BREAST CANCER PREDICTION

USING

PYTHON WITH MACHINE LEARNING

An internship report submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

S.Preethi 1215316662

Under the esteemed guidance of

V.S.L.Prasad

Python developer in Grepthor Software Solutions Pvt.Ltd

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

GITAM

(Deemed to be University)

VISHAKAPATNAM

MAY-JUNE 2019

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GITAM INSTITUTE OF TECHNOLOGY

GITAM

(Deemed to be University)

DECLARATION

We, hereby declare that the internship review entitled "**BREAST CANCER PREDICTION USING PYTHON WITH MACHINE LEARNING**" is an original work done in the Department of Computer Science and Engineering, GITAM Institute of Technology, GITAM (Deemed to be University) submitted in partial fulfillment of the requirements for the award of the degree of B.Tech. in Computer Science and Engineering.

The work has not been submitted to any other college or University for the award of any degree or diploma.

Date:

1215316662 **PREETHI SANA**

GREPTHOR
SOFTWARE SOLUTIONS PVT LTD

Dear Sana Preethi ,

Greetings from **GREPTHOR SOFTWARE SOLUTIONS!**

Grepthor Software Solutions takes great pride in our ability to attract, recruit, and develop world-class talent. Our internship program, comprised of rising juniors and seniors in college, exemplifies our commitment to premier talent acquisition, and it provides our own internal aspiring talent the opportunity to recruit students from across the globe and leverage their experience to benefit both personal and professional growth. With an extensive list of functional areas to choose from including Java, Python, Web Development, Android Development, and Digital marketing, the program is designed to help students find their niche.

In reference to your application we would like to congratulate you on being selected for internship on **"Machine learning"** with **GREPTHOR SOFTWARE SOLUTIONS PVT LTD** based at the following address. Your internship is scheduled to start effectively from **May 6th,2019** for a period of 45 days. All of us at Grepthor Software Solutions are excited that you will be joining our team! As such, your internship will include training/orientation and focus primarily on learning and developing new skills and gaining a deeper understanding of concepts through hands-on application of the knowledge you learned in class.

The project details and technical platform will be shared with you on or before commencement of training.

You should report for training at the following address:

Reporting Office Address:

Plot No:170/67,171, 1st Floor Behind Cyber Towers Balaji Empire, Patrika Nagar, Madhapur, Hyderabad, Telangana 500081

Contact Person: HR

Contact Number- 7658975704

Email: info@grepthorsoftware.com

Again, congratulations and we look forward to working with you.

Yours sincerely,

GREPTHOR SOFTWARE SOLUTIONS PVT LTD.

<Signature Authority Name>

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GITAM INSTITUTE OF TECHNOLOGY

GITAM

(Deemed to be University)

CERTIFICATE

This is to certify that the internship report entitled "**BREAST CANCER PREDICTION USING PYTHON WITH MACHINE LEARNING**" is a bonafide record of work carried out by **PREETHI SANA (1215316662)** students submitted in partial fulfillment of requirement for the award of degree of Bachelors of Technology in Computer Science and Engineering.

SUPERVISOR **INTERNSHIP REVIEWER**

Mr . V.S.L.Prasad **Dr. P.Anuradha**

ACKNOWLEDGEMENTS:

The internship opportunity I had with Grepthor Software Solutions Pvt Ltd was a great chance for learning and professional development. Therefore, I consider myself a very lucky individual as I was provided with an opportunity to be a part of it. Bearing in mind firstly I am using this opportunity to express my deepest gratitude and special thanks to Mr. Prasad at Grepthor Software Solutions Pvt Ltd who inspite of being extraordinarily busy with his duties, took time out to hear, guide and keep me on the correct path and allowing me to carry out my work at their esteemed organization.

I express my deepest thanks to Mr. Prasad for taking part in useful decisions & giving necessary advice and guidance, as mentors. I choose this moment to acknowledge their contribution gratefully.

