# INTERDISCIPLINARY PROJECT REPORT

**at**

**Sathyabama Institute of Science and Technology (Deemed to be University)**

Submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Computer Science and Engineering

By

## PETA LAKSHMI SUDHEER

**(Reg.No. 40110938)**

****

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING SCHOOL OF COMPUTING**

**SATHYABAMA**

**INSTITUTE OF SCIENCE AND TECHNOLOGY**

### (DEEMED TO BE UNIVERSITY)

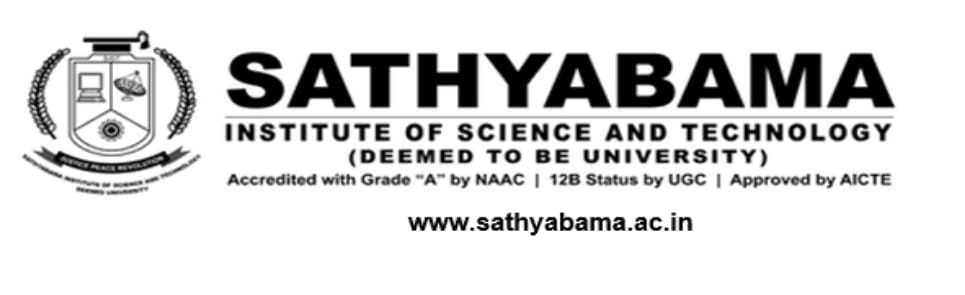
**Accredited with Grade “A” by NAAC | 12 B Status**

**by UGC | Approved by AICTE**

**JEPPIAR NAGAR, RAJIV GANDHISALAI,**

**CHENNAI – 600119**

**APRIL 2023**



|  |  |  |
| --- | --- | --- |
|  |  |  |



# DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

**BONAFIDE CERTIFICATE**

This is to certify that this Project Report is the Bonafide work of **PETA LAKSHMI SUDHEER (Reg. No: 40110938)** who carried out the project entitled **“Mammographic Mass DataSet using Decision Tree Classification on Machine Learning”** under my supervision from FEB 2023 to April 2023.

## Internal Guide

## Dr. GOPIKA, M.E., Ph.D.,

**Head of the Department**

**Dr. L.Lakshmanan, M.E., Ph.D.,**

## Submitted for Viva Voce Examination held on

**InternalExaminer ExternalExaminer**

**DECLARATION**

I, **PETA LAKSHMI SUDHEER** hereby declare that the project report entitled **Mammographic Mass Data using Decision Tree Classification on Machine Learning** done by me under the guidance of **Dr. GOPIKA, M.E.,Ph.D.,** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Computer Science and Engineering.

## DATE:

**PLACE: CHENNAI SIGNATURE OF THE CANDIDATE**

**ACKNOWLEDGEMENT**

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

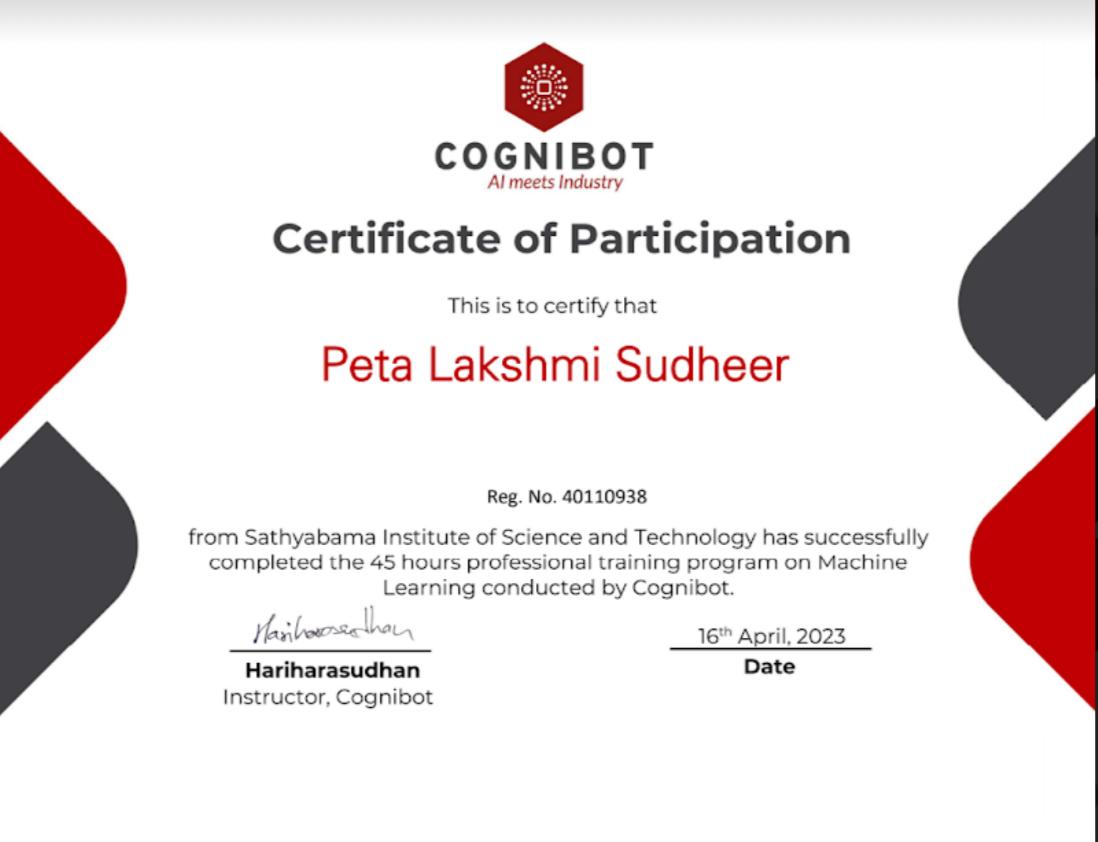
I convey my thanks to **Dr. T. Sasikala M.E., Ph.D**, **Dean**, School of Computing, **Dr. L. Lakshmanan, M.E., Ph.D., Heads of the Department** of **Computer Science and Engineering** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide

**Dr.GOPIKA**, **M.E., Ph.D.,** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

**TRAINING CERTIFICATE**



# ABSTRACT

Discrimination of benign and malignant mammographic masses based on BI-RADS attributes and the patient's age. Dataset characteristic are multivariate there are 961 instances, with 6 attributes . Based on the given characteristics of the given dataset, we need to predict the last variable by using the machine learning concept and the by using some methods like decision tree classification we can get the accuracy score,

so that we can find the last variable .Furthermore, a feature selection process can help to analyze the impact of the analytical tests,we used Decision Tree Algorithm in classification to find accuracy of the data .In this study we find the last variable by the Decision tree classification and find the accuracy by undergoing some other methods in machine learning.

i

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO** | **TITLE** | **PAGE NO** |
|  | ABSTRACT | i |
|  | LIST OF FIGURES | III |
|  | LIST OF TABLES | IV |
|  | LIST OF ABBREVATIONS | V |
| **1** | **INTRODUCTION** | **1** |
|  | 1.1 PROBLEM STATEMENT | 2 |
|  | 1.2 OBJECTIVE | 2 |
| **2** | **AIM , SCOPE AND LITERATURE SUMMARY** | 3 |
|  | 2.1 AIM | 3 |
|  | 2.2 SCOPE | 3 |
|  | 2.3 Literature summary | 3 |
| **3** | **DECISION TREE CLASSIFIER ALGORITHM** | 5 |
|  | 3.1 NAME OF THE ALGORITHM | 5 |
|  | 3.3 SOFTWARE AND LIBRARIES | 5 |
| **4** | **RESULTS AND DISCUSSION, PERFORMANCE ANALYSIS** | 6 |
|  | 4.1 RESULT AND DISCUSSION | 6 |
|  | 4.2 PERFORMANCE ANALYSIS | 7 |
| **5** | **SUMMARY AND CONCLUSIONS** | 8 |
|  | **REFERENCES** | 21 |
|  | **APPENDIX** | 9-20 |

**TABLE OF CONTENTS**

ii

# LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| **FIGURE NO** | **FIGURE NAME** | **PAGE NO** |
| 5.1 | Describing data set | 13 |
| 5.2 | Decision Tree Classifier | 16 |
| 5.3 | Confusion Matrix | 18 |
| 5.4 | Classification Report | 19 |
| 5.5 | Graph of Logistic Regression | 20 |

# 

# 

# 

# iii

# 

# List of Tables

# Table no Name Page No

# 4.1 Classification report 6

# 5.1 Screenshot of top 10 rows 11

iv

# LIST OF ABBREVIATIONS

**ABBREVIATION**  **EXPANSION**

ML Machine Learning

DTC Decision Tree Classification

MSE False Positive Rate

TPR True Positive Rate

ROC Receiver operating

Characteristic

v

**CHAPTER 1**

**INTRODUCTION**

Breast cancer is a life killing disease through the existence of debilitating growths Influencing ladies mostly after the age of 30 all over the world. Early diagnosis of the breast cancers by the radiologist reduces the death rate globally. Many techniques are available for the detection of breast cancers among which digital mammography is the familiar and successful technique currently used by the radiologists. Mammograms are collected from patients who are suspected for breast cancers mostly as full field mammograms where the image detection and classification are high due to the high image quality. Mammography cannot stop or decrease breast cancer can be supportive only in detecting the breast cancer at early stages to increase the survival rate.

