International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-11 Issue-1, May 2022

A Predictive Model of Stroke Diseases using Machine Learning Techniques

Alaa Ghannam, Jaber Alwidian

Abstract: ***Due to rapid changing in human lifestyles, a set of biological factors of human lives has changed, making people more vulnerable to certain diseases such as stroke. Stroke is a life-threatening disease leading to a long-term disability. It's now a leading cause of death all over the word. As well as it's the second leading cause of death after ischemic heart disease in Jordan. Stroke detection within the first few hours improves the chances to prevent complications and improve health care and management of patients. In this study we used patient's information that are believed to be related to the cause of stroke and applied machine learning techniques such as Naive Bayes, Decision Tree, and KNN to predict stroke. Orange software is used to automatically process data and generate data mining model that can be used by health care professionals to predict stroke disease and give better treatment plan. Results show that decision tree classifier outperformed other techniques with accuracy level of 94.2%.***

Keywords ***Data Mining, Classification, Stroke, Healthcare, Machine Learning.***

I. INTRODUCTION

Data mining and machine learning are a growing field of computer science including the improvement of calculations that figure out how to make forecasts dependent on information, they have various rising applications in the field of bioinformatics. (Aishwarya Roy, 2018). Information gain from health data may lead to innovative solution or better treatment plan for patients. In order to gain knowledge intelligently from stroke data, a data mining techniques are applied to automatically process data and generate data mining model that can be used by health care professionals to predict the symptoms of stroke. (Ohoud. M, 2018) A large number of people lose their life due to stroke, it's now the leading cause of death all over the world, an information from the Jordanian ministry of health (2007), shows that cardiovascular (CVD) diseases are the most common cause of death in Jordan, among CVDs, stroke is second leading cause of death after ischemic heart disease and is responsible for 30% of the CVD deaths in Jordan. The earlier a stroke is detected, the better the odds of preventing symptoms and improving patient safety and treatment.

Manuscript received on 21 March 2022.
Revised Manuscript received on 05 April 2022.
Manuscript published on 30 May 2022.

* Correspondence Author

Alaa Ghannam*, MBA Student, Talal Abu Ghazaleh University Collage for Innovations (TAGUCI), Amman, Jordan. E-mail: alaa.shghannam@gmail.com

Jaber Alwidian, Assistant Professor, Department of Data Science and Artificial Intelligence, University of Petra (UOP), Amman, Jordan. E-mail: Jaber.alwidian@uop.edu.jo

The machine learning algorithm helps to understand the relation between patient's medical history and symptoms to predict strokes. Stroke happens mainly due to people's lifestyle changing factors such as high blood sugar, heart disease, obesity, diabetes. Other demographic factors showed direct relation with stroke occurrence such as type of works, whether the patient is/was working in public or private sector, place of residency (city or country side), and marital status. (Eshah, 2013) In this study, we have used several machine learning algorithms to predict stroke like naïve Bayes, KNN, and decision tree. A ready data set was used to train the machine and generate data mining model.

II. PROBLEM STATEMENT

The problem statement of this study is applying data mining classification algorithms to help physicians predict stroke disease among different patient's symptoms early and clearly.

III. LITERATURE REVIEW

3.1 Related Works

Many works have been done in the field of predicting biomedical diseases using data science, in this section we will shed light on some previous works done by researcher for the same concept of predicting stroke diseases using data mining algorithm.

A.Sadha et al. in his work, applied a predictive model for stroke disease and adopted classification algorithm like, Decision Tree, Naive Bayes and Neural Network. A set of attributes were used that include patient information like age, blood pressure, hypertension, and patient history information. Principle component analysis algorithm was used for reducing the attributes. Feature subset selection is used for feature reduction to remove the irrelevant data and choose the data which relates to stroke disease. In his study sensitivity and accuracy indicators for evaluation the performance of the 3 algorithms were used, results show that neural network performance is having more accuracy, comparing with other two classification methods.

