

Dr. Dale E. Parson, Assignment 4, Using Naïve Bayes, BayesNet, and additional classification models to correlate several stream-sampling attributes with dissolved oxygen in numeric milligrams per liter. Due by 11:59 PM on Friday April 23 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.

The goals of this assignment include comparisons of machine models, discretization strategies, and effects of partially redundant attributes. The Assignment uses the RANDomly down-sampled 468,015 water sampling instances from Assignment 3, with 49,066 training instances and 418,949 disjoint testing instances. Please consult [my solution to Assignment 3](#) and related Zoom video recording from April 6 for the rationale behind using the RAND training and testing data..

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 previous assignments.
cd ./DataMine
cp ~parson/DataMine/csc458bayes4sp2021.problem.zip csc458bayes4sp2021.problem.zip
unzip csc458bayes4sp2021.problem.zip
cd ./csc458bayes4sp2021
```

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/csc458bayes4sp2021.problem.zip
csc458bayes4sp2021.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 files README.txt and USGS_PA_STREAM_2012_NOMINAL.arff by the deadline using the D2L Assignment page 2.

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"
*****
```

Questions Q1 through Q10 are worth 10% of the assignment each.

STEP 1: Load USGS_PA_STREAM_2012_CLASSIFICATION_RAND_TRAIN.arff into Weka. This is the randomly shuffled dataset from Assignment 3 with 49,066 training instances in this training file and with 418,949 testing instances in USGS_PA_STREAM_2012_CLASSIFICATION_RAND_TEST.arff. While some 10-fold cross-validation tests had better results using the other approaches of user-selected SITE or SORT-on-time down-sampling in Assignment 3, RAND had the least degradation and the best overall accuracy when switching to testing with an external dataset. Testing on the training data often leads to bias towards that data and over-fitting models to that data.

Note that in this difference check for attributes, the training and testing files have the same attribute type definitions except that site_no has been removed from the training data to eliminate trivial memorization of the (site_no, MinuteOfYear) -> OxygenMgPerLiter mappings. We want to build models that investigate the physical and biological properties of the water and do not simply memorize samples. It was necessary to discretize OxygenMgPerLiter into 10 non-equal-frequency bins **before** partitioning the data into training and testing files because the lower and upper bounds on numeric OxygenMgPerLiter values determine the size of the nominal intervals. If these bounds had been different between early-partitioned training and testing data (likely), they would have discretized into incompatible nominal attributes.

```
$ diff USGS_PA_STREAM_2012_CLASSIFICATION_RAND_TRAIN.arff USGS_PA_STREAM_2012_CLASSIFICATION_RAND_TEST.arff
2c2,3
< % ARFF file generated @ 2021-04-03 17:43:20.792033
---
> % ARFF file generated @ 2021-04-03 17:43:21.520290
> @attribute site_no
{01454700,01460200,01465798,014670261,01467042,01467048,01467086,01467087,01467200,014721
04,01473500,01473900,01474000,01474500,01475530,01475548,01477050,01480617,01480870,014810
00,01540500,01542500,01544500,01548303,01549700,01567000,01570500,03007800,03025500,030340
00,03036000,03039035,03039036,03039040,03039041,03049640,03073000,03075070,03077100,030810
00,03082500,03083500,03105500,03106000,03108490,04213152}
16,49081c17,418965
< 7.5,8.9,87,763,spring,evening,4,1335,105,132375,132375,"\"(10.45-12.24)\""
...
```

Here are the nominal bins for OxygenMgPerLiter. Bayesian and other models of this assignment require classification of nominal values. Other attributes can be nominal or numeric. We are using the attributes of Assignment 3 without discretization of other numeric attributes.

@attribute OxygenMgPerLiter

```
{'(-inf-3.29]','(3.29-5.08]','(5.08-6.87]','(6.87-8.66]','(8.66-10.45]','(10.45-12.24]','(12.24-14.03]','(14.03-15.82]','(15.82-17.61]','(17.61-inf)'}
```

Use 10-fold cross-validation testing for the initial questions.

Q1: Run the **bayes -> NaiveBayes**, **bayes -> BayesNet**, and **trees -> J48** classifiers with their default parameters against the training data. Copy & paste only the following measures for each. In terms of kappa and the pasted error measures, which of the three is the most accurate?

