Breast cancer detection through classification algorithms

Group ID: SM2020CSE47505

Jannatul Naim (2017-2-60-029)
Syed Reshad Faysal (2017-2-60-049)
MD Nayeem (2017-2-60-0028)

September 26, 2020

1. Introduction

Breast Cancer is the most common cancer among women, every year a large number of women have to suffer badly for the late identification of these diseases. Throughout this project, we have predicted patient breast cancer status (affected or not affected) by applying a machine learning approach. We apply some classification algorithms (support vector machine, Naive Bayes, Logistic regression) on our data set and analyze the output result of each algorithm. Finally, by interpreting these classifier algorithms' output results, we have found the best classifier algorithm among those to Predict Breast Cancer.

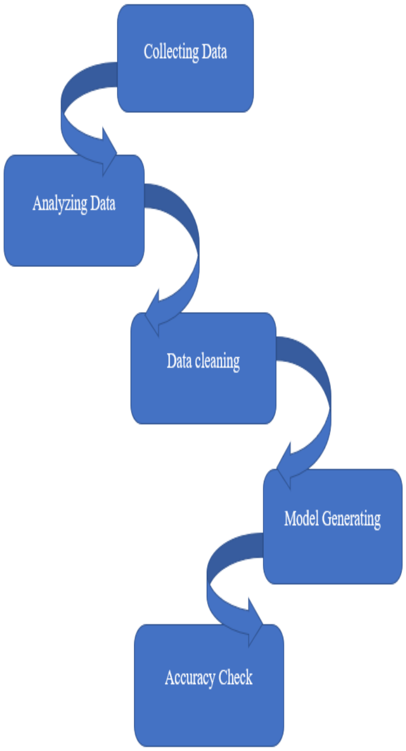
1.2 Motivation

Breast Cancer is the most common cancer among women, every year a large number of women have to suffer badly for the late identification of these diseases. As the machine learning approach has partly used all sectors in our life and it makes our life easier, on that aspect hopefully we can get a better way to treat that problem (Breast Cancer) by applying the machine learning approach. We hope the matching approach effectively much better to identify breast cancer than the traditional way and it should be much reliable. In this aspect, we choose this project.

1.4 Necessity

Necessity of this project is, people able to gain prior knowledge and get basic intuition about Machine learning classifier algorithms, which knowledge they can apply on any type of classification problem. In future,they can take a quick decision about applying and choosing suitable classification algorithm on other problems.

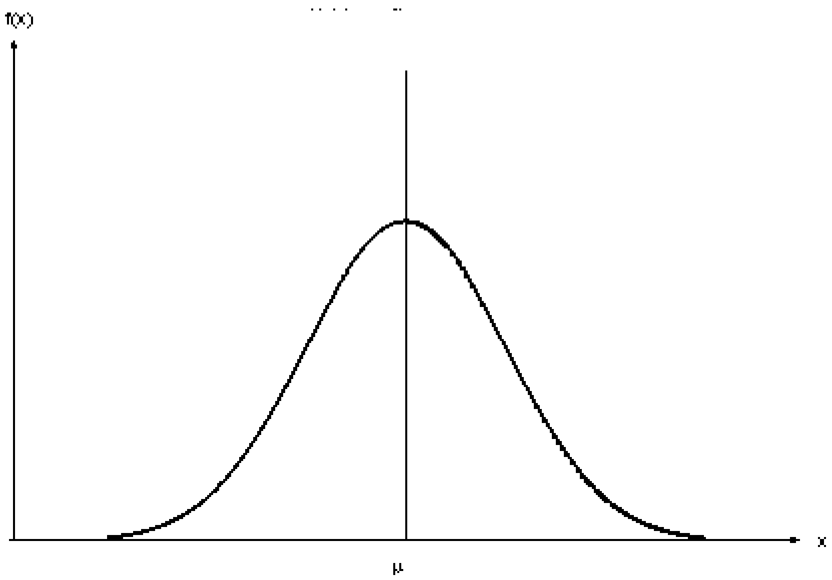
2. Methodology:



To developed our machine learning model, we have collected data set from UC Machine Learning Repository. Then we analyze our data set the attribute to identify suitable machine learning algorithms. In the coding section, we use some python library which are pandas, numpy, seaborn, matplotlib, sklearn, etc. As, most of the data set attribute contains categorical and continuous type data and sci-kit learn library cannot handle these type of data, so we encode our data set attributes value to numerical values using a function called “LabelEncoder()”. The attribute that we want to convert to a numeric is given as a function’s parameter of the ”levelencoder()”. We delete one of the attribute column (InvNodes) from our data set as there was some missing information. After completing the preprocessing data set, we select some classifier algorithms which are Gaussian naive Bayes, logistic regression, linear support vector machine, Gaussian kernel support vector machine, sigmoid kernel support vector machine for analysis their output result. We only choose these classifier algorithms because we have some prior knowledge of these algorithms. As most of the attribute values are continuous types in our data set, that’s why we choose the Gaussian naive Bayes algorithm. There is also another type of naive Bayes algorithm like barn naive Bayes, multinomial naive Bayes. As we know barn naive Bayes works well in discrete type values and multinomial naive Bayes works well in recurrent type value, that's why we use the gaussian naive Bayes algorithm only. From the regression classifier section, we choose the Gaussian naive Bayes algorithm. There is also another type of naive Bayes algorithm like barn naive Bayes, multinomial naive Bayes. As we know barn naive Bayes works well in discrete type values and multinomial naive Bayes works well in recurrent type value, that's why we use the gaussian naive Bayes algorithm only. From the regression classifier section, we choose logistic regression as our problem is a binary type classification problem. From support vector machine algorithms we work with linear SVM as data distribution of class attribute little bit linearly separable. Gaussian SVM kernel also a good classifier for any type of class data distribution, that's why we select it for our analysis purpose. Provides train data as a parameter in the fit function to fit the model and test data as a parameter in the predict function to predict. After applying all of these classifiers in our data set, we have got a confusion matrix for each of those. Then we calculate precision, recall,f1 score, accuracy, and compare these value to each others algorithms.

3. Implementation

Naïve Bayes:
Naïve Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the bayes theorem. Thisalgorithm has 3 function based on data type. Here we used gaussian naïve bayes. When the predictors take up a continuous value and are not discreate, we assume that these values are sampled from a gaussian distribution.



SVM:
A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. SVM has different type of kernel. Here we use RBF kernel. In the RBF kernel SVM, we construct the kernel matrix by computing the pair-wise distances between the training points, which makes it non-parametric.It is used to perform transformation, when there is no prior knowledge about data.
Logistic regression:
Logistic regression is a very powerful modeling tool, is a generalization of linear regression. Logistic Regression is used to assess the likelihood of a disease or health condition as a function of a risk factor (and covariates). Both simple and multiple logistic regression, assess the association between independent variable(s) (Xi) – sometimes called exposure or predictor variables — and a dichotomous dependent variable (Y) – sometimes called the outcome or response variable. It is used primarily for predicting binary or multiclass dependent variables.

3.1 Data collection:

The data used in this study are provided by the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. My special thanks go to M. Zwitter and M. Soklic for providing the data for this research work. The data set has 10 attributes and total 286 rows, we restricted testing to these same attributes (see Table 1) and contain the following variables.
1. Class: no-recurrence(no) or recurrence(yes) depending reappearing symptoms of breast cancer in the patients after treatment.
2. Age: patient's age at the time of diagnosis.

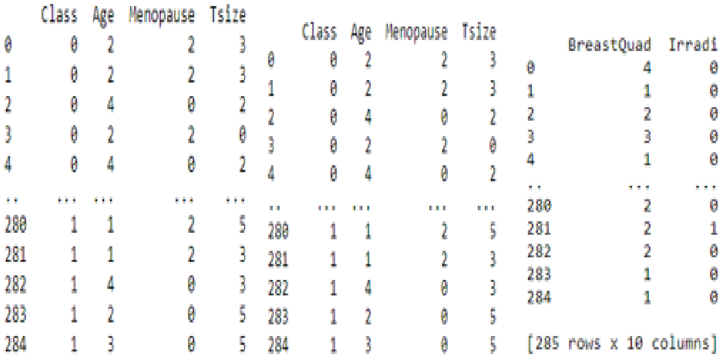
3. Menopause: menopause status of the patient at the time of diagnosis;
4. Tumor size: tumor size (in mm).
5. Inv-nodes: range 0 - 39 of axillary lymph nodes showing breast cancer at the time of histological examination.
6. Node caps: penetration of the tumor in the lymph node capsule or not.
7. Degree of malignancy: range 1-3 the histological grade of the tumor. That are
grade: 1 predominantly that consist of cancer cells.
grade: 2 neoplastic that consist of usual characteristics of cancer cells.
grade: 3 predominately that consist of cells that are highly affected.
8. Breast: breast cancer may occur in either breast.
9. Breast quadrant: if the nipple consider as a central point the breast may be divided into four quadrants;
10. Irradiation: patient's radiation (x-rays) therapy history.

