Dogo Rangsang Research Journal
ISSN : 2347-7180
UGC Care Group I Journal
Vol-09 Issue-01 No. 01 : 2022

BREAST CANCER PREDICTION USING MACHINE LEARNING

Sk. Ahmed Mohiddin[1], T. Pooja Sri[2], K. Sharmila[3], P. Bhavya[4],

G. Chandra Gupta[5] . [1]Asst Professor.[2, 3, 4, 5] Students

Dept Of Computer Science And Engineering

Sri Vasavi Institute Of Engineering And Technology, Pedana, A. P, India

ABSTRACT:

Machine learning is a technology which allows a software program to become more accurate at pretending more accurate results without being explicitly programmed and also ML algorithms uses historic data to predict the new outputs.

Because of this ML gets a distinguish attention. Now a day's prediction engine has become so popular that they are generating accurate and affordable predictions just like a human, and being using industry to solve many of the problems.

Breast cancer is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today's society. The early diagnosis can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Thus, the correct diagnosis of Breast cancer and classification of patients is the subject of much research. Because of its unique advantages in critical features detection from complex Breast Cancer Datasets, machine learning is widely recognized as the methodology of choice in Breast Cancer pattern classification and forecast modelling.

Keywords:

Machine Learning, Random Forest Classifier, Model Selection, Classification, Prediction, Supervised Learning

INTRODUCTION:

Breast cancer is a disease in which cells in the breast grow out of control. There are different kinds of breast cancer. The kind of breast cancer depends on which cells in the breast turn into cancer. Breast cancer can begin in different parts of the breast. A breast is made up of three main parts: lobules, ducts, and connective tissue. The lobules are the glands that produce milk. The ducts are tubes that carry milk to the nipple. The connective tissue (which consists of fibrous and fatty tissue) surrounds and holds everything together. Most breast cancers begin in the ducts or lobules. Breast cancer can spread outside the breast through blood vessels and lymph vessels. When breast cancer spreads to other parts of the body, it is said to have metastasized.

Machine learning is one of the applications of artificial intelligence (AI) that provides computers, the ability to learn automatically and improve from experience instead of explicitly programmed. It focuses on developing computer programs that can access data and use it to learn from themselves. The main aim is to allow computers to learn automatically without human intervention and also adjust actions accordingly. Breast

