**Exploring Machine Learning Techniques for Stroke Prediction and Prevention**

Ms. M. Sindhu

Assistant Professor, Department of Computer Applications

Kongu Engineering College

Perundurai, India.

sindhuitsri@gmail.com

Nisanth G

Department of Computer Applications

Kongu Engineering College

Perundurai,India [nisanthg1010@gmail.com](mailto:nisanthg1010@gmail.com)

Preteka A T

Department of Computer Application

Kongu Engineering College

Perundurai, India

[pretekaa@gmail.com](mailto:dhanushreed.22bcr@kongu.edu)

Dr.P.Mahesh Kumar,

Assistant Professor,

Department of Mathematics

Kongu Arts and Science College, Erode..

<mathsmaheshkumar@gmail.com>

Varshini T

Department of Computer Application

Kongu Engineering College

Perundurai, India

[varshinithirunavukkarasu206@gmail.com](mailto:varshinithirunavukkarasu206@gmail.com)

***Abstract*—** **Heart stroke is still a major cause of death, hence effective prediction and prevention techniques are required. Using a variety of classifiers, this study investigates the effectiveness of machine learning algorithms in the prediction and prevention of stroke. A dataset from Kaggle was used to train and assess the classifiers, and the results showed accuracy rates ranging from 49.85% to 99.63%. A few classifiers stood out for their exceptional performance, including the Random Forest and Decision Trees classifiers, which had respective accuracies of 99.63% and 100%. In contrast, strokes were difficult to forecast with accuracies of 17.44% and 49.85% for Linear Regression and Lasso Regression, respectively. These results highlight the promise of machine learning algorithms for stroke prevention and prediction, highlighting the necessity of choosing the right models to attain the best possible accuracy and** **effectiveness in applications related to healthcare.**

***Keywords*** ***-*** ***Stroke- KNN (K-Nearest Neighbor)- SVM (Support Vector Machine)- Hyperparameter- Grid Search- Machine Learning- Naive Bayes- Decision Tree- Random Forest- Gradient Boosting Machines (GBM)- Neural Networks- K-Means Clustering- Hierarchical Clustering- Principal Component Analysis (PCA)- Singular Value Decomposition (SVD)- Logistic Regression-Lasso Regression-Gaussian Mixture Model.***

1. **INTRODUCTION**

Stroke poses a serious healthcare concern since it can cause severe sickness, disability, or even death owing to disruption of the blood supply to the brain. One of the leading causes of death in the globe is stroke. Given its quick start and serious effects, the importance of early stroke detection and intervention cannot be emphasized. New developments in technology, especially in the field of machine learning (ML), present exciting opportunities to improve stroke prevention and prediction methods. Researchers can use machine learning (ML) algorithms to detect patterns and risk variables linked to stroke occurrence by analyzing large datasets. This allows for preventative interventions aimed at reducing the impact of stroke.

Many studies have looked into the use of ML approaches for stroke prevention and prediction, evaluating accuracy and efficacy using a variety of classifiers and datasets. These studies frequently make use of datasets from websites such as Kaggle, which offer extensive collections of medical data for ML model testing and training. The goal of the research is to determine which classifiers—Random Forest, Gradient Boosting, Support Vector Machines (SVM), and Neural Networks—are the best at predicting strokes. By using strategies including feature selection and data balancing, this research aim to improve forecast accuracy and maximize model performance.

The results of these investigations highlight how ML algorithms have the potential to completely transform efforts to predict and prevent strokes. Some classifiers, like the Random Forest and Decision Trees classifiers, are particularly good at classifying data; they can even achieve 100% accuracy. On the other hand, certain models, such as Lasso Regression and Linear Regression, have trouble correctly predicting strokes, which emphasizes the significance of choosing the right algorithms for certain healthcare applications. Additionally, the creation of user-friendly software, like mobile apps, improves the usability and accessibility of machine learning-based techniques for predicting stroke, allowing for early intervention and bettering patient outcomes.

**II. LITERATURE REVIEW**

The third most common cause of mortality worldwide is stroke. It is a serious medical condition that can lead to death, severe illness, or disability as a result of cutting off blood supply to the brain. To start treatment early and reduce the death rate, a precise stroke prediction is essential. In order to more precisely identify strokes with unbalanced data, this study suggests a machine learning method. In order to balance the data, the Random Over Sampling (ROS) technique was applied in this work. This paper analyzes eleven classifiers: Gradient Boosting, Multi-Layer Perception, Support Vector Machine, Random Forest, K-nearest Neighbour, Decision Tree, Naïve Bayes, Voting Classifier, AdaBoost, and Nearest Centroid. Before balancing the data, ten classifiers exhibit more than 90% accurate results; four classifiers, utilizing the oversampling method, produce more than 96% correct results. To improve the outcomes, cross-validation and hyperparameter tuning are applied to every model. Moreover, machine learning model performance is assessed using Accuracy, F1-Measure, Precision, and Recall. According to the findings, the Support Vector Machine has the maximum accuracy of 99.99%, along with 99.99% recall, 99.99% precision, and 99.99% F1-measure values. With a 0.001% inaccuracy, Random Forest attains the second-highest accuracy of 99.87%. Furthermore, a user-friendly mobile app and a user-friendly online app are developed using the most precise model.[1]

One of the leading illnesses that significantly affects the death rate is still heart stroke. In contrast to other illnesses that can be identified and managed, a heart attack typically happens quickly and requires immediate attention. The accuracy of diagnosis and analysis is increasing in the medical area due to technological advancements. If specific characteristics are properly analyzed, stroke can be predicted well in advance of onset, potentially saving lives. This study presents one such attempt utilizing machine learning algorithms. We obtained a dataset from Kaggle. Nine machine learning algorithms, including linear discriminant analytics, logistic regression, Gaussian Naive Bayes, support vector machines, K-nearest Neighbors classifier, Random forest classifier, bagging classifier, Ada boost classifier, and gradient boosting classifier, were used to analyze the dataset's attributes. In order to accurately predict heart stroke in patients, the pattern of attributes according to the dataset provided was tracked. To facilitate additional analysis and comparison, the experimental data were split into training and testing datasets. After the data was analyzed, it was determined that the Random Forest algorithm, which had an accuracy rate of 95.10%, was reliable.[2]

In the US, stroke is the main cause of significant long-term impairment and the third most common cause of death. For the purpose of early intervention and therapy, accurate stroke prediction is extremely important. In this article, we use the Cardiovascular Health article (CHS) dataset to compare the Cox proportional hazards model with a machine learning strategy for stroke prediction. We specifically take into account the prevalent issues with feature selection, prediction, and data imputation in medical datasets. We provide a novel automatic feature selection approach that uses the conservative mean, our suggested heuristic, to choose robust features. Our suggested feature selection method, when paired with Support Vector Machines (SVMs), outperforms the L1 regularized Cox model and the Cox proportional hazards model in terms of area under the ROC curve (AUC). In addition, we introduce a margin-based censored regression algorithm that outperforms the Cox model in concordance index by fusing the idea of margin-based classifiers with censored regression. Overall, our method performs better than the state-of-the-art using the concordance index and AUC metrics. Furthermore, prospective risk variables that have not been found by conventional methods have also been uncovered by our analysis. Our approach can be used for clinical prediction of various diseases when risk variables are poorly known and missing data are widespread.[3]

A stroke is a medical emergency that happens when the blood supply to a portion of the brain is interrupted. When brain cells are denied the oxygen and glucose necessary for survival, they perish. It is the second most common disease in the world and can be deadly if left untreated. A stroke could be avoided with early intervention and prediction. Algorithms for machine learning are revolutionizing healthcare and are frequently utilized in the early detection of illness. This work focuses on applying machine learning techniques to predict the early occurrence of stroke. In terms of early stroke prediction, this research proves the advantage of Ada Boost over other popular classification techniques.[4]

A stroke is a medical disorder where the brain's blood arteries are torn, causing damage. It may also happen if the brain's supply of blood and other nutrients is cut off. The World Health Organization (WHO) states that stroke is the primary cause of death and disability worldwide. Relatively little study has been done on the risk of brain stroke, while the majority of studies has focused on heart stroke prediction. Because of this, a lot of machine learning models are created to predict the chance of having a brain stroke.This study trained five distinct models for precise prediction using a variety of physiological factors and machine learning algorithms, including Naïve Bayes classification, K-Nearest Neighbors, Support Vector Machine, Decision Tree classification, and Logistic Regression. Naïve Bayes is the algorithm that did the best on this task, with an accuracy of about 82%.[5]

The majority of strokes happen as a result of an unanticipated blockage in the heart and brain's pathways. Stroke risk can be reduced by being aware of the various warning indicators of stroke early on. Using various machine learning techniques in conjunction with the presence of hypertension, body mass index, heart disease, average glucose level, smoking status, history of stroke, and age, this research work suggests an early prediction of stroke disorders. Ten distinct classifiers have been trained with these high feature attributes in order to predict the stroke: KNeighbors Classifier, Gradient Boosting Classifier, XGBoost Classifier, Quadratic Discriminant Analysis, Decision Tree Classifier, AdaBoost Classifier, Gaussian Classifier, Logistics Regression, and Stochastic Gradient Descent. Next, the weighted voting method is used to aggregate the basic classifier outputs. Furthermore, a 97% accuracy rate was attained in the suggested investigation, with the weighted voting classifier outperforming the base classifiers. The greatest accuracy for predicting strokes is provided by this model. Additionally, the weighted voting classifier's area under the curve value is large. The weighted classifier has the lowest false positive and false negative rates when compared to other models. Consequently, weighted voting is nearly ideal as a classifier for stroke prediction, which doctors and patients can use to prescribe and identify a possible stroke early on.[6]

A stroke happens when there is a reduction or cessation of blood flow to the brain, depriving the brain tissue of vital nutrients and oxygen. Stroke disease treatment is essential. For this reason, early intervention and therapy are greatly aided by the prediction of stroke. It is possible to forecast a stroke by examining various warning indicators. In this experiment, we use different machine learning techniques to build a process of stroke risk prediction using our dataset. In this study, data imputation, feature selection, and data preprocessing are regarded as the first tasks. Certain physical characteristics, like age, gender, BMI, heart disease, and hypertension, are taken into account while training and testing models. To forecast the risk possibility of stroke, the following classifiers are used in this study: XGBoost, adaBoost, artificial neural network, decision tree, k-nearest neighbor (KNN) classifier, random forest, stochastic gradient descent (SDG), support vector machine (SVM), and random forest. Next, these eight conventional classifiers are subjected to a voting classifier implementation. Using voting classifiers to analyze several machine learning methods, we discovered that 98% of the models could accurately predict the risk factor for stroke.[7]

This paper describes a prototype using machine learning methods and text mining tools to classify strokes. With the use of properly trained machine learning algorithms, machine learning may be portrayed as an important asset in fields such as data management, surveillance, and medical. The use of data mining tools in this work provides an overview of information tracking from both a syntactic and semantic perspective. The plan is to use the case sheets to mine the symptoms of the patients and use the collected data to train the system. The case files of 507 patients were gathered from Sugam Multispecialty Hospital in Kumbakonam, Tamil Nadu, India, during the data gathering phase. After that, the case sheets were mined using maximum entropy and tagging techniques. The suggested stemmer then extracts the shared and distinct set of properties to categorize the strokes. Following processing, the data were fed into a number of machine learning methods, including random forests, boosting and bagging, support vector machines, and artificial neural networks. Compared to the other techniques, artificial neural networks trained using a stochastic gradient descent methodology performed better, achieving a 95% classification accuracy and a 14.69 standard deviation..[8]

A stroke is a medical condition where the brain is damaged due to a rupture in one of the blood vessels in the brain. There may be symptoms if the brain's blood and other nutritional supply are cut off. The World Health Organization (WHO) states that stroke is the leading cause of death and disability worldwide. Reducing the severity of a stroke can be achieved by early detection of its warning symptoms. A variety of machine learning (ML) models have been created to forecast the chance of a brain stroke. -This study trains four distinct models using a variety of physiological characteristics and machine learning methods, including Logistic Regression (LR), Decision Tree (DT) Classification, Random Forest (RF) Classification, and Voting Classifier. With an accuracy of almost 96%, the Random Forest algorithm proved to be the most effective for this particular assignment. The open-access Stroke Prediction dataset served as the basis for the method's development. The accuracy % of the models employed in this study is considerably higher than that of earlier research, suggesting the increased reliability of the models used in this examination. Their robustness has been demonstrated by many model comparisons, and the study analysis allows for the deduction of the scheme.[9]

One of the deadly brain disorders can kill a person within three to ten hours is stroke. Nonetheless, the majority of stroke deaths can be avoided by recognizing the type of stroke and responding quickly to it using intelligent healthcare systems. In this work, a machine learning model is used to predict if a patient will have a stroke. The Random Forest classifier beats state-of-the-art methods including K-NN, Decision Tree Classifier (DTC), and Logistic Regression. We use datasets with 5110 observations and 12 attributes to conduct our studies. Additionally, we used feature approaches to balance the datasets and EDA for preprocessing. Lastly, a mobile app that runs on the cloud gathers user data in order to evaluate and warn the user about the likelihood of a stroke with 96% precision, 96% recall, and 96% F1-score accuracy. With the minimal information required, this user-friendly technology can literally save a person's life by giving them an urgent warning that they can access from any location using a mobile device.[10]

This work uses several machine learning classification techniques to develop a supervised model based on machine learning that can predict the occurrence of a stroke in the near future depending on specific parameters. The model's projections have the potential to save many deaths or provide guidance on how individuals can mitigate the risk. The models derived from this study do not replace doctors; rather, they are merely a tool for medical professionals to utilize. A dataset including the characteristics or factors that influence stroke disease was used to train the model. To determine how much a given factor influences the target feature (having a stroke) or whether it affects other features, the correlation values were computed. Ultimately, a series of samples were used to test the model and determine how accurate the trained model was. Finally, a variety of models were generated using various algorithms (classifiers); nevertheless, the Random Forest classifier provides the basis for the model that yielded the highest results in terms of accuracy, precision, recall, and F1-score, which ranged from 94% to 95%.[11]

A stroke is a medical emergency brought on by bleeding or blood clots that stop the blood supply to a portion of the brain. With a 5.5 million yearly mortality rate, it ranks as the second most common cause of death worldwide. Over 15 million people worldwide suffer a stroke each year, and one person dies from a stroke every four minutes. Since a bad lifestyle is typically the cause of a stroke, it can be prevented in up to 80% of instances. As a result, stroke prediction becomes essential and ought to be applied to stop stroke-related irreversible harm. Several machine learning models, including Gaussian Naive Bayes, Logistic Regression, Decision Tree Classifier, K-Nearest Neighbors, AdaBoost Classifier, XGBoost Classifier, and Random Forest Classifier, were used in the current study to predict strokes. The comparison of every machine learning algorithm is presented in the study. The results of the analysis showed that the Random Forest Classifier, XGBoost, and AdaBoost produced the fewest inaccurate predictions and had the highest accuracy scores (95%, 96%, and 97%, respectively). As a result, they were the most appropriate model for stroke prediction and could be practically applied by doctors to forecast stroke in practice.[12]

When blood flow to a portion of the brain is suddenly cut off, a stroke results. Brain damage develops based on the part of the brain damaged, since brain cells gradually die in the absence of blood flow. Early symptom detection can provide a great deal of useful information for predicting stroke and encouraging a healthy lifestyle. In this study, a number of models are created and assessed using machine learning (ML) in order to create a strong framework for the long-term risk prediction of stroke incidence. This study's primary contribution is a stacking technique that performs well and is supported by a number of metrics, including accuracy, precision, recall, F-measure, and AUC. The experiment's findings demonstrated that, with an accuracy of 98%, an AUC of 98.9%, an F-measure, precision, and recall of 97.4%, the stacking classification works better than the other approaches.[13]

**III. PROPOSED WORK**

The suggested method involves using the stroke dataset to predict strokes in individuals by using a variety of machine learning strategies, such as Decision Trees, Random Forest, Naive Bayes, K-Nearest Neighbors (KNN), Gradient-Boosting Machines (GBM), Neural Networks, and Logistic Regression, along with fine-tuning techniques like grid and random search for KNN and GBM, respectively. Preprocessing the data, dividing it into training and testing sets, training each model, assessing its performance with metrics like accuracy and F1-score, fine-tuning hyperparameters as needed, contrasting the results, and finally choosing the best algorithm to implement in practical situations are all steps in the workflow.

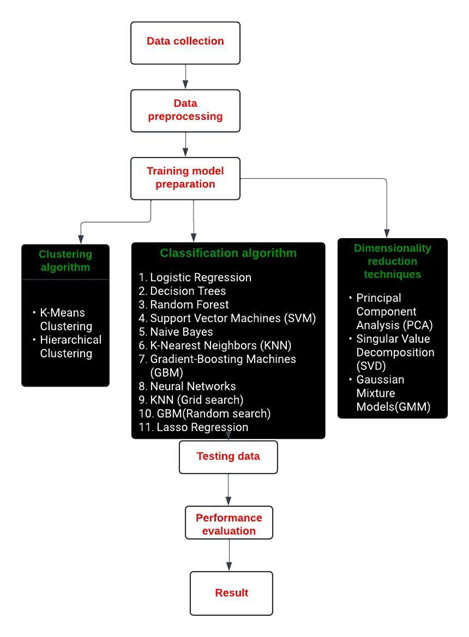


Figure 1. Flow diagram of Proposed System

**A. Dataset Collection**

The Stroke dataset which was taken from the Kaggle. The datasets consist of 40911 records of data which are divided into two class labels positive and negative classes and 11 attributes.