(A.Sudha, 2012). K.Priya et al. in his work, used Hybrid Neuro-Genetic approach to predict stroke disease, the study utilizes Artificial Neural Network based to predict stroke disease by improving the accuracy with higher consistent rate using optimized hidden neurons. This algorithm determines the attributes involving more towards the prediction of stroke disease. A data of 300 patients was collected among that 180 patients having disease. (K. Priya, 2013). S.

Ashkokan et al. designed a web application to predict stroke occurrence using R language in RStudio for conducting data analysis and for the construction of the predictive application. The user is able to enter their information to test themselves for stroke occurrence. Feature selection and as the variables gained are used to choose the most efficient and accurate variables required to predict stoke. Then different algorithms to perform the predictive modeling are applied such as Random Forest, Decision tree, Logistic Regression and support Vector Machines, the web application is programmed to process users inputs to predict stroke using the most accurate model. (Soodamani A., 2020). Duen Y. et al. used the classification technology to build a predictive model to improve the accuracy of cerebrovascular disease diagnosis (stroke is one of the cerebrovascular diseases), the study adopted three classification algorithms: Bayesian classifier decision tree and back propagation neural network. Based on the accuracy and sensitivity of each technique, decision tree constructed model gives the optimum predictive model for cerebrovascular diseases diagnosis comparing with the other two techniques. In this study 16 classification rules were extracted, five specialized physicians assessed and tested these rules and confirmed their usefulness. (Duen Y., 2011).
Data analytics is the process of analyzing raw data to find trends and derive conclusions; people are often expected to make mistakes during traditional analysis, especially when there are a lot of data that need to understand the relations among them. But in medical field mistakes means lives. Machine learning techniques can improve the efficiency of systems to better predict results. (Ohlhorst, 2013).

3.2 Data mining & Machine Learning

Data mining is the most significant application for Machine Learning; the method of collecting & extracting information from vast volumes of data is known as data mining. The information or expertise obtained in this manner can be applied to a variety of applications, especially in the medical and health care fields, where early detection of disease enhances curing hopes.

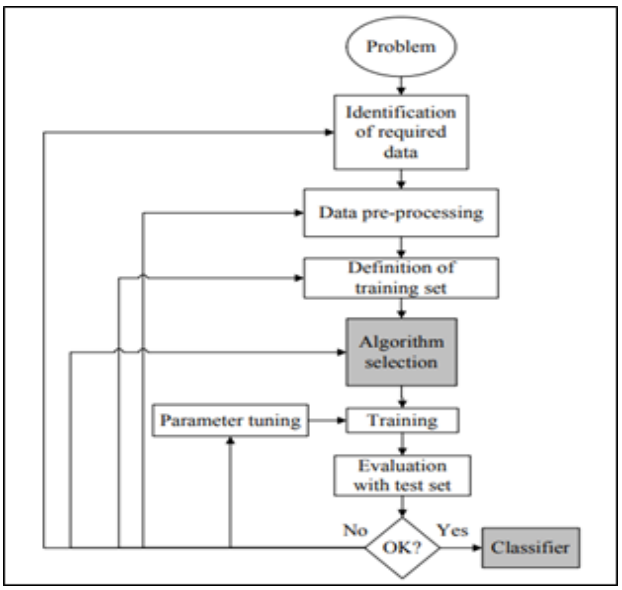

Fig.1: The Process of Supervised Machine Learning

In any dataset used by machine learning, there are many instances; every instance is represented using a set of features. The features may be categorical, continuous, or binary. If the instances in the dataset have a define class (the class which will define the results of a set of attributes together), then the learning is called supervised as shown in ***(Figure 1)*** which describes the process of supervised machine learning, when the attributes are not defined with a class; it's called unsupervised learning (Jain, 1999).
Popular data mining approaches are: Classification, Clustering, Association Rules, Regression and Associative Classification. For the purpose of our study, and as the class in our data is defined (has a stroke or not having a stroke); the best approaches to adopt is the classification techniques.

IV. CLASSIFICATION

The most popular supervised learning in data mining strategy is classification, which uses a collection of pre-classified samples to construct a model that can identify a wide set of data. This method allows the extraction of valuable data knowledge.