Attributes	Values
Class	Yes, no
age	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
Menopause	lt40, ge40, premeno
Tsize	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44,45-49, 50- 54, 55-59
InvNodes	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26,27-29, 30-32,
NodeCaps	yes, no
DegMalig	1, 2, 3
Breast	left, right
BreastQuad	left-up, left-low, right-up, right-low, central
Irradi	yes, no

Table 1:

3.2 Data Processing

In our machine learning model, we used a data set file whose name is "cancer.csv" and it's attributed contains categorical and continuous type values. As we use scikit learn library for developing our machine learning model and scikit learn library cannot work with categorical and continuous-type data, so we encode our data to numerical format by using "LabelEncoder()" function. Shown in Figure 3.

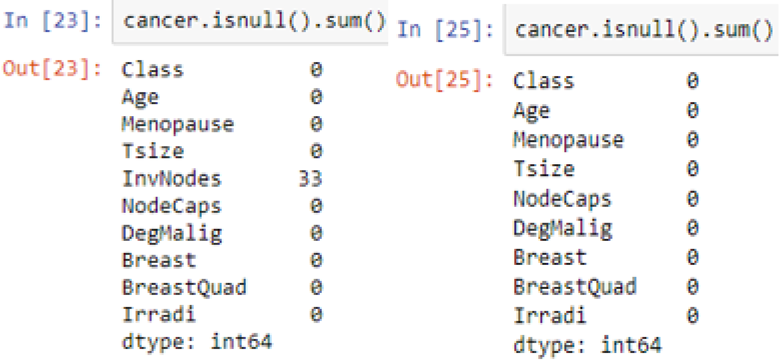


```
     Class  Age  Menopause  Tsize
0        0    2          2      3
1        0    2          2      3
2        0    4          0      2
3        0    2          2      0
4        0    4          0      2
..     ...  ...        ...    ...
280      1    1          2      5
281      1    1          2      3
282      1    4          0      3
283      1    2          0      5
284      1    3          0      5
```

```
     Class  Age  Menopause  Tsize
0        0    2          2      3
1        0    2          2      3
2        0    4          0      2
3        0    2          2      0
4        0    4          0      2
..     ...  ...        ...    ...
280      1    1          2      5
281      1    1          2      3
282      1    4          0      3
283      1    2          0      5
284      1    3          0      5
```

```
     BreastQuad  Irradi
0             4       0
1             1       0
2             2       0
3             3       0
4             1       0
..          ...     ...
280           2       0
281           2       1
282           2       0
283           1       0
284           1       0

[285 rows x 10 columns]
```

Also, there was some missing and noise data in our data set in a particular attribute whose name is “InvNode”. As there has no particular significance for this attribute so we remove that attribute from our data set. Sown in figure 4.



```
In [23]: cancer.isnull().sum()

Out[23]: Class          0
         Age            0
         Menopause      0
         Tsize          0
         InvNodes      33
         NodeCaps       0
         DegMalig       0
         Breast         0
         BreastQuad     0
         Irradi         0
         dtype: int64
```

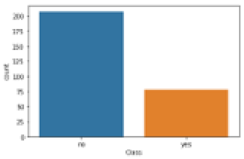
```
In [25]: cancer.isnull().sum()

Out[25]: Class         0
         Age           0
         Menopause     0
         Tsize         0
         NodeCaps      0
         DegMalig      0
         Breast        0
         BreastQuad    0
         Irradi        0
         dtype: int64
```

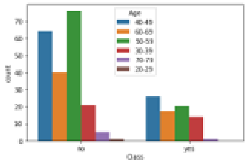
3.3 Model Development:

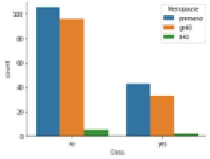
To practice of machine learning algorithms,Anaconda software used. All experiments described in this report were performed using libraries from Python machine learning environment. Python is a workbench for machine learning that is intended to aid in the application of machine learning techniques to a variety of realworld problems. Naïve Bayes, SVM and Logistic Regression have been used to predict attributes such as age, menopause, tumor-size, inv-nodes, node-caps, degmalig, breast, breast-quad, irradiation and class for chances of a patient getting breast cancer.

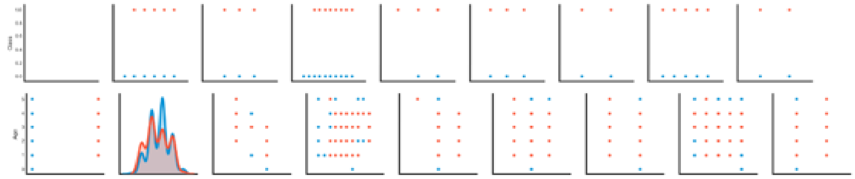
3.4 Results Analysis:

Initially, using machine learning algorithms analysis made with obtained experimental results from classifications on breast cancer data set. Result, after analyzing breast cancer data set visually using different attributes and figure out the distribution of values as shown in figure 5.





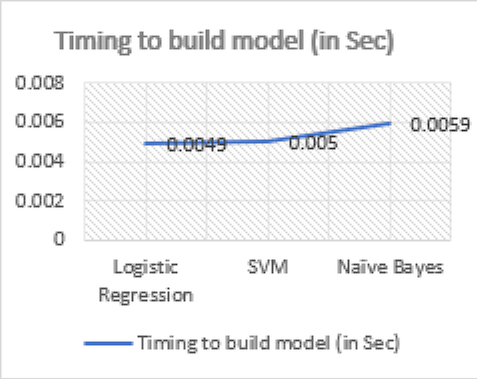


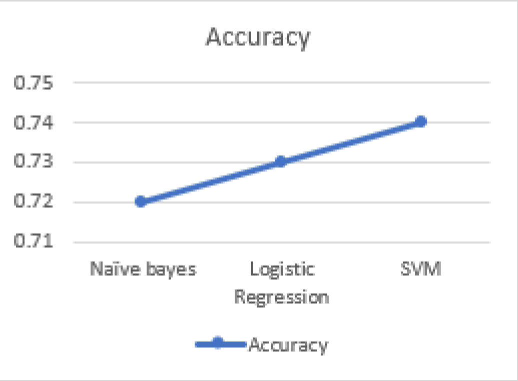


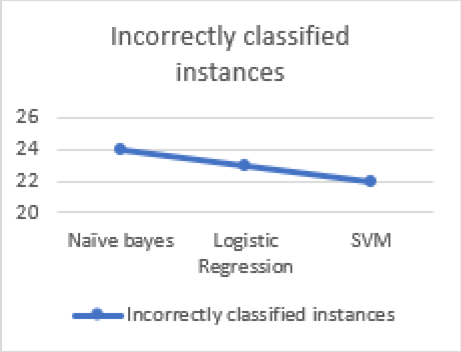
In our experimental result section we have carried out some experiments in order to evaluate the performance of different algorithms for predicting breast cancer in order to time to build a model and accuracy in Table 2. From above table we can conclude that SVM is more accurate classifier in comparison of others also it can be easily seen that it has highly classified correct instances as well as incorrectly classified instance than Naïve Bayes and Logistic Regression (see figure 6).

Evaluation Criteria	Naïve bayes	SVM	Logistic Regression
Timing to build model (in Sec)	0.0059872	0.0050223	0.0049856
Accuracy	0.72	0.74	0.73
Incorrectly classified instances	24	22	23

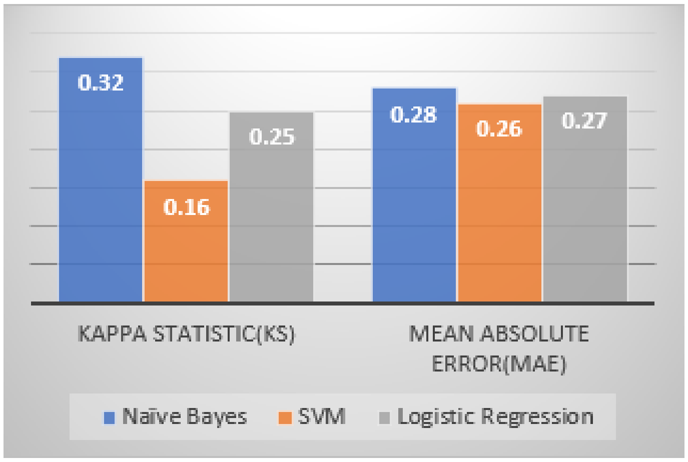
Table 2:





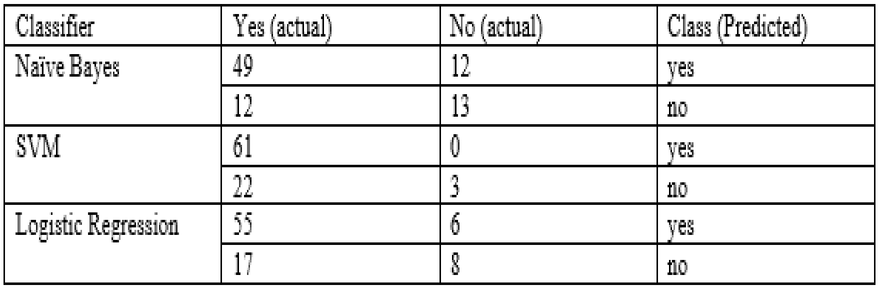


Kappa statistic and mean absolute error will be in numeric value only. The results of the simulation are shown in Figure 7.



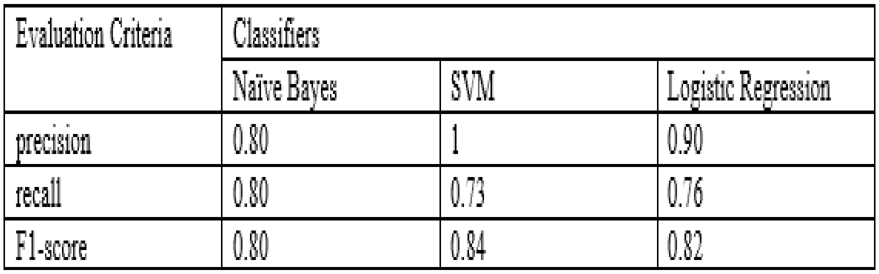
The performance of the learning techniques is highly dependent on the nature of the training data. Confusion matrices are very useful for evaluating classifiers. The columns represent the predictions, and the rows represent the actual class.

To evaluate the robustness of classifier (see Table 3).



Classifier	Yes (actual)	No (actual)	Class (Predicted)
Naïve Bayes	49	12	yes
	12	13	no
SVM	61	0	yes
	22	3	no
Logistic Regression	55	6	yes
	17	8	no

Once Predictive model is created, it is necessary to check how accurate it is, the accuracy of the predictive model is calculated based on the precision, recall values of classification matrix. Table 4 below shows the precision, recall and F1-score for Naïve Bayes, SVM, Logistic Regression.



Evaluation Criteria	Classifiers		
	Naïve Bayes	SVM	Logistic Regression
precision	0.80	1	0.90
recall	0.80	0.73	0.76
F1-score	0.80	0.84	0.82

Table 4: Classification Report
From above table we can conclude that SVM is more accurate classifier in comparison of others.

4. Conclusion:

In this report, we report on a research effort where we developed several prediction models for breast cancer. Specifically, we used three popular machine learning algorithms: Naïve Bayes, SVM and Logistic Regression. The best algorithm based on the patient's data is SVM Classification with accuracy of 0.74 and the total time taken to build the model is at 0.005 seconds. These results suggest that among the machine learning algorithm tested, SVM classifier has the potential to significantly improve the conventional classification methods used in the study.

4.1 Challenges:

In coding section of this project we use scikit learn building library for applying algorithms in our data set. For this reason we have to work with default function parameters for any algorithm. Another limitation of scikit learn library is it can't work with categorical and continuous type value directly. So a significant portion of our work is highly depended on Building Function Library.

4.2 Limitations:

In the coding section of this project, we use scikit learn building a library for applying algorithms in our data set. For this reason, we have to work with default function parameters for any algorithm. Another limitation of scikit learns library is it can't work with categorical and continuous type value directly. So a significant portion of our work is highly dependent on Building Function Library.

5. References:

1. Witten, I.H. and Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann (2005).
2. Review of cancer from perspective of molecular.
3. Tsirogiannis, G.L, Frossyniotis, D, Stoitsis, J, Golemati, S, Stafylopatis, A Nikita,K.S," Classification of Medical Data with a Robust Multi-Level Combination scheme", IEEE international joint Conference on Neural Networks.
4. My Chau Tu, Dongil Shin, Dongkyoo Shin,"Effective Diagnosis of Heart Disease through Bagging Approach", 2nd International Conference on Biomedical Engineering and Informatics,2009.
5. V. Chauraisa and S. Pal, "Early Prediction of Heart Diseases Using Data Mining Techniques", Carib.j.SciTech,Vol.1, pp. 208-217, 2013.
6. V. Chauraisa and S. Pal, "Data Mining Approach to Detect Heart Diseases", International Journal of Advanced Computer Science and Information Technology (IJACSIT),Vol. 2, No. 4,2013, pp 56-66.
7. https://stackabuse.com/
8. https://www.datacamp.com/
9. https://jakevdp.github.io/