International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Predicting Breast Cancer Using Gradient Boosting Machine

Sahr Imad Abed

Electrical and Electronic Engineering Altinbas University, Instanbul-Turkey

Abstract: ***Breast Cancer is an uncontrolled growth in the breast. Breast cancer is a primary cause of death in women globally. The amount of death can be reduced by eliminating the accuracies in the diagnosis of the disease. The increase in accuracy of diagnosis can be increased through the predictive technology developed using the Gradient Boosting Machine. The prediction will improve the quality of the treatment process and the survivability rate of the patients. In this paper, we propose a system that will be used for predicting breast cancer via the use of classifiers and machine learning algorithm. The system is intended to be user- friendly and cost-effective to contribute to the fight against this deadly disease. The system will estimate the risk of prevalent breast cancer in the early stage of development. Ultimately, the results of the system will be compared with the pertinent medical results of each patient. The practical part has illustrated that GBM algorithm performed better than the other models. The GBM algorithm had better specificity, accuracy, and sensitivity.***

Keywords: Gradient Boosting Machine, Prediction, Algorithm, Machine Learning

1. Introduction

Breast cancer is a killer disease, account for 25% of women deaths globally [1]. The early diagnosis and prognosis of breast cancer have become a vital element in cancer treatment and research. The primary challenge in the cancer treatment and management is the classification and prediction of breast cancer in patients into appropriate risk groups of breast cancer. The exercise of classifying the patient will allow treatment and follow-up. The process of risk assessment and prediction is essential in the optimization of the patient's health and the application of medical resources while ensuring that cancer does not reoccur [1]. The classification of patients onto either high or low risk has contributed to biomedical and bioinformatics studying the use of machine learning methods. Machine learning describes a suite of techniques that categories different predictive analytics. The machine learning technique utilizes a variety of probabilistic, statistical and optimization approaches that allow computers to learn and detect a certain pattern in any complex datasets. The machine language techniques are well-suited for medical applications especially due to the increased complexity of the proteomic and genomic information in this field. Machine learning has become increasingly used in cancer diagnosis and detection [2]. Moreover, the different algorithms used in machine learning have been used for breast cancer prognosis and prediction. This development has led to the growth of predictive medicine with a personal touch. The use of the gradient boosting machine for predictive analysis of breast cancer will be vital for the patients and the physician simultaneously. The patients will be diagnosed early ensuring treatment while the physician will be in a position to administer best treatment decisions for their patients.

The fundamental objective of breast cancer prognosis and prediction are different from the expectation of the diagnosis and detection. The prognosis processes have the following purposes: 1. The prediction of the breast cancer susceptibility or risk assessment [3] 2. The prediction of breast cancer recurrence [4] and 3. The prediction of breast cancer survivability [5]. The first objective outlines the need of an individual trying to predict the likelihood of developing breast cancer before the actual occurrence of the disease. Secondly, there is the need to predict the likelihood of redeveloping breast cancer after the disease has occurred and resolved. The final objective is aimed at predicting the outcome of the disease, for instance, the life expectancy after the diagnosis of the disease. This paper will focus on the application of a Gradient Boosting Machine algorithm in the prediction of the likelihood of the recurrence of cancer. The paper will provide a background into the topic with a primary focus on the practical part. The practical part will apply the algorithm on the various data collected in the different database.

2. Material and Methods

The Gradient Boosting Machine is proposed to allow in the observation of massive sores and abnormal growth in the breasts. The process will assist the radiologist in carrying out mammography to be able to identify the possibility of breast cancer development. The Gradient Boosting Machine technique requires a colossal amount of data from a complex database. The following objectives will be vital for the practical implementation of the theory work; first, there is the need to apply the fundamental concepts of machine learning from the dataset collected from the databases. The practical part will also allow the evaluation and interpretation of the results and justification of the interpretation based on the dataset. The analysis was divided into four phases

- Identifying the problem and the associative datasources
- Data analysis
- Preprocessing theData
- Construction of model to predict whether the breast tissue indicate a malignant or benign

3. Procedure

The data analysis process will be implemented as illustrated below

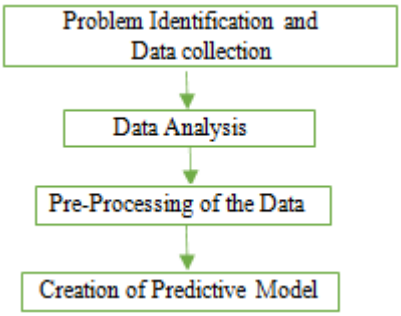

Figure 1: The procedure of data manipulation in the Gradient Boosting Machine

a) Identifying the problem

Breast cancer is the most form of malignancy in women, forming a third of all cancer diagnosed in women globally. The disease is caused by abnormal cell growth in the breast tissue known as a tumor. The tumor will not necessarily mean cancer as it can either be benign which is not cancerous or pre-malignant which is pre-cancerous or a malignant tumor which is cancerous. Therefore, based on the above classification the expected results will use two forms of training

1 = Malignant (cancerous)- illustrating the presence of breast cancer

0 = Benign (not cancerous) which will demonstrate the absence of breast cancer.

The label of the data will be discrete the prediction will thus be in the two categories. In machine learning discrete data is typically a problem; therefore, the goal will be to predict the benign and malignant tumor then illustrate the possibility of re-occurrence and non- recurrence of the malignancies after a certain period. The achievement of the prediction will thus apply the Grade Boosting Machine for the classification and fitting of the functions for discrete data. The breast cancer datasets were collected from various repositories as illustrated in the code below;

```
#load libraries
                                # linear algebra
import numpy as np

# Read the file "data.csv" and print the contents.

#!cat data/data.csv
```

The initialization will load the machine learning repositories for the datasets with 14 attributes and around 48842 instances that will define the model. The hypothesis developed for this phase is that there is a probability of predicting recurrence and non-recurrence malignancy in a certain period using the different breast cancer tumor attribute of the malignant or benign. In this phase a binary classifier that will be used in determining the breast cancer dataset information to classify the tumor.

b) The Data Type

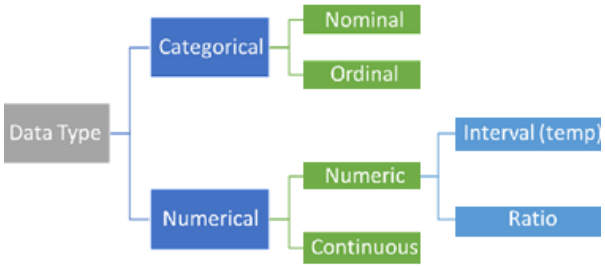

Figure 2: The data transformation Chart [2]

The data obtained from the database will provide numerical therefore it will be vital in transforming the data to categorical and back to numerical to ensure continuation in the analysis of the data.

The encoding process will entail the transformation of categorical data to numerical data such from malignant and benign to 1 and 0 for easy classification and application of predictive classification using the gradient boosting machine. This type of encoding is known as binary encoding where the categorical data is transformed to 1 or 0.

c) Exploratory Data Analysis (EDA)

The primary goal of exploration is a vital step that is used for feature engineering and data acquisition. The process data exploration is carried out before the modeling. The exploration of the data will allow the researcher to understand the data and its associative nature. The assumptions made with regards to the data will be accurate and precise. The results of EDA will provide the structure of the data, the presence of extreme value and distribution of certain values. The EDA will further allow the formulation of the interrelationships within the data sets. The following libraries will be used for data analysis

Table 1: The Libraries used in python for the Gradient Boosting Machine [6]

Library	Meaning	Explanation
NumPy	Numerical Python	Utilizes n-dimensionally array for basic algebraic functions, random numbers, Fourier transformand integration
SciPy	Scientific Python	Built on the NumPy used for high level scientific and engineering science such as discrete transform, linear algebra, optimization, discrete Fourier transform and Sparse Matrices
Matplotlib		Library for plotting a variety of graphs such as histograms, heat plots etc.
Pandas		Used for the analysis and manipulation of structured data.
Scikit learn		Create tools for machine learning such as statistical modelling, regression, classification and

		dimensionality reduction etc.
Statsmodels	Statistical modeling	Used for explore data, perform statistical test and estimate statistical model.
Seaborn	Statistical data Visualization	Used for informative statistical graphics in python
Bokeh		It is a library used to create interactive plots, data application and dashboards on web-browsers.
Blaze		Utilized to extend the functionalities of the NumPy.
Scrapy	Web crawling	Utilized for extraction of certain data pattern. It goes through web pages of a website gathering data.
SymPy	Symbolic Computation	Used for symbolic arithmetic

