Project Report: Los Angeles Crime Code Predictive Model

Luke Flecker and Tanush Kallem

10/21/2024

Table of Contents

Part 1 – Statement/Project Goal

The goal of this project is to develop a machine learning model that accurately predicts crime codes based on various features in the Los Angeles crime dataset. This dataset contains detailed records of criminal incidents in Los Angeles since 2020, including information about the type of crime, location, date, and time. By leveraging this data, we aim to build a predictive model that can identify the likely crime code for a given set of features, despite the inherent challenges posed by the dataset's inconsistencies, such as inaccuracies in location data and the potential for transcription errors.

This project has many practical implications for benefitting public safety and resource allocation in Los Angeles. Predicting crime codes with high accuracy can assist law enforcement agencies in identifying trends, allocating resources more efficiently, and potentially preventing crimes by recognizing patterns in criminal activity. Moreover, this project can provide insights into the types of crimes that are more likely to occur under specific conditions, contributing to a more impactful approach to crime prevention and community safety initiatives.

Part 2 – Description of Dataset

This dataset reflects incidents of crime in the City of Los Angeles dating back to 2020. This data is transcribed from original crime reports that are typed on paper and therefore there may be some inaccuracies within the data. Some location fields with missing data are noted as (0°, 0°). Address fields are only provided to the nearest hundred block in order to maintain privacy.

None of the attributes have widely skewed data and most of the bar graphs in WEKA are flat and even. The class, crime codes (crm_cd), has a nicely uniform distribution ranging from 110 to 956. The lower the number the more severe the crime. Overall this data looks promising and easy to work with in training and testing.

Class: Crm Cd (crm_cd)
Dimension before Preprocessing: 27
Instances before Preprocessing: 978628

1. DR_NO (dr_no) - Description: Official file number for each crime report, made up of a 2-digit year, area ID, and 5 digits. **Data Type**: Text. **Missing Values:** 0

2. Date Rptd (date_rptd) - Description: Date when the crime was reported (MM/DD/YYYY). **Data Type**: Floating Timestamp. **Missing Values:** 0

3. DATE OCC (date_occ) - Description: Date when the crime occurred (MM/DD/YYYY). **Data Type**: Floating Timestamp. **Missing Values:** 0

4. TIME OCC (time_occ) - Description: Time the crime occurred, recorded in 24-hour military time. **Data Type**: Text. **Missing Values:** 0

5. AREA (area) - Description: LAPD Geographic Area number (1-21) where the crime occurred. **Data Type**: Text. **Missing Values:** 0

6. AREA NAME (area_name) - Description: Name designation of the LAPD Geographic Area or Patrol Division. **Data Type**: Text. **Missing Values:** 0

7. Rpt Dist No (rpt_dist_no) - Description: Four-digit code representing a sub-area within a Geographic Area. **Data Type**: Text. **Missing Values:** 0

8. Part 1-2 (part_1_2) - Description: Categorizes the type of crime as Part 1 (serious) or Part 2 (less serious). **Data Type**: Number. **Missing Values:** 0

9. Crm Cd (crm_cd) - Description: Unique code indicating the specific crime committed. **Data Type**: Text. **Missing Values:** 0

10. Crm Cd Desc (crm_cd_desc) - Description: Description of the Crime Code provided in Crm Cd. **Data Type**: Text. **Missing Values:** 0

11. Mocodes (mocodes) - Description: Modus Operandi codes describing suspect actions during the crime. **Data Type**: Text. **Missing Values:** ~98778

12. Vict Age (vict_age) - Description: Age of the victim involved in the crime, recorded as a two-character numeric value. **Data Type**: Text. **Missing Values:** 0

13. Vict Sex (vict_sex) - Description: Sex of the victim (F - Female, M - Male, X - Unknown). **Data Type**: Text. **Missing Values:** ~64059

14. Vict Descent (vict_descent) - Description: Descent or ethnic background of the victim, represented by a code. **Data Type**: Text. **Missing Values:** ~64059

15. Premis Cd (premis_cd) - Description: Code indicating the type of structure, vehicle, or location where the crime occurred. **Data Type**: Number. **Missing Values:** 0

16. Premis Desc (premis_desc) - Description: Description corresponding to the Premise Code provided in Premis Cd. **Data Type**: Text. **Missing Values:** 0

17. Weapon Used Cd (weapon_used_cd) - Description: Code indicating the type of weapon used in the crime. **Data Type**: Text. **Missing Values:** ~773109

18. Weapon Desc (weapon_desc) - Description: Description corresponding to the Weapon Used Code provided in Weapon Used Cd. **Data Type**: Text. **Missing Values:** ~773109

19. Status (status) - Description: Current status of the case (IC is the default). **Data Type**: Text. **Missing Values:** 0

20. Status Desc (status_desc) - Description: Description of the status code provided in Status. **Data Type**: Text. **Missing Values:** 0

21. Crm Cd 1 (crm_cd_1) - Description: Primary and most serious crime code associated with the incident. **Data Type**: Text. **Missing Values:** 0

22. Crm Cd 2 (crm_cd_2) - Description: Code for an additional crime, less serious than the primary crime in Crm Cd 1. **Data Type**: Text. **Missing Values:** ~913452

23. Crm Cd 3 (crm_cd_3) - Description: Code for a third crime, less serious than those in Crm Cd 1 and 2. **Data Type**: Text. **Missing Values:** ~977022

24. Crm Cd 4 (crm_cd_4) - Description: Code for a fourth crime, less serious than those in Crm Cd 1, 2, and 3. **Data Type**: Text. **Missing Values:** 978600

25. LOCATION (location) - Description: Street address of the crime incident, rounded to the nearest hundred block. **Data Type**: Text. **Missing Values:** 0

26. Cross Street (cross_street) - Description: Cross street closest to the crime location, complementing the LOCATION field. **Data Type**: Text. **Missing Values:** ~922743

27. LAT (lat) - Description: Latitude coordinate of the crime location. **Data Type**: Number. **Missing Values:** 0

28. LON (lon) - Description: Longitude coordinate of the crime location. **Data Type**: Number. **Missing Values:** 0

Part 3 – Data Preprocessing

Part 3.1 Data Cleanup

Before we started analyzing our data, we wanted to clean and preprocess our data through various techniques such as redundancy and correlational analysis.

We first removed **Crm Cd 2**, **Crm Cd 3**, and **Crm Cd 4** due to the fact that all three of these instances had missing values that had a percentage greater than 70%. These attributes represent additional crimes to the ones seen in **Crm Cd**, however, due to the low number of values in **Crm Cd 2**, and no values in **Crm Cd 3** and **Crm Cd 4**, these attributes provide very little data.

The attribute **Cross Street** was also removed as it had a missing value percentage greater than 70%, which was 84%. This data type represents the cross street of a rounded address, however it is not needed as there are other attributes that provide location based data

We then started replacing missing values with "**Weapon Used Cd**". Since these were codes but ranked, we used median. Weka only does mean or mode, so the median was calculated with a simple python script:

```
import pandas as pd
import numpy as np
data = pd.read_csv('/content/drive/MyDrive/ML 1/Q1 Lab/newData.csv')
data['Weapon Used Cd'] = pd.to_numeric(data["'Weapon Used Cd'"],
errors='coerce')
median_value = data["Weapon Used Cd"].median()
print(median_value)
```

Then we used **ReplaceMissingWithUserConstant**, with **numericReplacementValue** being **400**. We then used **ReplaceMissingValues** to replace all the other missing values. This was using mode for nominal data and mean for numerical data which best fit our attributes. **Vict Age** had hidden missing values, as sometimes the age would be 0 or -1. The data did not explain this, but we believed that they were instances where the age was not reported. To fix this, we wrote a script that would calculate the median age that would replace values less than or equal to 0.

```
import pandas as pd

dataset = pd.read_csv("/content/drive/MyDrive/ML 1/Q1 Lab/cleanData.csv")


age_column = dataset.iloc[:, 4]

valid_ages = age_column[age_column > 0]
median_age = int(valid_ages.median())

dataset.iloc[:, 4] = age_column.apply(lambda age: median_age if age <= 0
else age)

dataset.iloc[:, 4] = dataset.iloc[:, 4].astype(int)

dataset.to_csv("/content/drive/MyDrive/ML 1/Q1 Lab/cleanData.csv",
index=False)
```

Vict Sex also had hidden missing values. In the description it denotes 'X' as gender not reported. Our data also had one instance where the gender was 'H'. The dataset did not describe the 'H', so we attributed it to human error and counted it as a hidden missing value. To replace these missing values, we created a python script that would calculate the mode based on the count of 'F' and 'M'. Then replaced the hidden missing values with that mode.

```
import pandas as pd

dataset = pd.read_csv("/content/drive/MyDrive/ML 1/Q1 Lab/cleanData.csv")

column_name = dataset.columns[5]

valid_genders = dataset[dataset[column_name].isin(['F', 'M'])]

mode_gender = valid_genders[column_name].mode()[0]

dataset[column_name] = dataset[column_name].replace({'H': mode_gender,
'X': mode_gender})

dataset.to_csv("/content/drive/MyDrive/ML 1/Q1 Lab/cleanData.csv",
index=False)
```

Vict Descent had similar hidden missing values, where missing values were filled with an ‘X’. We used a similar script to replace all these hidden missing values with the mode.

```
import pandas as pd
from statistics import mode

dataset = pd.read_csv("/content/drive/MyDrive/ML 1/Q1 Lab/cleanData.csv")

ethnicity_column = dataset.iloc[:, 6]
non_missing_ethnicities = ethnicity_column[ethnicity_column != 'X']

mode_ethnicity = mode(non_missing_ethnicities)

dataset.iloc[:, 6] = dataset.iloc[:, 6].replace('X', mode_ethnicity)

dataset.to_csv("/content/drive/MyDrive/ML 1/Q1 Lab/cleanData.csv",
index=False)
```

After doing more research about **Part 1-2**, we discovered that it classified crimes in either 1 or 2, with 2 being worse crimes. This was a direct correlation to our class, and so we decided to remove this attribute as it is a clear indicator of the severity of the crime.

We then moved **Crm Cd** to the end to attribute it as our class feature.

Crm Cd came with 91 unique numerical values. In essence the lower the crime code the worse the crime is. Thus to make this data nominal, we did equal width binning with 10 bins. This way, we can have a rank of 1-10 of the crimes done, with 1 being the worst crime and 10 being a small crime. Weka was not discretizing this data, so we wrote a simple python script to assign values of 1-10, then used **NumerictoNominal** on the class attribute. After running this on the crime codes, there were no crimes classified as **6**, so WEKA removed the label entirely, leaving **9** unique labels (1, 2, 3, 4, 5, 7, 8, 9, and 10).

```python
import pandas as pd
import numpy as np

dataset = pd.read_csv("/content/drive/MyDrive/ML 1/Q1 Lab/cleanData.csv")

last_column = dataset.columns[-1]
bins = 10
bin_labels = range(1, bins+1)
dataset[last_column] = pd.cut(dataset[last_column], bins=bins,
labels=bin_labels)
output_file = "/content/drive/MyDrive/ML 1/Q1 Lab/newData_binned.csv"
dataset.to_csv(output_file, index=False)
```

Part 3.2 Data Reduction.

We first removed the attribute **Crm Cd 1**, as in the data description, it is stated that **Crm Cd 1** and **Crm Cd** are the same attributes.

The attributes **AREA NAME**, **Crm Cd Desc**, **Premis Desc**, **Weapon Desc**, and **Status Desc** were removed. This is because all these attributes are all descriptions of other attributes in the dataset, such as **Premis Desc** defining **Premis Cd** which is already provided.

The attribute **LOCATION** was removed from the dataset as it could be derived from the **LAT** and **LONG** that are already present.

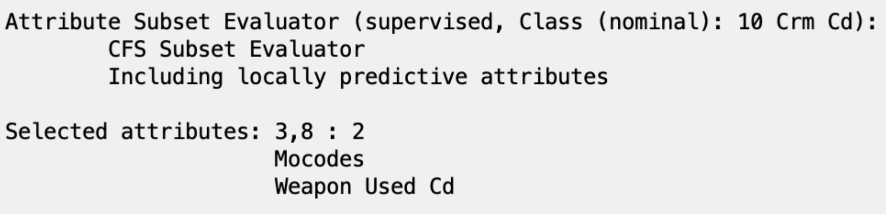
The attribute **Mocodes** was given as a string. However, it was described as a series of codes which each described activities correlated to the suspect. After taking a look at the codes they are closer to nominal, thus, the **StringToNominal** feature was used to convert **Mocodes** to a nominal data type. This allows us to perform correlation analysis as now all our data is Numerical or Nominal.

Part 4 – Attribute Selection and Classifier Models

Part 4.1 Attribute Selection

We then used 5 attribute selectors, four from weka and one that was our personal selection. We then created 5 separate datasets, each with chosen attributes the algorithms have selected.

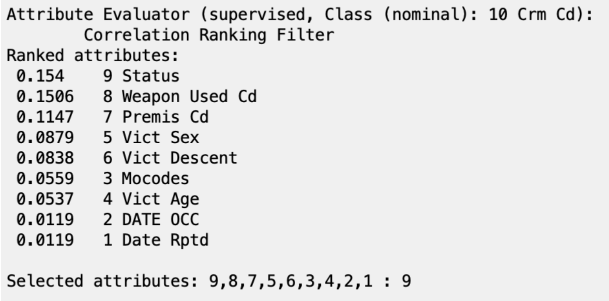
CfsSubsetEval:



```
Attribute Subset Evaluator (supervised, Class (nominal): 10 Crm Cd):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 3,8 : 2
                     Mocodes
                     Weapon Used Cd
```

We kept the **Mocodes** and **Weapon Used Cd** attributes based on the above results.

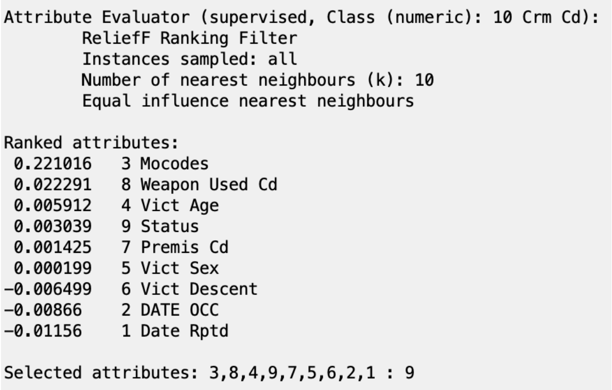
CorrelationAttributeEval:



```
Attribute Evaluator (supervised, Class (nominal): 10 Crm Cd):
        Correlation Ranking Filter
Ranked attributes:
 0.154    9 Status
 0.1506   8 Weapon Used Cd
 0.1147   7 Premis Cd
 0.0879   5 Vict Sex
 0.0838   6 Vict Descent
 0.0559   3 Mocodes
 0.0537   4 Vict Age
 0.0119   2 DATE OCC
 0.0119   1 Date Rptd

Selected attributes: 9,8,7,5,6,3,4,2,1 : 9
```

Based on the above results, we chose a cutoff value of **0.06**. This resulted in us keeping **Status**, **Weapon Used Cd**, **Premis Cd**, **Vict Sex**, and **Vict Descent**.

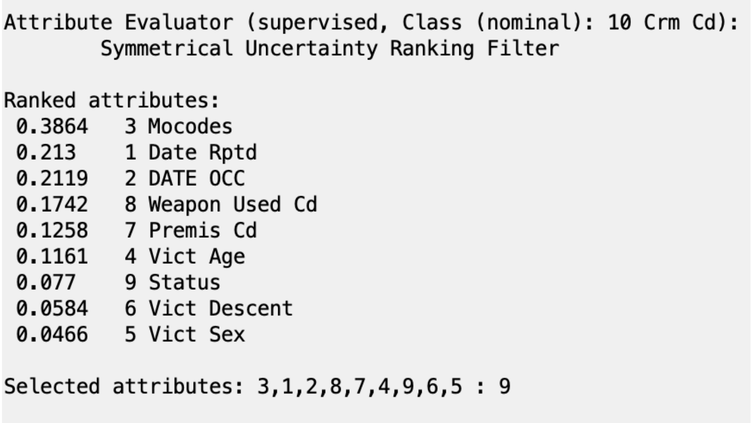
ReliefFAttributeEval:



```
Attribute Evaluator (supervised, Class (numeric): 10 Crm Cd):
        ReliefF Ranking Filter
        Instances sampled: all
        Number of nearest neighbours (k): 10
        Equal influence nearest neighbours

Ranked attributes:
 0.221016   3 Mocodes
 0.022291   8 Weapon Used Cd
 0.005912   4 Vict Age
 0.003039   9 Status
 0.001425   7 Premis Cd
 0.000199   5 Vict Sex
-0.006499   6 Vict Descent
-0.00866    2 DATE OCC
-0.01156    1 Date Rptd

Selected attributes: 3,8,4,9,7,5,6,2,1 : 9
```

Based on the above results, we chose a cutoff value of **0.001**. This resulted in us keeping **Mocodes**, **Weapon Used Cd**, **Vict Age**, **Status**, and **Premis Cd**.

SymmetricalUncertAttributeEval:



```
Attribute Evaluator (supervised, Class (nominal): 10 Crm Cd):
        Symmetrical Uncertainty Ranking Filter

Ranked attributes:
 0.3864   3 Mocodes
 0.213    1 Date Rptd
 0.2119   2 DATE OCC
 0.1742   8 Weapon Used Cd
 0.1258   7 Premis Cd
 0.1161   4 Vict Age
 0.077    9 Status
 0.0584   6 Vict Descent
 0.0466   5 Vict Sex

Selected attributes: 3,1,2,8,7,4,9,6,5 : 9
```

Based on the above results, we chose a cutoff value of **0.15**. This resulted in us keeping **Mocodes**, **DateRptd**, **DATE OCC**, and **Weapon Used Cd**.

For this algorithm, we chose a cutoff value of 0.15, which made us remove the attributes 8, 5, 10, 7, 6.

Personal Selection:

Below are the attributes we decided to **remove** and our justification for why:

1. **Status**: This attribute describes the status of the case. However, we felt like the status of the case wouldn't have a strong relationship with the severity of the crime, as there were other external factors that could affect the status of the case such as the efficiency of the officers assigned it.
2. **Weapon Used Cd**: We felt like many of the weapons used were STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE, and thus the data did not have any number of reports equal to the use of strong arm and felt like a weak attribute for us.
3. **Date Rptd**: We felt that this attribute was unnecessary, as DATE OCC would give us a more accurate timeline of when the crime happened compared to the date that it was reported.

Train-Test-Validation Split

We created 5 sets of train, test, validation splits based on each selection of attributes. To create a train-test-validation split in our dataset, we utilized sklearn and wrote a small python script that would create the separate files and save them.

```
import pandas as pd
from google.colab import files

import sklearn
from sklearn.model_selection import train_test_split
path = '/content/drive/MyDrive/ML 1/Q1 Lab/'
attribute = "SymmetricalUncertAttributeEval/"
dataset = pd.read_csv(path + attribute + "dataset.csv")

trainSet, tempSet = train_test_split(dataset, test_size=0.3,
random_state=42)
valSet, testSet = train_test_split(tempSet, test_size=0.5,
random_state=42)

trainSet.to_csv(path + attribute +'trainSet.csv', index = False)
testSet.to_csv(path + attribute + 'testSet.csv', index = False)
valSet.to_csv(path + attribute + 'valSet.csv', index = False)
```

Part 4.2 Classifier Models

1. **NaiveBayes Classifier (Bayes)**
   The Naive Bayes classifier is a probabilistic model that predicts class membership by applying Bayes' theorem, assuming that features are independent given the class. Despite the independence assumption, it often performs well in various tasks like text classification.

2. **DecisionTable Classifier (Rules)**
   The DecisionTable classifier builds a simple decision table that evaluates multiple attribute-value pairs to make predictions. It constructs rules for each combination of attributes, where internal nodes represent tests, and leaves predict the class based on the majority rule.

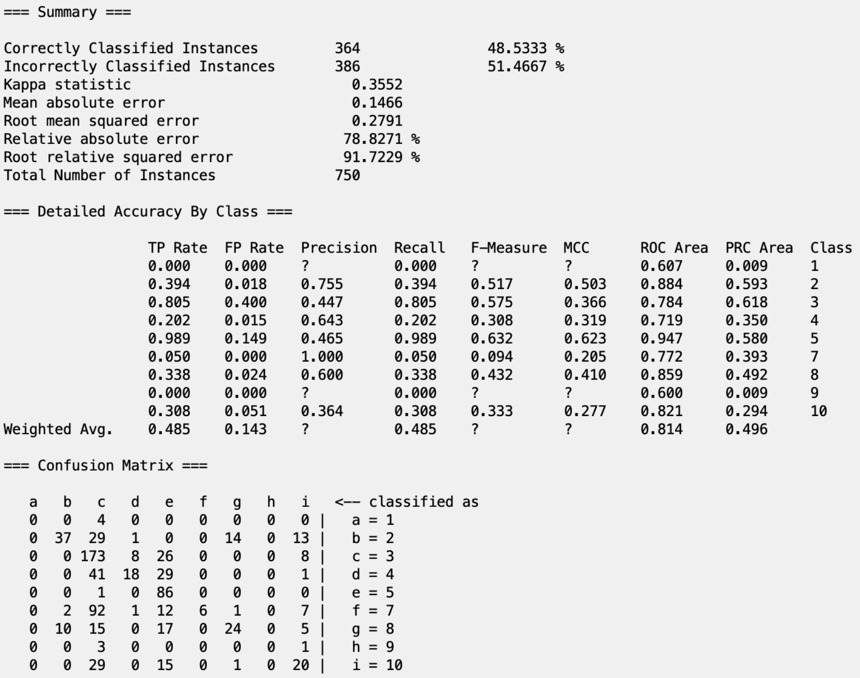
3. **J48 Classifier (Tree)**
   The J48 classifier is an implementation of the C4.5 algorithm, a popular decision tree technique. It recursively splits data based on attribute values to build a tree, with each leaf representing a class. It’s effective for both classification and feature selection.

