International Journal of Computing, Programming and Database Management

E-ISSN: 2707-6644
P-ISSN: 2707-6636
IJCPDM 2020; 2(2): 04-10
Received: 06-03-2021
Accepted: 06-05-2021

Divya G
Department of Computer Science, SDHR College, Tirupati, Andhra Pradesh, India

Boyella Mala Konda Reddy
Assistant Professor, Department of Computer Science, SDHR College, Tirupati, Andhra Pradesh, India

Corresponding Author:
Divya G
Department of Computer Science, SDHR College, Tirupati, Andhra Pradesh, India

Performance measure of breast cancer prediction using decision tree approach

Divya G and Boyella Mala Konda Reddy

DOI: https://doi.org/10.33545/27076636.2021.v2.i2a.25

Abstract
This paper investigations choice tree calculation for Breast disease discovery. The effectiveness of choice tree calculation can be broke down dependent on their precision and the quality choice measure utilized. The paper likewise gives a thought of the trait choice measure utilized by different choice tree calculation utilizes data gain and GINI Index as the quality choice measure. In this paper, the expectation of Decision Tree characterization is evaluated using two property trait choice decision measures for Breast Cancer sickness dataset. Choice tree uses separate and vanquish framework for the fundamental learning technique. From the result examination we can reason that the execution of Decision Tree grouping relies upon the trademark quality choice decision measures. Choice Tree is significant since improvement of decision tree classifiers doesn't need any territory learning. The essential objective is to produce a capable assumption show for Breast Cancer sickness expectation returns with high precision.

Keywords: Breast cancer prediction, GINI Index, expectation

1. Introduction

Breast cancer malignant growth is quite possibly the most deadly and heterogeneous illness in this current time that causes the passing of colossal number of ladies everywhere on the world. It is the second biggest illness that is capable of ladies demise [1]. There are different AI and information mining calculations that are being utilized for the forecast of bosom malignant growth. Finding the generally reasonable and fitting calculation for the expectation of bosom malignant growth is one of the significant assignments. Breast cancer disease is begun through threatening tumors, when the development of the cell gained out of power [2, 9]. A great deal of greasy and sinewy tissues of the bosom starts strange development that turns into the reason for bosom malignant growth. The disease cells spread all through the tumors that cause various phases of malignant growth.

The grouping of Breast Cancer sickness expectation has become an undeniably difficult issue, because of late advances in information assortment and clinical mining innovation. The Clinical associations have gathered enormous amounts of data about patients and illnesses. In this Paper we look at key parts of the Decision Tree Classification in Breast Cancer determination.

2. Classification Overview

Grouping is the handling of tracking down a bunch of models (or capacities) which depict and recognize information classes or ideas. Building quick and exact classifiers for enormous informational collections is a significant undertaking in information mining and AI [5]. Arrangement is the most usually applied information mining strategy, and utilizes a bunch of pre-ordered guides to build up a model that can characterize the number of inhabitants in records on the loose.

Order is a two-venture measure. In the initial step, which is known as the learning step, a model that portrays a foreordained arrangement of classes or ideas is worked by dissecting a bunch of preparing data set examples. Each example is accepted to have a place with a predefined class. In the subsequent advance, the model is tried utilizing an alternate informational collection that is utilized to assess the grouping precision of the model. On the off chance that the precision of the model is viewed as adequate, the model can be utilized to order future information cases for which the class name isn't known. There are a few calculations that can be utilized for arrangement, for example, choice tree, Support Vector

Machines, Bayesian techniques, rule-based calculations, and Neural Networks [6].

The goal is to utilize the preparation informational collection to construct a model of the class mark dependent on different qualities with the end goal that the model can be utilized to order new information not from the preparation informational index ascribes. With order, the produced model will actually want to foresee a class for given information relying upon recently took in data from recorded information.

The arrangement of Breast Cancer sickness expectation has become an inexorably difficult issue, because of ongoing advances in information assortment and Medical mining innovation. The Clinical associations have gathered huge amounts of data about patients and illnesses. In this paper we analyze crucial parts of the SVM arrangement in clinical finding.

3. Decision Tree Classification

A Decision Tree is a tree-like diagram comprising of inner hubs which address a test on a trait and branches which indicate the result of the test and leaf hubs which mean a class mark [7]. The arrangement rules are framed by the way chose from the root hub to the leaf. To isolate each information, first the root hub is picked as it is the most conspicuous quality to isolate the information. The tree is built by recognizing ascribes and their related qualities which will be utilized to investigate the information at each halfway hub of the tree. After the tree is shaped, it can prefigure recently coming information by navigating, beginning from a root hub to the leaf hub visiting every one of the inner hubs in the way relying on the test states of the traits at every hub. The primary benefit of utilizing choice trees rather than other characterization strategies is that they give a rich arrangement of decides that are straightforward.

Choice tree is a various leveled information structure that addresses information through a separation and overcome procedure. In characterization, the objective is to gain proficiency with a choice tree that addresses the preparation information to such an extent that names for new models can be resolved. The primary targets of choice tree classifiers are: 1) to order effectively however much of the preparation test as could be expected; 2) sum up past the preparation test so concealed examples could be characterized with as high of a precision as could really be expected.

3.1. Attribute Selection Measures

For choosing the parting model that "best" isolates the information segment, D, of class-named preparing tuples into singular classes, we utilized characteristic determination measure which is heuristic for such choice. If we somehow managed to part D into more modest segments as per the results of the parting basis, in a perfect world each parcel would be unadulterated (i.e., all the tuples that fall into a given segment would have a place into a similar class) [6, 7]. The consequence of this situation is really the "awesome" measure of the multitude of standards taken. Trait choice measure decides how to part the tuples at a given hub and are along these lines otherwise called dividing rules. The dividing properties can be ceaseless esteemed or it tends to be confined to twofold trees. For nonstop esteemed traits, a split point should be resolved as a feature of the parting measure while for the twofold trees a parting subset should be resolved. The tree hub for parcel is named with the parting rule, branches are developed for every result of standard and the tuples are divided appropriately. The most well-known property choice measures are – Entropy (Information Gain), Gain Ratio and Gini Index.

3.1.1 Entropy

Entropy is a measure of uncertainty associated with a random variable. The entropy increases with the increase in uncertainty or randomness and decreases with a decrease in uncertainty or randomness. The value of entropy ranges from 0-1.

Entropy (D) = $\sum_{i=0}^{c} -pi\ log2(pi)$

Where pi is the non-zero probability that an arbitrary tuple in D belongs to class C and is estimated by |Ci,D|/|D|. A log function of base 2 is used because as stated above the entropy is encoded in bits 0 and 1.

3.1.2 Information Gain

ID3 uses information gain as its attribute selection measure. Claude Shannon studied the value or "information content" of messages and gave information gain as a measure in his Information Theory [7]. Information Gain is the difference between the original information gain requirement (i.e. based on just the proportion of classes) and the new requirement (i.e. obtained after the partitioning of A).

Gain (D, A) = Entropy (D)- $\frac{|Dj|}{|D|}$ Entropy (Dj)

Where,
D: A given data partition
A: Attribute
V: Suppose we partition the tuples in D on some attribute A having v distinct values
D is split into v partition or subsets, $\{D_1, D_2, D_j\}$ were D_j contains those tuples in D that have outcome a_j of A.The attribute that has the highest information gain is chosen.

3.1.3 Gini Index

Gini Index is an attribute selection measure used by the CART decision tree algorithm. The Gini Index measures the impurity D, a data partition or set of training tuples as:

Gini (D) = 1- $\sum_{i=1}^{m}$ pi^2

Where pi is the probability that a tuple in D belongs to class Ci and is estimated by |Ci,D|/|D|. The sum is computed over m classes. The attribute that reduces the impurity to the maximum level (or has the minimum gini index) is selected as the splitting attribute.

4. Experimental Results

The Decision Tree Classification have been tried different things with information taken from the UCI Machine Learning Repository [9] and utilized the Python Language to explore the Decision Tree calculation. The Python Scikit-learn is a bundle for information order, relapse, bunching and perception. The dataset we utilized in our examination is momentarily portrayed in Table-1. The information is isolated in two sets. The preparation set is 70% and the excess 30% are utilized for testing. The investigations were directed with complete list of capabilities.

Table 1: provides the attribute information of UCI data

SNO	Datasets	Features	Instances	Class
1	Wisconsin Breast cancer	11	699	2

The Breast disease Data set has 683 lines and 10 sections. In characterization issues how class marks are dispersed. So in this information there are two class names i.e., The Benign class has 444 and malignant class has 239. Spellbinding measurements can give us an extraordinary understanding into the state of each characteristic of information. So we can outlines of the bosom malignancy information as demonstrated in the table-2

Thickness plots are another method of finding out about the dispersion of each trait as demonstrated in figure-1. The plots appear as though a preoccupied histogram with a smooth bend drawn through the highest point of each container, similar as your eye attempted to do with the histograms. We can see the appropriation for each trait is clearer than the histograms.

Table 2: Descriptive statistics of shape of each attribute.

	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bland Chromatin	Normal Nucleoli	Mitoses	Class
Count	683.000	683.000	683.000	683.000	683.000	683.000	683.000	683.000	683.000
Mean	4.442	3.151	3.215	2.830	3.234	3.445	2.870	1.603	2.700
Std	2.821	3.065	2.989	2.865	2.223	2.450	3.053	1.733	0.955
Min	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	2.000
25%	2.000	1.000	1.000	1.000	2.000	2.000	1.000	1.000	2.000
50%	4.000	1.000	1.000	1.000	2.000	3.000	1.000	1.000	2.000
75%	6.000	5.000	5.000	4.000	4.000	5.000	4.000	1.000	4.000
Max	10.000	10.000	10.000	10.000	10.000	10.000	10.000	10.000	4.000

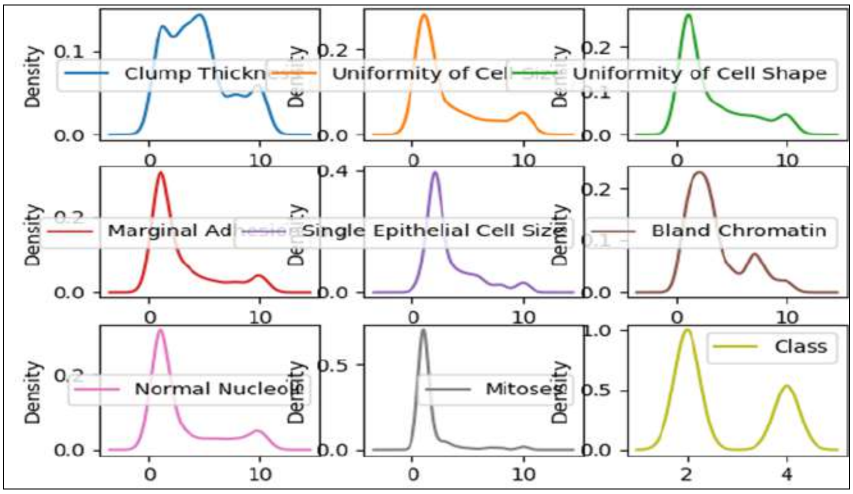


Fig 1: Density plot of Data distribution of each attribute

i. Measures for performance evaluation

There are many factors that affect the performance of a classifier. There exist different measures that can be used to evaluate the performance of a classifier, such as Accuracy, sensitivity, specificity, precision and recall, etc. These metrics are traditionally defined for a binary classification task with positive and negative classes. That is:

Accuracy: Accuracy is a measure which determines the probability that how much results are accurately classified.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

Sensitivity: Sensitivity is a measure which determines the probability of the results that are true positive such that person has the disease.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \qquad (2)$$

Specificity: Specificity is a measure which determines the probability of the results that are true negative such that person does not have the disease.

$$\text{Specificity} = \frac{TN}{TN + FP} \qquad (3)$$

Precision: Precision represents how precise the classifier predictions are since it shows the amount of true positives that were predicted out of all positive labels assigned to the instances by the classifier. Precision is the proportion of positive predictions that are correct

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (4)$$

Recall: Recall is the proportion of positive samples that are correctly predicted positive. It shows the amount of truly predicted positive classes out of the amount of total actual positive classes.

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (5)$$

Where,

- True positive (TP) = number of positive samples correctly predicted.
- False negative (FN) = number of positive samples wrongly predicted.
- False positive (FP) = number of negative samples wrongly predicted as positive.
- True negative (TN) = number of negative samples correctly predicted.

Confusion matrix is a visualization tool which is commonly used to present the accuracy of the classifiers in classification that assist with performance evaluation purposes which consist of the concepts defined above measurements. This is illustrated in table-3. It is used to show the relationships between outcomes and predicted classes.

Table 3: Confusion matrix

		Predicted	
		Positive	**Negative**
Actual Class	Positive	TP	FN
	Negative	FP	TN

ii. Results

The confusion matrix of Decision Tree classification method is presented in the table-4 of Wisconsin Breast cancer Data. The values to measure the performance of the methods (i.e. accuracy, sensitivity and specificity) are derived from the confusion matrix and shown in table-5 and same shown in graphical representation in figure-2.

Table 4: Confusion Matrix of Wisconsin Breast cancer Test Data

GINI Method Wisconsin Breast cancer Test Data (205)			
		Predicted	
		Benign	**Malignant**
Actual Class	Benign	125	8
	Malignant	3	69
ENTROPY Method Wisconsin Breast cancer Test Data (205)			
		Predicted	
		Benign	**Malignant**
Actual Class	Benign	126	7
	Malignant	2	70

Based on the above confusion matrices, we calculated Accuracies, Sensitivity, Specificity, Precision and Recall as shown in table-5 same shown in graphical representation in figure-2.

Table 5: Performance of Decision Tree Method

S No	Method	Accuracy	Sensitivity	Specificity	Precision	Recall
1	GINI	94.66	97.65	89.61	95	95
2	ENTROPY	95.6	98.43	97.22	96	96

