**Assignment #1**

Student Name

Date

**Assignment #1**

**Binary Classification**

**Dataset Analysis**

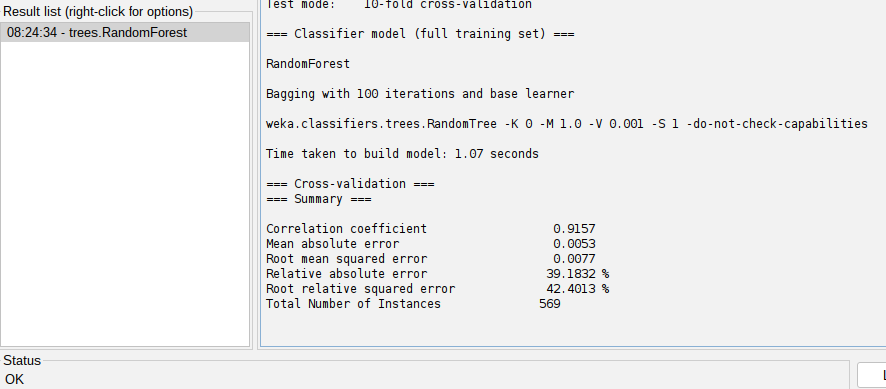
Breast Cancer Wisconsin (Diagnostic) is one of the widely used datasets to apply in binary classification issues where the primary activity is to predict based on several characteristics of cell nuclei about the tumour if it is benign (B) or malignant (M) based on digitized images of fine needle aspirate (FNA) specimens of breast mass. The database consists of 569 cases, all patient cases, and contains 30 real-valued features obtained from each image. The attributes are measurements such as radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, and fractal dimension. For each attribute, three values are calculated: the mean, the standard error, and the worst, which give rise to the 30 input attributes. No values are missing, and all of the attributes are numeric, meaning that the dataset is ready to be processed by classification algorithms without imputation or encoding being needed.

**Data Preparation**

In data preparation, the Breast Cancer Wisconsin (Diagnostic) data set was downloaded in CSV format and loaded into WEKA. Redundant ID feature was discarded, and the class attribute status of the target feature diagnosis was assigned. No subsequent cleaning was performed as features were numeric and the data had no missing values. The purified dataset was subsequently saved in the ARFF (Attribute-Relation File Format) to be compatible with the WEKA tool for future modeling and analysis.

**Model Development**

The screenshot below shows the results of executing a random forest on the data in Weka.



*Figure 1: Weka results*

Using the Random Forest classifier in WEKA with 100 trees and 10-fold cross-validation, the model performed very well in predicting breast cancer diagnosis. The model's correlation coefficient was very high at 0.9157, which suggests a very good relationship between the actual and predicted class labels. The mean absolute error (MAE) was very low at 0.0053, while the root mean squared error (RMSE) was 0.0077, which shows very little prediction error. Despite having 32 features and 569 samples, the model was trained efficiently in a little over one second. The relatively low relative absolute error (39.18%) and root relative squared error (42.40%) also confirm the precision and stability of the model. Overall, the Random Forest model proved to be highly efficient for binary classification in this medical diagnosis data.

**Multi-class Classification**

**Dataset Analysis**

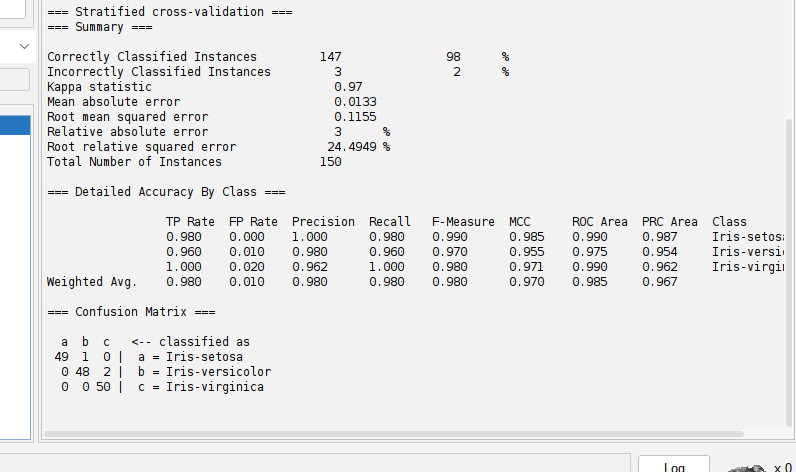
The Iris dataset is a typical benchmark dataset for multi-class classification tasks. It has 150 instances, each of which is an example of an iris flower belonging to one of three species: Iris setosa, Iris versicolor, and Iris virginica. The dataset has four numeric attributes: sepal length, sepal width, petal length, and petal width, all measured in centimeters. All examples are labeled with a target variable for the flower type, therefore, it is a three-class classification problem. The data is balanced since there are 50 samples for each class, and there are no missing values, hence, it is ideal for machine learning to operate on. Originally proposed by Sir Ronald A. Fisher in 1936, this database is still popular for testing classification algorithms since it is easy, well-structured, and possesses nicely separated class boundaries in feature space.

**Data Preparation**

For data preparation, the Iris dataset was loaded directly into the WEKA Explorer because it is already present in .arff format, and there is little preprocessing involved. All four features, sepal length, sepal width, petal length, and petal width, are numeric and clean with no missing values. The target characteristic, class, which identified the iris flower type, was correctly classified as the class attribute by WEKA. Because the dataset was well structured and balanced, there was no need for normalization, transformation, or feature deletion, and it could be used directly to train and test a model.

**Model Development**

The following are the results of the J48 decision tree model in WEKA.

*Figure 2: Decision tree classifier results*

The J48 decision tree classifier performed well on the Iris dataset with a total accuracy of 98%. The model correctly predicted 147 instances out of 150 and misclassified 3 instances, with a Kappa statistic of 0.97, indicating strong agreement between predicted and actual class labels. The tree structure was simple, consisting of only 5 nodes and 3 leaves, making it extremely interpretable. The most important evaluation metrics of precision, recall, and F-measure were also excellent for all classes, with class Iris-setosa achieving a perfect precision and recall of 1.0. The mean absolute error was very small at 0.0133, and the root mean squared error was 0.1155, which further speaks to the strength of the model. The ROC area and PRC area were both close to 1, showing excellent discrimination power for all classes.

**Regression**

**Dataset Analysis**

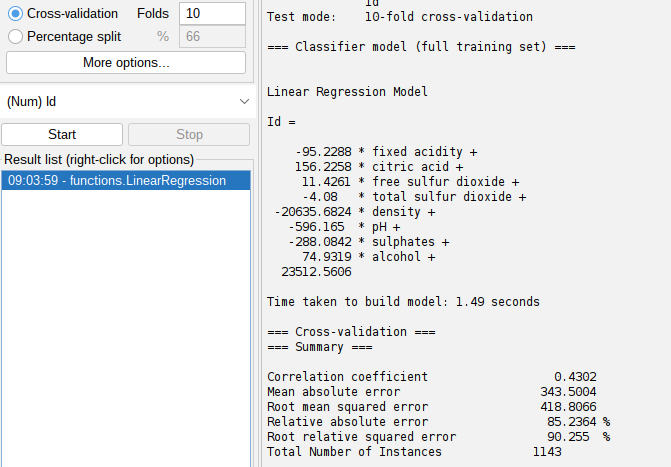
The Wine Quality dataset consists of red wine examples of the Portuguese "Vinho Verde" variety and is utilized for predicting wine quality from measurable physicochemical features. The dataset includes 11 input variables based on lab tests, e.g., acidity, sugar content, sulfur dioxide content, pH, and alcohol levels, and one output variable, quality, which is a 0-10 rating based on sensory evaluation by wine panelists. The data set is a regression problem since it is distributed in an unbalanced manner across classes, with most wines falling in the middle and fewer at high or low quality. Although simple, the problem is complex due to the nonlinear relationships and limited sample size. This makes it an ideal dataset for checking the efficacy of machine learning algorithms in capturing fine variations in product quality.

**Data Preparation**

The data was first cleaned and examined for missing or inconsistent values. The data, originally in CSV format, was then converted into ARFF (Attribute-Relation File Format) form that is compatible with WEKA. All attributes were assigned their respective numeric types, and the target variable, quality, was retained in numeric form for regression analysis. The ARFF file generated was then imported into WEKA to create the model.

**Model Development**

Below are the results of the Linear Regression model.



*Figure 3: Linear Regression results*

The correlation coefficient of 0.4302 of the linear regression model developed for the Wine Quality dataset in WEKA reflected a moderate positive relationship between actual and predicted wine quality ratings. This shows that the model captures some patterns within the data, but not with the precision necessary to generate very precise predictions. Root Mean Squared Error (RMSE) was 418.81 and Mean Absolute Error (MAE) was 343.50, both of which indicate a large margin of error in the model's estimations. The Relative Absolute Error and Root Relative Squared Error were 85.24% and 90.26%, respectively, showing that the model's predictions are away from the actual values. These findings suggest that linear regression alone may not suffice due to the imbalance and complexity of the data and that more advanced or non-linear models would potentially yield better predictive power.