WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Innovative Approaches for Heart Stroke Risk Assessment: A Comparative Study of Supervised and Unsupervised Learning Models

Kankanala L S V S Nagamani Charan[1], Rasagna T[2], Eturu Harshith[3]

[1,2]Department of Computer Science, KIIT, Bhubaneshwar
[3]Department of Computer Science, Vellore Institute of Technology, AP

ABSTRACT:

Heart strokes remain a significant global health concern, necessitating novel approaches to enhance risk assessment and early intervention. This research presents innovative methods by conducting a comprehensive comparative analysis of supervised and unsupervised learning models for heart stroke risk assessment.

Supervised learning models leverage labeled data to predict stroke risk based on comprehensive patient information, while unsupervised models delve into data-driven pattern discovery without pre-defined labels. Through the rigorous evaluation of real-world patient datasets, this study not only assesses the performance of both approaches but also elucidates their respective strengths and limitations.

The findings of this research offer valuable insights to healthcare practitioners and researchers, enabling a deeper understanding of the effectiveness of supervised and unsupervised learning models in predicting heart strokes. Such insights are critical in advancing stroke prevention strategies, optimizing patient care, and ultimately mitigating the devastating impact of heart strokes on individuals and healthcare systems worldwide.

Keywords: *- KNN, SVM, Random Forest, XG Boosting, and CatBoost classifier.*

I. INTRODUCTION

A stroke typically manifests when there is an interruption or reduction in the blood supply to a specific region of the brain, thereby impeding the delivery of vital oxygen and nutrients to the brain. Consequently, the outcome is the gradual degeneration of brain cells. The initial indicators of a stroke encompass symptoms such as paralysis, sensory numbness, impaired vision, cephalalgia, and nausea. According to the World Health Organization (WHO), stroke ranks second as a leading cause of mortality worldwide, accounting for 11% of global fatalities.

Based on a range of parameters, including gender, age, presence of hypertension, glucose level, and smoking status, a classification problem can be employed to determine the likelihood of an individual experiencing a stroke. Based on the dataset provided, it is possible to make predictions for a new record by training the data using suitable machine learning algorithms.

In order to ascertain the likelihood of an individual experiencing a heart stroke, various factors such as age, gender, average glucose level, smoking status, body mass index, work type, and residence type are taken into consideration. In order to accurately forecast and ascertain outcomes, it is imperative to employ suitable, effective, and dependable machine learning algorithms. In order to identify regions of elevated risk within the specified parameters, with the objective of providing healthcare recommendations to individuals who are at a heightened likelihood of experiencing a stroke.

II. BASIC CONCEPTS

Pandas

The Python Pandas library is widely recognized and highly regarded for its robust capabilities in the realm of data manipulation and analysis. The software provides a range of data structures and functions that enable efficient manipulation of data. Pandas, which is constructed upon the foundation of the NumPy library, serves as a crucial instrument for professionals in the field of data science and analysis.

NumPy

The NumPy library, also known as Numerical Python, is widely utilized for executing scientific computations within the Python programming language. The software offers a streamlined and user-friendly approach to manipulating arrays and matrices while also offering a diverse array of mathematical functions.

Matplotlib

Matplotlib, a widely utilized data visualization library in the Python programming language, offers a diverse range of tools to generate visually appealing and precise plots, charts, and figures. The software provides a diverse array of customization choices, enabling users to generate refined visual representations suitable for a multitude of scientific and engineering purposes. Matplotlib offers a diverse selection of plotting functions that empower users to generate an array of chart types, encompassing line plots, scatter plots, bar plots, histograms, and additional options.

Sklearn

Scikit-learn, alternatively referred to as Sklearn, is a widely utilized machine learning library that is constructed upon the foundation of Python's scientific computing stack. The platform offers a diverse selection of algorithms for both supervised and unsupervised learning tasks, encompassing classification, regression, clustering, and dimensionality reduction techniques. The design of Scikit-learn places a strong emphasis on facilitating ease of use and enhancing readability

Tensorflow

TensorFlow is a machine learning library that has been developed by the Google Brain team and is available as open source software. The platform offers a diverse range of tools and frameworks that facilitate the construction and implementation of machine learning models across a multitude of applications, including but not limited to image recognition, natural language processing, and speech recognition. TensorFlow has been specifically engineered to possess a high degree of flexibility and extensibility, rendering it highly appropriate for the purposes of research and development.