Classification technique is a two-step process:

1. A *Learning* step: a training data are analyzed by a classification algorithm. Here, the class label attribute is having a stroke or not, and the learned model or classifier is represented in the form of classification rules.
2. A *Classification* step: part of the data are used as a test data to predict the accuracy of the classification rules, if the accuracy is acceptable, the rules can be applied to classify the new data tuples. (Jiawei Han, 2012).

A ready dataset was used in this study to predict stroke. A set of attributes that believed to have direct effect on stroke diagnosis are provided, the last column which indicates (has a stroke/ doesn't have a stroke) is the class label attribute.

4.1.1 Classification Techniques:

In this study, three classification techniques were used: Decision Tree, Naïve Bayes and K-Nearest Neighbor. A classifier evaluation measures will be used to compare classification techniques applied to data under study.

4.1.2 Naive Bayes

Naive Bayes is a classification technique based on Bayes' theorem, it predicts class membership probabilities to create classification model. All naive Bayes classifiers presume that the value of a given attribute is independent of the value of any other attributes for a given class.

4.1.3 K-Nearest Neighbor (kNN)

The KNN algorithm is a simple, direct supervised learning algorithm that can be applied to both regression and classification problems. The KNN algorithm suggests that objects that are identical are close together. It makes an "accurate judgment" on how an unclassified point could be classified using test results.

Retrieval Number: 100.1/ijrte.A69000511122
DOI: 10.35940/ijrte.A6900.0511122
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

4.1.4 Decision tree

Decision tree is a supervised predictive model, each branch of the tree is viewed as a classification problem, and the leaves of the trees are viewed as divisions of the dataset relevant to that given criterion. This methodology is mostly used for exploratory analysis, data pre-processing, and prediction work.

(Figure .2) below shows a synopsis of the decision tree showing stroke disease probabilities based on given attributes.

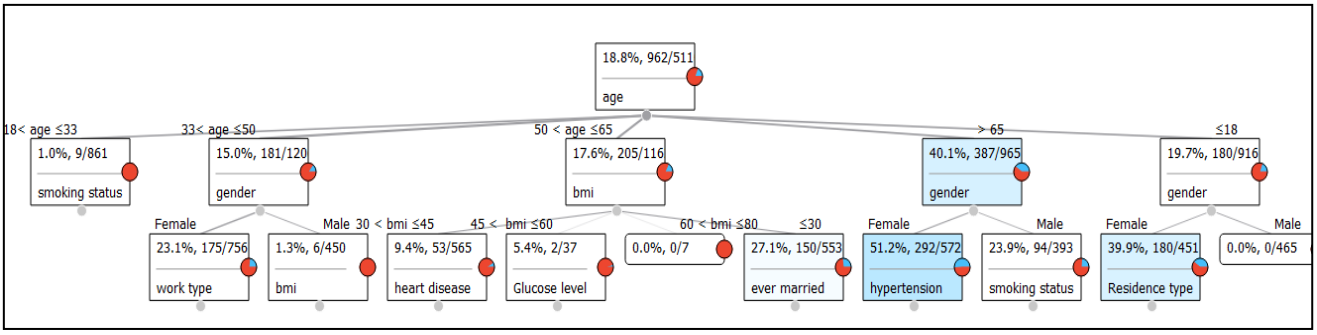

Fig. 2: Synopsis of the decision tree showing Stroke disease probabilities

V. EXPERIMENTAL STUDY:

A set of published data for patients' information needed for stroke prediction was downloaded from www.Kaggle.com and processed using Orange Tool. Based on input parameters such as gender, age, various diseases, and smoking status, this dataset is used to predict if a patient is likely to have a stroke or not.

The dataset contains 5110 instances and 11 attributes. Three classification algorithms were used; Decision Tree, Naïve Bayes and KNN. A survey was distributed to collect new dataset as an application for prediction of stroke disease. The prediction will be applied for the created models from the three classification techniques that can be used by health care professionals to predict stroke disease and give better treatment plan.