The code for loading these libraries will look like

```
from scipy.stats import norm
import seaborn as sns #visualization


plt.rcParams['figure.figsize'] = (15,8)
plt.rcParams['axes.titlesize'] ='large'
```

The data collected will be uni-variate or bivariate thus different plots will be used to represent the data. The box and histogram plots will be used for the representation of the univariate data. A scatter plot will be used to represent the relationship between two variables, i.e., bivariate. The code to generate these plots is as illustrated below

1. **Univariate dataset**

```
#Plot histograms of CUT1 variables
hist_mean=data_mean.hist(bins=10, figsize=(15, 10),grid=False,)

#Any individual histograms, use this:
#df_cut['radius_worst'].hist(bins=100)
```

2. Bivariatedata

```
X_pca = pca.transform(Xs)

PCA_df= pd.DataFrame()

PCA_df['PCA_1'] =X_pca[:,0]
PCA_df['PCA_2'] =X_pca[:,1]

plt.plot(PCA_df['PCA_1'][data.diagnosis == 'M'],PCA_df['PCA_2'][data.diagnosis == 'M'],'o', alpha = 0.7, color =
'r')
plt.plot(PCA_df['PCA_1'][data.diagnosis == 'B'],PCA_df['PCA_2'][data.diagnosis == 'B'],'o', alpha = 0.7, color =
'b')

plt.xlabel('PCA_1')
plt.ylabel('PCA_2')
plt.legend(['Malignant','Benign'])
plt.show()
```

Once the data was analyzed then the predictive model was created in Gradient Boosting Machine (GBM). The predictive model followed the following step to extract features.

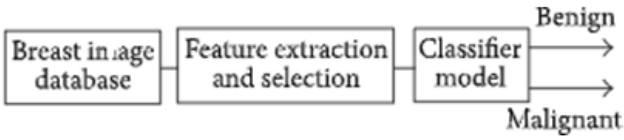

Figure 3: The Breast Image Classification model in GBM [10]

Figure 4: A mammogram benign images [9]

4. Result

For the approval of the methodology used in the predictive analysis of the dataset

Figure 5: A mammogram malignant images [9]

Results for the Univariate Dataset

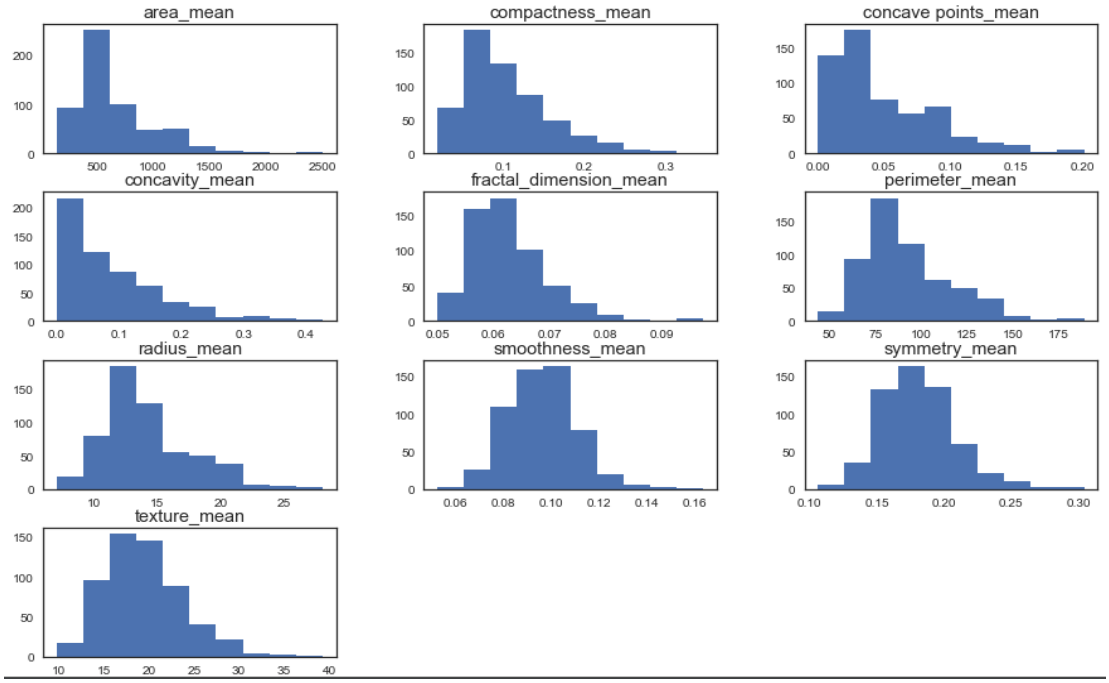

Figure 6: The histogram for the univariate variable

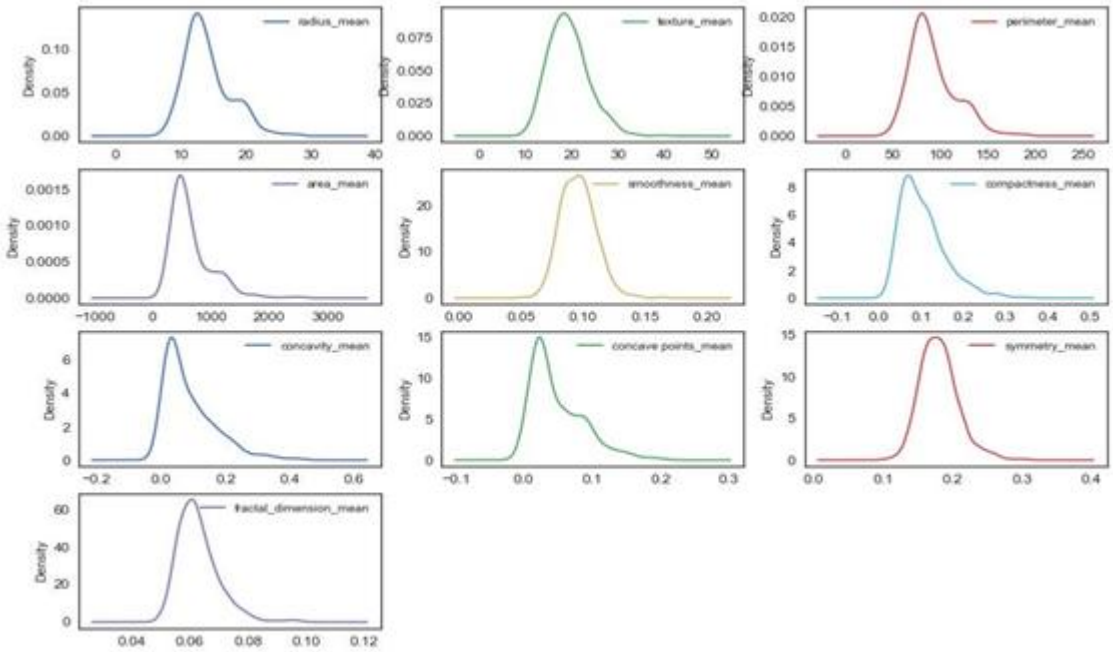

Figure 7: The Density plot for the univariate dataset

Results for the Bivariate Dataset

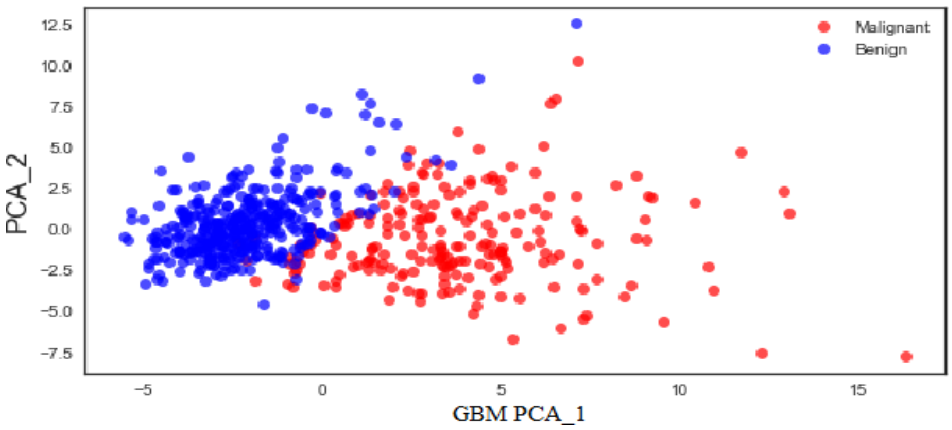

Figure 8: The Variants in each PC

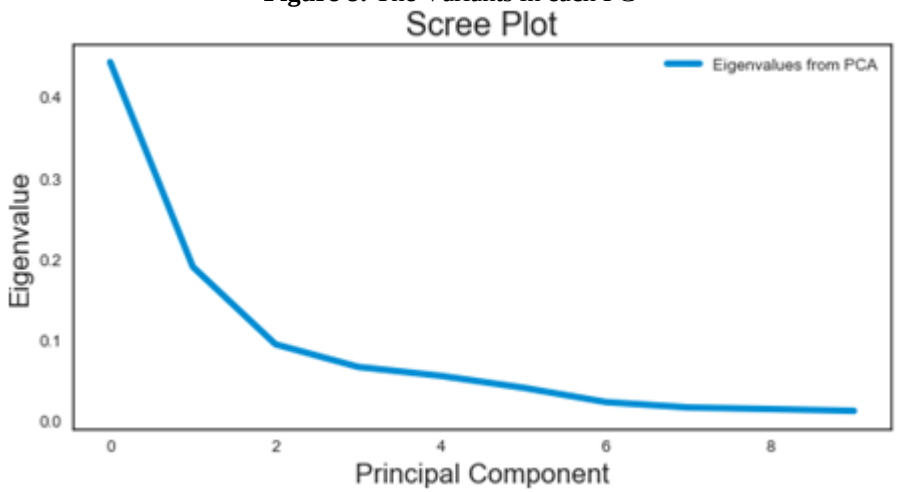

Figure 9: The Scree Plot to indicate the component of the data

Accuracy

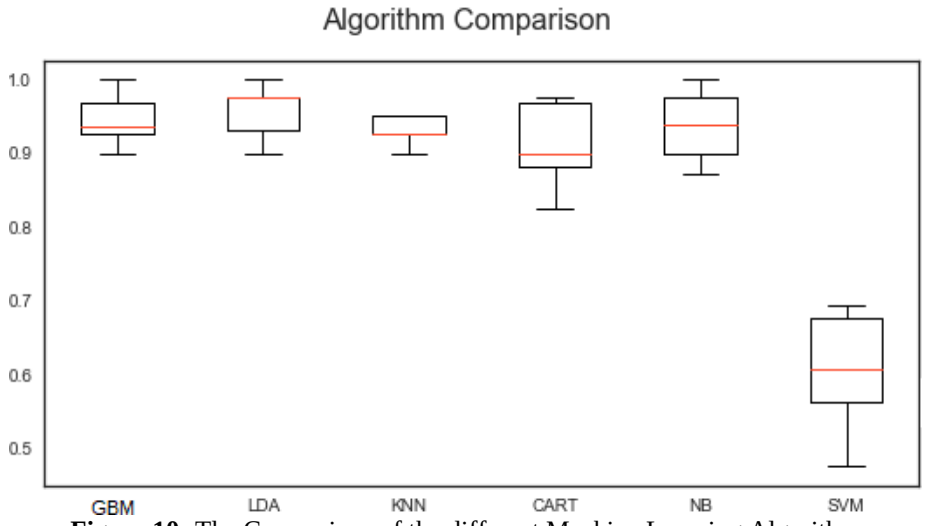

Figure 10: The Comparison of the different Machine Learning Algorithm

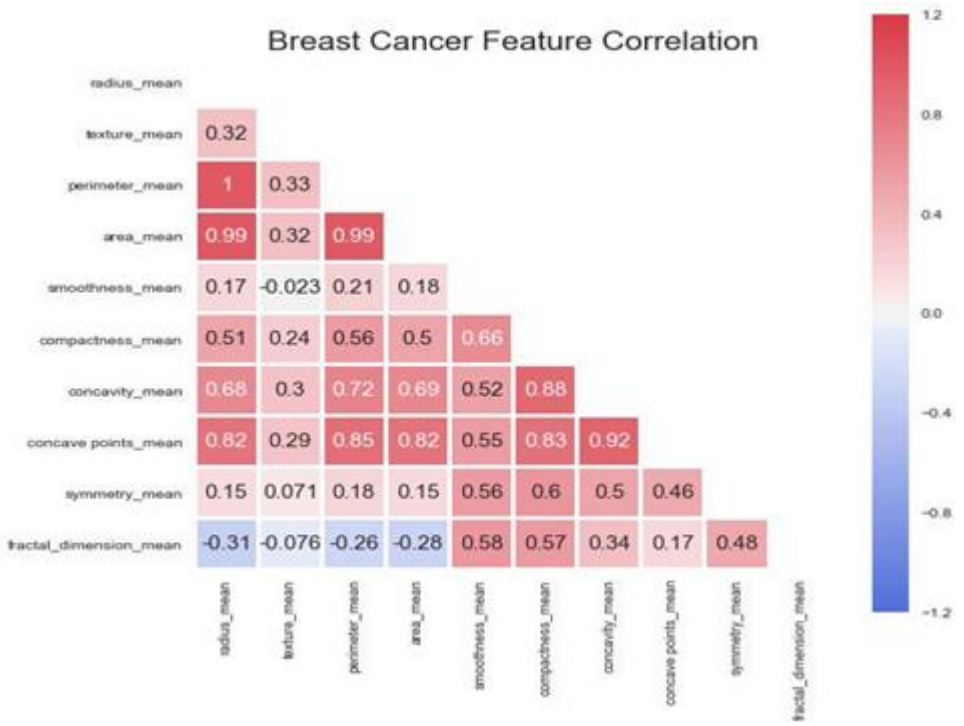

Figure 11: The Correlation of the breast cancer features

5. Discussion

The scree plot illustrated that change in the slope occurred at component 2 which is called the elbow. The argument from the plot would be to advocate for the retention of the first three components. Additionally, figure 8 illustrate that the linear PCA the transformation of