Dr. Dale E. Parson, Assignment 4, Using Weka J48 trees, NaiveBayes, and BayesNet to thin the attribute set of weather and Hawk Mountain data in anticipation of Final Assignment 5.
Due by 11:59 PM on Wednesday December 4 via make turnitin. I need to send my solution to the class in a timely manner so students can prepare for the Final Assignment 5. Therefore, I will NOT accept solutions to this Assignment 4 after 9 AM on Friday December 6.

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.
cp  ~parson/DataMine/csc458fall2019assn4.problem.zip DataMine/csc458fall2019assn4.problem.zip
cd  ./DataMine
unzip  csc458fall2019assn4.problem.zip
cd  ./csc458fall2019assn4
```

EDIT THE SUPPLIED 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.

Background: In Assignment 3 we compressed the value range of the BW attribute by using
BWbins = AddExpression

```
 ifelse(a26=0,0,
        ifelse(a26=1,1,
               ifelse(a26=2,2,ifelse(a26<30,3,
                      ifelse(a26<200,4,
                             ifelse(a26<1000,5,6))))))
```

to create custom discrete attribute BWbins. That is a lossy compression since it loses fine-grain numeric resolution; it treats BW subranges derived from time(X) -> BW(Y) graphs as informal clusters.

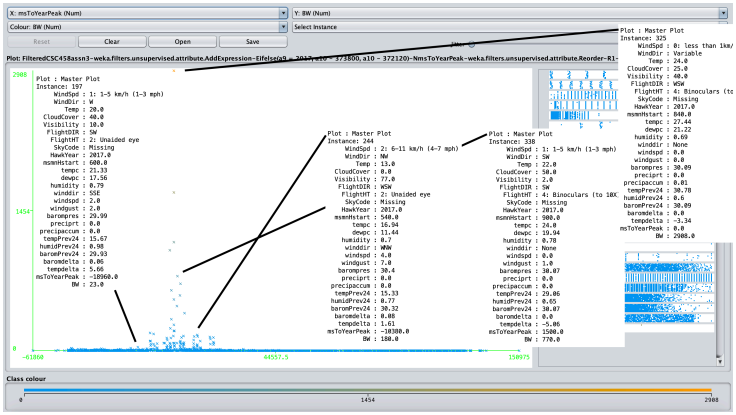

Figure 1: Partial illustration of basis for BWbins from Assignment 3

However, the BWbins histogram is still dominated by the 0-count bin.

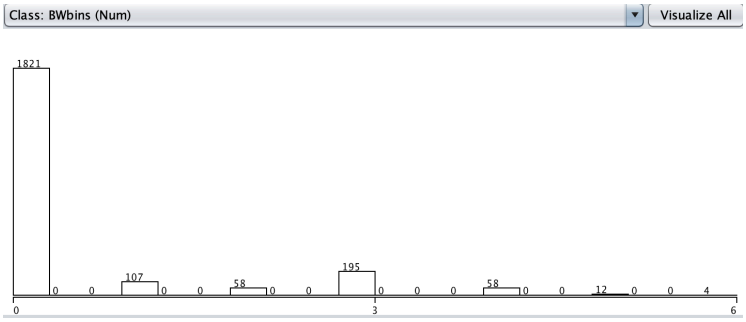

Figure 2: Bin 0 dominates BWbins

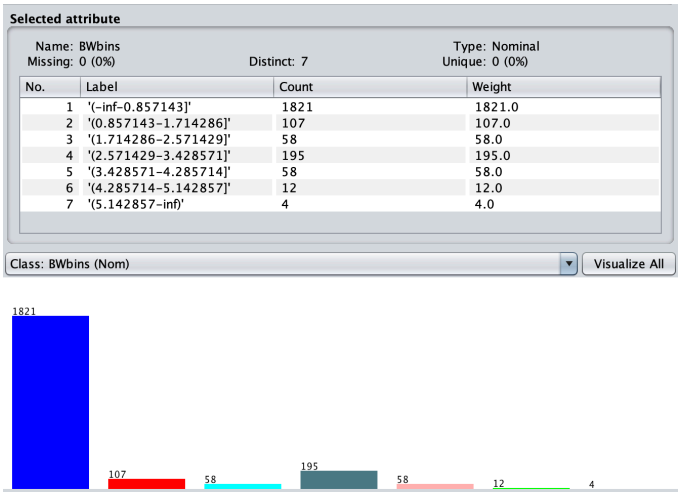

Figure 3: BWbins after Discretization into 7 bins in Q2 below

Final Assignment 5 will focus on compressing multiple contiguous BW==0 observation hours into single instances, in order to abstract them into "nothing happening" periods of time. That will reduce the number of BW==0 instances significantly. The current Assignment 4 sets out to explore two things A) NaiveBayes and BayesNet *statistical modeling mechanisms* as compared with J48 *information-entropic modeling*, and LinearRegression and M5P *linear modeling*, that we used in previous assignments, and B) investigation into whether we can stop joining Hawk Mountain data to separate weather station data such as the Hamburg Weather Underground data that we used this semester. The reason for (B) is that this work, if it goes forward after this semester, will incorporate many more years of Hawk Mountain data, and will incorporate additional observations sites to the north. If it is necessary to obtain separate weather data for (each year X each raptor observation site), this weather requirement will more than double the amount of work. It will be necessary to find additional sites and years of weather data amenable to automatic web scraping, to write scripts to clean and merge this data with Hawk Mountain data, and to analyze the reliability of this weather data. This assignment attempts to determine whether future efforts can concentrate solely on raptor observation site data.

STEP1: Load **HawkData20172018Assn4NumericsDeltas.arff** into Weka. This ARFF file is a modified variant of HawkData20172018.arff from Assignment 3. It contains the following

attributes. Note the H tag for Hawk Mountain Data, and the W tag for Weather Underground data from Hamburg.

WindSpd H North lookout wind speed as a nominal value, via portable anemometer
{'0: less than 1km/h (Calm)', '1: 1-5 km/h (1-3 mph)', '2: 6-11 km/h (4-7 mph)',
'3: 12-19 km/h (8-12 mph)', '4: 20-28 km/h (13-18 mph)',' '5: 29-38 km/h (19-24 mph)',
'6: 39-49 km/h (25-31 mph)', '7: 50-61 km/h (32-38 mph)', '8: 62-74 km/h (39-48 mph)',
'9: Greater than 75 km/h'}
WindDir H North lookout wind direction
{Variable,WNW,NW,SE,E,S,ESE,SW,SSW,N,NNW,NE,ENE,W,WSW,NNE,SSE}
HTemp H North lookout Celsius temperature
CloudCover H North lookout cloud cover, units of measure unknown
Visibility H North lookout visibility, units of measure unknown
FlightDIR H Raptor nominal flight direction (SE, etc.), same value set as WindDir
FlightHT H Raptor flight height as a nominal value
{'0: Below eye level', '1: Eye level to 30m', 2: Unaided eye', '3: limit of unaided vision',
'4: Binoculars (to 10X)', '5: At limit of binoculars (10X)', '7: Variable',(none)}
SkyCode H
{'0: Clear', '1: Partly Cloudy', '2: Mostly Cloudy', '3: Overcast',
'4: wind driven sand, snow, dust', '5: Fog or Dense Haze',
'6: Drizzle','7: Rain', '8. Snow'}
HawkYear H 2017 or 2018 for this dataset
msmnHstart H Minutes since observation day's previous midnight (00:00) for hawkStart.
WTemp W Weather station temperature in Celsius.
dewpc W Weather station dew point in Celsius.
humidity W Weather station % humidity as a fraction of 1.0.
winddir W Weather station wind direction as nominal.
{West,None,WNW,WSW,East,NNE,SE,SSW,NW,SW,South,SSE,NE,North,
ESE,ENE,NNW}
windspd W Weather station wind speed in MPH.
windgust W Weather station wind gust speed in MPH.
barompres W Weather station barometric pressure in inches.
preciprt W Weather station precipitation in inches.
precipaccum W Weather station accumulated precipitation in inches.
humidPrev24 W Weather station % humidity taken ~ 24 hours before this record.
baromPrev24 W Weather station barometric taken ~ 24 hours before this record.
baromdelta W Change in barometer in past 24 hours.
tempdelta W Change in temperature in past 24 hours.
msToYearPeak H Minutes to BW peak count for this year, from Assignment 3.
WindSpdMin H Minimum numeric value of WindSpd attribute range above.
WindSpdMean H Center numeric value of WindSpd attribute range above.
WindSpdMax H Maximum numeric value of WindSpd attribute range above.
WindDirNum H Compass point in degrees of WinDir, 0=N, 90=E, 180=S, 270=W
FlightDIRNum H Compass point in degrees like WindDirNum.
FlightHTNum H N: value N from FlightHT nominal code above.
SkyCodeNum H N: value N from SkyCode nominal code above.

WwinddirNum	W	Compass point in degrees as above.
HTempPrev72	H	HTemp 72 hours earlier.
WTempPrev72	W	WTemp 72 hours earlier.
HTempPrev48	H	HTemp 48 hours earlier.
WTempPrev48	W	WTemp 48 hours earlier.
HTempPrev24	H	HTemp 24 hours earlier.
WTempPrev24	W	WTemp 24 hours earlier.
HTempDelta72	H	These are the temperature changes from previous 72 or 48 or 24
WTempDelta72	W	hours for respective HTemp and WTemp values.
HTempDelta48	H	
WTempDelta48	W	
HTempDelta24	H	
WTempDelta24	W	
BW	H	Broad-winged Hawk count for that observation interval.
