Introduction to Data Science

Tshifaro Justin Ndivhuwo(SUQG0J)

December 2021

Abstract

The aim of this report is to demonstrate the results of Classification, Clustering and Frequent pattern mining performed on the Breast cancer data set.

Classification, Clustering KMeans

1 Introduction

Machine learning has became a crucial tool in most industries that exist today, this include medical industries.Using known gynaecological measurements and other cancer traits, Machine Learning algorithms can be utilized to predict women at risk of recurrence. This approach is non-invasive, low-cost, faster and can be used to identify asymptomatic women at risk of cancer recurrence in the future. Data can be used to predict the results of many diseases. In this case we will try to predict the recurrence of cancer. Recurrence means the cancer came back after a successful treatment. We will discuss the important algorithms for Classification, Clustering and Frequent pattern mining.

2 Data set Information

Understanding the data set is the first step to be executed. We have the data set breast-cancer that was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia.Thanks go to M. Zwitter and M. Soklic for providing the data.

The breast cancer data set has 286 instances and 10 Attributes including the class attribute. The data had few missing values which are denoted by "?". Most of the attributes are intervals(non numeric). The table shows information about the data set.

Attributes	Attributes Details	Values
Class	Reappearing of breast cancer after treatment	no-recurrence-events, recurrence-events.
age	Age of the patient when tumor was discovered	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.
menopause	whether the patient is pre- or postmenopausal at time of diagnosis	lt40, ge40, premeno
tumor-size	The size of the tumour	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44,45-49, 50-54, 55-59.
inv-nodes	the number of auxiliary lymph nodes	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.
node-caps	Has the tumour reached Node caps	yes, no
deg-malig	Degree of malignancy	1, 2, 3.
breast	The side of the breast	left, right.
breast-quad	Breast quadrant	left-up, left-low, right-up,right-low, central.
irradiat	Irradiation	yes, no.

Breast cancer data set information

Data visualization is the representation of data or information in a graph, chart, or other visual format.Data visualization gives us a clear idea of what the information means by giving it visual context through maps or graphs. This makes the data more natural for the human mind to understand and therefore makes it easier to identify patterns, and outliers within large data sets From the above figures we can see and learn that the Breast cancer data set was not balanced. We can clearly see that there are more non-recurrence cases than recurrence, this was achieved by plotting "class" vs all attributes.

