Imam Mohammad Ibn Saud Islamic University

College of Computer and Information Sciences

Information Management Department

IM472

Data Analytics and Metrics Project

Student name	Student number
Nwara Aljoufi	440019357

Instructor name: Dr.Nouf Alshareef

Project tasks:

Task 1 Choosing Dataset:

- Choose a dataset from one of the repositories, most of them are ready for WEKA https://archive.ics.uci.edu/ml/datasets.php, http://www.is.umk.pl/~twin/dload_data.html
- You should select at least one dataset; a dataset should contain at least 100 instances.
- Justify your choice, why you choose them.

I found the topic is interesting since Breast cancer is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today's society.
The early diagnosis can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. So, correct diagnosis of Breast cancer and classification of patients into malignant or benign groups is the subject of much research.

- Set a clear objective for your experiment and what you exactly want to discover.

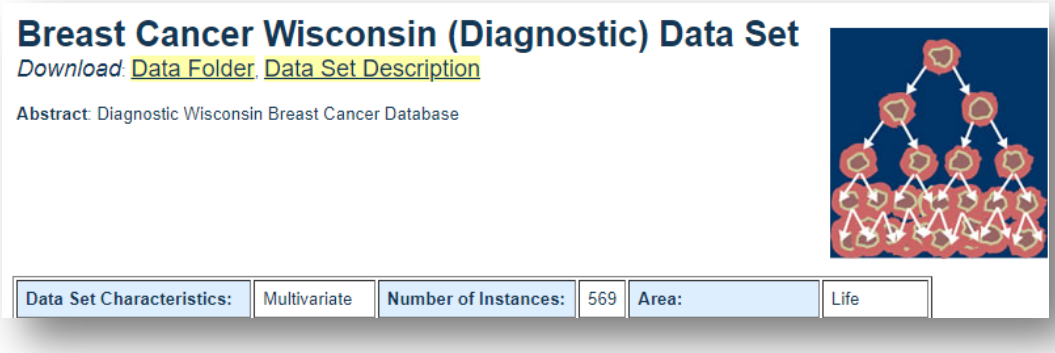
Classification and analysis data are an effective way to classify data. Especially in medical field, where those methods are widely used in involves the diagnoses the diseases correctly and providing the right advises.In this dataset we aims to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection. The goal is to classify whether the breast cancer is benign or malignant. To achieve this i have used machine learning classification methods to fit a function that can predict the discrete class of new input.

Task 2 Preparing data after pre-processing:

Download the data, and answer the following questions:

a. How many instances does the dataset contain?

My dataset contain 569 instances(rows).



b. How many attributes does the dataset contain? List the attributes and give the attribute data type.

Number of attributes in our dataset is 12

Attributes	Types
ID number	Categorical
Diagonsis(Malignant-Bengin)	Categorical(Boolean)
radius	Continuous
texture	Continuous
perimeter	Continuous
area	Continuous
smothness	Continuous
compactness	Continuous
concavity	Continuous
conave points	Continuous
symmetry	Continuous
fractal dimension	Continuous

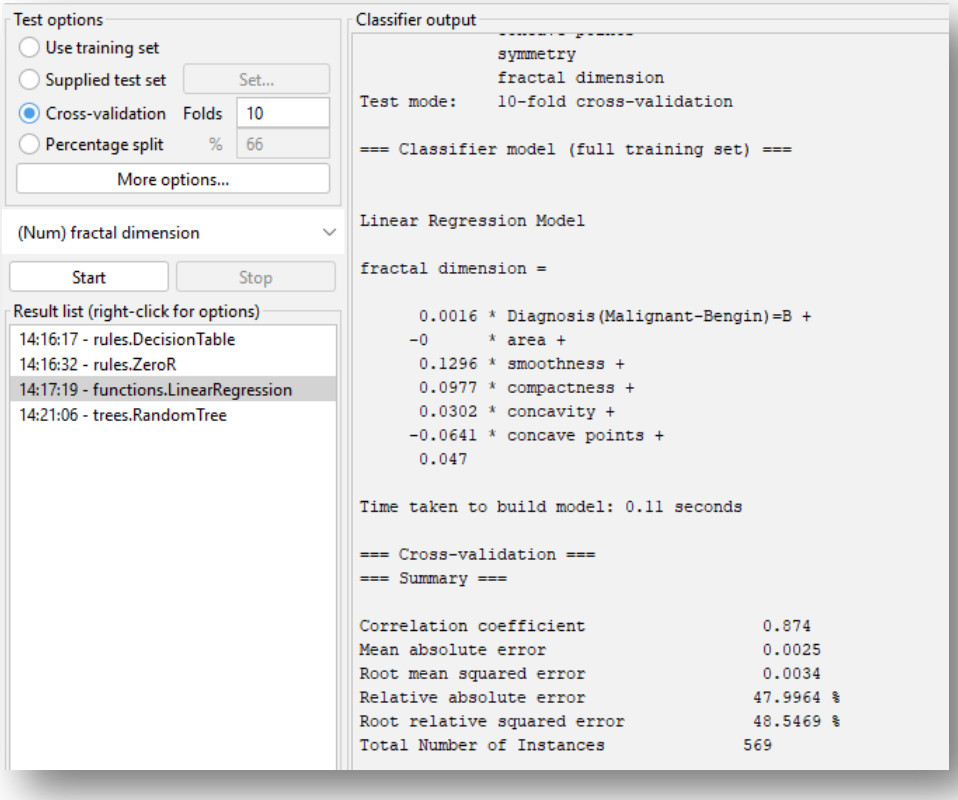
c. How many classes (classification label) does the dataset contain? List the classes

