VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

REVIEW – 1

NAME: G BHARGAV

REG NO: 15MIS0115

GUIDE: BRINDHA. K

COURSE NAME: SOFTWARE DESING AND DEVELOPMENT

COURSE CODE: SWE3004

FALL SEMSETER 2019-2020

TITLE:

A MACHINE LEARNING FRAMEWORK FOR PREDICTION OF BREAST CANCER

ABSTRACT

A breast Cancer is a malignant tumor that starts from cells of the breast. A malignant tumor is a group of cancer cells that may grow into (invade) surrounding tissues or spread (metastasize) to distant areas of the body. Breast Cancer occurs mainly in omen, but men can get it, too. Many people do not realize that men have breast tissue and that they can develop breast cancer. Male breast cancer is a relatively rare cancer in men that originates from the breast. As it presents a similar pathology as female breast cancer. Male breast cancer remains under diagnosed and, due to delays in diagnosis, is often also undertreated. The investigation and management of male breast cancer are based on studies on female patients.

Detection of breast cancer metastasis at the earliest stage is important for the management and prediction of breast cancer progression. Emerging Techniques using the analysis of circulating tumor cells promising results in predicting and identifying the early stages of breast cancer metastasis in patients.

LITERATURE SURVEY

1. Wisconsin University breast cancer database was analyzed by naïve Bayes prediction algorithm and naïve Bayes classification algorithms. So that algorithms are used to predict and classify whether the tumor is malignant. So data sets were chosen randomly. At the final naïve Bayes classification algorithm was shown that 85-95 percent is correctly classified. Two various data sets from Wisconsin breast cancer have been evaluated by different data mining algorithms. The out come that Random Forest model shows the highest classification accuracy (99.48 %) and when compared with the previous works, the new approach and methodology have come with highest performance and accuracy

2. Jimin Guol, Benjamin C.M. Fung, Farkhund Iqbal implemented decision tree algorithm decision tree algorithm with breast cancer data sets that get from Leiden University Medical Center. The data sets have 574 patients who have got surgery at that hospital. So they generate the recurrence of breast cancer by a decision tree algorithm within three years of initial diagnosis. The classifier predicted 70% accuracy. For the independent classifier of 65 patients the classifier exactly predicts the recurrence of the disease in 55 patients. The classifier also

separates patient into two based on their disease characteristic and their relevance of early relapse.

3. Ahmed Iqbal Pritom, shahed anzarus sababa, Md. Ahadur Rahman Munshi, Shihabuzzaman Shihab predicts whether the breast cancer is recurrent or not. They have used data sets from Wisconsin data sets of the UCI machine learning repository that have 35 attributes. After implementation of algorithms like C4.5 Decision Tree, Naïve Bayes and Support vector Machine (SVM) classification algorithm was implemented. The outcome of these SVM, Naïve Bayes and C4.5 has 75.75%, 67.17% and 73.73% respectively.

4. Uma Ojha and Dr. Savita Goel were also discussed about the study on the prediction of breast cancer recurrence using data mining techniques. The research was applied by both clustering and classification algorithms. The results show that decision tree and Support vector machine (SVM) came out with the best predictor 80% accuracy.

5. Bojana R. Andjelkovic Cirkovic, Aleksandar M. Cvetkovic, Srdjan M. Ninkovic, and Nenad D. Filipovi presents the application of data mining on estimation of survival rate and disease relapse for breast cancer patients. A data set that was taken from the Clinical Center of Kragujevac is evaluated by some classification algorithm. Based on selected data sets naïve Bayes algorithm was selected as an algorithm which have higher accuracy on the basis of the 5 year survival rate.

6. The research paper done by **Joana Diz Goreti Marreiros & Alberto Freitas** presents new computer based diagnosis system. By using this technique false positive diagnosis test can be reduced. After data sets analyzed nave bayes algorithm come with higher accuracy than Random forest.

7. Qi Fan,change-jie zhu and liu yin used different types of data mining techniques in order to predict the recurrence of breast. In this paper they researcher uses SEER data sets and applied a new classification method in order to predict the recurrence of this disease. After preprocessing of data sets the researcher applied several algorithms, so that the decision tree (c5) algorithms come with better performance.

