https://doi.org/10.33472/AFJBS.6.1.2024.429-435

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

K-Nn Version Metrics for Predicting Breast Cancer Survival

Dr. S. Bharathi[1], Poongodi.D[2*], Krithika.L[3]

[1]Assistant Professor, Department of Mathematics, Bharathiar University PG Extension and Research Center, Perundurai, Erode, Tamilnadu, India.

[2*]Research Scholar, Department of Mathematics, Bharathiar University PG Extension and Research Center, Perundurai, Erode, Tamilnadu, India.

[3]Research Scholar, Department of Mathematics, Bharathiar University PG Extension and Research Center, Perundurai, Erode, Tamilnadu, India.

bharathikamesh6@gmail.com[1], keerthilakshminarayanan@gmail.com[3]

Corresponding Email: dpoongodi1997@gmail.com*

Article Info

Volume 6, Issue 1, January 2024

Received: 09 December 2023

Accepted: 01 January 2024

Published: 28 January 2024

doi: 10.33472/AFJBS.6.1.2024.429-435

ABSTRACT:

Breast cancer is one of the most frequent and having a high mortality rate among women. Early detection of this enhances survival rates from 56% to more than 86%. As a result, an accurate and dependable approach is required. Predicting cancer survival is becoming an increasingly difficult task in medicine. Researchers have widely employed machine learning methods to address this difficulty. The k-nearest neighbour (KNN) technique is the most widely utilized among the various machine learning algorithms. This article examines the performance of various KNN versions (Classic one, adaptive, locally adaptive, k-means clustering, fuzzy, mutual, ensemble, Hassanat, and generalised mean distance) in predicting breast cancer survival. This study carried out massive implementations and experiments using Haberman's Survival Data Set, which was collected from Kaggle, to analyze these variants. For comparison analysis, we took into account the accuracy, precision, and recall performance metrics. Based on performance metrics, this study determines that the Hassanat KNN version performed the best, followed by the ensemble approach KNN. Based on four performance criteria (Accuracy, F1 score, Precision, and Recall) for survival prediction, the presented research summarizes which KNN variation is the most promising candidate to pursue. The results of this study could be utilized by beneficiaries and healthcare researchers to choose the best KNN variation, for predicting breast cancer survival.

Keywords: Breast Cancer- Survival- k nearest neighbourhood - KNN versions-Performance Measures

1. Introduction

Cancer is a formidable disease, and breast cancer stands out as one of the most frequently encountered cancers impacting women's health globally[17]. It is associated with a high mortality rate across different populations. Despite the progress made through recent technological advancements that have somewhat lowered mortality rates, there is an immediate requirement for the development of predictive models that can detect both the onset and recurrence of breast cancer in its early phases[2]. Conventional statistical techniques and software have long been employed to identify the determinants influencing breast cancer survival rates. While these traditional approaches fulfil their intended functions to a degree, they lack the flexibility and robustness necessary to accommodate new variables[20]. Consequently, the development of rapid and resilient computational models has become a primary focus for researchers and scientist. The survival rate for breast cancer is notably higher than that of other tumours, as individuals can survive without breast tissue[6]. Conventional methods have relied on different pathological traits, including hormone receptor expression, tumour size, and nuclear grade[7]. However, the introduction of powerful and efficient computational algorithms has facilitated the examination of high-dimensional parameters and the selection of key features by researchers[8,13]. This study explored the potential applications of the current approach in clinical research. Also, this paper discusses how breast cancer survival rates can be detected using different versions of KNN algorithm.

K-Nearest Neighbour

A supervised machine learning approach that is mostly used for classification is the k-nearest-neighbour (KNN) algorithm. It is extensively utilized in the prognosis of diseases[19]. By considering the features and labels of the training data, the supervised KNN algorithm predicts the classification of unlabelled data[3].By considering the k nearest training data points (neighbors), which are the ones closest to the query it is testing, the KNN technique can typically categorize datasets using a training model that is comparable to the testing question. Ultimately, the algorithm employs a majority voting mechanism to determine which classification should be finalized. Because of its highly flexible and simple-to-understand design, the KNN algorithm is one of the most popular and basic types of machine learning algorithms used in classification tasks[15]. The approach is well known for helping with regression and classification problems for a variety of data types, including ranges, sizes, label counts, noise levels, and contexts[22].

K-NN and its versions

Classic KNN Algorithm

A supervised machine learning technique that is primarily utilized for classification is the traditional KNN algorithm[14]. The technique uses a variable parameter called k, which stands for "nearest neighbour" in English. Finding the closest data point or neighbors for a query from a training dataset is how the KNN algorithm operates. Based on the closest distances from the query point, the closest data points are located. It locates the k closest data points and then uses a majority voting mechanism to determine which class showed up most frequently. The final categorization for the query is determined by looking at the class that showed up the most.