Regular screening can be a successful strategy to identify the early symptoms of breast cancer in mammographic images. This examination also ensures other pathologies detection suggesting cancer nature as being benign, malignant, or normal. The most important improvement is breast imaging which is possible due to the advancement in digital mammography [6]. Medical images classification can play an important role in diagnostic and teaching purposes in medicine. It is also a form of data analysis that extracts models describing important data classes. Numerous methods have been created to classify masses into benign and malignant categories by using the different classification method. In the researchers proposed method aims to apply image mining for breast mammograms to detect and classify the cancerous tissue without any help of radiologist or medical specialist. A total of twenty-six features including GLCM features and histogram intensity features were extracted. A dataset of 961people mammographic masses based on BI-RADS attributes and the patient's age was collected In the researchers performed a comparative study on the performance of binary classifiers. They have used dataset with 6 attributes and not the breast tissue dataset.

Moreover, they have not brought out the effect of feature selection in classification. Their experimental study was restricted to four classification algorithms viz. ID3, C4.5, K Nearest Neighbours (K-NN) and Support Vector Machines (SVM). Classification methods are one of the most fundamental and important tasks in mining and machine learning. Many of the researchers performed experiments on medical datasets using decision tree classifiers. The aim of the study is to determine the best decision tree classifier for medical datasets classification. In researchers analysed the performance of decision tree classifiers on various medical datasets in terms of accuracy and time complexity which proved that CART is the best. More recent research presented in concerned the identification of breast cancer patients for whom chemotherapy

**1.1 Problem Statement :**

* Mammography is the most effective method for breast cancer screening available today. However, the low positive predictive value of breast biopsy resulting from mammogram interpretation leads to approximately 70% unnecessary biopsies with benign outcomes. To reduce the high number of unnecessary breast biopsies, several computer-aided diagnosis (CAD) systems have been proposed in the last years. These systems help physicians in their decision to perform a breast biopsy on a suspicious lesion seen in a mammogram or to perform a short term follow-up examination instead.

**1.2 Objective:**

* This project deals with classifying Mammographic masses as benign or malignant using different Machine Learning algorithms including SVM, Logistic Regression, Decision Trees, Naive Bayes, Artificial Neural Network and many more.
* ROC curves are plotted for each to identify the best classification algorithm for the problem

**CHAPTER 2**

**AIM , SCOPE AND LITERATUTE SUMMARY**

**2.1 AIM :**

* The objective of the assignment is to train a machine learning algorithm to predict the last variable data in mammogram classification by using some machine learning classifications .

**2.2 SCOPE :**

* This prediction can help the doctor to find the missing values of the attributes so he can suggest the necessary steps to overcome breast cancer.

**2.3** **LITERATURE SUMMARY:**

Breast cancer is one of the most common forms of cancer among women. Early detection of breast cancer is essential for effective treatment and higher survival rates. Mammography is the most commonly used screening tool for breast cancer, and it involves the use of X-rays to produce images of the breast. However, mammography images can be difficult to interpret, and radiologists often rely on visual analysis to detect abnormalities.

In recent years, machine learning algorithms have been developed to aid in the interpretation of mammography images. These algorithms can analyze large amounts of data and identify patterns that may be difficult for a human observer to detect. This literature summary will discuss the use of machine learning algorithms for the classification of mammography masses.

One study by Wang et al. (2020) used a deep learning algorithm to classify mammography masses as benign or malignant. The algorithm was trained on a dataset of 12,753 mammography images and achieved an accuracy of 87.7% on an independent test set of 1,510 images. The authors noted that the deep learning algorithm outperformed traditional machine learning algorithms, such as support vector machines and random forests.

Another study by Wang et al. (2019) used a convolutional neural network (CNN) to classify mammography masses as benign, malignant, or normal. The CNN was trained on a dataset of 5,470 mammography images and achieved an accuracy of 89.1% on an independent test set of 2,738 images. The authors noted that the CNN outperformed traditional machine learning algorithms, such as logistic regression and decision trees.

A study by Bria et al. (2017) used a support vector machine (SVM) to classify mammography masses as benign or malignant. The SVM was trained on a dataset of 266 mammography images and achieved an accuracy of 84.6% on an independent test set of 102 images. The authors noted that the SVM was able to identify subtle differences between benign and malignant masses that were difficult for human observers to detect.

In conclusion, machine learning algorithms have shown promise in the classification of mammography masses. Deep learning algorithms, such as convolutional neural networks, have been shown to outperform traditional machine learning algorithms, such as support vector machines and random forests. With further development and refinement, these algorithms may eventually be integrated into clinical practice to aid radiologists in the detection of breast cancer.

**CHAPTER 3**

**DECISION TREE CLASSIFIER ALGORITHM USED**

**3.1** **NAME OF THE ALGORITHM :**

* The algorithm we used here is DECISION TREE – CLASSIFICATION.
* To determine the feature importance, we used feature selection method.

**3.2** **DECISION TREE – CLASSIFICATION :**

Decision tree classification is a machine learning algorithm used for supervised learning tasks, specifically for classification problems. It builds a tree-like model of decisions and their possible consequences. The algorithm starts with a single node and then recursively splits the data into subsets based on the most significant feature that differentiates them, until each subset is as homogeneous as possible in terms of the target variable.

The decision tree is built by selecting the attribute that gives the highest information gain, which is a measure of the degree to which a given attribute reduces the entropy (uncertainty) in the data set. The goal is to create a tree that generalizes well to new data by avoiding overfitting, which occurs when the tree is too complex and fits the training data too closely.

Once the decision tree is built, it can be used to predict the target variable for new input data by following the path through the tree that corresponds to the values of the input features. The leaves of the tree correspond to the predicted class for that input.

Decision trees have several advantages, including being interpretable, easy to understand and visualize, and able to handle both numerical and categorical data. However, they can be prone to overfitting, and small changes in the data can lead to significant changes in the resulting tree.

**3.3 Software and Libraries :**

* The software used is Jupyter Notebook from Anaconda 3.
* Libraries used in this project are pandas, sklearn, matplotlib, and seaborn

**CHAPTER 4**

**RESULTS AND DISCUSSION, PERFORMANCE ANALYSIS**

# 4.1 RESULTS AND DISCUSSIONS :

**Results**:

We got the last variable accuracy using machine learning methods like decision tree classification ,Logisitic regression

From the predicted accuracy score the doctor can suggest the best

methods for treatment .

Decision tree classification method were directly applied to the given data set used in the mammogram classification. In addition, we also use confusion matrix, feature selection and logistic regression for the perfect accuracy score.

We also got the classification report it was shown below in the table 4.1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.86 | 0.84 | 0.85 | 85 |
| 1 | 0.83 | 0.85 | 0.84 | 81 |
| accuracy |  |  | 0.84 | 166 |
| macro avg | 0.84 | 0.84 | 0.84 | 166 |
| weighted avg | 0.84 | 0.84 | 0.84 | 166 |

# Table4.1: Classification report

# DISCUSSIONS:

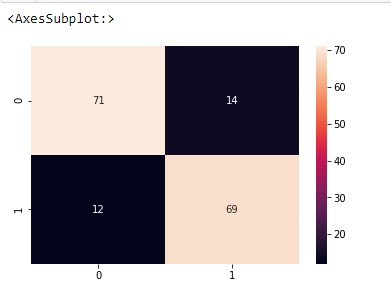
Mammography is the most effective method for breast cancer screening available today. However, the low positive predictive value of breast biopsy resulting from mammogram interpretation leads to approximately 70% unnecessary biopsies with benign outcomes. To reduce the high number of unnecessary breast biopsies, several computer-aided diagnosis (CAD) systems have been proposed in the last years. These systems help physicians in their decision to perform a breast biopsy on a suspicious.

This data set can be used to predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age.It contains a BI-RADS assessment, the patient's age and three BI-RADS attributes

together with the ground truth (the severity field) for 516 benign and 445 malignant masses that have been identified on full field digital mammograms collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006.Each instance has an associated BI-RADS assessment ranging from 1 (definitely benign) to 5 (highly suggestive of malignancy) assigned in a double-review process by physicians. Assuming that all cases with BI-RADS assessments greater or equal a given value (varying from 1 to 5), are malignant and the other cases benign,sensitivities and associated specificities can be calculated. These can be an indication of how well a CAD system performs compared to the radiologists.

**4.2 PERFORMANCE ANALYSIS:**

From the graphs they can able to analyse the Receiver Operating Characteristic We can also use the confusion matrix to find the accuracy



**CHAPTER 5**

**SUMMARY AND CONCLUSIONS**

* Mammography is the most reliable and the most effective technique for detection of breast cancer at initial stage. The manual reading of mammograms, results in variation of 65% to 85% in diagnosis because of a number of reasons like; inadequate contrast, experience of radiologist, and other human factors. This may results into missed lesions or unnecessary biopsy tests. So computer aided diagnosis has come up as an aid to radiologist and can be used for having the second opinion. In this paper a methodology for classification of benign mass and malignant mass has been proposed. Hierarchical centroid method which iteratively subdivides image is proposed to be used along with other shape based features. Various shape features and Haralick texture features are extracted for each region of interest. Differential Evolution based Feature Selection method has been used for feature subset selection. Using Support Vector Machine as classifier, an accuracy of 85% has been achieved.
* Each instance has an associated BI-RADS assessment ranging from 1 (definitely benign) to 5 (highly suggestive of malignancy) assigned in a double-review process by physicians. Assuming that all cases with BI-RADS assessments greater or equal a given value (varying from 1 to 5), are malignant and the other cases benign, sensitivities and associated specificities can be calculated. These can be an indication of how well a CAD system performs compared to the radiologists.
* Class Distribution: benign: 516; malignant: 445