III. Models Used:

K-Nearest Neighbours

The utilization of the K-Nearest Neighbors algorithm is a very suitable approach for predicting the occurrence of a stroke in an individual, as it relies on the concept of proximity or closeness to create accurate classification determinations. When specific parameters, such as blood sugar or BMI in relation to age, are considered, a substantial number of data points exhibit proximity to each other when plotted on a graph, corresponding to the occurrence or absence of strokes. Given a novel data entry, including variables such as age, body mass index (BMI), and blood sugar levels, it becomes feasible to ascertain the degree of proximity to a specific class in relation to another, hence facilitating the appropriate classification of said data entry.

SVM (Support Vector Machine)

The Support Vector Classifier technique is particularly well-suited for classification tasks involving a binary target variable, where there are two distinct classes. In this particular scenario, where the output is classed as either "high chances of a stroke" or "low chances of a stroke," the Support Vector Machine (SVM) employs a technique known as feature mapping to transform the data into a higher-dimensional space. This transformation enables the SVM to classify data points accurately, even in cases where they may not exhibit linear separability. In instances of non-linear separability, the separator can be represented as a hyperplane that effectively distinguishes the data points.

Random Forest Classifier

Although the decision tree classifier exhibits high accuracy, the random forest algorithm, despite its computing complexity, may offer superior effectiveness. The approach employs a combination of many decision trees, functioning as an ensemble. Each tree within the ensemble is constructed using a data sample that is selected from the training set with replacement, known as the bootstrap sample.

The utilization of this technique becomes advantageous in both classification and regression scenarios. In this particular scenario, a classification problem demonstrates the capability to effectively process substantial volumes of data, exhibiting superior resilience and precision compared to the decision tree algorithm.

CatBoost classifier

The CatBoost algorithm is a machine learning technique that utilizes gradient boosting on decision trees. This can be classified as an ensemble learning procedure. During the training process, a successive construction of decision trees takes place. Subsequent trees are constructed with less loss, heightened accuracy, and improved learning in comparison to the preceding trees. The CatBoost algorithm employs the boosting methodology to iteratively construct decision trees, with each subsequent tree benefiting from the information gained by its predecessor.

It is advantageous over random forest as it is great for processing categorical features.

XGBoost classifier

Similar to CatBoost, XGBoost is also an ensemble learning technique that exhibits unparalleled speed and performance, routinely surpassing alternative algorithms designed for supervised learning tasks. The approach involves the aggregation of many decision trees, referred to as base learners, which exhibit low bias and high variance characteristics. CART trees, also known as Classification and Regression Trees, exhibit slight deviations from conventional decision trees.

Instead of incorporating a solitary decision within each "leaf" node of the initial decision tree, the leaf nodes now encompass real-value scores that indicate the likelihood of the instance belonging to a specific category.

LGBM (Light Gradient Boosting Machine) Classifier

LightGBM is a rapid, distributed, and high-performance gradient boosting system that relies on decision tree methods. It is commonly employed for many machine learning applications, including ranking and classification. The training pace of the model is accelerated, and its performance closely resembles that of XG Boost. The approach utilizes two methodologies: Gradient-Based One-Side Sampling (GOSS) for preserving information accuracy and Exclusive Feature Bundling (EFB) for minimizing the number of effective features

K Means Clustering

The K-Means algorithm is a popular clustering technique in machine learning and data mining. Clustering is classified as an unsupervised machine learning approach. The data is partitioned into distinct clusters according to the categories of the dependent variable in the dataset. The aforementioned technique ensures convergence and exhibits adaptability to novel instances within the test dataset. Nevertheless, in this particular instance, it has been empirically demonstrated that the aforementioned algorithm exhibits a higher degree of inaccuracy when compared to alternative supervised learning algorithms.

