Text

Description automatically generated with medium confidence

Machine learning foundation

Project

Name: Mohammed Yahya Al Sultan

Registraion number: 11908184

Section:KM046

Email: mohammed20yahya@gmail.com

Phone number: 7696051325

P

Abstract:

Breast Cancer screening is an essential way to allow prediction in early time. Thus ,to detect the malicious disease in premature time. It is also used to improve the treatment by determining which symptom are contributing the most to the detection of Breast Cancer. Stable model for prediction is to be made using routine consultation such as blood analysis that can be effectively contributing to the model that is to be made. By using the dataset provided in UCI platform I could make the prediction model using different algorithm and tools[1].

Breast Cancer Coimbra is a type of classification problems at which to predict the presence or not of the Breath Cancer, using two labels (1: to indicate Healthy controls and 2: Patients). Here, there are ten predictors which are all [1]quantitively measured to indicate the existence of this type of cancer. The quantitative Attributes includes: Age ,BMI ,Glucose ,Insulin ,HOMALeptin , Adiponectin, Resistin ,MCP-1.

The goal here is to assess a model for prediction that can be used as biomarker of breast cancer, based on anthropometric along with parameters that can be collected in routine blood analysis. This prediction can be done by using different classification algorithm such as Perceptron, linear regression, Random\_forest .The dataset provided in all the model is to be pre-processed in order to enhance the result that is desired[2].

**Pre-process the dataset**

**Handling categorical data**

**Removing irrelevant features**

**Handling Missing values**

**Splitting the data into training and testing**

**Deploy feature engineering**

**Train the model**

No

yes

**Is trained**.

**Methodology:**

1. installing the datasets from UCI web page
2. importing the dataset to our program along with all required libraries such as numpy,pandas etc.
3. By initial visualizing of our dataset ,we see that this dataset does not have any categorical data .we also find out that this type of data requires classification algorithm with target of two possible values 1,2
4. Dividing the whole dataset into two parts (input and target).
5. Further splitting the data into training part and testing part by using sklearn.model\_selection with the function of train\_test\_split
6. Framing a training model by using different classification algorithms such as (perceptron,linearregression,support vector machine ,kneighborclassifier,gaussianNB,etc)
7. Verfy all the algorithms’ accuracy and time spent in training the model

8-Apply feature Engineering including feature selection. By using Randomforest we examine the importance of all features

9-Using feature selection such select\_from\_model

10-apply feature extraction including PCA(principle component analysis ,which is unsupervised),and also LDE (linear discriminant analysis ,which a supervised method)

11-verify different accuracy measurement techniques including precision, accuracy ,r2 ,etc.

12-using roc\_curve,auc for further visualization.

13-Using pipelining and visualize the accuracy

14-Use Hyper tunning to test the value of parameter the best fit the model in order to increase performance

**Result and discussion:**

A model is to be made for the detection of breast cancer using Breast Cancer Coimbra datasets ,where it contains of 10 columns and 115 samples.

![A picture containing text, newspaper

Description automatically generated]()

![Chart, bar chart, histogram

Description automatically generated]()

This dataset is of a classification model, where it is being passed through several classifier and each of which is given different accuracy and time of execution, as follows:

![Chart

Description automatically generated]()

In terms of time

feature engineering in the essence of knowing which features are contributing more than others.Hence, they are chosen in order to reduce the complexity of the model.

By using features important method we can see that feature number 2 is the highest contributing column in our dataset .Therefore, it is can be chosen along with other high importance features.

![Chart, bar chart

Description automatically generated]()

![Calendar

Description automatically generated with medium confidence]()

After the training process, I realized that linear Regression is given the highest accuracy and hence I further used feature engineering along with this model to gain better results.

|  |  |
| --- | --- |
| Feature engineering model | Accuracy gained in linear Regression model |
| Model selection with 4 features | 71% |
| PCA with 4 components | 65% |
| PCA with 7 components | 85% |
| LDA with 1 component | 77% |
|  |  |
|  |  |
|  |  |

Be further exploring the different performance measurement of this model we get:

![A picture containing square

Description automatically generated]()![A picture containing text, receipt

Description automatically generated]()

Further visualising can be done using AUC(area under curve) and ROC(received operating characteristic).

It shows that the are under curve=23%.

![Chart, line chart

Description automatically generated]()

Conclusion:

The model containing the dataset of Breast Cancer Coimbra has been trained using different classification algorithms, each of which has given different complexity in terms of time as well as accuracy .as a result. Linear regression has shown the highest accuracy and the decision tree is the lowest .In terms of time complexity random forest was had maximum value. The model has been passed through different feature engineering methods. By using select\_from\_model method ,we can say that features number ,3,5,8 are contributing the most into the learning process. Also, by using PAC ,LBA I could extract the data into new format and the accuracy has increases slightly.

**References**

1. Using Resistin, glucose, age and BMI to predict the presence of breast cancer 2. Link : https://bmccancer.biomedcentral.com/articles/10.1186/s12885-017-3877-1
2. Crisóstomo J, et al. Hyperresistinemia and metabolic dysregulation: the close crosstalk in obese breast cancer. Endocrine. 2016;53(2):433-42.

Link for dataset:

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra