International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2019): 7.583

Cancer Prediction Using Machine Learning Algorithms

Mohit Agrawal

Student, Computer Science and Engineering, IMSEC Ghaziabad, UP, India, 201009

Abstract: ***The main objective of this project to build the model for predicting cancer using a support vector machine classifier algorithm and compare the accuracies on different kernels and apply the various parameters on the efficient one kernel. The cancer dataset will be imported from the scikit-learn library. Cancer has been characterized as a heterogeneous disease consisting of many various subtypes. The soon diagnosis and prognosis of a cancer type have becomes a need in cancer research, as it can facilitate the subsequent clinical management of patients. these technique include Artificial Neural Network, Bayesian Networks, Support Vector Machines and Decision Trees have been widely apply in cancer research for the development of predictive prototype, results in effective and accurate decision making. Even though it is obvious that the use of ML methods can improve our understanding of cancer progression, an appropriate level of validations is needed in order for these methods to be Consider in the everyday clinical practice. In this work, we present a review of Machine learning approaches employed in the modeling of cancer progression. The predictive models discussed here are based on various supervised Machine learning techniques as well as on different input features and data samples. The using Algorithm KNN (K Nearest Neighbors), SVM(Support Vector Machine), LR(Logistic Regression), NB(Naïve Bayes) and also evaluate and compare that the classification of accuracy, precision, recall, f1-score.the UCI machine learning dataset will be partitioned as 75% for training phase and 25% for the testing phase and then apply all algorithm is best performance of All parameter.***

Keywords: Cancer data set, Support Vector Machine, Kernels, Labels, Target Values

1. Introduction

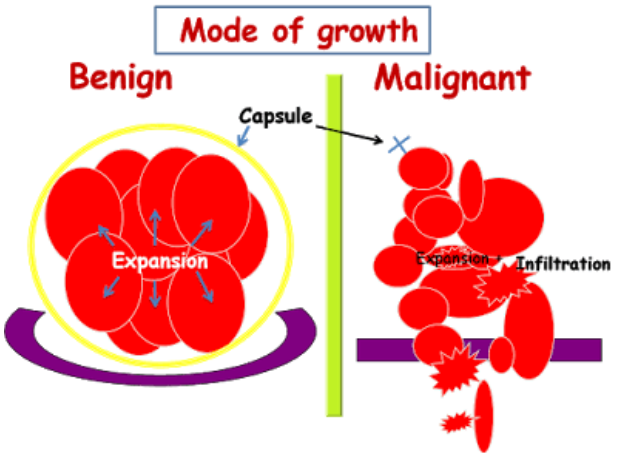
Cancer patient is the most common from of cancer along with lung and bronchus cancer, prostate cancer, colon cancer among others cancers. Cancer is prevalent cause of deaths and only type of cancer that is widespread among women in UAE and Worldwide. Cancer causes are involve family history, obesity hormones, radiation therapy even factors. Given the importance of personalized medicine and the growing trend on the appliance of ML techniques, we here present a review of studies that make use of these methods regarding the cancer predictions. In these studies prognostic and predictive feature are considered which may be independent of some treatment or are integrated in order to guides of therapy for cancer patients, respectively. Estimated numbers of new cancer cases 1762450 and Estimated numbers of deaths 606880 in 2019 (United States) [4]. It is clear that the application of ML methods could improve the accuracy of cancer sensitivity, recurrence and multiple predictions. Based On The accuracy of cancer prediction outcome has significantly improved by 15%–20% the last years, with the application of ML techniques. Woman cancer deaths are 15 %. While cancer rates are higher among women in more developed regions rate are increasing in every region globally [1].

However, the application of feature selection techniques may result in specific fluctuations concerning the creation of predictive feature lists. In the present work only studies that employed ML techniques for modeling cancer diagnosis and prognosis are presented.

The various types of classification algorithm namely KNN(K Nearest Neighbors), SVM (Support Vector Machine), LR(Logistic Regression), NB (Naïve Bayes) as also evaluate and compare the performance of the various classifiers in terms of Accuracy, precision, recall, f1-score.this paper provide an overview of the state of art ML technique for cancer detection.

2. Machine Learning Algorithms

Machine learning algorithm where the cancer data set is loaded, features have to be extracted and classification model can be trained and used Prediction of Malignant and Benign. A benign cancer that does not invade its surrounding tissue or spread around body and A Malignant cancer that may invade its surrounding tissue or spread around the body.



2.1 K nearest Neighbors [KNN]

K Nearest Neighbors algorithms work According to the similar features of the neighbor's data point. In this algorithm we used features similarity and predicted the value of new data point and assign the value based on how closely at matched we points in the training set it used to identify weather is patient having cancer or not. For implementation is the best example this Algorithm.

2.2 Support Vector Machine [SVM]

Support Vector Machine is supervised learning classification technique which is widely used in the field of cancer and prognoses and diagnosis. It this technique select critical sample from all class, these class are known as support vector and separating the classes by linear function. these function divides class broadly as possible using these support vector according to this technique a mapping between input vector to a dimensionality space is made using SVM that purpose to find the most suitable hyperplane that also divides the data set into class. The purpose this classifier to maximize the distance between the decision hyperplane and nearest data point, that distance is called the marginal distance, the finding the best suited hyperplane.

2.3 Naïve Bayes [NB]

Navie Bayes is based on Bayes theorem with independence among predictors and it is a classification technique it means, this classifier assumes that the presence of features in a particular class is not related to the presence of any other features if these features are depend on each other then these properties contribute to the probability of the class independently and that is the men reason for calling this classification technique is "NAÏVE". Navie Bayes is "Naïve" because it Assume that feature of a measurement are not depend of each other. Naïve Bayes is naïve because it almost never true. It is easy to bold and particularly useful for large data set. Along with simplicity, Naive Bayes is known as sophisticated classification method.

2.4 Logistic Regression [LR]

Logistic Regression belongs to the group of liner classifier and it is a fundamental classification technique it is similar to polynomial and liner regression it is fast and relatively uncomplicated it provide accurate and passed interpreted the results and it is also called binary classification we can use this regression to solve the multiclass problem it is different from liner regression because liner regression concerned with the predictive values. Logistic regression models falls to response into a special category. A logistic regression model can be able to solve the situation where the output can take only two value (0and 1) with help of sigmoid function.

3. Methodology

The cancer dataset from UCI and used jupyter Notebook as the platform for the coding purpose. Our classification involves use of classification techniques like KNN (K Nearest Neighbors), SVM (Support Vector Machine), LR (Logistic Regression), NB (Naïve Bayes) Algorithms.

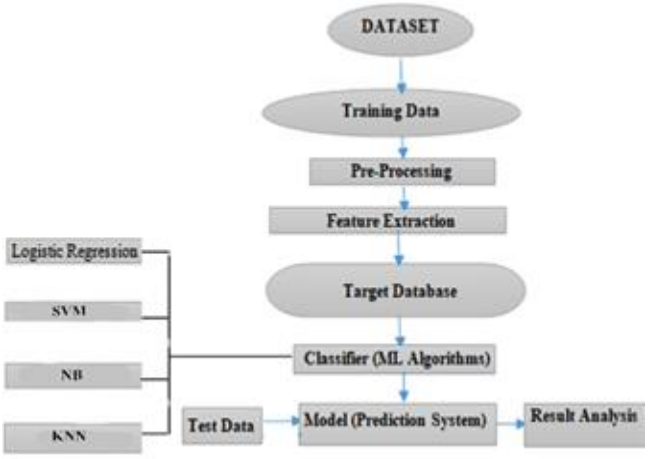


Figure: Block Diagram of Cancer Prediction

3.1 Features

In this cancer dataset have multiple value they used to classify the normal person and Tumor patient in this data set have different features these are defined as below:

3.1.1. Mean Radius:-The mean radius is signify of distance from center to point on signify Parameters that mean Radius using The Tumors Value Evaluation as accurate value predicted.

3.1.2. Mean Smoothness:- The mean smoothness is import the Machine learning library numpy as np. The smoothness is modification in radius lengths.

3.1.3. Mean Texture:- The Mean Texture is the standard form of the scale Gray value That the Scale Gray value is predicted dataset is used the Tumors value. Any others type of the Appearance use then the mean parameter, mean area, mean compactness, mean concavity, area error, worst compactness, worst fractal dimension etc.