Agglomerative Clustering

Agglomerative Clustering is classified as a form of hierarchical clustering technique. The technique employed is an unsupervised machine learning approach that partitions the population into multiple clusters, where data points within the same cluster exhibit higher similarity, while data points across different clusters demonstrate dissimilarity. In this study, a bottom-up methodology was employed, whereby each individual data point was first considered as a separate cluster. Subsequently, clusters were progressively merged as one ascended the hierarchical structure. The implemented version of K-means in the project exhibits lower accuracy compared to this alternative approach, while it still falls short of the accuracy achieved by other supervised learning methods.

IV. Data Selection and Analysis:

- **Data Selection:**

The dataset provided is comprehensive in nature, encompassing a wide range of parameters and factors that may contribute to the likelihood of an individual experiencing a stroke.

The attributes that are deemed significant in determining the same include gender, average glucose levels, BMI, age, hypertension, history of heart disease, and smoking status. However, it is important to acknowledge that work and lifestyle do play a significant role.

This is a detailed dataset that includes all kinds of parameters and factors that can play a role in determining whether a person will have a stroke.

Here are what we consider to be the most important among the attributes that determine the same, include the gender, average glucose, BMI, age, hypertension, heart disease history, and smoking status although work and lifestyle also play a role.

- **Data description:**

id	Integer
gender	Categorical: male, female or other
age	Float
hypertension	Categorical: 1 or 0
heart_disease	Categorical: 1 or 0
ever_married	Categorical: 1 or 0
work_type	Categorical: Private, Self employed, Govt job, Never_worked, children
Residence_type	Categorical: Urban, Rural
avg_glucose_level	Float
bmi	Float
smoking_status	Categorical: never smoked, formerly smoked, smoking, unknown
stroke	Categorical: 1 or 0

- **Model:**

Since the output variable is stroke, whose values are either yes or no, it is a binary classification problem. So we use logistic regression

Logistic regression is a type of regression we can use when the response variable is binary.

To evaluate the quality of a logistic regression model, create a confusion matrix, which is a 2 × 2 table that shows the predicted values from the model vs. the actual values from the test dataset.

