Deep Learning Prediction Model for predicting heart stroke using the combination Sequential Row Method integrated with Artificial Neural Network

T. Swathi Priyadarshini[a] , Mohd Abdul Hameed[b] , Balagadde Ssali Robert[c]
Department of computer science and engineering[a] , Department of computer science and engineering, University college of Engg, Osmania University, Hyderabad, Telangana, India[b], Department of computer science, school of mathematics and computing, Kampala International University[c].

Abstract

Stroke is an important health outcome in terms of morbidity, disability, mortality, and social and economic costs [2-3]. Method, in this paper, we build a smart prediction model which predicts whether a patient is at high risk or low risk of heart stroke with an early intervention by classifying the patient's records into one of the binary classes. We are using K-means Clustering technique and Classification techniques and to enhance the performance in predicting accurate values, we are integrating all 3 classification algorithms – Naive Bayes NB, Decision Tree DT and Artificial Neural ANN Network, with Sequential Row initial centroid selection methods of K-means clustering algorithm.Comparison analysis of each model is determined by calculating Sensitivity, Specificity and Accuracy using Confusion Matrix of each one. We also plotted ROC and AUC score as final assessment, in choosing the best prediction model. Conclusions, we developed smart and highly accurate predicting heart stroke system. The most effective model is Sequential Row Method integrated with Artificial Neural Network Classifier with accuracy score of 96% and area under the ROC curve (AUC) score of 1.

KEYWORDS Machine Learning, Naïve Bayes, Decision Tree, Artificial Neural Networks, K-Means Clustering, Initial Centroid Selection Methods, Heart Stroke Diagnosis.

1. Introduction :[4]Stroke is the second-leading cause of death in the world but has dropped to fourth in the United States, behind CHD, cancer, and chronic respiratory disease. The World Health Organization suggests 5.5 million deaths of stroke in 2002 (≈1 every 6 seconds). [8-12]Machine learning in healthcare can therefore be used to identify and understand emerging health trends in large populations and datasets. This helps health institutions and public bodies make public health interventions at scale. The classification algorithms i.e. Artificial Neural Networks, Naïve Bayes and Decision Tree are used for building prediction systems in identifying the risk of heart stroke considering all risk factors. [7] The Clustering algorithm i.e. K-means is used to organize the large dataset of heart disease patient records which helps in enhancing the accuracies of the classifiers.

2. Naïve Bayes (NB): [5-6]Naïve Bayes is a conditional probability model. Given a problem instance to be classified, represented by a vector X = (X1… Xn). The vector I a set of infinite features (independent variables), which are assigned to the instance probabilities.

$$P\ (Ck|\ X1,\ldots..,Xn)$$

For each of k possible outcomes or classes Ck. This is a problematic, if it has large set of features. Using Baye's theorem, the conditional probability can be decomposed as:

$$P\ (Ck|X) = P(Ck)P(X|Ck)\ /\ P(X)$$

Using Bayesian probability terminology, the above equation can be written as:

Posterior = (Likelihood) * {Prior) / (Evidence)

[13]By using joint model for repeated applications, under the independence assumptions, the conditional distribution over the class variable C is given by:

$$\text{Posterior}=\frac{\text{Likelihood* Prior}}{\text{Evidence}}$$

$$P(C_j \mid A_1, A_2, \ldots, A_n) = \frac{\left(\prod_{i=1}^{n} P(A_i \mid C_j)\right) P(C_j)}{P(A_1, A_2, \ldots, A_n)}$$

Where the evidence P (Ai) is a scaling factor dependent only on A1,…., An, i.e. a constant if the values of the feature variables are known.

3. Decision Tree (DT):A decision tree is a classification algorithm which is represented as upside down tree structure.[14-16] Decision tree contains Root Node: it represents the data sample and this further gets divided into two or more homogeneous sets. Decision Node: it is a node when sub-node splits into further sub-nodes by applying decision rule. Leaf Node: nodes that are not split further. Splitting process: it process of dividing a node into two or more sub-nodes. Decision rules: we need to select the best decision rules to split the current node from list of decision rules. Decision tree recursively splits data until we are left with pure leaf nodes.

[17-20]Information Gain, Gini Impurity and Entropy: To obtain pure nodes, the Decision Tree should decide on optimal splits. An optimal split is that which has maximum value of Information Gain andminimum value ofGini Impurity.Entropy is measure of information contained in a state (measurement of disorder or measurement of impurity).

