CAP667 Justin Johnson

Dr Taghi Khoshgoftaar Z23136514

**Assignment 1**

**Prediction and Classification of Fault Prone Modules Using WEKA**

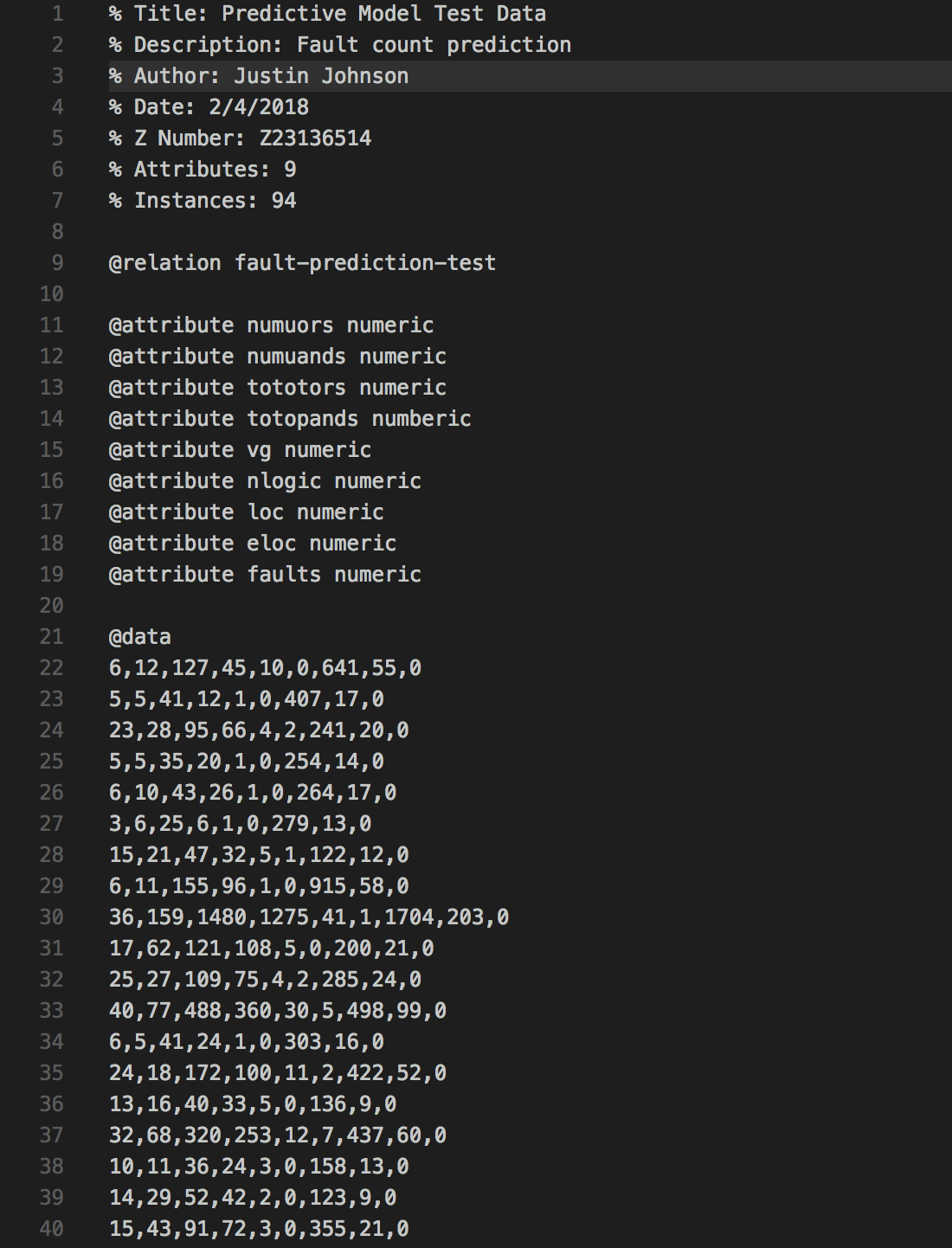
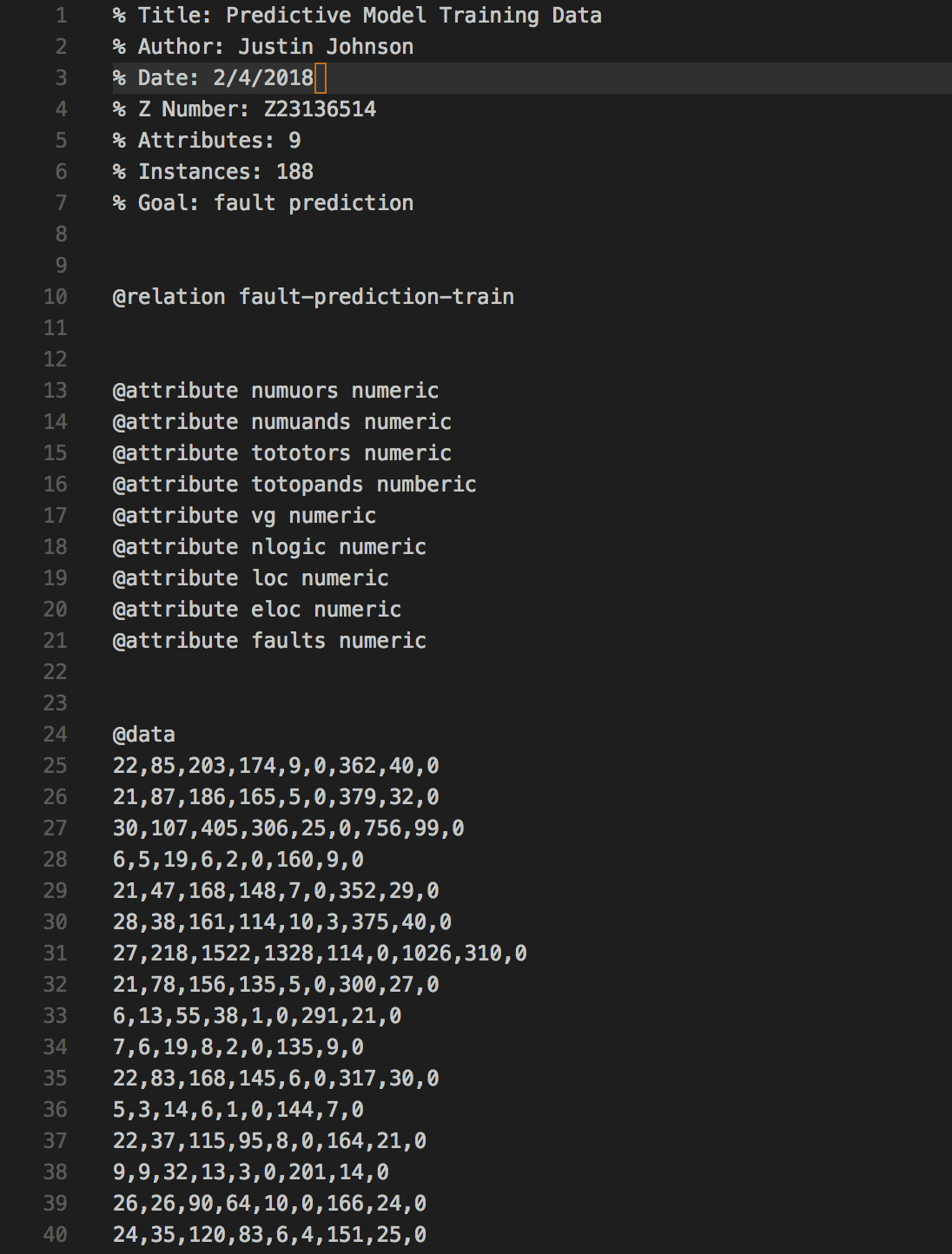
**Introduction)**

Labelled data that describes software modules with 9, one of which includes the total number of faults per module, was analyzed and prepared for WEKA analysis 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 (test.dat) is then used to evaluate the models and compare their results.

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

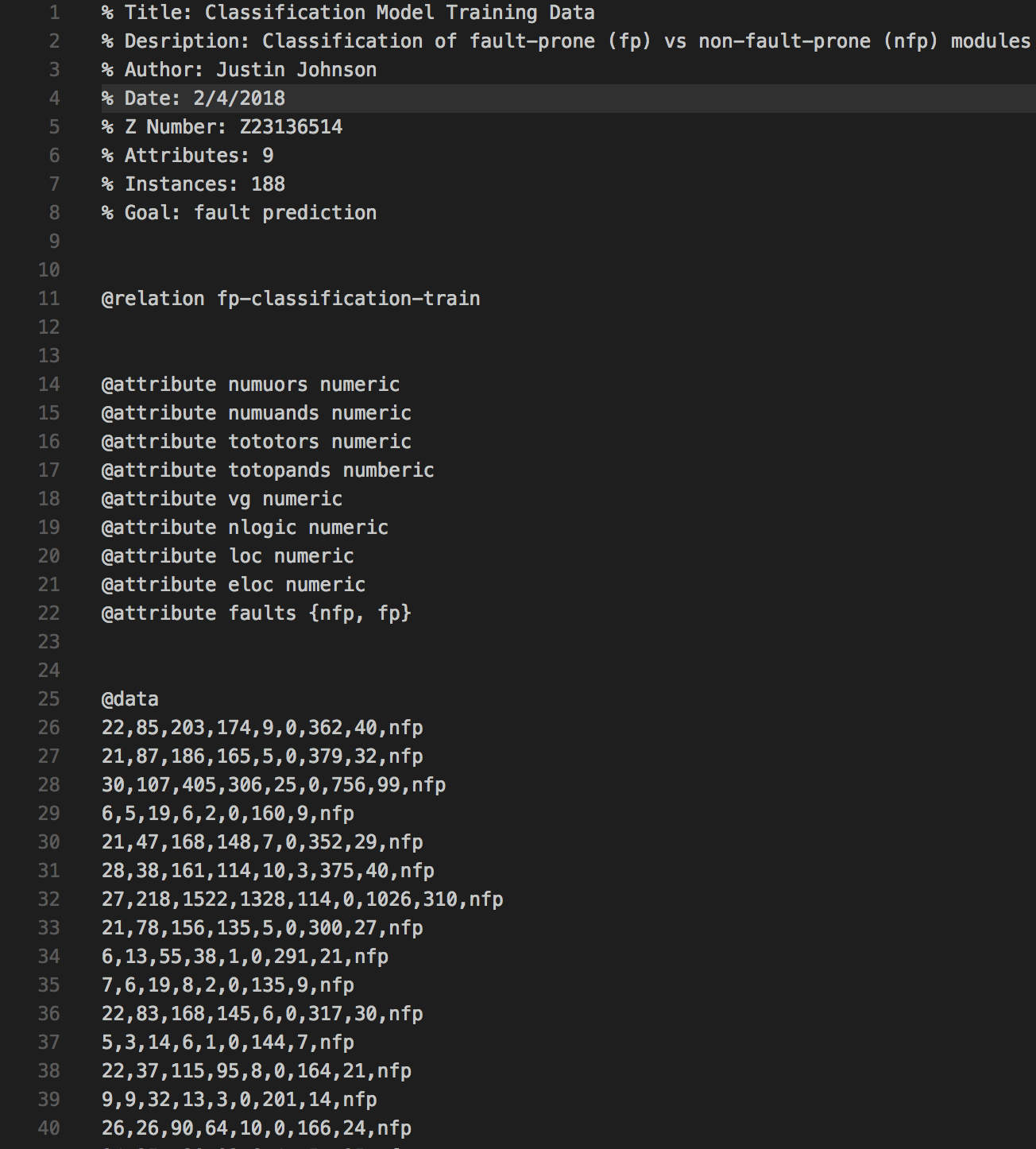
Two sets of data are created from the original raw data, one for predictive modelling and one for classification modelling. For both sets, the raw data format was converted to ARFF format by replacing the attribute spaces with commas, as ARFF requires comma delimiters.