Adaptive KNN (A-KNN)

A variation of the method called Adaptive[18,21] KNN concentrates on choosing the best k value for a testing data point. To find the ideal k value for every data point in the training dataset, it implements a different approach. Next, for each testing data point, the primary

algorithm locates its closest neighbour in the training dataset and takes on its k value. Using this inherited k value, the KNN variation continues to operate like the traditional KNN method to predict the result.

Locally adaptive KNN with Discrimination class (LA-KNN)

This variation[16] calculates the ideal k value by taking into account data from discrimination classes. The discrimination class idea takes into account the number and distribution of neighbors in the k-neighbourhood around a specific testing data point who belong to the majority class and those who belong to the second majority class. To define discrimination classes, the algorithm goes through a number of processes. Following the selection of one of those classes, a ranking table with various k values, Centro centric distances, and ratios is formed.

Fuzzy KNN (F-KNN)

The membership assignment principle serves as the foundation for the fuzzy[12] KNN Algorithm. The modification considers the k nearest neighbors of a testing dataset from the training dataset, much like the original KNN algorithm. Subsequently, it allocates "membership" values to every class included in the list of k's nearest neighbours. A fuzzy math approach based on the weight of each class is used to compute the membership values. The categorization result is then chosen for the class with the highest membership.

K-means clustering-based KNN (KM-KNN)

The KNN version that is based on clustering combines the widely used methods of k-means and 1NN. This deviation organizes the training dataset based on a predetermined variable (the number of clusters) using the k-means method[4]. After that, it determines the centroids of every cluster, creating a fresh training dataset with the centroids of every cluster. Using this new training dataset, the 1NN algorithm is run, and the single nearest neighbour is selected for classification.

Weight adjusted KNN (W-KNN)

The implementation of attribute weighting is the main objective of this KNN algorithm[10] version. Initially, the kernel function is used by this approach to apply a weight to each training data point. The purpose of this weight assignment is to allocate greater weight to points that are closer together and less weight to ones that are farther away. Any function that lowers the value as the distance rises can be applied as a kernel function. Next, the output class of a particular testing data point is predicted using the frequency of all nearest neighbors. This KNN variation takes into account the significance of various features for classification while generating the kernel function for a dataset that has numerous attributes.

Hassanat distance KNN (H-KNN)

An algorithm based on the distance measurement formula is the Hassanat[1] KNN algorithm. This modification suggests a more sophisticated method for calculating the separation between two data points while adhering to the KNN algorithm's basic structure. The utilization of maximum and minimum vector points, as in weight attributions in other versions, is the basis of the new distance calculation known as the Hassanat distance. Like in the original KNN method, this variant's Hassanat distance metric determines a testing query's closest neighbours and applies the majority voting procedure.

Generalised mean distance KNN (GMD-KNN)

The primary applications of local vector constructions and repeated generalized[9] mean distance calculations are the focus of the generalized mean distance KNN variant. Sorted lists of each class's k nearest neighbors are first stored to facilitate the algorithm's operation. Following the conversion of each list to a local mean vector, the mean distance is calculated iteratively until a final value for each class's distance from the testing query is obtained. For the testing query, the class with the least distance is therefore considered to be the right prediction.

Mutual KNN (M-KNN)

The mutual neighbours principle is the primary objective of the mutual[5] KNN algorithm. In order to change the training dataset, the method first eliminates sets from it that do not share any k nearest neighbors with the other sets. This results in a training dataset that has been trimmed and has fewer anomalies and noise. Then, the algorithm discovers the k nearest neighbors of the training dataset using the testing dataset, and it finds the k nearest neighbors of the testing dataset's nearest neighbors. This makes it possible for the algorithm to identify their shared nearest neighbors, who can then be evaluated as potential candidates for classification. The majority voting mechanism is used to classify the testing datasets.

Ensemble approach KNN (EA-KNN)

The KNN variation, to solve the issue of having a fixed "k" value for classification, EA-KNN[11] is based on an ensemble technique. The k-nearest neighbors of a testing query are found by this approach using a K max value of n, where n is the size of the training dataset. Following that, weight summation operations are applied to the list of closest neighbors, which is sorted based on distance. In order to complete the weight summation procedures, an inverse logarithm function for "k" values is added repeatedly for values ranging from 1 to K max in increments of 2. Subsequently, the class with the highest weight summation is considered the anticipated classification for the testing inquiry.

Table1. Comparison of Performance Measures among different KNN variants.

