ICS2203/ARI2203

NLP Methods and Tools Assignment

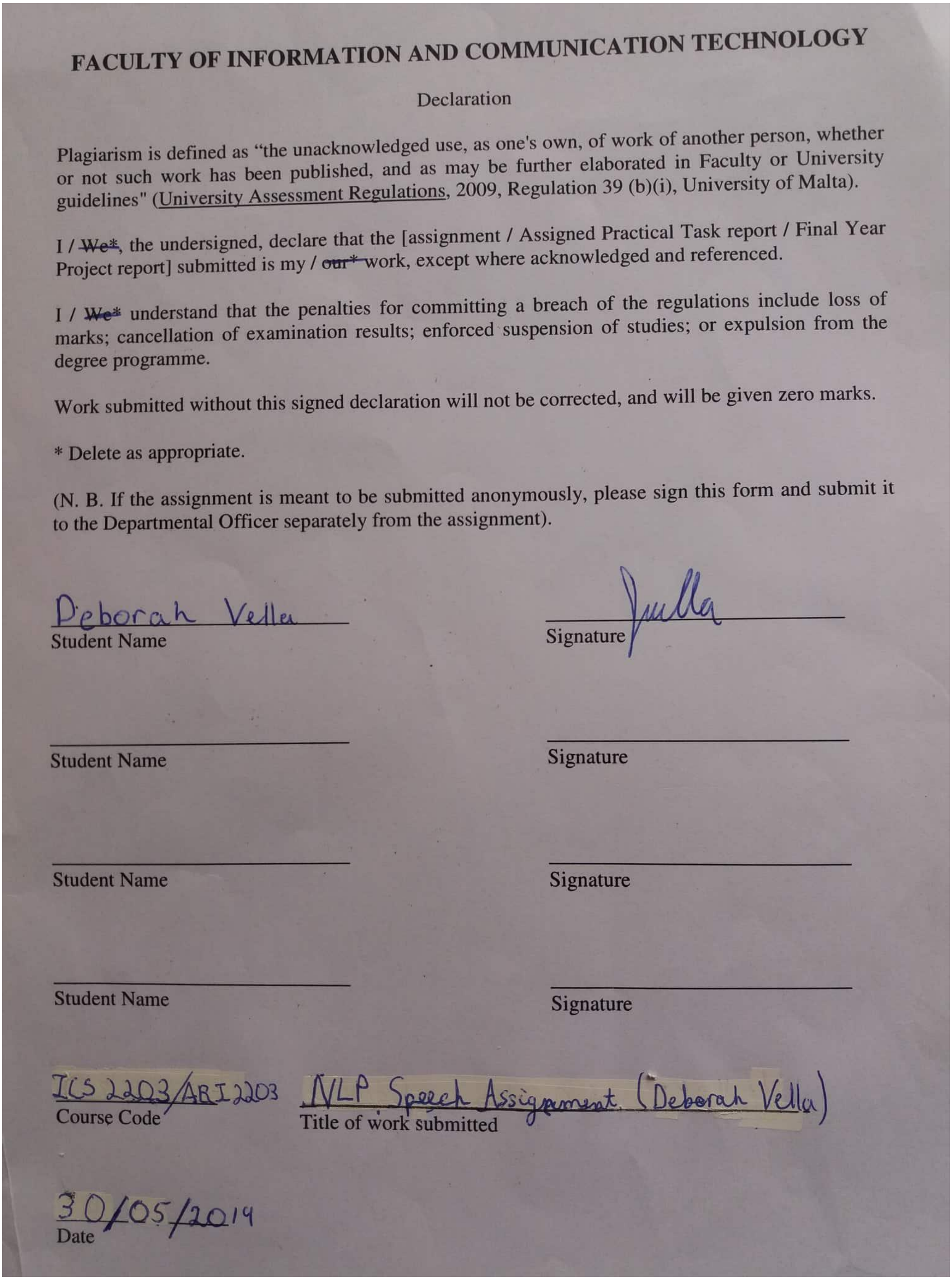
Speech Phoneme Analysis and Classification

Name: Deborah Vella

Course: Artificial Intelligence

Table of Contents

Plagiarism Declaration form



FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Declaration

Plagiarism is defined as "the unacknowledged use, as one's own, of work of another person, whether or not such work has been published, and as may be further elaborated in Faculty or University guidelines" (University Assessment Regulations, 2009, Regulation 39 (b)(i), University of Malta).

I / ~~We*~~, the undersigned, declare that the [assignment / Assigned Practical Task report / Final Year Project report] submitted is my / ~~our*~~ work, except where acknowledged and referenced.

I / ~~We*~~ understand that the penalties for committing a breach of the regulations include loss of marks; cancellation of examination results; enforced suspension of studies; or expulsion from the degree programme.

Work submitted without this signed declaration will not be corrected, and will be given zero marks.

* Delete as appropriate.

(N. B. If the assignment is meant to be submitted anonymously, please sign this form and submit it to the Departmental Officer separately from the assignment).