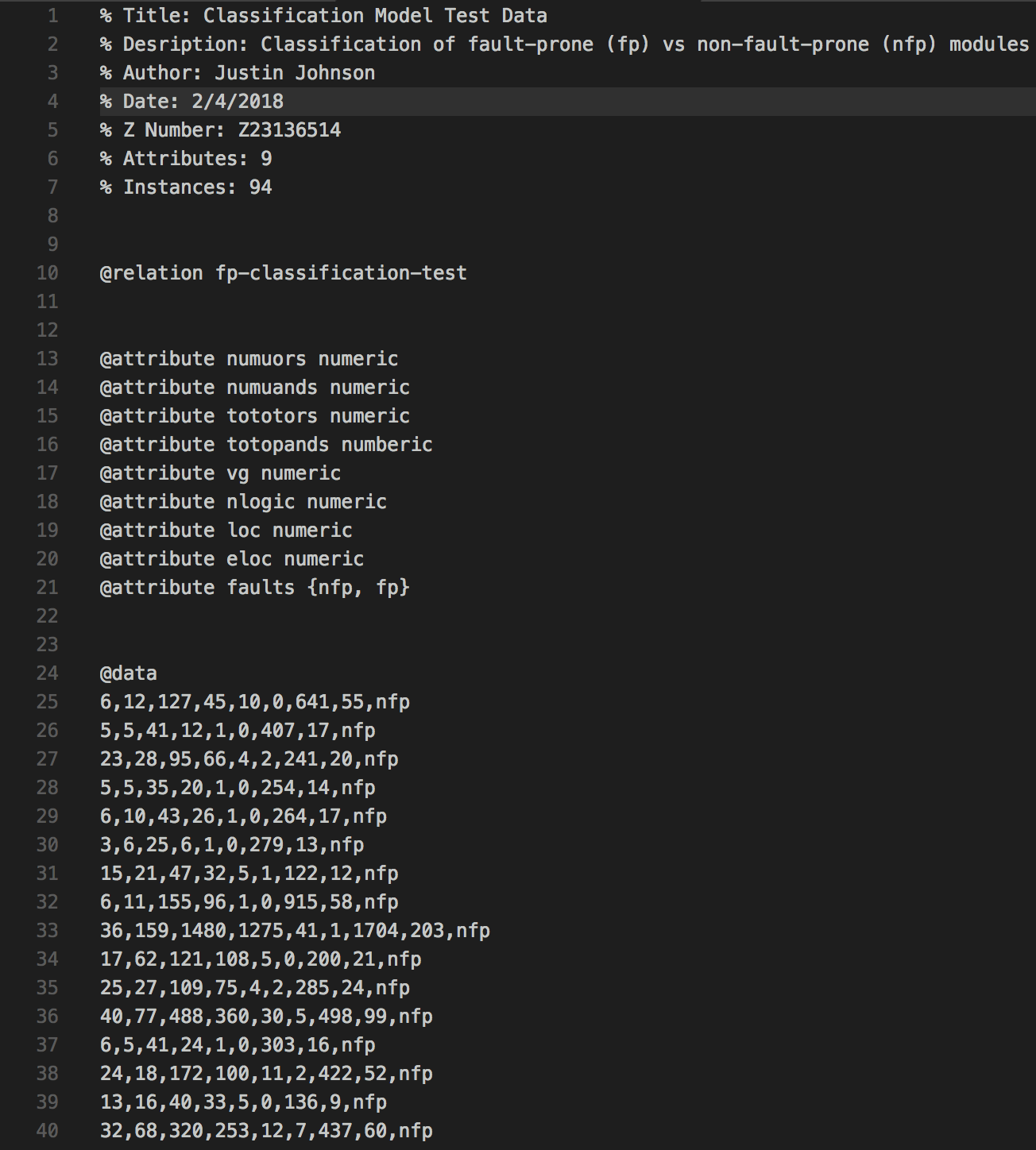
Set 1 did not require additional formatting. A relation was defined for both fit and test data, all 9 attributes were provided labels and marked as type numeric, and the data set was prefixed with the appropriate data label.



*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.





*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 will now be used to train Linear Regression and Decision Stump models. These models will be built and evaluated using 10-fold cross validation on the fit data. Finally, the test data set is used to validate the models.

**Linear Regression**

Three different feature selection options are used and compared: M5, Greedy, and No Attribute Selection.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **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 % |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **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 % |

Three model selection methods were tested: greedy, M5, and no selection.

Compare the models?

How many and which independent variables were selected?

Use Weka to perform comparison.

10-fold cross validation results:

Test Results:

**Decision Stymp**