International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 05 | May - 2021 ISSN: 2582-3930

Breast Cancer prediction using SVM Algorithm in Machine Learning

Janhavi Armarkar[1], Ayushi Jain[2], Mahima Hatewar[3], Mansi Bawne[4], Dhanashri Chopkar[5]

Dr. Vivek Deshmukh[6]

[1, 2, 3, 4, 5](*Student, Dept. of Electronics & Telecommunication Engineering,* SB Jain Institute of Technology Management*& Research, Nagpur, Maharashtra, India)*

[6](*Associate Professor, Dept. of Electronics & Telecommunication Engineering, SB Jain Institute Management & Research, Nagpur, Maharashtra, India)*

Abstract: Cancer is the 2nd source of death in the world. The main reason for this increased death rate is the delayed detection of cancerous tissue growth in a person. Nearly 60% of patients with breast cancer are diagnosed in advanced stages. The main objective of our paper is to enhance an image processing algorithm for an earlier finding of breast cancer. X-ray mammogram images that have been acquired are used as input images.

Keywords:*'data', 'target', 'frame', 'target names','DESCR','feature_names','filename'*

Introduction: A major health issue that arises these days had led to much advancement in the medical field, yet certain diseases remain incurable. Certain diseases even seem to be more challenging to guess what it is in the earlier stages. Among these many diseases (stroke, cancer, heart attack, viral hepatitis, chronic liver diseases, coronary artery disease, etc.), the death rate due to cancer is becoming more and more each year. Once tissue is affected by radiation, it takes almost 15 years to turn into cancerous tissue. Even though we have this time for perfect curability, many of the patients who have thiskind of basic abnormalities don't take it as a big deal since the symptoms are notindicating cancerous growth. Nearly 60% of patients with breast cancer are diagnosed in advanced stages. The growing cancer burden is due to several factors, including residence growth and aging as well as the changing commonness of certain causes of cancer connected to social and economic progress.

Methodology:

Data Reading: - Data preparation may be one of the most difficult steps in any machine learning project. The reason is that each dataset is different and highly specific to the project. Nevertheless, there are enough commonalities across predictive modelling projects that we can define a loose sequence of steps and subtasks that you are likely to perform. This process provides a context in which we can consider the data preparation required for the project, informed both by the definition of the project performed before data preparation and the evaluation of machine learning algorithms performed after.

Data pre-processing: - Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.When creating a machine learning project; it is not always a case that we come across clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put it in a formatted way. So for this, we use data pre-processing tasks.

- **Model building:-**

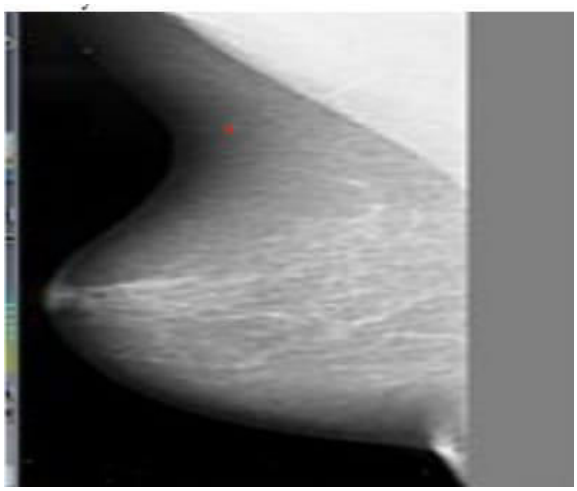
Machine learning consists of algorithms that can automatic analytical model building. Using algorithms that iteratively learn from data, machine learning models facilitate computers to findhiddeninsights from Big Datawithout being explicitly programmed where to look. This has given rise to a plethora of applications based on Machine learning.