Deborah Vella
Student Name

Signature

Student Name

Signature

Student Name

Signature

Student Name

Signature

ICS 2203/ARI 2203
Course Code

NLP Speech Assignment. (Deborah Vella)
Title of work submitted

30/05/2019
Date

Section 1

CSV File and Program Overview

A. CSV File

The csv file is sorted by gender. The first 75 rows are females' data, and the last 75 rows are the males' data. This way, the program can get the gender specific data (e.g males only) in an easy way. The accents chosen are brm_001, lvp_001, ilo_001, shl_001 and eyk_001. The extraction was done on the following phonemes: IH, EH and AA.

B. Program

The user is firstly prompted to choose on what type of data set (e.g females part only) the classification is going to happen. Then s/he is asked to input the percentage of the test size. Afterwards, the user has to input how many tests are going to be executed on the split data set. Now, the code starts looping through the number of test runs wanted and with each iteration the value of K is needed to be entered. The algorithm is executed on two different distance metrics each time. The outputs are a confusion matrix, classification report with f1-scores, and a plot of f1 against f2 against f3 displaying the training set and the results of the test set.

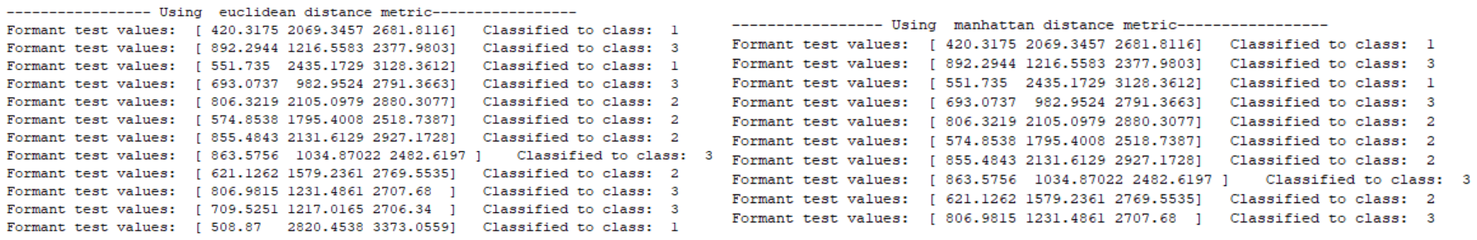
Section 2

Distance metrics used

For this assignment I decided to use three different distance metrics which are the Euclidean distance metric, the Manhattan distance metric and Chebyshev distance metric.

Manhattan Distance: sum(|x-y|) **Euclidean Distance:** sqrt(sum((x-y) ^ 2)) **Chebyshev Distance:** `max(|x-y|)`

The program was coded in such a way that, whenever the K-NN algorithm is to be executed, it loops twice using a different metric in each iteration. Figures 1 show parts of the output of the algorithm (with both distances).



```
----------------- Using  euclidean distance metric-----------------
Formant test values:  [ 420.3175 2069.3457 2681.8116]   Classified to class:  1
Formant test values:  [ 892.2944 1216.5583 2377.9803]   Classified to class:  3
Formant test values:  [ 551.735  2435.1729 3128.3612]   Classified to class:  1
Formant test values:  [ 693.0737  982.9524 2791.3663]   Classified to class:  3
Formant test values:  [ 806.3219 2105.0979 2880.3077]   Classified to class:  2
Formant test values:  [ 574.8538 1795.4008 2518.7387]   Classified to class:  2
Formant test values:  [ 855.4843 2131.6129 2927.1728]   Classified to class:  2
Formant test values:  [ 863.5756  1034.87022 2482.6197 ]    Classified to class:  3
Formant test values:  [ 621.1262 1579.2361 2769.5535]   Classified to class:  2
Formant test values:  [ 806.9815 1231.4861 2707.68  ]   Classified to class:  3
Formant test values:  [ 709.5251 1217.0165 2706.34  ]   Classified to class:  3
Formant test values:  [ 508.87   2820.4538 3373.0559]   Classified to class:  1
```

```
----------------- Using  manhattan distance metric-----------------
Formant test values:  [ 420.3175 2069.3457 2681.8116]   Classified to class:  1
Formant test values:  [ 892.2944 1216.5583 2377.9803]   Classified to class:  3
Formant test values:  [ 551.735  2435.1729 3128.3612]   Classified to class:  1
Formant test values:  [ 693.0737  982.9524 2791.3663]   Classified to class:  3
Formant test values:  [ 806.3219 2105.0979 2880.3077]   Classified to class:  2
Formant test values:  [ 574.8538 1795.4008 2518.7387]   Classified to class:  2
Formant test values:  [ 855.4843 2131.6129 2927.1728]   Classified to class:  2
Formant test values:  [ 863.5756  1034.87022 2482.6197 ]    Classified to class:  3
Formant test values:  [ 621.1262 1579.2361 2769.5535]   Classified to class:  2
Formant test values:  [ 806.9815 1231.4861 2707.68  ]   Classified to class:  3
```

Figure 1 Part of output using Euclidean distance (on the left). Part of output using Manhattan distance (on the right)

When conducting tests and analysis, both of the distance metrics were taken into consideration and when possible, they were compared with each other. This type of analysis is shown later on in this report.