I express my deepest thanks to Dr. Konala Thammi Reddy sir, HOD of C.S.E department, Gandhi Institute of Management and Technology(GITAM) for giving me a great opportunity to complete my internship in the company. I choose this moment to acknowledge his contribution too gratefully.

I would also like to thank Dr. P. Anuradha, A.M.C who helped us a lot in the successful completion of our internship and its report.

I perceive this opportunity as a big milestone in my career development. I will strive to use these gained skills and knowledge in the best possible way, and I will continue to work on their improvement, in order to attain desired career objectives. Hope to continue cooperation with all of you in the future.

Sincerely,

PREETHI SANA

TABLE OF CONTENTS

1. Abstract 6

2. About Organization 6

3. Schedule of the Internship(Training) 7

4. Internship Activities 9

4.1.Training 9

4.2.Role in Application Development 10

4.3.Methodology and functionalities 11

5. Outcomes 19

6. Assessment of Internship 22

7. References 23

1. ABSTRACT:

Cancer is the second cause of death in the world. 8.8 million patients died due to cancer in 2015. Breast cancer is the leading cause of death among women. Several types of research have been done on early detection of breast cancer to start treatment and increase the chance of survival. Most of the studies concentrated on mammogram images. However, mammogram images sometimes have a risk of false detection that may endanger the patient's health. It is vital to find alternative methods which are easier to implement and work with different data sets, cheaper and safer, that can produce a more reliable prediction.

This project proposes a hybrid model combined of several Machine Learning (ML) algorithms including Logistic Regression, Decision Tree (DT) for effective breast cancer detection. This study also discusses the datasets used for breast cancer detection and diagnosis. The proposed model can be used with different data types such as image, blood, etc.

2. ABOUT ORGANIZATION:

Grepthor is a customer oriented company offering Mobile applications, Software Development, Web Development, and Business Development Services in various technologies.

For the welfare of its Clients, Grepthor also additionally steps up for providing Digital promoting, Organic SEO, Ads on Google, Face book, LinkedIn and Various alternatives.

Founded in 2013 by three determined minds, Grepthor family has fully grown up to 700+ clients all over in India. Being a young IT body, Grepthor has clients in India and US.

Grepthor is the fastest growing offering win-to-win services in its lowest cost. We believe in offering services for building relationships rather than for making money. Clients will see crystal clear team working, proceedings, right after the project started.

3. SCHEDULE OF INTERNSHIP AND TRAINING:

The project assigned to me is based on Machine Learning Using Python. The joining date for the internship is 2/05/2019 and is scheduled to 15/06/2019.

Week 1:

- Introduction to Python Scripting.
- Platforms Used in Python Scripting.
- Python Software Installation.
- Python Applications in IDLE.
- Python Datatypes.

Week 2:

- Exception Handling in Python Scripting.
- String Handling.
- Python Modules.
- Functions in Python Scripting.
- Oops Application Configurations in Python Scripting.

Week 3:

- Python Variables and Methods in Class and Objects.
- Constructors.
- Working with dir functions and help functions.
- Database Configurations.
- Introduction of Python with Machine learning Implementations.

Week 4:

- Python Idle Installation.
- Working With Datasets using CSV.
- Working With Pandas Module.

Week 5:

- Working With Numpy Module.
- Working With Mat plot lib Module.
- Learning about Decision Algorithm.
- Learning about Logistic Regression.

Week 6:

- Working With Project Coding and Data Sets.

Week 7:

- Developing the code with the Data Sets using Machine Learning algorithms and finally finding out the accuracy of the people who are suffering with the cancer.

4. INTERNSHIP ACTIVITES

4.1. TRAINING:

INTRODUCTION:

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

Classification and data mining methods are an effective way to classify data. Especially in medical field, where those methods are widely used in diagnosis and analysis to make decisions.

There are some guidelines:

Mammography: The most important screening test for breast cancer is the mammogram. A mammogram is an X-ray of the breast. It can detect breast cancer up to two years before the tumor can be felt by you or your doctor**.** Women with age 40–45 or older who are at average risk of breast cancer should have a mammogram once a year. Women at high risk should have yearly mammograms along with an MRI starting at age 30.

PROBLEM STATEMENT:

Predict whether a tumour is *malignant(cancer)* or *benign(non Cancer)* based on two features,the mean radius of the tumour (radius mean) and its mean number of concave points (concave points mean).

The chances of an individual of getting a disease, such as breast cancer are more. Present women with breast cancer have no apparent risk factors. Some factors need to be known by women so that they can overcome their risk of breast cancer. Since causes of breast cancer are not fully known, Researchers believe that these risk factors increase (or decrease) the changes of developing breast cancer. Since breast cancer is a complex disease it is likely to be caused by a combination of risk factors. Some of the factors associated with breast cancer can't be changed (Non-preventable) like age, genetic factor, heredity. While making choices can change other factors (Preventable) like overweight, lack of exercises, smoking, harmone replacement therapy.

REVIEW OF LITERATURE:

A literature review showed that there have been several studies on the survival prediction problem using statistical approaches and artificial neural networks. However, we could only find a few studies related to medical diagnosis and recurrence using data mining approaches such as decision trees. Delen used artificial neural networks, decision trees and logistic regression to develop prediction models for breast cancer survival by analysing a large dataset, the SEER cancer incidence database. Lundin et al. used ANN and logistic regression models to predict 5, 10, and 15 -year breast cancer survival. They studied 951 breast cancer patients and used tumor size, axillary nodal status, histological type, mitotic count, tubule formation, and age as input variables . Pendharker et al. used several data mining techniques for exploring interesting patterns in breast cancer. In this study, they showed that data mining could be a valuable tool in identifying similarities (patterns) in breast cancer cases, which can be used for diagnosis, prognosis, and treatment purposes. These studies are some examples of researches that apply data mining to medical fields for prediction of diseases.

DATA COLLECTION

Data collection is the process of gathering and measuring information from the patient in the organization or from any other organization, in an established systematic fashion that enables one to answer stated research questions, test hypothesis, and evaluate outcomes. The data collection component of research is common to all fields of study including physical and social sciences, humanities, business, etc. While methods vary by discipline, the emphasis on ensuring accurate and honest collection remains the same. The data collection for cancer patients can be determined by considering the id type.

4.2 ROLE IN THE INTERNSHIP:

As a Machine Learning Trainee-cum-Intern, I was initially trained by our guide Mr.Prasad, in different concepts of Python for the first 2 weeks using the ‘IDLE’ or python version platform. In the third week, I was given access to Anaconda Jupiter installation, to deploy the programs and for predicting data. In the same week, all the interns were divided into teams of three. My teammates were B.Amuktha, G.Sannihitha. We were assigned with the project - “Breast Cancer Prediction Using Machine Learning”.

I was an active member of the group and participated in all the phases of the project, from research to deployment. All of us were instrumental in the completion of the project with an equal contribution at every stage. On the last day of my internship, our group had done a presentation with respect to our project. Our presentation was met with immense applause which concluded the sixth-week journey on a very happy note.

