International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 05 May 2024, pp: 972-980 www.ijaem.net ISSN: 2395-5252

# A Comparative Analysis of Machine Learning Models for Early Stroke Prediction

Emir Beba[1], Dželila Mehanović [1], Selma Vreto [1], Adnan Dželihodžić [2]

[1]*International Burch University, Faculty of Engineering, Natural and Medical Sciences, Sarajevo, Bosnia and Herzegovina.*
[2]*University of Zenica, Polytechnic Faculty, Zenica, Bosnia and Herzegovina.*
*Corresponding Author: Emir Beba*

---

Date of Submission: 18-05-2024 Date of Acceptance: 28-05-2024

---

**ABSTRACT**: To handle classification issues, this study offers a thorough examination of machine learning models, such as Random Forest, Support Vector Machines (SVM), and Logistic Regression, across various feature selection techniques. The primary focus lies on comparing model performances using accuracy and F1 Score metrics, particularly under conditions of class imbalance, which isa common challenge in practical machine learning applications.

Through a detailed analysis, this paper investigates the impact of feature selection techniques such as Mutual Information, ANOVA F-value, and Model-based selection using Random Forest on the predictive capabilities of each model.

The findings reveal that while Random Forest generally offers superior accuracy, peaking at 0.9207 with Mutual Information and F Classif methods, its F1 Scores suggest room for improvement in identifying the minority class. Conversely, Logistic Regression and SVM exhibit consistent accuracy, with Logistic Regression achieving its highest F1 Score of 0.2889 with Random Forest feature selection. These nuanced performance variations captured by the F1 Score highlight the complex interplay between model selection, feature selection, and performance evaluation metrics. The study emphasizes the importance of considering both accuracy and F1 Score in model evaluation to ensure a balanced assessment of model performance, especially in imbalanced datasets.

The results advocate for a strategic approach to model and feature selection, tailored to specific dataset characteristics, to enhance model performance and reliability in predicting outcomes across diverse applications.

**KEYWORDS:**Stroke prediction, Machine learning algorithms, Dataset, Features.

## I. INTRODUCTION

A brain clot or ruptured blood artery in the brain interrupts blood flow to a portion of the brain, resulting in a stroke, also known as a brain attack. Certain brain regions suffer harm or even die in both situations [1]. Stroke is one of the leading causes of death and disability. It is a medical emergency and life-threatening situation, so urgent treatment is crucial. Early treatment can reduce brain damage and many other complications [2].

According to the World Health Organization, each year, 15 million individuals globally experience a stroke. Among these cases, 5 million result in fatality, while an additional 5 million individuals are left with enduring disabilities, creating challenges for their families and communities [19]. Stroke, particularly prevalent among individuals over 55 years of age, poses a significant threat, leading not only to substantial healthcare expenses but also to long-term disabilities and fatalities.

The focus on stroke prediction was primarily prompted by the imperative to provide individuals with insights into their health status, potentially warning them about necessary habit changes or lifestyle modifications.

The aspiration to prevent potential strokes and decrease associated mortality and disability rates became the fundamental drive of this work. By creating a prediction model that might greatly improve preventative treatment and improve the quality of life for those who are at risk of stroke, the project sought to address this important healthcare issue.Computers can recognize patterns and forecast outcomes from data without explicit programming thanks to a subfield of artificial intelligence called machine learning. By evaluating patient data to identify risk variables and estimate the chance of stroke development, it is widely used in healthcare settings, including stroke prediction models [32].

The ultimate goal of the research was to create a prediction model that will be essential in spotting possible stroke cases and allowing for prompt preventative actions to lessen the effects and severity of this potentially fatal illness. Furthermore, the research aimed to contribute to the advancement of medical science by leveraging machine learning techniques to enhance stroke prediction accuracy and enable proactive intervention strategies in clinical settings.

The rest of the work is organized as follows: literature review gives an overview of related literature, methodology section gives an overview of data collection and research instruments, data and findings section gives an overview of the dataset that will be used. After all of that, comes conclusion, where we conclude our work.

