**Assignment 1 – Machine Learning (Blended Del)**

**Student No.: 23211267**

**Name: Conor Heffron**

**Module Code: COMP47460**

**1.1.**

**Examine the dataset carefully in the Weka Explorer. You should investigate some of the [5] filters in Weka which allow you to normalise and/or clean the dataset as appropriate. Describe the data preparation/cleaning steps you took if any (e.g., Your description may look like ”I did min-max normalisation of feature X using the minimum on the feature values in training examples and maximum of feature values over all labelled examples. I manually removed feature Y because …”).**

* **Note: Please see uploaded python scripts for incremental output csv files and console output outlined below. In addition, I have taken one of the earlier output files and updated with excel functions to ensure code matches manually calculated results.**
* I loaded the dataset into Weka and exported to csv format.
* There were 2115 records or instances loaded from the \*.arff file with 22 attributes or features.
* Straight away I could see ‘severe\_decelerations’ was a very limited feature with only 2 distinct values where a non-zero value only occurs 7 times as noted by the python scripts output below (z sum = 7.0 after normalising to range [0, 1]) so I removed this feature from further analysis in Weka and python script runs.
* I think the histogram features would normally be dropped but some of them appear to be quite varied so this makes for a good data sample.

A screenshot of a computer

Description automatically generated



‘mean\_value\_of\_long\_term\_variability’ has the most distinct values at 249 which was something to note.

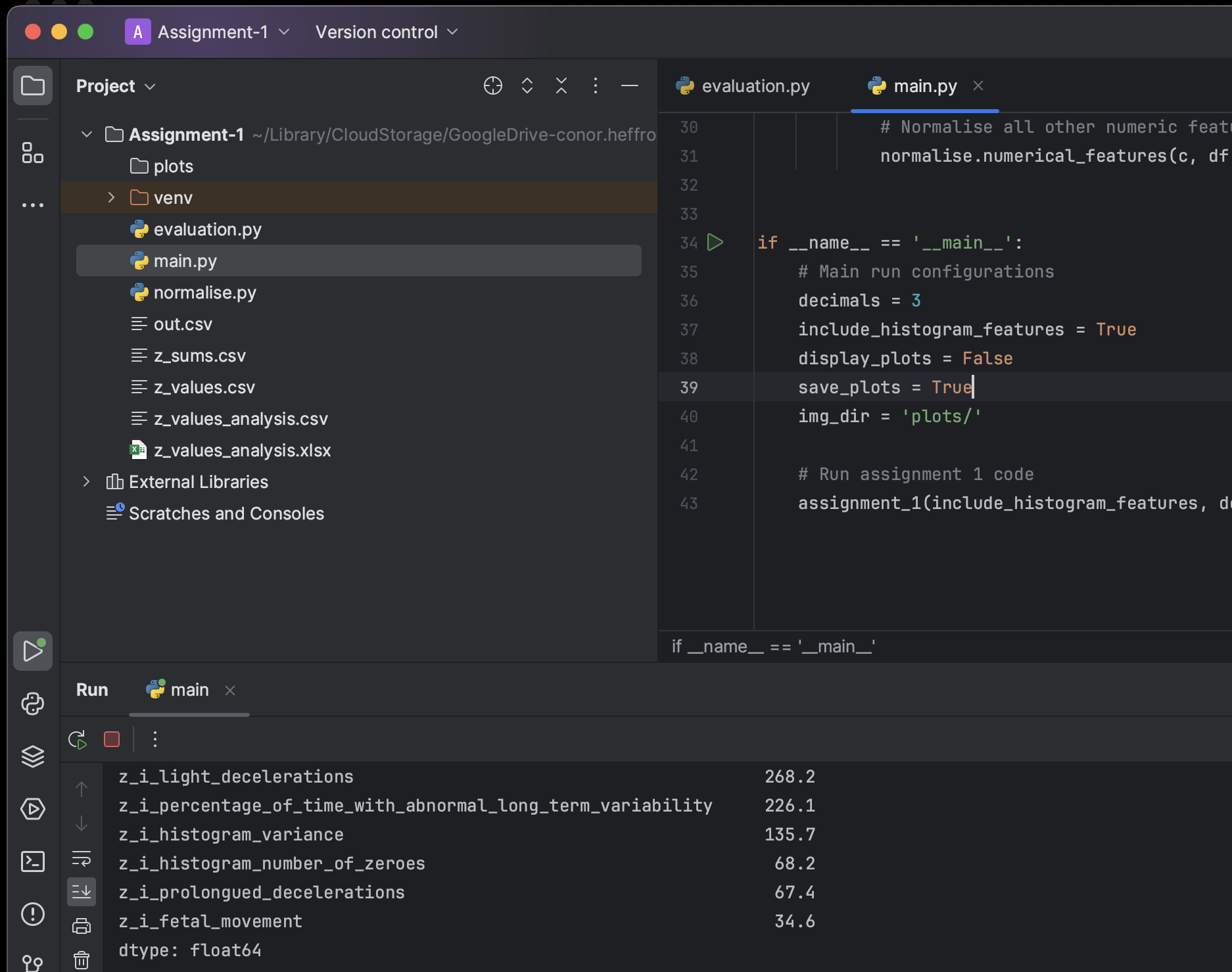


A screenshot of a computer

Description automatically generated

Then I wrote Python code to clean and normalise the data especially the numerical features.

I then calculated the z value for each and ordered by z sum as the final program output to see the key features from this process.





**Note:** From top left above marked with red:

* plots output directory
* python scripts,
* incremental output files
* manual excel analysis for comparison
* Main configurations for python run script

**Main Program Console Out:**

------------------------------------------------

Min of baseline\_value is: 106

Max of baseline\_value is: 160

Mean of baseline\_value is: 133.30070921985816

Median of baseline\_value is: 133.0

Standard Deviation of baseline\_value is: 9.837149600812499

Zi sum value baseline\_value is: 1074.6

------------------------------------------------

Min of accelerations is: 0.0

Max of accelerations is: 0.019

Mean of accelerations is: 0.0031787234042553194

Median of accelerations is: 0.002

Standard Deviation of accelerations is: 0.003863961682425237

Zi sum value accelerations is: 367.79999999999995

------------------------------------------------

Min of fetal\_movement is: 0.0

Max of fetal\_movement is: 0.481

Mean of fetal\_movement is: 0.009526713947990545

Median of fetal\_movement is: 0.0

Standard Deviation of fetal\_movement is: 0.046782681829283676

Zi sum value fetal\_movement is: 34.599999999999994

------------------------------------------------

Min of uterine\_contractions is: 0.0

Max of uterine\_contractions is: 0.015

Mean of uterine\_contractions is: 0.004373049645390071

Median of uterine\_contractions is: 0.004

Standard Deviation of uterine\_contractions is: 0.002947399462280164

Zi sum value uterine\_contractions is: 616.8

------------------------------------------------

Min of light\_decelerations is: 0.0

Max of light\_decelerations is: 0.015

Mean of light\_decelerations is: 0.001888888888888889

Median of light\_decelerations is: 0.0

Standard Deviation of light\_decelerations is: 0.0029635828830450834

Zi sum value light\_decelerations is: 268.20000000000005

------------------------------------------------

~~Min of severe\_decelerations is: 0.0~~

~~Max of severe\_decelerations is: 0.001~~

~~Mean of severe\_decelerations is: 3.309692671394799e-06~~

~~Median of severe\_decelerations is: 0.0~~

~~Standard Deviation of severe\_decelerations is: 5.744822913623554e-05~~

~~Zi sum value severe\_decelerations is: 7.0~~

------------------------------------------------

Min of prolongued\_decelerations is: 0.0

Max of prolongued\_decelerations is: 0.005

Mean of prolongued\_decelerations is: 0.00015933806146572102

Median of prolongued\_decelerations is: 0.0

Standard Deviation of prolongued\_decelerations is: 0.0005913692853366418

Zi sum value prolongued\_decelerations is: 67.4

------------------------------------------------

Min of abnormal\_short\_term\_variability is: 12

Max of abnormal\_short\_term\_variability is: 87

Mean of abnormal\_short\_term\_variability is: 46.95035460992908

Median of abnormal\_short\_term\_variability is: 49.0

Standard Deviation of abnormal\_short\_term\_variability is: 17.162584263361758

Zi sum value abnormal\_short\_term\_variability is: 983.9

------------------------------------------------

Min of mean\_value\_of\_short\_term\_variability is: 0.2

Max of mean\_value\_of\_short\_term\_variability is: 7.0

Mean of mean\_value\_of\_short\_term\_variability is: 1.3339952718676122

Median of mean\_value\_of\_short\_term\_variability is: 1.2

Standard Deviation of mean\_value\_of\_short\_term\_variability is: 0.8837077165359826

Zi sum value mean\_value\_of\_short\_term\_variability is: 345.20000000000005

------------------------------------------------

Min of percentage\_of\_time\_with\_abnormal\_long\_term\_variability is: 0

Max of percentage\_of\_time\_with\_abnormal\_long\_term\_variability is: 91

Mean of percentage\_of\_time\_with\_abnormal\_long\_term\_variability is: 9.823167848699764

Median of percentage\_of\_time\_with\_abnormal\_long\_term\_variability is: 0.0

Standard Deviation of percentage\_of\_time\_with\_abnormal\_long\_term\_variability is: 18.391154458995572

Zi sum value percentage\_of\_time\_with\_abnormal\_long\_term\_variability is: 226.10000000000002

------------------------------------------------

Min of mean\_value\_of\_long\_term\_variability is: 0.0

Max of mean\_value\_of\_long\_term\_variability is: 50.7

Mean of mean\_value\_of\_long\_term\_variability is: 8.19886524822695

Median of mean\_value\_of\_long\_term\_variability is: 7.4

Standard Deviation of mean\_value\_of\_long\_term\_variability is: 5.636014802636036

Zi sum value mean\_value\_of\_long\_term\_variability is: 339.70000000000005

------------------------------------------------

Min of histogram\_width is: 3

Max of histogram\_width is: 180

Mean of histogram\_width is: 70.49267139479906

Median of histogram\_width is: 68.0

Standard Deviation of histogram\_width is: 38.95078910381832

Zi sum value histogram\_width is: 807.9000000000001

------------------------------------------------

Min of histogram\_min is: 50

Max of histogram\_min is: 159

Mean of histogram\_min is: 93.54704491725768

Median of histogram\_min is: 93.0

Standard Deviation of histogram\_min is: 29.580143601297536

Zi sum value histogram\_min is: 844.7

------------------------------------------------

Min of histogram\_max is: 122

Max of histogram\_max is: 238

Mean of histogram\_max is: 164.03971631205673

Median of histogram\_max is: 162.0

Standard Deviation of histogram\_max is: 17.944182664569215

Zi sum value histogram\_max is: 768.3

------------------------------------------------

Min of histogram\_number\_of\_peaks is: 0

Max of histogram\_number\_of\_peaks is: 18

Mean of histogram\_number\_of\_peaks is: 4.069976359338061

Median of histogram\_number\_of\_peaks is: 3.0

Standard Deviation of histogram\_number\_of\_peaks is: 2.947638560550025

Zi sum value histogram\_number\_of\_peaks is: 492.3

------------------------------------------------

Min of histogram\_number\_of\_zeroes is: 0

Max of histogram\_number\_of\_zeroes is: 10

Mean of histogram\_number\_of\_zeroes is: 0.3224586288416076

Median of histogram\_number\_of\_zeroes is: 0.0

Standard Deviation of histogram\_number\_of\_zeroes is: 0.7041179104063311

Zi sum value histogram\_number\_of\_zeroes is: 68.20000000000002

------------------------------------------------

Min of histogram\_mode is: 60

Max of histogram\_mode is: 187

Mean of histogram\_mode is: 137.44586288416076

Median of histogram\_mode is: 139.0

Standard Deviation of histogram\_mode is: 16.403044334341104

Zi sum value histogram\_mode is: 1289.5

------------------------------------------------

Min of histogram\_mean is: 73

Max of histogram\_mean is: 182

Mean of histogram\_mean is: 134.60614657210402

Median of histogram\_mean is: 136.0

Standard Deviation of histogram\_mean is: 15.617939011476754

Zi sum value histogram\_mean is: 1198.4999999999998

------------------------------------------------

Min of histogram\_median is: 77

Max of histogram\_median is: 186

Mean of histogram\_median is: 138.09078014184396

Median of histogram\_median is: 139.0

Standard Deviation of histogram\_median is: 14.48426821225838

Zi sum value histogram\_median is: 1184.8

------------------------------------------------

Min of histogram\_variance is: 0

Max of histogram\_variance is: 269

Mean of histogram\_variance is: 18.846808510638297

Median of histogram\_variance is: 7.0

Standard Deviation of histogram\_variance is: 29.03532873969992

Zi sum value histogram\_variance is: 135.7

------------------------------------------------

Min of histogram\_tendency is: -1

Max of histogram\_tendency is: 1

Mean of histogram\_tendency is: 0.32056737588652484

Median of histogram\_tendency is: 0.0

Standard Deviation of histogram\_tendency is: 0.6107888873223769

Zi sum value histogram\_tendency is: 1396.5

------------------------------------------------

z\_i\_fetal\_health 1499.7

**z\_i\_histogram\_tendency 1396.5**

**z\_i\_histogram\_mode 1289.5**

**z\_i\_histogram\_mean 1198.5**

**z\_i\_histogram\_median 1184.8**

**z\_i\_baseline\_value 1074.6**

**z\_i\_abnormal\_short\_term\_variability 983.9**

**z\_i\_histogram\_min 844.7**

z\_i\_histogram\_width 807.9

z\_i\_histogram\_max 768.3

z\_i\_uterine\_contractions 616.8

z\_i\_histogram\_number\_of\_peaks 492.3

z\_i\_accelerations 367.8

z\_i\_mean\_value\_of\_short\_term\_variability 345.2

z\_i\_mean\_value\_of\_long\_term\_variability 339.7

z\_i\_light\_decelerations 268.2

z\_i\_percentage\_of\_time\_with\_abnormal\_long\_term\_variability 226.1

z\_i\_histogram\_variance 135.7

z\_i\_histogram\_number\_of\_zeroes 68.2

z\_i\_prolongued\_decelerations 67.4

z\_i\_fetal\_movement 34.6

~~z\_i\_severe\_decelerations 7.0~~

dtype: float64

Process finished with exit code 0

* ‘fetal\_health’ value is the categorical feature that was originally used to classify the data so I have excluded this as a key feature but more like a classifier field.
* The 7 key features thereafter and info are highlighted in yellow.
* The next 3 fields highlighted in orange were non histogram related fields.
* The features with z sum less than 40 are marked in blue. I removed ‘severe\_decelerations’ altogether as I don’t think it’s even close to the threshold to be very useful. However out of the seven non zero values there were 6 records with the same fetal health classifier assigned but this is probably coincidence as the data sample is too small.

Output Verified in Excel Worksheet after manually manipulating incremental csv output files:

A table with numbers and letters

Description automatically generated

**1.2.**

**Use 3 classifiers [k-NN, Naive Bayes, Decision Tree] we have discussed in the module [20] and compare the accuracy of the Fetal Health predictions. You should choose the accuracy measures, explain your choices, and discuss some reasons for the different accuracy values. What is the optimal value of k for the k-NN classifier?**

**Do you get better accuracy if you use 1/d weighting? Do the results change if you use k = 3 or k = 5-fold cross-validation?**

* Naïve Bayes is least accuracte.
* KNN and decision tree are close in accuracy (over 95% so good options).
* Running with k=1, then k=3 and k=5 improved accuracy each time when updating to use 1/distance weighting.

Naive Bayes

A screenshot of a computer

Description automatically generated



Decision Tree

A screenshot of a computer

Description automatically generated



k-NN

A screenshot of a computer

Description automatically generated



**1.3.**

**i. Plot the ROC curves for 3 different classification models (use Weka for this). [15]**

**ii. What do you learn from these ROC curve? Include the AUC in your discussion.**

**iii. Which classifier/configuration is best suited for this task? Are you satisfied with the performance?**

Weka Plots and visualisation notes:

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

**1.4.**

**This dataset has many features. Carefully identify the most discriminating features to [20] predict Fetal Health.**

Top 14 features after normalisation of numerical features (top 6 non histogram features in bold / red)

1. histogram\_tendency
2. histogram\_mode
3. histogram\_mean
4. histogram\_median
5. **baseline\_value**
6. **abnormal\_short\_term\_variability**
7. histogram\_min
8. histogram\_width
9. histogram\_max
10. **uterine\_contractions**
11. histogram\_number\_of\_peaks
12. **accelerations**
13. **mean\_value\_of\_short\_term\_variability**
14. **mean\_value\_of\_long\_term\_variability**

**1.5.**

**Use the following methods in Weka to find the top 5 features:**

**• Wrapper + forward search**

A screenshot of a computer

Description automatically generated



1. accelerations

2. prolonged\_decelerations

3. abnormal\_short\_term\_variability

4. mean\_value\_of\_short\_term\_variability

5. percentage\_of\_time\_with\_abnormal\_long\_term\_variability

~~6. histogram\_mean~~

**• Wrapper + backwards search**

A screenshot of a computer program

Description automatically generated



1. accelerations

2. prolonged\_decelerations

3. abnormal\_short\_term\_variability

4. mean\_value\_of\_short\_term\_variability

5. percentage\_of\_time\_with\_abnormal\_long\_term\_variability

~~6. histogram\_mean~~

**• Information Gain Discuss the differences in the selected sets of features.**

A screenshot of a computer

Description automatically generated



1. mean\_value\_of\_short\_term\_variability

2. percentage\_of\_time\_with\_abnormal\_long\_term\_variability

3. abnormal\_short\_term\_variability

4. histogram\_mean

5. histogram\_variance

~~6. histogram\_mode~~

**1.6.**

**Evaluate the performance of the [Decision Tree, Naive Bayes, k-NN] classifiers with the [20] top 5 features selected in the previous part. Comment on the differences you observe in the performance.**

Wrapper search forward and backward produce the same top 5 fields.

IG produces are different and more accurate top 5 feature list for what I have compared to up to this point.

**1.7.**

**Use Weka to extract the top 2 Principal Components. Evaluate the performance of the [20] [Decision Tree, Naive Bayes, k-NN] classifiers with these principal components and comment on any differences with previous methods. Visualise the principal components (you can add a screenshot of Weka) and comment on how they match with your expectations, having run various models and feature selections.**

A screenshot of a computer

Description automatically generated

