DOI: 10.54254/2755-2721/15/20230833

Comparison the effects of KNN and linear regression models in lung cancer prediction

Yi Zhou

Bishop Allen Academy, Ontario, M8Y 2T3, Canada.

zhouy014@tcdsb.ca

Abstract. Lung cancer has a range of major factors like smoking, yellow gingers, anxiety, etc. now, the problem of this research is that prediction for lung cancer. Prediction for lung cancer is a complex problem that is not suitable for human prediction. This research using a dataset was from Kaggle. There are 16 rows and 309 columns. To determine the k nearest neighbors (KNN) algorithm and linear regression algorithm, which one is better for prediction for lung cancer, and which coefficient will be best effective. This research uses mixed method research. In this work, when the K of the KNN algorithm equals 7 or 2, the effectiveness of the KNN model is best, when the alpha of the linear regression algorithm equals 20, the effectiveness of the linear regression model is best. The KNN model is better than the linear regression model, though the difference is negligible. In the future, more emphasis can be placed on using a wider range of algorithms or using more extensive and generalized dataset, as well as assessing the efficiency of the algorithm on larger datasets.

Keywords: lung cancer prediction, machine learning, KNN.

1. Introduction

Lung cancer, one of the most widespread and hazardous tumours in the world, has attracted attention for its relevance [1,2]. The principal factors linked to the occurrence of lung cancer involve smoking, chronic lung disease, among others [3]. The main treatment modalities for lung cancer include surgery, radiation therapy, chemotherapy, and immunotherapy [4].

Machine learning is the branch of artificial intelligence, it can find the formula and the law in the dataset. it will learn and improve in the large data. And then, they will be used to predict, classify, and make decisions.

Machine learning has a unique advance in in prediction problems in big data. It very easy that to process large amounts of data. For prediction of lung cancer that have utmost large, and many factors, machine learning is suited by. They can be based on large dataset, use quantitative analysis like mathematical statistics. There are more effective than human [5,6]. Human cannot use the large data, and quantitative analysis to predict the likelihood of the lung cancer.

It is important to research it for the health and safety of humans. In order to more batter for predicts the likelihood of one person who has the probability to get lung cancer in the future. Therefore, the research will discuss that two predict model about KNN model and linear regression model. They will be compared with the effect of those two models, and the effect of each different parameter of those models.

2. Method

2.1. Dataset

This research has been using the data of lung cancer in a website Kaggle that was updated by Mysar Ahmad Bhat. The dataset has 15 characteristic variables including gender (M/F), age, fatigue, anxiety, peer pressure, smoking, swallowing difficulty, chest pain, yellow gingers, wheezing, alcohol consuming, coughing, chronic disease, allergy, shortness of breath, and the label that whether has cancer (YES/NO). Then, the data has been pre-processed, by mapping discrete values to a specific number, such as mapping M to 1, F to 2, YES to 1 and NO to 2.

2.2. KNN model

K-Nearest Neighbors Algorithm: An Overview of the Procedure and Its Applications K-Nearest Neighbors is a popular machine learning strategy for classification and regression applications (KNN). It is a non-parametric, lazy learning approach that makes no assumptions about the underlying distribution of the data. [7].

The K-Nearest Neighbor (KNN) approach is a well-known supervised machine learning methodology for classification and regression applications. It is a simplistic algorithm that is easy to use, but it has some limitations, such as high computational cost and the curse of dimensionality [8].

It's more like a classifier that puts a label on an unknown thing. It's easy and easy to understand, but it is a lazy algorithm, which requires a large amount of memory. And it is computationally intensive and has low performance when classifying test samples; by the way it also has Poor interpretability. The results are highly logical and interpretable, but it is difficult to express highly complex data.

2.3. Linear regression model

Goodfellow et al. discuss linear regression as a simple and widely used method for regression tasks. They highlight its ability to a linear connection between the variables in the input and output, as well as its interpretability and ease of implementation. However, they also note that Linear regression is used to presume that input and output variables have a linear relationship, which may not hold in many real-world scenarios. Additionally, linear regression can be sensitive to outliers in the data and may not perform well in situations with high-dimensional input spaces [9].

Kutner et al. describe linear regression as a flexible and powerful tool for modelling relationships between variables. They note its ability to handle both quantitative and qualitative predictors and its simplicity in interpretation. However, they also discuss the potential for overfitting in linear regression models, particularly when the number of predictors is large relative to the sample size. They also note that in linear regression, input and output variables are believed to have a linear relationship, and that violations of this assumption can lead to poor model performance [10].

Using supervised machine learning, the linear regression technique can forecast continuous numerical variables. The algorithm establishes a linear model to describe the relationship between the independent and dependent variables. The advantages of linear regression are strong interpretability and generalization ability, but its disadvantage is that it may perform poorly on non-linear problems.

2.4. Evaluation metrics

This reach has use accuracy, precision, recall, F1 score. They are jointly used to measure the performances of the model. The accuracy represents the proportion of samples that were correctly predicted to all samples. Precision is the correct positive forecasts divided by all positive forecasts. The recall is the proportion of accurately predicted positive samples to all positive samples. The F1 score is the weighted harmonic mean of recall and accuracy.

3. Result

3.1. Comparison between KNN and linear regression

In this reach. The algorithm uses KNN model and linear regression model to predict the probability of lung cancer, as demonstrated in Table 1. Obviously, the effect of linear regression is better than another one.

Table 1. Comparison results between KNN and linear regression.

algorithm	accuracy	precision	Recall	F1_score
KNN	95.16%	98.31%	96.67%	97.48%
L-R	96.77%	98.33%	98.33%	98.33%

The advantage of KNN algorithm is that it is simple to use, flexible in handling outliers, and performs well when the training dataset is large. However, the KNN algorithm has a high computational complexity, leading to poor performance when processing large datasets. Additionally, because the KNN algorithm is an instance-based learning algorithm, it cannot abstract and generalize features, which may result in overfitting.