BWbins	H	BW compressed numeric value per AddExpression above.

Each of Q1 through Q15 is worth 6.66% of this assignment. There is no ARFF file to turn in.

Q1: Remove attribute BW. BWbins is our class attribute for this assignment. Run LinearRegression and M5P, and paste the following result values into Q1 in README.txt. All testing in Assignment 4 uses 10 fold cross-validation, i.e., no external test dataset.

LinearRegression
Correlation coefficient n.n
Relative absolute error n.n %
Root relative squared error n.n %
Total Number of Instances 2255

M5P
Number of Rules : N
Correlation coefficient n.n
Relative absolute error n.n %
Root relative squared error n.n %
Total Number of Instances 2255

Q2: **Unsupervised -> attribute -> Discretize** BWbins into 7 bins with useEqualFrequency=False and ignoreClass=True. Be very careful to Discretize ONLY the BWbins attribute. We will UNDO this step later. Make sure the 7 discretized bins have the same instance counts as their pre-Discretize numeric bins in the Preprocessor. See Figure 3 above. Run NaiveBayes, BayesNet, and J48, and paste the following result values into Q2 in README.txt.

NaiveBayes
Correctly Classified Instances N n.n %
Kappa statistic n.n
Total Number of Instances 2255

BayesNet
Correctly Classified Instances N n.n %
Kappa statistic n.n
Total Number of Instances 2255

J48
Correctly Classified Instances N n.n %
Kappa statistic n.n
Total Number of Instances 2255

Q3: UNDO to restore BWbins to its numeric state. Remove (we will UNDO) all attributes tagged with W for weather station data in the attribute list starting on page 2. We are trying to determine the effectiveness of skipping weather data collection outside the observation sites. Run LinearRegression and M5P, and paste the following result values into Q3 in README.txt.

LinearRegression
Correlation coefficient n.n
Relative absolute error n.n %
Root relative squared error n.n %
Total Number of Instances 2255

M5P
Number of Rules : N
Correlation coefficient n.n
Relative absolute error n.n %
Root relative squared error n.n %
Total Number of Instances 2255

Q4: In going from Q1 (all attributes) to Q3, which models (from LinearRegression and M5P) showed improvement in terms of correlation coefficient and the % error measures? Which degraded in performance, if any, and if so, was the degradation > 4% for any of the correlation coefficient and the % error measures? What do these changes tell you about the importance of collecting weather station data at a separate location (Hamburg) from the raptor data collection site?

Q5: Discretize BWbins into 7 bins with useEqualFrequency=False and ignoreClass=True as in Q2. Run NaiveBayes, BayesNet, and J48, and paste the following result values into Q5 in README.txt.

NaiveBayes
Correctly Classified Instances N n.n %
Kappa statistic n.n
Total Number of Instances 2255

BayesNet

Correctly Classified Instances		N	n.n %
Kappa statistic	n.n		
Total Number of Instances		2255	

J48

Correctly Classified Instances		N	n.n %
Kappa statistic	n.n		
Total Number of Instances		2255	

Q6: In going from Q2 (all attributes) to Q5, which models (from NaiveBayes, BayesNet, and J48) showed improvement in terms of % Correct Instances and kappa? Which degraded in performance, if any, and if so, was the degradation > 4% for any of the % Correct Instances and kappa error measures? What do these changes tell you about the importance of collecting weather station data at a separate location (Hamburg) from the raptor data collection site?

Q7: For the model in Q6 that degraded in performance, why would eliminating the weather station (W) data make that model worse, while making the other models better?

Q8: Copy and paste only these headings and rows from the NaiveBayes conditional probability table in the output, going left-to-right from BWbins=0, 1, 2, 3, 4, 5, 6. Note that NaiveBayes (and BayesNet) temporarily discretizes the non-target attributes into bins with fraction-bearing boundaries, but we know that BWbins is in the integer range [0, 6]. Look at each column separately. For a majority of columns, which WindDir value (row) has the highest count in its column? Are there substantial exceptions? Note that BWbins=6 gets a low count here because, even though there are many BW raptors in a given instance with BWbins=6, there are not many instances. Each number in this table is the number of instances classified in its column's bins.

Naive Bayes Classifier

	'(-inf-0.857143]'	'(0.857143-1.714286]'	'(1.714286-2.571429]'	'(2.571429-3.428571]'	'(3.428571-4.285714]'	'(4.285714-5.142857]'	'(5.142857-inf)'
	(0.81)	(0.05)	(0.03)	(0.09)	(0.03)	(0.01)	(0)
WindDir							
Variable							
WNW							
NW							
SE							
E							
S							
ESE							
SW							
SSW							
N							
NNW							
NE							
ENE							
W							
WSW							
NNE							
SSE							

Q9: Note that BWbins=0 (the first column) is dominated by this wind direction. BWbins=0 occurs when there are 0 BW raptor sightings during that time interval. What does the fact that this wind direction dominates the BWbins=0 (0 raptors) column, and its agreement with essentially all of the other columns, tell you about prevailing winds at North Lookout and the selection of this observation site?