5.1 Orange Tool

Orange is an open source machine learning and data visualization tool, its built using Visual programming and Python programming language. The orange tool has a canvas user interface where the user can drag and drop widgets and construct data analysis frameworks. Data mining widgets provides essential functionalities such as reading data, showing tables, selecting features, training predictors, comparing learning algorithm, visualization elements.

Via a template-based system, the Orange tool offers 99 percent of an innovative analytical approach, speeding implementation and reducing mistakes by nearly removing the need to write code. (G., 2019)

5.2 Dataset Details

Stroke is the second leading cause of death worldwide, accounting for about 11% of all deaths, according to the World Health Organization (WHO).

This study used published data for stroke prediction "Stroke Prediction Dataset" from repository database www.Kaggle.com, the data contains 5110 instances and 11 attributes .Attributes are varied from medical information such as (hypertension, heart disease, glucose level, body mass index) to non-medical information that is believed to be relevant to cause stroke, such as (age, gender, work type, resident type, smoking status).

(Table 1) below shows the description of each attribute used to predict stroke.

5.3 Dataset Pre-processing

The dataset is pre-processed to make it ready for mining process, missing data was handled by matching missing attributes in the same class, to avoid overfitting; continuous attributes were discretized into categorical attributes using entropy –gain discretization method.

Examples of discretization step are:

1. **Age:** this attribute was discretized for 5 age categories (≤18, 18< age ≤33, 33< age ≤50, 50 < age ≤65 & > 65) .
2. **Average glucose level** in blood was discretized into two categories (high / low).
3. **Body mass index** was discretized into 5 categories (≤30 , 30 < bmi ≤45 , 45 < bmi ≤60 , 60 < bmi ≤80 & >80).

Table 1 Attributes Description

#	Attribute	Description
1.	Patient ID	Unique Identifier
2.	Gender	Male / Female
3.	Age	(≤18, 18< age ≤33, 33< age ≤50, 50 < age ≤65 & > 65)
4.	Hypertension	Yes/No
5.	Heart disease	Yes/No
6.	Ever married	Yes/No

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

7.	Work type	Children/ Govt job/ Never worked/ Private/ or "Self-employed"
8.	Residence type	Rural/ Urban
9.	Glucose level	Average glucose level in blood; it was discretized into two categories high / low.
10.	BMI: body mass index	(≤30 , 30 < bmi ≤45 , 45 < bmi ≤60 , 60 < bmi ≤80 & >80)
11.	Smoking status:	Ex-smoker/ never smoker / current smoker
12.	Stroke	Yes/No (Class Attribute)

5.4. Models Evaluation and Selection

Using Orange Tool, we've trained the available data file in order to build a learning model that predicts future unknown instances as shown in (***Figure 3***).

After processing the data using the three classification techniques; Decision tree, Naïve Bayes and KNN, the models were evaluated and best model was selected for prediction stroke based on cross validation on 10 folds sampling method.

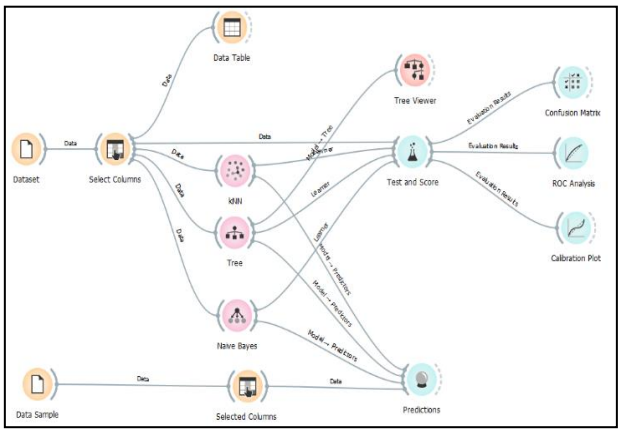

Fig.3: Dataset training using 3 classification techniques

Cross validation as shown in ***(Figure .4)*** below; splits the data into a given number of folds (here 10). The algorithm is tested by excluding examples from one fold at a time; the model is derived from other folds, and the examples from the excluded fold are categorized. This is done with each of the folds. Data folding can be controlled by parameters like ensuring that each fold has the same amount of Instances with a given categorical value, such as the class label .This is called **stratified** cross-validation.

