Comparing the Random Forest Algorithm with K-Nearest Neighbor Algorithm for a Novel Stroke Prediction

Kamal nadh P[1] Dr.Saraswathi S[2]

[1]Research scholar, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India, Pincode:602105.

[2*]Project Guide, Corresponding Author, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India, Pincode: 602105.

ABSTRACT

Aim: The main objective of this article is to improve the accuracy in stroke prediction using a Novel K-Nearest Neighbor machine learning algorithm compared with Random Forest algorithm. **Materials and methods:** The two groups that have been used in this paper are Novel K-Nearest Neighbor algorithm and Random Forest algorithm .The dataset consists of over 5000 records of patients' medical and personal records. Here the pretest power analysis was carried out 80% and CI value is 95%, the sample size for the two groups N=10 iterations. **Result:** The Novel K-Nearest Neighbor Algorithm 95.70 % in predicting the stroke prediction dataset used whereas Random forest 94.80%. There exists a statistically insignificant difference between the two groups (p=0.204; p>0.05). Hence Random forest is better than K-nearest neighbor. **Conclusion:** The performance of Novel K-Nearest Neighbor algorithm is better when compared with Random forest is terms both precision and accuracy

Keywords: Machine learning, Novel K-Nearest Neighbor, Random Forest, Stroke prediction, Classification, Decision Tree.

INTRODUCTION

A stroke is a major health related challenge. strokes also known as cerebrovascular accidents consist of neurological disease.stroke occurs is the blood vessel that supplies blood to the brain this medical condition causes disability and mortality amongst adults worldwide. Annually Stroke affects 16 million individuals worldwide and society costs (Sirsat, Fermé, and Câmara 2020). Severely imbalanced classes could be effectively avoided, which is crucial for accurate prediction (Kelleher, Namee, and D'Arcy 2020). Stroke tratement is not risk -free , physicians with treatment when the potential benefits outweigh the perceived risk (Lin et al. 2020)).

In this research of estimating the Stroke Prediction of paper published in IEEE Xplore 98 articles and the number published in google scholar are 1800 (Kelleher, Namee, and D'Arcy 2020). In paper Electronic medical claims database present it self as a valuable data source due to its large scaled and longitudnal nature so of data collection process along the variety in the record patients health -related information with variety in the record patients health related information

synthesized performance of these prediction methods improved significantly after the data balancing process, indicating that it is unwise to perform prediction with an imbalanced dataset with data balancing techniques (Monteiro et al. 2018). We present a complete list of the numerous risk variables for stroke prediction. We analyze the various factors present in the present in the Electronic Health record records of patients, to identify the most important factor for s stroke prediction we are used for dimension reduction techniques to identify patterns (Sirsat, Fermé, and Câmara 2020; Bandi, Bhattacharyya, and Midhunchakkravarthy 2020). In paper (Liu, Fan, and Wu 2019; Lin et al. 2020) Machine learning algorithm such as Decision Treealgorithm such as support vector machine, Random forest, Logistic regression as 89%. In the study (Yilmaz et al. 2021) Logistic Algorithm, Support vector machine, Decision Tree, Random Forest 93%. The explainable model appoarch was conducted with Random forest and K-nearest neighbor machine having an accuracy 90%, Random forest 86% respectively. In paper (Liu, Fan, and Wu 2019; Lin et al. 2020) Machine learning algorithm such as Decision Tree algorithm such as support vector machine,Random forest,Logistic regression as 89%. In the study (Monteiro et al. 2018; Lin et al. 2020), Logistic Algorithm, Support vector machine, Decision Tree, Random Forest 93%. (Yilmaz et al. 2021); (Bandi, Bhattacharyya, and Midhunchakkravarthy 2020). The design of a classification for associated medical data visualization and management for neurologists in a stroke clustering and prediction system. (Venu and Appavu 2021; Gudipaneni et al. 2020; Sivasamy, Venugopal, and Espinoza-González 2020; Sathish et al. 2020; Reddy et al. 2020; Sathish and Karthick 2020; Benin et al. 2020; Nalini, Selvaraj, and Kumar 2020)

Previously our team has a rich experience in working on various research projects across multiple disciplines(Venu and Appavu 2021; Gudipaneni et al. 2020; Sivasamy, Venugopal, and Espinoza-González 2020; Sathish et al. 2020; Reddy et al. 2020; Sathish and Karthick 2020; Benin et al. 2020; Nalini, Selvaraj, and Kumar 2020).From the study it is necessary to improve the accuracy to predict the stroke. The main objective of this article is to improve the accuracy in stroke prediction using a Novel K-Nearest Neighbor machine learning algorithm compared with Random Forest algorithm.

MATERIALS AND METHODS

The experiment was conducted at Machine Learning Laboratory, Department of Computer Science & Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai. Two supervised machine learning algorithms were considered as two groups, the Novel K-Nearest Neighbor algorithm and Random Forest algorithm on performing N= 10 iteration on each algorithm with sample size can be N=10 to identify various sales. The G-power test which is an estimate of the statistical power of statistical test 80% alpha error rate is a type -Error (Sirsat, Fermé, and Câmara 2020). Considered as 0.05 which gives the difference between two algorithms

considered. The Enrollment ratio in the research is about 1 based on the input data. The dataset collected from kaggle web site and dataset contains 11 clinical parameters for predicting stroke events, including gender, age, hypertension, heart disease, marital status, worktype, residence type, glucose level, smoking, heart disease, bmi. There are 5110 occureences of patient information in the dataset.

Pseudocode for k-nearest neighbor Algorithm

Input: Training dataset
Output: Classifier accuracy

1.Load the training and test data
2.choose the value of k
3 the calculated n Euclidean distance in non decreasing order
4.Among these k -nearest neighbors,count the number of the data points in each category
5. Assign the new data points to that category for which the number of the neighbor is maximum

K-Nearest Neighbor

K-nearest neighbor algorithm is supervised machine learning algorthium That can be used to solve both classification algorithms. It is used to predict the correct class for the test data by calculating the distance between test and all the traning point

Random forest Algorithm

Random forest is the supervised learning technique that can be used for both classification and regression problems. The Random forest fits the number of decision tree classifiers on various samples of the dataset and uses averaging to improve the prediction accuracy and control over-fitting. The sub-sample size is controlled with a max-sample parameter if bootstrap=true(default), otherwise the whole dataset to bulid each tree. The Gini index shows that or mentioned below formula (1) decides how nodes on a decision tree branch.

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2 \quad \text{------(1)}$$

Pseudocode for Random forest Algorithm

Input: Training dataset
Output: Classifier accuracy

1. Import and read the dataset
2. Select the feature randomly
3. Generate the RF classifier criterion as a parameter
4. gini was used as a parameter value
5. Construct a decision tree using RF classifier criterion as a parameter for the result for every sample
6. Voting was performed for every predicted result.
7. Most voted prediction results were selected as the final out come

The platform used to evaluate the machine learning algorithm was google colab. The hardware used to perform the work was Intel-5 with a RAM size of 8GB. The system type used a 32 bit Windows OS X based processor with an HDD of 1TB. The operating system used was Windows 10, and the tool used was google collab with python programming language. We had different package and libaray like Numpy, Matplot, Sklearn, seaborn that are present with python programming language.The testing procedure was split the data into train and test data and then implemented the machine learning classifier to build and train model on our data After training the prediction are made and the performance of the model is evaluated using avaliable metrics.The dataset for stroke prediction which is collected from kaggle. Data preprocessing was performed to gain some context on the data using statistical analysis techniques. Data cleaning method to gain some context about the data using statistical analysis techniques. Data cleaning methods such as removing unnecessary attributes, filling missing values are done. Data exploration gives us some context and valuable insight into the dataset. The comparison of K-nearest neighbor and Random forest

Statistical Analysis:

The analysis was done using IBM SPSS(Statistical Package for social science) version 25 (Sirsat, Fermé, and Câmara 2020). The independent variables were gender, ever-married, work_type, residence-type and the dependent variables were heart_disease, bmi and hypertension, average_gulcose_levels. For Random forest, Support vector machines were done with a maximum of 20 samples and for each iteration the predicted accuracy was noted for analyzing accuracy the value obtained from the iteration Independent Sample T-test was performed.

RESULTS

In Table 1 it was observed that the K-Nearest Neighbor algorithm gave us an accuracy of 95.70% Significantly better than Random Forest gave an accuracy of 94.80%. Each algorithm has 10 iterations for each algorithm and the accuracy varies for different test sizes in decimal values.

Table 2 represents the mean of the K-Nearest Neighbor (93.85%) which is better compared with the Random Forest (91.03%) with a standard deviation of 0.380 and 0.367 respectively also obtained a standard error mean rate of 0.120 whereas the Random Forest algorithm obtained an error mean rate of 0.84. The significance value 0.204.

Table 3. The first step processing is done for converting the raw data into understandable data. Since the original dataset contains null values and errors, it has been preprocessed. The original data has been resized and made ready for the next stage, as the network allows. the result for prediction of stroke prediction for K-Nearest Neighbor with an accuracy of approximately 95.70% and Random forest Algorithm 94.80%. There exists a statistically insignificant difference between the two groups ($p=0.204$; $p>0.05$). By comparing the accuracies conducted, that K-Nearest Neighbor algorithm achieved a

better performance than random forest shown in figure1

DISCUSSION

The predicted accuracy by the K-nearest neighbor has found 95.70% against the Random Forest algorithm has found 94.80%. Table 3 has calculated two tailed significant values and there exists a statistically insignificant difference between the two groups ($p=0.204$; $p>0.05$) and the result corresponds to equal variances assumed for analysis. Hence this research study found that the Novel K-nearest neighbor seems to be significantly better than the Random Forest.

There are similar papers on stroke prediction using machine learning (Boukhennoufa et al. 2022) for concepts that are used in reserach . This used to the wearable sensor is used and its location ,exercises tracked ,participants ,selected features of machine learning used for performance classifiction divided three categories based on the assessment type, namely activity recognition ,motor classification and clinical classificationIn Support of this research, for analysis of the data different types of machine learning algorithms are such as K-nearest neighbor , Random forest, Decision Tree as10 samples (Lin et al. 2020). The Random forest obtained accuracy of 91% (Sirsat, Fermé, and Câmara 2020) the K-nearest neighbor algorithm 90% basic literature survey survey, it is evident that the K-nearest neighbor algorithm performs better than Random forest, machine learning such as Decision Tree. GBDT is trained using 100 boost and also utilizes other ML-based classification, such as GBDT, LR and SVM (all implemented in python 2.7 using the scikit-learn version 0.18.0 packages).

GBDT is trained using 100 boosted trees and binomial loss function. L 2 - regularization is used with the strength set at 1.0 for the LR method. For SVM, we use the linear kernel suitable for high dimensionality of our feature space. The tuning dataset (around 5% of total patients) is used to adjust all the hyper-parameters in these algorithms. Although the results of the study are better in both experimental and statistical analysis, there are certain limitations in the work. The evaluation of accuracy cannot provdied a better outcome on a larger data set. Moreover in K-nearest neighbor the mean error appears to be higher than support vector machines. It would be better if the mean error can be reduced to a considerable extent. However, the work can be enhanced by applying optimization algorithm techniques,to achieve better accuracy and less mean error. A feature selection algorithm can be used before classification to improve the accuracy of the classifier.

CONCLUSION

The results obtained show that the Novel K-Nearest Neighbor has found 95.70% of accuracy on the provided dataset Random Forest 94.80%. As a result ,the accuracy of the Novel K-Nearest Neighbor appears to be much better than the Random forest algorithm in this research study.

DECLARATIONS

Conflict of interests

No conflict of interest in this manuscript.

Author Contribution

Author PK was involved in data collection, data analysis, and manuscript writing. Author S.Saraswathi was involved in conceptualization, data validation, and critical review of the manuscript

Acknowledgment

The authors would like to express their gratitude towards the Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding

We thank the following organizations for providing financial support that enabled us to complete this study:

1. Sri Cube Innovations Pvt Ltd,Vijayawada.
2. Saveetha School of Engineering
3. Saveetha University
4. Saveetha Institute of Medical and Technical Science

REFERENCES

1. Bandi, Vamsi, Debnath Bhattacharyya, and Divya Midhunchakkravarthy. 2020. "Prediction of Brain Stroke Severity Using Machine Learning." *Revue d'Intelligence Artificielle*. https://doi.org/10.18280/ria.340609.
2. Benin, S. R., S. Kannan, Renjin J. Bright, and A. Jacob Moses. 2020. "A Review on Mechanical Characterization of Polymer Matrix Composites & Its Effects Reinforced with Various Natural Fibres." *Materials Today: Proceedings* 33 (January): 798–805.
3. Boukhennoufa, Issam, Xiaojun Zhai, Victor Utti, Jo Jackson, and Klaus D. McDonald-Maier. 2022. "Wearable Sensors and Machine Learning in Post-Stroke Rehabilitation Assessment: A Systematic Review." *Biomedical Signal Processing and Control*. https://doi.org/10.1016/j.bspc.2021.103197.
4. Gudipaneni, Ravi Kumar, Mohammad Khursheed Alam, Santosh R. Patil, and Mohmed Isaqali Karobari. 2020. "Measurement of the Maximum Occlusal Bite Force and Its Relation to the Caries Spectrum of First Permanent Molars in Early Permanent Dentition." *The Journal of Clinical Pediatric Dentistry* 44 (6): 423–28.
5. Kelleher, John D., Brian Mac Namee, and Aoife D'Arcy. 2020. *Fundamentals of Machine Learning for Predictive Data Analytics, Second Edition: Algorithms, Worked Examples, and Case Studies*. MIT Press.
6. Lin, Ching-Heng, Kai-Cheng Hsu, Kory R. Johnson, Yang C. Fann, Chon-Haw Tsai, Yu Sun, Li-Ming Lien, et al. 2020. "Evaluation of Machine Learning Methods to Stroke Outcome Prediction Using a Nationwide Disease Registry." *Computer Methods and Programs in Biomedicine* 190 (July): 105381.
7. Liu, Tianyu, Wenhui Fan, and Cheng Wu. 2019. "A Hybrid Machine Learning Approach to Cerebral Stroke Prediction Based on Imbalanced Medical Dataset." *Artificial Intelligence in Medicine* 101 (November): 101723.
8. Monteiro, Miguel, Ana Catarina

Fonseca, Ana Teresa Freitas, Teresa Pinho E Melo, Alexandre P. Francisco, Jose M. Ferro, and Arlindo L. Oliveira. 2018. "Using Machine Learning to Improve the Prediction of Functional Outcome in Ischemic Stroke Patients." *IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM* 15 (6): 1953–59.

9. Nalini, Devarajan, Jayaraman Selvaraj, and Ganesan Senthil Kumar. 2020. "Herbal Nutraceuticals: Safe and Potent Therapeutics to Battle Tumor Hypoxia." *Journal of Cancer Research and Clinical Oncology* 146 (1): 1–18.
10. Reddy, Poornima, Jogikalmat Krithikadatta, Valarmathi Srinivasan, Sandhya Raghu, and Natanasabapathy Velumurugan. 2020. "Dental Caries Profile and Associated Risk Factors Among Adolescent School Children in an Urban South-Indian City." *Oral Health & Preventive Dentistry* 18 (1): 379–86.
11. Sathish, T., and S. Karthick. 2020. "Gravity Die Casting Based Analysis of Aluminum Alloy with AC4B Nano-Composite." *Materials Today: Proceedings* 33 (January): 2555–58.
12. Sathish, T., D. Bala Subramanian, R. Saravanan, and V. Dhinakaran. 2020. "Experimental Investigation of Temperature Variation on Flat Plate Collector by Using Silicon Carbide as a Nanofluid." In *PROCEEDINGS OF INTERNATIONAL CONFERENCE ON RECENT TRENDS IN MECHANICAL AND MATERIALS ENGINEERING: ICRTMME 2019*. AIP Publishing. https://doi.org/10.1063/5.0024965.
13. Sirsat, Manisha Sanjay, Eduardo Fermé, and Joana Câmara. 2020. "Machine Learning for Brain Stroke: A Review." *Journal of Stroke and Cerebrovascular Diseases: The Official Journal of National Stroke Association* 29 (10): 105162.
14. Sivasamy, Ramesh, Potu Venugopal, and Rodrigo Espinoza-González. 2020. "Structure, Electronic Structure, Optical and Magnetic Studies of Double Perovskite Gd2MnFeO6 Nanoparticles: First Principle and Experimental Studies." *Materials Today Communications* 25 (December): 101603.
15. Venu, Harish, and Prabhu Appavu. 2021. "Experimental Studies on the Influence of Zirconium Nanoparticle on Biodiesel–diesel Fuel Blend in CI Engine." *International Journal of Ambient Energy* 42 (14): 1588–94.
16. Yilmaz, Yusuf, Robert Carey, Teresa M. Chan, Venkat Bandi, Shisong Wang, Robert A. Woods, Debajyoti Mondal, and Brent Thoma. 2021. "Developing a Dashboard for Faculty Development in Competency-Based Training Programs: A Design-Based Research Project." *Canadian Medical Education Journal* 12 (4): 48–64.

TABLES AND FIGURES

Table 1. Accuracy and precision of the stroke prediction using test data with Novel K-NearestNeighborAlgorithm with Support Vector Machine algorithm for the different iterations. The result for prediction of stroke prediction for Novel K-NearestNeighbor algorithm with an accuracy of approximately 95.70 % and Support Vector Machine Algorithm 94.8 %

Sample(N)	Dataset size/Rows in %	KnnAccuracy	Random forestAlgorithm Accuracy in %
1	5110	95.70	94.8
2	5100	95.63	94.6
3	5000	95.60	93.7
4	4950	95.20	93.2
5	4890	95.10	93.1
6	4700	94.50	90.00
7	4600	93.10	89.00
8	4500	92.60	88.90
9	4600	91.00	87.00
10	4300	90.10	86.00

Table 2. Statistical results of K-Nearest Neighbor (KNN) and Random Forest (RN). Mean Accuracy Value, Standard deviation and Standard Error Mean for 10 iterations. It is observed the mean accuracy and standard deviation for K-nearest neighbor Algorithm is 93.85 % and 0.380 and for Random forest Algorithm is 91.03 and 0.367.

	Algorithm	N	Mean	Standard Deviation	Standard Mean Error
Accuracy	K-NN	10	93.85	.380	0.120
Accuracy	RN	10	91.03	.367	0.84

Table 3. The independent sample t-test of the significance level KNN and Random Forest algorithms results with two tailed significant values(p=0.204). There exists a statistically insignificant difference between the two groups (p=.204; p>0.05).

		Levene's Test for Equality of Variances (1)	**Levene's Test for Equality of Variances (2)**	**T-test for Equality of Means (3)**	**T-test for Equality of Means (4)**	**T-test for Equality of Means (5)**
		F	Sig.	Std.Error Difference	95% Confidence lower	95% Confidence upper
Accuracy	Equal variances assumed	1.733	.204	.136	.207	.940
	Equal variances not assumed			.136	.19	.942

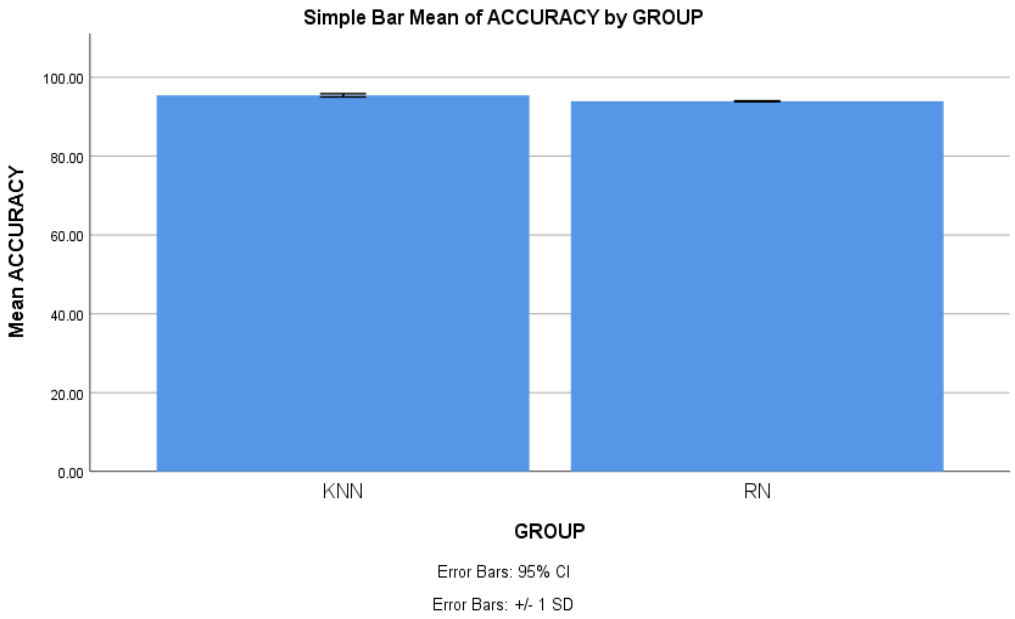

Fig 1. Prediction accuracy for the two algorithms.The accuracy of the K-Nearest Neighboris better than the accuracy of Random forest X Axis: K-Nearest Neighbor vs Random forest algorithm Y Axis: Mean accuracy of detection +/-1SD. The accuracy of the K-Nearest Neighbor Random Forest algorithm is 93.85 % and 91.03 %.