Identification and Augury of Chronic Disease by Logit Regression and Machine Learning

Radhesh Pandey and Kamal Srivastava

May 22, 2022

Identification and augury of Chronic Disease by Logit Regression and Machine Learning

Radhesh Pandey, Kamal Kumar Srivastava

Department Of Information Technology, ,Babu Banarasi Das Northern India Institute of Technology
Lucknow, Uttar Pradesh
[1]radhesh185@gmail.com
[2]2007.srivastava@gmail.com

Abstract— **Every year, a lot of deaths occurred from different chronic diseases in which Cancer is one of the most harmful diseases. It is the most common type diseases who are leading cause of death in the people worldwide. Cancer happens only when abnormal cells produce and replicate in an uncontrolled way in a explicit part of the body. These cancer cells can invade and destroy surrounding healthy tissue, including organs. The prediction and diagnosis of chronic diseases are challenge for healthy life and researcher on early stage of Cancer also. Therefore, high accuracy in predicting chronic disease is more important for treatment in all aspects for the patients. Here we use the Machine learning Concept and its significant contribution to the predicting and early detection of type of chronic disease (Cancer). Breast cancer is the second most common cancer in women after skin cancer. Cancer occurs as a result of mutations, or abnormal changes, in the genes responsible for regulating the growth of cells and keeping them healthy. To detect it on early stages we have use machine learning algorithms, Support Vector Machine (SVM), Random Forest, Logistic Regression, and Logit Regression, for breast cancer. Using this we are able to predict and diagnose breast cancer on early stages and find the effectiveness in terms of confusion matrix, accuracy, and accuracy. With the help of Logit regression the outperform gets accuracy (93.2%).**

Keywords— Logit Regression, Prediction, Malignant, Benign.

I. Introduction

Here we are talking about cancer. Unfortunately, people who have had one cancer are more probable to get a second cancer, which may be the same or different to their first cancer. Chemotherapy and radiotherapy further increase this risk, if it has not observed very carefully. It has been considered carefully when your initial treatment is planned. That's why many researchers are working on this research and we have also tried to find the breast cancer prediction on early stage so that many lives could be saved.

Breast cancers are the most common cancers in women and the second main motive of loss of life for women after lung cancers. Breast cancers have surpassed lung cancers with a predicted 2.3 million more breast cancer suffers people (11.7%), observed through lung (11.4%). Breast cancer is caused by a genetic abnormality However, only 5-10% of cancers are due to an abnormality inherited from your mother or father. Instead, 85-90% of breast cancers are due to genetic abnormalities that happen as a result of the aging process and the "wear and tear" of life in general. Doctors can effortlessly stumble on breast cancer using Breast ultrasound, Magnetic resonance imaging (MRI), Biopsy. Primarily based on these test effects, the health practitioner may also suggest in addition tests or treatments. Beside this Early detection may play an important role for breast cancers identification and their cure on early stage. If the danger of cancer is anticipated then the probabilities of survival for the patient may also increase. Another manner to diagnose breast cancers is to apply machine learning algorithms to expect an odd tumour. Therefore, studies become performed to determine the ideal diagnosis and type of patients into malignant and benign classes.[1]

The 2 most important kinds of cell characteristics are as follows: -

a. Genuine cancers cells-loose cells can not spread or broaden very slowly. And if doctors dispose of them, they cannot motive any damage to the human body.[5]

b. Unstable cells: - are cancerous and might unfold unexpectedly with inside the frame.[5]

Breast cancers cells typically form a tumour that is not often visible at the x-ray or that appears as a lump. even as it spreads outside the breast through the blood vessels and lymph nodes, it will become a prime breast cancers. Survival prices of breast cancer are very excessive if the cancer is not recognized early.