Section 3

How does performance change with different values of K?

To check how the K value affects the algorithm's result, it was decided to **test the classification 7 times**. With each test, the same training and testing sets were used with the same splitting percentage. This way I could compare how the preciseness changes, given the same sets as input, but a different K value each time. With every test, the f1-score of the macro average was added to the tables below. The f1-score should be able to give an indication of how accurate the results are. These tests were applied on the whole dataset (i.e. both males and females).

When using Euclidean distance metric

K-Value	1	2	3	4	5	6	7
F1-score	0.84	0.9	0.9	0.92	0.85	0.92	0.9

When using Manhattan distance metric

K-Value	1	2	3	4	5	6	7
F1-score	0.87	0.9	0.92	0.92	0.88	0.9	0.9

When using Chebyshev distance metric

K-Value	1	2	3	4	5	6	7
F1-score	0.81	0.9	0.9	0.92	0.85	0.9	0.85

The highest f1-score in all tables is 0.92. When using Euclidian distance, it looks like when using K-value 4 or K-value 6, the results are the most precise. On the other hand, when using the Manhattan distance, the K-values 3 and 4 give back the best answer. In contrast, Chebyshev only has one highest f1-score that is when K=4. It can also be concluded that, having K=4 as one of the best K values is in common with all three metrics. This may imply that having K being equal to 4 is the best K to be used whenever using all the metrics above.

I also created tables containing the f1-score, of a particular class given a particular K-value. These tables are shown below.

When using Euclidean distance metric

K-value	1	2	3	4	5	6	7
Class 1 (IH) f1-score	0.87	0.9	0.87	0.9	0.79	0.9	0.87
Class 2(EH) f1 score	0.73	0.83	0.83	0.87	0.77	0.87	0.83
Class 3(AA) f1 score	0.92	0.95	1	1	1	1	1

When using Manhattan distance metric

K-value	1	2	3	4	5	6	7
Class 1 (IH) f1-score	0.87	0.9	0.9	0.9	0.83	0.87	0.87
Class 2(EH) f1 score	0.78	0.83	0.87	0.87	0.8	0.83	0.83
Class 3(AA) f1 score	0.96	0.95	1	1	1	1	1

When using Chebyshev

K-value	1	2	3	4	5	6	7
Class 1 (IH) f1-score	0.83	0.88	0.87	0.9	0.79	0.87	0.79
Class 2(EH) f1 score	0.70	0.87	0.83	0.87	0.77	0.83	0.77
Class 3(AA) f1 score	0.92	0.95	1	1	1	1	1

When comparing the above tables with each other, it can be noted that the f1-scores do not vary much from one metric to another. It is also shown that, class 3 returns the most highly accurate results out of all the classes. In fact, its f1-score reaches 1 from K-value 3 and above, in all metrics. On the other hand, class 2 returns the least accurate results, as in all cases, with each K-value, the f1-score is always less than the f1-scores of the other two classes. Once again, the k-value 4 gives the same f1-scores, hence, it could be considered as being the best K, while K=1 could be considered as the worst K.

Section 4

How performance changes when using data of a single gender, or data of both genders.

To analyse this part, the program was **run 5 times**, applying classifications on all data sets (both genders & genders on their own) in every test. Five different test set percentages were chosen for all five experiments. The results analysed in this section are the plots and the classification reports of the Euclidean distance only.

A. Plots

The circles on the plots represent the training data, while the stars represent the test data. The colours show what class the points are a part of or predicted to be a part of. Figures 2-6 below have three different plots for the different genders. Only the Euclidean distance plots are shown below. (One might need to zoom in on the plots to be able to distinguish between circles and stars)