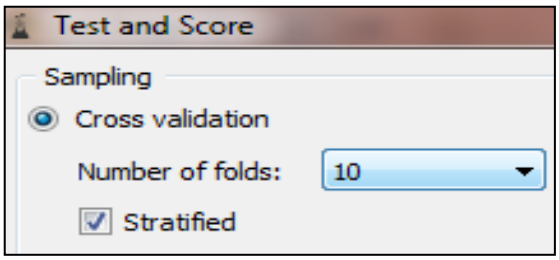

Fig.4: Cross Validation

The number/proportion of instances between the expected and actual class is given by the Confusion Matrix. The resulting instances are fed into the output signal by selecting the elements in the matrix. This allows us to see which individual cases were misclassified and how they were misclassified.

(***Table 2***) shows a basic confusion matrix for a two- class case; Positive and Negative.

Table 2: Confusion matrix of the actual and predicted class label

		Predicted class: Positive	Predicted class: Negative
Actual class	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

The classifier evaluation measures are:

1. ***Accuracy***: is the proportion of correctly classified examples , as shown in ***(Equation 1)*** below:

$$\text{Accuracy} = \frac{TN+TP}{TN+FP+FN+TP} \quad \text{....Equation} \quad (1)$$

2. ***Recall (Sensitivity)***: is the proportion of true positives among all positive (actual) instances in the data , as shown in ***(Equation 2)*** below:

$$TPR = \frac{TP}{FN+TP} \text{....Equation} \quad (2)$$

3. ***Specificity (True Negative Rate)***: is the proportion of true Negative among all positive (actual) instances in the data ,as ***shown*** in ***(Equation 3)*** below:

$$TNR = \frac{TN}{TN+FP} \text{....Equation} \quad (3)$$

4. ***Precision:*** is the proportion of true positives among instances classified (predicted) as positive , as shown in ***(Equation 4)*** below:

$$\text{Precision} = \frac{TP}{FP+TP} \text{....Equation} \quad (4)$$

5. ***F1***: is the weighted harmonic mean of precesion and recall; it accounts for both false positives and false negatives , as shown in ***(Equation 5)*** below:

$$F1 = 2 \text{ x} \frac{precision*recall}{precision+recall} \text{....Equation} \quad (5)$$

VI. RESULTS & DISCUSSION

6.1 Confusion matrices

In (***Tables 3, 4 & 5***) below we can see how different classification techniques can generate different predictions; for example Decision tree predicted 835 "Yes" class from actual of 962 "Yes" class instances, kNN predicted 785 while Naïve Bayes made only 228 correct predictions for the "Yes" class.

Retrieval Number: 100.1/ijrte.A69000511122
DOI: 10.35940/ijrte.A6900.0511122
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Table 3: Confusion matrix for Decision Tree Model

		Predicted class		
		Yes	**No**	∑
Actual class	**Yes**	835	127	962
	No	169	3979	4148
	∑	1004	4106	**5110**

Table 4: Confusion matrix for kNN Model

		Predicted class		
		Yes	**No**	∑
Actual class	**Yes**	785	177	962
	No	302	3846	4148
	∑	1087	4023	**5110**

Table 5: Confusion matrix for Naïve Bayes Model

		Predicted class		
		Yes	**No**	∑
Actual class	**Yes**	228	734	962
	No	167	3981	4148
	∑	395	4715	**5110**

6.2 Evaluation Measures

The following ***(Table 6)*** & ***(Figure 5)*** summaries evaluation measures for the three classification techniques.

Table 6: Test & Score (Evaluation Results)

Model	AUC	CA	F1	Precision	Recall
Tree	0.935	0.942	0.849	0.832	0.868
kNN	0.947	0.906	0.766	0.722	0.816
Naïve Bayes	0.805	0.824	0.336	0.557	0.237

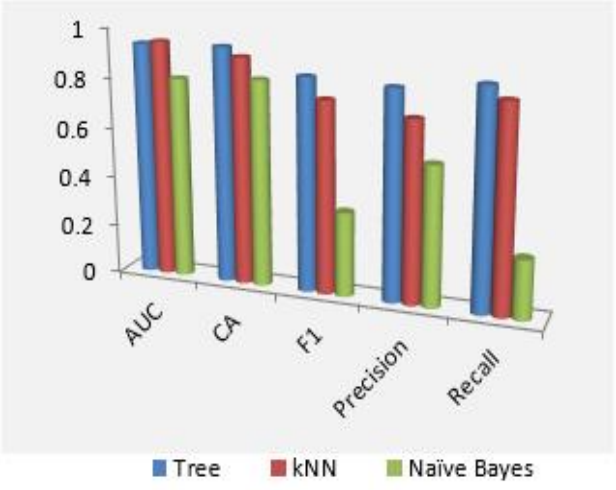

Figure 5: Test & Score

6.2.1 Classification Accuracy (CA)

Accuracy is the number of correct forecasts divided by the sample size. By displaying the likelihood of the true value of the class label, accuracy estimates the algorithm's effectiveness. Below (***Table 7***) compares classification errors based on calculated classification accuracy for the three classification techniques.

Table 7: Classification Accuracy

Classifier	Classification Accuracy (CA)	Error rate %
Decision Tree	94%	6%
kNN	91%	9%
Naive Bayes	82%	18%

As indicated above, Decision tree gave correct predictions in 94% of the total cases; kNN model generated 9% misclassified instances while Naive Bayes misclassification rate was the highest with 18% error.

6.2.2 Precision

Precision is the ability of a classification model to identify only the relevant data points; it assesses the predictive power of the algorithm. We can observe from ***(Table 6)*** that the three classifiers precision was lower than accuracy that indicates the weight of false positives in the predicted model. Also we can conclude that Naïve Bayes is a low precision model as the false "Yes" labeled instances is around 43% of the predicted classes. Best precision model was decision tree ;as around 83% of the predicted "Yes" labeled instances were actually "Yes" class instances.

6.2.3 Recall (Sensitivity)

Recall is the percentage of positive examples, from the entire set of actual positive examples, the model was able to identify. (It works on the class level) In our case, Decision tree algorithm succeeded in predicting the class in 86.8% of the cases of the yes "stroke" class instances (Sensitivity) as indicated in ***(Table 6)*** while the percentage was 95.9% in predicting the "no stroke" class (Specificity).

kNN model sensitivity was 81.6% & model specificity was 93.3% , Naïve Bayes model generated 23.7% sensitivity and 96% specificity , that means that Naïve Bayes model was good at predicting cases with no disease (no stroke diagnose true guesses) while its ability to detect disease was very low.

6.2.4 F1 Score

The harmonic correlation of precision and recall is the F1-score. Accuracy is a good measure used when the True Positives and True negatives are more important while F1-score is used when the False Negatives and False Positives are crucial, When the class distribution is equivalent, accuracy can be used, but F1-score is a better metric when there are imbalanced classes, as in the example above, our results shows very low F1 score of the Naïve Bayes algorithm 33.6% due it's low recall value 23.7% ***(Table 6).***

6.2.5 AUC

AUC is the area under the ROC (Receiver Operator Characteristic) curve that represents the tradeoff between Recall (TPR) and Specificity (FPR); the higher the AUC, the better the performance of The model's in differentiating between positive and negative classes.

It is evident from the plot ***(Figure 6)*** below that the AUC for the kNN ROC curve is higher than that for the Decision Tree ROC curve. Therefore, we can say that kNN algorithm did the best job of classifying the positive class in the dataset while Naïve Bayes algorithm has the highest classification errors.