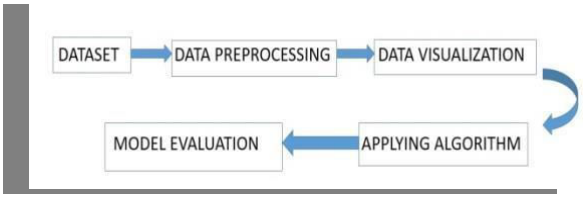
- **Model evaluation:-**

The above issues can be handled by evaluating the performance of a machine learning model, which is an integral component of any data science project. The model evaluation aims to estimate the generalization accuracy of a model on future (unseen/out-of-sample) data. Methods for evaluating a model'sperformance are divided into 2 categories: namely, holdoutand Cross-validation.Both methods use a test set (i.e. data not seen by the model) to evaluate model performance. It's not recommended to use the data we used to build the model to evaluate it. This is because our model will simply remember the whole training set, and will therefore always predict the correct label for any point in the training set. This is known as over fitting.

Proposed Work:

Block Diagram of proposed work





Machine learning: - Machine Learning is a process that machines (computers) are trained with data to decide for similar cases. ML is employed in various applications, such as object recognition, network, security, and healthcare. [1] There are two ML types i.e. single and hybrid methods like ANN, SVM, Gaussian Mixture Model (GMM), K-Nearest Neighbor (KNN), Linear Regressive Classification (LRC), Weighted Hierarchical Adaptive Voting Ensemble (WHAVE), etc.

Firstly we understood the basic libraries and its basic use in these algorithms: - # import libraries import

pandas as pd: - for data manipulation or analysis

import numpy as np: - for numeric calculation import matplotlib.pyplot as plt: for data visualization

import seaborn as sns: - for data visualization.

Then we understood the basic concepts of data analysis they are given as below:

- **Data load: -** Data loading refers to the "load" component of ETL. After data is retrieved and combined from multiple sources (extracted), cleaned and formatted (transformed), it is then loaded into a storage system, such as a cloud data warehouse.

- **Data manipulation: -** Data manipulation can be interpreted as the process of changing data in order to make it easier to read or be more organized. For an instance, the information in the data

becomes easier to locate by arranging its log of data in alphabetical order thereby by presenting individual entries for each of it.
Also Data manipulation applications are also used in the websites to help the owners view their most popular pages

Create data frame: - A data frame is a table or a two-dimensional array-like structure in which each column contains values of one variable and each row contains one set of values from each column.

Following are the characteristics of a data frame:

- The column names should be nonempty.
- The row names should be unique.
- The data stored in a data frame can be of numeric, factor or character type.
- Each column should contain same number of data items.

```
In [15]: # Head of cancer DataFrame
         cancer_df.head(20)
```

Out[15]:

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst texture	worst perimeter	worst area	worst smoothness
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	0.2419	0.07871	...	17.33	184.60	2019.0	0.162
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	0.1812	0.05667	...	23.41	158.80	1956.0	0.123
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	0.2069	0.05999	...	25.53	152.50	1709.0	0.144
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	0.2597	0.09744	...	26.50	98.87	567.7	0.209
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	0.1809	0.05883	...	16.67	152.20	1575.0	0.137
5	12.45	15.70	82.57	477.1	0.12780	0.17000	0.15780	0.08089	0.2087	0.07613	...	23.75	103.40	741.6	0.179
6	18.25	19.98	119.60	1040.0	0.09463	0.10900	0.11270	0.07400	0.1794	0.05742	...	27.66	153.20	1606.0	0.144
7	13.71	20.83	90.20	577.9	0.11890	0.16450	0.09366	0.05985	0.2196	0.07451	...	28.14	110.60	897.0	0.165
8	13.00	21.82	87.50	519.8	0.12730	0.19320	0.18590	0.09353	0.2350	0.07389	...	30.73	106.20	739.3	0.170
9	12.46	24.04	83.97	475.9	0.11860	0.23960	0.22730	0.08543	0.2030	0.08243	...	40.68	97.65	711.4	0.185

Data visualization:-

It's hard to think of a professional industry that doesn't benefit from making data more understandable. Every STEM field benefits from understanding data—and so do fields in government, finance, marketing, history, consumer goods, service industries, education, sports, and so on.

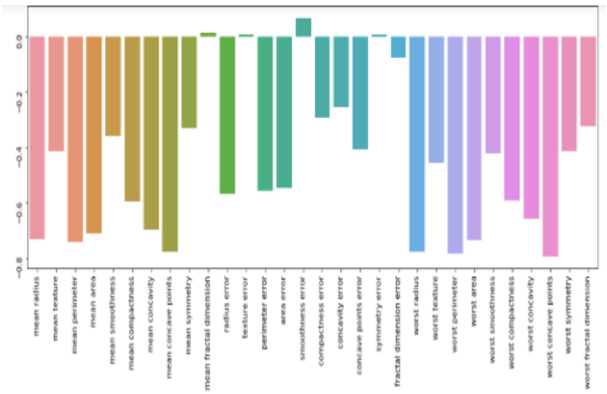
While we'll always wax poetically about data visualization (you're on the Tableau website, after all) there are practical, real-life applications that are undeniable. And, since visualization is so prolific, it's also one of the most useful professional skills to develop. The better you can convey your points visually; whether in a dashboard or a slide deck, the better you can leverage that information.

The concept of the citizen data scientist is on the rise. Skill sets are changing to accommodate a data-driven world. It is increasingly valuable for professionals to be able to use data to make decisions and use visuals to tell stories of when data informs who, what, when, where, and how. While traditional education typically draws a distinct line between creative storytelling and technical analysis, the modern professional world also values those who can cross between the two: data visualization sits right in the middle of analysis and visual storytelling.

The correlation graph: - Correlation plots help you to visualize the relationships between a set of quantitative variables by displaying their correlations using color or shading.

The ggcorplot function in the ggcorrplot package can be usedto visualize these correlations. By default, it creates a ggplot2 graph were darker red indicates stronger positive correlations, darker blue indicates stronger negative correlations and white indicates no correlation.

Corr can compute the correlations between variables, which are stored in matrix form in the output data set.



Splitting data into test and train data:-

As can be seen in the screenshot below, the data is located in the generated data folder.

We also want to save the train and test data to this folder, once these files have been created. Additionally, the script runs in the prepare_ml_data.py file which is located in the prepare_ml_data folder.

Splitting and saving: - Now, we have the data ready to split it. Luckily, the train_test_split function of the sklearn library is able to handle Pandas Data frames as well as arrays. Therefore, we can simply call the corresponding function by providing the dataset and other parameters, such as following:

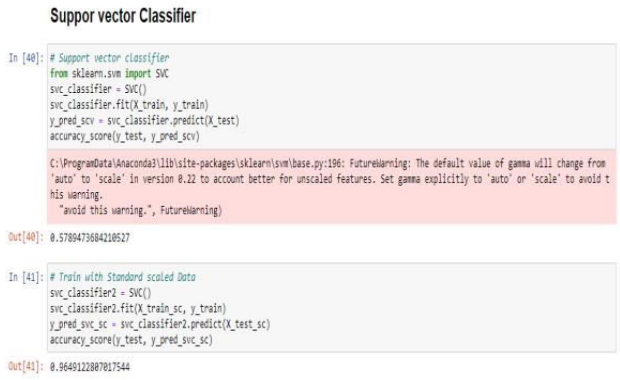
- **Test_Size**: This parameter represents the proportion of the dataset that should be included in the test split. The default value for this parameter is set to 0.25, meaning that if we don't specify the test_size, the resulting split consists of 75% train and 25% test data.
- **Random_State**: This parameter controls the shuffling applied to the data before the split. By defining the random state we can reproduce the same split of the data across multiple function calls.
- **Shuffle**: This parameter indicates whether the data should be shuffled before splitting. Since our dataset is ordered by genre, we definitely want to shuffle it. Otherwise the train and test set would not contain the same genres.

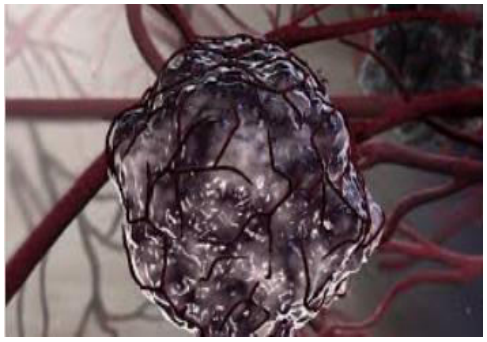
The corresponding data files can now be used to for training and evaluating text classifiers (depending on the model though, maybe additional data cleaning is required).

