International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 4 | ISSN : 2456-3307

Implementation of Dynamic Bayseian Classifier for Cancer Prediction

Trupti Brahmapurikar[1], Amrapali Patil[1], Swati Dharale[1], Vishakha Patil[1], Prof. Alok Chauhan[2]

[1]BE Scholars, Department of Information Technology, Rajiv Gandhi College of Engineering and Research, Nagpur, Maharashtra, India
[2]Assistant Professor, Department of Information Technology, Rajiv Gandhi College of Engineering and Research, Nagpur, Maharashtra, India

ABSTRACT

Cancer is one of the real issue today; diagnosing cancer in prior stage is yet trying for specialists. Recognizable proof of hereditary and ecological variables is critical in creating novel strategies to identify and avert cancer. Along these lines, a novel multi layered strategy-joining clustering and decision tree procedure is utilized to manufacture a cancer risk prediction system. The proposed system is predicts lung, bosom, oral, cervix, stomach and blood cancers and it is easy to use and cost sparing. This examination utilizes data mining strategies, for example, classification, clustering and prediction to distinguish potential cancer patients. We have proposed this cancer prediction system in view of data mining strategies. This system evaluates the risk of the bosom cancer in the prior stage. This system is approved by contrasting its anticipated outcomes and patient's earlier medical data. The fundamental point of this model is to give the prior notice to the clients and it is likewise fetched proficient to the client. At last, a prediction system is created to break down risk levels, which help in guess. This examination helps in location of a man's inclination for cancer before going for clinical and lab tests which is cost and tedious.

Keywords: Cancer, Data Mining, Clustering, Classification, Decision Tree.

I. INTRODUCTION

Cancer is a conceivably lethal infection caused primarily by ecological variables that transform qualities encoding basic cell-administrative proteins. The resultant atypical cell conduct prompts extensive masses of irregular cells that devastate encompassing typical tissue and can spread to key organs bringing about scattered infection, normally a harbinger of impending patient demise. All the more essentially, globalization of unfortunate ways of life, especially cigarette smoking and the selection of numerous highlights of the advanced Western eating regimen (high fat, low fiber content) will expand cancer frequency.

Data mining strategy includes the utilization of advanced data examination devices to find beforehand obscure, legitimate examples and connections in vast data set. These apparatuses can incorporate factual models, scientific calculation and machine learning strategies in early recognition of cancer. In classification taking in, the learning plan given an arrangement of grouped cases from which it relied upon to take in a method for ordering concealed cases. In affiliation adapting, any relationship among highlights is looked for, not only ones that anticipate a specific class esteem. In clustering, gatherings of illustrations that have a place together are looked for. In numeric prediction, the result to be anticipated is not a discrete class yet a numeric amount. In this investigation, to arrange the

data and to mine regular examples in data set Decision Tree calculation is utilized.

Data Mining strategies are executed together to make a novel technique to analyze the presence of cancer for a specific patient. When starting to deal with a data mining issue, it is first important to unite every one of the data into an arrangement of occurrences. Coordinating data from various sources more often than not introduces numerous difficulties. The data must be collected, coordinated, and tidied up. At that point, no one but it can be utilized for handling through machine learning procedures. Doctors and patients alike to effectively know a man's cancer status and seriousness without screening them for testing cancer can utilize this created system. Additionally it is valuable to record and spare extensive volumes of touchy data, which can be utilized to pick up information about the malady and its treatment.

II. REVIEW OF LITERATURE

Wisconsin University breast cancer database was dissected by innocent Bayes prediction calculation and credulous Bayes classification calculations. With the goal that calculations are utilized to foresee and group whether the tumor is harmful or benign[1]. So data sets were picked arbitrarily. At the last guileless Bayes classification calculation was demonstrated that 10-15 percent is wrongly ordered and 85-95 percent is accurately arranged. Two different data sets from Wisconsin breast cancer have been assessed by various data mining calculations. The result that Rotation Forest model demonstrates the most noteworthy classification precision (99.48 %) and when contrasted and the past works, the new approach and procedure have accompanied most noteworthy execution and accuracy[2].

Jimin Guo1, Benjamin C. M. Fung, Farkhund Iqbal actualized decision tree calculation with breast cancer data sets that get from Leiden University Medical Center [3].The data sets have 574 patients who have surgery at that healing facility. So they create the repeat of breast cancer by a decision tree calculation inside three years of introductory analysis. The classifier anticipated 70% exactness. For the autonomous classifier of 65 patients the classifier precisely predicts the repeat of the illness in 55 patients. The classifier likewise isolates understanding into two in light of their malady trademark and their importance of early backslide Research paper done by Ahmed Iqbal Pritom,shahed anzarus sababa, Md. Ahadur Rahman Munshi, Shihabuzzaman Shihab predicts whether the breast cancer is intermittent or not[4]. They have utilized data sets from Wisconsin data sets of the UCI machine learning storehouse that have 35 qualities. After usage of calculations like C4.5 Decision Tree, Naive Bayes and Support Vector Machine (SVM) classification calculation was actualized. The result of these SVM, Naïve Bayes and C4.5 has 75.75 %, (67.17 %) and (73.73 %) individually.

Uma Ojha and Dr. Savita Goel were additionally examined about the investigation on the prediction of breast cancer repeat utilizing data mining methods. The examination was connected by both clustering and classification calculations. The outcomes demonstrate that decision tree and Support vector machine (SVM) turned out with the best indicator 80% precision.

Bojana R. Andjelkovic Cirkovic, Aleksandar M. Cvetkovic, Srdjan M. Ninkovic, and Nenad D. Filipovi presents the use of data mining on estimation of survival rate and sickness backslide for breast cancer patients. A data set that was taken from the Clinical Center of Kragujevac is assessed by some classification calculation. In light of chose data sets gullible Bayes calculation was chosen as a calculation which have higher exactness based on the 5 year survival rate. The examination paper done by Joana Diz Goreti Marreiros and Alberto Freitas shows new PC based analysis system[5]. By utilizing this

procedure false positive analysis test can be decreased. After data sets broke down nave bayes calculation accompany higher exactness than Random backwoods.

Qi Fan,change-jie zhu and liu yin utilized diverse kinds of data mining procedures so as to foresee the repeat of breast[6]. In this paper they specialist utilizes SEER data sets and connected another classification strategy keeping in mind the end goal to anticipate the repeat of this ailment. In the wake of preprocessing of data sets the specialist connected a few calculations, so the decision tree (c5) calculations accompany better execution.

Dursun delen, Glenn walker And Amit Kadam utilized diferent data mining systems for prediction of survivability of breast cancer. Data mining, classification calculations, for example, fake neural system and decision tree alongside strategic relapse to build up a model for breast cancer survivability. In light of this paper decision tree calculation (c5) was accompanying better. execution and anticipated by more precision 93.6% and manufactured neural system indicates second execution 91.2% and calculated relapse go to the most noticeably awful of the three 89.2%.

A phase prescient model for breast cancer survivability exhibited by Rohit j and Ramya Nadig. In the paper they were utilized distinctive calculation all together anticipate the breast cancer feasibility. The assessment was done based the phase of the breast cancer. Three machine-learning calculations were connected with a specific end goal to anticipate breast cancer survivability. These data sets assessed by classification calculations, for example, innocent bayes, strategic relapse and decision tree to anticipate breast cancer survivability.

TABLE I

REVIEW OF ALGORITHM & APPROACHES

Title	Methodology	Results
Analysis of Efficiency of Classification and Prediction Algorithms (Naïve Bayes) for Breast Cancer Dataset[1]	Naïve Bayes classification algorithm and naïve Bayes prediction algorithms	The Naïve Bayes classification algorithm has Highest accuracy value 89 %-95%
Breast cancer diagnosis using GA feature selection and Rotation Forest[2]	Rotation forest model	Rotation forest shows highest classification accuracy (99.48%)
Revealing determinant factors for early breast cancer recurrence by a decision tree [3]	Decision Tree	The classifier predicts for whether a patient developed early disease recurrence; and is estimated to be about 70% accurate
Predicting Breast Cancer Recurrence using Effective Classification and Feature Selection technique[4]IEEE 2016	Decision tree c4.5, support Vector machine, naïve Bayes algorithms	Support vector machine provide better performance before and after attribute selection.

Prediction model for estimation of survival rate and relapse for breast cancer patient[7]	ANN, SVM , LG, DT, NB	ANN= 0.9315 SVM = 0.952 LR = 0.911 DT= 0.9657 NB= 0.856 so that decision tree algorithm has achieved the best performance in a classification task considering following parameters AC =0.9657,SENS =0.991 SPECI =0.889
Predicting breast cancer survivability [8]	Decision tree,(DT) Artificial neural network , Logistic regression	Decision tree =93.6 ANN = 91.2 Logistic regression = 89.2
Using three machine learning techniques for predicting breast cancer recurrence [9]	Decision tree, Artificial neural network, Support vector machine	DT =0.936, ANN =0.947, SVM = 0.957. so that SVM have higher accuracy than other algorithms
Hybrid computer aided diagnosis system for prediction of breast cancer recurrence using optimized ensemble learning [10]	SVM Decision tree multilayer perception (MLP)	Support vector machine give a more accurate value 78%

III. PROPOSED SYSTEM

In this work, an engineering data mining method based cancer prediction system consolidating the prediction system with mining innovation utilized. In this model, we have utilized one of the classification calculations called decision tree. Once the client goes into the cancer prediction system, they have to answer the questions, identified with hereditary and non-hereditary variables. At that point, the prediction system relegates the risk an incentive to each inquiry in view of the client reactions. Once the risk esteem anticipated, the scope of the risk can be controlled by the prediction system. It has four levels of risk like low level, middle level, abnormal state and abnormal state. In view of the anticipated risk esteems, the scope of risk will be appointed.

Algorithm

Step 1: Enter the text

Step 2: Predicting system will checks for the condition.

Step 3: System predicts the values based on the user answers.

Step 4: The range of the risk is determined based on the predicted value.

Step 5: If the value is ≤18 the risk is considered as a low risk.

If the value is >18 and ≤ 21 the risk is considered as an intermediate risk.

If the value is >21 and ≤28 is considered as a high risk.

If the value is > 28 is considered as a very high risk.

Step 6: The user data is stored in database.

Step 7: The result obtained with the reference values of the database.

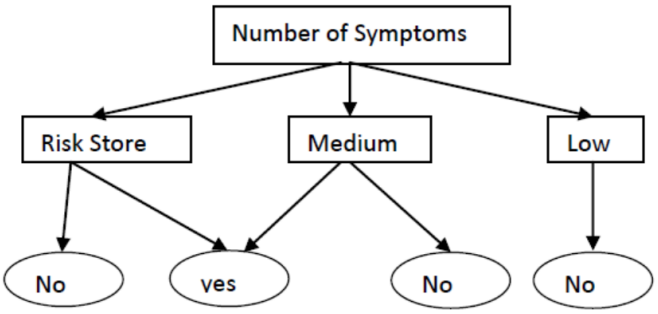
A decision tree is a stream graph like tree structure, where each interior hub signifies a test on a characteristic, each branch speaks to a result of the test and each leaf hub holds a class mark.

The best most hub is the root hub. The property estimation of the data tried against a decision tree. A way followed from root to leaf hub, which holds the class prediction for that data. Decision trees can be effectively changed over into classification rules. This decision tree utilized to create visit designs in the dataset.

The data and thing sets that happen as often as possible in the database known as regular examples. The successive examples that most essentially identified with particular cancer composes and are useful in anticipating the cancer and its write known as significant continuous example.

Utilizing this noteworthy examples produced by decision tree the data set is bunched as needs be and risk scores are given.

- *If symptoms = none and risk score x = <35 then result = you may not have cancer, tests = do simple clinical tests to confirm.*
- *If symptom= related to chest and shoulder and risk score x >= 40 then result = you may have cancer, cancer type may be= chest, tests = take CT scan of chest.*
- *If symptom= related to head and throat and risk score x >= 40 then result = you may have cancer, cancer type = oral, tests = biopsy of tongue and inner mouth.*
- *Else symptom= other symptoms and risk score x >= 40 then result = you may have cancer, cancer type = leukemia, tests = biopsy of bone marrow.*
- *Else if symptom= related to stomach and risk score x >= 45 then result = you may have cancer, cancer type = stomach, tests = endoscopy of stomach*
- *If symptom= related to breast and shoulder and risk score x >= 45 then result = you may have cancer, cancer type = breast, tests= mammogram and PET scan of breast*
- *If symptom= related to pelvis and lower hip and risk score x >=55 then result = you may have cancer, cancer type = cervix, tests = do pap smear test*
- *Based on the above mentioned rules and the calculated risk scores the severity of cancer is known as well as some tests were prescribed to confirm the presence of cancer.*



IV. CONCLUSION

Cancer is conceivably lethal sickness. Recognizing cancer is as yet trying for the specialists in the field of medication. Indeed, even now the real reason and finish cure of cancer isn't created. Identification of cancer in prior stage is treatable. In this work we have built up a system called data mining based cancer prediction system. The primary point of this model is to give the prior notice to the clients and it is likewise cost and efficient advantage to the client. It predicts three particular cancer risks. In particular, Cancer prediction system gauges the risk of the bosom, skin, and lung cancers by examining various client gave hereditary and non-hereditary elements. This system is approved by contrasting its anticipated outcomes and the patient's earlier medical record and furthermore this is examined utilizing weka system. This prediction system is accessible in on the web, individuals can without much of a stretch check their risk and make suitable move in view of their risk status. The execution of the system is superior to the current system.

V. REFERENCES

[1] G. D. Rashmi, A. Lekha, and N. Bawane, "Analysis of efficiency of classification and prediction algorithms (Naïve Bayes) for Breast Cancer dataset," 2015 Int. Conf. Emerg. Res. Electron. Comput. Sci. Technol., pp. 108–113, 2015.

[2] E. Aličković and A. Subasi, "Breast cancer diagnosis using GA feature selection and Rotation Forest," Neural Comput. Appl., vol. 28, no. 4, pp. 753–763, 2017.

[3] J. Guo et al., "Revealing determinant factors for early breast cancer recurrence by decision tree," Inf. Syst. Front., 2017.

[4] A. I. Pritom, "Predicting Breast Cancer Recurrence using effective Classification and Feature Selection technique," pp. 310–314, 2016.

[5] J. Diz, G. Marreiros, and A. Freitas, "Applying Data Mining Techniques to Improve Breast Cancer Diagnosis," J. Med. Syst., vol. 40, no. 9, 2016.

[6] Q. Fan, C. Zhu, and L. Yin, "Predicting Breast Cancer Recurrence Using Data Mining Techniques," pp. 310–311, 2010.

[7] B. R. A. Cirkovic, A. M. Cvetkovic, S. M. Ninkovic, and D. Nenad, "Prediction Models for Estimation of Survival Rate and Relapse for Breast Cancer Patients."

[8] R. J. Kate and R. Nadig, "Stage-specific predictive models for breast cancer survivability," Int. J. Med. Inform., vol. 97, pp. 304–311, 2017.

[9] A. LG and E. AT, "Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence," J. Heal. Med. Informatics, vol. 4, no. 2, pp. 2–4, 2013.

[10] M. R. Mohebian, H. R. Marateb, M. Mansourian, M. A. Mañanas, and F. Mokarian, "A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) Using Optimized Ensemble Learning," Comput. Struct. Biotechnol. J., vol. 15, pp. 75–85, 2017.

[11] C. Shah and A. G. Jivani, "Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction," 2013 Fourth Int. Conf. Comput. Commun. Netw. Technol., pp. 1–4, 2013.

[12] lulu wang. "Early Diagnosis of Breast Cancer", Sensors, 2017

[13] https://www.healthcatalyst.com/data-mining-in-healthcare