3.2 Labels

There are two types of tumors predicted Malignant and Begin Malignant Tumors Carcinogenic. Then the cells grow out of control. If the Cells Continue To Develop and lay out, the infection can become serious and Begin Tumors is Not Carcinogenic. Then don't use the Invade nearby tissue or layout any other parts of the body.

3.3 Output

The value 0 and 1 defines the prediction as 0 as malignant tumors and 1 is Begin Tumors then value Using the methods and Tables. The dataset is collected of 152528 data with 16 key attribute. A class varying is also considered mentioning, namely persistence to patients that had not Live and those that had pull through.

3.3.1 Dimensionally Reduction

This is process in which the number of variables is reduced to a dataset of variables by removing these less significant in predicting the outcome. Dimensionally reduction is correctly data defined.

3.3.2 Subset Selection

Feature selection is finding the subset of All algorithms KNN (K Nearest Neighbors), SVM (Support Vector Machine), LR (Logistic Regression), NB (Naïve Bayes) features by different approaches based on the predicted error and Accuracy.

3.3.3 Point Factor

Point Factor is transformation of high dimensional space data to a lower dimensional space (some attributes).the dataset used in the research paper is multidimensional dataset with 40 attribute, which are related to Accuracy, precision, recall, f1-score, Malignant and Begin selection of features by application of feature selection is a complex tasks.

3.3.4 Portrait Choice

The most exciting phase in building any machine learning model is selection of algorithm. We can use more than one kind of data mining techniques to large datasets. Supervised learning is the method in which the machine is trained on the data which the input and output are well labeled. The model can learn on the training data and can process the future data to predict outcome. In our dataset we have the Dependent variable or outcome variable. Y having only two set of values, either M (Malignant) or B (Begin). So we apply classification algorithms under supervised learning on it. We have chosen four different types of classification algorithms in Machine Learning.KNN (K Nearest Neighbors), SVM (Support Vector Machine), LR (Logistic Regression), NB (Naïve Bayes).

4. Database

The Research paper is based on a 2 data sets that is published is available from the UCI machine learning. The data set consist of several hundred human cell sample records, using data set SVM classification and other data used in the four algorithm comparisons in best algorithm predicted the data. The following data having attributes. Id Number, Clump Thickness, Uniformity of cell size, Mitoses, bland Chromatin, Normal Nucleoli, Marginal Adhesion. The ID NUMBER Attribute is the Patient identifiers. The characteristics of each cell from each patient are contained in other Attribute Clump Thickness to Mitoses the value range 1 to 10 with being 1 close and closest value find begin.

The data set [4] used in SVM classification Algorithm that is Attribute Define mean radius, smoothness concavity and Malignant begin as the predicted dataset and find best accuracy, mean square error, R2 square error. The dataset is used in world health organization[1].

5. Literature Review

Benbrahim [12] uses classification experiment that shows the maxima best accuracy 96.49% that was Achieved by the neural network algorithm.

Deepika et. al. [4] uses two classification algorithms Naive Bayes and Multi Layer Perceptron and after analyzing the performance of both algorithm found that Naïve Bayes gives the more accurate results.

Mariam et. al. [8] uses two different classifiers namely Naive Bayes and K Nearest Neighbors for breast cancer classification on comparing accuracy using cross validation and KNN achieved that 97.51% accuracy with lowest error rate then Naïve Bayes Classifier 96.19% accuracy.

Aruna [9] used Naive Bayes, Support Vector Machine, and K Nearest Neighbors to categorize a Wisconsin cancer dataset and obtain the best results by using K Nearest Neighbors with an accuracy score of 96.99%.

Ajani [10] "Data mining techniques to resolve cancer survival and Prediction with 95.98% an Accuracy".

6. Proposed Work

The classifier accuracy is a measure of how the classifier could predict cases into the category.it is correct prediction dived by total number of instances.then it is not optimal method to compare different classifier but may give an overview of the class, accuracy is calculated using the equations.

7. Observations

Tables 1: Accuracy without Applying Any Parameters

S/n.	Kernel	Accuracy	Mean square error
1.	Linear	0.965034958041965	0.18693662259964105
2.	Gaussian	0.8811188811188811	0.3447914135838056
3.	Sigmoidal	0.5244755244755245	0.6895828271676112
4.	polynomial	0.8811188811188811	0.3447914135838056

Most efficient result is in kernel="LINEAR"

In table 1 Accuracy results we can observe that linear kernel has the very best Accuracy, Gaussian and Polynomial kernel Accuracy is same and the Sigmoidal kernel Accuracy is worst then all its better to the Linear kernel.