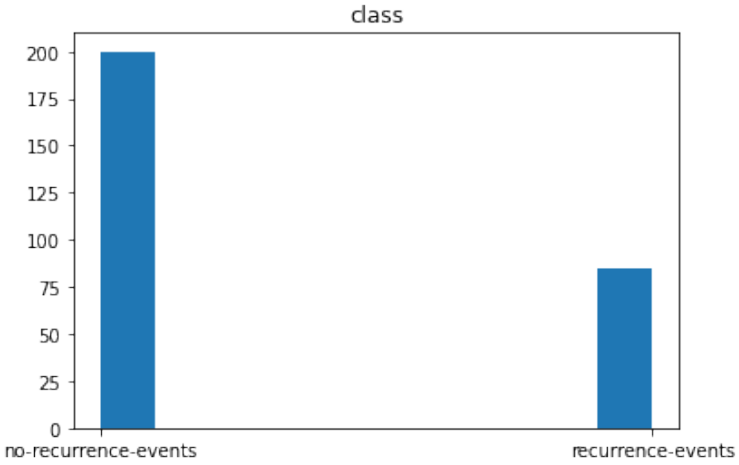

Figure 1: non-recurrence events and recurrence events

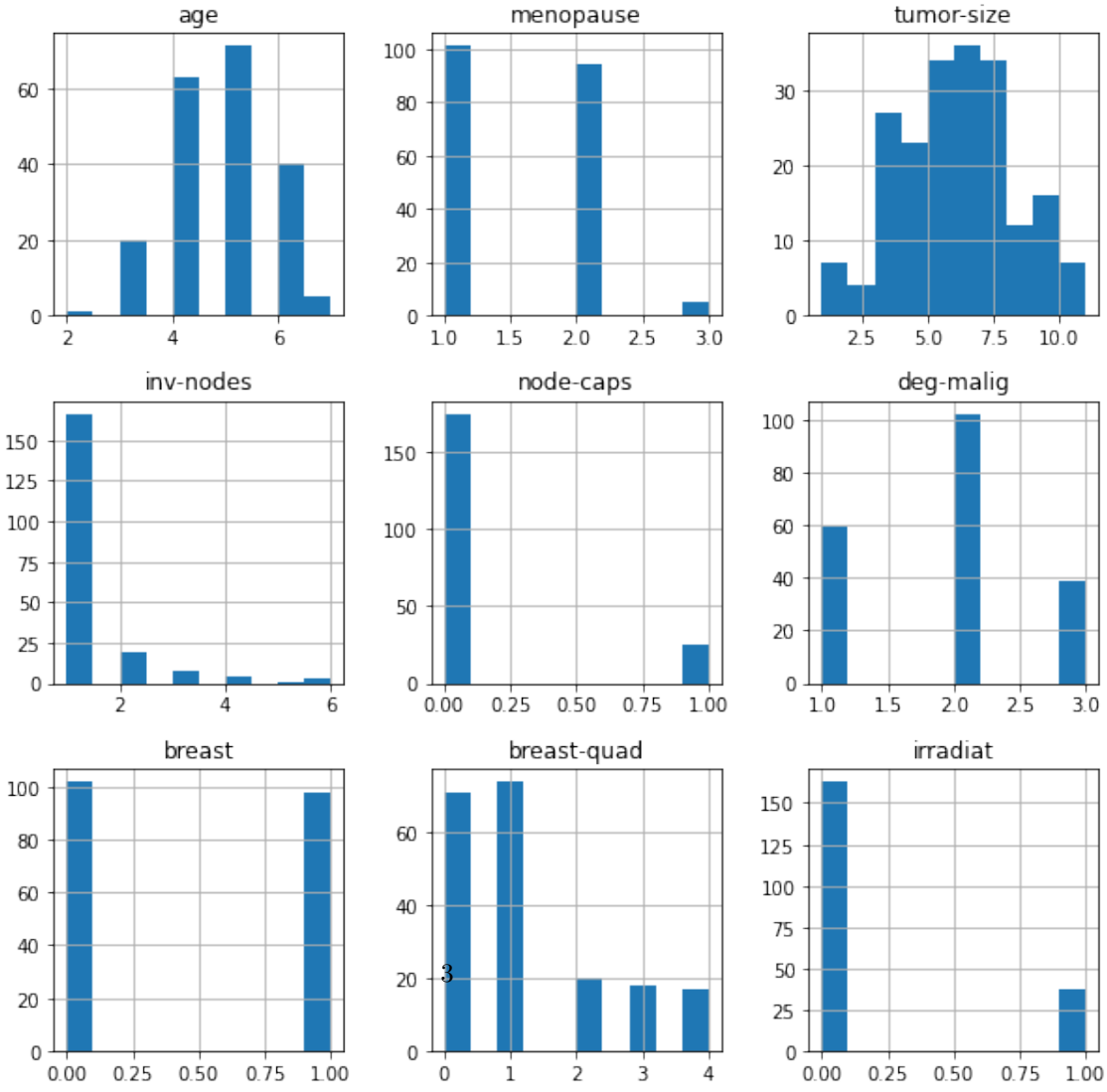

Figure 2: Class vs attributes

3 Data Cleaning and processing

Data processing is important in data mining as it provides better results that are, more accurate and reliable. The aim is to organise the data and covert it into its best organised form

The data set had few missing attributes values in node-caps and breast-quad, the missing attribute were replaced by the the value that appeared frequently. This idea limit the chances of completely changing the data set. An important part of Data analysis is analyzing duplicate Values and removing them, in order to get more accurate results.

Most variables in the data set were provided as strings and they had to be converted into numerical values for analysis. The table below demonstrate the numerical value representation

Attributes	Numerical Values
Class	no-recurrence-events = 1, recurrence-events = 0.
age	10-19 = 1, 20-29 = 2, 30-39 = 3, 40-49 = 4, 50-59 = 5, 60-69 = 6, 70-79 = 8, 80-89 = 9, 90-99 = 9.
menopause	premeno = 1, ge40 = 2, lt40 = 3,
tumor-size	0-4 = 1, 5-9 = 2, 10-14 = 3, 15-19 = 4, 20-24 = 5, 25-29 = 6, 30-34 = 7, 35-39 = 8, 40-44 = 9, 45-49 = 10, 50-54 =11, 55-59 = 12.
inv-nodes	0-2= 1, 3-5= 2, 6-8= 3, 9-11 = 4, 12-14= 5, 15-17 = 6,18-20 = 7, 21-23 = 8, 24-26 = 9, 27-29 = 10, 30-32 = 11, 33-35 = 12, 36-39 = 13.
node-caps	yes = 1, no = 0
deg-malig	1, 2, 3.
breast	left = 0, right = 1.
breast-quad	left-up = 0, left-low = 1, right-up = 2,right-low =3, central = 4.
irradiat	yes = 1, no = 0.