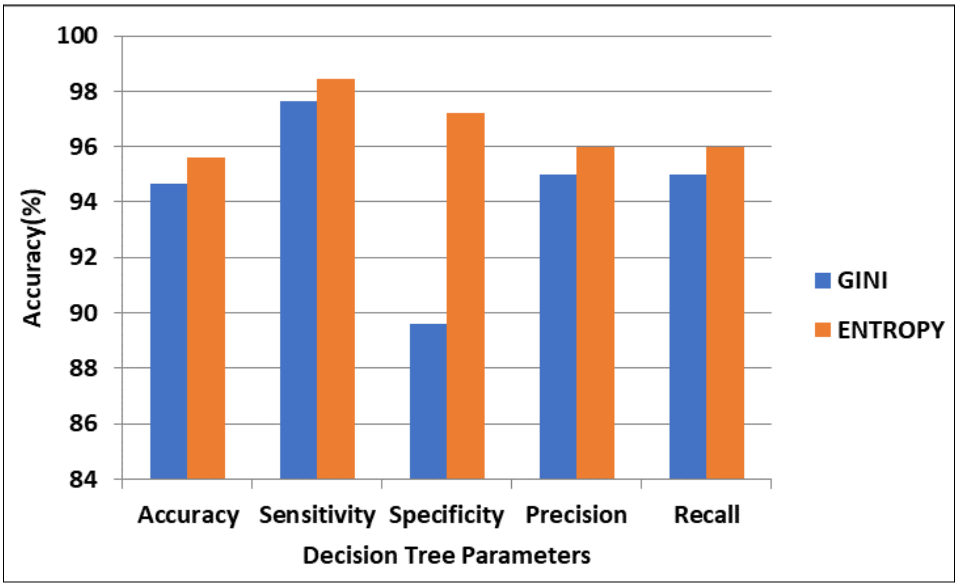
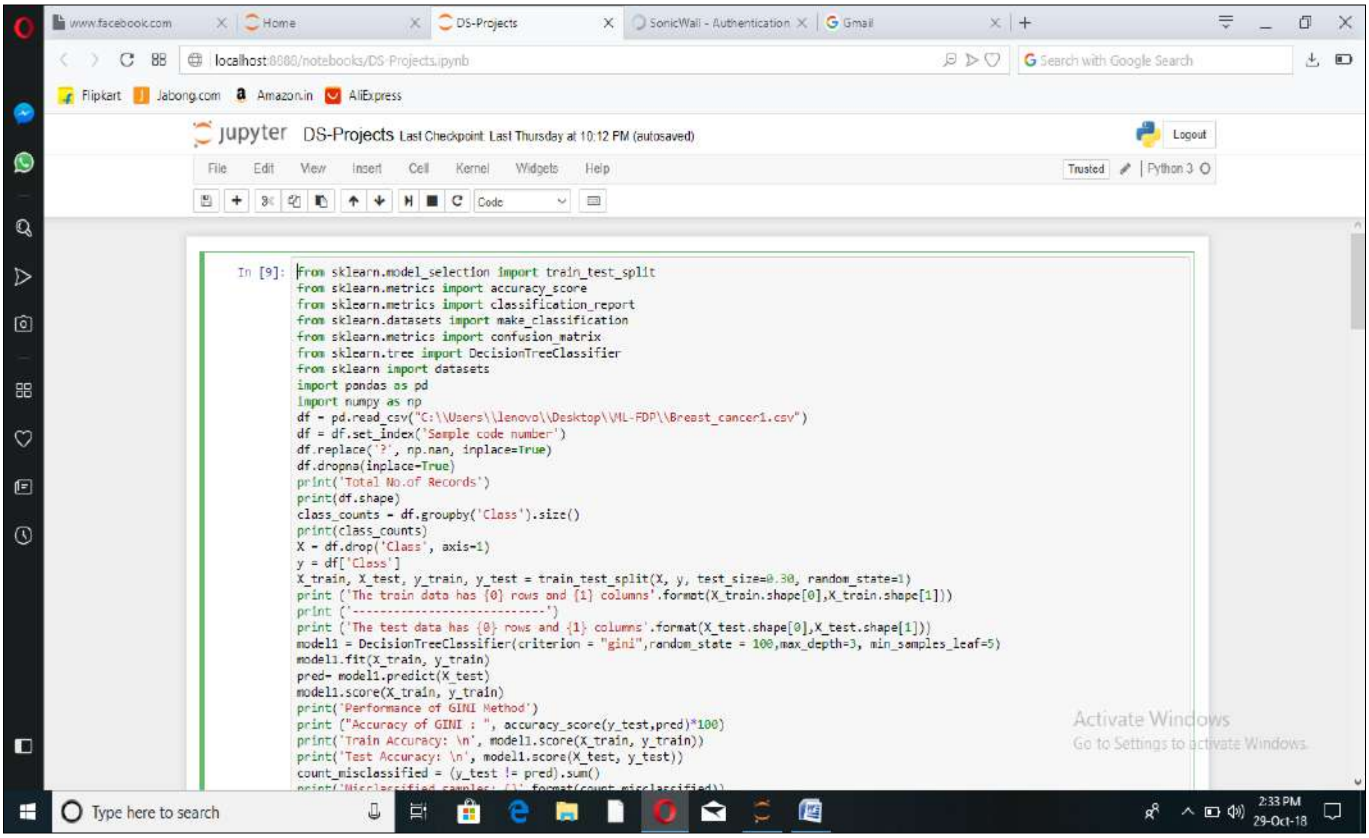


Fig 2: Performance of Decision Tree Method

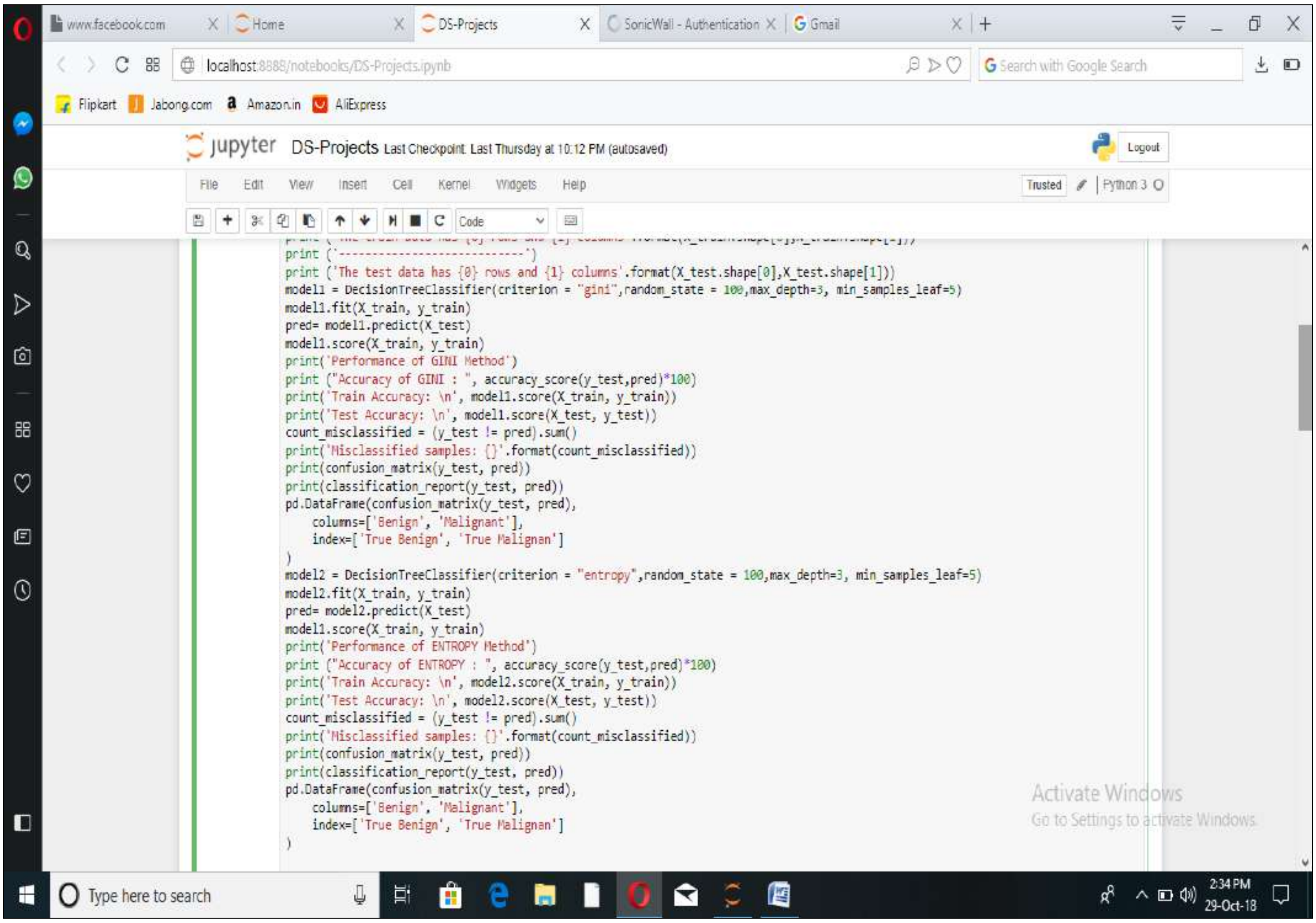
It might be found in the figure-2 that the Decision Tree estimation using Entropy quality assurance measure has achieved 95.6% of precision, while a comparable using Gini decision measure got 94.66% of exactness. So the Decision tree gathering with Entropy assurance measure in all execution metric characteristics like precision, survey are high appeared differently in relation to Gini decision measure.

4.2 Screen Shots



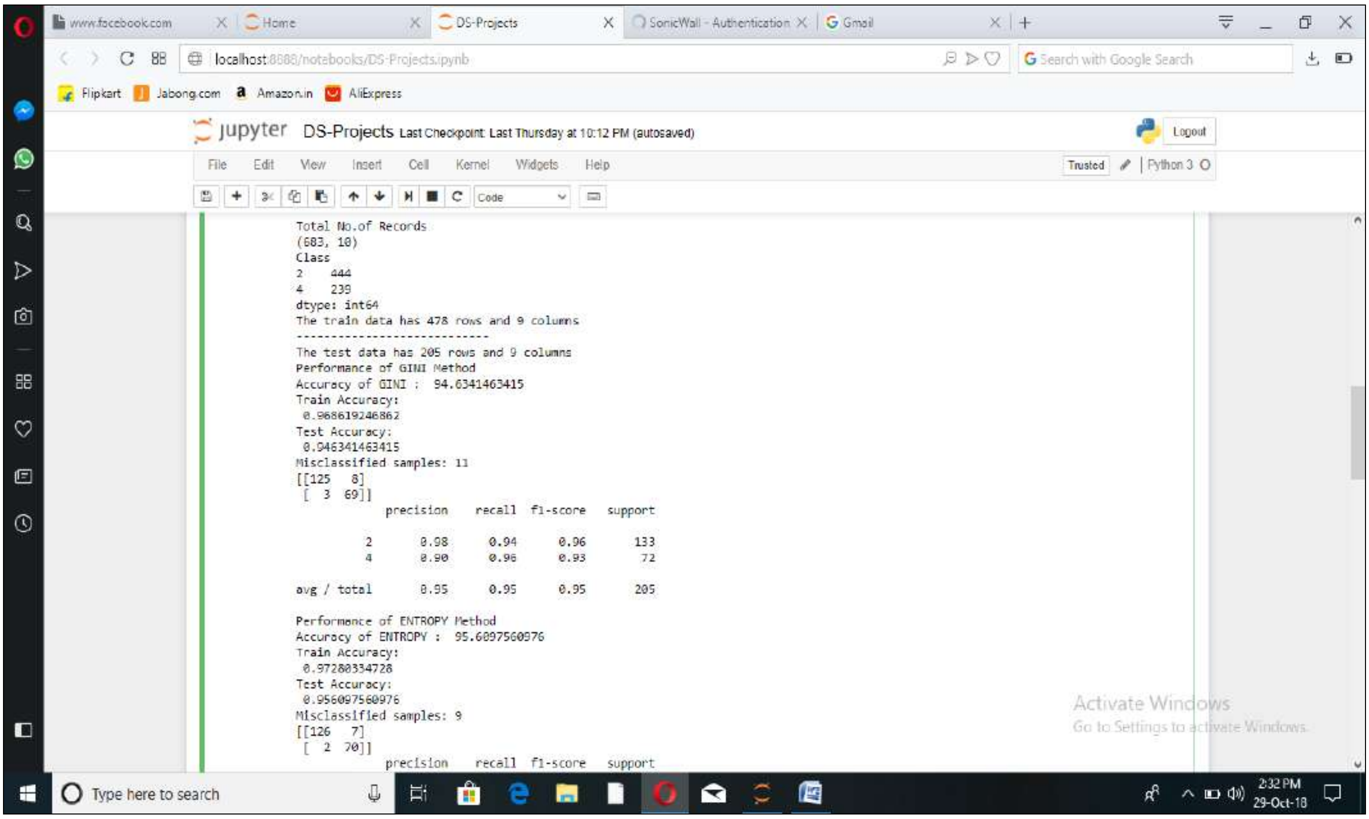
jupyter DS-Projects Last Checkpoint: Last Thursday at 10:12 PM (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Code
Logout
Activate Windows
Go to Settings to activate Windows.

```
In [9]: from sklearn.model_selection import train_test_split
        from sklearn.metrics import accuracy_score
        from sklearn.metrics import classification_report
        from sklearn.datasets import make_classification
        from sklearn.metrics import confusion_matrix
        from sklearn.tree import DecisionTreeClassifier
        from sklearn import datasets
        import pandas as pd
        import numpy as np
        df = pd.read_csv("C:\\Users\\lenovo\\Desktop\\ML-FDP\\Breast_cancer1.csv")
        df = df.set_index('Sample code number')
        df.replace('?', np.nan, inplace=True)
        df.dropna(inplace=True)
        print('Total No.of Records')
        print(df.shape)
        class_counts = df.groupby('Class').size()
        print(class_counts)
        X = df.drop('Class', axis=1)
        y = df['Class']
        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=1)
        print ('The train data has {0} rows and {1} columns'.format(X_train.shape[0],X_train.shape[1]))
        print ('-----------------------------')
        print ('The test data has {0} rows and {1} columns'.format(X_test.shape[0],X_test.shape[1]))
        model1 = DecisionTreeClassifier(criterion = "gini",random_state = 100,max_depth=3, min_samples_leaf=5)
        model1.fit(X_train, y_train)
        pred= model1.predict(X_test)
        model1.score(X_train, y_train)
        print('Performance of GINI Method')
        print ("Accuracy of GINI : ", accuracy_score(y_test,pred)*100)
        print('Train Accuracy: \n', model1.score(X_train, y_train))
        print('Test Accuracy: \n', model1.score(X_test, y_test))
        count_misclassified = (y_test != pred).sum()
        print('Misclassified samples: {}'.format(count_misclassified))
```



