ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

# Predicting Cancer Using Linear Regression and Naïve Bayes

Pooja Bhati [1], Dr. Dinesh Kumar [2]

M.Tech (Pursuing), Dept. of CSE, SRCEM at Palwal, Haryana, India [1]

Professor & HOD, Dept. of CSE, Dept. of CSE, SRCEM at Palwal, Haryana, India[2]

**ABSTRACT:** In this scheme, we proposed linear regression and Naïve Bayes Classification to covenant with two bioinformatics troubles, i.e., cancer identification and forecasting the phase where the patient is standing on and genetic material appearance statistics and secondary structure prediction that the patient is in which stage of cancer . For the problem of cancer diagnosis, the Naïve Bayes the probabilistic approach is used under the scheme to achieve highly accurate results with fewer classified and regression attributed will be compared to formerly projected approach whereas the dataset will be regressed with Linear Regression thus resulting the confusion matrix and based on probable segregated attributes to extract the desired models which will be drawn with accuracy and effectively for forecasting the venial or cancer.

**KEYWORDS**: Linear Regression, Naïve Bayes, Large B-Cell Lymphoma, and Small Round Blue Cell Tumors.

## I. INTRODUCTION

Cancer has been characterized as a amalgamate ache consisting of abounding altered subtypes. The aboriginal analysis and cast of a blight blazon accept become a call in blight research, as it can facilitate the consecutive analytic administration of patients. The accent of classifying blight patients into top or low accident groups has led abounding analysis teams, from the biomedical and the bioinformatics field, to abstraction the appliance of apparatus acquirements (Machine Learning) methods. Therefore, these techniques accept been activated as an aim to archetypal the progression and analysis of annihilative conditions. In addition, the adeptness of Machine Learning accoutrement to ascertain key appearance from circuitous datasets reveals their importance. A array of these techniques, including Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Accommodation Trees (DTs) accept been broadly activated in blight analysis for the development of predictive models, dependable and authentic accommodation making. Even admitting it is axiomatic that the use Machine Learning methods can advance our compassionate of blight progression, an adapted akin of validation is bare in adjustment for these methods to be advised in the accustomed analytic practice. In this work, we present a analysis of Machine Learning approaches using Linear Regression and Support Vector Machine to blight progression and forecasting for cancer prediction. However under the scheme predictive models discussed actuality is based on assorted supervised Machine Learning techniques as able-bodied as on altered ascribe appearance and abstracts samples. Given the growing trend on the appliance of Machine Learning methods in blight research and apply these techniques as an aim to accommodate outcomes in accurate manner. Machine Learning is the way against breaking down advice from alternating credibility of appearance and burden it into admired abstracts - abstracts that can be activated to aggrandize balance in corresponding context, espionage attributes and covariance and contra-variance, or both. Machine Learning programming is one of assorted analytic apparatuses for breaking down information. It enables practitioners to breach down advice from a advanced ambit of abstracts or edges, adjustment it, and abbreviate the access distinguished. In fact, advice mining is the way against advertent access or examples a part of abounding fields in all-embracing datasets and databases.

The broadly utilized and understood Machine Learning functionalities are characterization and discrimination, content based examination, association analysis, categorization and prediction, outlier analysis, evolution analysis.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

Arrangement calculations for the most part require a satisfactory and delegate set of preparing information to produce a suitable choice limit among various classes. This prerequisite still holds notwithstanding for outfit (of classifiers) based methodologies that resample and reuse the preparation information and knowledge base. In any case, securing of such information for true applications is frequently costly and tedious. Thus, it isn't phenomenal for the whole informational collection to slowly end up noticeably accessible in little groups over some undefined time frame. In such settings, a current classifier may need to take in the novel or supplementary data content in the new information without overlooking the already obtained learning and without expecting access to beforehand observed information. The capacity of a classifier to learn under these conditions is normally alluded to as incremental learning. On the other hand, in numerous applications that call for mechanized basic direction, it isn't surprising to get information got from various sources that may give integral data. An appropriate mix of such data is known as combination, and can prompt enhanced precision of the characterization choice contrasted with a choice in view of any of the individual information sources alone. Subsequently, both incremental learning and information combination include gaining from various arrangements of information using Machine Learning. In the event that the back to back informational collections that later wind up noticeably accessible are gotten from various sources as well as comprise of various highlights, the incremental learning issue transforms into an information combination issue. Perceiving this theoretical comparability, we propose an approach in view of a troupe of classifiers initially created for incremental learning as an option and outrageously well-performing way to deal with information combination to forecast the cancer and symptom even with precaution analysis. However, strategies accessible to take care of information mining issues are arrangement, affiliation run mining, time arrangement examination, bunching, rundown, and succession disclosure. Out of these machines learning lead techniques are prominent and all around inquired about information digging strategy for finding fascinating relations between factors in expansive databases. There are different affiliation control Machine Learning calculations like SVM, Association, Clustering and various approaches or relationship among information in extensive volume of dataset extraction. The greater part of the past examinations for visit itemsets age embraces an appropriate calculation that has exponential multifaceted nature (high execution time). In this examination and scheme, we propose a calculation that will diminish execution time by methods for creating item sets dynamically from static database specifically for cancer forecasting and remedial for precaution in context to cancer.

## II. RELATED WORK

**Priyanga and Prakasam** in their paper, they proposed a malignancy expectation framework in view of information mining innovation. Here they gather the client's hereditary and non-hereditary factor, which is predicts the bosom malignancy at beginning period. Primary downside of this framework is practical to the client. Weka framework is utilized to investigate the restorative data. Once the characteristics are finished, at that point the scope of the hazard can be dictated by the forecast framework. Here we are having four levels low level, middle of the road level, abnormal state and abnormal state. The above framework was effectively connected with the datasets of bosom malignancy, it gives better precision level contrasting with the current framework. This framework gives prior stage cautioning to the clients, cost and time advantages to the client.

**PriyanGopala Krishna and Murthy Nookala** made a near report about the 14 distinctive characterization calculations by utilizing the 3 distinct sorts of malignancy informational collections. The vast majority of the calculations give better outcome when size of the traits is expanded. However exactness level is relies upon the sort of the datasets to be utilized. At long last they understand that calculations are not gives the better exactness level, client endeavor to pick the best informational technique predict the cancer

**Cheng-Mei Chen and Chien Yen Hsu**, they proposed survival expectation demonstrate for liver growth utilizing information mining systems. They gather dataset from the medicinal server farm in North Taiwan between the years 2004 and 2008. They remove nine factors to liver growth survival. Counterfeit Neural Network(ANN) and Class and Regression Tree(CART) were utilized forecast display. The model was tried under three conditions: One Variable(Clinical Stage), Six huge variable and every one of the nine variable (Both huge and Non huge). The outcomes demonstrates that ANN display with nine sources of info gives 0.915% precision, 0.87% affectability and

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

0.88% specificity individually. At long last they reason that ANN demonstrate gives more exactness than the CART show.

## III. PROPOSED METHODOLOGY

**Linear Regression:** Linear Regression Technique is basically a Linear Approach which is used in order to model the relationship that exists between scalar dependent variables and also between variables more than pone termed as independent variable. Equation of Linear Regression Form expressed whereas The righteousness of fit character for the model calibrations are obtainable in below equation, and the librated coefficients are shown below. However, presents standard error ($S_e$) calculated as:-

$$S_e = \sqrt{\frac{1}{n-m}\sum(y-\hat{y})^2}$$

where $n$ is the number of observations,
$m$ is the number of coefficients or exponents being calibrated,
$y$ is the observed discharge (from the PeakFQ output), and
$\hat{y}$ is the predicted output calibrated by the regression tool.

Standard deviation ($S_y$) is calculated as

$$, = \sqrt{\frac{1}{n-1}\sum(y-\bar{y})^2},$$

where $\bar{y}$ is the mean of the discharges for the return period ($T$).

Explained variance ($R^2$) is calculated as

$$R^2 = \frac{1}{n^2 \cdot S_e^2 \cdot S_x^2}\left[\sum(y-\hat{y})\cdot(y-\bar{y})\right]^2$$

where

$$S_x = \sqrt{\frac{1}{n-1}\sum(\hat{y}-\bar{\hat{y}})^2},$$

in which $\bar{\hat{y}}$ is the mean of the predicted discharges for the return period

**Naïve Bayes** : Other learning algorithms eliminate those hypotheses, which are not consistent to an training example. Whereas the Bayesian Learning just reduces the probability of an inconsistent hypothesis. This gives the Bayesian Learning a bigger flexibility. The Bayesian Learning Algorithms combine training data with a priori knowledge to get the a posterior probability of an hypothesis. So it is possible to figure out the most probable hypothesis according to the training data. The basis for all Bayesian Learning Algorithms is the Bayes Rule.

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

P(h) = prior probability of hypothesis h
P(D) = prior probability of training data D

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

P(h|D) = probability of h given D
P(D|h) = probability of D given h

## IV. IMPLEMENTATION AND SIMULATION RESULTS

Confusion Matrix pertaining set of predictors among which some are significant in terms of mean and variance than others where estimating the predictors if the same significant predictors before prediction, pose as significant after prediction generated through the linear regression bases on dataset characteristics or attributes i.e Age, Obstruct, Performance, Adherence, Nodes, Status, Extent and Surgeries etc.

**Confusion Matrix based on Linear Regression**

| Rx | Gender | Age | Obstruct | Performance | Adherence | Nodes | Status | Difference | Extent | Surg. | Node 4 | Time | Etype |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Lev | 0 | 27 | 3.2364 | -1.3884 | -1.848 | 2.0 | 0.0 | 6.0 | 1.5701 | 0.0 | 2.0 | 1038.0 | 2.0 |
| Lev | 0 | 33 | 1.9073 | 0.078 | 0.1709 | 18.0 | 0.0 | 10.0 | 3.649 | 0.0 | 4.0 | 703.0 | 4.0 |
| Lev | 0 | 34 | -1.1892 | -40.9615 | 42.1512 | 6.0 | 0.0 | 2.0 | 8.2314 | 0.0 | 2.0 | 589.0 | 2.0 |
| Lev | 0 | 36 | -1.32 | 0.306 | 1.014 | 14.0 | 0.0 | 2.0 | 7.6766 | 0.0 | 2.0 | 813.0 | 2.0 |
| Lev | 0 | 37 | 9.1959 | -6.702 | -2.4937 | 2.0 | 0.0 | 4.0 | 2.3736 | 0.0 | 2.0 | 2613.0 | 2.0 |
| Lev | 0 | 38 | 5.549 | -1.8524 | -3.6964 | 12.0 | 0.0 | 8.0 | 6.5635 | 2.0 | 4.0 | 889.0 | 4.0 |
| Lev | 0 | 39 | 4.7394 | -0.8572 | -3.8823 | 8.0 | 0.433 | 8.0 | 4.9504 | 2.0 | 4.0 | 2069.0 | 4.0 |
| Lev | 0 | 41 | -0.6233 | 6.3963 | 4.2939 | 6.0 | 0.433 | 6.0 | 9.9002 | 2.0 | 4.0 | 1359.0 | 4.0 |
| Lev | 0 | 42 | 8.9509 | 3.6829 | -2.1713 | 20.0 | 0.0 | 6.0 | 8.6521 | 0.0 | 4.0 | 2231.0 | 4.0 |
| Lev | 0 | 43 | -0.2711 | -0.1969 | 0.4681 | 22.0 | 0.0 | 8.0 | 6.7139 | 2.0 | 4.0 | 492.0 | 4.0 |
| Lev | 0 | 44 | -10.6248 | 2.174 | 8.4507 | 10.0 | 0.0 | 4.0 | 3.4003 | 2.0 | 2.0 | 422.0 | 2.0 |
| Lev | 0 | 45 | 0.7736 | -1.7343 | 0.9607 | 30.0 | 0.4714 | 12.0 | 9.3145 | 0.0 | 6.0 | 1239.0 | 6.0 |
| Lev | 0 | 46 | 3.9708 | -1.3072 | -2.6636 | 20.0 | 0.5 | 8.0 | 5.8587 | 0.0 | 4.0 | 1442.0 | 4.0 |
| Lev | 0 | 47 | 1.6512 | -5.7003 | 4.0492 | 20.0 | 0.5 | 8.0 | 6.7573 | 0.0 | 4.0 | 1989.0 | 4.0 |
| Lev | 0 | 49 | -5.8975 | -1.1268 | 7.024 | 2.0 | 0.0 | 4.0 | 2.6901 | 0.0 | 2.0 | 2789.0 | 2.0 |
| Lev | 0 | 50 | 2.9092 | 5.5844 | 2.1703 | 12.0 | 0.3727 | 10.0 | 15.3089 | 4.0 | 6.0 | 2620.0 | 6.0 |
| Lev | 0 | 51 | 2.1705 | -0.4661 | 7.9235 | 22.0 | 0.5 | 10.0 | 4.9983 | 0.0 | 4.0 | 1650.0 | 4.0 |

Figure 1: Confusion Matrix Generated after Regression The Attributes Based on Liner Regression

Up to now we wondered which hypothesis is the most probable for a given dataset based on confusion matrix, But the question what Naive Bayesian Classifiers are about is which classification is the most probable for this new instance if we have a look at the training data.For example an instance of an patient = (age,sex,weight) could be (45,male,120kg). An Naive Bayes System could calculate values for the following two classifications “cardiac insufficiency” and “no cardiac insufficiency” according to the available training data. Then the classifier with the biggest value rules the hypothesis. Whereby the conditional independence of the attributes of the instances is required for the use of Naive Bayesian Classifiers.

How does it look brought into a formula?

1. X be a set of instances xi = (a1,a2,…,an)
2. V be a set of classifications vj

Naive Bayes assumption:

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

**International Journal of Innovative Research in Computer and Communication Engineering**

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

**Website: www.ijircce.com**

**Vol. 6, Issue 5, May 2018**

$$v = \max_{vj \in V} P(v_j \mid a_1, a_2 \ldots a_n)$$

$$= \max_{vj \in V} \frac{P(a_1, a_2 \ldots a_n \mid v_j) P(v_j)}{P(a_1, a_2 \ldots a_n)}$$

$$= \max_{vj \in V} P(a_1, a_2 \ldots a_n \mid v_j) P(v_j)$$

$$P(a_1, a_2 \ldots a_n \mid v_j) = \prod_i P(a_i \mid v_j)$$

**This leads to the following algorithm:**

Naive_Bayes_Learn (examples)
  for each target value vj
    estimate P(vj)
  for each attribute value ai of each attribute a
    estimate P(ai | vj )

Classify_New_Instance (x)

| Category | Age K | Obstruct K | Performance K | Adhere K | Nodes K | Extent K |
|---|---|---|---|---|---|---|
| Obs | 18 | 1.0706040899999998 | 0.04359743999999921 | 0.33977241000000014 | 0.0 | 0.005387766364928215 |
| No Symptom Found | | | | | | |
| Obs | 22 | 0.9161361224999993 | 0.16044030250000008 | 0.6456926025000005 | 0.0 | 0.008530354492819248 |
| No Symptom Found | | | | | | |
| Obs | 25 | 1.917394090000002 | 3.0705552900000015 | 0.30802500000000066 | 0.0 | 0.18518552664192978 |
| No Symptom Found | | | | | | |
| Lev+5FU | 26 | 0.03186224999999965 | 0.3335062500000001 | 2.8397305224999982 | 0.0 | 0.12066345602724082 |
| No Symptom Found | | | | | | |

Figure 2: Results depicting vide Naïve Classification from Confusion Matrix Dataset.

## V. CONCLUSION AND FUTURE WORK

Using the above amalgamated technique the resultant values and derived with more accuracy where as the linear regression produced the confusion matrix thus resulting the compact the precise formation of weights based on characteristics and attributes where as Naïve Bayes classifies the estimation of probabilistic model where scheme can define the range in which the prediction can be made more perfectly based on type of categorization in cancer prediction respectively. For the future work the same the same can be implements on gigantic database vide hadoop where map and reduce will cut short the datasets in small proportions and parallel execution can be performed for quick and prompt results.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

**International Journal of Innovative Research in Computer and Communication Engineering**

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

*Website: www.ijircce.com*

**Vol. 6, Issue 5, May 2018**

## REFERENCES

1. Jiewai Han and Micheline Kamber ,"Data Mining Concepts and Techniques", second editionJiewai Han and Micheline Kamber ,"Data Mining Concepts and Techniques", second edition
2. A. Kannan, Dr. V. Mohan, Dr. N. Anbazhagan, "Image Clustering and Retrieval using Image Mining Techniques", IEEE International Conference on Computational Intelligence and Computing Research, 2010.
3. Maciej Komosinski and Krzysztof Krawiec, "Evolutionary Weighting of image features for diagnosing of CNS tumors", Artificial Intelligence In Medicine• June 2000.
4. Aboul Ella Hassanien, Ajith Abraham, James F. Peters, Gerald Schaefer, Christopher Henry, "Rough Sets and Near Sets in Medical Imaging: A Review", IEEE Trans. On Information Technology In iomedicine, Vol. X, No. X, Nov. 2008.
5. Yudel Gómez, Rafael Bello, Amilkar Puris, María M. García, Ann Nowe, "Two Step Swarm Intelligence to Solve the FeatureSelection Problem" , Journal of Universal Computer Science, vol. 14, no. 15 (2008), 2582-2596.
6. Bing Liu, " Web Data Mining Exploring Hyperlinks, Content and Usage Data".
7. A.E. Hassanian, " Rough set approach for attribute reduction and Rule generation: a case of patients with suspected breast cancer, J.Am. Soc.Inform.Sci Technol.2004.
8. S. Tsumoto, "Mining Diagnostic rules from clinical databases using rough sets and medical diagnostic model", Inform.Sci. 162(2004)
9. J.Komorowski, A.Ohrn," Modelling Prognostic Power of Cardiac tests using Roughsets", Artif,Intell.Med.15(1999).
10. Li Q., Li F., Shiraishi J., Katsuragawa S., Sone S. And Doi K., \Investigation Of New Psychophys-Ical Measures For Evaluation Of Similar Images On Thoracic Computed Tomography For Distinctionbetween Benign And Malignant Nodules"", Medical Physics,Vol. 30, No.30, Pp. 2584-2593
11. Muller H., Michoux N., Bandond D. and Geissbuhler A., \A review of content-based image retrieval systems in medical applications-clinical bene¯ts and future directions"", Int J Med Informatics, vol. 73, pp. 1-23, 2004.
12. Kawata Y., Niki N., Ohmatsu H., Kusumoto M., Kakinuma R., Yamada K., Mori K., Nishiyama H., Eguchi E., Kaneko M., and Moriyama N., Pulmonary nodule classi¯cation based on nodule retrieval from 3-D thoracic CT image database, Medical Image Computing and ComputerAssisted Intervention (MICCAI 2004).
13. Lam M., Disney T., Raicu D. S., Furst J. and Channin D. S., \BRISC-An open source pulmonary nodule image retrieval framework"", Journal of digital imaging, 2007.
14. .Lens MB, Dawes M. Global perspectives of contemporary epidemiological trends of cutaneous malignant melanoma. Br J Dermatol. 2004;150:179– 85. doi: 10.1111/j.1365-2133.2004.05708.x. [PubMed] [Cross Ref]
15. Schaffer JV, Rigel DS, Kopf AW, Bolognia JL. Cutaneous melanoma: past, present, and future.J Am Acad Dermatol. 2004;51:S65–S69. doi: 10.1016/j.jaad.2004.01.030. [PubMed][Cross Ref]
16. Fiona J. Gilbert, F.R.C.R., Susan M. Astley, Ph.D., Maureen G.C. Gillan, Ph.D., Olorunsola F. Agbaje, Ph.D.,Matthew G. Wallis, F.R.C.R., Jonathan James, F.R.C.R., Caroline R.M. Boggis, F.R.C.R., Stephen W. Duffy, M.Sc.,for the CADET II Group (2008). Single Reading with Computer-Aided Detection for Screening Mammography, The New 2265 International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 4, April - 2013 ISSN: 2278-0181 www.ijert.org IJERT England Journal of Medicine, Volume 359:1675- 1684 Full text ^ Effect of Computer-Aided Detection on Independent Double Reading of Paired ScreenFilm and Full-Field Digital
17. Screening Mammograms Per Skaane, Ashwini Kshirsagar, Sandra Stapleton, Kari Young and Ronald A. Castellino^ Taylor P, Champness J, GivenWilson R, Johnston K, Potts H (2005). Impact of computer-aided detection prompts
18. On the sensitivity and specificity of screening mammography.Health Technology Assessment 9(6), 1-70.^ Fenton JJ, Taplin SH, Carney PA, Abraham L, Sickles EA, D'Orsi C et al. Influence of computer aided detection.
19. Performance of screening mammography. N Engl J Med 2007 April 5;356(14):1399-409. Full text^ Taylor P, Potts HWW (2008). Computer aids and human second reading as interventions in screening
20. Mammography: Two systematic reviews to compare effects on cancer detection and recall rate. European Journal of Cancer. doi:10.1016/j.ejca.2008.02.016 Full text ^ http://www.cancer.org/downloads/CRI/6976.00.pdf
21. Wu N, Gamsu G, Czum J, Held B, Thakur R, Nicola G: Detection of small pulmonary nodules using direct digital
22. Radiography and picture archiving and communication systems. J Thorac Imaging. 2006 Mar;21(1):27-31. PMID 16538152
23. xLNA (x-Ray Lung Nodule Assessment)
24. Petrick N, Haider M, Summers RM, Yeshwant SC, Brown L, Iuliano EM, Louie A, Choi JR, Pickhardt PJ. CT colonography with computeraided detection as a second reader: observer performance study. Radiology 2008 Jan;246(1):148- 56. Erratum in: Radiology. 2008 Aug;248(2):704. PMID 18096536
25. Halligan S, Altman DG, Mallett S, Taylor SA, Burling D, Roddie M, Honeyfield L, McQuillan J, Amin H, Dehmeshki J. Computed tomographic colonography: assessment of radiologist performance with and without computer-aided detection. Gastroenterology 2006 Dec;131(6):1690-9. Epub 2006 Oct 1. PMID 17087934
26. R. Agrawal, T. Imielinski, and A. N. Swami, ―Mining association rules between sets of items in large databases, in" Proc. 1993 ACMSIGMOD Int. Conf. Manage. Data – SIGMOD 93 SIGMOD 93, Washington, DC, 1993, pp. 207–216.
27. M.X. Ribeiro, A.J.M. Traina, C.T. Jr., N.A. Rosa, P.M.A. Marques, How to improve medical image diagnosis through association rules: The idea method, in: The 21th IEEE International Symposium on Computer Based Medical Systems, Jyvaskyla, Finland, 2008, pp. 266–271.