Q10: Copy and paste only these headings and rows from the NaiveBayes conditional probability table in the output, going left-to-right from BWbins=0, 1, 2, 3, 4, 5, 6. Can you see a pattern in temperature change in the last 72, 48, and 24 hours for the three rightmost columns, i.e., BWbins = 4, 5, and 6, in going left-to-right? How does this pattern relate to the abstract in this paper: https://faculty.kutztown.edu/parson/fall2019/034.pdf ?

Naive Bayes Classifier

	'(-inf-0.857143]'	'(0.857143-1.714286]'	'(1.714286-2.571429]'	'(2.571429-3.428571]'	'(3.428571-4.285714]'	'(4.285714-5.142857]'	'(5.142857-inf)'
	(0.81)	(0.05)	(0.03)	(0.09)	(0.03)	(0.01)	(0)
HTempDelta72							
mean							
HTempDelta48							
mean							
HTempDelta24							
mean							

Q11: Remaining attributes in **SET A**:

WindSpd, WindDir, FlightDIR, FlightHT, SkyCode

are redundant with this **SET B**:

WindSpdMin, WindSpdMean, WindSpdMax, WindDirNum, FlightDIRNum, FlightHTNum, SkyCodeNum

Temporarily Remove the attributes from the above **SET A**. Run NaiveBayes, BayesNet, and J48, and paste the following result values into Q11 in README.txt.

NaiveBayes
Correctly Classified Instances N n.n %
Kappa statistic n.n
Total Number of Instances 2255

BayesNet
Correctly Classified Instances N n.n %
Kappa statistic n.n
Total Number of Instances 2255

J48
Correctly Classified Instances N n.n %
Kappa statistic n.n
Total Number of Instances 2255

Q12: In going from Q5 (all H attributes) to Q11 (eliminated SET A), which models (from NaiveBayes, BayesNet, and J48 showed improvement in terms of % Correct Instances and kappa? Which degraded in performance? What do these changes tell you about using the original, nominal attributes?

Q13: Execute UNDO to restore the SET A attributes, then Remove the SET B attributes. Run NaiveBayes, BayesNet, and J48, and paste the following result values into Q13 in README.txt.

NaiveBayes
Correctly Classified Instances N n.n %
Kappa statistic n.n
Total Number of Instances 2255

BayesNet
Correctly Classified Instances N n.n %
Kappa statistic n.n
Total Number of Instances 2255

J48
Correctly Classified Instances N n.n %
Kappa statistic n.n
Total Number of Instances 2255

Q14: Comparing NaiveBayes, BayesNet, and J48 results for Q5 (all H attributes), Q11 (removed SET A), and Q13 (removed SET B), is there a clear "winner" among these attribute configurations. Do you recommend using the Q5, the Q11, or the Q13 attributes (choose one if possible)? Explain your choice. If there is no clear choice, explain that. Any answer with a justification in your answer is OK.

Q15: With the Q13 attributes still in place (SET B is still Removed), use filter **SUPERVISED -> attribute -> Discretize** to discretize **ALL** attributes except BWbins. This **supervised filter** attempts to correlate its discretization of numeric attributes with the BWbins class attribute bins, even before Classification. Run NaiveBayes, BayesNet, and J48, and paste the following result values into Q15 in README.txt. Compare these to the Q13 results that used numeric non-target attributes.

NaiveBayes
Correctly Classified Instances N n.n %
Kappa statistic n.n
Total Number of Instances 2255

BayesNet
Correctly Classified Instances N n.n %
Kappa statistic n.n
Total Number of Instances 2255

J48
Correctly Classified Instances N n.n %
Kappa statistic n.n
Total Number of Instances 2255

I considered having a K-means clustering question, but even with 32 clusters, 28 of them have BWbins=0, with 3 having BWbins=3, and 1 with BWbins=1. The BW=0 instances continue to dominate the dataset. My optimistic plan is to compress each contiguous sequence of BW=0 instances into a single instance for Assignment 5, to improve resolution on the non-0 BW cases with just discarding BW=0 data, for example for its timing value and potential value in transitions in and out of the BW=0 state.

When you have completed all of your work and double-checked the assignment requirements, and your **README.txt** that answers Q1 through Q15 is sitting in your **csc458fall2019assn4/** directory, then run **make turnitin** by the due date. **Due by 11:59 PM on Wednesday December 4 via make turnitin.** Late assignments lose 10% per day late, and I will not accept an assignment after 9 AM on Friday December 6.