Case 1: test set =0.25%

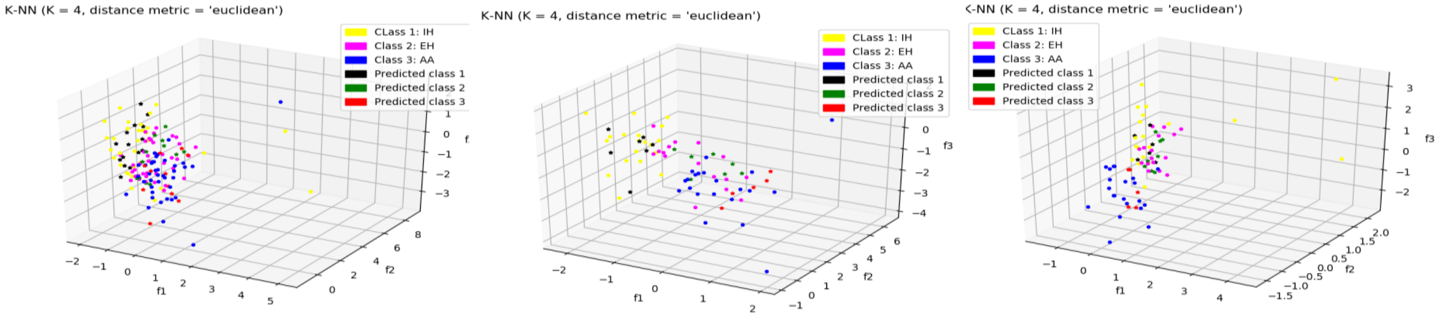


Figure 2 1st plot: both genders, 2nd plot: Females only, 3rd plot: males only

Case 2: test set= 0.5%

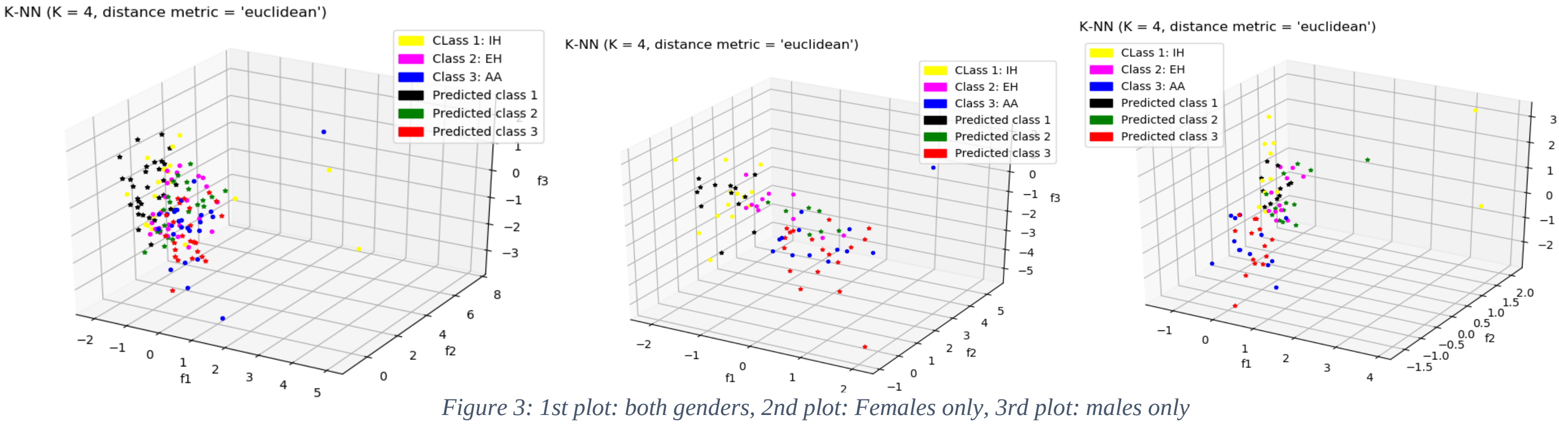


Figure 3: 1st plot: both genders, 2nd plot: Females only, 3rd plot: males only

Case 3: test set=0.15%

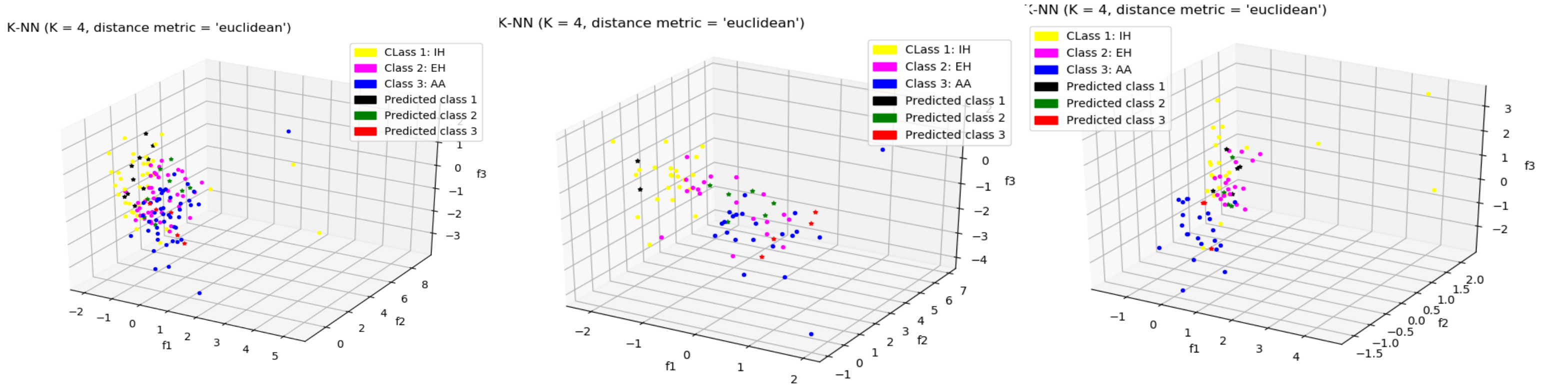


