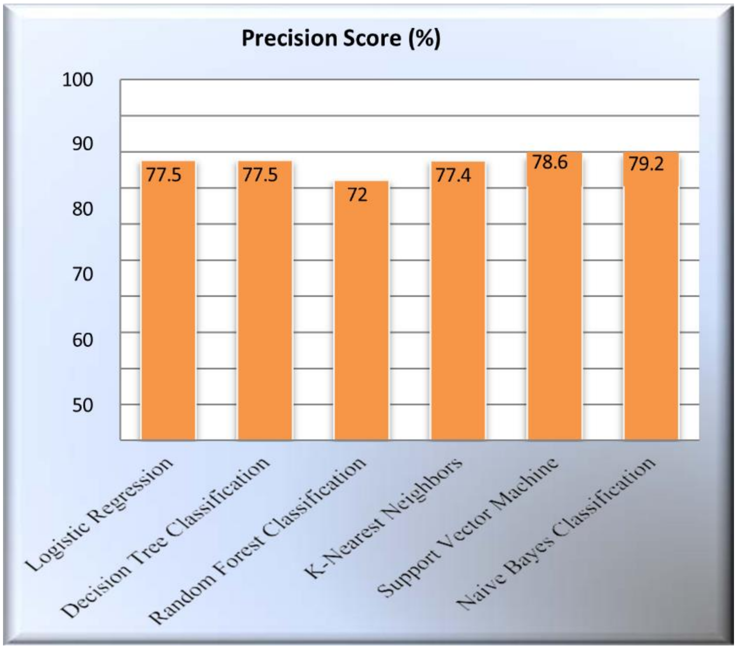

Figure 12: Comparing the Precision Scores of ML Algorithms

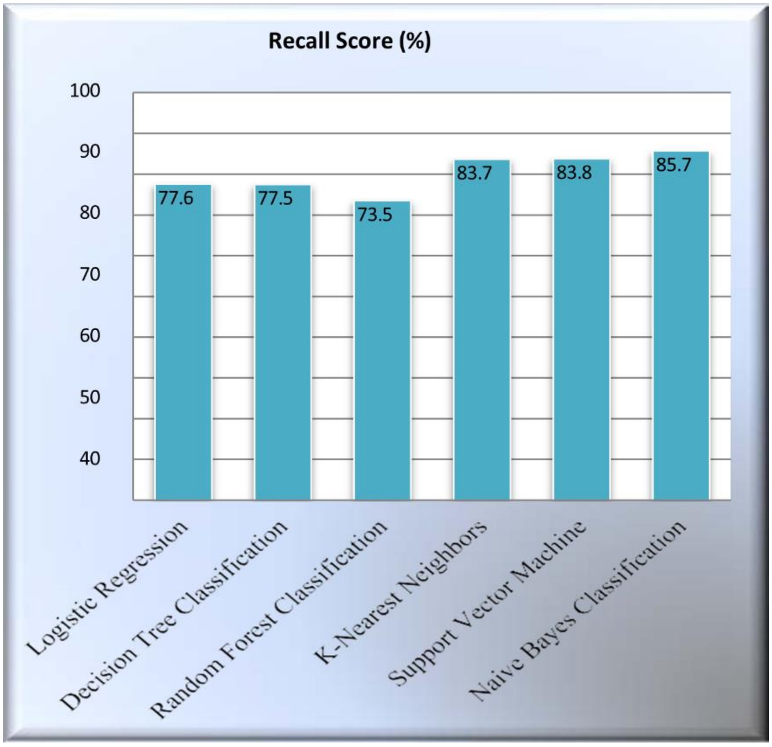

Figure 13: Comparing the Recall Scores of ML Algorithms

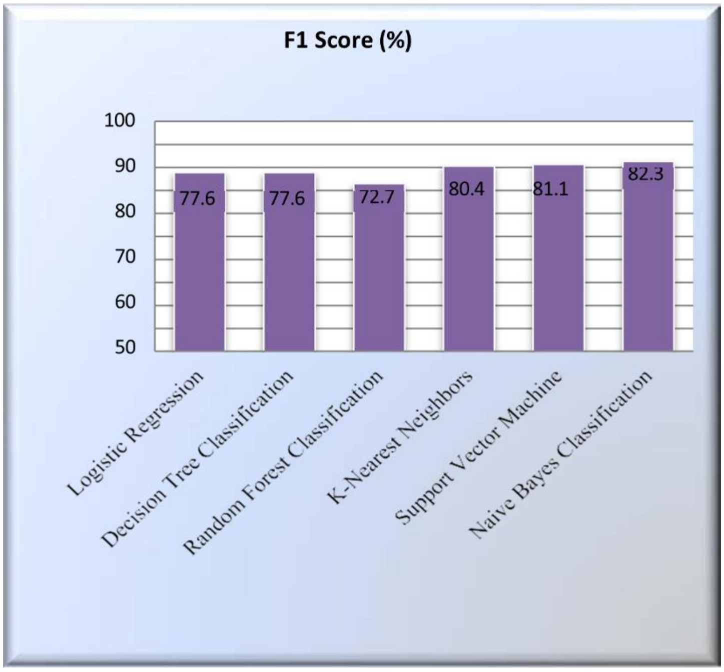

Figure 14: Comparing the F1 Scores of ML Algorithms

This paper investigates the implementation of various machine learning algorithms on the given dataset. Future work could extend this project by incorporating neural networks to train the model, allowing for a more comprehensive performance comparison using additional accuracy metrics. Currently, this study focuses on textual data, which may not always provide the most accurate predictions of stroke. A more effective approach could involve using datasets that include medical images, such as CT scans of the brain, to improve the accuracy of predicting stroke in the future.

References:

[1] Olajide, A. O. "Life after stroke: more than a survival of the fittest." (2021).

[2] Nugroho, Syamsul. *PERBANDINGAN METODE FUZZY K-NEAREST NEIGHBOR DAN NEIGHBOR WEIGHTED K-NEAREST NEIGHBOR UNTUK DETEKSI PENYAKIT STROKE*. Diss. University of Technology Yogyakarta, 2020.

[3] Pradeepa, S., Manjula, K. R., Vimal, S., Khan, M. S., Chilamkurti, N., & Luhach, A. K. (2023). DRFS: detecting risk factor of stroke disease from social media using machine learning techniques. *Neural Processing Letters*, 1-19.

[4] Singh, M. Sheetal, Prakash Choudhary, and Khelchandra Thongam. "A comparative analysis for various stroke prediction techniques." *Computer Vision and Image Processing: 4th International Conference, CVIP 2019, Jaipur, India, September 27–29, 2019, Revised Selected Papers, Part II 4*. Springer Singapore, 2020.

[5] Bandi, V., Bhattacharyya, D., & Midhunchakkravarthy, D. (2020). Prediction of Brain Stroke Severity Using Machine Learning. *Rev. d'Intelligence Artif.*, *34*(6), 753-761.

[6] Van Os, H. J., Ramos, L. A., Hilbert, A., Van Leeuwen, M., Van Walderveen, M. A., Kruyt, N. D., ... & Mr Clean Registry Investigators. (2018). Predicting outcome of endovascular treatment for acute ischemic stroke: potential value of machine learning algorithms. *Frontiers in neurology*, *9*, 784.

[7] Dataset named 'Stroke Prediction Dataset' from Kaggle: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.

[8] Ma'rifah, H., Wibawa, A. P., & Akbar, M. I. (2020). Klasifikasi artikel ilmiah dengan berbagai skenario preprocessing. *Ekonomi Bisnis*, *29*, 23-01.

[9] Analytics Vidhya. 2020. *Categorical Encoding One Hot Encoding vs Label Encoding*. Available at: https://www.analyticsvidhya.com/blog/2020/03/one-hot-encodingvshttps://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn/labelencoding-using-scikit-learn/.

[10] Rizal, A. A., & Soraya, S. (2018). Multi time steps prediction dengan recurrent neural network long short term memory. *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, *18*(1), 115-124.

[11] Documentation for Logistic Regression from Scikit-learn.org.

[12] Grigoryev, S. G., Lobzin, Y. V., & Skripchenko, N. V. (2016). The role and place of logistic regression and ROC analysis in solving medical diagnostic task. Journal Infectology, 8(4), 36-45.

[13] Documentation for Decision Tree Classification from Scikit-learn.org.

[14] Wu, Y., Xia, Z., Feng, Z., Huang, M., Liu, H., & Zhang, Y. (2024). Forecasting Heart Disease Risk with a Stacking-Based Ensemble Machine Learning Method. *Electronics*, *13*(20), 3996.

[15] Documentation for Random Forest Classification from Scikit-learn.org.

[16] Shimizu, G. Y., Schrempf, M., Romão, E. A., Jauk, S., Kramer, D., Rainer, P. P., ... & de Azevedo-Marques, P. M. (2024). Machine learning-based risk prediction for major adverse cardiovascular events in a Brazilian hospital: Development, external validation, and interpretability. *PloS one*, *19*(10), e0311719.

[17] Documentation for K-Nearest Neighbor from Scikit-learn.org.

[18] Arshad, H. (2024). The Wine Quality Prediction Using Machine Learning. *Journal of Innovative Computing and Emerging Technologies*, *4*(2). Documentation for Support Vector Machine from Scikit-learn.org.

[19] Documentation for Support Vector Machine from Scikit-learn.org.

[20] Feng, Qiya. "Investigation Based on Machine Learning Algorithms." 342.

[21] Documentation for Naïve Bayes Classification Algorithm from Scikit- learn.org.

[22] SALIHU, S. A., OWOYEMI, O. A., & SALIU, K. B. (2024). Performance Analysis of Some Machine Learning Algorithms in Prediction of Heart Disease. In *Conference Organising Committee* (p. 169).