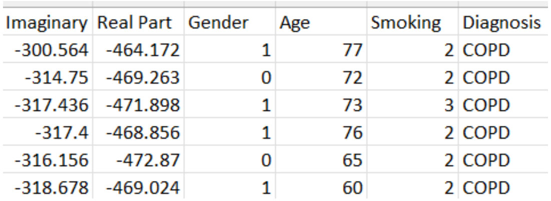
Data Mining - COP259

Coursework assignment

Credit value: 100% of the module

Part 1 – Pre-Processing

The data information was thoroughly analysed and modified on excel before loading the Exasens data set onto Weka.

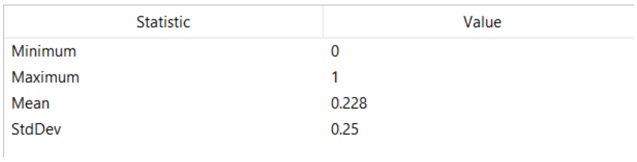


Imaginary	Real Part	Gender	Age	Smoking	Diagnosis
-300.564	-464.172	1	77	2	COPD
-314.75	-469.263	0	72	2	COPD
-317.436	-471.898	1	73	3	COPD
-317.4	-468.856	1	76	2	COPD
-316.156	-472.87	0	65	2	COPD
-318.678	-469.024	1	60	2	COPD

Figure 1- Modified Exasens dataset

Figure 1 shows a glimpse of the modified dataset, in this columns, ID, Imaginary Part – Min, and Real Part – Min was removed. Individual ID is unnecessary for classifying the data, as this attribute gives no correlation to the medical conditions. As this is clinical data, and can be patient specific and critical, all rows with data missing were deleted, this prevents creating a model without actual data, which can be risky, this leaves data set with 100 instances.

Diagnosis was moved to the last column, for personal preference as this is the class attribute, and this helps in Weka, as it identifies the last column as class automatically. Smoking attribute has already been numerically categorized, 1 – Non-Smoker, 2 – Ex-Smoker and 3 – Active Smoker. These steps were required before the pre-processing in Weka, as this may have been time consuming, and speeds up the learning process.



Statistic	Value
Minimum	0
Maximum	1
Mean	0.228
StdDev	0.25

Figure 3- Normalised Data - Imaginary Part

First pre-processing tool used was to **normalize** the data where min and max range are set to 0-1 respectively example shown in **Figure 2**, this was done to make the variables with different scales comparable and for ease of estimation whilst still preserving the shape of the original distribution, this also helps with duplicate data and minimise redundancy.

Figure 3 gives, some evidence of which attributes correlate together giving certain results. Important one to highlight is that the Saliva Permittivity – Imaginary Part, seems to not correlate well with our class (Diagnosis) which we are trying to classify, and neither does gender.

Standardising gave similar results except; this is when the mean is set to 0 and standard deviation is 1. Standardisation was chosen for the data set as it is more robust to any outliers, as in medical data all data is critical.

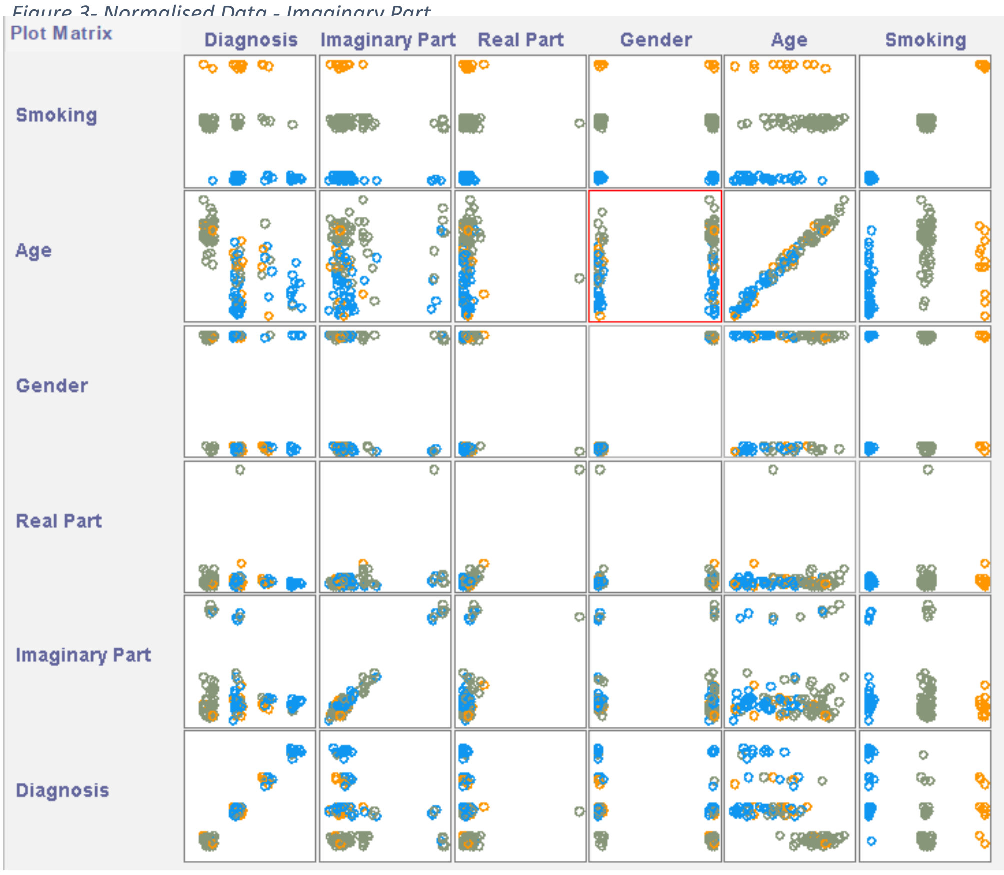
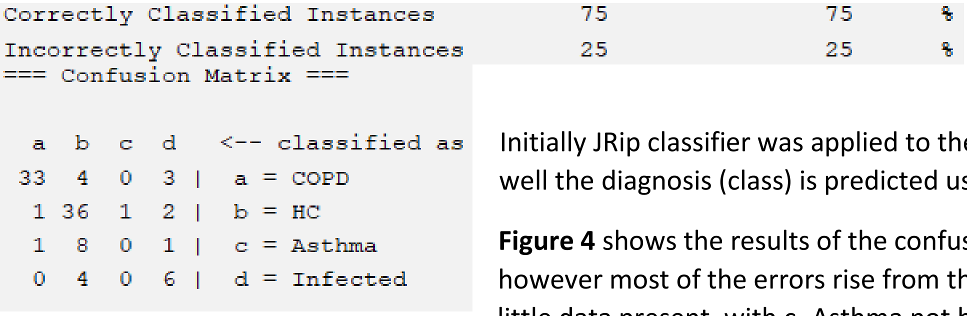


Figure 2- Plot Matrix



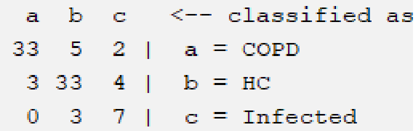
```
Correctly Classified Instances          75               75      %
Incorrectly Classified Instances        25               25      %
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 33  4  0  3 |  a = COPD
  1 36  1  2 |  b = HC
  1  8  0  1 |  c = Asthma
  0  4  0  6 |  d = Infected
```

Figure 4- Unfiltered - JRip Confusion Matrix

Initially JRip classifier was applied to the standardised data, to see how well the diagnosis (class) is predicted using all the attributes.

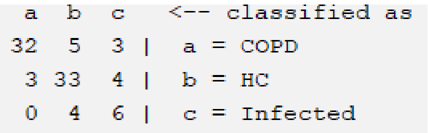
Figure 4 shows the results of the confusions matrix with 75% accuracy, however most of the errors rise from the diagnosis attributes with very little data present, with c- Asthma not being identified correctly at all. Likewise with d- Infected with 40% error. These may have to be removed for better learning from Weka.



```
  a  b  c   <-- classified as
 33  5  2 |  a = COPD
  3 33  4 |  b = HC
  0  3  7 |  c = Infected
```

Figure 5- Refined - JRip Confusion Matrix

After the removal of the 10 instances of Asthma as it cannot be predicted at all, the new model shows a much higher prediction rate, with 81% accuracy as calculated from **Figure 5**, this is better as the machine isn't influenced by the little data from the deleted attribute.

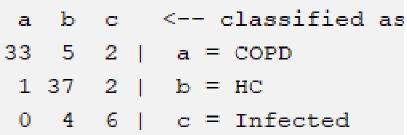


```
  a  b  c   <-- classified as
 32  5  3 |  a = COPD
  3 33  4 |  b = HC
  0  4  6 |  c = Infected
```

Figure 6- Attribute Selection- JRip Confusion Matrix

Supervised Attribution tool was used, this filters out which attributes participate most in the classifications and to determine if the prediction can be refined further. This took away the two attributes which was predicted using the plot matrix, gender, and Imaginary part.

However, this reduces the prediction model down to 78% correctly identified, this maybe influence of not enough data, other some attribute deleted might actually be useful, so a manual attribute selected is required for refinement.



```
  a  b  c   <-- classified as
 33  5  2 |  a = COPD
  1 37  2 |  b = HC
  0  4  6 |  c = Infected
```

Figure 7- Manual Attribute Selection- JRip Confusion Matrix

Manual Selection was done by observation of the Plot matrix and trial and error, and an optimal accuracy was reached of 84% correctly identified as shown in **Figure 7**, this was achieved by the removal of Saliva Permittivity- Imaginary Part. This proves that gender which was removed in the Attribute selection in the section above does play a slight role to the correctly identifying the classification.

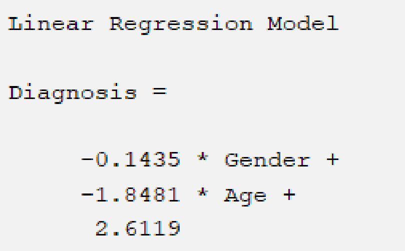
It was decided this was the most optimal method, standardising, and using the following attributes, Real Part, Gender, Age, Smoking and Diagnosis (class).

Part 2 – Ranking

Linear Regression

For Weka to create a linear regression equation all attributes need to be numerical, therefore a copy of the refined data was made, and the diagnosis feature was numerated accordingly, 1- COPD, 2 – HC and 3 – Infected.

A testing and training data set was created for the linear regression where 18 instances were used for the testing. The equation provided was then used on excel to see the success.



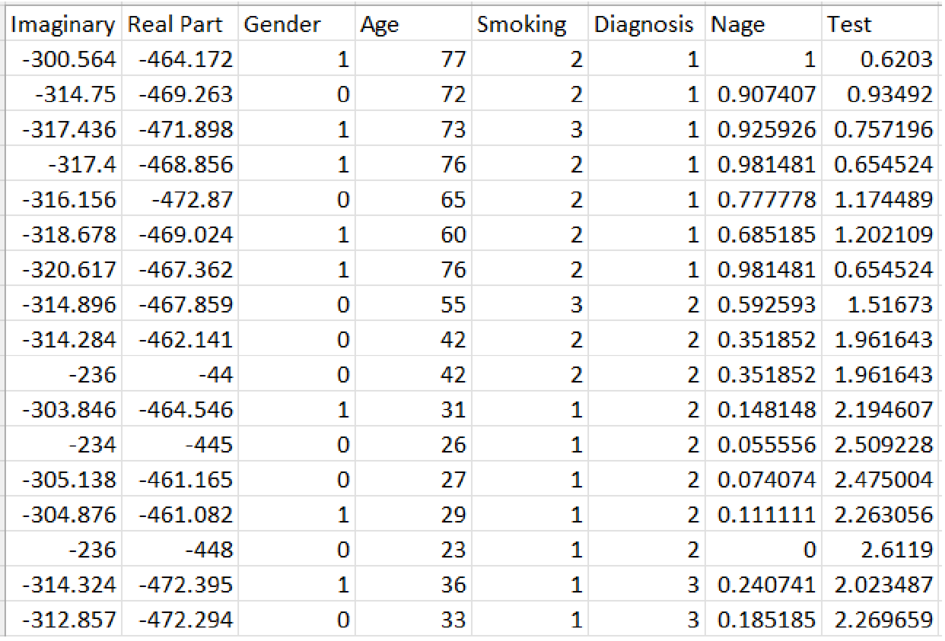
```
Linear Regression Model

Diagnosis =

     -0.1435 * Gender +
     -1.8481 * Age +
      2.6119
```