NaiveBayes

Correctly Classified Instances	N	n.n %
Kappa statistic	n.n	
Mean absolute error	n.n	
Root mean squared error	n.n	
Total Number of Instances	49066	

BayesNet

Correctly Classified Instances	N	n.n %
Kappa statistic	n.n	
Mean absolute error	n.n	
Root mean squared error	n.n	
Total Number of Instances	49066	

J48

Correctly Classified Instances	N	n.n %
Kappa statistic	n.n	
Mean absolute error	n.n	
Root mean squared error	n.n	
Total Number of Instances	49066	

NaiveBayes

Correctly Classified Instances	21407	43.629 %
Kappa statistic	0.2906	
Mean absolute error	0.1209	
Root mean squared error	0.2662	

Total Number of Instances 49066

BayesNet

Correctly Classified Instances 27708 56.4709 %
Kappa statistic 0.4333
Mean absolute error 0.0923
Root mean squared error 0.2531
Total Number of Instances 49066

J48 (most accurate of the three)

Correctly Classified Instances 39456 80.4141 %
Kappa statistic 0.7412
Mean absolute error 0.0541
Root mean squared error 0.1796
Total Number of Instances 49066

STEP 2: Inspect the Naïve Bayes conditional probability table with this heading.

Naive Bayes Classifier

Attribute	Class '(-inf-3.29]' (0)	'(3.29-5.08]' (0.01)	'(5.08-6.87]' (0.09)	'(6.87-8.66]' (0.28)	'(8.66-10.45]' (0.32)	'(10.45-12.24]' (0.21)	'(12.24-14.03]' (0.08)	'(14.03-15.82]' (0.01)	'(15.82-17.61]' (0)	'(17.61-inf)' (0)
-----------	-------------------------------	-------------------------	-------------------------	-------------------------	--------------------------	---------------------------	---------------------------	---------------------------	------------------------	----------------------

Each table entry shows the conditional probability $P(\text{OxygenMgPerLiter} \mid \text{non-target-attribute-value})$ between OxygenMgPerLiter in that column and a non-target-attribute-value in that row. In the case of nominal attributes such as TimeOfYear and TimeOfDay the counts are one greater than the actual count of instances containing the paired value of OxygenMgPerLiter and the non-target-attribute-value because Weka's Naïve Bayes adds 1 to each such pair count to avoid a divide by 0 in computing probabilities.

Attribute	Class '(-inf-3.29]' (0)	'(3.29-5.08]' (0.01)	'(5.08-6.87]' (0.09)	'(6.87-8.66]' (0.28)	'(8.66-10.45]' (0.32)	'(10.45-12.24]' (0.21)	'(12.24-14.03]' (0.08)	'(14.03-15.82]' (0.01)	'(15.82-17.61]' (0)	'(17.61-inf)' (0)
TimeOfYear	1.0	1.0	1.0	24.0	289.0	686.0	260.0	69.0	3.0	1.0
winter	3.0	250.0	1592.0	3632.0	3667.0	2250.0	822.0	27.0	3.0	1.0
TimeOfDay	1.0	111.0	648.0	2299.0	4383.0	3162.0	1375.0	212.0	34.0	5.0
morning	1.0	114.0	727.0	3509.0	4324.0	2580.0	875.0	80.0	16.0	2.0
afternoon	4.0	261.0	1611.0	4345.0	3369.0	2131.0	659.0	19.0	1.0	1.0
evening	9.0	736.0	4578.0	13785.0	15743.0	10123.0	3731.0	338.0	54.0	9.0
night										
[total]										

Correlation of numeric non-target-attribute-values to discretized OxygenMgPerLiter uses mean values for those attributes found via an internal discretization of Naïve Bayes.

Attribute	Class '(-inf-3.29]' (0)	'(3.29-5.08]' (0.01)	'(5.08-6.87]' (0.09)	'(6.87-8.66]' (0.28)	'(8.66-10.45]' (0.32)	'(10.45-12.24]' (0.21)	'(12.24-14.03]' (0.08)	'(14.03-15.82]' (0.01)	'(15.82-17.61]' (0)	'(17.61-inf)' (0)
pH	7.24	7.1452	7.4501	7.467	7.5519	7.6172	7.8579	8.5868	8.9022	8.98
mean	0.2871	0.342	0.3136	0.4162	0.4509	0.5264	0.6329	0.4493	0.2583	0.04
std. dev.	5	717	3701	11956	14429	8772	3444	317	46	5
weight sum	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
precision										

Q2: For Q2 ignore the table columns with (0) immediately below the class label such as '(-inf-3.29]' because the (0) indicates that the model found less than a 1% (0.01) correlation between the attribute values of the rows and that range of OxygenMgPerLiter values. While ignoring those columns, which **numeric**

attributes' **mean** values increase or decrease monotonically as you read across their rows from left to right (i.e., they never change direction when decreasing or increasing)? ["A monotonic function is a function which is either entirely nonincreasing or nondecreasing."](#)

pH is the only one. Underlines below show 3-column non-monotonic changes.

Attribute	'(-inf-3.29]' (0)	'(3.29-5.08]' (0.01)	'(5.08-6.87]' (0.09)	'(6.87-8.66]' (0.28)	'(8.66-10.45]' (0.32)	'(10.45-12.24]' (0.21)	'(12.24-14.03]' (0.08)	'(14.03-15.82]' (0.01)	'(15.82-17.61]' (0)	'(17.61-inf)' (0)
pH mean	7.24	7.1452	7.4501	7.467	7.5519	7.6172	7.8579	8.5868	8.9022	8.98
TempCelsius mean	19.4897	26.1429	24.4525	21.1923	16.4829	10.623	7.2783	9.9282	13.6448	30.0621
Conductance mean	393.4736	327.0547	439.7131	343.0055	304.5396	289.305	307.83	463.5201	660.8051	642.397
DischargeRate mean	1289.6719	2369.6617	1199.6793	1546.7746	1836.5075	2192.4666	1580.0603	819.9079	353.2142	531.7068
month mean	9.4	7.7746	7.4563	7.6862	7.9279	8.646	9.8178	7.8174	7.44	6.8
MinuteOfDay mean	340.5508	558.577	568.4771	659.7214	749.8306	747.7979	767.2077	902.8278	967.3093	934.7034
MinuteFromMidnite mean	342.7606	354.3616	348.6087	324.6939	368.2642	379.4404	399.8383	455.1236	445.7915	502.9859
MinuteOfYear mean	385395.3399	316939.9815	304204.4987	315874.3655	326517.8851	355602.7092	411097.5861	323079.4457	306421.9096	262727.4115
MinuteFromNewYear mean	141642.7811	209172.6139	209833.6062	190098.5592	157797.683	108546.2065	73525.5919	94798.9073	125176.1052	251496.7348

Q3: For a numeric attribute identified in your answer to Q2, does it have a strong correlation to OxygenMgPerLiter, either positive or negative correlation, in the LinearRegression formula for normalized non-target attributes of USGS_PA_STREAM_2012_REGRESSION_RAND_TRAIN at the bottom page 10 in my solution to Assignment 3? Explain your answer.

Yes. pH has the second-highest magnitude coefficient in that formula.
USGS_PA_STREAM_2012_REGRESSION_RAND_TRAIN formula:

OxygenMgPerLiter =
5.9934 * pH +
-7.693 * TempCelsius +
-2.0454 * Conductance +
-1.5807 * DischargeRate +
0.5271 * TimeOfYear=spring,autumn,winter +
-0.9279 * TimeOfYear=autumn,winter +
1.4536 * TimeOfYear=winter +
-0.1609 * TimeOfDay=morning,evening,afternoon +
0.7295 * TimeOfDay=evening,afternoon +
0.2125 * TimeOfDay=afternoon +
1.81 * month +
0.547 * MinuteFromMidnite +
-0.8017 * MinuteOfYear +
0.1997 * MinuteFromNewYear +
9.2015

Q4: The attribute that has the highest magnitude (i.e., highest absolute value) coefficient (multiplier) in the LinearRegression formula of Q3 does not vary monotonically in your answer to Q2. What attribute has the highest magnitude (i.e., highest absolute value) coefficient in Q3? From our previous discussions and assignments, what most probably accounts for the non-monotonic interval in that attribute's mean values reading left-to-right?

TempCelsius. Its non-monotonic increase from 7.2783 to 9.9282 degrees C in going from OxygenMgPerLiter range '(12.24-14.03]' to '(14.03-15.82]' is probably due to the exponential spike in

OxygenMgPerLiter in late June to early July from exponential plant growth, e.g., for the peaks that we took out for the Susquehanna River at Danville.

STEP 3: Statistical algorithms such as Naïve Bayes assume statistical independence of non-target attributes in calculating joint probabilities. Temporarily remove attributes TimeOfYear, TimeOfDay, month, MinuteFromMidnite, and MinuteFromNewYear, since they are redundant with MinuteOfDay and MinuteOfYear. This leaves the 7 attributes of Figure 1. MinuteOfDay is slightly redundant with MinuteOfYear, since you can derive the former from the latter, but it provides correlation with diurnal patterns in oxygen levels not easily seen using MinuteOfYear.

No.	Name
1	<input checked="" type="checkbox"/> pH
2	<input type="checkbox"/> TempCelsius
3	<input type="checkbox"/> Conductance
4	<input type="checkbox"/> DischargeRate
5	<input type="checkbox"/> MinuteOfDay
6	<input type="checkbox"/> MinuteOfYear
7	<input type="checkbox"/> OxygenMgPerLiter

Figure 1: Non-redundant attributes of STEP 3.

Q5: As in Q1, Run the **bayes -> NaiveBayes**, **bayes -> BayesNet**, and **trees -> J48** classifiers with their default parameters against this reduced training data. Copy & paste only the following measures for each. In terms of **kappa**, which of the three improved? Did the improver(s) improve enough to offset the reduction in kappa in the other(s), i.e., did it/they have a more numerically significant improvement than the other(s) reduction in kappa? Explain your answer.

NaiveBayes

Correctly Classified Instances	N	n.n %
Kappa statistic	n.n	
Mean absolute error	n.n	
Root mean squared error	n.n	
Total Number of Instances	49066	

BayesNet

Correctly Classified Instances	N	n.n %
Kappa statistic	n.n	
Mean absolute error	n.n	
Root mean squared error	n.n	
Total Number of Instances	49066	

J48

Correctly Classified Instances	N	n.n %
--------------------------------	---	-------

Kappa statistic	n.n
Mean absolute error	n.n
Root mean squared error	n.n
Total Number of Instances	49066

NaiveBayes Improved more than 10% of its prior kappa.

Correctly Classified Instances	26876	54.7752 % (up from 43.629 %)
Kappa statistic	0.4057	(up from 0.2906)
Mean absolute error	0.1198	(down from 0.1209)
Root mean squared error	0.2454	(down from 0.2662)
Total Number of Instances	49066	

BayesNet Improved more than 10% of its prior kappa.

Correctly Classified Instances	29890	60.9179 % (up from 56.4709 %)
Kappa statistic	0.4828	(up from 0.4333)
Mean absolute error	0.0926	(up slightly from 0.0923)
Root mean squared error	0.2301	(down from 0.2531)
Total Number of Instances	49066	

J48 Degraded less than 10% of its prior kappa. Yes, improvement offsets J48 drop.

Correctly Classified Instances	39305	80.1064 % (down slightly from 80.4141 %)
Kappa statistic	0.7371	(down slightly from 0.7412)
Mean absolute error	0.0548	(up slightly from 0.0541)
Root mean squared error	0.1797	(up very slightly from 0.1796)
Total Number of Instances	49066	

STEP 4: Execute to Undo get back all attributes or exit Weka and reload the TRAINing ARFF file to get to the non-redundant nominal attributes in Figure 2. Part of the intent here is to move towards Minimal Description Length (MDL) of models without sacrificing more than 10% of accuracy in terms of kappa. TimeOfDay has only 4 nominal values and month has only 11 discrete numeric values, as compared with 10 values for each of the numeric attributes of Q2 and Q5. You will want to save the training set with the attributes of Figure 2 into a temporary ARFF file with its own name since that will be your training set for the remainder of the Assignment, which you can re-load after taking breaks.

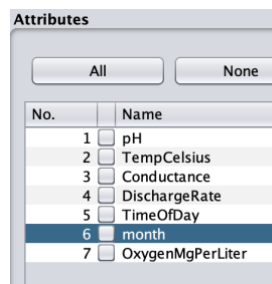


Figure 2: Coarse-grain temporal values TimeOfDay (nominal) and month (discrete integers)

Q6: As in Q5, Run the **bayes -> NaiveBayes**, **bayes -> BayesNet**, and **trees -> J48** classifiers with their default parameters against this reduced training data. Copy & paste only the following measures for each. In terms of **kappa**, which of the three improved?

NaiveBayes

Correctly Classified Instances	N	n.n %
Kappa statistic	n.n	
Mean absolute error	n.n	
Root mean squared error	n.n	
Total Number of Instances	49066	

BayesNet

Correctly Classified Instances	N	n.n %
Kappa statistic	n.n	
Mean absolute error	n.n	
Root mean squared error	n.n	
Total Number of Instances	49066	

J48

Correctly Classified Instances	N	n.n %
Kappa statistic	n.n	
Mean absolute error	n.n	
Root mean squared error	n.n	
Total Number of Instances	49066	

NaiveBayes Kappa improved slightly.

Correctly Classified Instances	27232	55.5008 % (up from 54.7752 %)
Kappa statistic	0.4138	(up from 0.4057)
Mean absolute error	0.1192	(down from 0.1198)
Root mean squared error	0.2438	(down from 0.2454)
Total Number of Instances	49066	

BayesNet Kappa down slightly.

Correctly Classified Instances	29427	59.9743 % (down from 60.9179 %)
Kappa statistic	0.4679	(down from 0.4828)
Mean absolute error	0.0963	(up from 0.0926)
Root mean squared error	0.2324	(up from 0.2301)
Total Number of Instances	49066	

J48 Kappa down slightly.

Correctly Classified Instances	38339	78.1376 % (down from 80.1064 %)
Kappa statistic	0.7111	(down from 0.7371)
Mean absolute error	0.0617	(up from 0.0548)
Root mean squared error	0.186	(up from 0.1797)
Total Number of Instances	49066	

STEP 5: Run BayesNet classification repeatedly after incrementing the searchAlgorithm -> maxNrOfParents by 1 each time until kappa hits a plateau, levelling off at a higher value than when maxNrOfParents is 1.

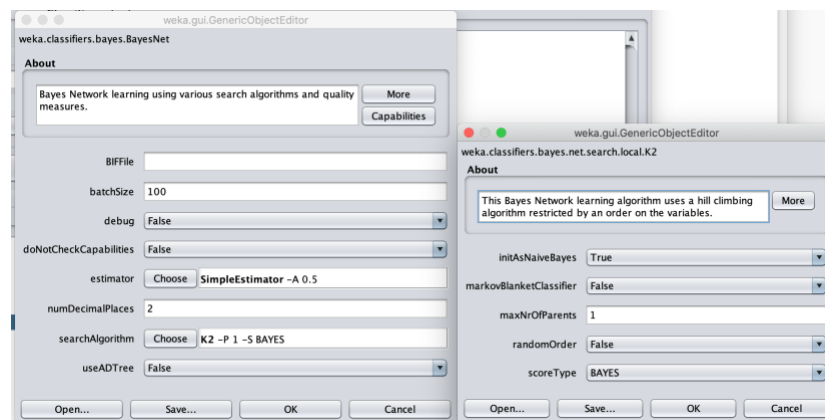


Figure 3: The searchAlgorithm -> maxNrOfParents parameter of BayesNet.

Q7: What is the smallest value for maxNrOfParents that hits this plateau in kappa value in STEP 5. Copy & paste the following measures for BayesNet with this maxNrOfParents value, noting increase or decrease in kappa for BayesNet from Q6.

BayesNet

Correctly Classified Instances	N	n.n %
Kappa statistic	n.n	
Mean absolute error	n.n	
Root mean squared error	n.n	
Total Number of Instances	49066	

BayesNet Kappa improved more than 10%. maxNrOfParents = 3

Correctly Classified Instances	36960	75.3271 % (up from 59.9743 %)
Kappa statistic	0.674	(up from 0.4679)
Mean absolute error	0.0611	(down from 0.0963)
Root mean squared error	0.1873	(down from 0.2324)
Total Number of Instances	49066	

Q8: NaiveBayes of Q6 gives the following table entries for TimeOfDay.

Attribute	Class '(-inf-3.29]' (0)	'(3.29-5.08]' (0.02)	'(5.08-6.87]' (0.09)	'(6.87-8.66]' (0.28)	'(8.66-10.45]' (0.32)	'(10.45-12.24]' (0.21)	'(12.24-14.03]' (0.08)	'(14.03-15.82]' (0.01)	'(15.82-17.61]' (0)	'(17.61-inf)' (0)
TimeOfDay										
morning	3.0	250.0	1592.0	3632.0	3667.0	2250.0	822.0	27.0	3.0	1.0
afternoon	1.0	111.0	648.0	2299.0	4383.0	3162.0	1375.0	212.0	34.0	5.0
evening	1.0	114.0	727.0	3509.0	4324.0	2580.0	875.0	80.0	16.0	2.0
night	4.0	261.0	1611.0	4345.0	3369.0	2131.0	659.0	19.0	1.0	1.0

Looking at the TimeOfDay counts for the higher dissolved oxygen columns starting at (8.66-10.45] mg. per liter and going right, do you see a pattern in the highest TimeOfDay row for each column? How does it relate to your prior readings?