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Where 'Pi' is simply the frequent probability of an element/class 'i' in our dataset.Information Gain of each split is calculated by subtracting entropy of parent node with each child node.

Information Gain = entropy (parent) – [average entropy (children)]

3. Artificial Neural Networks (ANN): An Artificial neural network model is a collection of connected network of neurons like those in human brain.[21-22]Each neuron does simple mathematical operations. It receives data from other neurons. Neurons are placed in layers; a neuron from one layer receives data from neuron of other layers, modifies it and sends data to neuron of other layers. A neural network is madeup of one or more layers. The first layer is called the input layer; it receives data from outside world (for example an image or text).The last layer is called output layer. [23]The data from neurons in output layer is read and used as the output of network. The layers between input and output layers are called hidden layers where the actual processing of the inputs is done and outputs are generated.

[24]The general model of artificial neural network, the net input can be calculated as follows −

yin=x1.w1+x2.w2+x3.w3…xm.wmyin=x1.w1+x2.w2+x3.w3…xm.wm

i.e., Net input yin=∑mixi.wi

The output can be calculated by applying the activation function over the net input.

Y=F(yin)

5. K-means Clustering: K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belong to only one group that has similar properties. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training. It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the datapoints and their corresponding clusters. The measure of distance is generally Euclidean in *k*-means, which, given 2 points in the form of (x, y), can be represented as:

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

6. K-means initial centroid selection method - Sequential Row Method (SRM): Standard K-means clustering algorithm randomly assigns centroids for the K clusters. Since the grouping of similar datapoints into clusters will be done based on the initial centroids, the randomly initialized centroids may not be optimal choice and may stuck and fail to meet the convergence point. So it is advised to initialize distinct and optimal centroids [22]. Hence, we have developed a new derivative of K-means initial centroid selection method, Sequential Row Method (SRM).In this method, records of given dataset are chosen sequentially as centroids of each cluster. The x and y coordinates of a centroid are the values of all attributes, considered from each sequentially chosen record.

In generating the initial k centroids using sequential row method the following equations are used:

I = 1, 2, 3… N. I take value sequentially

Ci = X (I),

Cj = Y (I)

The initial centroid is C (Ci,Cj). N is the number of instances in the training dataset. X (I) and Y (I) are the values of the attributes X and Y receptively for the row I.

7.1 Proposed Methodology and Experimental Results

7.1.1 Integration of Sequential Row initial centroid selection Method of K-means clustering algorithm with Naive Bayes, Decision Tree and Artificial Neural Network Classification algorithms (SRM – Naïve Bayes), (SRM – Decision Tree), (SRM – ANN):

Performance Analysis of SRM – Naive Bayes, Decision Tree and Artificial Neural Network

Confusion Martix also called error matrix, is a 2x2 matrix, by evaluating the performance of a classification model.The matrix compares the actual target values with those predicted by the machine learning model. From Confusion Matrix, we calculated the erformance metrics.

Performance Metrics

Sensitivity	TP/TP+FN	0.90
Specificity	TN/TN+FP	0.31
Accuracy	TP/TP+FN	0.82

Table 1: Performance Metrics of SRM-NB

Table 1 shows that, SRM-NB model is detecting 90% of patient's positive condition and detecting 31% of patient's negative condition. SRM-NB model makes 82% correct predictions about patient's condition.

Sensitivity	TP/TP+FN	0.89
Specificity	TN/TN+FP	0.76
Accuracy	TP/TP+FN	0.88

Table2: Performance Metrics of SRM-DT

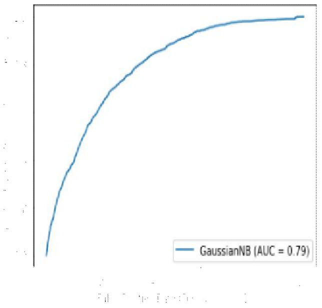
Table 2 shows that, SRM-DT model is detecting 89% of patient's positive condition and detecting 76% of patient's negative condition. SRM-DT model makes 88% correct predictions about patient's condition.

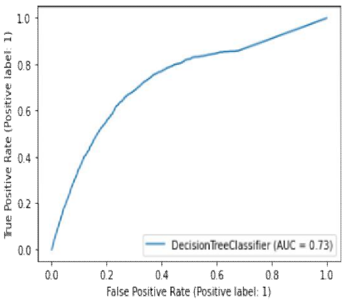
Sensitivity	TP/TP+FN	1
Specificity	TN/TN+FP	0.88
Accuracy	TP/TP+FN	0.92

Table 3: Performance Metrics of SRM-ANN

Table 3 shows that, SRM-ANN model is detecting 100% of patient's positive condition and detecting 88% of patient's negative condition. SRM-ANN model makes 92% correct predictions about patient's condition.