Figure 8- Linear Regression Model

The equation given shows the following equation, **Figure 8**. Linear regression equation suggests that only 2 of the attributes are appropriate for classifying the diagnosis, which are Gender and Age, these are ranked as well. This shows that the highest coefficient is the best predictor, which is the Age, this gives 60% error. Trying this on excel fails considerably. Linear regression is not suitable for this classification of data. Age is a good predictor and is expected to highly influence the diagnosis, but the some of the other attributes should be expected in the equation.



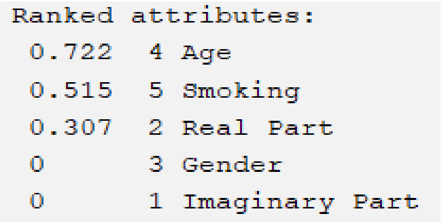
Imaginary	Real Part	Gender	Age	Smoking	Diagnosis	Nage	Test
-300.564	-464.172	1	77	2	1	1	0.6203
-314.75	-469.263	0	72	2	1	0.907407	0.93492
-317.436	-471.898	1	73	3	1	0.925926	0.757196
-317.4	-468.856	1	76	2	1	0.981481	0.654524
-316.156	-472.87	0	65	2	1	0.777778	1.174489
-318.678	-469.024	1	60	2	1	0.685185	1.202109
-320.617	-467.362	1	76	2	1	0.981481	0.654524
-314.896	-467.859	0	55	3	2	0.592593	1.51673
-314.284	-462.141	0	42	2	2	0.351852	1.961643
-236	-44	0	42	2	2	0.351852	1.961643
-303.846	-464.546	1	31	1	2	0.148148	2.194607
-234	-445	0	26	1	2	0.055556	2.509228
-305.138	-461.165	0	27	1	2	0.074074	2.475004
-304.876	-461.082	1	29	1	2	0.111111	2.263056
-236	-448	0	23	1	2	0	2.6119
-314.324	-472.395	1	36	1	3	0.240741	2.023487
-312.857	-472.294	0	33	1	3	0.185185	2.269659

Figure 9- Linear Regression Equation Applied on Test Data

The equation was entered into excel on the testing data, **Figure 9**, the Normalised age was required, and Gender coefficient is irrelevant in the equation so can be ignored, just to understand if the classification would work. The results predicted 4 wrong classifications, most were from diagnosis class 3, this is probably due to the lack of data for that group (Infected) so not a lot can be learned about it to predict well. However, this data shows about 78% success in this testing data, so age is a very good predictor.

InfoGainAttributeEval

This was the next chosen method for ranking the attributes for classification. This was performed on the original data, where class is still nominal



```
Ranked attributes:
 0.722   4 Age
 0.515   5 Smoking
 0.307   2 Real Part
 0       3 Gender
 0       1 Imaginary Part
```

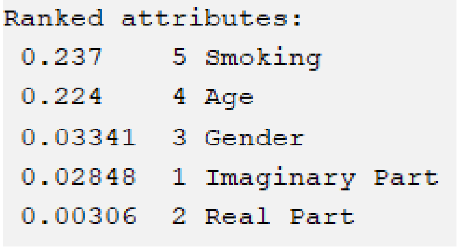
Figure 10- 'InfoGainAttributeEval' Ranker

This ranker, **Figure 10**, shows that 3 attributes play a big decent part in influencing the diagnosis class. Ranked top 3 are Age, Smoking and Saliva Permittivity- Real part. The coefficients show the relative influence compared to the rest of the attributes, with Gender and Imaginary Part having value of 0, meaning it is irrelevant.

This is quite close to what was predicted, age should be the leading contributor to the prediction model and smoking also influences close by. However, in the confusion matrix in Part 1, it was revealed that inclusion of gender actually helped predict better and increase the percentage of correct identifications, so this data in the ranked table may be a bit contradicting.

ReliefAttributeEval

A final ranker method was carried out, this was the **ReliefAttributeEval** evaluator, the class is nominal.



```
Ranked attributes:
 0.237     5 Smoking
 0.224     4 Age
 0.03341   3 Gender
 0.02848   1 Imaginary Part
 0.00306   2 Real Part
```

Figure 11 - 'ReliefAttributeEval' Ranker

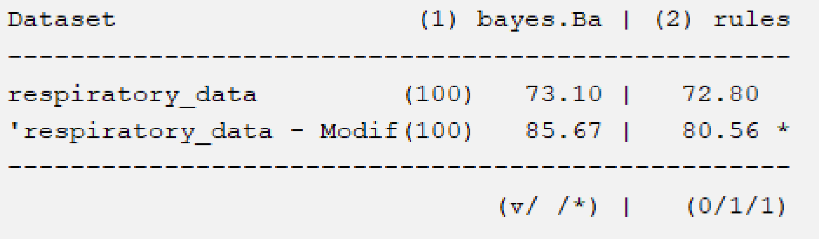
This ranker, **Figure 11**, shows the order of ranking of which elements influence the identification of the diagnosis the most. The coefficients determine the weightings of the attribute's contribution. This ranker shows Smoking is top closely followed by Age, not much difference between them, Gender and Imaginary Part seem to be close together as well with their coefficient values, and this model believes that the Real Part for the Saliva Permittivity is almost irrelevant.

Final decision was made overall the best attribute that determines the Diagnosis (class) is Age, as this has been ranked highly in all of the ranker tests, and the coefficient of the Age is always high relatively. This is as predicted, from Part 1. Another attribute to include is Smoking, as this was present in the final two ranking methods, with decent influence on the classification, this is a good predictor and was expected as it is known for respiratory problems in medical background. Based on the results only 3 attributes seem most important, Age, Smoking and Real Part.

Part 3 – Classification

A new data set was made, this is the modified data set which consists of the 3 attributes as well as the Diagnosis (class) which is believed to be the best ranker as given from the results above. This data was compared against the original data of 100 instances to determine if the modified data performs better and produces a higher certainty of predicted results.

The two chosen classifiers are as listed, the base is 'BayesNet', and the other is 'OneRule'. These were used in Experimenter mode in Weka, which performs the T-tests automatically. As this is a medical dataset, a good level of confidence would be from 75-85%, so this was the changed in the settings to give a good evaluation of the classifiers and making a good comparison.



```
Dataset                     (1) bayes.Ba | (2) rules
--------------------------------------------------
respiratory_data           (100)   73.10 |   72.80
'respiratory_data - Modif(100)   85.67 |   80.56 *
--------------------------------------------------
                               (v/ /*) |   (0/1/1)
```

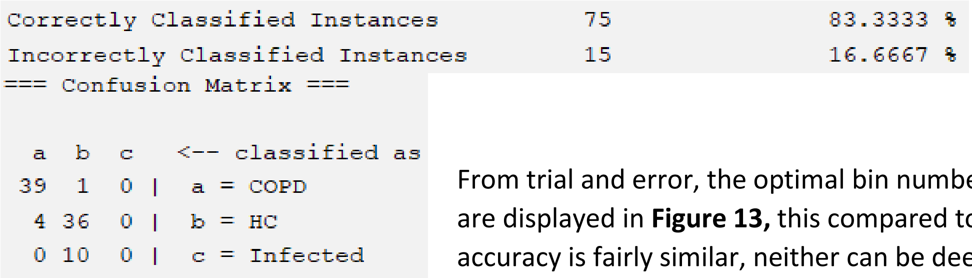
Figure 12- T-Test Table

Figure 12 shows the results of the T-test at 80% confidence level, which is adequate for medical data. Initially values show that 'BayesNet' classifier performs better that the 'OneRule', the percentage difference and also the star on the modified data set highlights this. This is understandable as, there are many complicated attributes so one rule may not be a good classification method. In the modified data the base classifier (BayesNet) performs significantly better, more that 5% than the 'OneRule'.

Another focal point would be that the performance of the modified data was substantially better than the original data, in both classification methods. The table shows that the modified data performs about 12-13% better than the original data with 80% confidence. This result proves that the methods researched in the previous parts and predictors highlighted were indeed good evaluators for the classification of the Diagnosis.

