(IJRMST) 2019, Vol. No. 7, Jan-Jun **e-ISSN: 2455-5134, p-ISSN: 2455-9059**

DEVELOPING AN INTEGRATED SYSTEM COMBINING DIFFERENT CLASSIFICATION TECHNIQUES OF MACHINE LEARNING IN THE EARLY DIAGNOSIS OF BREAST CANCER

Ruchika Chakravarti

Delhi Technological University (DTU), New Delhi

ABSTRACT:

Analysis of Data assumes fundamental parts in conclusion and treatment in the medical care area. To empower professional independent direction, process gigantic volumes of information with AI procedures to create apparatuses for the expectation and characterization of Breast Cancer reports of 1 million cases each year. We have proposed a forecast model, which is explicitly intended for the expectation of Breast Cancer utilizing Machine learning calculations, a Decision tree classifier, Naïve Bayes, SVM and K-Nearest Neighbour calculations. The model predicts the sort of cancer. The growth can be harmless (noncancerous) or dangerous (dangerous). The model purposes directed learning, an AI idea where we give subordinate and free segments to machines. It utilizes a characterization procedure which predicts the kind of cancer.

I. INTRODUCTION

Breast disease is among the unavoidable tumours tracked down in ladies. It is a disease in which the bosom cells develop unusually. Female breast malignant growth has outperformed cellular breakdown in the lungs as the most ordinarily analyzed disease worldwide, with an expected 2.3 million new cases (11.7%)[1]. 10% of the breast disease cases are genetic, and the other 90% are connected with the way of life factors. A huge expansion in bosom disease occurrence rates was seen in 15 PBCRs in females. Most patients went through multimodality treatment, and 97.7% were epithelial growths. Israel (84.6) had Asia's most significant occurrence of bosom malignant growth. In India, the Hyderabad locale (48.0) had the most noteworthy occurrence rate[2]. As indicated by a report distributed by National Cancer Registry Program (NCRP), malignant growth cases are supposed to increment from 13.9 lakh in 2020 to 15.7 lakh by 2025, expecting a 20 per cent increment in general [3]. It can keep normal diseases from being deadly whenever treated early. A bosom disease finding made early can prompt Viable treatment. The examination plans to order the patients in Malignant and Benign kinds of growths by arrangement methods, accomplishing higher exactness. The dataset is from the Kaggle site. We have utilized regulated realizing, an AI idea where we give reliant and free sections to the machine for learning. After the educational experience, the machine will predict the incentive for the reliant variable for a given contribution to the type of a free factor. The arrangement strategies for identifying the cancer are, Decision tree, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Naïve Bayes (NB) arrangement in Jupyter notebook Information perception.

II. BACKGROUND STUDY: MACHINE LEARNING ALGORITHMS

The accompanying four AI arrangement methods are utilized:

(IJRMST) 2019, Vol. No. 7, Jan-Jun **e-ISSN: 2455-5134, p-ISSN: 2455-9059**

A. Decision Tree Classifier

In 1980, J. Ross Quinlan created ID3(Iterative Dichotomiser), a choice tree calculation. This classifier is an

illustration of directed AI. A choice tree works on potential answers for a choice in light of specific circumstances. It characterizes conditions at each hub to track down an answer.

Calculation

1) Starts at the root hub

2) Root esteem contrasted and recorded genuine dataset characteristic

3) Jump to the following hub, given the correlation

4) Compare quality worth with sub-hub worth and bounce likewise

5) Repeat till the left hub of the tree is reached.

B. Naive Bayes

The gullible Bayes characterization procedure depends on the Bayes' Theorem. It is named 'guileless' because this calculation expects each information

variable to be autonomous. It is a fast and straightforward ML calculation used to foresee a class of typically enormous classes.

When a class variable is given, the Naïve Bayes classifier expects that an element's presence or nonattendance is irrelevant to the presence

or, again, the nonappearance of different highlights. It is very helpful and successful when complex issues are involved.

Calculation

1) Separate the preparation information by class

2) Calculate the mean and standard deviation for each quality

3) Summarize and coordinate the dataset by class

4) Calculate the Gaussian Probability Density capability

5) Lastly, compute the class probabilities

D. K-Nearest Neighbor

K-Nearest Neighbor is a Machine Learning calculation which has a place with the Supervised Learning strategy. K-NN calculation can take care of grouping and relapse issues however is principally utilized for order issues. It is a non-parametric grouping strategy which doesn't make suppositions given basic information. It is likewise a lethargic student calculation because the preparation set isn't advanced immediately. It stores the dataset, and afterwards, it follows up on it at grouping time. KNN stores datasets during the preparation stage, and when it gets new datasets, it sorts them into a comparative classification to the new information. In the K-NN the calculation expects the similitude between another case and accessible information, and the new case is set in the classification with the most closeness to the accessible cases. Every accessible information is put away, and new information focuses are characterized given similitude. Subsequently, when new information shows up, the K-NN calculation can effectively arrange it into a proper class. Calculation

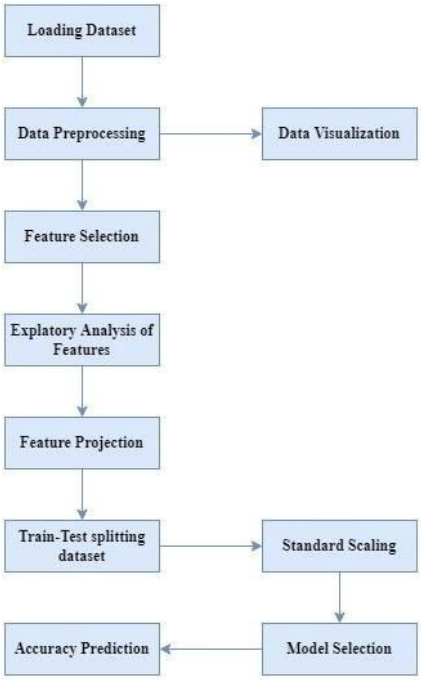
1) Load preparing and testing information from the dataset

2) Choose the worth of the closest data of interest (K)

3) For every data of interest, compute Euclidean (distance among testing and columns of preparing information) distance and sort them ascendingly

4) Choose the top K column from the arranged cluster and designate a class to test point in light of the most continuous class

III. PROPOSED METHODOLOGY



A. Stage 1: Loading Dataset

In this stage, the information from the Breast Cancer Wisconsin (Diagnostic) Data Set is stacked into the jupyter note pad for additional investigation.

B. Stage 2: Data Pre-processing

Information pre-handling alludes to controlling datasets to guarantee that main important information is utilized further for grouping. The dataset utilized may contain credits that are not helpful for arrangement and can cause mistakes in the model. Consequently, information pre-handling is vital as it sorts this crude information into valuable information. The crude information is changed into a justifiable arrangement for future investigation. This is a vital stage as the model's exhibition relies upon the information utilized.

C. Stage 3: Data Visualization

This information that has been pre-handled is currently imagined utilizing seaborn and matplotlib libraries of Python. Information representation helps better comprehension as it interprets the huge datasets into realistic and metric models, which are simple for the human brain to understand. In the accompanying examination, the framework is utilized to

check whether there are any invalid qualities in the dataset. Likewise a co-connection between the various properties is plotted for representation purposes.

D. Stage 4: Feature Selection

Include Selection is the cycle where the highlights which contribute most to your expectation variable or result are chosen physically.

In the dataset, there are 31 sections containing various highlights that could be essential for examination. So certain elements that don't assist with foreseeing exactness are dropped and not utilized.

E. Stage 5: Exploratory Analysis Of Features

Exploratory Data Analysis is the cycle of reviewing information to track down examples and inconsistencies. Utilizing different graphical representations and outline measurements, a theory is made to study and investigate suspicions. In this paper, the quantity of harmless and threatening cancers in the dataset's conclusion segment is counted and outwardly addressed utilizing components like outlines and charts.

F. Stage 6: Feature Projection

This paper utilizes the element projection method to develop information change and delegate learning further. Include projection or element extraction alludes to the direct mix of new highlights utilizing existing highlights. Include projections to change the high layered information to those with fewer characteristics.

G. Stage 7: Train-Test Splitting Dataset

The train-test parting strategy is utilized to assess and investigate the exhibition of any AI model. This method is utilized for both characterization and relapse issues. In this model, grouping methods have been assessed, where the information is parted into two subsets-preparing and testing. Then the applicable information is tried to get the important results.

H. Stage 8: Standard Scaling

Normalization is a scaling method that rescales the circulation upsides of a dataset. It tends to be helpful when information follows Gaussian appropriation. In the model, the scaler is fitted on preparing information and is then used to change prepared information. This assists with staying away from any information spillage all through the model preparation process.

I. Stage 9: Model Selection

In model choice, fitting AI models are chosen for preparing the dataset to obtain functional outcomes. Our paper applies the model choice interaction across various models like SVM, K-NN, Naïve Bayes, and Decision Tree Classifiers. In our dataset, we have a reliant variable Y with just two arrangements of values: Malignant (M) and Benign(B). Our model deals with this arrangement of values for anticipating the result in light of the different order calculations.

J. Stage 10: Accuracy Prediction

This paper predicts the exactness for characterizing the kind of cancer in the dataset. Exactness is the level of accurately grouped occasions for the prepared dataset, particularly TN, TP, FN, and TP. It is the proportion of the number of right forecasts to the number of complete expectations. The misclassifications in the dataset are determined, which comprehends why such precision has been anticipated.

IV. EXECUTION EVALUATION

In execution assessment, we check the various accuracy anticipated utilizing the different grouping calculations. Carried out this work in the JUPYTER note pad of ANACONDA guide in 1.6 GHz Dual-Core Intel Core i5 with 8GB memory. We utilized four calculations, specifically SVM, K-NN, Naïve Bayes and Decision Tree, for making the expectation.

The information was part of utilizing the train-test parting strategy into a 70-30% model for preparing and testing. The information was then standard scaled for examination.Utilized an alternate pipeline strategy to make initial forecasts on the dataset. It had the accompanying forecasts:

Algorithm	Accuracy
SVM	96.808
KNN	95.744
Naïve Bayes	93.617
Decision Tree	95.213

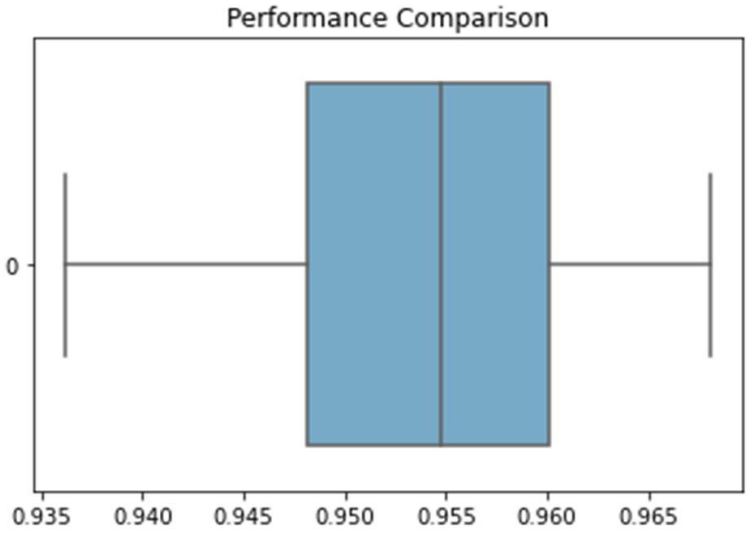


Fig 2: Boxplot of performance comparison

Completed the various calculations and anticipated the preparation and testing dataset.

Later an order report was ready, which comprised information exactness for harmful and harmless cancers. Accuracy, support and F1-score for both 0-Benign and 1-Malignant were determined. The F1 score is the typical accuracy and review score for threatening and harmless cancers. Misleading up-sides (FP) and bogus negatives (FN) are considered when computing this score. The table underneath addresses the result assembled from various characterization calculations. It specifies the weighted normal.

Algorithm	Accuracy	Precision	Recall	Support	F1-score
SVM	97.8723	0.98	0.98	188	0.98
KNN	97.3753	0.97	0.97	188	0.97
Naïve Bayes	93.617	0.94	0.94	188	0.94
Decision Tree	96.808	0.97	0.97	188	0.97

• Exactness: Determines the best model for perceiving designs in a dataset

• Accuracy: Number of TP by all-out number of positive forecasts

(IJRMST) 2019, Vol. No. 7, Jan-Jun **e-ISSN: 2455-5134, p-ISSN: 2455-9059**

• Review: Measure to recognize TP accurately

• Support: Number of right examples in each class of target values

• F1-score: Weighted average accuracy score and review plotted a disarray network to work out the errors in each model. A disarray network is utilized to assess and describe the exhibition of a classifier.

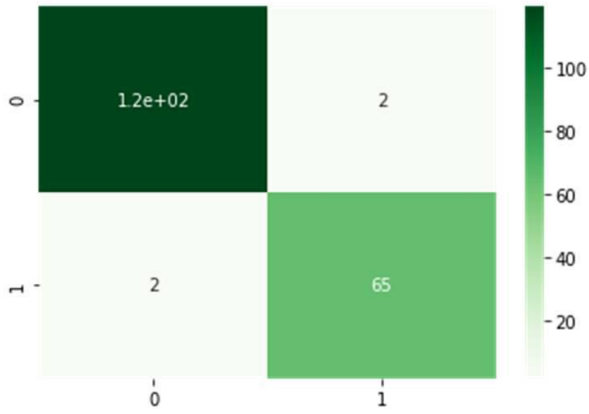
•True Positive (TP): Correctly predicts positive class

• False Positive (FP): Incorrectly predicts positive class

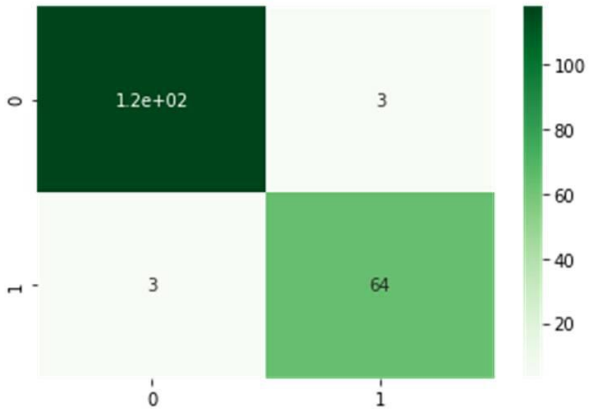
• False Negative (FN): Incorrectly predicts misleading class

• True Negative (TN): Correctly predicts bogus class

The SVM calculation showed minimal errors of 4 as 2 were FP, and the other 2 were FN. We can see the errors in the underneath referenced charts of the disarray framework.



Fig(3) SVM



Fig(4) K-NN

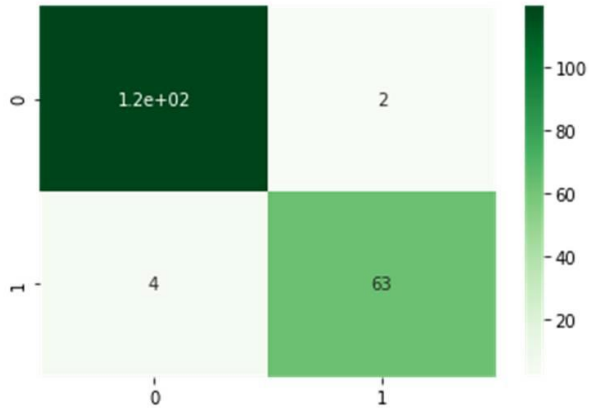


Fig (5) Decision Tree

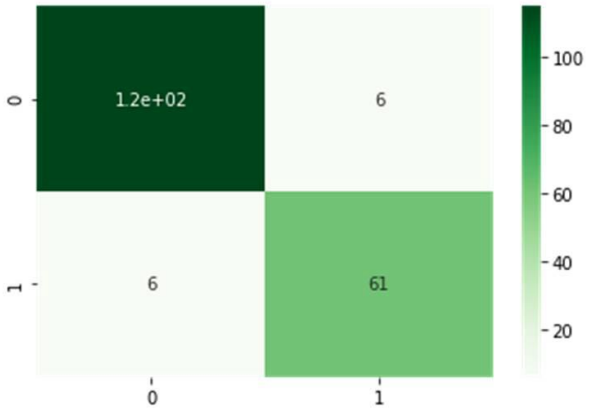


Fig (6) Naïve Bayes

V. CONCLUSION

This paper uses unique classifiers to examine a Breast Cancer Wisconsin (Diagnostic) Data Set with 32 characteristics and predicts whether the growth is harmless or threatening. The highest precision is obtained utilizing the SVM model with a precision of 97.87% for the straight bit. K-NN model showed an exactness of 97.37%, while Naïve Bayes showed 93.61% precision, and the Decision Tree showed 96.81% exactness. We can sum up by saying that the SVM model

showed the most proficient and successful precision out of the relative multitude of 4 models utilized for arranging harmless and dangerous growths.

REFERENCES

1. Hyuna Sung, PhD1; Jacques Ferlay, MSc, ME2; Rebecca L. Siegel, MPH1; Mathieu Laversanne, MSc2; Isabelle Soerjomataram, MD, MSc, PhD2; Ahmedin Jemal, DMV, PhD1; Freddie Bray, BSc, MSc, PhD2, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries", CA CANCER J CLIN 2021;71:209–249, Volume 71, Number 3 May/June 2021, p. 209
2. Report of National Cancer Registry Programme (ICMR-NCDIR), Bengaluru, India 2020, Chp. 7, p. xvi
3. Ritu Madhan, Shivani Kalariya, "Design and Development of Prosthetic Brassieres for Breast Cancer Patients", Journal of Scientific Research Institute of Science, Banaras Hindu University, Varanasi, India, Volume 65, Issue 4, 2021
4. Muktevi Srivenkatesh, "Prediction of Breast Cancer Disease using Machine Learning Algorithms", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-4, February 2020
5. Ramik Rawal, "BREAST CANCER PREDICTION USING MACHINE LEARNING", Journal of Emerging Technologies and Innovative Research, May 2020, Volume 7, Issue 5
6. Gaurav Singh, "Breast Cancer Prediction Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Volume 6, Issue 4 Page Number: 278-284, July-August 2020
7. Min-Wei Huang,Chih-Wen Chen, Wei-Chao Lin, Shih-Wen Ke,Chih-Fong Tsai, "SVM and SVM Ensembles in Breast Cancer Prediction", PLoS ONE 12(1): e0161501.
8. Deepika Verma, Nidhi Mishra, "Comparative analysis of breast cancer and hypothyroid dataset using data mining classification techniques", 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)
9. Ch. Shravya, K. Pravalika, Shaik Subhani, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-6, April 2019.
10. Wang Haifeng; Yoon Sang Won, "Breast Cancer Prediction Using Data Mining Method", IIE Annual Conference. Proceedings; Norcross (2015): 818-828