DOI : 10.36893.DRSR.2022.V09I01N01.882-891

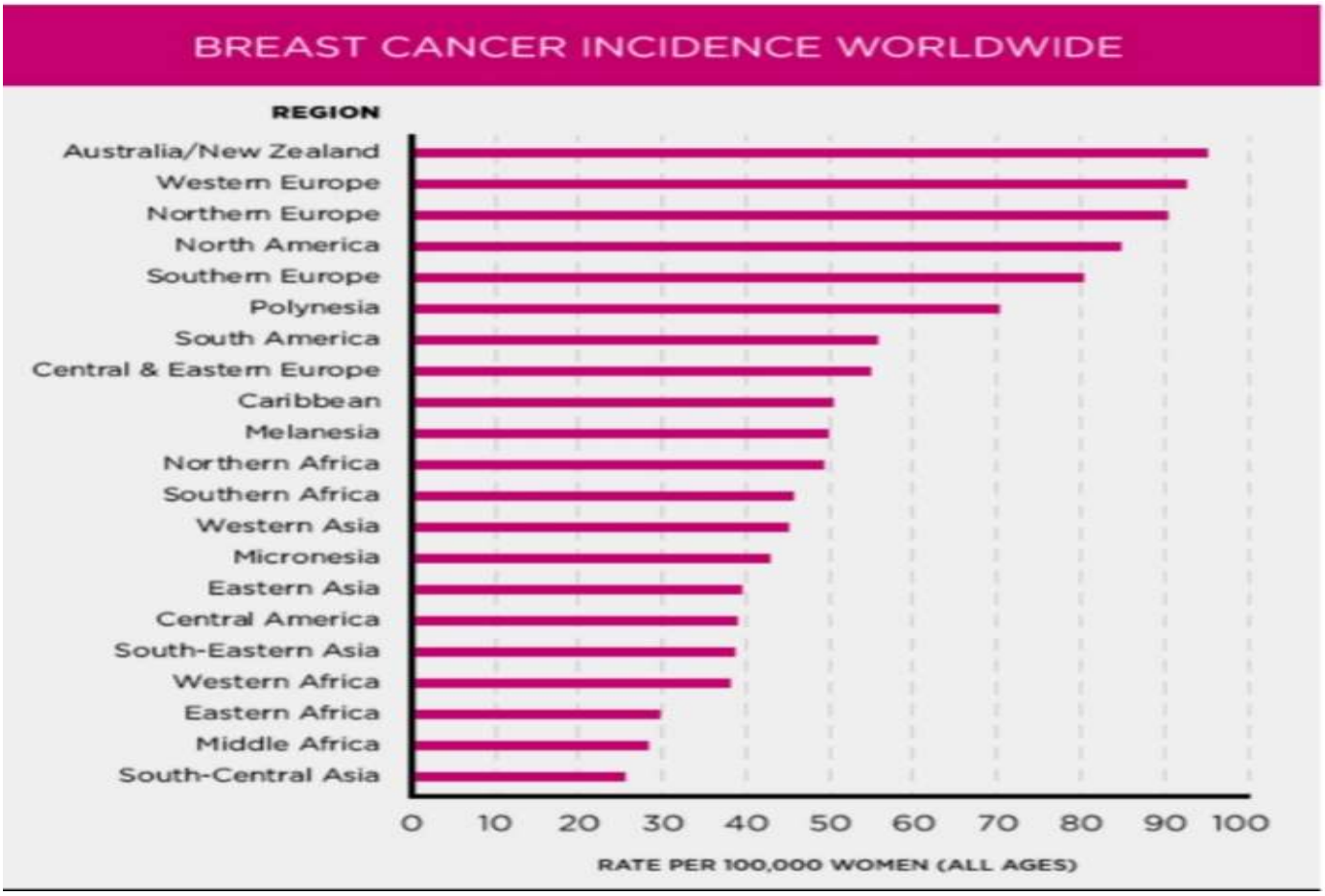
cancer has a great deal of attention in medical research. The diagnosis of breast cancer is a challenging task, which can offer automated prediction about the condition of patient so that further treatment can be made effective. Classification algorithms are very important category of supervised machine learning algorithms. These algorithms require a very large training set. These training data sets are consisting of many features or attributes which describe the individual sample. Since we are doing supervised learning algorithm. All the training set are labelled correctly. The classification algorithms such as Decision Tree, Naive Bayes, Random Forest Classifier, K-Nearest Neighbours, Logistic Regression and Support Vector Count (SVC), develop model with these data with many different parameters. When we have a new unlabelled sample, we can use the model to predict the label of the new sample. These techniques are used for disease diagnosis to help doctor to effectively label the new case.

LITERATURE SURVEY:

Statistics show a breast cancer prevalence rate of 25.8 women per 100000 women in India. Considering the population of 1.3 billion Indians with half female population, this amounts to almost 350000 women living with breast cancer in India for all ages combined as per a study done in 2017. This is projected to increase to a staggering 1800000 by early 2020 as per predictions as there is increasing awareness and detection in the population. Thus, this shows us the scale and magnitude of the disease and is the number 1 cancer among women in India surpassing cervical cancer with 25 – 32% of all cancers affecting women being breast cancers. Thiruvananthapuram, Chennai, Delhi and Mumbai have shown the highest prevalence of breast cancer in the Country. A few other worrying statistics have shown that the incidence of breast cancer in populations below 50 years of age has doubled as compared to 25 years ago along with an increase in mortality (risk of death) of breast cancer by approximately 13% as compared to 5 years back. Almost 45-50 % of patients in India present in the advanced stages (Stage 3 and 4) as per another study. All these facts and figures entail that India still has enormous scope to improve the detection and quality of treatment that can be given to breast cancer patients.

Machine learning products Madhu Kumari and Virendra Singh [3] used many approaches for detecting breast cancer by means rough sets and Back propagation neural network, Feature extraction using a hybrid of k-means and SVM, LR Neural Network and Decision tree. Highest accuracy is 98.6% for rough sets and back propagation. Hiba Asria, Hajarmousannifb, Hassan Al Moatassimec, Thomas Noeld [4] compared 4 machine learning algorithms they are support vector machine, Naïve Bayes, k-nearest neighbour’s, decision tree. Highest accuracy is 97.3% for support vector machine.