**B.Data Preprocessing**

The first and most vital phase is data preprocessing. Data preparation is the process of cleansing data so that it can be used by machine learning algorithms. A lot of medical-related data has noise and missing values. It may also be in formats that can’t be processed directly by machine learning models. Cleaning the data also improves the model’s accuracy and performance. First, we needed to find the missing value in the datasets available or not. In diabetes datasets, most of the attributes are categorical, so the categorical values must be converted into numerical values. This allows us to predict good results using various machine-learning techniques.

**C.Training Data and Test Data**

The datasets are then broken down into training and test data. 80% operated for training and the testing data is the other 20%. Each model has the training data added to it, and then the predicted results are compared with the testing data to see how accurate they are. We also look at the confusion matrices and compare performance measures like precision, recall, accuracy, and F1-Score measures.

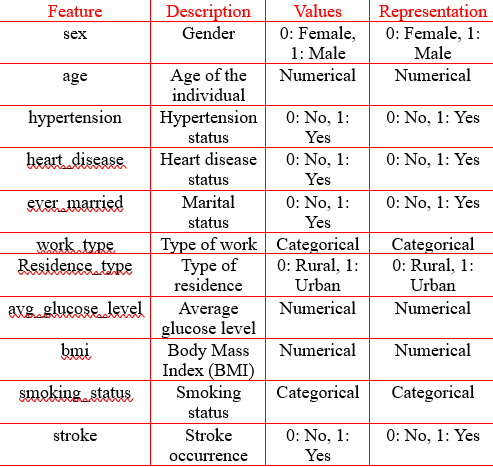


TABLE 1 -STROKE DATASET

**D.Methodology**

The process of assessing each machine learning algorithm's effectiveness against a stroke dataset is a component of the suggested technique for stroke detection. The dataset is first preprocessed in order to manage missing values, normalize features, and, if applicable, encode categorical variables. The dataset is then divided into training and testing sets. Then, the following algorithms are used: GBM with random search, KNN with hyper-parameter grid search, Decision Trees, Random Forest, Support Vector Machines (SVM), Naive Bayes, K-Nearest Neighbors (KNN), Gradient-Boosting Machines (GBM), Neural Networks, and Logistic Regression. A algorithm's performance is evaluated using measures like F1-score, accuracy, precision, and recall. To maximize the performance of algorithms like KNN and GBM, hyperparameter adjustment is done. Ultimately, the algorithm that shows the best performance metrics on the stroke dataset is determined to be the most appropriate for detecting stroke in individuals.

**CLASSIFICATION ALGORITHMS**

**1. Support Vector Machine**

SVM is classified as supervised learning strategy. It can be further classified as a classification algorithm and a regression algorithm, however, SVM is primarily employed in classification algorithms.

SVM divides various target classes into hyperplanes in n-dimensional space or multidimensional space. The primary pursuit of the SVM strategies is to specify a perfect decision boundary separating two or more classes by a maximum margin. This allows the placement of new data points in the relevant class.

The data points in the hyperplane are called vectors. The kernel makes it super simple to work with large-scale data. There are excess SVM kernel functions that can utilized to turn non-linear data into linear data. There are various kinds of kernel functions accessible.

• Radial Basis Function (RBF)

• Sigmoid

• Polynomial

• Linear

**Radial Basis Function Kernel (RBF)**

It's one of the most well-known and used kernel functions in SVM. It's often used with non-linear data and helps with proper separation when you don't have any prior knowledge of the data.

**Linear Kernel**

Linear kernel functions, which are typically one-dimensional in nature and the most fundamental type of kernel. When the number of features is large, linear kernel functions tend to be faster than other functions..

**Polynomial Kernel**

Polynomial kernels, which are non-linear kernels employed in machine learning, essentially transform the data from one dimension to another employing a polynomial function.

**Sigmoid Kernel**

The kernel function is the most widely used type of neural network. It looks similar to a two-layer perceptron model inside a neural network and acts as the neuron activation function.

**Algorithm For SVM**

1) Plotting data points based on class labels.

2) Draw the N digit of possible hyperplanes.

3) choose the most satisfactory hyperplane that divides the class as maximum margin.

4) Figuring out how far away two planes are from each other is called the maximum margin or decision boundary.



Figure.2 - Support Vector Machine

**2. K-Nearest Neighbor**

K-NN is a most popularized and fundamental ML strategy that uses supervised learning techniques. It detects similarities between freshly received instances/data and existing cases and classifies them accordingly. KNN saves all datasets and sorts new test values deployed to their closeness, allowing new data to be effortlessly sorted toward a well-fitted class. K-NN is a non-parametric strategy, which symbolizes that it creates no speculation regarding the data it is working with. It is also comprehended as a "lazy learner" since it does not immediately practice from a learning sample, instead holds data samples and enacts actions on them during the learning phase.

**Algorithm For KNN**

STEP 1 - Pick the k digit of nearest neighbor.

STEP 2- Computing the k-nearest neighbor by using Euclidean distance.

STEP 3- The K-nearest neighbors are stated by the Euclidean distance computed.

STEP 4- Estimate the amount of data samples within each class in the k neighbors.

STEP 5- Test provided data with training data and determine the majority of class labels.

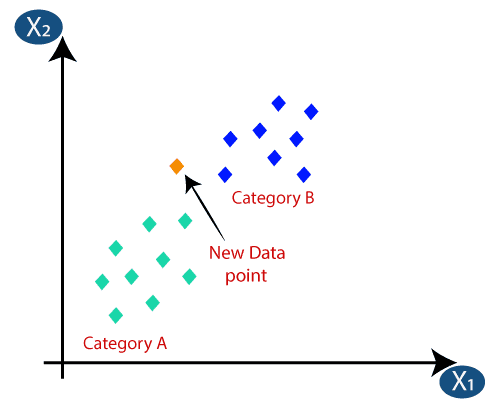


Figure.3 - K- Nearest Neighbor

**K-NN (With Grid Search Optimization Techniques)**

Grid search is one of the simplest algorithms you can use to get hyperparameters fine-tuned. We divide up the hyperparameter domain into different grids. Then, we try out all the possible combinations of parameters from each grid. Finally, we cross-test some performance metrics against each grid to see which one gets the most from the average of the cross-validated parameters.

Grid search is an algorithm that goes through all the combinations and finds the best one in the domain, but the downside is that it's really slow. It takes a lot of time to check every combination in the space, which isn't always available and each point in the grid requires k-fold cross-validation, which means you need to train your model for k times. So, tuning your model's hyperparameters can be complicated and expensive. But if you want to find the best combination of hyper-parameter values, grid search is a good option.

**3. Decision Tree**

The Decision Tree is a widely used supervised learning strategy that is easy to implement and interpret. It tree format that consists of leaf nodes, branches, and internal nodes. In a classification or else the regression exceptions, leaf nodes represent the result or constrained variable, while inner nodes represent the unconstrained variables. The branches indicate the decision rules that will be followed as decision-making progresses through the tree. Decision-making initiates at the core of the tree and progresses through the tree before a decision being reached. Each non-coastal node comprises a selection procedure that is employed when correlating a feature from data samples with a tree. The nodes are assigned attributes Gini, information gain, and entropy.



Figure.4 - Decision Tree

**4. Bernoulli Naive Bayes**

The Naive Bayesian Classification technique is a supervised learning strategy that originated from the Bayesian distribution theorem. It is designed to solve classification problems and is considered to be a precise and efficient Classification algorithm for machine-learning standards. It is competent in producing rapid projections and provides better results when datasets are of high dimension.

The Naive Bayesian variant of the Naive Bayesian theorem, developed by Bernoulli, is employed in machine learning. This variant is particularly useful when the dataset is a binary distribution in which the output label is either this or that. This algorithm has the advantage of only accepting binary features, such as

• 0/1

• YES/NO

• TRUE/FALSE

• POSITIVE/NEGATIVE

**Bernoulli Naive Bayes Formula**

P (Ai / B) = P (i / B) Ai + (1 – P (i/ B)) (1 - Ai)

P (A | B) is the probability that Ai is going to happen if B has already happened. i is the event that x holds either 0 or 1.

**E.Build Model for Proposed Methodology**

STEP 1: Importing the libraries and datasets.

STEP 2: Data preprocessing is needed to remove the missing data.

STEP 3: The data needs to be divided into testing data and training data.

STEP 4: Training data incorporates with 80% of the total data and the testing data incorporates 20%.

STEP 5: Choose the algorithm that ought to be performed.

STEP 6: Utilize the test set to assess the Classifier model for the machine learning algorithm.

STEP 7: Comparison of the performance evaluation.

STEP 8: After analysis, select the algorithm that performs the best based on various parameters.

**F.Performance Evaluation**

There are a variety of machine learning models available that can be utilized to assess a model's performance. There is a range of evaluation metrics that can be applied to the model, including accuracy, recall, F1-score, and precision were calculated employing confusion metrics.

