**Assignment #1: UCI Dataset Analysis and Model Development**

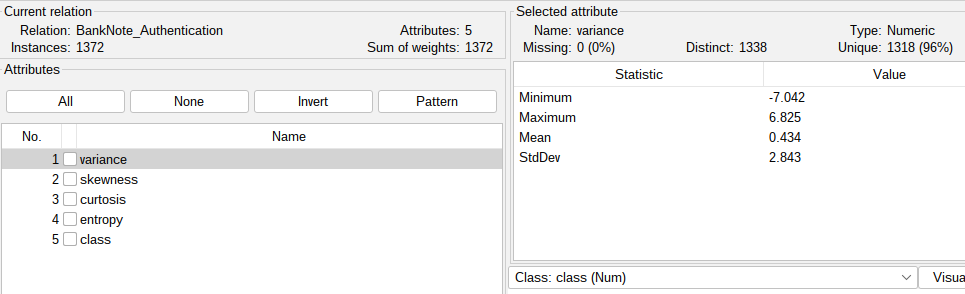
Student Name

Date

**Binary Classification**

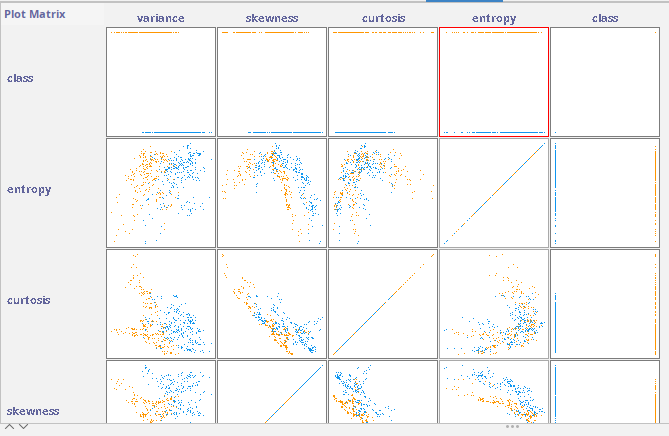
In this task, we aim to explore binary classification using datasets in the UCI Machine Learning Repository. The following report displays the methodology and results for each dataset and task type. The dataset used for the task is Banknote Authentication.

The dataset contains 1,372 cases, each of which represents features extracted from banknote images using wavelet transforms. There are four numeric attributes: variance, skewness, kurtosis, and wavelet-transformed image entropy. The target attribute, class, is binary—authentic banknote (0) or forged banknote (1). The objective is to train a classifier model that can accurately distinguish authentic from counterfeit banknotes based on these statistical features.



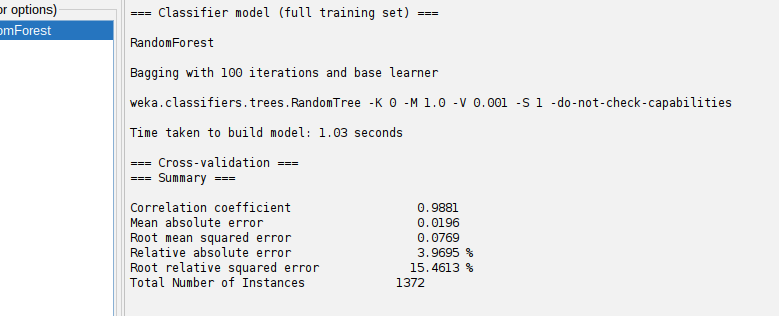
*Figure 1: Banknote dataset loaded in WEKA.*

For data preparation, the dataset was first downloaded from the UCI Machine Learning Repository in CSV format. It was then transformed into ARFF using the internal ArffViewer tool of WEKA so that it becomes compatible with the WEKA environment. The four numeric attributes were preserved in the conversion process, and the class attribute was converted into a nominal attribute with two values: 0 for genuine and 1 for forged. This file in ARFF format was then used for testing and training models in WEKA.



*Figure 2: Visualization of attributes*

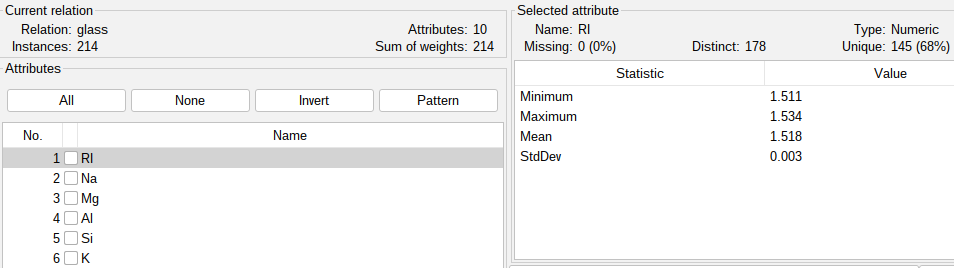
The Random Forest model performed better on the Banknote Authentication dataset. Using 10-fold cross-validation, the model demonstrated a very high correlation coefficient of 0.9881, indicating strong predictive accuracy. The mean absolute error (0.0196) and root mean squared error (0.0769) were both low, indicating minimal deviation from predicted to actual values. These metrics suggest that the model is highly reliable for the binary classification of banknotes as authentic or counterfeit.



*Figure 3: Model results*

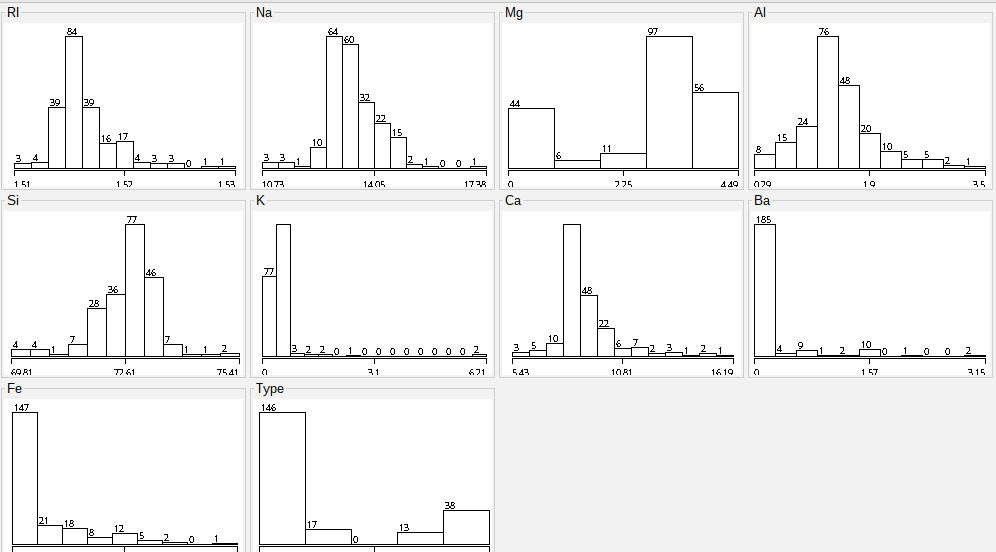
Also, the relative absolute error (3.97%) and the root relative squared error (15.46%) validate the robustness of the model.

**Multi-class classification**

The selected dataset is the UCI Machine Learning Repository's Glass Identification dataset, where the goal is to solve a multi-class classification issue. It is supposed to classify the type of glass correctly based on its physical and chemical properties. It consists of 214 instances and 10 attributes, although the ID attribute was left out for analysis. The other nine attributes are numeric and represent the concentration (weight percent) of various elements in the glass: refractive index (RI), Sodium (Na), Magnesium (Mg), Aluminum (Al), Silicon (Si), Potassium (K), Calcium (Ca), Barium (Ba), and Iron (Fe). 

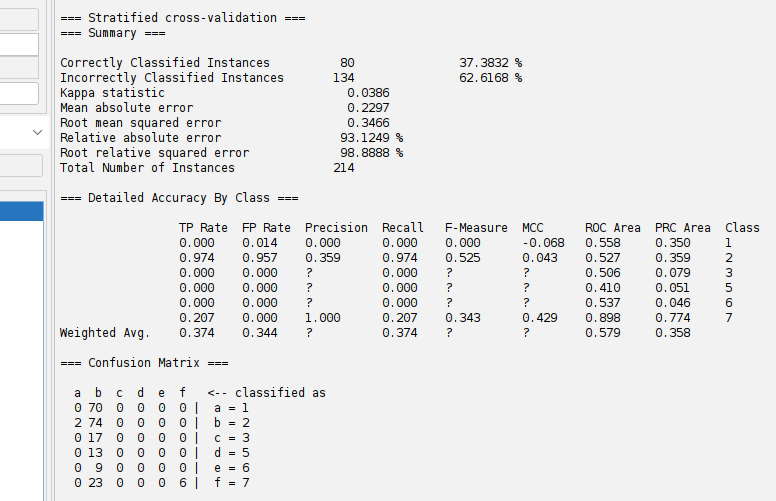
*Figure 4: Glass identification dataset*

The type of glass is the target attribute and a discrete nominal attribute with seven possible classes: building windows (float processed and non-float processed), vehicle windows (float processed), containers, tableware, and headlamps.



*Figure 5: Visualization of attributes*

For data preparation, the Glass Classification data were first converted from CSV to ARFF format to convert it into a format compatible with the WEKA tool. Since the class attribute (glass type) was originally numeric, it was specially converted to nominal to enable correct processing during classification. All other attributes were retained in their numerical format, and the resulting ARFF file was structured with clear attribute types and the class attribute at the end.

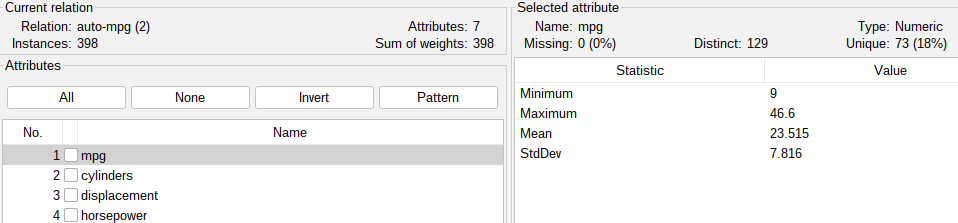


*Figure 6: Model results*

The J48 decision tree classifier was applied to the Glass Classification dataset using 10-fold cross-validation. The model was only 37.38% accurate and correctly classified only 80 instances out of a total of 214. The confusion matrix and full accuracy information indicate that the classifier worked poorly for most of the classes, especially classes 1, 3, 5, and 6, and the true positive rate was 0%, which means that no instances in those classes were correctly classified. Class 2 worked best with a TP rate of 0.974 but at the cost of having extremely high false positives in other classes. The Kappa statistic was very low (0.0386), showing poor agreement between the predictions and actual classifications. The mean absolute error (0.2297) and root mean squared error (0.3466) also bear testimony to the model's limitation.

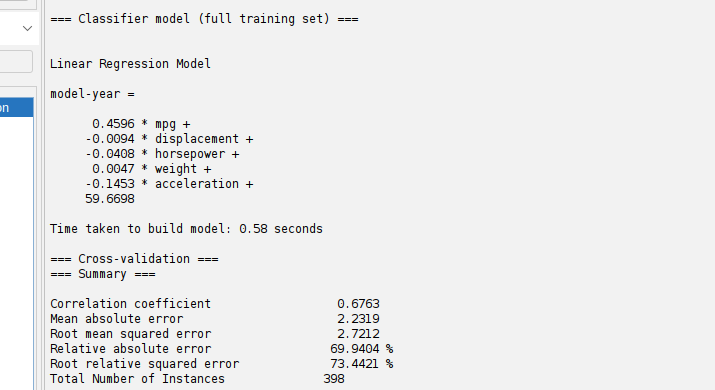
**Estimation (regression)**

The selected regression dataset is the UCI Machine Learning Repository's Auto MPG dataset, which contains specifications of various cars and aims to forecast their gas mileage in miles per gallon (mpg). The data is made up of 398 instances and 9 attributes, 8 of which are input attributes such as cylinders, displacement, horsepower, weight, acceleration, model year, and origin, while the car name is a string identifier and is generally not used in modeling. The target attribute is mpg, which is continuous; hence, this is a regression problem. There are a few missing values for the horsepower attribute, which have to be handled in preprocessing.



*Figure 7: auto dataset*

Missing values in the horsepower feature were handled for the preparation of data by replacing them with the mean to maintain data consistency. The car name feature is instance-specific and non-numeric, so it was eliminated since it does not contribute to the regression problem. The other features were kept in numeric format, and normalization was performed to avoid features of a higher scale dominating the regression model.



*Figure 8: Linear regression model results*

The "car.model.year", which contains the predicted values and the actual values, had a correlation coefficient of 0.6763, indicating that there is a moderate positive relationship between the predicted and actual values. Mean absolute error was 2.2319 and root mean squared error was 2.7212, indicating that the model is a fairly good fit but can be improved upon. The relative absolute error (69.94%) and root relative squared error (73.44%) further support that while the model captures general trends, its predictive accuracy is not perfect.