Part 4 – Discretisation

To achieve equal width binning the data must be discretized (unsupervised). The method involves splitting numeric values into a range of data from the minimum and maximum values. This was done in Weka to produce a modified data set. The number of bins was changed several times ranging from 1-10, 20- 100 (intervals of 10) to find the optimum bin number. This was verified by the best classifier tool used above which was 'BayesNet'.



```
Correctly Classified Instances          75               83.3333 %
Incorrectly Classified Instances        15               16.6667 %
=== Confusion Matrix ===

  a  b  c   <-- classified as
 39  1  0 |  a = COPD
  4 36  0 |  b = HC
  0 10  0 |  c = Infected
```

Figure 13- Equal-Width Binning

From trial and error, the optimal bin number was found to be 5. The results are displayed in **Figure 13,** this compared to the result above of 85.67% accuracy is fairly similar, neither can be deemed superior to the other, as the figures are very close. However the confusion matrix for the euqal-width binning shws that it cannot identify the Indected diagnsoed patients, this is probably due to the discretisation range which cannot adjust for the limited data it has on the Attribute.

Binning strategy chosen has a spectrum of merits. Initially it is a unsupervised method, therefore it is oblivious to the Diagnosis (class) attribute, it finds a pattern in data of continuous variables which is easy to infer and examine.

A small number of bins was used which improved the understanding of the attribute but gives the machine less ability to learn this was evident in mispredicting the infected column. Binning the data set ipproves the precison of the models that are made and predicted, by reducing the noise and no-linear data in data set, this strategy is very effiecient for identifying outliers, unacceptable or missing values in a data set. Hoever causes uneven distribution of data.

Another binning method is equal-frequency binning , this produces equal distribution but not equal intervals. However a setback is that binning leads to loss of information from the data, and bad boundaries so wrong clasiifications.

Part 5 – Clustering

Clustering would initially give an insight of the data, when nothing is known, just to evaluate how everything would fit together. However, after the classifications part it is used as a further analysing and evaluating method.

The original data consists of 3 Diagnosis- COPD, HC and Infected (Asthma was removed). Therefore, when simple unsupervised K-means clustering method was used, number of clusters were set to 3. This is not a classifying method but can be used to identify the success of other classifiers used in the report above.

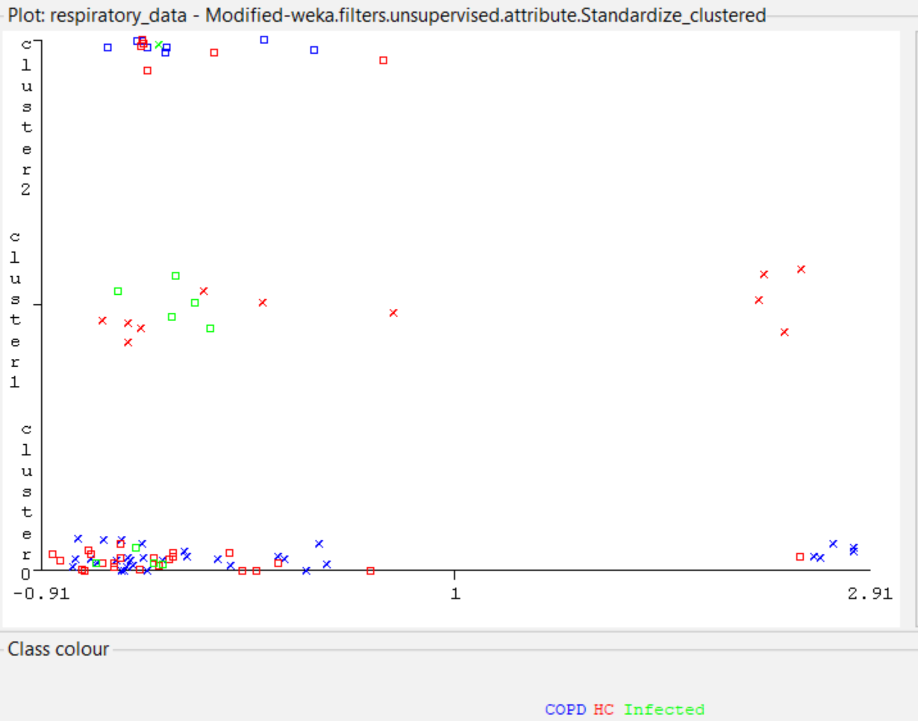

Plot: respiratory_data - Modified-weka.filters.unsupervised.attribute.Standardize_clustered