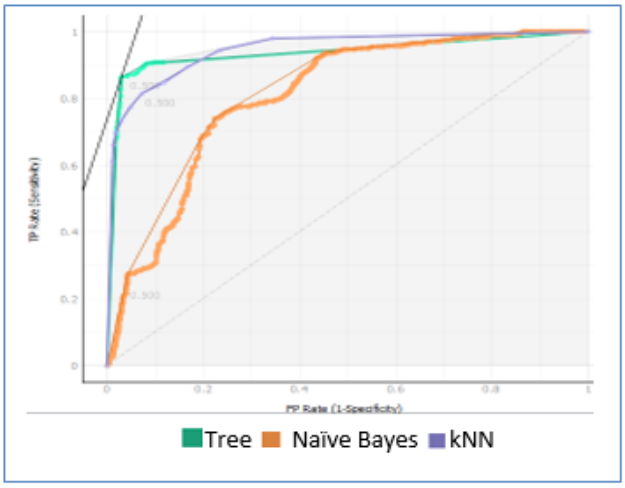

Fig .6: ROC Analysis

6.2.6 Calibration Curve

It is essential to have a model with good accuracy but also that is well calibrated.

Calibration curve is "a display of the match between classifiers' predictions probability and actual class probabilities" When a model goes below the diagonal, the model is over-forecasting, above the diagonal, the model is under-forecasting.

From the below plot ***(Figure 7)*** we can tell that the kNN & decision tree models are over-forecasting the "Yes" class as total number of predicted yes class is more than actual "Yes" class instances.

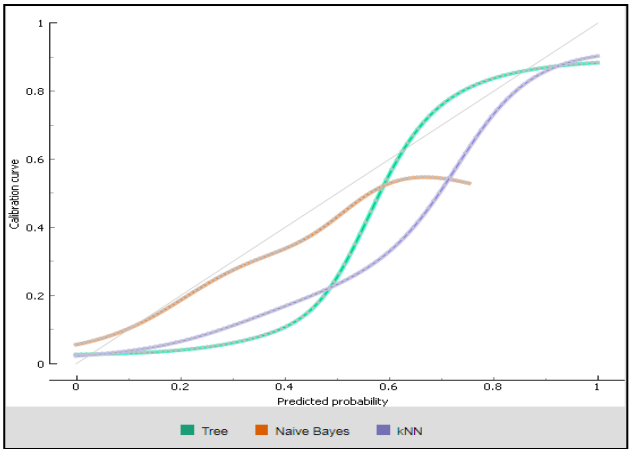

Fig .7: Calibration Curve

VII. CONCLUSION & FUTURE WORK

This paper has provided the detailed review of classification techniques for diagnosis of stroke disease, the study has been examined with the Decision tree, Naïve Bayes, KNN, methods using 5110 samples from the stroke disease data set. The observations were analyzed and discussed. In the discussion it was found that Decision tree algorithm had maximum accuracy (94.2%) and minimum error rate. Also it was noticed that AUC value for the kNN model was better than other models indicating that kNN is the best classifier in terms of distinguishing between positive & negative classes. When comparing algorithms; precision, recall and F-measure values were connected together and have maximum value in decision tree model. Application was created based on anonymous survey covering same studied attributes on real persons in order to predict the probability of having a stroke; their answers for questions were processed using Orange predictions tool. Predicted results were classified in 2 categories, yes if expected to have stroke and no if not. It was noticed that (51.2%) of the females above 65 years have a higher tendency to have stroke if compared with males in the same age category or males & females in other age categories. For the future work, enhancement in classification procedures would improve accuracy and, as a result, models ability to be used as computer-aided diagnosis, where more robust approaches are being developed.

