International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Cancer Prediction Using Machine Learning Algorithms

Kotla Lakshmi Narayana

Department of Information Technology, GMR Institute of Technology

ABSTRACT:

Cancer is a condition where specific body parts cells grow in a uncontrol manner. The cancerous cells can foray and destroy girding healthy towel, including organs. There are more than 200 differing types of cancer, and each is diagnosed and treated in an exceedingly way. Cancer has identified a various condition of several various subtypes. The timely webbing and course of treatment of a cancer form is now a demand in early cancer exploration because it supports the medical treatment of cases. Many research teams studied the applying of Machine Learning methods within the field of biomedicine and bioinformatics within the classification of individuals with cancer across high-risk or low-risk categories. These techniques have therefore been used as a model for the event and treatment of cancer. It's important that ML instruments are able of detecting crucial features from complex datasets. The methods are widely used for the event of predictive models for predicating a cure for cancer are random forest (RF), support vector machine (SVMs) and decision trees (DTs), etc. In this, compare the performances of different machine learning algorithms on cancer prediction.

Keywords: Cancer, Machine Learning, Random Forest, Support Vector Machines, Decision tree.

1. INTRODUCTION

Cancer may be a disease during which a number of the body's cells grow uncontrollably and spread to other parts of the body. Cancer could be a global health concern that causes death worldwide, in keeping with the globe Health Organization. Cancer can start almost anywhere within the physical body, which is formed of trillions of cells. the number of Indians affected by cancer is projected to extend to 29.8 million in 2025 from 26.7 million in 2021. Early cancer diagnosis is related to significantly higher survival rate and lower mortality and associated costs. Early-stage cancers require less complex treatment regimens and reduced hospital utilization, leading to reduced healthcare costs, whereas late-stage cancers require complex multimodal management, several rounds of extremely expensive drugs over significant periods of your time, and also the treatment of recurrences, equating to a staggering economic burden. Therefore, the importance of early diagnosis can't be overestimated. Treating cancer early has significant cost saving benefits. Therefore, we'd like predict within the initial stage itself to save lots of more lives from this dangerous disease. There are many Machine Learning methods to predict the disease. Past research papers said that supported dataset accuracy changed for the identical algorithm also.

2. LITERATURE SURVEY

[1]. Alfayez, A. A., Kunz, H., & Lai, A. G. (2021). Predicting the risk of cancer in adults using supervised machine learning: A scoping review. BMJ open, 11(9), e047755.The study demonstrates the aptitude of ML models to predict the number of patients is tormented by cancer which is presently most typical disease within the world. There are many methods utilized in the prediction like Logistic Regression (LR), Decision Tree (DT), Naïve Bayes (NB), Support Vector Machines (SVM). Noted accuracies are 0.61,0.51,0.59 and 0.52 for LR, DT, NB and SVM respectively.

[2]. Jasim, M., Machado, L., Al-Shamery, E. S., Ajit, S., Anthony, K., Mu, M., & Agyeman, M. O. (2022). A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction IEEE Access. The check captures multiple aspects of machine literacy associated cancer studies, including cancer bracket, cancer vaticination, identification of biomarker genes, microarray, and RNA-Seq data. The accuracy achieved was 0.82, 0.80,0.79 and 0.76 for MDNN, SVM, RF and LR independently. Deep learning approaches are overcoming the problems of traditional machine learning techniques in analyzing organic phenomenon data for cancer.

[3]. Mohit Agrawal," Cancer Prediction Using Machine Learning Algorithms", International Journal of Science and Research (IJSR), Volume 9 Issue 8, August 2020. Used algorithms are KNN (K-Nearest Neighbors), SVM (Support Vector Machine), LR (Logistic Regression), NB (Naïve Bayes) and estimate and compare that the bracket of delicacy, perfection, recall, f1- score. K- Nearest Neighbors (KNN) gives most Accurate Algorithm having classified the samples with 98% in the confirmation. KNN Algorithm give the best result with maximum accuracy.

[4]. Shaikh, F. J., & Rao, D. S. (2022). Prediction of cancer disease using machine learning approach. Accoutrements moment Proceedings, 50, 40- 47. ML & DL approaches employed in cancer progression modeling are reviewed. The issues have proven that DT was the well-conditioned performing classifier for prognosticating the complaint with an delicacy of 93.6 % on the holdout sample; ANN was standing

the runner-up position with an accuracy of 91.2%. Also, logistic regression has attained the delicacy of 89.2. SVM acquired high accuracy, next thereto ANN and DT attained more accuracy.

[5]. Raihan Rafique, S.M. Riazul Islam, Julhash U. Kazi "Machine Learning in the Cancer Prediction", 2021. They aim to classify tumor into malignant or benign tumor using different features from several cell images. They used K-nearest neighbor (KNN), Support vector machine (SVM) and Naive Bayes (NB) with 150 images database. They used Convolutional Neural Network (CNN) for image recognition and classification. Then they have implemented a neural network algorithm and achieved 87.64% accuracy.

3. METHODOLOGY

This is the classification model in which the goal is to predict the discrete value like {0,1} or (yes, no) or (spam, not spam). Here we are predicting whether the person is having cancer or not as like as (yes or no). There are three types of classifications:

1. Binary Classification: Classification task with two possible outcomes.
2. Multi-class Classification: Classification with more than two classes.
3. Multi-label Classification: Classification task where each sample is mapped to set of target labels (more than one class).

Methods used for classification are:

1. Logistic Regression
2. Naive Bayes
3. Support Vector Machine
4. Decision Tree
5. Random Forest
6. K-Nearest Neighbors

Logistic Regression:

Logistic regression is a machine learning algorithm for classification. In this algorithm, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function that is shown in below figure 1.

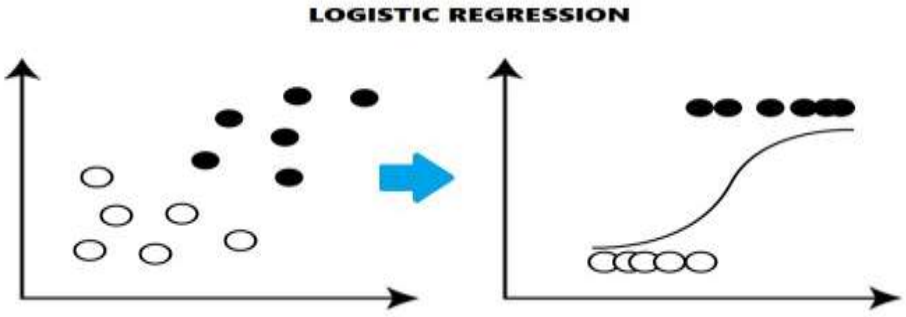

Figure 1. Logistic Regression

It is designed for this purpose (classification), and is most useful for understanding the influence of several independent variables on a single outgrowth variable. It Works only when the predicted variable is binary, assumes all predictors are independent of each other and assumes data is free of missing values.

Naive Bayes:

Naive Bayes algorithm based on Bayes' theorem with the assumption of independence between every pair of features. Naive Bayes classifiers work well in many real-world situations such as document classification and spam filtering shown in below figure 2.

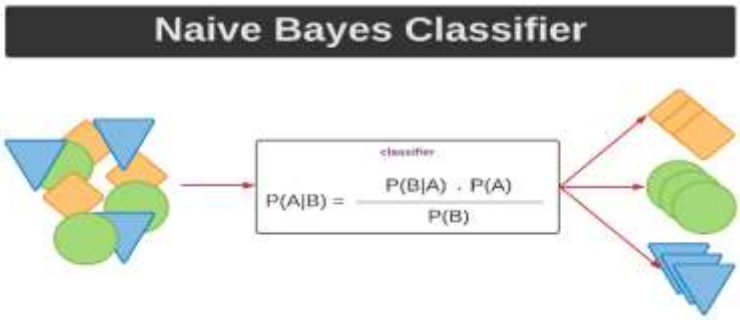

Figure 2. Naïve Bayes Classifier

This algorithm requires a small amount of training data to estimate the necessary parameters. Naive Bayes classifiers are extremely fast compared to more sophisticated styles. Naive Bayes is known to be a bad estimator.

Support Vector Machine:

Support vector machine is a representation of the training data as points in space separated into orders by a clear gap that's as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall shown in below figure 3.

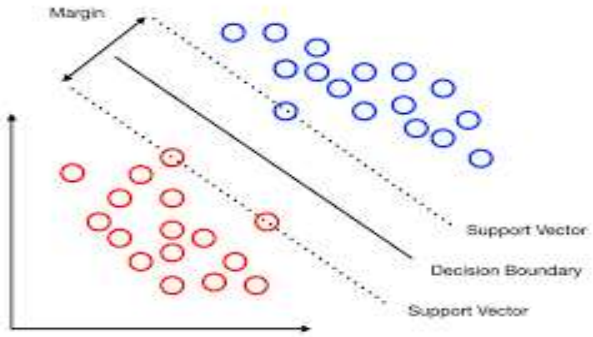

Figure 3. Super Vector Machine

Effective in high dimensional spaces and uses a subset of training points in the decision function so it is also memory efficient. The algorithm doesn't directly give probability estimates, these are calculated using an precious five-fold cross-validation.

Decision Tree:

Given a data of attributes together with its classes, a decision tree produces a sequence of rules that can be used to classify the data. Decision Tree is simple to understand and visualize, requires little data preparation, and can handle both numerical and categorical data.

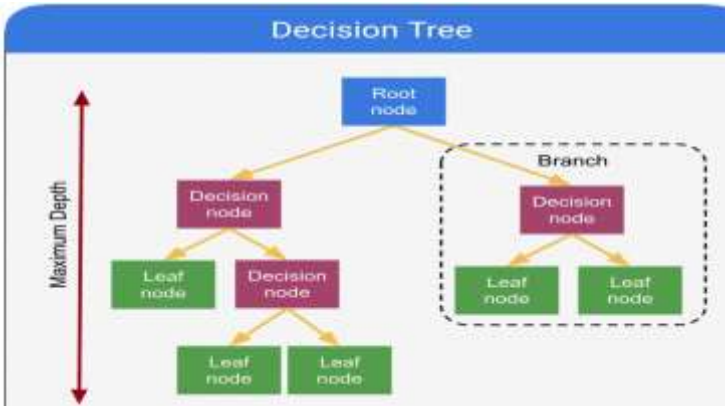

Figure 4. Decision Tree

Decision tree can create complex trees that do not generalize well, and decision trees can be unstable because small variations in the data might affect in a fully different tree being generated.

Random Forest:

Random Forest classifier is a meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses average to ameliorate the prophetic delicacy of the model and controls over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement and is shown in below figure 5.

Figure 5. Random Forest Classifier

Reduction in over-fitting and random forest classifier is more accurate than decision trees in most cases. Slow real time vaticination, delicate to apply, and complex algorithm.

K-Nearest Neighbors:

Neighbors based classification is a type of lazy learning as it does not essay to construct a general internal model, but simply stores cases of the training data. Classification is computed from a simple majority vote of the k nearest neighbors of each point that is shown in below figure 6.

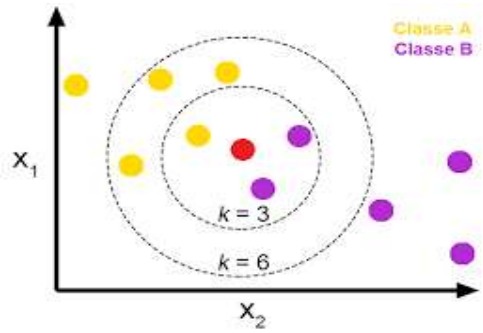

Figure 6. K Nearest Neighbor

This algorithm is simple to implement, robust to noisy training data, and effective if training data is large. Need to determine the value of K and the calculation cost is high as it needs to cipher the distance of each case to all the training samples.