the subspace from 3D to 2D illustrating that the data was well separated [5]. The gradient boosting machine performed better in this analysis as the classifiers were different from the base classifiers thus the training error in the base classifier did not affect the result of the iterations. This section will analyze the results of the simulations [6]. The principal component was used for determining the classifiers in the GBM method. The components were used for the creation of the different number of boosting iterations. The practical advantage of the approach combined with the base classifier increasing as the subsequent base would produce a similar result with the prediction results. The performance of the Gradient Boosting Machine in the simulation would be closely related with the mechanism used in updating the score of the feature extracted from the images. The update in the algorithm is certainly reliant on the data used for training the classifiers [7]. The data in this experiment needed to be compressed to reduce overhead and increase the time for the iteration reducing the performance of the algorithm.

6. Conclusion

The prediction and prognosis of breast cancer have been identified as an important thing that can be utilized in the improvement of the treatment process. The predictive medicine will improve the quality of medical services offered and the survivability rate of the patients. This paper presents the Gradient Boosting Machine algorithm for efficient prediction and diagnosis of breast cancer in the early stages. The Gradient Boosting Machine has introduced new computational biology for the classification of risk assessment of breast cancer as high or low. The technique will also reduce the cost of the test and diagnosis in the late stages. The proposed system will be utilized to predict and diagnose breast cancer to ensure the appropriate treatment for the patients.

