CAP667 Justin Johnson
Dr Taghi Khoshgoftaar Z23136514

Assignment 1
Prediction and Classification of Fault Prone Modules Using WEKA

Introduction)

Labelled data describing software modules with 9 software process attributes, one of which includes the total number of faults per module, is analyzed and prepared for WEKA modelling by converting to the ARFF file format. The training data (fit.dat), containing 188 instances is used to train both predictive and classification models using WEKA. The test data's 94 instances (test.dat) is then used to evaluate the models and compare results.

Part 1-A) Preparing Data Sets - Converting to ARFF

Two sets of data are created from the original raw data, one for quantitative prediction and one for classification. For both sets, the raw data format was converted to ARFF format by replacing the attribute space delimeters with commas, as ARFF requires comma delimiters.

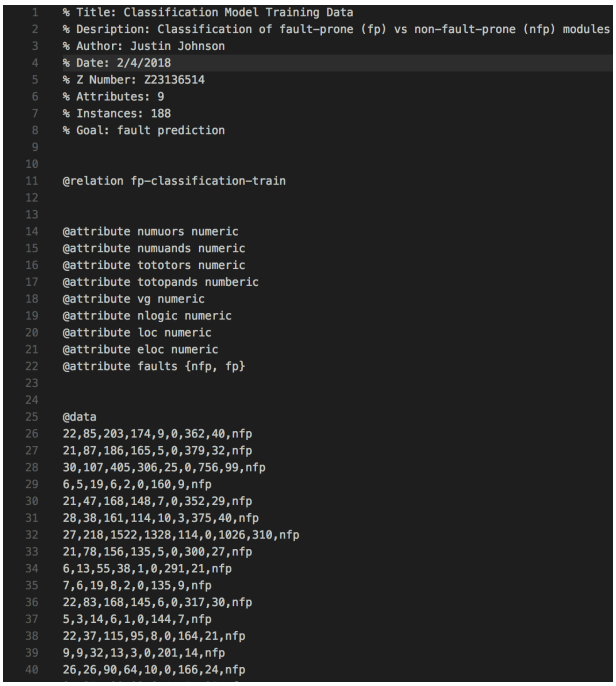
Set 1 did not require additional formatting. A relation was defined for both fit and test data, all 9 attributes were provided labels and marked as type numeric, and the data set was prefixed with the appropriate data label.

```
% Title: Predictive Model Training Data
% Author: Justin Johnson
% Date: 2/4/2018
% Z Number: Z23136514
% Attributes: 9
% Instances: 188
% Goal: fault prediction


@relation fault-prediction-train


@attribute numuors numeric
@attribute numuands numeric
@attribute tototors numeric
@attribute totopands numberic
@attribute vg numeric
@attribute nlogic numeric
@attribute loc numeric
@attribute eloc numeric
@attribute faults numeric


@data
22,85,203,174,9,0,362,40,0
21,87,186,165,5,0,379,32,0
30,107,405,306,25,0,756,99,0
6,5,19,6,2,0,160,9,0
21,47,168,148,7,0,352,29,0
28,38,161,114,10,3,375,40,0
27,218,1522,1328,114,0,1026,310,0
21,78,156,135,5,0,300,27,0
6,13,55,38,1,0,291,21,0
7,6,19,8,2,0,135,9,0
22,83,168,145,6,0,317,30,0
5,3,14,6,1,0,144,7,0
22,37,115,95,8,0,164,21,0
9,9,32,13,3,0,201,14,0
26,26,90,64,10,0,166,24,0
24,35,120,83,6,4,151,25,0
```

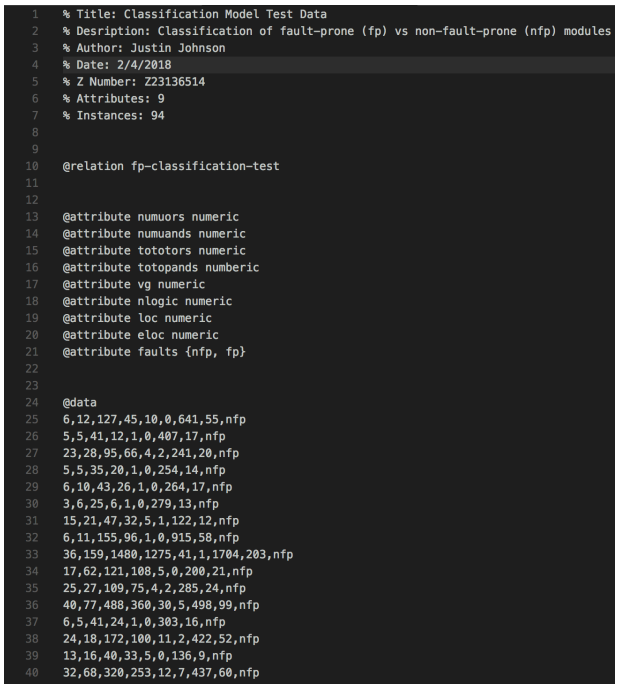
```
% Title: Predictive Model Test Data
% Description: Fault count prediction
% Author: Justin Johnson
% Date: 2/4/2018
% Z Number: Z23136514
% Attributes: 9
% Instances: 94

@relation fault-prediction-test

@attribute numuors numeric
@attribute numuands numeric
@attribute tototors numeric
@attribute totopands numberic
@attribute vg numeric
@attribute nlogic numeric
@attribute loc numeric
@attribute eloc numeric
@attribute faults numeric

@data
6,12,127,45,10,0,641,55,0
5,5,41,12,1,0,407,17,0
23,28,95,66,4,2,241,20,0
5,5,35,20,1,0,254,14,0
6,10,43,26,1,0,264,17,0
3,6,25,6,1,0,279,13,0
15,21,47,32,5,1,122,12,0
6,11,155,96,1,0,915,58,0
36,159,1480,1275,41,1,1704,203,0
17,62,121,108,5,0,200,21,0
25,27,109,75,4,2,285,24,0
40,77,488,360,30,5,498,99,0
6,5,41,24,1,0,303,16,0
24,18,172,100,11,2,422,52,0
13,16,40,33,5,0,136,9,0
32,68,320,253,12,7,437,60,0
10,11,36,24,3,0,158,13,0
14,29,52,42,2,0,123,9,0
15,43,91,72,3,0,355,21,0
```

Training data (left img) and test data (right img) for predictive modelling (Set 1) after converted to ARFF format.

Since Set 2 is used for classification, preparing the data required an extra step. Per assignment instructions, a module is labelled as fault-prone if it contains 2 or more faults. Therefore, the 9th attribute (total number of faults) is converted to a nominal value of ‘fp’ or ‘nfp’, depending on the total number of faults on the given instance.



```
% Title: Classification Model Training Data
% Desription: Classification of fault-prone (fp) vs non-fault-prone (nfp) modules
% Author: Justin Johnson
% Date: 2/4/2018
% Z Number: Z23136514
% Attributes: 9
% Instances: 188
% Goal: fault prediction


@relation fp-classification-train


@attribute numuors numeric
@attribute numuands numeric
@attribute tototors numeric
@attribute totopands numberic
@attribute vg numeric
@attribute nlogic numeric
@attribute loc numeric
@attribute eloc numeric
@attribute faults {nfp, fp}


@data
22,85,203,174,9,0,362,40,nfp
21,87,186,165,5,0,379,32,nfp
30,107,405,306,25,0,756,99,nfp
6,5,19,6,2,0,160,9,nfp
21,47,168,148,7,0,352,29,nfp
28,38,161,114,10,3,375,40,nfp
27,218,1522,1328,114,0,1026,310,nfp
21,78,156,135,5,0,300,27,nfp
6,13,55,38,1,0,291,21,nfp
7,6,19,8,2,0,135,9,nfp
22,83,168,145,6,0,317,30,nfp
5,3,14,6,1,0,144,7,nfp
22,37,115,95,8,0,164,21,nfp
9,9,32,13,3,0,201,14,nfp
26,26,90,64,10,0,166,24,nfp
```



```
% Title: Classification Model Test Data
% Desription: Classification of fault-prone (fp) vs non-fault-prone (nfp) modules
% Author: Justin Johnson
% Date: 2/4/2018
% Z Number: Z23136514
% Attributes: 9
% Instances: 94


@relation fp-classification-test


@attribute numuors numeric
@attribute numuands numeric
@attribute tototors numeric
@attribute totopands numberic
@attribute vg numeric
@attribute nlogic numeric
@attribute loc numeric
@attribute eloc numeric
@attribute faults {nfp, fp}


@data
6,12,127,45,10,0,641,55,nfp
5,5,41,12,1,0,407,17,nfp
23,28,95,66,4,2,241,20,nfp
5,5,35,20,1,0,254,14,nfp
6,10,43,26,1,0,264,17,nfp
3,6,25,6,1,0,279,13,nfp
15,21,47,32,5,1,122,12,nfp
6,11,155,96,1,0,915,58,nfp
36,159,1480,1275,41,1,1704,203,nfp
17,62,121,108,5,0,200,21,nfp
25,27,109,75,4,2,285,24,nfp
40,77,488,360,30,5,498,99,nfp
6,5,41,24,1,0,303,16,nfp
24,18,172,100,11,2,422,52,nfp
13,16,40,33,5,0,136,9,nfp
32,68,320,253,12,7,437,60,nfp
```

Training data (left img) and test data (right img) after being prepared for classification. Note the last attribute has been converted to a nominal value of either nfp or fp. The label was determined by comparing total number of faults to a threshold of 2, such that greater than 2 faults is labelled fault-prone.

Part 1-B) Linear Regression & Decision Stump Training and Testing

The data sets from part 1-A are now used to train Linear Regression and Decision Stump models. Trained models will first be validated using 10-fold cross validation with the training data, and then they will be validated using a separate test data set.

Linear Regression

For the linear regression prediction models, three different feature selection options are used, and the model's results are compared. The three variants compared include: M5, Greedy, and No Attribute Selection.

Linear Regression: 10-Fold Cross Validation With Fit Data Results					
	Correlation Coefficient	Mean Absolute Error	Root Mean Sqrd Error	Relative Absolute Error	Root Relative Sqrd Error
M5	0.7935	1.7017	2.8612	58.8734 %	61.3972 %
Greedy	0.7961	1.6939	2.8425	58.6027 %	60.9977 %
No Attr Selection	0.7969	1.6902	2.8362	58.4755 %	60.8616 %

Linear Regression: Model Validation With Test Data Results					
	Correlation Coefficient	Mean Absolute Error	Root Mean Sqrd Error	Relative Absolute Error	Root Relative Sqrd Error
M5	0.8290	1.8376	3.7324	58.6423 %	63.7000 %
Greedy	0.8314	1.8383	3.6895	58.6625 %	62.9680%
No Attr Selection	0.8290	1.8377	3.7317	58.6426 %	63.6881 %

Resulting Linear Regression Models

Greedy 10-Fold CV

```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


Linear Regression Model

faults =

     -0.0482 * numuors +
      0.0336 * numuands +
     -0.0021 * tototors +
     -0.0337 * vg +
      0.2088 * nlogic +
      0.0019 * loc +
     -0.3255
```

Greedy Separate Test Data

```
=== Classifier model (full training set) ===


Linear Regression Model

faults =

     -0.0482 * numuors +
      0.0336 * numuands +
     -0.0021 * tototors +
     -0.0337 * vg +
      0.2088 * nlogic +
      0.0019 * loc +
     -0.3255
```

The Greedy feature selection algorithm produced Linear Regression models that include 7 attributes and an intercept. The 'totoponds' (total number of operands) attribute is not included in the models produced by Greedy algorithm. The 'nlogic' attribute is given the most positive weight in predicting the total number of faults.

M5 10-Fold CV

```
=== Classifier model (full training set) =

Linear Regression Model

faults =

     -0.0516 * numuors +
      0.0341 * numuands +
     -0.0027 * totoTors +
     -0.0372 * vg +
      0.2119 * nlogic +
      0.0018 * loc +
      0.005  * eloc +
     -0.3091
```

M5 Separate Test Data

```
Linear Regression Model

faults =

     -0.0516 * numuors +
      0.0341 * numuands +
     -0.0027 * totoTors +
     -0.0372 * vg +
      0.2119 * nlogic +
      0.0018 * loc +
      0.005  * eloc +
     -0.3091
```

Similar to the Greedy algorithm, the M5 feature selection algorithm also produced models that include 7 of 8 attributes, leaving out the 'totopands' attribute. Again, the resulting model gives the greatest weight to the 'nlogic' attribute.

No Attr Selection CV

```
=== Classifier model (full training set) ===

Linear Regression Model

faults =

     -0.0517 * numuors +
      0.0341 * numuands +
     -0.0026 * totoTors +
     -0      * totopands +
     -0.0372 * vg +
      0.2118 * nlogic +
      0.0018 * loc +
      0.005  * eloc +
     -0.309
```

No Attr Selection Test Data

```
Linear Regression Model

faults =

     -0.0517 * numuors +
      0.0341 * numuands +
     -0.0026 * totoTors +
     -0      * totopands +
     -0.0372 * vg +
      0.2118 * nlogic +
      0.0018 * loc +
      0.005  * eloc +
     -0.309
```

When using no attribute selection, of course all 8 attributes are included in the resulting model. The trained model gave the 'totopands' attribute a weight of 0, which has the same effect as removing the attribute all together. Again the 'nlogic' attribute was given the greatest positive weight for predicting the total number of faults in a given software module.

Decision Stump Tree

Next the data prepared for classification is used to create a Decision Stump Tree model. Below are the results obtained with 10-fold cross validation (training data) and the results obtained by validating the trained model with the test data.

Decision Stump Tree: 10-Fold Cross Validation vs Separate Test Data Summary					
	Kappa Statistic	Mean Absolute Error	Root Mean Sqrd Error	Relative Absolute Error	Root Relative Sqrd Error
10-fold CV	0.5779	0.2425	0.3735	58.4246 %	82.0818 %
Separate Test Data	0.6254	0.2321	0.3427	55.6499 %	74.9407 %

The error results from 10-fold cross validation and the separate test data set validation are very similar, differing by just a few points in all error calculations.

Decision Stump Tree: 10-Fold Cross Validation vs Separate Test Data - Detailed Accuracy By Class						
	10-Fold Cross Validation			Test Data Validation		
	nfp	fp	Weighted Avg	nfp	Fp	Weighted Avg
TP Rate	0.797	0.836	0.809	0.758	0.964	0.819
FP Rate	0.164	0.203	0.175	0.036	0.242	0.097
Precision	0.922	0.630	0.836	0.980	0.628	0.875
Recall	0.797	0.836	0.809	0.758	0.964	0.819
F-Measure	0.855	0.719	0.815	0.855	0.761	0.827
MCC	0.591	0.591	0.591	0.663	0.663	0.663
ROC Area	0.833	0.833	0.833	0.861	0.861	0.861
PRC Area	0.918	0.564	0.814	0.913	0.616	0.825

Cross Validation Confusion Matrix:

```
=== Confusion Matrix ===

   a   b   <-- classified as
 106  27 |   a = nfp
   9  46 |   b = fp
```

Test Data Confusion Matrix:

```
=== Confusion Matrix ===

  a  b   <-- classified as
 50 16 |  a = nfp
  1 27 |  b = fp
```

Cross Validation Results:
Type I error = 0.203 Type II error = 0.164

Test Data Validation Results:
Type I error = 0.242 Type II error = 0.036

During cross validation, the decision stump model had fewer false positives, but many more false negatives. The decision stump model performed better on the test data, resulting in significantly less false negatives, a smaller Type II error.

Conclusions)

Software module data characterized by 9 software process attributes was effectively used to train quantitative prediction models (Linear Regression) and classification models (Decision Stump Tree). The attribute of interest is the total number of faults. Linear regression models are trained to predict the total number of faults on new software module instances. In order to classify modules as fault prone or non-fault prone using the Decision Stump tree, training data instances were labelled as fault prone if the instance contained 2 or more faults, otherwise it was labelled as non-fault prone.

Two feature selection algorithms (Greedy and M5) were applied to the Linear Regression model training. Both resulting models excluded the totoponds attribute from the model. The third iteration used no feature selection, and the resulting model gave the totoponds attribute a weight of 0, removing it's influence from the predictor model. All three iterations returned similar error rates, as expected by previous explanation of feature removal. The cross validation and test data results were similar.

Classification was completed using Decision Stump Tree and 10-fold cross validation results were compared to validation with a the test data. The model performed better against the test data, achieving a significantly lower Type II error.