Converting Strings to Numerical Values

4 Methods

Machine learning has different methods. In this project we will only explore classification and clustering

4.1 Classification

First we need to understand what is classification and how it works. This data mining technique is used to sort things in data sets into classes or groups. It aids in precisely predicting the behavior of entities within the group. It's a two-part procedure:

1. Training step :A classification algorithm analyzes a training set to create the classifier.

2. Classification step : The accuracy or precision of the classification rules is estimated using test data.

In this case only Decision tree classifier, K-Neighbors classifier and Random forest classifier.

4.1.1 Decision Tree Classifier

Decision Tree is a Supervised learning technique that is mostly used for classification, but can also be used for regression. It has internal nodes that represents feature of the data set, branches that represent decision rules and leaf nodes that represent outcomes. It's termed a decision tree because it starts with a root node and grows into a tree-like structure.

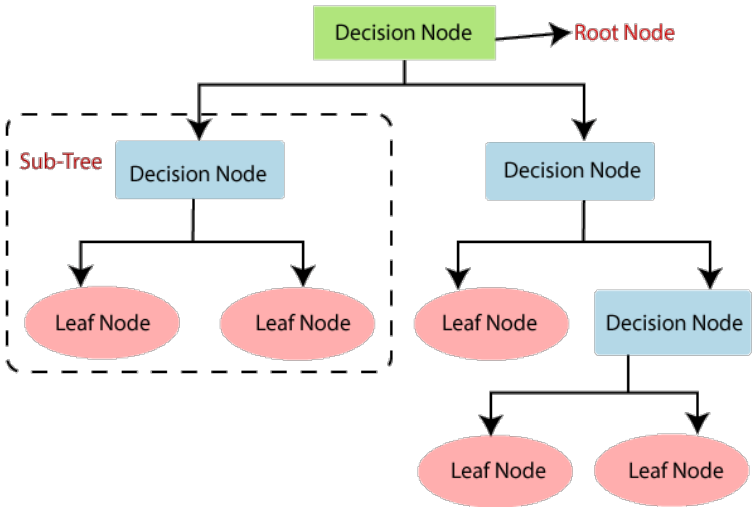

Figure 3: Decision Tree

Decision Trees are designed to mirror human decision-making abilities, so they are simple to understand.

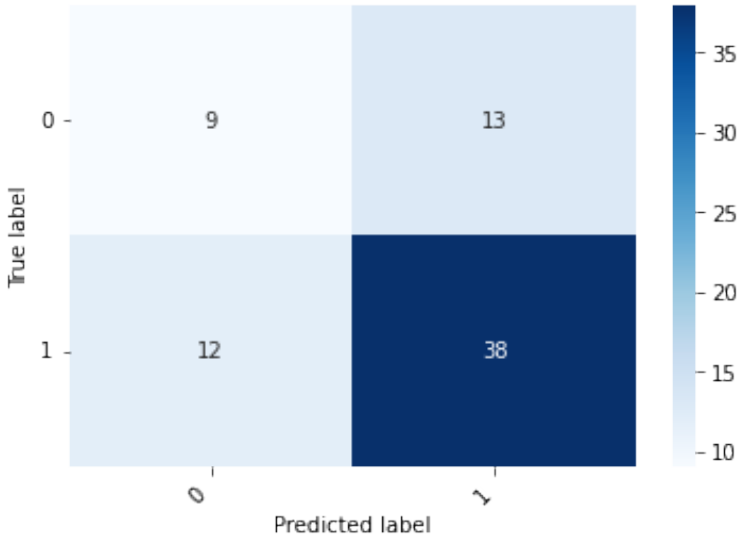

Figure 4: Decision tree results

Decision tree accuracy was 65,27% which is quite low. The model did not work well on the breast cancer data set.The number 38 corresponds to the number of customers who were correctly predicted by the model to no-recurrence-events, which means cancer has not came back, while number 9 corresponds to the number of customers that the model correctly predicted recurrence-events.

4.1.2 k-Nearest Neighbors

k-Nearest Neighbors is a supervised learning model that is used to solve the classification model problems. K-nearest neighbor or K-NN algorithm basically creates an imaginary boundary to classify the data. Whenever new data comes the model will try to predict and classify it to the nearest boundary line.

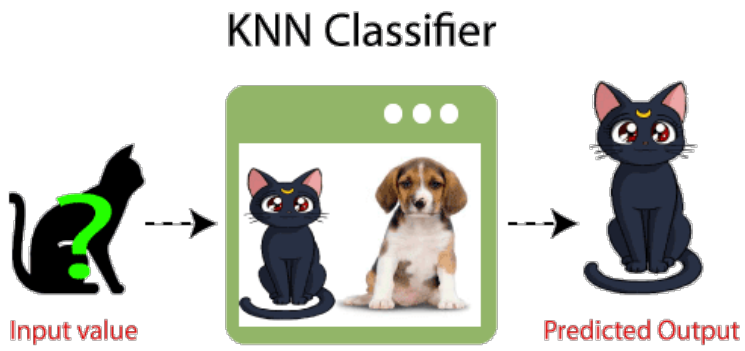

Figure 5: k-Nearest Neighbors

k-Nearest Neighbors accuracy was 66.67% which is quite low as well. The model did not work well on the breast cancer data set

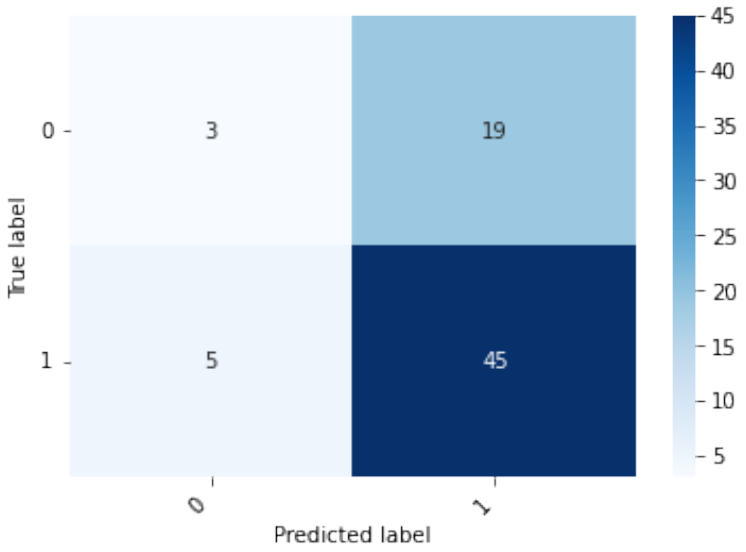

Figure 6: k-Nearest Neighbors results

45 correspond to no-recurrence-events while 3 correspond to recurrence-events. Which means most patient are most likely not to have any recurrence.

4.1.3 Random Forest Algorithm

Random Forest Algorithm is a well known supervised learning technique, it can be used for for both classification and regression. It is base on concept of ensemble learning. Ensembles means combining multiple models

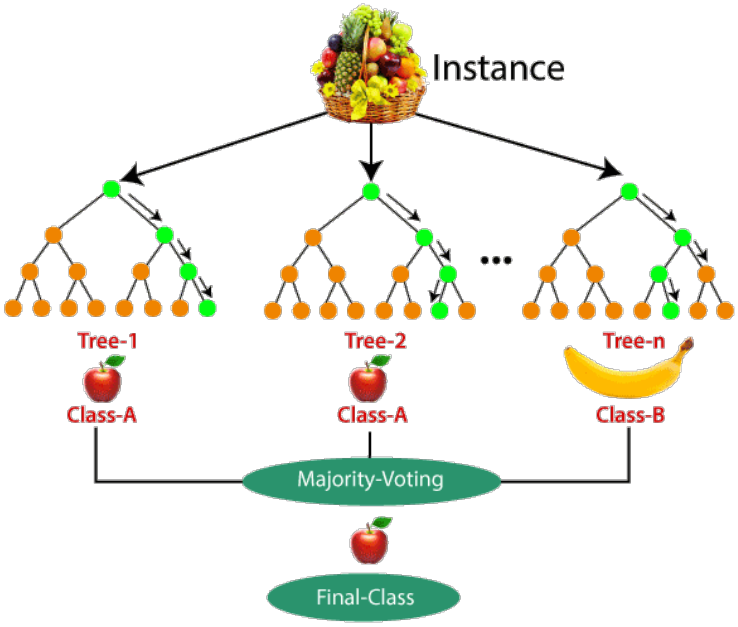

Figure 7: Random Forest Algorithm

Random Forest Algorithm achieved an accuracy of 76.39%, which is better than both Decision tree and K-NN.