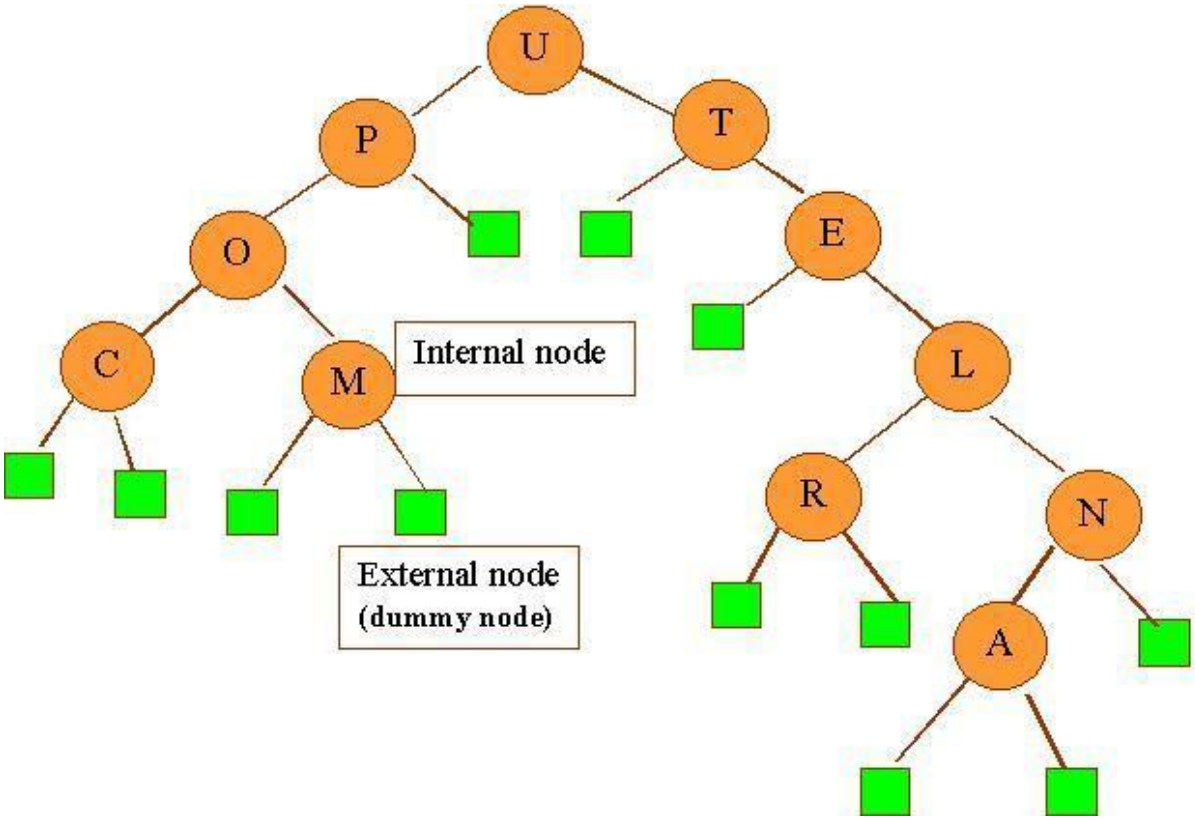
4.3. METHODOLOGIES AND FUNCTIONALTIES:

This project aims to develop a good classifier model to offer accurate prediction of breast cancer. In order to do this, we need some previous data sets .There are various attributes and features that have impact set of data leading to complexity in analyzing it. The focus of this research is to build such a model which has the capability of dealing with high complexity and gives accurate results irrespective of the magnitude of data set . The dataset used in research is publicly available and was created by Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin, USA.

The research includes the evaluation of features like:
Radius,Texture,Perimeter,Area,Smoothness,Compactness,Concavity,Concave points,Symmetry,Fractal dimension.Once the data collection was over,we proceeded data using decision tree classification technique for the prediction.

Decision Tree Classifier:

Decision trees are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter. The tree can be explained in two entities, namely decision nodes and leaves. The leaves are the decision or final outcomes and the decision nodes are the place where the data is split.



Phase 1:Data Preparation:

The dataset used in this research is publicly available and was created by Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin, USA.

Attribute Information:

1. ID number 2) Diagnosis (M = malignant, B = benign)

Ten real-valued features are computed for each cell nucleus:

1. radius (mean of distances from center to points on the perimeter)
2. texture (standard deviation of gray-scale values)
3. perimeter
4. area
5. smoothness (local variation in radius lengths)
6. compactness (perimeter² / area—1.0)
7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry

10. fractal dimension (“coastline approximation”—1)

1.read_csv( );

read_csv is an important pandas function to read csv files and do operations on it.

```
df = pd.read_csv("Wisconsin_Breast_Cancer_Dataset.csv")
```