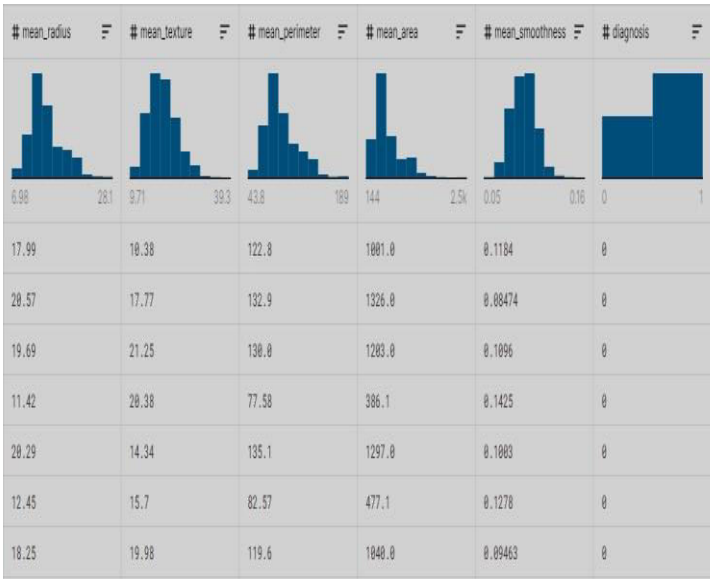
To deal with the growing burden of breast cancers, it's miles crucial to make improvement in getting access to early detection, nicely timed treatment and care of breast cancer now become most important factor for survival of cancerous .[1]

II Features of cancer tumours

The foremost functions that have been related to breast most cancers tumours are:

- Genetic
- size of the tumour
- Radius
- Concavity

For early detection of the breast cancer we have taken dataset for the evaluation which has been taken from the Kaggle. It consists of a complete of 6 columns and 570 rows. Analysis of the data and the final results that represents 0 as Benign and 1 as Malignant.[5]



# mean_radius	# mean_texture	# mean_perimeter	# mean_area	# mean_smoothness	# diagnosis
17.99	10.38	122.8	1001.0	0.1184	0
20.57	17.77	132.9	1326.0	0.08474	0
19.69	21.25	130.0	1203.0	0.1096	0
11.42	20.38	77.58	386.1	0.1425	0
20.29	14.34	135.1	1297.0	0.1003	0
12.45	15.7	82.57	477.1	0.1278	0
18.25	19.98	119.6	1040.0	0.09463	0

Table.1

3) METHODOLOGY

1) Problem Statement

The problem statement is to detect breast most cancers from the tumour cell at an early level to save a person's life. It predicts whether or not or now no longer the tumour cells are Malignant (Cancerous) or Benign (Non-cancerous) with the assist of a historical dataset.

2) Technology Used:-

•Machine Learning is used to pre-process the dataset using the Supervisedlearning method

•Jupyter Notebook is used because the platform for evaluation.

•Python language is used as it is taken into consideration quality for machine learning knowledge of responsibilities.

•For growing an powerful person interface Tkinter library is used.

•representation of graphs is finished by using the use of Matplotlib.

•Pandas and Numpy for getting access to and clean manipulation of raw_dataset.

•Logit Regression

Use of Supervised learning and historical datasets:-

We've used supervised learning as it may work based on historical and categorized datasets. in the case of breast cancer prediction, these datasets are preferred, as the system can advantage as a whole lot experience on these datasets in place of present or any short time period datasets, because short time period or small datasets can't be able to offer a hundred percentage precision and accuracy in prediction and case of cancers, precision and accuracy is the soul.

The Historical datasets which includes a big amount of information have probable higher possibilities to provide effects with better accuracy and precision than short-term datasets. Also within the case of cancer prediction, there is a minor differences between tumours and cancerous. Some outside parameters are usually wished that's why right here the usage of deep learning or unsupervised learning is not efficient in phrases of results and schooling purposes. [9]

Machine Learning:-

Machine learning is an application of synthetic intelligence (AI) that gives systems the capacity to routinely learn and enhance from experience without being explicitly programmed. Machine learning makes a speciality of the improvement of computer programs that can get entry to statistics and use it to examine for themselves. The manner of learning begins with observations or information, which includes examples, direct experience, or training, to look for patterns in information and make higher selections inside the future based totally on the examples that we provide.
here we are using the machine learning approach to create and teach the model with the intention to itself be able to predict whether the tumour is benign or malignant primarily based on a few parameters.[7]

There are three subclasses of machine learning:

1) Supervised learning
2) Unsupervised learning
3) Reinforcement learning

4) Logit Regression & Statsmodel.api

Stats model is a Python module that provides classes and functions for the estimation of many exclusive statistical models, as well as for engaging in statistical assessments and statistical facts exploration. An intensive listing of result facts is available for every estimator. The results are tested against current statistical programs to ensure that they're accurate.[4]

In case of our model the result of stats model's statistics is:

Optimization terminated successfully when Current function value: 0.127326 Iterations 10 .

Logit Regression Results

Dep. Variable:	diagnosis	No. Observations:	455
Model:	Logit	Df Residuals:	449
Method:	MLE	Df Model:	5
Date:	Thu, 23 Feb 2022	Pseudo R-squ.:	0.8056
Time:	19:09:18	Log-Likelihood:	-57.933
converged:	True	LL-Null:	-297.99
Covariance Type:	nonrobust	LLR p-value:	1.566e-101

Table.2

we has used the stats model due to the fact stats model supports the models by way of R style formulation and pandas data frame and it is straightforward to apply and provides the statistics and whole analytical view of data frame in a single pass, which is not furnished by other libraries.

After importing the statsmodel.api for creating useful regression we first needs to create a steady variable by the usage of “statsmodel.add_constant()” , and assign it to some other variable.

After that we has fitted the variables into model using “logit().fit()” , right here Logit is used as for “Logistic Regression”.

Logistic regression is the kind of regression analysis used to discover the probability of a positive event occurring. it's far the satisfactory suited type of regression in the cases wherein we have a categorical structured variable which can take handiest discrete values, It applies feasible results which are not numerical but alternatively specific. Within the same way like we

Can used in linear regression with the help of dummies. This sort of regression is non-linear in nature.

Through the usage of above idea as a base the Logit model is shaped, basically if we take the Log in both the perimeters of logistic model equation then we are able to acquired a very precise equation which is an equation of Logit version.[4]

$$\mathbf{Log(odds)=\beta_0 +\beta\,_1\beta x_1 +\ldots\ldots\ldots\ldots+ \beta_n\, X_n}$$

The above equation is optimized version of logistic regression equation due to which it is simple to calculate by a system and in understandable fashion. Due to those features we've included the logit version instead of the logistic version.

5)Preprocessing and model creation

Data Analysis

Before going to analysis of the data set in to our model earlier first of all we need to preprocess the records, to get the preferred outcome.

The dataset may be incomplete have a few missing attribute values, or have simplest mixture facts. So, there is a need to pre-procedure the medical dataset which has foremost attributes consisting of identification, prognosis, and different actual-valued capabilities which can be computed for each cell nucleus like radius, texture, parameter, smoothness, area, and so forth. [11]

Detecting Multicollinearity using VIF via Statsmodel

Multicollinearity may be detected thru diverse methods. In this project, we used the maximum not unusual place one – VIF (Variance Inflation Factors), as Sklearn has no approach to discover multicollinearity.

VIF determines the power of the correlation between the unbiased variables. it is predicted via taking a variable and regressing it towards each other variable. It produces a measure of how plenty larger the square root of a standard errors of an estimate is, as compared to the scenario in which the variable is absolutely uncorrelated with the. opposite predictors

R2 value is determined to find out how well an independent variable is described by the other independent variables. A high value of R2 means that the variable is highly correlated with the other variables.[8]

$$\mathbf{VIF=1/(1\text{-}R^2)}$$

So, the closer the R^2 value to 1, the higher the value of VIF and the higher the multicollinearity with the particular independent variable.

	mean_radius	mean_texture	mean_perimeter	mean_area	mean_smoothness	diagnosis
0	17.99	10.38	122.80	1001.0	0.11840	0
1	20.57	17.77	132.90	1326.0	0.08474	0
2	19.69	21.25	130.00	1203.0	0.10960	0
3	11.42	20.38	77.58	386.1	0.14250	0
4	20.29	14.34	135.10	1297.0	0.10030	0

Fig 1- Data Head

Sr.no	VIF	features
0	3526.872924	mean_radius
1	22.329552	mean_texture
2	4527.360366	mean_perimeter
3	89.413301	mean_area
4	49.764471	mean_smoothness

Table3- Detection using VIF

1) Heatmap:-

Looking at the matrix, you can quickly see the existence of multicollinearity between other variables. For example, the radius mean column has a relationship of 1 and 0.99 with respect to the perimeter mean and area_mean columns, respectively. This is because the three columns contain the same information: the apparent length of the observer (cell). Therefore, when proceeding with the analysis, you need to select only one of the three columns. Another place where the Multi Colliery appears is below the "center" and "worst" columns. For example, radius_mean has a 0.97 link correlation with the radius_worst column. In fact, each of the 10 key characteristics has the highest (0.7 to 0.97) correlation between the mean and worst columns. This is unavoidable because the "worst" column is not really a small set of "medium" columns. The "worst" column is also the "average" of some values (three large values between comments). When training your model, you'll need to remove the "worst" column of scores and best identify the "middle" column. Some positive correlation characteristics:

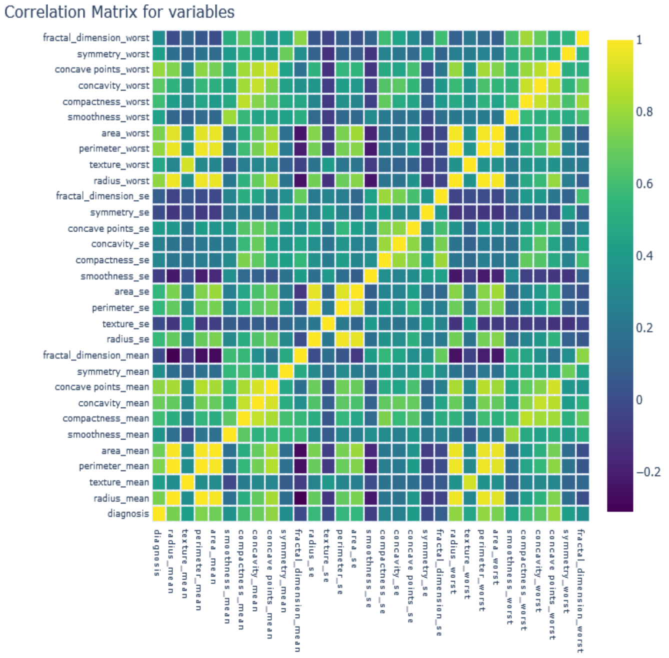


Fig 2) Correlation by Heatmap.

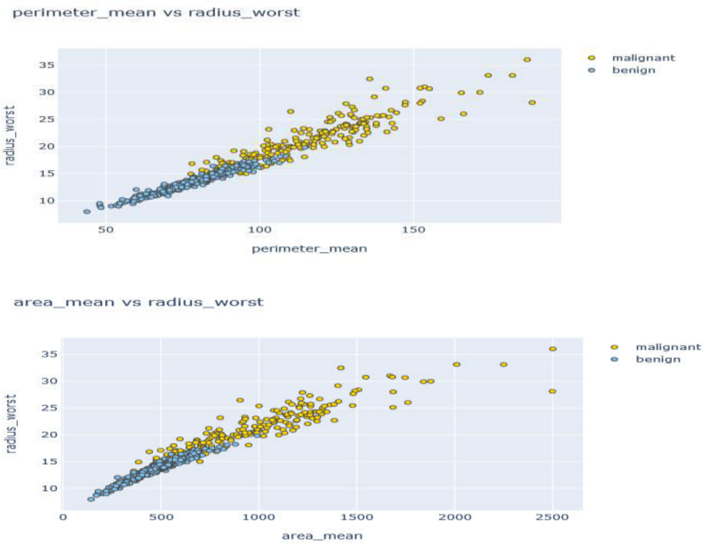


Fig 3) Correlation by Heatmap.

2.)Accuracy without removing correlated variables:-

- **Logistic Regerssion:-**

accuracy			0.96	228
macro avg	0.96	0.96	0.96	228
weighted avg	0.96	0.96	0.96	228

TABLE.4

	precision	**recall**	**f1-score**	**support**
0	0.97	0.97	0.97	148
1	0.94	0.95	0.94	80

TABLE.5

1) **Confusion Matrix:-**

```
[[143   5]
[4    76]]
```

SUPPORT VECTOR CLASSIFICATION:-

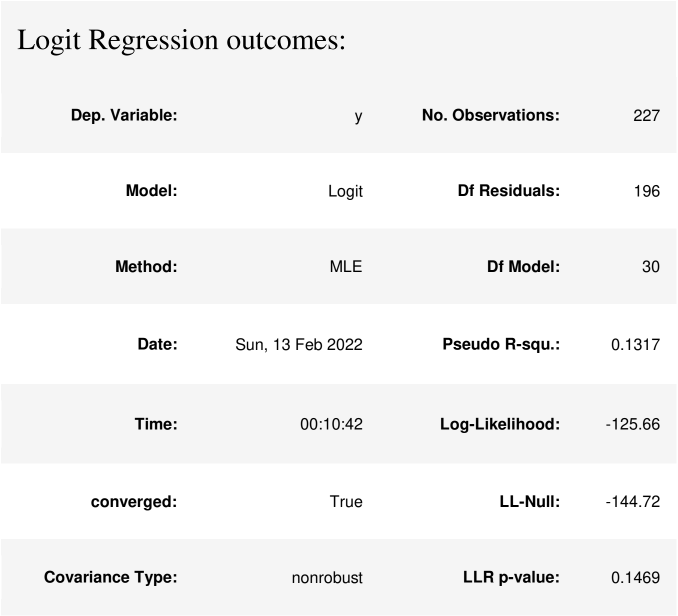
Logit Regression :-

	precision	recall	f1-score	support
0	0.96	0.97	0.96	219
1	0.94	0.93	0.93	123

Table 6

accuracy			0.95	342
Macro avg	0.95	0.95	0.95	342
weighted av g	0.95	0.95	0.95	342

Table 7



Logit Regression outcomes:

Dep. Variable:	y	**No. Observations:**	227
Model:	Logit	**Df Residuals:**	196
Method:	MLE	**Df Model:**	30
Date:	Sun, 13 Feb 2022	**Pseudo R-squ.:**	0.1317
Time:	00:10:42	**Log-Likelihood:**	-125.66
converged:	True	**LL-Null:**	-144.72
Covariance Type:	nonrobust	**LLR p-value:**	0.1469

TABLE.8

So From below results it is clear that Logistic regression & SVM unable to specify the multicolinearity in a dataset and provide higher accuracy which is not suitable for the precise models, but in case of LOGIT regression it clearly provides the lesser value of R^2 with the dataset having multicolinearity.

After Pre-processing :-

- **Logistic Regression:-**

	precision	recall	f1-score	support
0	0.91	0.95	0.94	213
1	0.91	0.86	0.89	129
accuracy		0.91	342	
macro avg	0.91	0.91	0.90	342
weighted avg	0.91	0.92	0.91	342

TABLE 9

Confusion Matrix:-

```
[ [203  10]
 [18 111]]
```

- **SVM Classifier:-**

TABLE 10

	precision	recall	f1-score	support
0	).90	0.95	0.94	213
1	0.90	0.86	0.89	129
accuracy		0.90	342	
macro avg	0.90	0.91	0.89	342
weighted avg	0.90	0.92	0.90	342

Table. 11

Logit : Optimization terminated successfully. Current function value: 0.127326 Iterations 10

Logit Regression Experimental Results

Dep. Variable:	**diagnosis**	**No. Observations:**	**455**
Model:	Logit	Df Residuals:	449
Method:	MLE	Df Model:	5
Date:	Thu, 23 Feb 2022	Pseudo R-squ.:	0.8056
Time:	19:09:18	Log-Likelihood:	-57.933
converged:	True	LL-Null:	-297.99
Covariance Type:	nonrobust	LLR p-value:	1.566e-101

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-0.3257	0.580	-0.561	0.575	-1.463	0.811
x1	23.1784	7.679	3.018	0.003	8.128	38.229
x2	-1.7208	0.328	-5.244	0.000	-2.364	-1.078
x3	-16.5433	5.035	-3.286	0.001	-26.412	-6.674
x4	-14.5570	5.851	-2.488	0.013	-26.024	-3.090
X5	-1.6819	0.336	-5.000	0.000	-2.341	-1.023

Table.12

Confusion Matrix And Accuracy:-

```
(array ([[148., 17.], [ 13., 277.]]),
        94.4065934065934)
```

Algorithm used to find CM and Accuracy in Logit Regression:-

```
defconfu(data,actual,model):
 pred=model.predict(data)
 bins=np.array([0,0.5,1])
 cm=np.histogram2d(actual,pred,bins=bins)[0]
 accuracy=(cm[0,0]+cm[1,1])/cm.sum()*100
 returncm,accuracy
```

Data Splitting:-

After the successful processing of null statistics. The dataset is split into parts i.e. test data and train data. In our paper, seventy-five% of records are trained records, and 25% of information is test data. Test_train_split is used to split the data.

Model Selection

this is the most crucial phase of evaluation where numerous models can be used. We've analyzed diverse models and primarily based on the accuracy we carried out the fine one i.e. Logit Regression model.
Following are some models which we've analyzed:

Logistic Regression

Logistic regression is the type of regression analysis used to find the probability of a certain event happening. It is the best-suited type of regression in the cases where we have a categorical dependent variable that can take only discrete values, It applies possible outcomes which are not numerical but rather categorical. In the same way, we can use linear regression with the help of dummies. This type of regression is non-linear.

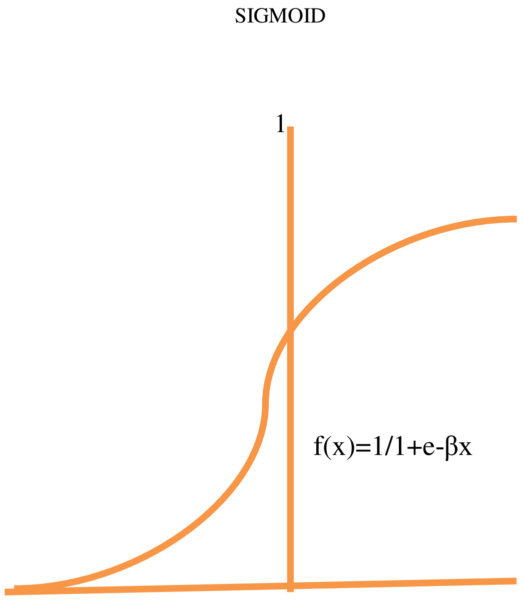


FIGURE-4 LOGISTIC REGRESSION

Logit Regression

The Logit version is formed if we take the Log in both the perimeters of the logistic model equation then we will obtain a totally precise equation that's an equation of the Logit model.

$$Log(odds)=\beta_0 + \beta_1 x_1 +\ldots\ldots+ \beta_n x_n$$

The above equation is simple to calculate and in a comprehensible manner and for this project it supplied a totally particular end result due to these capabilities we've got covered the logit version instead of the logistic model from statsmodel.*[4]*

Random Forest Classification

The random forest classifier or regress or is a bagging method, which means it includes numerous base learner models, In random forests, these models are known as decision trees, It took a few characteristic samples and provide that to one of the decision trees the same manner with alternative is done with all of the trees it incorporates, and the very last output is the aggregated to get a majority vote for the final results from that bootstrap of the selection trees. Means With random forest, you may also cope with regression tasks through using the algorithm's regression.[9]

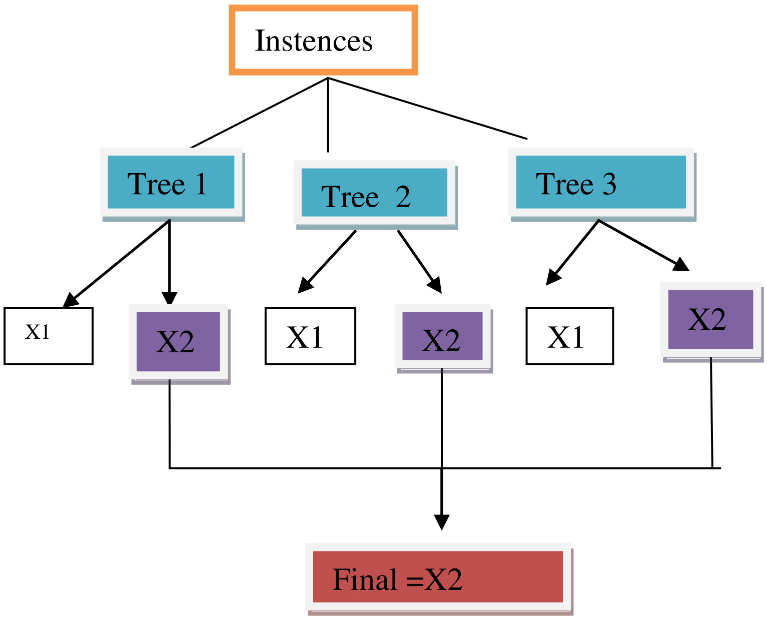


Fig. 5 Random Forest classifier

6) EXPERIMENTAL RESULT

After analyzing all **3** models we have a tendency to got our final model which is able to be used for prediction. The accuracy of Logit Regression was found to be additional correct than the opposite2 models. The accuracy of Logit regression is **94.40%.**

Logit Regression Result:

Optimization terminated successfully. Current function value: 0.127326 Iterations 10

Dep. Variable:	diagnosis	No. Observations:	455
Model:	Logit	Df Residuals:	449
Method:	MLE	Df Model:	5
Date:	Thu, 23 Feb 2022	Pseudo R-squ.:	0.8056
Time:	19:09:18	Log-Likelihood:	-57.933
converged:	True	LL-Null:	-297.99
Covariance Type:	nonrobust	LLR p-value:	1.566e-101

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3257	0.580	-0.561	0.575	-1.463	0.811
x1	23.1784	7.679	3.018	0.003	8.128	38.229
x2	-1.7208	0.328	-5.244	0.000	-2.364	-1.078
x3	-16.5433	5.035	-3.286	0.001	-26.412	-6.674
x4	-14.5570	5.851	-2.488	0.013	-26.024	-3.090
X5	-1.6819	0.336	-5.000	0.000	-2.341	-1.023

Table.13

7) CONCLUSION

Breast cancer is the most typical cancer and a type of Chronic Disease in women's and also the second main reason behind the cancer death in women. Once the first symptoms of breast cancer are ignored, the patient would possibly find herself with forceful consequences in her health carcinoma are often unbroken in restraint when it's detected early. Several studies focus primarily on the applying of classification techniques to breast cancer prediction, but here after comparing with the three models 1)SVM,2)Random Forest 3) Logit Regression we found that the logit provides the precise details about the dataset that whether the data is preprocessed or not and then it provides the result about the states of cancer tumours with an accuracy of 94.40% without any overfitting and underfitting condition in comparison of SVM and Random Forest classifier which provides the accuracy of 96% each in case of unprocessed data and 93.50% in preprocessed condition .Even it is look like they are providing the higher accuracy but this result can be harmful for patients as if data is unprocessed and model got trained then in that case there are higher chances to get wrong result that's why it is important to use a model which can handle the unprocessed data on its own . It's been ascertained that a decent dataset provides higher accuracy. The choice of acceptable algorithms with a decent local dataset can result in the event of prediction systems. These systems will assist in correct treatment ways for a patient diagnosed with carcinoma.

REFRENCES:

[1] Mayo Clinic. Breast Cancer: Symptoms and causes - [Internet]. Mayo Clinic. 2016. Available from:https://www.mayoclinic.org/diseases...../breast-cancer/symptoms-causes/syc-20352470

[2]https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.21660

[3]https://www.kaggle.com/c/1056lab-breast-cancer-diagnosis

[4] Seabold, Skipper, and Josef Perktold. "statsmodels: Econometric and statistical modeling with python." Proceedings of the 9th Python in Science Conference. 2010.

[5]American Cancer Society. How Common Is Breast Cancer? https://www.cancer.org/cancer/breast-cancer/about/howcommon-is-breast-cancer.html (2018).

