Applying machine learning models to breast cancer prediction problem

Ziqi Mai

An De College, Xi'an University of Architecture and Technology, Xi'an, 710311, China

maiziqi@xauat.edu.cn

Abstract. Cancer has become the number one killer of human life and health. Therefore, a model that can predict cancer is able to help doctors to diagnose whether a patient has cancer or not, which can boost the accuracy of the diagnosis and enhance diagnostic efficiency, thus reducing the chance of misdiagnosis and other situations. This paper focuses on breast cancer prediction and adopted three machine learning based methods, including logistic regression, K-Nearest Neighbor, and decision tree models to build automatic solutions and investigate which model is more suitable for such a simple prediction problem. In this study, the detailed features, data collection and pre-processing approaches are presented to better understand such medical data. Then extensive experiments show that the accuracy scores of the three models are 97.08%, 94.89%, and 93.43%, respectively. Through comparison, it is concluded that the logistic regression model achieves the best performance for the breast cancer prediction task.

Keywords: classification problems, breast cancer predictions, logistic regression.

1. Introduction

When people's standard of living is increasing, more and more people have become concerned not only about how to live, but also how to live better, while disease is clearly a major obstacle to achieving a better life, with an increasing proportion of personal consumption being spent on medical care. In this context, precision treatment and predictive diagnosis have become hot topics today.

Among the many illnesses, cancer is a devastating disease that is rapidly overtaking all others as the leading cause of death in the globe. The IARC group claims, just in 2020 alone, nearly 10 million people will die of cancer worldwide [1]. The data shows that the most common deceases include breast cancer, lung cancer, colon and rectal cancer, prostate cancer, skin cancer (non-melanoma), and stomach cancer [1]. Table 1 shows some of the data from the IARC statistics.

Table 1. Number of global cases of cancer in 2020 from IARC.

Type of cancer	Quantity (/Ten thousand)
Breast cancer	226
Lung cancer	221
Colorectal and rectal cancer	193
Prostatic cancer	141

Table 1. (continued).

Skin cancer (non melanoma)	120
Gastric cancer	109
Total number	1010

There is no doubt that cancer is seriously threatening human life and health, therefore, the danger of cancer deserves high attention. It is urgent to research a model that can predict cancer, which can help doctors to judge patients' cancer more easily, improve the efficiency of diagnosis, avoid delaying patients' treatment due to long diagnosis time or wrong diagnosis as much as possible, and contribute to the life and health of contemporary people.

This study focuses on patient cancer prediction given the social setting. After determining the direction of the research, the next step is to analyse the nature of the problem. The process of predicting cancer is roughly as follows. Firstly, cancer prediction for a sample is done by examining a series of physical examinations of the sample, especially the tumor growth, and then a series of data are obtained and further analysed to determine whether the sample is in a healthy state (the tumor is benign) or a non-healthy state (if the tumor is malignant, the sample has cancer). Further analysis can be done not only to determine whether the sample has cancer or not, but also to analyse the type of cancer the patient has. As an example, Hossam H. Sultan and Nancy M. Salem et al., in their experiments on the classification of brain tumor assessment, divided the results into several categories such as meningioma, glioma, and pituitary tumor [2]. Therefore, the results of predicting cancer in a sample can be divided into two cases: in the general case, there are only two results: yes or no. In the special case, there are multiple results: no cancer and some type of cancer (breast cancer, lung cancer or stomach cancer, etc.). In fact, the special case is just a splitting of the "yes" result into various types of cancer on the basis of the general case.

This is one process of the classification problem in machine learning: input a series of data used to identify the type and output the category [3]. In summary, the problem of predicting a patient's cancer is a typical classification problem, and the two cases above correspond to dichotomous and multi-classification problems, respectively. The binary classification problem is commonly known as a classification problem with only two categories, such as whether it is spam or malignant tumor, which is a black-or-white problem.

Compared with multi-category problems, the data requirements for binary classification problems are not very strict. For the reason of convenience of data collection and good data processing of the binary classification problem, this experiment only discusses the binary classification problem, i.e., only predicting whether the sample has cancer or not. And since each cancer is judged by different criteria, the most frequently diagnosed type of cancer in the world today - breast cancer - is selected here. As illustrated in table1, the number of new cases of breast cancer in the world reached a staggering 2.26 million in 2020 alone, surpassing the number of new cases of lung cancer worldwide for the first time and becoming the number one cancer type in the world. The most prevalent illness that affects the health of middle-aged women, breast cancer has emerged as a severe threat to women's wellbeing [4]. The current methods of predicting breast cancer are mainly based on imaging and histological analysis, which are not only time-consuming and require doctors' extensive experience. Therefore, it is important to develop a predictive model based on clinical indicators for early breast cancer detection and therapy. In this experiment, it can use three methods, logistic regression, K-Nearest Neighbor (i.e., KNN), and decision tree, to build a prediction model that predicts breast cancer based on clinical indicators. This experiment collects a set of sample data for the experiment, constructs a correlation model, and compares the results with the diagnostic results to determine which of these three methods is more appropriate for building a predictive model. The major goal of this experiment is to find a better technique for building breast cancer prediction models in order to increase the accuracy of early detection of breast cancer.

2. Method

After determining the research topic, the formal prediction model construction is conducted. The following steps are the procedures of model construction:

1. data collection
2. data analysis
3. data pre-processing
4. prediction model training, i.e., logistic regression, KNN, decision tree, and prediction using the trained model.

2.1. Data collection

The data used in this experiment comes from the UCI Machine Learning Repository[1], a machine learning community that provides a large number of datasets. The dataset adopted is a set of breast cancer samples from Wisconsin, which contains a total of 699 data instances. Detailed features are shown in the part (a) and (b) of table 2:

Table 2 (a). Breast cancer dataset from Wisconsin.

Sample ID	Sample code number	Clump Thickness	Uniformity of Cell Size	Marginal Adhesion	Uniformity of Cell Shape
0	1000025	5	1	1	1
1	1002945	5	4	5	4
…	…	…	…	…	…
697	888820	5	10	10	3
698	897471	4	8	5	8

Table 2 (b). Breast cancer dataset from Wisconsin.

Sample ID	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
0	2	1	3	1	1	2
1	7	10	3	2	1	2
…	…	…	…	…	…	…
697	7	3	8	6	1	5
698	4	5	10	4	1	4

2.2. Data analysis

As can be seen from the part (a) and (b) of table 2, each sample has 11 attributes, namely Sample ID, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses, and Class.

(1) Sample code number: This attribute is just a general number to distinguish each sample, and is useless in prediction experiments.

(2) Clump Thickness: This property is one of the criteria for determining whether a sample has cancer or not.

(3) Uniformity of Cell Size: This attribute is used to determine whether there is any abnormal cell growth in the sample by the average growth of various cells, such as breast epithelial cells, and is usually combined with Uniformity of Cell Shape.

[1] https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/

(4) Uniformity of Cell Shape: This attribute is also a marker to determine cell carcinogenesis, but it cannot be used as the only marker to determine carcinogenesis because cells may change their shape under the condition of aging, so it needs to be combined with Uniformity of Cell Size.

(5) Marginal Adhesion: This attribute is the ability to determine the metastasis of cells, if it is too high, the chance of cancer is greatly increased.

(6) Single Epithelial Cell Size: Breast cancer is caused by the canceration of mammary epithelial cells. Thus, it is necessary to do a detailed examination of the growth of the cells.

(7) Bare Nuclei: This property refers to cells with nuclei that are exposed for some reason, which helps to determine if the cells are necrotic.

(8) Bland Chromatin: The chromatin of cancerous cells is irregularly granular or clumped with clear interstices, sometimes in the form of ink dots. The probability of the cells' cancer can be determined by comparing with the chromatin of normal cells.

(9) Normal Nucleoli: Since cell carcinogenesis can lead to larger and more nucleoli, whether a cell has normal nucleoli or not is a criterion for determining cell carcinogenesis.

(10) Mitoses: It can be determined whether a sample cell is normal or not by observing whether it is able to mitigate normally.

(11) Class: This attribute is actually the target output. There are only two values, i.e., 2 and 4, which represent that the tumor of the sample is benign (meaning that the sample is not cancerous) and the tumor of the sample is malignant (meaning that the sample has cancer), respectively.

It can be found that among the above 11 attributes, Sample code number is the attribute that cannot be used in the prediction model. Therefore, it needs to be excluded from the experiment. The remaining 10 attributes can be divided into two parts, "Class" alone forms one part, which is the expected result needed for the prediction model, while the remaining 9 items are the other part, which are all describing the tumor and cellular characteristics of the sample, and the comprehensive analysis of a set of 9 items of data can lead to the corresponding sample's "Class".

After determining the basic structure of the data used, it is necessary to determine the machine learning methods. Machine learning is the process of using existing data (experience), finding patterns in it, arriving at some kind of model, and using that model to make predictions on unseen data [4].

For classification problems, the specific behaviour of machine learning is to learn from existing data sets, find patterns in them, and classify unseen data. Machine learning is divided into supervised and unsupervised learning according to the difference of training set data.

2.2.1. Supervised learning. Supervised learning allows a computer to learn a function (Mapping Relationships) from a given training dataset that can be utilized to forecast outcomes when new information comes in, and this training set is necessary to include inputs and outputs, which may also be described as characteristics and targets [5]. The targets in the training set are manually annotated.

In brief, there exists a set of variables as inputs that have different effects on the output, which are then used to predict the output values [6].

Supervised learning is one of the most frequent classification issues. An ideal model is trained using the already-existing training samples (i.e., known data and their associated outputs), which is then utilized to map all the inputs to the outputs and to make straightforward classification decisions regarding the outputs.

2.2.2. Unsupervised learning. Unsupervised learning has no obvious consequences, as the inputs are not labeled, unlike supervised learning.[2] The class of the sample data is unknown, and it is necessary to classify (clustering) the sample set based on the similarity between samples in an attempt to minimize the intra-class gap and maximize the inter-class gap. The classifier design can only be learned from the original sample set without sample labels because, in many circumstances, in

[2] https://developer.ibm.com/articles/cc-machine-learning-deep-learning-architectures/

actuality, the labels of the samples are not known in advance, i.e., there is no category corresponding to the training data.

The difference between the two is that: 1. Supervised learning methods must have a training set and test samples. The law is found in the training set, and the law is used for the test samples [7]. The unsupervised learning method does not have a training set, but only a set of data, and looks for patterns within that set of data. 2. The supervised learning method is to identify things, and the result of the identification is shown by adding labels to the data to be identified. In the case of unsupervised learning, only the data set itself needs to be examined, and there are no labels in advance. 3. The unsupervised learning method is looking for regularities in the data set, and such regularities do not necessarily have to achieve the purpose of dividing the data set. The unsupervised learning method looks for regularities in the dataset, which do not necessarily lead to a division of the dataset, that is, they do not necessarily "classify" it.

For the dataset used in this experiment, the data has been clearly divided into two parts, corresponding to the input and the label, so the machine learning used in this experiment is supervised learning.

2.3. Data pre-processing

This experiment has been determined to use supervised learning, and supervised learning should be accompanied by a training set and a test sample (test set), then the 699 pieces of data need to be partitioned. But before that, it needs to be confirmed whether this dataset can be used for every piece of data. This is because raw data without any processing often contain many errors that are not conducive to further experiments [8]. There exist some samples with unknown data in the data file, representing by "?" instead. Due to the missing data, these data can neither be used for training the model nor for the final prediction, so this part of the sample needs to be excluded. The pandas library is a powerful Python data analysis support library that provides high-performance, easy-to-use data structure and data analysis tools. Specifically, the replace() and dropna() functions are used to conduct data pre-processing. To facilitate computer identification of missing data, The replace() function is first called to replace the "?" in the data with "NAN". The replace() function is again used to replace the "?" in the data with "NAN", and use dropna() to delete the rows (samples) containing NAN.

After removing the samples with missing values, there are still 683 samples within the dataset, and the following step is to split the 683 data into a training set and a test set. For convenience, the whole data set is divided into two groups: feature values and target values, which are stored in x and y respectively. To be specific, the feature dataset and target dataset are divided into training set and test set, respectively, where the test set accounts for 20% of the medium dataset.

After completing this step, the grouping of data has been completed, and generally speaking, the data can be used for model training at this point. However, there is still a risk that can be easily overlooked at this point. Table 2 shows that the transactions described by the 9 feature values are completely different, which leads to differences in their units. When the feature values have different scales and units, it will have an impact on the data analysis, which may amplify the error in the final prediction. Moreover, if the difference is too large, it may even make the algorithm unable to learn the other features, resulting in learning failure. Thus, it is also necessary to normalize the feature values.

Standardization refers to the dimensionless conversion of data to the same specification for data of different scales as a solution to the problem of comparability between data [9]. This is performed by mapping the data to a uniform interval, [0,1] by default [9].

In this work, the StandarScaler() function of the sklearn library (version 1.0.2) is used to normalize the feature values. Since then, the processing of the data has been basically completed, and the model can be trained.

2.4. Prediction model training

2.4.1. Logistic regression. Although logistic regression has the word “regression” in it, it is a classical algorithm for solving classification problems. However, the linear regression method, a “true” regression method, cannot be used to solve classification problems, since classification problems have only a limited number of discrete outcomes, and can use 0, 1 (in the binary cases) or 0, 1, 2, etc. (in the multi-classification cases) to represent different classifications, this is non-linear [10]. The linear function may have these values, but it must also contain other values in addition to these values, and the values of the linear function are continuous. Therefore, linear regression cannot be used to solve classification problems.

The distinction between logistic regression and regression: regression examines the association between the independent variable (characteristics) and the dependent variable (target). Logistic regression uses a sigmoid function to restrict the dependent variable (target) to values between 0 and 1, while the dichotomous classification problem goes a step further by classifying values less than 0.5 as 0 and values greater than 0.5 as 1 individually. Logistic regression can be applied to classification problems because the output is discrete, but the output of regression problems is continuous, so classification problems can be solved by logistic regression instead of regression models.

The graph of the sigmoid function is an S-shaped curve, as shown in figure 1. It always returns a probability value between 0 and 1. The sigmoid function is used to convert an expected value into a probability. This function converts any real number to a number between 0 and 1. The sigmoid function is used to convert predictions into probabilities in machine learning. The following is the formula for sigmoid function:

$$f(z) = (1 + \mathrm{e}^{-z})^{-1} \tag{1}$$

$$\mathrm{z} = \mathrm{w}^{\mathrm{T}}\mathrm{x} + \mathrm{w}_0 \tag{2}$$

where *x* is input features, *w* and w_0 are model parameters.

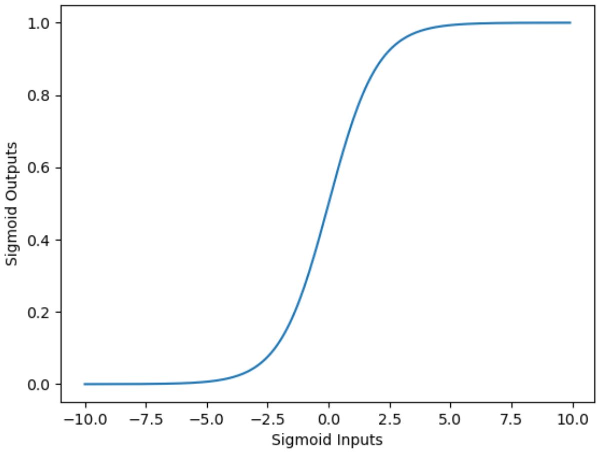

Figure 1. The curve of sigmoid function.

Indeed, the sigmoid function is a conversion of an actual number to a number between 0 and 1. Therefore, one can simply think of logistic regression as the combination of linear regression and the sigmoid function. The result of linear regression is a variety of values, and the result of the regression function can be converted to a number between 0 and 1 by using the sigmoid function, which is the result as expected, i.e., the result of logistic regression (0-1). Further processing can cast the final result values below 0.5 as label 0 and those in excess of 0.5 as label 1 [11].

The above is the theoretical basis of logistic regression. In the concrete implementation, the LogisticRegression() model can be directly imported from the sklearn library and trained over the training set for training. Upon completion of the training, the predict() function is used to make predictions about the test set. The predicted results are shown in table 3:

Table 3. The prediction results of logistic regression method.

The number of 2 (benign tumor)	The number of 4 (malignant tumor)	sum total
89	48	137

The number of samples with cancer, the number of samples without cancer and the total number of test sets make up the three prediction data. Table 4 displays the actual outcome:

Table 4. Actual results of the sample.

The number of 2 (benign tumor)	The number of 4 (malignant tumor)	sum total
91	46	137

2.4.2. K-Nearest neighbor. The KNN algorithm is a classification algorithm associated with supervised learning. The core idea is that in a certain data set, if a sample and k other samples in the data set have many characteristics and most of these k samples fall into the same group, then this sample also falls within that category, that is, the category of the nearest one or more samples determines the category of the sample that needs to be classified. Specifically in two dimensions, after determining a value of k, the k nearest points to the point to be classified are selected, and these points may belong to different categories, then the point to be classified belongs to the category that contains the most points.

Regarding the calculation of the distance, the Euclidean distance is usually used [12].

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2 \ldots} \tag{3}$$

Here x, y, z correspond to different eigenvalues. By this method the distances between the samples to be classified and the known samples are calculated and sorted and wait for the k-value to be determined.

Therefore, it can be found that the selection of k value is very important. k indicates the number of neighbors, if k is small, the complexity of the model will increase and the training error will decrease. if k is large, the complexity of the model will decrease and the training error will increase. It is not preferable to choose a smaller value of k because doing so will result in overfitting, while choosing a bigger number will result in underfitting.

In this experiment, the k-fold cross-validation method and learning curve are used to obtain the optimal k value. Cross-validation, as the name suggests, it means the repeated use of data, segment the resulting sample data and combine them in different sets of training and testing, utilizing the training set to form the model, using the set of tests to assess whether the model's prediction is right or wrong. In that framework, the several different training sets and test sets can be obtained, and a sample in one training set may become a sample in the next test set, which is called "crossover".

The k-fold cross-validation involves randomly dividing the data set into k identical-sized, mutually exclusive subsets, and then randomly select one copy of the data as the test set and the rest k-1 copies as the training set, and then repeat the series of operations. After a number of rounds (less than k), the result of averaging k times is taken as the final average accuracy of the model.

As for the problem of selecting the optimal k value, this experiment chooses the optimal k value on the interval of [1,20]. The selection method is to do cross-validation for each k-value through a loop operation to find out the average accuracy corresponding to these 20 k-values, and compare the k at the maximum average accuracy, and use it for the final model construction. And in the loop, this experiment uses 5-fold cross-validation, dividing the training set into 5 equal parts, in turn, as a sub-test set (which should be better called the validation set at this point), and the remaining sub-training set is still used to train the validation model, and finally the average accuracy and variance are obtained. The cross_val_score function from sklearn is used here.

The KNN model is trained with k values ranging from 1 to 20 with step of 1 using the cross-validation approach, and it is required to choose an appropriate k value.

After performing the above operation for all the candidate values of k to obtain the corresponding mean accuracy and variance, these data can be demonstrated through a learning curve. The designed learning curve is presented in figure 2:

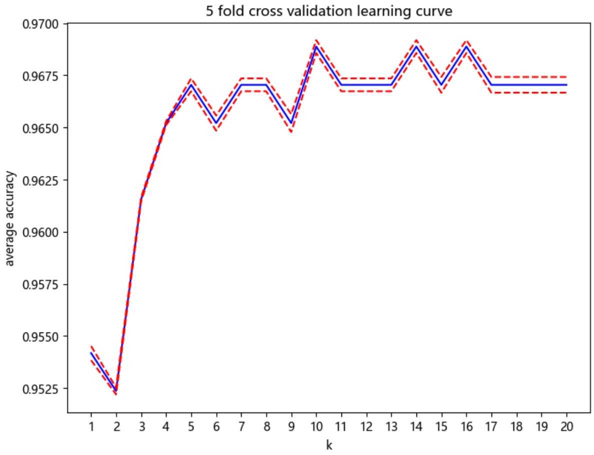

Figure 2. The learning curve of 5-fold cross-validation.

The optimal value of k is the one that has the maximum average accuracy, where k is set to 14 and the average accuracy of cross-validation settings is 96.89%. Thus, the value of k is set to 14 to train the KNN model and make predictions. The predicted results are shown in table 5:

Table 5. The prediction results of K-nearest neighbor algorithm.

The number of 2 (benign tumor)	The number of 4 (malignant tumor)	sum total
84	53	137

And the real situation is listed in table 6:

Table 6. Actual results of the sample.

The number of 2 (benign tumor)	The number of 4 (malignant tumor)	sum total
79	58	137

2.4.3. Decision tree. A decision tree is a tree-based model for solving classification and regression problems. It splits a large problem into many smaller problems each of which is associated with a decision. This is an inductive learning algorithm that emphasizes the transformation of apparently messy and congested known data into a tree structure that is capable of predicting unknown data. The tree structure consists of a tree with a root node, inner nodes and leaf nodes, where a decision-making rule can be represented by a path from the root node to a leaf [13].

A decision tree actually represents a mapping relationship between input features and output values, in which each inner node represents a feature and each leaf node represents a class. Decision tree learning essentially generalizes a set of classification rules, or a sequence of eigenvalue classifications, from a training set to improve the purity of disordered data and make it as orderly as possible. Then how to select the attributes is an important issue in building a decision tree. In this experiment, the CART tree, in short for categorical regression tree is adopted, where dichotomy is used at each node, i.e., classification regression tree is a binary tree [14]. The acronym CART tree indicates that it can be used simultaneously for both regression and classification. Since this experiment deals with a classification problem and the variables that depend on the dataset are discrete values, the CART classification tree is adjusted using the category with the greatest likelihood of the leaf node as the expected category for that node [14].

This experiment uses the Gini index to measure the selection of attributes, which is a criterion for finding the optimal classification of attributes. The Gini index means the standard uncertainty, which is commonly known as the probability that a randomly selected sample in the sample set is divided wrongly. A smaller Gini index refers to a smaller the probability that a sample is wrongly selected, and higher purity of the data set. If all samples in the set are of one category, the Gini index is 0. The Gini index is calculated by the following formula:

$$Gin\mathrm{i} = 1 - \sum_{n=1}^{N} p_n^2 \tag{4}$$

The meaning of the parameter $\mathrm{p_n}$ is the proportion of samples belonging to the n-th type in the sample set, which can also be described as the probability of occurrence. Specifically, in this paper, there are two categories of having cancer and not having cancer. The calculation is made by dividing the number of occurrences of each category by the total number of samples to obtain the p-value of each category, and then calculating it according to the formula to obtain the Gini coefficient. After calculating the Gini index of all nine feature values, the feature with smallest Gini value is selected as the root node for the initial decision. On this basis, the Gini index of the remaining 8 feature values is calculated again, and the smallest one is again selected as the internal node for the decision tree. Such a step is repeated until the target value is obtained by completing the decision making of 9 feature values, so that a set of classification rules is found.

In this experiment, the tree module in sklearn is used. As with the two methods above, the DecisionTreeClassifier() model is imported directly and then trained over the training set. Afterward, the predicted outcomes of the test set are listed in table 7:

Table 7. The prediction results of decision tree.

The number of 2 (benign tumor)	The number of 4 (malignant tumor)	sum total
84	53	137

The statistics of the ground-truth labels is listed in table 8:

Table 8. Actual results of the sample.

The number of 2 (benign tumor)	The number of 4 (malignant tumor)	sum total
79	58	137

The decision tree is then sketched. Figure 3 demonstrates a succinct five-layer decision tree:

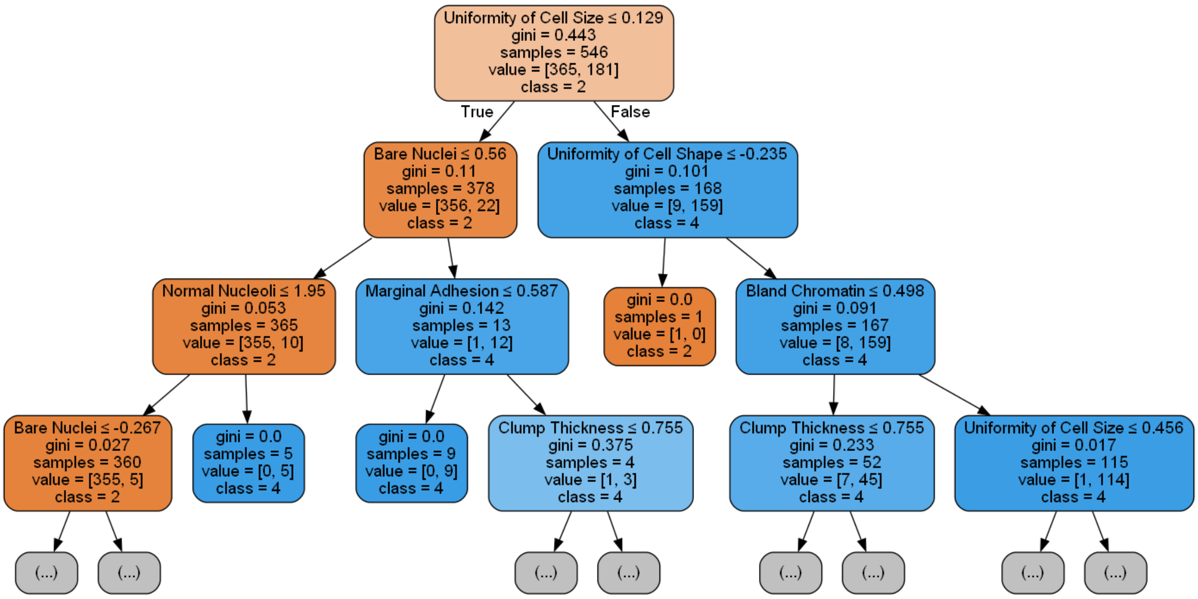

Figure 3. The results of decision tree model.

3. Results

The above three models are further evaluated using two evaluation metrics, including confusion matrix and accuracy.

The confusion matrix is the most straightforward and logical technique to assess the accuracy of a typed model. It is a standard format for accuracy evaluation shown as a matrix. A number of secondary indicators, including accuracy, recall, and precision, can be found based on the confusion matrix [15]. The principle of confusion matrix is displayed in table 9.

Table 9. The schematic diagram of confusion matrix.

Confusion Matrix		True Value	
		Positive	Negative
Predicted value	Positive	TP	FP
	Negative	FN	TN

True Positive (TP): True positive class. The sample’s true class is the positive class, and the model recognition’s outcome is also the positive class.

False Negative (FN): False negative class. The sample actually belongs to a positive class, but the model perceives it as a negative class.

False Positive (FP): False positive class. Although the sample actually belongs to a negative class, the model interprets it as a positive class.

True Negative (TN): True negative class. The model understands that the sample actually belongs to a negative class, which is also its true class.

The confusion matrices for the three models are illustrated as figure 4, figure 5 and figure 6, as below:

(1) Logistic Regression

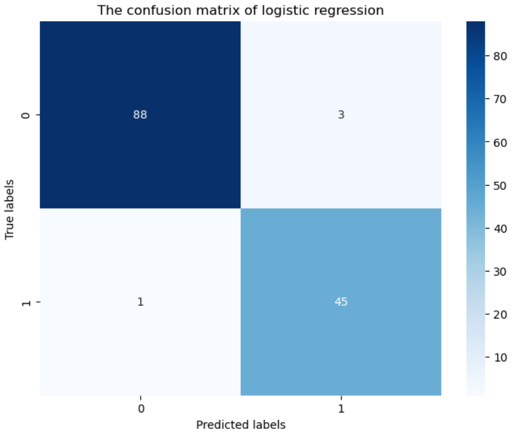

Figure 4. The confusion matrix of logistic regression model.

(2) KNN

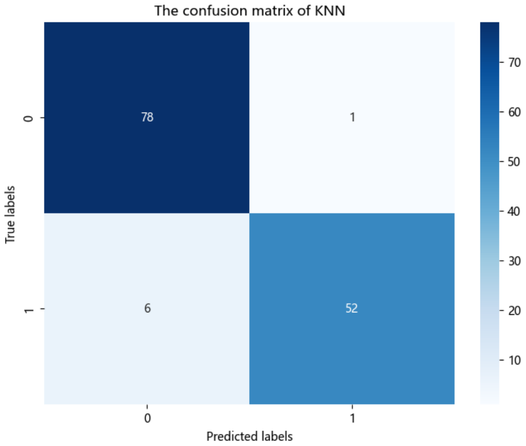

Figure 5. The confusion matrix of KNN model.

(3) Decision Tree

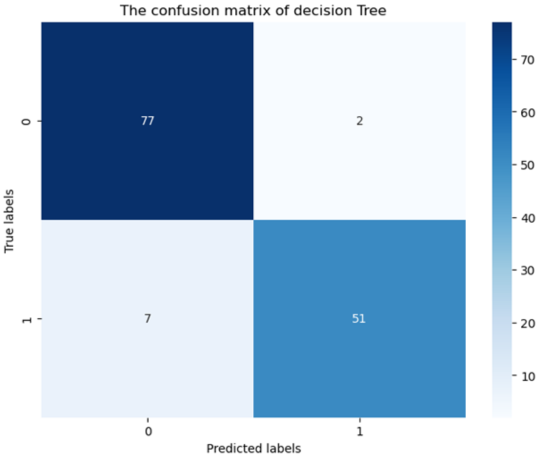

Figure 6. The confusion matrix of decision tree model.

Accuracy is a secondary evaluation metric derived from the confusion matrix. It can be used to illustrate the model's correctness, or the ratio of the number of accurate models to the total samples. Generally speaking, a model's effectiveness increases with its accuracy.

The accuracy calculation formula is as follows:

$$\text{Accuracy} = \frac{(\text{TP+TN})}{(\text{TP+FN+FP+TN})} \tag{5}$$

Table 10 shows that, despite the little differences between the three models, it is concluded that the logistic regression model achieves the highest accuracy score, followed by KNN, while the decision tree performs the worst.

Table 10. The accuracy of the three models.

Model	Accuracy
Logistic Regression	97.08%
KNN	94.89%
Decision tree	93.43%

4. Discussion

The accuracy scores for each model have been derived. When constructing the decision tree model, it is observed that for each decision, the Gini index of each eigenvalue is calculated and compared, where the smallest Gini index is selected for the current step. Therefore, it can be concluded that the impact of each eigenvalue on the final target value might be different. Here, the percentage of influence of each eigenvalue on the experimental results among all nine eigenvalues can be discussed, which is also known as the feature weight (feature importance). In this experiment, the influence of each feature on the results is calculated only in the logistic regression model.