jupyter DS-Projects Last Checkpoint: Last Thursday at 10:12 PM (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Code
Logout
Activate Windows
Go to Settings to activate Windows.

```
print ('The train data has {0} rows and {1} columns'.format(X_train.shape[0],X_train.shape[1]))
print ('-----------------------------')
print ('The test data has {0} rows and {1} columns'.format(X_test.shape[0],X_test.shape[1]))
model1 = DecisionTreeClassifier(criterion = "gini",random_state = 100,max_depth=3, min_samples_leaf=5)
model1.fit(X_train, y_train)
pred= model1.predict(X_test)
model1.score(X_train, y_train)
print('Performance of GINI Method')
print ("Accuracy of GINI : ", accuracy_score(y_test,pred)*100)
print('Train Accuracy: \n', model1.score(X_train, y_train))
print('Test Accuracy: \n', model1.score(X_test, y_test))
count_misclassified = (y_test != pred).sum()
print('Misclassified samples: {}'.format(count_misclassified))
print(confusion_matrix(y_test, pred))
print(classification_report(y_test, pred))
pd.DataFrame(confusion_matrix(y_test, pred),
    columns=['Benign', 'Malignant'],
    index=['True Benign', 'True Malignan']
)
model2 = DecisionTreeClassifier(criterion = "entropy",random_state = 100,max_depth=3, min_samples_leaf=5)
model2.fit(X_train, y_train)
pred= model2.predict(X_test)
model1.score(X_train, y_train)
print('Performance of ENTROPY Method')
print ("Accuracy of ENTROPY : ", accuracy_score(y_test,pred)*100)
print('Train Accuracy: \n', model2.score(X_train, y_train))
print('Test Accuracy: \n', model2.score(X_test, y_test))
count_misclassified = (y_test != pred).sum()
print('Misclassified samples: {}'.format(count_misclassified))
print(confusion_matrix(y_test, pred))
print(classification_report(y_test, pred))
pd.DataFrame(confusion_matrix(y_test, pred),
    columns=['Benign', 'Malignant'],
    index=['True Benign', 'True Malignan']
)
```

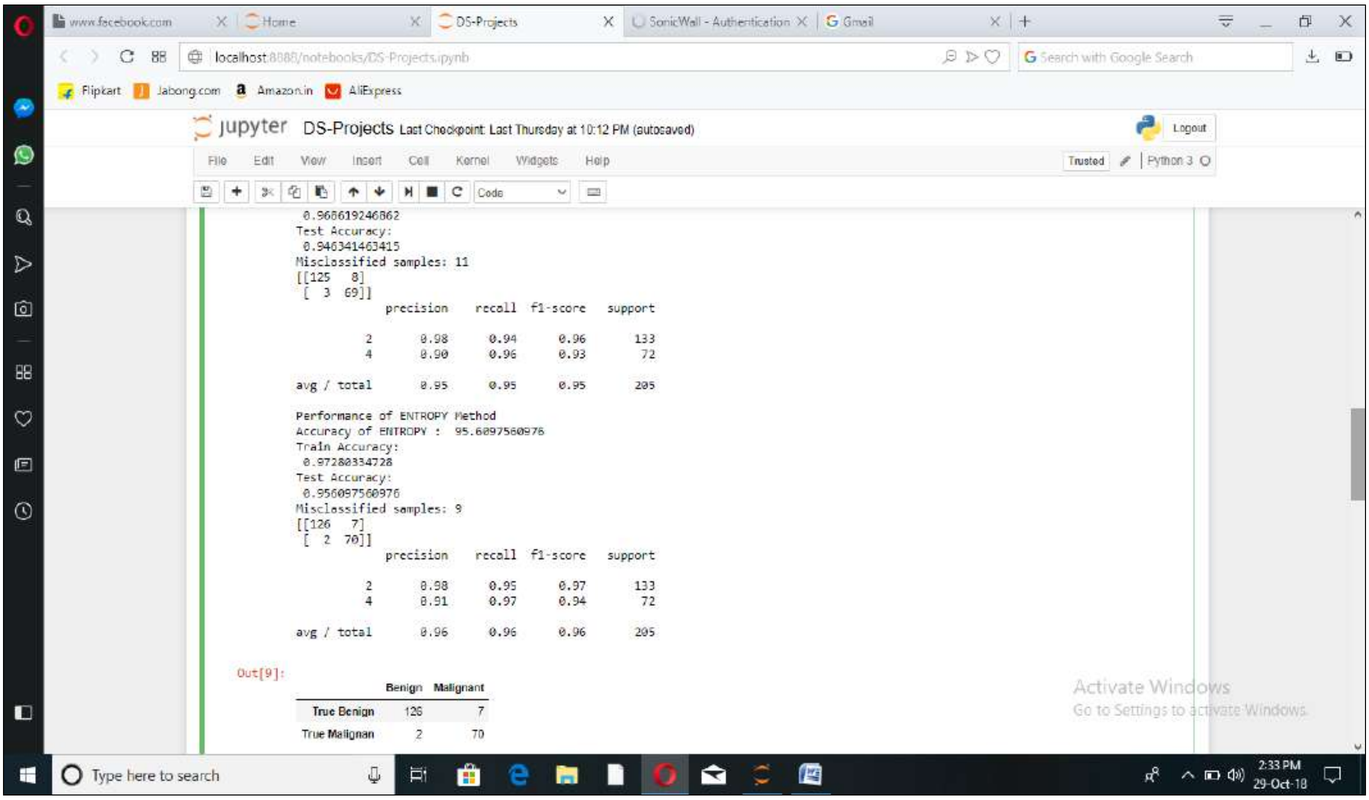

jupyter DS-Projects Last Checkpoint: Last Thursday at 10:12 PM (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Logout

```
Total No.of Records
(683, 10)
Class
2    444
4    239
dtype: int64
The train data has 478 rows and 9 columns
----------------------------
The test data has 205 rows and 9 columns
Performance of GINI Method
Accuracy of GINI :  94.6341463415
Train Accuracy:
 0.968619246862
Test Accuracy:
 0.946341463415
Misclassified samples: 11
[[125   8]
 [  3  69]]
             precision    recall  f1-score   support

          2       0.98      0.94      0.96       133
          4       0.90      0.96      0.93        72

avg / total       0.95      0.95      0.95       205

Performance of ENTROPY Method
Accuracy of ENTROPY :  95.6097560976
Train Accuracy:
 0.97280334728
Test Accuracy:
 0.956097560976
Misclassified samples: 9
[[126   7]
 [  2  70]]
             precision    recall  f1-score   support
```


jupyter DS-Projects Last Checkpoint: Last Thursday at 10:12 PM (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Logout

```
 0.968619246862
Test Accuracy:
 0.946341463415
Misclassified samples: 11
[[125   8]
 [  3  69]]
             precision    recall  f1-score   support

          2       0.98      0.94      0.96       133
          4       0.90      0.96      0.93        72

avg / total       0.95      0.95      0.95       205

Performance of ENTROPY Method
Accuracy of ENTROPY :  95.6097560976
Train Accuracy:
 0.97280334728
Test Accuracy:
 0.956097560976
Misclassified samples: 9
[[126   7]
 [  2  70]]
             precision    recall  f1-score   support

          2       0.98      0.95      0.97       133
          4       0.91      0.97      0.94        72

avg / total       0.96      0.96      0.96       205
```

Out[9]:

	Benign	Malignant
True Benign	126	7
True Malignan	2	70

5. Conclusion

In this investigation work, we surveyed the execution of Decision Tree assumption procedure for portrayal of Breast Cancer sickness forecast with two property decision extents of Gini and Entropy. Request precision is significantly penniless upon the Decision Tree is trademark assurance measure models for smoothing out. We survey the execution of the Decision Tree Classification technique with the two standard extents of Entropy and Gini record, similar to the precision, Precision and Recall of the model. The assessments performed using this dataset as data has achieved a structure giving high affirmation pace of Breast Cancer sickness expectation. The target to get high accuracy of gauge is fulfilled by Decision Tree using Entropy measure.

References

1. American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/)

2. Bellachia A, Guvan E. “Predicting breast cancer survivability using data mining techniques”, Scientific Data Mining Workshop, in conjunction with the 2006 SIAM Conference on Data Mining, 2006.
3. Ravi Kumar G, Nagamani K, Anjan Babu G. “A Framework of Dimensionality Reduction Utilizing PCA for Neural Network Prediction”, Lecture Notes on Data Engineering and Communications Technologies, ISBN Springer Nature Singapore Pte Ltd. 2020;978-981-15-0977-3, 37:173-180
4. Ravi Kumar G, Venkata Sheshanna Kongara, Dr. G. A. Ramachandra, “An Efficient Ensemble Based Classification Techniques for Medical Diagnosis”, International Journal of Latest Technology in Engineering, Management and Applied Sciences, Volume II, Issue VIII, Pages: 5-9, ISSN-2278-2540, 2013.
5. Witten H, Frank E. “Data mining: practical machine learning tools and techniques with Java implementations”, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000.
6. Ian Witten H, Eibe Frank. Data Mining: Practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann, 2005.
7. Han J, Kamber M. Data Mining Concept and Technology, 3th ed., San Francisco: Morgan Kaufmann 2012;8:330-348.
8. Khourdifi Y, Bahaj M. ‘‘Applying best machine learning algorithms for breast cancer prediction and classification,’’ in Proc. Int. Conf. Electron., Control, Optim. Comput. Sci. (ICECOCS), 2018, 1-5.
9. UCI machine learning repository. http://archive.ics.uci.edu/ml/.