References

[1] A. Al Nahid and Y. Kong, "Involvement of Machine Learning for Breast Cancer Image Classification: A Survey," Computational and Mathematical Methods in Medicine, vol. 3, no. 3781951, pp. 1-29,2017.
[2] M. Deepika, M. L. Gladence and M. R. Keerthana, "A Review on Prediction Of Breast Cancer Using Various Data Mining Techniques," Research Journal of Pharmaceutical, Biological and Chemical Science, vol. 7, no. 1, pp. 808-813,2016.
[3] Q. Zhu, S. Tannenbaum, S. H. Kurtzman, P. DeFusco, A. Ricci Jr, H. Vavadi, F. Zhou, C. Xu, A. Merkulov, P. Hegde, M. Kane, L. Wang and K. Sabbath, "Identifying an early treatment window for predicting breast cancer response to neoadjuvant chemotherapy using immunohistopathology and hemoglobin parameters," Breast Cancer Research, vol. 20, no. 56, pp. 1-17,2018.
[4] P. H. Abreu, M. S. Santos, M. H. Abreu, B. Andrade and D. C. Silva, "Predicting Breast Cancer Recurrence Using Machine Learning Techniques: A Systematic Review," ACM Computing Surveys, vol. 49, no. 3,2016.
[5] S. M. Rostami, M. R. Parsaei and M. Ahmadzadeh, "A survey on predicting breast cancer survivability and its challenges," UCT Journal of Research in Science, Engineering and Technology, pp. 37-42,2016.
[6] A method to select a good setting using python open source provided by the open source community. In Biomedical and Health Informatics (BHI),

"https://pythontips.com/2013/07/30/20-python-libraries-you-cant-live-without/"

[7] R. Blagus and L. Lusa, "Boosting for high-dimensional two-class prediction," BMC Bioinformatics, vol. 16, no. 300,2015.

[8] Y. Chen, Z. Jia, D. Mercola and X. Xie, "A Gradient Boosting Algorithm for Survival Analysis via Direct Optimization of Concordance Index," Computational and Mathematical Methods in Medicine, vol. 2013, no. 873595, pp. 1-8,2013.

[9] Z. Beheshti, S. M. Hj. Shamsuddin, E. Beheshti, and S. S. Yuhaniz. Enhancement of mammo-graphs provided for the swarm optimization for medical diseases diagnosis. Soft Computing, 18(11):2253–2270,2014.

[10] M. Lichman. UCI Machine Learning Repository for feature extraction. http://archive.ics.uci.edu/ml, 2013. [Accessed:2015-03-30].

Appendix

Spot Check Algorithm

```
# Spot-Check Algorithms
models = []
models.append(( 'GBM' , Gradient Boosting Machine()))
models.append(( 'LDA' , LinearDiscriminantAnalysis()))
models.append(( 'KNN' , KNeighborsClassifier()))
models.append(( 'CART' , DecisionTreeClassifier()))
models.append(( 'NB' , GaussianNB()))
models.append(( 'SVM' , SVC()))

# Test options and evaluation metric
num_folds = 10
num_instances = len(X_train)
seed = 7
scoring =  'accuracy'

# Test options and evaluation metric
num_folds = 10
num_instances = len(X_train)
seed = 7
scoring =  'accuracy'
results = []
names = []
for name, model in models:
 kfold = KFold(n=num_instances, n_folds=num_folds, random_state=seed)
 cv_results = cross_val_score(model, X_train, y_train, cv=kfold, scoring=scoring)
 results.append(cv_results)
 names.append(name)
 msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
 print(msg)
print('-> 10-Fold cross-validation accurcay score for the training data for six
classifiers')
```

Matrix Correlation

```
# plot correlation matrix
import pandas as pd
import numpy as np
import seaborn as sns
from matplotlib import pyplot as plt

plt.style.use('fivethirtyeight')
sns.set_style("white")

data = pd.read_csv('data/clean-data.csv', index_col=False)
data.drop('Unnamed: 0',axis=1, inplace=True)
# Compute the correlation matrix
corr = data_mean.corr()

# Generate a mask for the upper triangle
mask = np.zeros_like(corr, dtype=np.bool)
mask[np.triu_indices_from(mask)] = True

# Set up the matplotlib figure
data, ax = plt.subplots(figsize=(8, 8))
plt.title('Breast Cancer Feature Correlation')

# Generate a custom diverging colormap
cmap = sns.diverging_palette(260, 10, as_cmap=True)

# Draw the heatmap with the mask and correct aspect ratio
sns.heatmap(corr, vmax=1.2, square='square', cmap=cmap, mask=mask,
            ax=ax,annot=True, fmt='.2g',linewidths=2)
```