**APPENDIX**

**5.1 DATA SET**

Mammography is the most effective method for breast cancer screening available today. However, the low positive predictive value of breastbiopsy resulting from mammogram interpretation leads to approximately 70% unnecessary biopsies with benign outcomes. To reduce the high number of unnecessary breast biopsies, several computer-aided diagnosis (CAD) systems have been proposed in the last years. These systems help physicians in their decision to perform a breast biopsy on a suspicious lesion seen in a mammogram or to perform a short term follow-up examination instead.

This data set can be used to predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age. It contains a BI-RADS assessment, the patient's age and three BI-RADS attribute together with the ground truth (the severity field) for 516 benign and 445 malignant masses that have been identified on full field digital mammograms collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006

Each instance has an associated BI-RADS assessment ranging from 1 (definitely

benign) to 5 (highly suggestive of malignancy) assigned in a double-review

process by physicians. Assuming that all cases with BI-RADS assessments

greater or equal a given value (varying from 1 to 5), are malignant and the other

cases benign, sensitivities and associated specificities can be calculated. These

can be indication of how well a CAD system performs compared to the

radiologists.

Class Distribution: benign: 516; Malignant: 445

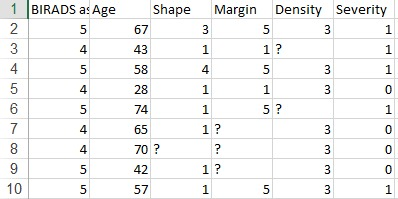
how well a CAD system performs compared to the radiologists.

Class Distribution: benign: 516; malignant: 445

It can be used to predict the severity (benign or malignant) of a mammographic mass from BI-RADS attributes and the patient’s age. Number of Attributes: 6 (1 goal field: severity, 1 non-predictive: BI-RADS, 4 predictive attributes

# Attribute Information:

1. BI-RADS assessment: 1 to 5 (ordinal)
2. Age: patient’s age in years (integer)
3. Shape (mass shape): round=1, oval=2, lobular=3, irregular=4 (nominal)
4. Margin (mass margin): circumscribed=1, microlobulated=2, obscured=3, ill-defined=4, spiculated=5 (nominal)
5. Density (mass density): high=1, iso=2, low=3, fat-containing=4 (ordinal)
6. Severity: benign=0 or malignant=1 (binomial)



**Table 5.1 Screenshot of top 10 rows of the datasets**

#### **Missing Attribute Values:**

#### - BI-RADS assessment: 2

#### - Age: 5

#### - Shape: 31

#### - Margin: 48

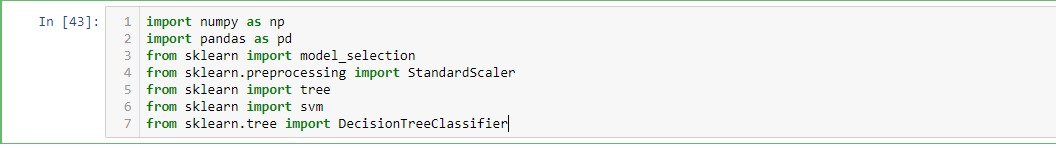
#### - Density: 76

#### - Severity: 0

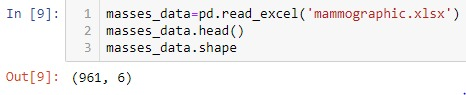
# 5.2 PREPROCESSING

Data normalization is one of the preprocessing approaches where the data is scaled or transformed to obtain an equal contribution from each characteristic, thus translating into a significant improvement in the performance of Machine Learning algorithms. In this work, they contain numerical datasets in integers values.

Create a new IPython Notebook and insert the below code to import the necessary modules. In case you get any error, do install the necessary packages using pip



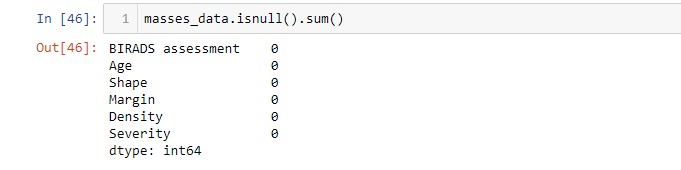
The given data set which is the form of xcel sheet is imported .Read the data using pandas into a dataframe. To check the top 5 rows of the dataset, use df.head(). You can specify the number of rows as an argument to this function in case you want to check different number of rows. BI-RADS attribute has been given as non-predictive in the dataset and so it won’t be taken into consideration.