DOI : 10.36893.DRSR.2022.V09I01N01.882-891



PROPOSED:

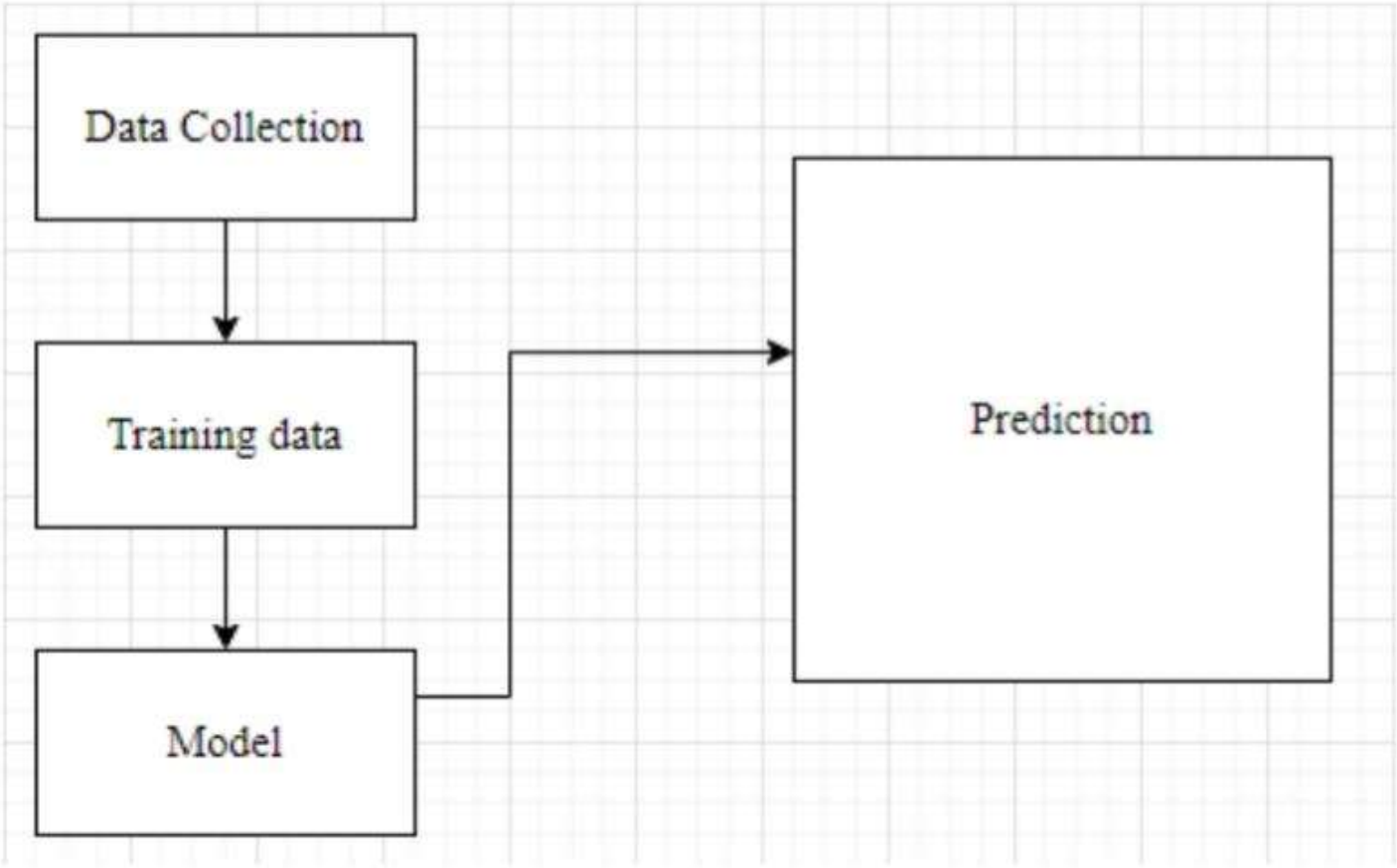
To identify breast cancer symptoms at early stage to save someone's life by using data mining techniques and machine learning models on dataset.

Advantages:

1. Generate accurate and efficient results.
2. Computation time is greatly reduced.
3. Reduces manual work.
4. Efficient further treatment.
5. Automated prediction.