**Accuracy Score**

Accuracy is one of the simplest evaluation metrics to evaluate different types of machine learning algorithms.

Accuracy = (TP+TN)/(TP+FP+FN+TN)

**Precision**

This shows the rate of the total anticipated samples that were correctly predicted to be positive. It can be expressed as follows:

Precision = TP/(FP+TP)

**Recall**

Recall shows the rate of total diabetic samples that stood correctly classified as diabetic.

Recall = TP/(FN+TP)

**F1-Score**

F1-Measure is the total number of times the model performed correctly on the test data set according to the confusion matrix.

F1-Score = 2TP/(FN+FP+2TP)

Where,

|  |
| --- |
| TN =>True Negative |
| TP =>True Positive |
| FN =>False Negative |
| FP =>False Positive. |

**5. Logistic Regression**

1.For binary classification tasks, such as those with a category target variable and two alternative outcomes (e.g., 0 or 1, yes or no), statistical regression is employed.

2.The logistic function, which converts any input to a number between 0 and 1, is used to model the likelihood that a given input belongs to a specific class.

3.The premise of logistic regression is that there is a linear relationship between the input variables and the outcome's log-odds.

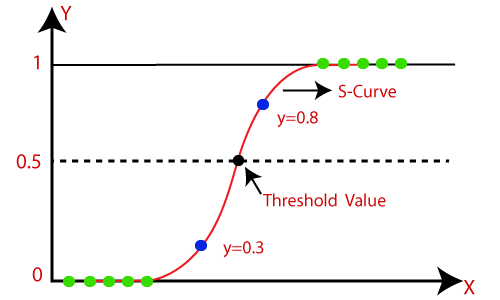


Figure.5 – Logistic Regression

**6. Random Forest**

• This ensemble learning technique is applied to both regression and classification problems.

• During training, it builds several decision trees, from which it produces the mode (classification) or average prediction (regression) of each tree.

• By training each tree using a bootstrap sample of the training data and taking into account a random subset of features at each split, Random Forest adds randomization to the system.

• When compared to individual decision trees, it reduces the likelihood of overfitting and increases prediction accuracy and generalization.

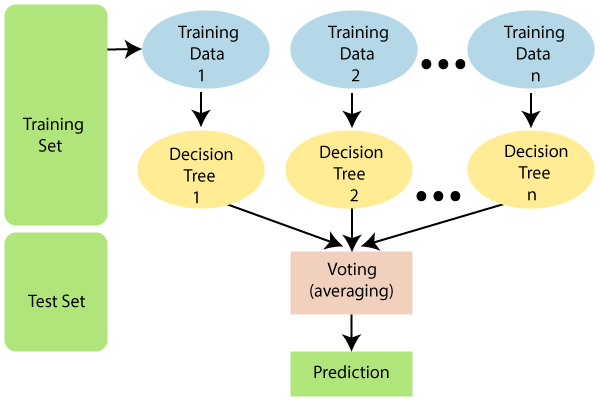


Figure.6 – Random Forest

**7. Neural Networks**

Artificial Neural Networks (ANNs), often known as neural networks, are a class of models that draw inspiration from the biological neural networks seen in the human brain.

• They are made up of interconnected layers of nodes, or neurons, each of which carries out a straightforward computation and then transfers its result to the one above it.

• Through the processes of forward propagation, which involves sending input data through the network to generate predictions, and backward propagation, which involves modifying the network's parameters in response to prediction failures, neural networks can discover intricate patterns in data.

•Their versatility allows them to be used for a wide range of tasks, including as image and speech recognition, regression, and classification.

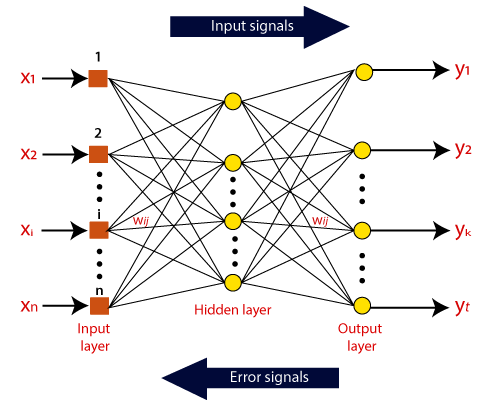


Figure.7 – Neural Networks

**8. Gradient-Boosting Machines (GBM)**

• This additional ensemble learning method is mainly applied to tasks related to classification and regression.

• GBM successively constructs an ensemble of weak learners, usually decision trees, whereby each new tree fixes the mistakes produced by its predecessor.

• It works by iteratively fitting new models to the residual errors of the prior models to minimize a loss function (such as the mean squared error for regression or the deviation for classification).

• GBM is a well-liked option in machine learning contests and practical applications because of its strong prediction accuracy and resilience to overfitting.

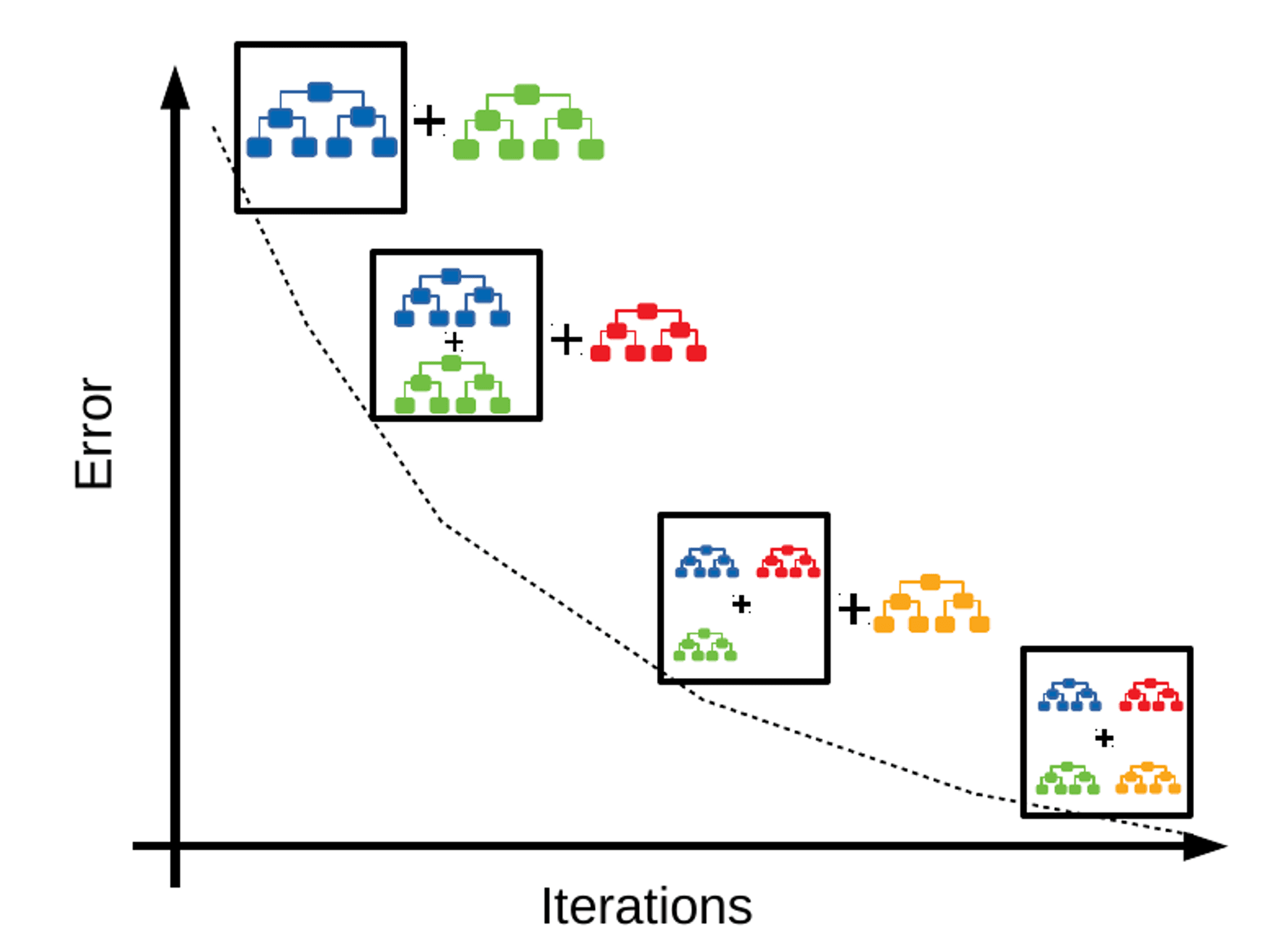


Figure.8 – Gradient Boosting

**CLUSTERING ALGORITHMS**

**K-Means Clustering**

For applications involving grouping, one common unsupervised machine learning technique is clustering.

A dataset is divided into 'k' unique, non-overlapping clusters, where 'k' is an integer that the user specifies

.Using an iterative process, the algorithm places each data point in one of the 'k' clusters according to how similar their features are.

The goal is to reduce the within-cluster variance, which is typically calculated by adding the squared distances between each data point and the cluster centroid.

K-Means is widely utilized in many different applications, including picture compression, anomaly detection, and customer segmentation. It is also comparatively quick and scalable.



Figure.9 – K-Means Clustering

**Hierarchical Clustering**

Another unsupervised learning technique for assembling comparable data points into groups is hierarchical clustering.

Hierarchical Clustering does not require the user to predetermine the number of clusters, in contrast to K-Means.

Iteratively joining or dividing clusters according to the proximity of data points results in a hierarchical tree of clusters, or dendrogram.

Agglomerative (bottom-up) and divisive (top-down) hierarchical clustering are the two primary forms. Whereas divisive clustering starts with all the data points in a single cluster and splits them recursively, agglomerative clustering starts with each data point as a separate cluster and merges them iteratively.

Because of its adaptability and ability to be shown as a dendrogram, hierarchical clustering is a valuable tool for exploratory data analysis and comprehending the connections between different data points. Still, For big datasets, it could be computationally costly though.

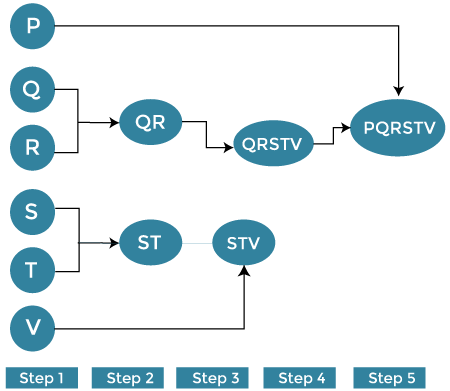


Figure.10 – Hierarchical Clustering

**DIMENSIONALITY REDUCTION TECHNIQUES**

**Principal Component Analysis (PCA)**

A dimensionality reduction method called principal component analysis is used to move high-dimensional data into a lower-dimensional space while keeping the most crucial information.

The original data is projected onto the directions (principal components) that show the greatest variation in the data, as determined using PCA.

The principle components rank according to how much of the variance in the data they can explain, and they are orthogonal to one another.

PCA lowers the dimensionality of the dataset by choosing a subset of the principle components that account for the majority of the variance in the data.

PCA is frequently used as a preprocessing step for several machine learning methods, as well as for exploratory data analysis and visualization.

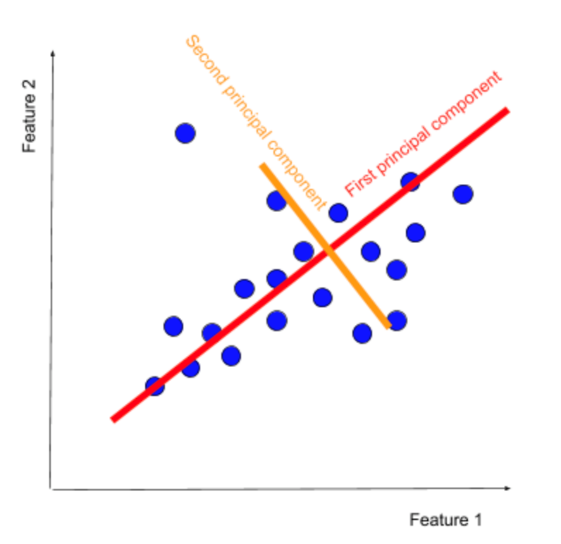


Figure.11 – Principal Component Analysis (PCA)

**Singular Value Decomposition (SVD)**

Singular Value Decomposition (SVD) is a matrix factorization technique that breaks down a matrix into three components:U, Σ,and 𝑉T.

These components represent orthogonal matrices and a diagonal matrix of singular values, respectively. Widely used in various applications such as dimensionality reduction and image compression, SVD is akin to PCA but offers numerical stability and versatility, accommodating matrices of any size**.**

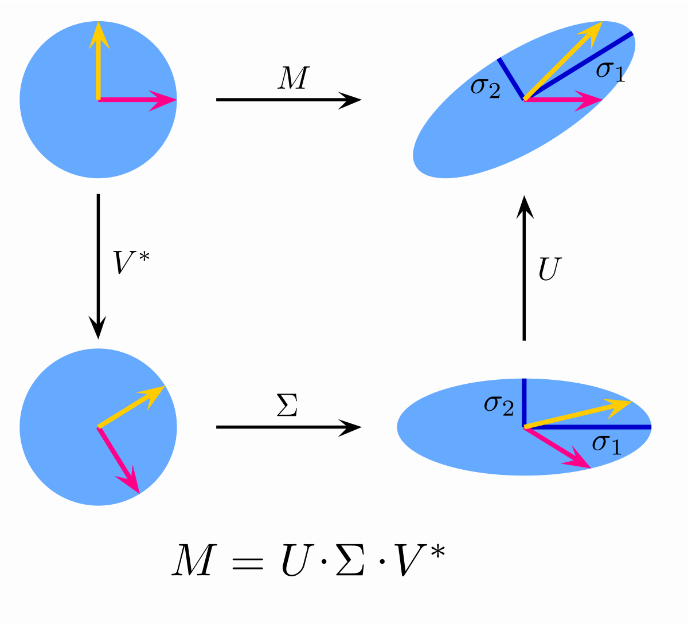


Figure.12 – Singular Value Decomposition (SVD)

**Gaussian Mixture Model (GMM)**

A probabilistic model used for tasks involving density estimation and grouping is the Gaussian Mixture Model (GMM). A weighted sum of several Gaussian distributions, each with a mean and covariance matrix, is used to describe the data distribution.

Each data point in a GMM is believed to have come from one of the Gaussian distributions, and the component weights indicate the likelihood that each Gaussian created the data point. In terms of math, a GMM is defined as:

*p*(**x**)=∑*k*=1*K*​*πk*​N(**x**∣***μ****k*​,**Σ***k*​)

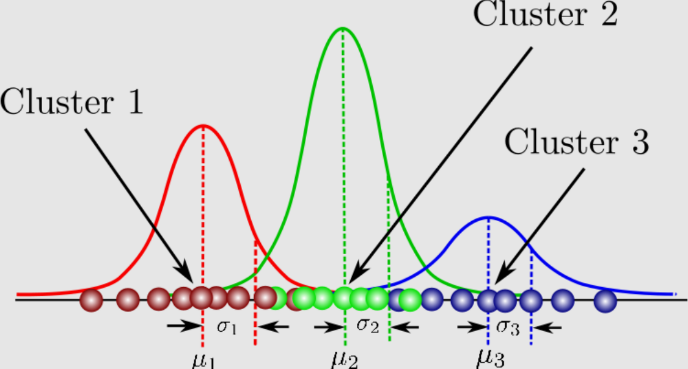


Figure.13 – Gaussian Mixture Model (GMM)

**IV. RESULTS**

In this Research work, the Stroke projection utilizes machine-learning algorithms to be implemented. In this work, we use Google Colab and Kaggle datasets to predict a patient has diabetes or not.

After performing statistical calculations, the program will gain a deeper understanding of the data. The data is split into an 80:20 portion, 80% of data is employed for training data, and 20 % of data is employed for testing data.

Out of all the methods that have been studied, Random Forest and Decision Tree classifiers perform exceptionally well in terms of accuracy, precision, recall, and F1-score. K-Nearest Neighbors (KNN) exhibits superior accuracy, precision, recall, and F1-score in comparison. While Neural Network and Logistic Regression classifiers perform significantly lower in terms of accuracy, precision, recall, and F1-score, Support Vector Machines (SVM) and Naive Bayes classifiers show middling performance. Across all metrics, the Lasso Regression classifier performs the worst.

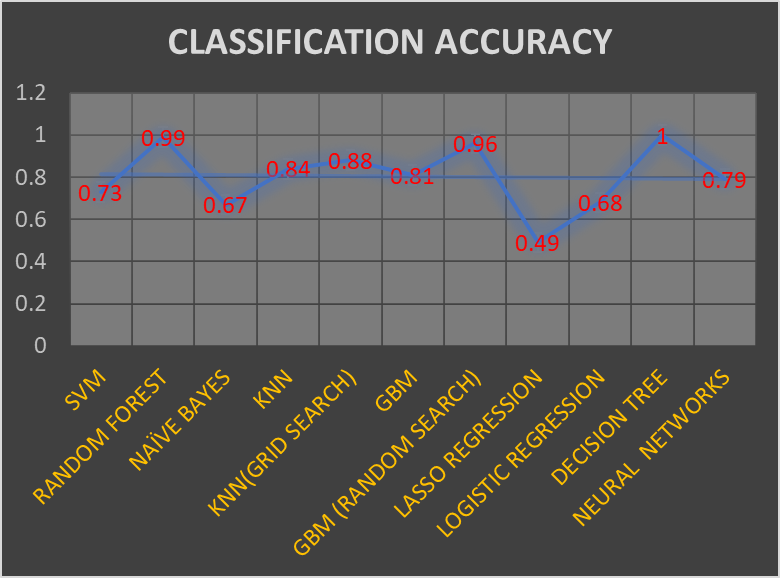


Figure.14 – Classification Accuracy

There are different evaluation metrics available like recall, precision, f1-score, and accuracy. Comparison of the metrics that are shown below Figure.11 - Evaluation Metrics

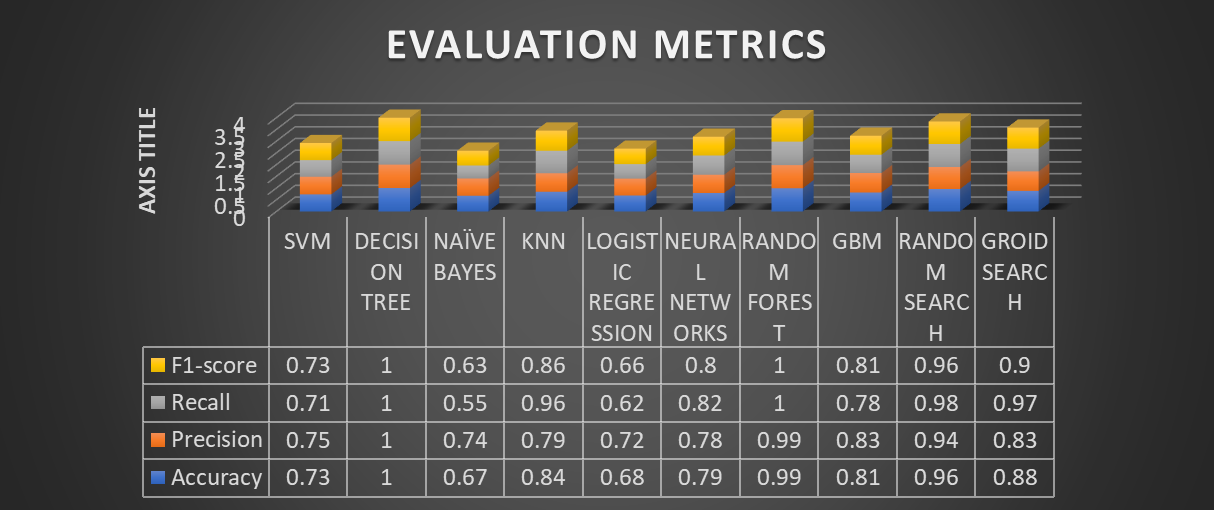


Figure.15- Evaluation Metrics

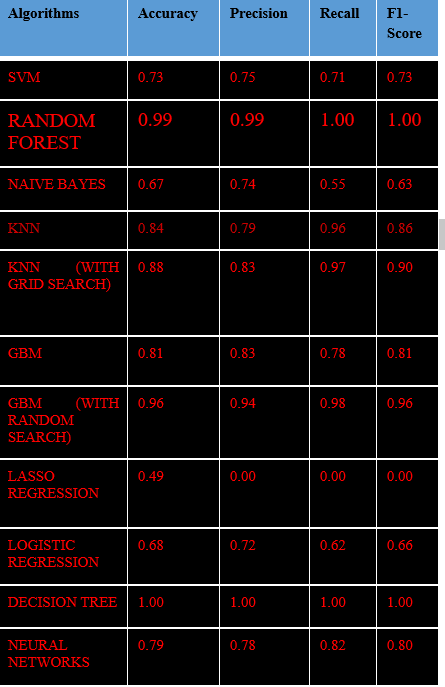


TABLE 2 – Evaluation Metrics Values

**CONFUSION MATRIX FOR CLASSIFICATION ALGORITHMS**

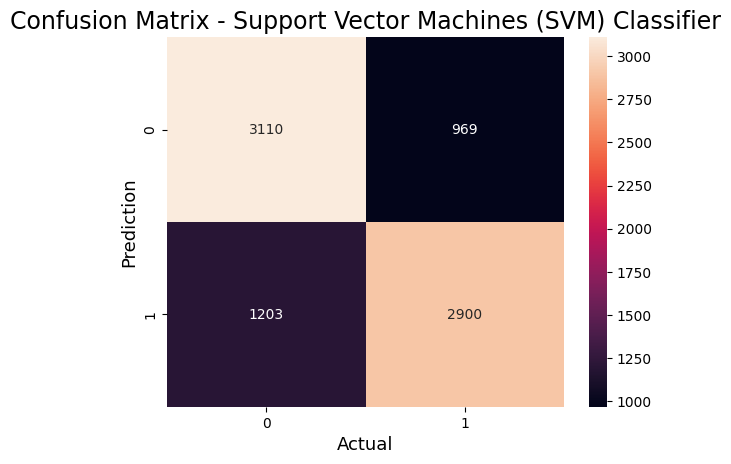


Figure.16 - Matrix for Support Vector Machine

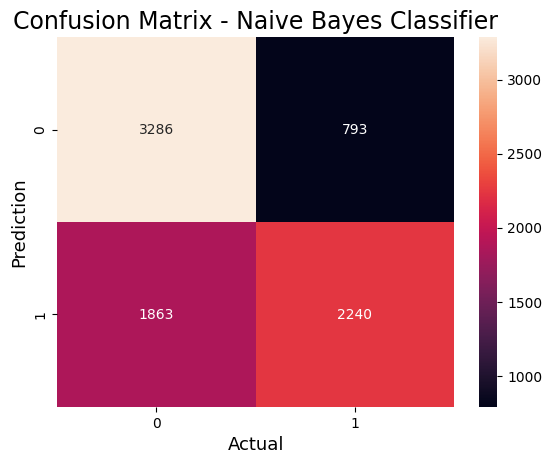


Figure.17 - Matrix for Naive Bayes

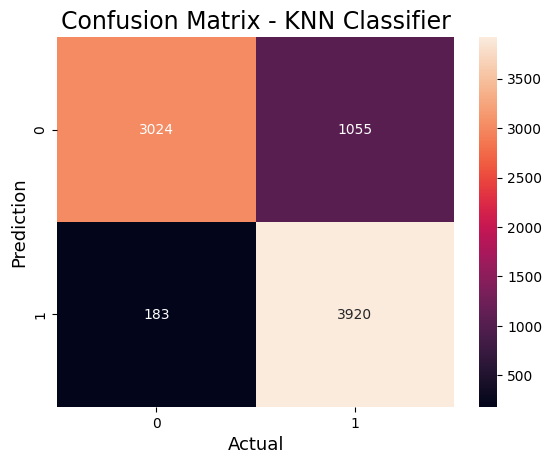


Figure.18 - Matrix for K-Nearest Neighbor

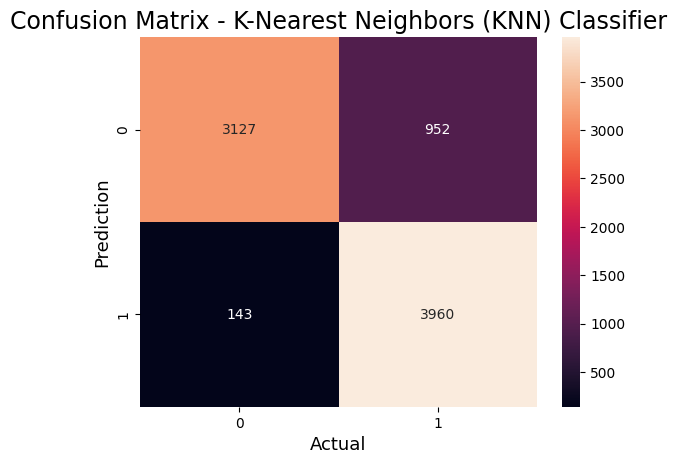


Figure.19 - Matrix for KNN (GRID SEARCH)

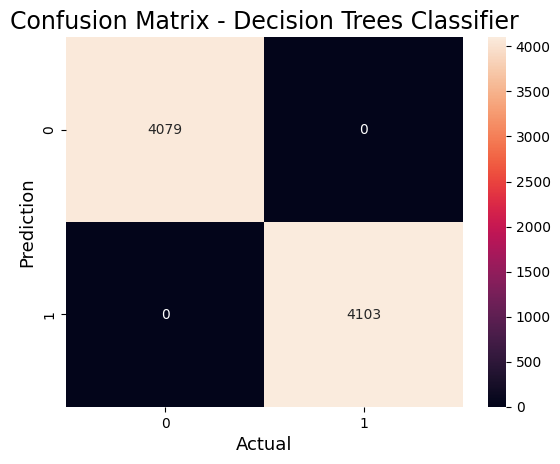


Figure.20 - Matrix for Decision tree

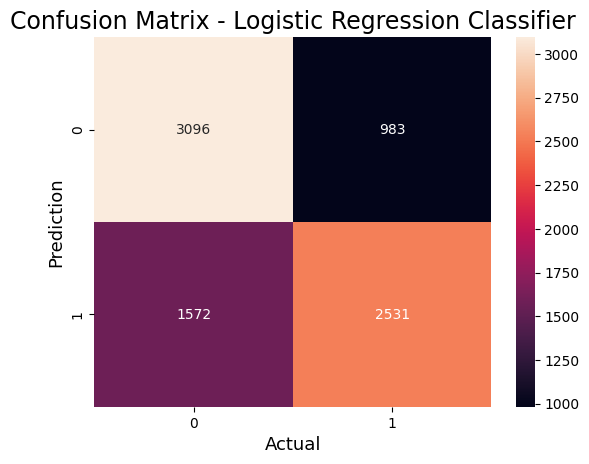


Figure.21 - Matrix for Logistic Regression

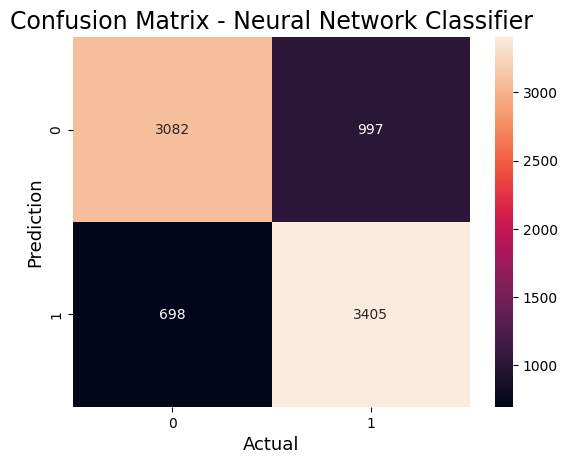


Figure.22 - Matrix for Neural Network

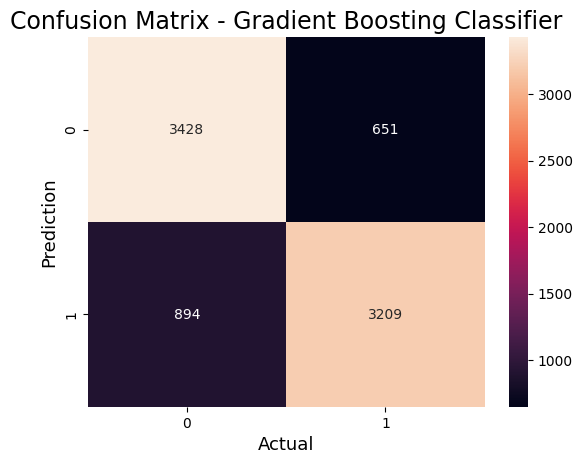


Figure.23 - Matrix for GBM

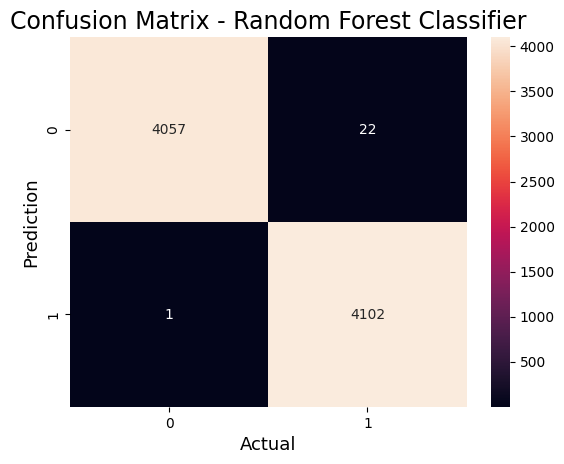


Figure.24 - Matrix for Random Forest

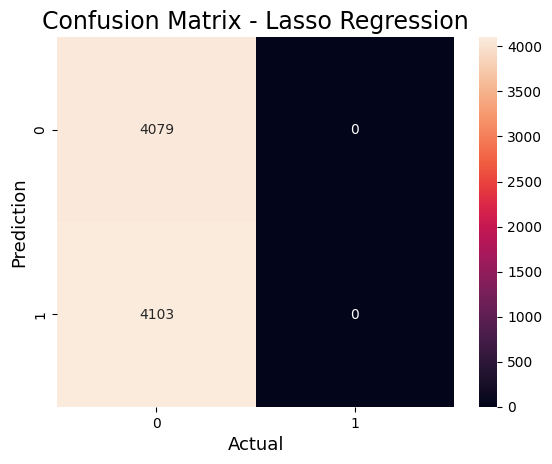


Figure.25 - Matrix for Lasso Regression

**CLUSTERING ALGORITHMS OUTPUT**

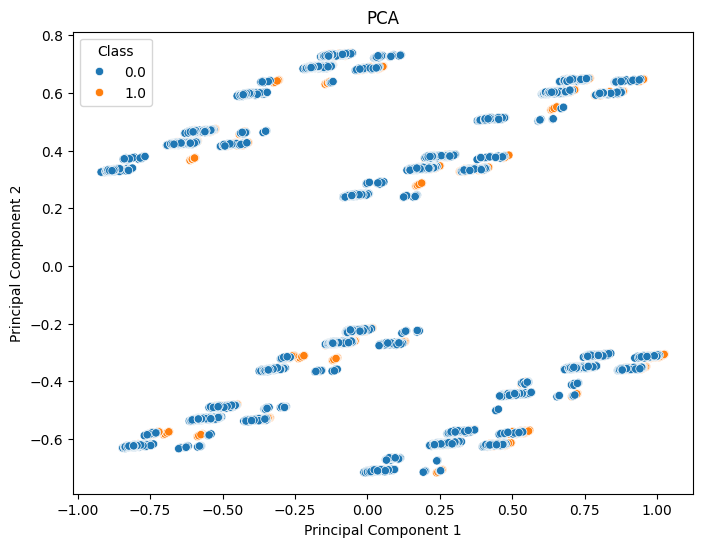


Figure.26– K-Means Clustering

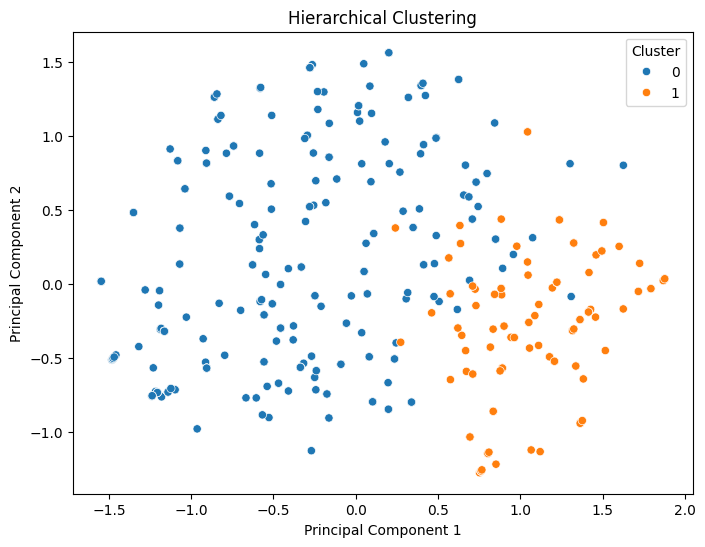


Figure.27 – Hierarchical Clustering

**DIMENSIONALITY REDUCTION TECHNIQUES OUTPUT**

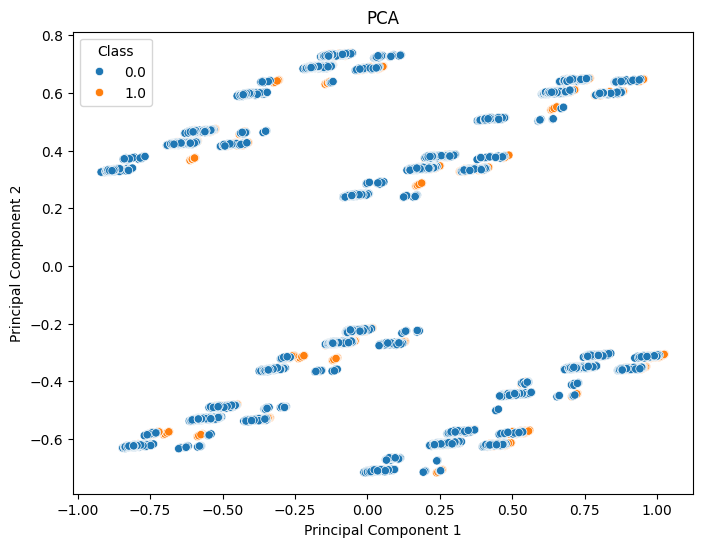


Figure.28 – Principal Component Analysis (PCA)

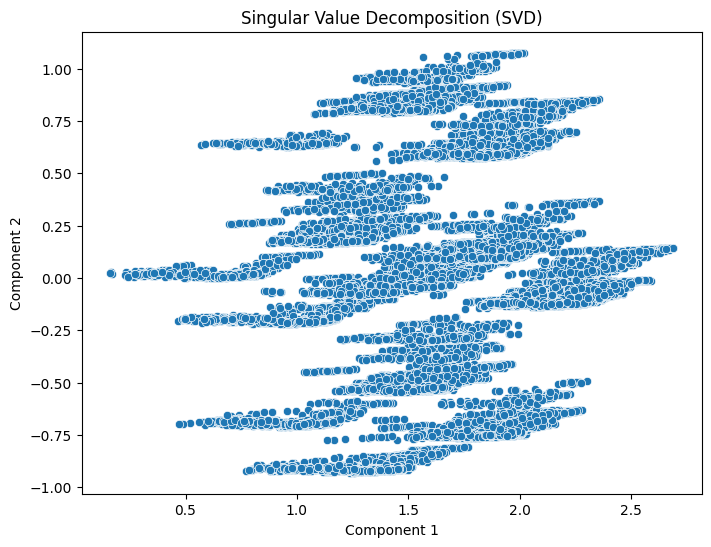


Figure.29 – Singular Value Decomposition (SVD)

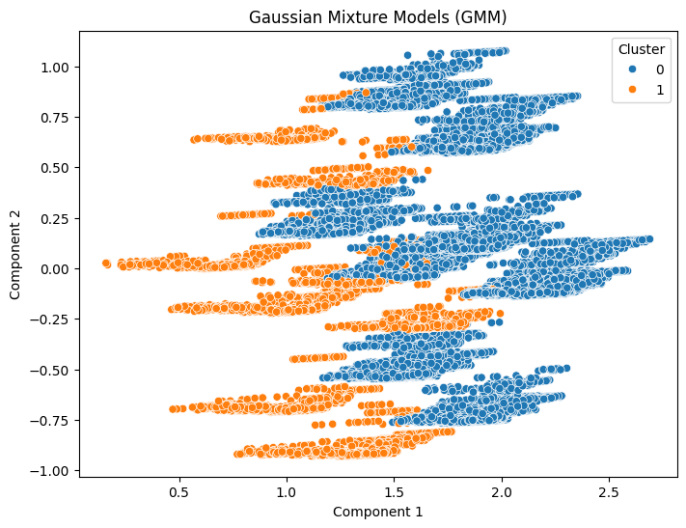


Figure.30 Gaussian Mixture Model (GMM)

**V. CONCLUSION AND DISCUSSION**

It is clear from the analysis of different machine learning algorithms for stroke illness prediction that Random Forest and Decision Tree classifiers perform exceptionally well, attaining ideal scores in F1-score, accuracy, precision, and recall. These models show strong capacities to represent intricate relationships in the dataset, which makes them excellent choices for challenges involving the prediction of strokes. Support Vector Machines (SVM) perform rather well; however, they are not as accurate as Random Forest and Decision Trees. Although they perform somewhat poorly, K-Nearest Neighbors (KNN) and Naive Bayes classifiers are good substitutes for predicting stroke. Though marginally less so than Decision Trees and Random Forest, Neural Network and Gradient Boosting models also demonstrate encouraging outcomes. However, the low performance of the Lasso Regression and Logistic Regression models indicates their limited applicability for this task. In general, Decision Trees and Random Forest are the most competitive models for predicting stroke disease; however, other models may be able to meet certain needs and limitations. These models may benefit from additional optimization and fine-tuning to improve their performance in practical settings.

REFERENCES

[1] N. Biswas, K. M. M. Uddin, S. T. Rikta, and S. K. Dey, “A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach,” *Healthcare Analytics*, vol. 2, p. 100116, Nov. 2022, doi: 10.1016/j.health.2022.100116.

[2] D. Paikaray and A. K. Mehta, “An Extensive Approach Towards Heart Stroke Prediction Using Machine Learning with Ensemble Classifier,” in *Proceedings of the International Conference on Paradigms of Communication, Computing and Data Sciences*, M. Dua, A. K. Jain, A. Yadav, N. Kumar, and P. Siarry, Eds., in Algorithms for Intelligent Systems. , Singapore: Springer Singapore, 2022, pp. 767–777. doi: 10.1007/978-981-16-5747-4\_66.

[3] A. Khosla, Y. Cao, C. C.-Y. Lin, H.-K. Chiu, J. Hu, and H. Lee, “An integrated machine learning approach to stroke prediction,” in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, Washington DC USA: ACM, Jul. 2010, pp. 183–192. doi: 10.1145/1835804.1835830.

[4] V. JalajaJayalakshmi, V. Geetha, and M. M. Ijaz, “Analysis and Prediction of Stroke using Machine Learning Algorithms,” in *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, Coimbatore, India: IEEE, Oct. 2021, pp. 1–5. doi: 10.1109/ICAECA52838.2021.9675545.

[5] G. Sailasya and G. L. A. Kumari, “Analyzing the Performance of Stroke Prediction using ML Classification Algorithms,” *IJACSA*, vol. 12, no. 6, 2021, doi: 10.14569/IJACSA.2021.0120662.

[6] M. U. Emon, M. S. Keya, T. I. Meghla, Md. M. Rahman, M. S. A. Mamun, and M. S. Kaiser, “Performance Analysis of Machine Learning Approaches in Stroke Prediction,” in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India: IEEE, Nov. 2020, pp. 1464–1469. doi: 10.1109/ICECA49313.2020.9297525.

[7] Md. M. Islam, S. Akter, Md. Rokunojjaman, J. H. Rony, A. Amin, and S. Kar, “Stroke Prediction Analysis using Machine Learning Classifiers and Feature Technique,” *IJECS*, vol. 1, no. 2, pp. 57–62, Dec. 2021, doi: 10.24042/ijecs.v1i2.10393.

[8] P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi, R. Patan, P. Jayaraman, and R. Manikandan, “RETRACTED ARTICLE: Classification of stroke disease using machine learning algorithms,” *Neural Comput & Applic*, vol. 32, no. 3, pp. 817–828, Feb. 2020, doi: 10.1007/s00521-019-04041-y.

[9] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, and M. Monirujjaman Khan, “Stroke Disease Detection and Prediction Using Robust Learning Approaches,” *Journal of Healthcare Engineering*, vol. 2021, pp. 1–12, Nov. 2021, doi: 10.1155/2021/7633381.

[10] Md. M. Islam, S. Akter, Md. Rokunojjaman, J. H. Rony, A. Amin, and S. Kar, “Stroke Prediction Analysis using Machine Learning Classifiers and Feature Technique,” *IJECS*, vol. 1, no. 2, pp. 57–62, Dec. 2021, doi: 10.24042/ijecs.v1i2.10393.

[11] H. Al-Zubaidi, M. Dweik, and A. Al-Mousa, “Stroke Prediction Using Machine Learning Classification Methods,” in *2022 International Arab Conference on Information Technology (ACIT)*, Abu Dhabi, United Arab Emirates: IEEE, Nov. 2022, pp. 1–8. doi: 10.1109/ACIT57182.2022.10022050.

[12] S. Gupta and S. Raheja, “Stroke Prediction using Machine Learning Methods,” in *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India: IEEE, Jan. 2022, pp. 553–558. doi: 10.1109/Confluence52989.2022.9734197.

[13] E. Dritsas and M. Trigka, “Stroke Risk Prediction with Machine Learning Techniques,” *Sensors*, vol. 22, no. 13, p. 4670, Jun. 2022, doi: 10.3390/s22134670.

[14] Prabavathi G.T. & Sindhu M, “A Review on Heart Disease Predictive System Using Machine Learning Algorithms”, Shodha Prabha (UGC CARE Journal), 2022, vol. no. 47, pp 37-42.

[15] Prabavathi G.T. & Sindhu M, "DNFH: Data Pre-processing and Feature Selection Using Hybrid Methods For Heart Disease Prediction", Journal of Namibian Studies - History Politics Culture (ISSN print 1863-5954; ISSN online 2197-5523), Vol. 35 (2023): Special Issue 1, DOI https://doi.org/10.59670/jns.v35i.3775. (SCOPUS-Q2)

[16] Prabavathi G.T., & Sindhu M., “ A Novel Density-Based Clustering Algorithm For Predicting Cardiovascular Disease”, Journal of Tianjin University Science and Technology. ISSN:0493-2137. vol. no. 57 Issue: 2(2024), pp 264-275. (SCOPUS-Q3)

[17] Sindhu M, Prabavathi G.T, HEART DISEASE PREDICTION USING CARDIO CAPSULE NEURAL NETORK. (2024). Machine Intelligence Research, 18(2), 43-64.  https://machineintelligenceresearchs.com/index.php/MIR/article/view/136