Highest if afternoon, second highest for evening due to photosynthesis.

STEP 6: Clicking Alt (option) “Visualize graph” on the BayesNet Result list entry with the maxNrOfParents value of Q7 yields the following graph.

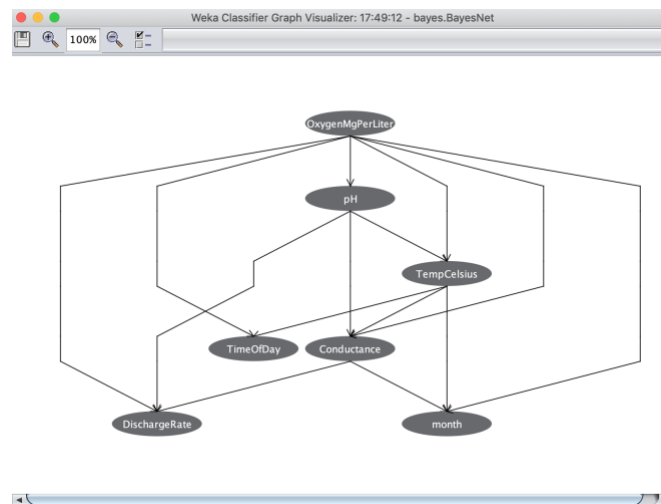


Figure 4: BayesNet graph of Q7.

We will discuss navigating this graph when we go over the solution to Assignment 4 in class. Compared to the rows and columns of the NaïveBayes correlation tables such as the one in Q8, interpreting BayesNets becomes as complicated as interpreting large decision trees. Basically trading intelligibility for accuracy.

Q9: Switch **TESTING** (do not change the training data) to use file USGS_PA_STREAM_2012_CLASSIFICATION_RAND_TEST.arff and run **bayes -> NaiveBayes** and **trees -> J48** classifiers as in Q6 and **bayes -> BayesNet** as in Q7 with its best maxNrOfParents parameter value. I recommend exiting Weka & restarting for each of these 3 tests, reloading the temporary ARFF file you saved in STEP 4; you must have the 49,066-instance training set with the attributes of Figure 2. Copy & paste only the following measures for each. How did kappa change as the result of going to testing with the non-training TEST dataset, compared to Q6 for NaiveBayes and J48 and to Q7 for BayesNet?

NaiveBayes

Correctly Classified Instances	N	n.n %
Kappa statistic	n.n	
Mean absolute error	n.n	
Root mean squared error	n.n	
Total Number of Instances	418949	

BayesNet with its best maxNrOfParents value

Correctly Classified Instances	N	n.n %
Kappa statistic	n.n	
Mean absolute error	n.n	
Root mean squared error	n.n	
Total Number of Instances	418949	

J48

Correctly Classified Instances	N	n.n %
Kappa statistic	n.n	
Mean absolute error	n.n	
Root mean squared error	n.n	
Total Number of Instances	418949	

NaiveBayes Kappa improved slightly. No over-fitting.

Correctly Classified Instances	231820	55.3337 % (on par with 55.5008 % x-validation)
Kappa statistic	0.412	(on par with 0.4138 from cross-validation)
Mean absolute error	0.1195	(on par with 0.1192 from cross-validation)
Root mean squared error	0.2443	(on par with 0.2438 from cross-validation)
Total Number of Instances	418949	

BayesNet with maxNrOfParents = 3 Minor improvement, no over-fitting.

Correctly Classified Instances	318199	75.9517 % (up from 75.3271 %)
Kappa statistic	0.6824	(up from 0.674)
Mean absolute error	0.0595	(down from 0.0611)
Root mean squared error	0.1852	(down from 0.1873)
Total Number of Instances	418949	

J48 No over-fitting.

Correctly Classified Instances	330378	78.8588 % (up slightly from 78.1376 %)
Kappa statistic	0.721	(up slightly from 0.7111)
Mean absolute error	0.0599	(down from 0.0617)
Root mean squared error	0.1828	(up slightly from 0.186)
Total Number of Instances	418949	

Q10: Which, if any, of the models of Q9 show the effects of overfitting their learning to training data when compared to kappa results for **bayes** -> **NaiveBayes** and **trees** -> **J48** from Q6 and **bayes** -> **BayesNet** from Q7?

None. Kappa was on par for NaiveBayes and slightly improved for BayesNet and J48 when going to external test data,

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.