Figure 4: 1st plot: both genders, 2nd plot: Females only, 3rd plot: males only

Case 4: test set = 0.35%

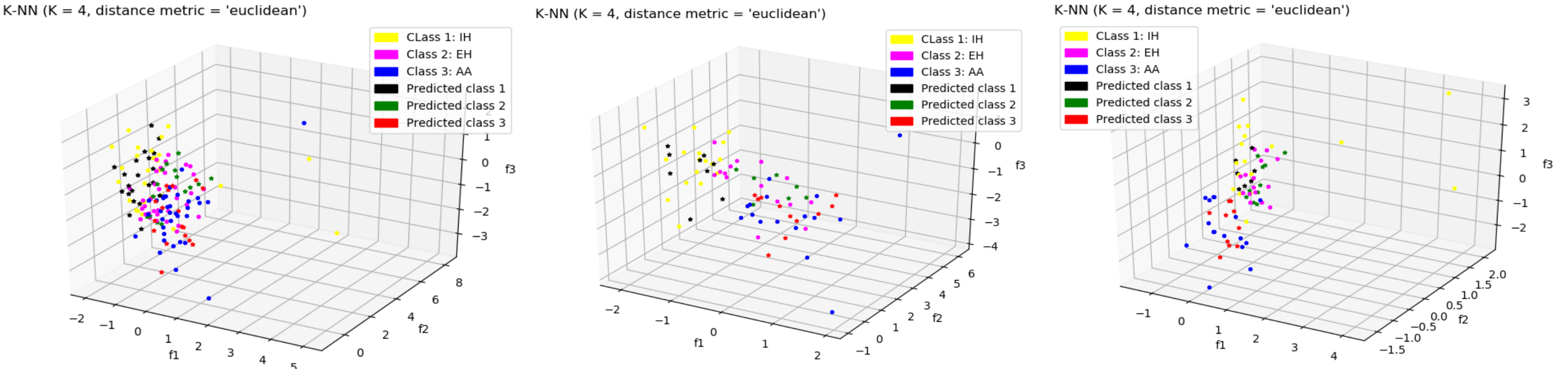


Figure 5: 1st plot: both genders, 2nd plot: Females only, 3rd plot: males only

Case 5: test set=0.4%

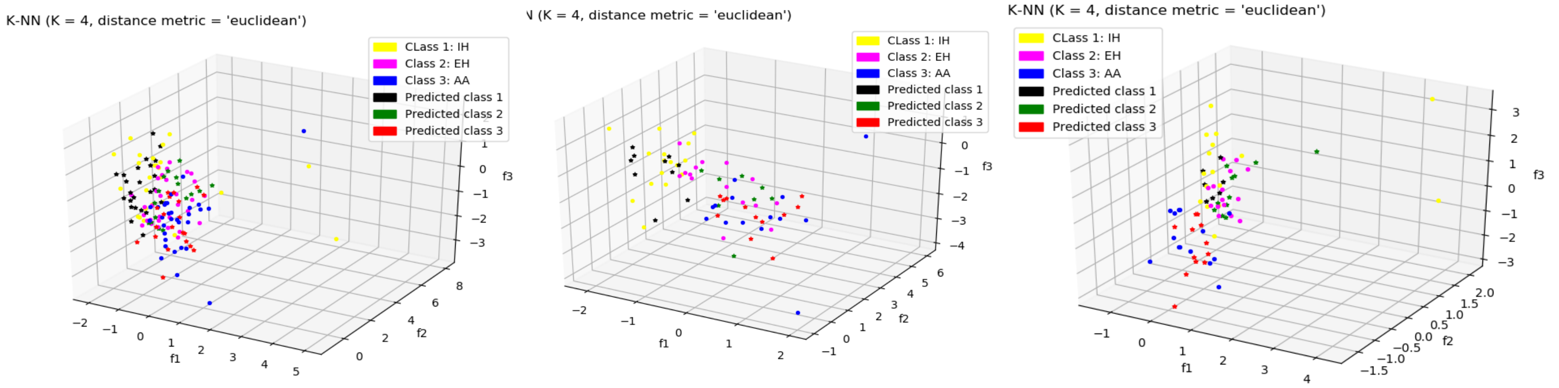


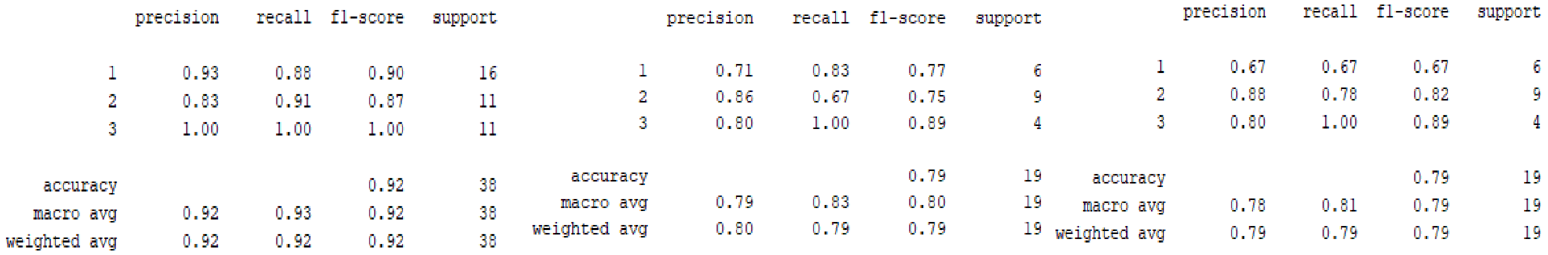
Figure 6: 1st plot: both genders, 2nd plot: Females only, 3rd plot: males only

From the plots displayed from figure 2 till figure 6, it can be noted that the female data on its own is more scattered than the others. One main reason is that, the females' formant 1 has a higher range of values than the formant 1s in the other data sets. Another noticeable feature is that, in each plot, there are classes which are overlapping each other which may result in an inaccurate class prediction. The male's and the genders together's points are more compact with each other. The single genders on their own have fewer training points which may affect the accuracy of the outcome more. The bigger the training set the better the performance should be.

B. Classification report

Below are five different classification reports that where outputted when testing all gender possibilities. These are results shown when using Euclidean distance, with 0.25% test data and K is equal to 4. I chose to test this on K being 4 because, in the analysis of section 3, it was concluded that K=4 is the best K for all distance metrics.

Case 1: test set=0.25%



```
              precision    recall  f1-score   support

           1       0.93      0.88      0.90        16
           2       0.83      0.91      0.87        11
           3       1.00      1.00      1.00        11

    accuracy                           0.92        38
   macro avg       0.92      0.93      0.92        38
weighted avg       0.92      0.92      0.92        38
```

```
              precision    recall  f1-score   support

           1       0.71      0.83      0.77         6
           2       0.86      0.67      0.75         9
           3       0.80      1.00      0.89         4

    accuracy                           0.79        19
   macro avg       0.79      0.83      0.80        19
weighted avg       0.80      0.79      0.79        19
```

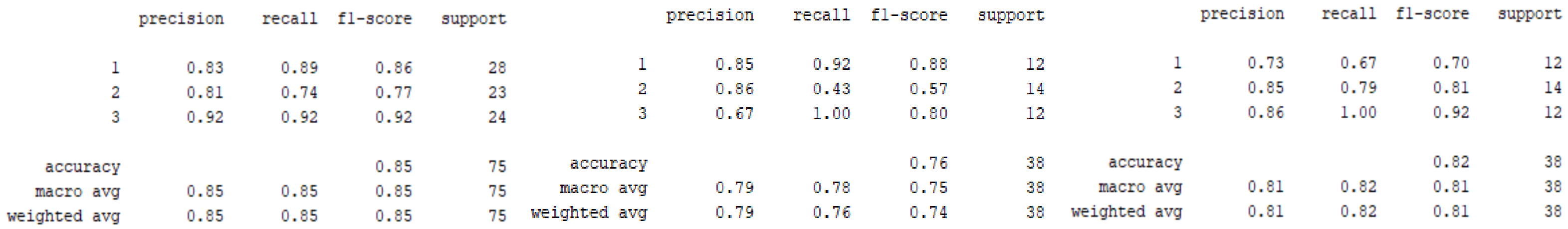
```
              precision    recall  f1-score   support

           1       0.67      0.67      0.67         6
           2       0.88      0.78      0.82         9
           3       0.80      1.00      0.89         4

    accuracy                           0.79        19
   macro avg       0.78      0.81      0.79        19
weighted avg       0.79      0.79      0.79        19
```

Figure 7: 1st report: Both Genders, 2nd report: Female only, 3rd report: Male only

Case 2: test set=0.5%



```
              precision    recall  f1-score   support

           1       0.83      0.89      0.86        28
           2       0.81      0.74      0.77        23
           3       0.92      0.92      0.92        24

    accuracy                           0.85        75
   macro avg       0.85      0.85      0.85        75
weighted avg       0.85      0.85      0.85        75
```

```
              precision    recall  f1-score   support

           1       0.85      0.92      0.88        12
           2       0.86      0.43      0.57        14
           3       0.67      1.00      0.80        12

    accuracy                           0.76        38
   macro avg       0.79      0.78      0.75        38
weighted avg       0.79      0.76      0.74        38
```

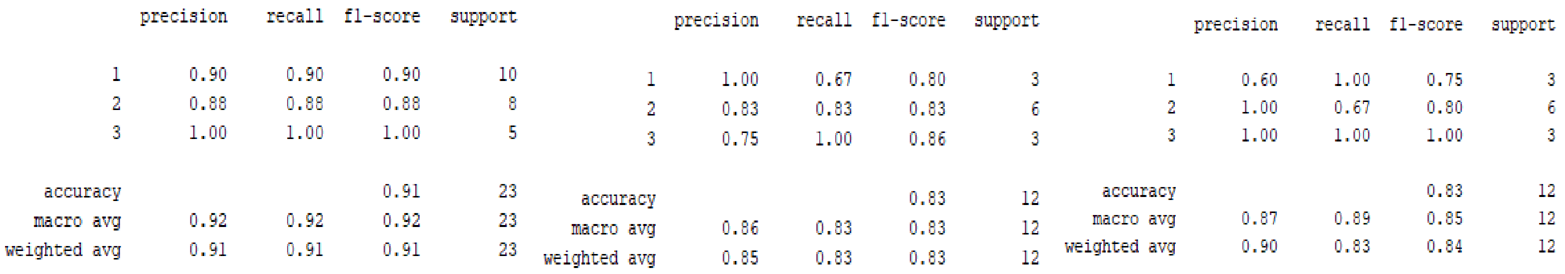
```
              precision    recall  f1-score   support

           1       0.73      0.67      0.70        12
           2       0.85      0.79      0.81        14
           3       0.86      1.00      0.92        12

    accuracy                           0.82        38
   macro avg       0.81      0.82      0.81        38
weighted avg       0.81      0.82      0.81        38
```

Figure 8: 1st report: Both Genders, 2nd report: Female only, 3rd report: Male only

Case 3: test set=0.15%



```
              precision    recall  f1-score   support

           1       0.90      0.90      0.90        10
           2       0.88      0.88      0.88         8
           3       1.00      1.00      1.00         5

    accuracy                           0.91        23
   macro avg       0.92      0.92      0.92        23
weighted avg       0.91      0.91      0.91        23
```

```
              precision    recall  f1-score   support

           1       1.00      0.67      0.80         3
           2       0.83      0.83      0.83         6
           3       0.75      1.00      0.86         3

    accuracy                           0.83        12
   macro avg       0.86      0.83      0.83        12
weighted avg       0.85      0.83      0.83        12
```

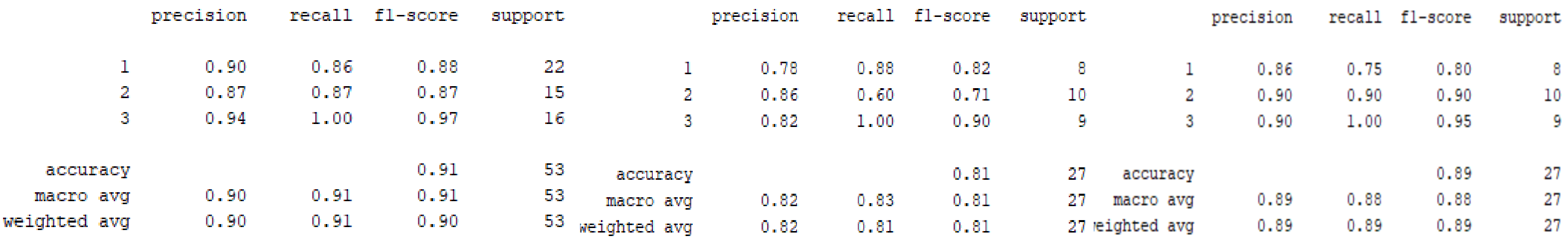
```
              precision    recall  f1-score   support

           1       0.60      1.00      0.75         3
           2       1.00      0.67      0.80         6
           3       1.00      1.00      1.00         3

    accuracy                           0.83        12
   macro avg       0.87      0.89      0.85        12
weighted avg       0.90      0.83      0.84        12
```

Figure 9: 1st report: Both Genders, 2nd report: Female only, 3rd report: Male only

Case 4: test set = 0.35%



```
              precision    recall  f1-score   support

           1       0.90      0.86      0.88        22
           2       0.87      0.87      0.87        15
           3       0.94      1.00      0.97        16

    accuracy                           0.91        53
   macro avg       0.90      0.91      0.91        53
weighted avg       0.90      0.91      0.90        53
```

```
              precision    recall  f1-score   support

           1       0.78      0.88      0.82         8
           2       0.86      0.60      0.71        10
           3       0.82      1.00      0.90         9

    accuracy                           0.81        27
   macro avg       0.82      0.83      0.81        27
weighted avg       0.82      0.81      0.81        27
```

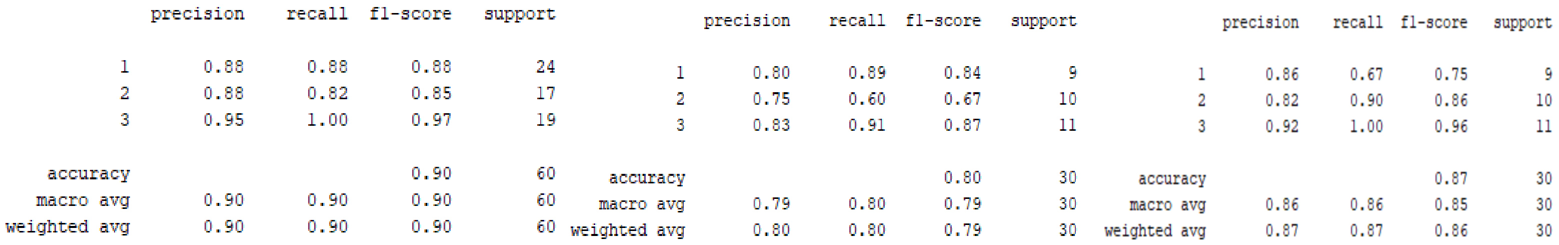
```
              precision    recall  f1-score   support

           1       0.86      0.75      0.80         8
           2       0.90      0.90      0.90        10
           3       0.90      1.00      0.95         9

    accuracy                           0.89        27
   macro avg       0.89      0.88      0.88        27
weighted avg       0.89      0.89      0.89        27
```

