Chi Nguyen

CS 699

Term Project

11/16/17

**Introduction**

Climate changing is one of the most debatable topics nowadays. While many things can contribute to it. CO2 emissions total is one of the most important factors as they caused the greenhouse effect in the Earth. There are several factors that contribute to the CO2 emissions such as the population, energy use, country’s wealth, etc. This project will focus in categorize the CO2 emissions using the CO2\_cat attribute in the climate change data. The categorize will be determined based on attributes such as “Energy use”, “GDP”, “Population”, “Urban population growth”, “Urban population”.

**Data information**

There are 35 attributes and 233 tuples total in the data.

**Country name:** Name of the country

**Agricultural land under irrigation percent of total agricultural land**: The percentage of agricultural land

**Cereal yield (kg/hectare**): The amount of cereal products such as rice, wheat, etc., yield

**Foreign direct investment net inflows percent of GDP**: How much investment from foreign countries the country receives

**Energy use per units of GDP**: How much energy is used in term of GDP

**Energy use per capita**: How much energy is used per capita

**CO2 emissions total**: The total amount of CO2 emission in one country

**CO­2 emission per capita**: The amount of CO2 emission per capita in one country

**GHG net emission**: The amount of greenhouse gas that has been removed by government

**Population in urban agglomerations**: The percentage of populations that live in city, town per country

**Nationally terrestrial protected area**: The protected area in the whole country

**Invest in energy with private participation**: The amount of energy that is not controlled by the government

**Invest in telecoms with private participation**: The amount of telecoms that owned by private

**Invest in transport with private participation**: The amount of transports that is controlled by private

**Invest in water sanity with private participation**: The amount of clean water that is responsible by private

**Public sector management institutions**: The sections that are managed by government

**Paved roads**: How many roads have been built in the country

**GDP**: Gross Domestic Product, the value of all the products by people or companies

**GNI per capita**: the total domestic and foreign output claimed by residents of a country

**Ratio of girls to boys in primary, secondary school percent**: The ratio between boys and girls in K-12

**Primary completion rate total**: The total number of people finished K-12

**Under five mortality rate per 1000**: The mortality rate for children under 5

**Access to improved water source**: The percentage of population that gets access to clean water

**Malaria incidence rate per 10000**: The rate of malaria per 10,000 people

**Access to improved sanitation**: The percentage of population that gets access to clean areas

**Child malnutrition underweight**: The percentage of children under 5 who are malnutrition

**Population living below 1.25 a day**: The percentage of population who are in poverty

**Population growth annual**: The percentage growth of population per year

**Population:** The total population of one country

**Urban population growth**: The growth percentage of population in urban areas

**Urban population**: The amount of people who live in urban area

**CO2\_cat:** CO2 categorize based on the amount of CO2 emission a country emits per year

**Data mining tools and algorithms**

**Attribute selection algorithms:**

Info Gain Attribute Evaluation

Info Gain Attribute Evaluation measure how each feature contributes in decreasing the overall entropy. A good attribute is an attribute that contains the most information or reduces the most the entropy.

Gain Ratio Attribute Evaluation

Evaluate the worth of an attribute by measuring the gain ratio with respect to the class.

Correlation Attribute Evaluation

Evaluate the worth of an attribute by measuring the correlation (Pearson’s) between it and the class.

One R Attribute Evaluation

Evaluate the worth of an attribute by using the OneR classifier.

Wrapper Subset Evaluation

Evaluate attribute sets by using a learning scheme. Cross validation is used to estimate the accuracy of the learning scheme for a set of attributes.

**Classifier algorithms:**

Naive Bayes

The Naïve Bayes classifier performs probabilistic prediction based on the probability values of class membership. Given a data sample, XX, whose class label is unknown, the classifier computes the probability of XX belonging to a class CC for each of the class labels. The classifier then predicts that XX belongs to the class with the highest probability.

Random Trees

Class for constructing a tree that considers K randomly chosen attributes at each node. Performs no pruning. Also has an option to allow estimation of class probabilities (or target mean in the regression case) based on a hold-out set (backfitting).

J48

Class for generating a pruned or unpruned C4.5 decision tree

Decision Table

Decision trees are a classification technique which builds a predictive model in the form of a tree. Each branch of the tree is based on a decision on one of the attributes.

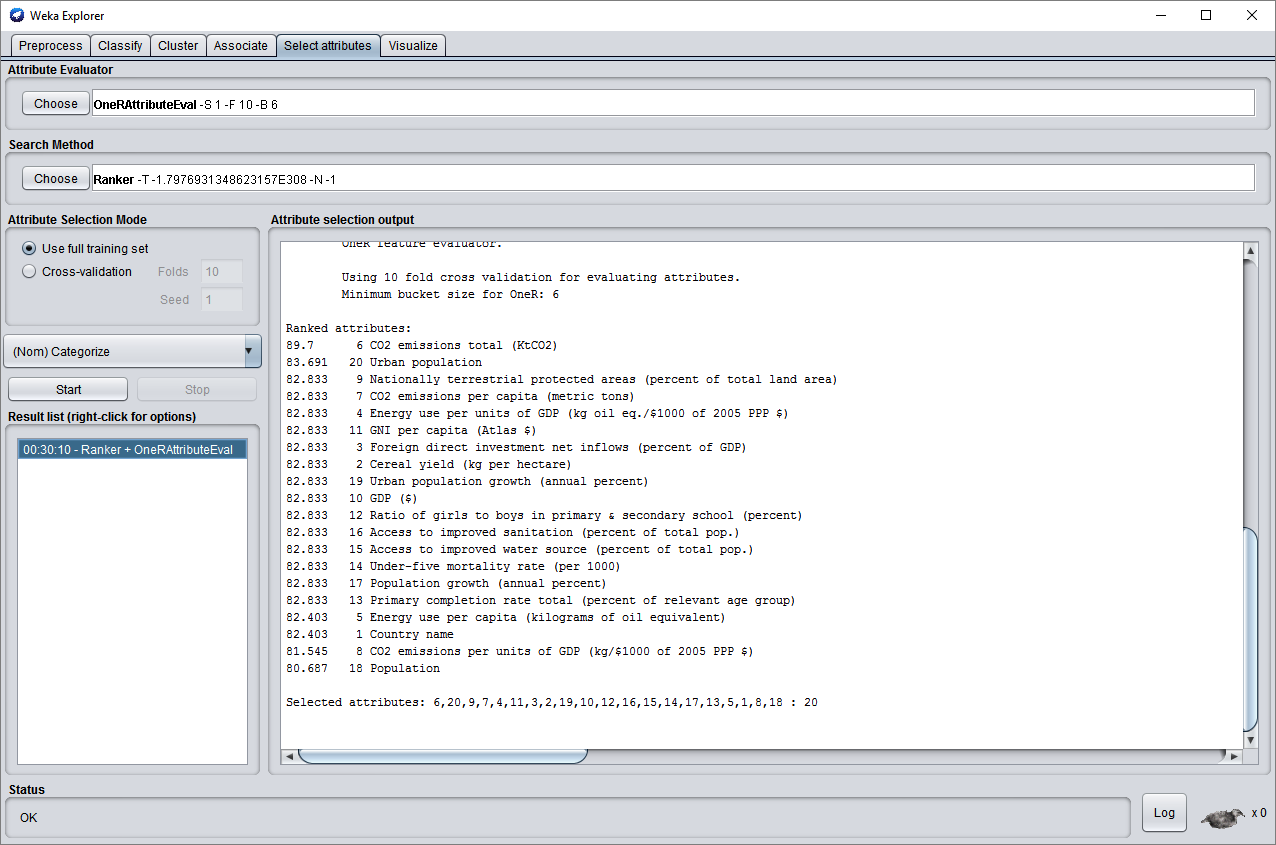
**Data Preprocessing**

Remove attributes that have missing data more than 50% because it will contribute nothing to the data and will create noise.

The categorize will be divided into 5 sections: Very low, Low, Medium, High, Very High, with the range between 0, 200,000, 500,000, 1,000,000, 3,000,000, 10,000,000.

**Data mining result and evaluation**

**Info Gain Attribute Evaluation**



**Naive Bayes**

=== Summary ===

Correctly Classified Instances 135 57.9399 %

Incorrectly Classified Instances 98 42.0601 %

Kappa statistic 0.1353

Mean absolute error 0.1686

Root mean squared error 0.367

Relative absolute error 133.7007 %

Root relative squared error 148.8022 %

Total Number of Instances 233

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.648 0.275 0.919 0.648 0.760 0.285 0.776 0.941 Very Low

0.286 0.302 0.086 0.286 0.132 -0.010 0.516 0.104 Low

0.333 0.062 0.125 0.333 0.182 0.170 0.772 0.264 Medium

0.000 0.004 0.000 0.000 0.000 -0.012 0.820 0.301 High

0.333 0.035 0.200 0.333 0.250 0.233 0.593 0.119 Very High

Weighted Avg. 0.579 0.258 0.777 0.579 0.652 0.245 0.749 0.808

**Random Trees**

=== Summary ===

Correctly Classified Instances 201 86.2661 %

Incorrectly Classified Instances 32 13.7339 %

Kappa statistic 0.4852

Mean absolute error 0.0805

Root mean squared error 0.2167

Relative absolute error 63.8721 %

Root relative squared error 87.8533 %

Total Number of Instances 233

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.969 0.425 0.917 0.969 0.942 0.621 0.831 0.934 Very Low

0.476 0.033 0.588 0.476 0.526 0.488 0.746 0.430 Low

0.000 0.013 0.000 0.000 0.000 -0.019 0.564 0.041 Medium

0.286 0.009 0.500 0.286 0.364 0.364 0.747 0.384 High

0.333 0.013 0.400 0.333 0.364 0.350 0.790 0.430 Very High

Weighted Avg. 0.863 0.356 0.838 0.863 0.848 0.578 0.813 0.836

**J48**

=== Summary ===

Correctly Classified Instances 219 93.9914 %

Incorrectly Classified Instances 14 6.0086 %

Kappa statistic 0.7934

Mean absolute error 0.0363

Root mean squared error 0.1522

Relative absolute error 28.8177 %

Root relative squared error 61.7273 %

Total Number of Instances 233

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.984 0.175 0.964 0.984 0.974 0.845 0.912 0.957 Very Low

0.714 0.009 0.882 0.714 0.789 0.776 0.859 0.778 Low

0.667 0.004 0.800 0.667 0.727 0.724 0.819 0.642 Medium

0.857 0.009 0.750 0.857 0.800 0.795 0.918 0.589 High

0.667 0.009 0.667 0.667 0.667 0.658 0.698 0.386 Very High

Weighted Avg. 0.940 0.146 0.939 0.940 0.938 0.829 0.900 0.907

**Decision Table**

=== Summary ===

Correctly Classified Instances 223 95.7082 %

Incorrectly Classified Instances 10 4.2918 %

Kappa statistic 0.8522

Mean absolute error 0.06

Root mean squared error 0.145

Relative absolute error 47.5984 %

Root relative squared error 58.7981 %

Total Number of Instances 233

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.990 0.150 0.970 0.990 0.979 0.876 0.943 0.979 Very Low

0.810 0.009 0.895 0.810 0.850 0.837 0.933 0.731 Low

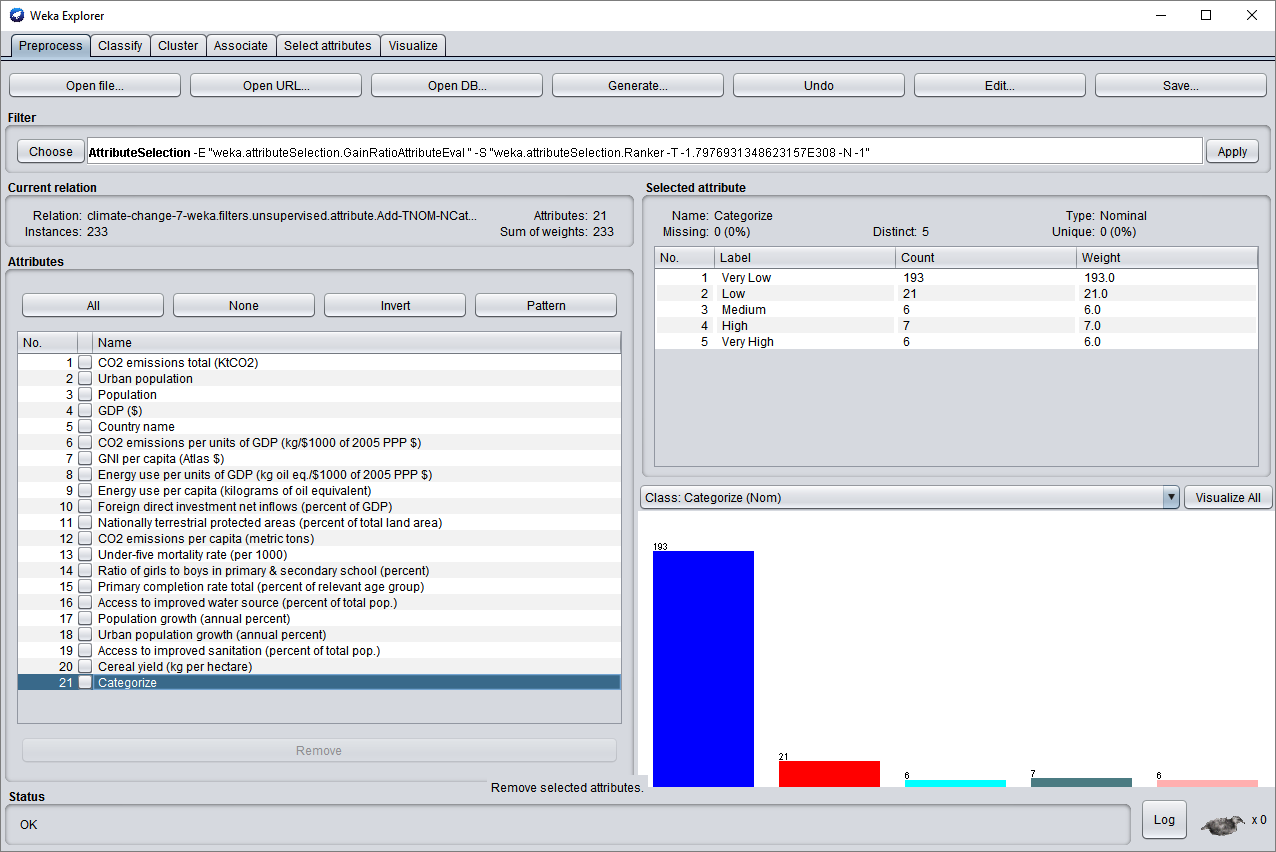
0.833 0.000 1.000 0.833 0.909 0.911 0.957 0.849 Medium

0.857 0.004 0.857 0.857 0.857 0.853 0.987 0.775 High

0.667 0.004 0.800 0.667 0.727 0.724 0.834 0.570 Very High

Weighted Avg. 0.957 0.125 0.956 0.957 0.956 0.869 0.941 0.937

**Gain Ratio Attribute Evaluation**



**Naive Bayes**

=== Summary ===

Correctly Classified Instances 135 57.9399 %

Incorrectly Classified Instances 98 42.0601 %

Kappa statistic 0.1353

Mean absolute error 0.1686

Root mean squared error 0.367

Relative absolute error 133.7007 %

Root relative squared error 148.8022 %

Total Number of Instances 233

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.648 0.275 0.919 0.648 0.760 0.285 0.776 0.941 Very Low

0.286 0.302 0.086 0.286 0.132 -0.010 0.516 0.104 Low

0.333 0.062 0.125 0.333 0.182 0.170 0.772 0.264 Medium

0.000 0.004 0.000 0.000 0.000 -0.012 0.820 0.301 High

0.333 0.035 0.200 0.333 0.250 0.233 0.593 0.119 Very High

Weighted Avg. 0.579 0.258 0.777 0.579 0.652 0.245 0.749 0.808

**Random Trees**

=== Summary ===

Correctly Classified Instances 190 81.5451 %

Incorrectly Classified Instances 43 18.4549 %

Kappa statistic 0.2715

Mean absolute error 0.0949

Root mean squared error 0.2387

Relative absolute error 75.2918 %

Root relative squared error 96.7738 %

Total Number of Instances 233

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.948 0.625 0.880 0.948 0.913 0.394 0.742 0.910 Very Low

0.190 0.057 0.250 0.190 0.216 0.152 0.640 0.175 Low

0.000 0.004 0.000 0.000 0.000 -0.011 0.705 0.094 Medium

0.143 0.004 0.500 0.143 0.222 0.256 0.821 0.326 High

0.333 0.018 0.333 0.333 0.333 0.316 0.762 0.233 Very High

Weighted Avg. 0.815 0.524 0.775 0.815 0.791 0.355 0.735 0.787

**J48**

=== Summary ===

Correctly Classified Instances 219 93.9914 %

Incorrectly Classified Instances 14 6.0086 %

Kappa statistic 0.7934

Mean absolute error 0.0363

Root mean squared error 0.1522

Relative absolute error 28.8177 %

Root relative squared error 61.7273 %

Total Number of Instances 233

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.984 0.175 0.964 0.984 0.974 0.845 0.912 0.957 Very Low

0.714 0.009 0.882 0.714 0.789 0.776 0.859 0.768 Low

0.667 0.004 0.800 0.667 0.727 0.724 0.819 0.642 Medium

0.857 0.009 0.750 0.857 0.800 0.795 0.918 0.589 High

0.667 0.009 0.667 0.667 0.667 0.658 0.698 0.386 Very High

Weighted Avg. 0.940 0.146 0.939 0.940 0.938 0.829 0.900 0.906

**Decision Table**

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 223 95.7082 %

Incorrectly Classified Instances 10 4.2918 %

Kappa statistic 0.8522

Mean absolute error 0.06

Root mean squared error 0.145

Relative absolute error 47.5984 %

Root relative squared error 58.7981 %

Total Number of Instances 233

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.990 0.150 0.970 0.990 0.979 0.876 0.943 0.979 Very Low

0.810 0.009 0.895 0.810 0.850 0.837 0.933 0.731 Low

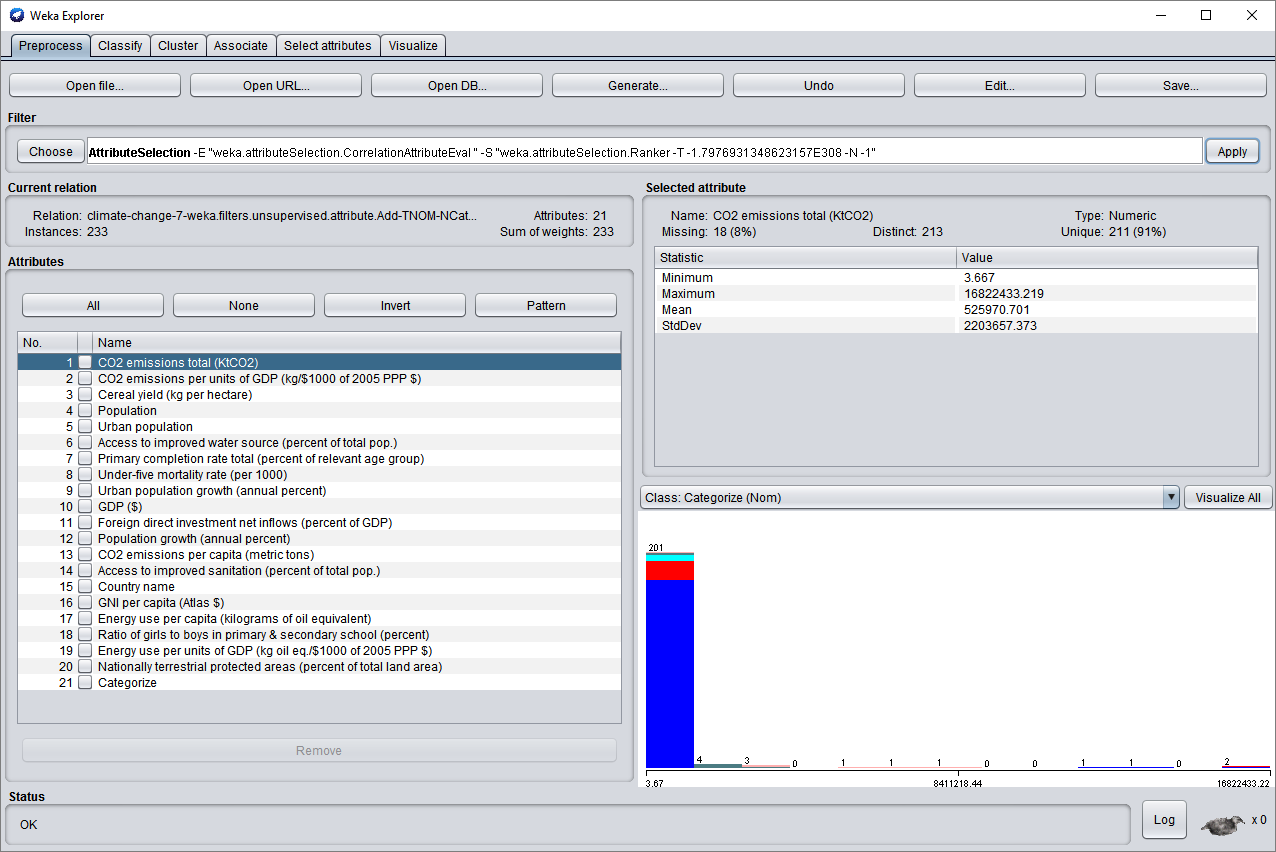
0.833 0.000 1.000 0.833 0.909 0.911 0.957 0.849 Medium

0.857 0.004 0.857 0.857 0.857 0.853 0.987 0.775 High

0.667 0.004 0.800 0.667 0.727 0.724 0.834 0.570 Very High

Weighted Avg. 0.957 0.125 0.956 0.957 0.956 0.869 0.941 0.937

**Correlation Attribute Evaluation**



**Naive Bayes**

=== Summary ===

Correctly Classified Instances 135 57.9399 %

Incorrectly Classified Instances 98 42.0601 %

Kappa statistic 0.1353

Mean absolute error 0.1686

Root mean squared error 0.367

Relative absolute error 133.7007 %

Root relative squared error 148.8022 %

Total Number of Instances 233

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.648 0.275 0.919 0.648 0.760 0.285 0.776 0.941 Very Low

0.286 0.302 0.086 0.286 0.132 -0.010 0.516 0.104 Low

0.333 0.062 0.125 0.333 0.182 0.170 0.772 0.264 Medium

0.000 0.004 0.000 0.000 0.000 -0.012 0.820 0.301 High

0.333 0.035 0.200 0.333 0.250 0.233 0.593 0.119 Very High

Weighted Avg. 0.579 0.258 0.777 0.579 0.652 0.245 0.749 0.808

**Random Trees**

=== Summary ===

Correctly Classified Instances 200 85.8369 %

Incorrectly Classified Instances 33 14.1631 %

Kappa statistic 0.4471

Mean absolute error 0.0868

Root mean squared error 0.218

Relative absolute error 68.7967 %

Root relative squared error 88.3974 %

Total Number of Instances 233

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.969 0.500 0.903 0.969 0.935 0.562 0.861 0.958 Very Low

0.429 0.042 0.500 0.429 0.462 0.414 0.825 0.369 Low

0.167 0.004 0.500 0.167 0.250 0.279 0.737 0.144 Medium

0.286 0.013 0.400 0.286 0.333 0.321 0.885 0.268 High

0.167 0.000 1.000 0.167 0.286 0.404 0.732 0.293 Very High

Weighted Avg. 0.858 0.419 0.844 0.858 0.840 0.530 0.852 0.846

**J48**

=== Summary ===

Correctly Classified Instances 219 93.9914 %

Incorrectly Classified Instances 14 6.0086 %

Kappa statistic 0.7934

Mean absolute error 0.0363

Root mean squared error 0.1522

Relative absolute error 28.8177 %

Root relative squared error 61.7273 %

Total Number of Instances 233

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.984 0.175 0.964 0.984 0.974 0.845 0.912 0.957 Very Low

0.714 0.009 0.882 0.714 0.789 0.776 0.859 0.778 Low

0.667 0.004 0.800 0.667 0.727 0.724 0.819 0.642 Medium

0.857 0.009 0.750 0.857 0.800 0.795 0.918 0.589 High

0.667 0.009 0.667 0.667 0.667 0.658 0.698 0.386 Very High

Weighted Avg. 0.940 0.146 0.939 0.940 0.938 0.829 0.900 0.907

**Decision Table**

=== Summary ===

Correctly Classified Instances 223 95.7082 %

Incorrectly Classified Instances 10 4.2918 %

Kappa statistic 0.8522

Mean absolute error 0.06

Root mean squared error 0.145

Relative absolute error 47.5984 %

Root relative squared error 58.7981 %

Total Number of Instances 233

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.990 0.150 0.970 0.990 0.979 0.876 0.943 0.979 Very Low

0.810 0.009 0.895 0.810 0.850 0.837 0.933 0.731 Low

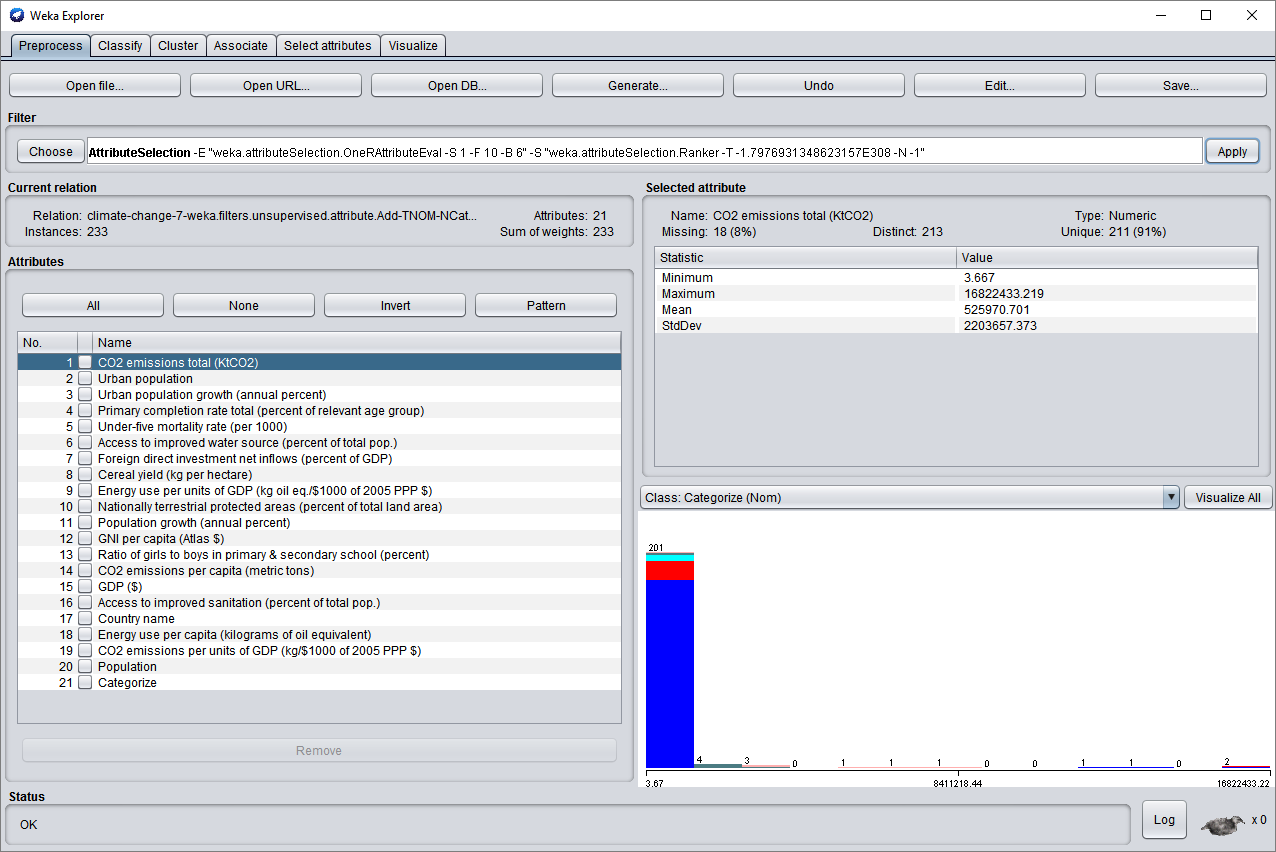
0.833 0.000 1.000 0.833 0.909 0.911 0.957 0.849 Medium

0.857 0.004 0.857 0.857 0.857 0.853 0.987 0.775 High

0.667 0.004 0.800 0.667 0.727 0.724 0.834 0.570 Very High

Weighted Avg. 0.957 0.125 0.956 0.957 0.956 0.869 0.941 0.937

**One R Attribute Evaluation**



**Naive Bayes**

=== Summary ===

Correctly Classified Instances 135 57.9399 %

Incorrectly Classified Instances 98 42.0601 %

Kappa statistic 0.1353

Mean absolute error 0.1686

Root mean squared error 0.367

Relative absolute error 133.7007 %

Root relative squared error 148.8022 %

Total Number of Instances 233

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.648 0.275 0.919 0.648 0.760 0.285 0.776 0.941 Very Low

0.286 0.302 0.086 0.286 0.132 -0.010 0.516 0.104 Low

0.333 0.062 0.125 0.333 0.182 0.170 0.772 0.264 Medium

0.000 0.004 0.000 0.000 0.000 -0.012 0.820 0.301 High

0.333 0.035 0.200 0.333 0.250 0.233 0.593 0.119 Very High

Weighted Avg. 0.579 0.258 0.777 0.579 0.652 0.245 0.749 0.808

**Random Trees**

=== Summary ===

Correctly Classified Instances 192 82.4034 %

Incorrectly Classified Instances 41 17.5966 %

Kappa statistic 0.2124

Mean absolute error 0.1015

Root mean squared error 0.2415

Relative absolute error 80.5009 %

Root relative squared error 97.9013 %

Total Number of Instances 233

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.974 0.725 0.866 0.974 0.917 0.371 0.726 0.907 Very Low

0.190 0.038 0.333 0.190 0.242 0.198 0.640 0.202 Low

0.000 0.004 0.000 0.000 0.000 -0.011 0.686 0.123 Medium

0.000 0.009 0.000 0.000 0.000 -0.016 0.771 0.169 High

0.000 0.004 0.000 0.000 0.000 -0.011 0.760 0.138 Very High

Weighted Avg. 0.824 0.604 0.748 0.824 0.781 0.324 0.719 0.781

**J48**

=== Summary ===

Correctly Classified Instances 219 93.9914 %

Incorrectly Classified Instances 14 6.0086 %

Kappa statistic 0.7934

Mean absolute error 0.0363

Root mean squared error 0.1522

Relative absolute error 28.8177 %

Root relative squared error 61.7273 %

Total Number of Instances 233

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.984 0.175 0.964 0.984 0.974 0.845 0.912 0.957 Very Low

0.714 0.009 0.882 0.714 0.789 0.776 0.859 0.768 Low

0.667 0.004 0.800 0.667 0.727 0.724 0.819 0.642 Medium

0.857 0.009 0.750 0.857 0.800 0.795 0.918 0.589 High

0.667 0.009 0.667 0.667 0.667 0.658 0.698 0.386 Very High

Weighted Avg. 0.940 0.146 0.939 0.940 0.938 0.829 0.900 0.906

**Decision Table**

=== Summary ===

Correctly Classified Instances 224 96.1373 %

Incorrectly Classified Instances 9 3.8627 %

Kappa statistic 0.8685

Mean absolute error 0.0558

Root mean squared error 0.1384

Relative absolute error 44.2205 %

Root relative squared error 56.1076 %

Total Number of Instances 233

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.990 0.125 0.974 0.990 0.982 0.892 0.945 0.980 Very Low

0.810 0.009 0.895 0.810 0.850 0.837 0.931 0.728 Low

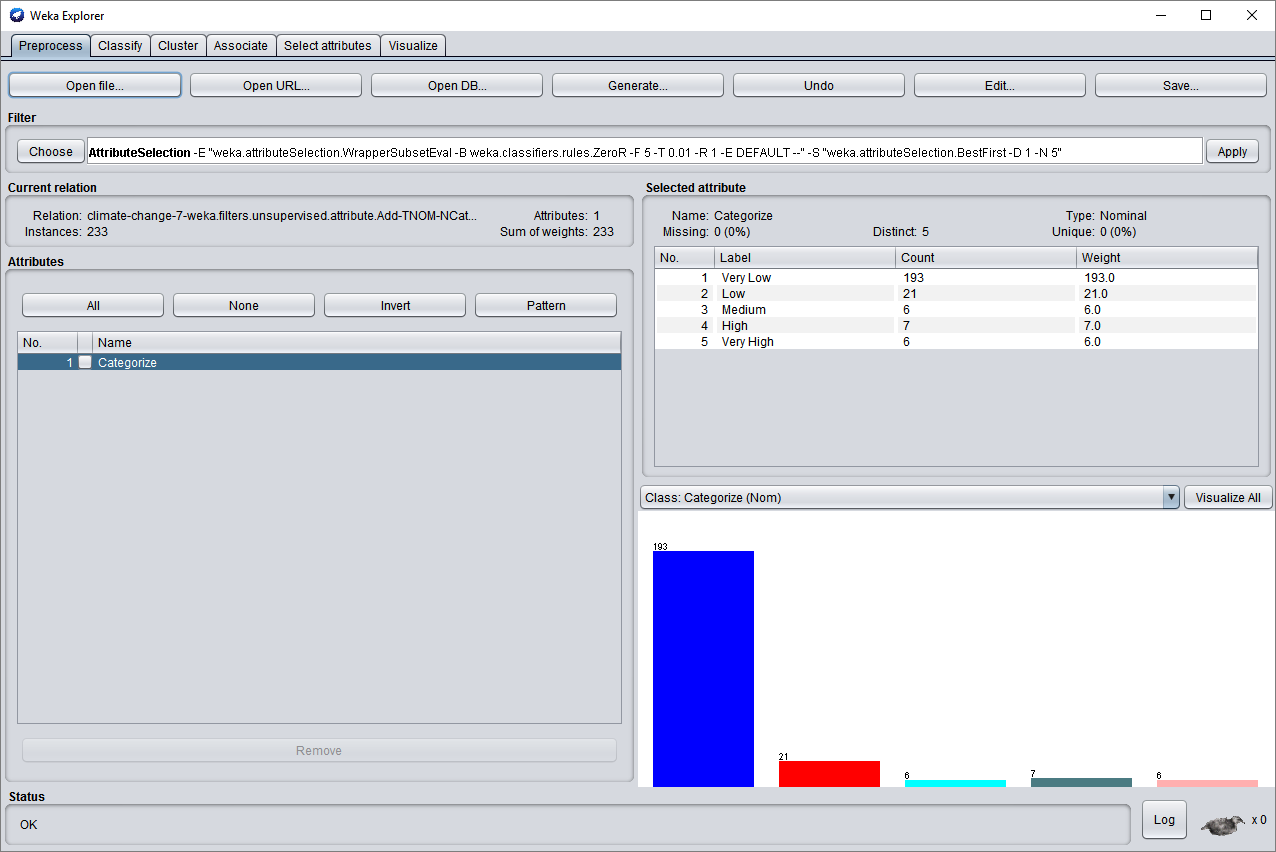
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 Medium