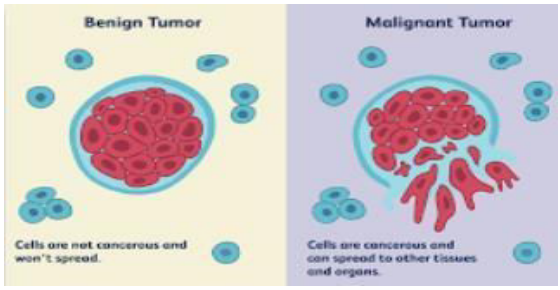
Feature scaling: - Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data preprocessing to handle highly varying magnitudes or values or units. If feature scaling isnot done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Finally, Support vector classifier: - Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

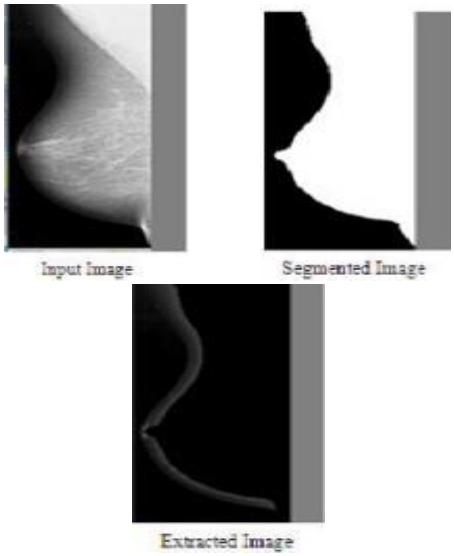
The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyper plane.







Result: The project we are dealing with is based upon machine learning and the main output of this project dependsupon accuracy. This accuracy also dependson the scalingofthe features (names of columns). Therefore, the accuracy of scaled data is far more than the unscaled data. Accuracy of unscaled data: - 56% Accuracy of scaled data: - 96% in this paper, detection of breast cancer was done using SVM algorithm. There are several other algorithms are also available such as ANN, DNN, KNN, etc. But as per the accuracy level is concerned, SVM provides a better result. The average accuracy was improved to 56% and the high accuracy changed to 96%.Finally, the simulated results show that the used methodologies provide a better classifier rate with a minimum error rate for all test samples.



Conclusion: The most important idea behind this project was to acquire an accurate result of the machine learning algorithms for the betterment of the medical sector and patients who are diagnosed with breast cancer. We have come to an agreement that the support vector machine algorithm has proven to be the best algorithm till now. But saying this we also know thatthe accuracy level varies from data sets and the featured scaling of the data. This study attempts to solve the problem of automatic detection of breast cancer using a machine learning algorithm.

REFERENCES:

1. "Latest Global Cancer Data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018", International Agency for Research on Cancer, World Health Organization, 12 September 2018.

2. M.M.Mehdy, E.E.Shair, and P.Y.Ng, "Artificial Neural Networks in Image Processing for Earlier Detection of Breast Cancer", Hindawi, Computational and Mathematical Methods in Medicine, Volume 2017, and Article ID 2610628.

3. Vishnukumar K.Patel, Prof.Syed Uvaid and Prof.A.C.Suthar, "Mammogram of Breast Cancer Detection Based Using Image Enhancement Algorithm", International Journal of Engineering Technology and Advanced Engineering, Volume 2, Issue 8, August 2012.

4. Melanie A. Sutton, "Breast Cancer Detection Using Image Processing Techniques", IEEE International Conference on Fuzzy System- February 2000.

5. Bhagyashri k Yadav, Dr. Prof. M. S. Panse, "Virtual Instrumentation Based Breast Cancer Detection and Classification Using Image-Processing", International Journal of Research and Scientific Innovation (IJRSI), Volume V, Issue IV, April 2018.