Figure 14- Clustering - Imaginary Part

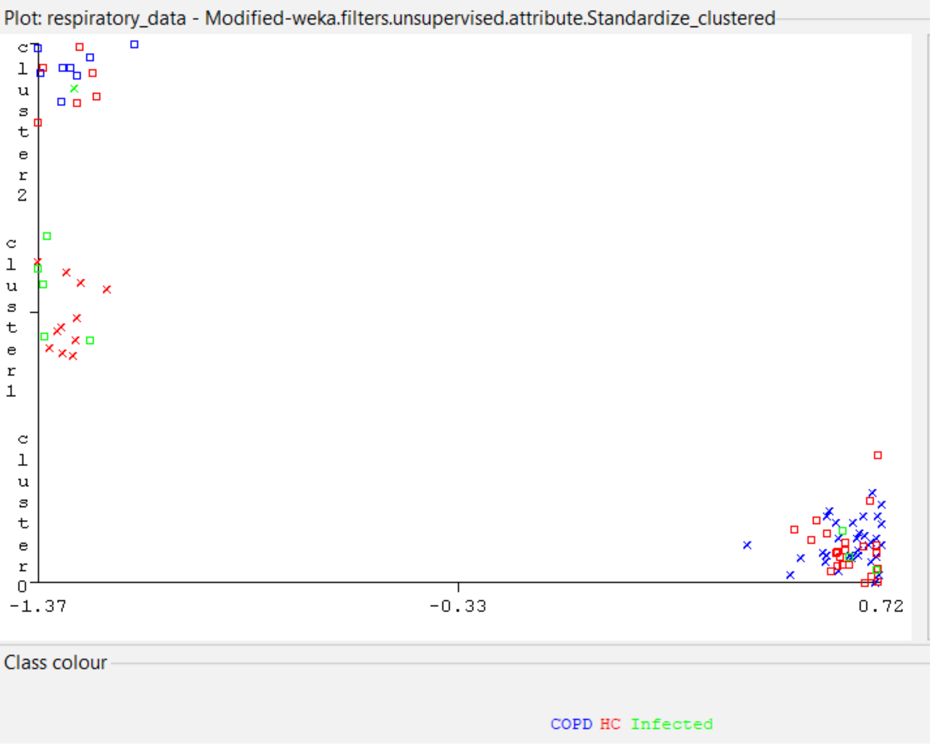

Plot: respiratory_data - Modified-weka.filters.unsupervised.attribute.Standardize_clustered

Figure 15- Clustering - Gender

Althoguh this is not a classification method, this cluttering tool is useful for understanding the performaces of the clasifiers that were used above.

Figures 14 and 15, show the clusterring of Saliva Permitivoty – Imaginary part and the Gender attribute cluseterring. These perforemd poorly in the classifiers as they didn't hold value to when classifying the Diagnosis. This graph shows that this is indeed representative of the results as there is no proper clustering of these to attributes, its all over the place. There is no evideces of actual clustering or corelation so this proves why these predictor would give very poor accury when applied for calssifications as no pattern can be seen for the machine to learn from and give valid results.

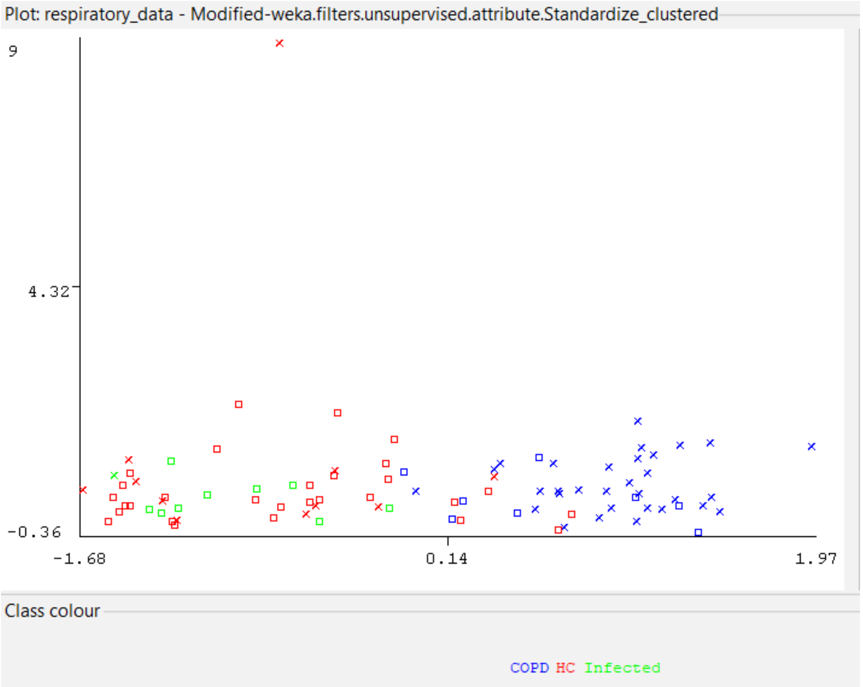


Figure 16- Clustering - Age v Real Part

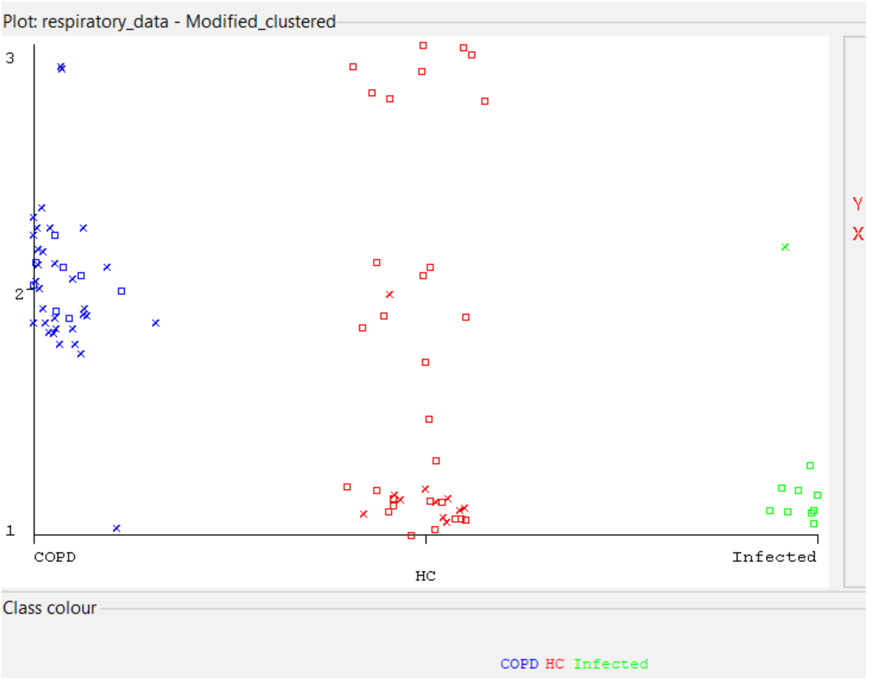


Figure 17- Clustering - Smoking v Diagnosis

On the contrary clustering proves the reasons for the well performing attributes such as Age and Real Part- Saliva Permitvity. The viusal graph in **Figures 16-17** plainly highlights much better clustering that seen before. This groups the data and reasures that the methods were corect above, when selecting and ranking the attributes.

Clear distinct clusterings shown in **Figure 16** are that with that younger patiets are helthier highlighted by red, and have better Saliva Permitivity. This is true as age casues healath related problems so the evideve can be validated .

Figure 17. highlights the smokers, wehre ex smokers who are older seem to have resiratory problems, and there is a contetrated cluster evident of that highlighted by by the blue

Clustering double checks the classifying methods, furhter proof is given in **Figure 16 and 17**, where the clustering of Infected Diagnosis (green) falls insided or close to the HC cluster, this shows why the confusion matrix is falsely identifying Infected as HC and predicting it wrong.