class label is the discrete attribute whose value you want to predict based on the values of other attribute *so in our dataset "Diagnosis*" is the column which we are going to predict.

Task 3 Investigating supervised algorithms:

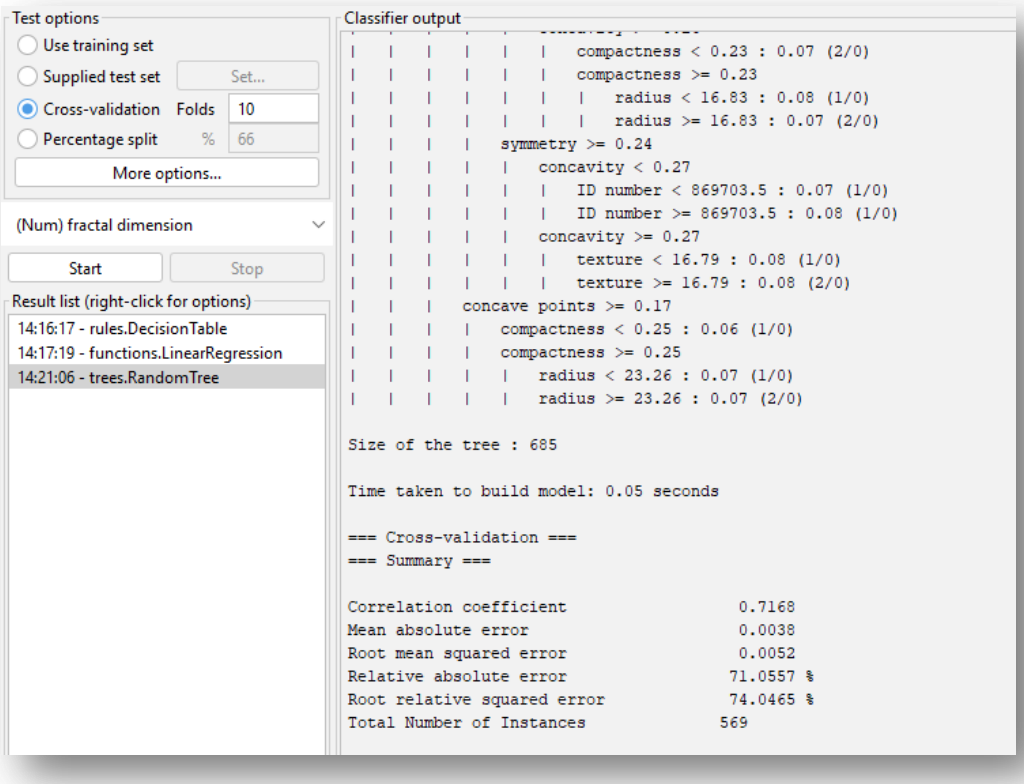
a. Select (supervised) classification algorithms and apply to your data.

We start using linear regression supervised.



b. Select another (supervised) classification algorithms from a different category and apply to your data.

Second choice is RandomTreee supervised algorithms.



c. How well do the selected classifiers perform?

- **In both classifiers have high corrleation cofficients since they're between 0.7 to 1.0 and the linear correlation coefficient when it is greater than zero it indicates a positive relationship between two normally distributed random variables.**

- **The labels on the test set are supposed to be the actual correct classification and both of them read correctly classified Instances with 569 rows.**

d. Compare the results of both algorithms

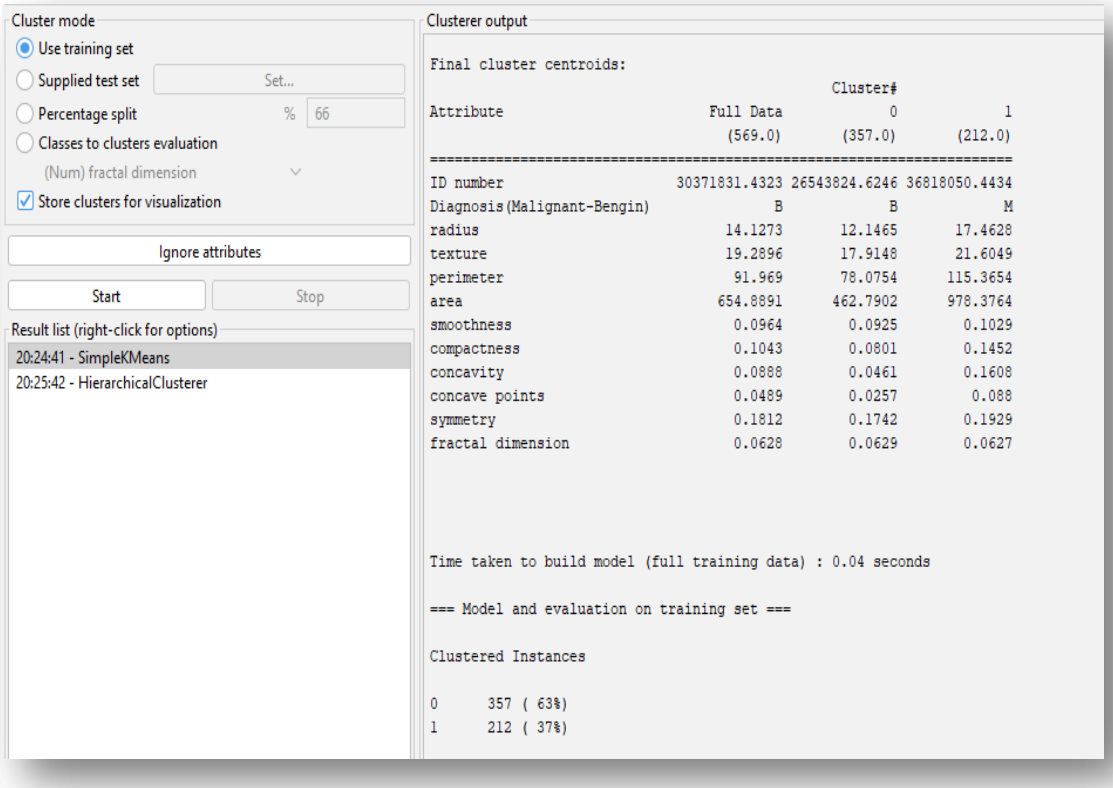
Error Measurement	Linear regression	RandomTree
Mean absoluate error	0.0025	0.7168
Root mean squared error	0.0034	0.0038
Relative absoluate error	47.9946%	71.0557%
Root relative squared error	46.5469%	74.0465%

In linear regression the measurement errors is less than in RandomTree so that's mean we can conclude that using linear regression is more efficient than RandomTree according to the selected criteria in table and gives the best results for our dataset.

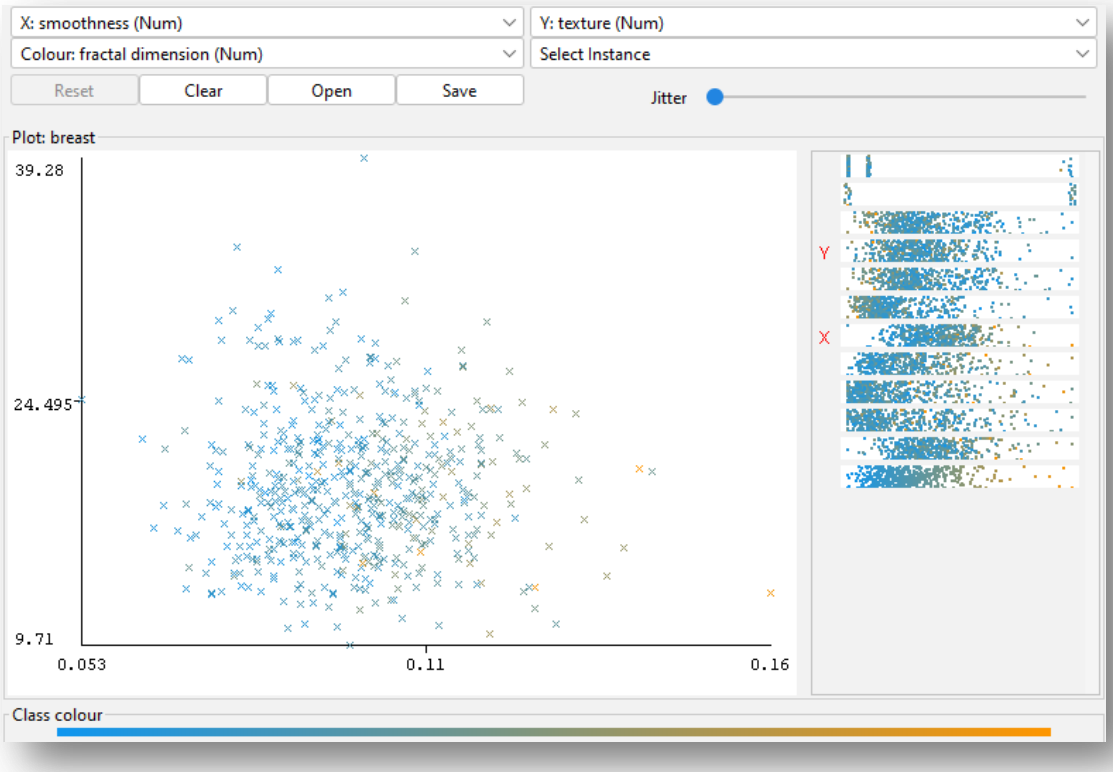
Task 4 Investigating unsupervised algorithms:

a. Select unsupervised algorithm and apply to your data.

I implement K-means clustering unsupervised algorithm.



Another way to grasp the characteristics of each cluster is to visualize them and below is the cluster between smothness and texture attributes.



b. How well does the selected method perform?

The accuracy results are shown in table.Simple k-means get the highest accuracy in a shorttime 0.04 seconds. In accordance with the obtained results, it can be told that, Simple k-means algorithm is the most proper clustering method for evaluation of the dataset class performance which says if the cancer 1 means the cancer is malignant and 0 means benign. We can identify that out of the 569 persons, 357 are labeled as B (benign) and 212 as M (malignant).

c. How does this compare to supervised classification?

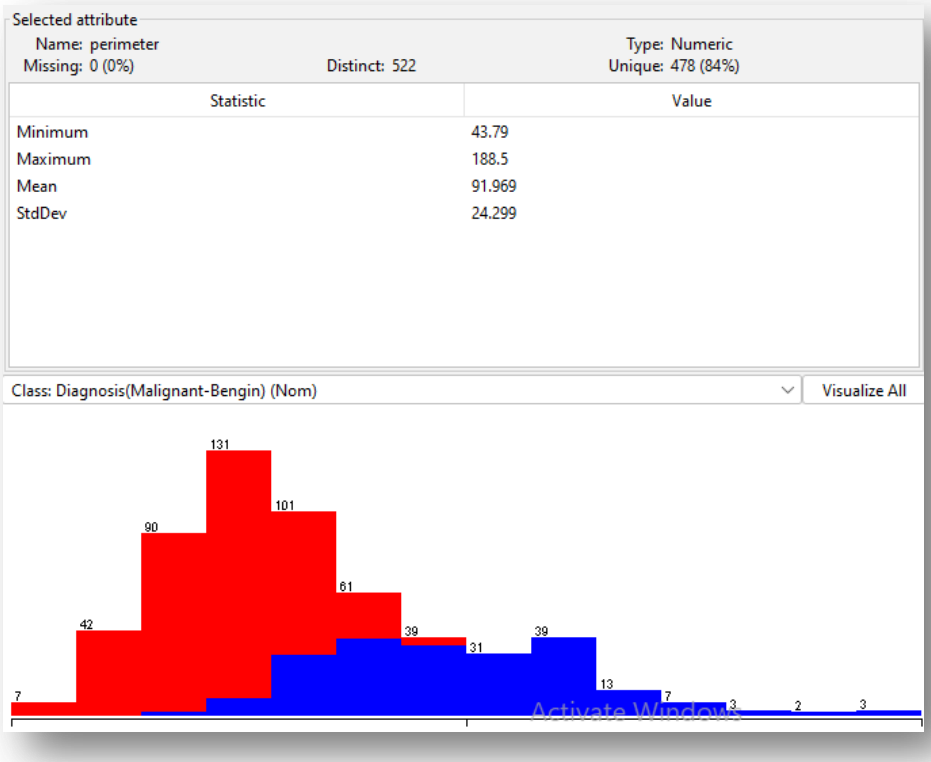
Task 5 Analyzing the results:

a. How useful is supervised and unsupervised learning for your dataset? (Did it meet your objectives?)

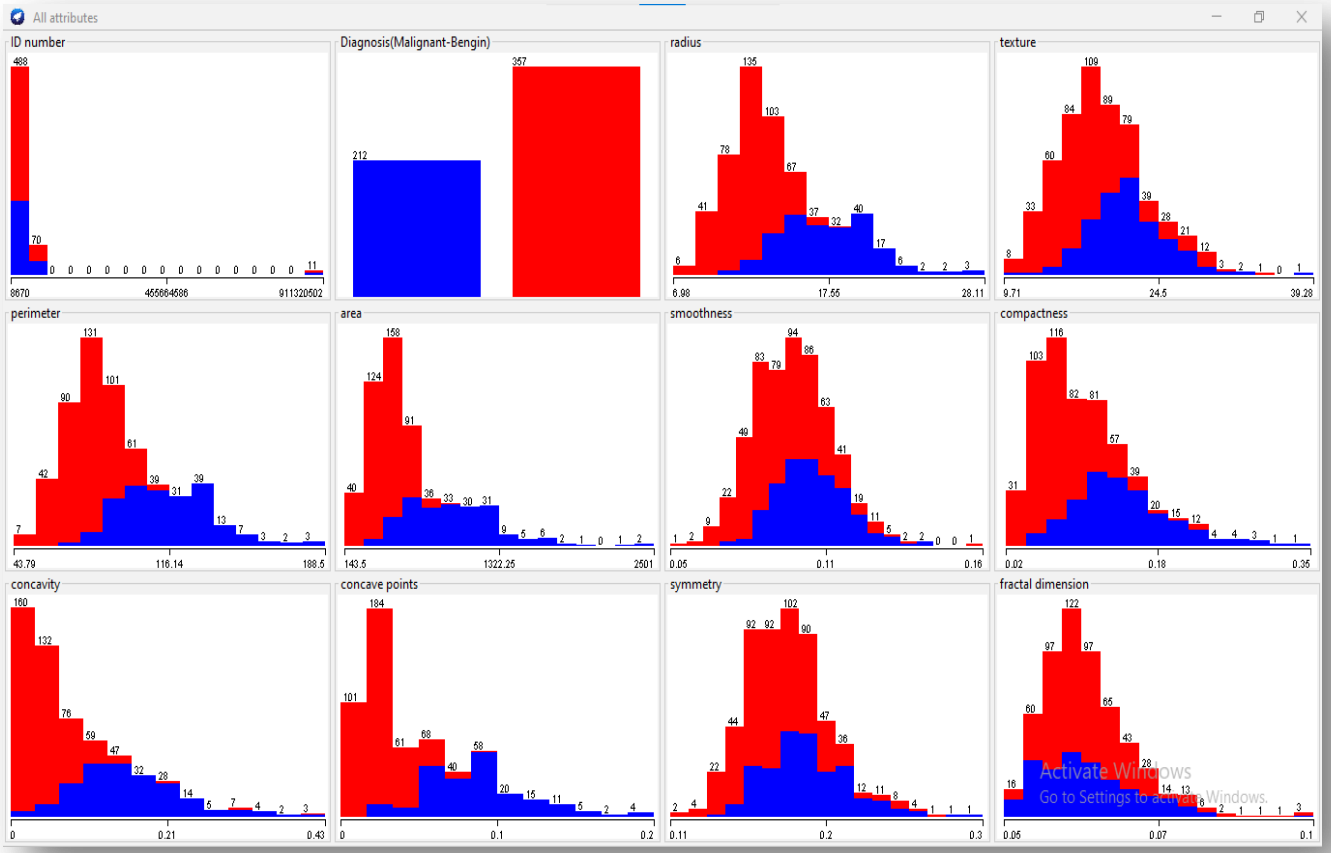
Patients with solid breast masses we found an easy-to-use graphical computer program which is capable of perform the analysis of cytological features based on a digital scan. The program uses a curve-fitting algorithm, to take the attributes in the sample under consideration, than it calculates the mean value, extreme value and standard error of each feature for the image and according to these the class "Diagnosis" is predicted.

b. What might the effect of missing data when classifying data?

Our dataset has 0 missing values as shown below and all other attributes is the same.



Our dataset attributes visualization



Refrences

https://archive-beta.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+diagnostic#Descriptive

https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/