[6] Oeffinger, K. C. et al. Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society. JAMA 314, 1599–1614 (2015).

[7] Glenn. V. Ostir, "Logistic Regression: A Nontechnical Review", American Journal of Physical Medicine & Rehabilitation., vol. 6, pp. 565-572, 2000.

[8] M Kalaiyarasi, R Dhanasekar, S Sakthiya Ram, P. Vaishnavi, "Classification of Benign or Malignant Tumour Using Machine Learning", IOP Conference Series: Materials Science and Engineering, vol. 995, pp. 012028, 2020.

[9] S. Nanglia, Muneer Ahmad, Fawad Ali Khan, N.Z. Jhanjhi, "An enhanced Predictive heterogeneous ensemble model for breast cancer prediction", Biomedical Signal Processing and Control, vol. 72, pp. 103279, 2022.

[10] GitanjaliWadhwa, Amandeep Kaur, Soft Computing and Signal Processing, vol. 1340, pp. 281, 2022.

[11] LeenaNesamani S., S. NirmalaSigirthaRajini, Handbook of Research on Innovations and Applications of AI, IoT, and Cognitive Technologies, pp. 204, 2021.

[12]Sultana, J.. (2018). Predicting Breast Cancer using Logistic Regression and Multi-Class Classifiers. International Journal of Engineering and Technology(UAE).

[13] Li J, Wong L. Rule-Based Data Mining Methods for Classification Problems in Biomedical Domains; 15th European Conference on Machine Learning (ECML) (2004).

[14] Kumar D, Beniwal S. Genetic Algorithm and Programming Based Classification: A Survey. Journal of Theoretical and Applied Information

Technology. 54:48–58 (2013).

[15] Mansuri AM, Verma M, Laxkar P. A Survey of Classifier Designing Using Genetic Programming and Genetic Operators. International Journal of Engineering Research and Reviews (IJERR) Vol. 2:16–22 (2014).

[16] Loh WY. Encyclopedia of Statistics in Quality and Reliability. Ruggeri, Kenett&Faltin, Wiley; Classification and Regression Tree Methods; pp. 315–323 (2008).

[17] Li Y, Zhu J. Analysis of array CGH data for cancer studies using fused quantile regression. Bioinformatics. Vol.23:2470–2476 (2007).

[18]-UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/ (5-15) several classification algo.

[19] Refaeilzadeh P., Tang L., Liu. H. Cross Validation. In Encyclopedia of Database Systems, 532538, Springer, U.S, (2009). [20]"WEKA Data Mining Book" (n.d.).

[20 Kusiak A. Decomposition in Data Mining: An Industrial Case Study in IEEE Transactions On Electronics Packaging Manufacturing, Vol. 23, No. 4, 87-97, (2000).

[21] leCessie, S., van Houwelingen, J.C. (1992). Ridge Estimators in Logistic Regression. Applied Statistics. 41(1):191-201.D. Aha, D. Kibler Instance-based learning algorithms. Machine Learning. 6:37-66 (1991).

[22]Aha D., Kibler D., Instance-based learningalgorithms. Machine Learning. 6:37-66 (1991).

[23] John G. Cleary, Leonard E. Trigg: K*: An Instance-based Learner Using an Entropic Distance Measure. In: 12th International Conference on Machine Learning, 108-114, (1995).

[24] Walter H. Delashmit and Michael T. Manry, 2005. Recent Developments in Multilayer Perceptron Neural Networks. Proceedings of the 7th Annual Memphis Area Engineering and Science Conference, MAESC. 699 (2005).

[25] John G. Cleary, Leonard E. Trigg: K*: An Instance-based Learner Using an Entropic Distance Measure. In: 12th International Conference on Machine Learning, 108-114, (1995).

[26] Walter H. Delashmit and Michael T. Manry, 2005. Recent Developments in Multilayer Perceptron Neural Networks. Proceedings of the 7th Annual Memphis Area Engineering and Science Conference, MAESC. 699 (2005).

- Mining Book" (n.d.) http://www.cs.waikato.ac.nz/~ml/weka/book.html.