0.857 0.004 0.857 0.857 0.857 0.853 0.989 0.780 High

0.667 0.004 0.800 0.667 0.727 0.724 0.837 0.576 Very High

Weighted Avg. 0.961 0.105 0.960 0.961 0.960 0.884 0.944 0.941

**Wrapper Subset Evaluation**



**Naive Bayes**

=== Summary ===

Correctly Classified Instances 193 82.8326 %

Incorrectly Classified Instances 40 17.1674 %

Kappa statistic 0

Mean absolute error 0.1261

Root mean squared error 0.2466

Relative absolute error 100 %

Root relative squared error 100 %

Total Number of Instances 233

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

1.000 1.000 0.828 1.000 0.906 0.000 0.495 0.827 Very Low

0.000 0.000 0.000 0.000 0.000 0.000 0.472 0.086 Low

0.000 0.000 0.000 0.000 0.000 0.000 0.291 0.026 Medium

0.000 0.000 0.000 0.000 0.000 0.000 0.344 0.026 High

0.000 0.000 0.000 0.000 0.000 0.000 0.294 0.022 Very High

Weighted Avg. 0.828 0.828 0.686 0.828 0.751 0.000 0.478 0.695

**Random Trees**

=== Summary ===

Correctly Classified Instances 193 82.8326 %

Incorrectly Classified Instances 40 17.1674 %

Kappa statistic 0

Mean absolute error 0.1261

Root mean squared error 0.2466

Relative absolute error 100 %

Root relative squared error 100 %

Total Number of Instances 233

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

1.000 1.000 0.828 1.000 0.906 0.000 0.495 0.827 Very Low

0.000 0.000 0.000 0.000 0.000 0.000 0.472 0.086 Low

0.000 0.000 0.000 0.000 0.000 0.000 0.291 0.026 Medium

0.000 0.000 0.000 0.000 0.000 0.000 0.344 0.026 High

0.000 0.000 0.000 0.000 0.000 0.000 0.294 0.022 Very High

Weighted Avg. 0.828 0.828 0.686 0.828 0.751 0.000 0.478 0.695

**J48**

=== Summary ===

Correctly Classified Instances 193 82.8326 %

Incorrectly Classified Instances 40 17.1674 %

Kappa statistic 0

Mean absolute error 0.1215

Root mean squared error 0.2465

Relative absolute error 96.3334 %

Root relative squared error 99.957 %

Total Number of Instances 233

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

1.000 1.000 0.828 1.000 0.906 0.000 0.495 0.827 Very Low

0.000 0.000 0.000 0.000 0.000 0.000 0.472 0.086 Low

0.000 0.000 0.000 0.000 0.000 0.000 0.291 0.026 Medium

0.000 0.000 0.000 0.000 0.000 0.000 0.344 0.026 High

0.000 0.000 0.000 0.000 0.000 0.000 0.294 0.022 Very High

Weighted Avg. 0.828 0.828 0.686 0.828 0.751 0.000 0.478 0.695

**Decision Table**

=== Summary ===

Correctly Classified Instances 193 82.8326 %

Incorrectly Classified Instances 40 17.1674 %

Kappa statistic 0

Mean absolute error 0.1259

Root mean squared error 0.2466

Relative absolute error 99.8691 %

Root relative squared error 99.9928 %

Total Number of Instances 233

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

1.000 1.000 0.828 1.000 0.906 0.000 0.497 0.827 Very Low

0.000 0.000 0.000 0.000 0.000 0.000 0.477 0.086 Low

0.000 0.000 0.000 0.000 0.000 0.000 0.291 0.026 Medium

0.000 0.000 0.000 0.000 0.000 0.000 0.344 0.026 High

0.000 0.000 0.000 0.000 0.000 0.000 0.370 0.024 Very High

Weighted Avg. 0.828 0.828 0.686 0.828 0.751 0.000 0.482 0.695

**Evaluation**

**Info Gain Attribute Evaluation**

The results show that for info gain attribute evaluation, Naïve Bayes has the lowest correctly classified instances with only 58%, while decision table has the highest one with 95.7%. J48 has second highest correctly classified instances with 93.99%. The TP rate between J48 and Decision Table are very close, 0.94 and 0.957, respectively. Random Tree is much lower comparing to the other models with 86.2%, but still much better than Naïve Bayes model. It showed that for Info gain attribute evaluation, the best classify model will be Decision Table, and if decision table is not available, J48 will be suffice.

**Gain Ratio Attribute Evaluation**

Since gain ratio attribute evaluation is the more enhance version of info gain attribute evaluation, it is expected that the models will work in similar manner, or better. Indeed, among 4 classify models, Decision Table still has the highest correctly classified instances. It also has the highest TP rate and lowest FP rate, 0.957 and 0.125, respectively. J48 is still the second highest one, while Naïve Bayes is still lowest. However, with this attribute evaluation, the Random Tree model actually preforms worse with 83.7%, while it was 86% with info gain attribute evaluation. It is also noteworthy that while Random Tree model has higher correctly classified instances than Naïve Bayes, the FP rate of Random Tree model is the highest, even double the FP rate of Naïve Bayes model, despite of the fact that it performs better overall. Therefore, Decision Table and J48 are the optimal choices for this attribute evaluation.

**Correlation Attribute Evaluation**

The correctly classified instances percentages of Decision Table and J48 remain the same with 95.7% and 93.99%, respectively. Naïve Bayes is still the worst model with only 57% of correctly classified instances. With correlation attribute evaluation, Random Tree model improved its correctly classified instances to 85%. While it is still lower than the info gain attribute evaluation, it is better than the gain ratio attribute evaluation. Its FP rate is certainly higher than the other models, however, it still performs relatively well.

**One R Attribute Evaluation**

Once again, the result shows that the correctly classified instances percentages of Decision Table and J48 remain the same. The Naïve Bayes model also remains the same with 57%. The only different is the Random Tree model again. With the correctly percentage of 84.5%, it does relatively better than in the gain ratio attribute evaluation, but still do worse than in info gain and correlation attribute evaluations. Hence, Decision Table and J48 work very well with one R attribute evaluation, while Random Tree can work but not the best, and Naïve Bayes would be the worst one to use.

**Wrapper Subset Evaluation**

Interestingly, for this evaluation, Naïve Bayes’ correct percentage jumps to 82.8%, while J48’s and Decision Table’s percentage drop to 82.8%. Random Tree model is also 82.8%. Since Decision Table has the highest ROC area (0.482), comparing to Naïve Bayes (0.478), Random tree (0.478), and J48 (0.478), it provides the most reliable result. Nevertheless, because the ROC area is much lower overall for all models (on other evaluation attributes, J48 and Decision Table are usually able to reach at least 0.9 for ROC area), this evaluation is not reliable, and should not be used for big data set as it will yield much higher errors than other evaluations because it has much more simple method.

**Conclusion**

Based on the results of 20 classifier models, it shows that Decision Table will give out the best results, J48 is next, and Naïve Bayes is the worst model to use. Any attribute evaluations except wrapper subset evaluation will give very high ROC area, reliable results using Decision Table and J48 models. If Random Tree model is desired, info gain attribute evaluation will be the best attribute evaluation to use. While Decision Table still the best in wrapper subset evaluation, and this evaluation improves Naïve Bayes model immensely, because of its very low ROC area, it is strongly recommended to stay away as it will not provide good results for large dataset.

There are some limitations in this project. For example, it is unclear how the population, GPD, or urban growth affect the classification. In addition, due to many data was missing from the dataset, it might create some noise during classification process. In the future, different attribute evaluations will be picked with same classification models, and different models.

**Citation**

Weka - Data Mining with Open Source Machine Learning Software in Java. (n.d.). Retrieved November 16, 2017, from http://weka.sourceforge.net/