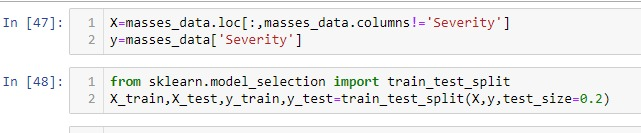
You can get a description of the data like values of count, mean, standard deviation as masses\_data\_describe().

# fig5.2:Describing data set

As you might have observed, there are missing values in the dataset. Handling missing data is something very important in data preprocessing. We fill out the empty values using the mean or mode of the column depending on the data analysis. For simplicity, as of now, you can drop the null values from the data



# From sklearn model selection import train\_test\_split , Create training and testing set using train\_test\_split. 20% of the data is used for testing and 80% for training.



# 5.3 CLASSIFICATION

**Decision Trees (DTs)** are a non-parametric supervised learning method used for [classification](https://scikit-learn.org/stable/modules/tree.html#tree-classification) and [regression](https://scikit-learn.org/stable/modules/tree.html#tree-regression). The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

# Some advantages of decision trees are:

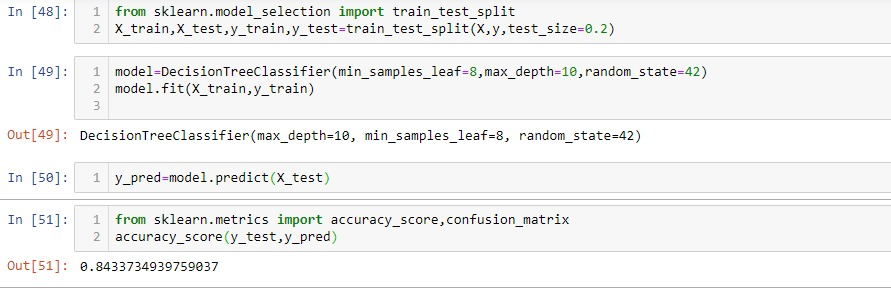
* Simple to understand and to interpret. Trees can be visualise
* Enquires little data preparation. Other techniques often require data normalisation, dummy variables need to be created and blank values to be removed. Note however that this module does not support missing values.
* The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree.
* Able to handle both numerical and categorical data. However scikit-learn implementation does not support categorical variables for now. Other techniques are usually specialised in analysing datasets that have only one type of variable. See algorithm for more information.
* Able to handle multi-output problems.

**The disadvantages of decision trees include:**

* Decision-tree learners can create over-complex trees that do not generalise the data well. This is called overfitting. Mechanisms such as pruning, setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree are necessary to avoid this problem.
* Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This problem is mitigated by using decision trees within an ensemble.
* Predictions of decision trees are neither smooth nor continuous, but piecewise constant approximations as seen in the above figure. Therefore, they are not good at extrapolation.
* The problem of learning an optimal decision tree is known to be NP-complete under several aspects of optimality and even for simple concepts. Consequently, practical decision-tree learning algorithms are based on heuristic algorithms such as the greedy algorithm where locally optimal decisions are made at each node. Such algorithms cannot guarantee to return the globally optimal decision tree. This can be mitigated by training multiple trees in an ensemble learner, where the features and samples are randomly sampled with replacement.

[Decision Tree Classifier](https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier) is a class capable of performing multi-class classification on a dataset.

To build a Decision Tree Classifier from the training set, we just need to use the function DecisionTreeClassifier() It has a certain number of parameters about which you can find on the scikit-learn documentation. For now, we would just use the default values of each parameter. Use predict() on the test input features X\_test to get the predicted values y\_pred. The function score() can be used directly to compute the accuracy of prediction on test samples.

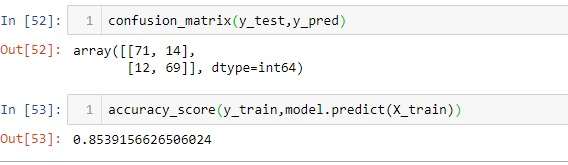


**Fig5.2 Decision tree Classifier**

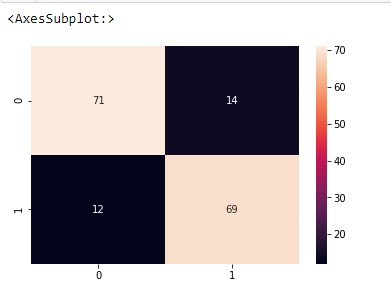
Here we got the accuracy\_score 0.843373. The DecisionTreeClassifier() without any tuning gives a result around 82% which we can say is not the worst.

# 5.4 Confusion matrix:

A confusion matrix is a table that is often used to describe the performance of a classification model(or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.



Here we find the accuracy score using the confusion matrix, we also displayed it in the below figure.



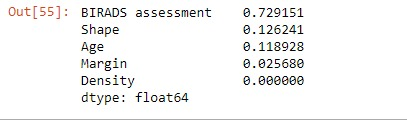
**Fig 5.3confusion matrix**

# 5.5 Feature selection:

**Feature selection** is the process of reducing the number of input variables when developing a predictive model.

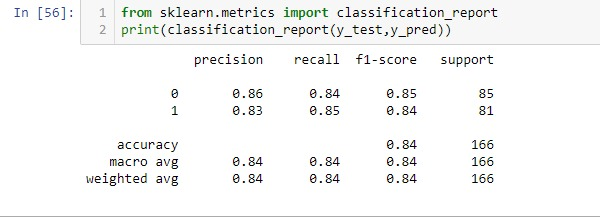
It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model.

We can see the feature score of the attributes below



# 5.6 Classification report:

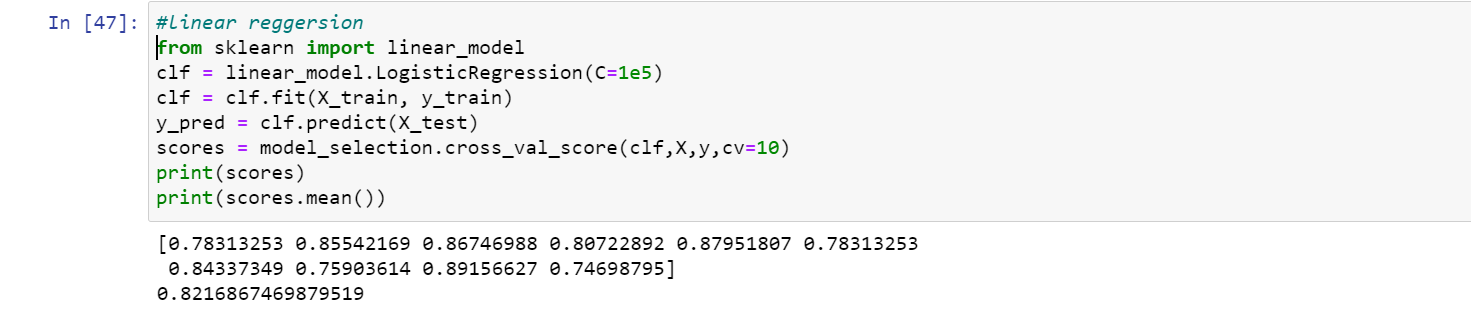
We gather all the attributes and forms a classification\_report of the missing attributes.



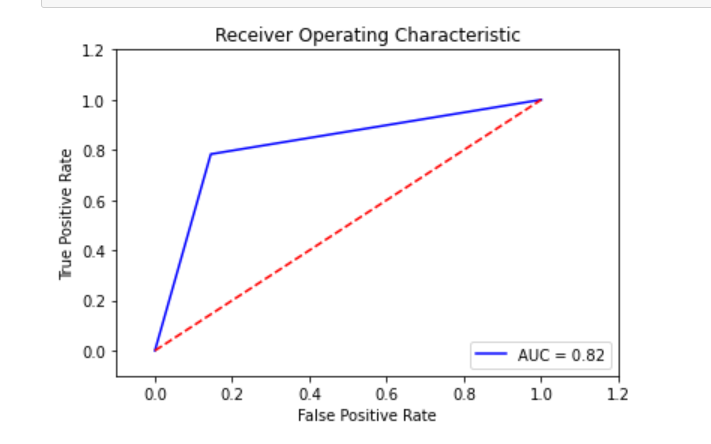
# Fig:5.4 classification Report

# 5.7 Logistic Regression:

In the multiclass case, the training algorithm uses the one-vs-rest (OvR) scheme if the ‘multi\_class’ option is set to ‘ovr’, and uses the cross-entropy loss if the ‘multi\_class’ option is set to ‘multinomial’. This class implements regularized logistic regression using the ‘liblinear’ library, ‘newton-cg’, ‘sag’, ‘saga’ and ‘lbfgs’ solvers. It can handle both dense and sparse input. Use C-ordered arrays or CSR matrices containing 64-bit floats for optimal performance; any other input format will be converted.Similar to the Decision Tree Classifier, we can also create Logistic Regression classifier. The function Logistic Regression() is used. The classifier is fitted on training set and similarly used to predict target values for the test set. It gives a mean score of 82%.Logistic Regression is used ,from sklearn we import linear\_model and print the score.mean(). Cross\_val\_score() function evaluates score uses cross-validation method. Cross-validation is used to avoid any kind of overfitting.



