Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/40/20230666

The application and analysis of the KNN algorithm in machine learning for breast cancer prediction

Hangyu Li

School of Electrical and Electronic Engineering, WenZhou University, Wenzhou, 325035, China

mmedinak84848@student.napavalley.edu

Abstract. When female patients are tested for cancer, breast cancer tends to be detected more often than other cancers. Breast cancer also accounts for a large percentage of deaths of female patients due to cancer. Timely prediction of whether one has breast cancer or not is of great significance in the treatment of breast cancer. In this study, K Nearest Neighbors (KNN) algorithm is used for data analysis as well as probability prediction to determine the probability of having breast cancer and compare it with traditional machine learning algorithms. The variation of accuracy versus K-value for a given dataset was experimentally represented by a graph. Also, the experiment gives and compares the accuracy of different learning algorithms in the same situation. According to the experimental results, the KNN algorithm tends to be more accurate than the other algorithms, with an accuracy of 98 percent. This turned out to be significantly higher than the remaining two algorithms. This experiment has significant implications for the healthcare industry. Because the use of this method can greatly improve the efficiency of doctors and the diagnosis of breast cancer, so that more people with the risk of breast cancer can be diagnosed and treated in a timely manner. In addition, hospitals can use the study to predict the incidence of breast cancer in a given area of the population.

Keywords: Breast Cancer, Prediction, KNN Algorithm, Healthcare Industry.

1. Introduction

Breast cancer is a malignant tumor that not only erupts in the breast tissue, but also spreads to surrounding areas, with areas such as the ductal lining of the breast being the hardest hit by its appearance. Breast cancer is much more prevalent in the female population than in the male population, and it is also the most commonly detected cancer that causes death in female patients. Among non-skin cancers, breast cancer is second only to lung cancer. It is the fifth most common cause of death due to cancer [1].

Manual prediction of breast cancer in women relies on the following factors: increasing age from year to year; the presence of a discontinuous mass; breast thickening; lymphadenopathy; and a mass $\geqslant$ 2 cm [2]. Also, breast cancer is influenced by genetic factors [3]. In clinical testing, the General Practitioner (GP) will be responsible for the assessment in the first instance, and only if there is a breast problem will that patient be offered a specialized mammogram. Other researchers have estimated that GPs will only see one case of breast cancer in a year out of 6 to 34 patients with breast problems [4-8]. Improving the success of breast cancer treatment and being able to detect that a patient has breast cancer

DOI: 10.54254/2755-2721/40/20230666

earlier is critical [2]. Most of the researchers have used a combination of linear and nonlinear or nonlinear and integrated algorithms for breast cancer prediction [9]. Artificial Neural Network (ANN) has the highest accuracy of 98.57% using the same dataset.[10]. Combining machine learning techniques with manual diagnosis can be of great help to doctors compared to manual-only medical diagnosis.

Prediction of breast cancer using K Nearest Neighbors (KNN) algorithm in machine learning is the main objective of this study. The first step is to process the imported data into a form that is convenient for subsequent algorithms, such as converting benign and malignant in the original data into corresponding individual numbers. Then random sampling is used to get an overall picture of the data. Next the relationship between features is visualized using heat maps and scatter plots so that the relationship between features can be visualized when outputting the results. After that divide the target vector and feature matrix and divide the training set and data set. The number of elements of the target variable in the test set is then obtained, again for subsequent use. Finally, the model is used to perform calculations and output the results. This study facilitates a more accurate prediction of the chances of getting breast cancer.

2. Methodology

2.1. Dataset description

Breast Cancer dataset from Kaggle [11]. It contains 31 variables, excluding the IDs in the first column, and is divided into three parts. The first part is the actual values, including: tumour nature (benign, malignant), breast lobe radius, surface texture means, lobe periphery, lobe mean area, smoothness means, compactness means, concave means, concave point means, symmetry means, and fractal dimension means. The second part consists of the standard errors of the above data. The third part consists of the worst values.