In comparison, a linear regression algorithm performs better in handling large-scale datasets due to its lower computational complexity. Additionally, the algorithm can abstract and generalize features to avoid overfitting. However, the linear regression algorithm struggles with outliers because it is based on the least squares method, which can be affected by outliers. Moreover, the performance of the algorithm may decrease when there are multiple features in the dataset.

Above all, KNN and linear regression algorithms have their respective downsides and benefits, as well as the selection of algorithm depending on the dataset's characteristics and the requirements of the problem. In this study, KNN and linear regression algorithms are chosen for prediction because of the relatively small dataset size. For larger datasets, linear regression algorithms may be more suitable.

Next, the impact of k value in KNN algorithm and the regularization coefficient alpha in the linear regression algorithm are further explored.

3.2. Effectiveness of k in KNN models

While using the KNN algorithm, the choice of K has a large effect with the result of prediction. So, there are a range of experiments to research the impact of different k values on algorithm performance and the best one to predict.

There are 11 groups for the K that 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 be used to research what impact of each K value and which one is the best for algorithm performance as demonstrated in Table 2. During evaluation the same metrics are leveraged.

Table 2. Effectiveness of K in KNN.

K	Accuracy	Precision	Recall	F1 Score
1	91.9%	98.2%	93.3%	95.7%
2	96.8%	98.3%	98.3%	98.3%
3	95.2%	98.3%	96.7%	97.5%
4	96.0%	98.3%	98.3%	98.3%
5	95.2%	98.3%	96.7%	97.5%
6	96.8%	96.8%	100%	98.4%
7	96.8%	98.3%	98.3%	98.3%
8	96.8%	96.8%	100%	98.4%
9	96.8%	96.8%	100%	98.4%
10	96.8%	96.8%	100%	98.4%
11	96.8%	96.8%	100%	98.4%

This table shows that when k larger than 7 and equal 6, the recall be 1 is best, but the precision be lowest value. But the K equals 7 and 2, those 4 values are relatively high. So, the best one is 7 or 2.

3.3. Effectiveness of alpha in linear regression models

While the linear regression algorithm, the choice of the regularization parameter has a large effect with the result of prediction. So, there are a range of experiments to research the impact of different regularization parameters on algorithm performance and the best one to predict.

Table 3. Effectiveness of alpha in linear regression model.

alpha	Accuracy	Precision	Recall	F1 Score
0 .01	96.77%	98.33%	98.33%	98.33%
0 .02	96.77%	98.33%	98.33%	98.33%
0 .05	96.77%	98.33%	98.33%	98.33%
0 .1	96.77%	98.33%	98.33%	98.33%
0 .5	96.77%	98.33%	98.33%	98.33%
1	96.77%	98.33%	98.33%	98.33%
5	96.77%	98.33%	98.33%	98.33%
10	96.77%	98.33%	98.33%	98.33%
20	98.39%	98.36%	100%	99.17%
50	96.77%	96.77%	100%	98.36%
100	96.77%	96.77%	100%	98.36%

While alpha is less than 20, the accuracy, precision, recall, F1-score is the same. When alpha equals 20, the accuracy, precision, recall, F1-score is the largest. But over the value, those are lower, except recall be 1.so the best value is 20.

4. Conclusion

This research uses mixed method research to determine that the K equal 7 or 2 the KNN model is best effective, the alpha equal 20 the linear regression has best effective, each model's difference is negligible, however the KNN model is a little bit more effective than the linear regression model. This research uses KNN model and linear regression model with L2 regularized to predict the probability of lung cancer. Above all, while the K equals 6 or 2, the performance of the KNN model is best. And while the alpha equals 20, the performance of linear regression is best. But each model's difference was negligible.

In future research, lung cancer prediction can be achieved using other methods such as deep learning and neural network learning. Then test the accuracy and precision for them. Consider the advantages and disadvantages of each algorithm. Identify the most promising algorithms. Additionally, the data can be changed and added, making the dataset more extensive. including lung cancer data from multiple countries to increase the generalizability and applicability of the results. Further investigation, testing, and analysis should be conducted to improve the accuracy of the conclusions. By the way, the KNN algorithm and linear regression algorithm can be researched again in larger dataset, determining that each algorithm whither will be good for prediction, maybe KNN algorithm will be less effective, since the KNN algorithm is bad in too large dataset.

References

[1] Zhou, B., Zang, R., Zhang, M., Song, P., Liu, L., et, al. (2022). Worldwide burden and epidemiological trends of tracheal, bronchus, and lung cancer: A population-based study. EBioMedicine, 78, 103951.

[2] Mathur, P., Sathishkumar, K., Chaturvedi, M., Das, P., Sudarshan, K. L., et, al. (2020). Cancer statistics, 2020: report from national cancer registry programme, India. JCO global oncology, 6, 1063-1075.

[3] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 71(3), 209-249.

[4] Siegel, R. L., Miller, K. D., & Jemal, A. (2019). Cancer statistics, 2019. CA: a cancer journal for clinicians, 69(1), 7-34.

[5] Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386.

[6] Bell, J. (2022). What is machine learning?. Machine Learning and the City: Applications in Architecture and Urban Design, 207-216.

[7] Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., et, al. (2020). An introduction to machine learning. Clinical pharmacology & therapeutics, 107(4), 871-885.

[8] Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 1-27.

[9] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

[10] Senter, H. F. (2008). Applied Linear Statistical Models . Michael H. Kutner, Christopher J. Nachtsheim, John Neter, and William Li. Journal of the American Statistical Association, 103, 880-880.