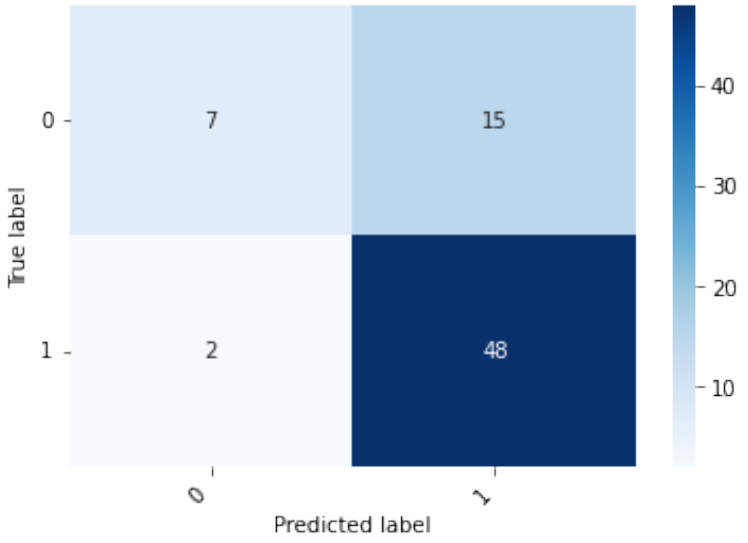

Figure 8: Random Forest Algorithm results

48 correspond to no-recurrence-events while 7 correspond to recurrence-events. Which means most patient are most likely not to have any recurrence.This model performed better than the rest

4.2 Clustering

Clustering, often known as cluster analysis, is a machine learning technique that groups unlabeled data into groups. It is definable. Clustering is very important for machine learning. For the breast cancer data set we used K-Means clustering algorithm, Agglomerative clustering, and BIRCH clustering. Scatterplot alongside Principal Component Analysis (PCA) was used to visualize the result of clustering

4.2.1 K-Means clustering algorithm

K-Means Clustering is an unsupervised learning approach used in machine learning to solve clustering problems. It groups the data that is not labeled into different clusters. K is the number of clusters that needs to be created K=2, means there will be two clusters.

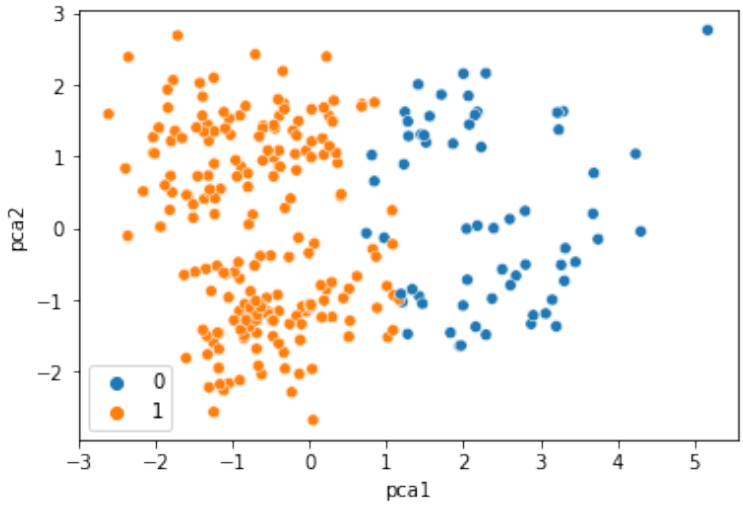

Figure 9: K-Means clustering algorithm Scatter plot

The scatter plot for K-Means shows that recurrences are more than no-recurrences. More patients will cancer will not reoccur after treatment.

4.2.2 Agglomerative clustering

One most frequent type of hierarchical clustering used to put objects in clusters based on their similarity is agglomerative clustering. AGNES is another name for it (Agglomerative Nesting).In this algorithm we don't have to define the number of clusters in advance.It start by treating each piece of data as a singleton cluster, then agglomerate pairs of clusters until all of them have been merged into a single cluster that contains all of the data.

