PREDICTING BREAST CANCER USING DATA MINING

Yash Rameshwar Bhise[1], Shashwat Gopal Vairale[2], Anushka Rajesh Ganjare[3]

[1,2,3]Student, Dept. of Computer Science Engineering, Prof. Ram Meghe Institute of Technology and Research, Maharashtra, India

---***---

Abstract - *Breast cancer is the second most driving malignant growth happening in ladies contrasted with every single other malignancy. Around 1.1 million cases were recorded in 2004. Watched pases of this cancer increase with industrialization and urbanization and also with facilities for early detection. It stays significantly more typical in high-salary nations yet is presently expanding quickly in middle and low-pay nations including inside Africa, a lot of Asia, and Latin America.*

Breast cancer is lethal in under portion all things considered and is the main source of death from disease in ladies, representing 16% of all disease passings around the world. The target of this is to exhibit a report on breast cancer where we exploited those accessible mechanical headways to create expectation models for breast cancer survivability. Three famous Data mining algorithms (Naïve Bayes, RBF Network, J48) are to build up the expectation models utilizing an enormous dataset (683 breast cancer cases).

The results (given normal precision Breast Cancer dataset) demonstrated that the Naïve Bayes is the best indicator with 97.36% exactness on the holdout test, RBF Network turned out to be the second with 96.77% precision, J48 turned out third with 93.41% precision.

***Key Words*: Benign, Malignant, Naïve Bayes, RBF network, J48**

1. INTRODUCTION

The number and the size of databases recording medical data are expanding quickly. Medical data, created from estimations, assessments, solutions, and so forth., are put away in various databases on a persistent premise. This gigantic measure of information surpasses the capacity of conventional strategies to examine and scan for intriguing examples and data that is covered up in them. Consequently, new systems and instruments for finding valuable data in this information storehouses are getting all the more requesting. Examining this information with new investigative strategies to discover fascinating examples and concealed information is the initial phase in expanding the conventional capacity of these information sources.

1.1 Data Mining:

Data mining and knowledge discovery in databases (KDD) are separating novel, reasonable and valuable data, information or examples from an enormous measure of accessible information. In different words, data mining has abilities for examining the huge datasets, finding startling or shrouded connections between different qualities and condensing the removed data increasingly reasonable and helpful to information clients or proprietors. In the conventional model for transforming data to knowledge, some manual examination and translation are executed. For instance, in medical centers, for the most part, specialists or experts physically examine ebb and flow patterns, infection and medicinal services information, at that point make a report and utilize this report for dynamic or making arrangements for clinical findings, medications and so forth. The issue of this sort of information examination is that this type of manual information investigation is moderate, costly, tedious, and profoundly emotional. In any case, KDD has different information preparing steps including:

- Selection: choosing objective or significant information dependent on the objective or Data mining task.

- Pre-preparing: evacuating missing, erroneous, uproarious, and conflicting or no quality information.

- Transformation: incorporates smoothing, collection, speculation or standardization, and characteristic/highlight choice.

- Data mining: applying data mining strategies or methods for extricating intriguing examples.

- Interpretation/Evaluation: incorporates factual approval, subjective survey and so on.

Data mining has two fundamental assignments:

- Predictive assignments: By applying different systems or calculations, it can settle on choices or anticipate the obscure or future estimations of different factors. These technique incorporate grouping, affiliation rule and so forth.

- Descriptive assignments: depict the information or discover human reasonable examples and present the outcomes in tables, graphs and so forth, which can be seen effectively by information proprietors or information clients.