8. Dursun delen, Glenn walker And Amit Kadam used diferent data mining techniques for prediction of survivability of breast cancer. Data mining,

classification algorithms such as artificial neural network and decision tree along with logistic regression to develop a model for breast cancer survivability. Based on this paper decision tree algorithm (c5) was coming with better. performance and predicted by more accuracy 93.6% and artificial neural network shows second performance 91.2% and logistic regression come to the worst of the three 89.2%.

9. A stage predictive model for breast cancer survivability presented by **Rohit j and Ramya Nadig**. In this paper they were used different algorithm in order predict the breast cancer solvability. The evaluation was done based the stage of the breast cancer. Three machine learning algorithms were applied in order to predict breast cancer survivability. These data sets was evaluated by classification algorithms such as naïve bayes, logistic regression and decision tree to predict breast cancer survivability.

10. Analysis of Efficiency of Classification and Prediction Algorithms (Naïve Bayes) for Breast Cancer Dataset presented by **liu yin and Bojana. R**. In this paper Naïve Bayes classification algorithm and naïve Bayes prediction algorithms. The result of the Naïve Bayes classification algorithm has highest accuracy value 89 %-95%

HARDWARE AND SOFTWARE REQUIREMENTS

1. Hardware Requirements: The hardware requirements for the project are as follows:

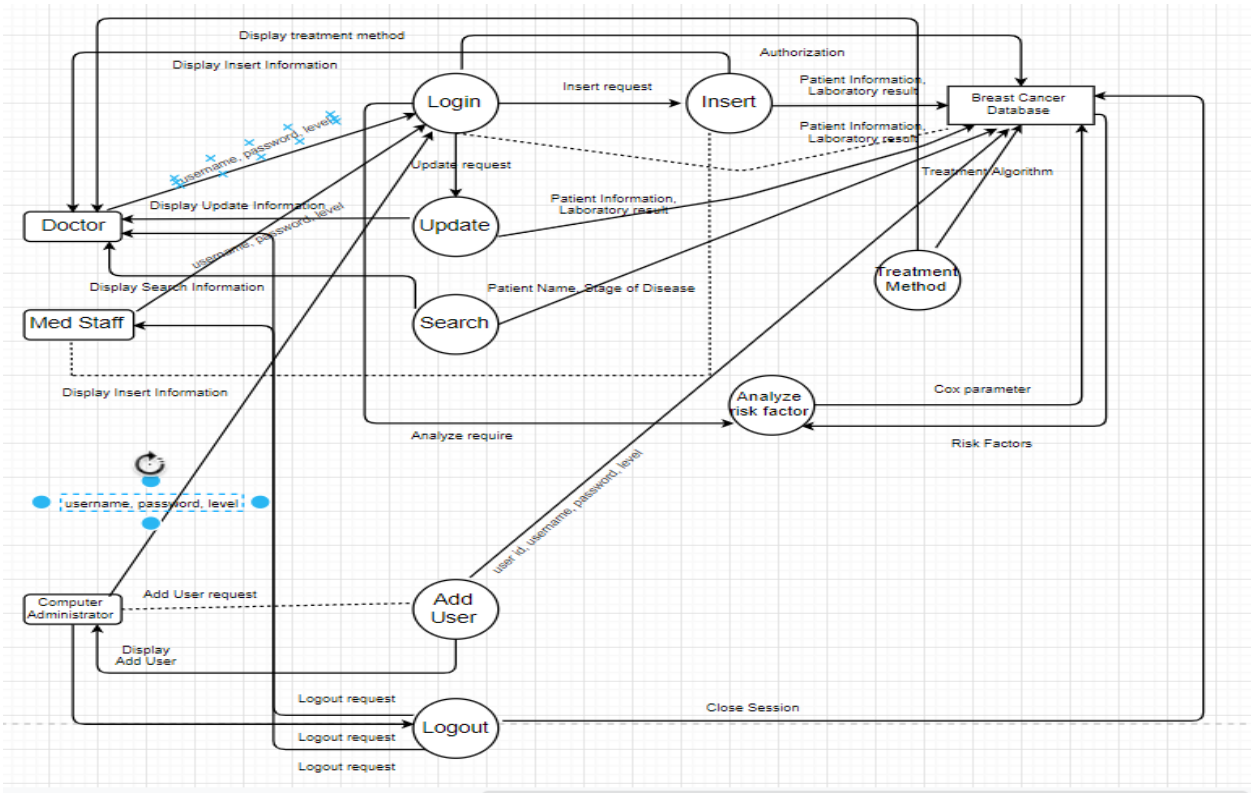
1. Processor : Any Update Processor
2. Ram : Min 4 GB
3. Hard Disk : Min 100 GB

2. Software Requirements: The software needed for the demonstration of the project are:

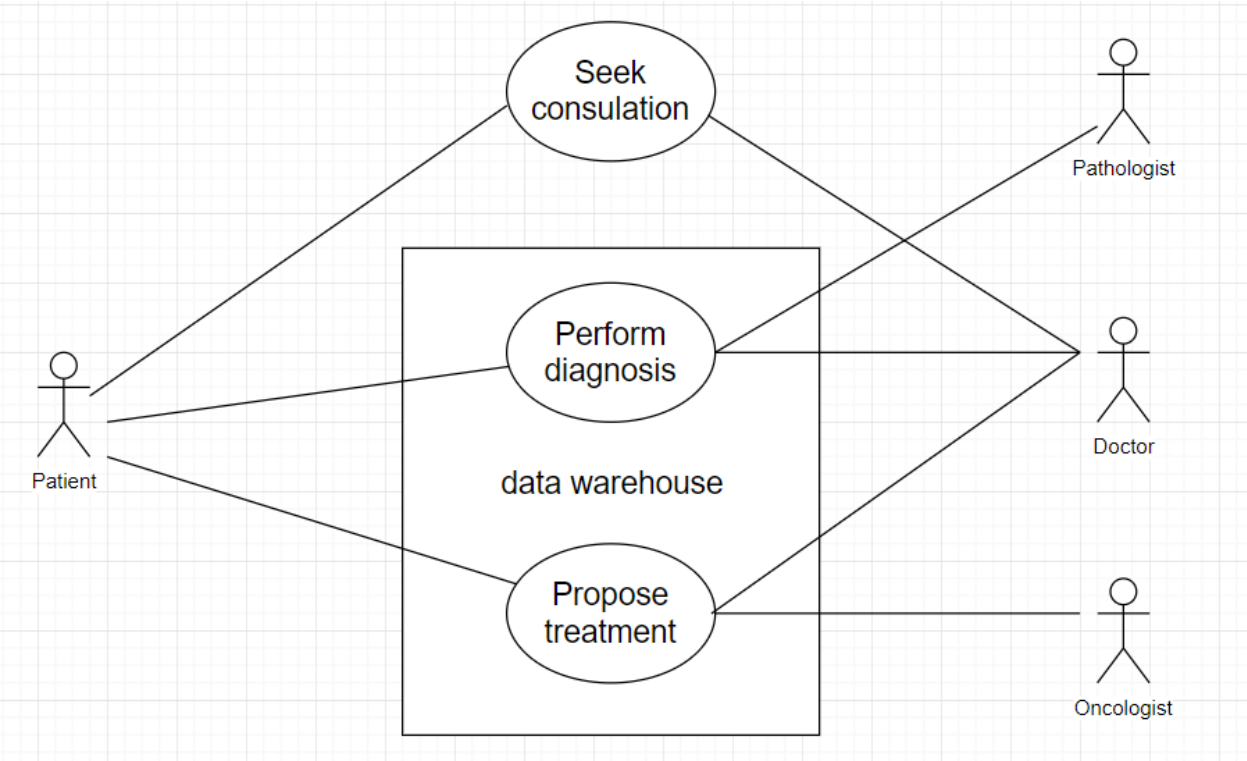
1. Operating System : Windows Family
2. Language : Python 3.6 or Anaconda
3. Other Tools : RStudio and Rapid Miner

DETAILED DESIGN:

- **SYSTEM ARCHITECTURE**



- **UML DIAGRAM / ER DIAGRAM**



- **MODULE DESCRIPTION**

❖ **Language Used : Python**
  - Breast Cancer is one of the most common cancers among women world-wide, representing the majority of new cancer cases and cancer-related death according to global statistic, making it a significant public health problem in today's society

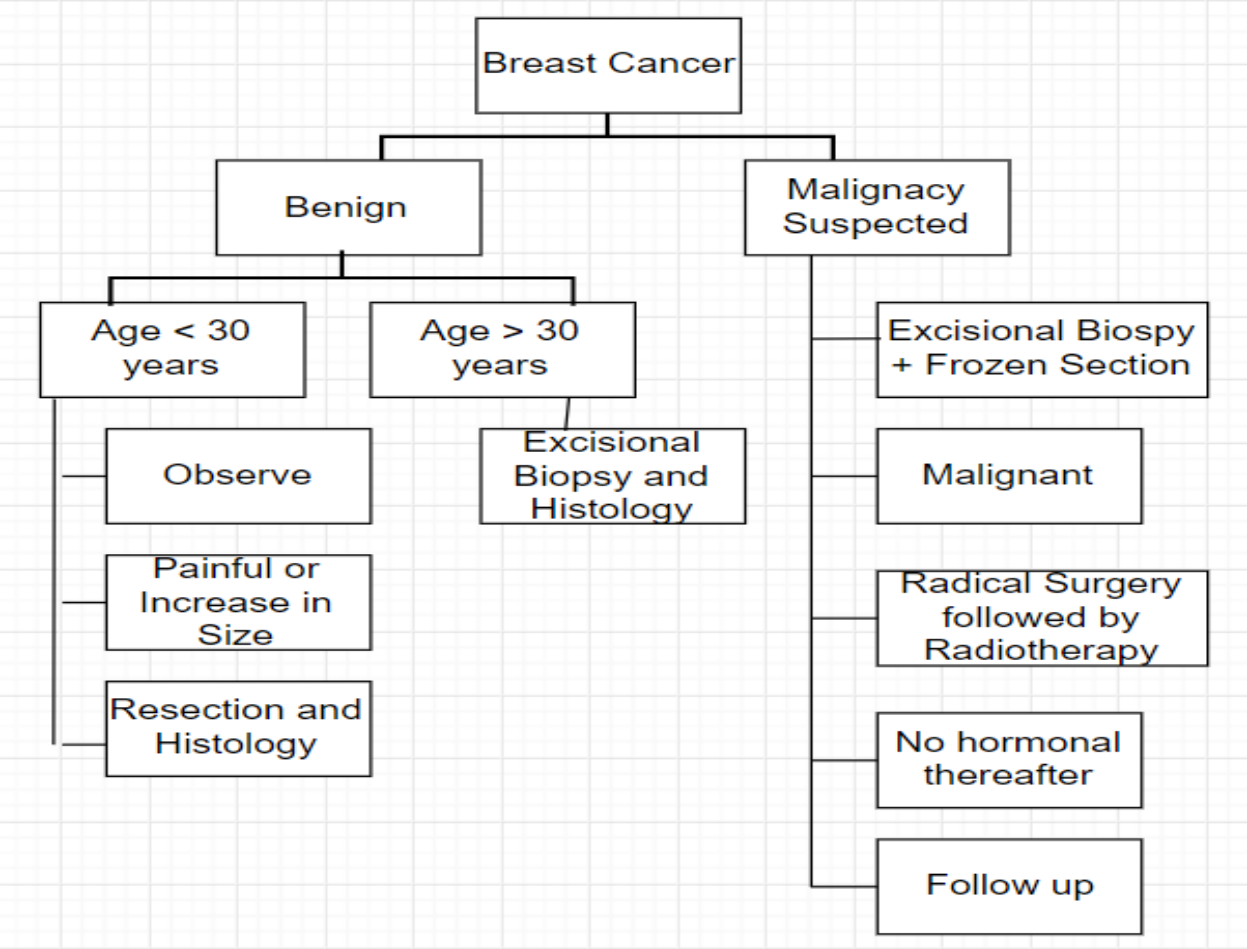
❖ **Insert**
  - Here this is the button where the Computer Administrator can Insert into Breast cancer database as a username, password, level, Stage of Disease

❖ **Update**
  - Here this is the button where the Computer Administrator, Doctor, Med Staff can Update into Breast cancer database as a username, password, level, Stage of Disease

❖ **Search**

- Here this is the button where the Computer Administrator, Doctor, Med Staff can Search into Breast cancer database as a User-Id or Username

- **DATABASE DESIGN / DFDs**



IMPLEMENTATION

```
import os
import numpy as np
import pandas as pd
import seaborn as sns
```

```
import datetime as dt
import matplotlib.pyplot as plt
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
```

---------- **READING THE DATA** ----------

```
data = pd.read_csv('C:/Users/Gurramkonda Bhargav/Desktop/breastcancer.csv')
```

---------- **OVERALL VIEW OF THE DATA** ----------

```
data.info()
```

---------- **CHECKING THE FIRST FEW ROWS OF THE DATA** ----------

```
data.head()
```