SYSTEM ARCHITECTURE:

DOI : 10.36893.DRSR.2022.V09I01N01.882-891



DATA SET:

DOI : 10.36893.DRSR.2022.V09I01N01.882-891

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1		mean radi	mean text	mean per	mean are	mean smc	mean cor	mean con	mean con	mean sym	mean frac	radius err	texture er	perimeter	area error	sm
2	0	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.0
3	1	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.0
4	2	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0
5	3	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	0
6	4	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0
7	5	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	0
8	6	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0.0
9	7	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0.0
10	8	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0.0
11	9	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039	23.94	0.0
12	10	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466	40.51	0.0
13	11	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9849	3.564	54.16	0.0
14	12	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	11.07	116.2	0.0
15	13	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078	2.903	36.58	0.0
16	14	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	2.061	19.21	0.0
17	15	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	0.37	1.033	2.879	32.55	0.0
18	16	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	0.4727	1.24	3.195	45.4	0.0
19	17	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	0.5692	1.073	3.854	54.18	0.0
20	18	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	0.7582	1.017	5.865	112.4	0.0
21	19	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766	0.2699	0.7886	2.058	23.56	0.0
22	20	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811	0.1852	0.7477	1.383	14.67	0.0
23	21	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905	0.2773	0.9768	1.909	15.7	0.0
24	22	15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032	0.4388	0.7096	3.384	44.91	0.0
25	23	21.16	23.04	137.2	1404	0.09428	0.1022	0.1097	0.08632	0.1769	0.05278	0.6917	1.127	4.303	93.99	0.0

IMPLEMENTATION:

- Data Collection.
- Data Preprocessing.
- Model Selection.
- Train the model.
- Evaluating model.
- Making Predictions.

Data Collection:

- The dataset used for the project is Wisconsin Breast Cancer Dataset collected from UCI repository.
- https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)
- Attributes in the dataset are:

1. Clump Thickness
2. Uniformity Cell Size
3. Uniformity Cell Shape
4. Marginal Adhesion
6. Bare Nuclei
7. Bland Chromatin
8. Normal Nucleoli
9. Mitoses

DOI : 10.36893.DRSR.2022.V09I01N01.882-891

5. Single Epithelial Cell Size 10.ClassLabel

Data Preprocessing:

- Loading the dataset and training the data to predict the result is called Data preparation
- Cleaning the data by filling missing data comes under Data training.
- Many Preprocessing techniques like Correlation, SMOTE, Filling missing values are used in data preparation.

Model Selection:

- Model selection is the process of choosing between different machine learning approaches - e.g. SVM, logistic regression, etc. - or choosing between different hyperparameters or sets of features for the same machine learning approach. Some of the models used in this project are:

1. Random Forest Classifier
2. Logistic Regression
3. Gaussian Naïve Bayes
4. SVC
5. K-Nearest Neighbors
6. Decision Tree

Train the model:

- Creating a train and test split of your dataset is one method to quickly evaluate the performance of an algorithm on your problem. The training dataset is used to prepare a model, to train it. We pretend the test dataset is new data where the output values are withheld from the algorithm.
- In Breast Cancer Detection project we have spitted original data into two Datasets. Training Data with 70% of original data, testing data with 30% of original data.

Evaluating Model:

- Model evaluation aims to estimate the generalization accuracy of a model on future (unseen) data.
- Methods for evaluating Model's performance are divided into 2 categories: Holdout and Cross-valida tion. Both methods use a test set (i.e. data not seen by the model) to evaluate model performance.

Comparison of Algorithms:

Comparison of algorithms by using K-Fold

DOI : 10.36893.DRSR.2022.V09I01N01.882-891

	Model	Train_Score	Test_Score	Accuracy
3	Random Forest	100.00	100.00	100.000000
0	Gaussian Naive Bayes	97.37	99.12	99.115044
4	Logistic Regression	94.96	99.12	99.115044
1	Decision Tree	96.71	98.23	98.230088
5	K-Nearest Neighbours	96.71	98.23	98.230088
2	Support Vector Machine	59.65	78.76	78.761062

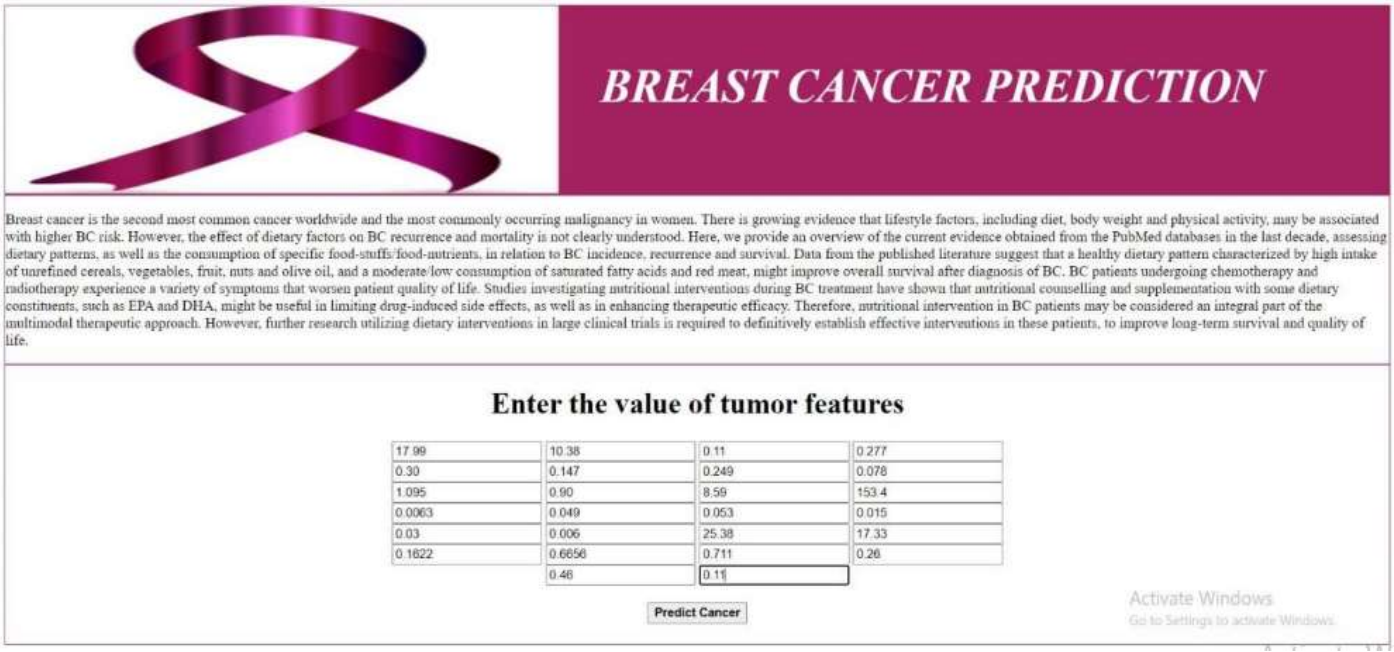
RESULTS:

HOME PAGE:

DOI : 10.36893.DRSR.2022.V09I01N01.882-891

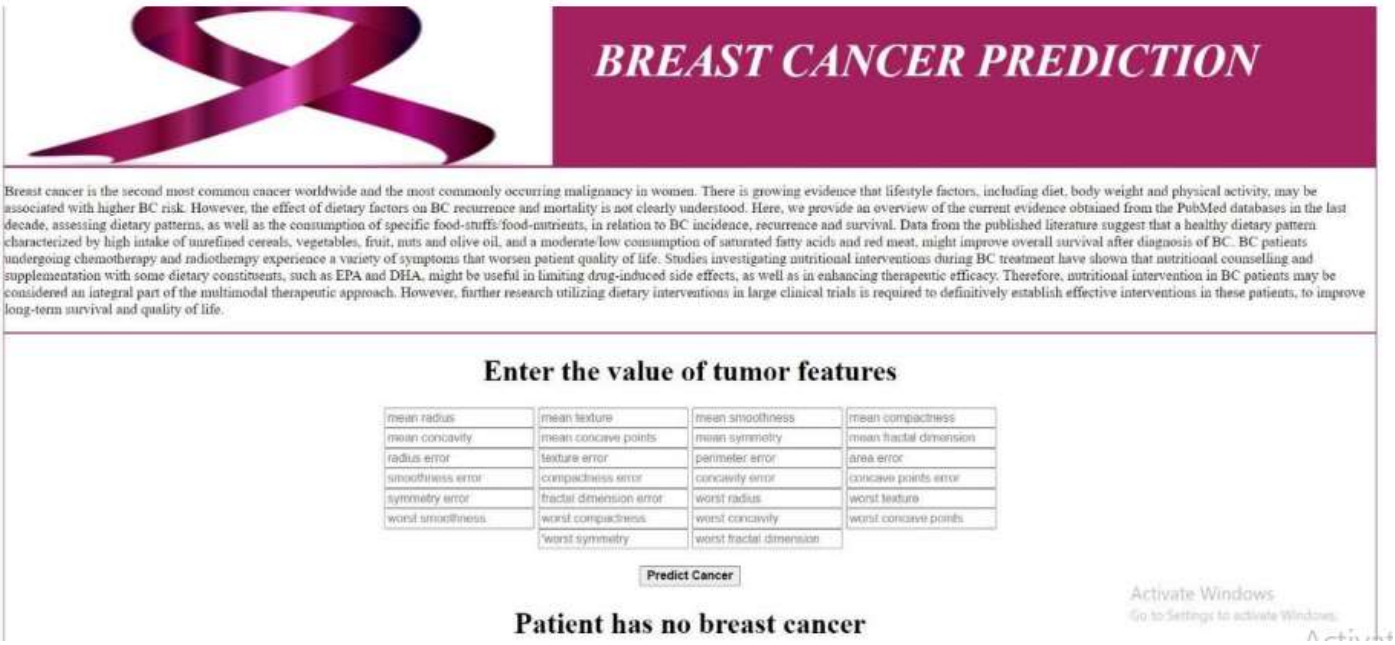
Has Breast Cancer:

Test Case-1:



DOI : 10.36893.DRSR.2022.V09I01N01.882-891

Has No Breast Cancer:



CONCLUSION:

We have used 6 algorithms like Decision Trees, Random Forests, Naive Bayes, SVC, KNN and Logistic Regression in-order to predict presence or absence of breast cancer using SMOTE Technique. The accuracy varies for different algorithms. The accuracy for Decision tree algorithm is 89.13%. The accuracy for Random Forest algorithm is 96.73%. The accuracy for Naive Bayes algorithm is 93.80%. The accuracy for KNN algorithm is 93.80%. The accuracy for Logistic Regression algorithm is 79.34%. The accuracy for SVM algorithm is 93.80%. The highest accuracy is given when we have used Random Forest algorithm using K-Fold Technique which is nearly 100.00%.

REFERENCES:

[1] Jemal A, Murray T, Ward E, Samuels A, Tiwari RC, Ghafoor A, Feuer EJ, Thun MJ. Cancer statistics, 2005. CA: a cancer journal for clinicians. 2005 Jan 1; 55(1):10-30.

[2] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999– 2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centres for Disease Control

[3] MadhuKumaria, VijendraSinghb Department of Computer Science and Engineering, The North Cap University, Sector 23A,Gurugram, Haryana, 2017, India.

[4] Hiba Asria, Hajarmousannifb, Hassaan Al Moatassimec, Thomas Noeld. The 6th International Symposium on Frontiers in Ambient and Mobile System (FAMS 2016).

[5] "UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) DataSet."[Online].Available:'https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29'.

DOI : 10.36893.DRSR.2022.V09I01N01.882-891

[6] Step 7 - Science Fair 2012: Project Sites.google.com, 2020. [Online].

[7] Data Pre-processing- an overview | ScienceDirect Topics, Sciencedirect.com

[8] Handling Missing Values in machine learning: part1, @george.drakos62.

[9] Machine Learning | Handling Imbalanced Data with SMOTE and Near Miss Algorithm in Python - GeeksforGeeks

[10] D.Lavanya, Dr.K.Usha Rani, Analysis of feature selection with classification: Breast cancer datasets",Indian Journal of Computer Science and Engineering (IJCSE),October 2011.

[11] In-Database Machine Learning 2: Calculate a correlation Matrix – A Data Exploration Post | Vertica

[12] D.Lavanya, Dr.K.Usha Rani,..," Analysis of feature selection with classification: Breast cancer datasets",Indian Journal of Computer Science and Engineering (IJCSE),October 2011.

[13] Logistic Regression for Machine Learning, machinelearningmastery.com

[14]Zhang,J.;Zulkernine,M.;Haque,A.Random-forestsbasednetworkintrusiondetectionsystems. IEEETrans.Syst. Man Cybern. 2008, 38, 649–659.

[15] Rish I. An empirical study of the naive Bayes classifier. IJCAI Work Empir methods ArtifIntell. 2001; 3(November):41-46.

[16]Noble WS. What is a support vector machine? Nat Biotechnol. 2006; 24(12):1565-1567. Doi: 10.1038/nbt1206-1565.

[17] Introduction to KNN, K-Nearest Neighbours: Simplified, analyticsvidhya.com

[18]Quinlan JR. C4.5: Programs for Machine Learning; 2014:302. https://books.google.com/books?hl=fr&lr=&id=b3ujBQAAQBAJ&pgis=1. Accessed January 5, 2016.

[19] Cross-validation: evaluating estimator performance – scikit-learn 0.22.2

[20] J. Han and M. Kamber,"Data Mining Concepts and Techniques", Morgan Kauffman Publishers, 2000

DOI : 10.36893.DRSR.2022.V09I01N01.882-891