V. Exploratory Analysis:

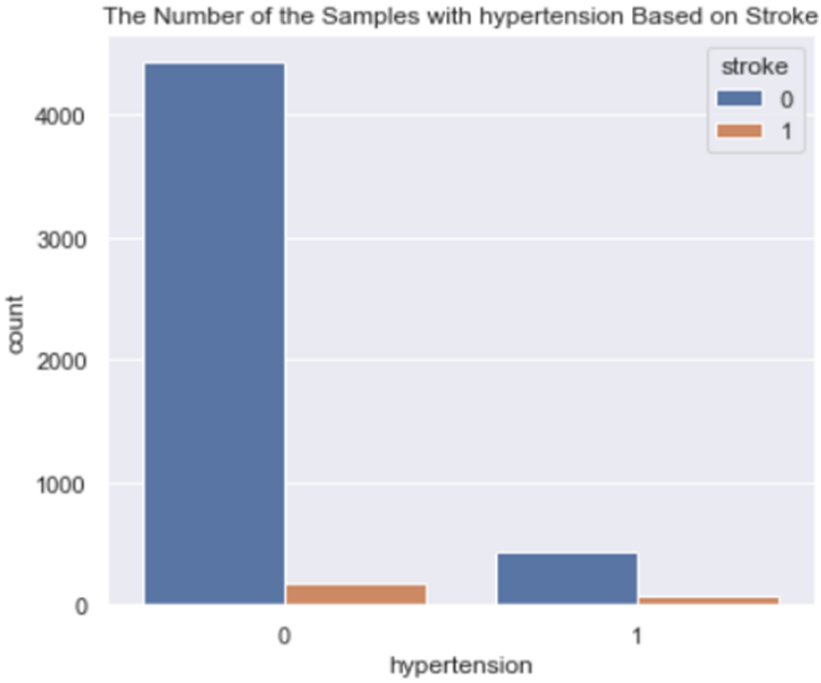

Fig. 2: The number of samples with hypertension Based on stroke

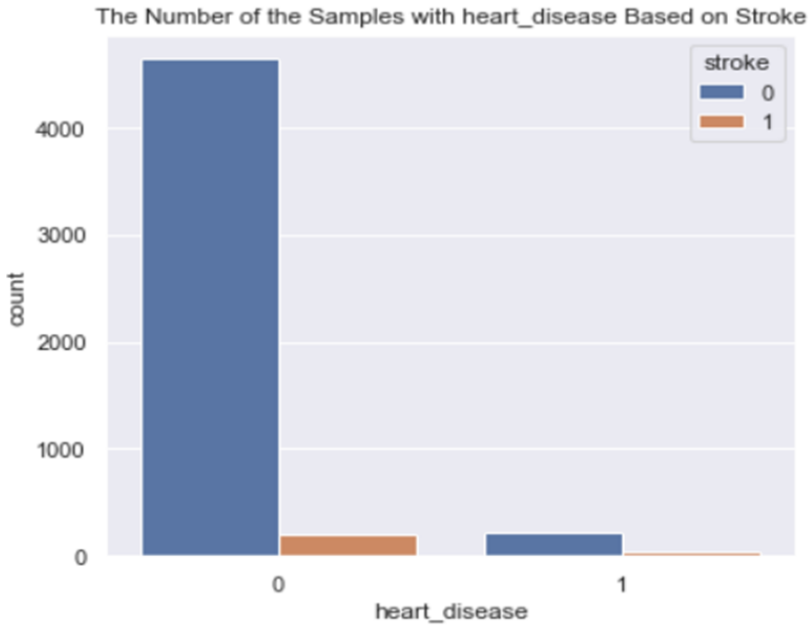

Fig. 3: The number of samples with heart disease Based on stroke

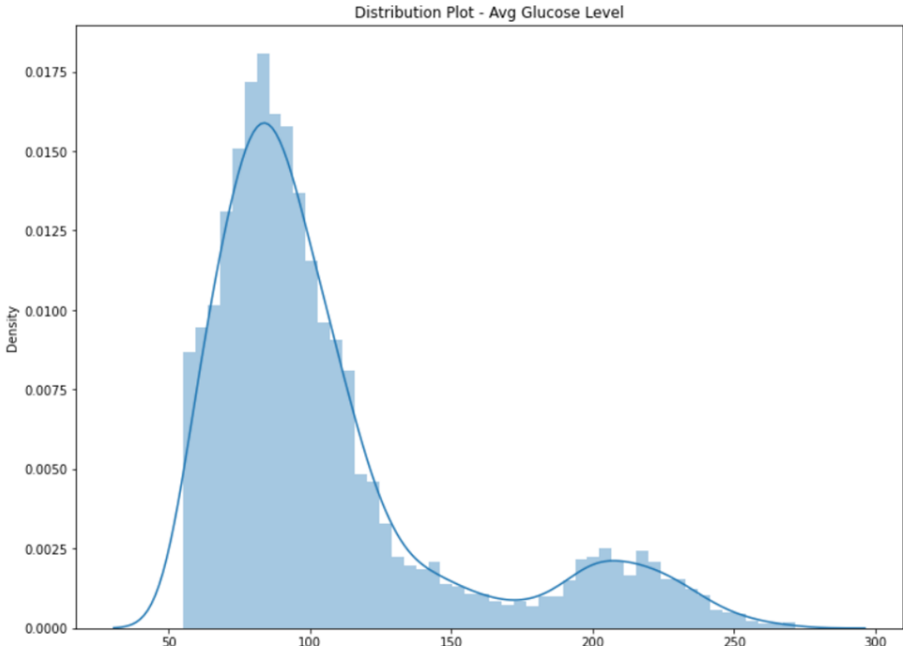

Fig. 4: Distribution Plot - Avg Glucose Level

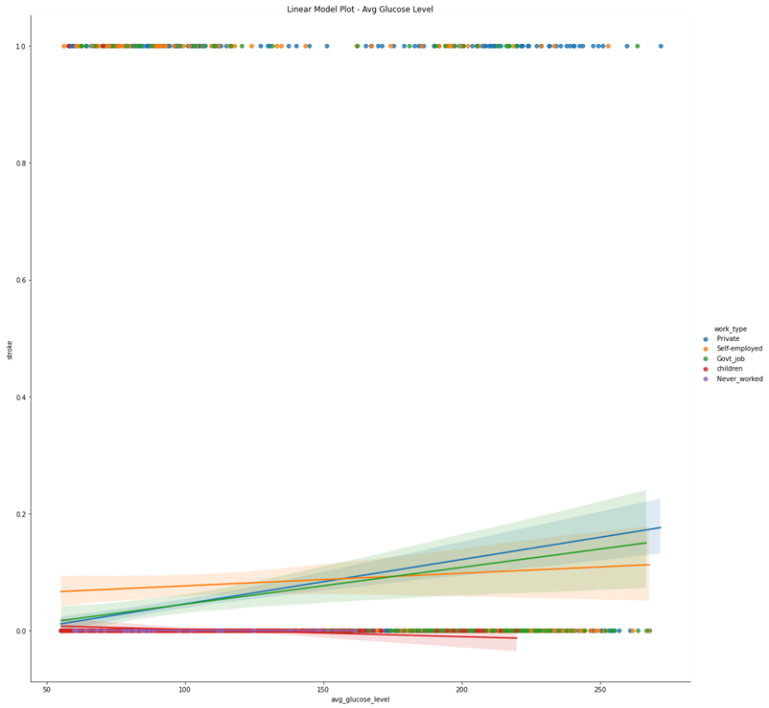

Fig. 5: linear model plot - Avg Glucose Level

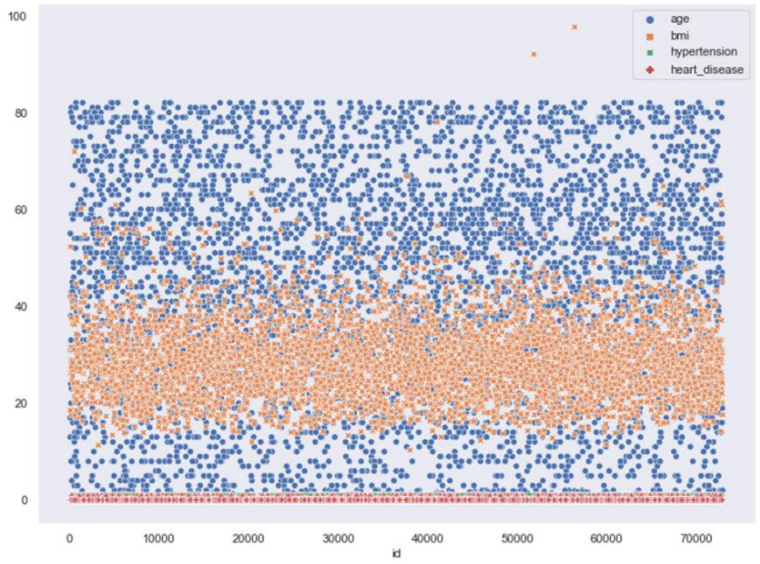

Fig. 6: Scatter plot

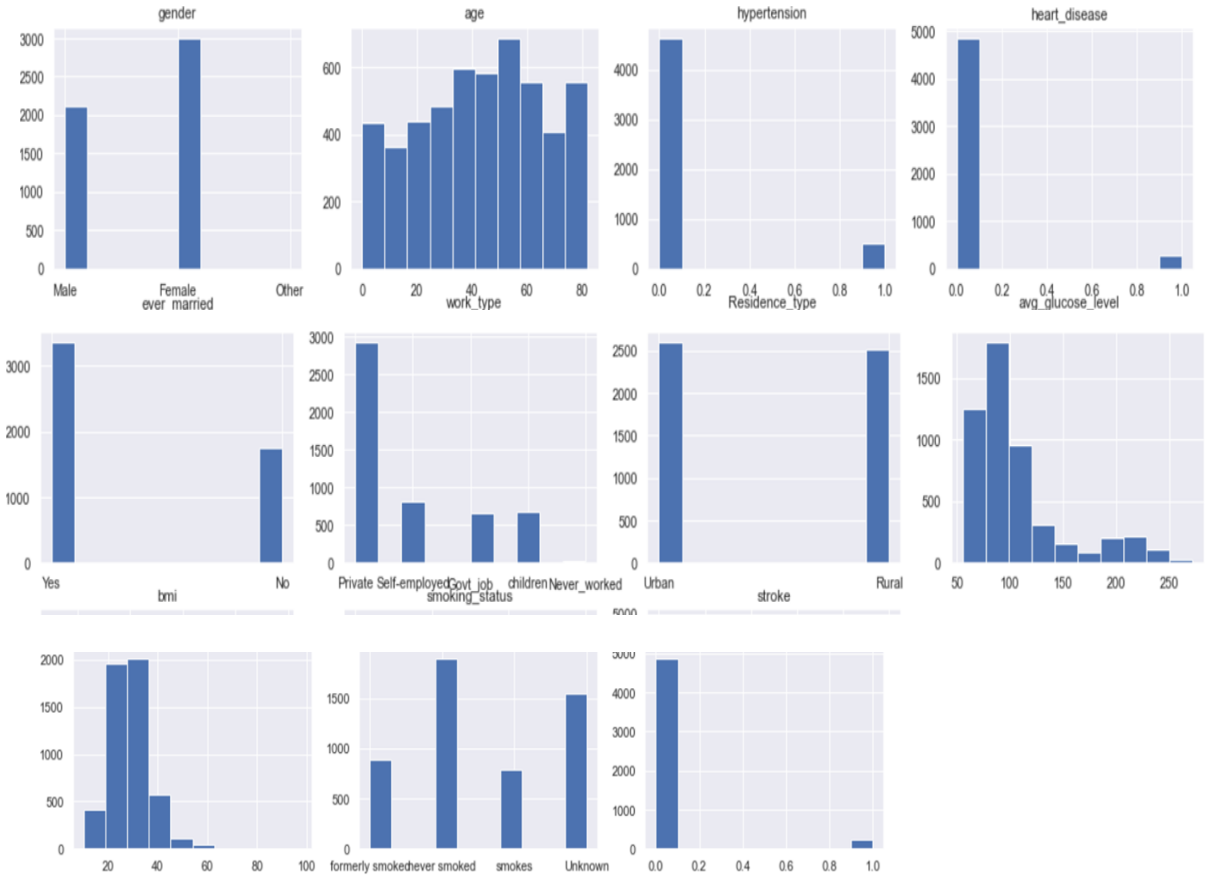

Fig. 7: Pair Plots

VI. Feature Engineering:

Dealing with missing and null data: When it comes to BMI, a significant number of records have values for BMI that are ambiguous. Due to the large number of entries with non-null BMI values, the blanks in the table have been populated with the mean of these values. Because of this, the model's capacity to produce accurate predictions is simplified and improved.

One Hot Encoding is a method that converts categorical data variables into a format that can be used by machine learning algorithms. This format can increase a model's ability to make accurate predictions and classifications. In this step, the categorical data is transformed into binary vectors, a kind of numerical data, producing a number of extra features that are dependent on the total number of distinct values contained in the categorical feature.

VII. Combined Analysis:

	ML Classification Algo	Rightly_Classified	Wrongly_Classified	Accuracy	Recall	Specificity
9	KNN (k=10)	0.947162	0.052838	0.947162	0.000000	1.000000
10	KNN (k=50)	0.947162	0.052838	0.947162	0.000000	1.000000
15	SVM (Support Vector Machine)	0.947162	0.052838	0.947162	0.000000	1.000000
3	XGBoosting	0.947162	0.052838	0.947162	0.000000	1.000000
4	Random Forest (Max 2 features in Bootstraping)	0.947162	0.052838	0.947162	0.000000	1.000000
5	Random Forest (Max 5 features in Bootstraping)	0.947162	0.052838	0.947162	0.000000	1.000000
14	KNN (k=4000)	0.947162	0.052838	0.947162	0.000000	1.000000
13	KNN (k=1000)	0.947162	0.052838	0.947162	0.000000	1.000000
12	KNN (k=500)	0.947162	0.052838	0.947162	0.000000	1.000000
11	KNN (k=200)	0.947162	0.052838	0.947162	0.000000	1.000000
0	LGBM Classifier	0.946184	0.053816	0.946184	0.074074	0.994835
2	CatBoost Classifier (300 iterations)	0.946184	0.053816	0.946184	0.037037	0.996901
7	Random Forest (Max 18 features in Bootstraping)	0.945205	0.054795	0.945205	0.055556	0.994835
6	Random Forest (Max 10 features in Bootstraping)	0.945205	0.054795	0.945205	0.000000	0.997934
1	CatBoost Classifier (1000 iterations)	0.944227	0.055773	0.944227	0.037037	0.994835
17	Agglomerative Clustering	0.628180	0.371820	0.628180	0.357143	0.639796
8	Gaussian Naive Bayes	0.375734	0.624266	0.375734	0.981481	0.341942
16	K-Means Clustering	0.138943	0.861057	0.138943	0.462963	0.120868

Fig. 8: Analysis of all models

VIII. Conclusion:

In conclusion, our study has provided a comprehensive examination of innovative approaches to heart stroke risk assessment through a comparative analysis of supervised and unsupervised learning models. We set out to address the critical need for more effective early detection and prevention strategies in the context of heart strokes, a significant global health concern.

Through our research, we observed that supervised learning models, relying on labeled data, demonstrated commendable predictive accuracy in assessing heart stroke risk. They proved to be particularly adept at leveraging patient information to make informed predictions.

An analysis of different supervised and unsupervised models was done, and the performance, computational efficiency, accuracy, and other parameters have been compared.

To conclude, the performances of KNN, SVM, Random Forest, and XGBoosting have been found to have significantly higher accuracy compared to other models. Hence, these supervised learning algorithms have been chosen over unsupervised learning models.

IX. Future Scope:

The study on innovative approaches for heart stroke risk assessment using supervised and unsupervised learning models opens up several promising avenues for future research and practical applications:

1. Hybrid Models: Researchers could explore the development of hybrid models that combine the strengths of supervised and unsupervised approaches. Combining predictive capabilities with pattern discovery could lead to more robust and accurate heart stroke risk assessment tools.

2. Feature Engineering: Further investigation into feature engineering techniques could enhance the performance of supervised models. Identifying and selecting the most relevant features from patient data could improve predictive accuracy.

3. Big Data and Deep Learning: As healthcare data continues to grow, the application of deep learning techniques to larger datasets holds promise. Deep learning models, such as neural networks, could extract intricate patterns from patient data for improved risk assessment.

4. Real-Time Risk Assessment: The development of real-time risk assessment systems that continuously monitor patient health data could be invaluable in early intervention. Integrating machine learning models into healthcare systems for ongoing risk evaluation is an exciting prospect.

5. Personalized Medicine: Tailoring risk assessment and intervention plans to individual patient profiles is a burgeoning field. Future research could focus on refining models to provide personalized stroke risk assessments and prevention strategies.

6. Interpretable AI: Developing models that provide interpretable explanations for their predictions is essential in healthcare. Ensuring the transparency and trustworthiness of AI-based risk assessment tools is crucial for widespread adoption.

7. Clinical Validation: Conducting extensive clinical trials and validations to assess the real-world effectiveness of these models is paramount. Collaboration between data scientists, clinicians, and healthcare institutions is essential to bridge the gap between research and clinical practice.

8. Ethical Considerations: As AI-based risk assessment tools become more prevalent, ethical and privacy considerations must be addressed. Research on responsible AI in healthcare, including data privacy and patient consent, is an ongoing area of concern.

9. Global Adoption: Scaling the deployment of these models to diverse healthcare systems worldwide is a significant challenge. Research on adapting and fine-tuning models for different healthcare contexts and populations is needed.

In summary, the future scope of research in heart stroke risk assessment using machine learning models is vast and holds the potential to transform how we approach stroke prevention and patient care. Continued collaboration between researchers, healthcare professionals, and technology experts will be essential to realizing the full potential of these innovative approaches.

References:

1. Prediction of Heart Stroke Using Support Vector Machine Algorithm(Research Gate)
2. Comparative Analysis and Implementation of Heart Stroke Prediction using Various Machine Learning Techniques(IJRASET)
3. anaconda.org/conda-forge/vadersentiment
4. kaggle.com/datasets/Early_Stroke_Prediction_Using_Machine_Learning
5. anaconda.org/anaconda/nltk