4. RESULT AND DISCUSSION

- There are three types of results like Low, Medium and High.
- These outputs are better in order to know the level of the disease.

After the study of various algorithms, the below are the accuracy obtained for each method shown in table 1. Also represent the accuracy in visual format shown in figure 7.

S.No	Method	Accuracy
1.	Logistic Regression	95%
2.	Naïve Bayes	96%
3.	Support Vector Machine	95%
4.	Decision Tree	93%
5.	Random Forest	94%
6.	K-Nearest Neighbours	97%

Table 1. ML Algorithms with their Accuracies

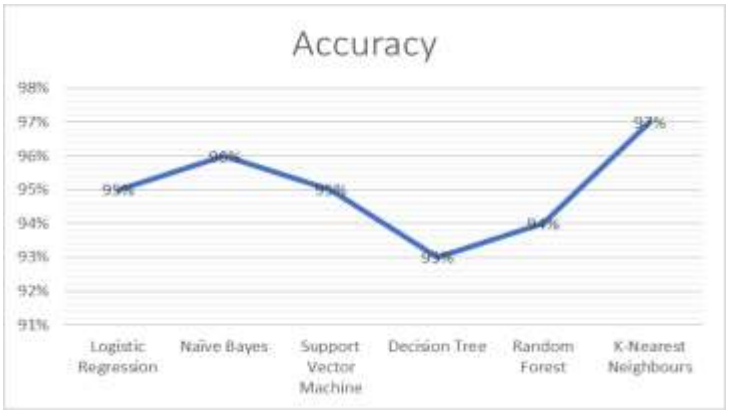

Figure 7. Graphical representation of their accuracies

5. CONCLUSION

In this study, compare the performances of different machine learning algorithms on cancer prediction. Specifically, identified a number of trends with respect to the types of machine learning methods being used, the types of training data being integrated, the kinds of endpoint prognostications being made. Among all K-Nearest Neighbor (KNN) gives the more accuracy in this cancer prediction. But in the KNN also, need to take care about the value of 'K', because based on the 'K' value accuracy differs more.

References

1. Alfayez, A. A., Kunz, H., & Lai, A. G. (2021). Predicting the risk of cancer in adults using supervised machine learning: A scoping review. BMJ open, 11(9), e047755.
2. Shaikh, F. J., & Rao, D. S. (2022). Prediction of cancer disease using machine learning approach. Materials Today: Proceedings, 50, 40-47.
3. Jasim, M., Machado, L., Al-Shamery, E. S., Ajit, S., Anthony, K., Mu, M., & Agyeman, M. O. (2022). A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction. IEEE Access
4. Raihan Rafique, S.M. Riazul Islam, Julhash U. Kazi "Machine Learning in the Cancer Prediction", 2021.
5. Mohit Agrawal, "Cancer Prediction Using Machine Learning Algorithms", International Journal of Science and Research (IJSR), Volume 9 Issue 8, August 2020.
6. Waseem, M. H., Nadeem, M. S. A., Abbas, A., Shaheen, A., Aziz, W., Anjum, A., ... & Shim, S. O. (2019). On the feature selection methods and reject option classifiers for robust cancer prediction. IEEE Access, 7, 141072-141082.
7. G. Sruthi, C. L. Ram, M. K. Sai, B. P. Singh, N. Majhotra and N. Sharma, "Cancer Prediction using Machine Learning," 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), 2022, pp. 217-21, doi: 10.1109/ICIPTM54933.2022.9754059
8. Sathe, P., Bombay, M., Mani, G. S., Kalathil, D., & Phadtare, A. (2008). Cancer Detection using Machine Learning.
9. Saba, T. (2020). Recent advancement in cancer discovery using machine learning: Methodical check of decades, comparisons, and challenges. Journal of Infection and Public Health, 13(9), 1274-1289.
10. Noor, M. M., & Narwal, V. (2017). Machine learning approaches in cancer detection and diagnosis: mini review. IJ Mutil Re App St, 1(1), 1-8.
11. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning operations in cancer prognostic and vaticination. Computational and structural biotechnology journal, 13, 8-17.