* We plot a graph with Receiver Operating Characteristic with True Positive Rate on Y-axis and False Positive Rate on X\_axis

**Fig5.5Graph of logistic regression**

**B.SOURCE CODE :**

#IMPORT LIBRARIRES

import numpy as np

import pandas as pd

from sklearn import model\_selection

from sklearn.preprocessing import StandardScaler

from sklearn import tree

from sklearn import svm

from sklearn.tree import DecisionTreeClassifier

#loading the dataset

masses\_data=pd.read\_excel('mammographic.xlsx')

#Getting the values from the date set

masses\_data.head()

#shape the dataset i.e no.of rws and columns

masses\_data.shape

#Describing the data set

masses\_data.describe()

masses\_data=masses\_data[(masses\_data!='?').all(axis=1)]

masses\_data.shape

#checking the null values

masses\_data.isnull().sum()

X=masses\_data.loc[:,masses\_data.columns!='Severity']

y=masses\_data['Severity']

from sklearn.model\_selection import train\_test\_split

X\_train,X\_test,y\_train,y\_test=train\_test\_split(X,y,test\_size=0.2)

model=DecisionTreeClassifier(min\_samples\_leaf=8,max\_depth=10,random\_state=42)

model.fit(X\_train,y\_train)

y\_pred=model.predict(X\_test)

from sklearn.metrics import accuracy\_score,confusion\_matrix

accuracy\_score(y\_test,y\_pred)

confusion\_matrix(y\_test,y\_pred)

accuracy\_score(y\_train,model.predict(X\_train))

import seaborn as sns

sns.heatmap(confusion\_matrix(y\_test,y\_pred),annot=True)

feature\_scores=pd.Series(model.feature\_importances\_,index=X\_train.columns).sort\_values(ascending=False)

feature\_scores

from sklearn.metrics import classification\_report

print(classification\_report(y\_test,y\_pred))

import matplotlib.pyplot as plt

from sklearn import metrics

def createROC(y\_test, y\_pred):

fpr, tpr, thresholds = metrics.roc\_curve(y\_test, y\_pred)

roc\_auc = metrics.auc(fpr,tpr)

plt.title('Receiver Operating Characteristic')

plt.plot(fpr, tpr, 'b',label='AUC = %0.2f'% roc\_auc)

plt.legend(loc='lower right')

plt.plot([0,1],[0,1],'r--')

plt.xlim([-0.1,1.2])

plt.ylim([-0.1,1.2])

plt.ylabel('True Positive Rate')

plt.xlabel('False Positive Rate')

plt.show()

createROC(y\_test, y\_pred)

tree.plot\_tree(model);

#linear reggersion

from sklearn import linear\_model

clf = linear\_model.LogisticRegression(C=1e5)

clf = clf.fit(X\_train, y\_train)

y\_pred = clf.predict(X\_test)

scores = model\_selection.cross\_val\_score(clf,X,y,cv=10)

print(scores)

print(scores.mean())

createROC(y\_test, y\_pred)

# REFERENCES:

[1] Cognibot Machine Learning classes by Hari.

[2] Data\_reference\_from:

<http://archive.ics.uci.edu/ml/datasets/mammographic+mass>

<https://www.sciencedirect.com/science/article/pii/S2352340920308222>

[3] Pareek, A. and S.M. Arora, Breast cancer detection techniques using medical image processing. Breast cancer, 2017.2(3).

[4] S. Punitha, S. Ravi, et al. Breast Cancer Detection using Classification Techniques in Digital Mammography: International Science Press, I J C T A, 9(7), 2016, pp. 3123-3134,ISSN: 0974- 5572.Curtis, C., et al., The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature, 2012. 486(7403): p.346-352.

[5] M. Hejmadi, Introduction to Cancer Biology, Bookboon, London, UK, 2nd edition, 2010.

[6] American Cancer Society, Breast Cancer Facts and Figures 2017-2018, American Cancer Society, Atlanta, GA, USA, 2017.

[7] American Cancer Society, Breast Cancer Facts and Figures 2019, American Cancer Society, Atlanta, GA, USA, 2019.

[8] O. Ginsburg, F. Bray, M. P. Coleman et al., “The global burden of women’s cancers: a grand challenge in global health,” The Lancet, vol. 389, no. 10071, pp. 847–860, 2017.View at: [Publisher Site](https://doi.org/10.1016/s0140-6736(16)31392-7) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=The%20global%20burden%20of%20women%E2%80%99s%20cancers:%20a%20grand%20challenge%20in%20global%20health&author=O.%20Ginsburg&author=F.%20Bray&author=M.%20P.%20Coleman%20et%20al.&publication_year=2017)