Out[4]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	conca
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.300
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.086
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.197
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.241
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.198

---------- **SUMMARY OF NUMERIC VALUES** ----------

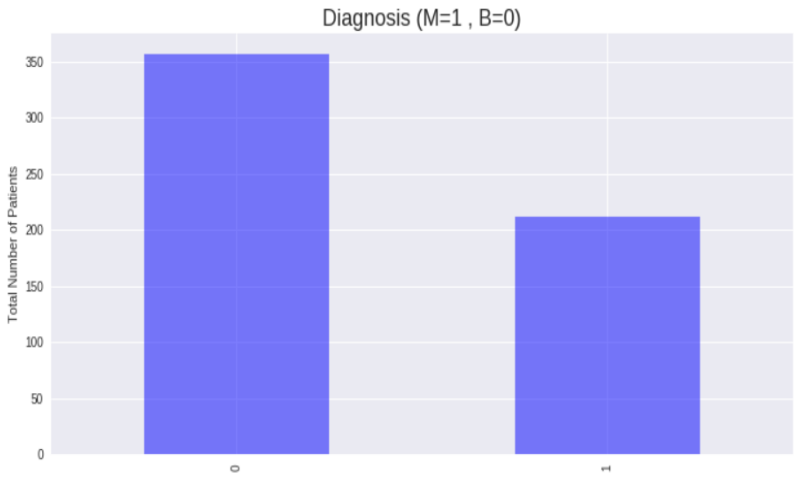
```
data.describe()
data.drop('id', axis = 1, inplace = True)
data.drop('Unnamed : 32', axis = 1, inplace = True)
data['diagnosis'] = data['diagnosis'].map({'M' : 1, 'B' : 0})
datas = pd.DataFrame(preprocessing.scale(data.iloc[:,1:32]))
datas.columns = list(data.iloc[:,1:32].columns)
datas['diagnosis'] = data['diagnosis']
```

Out[6]:

	id	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concav
count	5.690000e+02	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.00
mean	3.037183e+07	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.0887
std	1.250206e+08	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.0797
min	8.670000e+03	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.0000
25%	8.692180e+05	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.0295
50%	9.060240e+05	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.0615
75%	8.813129e+06	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.1307
max	9.113205e+08	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.4268

---------- **DIAGNOSIS** ----------

```
datas.diagnosis.value_counts().plot(kind='bar', alpha = 0.5, facecolor = 'b', figsize=(12,6))
plt.title("Diagnosis (M=1 , B=0)", fontsize = '18')
plt.ylabel("Total Number of Patients")
plt.grid(b=True)
```



---------- **SETTING UP THE TRAIN AND TEST DATA** ----------

```
from sklearn.model_selection import train_test_split, cross_val_score, cross_val_predict
from sklearn import metrics
predictors = data_mean.columns[2:11]
target = "diagnosis"
X = data_mean.loc[:,predictors]
y = np.ravel(data.loc[:,[target]])
# Split the dataset in train and test:
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

print ('Shape of training set : %i || Shape of test set : %i' % (X_train.shape[0],X_test.shape[0]))

print ('The dataset is very small so simple cross-validation approach should work here')

print ('There are very few data points so 10-fold cross validation should give us a better estimate')

```
Shape of training set : 455 || Shape of test set : 114
The dataset is very small so simple cross-validation approach should work here
There are very few data points so 10-fold cross validation should give us a better estimate
```

REFERENCES

- http://www.indianjcancer.com/article.asp?issn==0019-509X;year=2014;volume=51;issue=3;spage=200;epage=208;aulast=Datta
- http://info.cancerresearchuk.org/cancerstats/incidence/prevalence
- www.healthcareimprovementscotland.org/programmes/cancer/standards_for_cancer_services.aspx
- www.nice.org.uk/media/F66/90/BreastCancerQualityStandardFinal.pdf
- www.ncin.org.uk/view.aspx?rid=1043
- http://www.cancer.org/cancer/breastcancer/detailedguide /breast-cancer-diagnosis
- Chi C.L., Street W.H. and Wolberg W.H., “Application of Artifical Neural Network- based Survival Analysis on Two Breast Cancer Datasets”, Annual Symposium Proceedings / AMIA Symposium, 2007.
- William H Wolberg, Olvi Mangasarian, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA