"BREAST CANCER DISEASE PREDICTION: USING MACHINE LEARNING APPROACH"

Bhondve Arti T[1], Bhame Vaishali S[2], Kadam Aishwarya R[3], Kopnar Komal D[4]

[1,2,3,4]Department of Information Technology, SVPM's COE Malegaon(bk), Baramati

---

Abstract - Cancer is not a single disease, but rather many related diseases that all involve uncontrolled cellular growth and reproduction. Breast cancer is the leading cause of death in the developed world and second in the developing world, killing almost 8 million people a year. Since cancer is many diseases, treating an individual cancer requires knowing what abnormal behaviors are happening inside the cells. Machine learning is the subfield of computer science that studies programs that generalize from past experience. This project looks at classification, where an algorithm tries to predict the label for a sample. The machine learning algorithm takes many of these samples, called the training set, and builds an internal model. Machine learning is the study of algorithm and systems that improve their performance with experience. Machine learning is use to classify and predict the data. This model is use to predict by using machine learning approaches. This model is use to accurate analysis of medical data and early breast cancer disease prediction. Using KNN algorithm and decision tree, by clustering tumours are predicted breast cancer is benign or malignant.

Keywords— machine learning, healthcare, decision tree, big data, K-nearest neighbor algorithm.

1. Introduction

Over the past decades, a continuous evolution related to breast cancer research has been performed. Scientists applied different methods, such as screening in early stage, in order to find types of breast cancer before they cause symptoms. Moreover, they have developed new strategies for the early prediction of breast cancer treatment outcome. With the advent of new technologies in the field of medicine, large amounts of breast cancer data have been collected and are available to the medical research community. However, the accurate prediction of a breast cancer disease outcome is one of the most interesting and challenging tasks for physicians. Cancer or tumor is a group of disease that involve abnormal cell growth with the potential to spread to other parts of the body but not all tumors are cancerous. There are various types of cancer, including breast cancer, skin cancer, lung cancer, colon cancer and lymphoma. Breast cancer is the one of the popular and second leading cause of cancer death. It can be found in both men and women, but it is more common for women. It is important to all people especially women to be aware of changes in the breasts and to know the signs and symptoms of breast cancer. In this project, we propose K-NN and DT algorithms that functions is a reliable for cancer prediction. K-NN algorithm use for classification application of text categorization. The decision tree classifier (DTC) is one of the possible approaches to multistage decision making. decision tree algorithm can be use regression and classification problems. The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from training data. Decision tree classifier provide flexibility as well as accuracy and time / space efficiency. Decision tree provide more unified view of Decision tree classifier. Loaded Data set is not pre-processed. It is noisy , redundant and unreliable data so, data set pre-processing is the final training set. It is well known that data preparation and filtering steps take considerable amount of processing time in machine learning problems. Data pre-processing includes data cleaning, transformation, feature extraction and selection, etc.

In this work, we constructed an expert system called cancer prediction system which predict specific cancer "breast cancer" risk, it help the user to predict the cancer. It can save cost and time. This system help the people to know their cancer risk and it also help the people to take appropriate decision based on their cancer risk status.

2. Proposed system

We proposed the User/Patient is Register with their all Basic information .After successful login to the system the user will enter his/her all previous medical history. The user will enter their symptoms in the system and Syste will classify the symptoms and show the disease is benign or malignant. Using KNN algorithm and decision tree, by clustering tumours are predicted if it is benign or malignant. The user will get the prediction of breast cancer disease.

The aim of this study is to develop an early stage and malignant stage breast cancer detection system which can automatically classify abnormalities in patients lab analysed records. In this method, data pre-processing stage is to remove the noise . Data analyzation is used to predict the cancer is benign or malignant. We can insert new data of patient and get new analyzation data records. This system is affordable cost.