4. **Random Forest Classifier (Tree)**
   The Random Forest classifier generates a collection of decision trees, each trained on random subsets of the data and features. The final prediction is made based on the majority vote of these independent trees, offering robust performance and reduced overfitting compared to individual trees.

Part 5 – Results and Analysis

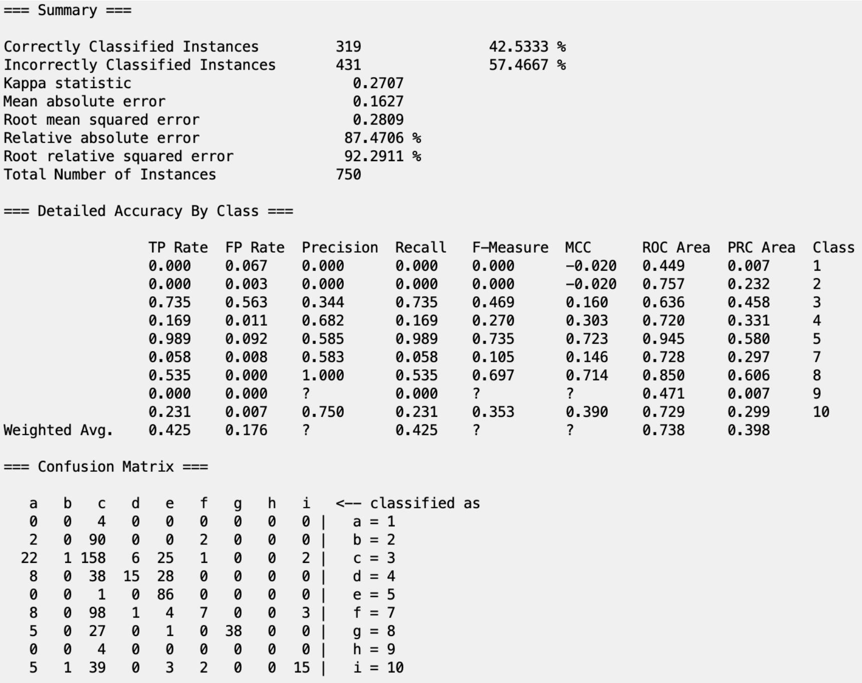
Part 5.1 - Results

CfsSubsetEval with Naive Bayes:



```
=== Summary ===

Correctly Classified Instances         364               48.5333 %
Incorrectly Classified Instances       386               51.4667 %
Kappa statistic                          0.3552
Mean absolute error                      0.1466
Root mean squared error                  0.2791
Relative absolute error                 78.8271 %
Root relative squared error             91.7229 %
Total Number of Instances              750     

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.607     0.009     1
                 0.394    0.018    0.755      0.394    0.517      0.503    0.884     0.593     2
                 0.805    0.400    0.447      0.805    0.575      0.366    0.784     0.618     3
                 0.202    0.015    0.643      0.202    0.308      0.319    0.719     0.350     4
                 0.989    0.149    0.465      0.989    0.632      0.623    0.947     0.580     5
                 0.050    0.000    1.000      0.050    0.094      0.205    0.772     0.393     7
                 0.338    0.024    0.600      0.338    0.432      0.410    0.859     0.492     8
                 0.000    0.000    ?          0.000    ?          ?        0.600     0.009     9
                 0.308    0.051    0.364      0.308    0.333      0.277    0.821     0.294     10
Weighted Avg.    0.485    0.143    ?          0.485    ?          ?        0.814     0.496     

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
   0   0   4   0   0   0   0   0   0 |   a = 1
   0  37  29   1   0   0  14   0  13 |   b = 2
   0   0 173   8  26   0   0   0   8 |   c = 3
   0   0  41  18  29   0   0   0   1 |   d = 4
   0   0   1   0  86   0   0   0   0 |   e = 5
   0   2  92   1  12   6   1   0   7 |   f = 7
   0  10  15   0  17   0  24   0   5 |   g = 8
   0   0   3   0   0   0   0   0   1 |   h = 9
   0   0  29   0  15   0   1   0  20 |   i = 10
```

CfsSubsetEval with DecisionTable:



```
=== Summary ===

Correctly Classified Instances         319               42.5333 %
Incorrectly Classified Instances       431               57.4667 %
Kappa statistic                          0.2707
Mean absolute error                      0.1627
Root mean squared error                  0.2809
Relative absolute error                 87.4706 %
Root relative squared error             92.2911 %
Total Number of Instances              750     

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.067    0.000      0.000    0.000      -0.020   0.449     0.007     1
                 0.000    0.003    0.000      0.000    0.000      -0.020   0.757     0.232     2
                 0.735    0.563    0.344      0.735    0.469      0.160    0.636     0.458     3
                 0.169    0.011    0.682      0.169    0.270      0.303    0.720     0.331     4
                 0.989    0.092    0.585      0.989    0.735      0.723    0.945     0.580     5
                 0.058    0.008    0.583      0.058    0.105      0.146    0.728     0.297     7
                 0.535    0.000    1.000      0.535    0.697      0.714    0.850     0.606     8
                 0.000    0.000    ?          0.000    ?          ?        0.471     0.007     9
                 0.231    0.007    0.750      0.231    0.353      0.390    0.729     0.299     10
Weighted Avg.    0.425    0.176    ?          0.425    ?          ?        0.738     0.398     

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
   0   0   4   0   0   0   0   0   0 |   a = 1
   2   0  90   0   0   2   0   0   0 |   b = 2
  22   1 158   6  25   1   0   0   2 |   c = 3
   8   0  38  15  28   0   0   0   0 |   d = 4
   0   0   1   0  86   0   0   0   0 |   e = 5
   8   0  98   1   4   7   0   0   3 |   f = 7
   5   0  27   0   1   0  38   0   0 |   g = 8
   0   0   4   0   0   0   0   0   0 |   h = 9
   5   1  39   0   3   2   0   0  15 |   i = 10
```

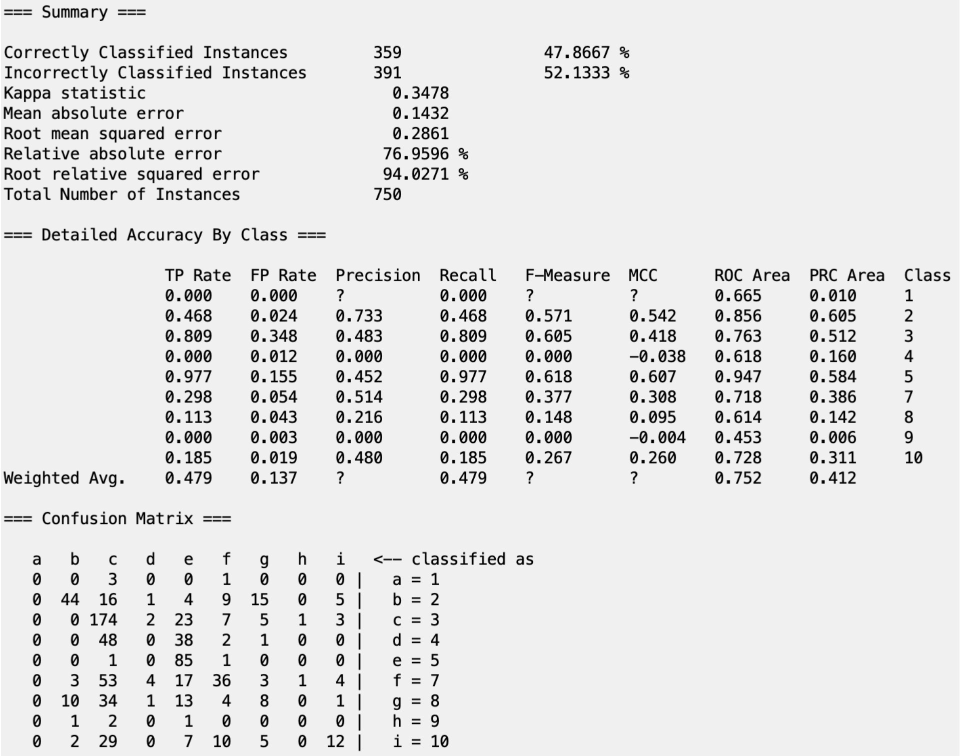
CfsSubsetEval with J48:

```
=== Summary ===

Correctly Classified Instances         421               56.1333 %
Incorrectly Classified Instances       329               43.8667 %
Kappa statistic                          0.4471
Mean absolute error                      0.1245
Root mean squared error                  0.254 
Relative absolute error                 66.9004 %
Root relative squared error             83.4622 %
Total Number of Instances              750     

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.678     0.010     1
                 0.543    0.024    0.761      0.543    0.634      0.601    0.878     0.534     2
                 0.828    0.421    0.442      0.828    0.576      0.369    0.801     0.616     3
                 0.225    0.012    0.714      0.225    0.342      0.363    0.786     0.376     4
                 0.989    0.092    0.585      0.989    0.735      0.723    0.945     0.580     5
                 0.124    0.003    0.882      0.124    0.217      0.299    0.810     0.404     7
                 0.507    0.000    1.000      0.507    0.673      0.694    0.865     0.635     8
                 0.000    0.000    ?          0.000    ?          ?        0.668     0.009     9
                 0.538    0.025    0.673      0.538    0.598      0.569    0.845     0.454     10
Weighted Avg.    0.561    0.138    ?          0.561    ?          ?        0.836     0.520     

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
   0   0   4   0   0   0   0   0   0 |   a = 1
   0  51  38   0   0   1   0   0   4 |   b = 2
   0   0 178   8  25   0   0   0   4 |   c = 3
   0   0  40  20  28   0   0   0   1 |   d = 4
   0   0   1   0  86   0   0   0   0 |   e = 5
   0   3  96   0   4  15   0   0   3 |   f = 7
   0  12  18   0   1   0  36   0   4 |   g = 8
   0   0   3   0   0   0   0   0   1 |   h = 9
   0   1  25   0   3   1   0   0  35 |   i = 10
```

CfsSubsetEval with Random Forest:

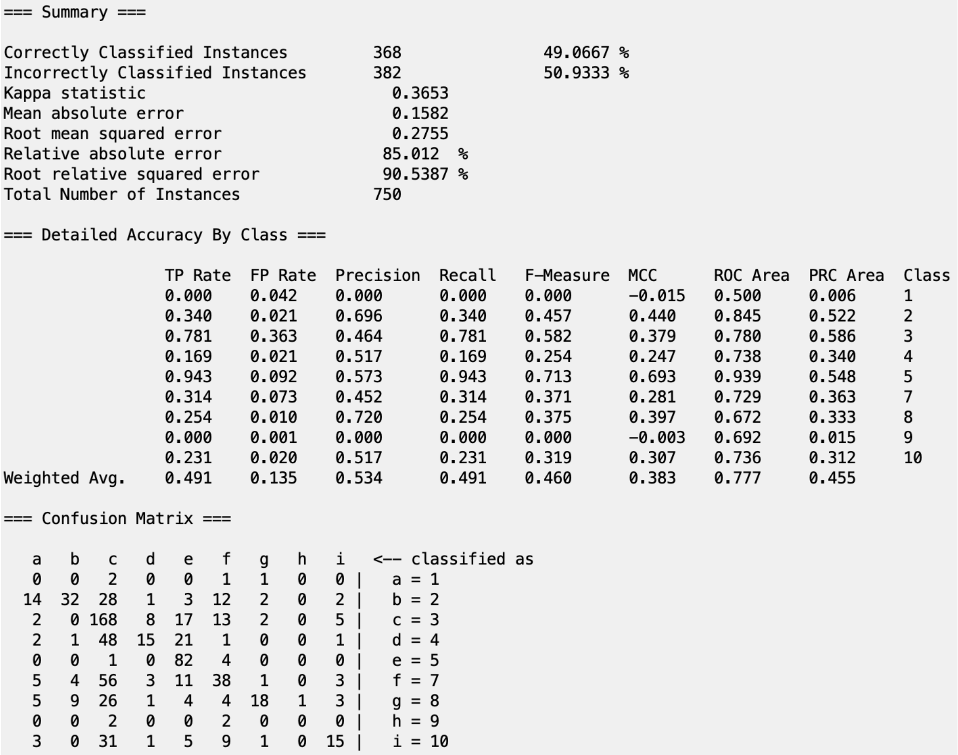
```
=== Summary ===

Correctly Classified Instances         359               47.8667 %
Incorrectly Classified Instances       391               52.1333 %
Kappa statistic                          0.3346
Mean absolute error                      0.1375
Root mean squared error                  0.2619
Relative absolute error                 73.9014 %
Root relative squared error             86.0563 %
Total Number of Instances              750     

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.678     0.011     1
                 0.000    0.002    0.000      0.000    0.000      -0.014   0.884     0.528     2
                 0.791    0.495    0.391      0.791    0.523      0.271    0.780     0.629     3
                 0.236    0.014    0.700      0.236    0.353      0.367    0.795     0.390     4
                 0.989    0.092    0.585      0.989    0.735      0.723    0.945     0.580     5
                 0.174    0.070    0.323      0.174    0.226      0.135    0.762     0.377     7
                 0.577    0.003    0.953      0.577    0.719      0.723    0.858     0.663     8
                 0.000    0.001    0.000      0.000    0.000      -0.003   0.673     0.009     9
                 0.308    0.012    0.714      0.308    0.430      0.439    0.751     0.366     10
Weighted Avg.    0.479    0.167    ?          0.479    ?          ?        0.815     0.515     

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
   0   0   4   0   0   0   0   0   0 |   a = 1
   0   0  77   0   0  15   1   1   0 |   b = 2
   0   0 170   9  25   7   0   0   4 |   c = 3
   0   0  39  21  28   0   0   0   1 |   d = 4
   0   0   1   0  86   0   0   0   0 |   e = 5
   0   0  93   0   4  21   1   0   2 |   f = 7
   0   0  24   0   1   4  41   0   1 |   g = 8
   0   0   3   0   0   1   0   0   0 |   h = 9
   0   1  24   0   3  17   0   0  20 |   i = 10
```

CorrelationAttributeEval with Naive Bayes:



```
=== Summary ===

Correctly Classified Instances         359               47.8667 %
Incorrectly Classified Instances       391               52.1333 %
Kappa statistic                          0.3478
Mean absolute error                      0.1432
Root mean squared error                  0.2861
Relative absolute error                 76.9596 %
Root relative squared error             94.0271 %
Total Number of Instances              750

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.665     0.010     1
                 0.468    0.024    0.733      0.468    0.571      0.542    0.856     0.605     2
                 0.809    0.348    0.483      0.809    0.605      0.418    0.763     0.512     3
                 0.000    0.012    0.000      0.000    0.000      -0.038   0.618     0.160     4
                 0.977    0.155    0.452      0.977    0.618      0.607    0.947     0.584     5
                 0.298    0.054    0.514      0.298    0.377      0.308    0.718     0.386     7
                 0.113    0.043    0.216      0.113    0.148      0.095    0.614     0.142     8
                 0.000    0.003    0.000      0.000    0.000      -0.004   0.453     0.006     9
                 0.185    0.019    0.480      0.185    0.267      0.260    0.728     0.311     10
Weighted Avg.    0.479    0.137    ?          0.479    ?          ?        0.752     0.412

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
   0   0   3   0   0   1   0   0   0 |   a = 1
   0  44  16   1   4   9  15   0   5 |   b = 2
   0   0 174   2  23   7   5   1   3 |   c = 3
   0   0  48   0  38   2   1   0   0 |   d = 4
   0   0   1   0  85   1   0   0   0 |   e = 5
   0   3  53   4  17  36   3   1   4 |   f = 7
   0  10  34   1  13   4   8   0   1 |   g = 8
   0   1   2   0   1   0   0   0   0 |   h = 9
   0   2  29   0   7  10   5   0  12 |   i = 10
```

CorrelationAttributeEval with DecisionTable



```
=== Summary ===

Correctly Classified Instances         368               49.0667 %
Incorrectly Classified Instances       382               50.9333 %
Kappa statistic                          0.3653
Mean absolute error                      0.1582
Root mean squared error                  0.2755
Relative absolute error                 85.012  %
Root relative squared error             90.5387 %
Total Number of Instances              750

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.042    0.000      0.000    0.000      -0.015   0.500     0.006     1
                 0.340    0.021    0.696      0.340    0.457      0.440    0.845     0.522     2
                 0.781    0.363    0.464      0.781    0.582      0.379    0.780     0.586     3
                 0.169    0.021    0.517      0.169    0.254      0.247    0.738     0.340     4
                 0.943    0.092    0.573      0.943    0.713      0.693    0.939     0.548     5
                 0.314    0.073    0.452      0.314    0.371      0.281    0.729     0.363     7
                 0.254    0.010    0.720      0.254    0.375      0.397    0.672     0.333     8
                 0.000    0.001    0.000      0.000    0.000      -0.003   0.692     0.015     9
                 0.231    0.020    0.517      0.231    0.319      0.307    0.736     0.312     10
Weighted Avg.    0.491    0.135    0.534      0.491    0.460      0.383    0.777     0.455

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
   0   0   2   0   0   1   1   0   0 |   a = 1
  14  32  28   1   3  12   2   0   2 |   b = 2
   2   0 168   8  17  13   2   0   5 |   c = 3
   2   1  48  15  21   1   0   0   1 |   d = 4
   0   0   1   0  82   4   0   0   0 |   e = 5
   5   4  56   3  11  38   1   0   3 |   f = 7
   5   9  26   1   4   4  18   1   3 |   g = 8
   0   0   2   0   0   2   0   0   0 |   h = 9
   3   0  31   1   5   9   1   0  15 |   i = 10
```

CorrelationAttributeEval with J48

```
=== Summary ===

Correctly Classified Instances         411               54.8    %
Incorrectly Classified Instances       339               45.2    %
Kappa statistic                          0.4387
Mean absolute error                      0.1383
Root mean squared error                  0.2702
Relative absolute error                 74.3136 %
Root relative squared error             88.784  %
Total Number of Instances              750

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.592     0.008     1
                 0.553    0.038    0.675      0.553    0.608      0.562    0.812     0.534     2
                 0.791    0.290    0.523      0.791    0.630      0.457    0.786     0.527     3
                 0.303    0.054    0.429      0.303    0.355      0.290    0.737     0.318     4
                 0.943    0.069    0.641      0.943    0.763      0.743    0.942     0.598     5
                 0.339    0.084    0.436      0.339    0.381      0.283    0.706     0.325     7
                 0.268    0.021    0.576      0.268    0.365      0.353    0.697     0.393     8
                 0.000    0.000    ?          0.000    ?          ?        0.621     0.009     9
                 0.308    0.015    0.667      0.308    0.421      0.421    0.756     0.350     10
Weighted Avg.    0.548    0.119    ?          0.548    ?          ?        0.776     0.445

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
   0   0   2   0   0   1   1   0   0 |   a = 1
   0  52  17   2   3  13   3   0   4 |   b = 2
   0   2 170  14  10  13   6   0   0 |   c = 3
   0   1  35  27  20   5   1   0   0 |   d = 4
   0   0   2   1  82   2   0   0   0 |   e = 5
   0   6  50  11   7  41   3   0   3 |   f = 7
   0  14  21   6   2   6  19   0   3 |   g = 8
   0   0   2   0   0   2   0   0   0 |   h = 9
   0   2  26   2   4  11   0   0  20 |   i = 10
```

CorrelationAttributeEval with Random Forest

```
=== Summary ===

Correctly Classified Instances         389               51.8667 %
Incorrectly Classified Instances       361               48.1333 %
Kappa statistic                          0.4096
Mean absolute error                      0.1368
Root mean squared error                  0.2761
Relative absolute error                 73.518  %
Root relative squared error             90.7362 %
Total Number of Instances              750

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.449     0.005     1
                 0.489    0.043    0.622      0.489    0.548      0.496    0.809     0.534     2
                 0.712    0.258    0.526      0.712    0.605      0.421    0.784     0.568     3
                 0.292    0.050    0.441      0.292    0.351      0.291    0.711     0.317     4
                 0.954    0.075    0.624      0.954    0.755      0.737    0.953     0.618     5
                 0.314    0.081    0.427      0.314    0.362      0.265    0.722     0.319     7
                 0.310    0.052    0.386      0.310    0.344      0.285    0.670     0.386     8
                 0.000    0.007    0.000      0.000    0.000      -0.006   0.770     0.038     9
                 0.323    0.031    0.500      0.323    0.393      0.358    0.741     0.386     10
Weighted Avg.    0.519    0.115    ?          0.519    ?          ?        0.772     0.461

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
   0   0   1   0   0   2   1   0   0 |   a = 1
   0  46  11   2   3  11  13   0   8 |   b = 2
   0   2 153  13  12  15  13   3   4 |   c = 3
   0   4  30  26  21   4   3   0   1 |   d = 4
   0   0   2   1  83   1   0   0   0 |   e = 5
   0   6  48  10   9  38   3   1   6 |   f = 7
   0  14  20   5   1   6  22   1   2 |   g = 8
   0   0   2   0   0   2   0   0   0 |   h = 9
   0   2  24   2   4  10   2   0  21 |   i = 10
```

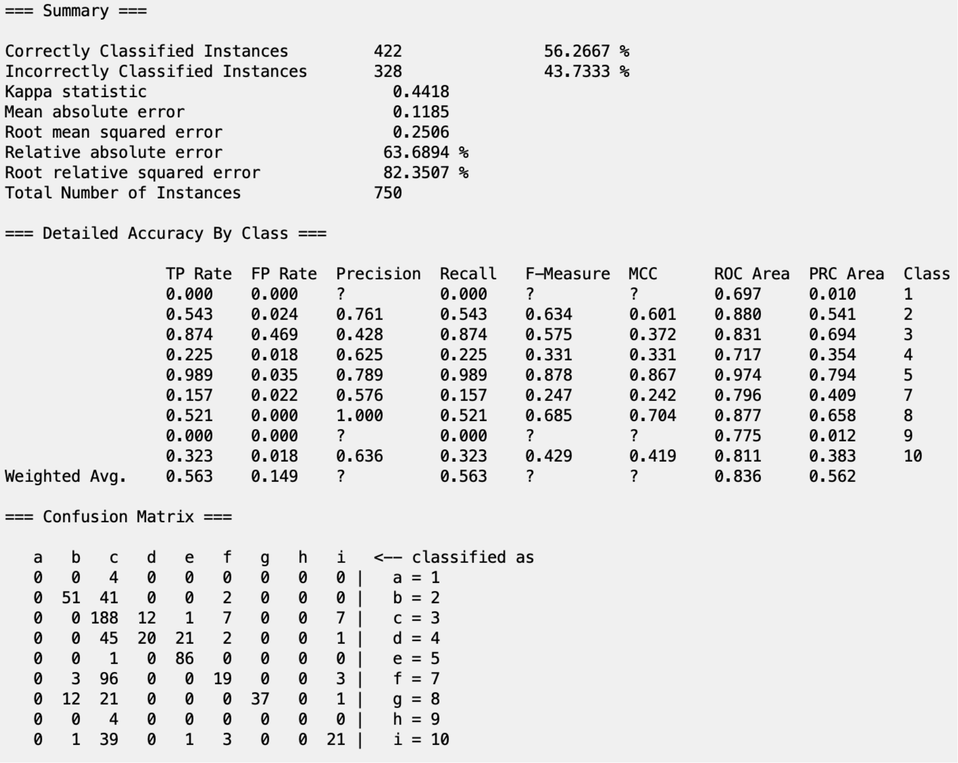
ReliefFAttributeEval with Naive Bayes:

```
=== Summary ===

Correctly Classified Instances         371               49.4667 %
Incorrectly Classified Instances       379               50.5333 %
Kappa statistic                          0.383
Mean absolute error                      0.1299
Root mean squared error                  0.277
Relative absolute error                 69.8416 %
Root relative squared error             91.0248 %
Total Number of Instances              750

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.688     0.010     1
                 0.447    0.024    0.724      0.447    0.553      0.524    0.885     0.628     2
                 0.637    0.262    0.495      0.637    0.557      0.352    0.798     0.651     3
                 0.169    0.026    0.469      0.169    0.248      0.229    0.713     0.356     4
                 0.989    0.204    0.389      0.989    0.558      0.551    0.975     0.768     5
                 0.331    0.046    0.580      0.331    0.421      0.362    0.800     0.484     7
                 0.451    0.040    0.542      0.451    0.492      0.447    0.850     0.491     8
                 0.000    0.001    0.000      0.000    0.000      -0.003   0.396     0.005     9
                 0.292    0.020    0.576      0.292    0.388      0.373    0.817     0.387     10
Weighted Avg.    0.495    0.118    ?          0.495    ?          ?        0.823     0.555

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
   0   0   3   0   0   1   0   0   0 |   a = 1
   0  42  11   0   9  10  16   0   6 |   b = 2
   0   0 137  13  51   6   5   0   3 |   c = 3
   0   0  38  15  32   3   1   0   0 |   d = 4
   0   0   1   0  86   0   0   0   0 |   e = 5
   0   4  52   2  16  40   2   1   4 |   f = 7
   0  10  10   1  15   2  32   0   1 |   g = 8
   0   1   0   0   3   0   0   0   0 |   h = 9
   0   1  25   1   9   7   3   0  19 |   i = 10
```

ReliefFAttributeEval with DecisionTable

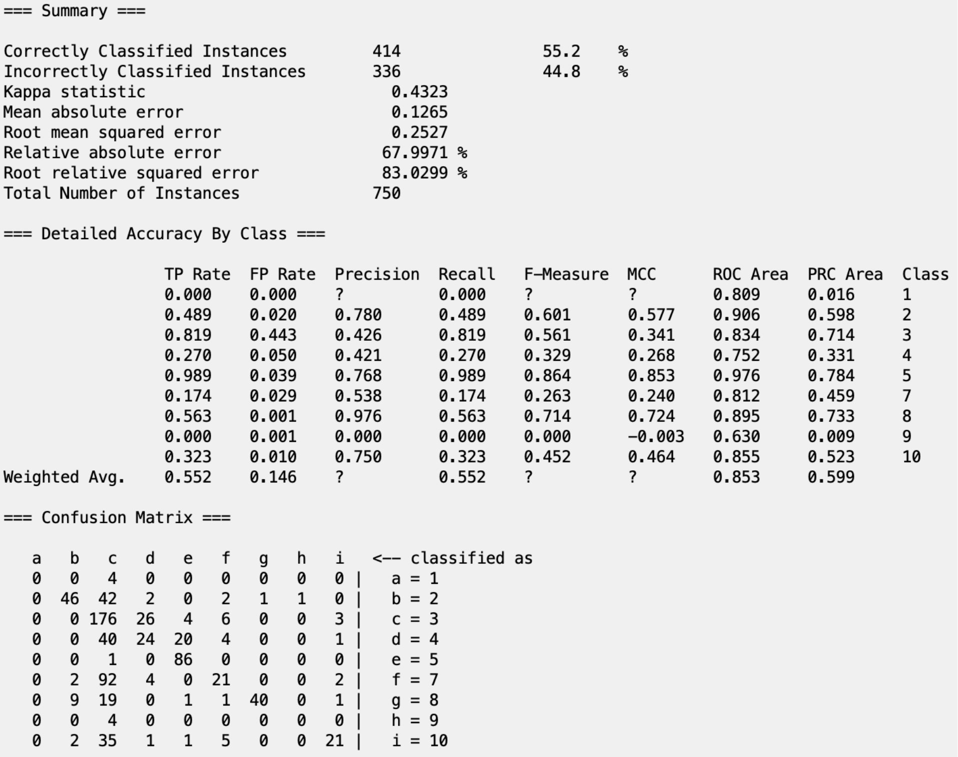
```
=== Summary ===

Correctly Classified Instances         319               42.5333 %
Incorrectly Classified Instances       431               57.4667 %
Kappa statistic                          0.2707
Mean absolute error                      0.1627
Root mean squared error                  0.2809
Relative absolute error                 87.4706 %
Root relative squared error             92.2911 %
Total Number of Instances              750

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.067    0.000      0.000    0.000      -0.020   0.449     0.007     1
                 0.000    0.003    0.000      0.000    0.000      -0.020   0.757     0.232     2
                 0.735    0.563    0.344      0.735    0.469      0.160    0.636     0.458     3
                 0.169    0.011    0.682      0.169    0.270      0.303    0.720     0.331     4
                 0.989    0.092    0.585      0.989    0.735      0.723    0.945     0.580     5
                 0.058    0.008    0.583      0.058    0.105      0.146    0.728     0.297     7
                 0.535    0.000    1.000      0.535    0.697      0.714    0.850     0.606     8
                 0.000    0.000    ?          0.000    ?          ?        0.471     0.007     9
                 0.231    0.007    0.750      0.231    0.353      0.390    0.729     0.299     10
Weighted Avg.    0.425    0.176    ?          0.425    ?          ?        0.738     0.398

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
   0   0   4   0   0   0   0   0   0 |   a = 1
   2   0  90   0   0   2   0   0   0 |   b = 2
  22   1 158   6  25   1   0   0   2 |   c = 3
   8   0  38  15  28   0   0   0   0 |   d = 4
   0   0   1   0  86   0   0   0   0 |   e = 5
   8   0  98   1   4   7   0   0   3 |   f = 7
   5   0  27   0   1   0  38   0   0 |   g = 8
   0   0   4   0   0   0   0   0   0 |   h = 9
   5   1  39   0   3   2   0   0  15 |   i = 10
```

ReliefFAttributeEval with J48



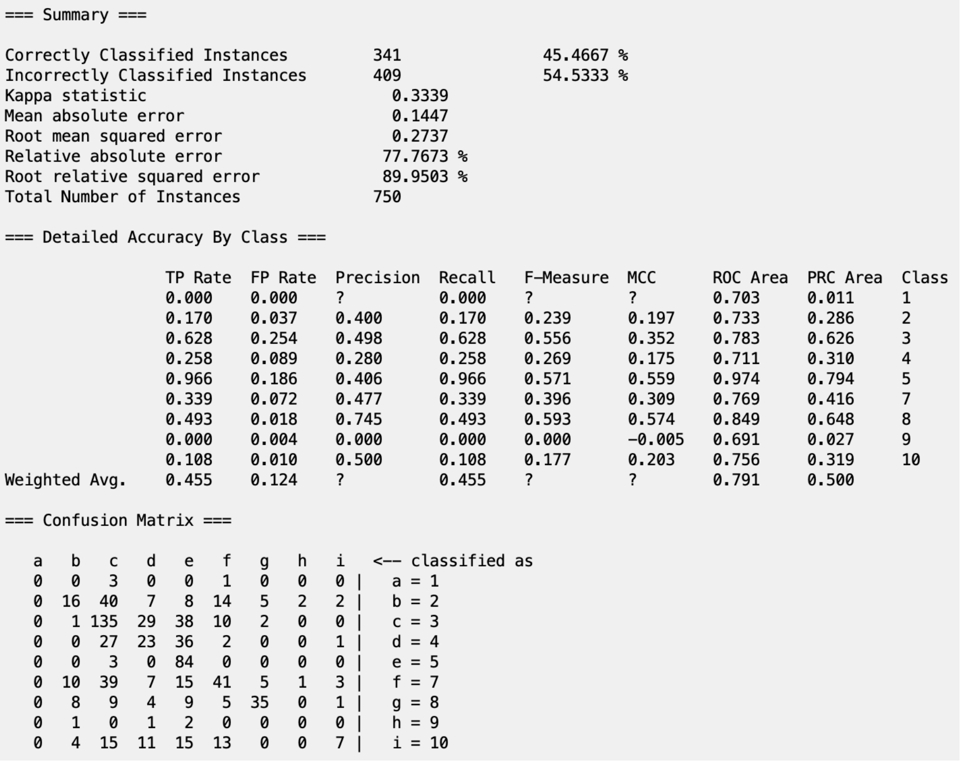
```
=== Summary ===

Correctly Classified Instances         422               56.2667 %
Incorrectly Classified Instances       328               43.7333 %
Kappa statistic                          0.4418
Mean absolute error                      0.1185
Root mean squared error                  0.2506
Relative absolute error                 63.6894 %
Root relative squared error             82.3507 %
Total Number of Instances              750

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.697     0.010     1
                 0.543    0.024    0.761      0.543    0.634      0.601    0.880     0.541     2
                 0.874    0.469    0.428      0.874    0.575      0.372    0.831     0.694     3
                 0.225    0.018    0.625      0.225    0.331      0.331    0.717     0.354     4
                 0.989    0.035    0.789      0.989    0.878      0.867    0.974     0.794     5
                 0.157    0.022    0.576      0.157    0.247      0.242    0.796     0.409     7
                 0.521    0.000    1.000      0.521    0.685      0.704    0.877     0.658     8
                 0.000    0.000    ?          0.000    ?          ?        0.775     0.012     9
                 0.323    0.018    0.636      0.323    0.429      0.419    0.811     0.383     10
Weighted Avg.    0.563    0.149    ?          0.563    ?          ?        0.836     0.562

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
   0   0   4   0   0   0   0   0   0 |   a = 1
   0  51  41   0   0   2   0   0   0 |   b = 2
   0   0 188  12   1   7   0   0   7 |   c = 3
   0   0  45  20  21   2   0   0   1 |   d = 4
   0   0   1   0  86   0   0   0   0 |   e = 5
   0   3  96   0   0  19   0   0   3 |   f = 7
   0  12  21   0   0   0  37   0   1 |   g = 8
   0   0   4   0   0   0   0   0   0 |   h = 9
   0   1  39   0   1   3   0   0  21 |   i = 10
```

ReliefFAttributeEval with Random Forest



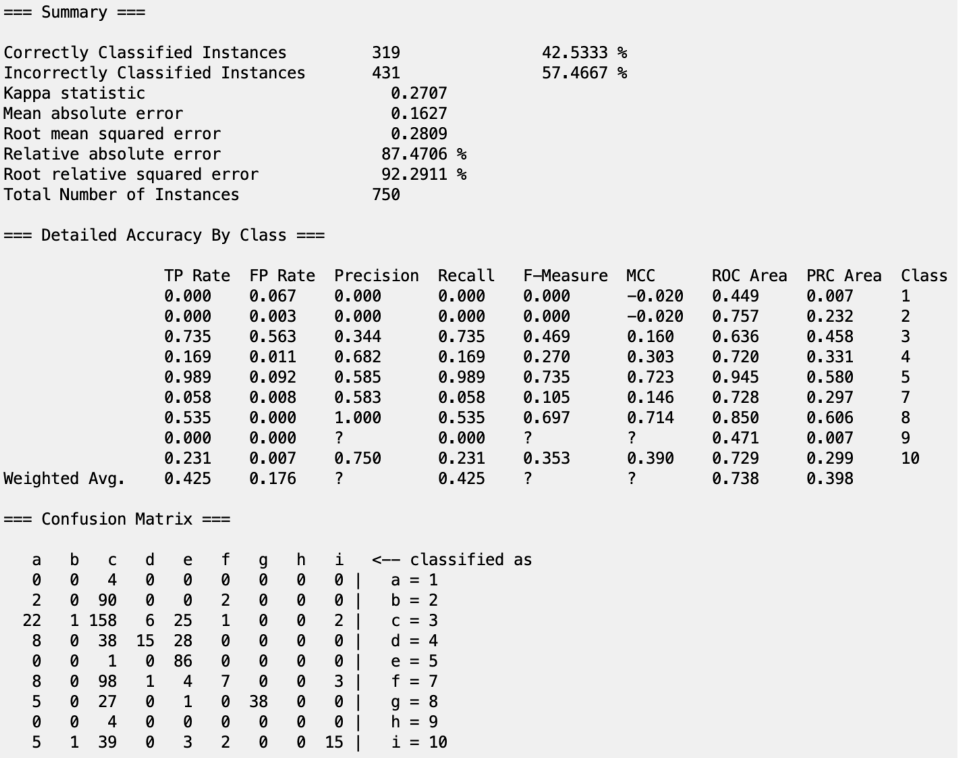
```
=== Summary ===

Correctly Classified Instances         414               55.2    %
Incorrectly Classified Instances       336               44.8    %
Kappa statistic                          0.4323
Mean absolute error                      0.1265
Root mean squared error                  0.2527
Relative absolute error                 67.9971 %
Root relative squared error             83.0299 %
Total Number of Instances              750

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.809     0.016     1
                 0.489    0.020    0.780      0.489    0.601      0.577    0.906     0.598     2
                 0.819    0.443    0.426      0.819    0.561      0.341    0.834     0.714     3
                 0.270    0.050    0.421      0.270    0.329      0.268    0.752     0.331     4
                 0.989    0.039    0.768      0.989    0.864      0.853    0.976     0.784     5
                 0.174    0.029    0.538      0.174    0.263      0.240    0.812     0.459     7
                 0.563    0.001    0.976      0.563    0.714      0.724    0.895     0.733     8
                 0.000    0.001    0.000      0.000    0.000      -0.003   0.630     0.009     9
                 0.323    0.010    0.750      0.323    0.452      0.464    0.855     0.523     10
Weighted Avg.    0.552    0.146    ?          0.552    ?          ?        0.853     0.599

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
   0   0   4   0   0   0   0   0   0 |   a = 1
   0  46  42   2   0   2   1   1   0 |   b = 2
   0   0 176  26   4   6   0   0   3 |   c = 3
   0   0  40  24  20   4   0   0   1 |   d = 4
   0   0   1   0  86   0   0   0   0 |   e = 5
   0   2  92   4   0  21   0   0   2 |   f = 7
   0   9  19   0   1   1  40   0   1 |   g = 8
   0   0   4   0   0   0   0   0   0 |   h = 9
   0   2  35   1   1   5   0   0  21 |   i = 10
```

SymmetricalUncertAttributeEval with Naive Bayes:



```
=== Summary ===

Correctly Classified Instances         341               45.4667 %
Incorrectly Classified Instances       409               54.5333 %
Kappa statistic                          0.3339
Mean absolute error                      0.1447
Root mean squared error                  0.2737
Relative absolute error                 77.7673 %
Root relative squared error             89.9503 %
Total Number of Instances              750

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.703     0.011     1
                 0.170    0.037    0.400      0.170    0.239      0.197    0.733     0.286     2
                 0.628    0.254    0.498      0.628    0.556      0.352    0.783     0.626     3
                 0.258    0.089    0.280      0.258    0.269      0.175    0.711     0.310     4
                 0.966    0.186    0.406      0.966    0.571      0.559    0.974     0.794     5
                 0.339    0.072    0.477      0.339    0.396      0.309    0.769     0.416     7
                 0.493    0.018    0.745      0.493    0.593      0.574    0.849     0.648     8
                 0.000    0.004    0.000      0.000    0.000      -0.005   0.691     0.027     9
                 0.108    0.010    0.500      0.108    0.177      0.203    0.756     0.319     10
Weighted Avg.    0.455    0.124    ?          0.455    ?          ?        0.791     0.500

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
   0   0   3   0   0   1   0   0   0 |   a = 1
   0  16  40   7   8  14   5   2   2 |   b = 2
   0   1 135  29  38  10   2   0   0 |   c = 3
   0   0  27  23  36   2   0   0   1 |   d = 4
   0   0   3   0  84   0   0   0   0 |   e = 5
   0  10  39   7  15  41   5   1   3 |   f = 7
   0   8   9   4   9   5  35   0   1 |   g = 8
   0   1   0   1   2   0   0   0   0 |   h = 9
   0   4  15  11  15  13   0   0   7 |   i = 10
```

SymmetricalUncertAttributeEval with DecisionTable



```
=== Summary ===

Correctly Classified Instances         319               42.5333 %
Incorrectly Classified Instances       431               57.4667 %
Kappa statistic                          0.2707
Mean absolute error                      0.1627
Root mean squared error                  0.2809
Relative absolute error                 87.4706 %
Root relative squared error             92.2911 %
Total Number of Instances              750

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.067    0.000      0.000    0.000      -0.020   0.449     0.007     1
                 0.000    0.003    0.000      0.000    0.000      -0.020   0.757     0.232     2
                 0.735    0.563    0.344      0.735    0.469      0.160    0.636     0.458     3
                 0.169    0.011    0.682      0.169    0.270      0.303    0.720     0.331     4
                 0.989    0.092    0.585      0.989    0.735      0.723    0.945     0.580     5
                 0.058    0.008    0.583      0.058    0.105      0.146    0.728     0.297     7
                 0.535    0.000    1.000      0.535    0.697      0.714    0.850     0.606     8
                 0.000    0.000    ?          0.000    ?          ?        0.471     0.007     9
                 0.231    0.007    0.750      0.231    0.353      0.390    0.729     0.299     10
Weighted Avg.    0.425    0.176    ?          0.425    ?          ?        0.738     0.398

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
   0   0   4   0   0   0   0   0   0 |   a = 1
   2   0  90   0   0   2   0   0   0 |   b = 2
  22   1 158   6  25   1   0   0   2 |   c = 3
   8   0  38  15  28   0   0   0   0 |   d = 4
   0   0   1   0  86   0   0   0   0 |   e = 5
   8   0  98   1   4   7   0   0   3 |   f = 7
   5   0  27   0   1   0  38   0   0 |   g = 8
   0   0   4   0   0   0   0   0   0 |   h = 9
   5   1  39   0   3   2   0   0  15 |   i = 10
```

SymmetricalUncertAttributeEval with J48

```
=== Summary ===

Correctly Classified Instances         360               48      %
Incorrectly Classified Instances       390               52      %
Kappa statistic                          0.3295
Mean absolute error                      0.1365
Root mean squared error                  0.2674
Relative absolute error                 73.3973 %
Root relative squared error             87.8526 %
Total Number of Instances              750

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.748     0.011     1
                 0.000    0.002    0.000      0.000    0.000      -0.014   0.777     0.260     2
                 0.823    0.570    0.367      0.823    0.508      0.239    0.722     0.586     3
                 0.225    0.014    0.690      0.225    0.339      0.354    0.754     0.365     4
                 0.989    0.092    0.585      0.989    0.735      0.723    0.945     0.580     5
                 0.132    0.005    0.842      0.132    0.229      0.298    0.748     0.343     7
                 0.577    0.001    0.976      0.577    0.726      0.733    0.866     0.623     8
                 0.000    0.001    0.000      0.000    0.000      -0.003   0.736     0.010     9
                 0.308    0.013    0.690      0.308    0.426      0.430    0.777     0.341     10
Weighted Avg.    0.480    0.178    ?          0.480    ?          ?        0.781     0.455

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
   0   0   4   0   0   0   0   0   0 |   a = 1
   0   0  90   0   0   2   1   1   0 |   b = 2
   0   0 177   9  25   0   0   0   4 |   c = 3
   0   0  40  20  28   0   0   0   1 |   d = 4
   0   0   1   0  86   0   0   0   0 |   e = 5
   0   0  98   0   4  16   0   0   3 |   f = 7
   0   0  28   0   1   0  41   0   1 |   g = 8
   0   0   4   0   0   0   0   0   0 |   h = 9
   0   1  40   0   3   1   0   0  20 |   i = 10
```

SymmetricalUncertAttributeEval with Random Forest

```
=== Summary ===

Correctly Classified Instances         354               47.2    %
Incorrectly Classified Instances       396               52.8    %
Kappa statistic                          0.3233
Mean absolute error                      0.1472
Root mean squared error                  0.2715
Relative absolute error                 79.1035 %
Root relative squared error             89.2096 %
Total Number of Instances              750

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.630     0.008     1
                 0.000    0.002    0.000      0.000    0.000      -0.014   0.755     0.262     2
                 0.786    0.536    0.371      0.786    0.504      0.231    0.759     0.610     3
                 0.292    0.067    0.371      0.292    0.327      0.251    0.696     0.273     4
                 0.908    0.059    0.669      0.908    0.771      0.747    0.965     0.720     5
                 0.157    0.025    0.543      0.157    0.244      0.229    0.781     0.412     7
                 0.577    0.001    0.976      0.577    0.726      0.733    0.879     0.704     8
                 0.000    0.001    0.000      0.000    0.000      -0.003   0.774     0.016     9
                 0.308    0.010    0.741      0.308    0.435      0.449    0.785     0.432     10
Weighted Avg.    0.472    0.174    ?          0.472    ?          ?        0.791     0.494

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
   0   0   4   0   0   0   0   0   0 |   a = 1
   0   0  83   4   0   5   1   1   0 |   b = 2
   0   0 169  25  13   4   0   0   4 |   c = 3
   0   0  39  26  21   2   0   0   1 |   d = 4
   0   0   2   5  79   1   0   0   0 |   e = 5
   0   0  95   3   3  19   0   0   1 |   f = 7
   0   0  27   2   0   0  41   0   1 |   g = 8
   0   0   4   0   0   0   0   0   0 |   h = 9
   0   1  33   5   2   4   0   0  20 |   i = 10
```

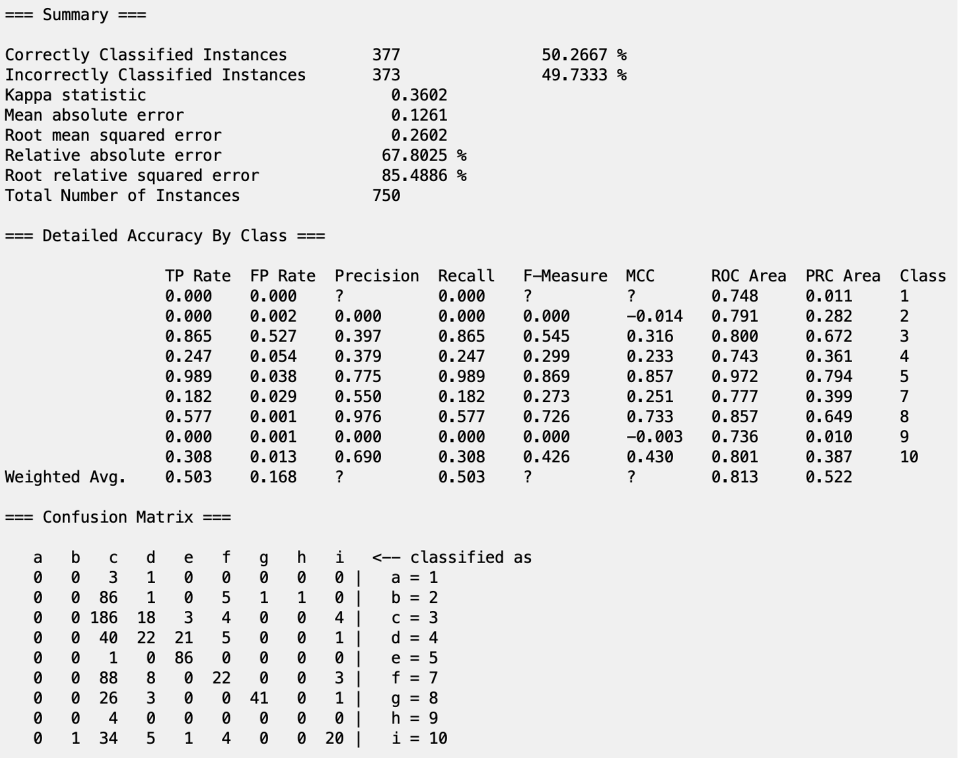
Correlation (Personal Selection) with Naive Bayes:

```
=== Summary ===

Correctly Classified Instances         385               51.3333 %
Incorrectly Classified Instances       365               48.6667 %
Kappa statistic                          0.393 
Mean absolute error                      0.1398
Root mean squared error                  0.2712
Relative absolute error                 75.1386 %
Root relative squared error             89.1296 %
Total Number of Instances              750     

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.892     0.086     1
                 0.170    0.011    0.696      0.170    0.274      0.306    0.789     0.388     2
                 0.795    0.323    0.497      0.795    0.612      0.428    0.807     0.651     3
                 0.281    0.056    0.403      0.281    0.331      0.264    0.647     0.323     4
                 0.989    0.142    0.478      0.989    0.644      0.635    0.979     0.788     5
                 0.322    0.075    0.453      0.322    0.377      0.286    0.783     0.403     7
                 0.507    0.000    1.000      0.507    0.673      0.694    0.844     0.685     8
                 0.000    0.000    ?          0.000    ?          ?        0.709     0.024     9
                 0.185    0.010    0.632      0.185    0.286      0.312    0.765     0.364     10
Weighted Avg.    0.513    0.130    ?          0.513    ?          ?        0.802     0.527     

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
   0   0   3   0   0   1   0   0   0 |   a = 1
   0  16  37   6  14  20   0   0   1 |   b = 2
   0   1 171  14  20   7   0   0   2 |   c = 3
   0   1  34  25  27   2   0   0   0 |   d = 4
   0   0   0   1  86   0   0   0   0 |   e = 5
   0   4  58   7   9  39   0   0   4 |   f = 7
   0   0  13   3  11   8  36   0   0 |   g = 8
   0   0   2   0   1   1   0   0   0 |   h = 9
   0   1  26   6  12   8   0   0  12 |   i = 10
```

Correlation (Personal Selection) with DecisionTable

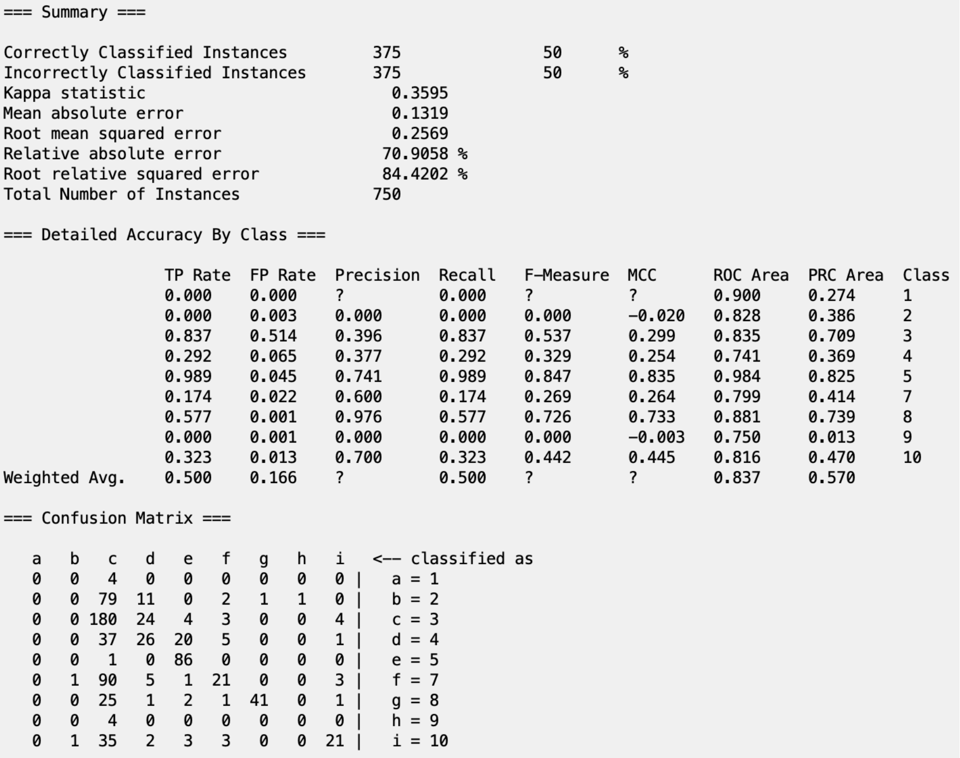
```
=== Summary ===

Correctly Classified Instances         319               42.5333 %
Incorrectly Classified Instances       431               57.4667 %
Kappa statistic                          0.2707
Mean absolute error                      0.1627
Root mean squared error                  0.2809
Relative absolute error                 87.4706 %
Root relative squared error             92.2911 %
Total Number of Instances              750     

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.067    0.000      0.000    0.000      -0.020   0.449     0.007     1
                 0.000    0.003    0.000      0.000    0.000      -0.020   0.757     0.232     2
                 0.735    0.563    0.344      0.735    0.469      0.160    0.636     0.458     3
                 0.169    0.011    0.682      0.169    0.270      0.303    0.720     0.331     4
                 0.989    0.092    0.585      0.989    0.735      0.723    0.945     0.580     5
                 0.058    0.008    0.583      0.058    0.105      0.146    0.728     0.297     7
                 0.535    0.000    1.000      0.535    0.697      0.714    0.850     0.606     8
                 0.000    0.000    ?          0.000    ?          ?        0.471     0.007     9
                 0.231    0.007    0.750      0.231    0.353      0.390    0.729     0.299     10
Weighted Avg.    0.425    0.176    ?          0.425    ?          ?        0.738     0.398     

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
   0   0   4   0   0   0   0   0   0 |   a = 1
   2   0  90   0   0   2   0   0   0 |   b = 2
  22   1 158   6  25   1   0   0   2 |   c = 3
   8   0  38  15  28   0   0   0   0 |   d = 4
   0   0   1   0  86   0   0   0   0 |   e = 5
   8   0  98   1   4   7   0   0   3 |   f = 7
   5   0  27   0   1   0  38   0   0 |   g = 8
   0   0   4   0   0   0   0   0   0 |   h = 9
   5   1  39   0   3   2   0   0  15 |   i = 10
```

Correlation (Personal Selection) with J48



```
=== Summary ===

Correctly Classified Instances         377               50.2667 %
Incorrectly Classified Instances       373               49.7333 %
Kappa statistic                          0.3602
Mean absolute error                      0.1261
Root mean squared error                  0.2602
Relative absolute error                 67.8025 %
Root relative squared error             85.4886 %
Total Number of Instances              750

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.748     0.011     1
                 0.000    0.002    0.000      0.000    0.000      -0.014   0.791     0.282     2
                 0.865    0.527    0.397      0.865    0.545      0.316    0.800     0.672     3
                 0.247    0.054    0.379      0.247    0.299      0.233    0.743     0.361     4
                 0.989    0.038    0.775      0.989    0.869      0.857    0.972     0.794     5
                 0.182    0.029    0.550      0.182    0.273      0.251    0.777     0.399     7
                 0.577    0.001    0.976      0.577    0.726      0.733    0.857     0.649     8
                 0.000    0.001    0.000      0.000    0.000      -0.003   0.736     0.010     9
                 0.308    0.013    0.690      0.308    0.426      0.430    0.801     0.387     10
Weighted Avg.    0.503    0.168    ?          0.503    ?          ?        0.813     0.522

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
   0   0   3   1   0   0   0   0   0 |   a = 1
   0   0  86   1   0   5   1   1   0 |   b = 2
   0   0 186  18   3   4   0   0   4 |   c = 3
   0   0  40  22  21   5   0   0   1 |   d = 4
   0   0   1   0  86   0   0   0   0 |   e = 5
   0   0  88   8   0  22   0   0   3 |   f = 7
   0   0  26   3   0   0  41   0   1 |   g = 8
   0   0   4   0   0   0   0   0   0 |   h = 9
   0   1  34   5   1   4   0   0  20 |   i = 10
```

Correlation (Personal Selection) with Random Forest



```
=== Summary ===

Correctly Classified Instances         375               50      %
Incorrectly Classified Instances       375               50      %
Kappa statistic                          0.3595
Mean absolute error                      0.1319
Root mean squared error                  0.2569
Relative absolute error                 70.9058 %
Root relative squared error             84.4202 %
Total Number of Instances              750

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.900     0.274     1
                 0.000    0.003    0.000      0.000    0.000      -0.020   0.828     0.386     2
                 0.837    0.514    0.396      0.837    0.537      0.299    0.835     0.709     3
                 0.292    0.065    0.377      0.292    0.329      0.254    0.741     0.369     4
                 0.989    0.045    0.741      0.989    0.847      0.835    0.984     0.825     5
                 0.174    0.022    0.600      0.174    0.269      0.264    0.799     0.414     7
                 0.577    0.001    0.976      0.577    0.726      0.733    0.881     0.739     8
                 0.000    0.001    0.000      0.000    0.000      -0.003   0.750     0.013     9
                 0.323    0.013    0.700      0.323    0.442      0.445    0.816     0.470     10
Weighted Avg.    0.500    0.166    ?          0.500    ?          ?        0.837     0.570

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
   0   0   4   0   0   0   0   0   0 |   a = 1
   0   0  79  11   0   2   1   1   0 |   b = 2
   0   0 180  24   4   3   0   0   4 |   c = 3
   0   0  37  26  20   5   0   0   1 |   d = 4
   0   0   1   0  86   0   0   0   0 |   e = 5
   0   1  90   5   1  21   0   0   3 |   f = 7
   0   0  25   1   2   1  41   0   1 |   g = 8
   0   0   4   0   0   0   0   0   0 |   h = 9
   0   1  35   2   3   3   0   0  21 |   i = 10
```

Part 5.2 - Analysis

In total, we created 20 different classifier models that each predicted crime codes of recorded crimes in Los Angeles. The classifier models and their accuracies are below:

CfsSubsetEval with Naive Bayes --- 48.53%

CfsSubsetEval with DecisionTable --- 42.53%

CfsSubsetEval with J48 --- 56.13%

CfsSubsetEval with Random Forest --- 47.87%

CorrelationAttributeEval with Naive Bayes --- 47.87%

CorrelationAttributeEval with DecisionTable --- 49.07%

CorrelationAttributeEval with J48 --- 54.80%

CorrelationAttributeEval with Random Forest --- 51.87%

ReliefFAttributeEval with Naive Bayes --- 49.47%

ReliefFAttributeEval with DecisionTable --- 42.53%

ReliefFAttributeEval with J48 --- 56.27%

ReliefFAttributeEval with Random Forest --- 55.20%

SymmetricalUncertAttributeEval with Naive Bayes --- 45.47%

SymmetricalUncertAttributeEval with DecisionTable --- 42.53%

SymmetricalUncertAttributeEval with J48 --- 48.00%

SymmetricalUncertAttributeEval with Random Forest --- 47.20%

Correlation (Personal Selection) with Naive Bayes --- 51.33%

Correlation (Personal Selection) with DecisionTable --- 42.53%

Correlation (Personal Selection) with J48 --- 50.27%

Correlation (Personal Selection) with Random Forest --- 50.00%

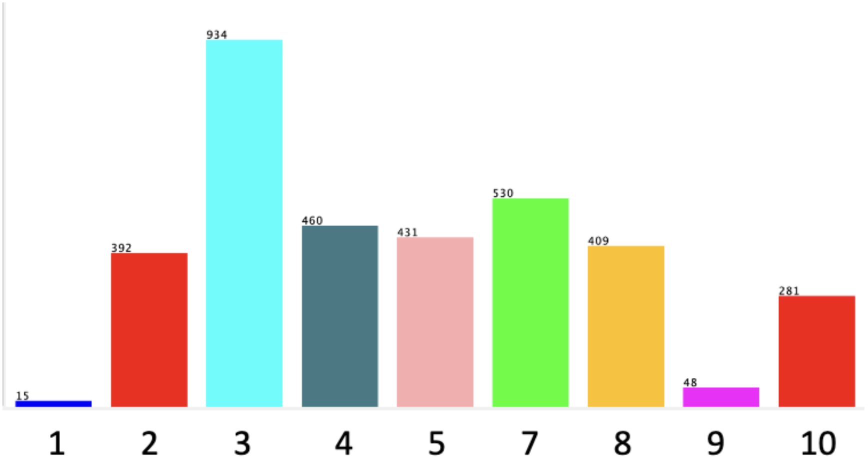
The five models with the highest accuracies are ranked below:

1. ReliefFAttributeEval with J48 --- 56.27%
2. CfsSubsetEval with J48 --- 56.13%
3. ReliefFAttributeEval with Random Forest --- 55.20%
4. CorrelationAttributeEval with J48 --- 54.80%
5. CorrelationAttributeEval with Random Forest --- 51.87%

Although the top accuracies hover around 56%, it is important to note that these accuracies may seem low at first glance but provide valuable insights when analyzed in depth. Crime classification is a complex task due to the inherent overlap and subtle differences between crime codes. In this case, achieving over 50% accuracy with numerous classes and highly imbalanced data is notable.

Further Accuracy Analysis

When we considered cases where predictions were off by only one crime code (i.e., classified one level above or below the actual crime code), the model's effective accuracy increased to **72.1%**. This suggests that while the model might not always predict the exact crime code, it often predicts a similar one, indicating the model's effectiveness in distinguishing between closely related crimes.



Moreover, when we excluded the most common crime classification, code 3, which skews the results due to its frequency (as seen in the graph above, with 3 being the cyan bar), the accuracy rose dramatically to **97.7%**. This shows that a disproportionate amount of the classification errors were concentrated around a single overrepresented crime code, and addressing this imbalance through further techniques like oversampling or undersampling could yield even better results.

In addition, the confusion matrix also shows that the inaccurate results the model predicted were close to the actual value. This is also seen by the classification having a Mean Absolute Error of 0.1185, which is very low and shows consistent grouping.

Precision and Recall

While accuracy gives us an overall picture, evaluating precision and recall offers further insights, especially for imbalanced datasets like this one. In this dataset, less frequent crime codes had much higher precision, indicating that our model is better at predicting less common crime codes (since they are less likely to be predicted incorrectly due to their rarity) but may struggle with recall, particularly for the most common crime codes.

For instance, with our highest accuracy classifier model, ReliefFAttributeEval with J48, crime code 3 exhibited high recall (**0.874**) but low precision (**0.428**), as it misclassified many instances from other classes as crime code 3. Conversely, for rarer crime codes, such as crime code 8, they exhibit higher precision (**1.000**) but lower recall (**0.521**). This is typical in situations where the model is better at identifying certain crime types but fails to catch others with similar frequency.

Future Improvements

Discretization: Further discretizing the data to simplify the classification problem could result in a significant improvement in accuracy. By consolidating similar crime codes or categorizing crimes into broader types, we could reduce the complexity of the problem.

Resampling Techniques: Addressing the imbalanced nature of the dataset through techniques like SMOTE (Synthetic Minority Over-sampling Technique) could improve the model's performance, especially for the less frequent crime codes.

Feature Engineering: Incorporating additional features, such as temporal patterns, geographic trends, or external data sources (e.g., weather conditions, major events), could improve predictive power.

Part 6 – Conclusion/How to Reproduce Our Model

As stated above, the J48 model with ReliefAttributeEval selector had the best results of the 20 runs for this project. We were successfully able to train and test a predictive classification model that predicted the Crime Codes of crimes in Los Angeles and feel confident about those results. Most of the inaccuracies came from predictions that were one off the correct prediction. This shows that at least we are able to have a general idea of how severe the crime was. However, we feel that we definitely have room for improvement. Perhaps a better grouping of attributes would have led to more accurate results. Future projects could lead to further performance analysis of more models as well as better selecting attributes to create models of higher accuracy. We are proud of being able to predict the codes as well as we did.

Steps to Reproduce Our Model: J48 model with ReliefAttributeEval Selector:

1. Open Weka and load the **trainSet.csv** from our directory.
2. Under the Preprocess tab, click Filter > Choose > filters > unsupervised > attribute, select NumerictoNominal.
3. Click on the white space and change attributeIndices to last which is the index of the class variable. Hit Apply.
4. Go to the Select attributes tab and choose the correct class – Crm Cd.
5. Select ReliefFAttributeEval as Attribute Evaluator and Ranker as Search Method and hit Start.
6. Take inventory of the features that have a ratio greater than 0.001 and keep those features and the class on the Preprocess tab. (Select ones to remove and click Remove).
7. Save this train dataset as **train_set.arff** for testing.
8. Repeat steps 1-3 for the **testSet.csv** from the ReliefAttributeEval folder and then limit the selection to the same features from step 6. Save this file as **test_set.arff** for testing.
9. **The above arff files can also be found in the ReliefAttributeEval folder in our directory.**
10. Open Weka Explorer and load the train_set.arff on Preprocess.
11. Click on the Classify tab and click "Supplied test set" under Test Options.
12. Load the test_set.arff and select the correct class – Crm Cd.
13. Select the class – Crm Cd on the Classify tab.
14. Select the J48 model under trees.
15. Click Start. This model can be already be found in our directory along with the other models: **Data/ReliefAttributeEval/Models/J48.model**

Part 7 - Team Members and Tasks Performed

Finding the Data & Building Proposal: Luke Flecker and Tanush Kallem
Preprocessing Initial Attempt: Luke Flecker and Tanush Kallem
Preprocessing & Project Update: Luke Flecker and Tanush Kallem
Non-Weka Attribute Selection Algorithm: Tanush Kallem
Attribute Selection Algorithms and Classifiers: Luke Flecker
Results Output: Luke Flecker and Tanush Kallem
Results Analysis: Luke Flecker and Tanush Kallem
Building Final Report: Luke Flecker and Tanush Kallem

Part 8 – Appendix and Sources

Google Drive Link: **Q1 Project**
Data Source Website: https://catalog.data.gov/dataset/crime-data-from-2020-to-present

References:

Weka Rules: http://weka.sourceforge.net/doc.dev/weka/classifiers/rules/DecisionTable.html