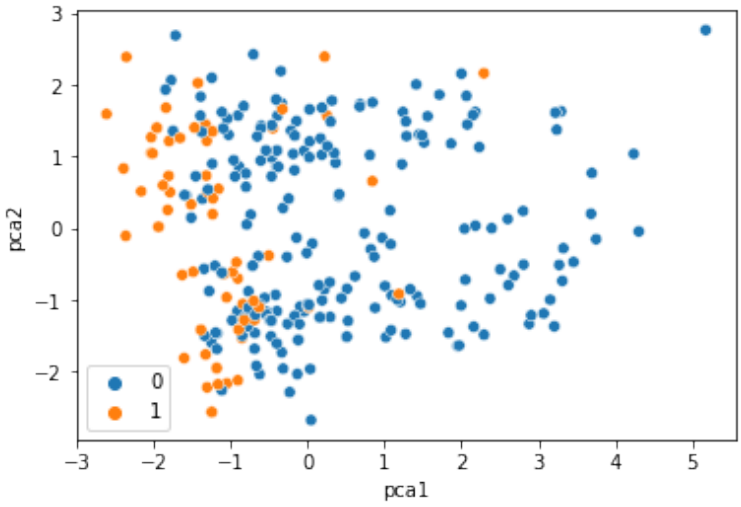

Figure 10: Agglomerative clustering Scatter plot

The results for agglomerative clustering shows recurrences to be more than no-recurrences. which means more patients will have cancer again, after treatment.

4.2.3 BIRCH Clustering

BIRCH stands for Balanced Iterative Reducing and Clustering using Hierarchies is Clustering algorithm that clusters huge data sets by first creating a tiny, compact summary of the large data set that maintains as much information as possible. Instead of clustering the whole data set, this smaller summary is clustered.

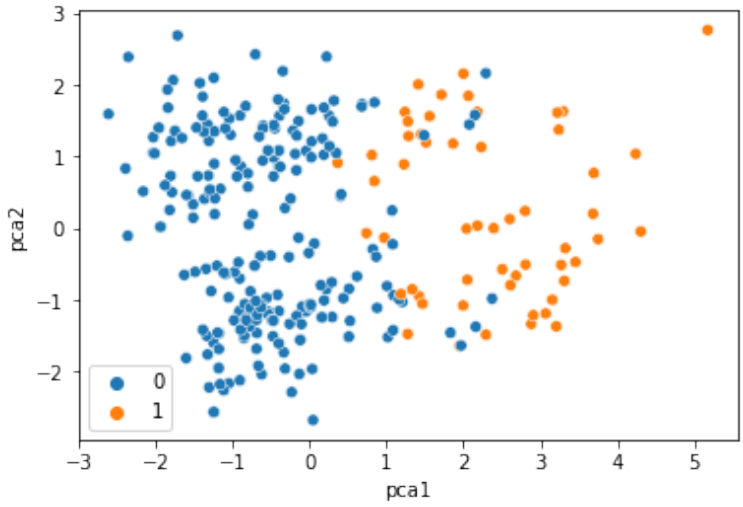

Figure 11: non-recurrence events and recurrence events

The results for birch shows recurrences to be slightly more than no-recurrences. More patients seems to have cancer again, after treatment.

5 Conclusion

In the processing phase, data was converted in numerical values, missing values were replaced by those that appeared most while duplicates were dropped. Three Classifying algorithm Decision tree classifier, K-Neighbors classifier, Random forest classifier were used for predictions and only Random forest classifier(76.39%) performed better than the rest, It had the highest accuracy. Three clustering algorithm were used, K-Means clustering algorithm, Agglomerative clustering, and BIRCH clustering. K-Means showed slightly high performance. There were 2 clusters and "no-recurrence-events" appeared to have more variables than recurrence, while the case is different in the other two algorithm.

For medical purpose, the prediction scores were quite low, it would put women with breast cancer in risk. The data set can be improved for better results as there are many important symptoms of breast cancer missing. Overall Machine Learning algorithms can be used to supplement the existing model of care for the early detection of Breast Cancer Recurrence Events.

References

[1] https://www.geeksforgeeks.org/ml-hierarchical-clustering-agglomerative-and-divisive-clustering/

[2] https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

[Nick McCullum] https://www.freecodecamp.org/news/how-to-build-and-train-k-nearest-neighbors-ml-models-in-python