CSC 458 Data Mining and Predictive Analytics I, Spring 2021

Dr. Dale E. Parson, Assignment 3, Using Weka linear regression, M5P trees, and additional machine learning (ML) models to correlate several stream-sampling attributes with dissolved oxygen in numeric milligrams per liter. Due by 11:59 PM on Friday April 2 via D2L. There will be a 10% per-day late penalty after that, and I cannot award points after I go over my solution during the following class period.

This Assignment focuses on two things: 1) using regression to predict numeric attribute values for OxygenMgPerLiter, and 2) investigating three techniques for down-sampling 468,015 water sampling instances from 2012 in PA to 49,066 training instances and 418,949 disjoint testing instances to determine the most effective technique in terms of validation within the training datasets and external testing against the testing datasets. I have written a Python script that extracts the 3 pairs of training-testing datasets in order to avoid low-level preprocessing in Weka and to encourage you to practice your code reading skills.

Perform the following steps to set up for this semester's projects and to get my handout. Start out in your login directory on csit (a.k.a. acad).

```
cd $HOME
mkdir DataMine # This should already be there from assignment 1.
cd ./DataMine
cp ~parson/DataMine/csc458regression3sp2021.problem.zip csc458regression3sp2021.problem.zip
unzip csc458regression3sp2021.problem.zip
cd ./csc458regression3sp2021
```

TO GET THE FILES FROM ACAD TO YOUR Mac OR Linux machine use this command line:

```
scp YOURLOGIN@acad.kutztown.edu://home/kutztown.edu/parson/DataMine/csc458regression3sp2021.problem.zip
csc458regression3sp2021.problem.zip
```

You can use the reverse command line to copy files from your machine into your acad account.

WINDOWS USERS should log into https://download.kutztown.edu/ , log in, go to Computer Science, and download WinSCP. It appears you can just go to http://winscp.net/eng/download.php To get it.

EDIT THE SUPPLIED FILE README.txt when the following questions starting at Q1 below. Keep with the supplied format, and do not turn in a Word or PDF or other file format. I will deduct 20% for other file formats, because with this many varying assignments being turned in, I need a way to grade these in reasonable time, which for me is a batch edit run on the **vim** editor. Please turn in your final **README.txt** by the deadline using the D2L Assignment page 3.

Running Weka

The best way to run Weka during the pandemic is to load it onto your PC or laptop, downloading the latest stable version (currently 3.8.5) from here.
https://www.cs.waikato.ac.nz/ml/weka/

I will prepare the assignments by clicking on the Weka icon, not by running the command line to expand Weka's memory. This way I will try to ensure that you will not run out of memory using the default amount. A student has thoughtfully supplied configuration steps for expanding available memory on the Mac, but I have not had the opportunity to find & test the equivalent on Windows, so for now I am limiting out training dataset size to that of Assignment 2. External testing datasets can be much bigger, since Weka does not read external test instances into memory all at one time.

If you do your work on campus Windows PCs, make sure to save your work on a USB drive that you remember to take with you. Campus PCs discard any file changes when you log out. Campus PC users can run S:\ComputerScience\WEKA\WekaWith4GBcampus, which contains this batch command:

```
java -Xmx4096M -jar "S:\ComputerScience\WEKA\weka.jar"
```

Here are the files in directory csc458regression3sp2021 in order of creation.

README.txt # The file in which you must answer questions. **Turn it in via D2L by the deadline**.
arfflib_3_1.py # My ARFF data manipulation library from previous semesters.
USGS_PA_STREAM_2012_NUMERIC.arff # Source data for assns. 2 & 3, OxygenMgPerLiter is numeric
csc458regression3sp2021.py # My script for extracting the following files from ^^ this ARFF file.
The keys for the following TRAIN and TEST files are as follows:
SITE extracts sample sites '01454700', '01467200', '01474500', '01570500' for training from
USGS_PA_STREAM_2012_NUMERIC.arff as in Assignment 2. These are the 4 training
sites identified as a good cross-section for training in fall 2018. The 4 sites provide
49066 training instances with 418949 remaining test instances.

You could have done this Weka by using **Filter -> unsupervised -> instance -> RemoveWithValues** and removing the 4 sites for TEST data and invertSelection for keeping them in TRAIN. I decided to automate this in Python.

USGS_PA_STREAM_2012_REGRESSION_SITE_TRAIN.arff
USGS_PA_STREAM_2012_REGRESSION_SITE_TEST.arff

SORT sorts all instances on MinuteOfYear as the primary sort key and MinuteOfDay as the
secondary sort key, then extracts the first 49066 instances as training instances and the
remaining 418949 instances as testing instances.

You could have done this in Weka by sorting by clicking in column headings in the Edit window, using shift-click to get descending order, and then using **Filter -> unsupervised -> instance -> RemovePercentage** to remove TEST instances from the TRAIN dataset and vice versa. That approach seemed too laborious and error prone to me.

USGS_PA_STREAM_2012_REGRESSION_SORT_TRAIN.arff
USGS_PA_STREAM_2012_REGRESSION_SORT_TEST.arff

RAND randomizes instance order and then extracts the first 49066 instances as
training instances and the remaining 418949 instances as testing instances.

You could have done this in Weka by using **Filter -> unsupervised -> instance -> Randomize**, and then using **Filter -> unsupervised -> instance -> RemovePercentage** to remove TEST instances from the TRAIN dataset and vice versa. That approach seemed too laborious and error prone to me.

USGS_PA_STREAM_2012_REGRESSION_RAND_TRAIN.arff
USGS_PA_STREAM_2012_REGRESSION_RAND_TEST.arff

The following attributes appear in the TEST ARFF files and all except site_no appear in the TRAIN files; site_no is eliminated from the TRAIN files to avoid trivial (site_no, datetime) -> OxygenMgPerLiter memorization.

site_no	The USGS site number for the sampling station.
pH	Base / acidity pH scale. Low numbers are acidic, with 7 being neutral.
TempCelsius	Water temperature in centigrade.
Conductance	Electrical conductance in microsiemens per centimeter at 25ºC.
DischargeRate	Flow discharge rate in cubic feet per second.
TimeOfYear	Nominal value for one of the seasons, derived from datetime.
TimeOfDay	Nominal value for one time of day of sample, derived from datetime.
month	Month as a number 2 through 12; there were no January measures.
MinuteOfDay	Number of minutes since the preceding midnight for this sample.
MinuteFromMidnite	Number of minutes to the closest midnight for this sample.
MinuteOfYear	Sampling time in minutes since the previous start of the year.
MinuteFromNewYear	Sampling time in minutes to the closest start of the year.
OxygenMgPerLiter	Dissolved oxygen in numeric milligrams per liter is our target attribute.

Analyzing the data. 5% for each of Q1 through Q20.

LINKED READINGS. These are the readings from Assignment 2. I am linking them in Assignment 3 for reference..
http://www.fondriest.com/environmental-measurements/parameters/water-quality/dissolved-oxygen/
Here is another, in case that site is down when you go to it.
https://water.usgs.gov/edu/dissolvedoxygen.html

In the following steps you will need to do the same model building and investigation for training datasets
USGS_PA_STREAM_2012_REGRESSION_SITE_TRAIN.arff
USGS_PA_STREAM_2012_REGRESSION_SORT_TRAIN.arff and
USGS_PA_STREAM_2012_REGRESSION_RAND_TRAIN.arff.
I recommend doing all of the SITE steps, then going back and doing the SORT steps, and then going back and doing the RAND steps, rather than constantly loading three different TRAIN files at each step. This approach will go faster and may be less error prone in terms of loading the wrong file. I also recommend exiting and re-starting Weka when you complete each of these 3 passes in order to recover memory. BE CAREFUL WHEN GOING FROM AN EXTERNAL TEST DATASET BACK TO 10-FOLD CROSS-VALIDATION, SINCE WEKA APPEARS TO ATTACH THE TESTING APPROACH TO EACH INDIVIDUAL MODELING ALGORITHM SUCH AS LinearRegression.

Q1: Run **rules -> ZeroR** classifier on your TRAIN dataset and paste the following measures into README.txt, supplying the actual values instead of N.N.

USGS_PA_STREAM_2012_REGRESSION_SITE_TRAIN

ZeroR predicts class value: N.N
Correlation coefficient N.N
Mean absolute error N.N
Root mean squared error N.N
Relative absolute error N.N %
Root relative squared error N.N %
Total Number of Instances 49066

USGS_PA_STREAM_2012_REGRESSION_SORT_TRAIN

Same measures as above in Q1.

USGS_PA_STREAM_2012_REGRESSION_RAND_TRAIN

Same measures as above in Q1.

Q2: What is the basis for the ZeroR prediction of the dissolved oxygen (DO OxygenMgPerLiter) class value?

Q3: Run functions -> LinearRegression regressor on your TRAIN dataset and paste the following measures into README.txt, supplying the actual values instead of N.N.

USGS_PA_STREAM_2012_REGRESSION_SITE_TRAIN

Correlation coefficient N.N
Mean absolute error N.N
Root mean squared error N.N
Relative absolute error N.N %
Root relative squared error N.N %
Total Number of Instances 49066

USGS_PA_STREAM_2012_REGRESSION_SORT_TRAIN

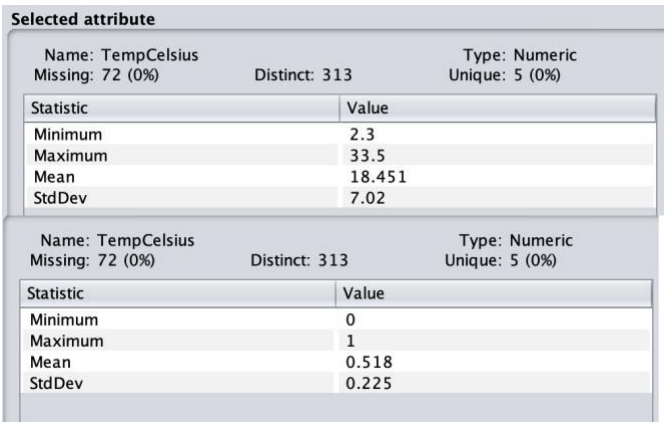
Same measures as above in Q3.

USGS_PA_STREAM_2012_REGRESSION_RAND_TRAIN

Same measures as above in Q3.

Q4: Run Preprocess -> Filter -> unsupervised -> attribute -> Normalize with its default parameters to normalize all attributes except target OxygenMgPerLiter. What is the default normalized range of each non-target attribute?

Q5: Before normalization USGS_PA_STREAM_2012_REGRESSION_SITE_TRAIN's TempCelsius shows a minimum value of 2.3, a maximum of 33.5, and a mean of 18.451. After Normalization it shows a min of 0, a max of 1, and mean of 0.518. How does Normalization arrive at the per-instance values for TempCelsius? You can describe it or give a formula for computing normalization in the resulting range.



Q6: Run functions -> LinearRegression regressor on your Normalized TRAIN dataset and paste the following measures into README.txt, supplying the actual values instead of N.N.

USGS_PA_STREAM_2012_REGRESSION_SITE_TRAIN

Correlation coefficient N.N
Mean absolute error N.N
Root mean squared error N.N
Relative absolute error N.N %
Root relative squared error N.N %
Total Number of Instances 49066

USGS_PA_STREAM_2012_REGRESSION_SORT_TRAIN

Same measures as above in Q6.

USGS_PA_STREAM_2012_REGRESSION_RAND_TRAIN

Same measures as above in Q6.

Q7: Do you see a substantial change in any of the accuracy / error measures going from Q3 to Q6, where "substantial" is a 10% or greater increase or decrease?

USGS_PA_STREAM_2012_REGRESSION_SITE_TRAIN change?

USGS_PA_STREAM_2012_REGRESSION_SORT_TRAIN change?

USGS_PA_STREAM_2012_REGRESSION_RAND_TRAIN change?

Q8: Copy & paste the entire normalized linear regression formula from Q7 below.

USGS_PA_STREAM_2012_REGRESSION_SITE_TRAIN formula:

OxygenMgPerLiter =
    Show sum of products + final constant coefficient.

USGS_PA_STREAM_2012_REGRESSION_SORT_TRAIN formula:

OxygenMgPerLiter =
    Show sum of products + final constant coefficient.

USGS_PA_STREAM_2012_REGRESSION_RAND_TRAIN formula:

OxygenMgPerLiter =
    Show sum of products + final constant coefficient.

Q9: Do you see a substantial change in any of the coefficients (multipliers) going from the unnormalized linear regression formula of Q3 to the normalized formula of Q6-Q8, where "substantial" is a 10% or greater increase or decrease? If there are any, give the name of one attribute whose coefficient has changed 10% or more.

USGS_PA_STREAM_2012_REGRESSION_SITE_TRAIN change?

USGS_PA_STREAM_2012_REGRESSION_SORT_TRAIN change?

USGS_PA_STREAM_2012_REGRESSION_RAND_TRAIN change?

Q10: A negative coefficient (multiplier) may show an inverse correlation between the multiplied attribute value and the target attribute value, where when one goes up the other goes down. A negative multiplier may also be a small correction to a bigger positive multiplier for a partially redundant attribute. In the formulas of Q8, is there any obvious attribute with an unambiguous negative correlation with target attribute OxygenMgPerLiter? (NOTE: A term like this in the formula – N.N * TimeOfYear=autumn,winter – indicates substituting 1 for "TimeOfYear=autumn,winter" when TimeOfYear is one of those values, else substituting 0.)

USGS_PA_STREAM_2012_REGRESSION_SITE_TRAIN attribute with negative correlation?

USGS_PA_STREAM_2012_REGRESSION_SORT_TRAIN attribute with negative correlation?

USGS_PA_STREAM_2012_REGRESSION_RAND_TRAIN attribute with negative correlation?

Q11: Run trees -> M5P model tree on your Normalized TRAIN dataset and paste the following measures into README.txt, supplying the actual values instead of N.N.

USGS_PA_STREAM_2012_REGRESSION_SITE_TRAIN

Correlation coefficient N.N
Mean absolute error N.N
Root mean squared error N.N
Relative absolute error N.N %
Root relative squared error N.N %
Total Number of Instances 49066

USGS_PA_STREAM_2012_REGRESSION_SORT_TRAIN

Same measures as above in Q11.

USGS_PA_STREAM_2012_REGRESSION_RAND_TRAIN

Same measures as above in Q11.

Q12: Run trees -> RandomTree regressor tree on your Normalized TRAIN dataset and paste the following measures into README.txt, supplying the actual values instead of N.N.

USGS_PA_STREAM_2012_REGRESSION_SITE_TRAIN

Correlation coefficient N.N
Mean absolute error N.N
Root mean squared error N.N
Relative absolute error N.N %
Root relative squared error N.N %
Total Number of Instances 49066

USGS_PA_STREAM_2012_REGRESSION_SORT_TRAIN

Same measures as above in Q12.

USGS_PA_STREAM_2012_REGRESSION_RAND_TRAIN

Same measures as above in Q12.

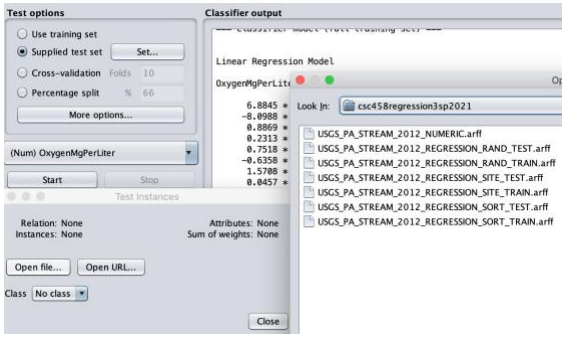
Q13: Noting my working definition of MDL (minimum description length) as being the most intelligible model for human interpretation with no worse than a 10% degradation from the most accurate model, which of the three models applied to the normalized attributes – LinearRegression, M5P, or RandomTree – gives the MDL model? Use Correlation Coefficient as the measure of accuracy. Explain your answer.

USGS_PA_STREAM_2012_REGRESSION_SITE_TRAIN MDL model out of the three?

USGS_PA_STREAM_2012_REGRESSION_SORT_TRAIN MDL model out of the three?

USGS_PA_STREAM_2012_REGRESSION_RAND_TRAIN MDL model out of the three?

Q14: For this one change from 10-fold cross-validation to an external test dataset corresponding to the training data per the next page. Give the indicated results of LinearRegression. **YOU MUST RELOAD THE UNNORMALIZED TRAIN FILES BECAUSE THE TEST FILES ARE NOT NORMALIZED!** I recommend exiting & re-starting Weka for Q14.



Unnormalized (original) USGS_PA_STREAM_2012_REGRESSION_SITE_TRAIN paired with USGS_PA_STREAM_2012_REGRESSION_SITE_TEST

Correlation coefficient	N.N
Mean absolute error	N.N
Root mean squared error	N.N
Relative absolute error	N.N %
Root relative squared error	N.N %
Total Number of Instances	418949

Unnormalized (original) USGS_PA_STREAM_2012_REGRESSION_SORT_TRAIN paired with USGS_PA_STREAM_2012_REGRESSION_SORT_TEST

Same measures as above in Q14.

Unnormalized (original) USGS_PA_STREAM_2012_REGRESSION_RAND_TRAIN paired with USGS_PA_STREAM_2012_REGRESSION_RAND_TEST

Same measures as above in Q14.

Q15: Which dataset, SITE, SORT, or RAND gives the most accurate measure for Correlation coefficient in Q14? Why? Inspect the TRAIN distributions for attribute TimeOfYear in the Weka Preprocess tab to help explain "Why?".

Q16: Which dataset, SITE, SORT, or RAND shows the least over-fitting to the training data in Q14? Explain your answer.

For Q17 through Q20 inspect the source code in csc458regression3sp2021.py to arrive at you answers. You do not need to look at any other Python source code files. Give brief, one or two sentence summaries.

Q17: How does csc458regression3sp2021.py extract USGS_PA_STREAM_2012_REGRESSION_SITE_TRAIN and USGS_PA_STREAM_2012_REGRESSION_SITE_TEST from USGS_PA_STREAM_2012_NUMERIC.arff?

Q18: How does csc458regression3sp2021.py extract USGS_PA_STREAM_2012_REGRESSION_SORT_TRAIN and USGS_PA_STREAM_2012_REGRESSION_SORT_TEST from USGS_PA_STREAM_2012_NUMERIC.arff?

Q19: How does csc458regression3sp2021.py extract USGS_PA_STREAM_2012_REGRESSION_RAND_TRAIN and USGS_PA_STREAM_2012_REGRESSION_RAND_TEST from USGS_PA_STREAM_2012_NUMERIC.arff?

Q20: Based on your answers to the above questions, especially Q15 and Q16, which of the three approaches to training models – SITE, SORT, or RAND – do you recommend using in Assignment 4? Why?

Make sure to **turn README.txt into the D2L Assignment 3 by the due date & time** to avoid a late penalty of 10% per day.