Table 2: Accuracy with Different value of C

C	Accuracy	Mean square error	R2 square error
1	0.965034965034965	0.18698939800169145	0.8489010989010989
0.1	0.958041958041958	0.20483662259967567	0.8186813186813187
0.01	0.958041958041958	0.20483662259967567	0.8186813186813187
0.001	0.958041958041958	0.20483662259967567	0.8186813186813187
0.0001	0.958041958041958	0.20483662259967567	0.8186813186813187

C=1 is the best Accuracy

When the Apply different value of C then we observe that C=1 Linear kernel give the best Accuracy we chooses the 1 value for C=0.965034965034965.

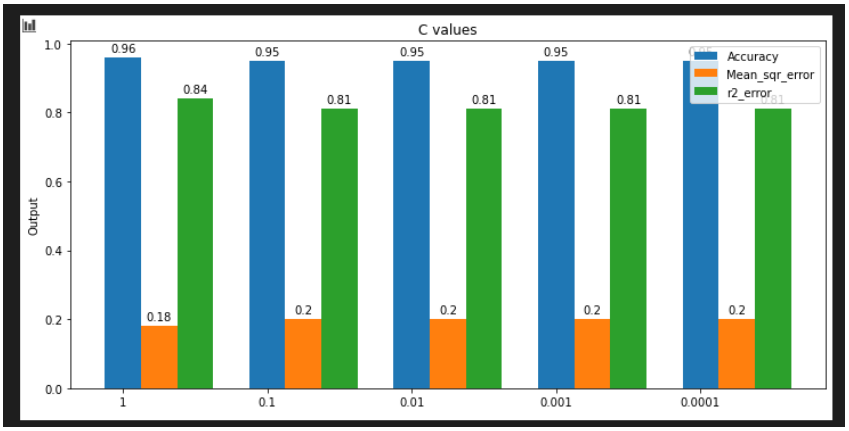


Figure 1

Graph of Accuracy, Mean square error, R2 square error

Table 3: Accuracy with Different value of Max iter

Max iter	Accuracy	Mean square error	R2 square error
0.1	0.6363636363636364	0.6030226891555273	-0.5714285714285716
0	0.6363636363636364	0.6030226891555273	-0.5714285714285716
-1	0.965034965034965	0.18698939800169145	0.8489010989010989
1	0.7832167832167832	0.46559984620188266	0.6318681318681307

Max_iter=-1 is the best Accuracy

When the Apply different value of Max iter then we observe that Max iter=-1 Linear kernel give the best Accuracy we chooses the 1 value for Max iter=0.965034965034965.

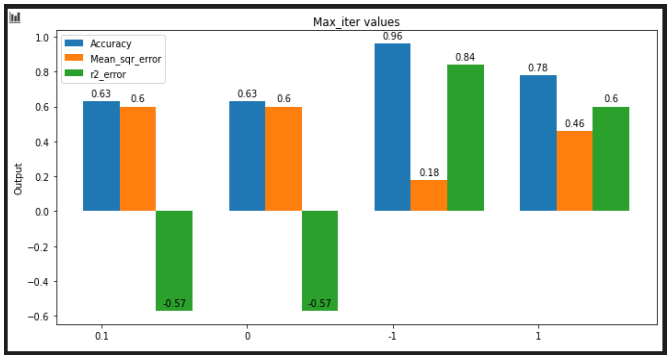


Figure 2

Figure 2 Graph of Accuracy, Mean square error, R2 square error

Table 4: Accuracy with Different value of Tol

Tol	Accuracy	Mean square error	R2 square error
0.001	0.94405594	0.23652495	0.76363636
0.1	0.93706293	0.25087260	0.73409090
1	0.92307692	0.27735009	0.67499999
5	0.61538461	0.62017367	-0.62500000
10	0.61538461	0.62017367	-0.62500000

Tol=0.001 is the best Accuracy

When the Apply different value of Tol then we observe that Tol=0.001 Linear kernel give the best Accuracy we chooses the 1 value for Tol=0.94405594.