The regression coefficients are further ranked and illustrated in figure 7:

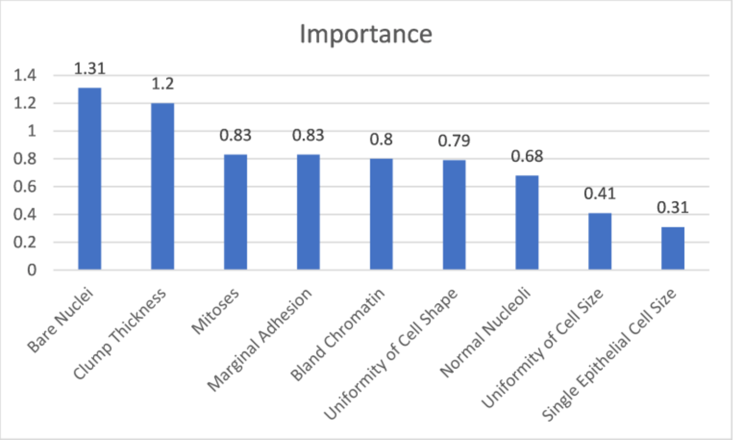

Figure 7. The histogram of feature importance ranking.

From the figure 7, it can be concluded that the effect of Bare Nuclei on the target value is the largest among the nine eigenvalues.

5. Conclusion

This paper presents three predictions solutions for breast cancer by training three models: logistic regression, KNN, and decision tree. From the results, logistic regression model is the best, KNN is the second best, and decision tree is the worst. Therefore, logistic regression is the best choice for the breast cancer prediction task. First of all, the design of this model fits well with the cancer prediction data. The main purpose of cancer prediction is to determine whether the patient has cancer or not, and the result is only yes or no. The sigmoid function used in logistic regression can summarize the function values as 0 and 1, which can represent yes and no respectively. In logistic regression, each feature value is independent of each other, so it is possible to use the function to make a more accurate representation of the feature value and the target value, which can be more visual and intuitive, and clearly present the change of the relationship. Second, the logistic regression model is clear, and the probability derivation behind it can stand up to scrutiny. The most important thing is that the implementation is simple and efficient. As for the remaining two models, KNN is an unsupervised algorithms, resulting in slower predictions than algorithms like logistic regression. It needs a pre-defined k-values and is computationally costly, especially when the number of features is quite large. Decision trees tend to produce an overly complex model, and the generalization performance of such a model is very poor. Furthermore, a small change of data could lead to the original model failing, since each feature value affects each other. Therefore, even though the principle of logistic regression is relatively simple, its model is very stable and efficient.

This experiment provides some data basis for the use of future cancer prediction models and helps to improve cancer models in the future. However, there are still shortcomings in this experiment. Firstly, the amount of available data for the experiment only contains only 683 instances, which greatly limits the reliability of the experiment. The experiment only focuses on breast cancer, but not other types of cancers in general. Applying larger and more diverse data to improve the generalization of the logistic regression model for cancer prediction will be conducted in the near future.

References

[1] Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I and Bray F 2020 Global Cancer Observatory: Cancer Today *International Agency for Research on Cancer.*

[2] Sultan H H, Salem N M and Al-Atabany W 2019 Multi-classification of brain tumor images using deep neural network IEEE access **7** p69215-25.

[3] Domingos P 2012 A few useful things to know about machine learning *Communications of the ACM* **55.10** p78-87.

[4] Kumari M and Singh V 2018 Breast cancer prediction system *Procedia computer science* **132** p371-376.

[5] Nasteski V 2017 An overview of the supervised machine learning methods *Horizons. b* **4** p51-62.

[6] Ziegel E R 2003 The elements of statistical learning *Technometrics* **45.3** p267-8.

[7] Learned-Miller E G 2014 Introduction to supervised learning *I: Department of Computer Science, University of Massachusetts* **3.**

[8] Kaleem A, Ghori K M, Khanzada Z and Malik M N 2011 Address standardization using supervised machine learning *interpretation* **1.2** p10.

[9] Ali P J M, Faraj R H, Koya E, Ali P J M and Faraj R H 2014 Data normalization and standardization: a technical report *Mach Learn Tech Rep* **1.1** p1-6.

[10] DeMaris A and Selman S H 2013 Logistic regression *Converting Data into Evidence: A Statistics Primer for the Medical Practitioner* p115-36.

[11] Khairunnahar L, Hasib M A, Rezanur R H B, Islam M R and Hosain M K 2019 Classification of malignant and benign tissue with logistic regression *Informatics in Medicine Unlocked* **16** p100189.

[12] Peterson L E 2009 K-nearest neighbor *Scholarpedia* **4.2** p1883.

[13] Song Y Y and Ying L U 2015 Decision tree methods: applications for classification and prediction *Shanghai archives of psychiatry* **27.2** p130.

[14] Daniya T, Geetha M, and Kumar K S 2020 Classification and regression trees with Gini index *Advances in Mathematics: Scientific Journal* **9.10** p8237-47.

[15] Liang J 2022 Confusion Matrix: Machine Learning *POGIL Activity Clearinghouse* **3.4.**