ROC curve and AUC score is a performance measurement for the classification problems and tells how much the model is capable of distinguishing between classes.Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease.





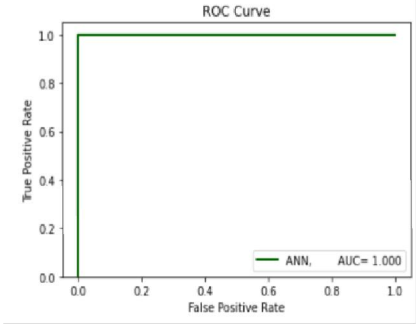


Table4: ROC curve SRM-NBTable5: ROC curve SRM-DTTable6: ROC curve SRM-ANN

Table 4 shows, SRM-NB model is capable of distinguishing the two classes with AUC score of 0.79
Table 5 shows, SRM-DT model is capable of distinguishing the two classes with AUC score of 0.73
Table 6 shows, SRM-DT model is capable of distinguishing the two classes with AUC score of 1.00

8. Comparison Analysis:

The comparison results of sensitivity, specificity, and accuracy in the diagnosis of heart Stroke using k-means clustering and three classifiers Naïve Bayes, Decision Tree and Artificial Neural network when integrated with Sequential Row initial centroids selection method are shown in Table 7.

	SRM – NB	SRM – DT	SRM – ANN
Sensitivity	0.90	0.89	1
Specificity	0.31	0.76	0.86
Accuracy	0.82	0.88	0.92
AUC score	0.79	0.73	1

Table 7: Comparison table of performance metrics of SRM integrated with NB, DT, ANN

Above comparison Table 7 shows that the model using Sequential Row Method when integrated with Artificial Neural Network has the ability to detect the high risk of heart stroke among patients accurately as its sensitivity score is 1 i.e. it is predicting the positive condition correctly. Also it has a better ability to detect the low risk of heart Stroke condition among the patients as it has specificity score of 0.86 which is the highest of the three models. The best accuracy achieved is by the model using Sequential Row Method integrated with Artificial Neural Network, showing accuracy of 92% as shown in Table 7. For choosing the best performing model, we obtain ROC curves and AUC scores of three models. With AUC score of 1, SRM-ANN is considered as the best performing model in early prediction of heart stroke of patients.

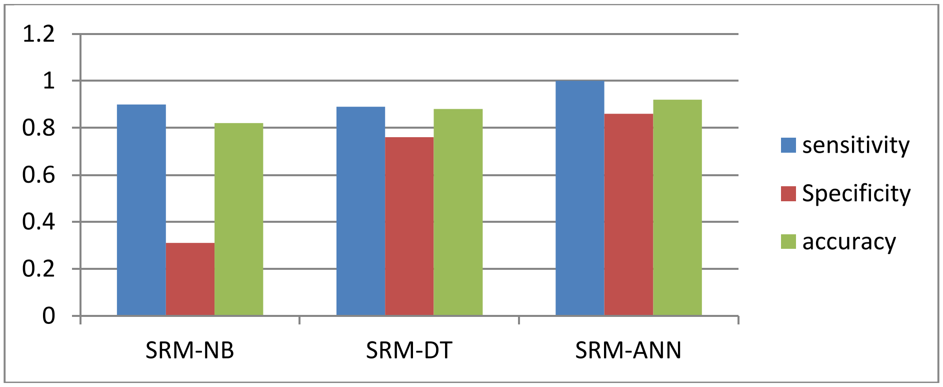


Table 8: Overall comparison of performance metrics of models using SRM integrated with NB, DT, ANN

When comparing integrating k-means clustering with Naïve Bayes, Decision Tree and Artificial Neural Network applied on the dataset, integrating k-means clustering using Sequential Row Method with Artificial Neural Network enhance the accuracy of Artificial Neural Network in diagnosing risk of heart stroke in patients as shown in overall comparison table 8.

9. Conclusion: In this paper, we are developing a smart prediction model for prognosis of risk of heart stroke of patients in heat stroke indicators dataset. We have used K-means clustering and Naïve Bayes, Decision Tree and Artificial neural network classifiers for building the prediction system. Since initial centroids have strong affect on the resulting clusters, we have investigated by integrating K-means initial centroid selction method, Sequential Row Method with three classifiers. Integrating initial centroid selection method enhances the accuracy of classifier. We developed three modls and compared them using performance metrics obtained by plotting confusion matrix of each model. As final assessment of performance evolution, we considered ROC curves and AUC score of the models. Since SRM-ANN model has sensitivity score as 1 and specificity score as 0.86 which is the highest of the three models and prove the model has the ability to classify patients into classes accurately. With accuracy 92% and AUC score of 1, SRM-ANN is a smart predicting system, developed for early intervention of heart stroke of patients.

REFERENCES

[1]https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2]https://www.who.int/nmh/publications/fact_sheet_cardiovascular_en.pdf

[3]https://www.ahajournals.org/doi/full/10.1161/str.0b013e31825bcdac#d3e2905

[4]Alpaydin, Ethem (2010). Introduction to Machine Learning. MIT Press. p. 9. ISBN 978-0-262-01243-0.

[5]Yusuf S, Reddy S, Ounpuu S, Anand S. Global burden of cardiovascular diseases: part I: general considerations, the epidemiologic transition, risk factors, and impact of urbanization. Circulation. 2001; 104: 2746– 2753. LinkGoogle Scholar

[6]O'Donnell MJ, Xavier D, Liu L, Zhang H, Chin SL, Rao-Melacini P, Rangarajan S, Islam S, Pais P, McQueen MJ, Mondo C, Damasceno A, Lopez-Jaramillo P, Hankey GJ, Dans AL, Yusoff K, Truelsen T, Diener HC, Sacco RL, Ryglewicz D, Czlonkowska A, Weimar C, Wang X, Yusuf S. Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. Lancet. 2010; 376: 112– 123. CrossrefMedlineGoogle Scholar

[7]Yusuf S, Hawken S, Ounpuu S, Dans T, Avezum A, Lanas F, McQueen M, Budaj A, Pais P, Varigos J, Lisheng L. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. Lancet. 2004; 364: 937– 952. CrossrefMedlineGoogle Scholar

[8]Heidenreich PA, Trogdon JG, Khavjou OA, Butler J, Dracup K, Ezekowitz MD, Finkelstein EA, Hong Y, Johnston SC, Khera A, Lloyd-Jones DM, Nelson SA, Nichol G, Orenstein D, Wilson PW, Woo YJ. Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association. Circulation. 2011; 123: 933– 944. LinkGoogle Scholar

[9]Mackay J, Mensah GA, Mendis S, Greenlund K; World Health Organization. The Atlas of Heart Disease and Stroke. Geneva, Switzerland: World Health Organization; 2004. Google Scholar

[12]Johnston SC, Mendis S, Mathers CD. Global variation in stroke burden and mortality: estimates from monitoring, surveillance, and modelling. Lancet Neurol. 2009; 8: 345– 354. CrossrefMedlineGoogle Scholar

[13]https://pubmed.ncbi.nlm.nih.gov/28287467/

[14]https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/heart-disease-stroke#:~:text=Having%20diabetes%20means%20that%20you%20are%20more%20likely,such%20as%20high%20blood%20pressure%20or%20high%20cholesterol.

[15]https://www.ahajournals.org/doi/10.1161/strokeaha.107.505867

[18]https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113
[19]https://www.cdc.gov/nchs/icd/icd9cm.htm#:~:text=The%20International%20Classification%20of%20Diseases%2C%20Ninth%20Revision%2C%20Clinical,associated%20with%20hospital%20utilization%20in%20the%20United%20States.
[20]https://ieeexplore.ieee.org/document/8947802
[21]Jordan, Michael I.; Bishop, Christopher M. (2004). "Neural Networks". In Allen B. Tucker (ed.). Computer Science Handbook, Second Edition (Section VII: Intelligent Systems). Boca Raton, Florida: Chapman & Hall/CRC Press LLC. ISBN 978-1-58488-360-9.
[22] https://www.geeksforgeeks.org/activation-functions-neural-networks/
[23]Jordan, Michael I.; Bishop, Christopher M. (2004). "Neural Networks". In Allen B. Tucker (ed.). Computer Science Handbook, Second Edition (Section VII: Intelligent Systems). Boca Raton, Florida: Chapman & Hall/CRC Press LLC. ISBN 978-1-58488-360-9.
[24]Artificial Neural Network - Basic Concepts (tutorialspoint.com)