KNN and its version	Accuracy	Precision	Recall	F1 Score
Classic KNN	0.70	0.65	0.70	0.67
Adaptive KNN	0.70	0.65	0.70	0.67
Locally Adaptive KNN	0.72	0.66	0.72	0.68
Fuzzy KNN	0.74	0.72	0.74	0.73
Weighted KNN	0.70	0.65	0.70	0.67
Hassanat KNN	0.75	0.71	0.75	0.72
GeneralisedMean Distance KNN	0.69	0.73	0.69	0.70
Mutual KNN	0.26	0.82	0.26	0.15
Ensemble Approach KNN	0.77	0.59	0.77	0.67
K means clustering KNN	0.77	0.59	0.77	0.67

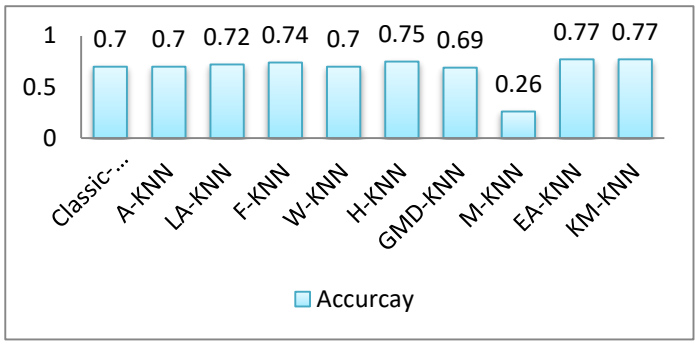

Figure 1. Accuracy comparison among KNN versions

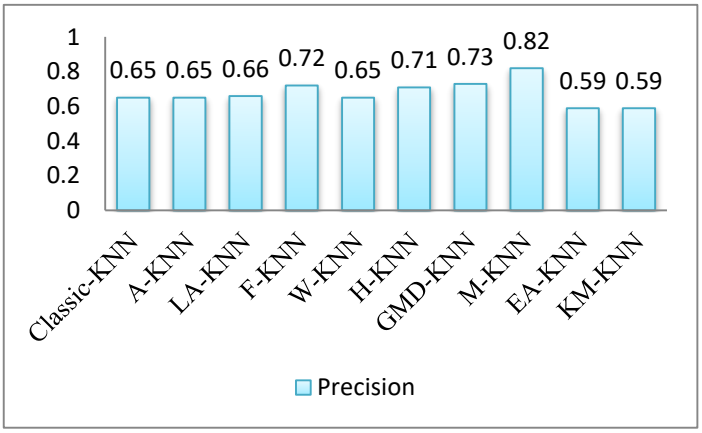

Figure 2. Precision comparison among KNN versions

Figure 3. Recall comparison among KNN versions

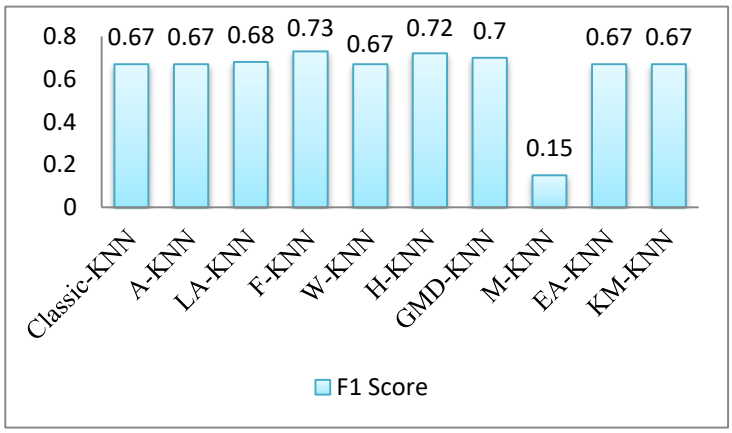

Figure 4. F1 Scores comparison among KNN versions

Software Used

For scripts that are implemented under Anaconda3, the Python programming language (version 3.7.1) was applied. The k-NN algorithm and its variants were applied using libraries from the scikit-learn package (version 0.20.1).

Availability of Dataset

The Haberman's Survival Data Set which has 306 instances with 4 attributes was taken which was available in Kaggle. It is open source and freely accessible to all users. This data encouraged improved prediction of outcomes.

2. Conclusion

Intelligent use of ML algorithms can optimize the impact in a highly supervised manner. Future difficulties will demand several approaches for training, testing, and validating the models rather than relying on a single solution. It will give the expert and researcher more flexibility when integrating these methods into the prediction of breast cancer. In this study, we employed various KNN algorithms to predict breast cancer survival. The accuracy, precision, recall, and F1-score of ML classifiers were also evaluated. Researchers are better able to make judgments and raise public awareness of breast cancer at an early stage thanks to the technology advancements and advances in fundamental principles in these machine learning methodologies. The development of this technology will help in determining cancer therapy more precise and targeted at a reasonable cost.

3. References

1. Alkasassbeh, M., Altarawneh, G. & Hassanat, A. On enhancing the performance of nearest neighbour classifiers using hassanat distance metric. Can. J. Pure Appl. Sci. 9, 1–6 (2015).
2. Boyle, P. (2012, August 1). Triple-negative breast cancer: Epidemiological considerations and recommendations. Annals of Oncology [Internet] [cited 2021 Aug 31], 23(SUPPL. 6), vi7–12. Available from: http://www.annalsofoncology.org/article/S0923753419376355/fulltext
3. Bzdok, D., Krzywinski, M. & Altman, N. Machine learning: supervised methods. Nat. Methods 15, 5–6 (2018).
4. Cherif, W. Optimization of K-NN algorithm by clustering and reliability coefficients: Application to breast-cancer diagnosis. Procedia Computer. Sci. 127, 293–299 (2018).
5. Dhar, J., Shukla, A., Kumar, M. & Gupta, P. J. A. P. A. A weighted mutual k-nearest neighbour for classification mining. (2020)
6. Feng, Y., Spezia, M., Huang, S., Yuan, C., Zeng, Z., Zhang, L., et al. (2018). Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis [Internet]. In Genes and diseases. Chongqing yi ke da xue, di 2 lin chuang xue yuan Bing du xing gan yan yan jiu suo [cited 2021 Mar 7] (Vol. 5, pp. 77–106). Available from: /pmc/articles/PMC6147049/
7. Ganggayah, M. D., Taib, N. A., Har, Y. C., Lio, P., & Dhillon, S. K. (2019, March 22). Predicting factors for survival of breast cancer patients using machine learning techniques.BMC Medical Informatics and Decision Making [Internet]. [cited 2021 Mar 7], 19(1), 48. Available from: https://bmcmedinformdecismak.biomedcentral.com/articles/https://doi.org/10.1186/s12 911-019-0801-4

8. Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. VSURF: An R package for variable selection using random forests [cited 2021 Aug 31]. Available from: http://CRAN.R-project.org/pac kage=VSURF
9. Gou, J. et al. A generalised mean distance-based k-nearest neighbour classifier. Expert Syst. Appl. 115, 356–372 (2019).
10. Han, E.-H. S., Karypis, G. & Kumar, V. In Pacific-Asia Conference on Knowledge Discovery and Data Mining. 53–65 (Springer).
11. Hassanat, A. B., Abbadi, M. A., Altarawneh, G. A. & Alhasanat, A. A. J. A. P. A. Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach. (2014).
12. Keller, J. M., Gray, M. R. & Givens, J. A. A fuzzy k-nearest neighbour algorithm. IEEE Trans. Syst. Man Cybern. 15, 580–585 (1985).
13. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015, January 1). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8–17.
14. Lopez-Bernal, D., Balderas, D., Ponce, P. & Molina, A. Education 4.0: Teaching the basics of KNN, LDA and simple perceptron algorithms for binary classification problems. Future Internet 13, 193–206 (2021)
15. Mahesh, B. Machine learning algorithms—a review. Int. J. Sci. Res. 9, 381–386 (2020).
16. Pan, Z., Wang, Y. & Pan, Y. A new locally adaptive k-nearest neighbour algorithm based on discrimination class. Knowl. Based Syst. 204, 106185 (2020).
17. Sharma, G. N., Dave, R., Sanadya, J., Sharma, P., & Sharma, K. K. (2010, April). Various types and management of breast cancer: An overview. Journal of Advanced Pharmaceutical Technology & Research [Internet] [cited 2021 Aug 31] 1(2), 109. Available from: /pmc/articles/PMC3255438/
18. Sun, S. & Huang, R. In 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery. 91–94 (IEEE).
19. Uddin, S., Khan, A., Hossain, M. E. & Moni, M. A. Comparing different supervised machine learning algorithms for disease prediction. BMC Med. Inform. Decis. Mak. 19, 1–16 (2019)
20. Vickers, A. J., & Cronin, A. M. (2010). Traditional statistical methods for evaluating prediction models are uninformative as to clinical value: Towards a decision analytic framework. Seminars in Oncology., 37(1), 31–38. 4.
21. Wettschereck, D. & Dietterich, T. G. In Advances in Neural Information Processing Systems, Vol. 6 184–184 (Morgan Kaufmann Publishers, 1994).
22. Zhang, S., Li, X., Zong, M., Zhu, X. & Cheng, D. Learning k for kNN classification. ACM Trans. Intell. Syst. Technol. 8, 1–19 (2017)