Scikit-Learn Implements Machine Learning Algorithms for Lung Cancer Analysis

G. Mahendran

M.Sc (CS)
Department of Computer Science
Sri Kaliswari College (Autonomous), Sivakasi, Tamilnadu

***Abstract*: Lung cancer is one of the foremost common and high forms of cancer. The 2 main varieties are small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). These varieties are diagnosed supported however the cells look underneath a microscope. Over 80th of all lung cancers belong to the non-small cell kind. It's the second commonest cancer in men and also the fifth most typical cancer in each men and women along. ML algorithms are people who will learn from information and improve from expertise, while not human intervention. Victimization Python and its open source libraries make it attainable for nearly anyone to use this approach. In this paper using some Supervised ML classification algorithms to create models to check the results and verify the lot of correct algorithmic program.**

***Index Terms*: Lung cancer, Jupyter notebook, Scikit-Learn, Python, Supervised ML Algorithms**

I. INTRODUCTION

Machine learning, a kind of AI that 'learns' because it identifies new patterns in information, allows information scientists to effectively pinpoint revenue opportunities and build methods to enhance client experiences exploitation information hidden in vast information sets. Select the proper algorithmic rule could be a key a part of any machine learning project in varied business applications is important. The two major kinds of carcinoma are non–small-cell lung cancer (about eighty fifth of all carcinomas) and small-cell lung cancer (about 15%).

Despite advances in early Detection and normal treatment, non–small-cell lung cancer is usually diagnosed at a complicated stage and contains a Poor prognosis. Any patient with carcinoma might want to think about participating in one in every of the various clinical trials that are underneath manner at any given time with the goal to seek out more practical treatments.

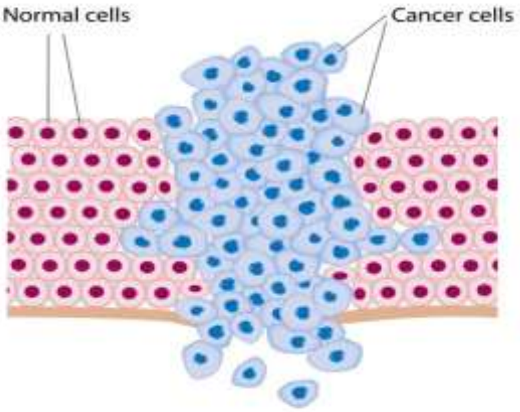


Figure.1. Cancer cells differentiate from Normal cells

II. RELATED WORK

In 2012 [1] We discussed and highlight the role that data mining approaches, particularly machine learning methods, can play to improve our understanding of complex systems such as tumor response to radiotherapy. Review of the current knowledge of genetic and signaling pathways in modulating tumor response to radiotherapy in non-small cell lung cancer as a case study of data mining application in the challenging cancer treatment.

In 2015 [2] this paper, involves a novel region based ACM with pre-processing technique. It is proposed for the segmentation phase of the lung nodule detection. From segmented images features are extracted and for classification SVM classifier is used. We have used the dataset of 12 images taken from LIDC.
In 2015[3] this paper presents, an overview of various cancer classification methods and evaluate these methods based on their classification accuracy, computational time and ability to reveal gene information.

In 2018[4] this system developing data analyses the past history of an individual's or a patient's behavior and also based on the result of simple diagnosing method (blood test) using Machine learning algorithm. It involves the process of identifying efficient algorithm to develop a system. The developed system might improve the survival rate of the patients.

In 2018[5] this paper, performed by using very basic steps to create a strong machine learning program which is able to identify the tumor as malignant or benign. In this study using KNN, Logistic Regression algorithms to compare results.

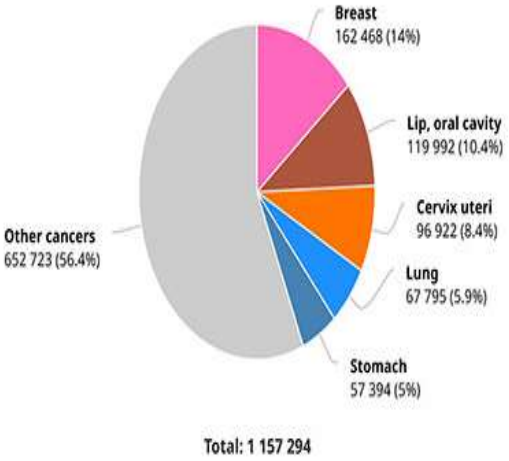


Figure.2. Number of new cases in 2018, Both sexes – All ages.
[source: globocan-2018]

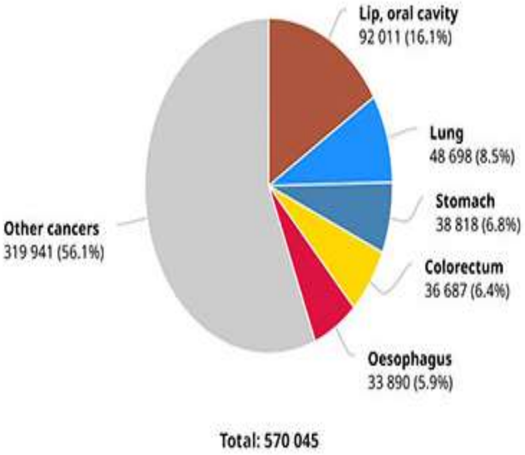


Figure.3. Number of new cases in 2018, males- All ages
[source: globocan-2018]

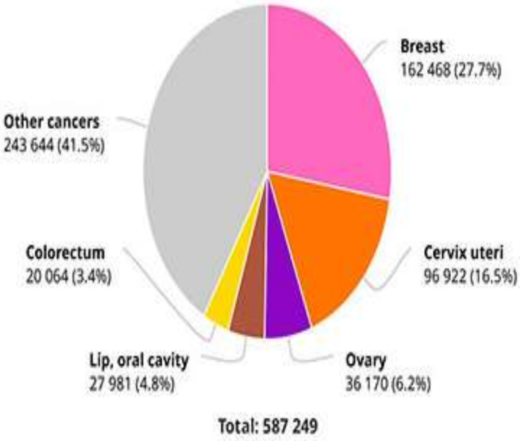


Figure.4. Number of new cases in 2018, females- All ages
[source: globocan-2018]

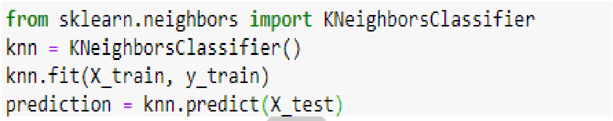
III. Methodology

The majority of sensible machine learning uses supervised learning. It is takes an input variable (x) and an output variable (Y) and used an algorithmic rule to find out the mapping perform from the input to the output. The goal is to approximate the mapping perform thus well that after have new knowledge or input (x) that just will predict the output variables (Y) for that data.

In this paper implements K-Nearest Neighbors, Naïve Bayes, Support Vector Machine, Random Forest, Decision Tree and Logistic Regression.

A. K-*Nearest* Neighbor Algorithm

Neighbors based classification is a type of lazy learning as it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the k nearest neighbors of each point.



```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier()
knn.fit(X_train, y_train)
prediction = knn.predict(X_test)
```

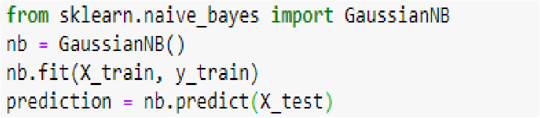
Figure.5. Predict knn classifier for lung cancer data

Advantages: This algorithm is simple to implement, robust to noisy training data and effective if training data is large.

Disadvantages: Need to determine the value of K and the computation cost is high as it needs to computer the distance of each instance to all the training samples.

B. Naïve Bayes Algorithm

Naive Bayes is 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.



```
from sklearn.naive_bayes import GaussianNB
nb = GaussianNB()
nb.fit(X_train, y_train)
prediction = nb.predict(X_test)
```

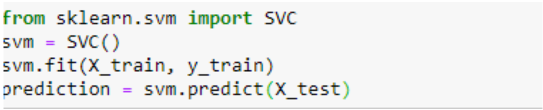
Figure.6. Predict Naïve Bayes for lung cancer data

Advantages: This algorithm requires a small amount of training data to estimate the necessary parameters. Naive Bayes classifiers are extremely fast compared to more sophisticated methods.

Disadvantages: Naive Bayes is is known to be a bad estimator.

C. Support Vector Machine

Support vector machine is a representation of the training data as points in space separated into categories by a clear gap that is 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.



```
from sklearn.svm import SVC
svm = SVC()
svm.fit(X_train, y_train)
prediction = svm.predict(X_test)
```

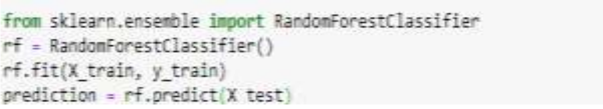
Figure.7. Predict SVM classifier for lung cancer data

Advantages: Effective in high dimensional spaces and uses a subset of training points in the decision function so it is also memory efficient.

Disadvantages: The algorithm does not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

D. 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 improve the predictive accuracy 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.



```
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier()
rf.fit(X_train, y_train)
prediction = rf.predict(X_test)
```

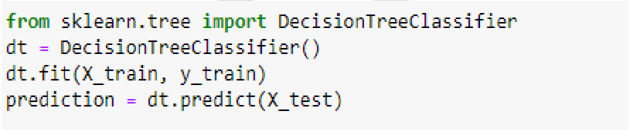
Figure.8. Predict Random Forest for lung cancer data

Advantages: Reduction in over-fitting and random forest classifier is more accurate than decision trees in most cases.

Disadvantages: Slow real time prediction, difficult to implement, and complex algorithm.

E. 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.



```
from sklearn.tree import DecisionTreeClassifier
dt = DecisionTreeClassifier()
dt.fit(X_train, y_train)
prediction = dt.predict(X_test)
```

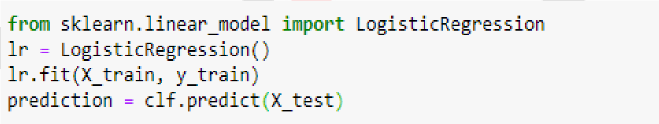
Figure.9. Predict Decision Tree for lung cancer data

Advantages: Decision Tree is simple to understand and visualize, requires little data preparation, and can handle both numerical and categorical data.

Disadvantages: Decision tree can create complex trees that do not generalise well, and decision trees can be unstable because small variations in the data might result in a completely different tree being generated.

F. 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 modeled using a logistic function.



```
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression()
lr.fit(X_train, y_train)
prediction = clf.predict(X_test)
```

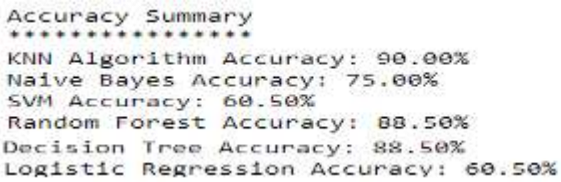
Figure.10. Predit LR Classifier for lung cancer data

Advantages: Logistic regression is designed for this purpose (classification), and is most useful for understanding the influence of several independent variables on a single outcome variable.

Disadvantages: Works only when the predicted variable is binary, assumes all predictors are independent of each other, and assumes data is free of missing values.

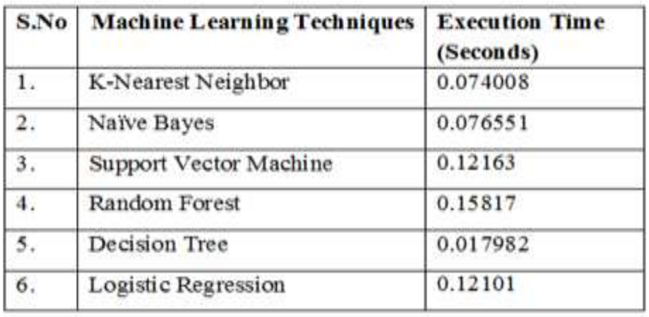
IV. ACCURACY SUMMARY & TIME TAKEN

In this module accuracy measures how often the classifier makes the correct prediction. Accuracy of classifier refers to the ability of classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.



```
Accuracy Summary
****************
KNN Algorithm Accuracy: 90.00%
Naive Bayes Accuracy: 75.00%
SVM Accuracy: 60.50%
Random Forest Accuracy: 88.50%
Decision Tree Accuracy: 88.50%
Logistic Regression Accuracy: 60.50%
```

Figure.11. Accuracy Summary



S.No	Machine Learning Techniques	Execution Time (Seconds)
1.	K-Nearest Neighbor	0.074008
2.	Naïve Bayes	0.076551
3.	Support Vector Machine	0.12163
4.	Random Forest	0.15817
5.	Decision Tree	0.017982
6.	Logistic Regression	0.12101

Figure.12. Execution Time Taken (in Seconds)

V. RESULTS AND DISCUSSION

In this research work will analyses the carcinoma malady predictions victimization completely different classification algorithms like K-Nearest Neighbor, Naïve Bayes, Support Vector Machine, Random Forest, Decision Tree and logistic Regression. Moreover, this information contains differing biological entities, genes or proteins, which implies that while knowledge discovery may be a massive part of bioinformatics, information management is additionally a primary concern.

The extensively immense science of information mining inside the domain of bioinformatics could be a proper ideal work because of the ever growing and developing scope of biological data. As this area of research is thus in depth it's apparent that attributes of biological databases propose an outsized quantity of challenges. As a result it's necessary for the future directions of research to adapt for the combination of latest bioinformatics databases so as to produce a lot of strategies of effective research.

VI. CONCLUSION AND FUTURE SCOPE

In this study performed by classification algorithms has been mostly studied by researchers within the field of machine learning, statistics, and databases. Varied classification techniques are projected within the past, just like the linear discrimination analysis, the Bayesian network, the decision tree strategies, etc. The K-Nearest Neighbor performed high accuracy of other ML algorithms accuracy and execution time taken for 0.074008 Seconds calculated by lung cancer dataset.

REFERENCES

[1] Issam E Naqa, "Machine learning methods for predicting tumor response in lung cancer", Wiley Interdisciplinary Reviews: Data Mining and Knowledge DiscoveryVolume 2 Issue 2, March 2012.

[2] Khan, Sajid Ali, Nazir, "Proficient lungs nodule detection and classification using machine learning techniques", Journal of Intelligent & Fuzzy Systems, vol. 28, no. 2, pp. 905-917, 2015.

[3] Amit Bhola and Arvind Kumar Tiwari, "Machine Learning Based Approaches for Cancer Classification Using Gene Expression Data", Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.3/4 and December 2015.

[4] Ram Nikhilesh.M, Indresh.M, "Supervised Deep Learning Approach To Identify the Lung Cancer", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 7, Special Issue 2, March 2018.

[5] Arushi Agarwal, Ankur Saxena, "Malignant Tumor Detection Using Machine Learning through Scikit-learn", International Journal of Pure and Applied Mathematics, Volume 119 No. 15, 2018.

[6] K. V. Bawane1, A. V. Shinde, "Diagnosis Support System for Lung Cancer Detection Using Artificial Intelligence", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 6, Issue 1, January 2018.

[7] Roseline Jecintha I, Poonguzhali.V, "Study on Data Mining Techniques for Cancer Prediction System", International Journal of Data Mining Techniques and Applications Volume: 07, Issue: 01, June 2018.

[8] Syed Abbas Ali, Fatima Waheed, Wajahat Rehman, Sallar Khan, "Comparative Analysis of Learning Algorithms for Lung Cancer Identification", Indian Journal of Science and Technology, Vol 11(27), July 2018.