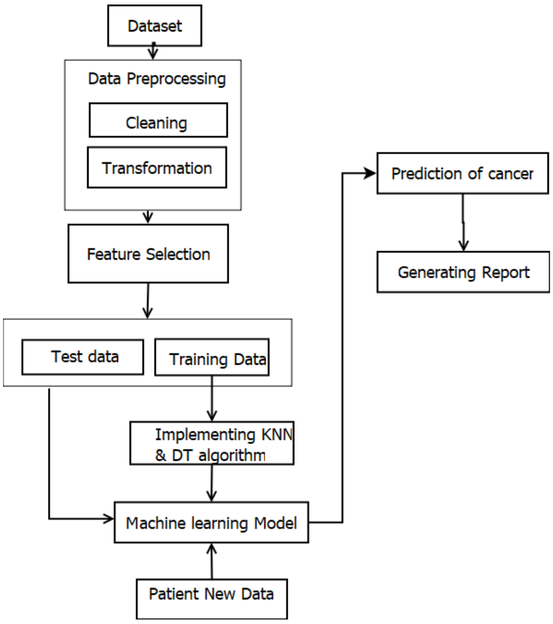


Fig 1. System Architecture

A Dataset is a collection of data. Load Dataset into the program. Data Pre-processing is a important step in machine learning process. Data pre-processing is a technique that is used to convert the raw data into a clean dataset. Data is clean through process such as handle the missing value ,noisy data or resolving the inconsistancies in the data. Data transformation is the process of converting data from one format or structure into another format or structure. It is primarily involves mapping how source data element will be changed for the destination.

Machine learning uses so called features (i.e. variables or attributes) to generate predictive models. Using a suitable combination of features is essential for obtaining high precision and accuracy. Because too many (unspecific) features pose the problem of overfitting the model, we generally want to restrict the features in our models to those, that are most relevant for the response variable we want to predict. Using as few features as possible will also reduce the complexity of our models, which means it needs less time and computer power to run and is easier to understand. Here select the features on the basis of Principal Component Analysis(PCA).

After feature selection dataset is split into training data and test data. Apply training dataset to the K-NN or DT algorithms. Applying the K-NN and DT algorithms generate the machine learning model. Send test data to the model and first check the performance. Identify which algorithm is best for our system and then enter new test data to that model and predict the cancer stage.

3.Working

Cancer is difficult to diagnose at early stages until it comes to stage III and IV. If cancer diagnose at stage III and IV, it will very dangerous to human life,many time human will be death. Nowadays, breast cancer can be found in both men and women, but it is more common for women. Breast cancer is one of the popular and second leading cause of cancer death in women. It is important to all people especially women to be aware of sign and symptoms of breast cancer. So this cancer prediction system which take a symptoms from human and predict cancer risk and it can also help the patient to the predict the breast cancer risk.

This system is expected to give an accurate prediction about the cancer based on the symptoms that user have enter the details that have be given. This system also expected to give accuracy of algorithm.

Algorithms:

1. KNN Algorithm (k-Nearest Neighbors algorithm):

The k-Nearest Neighbors algorithm is an easy algorithm to understand and to implement, and a powerful tool to have at your disposal. The model for KNN is the entire training dataset. When a prediction is required for a unseen data instance, the KNN algorithm will search through the training dataset for the k-most similar instances. The prediction attribute of the most similar instances is summarized and returned as the prediction for the unseen instance.

K-Nearest Neighbor Algorithm:

1. Calculate "d(x, xi)" i =1, 2, ..., n; where d denotes the Euclidean distance be-

tween the points.

2. Arrange the calculated n Euclidean distances in non-decreasing order.

3. Let k be a +ve integer, take the first k distances from this sorted list.

4. Find those k-points corresponding to these k-distances.

5. Let ki denotes the number of points belonging to the ith class among k points i.e. k 0

6. If ki¿kj i j then put x in class i.

2. Decision Tree Algorithm:

Decision Tree algorithm belongs to the supervised learning algorithms. decision tree algorithm can be use regression and classification problems. The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data(training data).

Decision Tree Algorithm: Pseudocode

1. Place the best attribute of the dataset at the root of the tree.

2. Split the training set into subsets. Subsets should be made in such a way that each subset contains data with the same value for an attribute.

3. Repeat step 1 and step 2 on each subset until you find leaf nodes in all the branches of the tree.

This system will focus on Registered User, Admin and System Scope:

Registered User:

- The user that want to check or predict the health status on cancer based on the symptoms that they have.
- The user need to register/sign up to be a member and then login to access the system.
- Do prediction of cancer using the system.

Admin:

- The person who will coordinate this system and update the system based on situation.