2.2. Proposed approach

The purpose of this experiment is to use the KNN model to create a reliable, intuitive predictive model to assist physicians in breast cancer prediction. Inputs were entered and pre-processed by removing the ID column from the data and replacing the values in the diagnosis column by replacing B with 1.0 and M with -1.0. Ten samples were then randomly selected to facilitate an understanding of the characteristics, distribution, and variation of the data. A heat map is then plotted to make it easier to see the correlation between the various features in the data set in a straightforward manner. The next step to visualize the relationship between the features is to draw a scatterplot, which will also be of great help in dealing with anomalous conditions at a later stage. This is followed by removing the diagnosis columns, creating the feature matrix and the target vectors, and in general pre-processing the data, followed by randomly disrupting the data and dividing it between the training and test sets. The purpose of disrupting the data is to prevent the introduction of undesirable influences and errors, and to make the model more accurate. The process is shown in the figure 1.

Finally, the KNN model is used for training and evaluation and the accuracy of the model is printed.

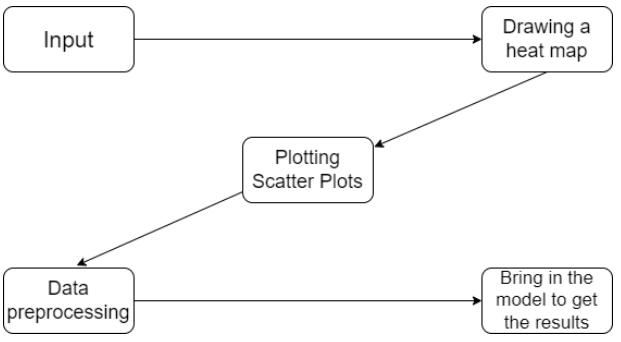

Figure 1. Illustration of the whole process.

DOI: 10.54254/2755-2721/40/20230666

2.2.1. Data pre-processing. Before processing, the file data is traversed once to ensure that it is loaded correctly. When processing, the irrelevant ID columns are deleted first, and in order to make the dataset more suitable for machine learning and reduce the computational burden, the values in the "diagnosis" column are replaced, replacing B (benign) with 1 and M (malignant) with -1. The next step is to divide the training and test sets. However, the data are randomly disrupted before division to ensure that the samples in both sets are randomly distributed and to avoid possible bias or overfitting problems. In order to keep the experimental results reproducible, so a random number seed was specified. Next divide the training and testing sets. Specify the feature matrix and feature vectors, using the "diagnosis" column as the target variable and the remaining columns as the feature matrix, and convert them to NumPy arrays. After completing the previous step the preprocessing ends here.

2.2.2. KNN. The principle of KNN is that when a value X needs to be predicted, the category to which X belongs is determined according to what the K nearest points to it are. KNN is a nonparametric, inert algorithmic model. Non-parametric means that the model does not make specific assumptions about the data or the problem, but relies on the data itself to perform the classification or regression task. Inert means that KNN, unlike logistic regression which requires extensive training on the data first, does not have an explicit process of skillful data, or the process is fast. The KNN framework is shown in Figure 2 and the main steps are K value selection and point distance calculation.

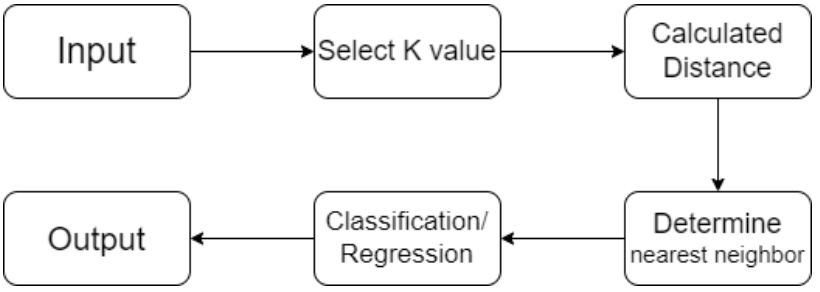

Figure 2. The pipeline of KNN.

Distance calculations are commonly Manhattan distance calculations and Euclidean distance calculations, and Euclidean distance calculations are usually used. The Euclidean formula for finding the distance between two points in a two-dimensional plane is as follows (x denotes coordinates in the horizontal direction and y denotes coordinates in the vertical direction):

$$\rho = \sqrt{(x_2 - x_1)^2 - (y_2 - y_1)^2} \tag{1}$$

Expanding to multi-dimensional space, the formula becomes:

$$d(x, y) := \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2}$$

The K-value was selected using cross-validation with values ranging from small to large, and then the validation set variance was calculated to finalize the K-value. After calculating the variance, it is obtained that as the K value increases, the error first decreases and then rises. The reason for the decrease is that there are more samples to draw from, and the reason for the increase can be compared to the analogy that there are 10 samples, and when the value of K increases to 9, there is no point in continuing to use the KNN. Therefore, when selecting the K value is usually selected a point K, the value of this point, whether increasing or decreasing, the error rate will increase. Because KNN is different from traditional model training and parameter tuning, it is a real-time judgment with the training tools that are already in place at the time of prediction, so KNN is not like traditional machine learning algorithms that have a displayed model.

DOI: 10.54254/2755-2721/40/20230666

2.3. Evaluation index

Accuracy denotes the prediction success rate of this model. The study uses a Boolean array to calculate the accuracy. The study first creates a Boolean array, and determine whether the prediction result is equal to the test label, will be equal to 1 is not equal to 0, and finally the Boolean array can be obtained by summing the number of correctly predicted samples. The accuracy can be obtained with the following formula:

$$Accuracy = \frac{Number\ of\ correctly\ predicted\ samples}{Total\ test\ sample\ size} \tag{3}$$

2.4. Evaluation index

This experiment is implemented using Visual Studio Code 1.81.1 and the python version provided therein as 3.11.4 64-bit. KNN models were referenced using the Scikit-learn library, Numpy and Pandas libraries were used for data processing, and Seaborn and Matplotlib libraries were used for visualization. The experiment is performed on a Windows system device equipped with Intel(R) Core (TM) i5-7300HQ CPU. The value of K in the KNN model is set to 12.

3. Result and discussion

The accuracy rate obtained from the final calculation of this model is 0.9824561403508771, which is approximated as 98%, as shown in the Table 1. In the case of unified dataset and preprocessing, the accuracy rate obtained by decision tree is 0.9385964912280702, which is approximated as 94%; the accuracy rate obtained by random forest in the same case is 0.9649122807017544, which is approximated as 96%. It can be seen that KNN performs better in these three models.

Table 1. The Comparison of the Different Model.

model	accuracy
decision tree	0.9385964912280702
random forest	0.9649122807017544
KNN	0.9824561403508771

Figure 3 represents the effect of K value selection on accuracy. The range of values is from 1 to 49, excluding 50. The vertical direction is the accuracy and the horizontal direction is the range of values of K. As shown in the Figure 3, the accuracy reaches the highest in the range when K is taken as 11 and 12, fluctuates when K is taken as 6 and 24, and changes more gently when K is taken as 1 to 3, 7 to 9, 15 to 22, and 25 to 49. Figure 3 indicates that choosing too small a value of K leads to overfitting, where the model is susceptible to noise in the data. Selecting too large a value for K can lead to underfitting, making the model too smooth to capture the complex relationships in the data. This can all lead to a lack of accuracy in the model. The model demonstrated greater predictive power when K values of 11 and 12 were selected, when a balance between overfitting and underfitting was achieved.

DOI: 10.54254/2755-2721/40/20230666

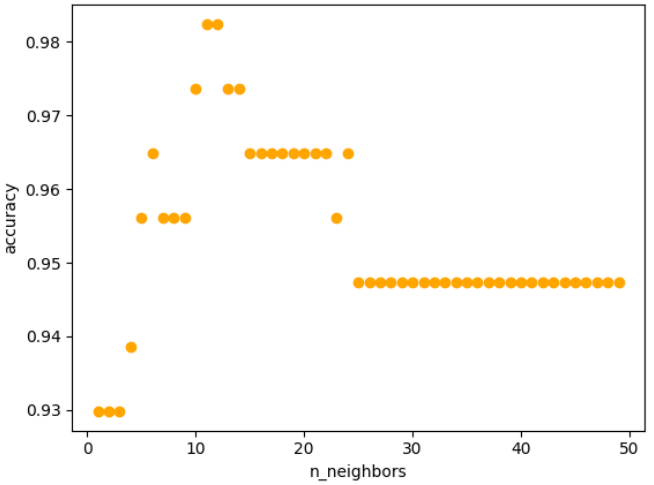

Figure 3. The result of the model.

These findings have significant implications for clinical prediction of breast cancer. First, the KNN model shows better prediction potential than the other two models, which helps doctors to make more accurate predictions. Second, it is easier and faster to use this prediction method than the traditional manual prediction method. Finally, if this technique is introduced and used in hospitals, it will greatly accelerate the efficiency of doctors, who will have more time to treat more patients. In compliance with the law, hospitals can keep statistics on all breast cancer patients to predict the number of people who are likely to suffer from breast cancer, so that medical preparations can be made in advance to prevent shortages of manpower, therapeutic equipment and medication.

4. Conclusion

The main objective of this study is to use a certain type of KNN in machine learning for breast cancer prediction in order to help doctors to make predictions more accurately and efficiently and to reduce the impact of breast cancer on patients' lives. The experiment uses heat map and scatter plot to visualize the relationship between the features and finally KNN model, decision tree model, random forest model is used for prediction and the accuracy of the models are compared. According to the experimental results the accuracy is obtained: KNN, 98%; Random Forest, 96%; Decision Tree, 94%, KNN model is significantly better than the remaining two models. The prediction speed of the machine model is significantly higher than the traditional manual prediction. This model helps to predict the probability of developing breast cancer more accurately and efficiently. For future research, the main directions for the future lie in the optimization of data processing and finding more appropriate machine models. Optimization of data processing can help to improve the prediction of abnormal data, which can be easily adapted to different countries, regions, and populations with different body types to ensure high accuracy of the model. Optimization of machine models provides great possibilities for further improvement of prediction accuracy.

References

[1] Sharma G N Dave R Sanadya J et al 2010 Various types and management of breast cancer: an overview Journal of advanced pharmaceutical technology & research 1(2): p 109

[2] McCowan C, Donnan P T Dewar J et al 2011 Identifying suspected breast cancer: development and validation of a clinical prediction rule British Journal of General Practice 61(586): pp e205-e214

[3] Kim G Bahl M 2021 Assessing risk of breast cancer: a review of risk prediction models Journal of breast imaging 3(2): pp 144-155

DOI: 10.54254/2755-2721/40/20230666

[4] Nichols S Waters W E Wheeler M J 1980 Management of female breast disease by Southampton general practitioners Br Med J 281(6253): pp 1450-1453
[5] Newton P Hannay D R Laver R 1999 The presentation and management of female breast symptoms in general practice in Sheffield Family Practice 16(4): pp 360-365
[6] Roberts M M Elton R A Robinson S E et al 1987 Consultations for breast disease in general practice and hospital referral patterns British Journal of Surgery 74(11): pp 1020-1022
[7] Edwards A G Robling M R Wilkinson C E et al 1999 The presentation and management of breast symptoms in general practice in South Wales The BRIDGE Study Group British Journal of General Practice 49(447): pp 811-812
[8] Samers J M Galetakis S Scott C J et al 2004 Breast cancer management: the perspective of general practitioners in inner and eastern Melbourne The Breast 13(6): pp 468-475
[9] Fatima N Liu L Hong S et al 2020 Prediction of breast cancer, comparative review of machine learning techniques, and their analysis IEEE Access 8: pp 150360-150376
[10] Islam M M Haque M R Iqbal H et al 2020 Breast cancer prediction: a comparative study using machine learning techniques SN Computer Science 1: pp 1-14
[11] Dataset https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset