EAST WEST UNIVERSITY

Course Title: Machine Learning

Course Code: CSE475

Section: 3

Semester: Summer 2021

Submitted to

Dr. Md Samiullah

Adjunct Faculty, Department of Computer Science and Engineering

Submitted by

Mahin Khan

ID:2017-2-63-001

Bibi Joynab
ID: 2017-3-60-031

Sadia Yasmin

ID:2017-3-60-004

Mehedi Hassan Majumder

ID:2017-3-60-034

Git Hub Link—

https://github.com/mahinkhan727/Breast-Cancer

Abstract—

Breast cancer is the most common invasive cancer in women and the second leading cause of cancer death in the world. Breast cancer is cancer that develops in breast cells. Typically, the cancer forms in either the lobules or the ducts of the breast. The World Health Organization was seen by recharging that top five cancers are most common `breast cancer. 627000 patients died in the world in 2018 for breast cancer. Breast cancer is the most vital cause of death among females. Some kinds of research have been done on early detection of breast cancer to start treatment and enhance the opportunity of living. However, mammogram reflection from time to time has a risk of wrong detection that may danger the patients' health. To signify to find substitute methods that are easier to materialize and perform more faithful predictions. This paper suggests a hybrid model mixing of some Machine Learning algorithms including Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest, AdaBoost, Principal Component Analysis (PCA), Gradient Boosting (GB), Gradient Descent (GD), Regression Tree (RT) and Decision Tree (DT) for effective breast cancer detection. This study also discusses the datasets used for breast cancer detection and diagnosis. The proposed model can be used with one data type.

Keywords—

Breast Cancer; Random Forest; Classification; CSV; Earlier Detection; Machine Learning

Git Hub Link—

https://github.com/mahinkhan727/Breast-Cancer

1. Introduction-

Breast cancer is one of the most dangerous disease in women. Breast cancer is the fatal disease. Day by day increasing the rate of death for breast cancer. Breast cancer is not properly detected for this reason the rate of death is increasing. Machine learning is an efficient way to detect breast cancer. The vital part of machine learning is the dataset used. The key objective of this paper is to propose random forest algorithm to detect breast cancer. This paper presents a detailed study of existing cancer detection models and presents the highly accurate and efficient results. For this reason data sets should be substantive as much as possible cause small changes can result in drastic changes.

2. BG study

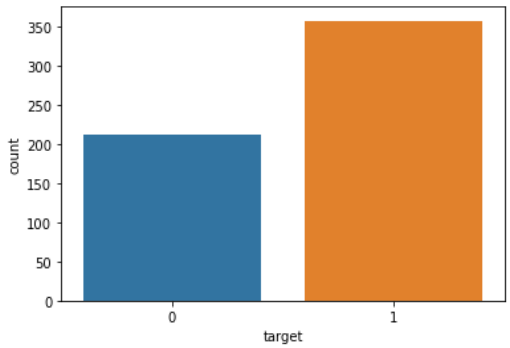
By 2014, there was a 24 percent increase (232,670) in new cases of female breast cancer. Now a day's breast cancer became a major issue of the world. By using our breast cancer detection technology it'll much more efficient way to detect the anomaly.

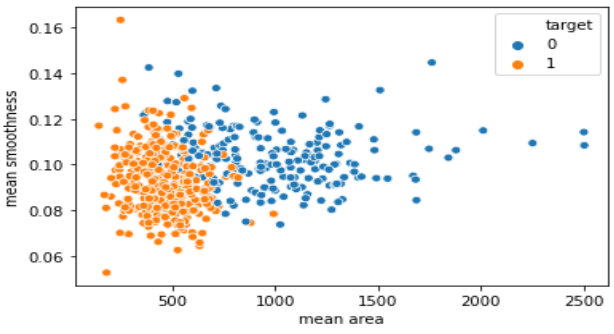
3. Idea + Implementation

Here, among all other machine learning approaches we found Random Forest Classification more suitable for finding out the results. Here in the dataset used for breast cancer detection there are 398 instances of one class and 171 instances of another class. The instances are described by 30 attributes, most of them are nominal. Random forest aggregates many decision trees rather than a single decision tree thus avoiding over fitting error along with error due to bias. Moreover, Random Forest algorithm has already been implemented in many cases to analyze a patient's medical history to identify diseases. Since Random forest classification can derive the result accurately and also has reliable classification performance, we have used this method. For training and testing we divided the dataset into 70% and 30% respectively.

The datasets consist of 569 number of Instances 30 number of Attribute data donated 2017-02-13. This is one of three domains which has repeatedly appeared in the machine learning literature. This data set includes 398 instances of one class and 171 instances of another class. The instances are described by 30 attributes, most of them are nominal. The data type is a string type with any numerical value.

Sample dataset target:





4. Experimentation

Using the **Random Forest algorithm** for 2000 estimators, we were able to produce an F1 accuracy score of more than 97%. Based on the confusion matrix, we could see that out of the 171 test data set, our algorithm produced 58 true positives and 3 false positives for no-recurrence-events class; 2 false negatives and 108 true negatives for the recurrence-events class.

Using the **Support Vector Machine** algorithm for 2000 estimators, we were able to produce an F1 accuracy score of more than 96%. Based on the confusion matrix, we could see that out of the 171 test data set, our algorithm produced 55 true positives and 6 false positives for no-recurrence-events class; 0 false negatives and 108 true negatives for the recurrence-events class.

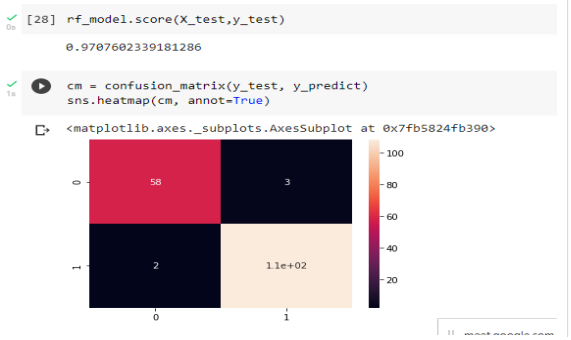


Fig: Random Forest model accuracy

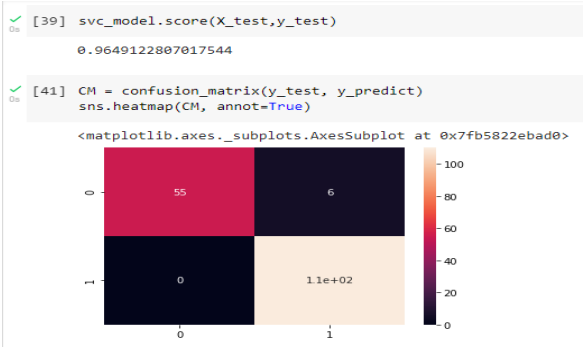


Fig: SVC model accuracy

5. CONCLUSION

Machine learning technique is the effects of these newness Are in going to more application domains progressively and the medic field is one of them. Decision making in medical Field is a big problem, sometimes it can be a trouble. The Machine learning.