System:

Login Module:

a) There is a registration and login access for user and admin to access this system.

Evaluation Module:

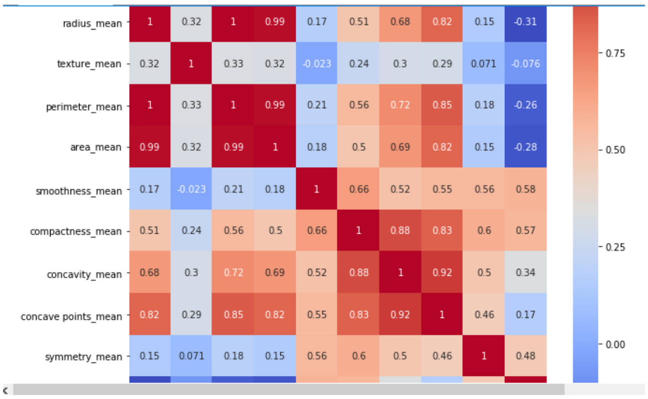
Registered User will answer and evaluate the questionnaires based on what this system provide to find out the result.

Domain System (cancer)

The result will generate based on the answer from Registered User and analyze it with Decision tree algorithm and KNN algorithm.

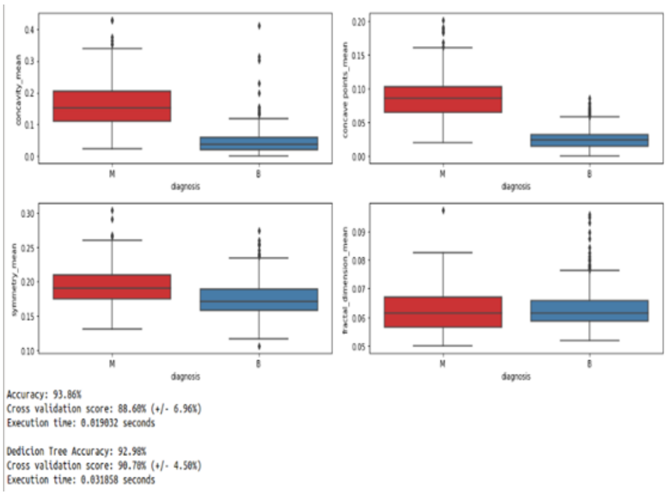
4. Experimental Study

The data has 569 diagnosis and 357malignant,212 benign.



radius_mean	1	0.32	1	0.99	0.17	0.51	0.68	0.82	0.15	-0.31
texture_mean	0.32	1	0.33	0.32	-0.023	0.24	0.3	0.29	0.071	-0.076
perimeter_mean	1	0.33	1	0.99	0.21	0.56	0.72	0.85	0.18	-0.26
area_mean	0.99	0.32	0.99	1	0.18	0.5	0.69	0.82	0.15	-0.28
smoothness_mean	0.17	-0.023	0.21	0.18	1	0.66	0.52	0.55	0.56	0.58
compactness_mean	0.51	0.24	0.56	0.5	0.66	1	0.88	0.83	0.6	0.57
concavity_mean	0.68	0.3	0.72	0.69	0.52	0.88	1	0.92	0.5	0.34
concave points_mean	0.82	0.29	0.85	0.82	0.55	0.83	0.92	1	0.46	0.17
symmetry_mean	0.15	0.071	0.18	0.15	0.56	0.6	0.5	0.46	1	0.48

Fig:2



```
Accuracy: 93.86%
Cross validation score: 88.60% (+/- 6.96%)
Execution time: 0.019032 seconds

Dedicion Tree Accuracy: 92.98%
Cross validation score: 90.78% (+/- 4.50%)
Execution time: 0.031858 seconds
```