1.2 Breast Cancer:

The organs and tissues of the body are comprised of little structure squares called cells. Cancer is a malady of these cells. Even though cells in each piece of the body may look and work unexpectedly, most fix and imitate themselves similarly. Typically, cells separate in an efficient and controlled way. Yet, on the off chance that for reasons unknown the procedure gains out of power, the cells continue separating and form into a bump called a tumor. Breast tumors are usually caused by an overgrowth of the cells lining the breast ducts. They can be either benign or malignant. In a benign tumor, the cells develop unusually and structure a protuberance. In any case, they don't spread to different pieces of the body as are not malignant growths. The most well-known kind of favorable benign breast tumor is known as a fibroadenoma. This may be precisely expelled to affirm the analysis. No other treatment is important. In a malignant tumor, the cancer growth cells can spread past the breast if they are left untreated. For instance, if a threatening tumor in the bosom isn't dealt with, it might develop into the muscles that lie under the breast. It can likewise develop into the skin covering the breast. Sometimes cells split away from the first (essential) malignant growth and spread to different organs in the body. They can spread through the circulation system or lymphatic framework. At the point when these cells arrive at another region, they may continue isolating and structure another tumor. The new tumor is regularly called an optional or metastasis. Breast cancer happens when cells inside the breast conduits and lobules become cancerous. Whenever got at the beginning time, breast cancer can frequently be restored. If the disease has spread to different zones of the body it can't generally be relieved, however it can regularly be adequately controlled for quite a while.

1.3 Risk factor associated with Breast Cancer:

Each lady needs to comprehend what she can do to bring down her danger of Breast cancer. A portion of the factors related to breast cancer:

• Family history: Women with close family members who have been determined to have breast cancer growth have a higher danger of building up the illness. In the event that you have had one first-degree female family member (sister, mother, girl) determined to have breast cancer, your hazard is multiplied.

• Genetics: About 5% to 10% of breast cancers are believed to be innate, brought about by anomalous qualities that went from parent to kid.

• Age: As with numerous different illnesses, your danger of breast cancer goes up as you get more established. Around two out of three obtrusive breast cancers are found in ladies 55 or more seasoned.

• Being a woman: Just being a lady is the greatest hazard factor for creating Breast cancer. There are around 190,000 new instances of obtrusive breast cancer and 60,000 instances of non-intrusive Breast cancer this year in American ladies. While men do create breast cancer, under 1% of all new Breast cancer cases occur in men. Around 2000 instances of breast cancer will be analyzed in American men this year.

• Personal history of breast cancer: If you have been determined to have breast cancer, you are three to multiple times bound to build up another malignant growth in the other breast or an alternate piece of a similar breast. This hazard is not the same as the danger of the first malignant growth returning (called danger of repeat).

• Being overweight: Overweight and large ladies have a higher danger of being determined to have breast cancer contrasted with ladies who keep up a solid weight, particularly after menopause. Being overweight likewise can expand the danger of breast cancer returning (repeat) in ladies who have had the illness.

• Breast feeding history: If a lady breastfeeds for longer than one year this may diminish breast cancer hazard.

• Radiation to chest or face before age 30: If you had radiation to the chest to treat another malignant growth (not Breast cancer, for example, Hodgkin's ailment or non-Hodgkin's lymphoma, you have a higher-than-normal danger of Breast cancer. In the event that you had radiation to the face at an immature to treat skin break out (something that is never again done), you are at higher danger of creating breast cancer further down the road.

• Race/ethnicity: White ladies are somewhat bound to create breast cancer than African American, Hispanic, and Asian ladies. However, African American ladies are bound to grow increasingly forceful, further developed stage breast cancer that is analyzed at a youthful age.

• Menstrual history: Women who began bleeding (having periods) more youthful than age 12 have a higher danger of bosom malignancy sometime down the road. The equivalent is valid for ladies who experience menopause when they are more seasoned than 55.

• Certain breast changes: If you have been determined to have certain favorable (not malignant growth) breast conditions, you may have a higher danger of Breast cancer.

There are a few kinds of benign breast conditions that influence Breast cancer chance.

• Pregnancy history: Women who have not had a full-term pregnancy or have their first youngster after age 30 have a higher danger of breast cancer contrasted with ladies who conceived an offspring before age 30.

• Drinking liquor: Research reliably shows that drinking mixed refreshments – brew, wine, and alcohol – builds a lady's danger of hormone-receptor-positive breast cancer.

• Having thick breast: Research has demonstrated that thick bosoms can be multiple times bound to create malignancy and can make it harder for mammograms to distinguish Breast cancer.

• Lack of exercise: Research shows a connection between practicing routinely at a moderate or extraordinary level for 4 to 7 h for every week and a lower danger of breast cancer.

• Smoking: Smoking causes various sicknesses and is connected to a higher danger of breast cancer in more youthful, premenopausal ladies. Research likewise has demonstrated that there might be connected between the overwhelming recycled smoke introduction and breast cancer growth hazard in postmenopausal ladies.

• Low of vitamin D levels: Research proposes that ladies with low degrees of nutrient D have a higher danger of breast cancer. Nutrient D may assume a job in controlling typical breast cell development and might have the option to prevent breast cancer cells from developing.

• Light exposure at night: The consequences of a few investigations propose that ladies who work around evening time – assembly line laborers, specialists, medical caretakers, and cops, for instance – have a higher danger of breast cancer contrasted with ladies who work during the day.

2. RELATED WORK

A few examinations have been accounted for that have concentrated on breast cancer survivals. These examinations have applied various ways to deal with the given issue and accomplished high arrangement correctnesses. Subtleties of a portion of the past research works are given in the accompanying:

Liu et al. utilized choice table (DT)- based prescient models for breast cancer survivability, presuming that the endurance pace of patients was 86.52%. They utilized the under-inspecting C5 strategy and stowing calculation to manage the imbalanced issue, along these lines improving the prescient exhibition on breast cancer.

Tan and Gilbert exhibited the convenience of utilizing troupe strategies in arranging microarray information and displayed some hypothetical clarifications on the presentation of gathering techniques. Therefore, they recommend that gathering AI ought to be considered for the undertaking of arranging quality articulation information for carcinogenic examples.

Chaurasia and Pal think about the presentation basis of directed learning classifiers, for example, Naïve Bayes, SVM-RBF piece, RBF neural systems, Decision Tree (Dt) (J48), and basic grouping and relapse tree (CART), to locate the best classifier in Breast cancer datasets. The exploratory outcome shows that the SVM-RBF kernel is more exact than different classifiers; it scores at the precision level of 96.84% in the Wisconsin Breast Cancer (unique) datasets.

Chaurasia and Pal offered three well known Data mining calculations: CART, ID3 (iterative dichotomized 3), and DT for diagnosing heart ailments, and the outcomes exhibited showed that CART acquired higher precision inside less time.

Chaurasia and Pal directed an analysis to recognize the most well-known Data mining calculations, executed in present-day Medical Diagnosis, and assess their exhibition on a few clinical datasets. Five calculations were picked: Naïve Bayes, RBF Network, Simple Logistic, J48, and Decision Tree. For the assessment, two Irvine Machine Learning Repository (UCI-UC) databases were utilized: heart disease and breast cancer datasets. A few exhibition measurements were used: percent of right arrangements, True/False Positive rates, the territory under the bend (AUC), exactness, review, F-measure, and a lot of mistakes.

Li et al.discovered many enhanced and noteworthy principles from high-dimensional profiling information and proposed accumulation of the separating intensity of these guidelines for dependable forecasts. The found guidelines are found to contain low-positioned highlights; these highlights are seen as now and again fundamental for classifiers to accomplish impeccable precision.

Delen et al. had taken 202,932 breast cancer patients' records, which then pre-arranged into two gatherings of "endure" (93,273) and "not endure" (109,659). The aftereffects of anticipating the survivability were in the scope of 93% precision.

Cao et al. proposed another choice tree-based group technique joined with highlight choice technique in reverse end system with stowing to discover the structure movement connections in the territory of chemometrics identified with the pharmaceutical industry.

3. WORKING

This uses three mainstream data mining algorithms each on breast cancer dataset, Naïve Bayes, RBF Network, and J48. One reason for picking the Naïve Bayes classification algorithm is because it is a basic yet ground-breaking model and it returns the expectation as well as the level of conviction, which can be valuable. RBF Network is utilized because of its focal points over customary multilayer perceptrons (MLPs), specifically quicker assembly, littler extrapolation mistakes, and higher unwavering quality. Radial basis function network (RBFN) is a class of single shrouded layer feed-forward system where the enactment capacities for concealed units are characterized as radially symmetric basis functions. J48 is an expansion of ID3. The extra highlights of J48 are representing missing qualities, decision trees pruning, persistent trait esteem ranges, derivation of rules. These classification algorithms are chosen since they are all the time utilized to look into purposes and can yield great outcomes. Additionally, they utilize various methodologies for creating the characterization models, which expands the odds for finding a forecast model with high classification accuracy.

3.1 Naïve Bayes:

Naïve Bayes is a machine learning algorithm for classification issues. It depends on Thomas Bayes' probability theorem. It is fundamentally utilized for content classification which includes high-dimensional training datasets. A couple of models are spam filtration, sentimental analysis, and characterizing news stories. It isn't just known for its straightforwardness yet additionally for its adequacy. It is quick to build models and make forecasts with the Naïve Bayes algorithm. Naïve Bayes algorithm is the algorithm that learns the likelihood of an item with specific highlights having a place with a specific group/class. To put it plainly, it is a probabilistic classifier. The Naïve Bayes algorithm is designated " naïve " because it makes the suspicion that the event of a specific component is independent of the occurrence of other features. The "Bayes" part alludes to the statistician and philosopher Thomas Bayes and the theorem was named after him, Baye's theorem, which is the base for the Naïve Bayes algorithm. It gives us a strategy to ascertain the contingent likelihood, for example, the probability of an occasion dependent on past information accessible on the occasions. All the more officially, Baye's theorem is expressed as the following equation

$$\mathrm{P}\ (A|B)= P(B|A)P(A)/\mathrm{P}(B)$$

§ $(A|B)$: Probability (conditional probability) of occurrence of event A given the event B is true.

§ (A) and (B): Probabilities of the occurrence of events A and B, respectively.

§ $(B|A)$: Probability of the occurrence of event B given that event A is true.

3.2 RBF network:

A radial basis function network is an artificial neural system that utilizes radial basis functions as activation functions. Radial basis functions are first presented in the arrangement of the genuine multivariable introduction issues. Broomhead and Lowe (1988), and Moody and Darken (1989) were the first to exploit the utilization of radial basis functions in the structure of neural systems. The yield of the network is a linear combination of radial basis functions of the inputs and neuron parameters. Radial basis function networks have numerous utilizations, including function approximation, time series prediction, characterization, and system control. Radial basis functions (RBF) network ordinarily have three layers: an input layer, a hidden layer with a non-linear RBF activation function and a linear output layer, as appeared in figure 3.2.1.

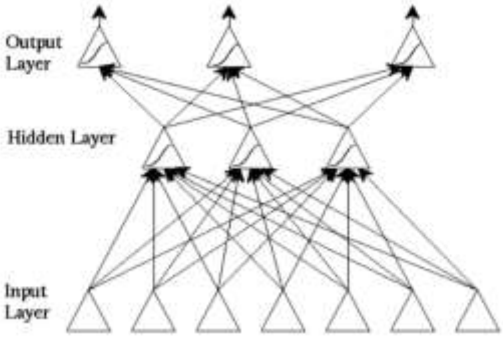

Figure 3.2.1: The structure of RBF neural network.

3.3 J48 Decision Tree:

J48 decision trees classifier is a basic decision learning algorithm, it acknowledges just categorical data for building a model. The basic idea of ID3 is to develop a decision tree by utilizing a top-down greedy hunt through the given set of training data to test each attribute at each node. It utilizes statistical property known as information gain to choose which attribute to test at every node in the tree. Information gain quantifies how well a given attribute isolates the preparation tests as per their classification. It is reasonable for taking care of both categorical as well as continuous data. A threshold value is fixed to such an extent that all the qualities over the limit are not thought about. The underlying advance is to ascertain data gain for each attribute. The attribute with the maximum gain will be favored as the root node for the decision tree. Given a set S of Breast cancer cases, J48 first grows an initial tree using the divide-and-conquer algorithm as follows: If all the cases in S have a place with a similar class or S is little, the tree is a leaf named with the most incessant class in S, in any case, pick a test depends on a single attribute with at least two results. Make this test the root of the tree with one branch for every result of the test, partition S into relating subsets S1, S2,......, Sn for a dataset containing n cases as indicated by the result for each case, and apply a similar method recursively to every subset. It utilizes a statistical property known as information gain to choose which attribute to test at every node in the tree. It quantifies how well a given attribute isolates the training samples as per their classification.

4. WISCONSIN BREAST CANCER DATASET

This has 699 cases (Benign: 458 Malignant: 241) of which 16 occasions has missing attributes estimations expelling that we have 683 cases of which 444 benign and 239 are malingnant. Features are processed from a digitized picture of a Fine Needle Aspiration (FNA) of a breast mass. Table 4.1 presents the description about the attributes of the WBC dataset.

Attributes	Domain
Sample code number	Id number
Clump thickness	1-10
Uniformity of cell size	1-10
Uniformity of cell shape	1-10
Marginal adhesion	1-10
Single epithelial cell size	1-10
Bare nuclei	1-10
Bland Chromatin	1-10
Normal nucleoli	1-10
Mitoses	1-10

Table 4.1: Breast Cancer Database.

5. PERFORMANCE

The irrefutable pioneer in lion's share of cases is the Naïve Bayes. By and by, generally speaking, execution is in every case better in contrast with different algorithms. With regards to the RBF Network, it wins the runner up as far as the performance. For a large portion of the databases and measurements, the outcomes picked up by this algorithm were somewhat more awful than for the Naïve Bayes in the greater part of the cases. At long last, the most noticeably worst outcomes were yielded by the J48. The explanation behind this may the idea of clinical data. Its intricacy and heterogeneity of estimations of attributes can frustrate Data mining.

The best is of the Naïve Bayes if there should be an occurrence of the breast database. However, the overall best algorithm is the Naïve Bayes, with the RBF Network being the second

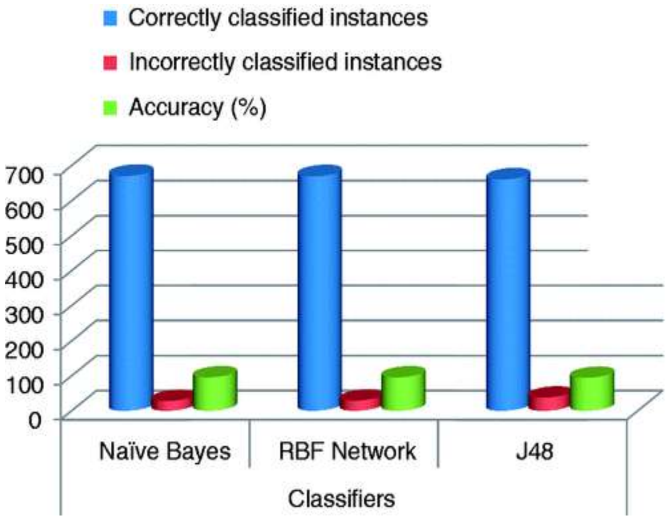

Figure 5.1: Comparative graph of different classifiers showing at different evaluation criteria.

Now we calculate the accuracy, sensitivity and specificity for the methods Naïve Bayes, RBF Network, and J48 using the formulae

Accuracy=TP+TN/TP+FP+FN+TN

Sensitivity=TP/TP+FN

Specificity=TN/TN+FP

The calculated values for the three methods Naïve Bayes, RBF Network, and J48 are listed in Table 5.1.

Methods	Accuracy(%)	Sensitivity(%)	Specificity(%)
Naïve Bayes	97.36	97.4	97.90
RBF Network	96.77	97.07	96.23
J48	93.41	93.4	90.37

Table 5.1: Accuracy, sensitivity, specificity of different methods.

The analysis conveyed intriguing outcomes. The best classifier is Naïve Bayes. Its overall performance ended up being the most noteworthy if there should be an occurrence of the greater part of the databases. This might be brought about by the idea of data being complex could have caused overtraining of the different algorithms. The runner up was won by the RBF network. Its general performance was just somewhat more regrettable than Naïve Bayes'. On the third was the J48.

6. ADVANTAGES AND DISADVANTAGES

6.1 Advantages:

• Marketing/Retail

Data mining enables promoting organizations to build models dependent on historical data to predict who will react to the new advertising efforts, for example, standard mail, internet showcasing effort… and so forth. Through the outcomes, advertisers will have a proper approach to deal with offering gainful items to focused clients. Data mining carries a lot of advantages to retail companies similarly to marketing. Likewise, it additionally helps the retail companies offer certain limits for specific items that will pull in more clients.

• Finance/Banking

Data mining gives financial institutions information about loan data and credit reporting. By building a model from recorded client's data, the bank, and financial institution can decide great and awful loans. Likewise, data mining banks detect fraudulent credit card transactions to protect credit card owners.

• Manufacturing

By applying data mining in operational engineering data, manufactures can identify flawed gear and decide ideal control parameters. For instance, semiconductor manufacturers have a test that even the states of manufacturers environments at various wafer creation plants are comparative, the nature of wafer is a ton the equivalent and some for obscure reasons even have deserted.

• Government

Data mining helps government agencies by burrowing and analyzing records of the financial exchange to build designs that can distinguish illegal tax avoidance or crimes.

6.2 Disadvantages:

- Protection Issues:

The worries about individual protection have been expanding hugely as of late particularly when the web is blasting with interpersonal organizations, internet business, gatherings, online journals. As a result of security issues, individuals fear their own data is gathered and utilized in a deceptive manner that possibly causing them a lot of difficulties. Businesses gather data about their clients from numerous points of view for understanding their buying practices patterns. Anyway, organizations don't keep going forever, every so often they might be gained by others or gone. Right now, the individual data they claim presumably is offered to others or spill.

- Security issues:

Security is a major issue. Businesses possess data about their representatives and clients including government managed savings number, birthday, finance and so forth. Anyway how appropriately this data is taken consideration is still in question. There has been a lot of cases that hackers got to and took huge data of clients from the big corporation, for example, Ford Motor Credit Company, Sony… with so much personal and financial data accessible, the credit card shoplifting and identity theft become a major issue.

- Misuse of information/inaccurate information:

Data is gathered through data mining planned for ethical purposes that can be misused. This data might be abused by unethical individuals or organizations to take advantage of helpless individuals or victimize a gathering of individuals.

7. CONCLUSION

Three prediction models are for breast cancer survivability on two parameters: benign and harmful malignancy cancer patients. Here, there are three famous data mining strategies: Naïve Bayes, RBF Network, and J48. Dataset (683 examples) is acquired from the UCI. Procedures like data selection, preprocessing, and transformation are applied to build up the prediction models. The performance demonstrated that the Naïve Bayes played out the best with a classification accuracy, RBF Network turned out to be second best with a classification accuracy, and the J48 turned out to be the third with classification accuracy. Notwithstanding the prediction model, leading sensitivity analysis and specificity analysis on Naïve Bayes, RBF Network, and J48 to pick up understanding into the general commitment of the independent factors to anticipate survivability.

REFERENCES

[1] Chaurasia, Vikas and Pal, Saurabh, Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability (June 29, 2017). International Journal of Computer Science and Mobile Computing IJCSMC, Vol. 3, Issue. 1, January 2014, pg.10 – 22. Available at SSRN: https://ssrn.com/abstract=2994925 | Google Scholor

[2] Chaurasia, V., Pal, S., &Tiwari, B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology, 119–126. https://doi.org/10.1177/1748301818756225 | Google Scholor

[3] Aruna, S., S. P. Rajagopalan, and L. V. Nandakishore. "Knowledge based analysis of various statistical tools in detecting breast cancer." Computer Science & Information Technology 2 (2011): 37-45.| Google Scholor

[4] El-Sebakhy, Emad A., et al. "Evaluation of breast cancer tumor classification with unconstrained functional networks classifier." IEEE International Conference on Computer Systems and Applications, 2006.. IEEE, 2006. | Google Scholor

[5] Gupta, Shelly, Dharminder Kumar, and Anand Sharma. "Performance analysis of various data mining classification techniques on healthcare data." International journal of computer science & Information Technology (IJCSIT) 3.4 (2011): 155-169. . | Google Scholor

[6] Fallahi, Amir, and ShahramJafari. "An expert system for detection of breast cancer using data preprocessing and bayesian network." International Journal of Advanced Science and Technology 34 (2011): 65-70.| Google Scholor

BIOGRAPHIES

Yash Rameshwar Bhise
Pursuing Bachelor of Engineering.
(Computer Science & Engineering).

Shashwat Gopal Vairale
Pursuing Bachelor of Engineering.
(Computer Science & Engineering).

Anushka Rajesh Ganjare
Pursuing Bachelor of Engineering.
(Computer Science & Engineering).