## II. LITERATURE REVIEW

There are research works that employ machine learning algorithms for stroke prediction. In this section we give an overview of some of those.

Sailasya and Kumari (2021) embarked on predicting stroke risks by employing six machine learning algorithms. Notably, Naïve Bayes classification emerged as the standout performer, boasting an accuracy of approximately 82% [10]. Harshitha and Gunjan Gupta in 2021, utilized five machine learning algorithms, spotlighting Random Forest with the highest accuracy at 95.5% [11].

Also in 2021, Geethanjali and Divyashree harnessed three algorithms: Support Vector Machine, Logistic Regression, and Decision Tree Classifier revealing that Logistic Regression and Support Vector Machine outperformed the Decision Tree Classifier with an accuracy of 95.49% [12].

Lastly, Monirul Islam (2021) demonstrated the supremacy of the Random Forest classifier, providing the highest accuracy at 96% for stroke prediction among the explored machine learning models [13].

Michael Wiryaseputra (2022) delved into the domain, scrutinizing four machine learning models: Decision Tree, Random Forest, XGBoost, and Logistic Regression. His findings underscore the remarkable accuracy of the Random Forest algorithm at 99.27% in stroke prediction [14].

Additionally, Neha Saxena (2022) explored five machine learning models, emphasizing the supremacy of Random Forest (98.56%) over Logistic Regression (76.96%) for stroke prediction [15].

Elias Dritsas and Maria Trigka (2022) comprehensively examined multiple ML algorithms, highlighting the stacking classification model's high predictive capability, exhibiting an AUC of 98.9% and accuracy of 98% [16]. Ghanipour and Soroush (2022) highlighted the significance of oversampling in stroke prediction, establishing correlations between high age, high average glucose level, or high BMI with stroke occurrence [17].

In 2023, Mohammed Guhdar optimized stroke prediction accuracy, demonstrating an 86% accuracy, outperforming similar models using logistic regression [18].

This is a tabular representation as an overview of the research works, as summarized in Table 1:

**Table1.**Overview of research works.

| Authors | Year | Algorithms | Results – Accuracy |
|---|---|---|---|
| Sailasya and Kumari | 2021 | Naïve Bayes | 82% |
| Harshitha and Gunjan Gupta | 2021 | Random forest | 95.5% |
| Geethanjali and Divyashree | 2021 | SVM, Logistic regression, Decision Tree classifier | LR & SVM: 95.49% |
| Monirul Islam | 2021 | Random forest | 96% |
| Michael Wiryaseputra | 2022 | Decision Tree, Random forest, XGBoost, Logistic regression | LR: 99.27% |
| Neha Saxena | 2022 | Random forest<br>Logistic regression | RF: 98.56%<br>LR: 76.96% |

| | | | |
|---|---|---|---|
| Elias Dritsas and Maria Trigka | 2022 | Stacking classification model | 98% |
| Ghanipour and Soroush | 2022 | N/A | Highlighted attribute correlations with stroke occurrence |
| Mohammed Guhdar | 2023 | Logistic regression | 86% |

## III. DATASET

The research draws upon a dataset available on Kaggle, consisting of a CSV file named "healthcare-dataset-stroke-data.csv."[3].

This dataset encompasses a diverse range of data types, including integers, booleans, strings, and floating-point values.

The dataset incorporates twelve distinct features relevant to stroke prediction, providing essential information on patient attributes crucial for risk assessment and predictive modeling, as summarized in Table 2:

**Table2.**Overview of dataset features relevant to stroke prediction.

| **Features** | **Values** | **Description** |
|---|---|---|
| id | Integer | Unique identifier |
| gender | 0=Male,<br>1 = Female,<br>2=Other | Patient'sGender |
| age | Integer | Patient'sAge |
| hypertension | Hypertension = ’1’<br>No Hypertension = ’0’ | Does the patient have high blood pressure? |
| heart_disease | No Heart Disease = ‘0’<br>Heart Disease = ‘1’ | Does the patient have any heart disease? |
| ever_married | It is represented by True or False | Has the patient ever been married? |
| work_type | Children = ‘0’<br>Government Job = ’1’<br>Never Worked = ’2’<br>Private = ’3’<br>Self Employed = ’4’ | Patient’s work type |
| residence_type | Rural = ’0’<br>Urban = ‘1’ | Patient’s residence type |
| avg_glucose_level | It is represented in Numeric | Average glucose level in blood |
| bmi | It is represented in Numeric | Body mass index |
| smoking_status | Formerly Smoked = ’0’<br>Never Smoked = ’1’<br>Smokes = ’2’<br>Unknown = ’3’ (No Information) | Patient’s smoking status |
| stroke | No stroke = ‘0’<br>Stroke = ‘1’ | Has the patient had a stroke before? |

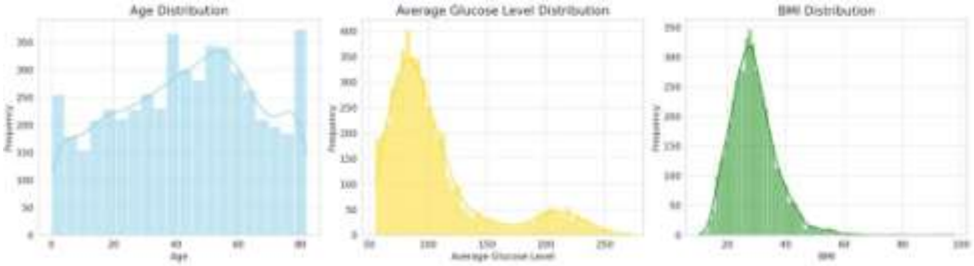

**Fig.1.**Distribution of Key Features: Age, Average Glucose Level, and BMI.

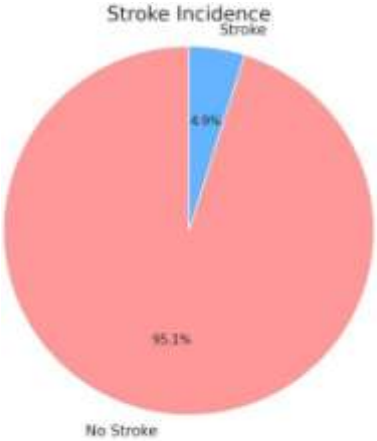

**Fig.2.**Stroke Incidence Distribution.

Column 'id' functions solely as a unique identifier and does not contribute to stroke prediction. Hence, it will be omitted from further examination.

Age distribution ranges from 0.08 to 82 years old, with a mean age of approximately

43.23 years. This suggests a wide range of ages among participants, from infants to elderly adults.

Hypertension and Heart Disease: About 9.75% of participants have hypertension, and 5.40% have heart disease. These conditions are relatively uncommon in the dataset. The average glucose level ranges from 55.12 to 271.74, with a mean of 106.15. This indicates a broad range of glucose levels among participants.

BMI values range from 10.3 to 97.6, with a mean of 28.89. This suggests a wide variation in body mass index among the participants, including underweight, normal, overweight, and obese individuals.

Approximately 4.87% of the participants have had a stroke, while the vast majority, 95.13%, have not. This distribution aligns with the expected rarity of stroke occurrences among the general population.

Given the imbalance in stroke incidence (with a much smaller proportion of stroke cases), care must be taken when analyzing the data or building predictive models to ensure that results are not biased towards the majority class.

## IV. METHODOLOGY

In this paper, the data described in the previous section were used. The process of the experiment begins with the loading of that data, then we approach the preprocessing of the data. In accordance with the analysis of the dataset, it is necessary to fill in the missing values. Also, a conversion of the categorical variable was performed using the get_dummies method [30], which generates dummy variables for each unique categorical attribute, enabling their inclusion in modeling. This transformation facilitates better understanding and interpretation of the data, contributing to their suitability for analysis and modeling.

Since this is a model that will use historical data, that is, supervised learning, it is necessary to separate the data set into training and test data. In this case, the 80/20 ratio is used, which means that 20 percent of the data will be used for testing using a random distribution of data. Features were then standardized to have a mean of 0 and a standard deviation of 1, given that classifiers are very sensitive to unscaled data.

In the analysis of the dataset, we found that the attribute used as a class or label is unbalanced where more than 95% of rows have one class. In order to solve the problem of the unbalanced test, the SMOTE technique (Synthetic Minority Oversampling Technique) [31] was applied, which creates synthetic samples of the minority class. The prepared data were then used as data for testing and evaluating three different models,i.e. classifiers: Logistic Regression, Support Vector Machines (SVM) and Random Forest.

First, the model was evaluated on the data without previously applied feature selection.

Then different methods of attribute selection were applied: mutual information, ANOVA f-value and Model-based selection using Random Forest.

The reason why we did the balancing of the dataset is that we can apply the accuracy metric. Accuracy [22] represents the proportion of correctly predicted observations (both true positives and true negatives) to the total observations in the dataset. It's one of the most intuitive and straightforward metrics for evaluating classification models.

F1 Score [23], another statistic, is employed. The F1 score offers a balance between recall and precision by taking the harmonic mean of the two. Recall, sometimes referred to as sensitivity, calculates the ratio of correctly predicted positive observations to all observations in the actual class, whereas precision is the ratio of correctly predicted positive observations to the total predicted positives.

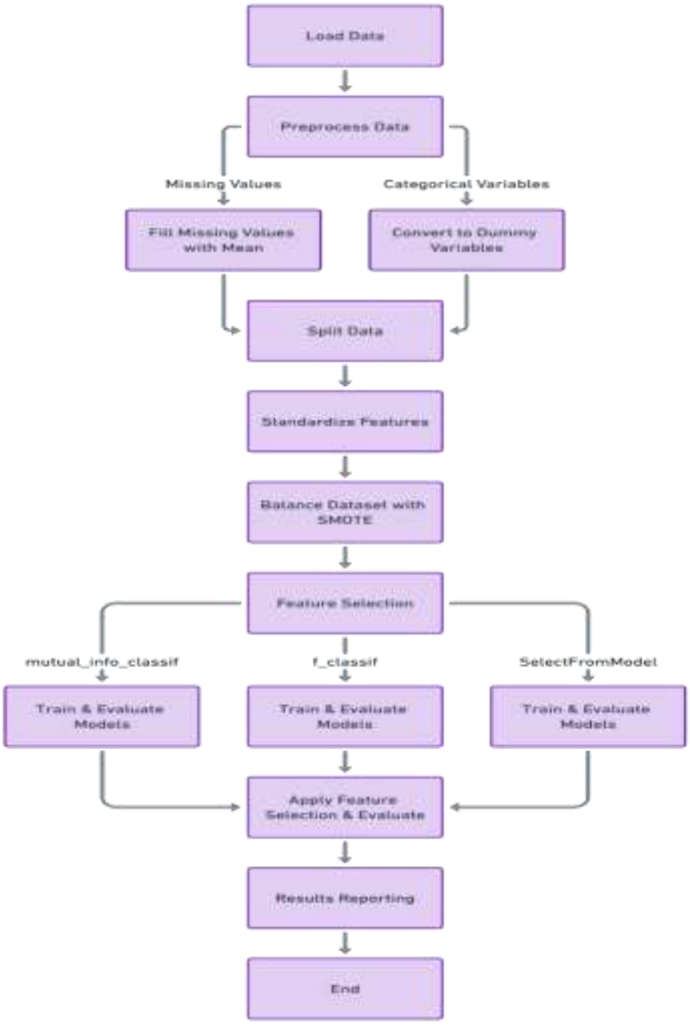

**Fig.3.**Aprocess of developing machine learning model for stroke prediction.

## V. FEATURE SELECTION AND CLASSIFIERS

Three different methods for feature selection, each leveraging a distinct approach to identify the most relevant features for the modeling process. These methods are:

1. Mutual Information - Measures how much one attribute informs us about the target variable. A higher value indicates a stronger connection, suggesting that the attribute might be more important for prediction. This method is good for detecting both linear and nonlinear relationships [24].
2. ANOVA F-value - Checks if there's a statistically significant difference between the mean values of the target variable for different values of an attribute. High F-values suggest that the attribute has a significant impact on the target variable. This method is particularly useful when the relationships between attributes and the target variable are linear [25].
3. Model-based Selection (SelectFromModel with RandomForestClassifier) Uses a machine learning model (in this case, Random Forest) to assess the importance of each attribute. Attributes that contribute most to the model's accuracy are considered important. This method is flexible as it can capture both linear and nonlinear relationships, depending on the chosen model. In short, each of these methods has its advantages and can be used in different scenarios: Mutual Information for all types of relationships, ANOVA F-value for linear relationships, and Model-based Selection for a flexible approach based on the model. In this paper, we included each of them in a model to try to improve performance of classification [26].

The selection of these methods was informed by their ability to handle both linear and nonlinear data structures, ensuring adequacy for our case. Furthermore, prior research and empirical evidence support the applicability of these methods in similar contexts, providing a robust foundation for their inclusion in our study.

In this paper, based on literature review, three classification models were employed. Description: Logistic Regression is a statistical method for predicting binary outcomes based on one or more independent variables. It estimates probabilities using a logistic function, which is particularly useful for binary classification tasks (e.g., spam or not spam). Despite its name, Logistic Regression is used for classification rather than regression tasks. Logistic Regression is straightforward to implement and can be trained efficiently. As a result, it's a natural place to start when the goal is to classify some data using a method that is both simple and interpretable [27].

Support Vector Machines are a powerful class of machine learning algorithms used for classification, regression, and outlier detection. As a supervised algorithm, it can solve both linear and nonlinear problems and work well for many practical problems. SVMs are often used in the finance industry due to how well they adapt to time series data, because it is helpful to denote linear regressions that separate data into up and down markets [28].

In order to do classification, regression, and other tasks involving ensemble learning, RandomForest builds a large number of decision trees during training. It then outputs the class that represents the mean prediction (regression) or mode of the classes (classification) of each individual tree. The tendency of decision trees to overfit to their training set is typically corrected by random forests.

"Overfitting" refers to when a model learns to perform well not just on known data (training) but on unknown data (testing) as well. "Trees" are what the RandomForests create. They're used to classify inputs into outputs. By creating many trees in the model and then having all the trees vote on the output, we are able to correct for overfitting and handle large datasets with higher dimensionality. Random forests can also rank the importance of variables in a classification [29].

## VI. RESULTS AND FINDINGS

The table compares three machine learning models - Logistic Regression, SVM (Support Vector Machine) and Random Forest across four feature selection techniques: None (no feature selection), Mutual Information, F Classif, and Random Forest. Performance metrics include Accuracy, indicating the overall proportion of correct predictions, and F1 Score, reflecting the balance between precision and recall, particularly important in scenarios with imbalanced classes.

**Table3.**Key metrics value per each algorithm and feature selection method.

| **FeatureSelection** | **Model** | **Accuracy** | **F1Score** |
|---|---|---|---|
| None | LogisticRegression | 0.7554 | 0.2857 |
| None | SVM | 0.8200 | 0.2137 |
| None | RandomForest | 0.9178 | 0.1064 |
| MutualInfo | LogisticRegression | 0.7554 | 0.2857 |
| MutualInfo | SVM | 0.8200 | 0.2137 |
| MutualInfo | RandomForest | 0.9207 | 0.1474 |
| FClassif | LogisticRegression | 0.7554 | 0.2857 |
| FClassif | SVM | 0.8200 | 0.2137 |
| FClassif | RandomForest | 0.9207 | 0.1099 |
| RandomForest | LogisticRegression | 0.7495 | 0.2889 |
| RandomForest | SVM | 0.8170 | 0.1834 |
| RandomForest | RandomForest | 0.9070 | 0.1284 |

After excluding the 'id' column, which solely functions as a unique identifier and does not contribute to stroke prediction, all other columns were utilized for analysis.

The Random Forest model consistently shows high accuracy across all feature selection methods, peaking at 0.9207 with Mutual Information and F Classif methods. This underscores Random Forest's robustness and its ability to handle both balanced and imbalanced datasets effectively. However, its F1 Scores, though higher in some instances (0.1474 with Mutual Information), suggest room for improvement in identifying the minority class. Stable Performance of Logistic Regression and SVM: Both models exhibit consistent accuracy across different feature selection methods, with SVM slightly outperforming Logistic Regression in this metric.

However, their F1 Scores, particularly for Logistic Regression, which peaks at 0.2889 with Random Forest feature selection, indicate variability in performance when it comes to classifying the minority class accurately. The feature selection method does not significantly affect the accuracy for any model but has a varied impact on the F1 Score. This suggests that while feature selection might not influence the overall predictive capability of a model, it can affect its ability to balance precision and recall. The choice of model appears to be crucial, with Random Forest generally offering the best accuracy.

However, when considering F1 Score, the differences among models become more nuanced, suggesting no one-size-fits-all solution. The lack of a clear pattern in performance improvement with different feature selection methods indicates that the effectiveness of these techniques may depend on the specific characteristics of the dataset and the nature of the classification problem. The varying F1 Scores across models and feature selection methods highlight the challenge of achieving high precision and recall simultaneously, especially in imbalanced datasets. This suggests a need for models and techniques that can better identify minority classes.

This analysis underscores the importance of evaluating multiple models and feature selection methods to identify the most effective combination for a given dataset. It also highlights the necessity of considering both accuracy and F1 Score to fully understand a model's performance, particularly in imbalanced classification scenarios.

Future work could explore more advanced ensemble methods, hyperparameter tuning, and alternative feature selection techniques to further optimize model performance.

## VII. CONCLUSION

In conclusion, this article delves into the intricacies of model selection, feature selection methods, and the balancing act between different performance metrics in the realm of machine learning. Through the comparative analysis of Logistic Regression, SVM and Random Forest across various feature selection techniques, we've uncovered valuable insights into their relative strengths and limitations.

The investigation revealed that while Random Forest consistently exhibits high accuracy, its performance, as measured by the F1 Score, suggests that there is still room for improvement in precisely identifying instances of the minority class. This observation underscores the critical importance of not only considering overall accuracy but also paying close attention to metrics like the F1 Score, especially in scenarios characterized by class imbalances.

Furthermore, the study highlights that the choice of feature selection method, while not drastically influencing model accuracy, can significantly impact the F1 Score. This finding suggests that the right feature selection technique can enhance a model's ability to achieve a harmonious balance between precision and recall, a crucial aspect in the effective classification of imbalanced datasets.

In essence, this article underscores the multifaceted nature of model evaluation, advocating for a holistic approach that considers a range of metrics and methodologies. As the field of machine learning continues to evolve, so too will our strategies for model selection and evaluation, guided by the overarching goal of developing robust, accurate, and equitable predictive models.

## REFERENCES

[1]. Centers for Disease Control and Prevention, 2024, “About Stroke”NHS, 2022,”Stroke”

[2]. Fedesoriano, 2021, “Stroke Prediction Dataset” Kaggle

[3]. National Heart, Lung, and Blood Institute, 2023, “Causes and Risk Factors”

[4]. Center for Disease Control and Prevention, 2023, “Know Your Risk for Strok”

[5]. Moira K. Kapral, Peter C. Austin, Geerthana Jeyakumar, Ruth Hall, Anna Chu, Anam M. Khan, Albert Y. jin, Cally Martin, Doug Manuel, Frank L. Silver, Richard H. Swartz, Jack V. Tu, 2019, “Rural-Urban Differences in Stroke Risk Factors, Incidence, and Mortality in People With and Without Prior Stroke”,(DOI:10.1161/CIRCOUTCOMES.118.004973)

[6]. Sandeep K. Dhaliwal, David C. Dugdale, Brenda Conaway, 2023, “Estimated average glucose (eAG)”, MedlinePlus

[7]. Heidi Moawad, Huma Sheikh, 2022, “Being Overweight Doubles Your Chances of Having a Stroke”

[8]. Johns Hopkins Medicine, “What is a Stroke?”

[9]. Gangavarapu Sailasya; and Gorli L Aruna Kumari, 2021, “Analyzing the Performance of Stroke Prediction using ML Classification Algorithms”, International Jounal of Advanced Computer Science and Applications (IJACSA), DOI: 10.14569/IJACSA.2021.0120662

[10]. Harshitha K V, Harshitha P, Gunjan Gupta, Vaishak P, Prajna K B, 2021, “Stroke Prediction Using Machine Learning Algorithms”

[11]. T M Geethanjali, Divyashree M D, Monisha S K, Sahana M K, 2021, “Stroke Prediction using Machine Learning”, No.d710-d717

[12]. Md. Monirul Islam, Sharmin Akter, Md. Rokunojjaman, Jahid Hasan Rony, Al Amin, Susmita Kar, 2021, “Stroke Prediction Analysis using Machine Learning Classifiers and Feature Technique”

[13]. Michael Wiryaseputra, 2022, “Stroke Prediction Using Machine Learning Classification Algorithm”, ISSN 2229-5518

[14]. Deep Singh Bhamra, Arvind Choudhary, Preet Maru, Neha Saxena, 2022, “BrainOK – Brain Stroke Prediction Using Machine Learning”, No.f143-f148

[15]. Elias Dritsas, Maria Trigka, 2022, “Stroke Risk Prediction with Machine Learning Techniques”, No. 13:4670

[16]. Samaneh Ghanipour, Yousefzadeh Soroush, 2022, “Stroke Prediction with Logistic Regression and assessing it using Confusion Matrix”

[17]. Mohammed Guhdar, Amera Ismail Melhum, Alaa Luqman Ibrahim, 2023, “Optimizing Accuracy of Stroke Prediction Using Logistic Regression”

[18]. World Health Organization, “Stroke, Cerebrovascular accident”

[19]. Wikipedia, 2024, “Logistic regression”

[20]. Scikit Learn, “Support Vector Machines”

[21]. Alex Yartsev, 2022, “Measures of test accuracy: sensitivity specificity and predictive value”

[22]. Stephen M. Walker II, “F-Score: What are Accuracy, Precision, Recall, and F1 Score?”

[23]. Jason Brownlee, 2020, “Information Gain and Mutual Information for Machine Learning”

[24]. Gurchetan Singh, 2024, “ANOVA: Complete guide to Statistical Analysis & Applications (Updated 2024)”

[25]. GeeksforGeeks, 2024, “Feature selection using SelectFromModel and LassoCV in Scikit Learn”

[26]. Kinza Yasar, George Lawton, Ed Burns, 2022, “Logistic Regression”

[27]. Sunil Ray, 2024, “Learn How to Use Support Vector Machines (SVM) for Data Science”

[28]. Madhavan Vivekanandan, 2023, “Random Forest – A multitude of Decision Trees”

[29]. GeeksforGeeks, 2020, “Python Pandas – get_dummies() method”

[30]. Sole Galli, 2023, “Overcoming Class Imbalance with SMOTE: How to Tacke Imbalanced Datasets in Machine Learning”

[31]. Mario Daidone, Sergio Ferrantelli, Antonino Tuttolomondo, 2024, “Machine learning applications in stroke medicine: advancements, challenges, and future prospectives”, doi: 10.4103/1673-5374.382228