2.info( ):

info( ) function is used to get a concise summary of the data frame. It comes really handy when doing exploratory analysis of the data.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 33 columns):
id                         569 non-null int64
diagnosis                  569 non-null object
radius_mean                569 non-null float64
texture_mean               569 non-null float64
perimeter_mean             569 non-null float64
area_mean                  569 non-null float64
smoothness_mean            569 non-null float64
compactness_mean           569 non-null float64
concavity_mean             569 non-null float64
concave points_mean        569 non-null float64
symmetry_mean              569 non-null float64
fractal_dimension_mean     569 non-null float64
radius_se                  569 non-null float64
texture_se                 569 non-null float64
```

Phase 2: Data Exploration:

We will be using **Jupiter** to work on this dataset. We will first go with importing the necessary libraries like Numpy, Pandas, and Matplot and import our dataset to Jupiter.

Radius-Mean:

Radius -Mean of distances from center to points on the perimeter.

Concave Points-Mean:

Concave Points -Number of concave portions of the contour.

Diagnosis:

Diagnosis determines whether a person’s cancer type is Malignant or Benign.

```
X = df[["radius_mean","concave points_mean"]]
y=df["diagnosis"]
```

```
X[15:25]
```

	radius_mean	concave points_mean
15	14.540	0.07364
16	14.680	0.05259
17	16.130	0.10280
18	19.810	0.09498
19	13.540	0.04781
20	13.080	0.03110
21	9.504	0.02076
22	15.340	0.09756
23	21.160	0.08632
24	16.650	0.09170

```
y[15:25]
```

```
15    M
16    M
17    M
18    M
19    B
20    B
21    B
22    M
23    M
24    M
Name: diagnosis, dtype: object
```

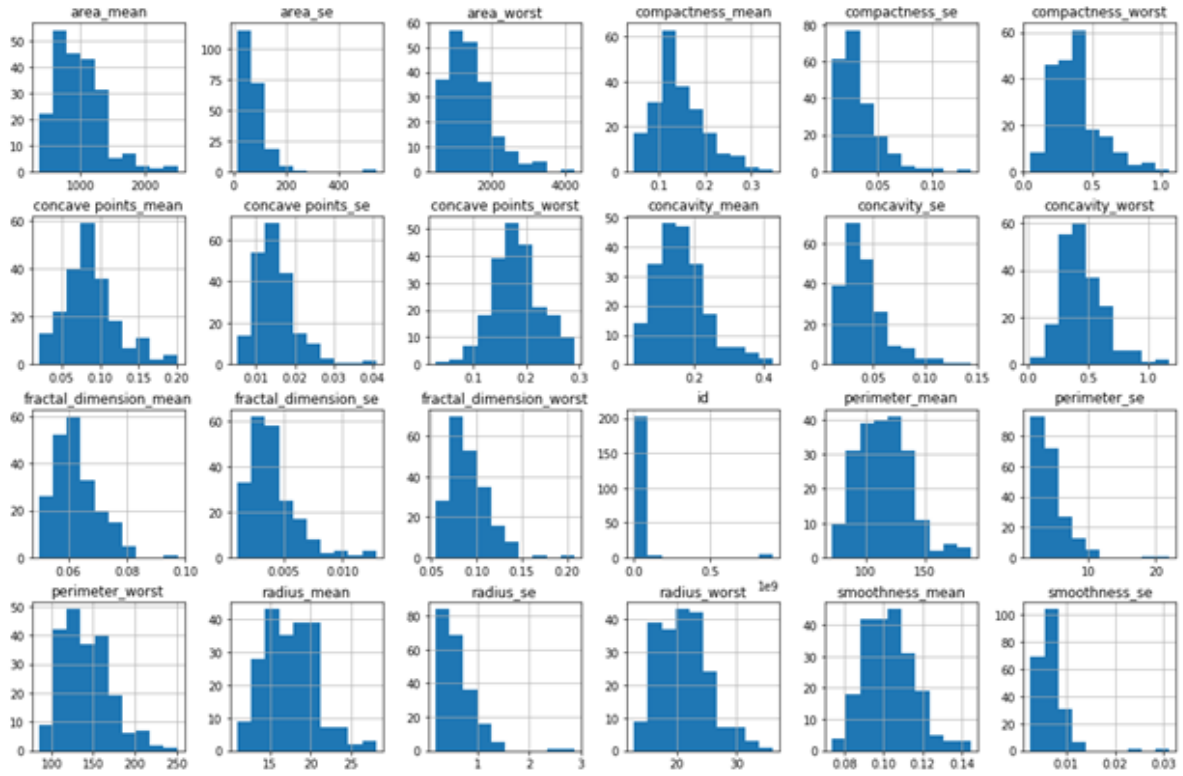


Fig : Visualization of Dataset

Based on the sample of the above we are fitting and training the data to x and y axis and then we are fitting the above data to a decision tree classifier algorithm.

Phase 3:Splitting the dataset

The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on this subset.

We will do this using SciKit-Learn library in Python using the train_test_split method.

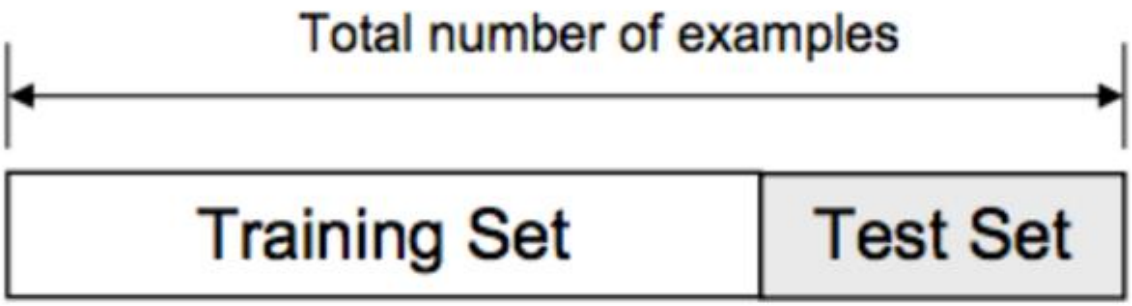


Fig: Training and test set

Shape():

Shape is a tuple that gives dimensions of the array.

```
print(X_test.shape)
print(y_test.shape)
```

```
(455, 2)
(455,)
```

Phase 4: Create model using criterion as entropy

Entropy:

A probability-based measure used to calculate the amount of uncertainty.

Max-Depth:

The **depth of a decision tree** is the length of the longest path from a root to a leaf.

Random _State:

Random_state , is used for initializing the internal **random** number generator, which will decide the splitting of data into train and test indices.

```
dt_entropy = DecisionTreeClassifier(max_depth=8,
                                    criterion='entropy',
                                    random_state=SEED)
```

```
# Fit dt_entropy to the training set
dt_entropy.fit(X_train, y_train)
```

```
DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=
8,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=1,
            splitter='best')
```

After taking the malignant and benign data, it can be trained and tested by using a mathematical concept called entropy with which we finally test the data using decision tree classifier and find out the accuracy as 0.89 in the sense 89 percent of members in that given d ataset are suffering with the cancer.

Phase 5: Create model using criterion as Gini

Gini index measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen.

```
dt_gini= DecisionTreeClassifier(max_depth=8,
                                criterion='gini',
                                random_state=SEED)
```

```
# Fit dt_entropy to the training set
dt_gini.fit(X_train, y_train)
```

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=8,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=1,
            splitter='best')
```

Phase 6: Compare Entropy and Gini Accuracy

```
# Use dt_entropy to predict test set labels
y_pred = dt_entropy.predict(X_test)
```

```
# Evaluate accuracy_entropy
accuracy_entropy = accuracy_score(y_test, y_pred)
```

```
accuracy_entropy
```

```
0.9298245614035088
```

```
# Use dt_entropy to predict test set labels
y_pred_gini = dt_gini.predict(X_test)
```

```
# Evaluate accuracy_gini
accuracy_gini = accuracy_score(y_test, y_pred_gini)
```

```
accuracy_gini
```

```
0.9298245614035088
Accuracy achieved by using entropy:  0.9298245614035088
Accuracy achieved by using the gini index:  0.9298245614035088
```

The two models achieve exactly the same accuracy. Most of the time, the gini index and entropy lead to the same results. The gini index is slightly faster to compute and is the default criterion used in the DecisionTreeClassifier model of scikit-learn.

Phase 7:
we retest the accuracy predition by finding the result again by using Logistic Regression algorithm**.**

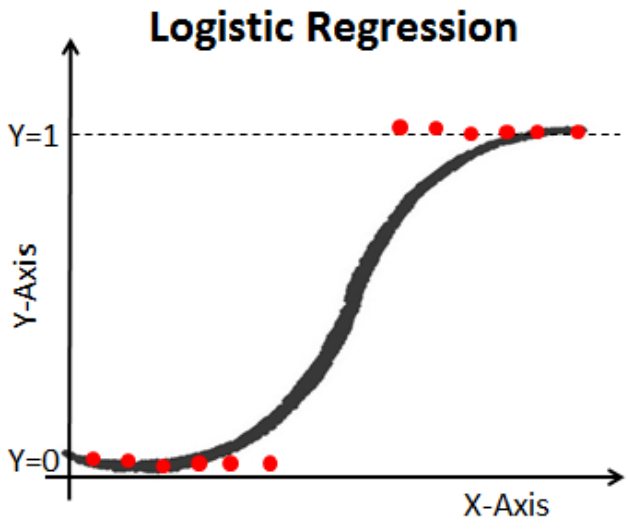
```
# Instatiate logreg
logreg = LogisticRegression(random_state=1)
```

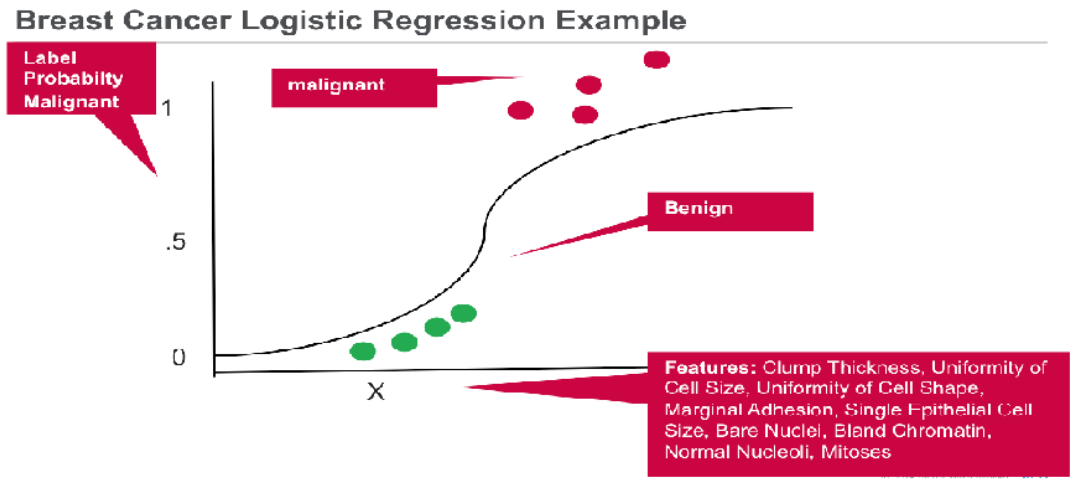
```
# Fit logreg to the training set
logreg.fit(X_train, y_train)
```

```
# predict
y_pred1 = logreg.predict(X_test)
```

```
acc1 = accuracy_score(y_test, y_pred1)
acc1
```

```
Out[23]: 0.9122807017543859
```





TESTING:

During testing, the implementation is tested against the requirements to make sure that the product is actually solving the needs gathered during the requirements phase. We have to test the project using testing tools automation technology called as selenium by using this, it can verify the bugs and solve the bugs.

OUTCOMES:

```
In [58]: accuracy_entropy
Out[58]: 0.9298245614035088
```

```
In [38]: accuracy_gini
Out[38]: 0.9298245614035088
```

```
Accuracy achieved by using entropy:  0.9298245614035088
Accuracy achieved by using the gini index:  0.9298245614035088
```

```
In [15]: # Compute test set accuracy
         acc = accuracy_score(y_test, y_pred)

In [16]: print("Test set accuracy: {:.2f}".format(acc))

         Test set accuracy: 0.89

In [17]: print(confusion_matrix(y_test,y_pred))

         [[65  7]
          [ 6 36]]

In [18]: (65+36)/114

Out[18]: 0.8859649122807017
```

Note: Not bad! Using only two features, your tree was able to achieve an accuracy of 89% :)

BY LOGISTIC REGRESSION:

```
In [22]: # predict
         y_pred1 = logreg.predict(X_test)

In [23]: acc1 = accuracy_score(y_test, y_pred1)
         acc1

Out[23]: 0.9122807017543859
```

By observing accuracies, On this data set which alogorithm is good ?

○ DT Classifier ◉ Logistic Regression

CONCLUSION:

There are different techniques that can be used for the prediction of breast cancer recurrence. In this project, I analyzed breast cancer data using two classification techniques to predict the recurrence of the cancer and then compared the results. The results indicated that Logistic Regression is the best classifier predictor with the test dataset, when compared to Decision tree classifier. Further studies should be conducted to improve performance of these classification techniques by using more variables and choosing for a longer follow-up duration.

ASSESMENT OF INTERNSHIP:

GREPTHOR
SOFTWARE SOLUTIONS PVT LTD

INTERN PERFORMANCE EVALUATION

1.	Name Of The Student	:	SANA PREETHI
2.	Regd.No	:	1215316662
3.	Duration Of Internship	:	45Days (2/May/2019 to 15/June/2019)
4.	Contact Person	:	Vasanthi Kuppam
5.	Quality Of Project & Attendance	:	27/30
6.	Domain	:	Python with Machine Learning
7.	Name of Project supervisor	:	V. S.L.Prasad
8.	Name Of the Organization / Address	:	Grepthor Software Solutions Pvt.Ltd. Plot No:170/67,171, 1st Floor Balaji Empire Patrika Nagar, Madhapur, Hyderabad, Telangana 500081

Grepthor Software Solutions Pvt. Ltd. Hyderabad

(Authorized Signatory)

040- 48528954 | info@grepthorsoftware.com | www.grepthorsoftware.com
ot No:170/67,171, 1st Floor, Balaji Empire, Behind Cyber Towers, Patrika Nagar, Madhapur, Hyderabad- 500081.

REFERENCES:

Published by Jason Brownlee in 2017: Available online in several editions with varying amounts of supplementary material, cost between $USD 37 and 237. The reviewed edition costs $USD 37, 163 pages.

Machine learning is the subject of a large and sophisticated professional literature, with excellent books for biomedical engineers [1, 2] as well as at least one excellent text available free online [3]. Machine learning, together with related topics such as data mining, provides a set of tools with a huge potential range of applications from improving medical diagnosis to optimization of cancer therapy. It has also been the subject of considerable hype in the popular literature.

(http://machinelearningmastery.com/about/) Brownlee describes himself as a software developer who initially taught himself machine learning "to figure this stuff out". He also is an active blogger on machine learning, and has written several books on the topic for novices, some available online at his website and others available through online stores such as Amazon. In a sense, Brownlee is one of us, with a Ph.D. (Swinburne University, Melbourne, Australia) and a thesis and academic publications on modeling of artificial immune systems.

Master Machine Learning Algorithms can be purchased online.
At https://machinelearningmastery.com/master-machine-learning-algorithms (accessed on 03.08.2017) at modest cost ($USD 37), which also includes 17 Excel spreadsheets to illustrate the main algorithms. His website offers 10 related books (including four at a more advanced level) that are tailored for around with. Brownlee frequently sends emails to a wide distribution list with interesting tutorial material about topics in machine learning.

The algorithms discussed include linear regression, logistic regression, discriminant analysis, classification and regression trees, Naive Bayes, k-nearest neighbours, support vector machines, decision trees. Introductory and concluding chapters discuss general aspects of machine learning, including problems of overfitting.