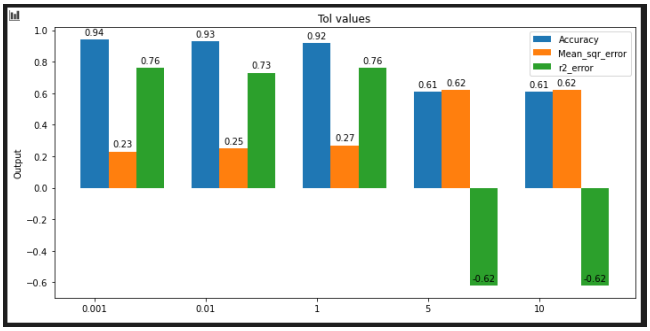


Figure 3

Graph of Accuracy, Mean square error, R2 square error

Degree, Gamma and Random state do not affect the Accuracy. A confusion matrix is a table is used to describe the performance of a classification model on a set of test data for true and false value known.

Physical Class (PC)	Negative Prediction (NP)	Positive Prediction (PP)
Positive Class(PC)	FN	TP
Negative Class(NC)	TN	FP

Table defined two classes that positive class and negative class. True positive and True negative are observed that are predicted underline the table and False positive and False negative observed that are predicted the table. All short forms defined as.
TP=True Positive
FN=False Negative
FP=False Positive
TN=True Negative

Accuracy = (True Negative + True Positive) / (True Negative+ True Positive +False Negative + False Positive)
All Algorithms value for Accuracy.

Algorithms	Accuracy
KNN	0.9795
LR	0.9567
NB	0.9623
SVM	0.9519

7.1 Recall

Recall, also known as sensibility, the observation table that is defined as positive prediction. Study, it is more important to correctly identify a malignant tumor than it is to incorrectly identify a benign one.
Recall = True Positive / (True Positive +False Negative)

Recall values for all four ML algorithms.

Algorithms	Malignant	Begin	Average
KNN	0.9423	1.0023	0.9793
LR	0.8825	0.9824	0.9324
NB	0.9299	0.9992	0.9645
SVM	0.9128	0.9935	0.9531

7.2 Precision

Precision, also commonly known as confidence, as the rate of both true positives and true negatives that have been identified as true positives.
Precision =True Positive / (True Positive + False Positive)
Precision values for all four Ml Algorithms

Algorithms	Malignant	Begin	Average
KNN	1.0021	0.9625	0.9823
LR	0.9625	0.9324	0.9474
NB	1.0011	0.9322	0.9666
SVM	0.9724	0.9566	0.9692

7.3 F1-Score

F1 Score is the weighted average of Precision and Recall. Therefore, take both FP and FN.
*FN (FALSE NEGATIVE)
*FP (FALSE POSITIVE)
F1-Score = 2× (Recall × Precision)
(Recall + Precision)
F1-Score values for all four Ml Algorithms

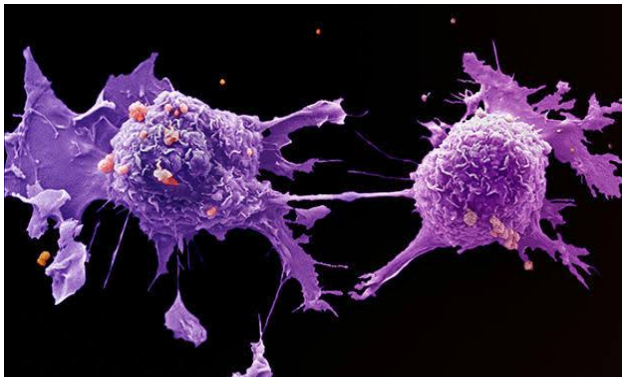
Algorithms	Malignant	Begin	Average
KNN	0.9799	0.9899	0.9849
LR	0.9611	0.9625	0.9618
NB	0.9866	0.9877	0.9871
SVM	0.9758	0.9785	0.9771

8. Results

All Algorithms used and the Accuracy parameters defined

Algorithms	Accuracy	Recall	Precision	F1-Score
KNN	0.9795	0.9793	0.9823	0.9849
LR	0.9567	0.9324	0.9474	0.9618
NB	0.9623	0.9645	0.9666	0.9871
SVM	0.9519	0.9531	0.9692	0.9771

K Nearest Neighbors (KNN) gives most Accurate Algorithm.KNN is Best accuracy as cancer prediction.



the fig in the results shows the prediction of cancer. K Nearest Neighbors (KNN) gives most Accurate Algorithm having classified the samples with 0.979&98% accuracy in the

validation. Naive Bayes (NB), Support Vector Machine (SVM) and Logistic Regression (LR) come in classification accuracy.KNN is Best accuracy as cancer prediction.

9. Conclusion

We discussed the concepts of ML while we outlined their application in Cancer prediction / prognosis. Most of the studies that have been proposed the last years and focus on the development of predictive models using supervised ML methods and classification algorithms aiming to predict valid disease outcomes. We observe that the KNN Algorithm give the best result with maximum accuracy. In feature I will modified this paper According to the new techniques. We chooses the best cancer predictive data set and used all important components or features used in this research paper.KNN algorithm is best because in this algorithm predicted the value on the basis of nearest point value.

10. Acknowledgement

I have finished this work under the guidance of Dr. Pankaj Agarwal (Professor & head) And Ms Sapna Yadav (Assistant Professor), Dept of CSE at IMSEC, Ghaziabad, Uttar Pradesh. I am doing an online Summer Internship on Machine Learning where I have learning various ML Algorithm from both of my mentors as a course instructor. This paper has been assigned as a project assignment for us. I would like to express my special thanks both of my mentors for inspiring us to complete work and write a paper. I would not lead way in writing the paper. I am extremely thankful for their valuable guidance and support on completion of this paper. I extend my gratitude to "IMS Engineering College, Ghaziabad, Uttar Pradesh" for giving me this opportunity. I also acknowledge with a deep sense of reverence, my gratitude towards my friends, parents and member of my family, who always supported me morally, mentally as well as economically.

References

[1] WHO- https://www.who.in/cancer/prevention.
[2] J.A. Cruz, D.S. Applications of machine learning in cancer prediction and prognosis Cancer Information (2006).
[3] M.J, N. Miller, K.M H.M. Heneghan as biomarkers and therapeutic targets in cancer Curr Opin Pharmacol, 10 (2010).
[4] E. Feuer, M. Reichman, L. Huang, A. Mariotto, J. J. Dignam, And L. Ries, , "Estimating cancer statistic and other-cause mortality in clinical trial and population-based cancer registry cohorts", *Wiley InterScience[Online]*, vol. 115, no. 22, August 2009.
[5] Mariam Amrane, Saliha Oukid, Ikram Gagaoua and Tolga Ensari, "Breast cancer classification using machine learning" 2018 Electric Electronics, Computer Science, Biomedical Engineering' Meeting (EBBT)
[6] Partridge AH And Azim HA Jr,. Biology of cancer in young women Cancer Res.2014;16(4):427.
[7] Levy-Lahad E And King MC, Levy-Lahad E, Lahad A Population-based screening for BRCA1 and BRCA2 2014 Lasker Award. Jama. 2014;312(11):1091–2.
[8] Tolga Ensari, Saliha, Gagaoua Oukid, Ikram and Mariam Amrane, "Breast cancer classification using ML Algorithm"2018 Computer Science, Biomedical Engineerings' Meeting (EBBT).
[9] Nandakishore, Aruna S, Rajagopalan S, , L, 'Knowledge based analysis of various statistical tools in detecting cancer"2011 Computer Science Information Technology.
[10] Ajani "Data mining techniques To resolve cancer survival and Prediction" Int. J.Computer Science 2015.
[11] S. Koscielny Why most gene expression signatures of tumors have not been useful in the clinic Sci Transl Med, 2 (2010.
[12] Benbrahim uses classification experiment that show the maxima best accuracy computer Science2016.
[13] H. Wolberg R. WIlliam (physician) http://archive.ics.uci.edu/ml/datasets.php University of Wisconsin Hospitals Madison, Wisconsin, USA.

Author Profile

Mohit Agrawal is B.Tech 3rd Year Student in the Dept. of C.S.E & Eng At IMS Eng Coll. Ghaziabad, UP, India 201009.