Fig:3

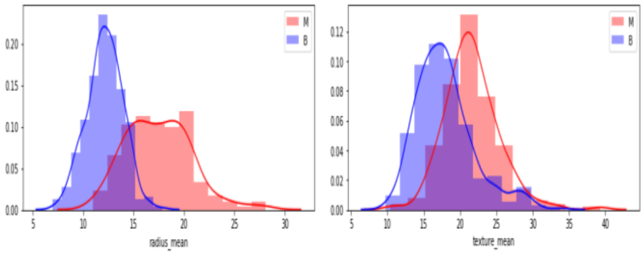


Fig:4

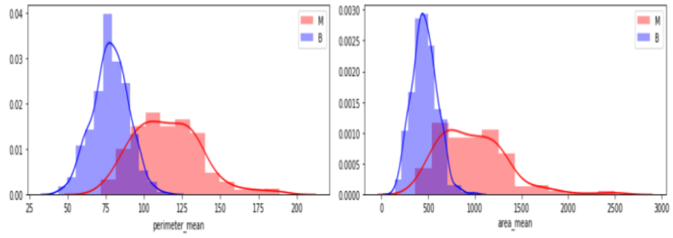


Fig:5

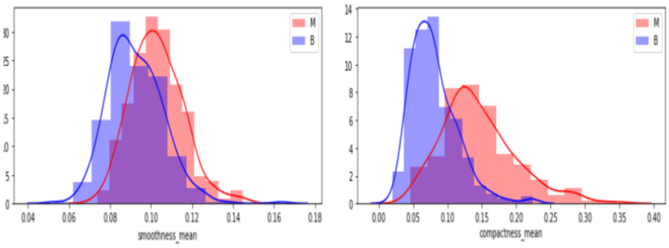


Fig:6

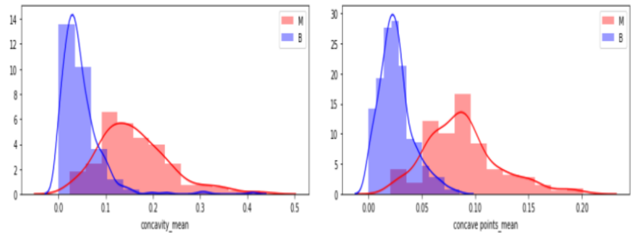
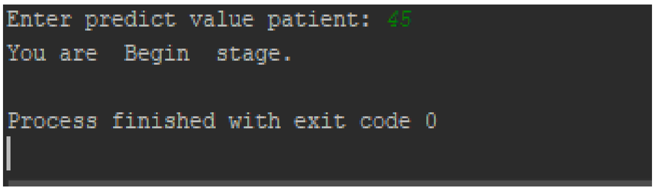
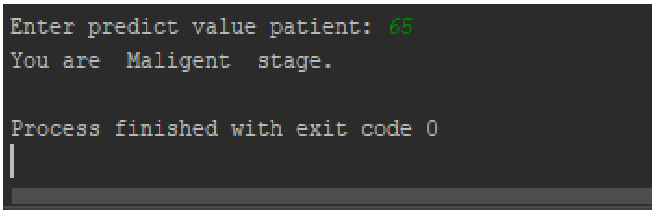


Fig:7



```
Enter predict value patient: 45
You are  Begin  stage.

Process finished with exit code 0
```



```
Enter predict value patient: 65
You are  Maligent  stage.

Process finished with exit code 0
```

5. Conclusion

The main purpose of this system is predict breast cancer disease by using machine learning approaches. Many time people may have different symptoms and it may not be detected easily whether they have breast cancer or not. This system can help in detecting breast cancer. This proposed system breast cancer-based on Changes in breasts for individual person to Determining cancer for human Body to prevent from major health risks.

References

[1] Vinitha S, Sweetlin S, Vinusha H and Sajini S "DISEASE PREDICTION USING MACHINE LEARNING OVER BIG DATA."

[2] Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang "Disease Prediction by Machine Learning over Big Data from Healthcare Communities."

[3] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas." Data Preprocessing for Supervised Leaning."

[4] A.Kousar Nikhath1, K.Subrahmanyam2, R.Vasavi3" Building a K-Nearest Neighbor Classifier for Text Categorization."

[5] T. Hannah Rose Esther , G. Kannan , Suresh Sagadavan , L. Sharmila "Intelligent and Effective Prediction of various diseases Prognosticating Naïve Bayes Classifier."

[6] S. Rasoul Safavian and David Landgrebe" A Survey of Decision Wee Classifier Methodology."

[7] PHILIP H. SWAIN, MEMBER,IEEE, AND HANS HAUSKA "The Decision Tree Classifier: Design and Potential."