Figure 10: 1st report: Both Genders, 2nd report: Female only, 3rd report: Male only

Case 5: test set =0.4%



```
              precision    recall  f1-score   support

           1       0.88      0.88      0.88        24
           2       0.88      0.82      0.85        17
           3       0.95      1.00      0.97        19

    accuracy                           0.90        60
   macro avg       0.90      0.90      0.90        60
weighted avg       0.90      0.90      0.90        60
```

```
              precision    recall  f1-score   support

           1       0.80      0.89      0.84         9
           2       0.75      0.60      0.67        10
           3       0.83      0.91      0.87        11

    accuracy                           0.80        30
   macro avg       0.79      0.80      0.79        30
weighted avg       0.80      0.80      0.79        30
```

```
              precision    recall  f1-score   support

           1       0.86      0.67      0.75         9
           2       0.82      0.90      0.86        10
           3       0.92      1.00      0.96        11

    accuracy                           0.87        30
   macro avg       0.86      0.86      0.85        30
weighted avg       0.87      0.87      0.86        30
```

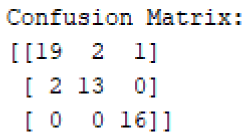
Figure 11: 1st report: Both Genders, 2nd report: Female only, 3rd report: Male only

When comparing the macro averages, it can be noted that when applying K-NN on both genders together, the f1-scores are higher. And, when applying the classification algorithm on one gender only, the f1-scores decrease in value. This holds for all the different cases shown above. When analysing the females only and the males only, one may realize that the f1-scores of the males are on the majority higher than those of females. This shows that, the male data set is inclined to give out better performance when using single gender only.

Section 5

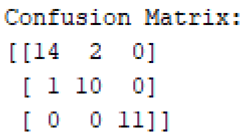
<u>Analysing the phonemes that produce the most confusion (based off of confusion matrices)</u>

The confusion matrices in figures 12-15 below are the output from the program. These were calculated on the data set containing both genders and while having K=4, because as discussed previously these give the best results. Therefore, I will be analysing these confusion matrices that were generated from the best data possible.



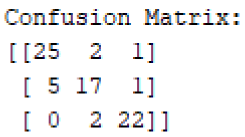
```
Confusion Matrix:
[[19  2  1]
 [ 2 13  0]
 [ 0  0 16]]
```

Figure 132: test set=0.35%



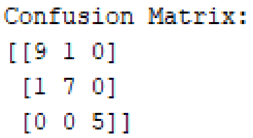
```
Confusion Matrix:
[[14  2  0]
 [ 1 10  0]
 [ 0  0 11]]
```

Figure 13: test set=0.25%



```
Confusion Matrix:
[[25  2  1]
 [ 5 17  1]
 [ 0  2 22]]
```

Figure 14: test set=0.5%



```
Confusion Matrix:
[[9 1 0]
 [1 7 0]
 [0 0 5]]
```

Figure 125: test set=0.15%

The below tables represent the same matrices above but with the phonemes, and totals listed on the sides to help visualize the information more.

N=53	Predicted: IH	Predicted: EH	Predicted: AA	
Actual: IH	19	2	1	22
Actual: EH	2	13	0	15
Actual: AA	0	0	16	16
	21	15	17	

Test set =0.35%

N=38	Predicted: IH	Predicted: EH	Predicted: AA	
Actual: IH	14	2	0	16
Actual: EH	1	10	0	11
Actual: AA	0	0	11	11
	15	12	11	

Test set = 0.25%

N=75	Predicted: IH	Predicted: EH	Predicted: AA	
Actual: IH	25	2	1	28
Actual: EH	5	17	1	23
Actual: AA	0	2	22	24
	30	21	24	

Test set=0.5%

N=23	Predicted: IH	Predicted: EH	Predicted: AA	
Actual: IH	9	1	0	10
Actual: EH	1	7	0	8
Actual: AA	0	0	5	5
	10	8	5	

Test set = 0.15%

These confusion matrices show that, the differences between the totals of the actual phonemes and the totals of the same predicted phonemes, are not huge. In fact, the highest difference from all cases is a difference of 2. This may imply that, overall the data is classified to the correct phoneme and errors are minimal occurrences.

Another feature that pops out is that, the phoneme ‘AA’ causes the least errors. This has already been noted when analysing the f1-scores of each class when given a particular k value.

In contrast, the other two phonemes cause most of the errors. They approximately caused the same number of errors each, so there is no significant difference between them. Hence, neither one of them should be considered as the phoneme that gives out the worst accuracy.

Conclusion

The program was tested 12 times to analyse how given particular inputs the performance changes:

Tested 7 times: To check how the K-values from 1 to 7 change the performance

Tested 5 more times: To check how different gender specific sets affect the performance.

All in all, the K-NN algorithm worked fine and gave out very good results and the errors were minimal.