REFERENCES

1. A.Sudha, P. N. (2012). Effective Analysis and Predictive Model of Stroke Disease using Classification Methods. India: International Journal of Computer Applications.
2. Aishwarya Roy, A. K. (2018). 'Stroke prediction using decision trees in artificial intellegence'. internationl journal of advance research, ideas and innovations in technology, volume 4, issue 2.
3. Duen Y., C. H. (2011). A Predictive Model for Cerebrovascular Disese Using Data Mining. Expert System With Application, 38. Retrieved from www.elsevier.com/locate/eswa .
4. Eshah, N. F. (2013). Knowledge of Stroke and Cerebrovascular Risk Factors Among Jordanian Adults. Journal of Neuroscience Nursing, 45(4).
5. G., A. (2019). Orange Tool Approach for Comparative Analysis of Supervised Learning Algorithm in the Classification Mining. Journal of Analysis and Computation, 12(1), 1-10.
6. Jain, A. M. (1999). Data clustering: A review, ACM Computing Surveys.
7. Jiawei Han, M. K. (2012). Data Mining, Concepts and Techniques (3rd ed.). USA: Elsevier Inc.
8. Jordanian Ministry of Health Publications. (2007). Behaviours and risk factors of non-communicable diseases in jordan. Jordan: Directorate of Disease Control and Prevention.
9. K. Priya, T. R. (2013). Predictive Model of Stroke Disease Using Hybrid Neuro-Ginitic Approach. India: International Journal of Engineering and Computer Science .
10. Ohlhorst, F. (2013). Big Data Analytics-Turning Big Data Into Big Money. Canada: John Wiley & Sons, Inc.
11. Ohoud. M, R. s. (2018). Prediciton of Stroke using Data Mining Classifiaction Techniques. International Journal of Advances Computer Science and Application, 9(1).
12. Roger, V. G.-J. (2011). Heart Disease and stroke Statistics Subcommittee. American Heart Association.
13. Soodamani A., S. G. (2020). An Effecrive Stroke Prediction System Using Predictive Models. International Research Journal of Engineering and Technology (IRJET), 07(03). Retrieved from www.irjet.net.
14. Stroke Prediction Dataset. Kaggle. Available at: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset [Accessed May 15, 2021].

15. Khosla, A., Cao, Y., Lin, C.C.Y., Chiu, H.K., Hu, J. and Lee, H., 2010, July. An integrated machine learning approach to stroke prediction. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 183-192)
16. Letham, B., Rudin, C., McCormick, T.H. and Madigan, D., 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. Annals of Applied Statistics, 9(3), pp.1350-1371.
17. Al-Buhairi, A.R., Phillips, S.J., Llewellyn, G. and Jan, M.M., 1998. Prediction of infarct topography using the Oxfordshire Community Stroke Project classification of stroke subtypes. Journal of Stroke and Cerebrovascular Diseases, 7(5), pp.339-343.
18. Rosado, J.T. and Hernandez, A.A., 2019, October. Developing a Predictive Model of Stroke using Support Vector Machine. In *2019 IEEE 13th International Conference on Telecommunication Systems, Services, and Applications (TSSA)* (pp. 35-40). IEEE.
19. Miranda, E., Aryuni, M. and Irwansyah, E., 2016, November. A survey of medical image classification techniques. In *2016 International Conference on Information Management and Technology (ICIMTech)* (pp. 56-61). IEEE.
20. Joshi, S. and Nair, M.K., 2018. Survey of Classification Based Prediction Techniques in Healthcare. *Indian Journal of Science and Technology*, *11*(15), pp.1-19.
21. Phyu, T.N., 2009, March. Survey of classification techniques in data mining. In *Proceedings of the international multiconference of engineers and computer scientists* (Vol. 1, No. 5).

AUTHORS PROFILES

Alaa Ghannam, is MBA student at Talal Abu Ghazaleh University Collage for Innovation, She received her B.Sc. degree in Chemical Engineering from Balqa Applied University in 2008, and she has 14 years experience in the fields of chemicals sales & technical support.

Jaber Alwidian, holds a PhD in Computer Science (The University of Jordan). He received his B.Sc. degree in Computer Information System from Philadelphia University and M.Sc. degree in Information System from the University of Jordan in 2005 and 2010, respectively. He has about seven years of work experience as a lecturer and two years as a big data scientist and consultant (INTRASOFT Middle East/big data department). His research interests are data mining, big data, data analytics, data engineering, bioinformatics, software engineering and image processing. He is implementer for more than 10 projects in data science and big data platforms.