EECS 349 Final Project: Day Zero Predictor
Group 63: Aamir Husain and Ahalya Mandana

Initial Phase: Aggregating the Data

The first step of our project involved building a dataset. Because our goal was to develop a model that incorporates new attributes compared to other models. After doing some initial research, we made a small list of attributes we thought would be useful in classifying a nation's stress level. Fortunately, we did not have to hunt down every data sample. The water consumption and usage database from AQUASTAT proved to be extremely comprehensive, however we were still tasked with consolidating multiple tables for the attributes we were interested in. The following attributes were used to generate our model:

Attribute	Unit	Description
Rainwater Harvesting Awareness	yes/no	determined by whether or not rainwater harvesting is widely practiced
Water Consumption per Capita	m^3/year/inhabitant	Total amount of water withdrawn per capita
Desalination Capacity	km^3/year	Fresh water produced using brackish or salt water
Water Dependency Ratio	%	Percentage of water that comes from other countries
Agricultural Water Withdrawal	%	Percentage of total water withdrawn used for agriculture
Industrial Water Withdrawal	%	Percentage of total water withdrawn used for industrial purposes
Municipal Water Withdrawal	%	Percentage of total water withdrawn used for municipal purposes
Water Stress Level	%	Water stress level measured by dividing total water withdrawal by the total water available minus any water needed for environmental flow. This was used to determine the class label for each sample
Total Land Cultivated	%	Percentage of the total land area of the country that has been cultivated
Annual Precipitation	mm/yr	Total depth of precipitation per year

Total Renewable Water Resources per Capita	m^3/year/inhabitant	The maximum theoretical yearly amount of water available per person for a country at a given moment

Data from our sources were presented in several different formats and contained extra information that was not useful for our application, so we developed data purification scripts to keep only what was necessary. `parse_csv` contains functions to parse raw CSV files from our sources into datasets that are compatible with our machine learning application. All the clean datasets were then merged into one master dataset using `master_gen.py`. This script also handles the best-fit model categorizing the output stress levels of each country into one of 6 classes and splits the whole dataset into testing and training subsets.

AQUASTAT provides nearly all of their values in 5 year increments, however the initial start time for one attribute might be different from another. This led to instances where half the attributes were present for one year and the remaining attributes were present for the year or two years after. As a result, we had large gaps of information for every documented year. To fix this issue, we shifted the years of some attributes so that all attributes would follow the same 5 year time step for each country. For example, an attribute that has a year label of 1991 would be shifted to 1990 unless a value was already present in the 1990 space. This is under the assumption that values are relatively steady over the course of a couple years.

Much of this data was difficult to measure and record for several countries, so there were many missing fields to deal with. To see what effect missing attributes had over a complete dataset, we generated two sample sets. The filled dataset uses linear regression to come up with a simple best-fit model for every attribute of every country. Any missing attributes are then populated using this model. Any negative values generated by the models are saturated at 0. The original dataset uses ridge regression to complete only the stress column - the value we use to classify our data. All other missing attributes are left as missing.

Pre-processing proved to be a very time consuming task, however it set us up for developing a high quality model. The final dataset contains a total of 1,986 samples including data on 179 countries from 1960 to 2014.

Results

Algorithm	Accuracy (%)
IBk	88.26
KStar	89.28
RandomForest	88.26

RandomTree	84.69
MultiClassClassifier	86.22
J48	84.69
LogitBoost	84.69
BayesNet	85.00
NaiveBayes	67.85
AdaBoostM1	75.51
ZeroR	75.51
MLP	75.51

NOTE: All models were generated using 10-Fold cross validation in Weka.

After generating these models, we converted our entire dataset into a test set, and we used the models to generate a whole new set of stress values. Once the new set of stress values were generated, we collected this data and used scikit-learn to perform linear regression on each country's stress data over the time period of 1960 to 2014. Using these models, we were able predict the year when stress crossed a critical level, for each country.

Analysis

Despite the high number of missing attributes, our model was able to predict Day Zero for all countries in the list within a reasonable range compared to other existing models. According to our model’s predictions, some countries appear to be rapidly approaching a water crisis, like Spain and Azerbaijan. Our model predicted that Spain would reach Day Zero by 2026, and Azerbaijan would reach Day Zero by 2065. For countries that have not been facing any water shortage so far, the Day Zero predictions show that they may only run out of water after many centuries, which seemed quite accurate.

Given the type of inputs and outputs of our dataset, we suspected using a decision tree or some instance based learner would be effective in classifying the data. We also predicted that a multilayer perceptron may do well with our wide range of attributes because of its flexibility. The results obtained for our project were very interesting. As expected, the best models were either trees or instance based learners. These algorithms did much better than ZeroR, showing that a decent model can certainly be generated with our sample size and attributes. Our best performer was the KStar model which is implemented in the visualization tool on the website. The MLP model performed much lower than expected, possibly because it was overfitting the training set. Further tuning could potentially improve this model, but there was unfortunately not enough time to test them all out due to the time it takes to generate just one model.

We observed that certain attributes like desalination capacity had very high information gain, when the decision tree model was built in Weka. This could be very useful information for countries that need to implement measures quickly to avert a water shortage.

Future Work

It is hard to deny that accurately determining Day Zero is influenced by many attributes that we have not even considered for our model. It is likely that many of these other attributes simply cannot be measured in a useful way. Our future plans to improve this model are to expand the list of attributes that have plentiful and reliable data so that we can get even more information about which factors are the most impactful in determining a country's water supply.

A second goal is to make the model more fine grained and include data on major cities rather than countries. Though our model can still provide useful information, it is still very generalized, especially for large countries that have varying climates, populations, and socioeconomic structures.

Finally, we would like to add more to our website and make it more interactive for users. Providing more information on each country and going into more detail on how to conserve water for a specific country can help raise awareness for areas that are in need.

Division of Work

Attribute Selection: Ahalya/Aamir
Data Gathering: Ahalya/Aamir
Data Purification: Aamir
Model to Predict Stress Level: Ahalya/Aamir
Analyzing and Selecting best Stress Level Predictor: Aamir
Model to Predict Day Zero: Ahalya
Website: Ahalya