Sophia Swanberg

Data Mining

4 May 2021

Final Project

Questions:

***What are some unexpected patterns found in this dataset?***

***What attributes may effect population growth across the 12 islands sampled?***

I chose to do these questions because I am curious about the population and environment of 12 tropical islands. Climate change is decimating these populations, so I would like to see if my questions can be answered based on the data and the algorithms collected.

Dataset:

This dataset is a culmination of different datasets found using the following links:

*Precipitation and temperature dataset:* <https://ourworldindata.org/country/marshall-islands>

*Cancer incidence, Deaths from air pollution, Prevalence of mental health disorders, Total population, and Asthma prevalence for each island can be found here:* <https://climateknowledgeportal.worldbank.org/download-data>

Originally, I had a dataset strictly on the Marshall Islands that had different attributes such as total fisheries production, infant mortality rate, etc. I then changed my dataset to expand across 12 different tropical islands, including the Marshall Islands, Puerto Rico, Samoa, the Bahamas, Kiribati, Barbados, Jamaica, Fiji, Tonga, Solomon Islands, New Zealand, and Madagascar. These islands chosen were not limited to what ocean they were found in, but they were chosen based on their island size and tropical climate.

Preprocessing:

This dataset needed some work. Specifically, the temperature and rainfall data needed to be converted from their monthly averages to their yearly averages, since the rest of the data was on that scale. To do so, I wrote a program in Python that would read in the year and the monthly average, add the averages together, then divide by 12, and finally output the data in the following format: year, average. To use the 2 algorithms, I had to convert the island’s name into a number, so I set the range from 1-12 and gave each island a random number from that range. Finally, I needed to add a class value to the dataset to use the ID3 algorithm that was based on population growth (did the population increase or decrease?). I wrote another Python program that would read in the population values from 1989-2016, take two values (exp: 1990’s population and 1989’s population), subtract the two and determine if the population increased or decreased based on the previous year. If the result was positive, the output value was a 0 for ‘yes’, and a 1 for ‘no’ if the output was negative.

Steps:

***Algorithms:*** The two algorithms chosen to answer these questions were agglomerative clustering with single linkage and ID3. I chose these two based on the questions asked; the first question suggests that clustering would be effective, since it will find unexpected patterns, and ID3 for the second, since I’d like to see what attributes have an effect on population increase or decrease (implies classification).

***Clustering:***

***-Program:*** The clustering program I used for this project was a clustering algorithm I wrote for homework 2, ‘clusterstudent.py’. I had to modify this program slightly because I am ran into a ‘divide by zero’ error (I believe this is because the island attributes are the same across the 26 years), so I added a conditional in the ‘nDistance’ method to return 0 if the parameter ‘nvalue’ was 0.

***Part 1:***

***-Data Entry:*** To read in this data, I had to additionally preprocess my dataset to include another column labeled ‘clustercode’, so I could give each tuple its own identifier to analyze the cluster output. To achieve this, I simply took two letters from each island name (like how we abbreviate states in the US), then gave each abbreviation a number 0-26. I then copied the preprocessed data from excel into the ‘clusterdata.txt’ file, *not* including the row which describes the column values.

***-Running the program:*** When you run ‘clusterstudent.py’, you will be prompted ‘Enter 1 for Single Linkage or 2 for Complete Linkage’. I entered 1, which ran the single linkage method. When I wrote this program, I originally printed out the matrix of calculated distances at every iteration, and I had to comment this out due to the size of the dataset; it crashed my computer. Instead, I printed out the array containing the newly formed cluster names after every iteration, and my result consists of the last clusters outputted.

***-Results:*** The first time I ran this program, I included the island code column in the dataset. I was unsure how this would skew the results since there were consistently identical values across the 26 years for each island. In the end, there were 3 clusters created after 322 iterations: (note: if you look in the 2nd page of the spreadsheet in this folder, you will find the dataset clustered like this so you can see the tuple clusters)

['MA22 MA3 NZ11 SI13 TO22 TO12 TO2 FI19 FI9 JA26 JA16 JA6 BR23 BR13 BR3 KI13 KI14 MA23 MA4 NZ12 SI14 SI15 MA24 MA5 NZ13 SI16 SI17 MA25 MA6 NZ14 SI18 TO23 TO13 TO3 FI20 FI10 FI0 JA17 JA7 BR24 BR14 BR4 KI15 KI16', 'MA26 MA7 NZ15 SI19 TO24 TO14 TO4 FI21 FI11 FI1 JA18 JA8 BR25 BR15 BR5 KI17 KI18 MA8 NZ16 SI20 TO25 TO15 TO5 FI22 FI12 FI2 JA19 JA9 BR26 BR16 BR6 KI19 KI20 MA9 NZ17 SI21 SI22 MA10 NZ18 SI23 SI24 MA11 NZ19 SI25 SI26NZ0 MA12 NZ20 NZ1 TO26 TO16 TO6 FI23 FI13 FI3 JA20 JA10 JA0 BR17 BR7 KI21 KI22 MA13 NZ21 NZ2 SI0 TO17 TO7 FI24 FI14 FI4 JA21 JA11 JA1 BR18 BR8 KI23 KI24', 'MA14 NZ22 NZ3 SI1 TO18 TO8 FI25 FI15 FI5 JA22 JA12 JA2 BR19 BR9 KI25 KI26 MA15 NZ23 NZ4 SI2 SI3 MA16 NZ24 NZ5 SI4 TO19 TO9 FI26 FI16 FI6 JA23 JA13 JA3 BR20 BR10 BR0 KI9 KI7 KI5 KI3 KI1 BA26 BA24 BA22 BA20 BA18 BA16 BA14 BA12 BA10 BA8 BA6 BA4 BA2 BA0 SA25 SA23 SA21 SA19 SA17 SA15 SA13 SA11 SA9 SA7 SA5 SA3 SA0 PR26 PR25 PR24 PR23 PR22 PR21 PR20 PR19 PR18 PR17 PR16 PR15 PR14 PR13 PR12 PR11 PR10 PR9 PR8 PR7 PR6 PR5 PR4 PR3 PR2 PR1 PR0 MI26 MI25 MI24 MI23 MI22 MI21 MI20 MI19 MI18 MI17 MI16 MI15 MI14 MI11 MI10 MI9 MI8 MI7 MI6 MI5 MI4 MI3 MI2 MI1 MI0 MI12 MI13 MA17 NZ25 NZ6 SI5 SI6 MA18 NZ26 NZ7 SI7 TO20 TO10 TO0 FI17 FI7 JA24 JA14 JA4 BR21 BR11 BR1 KI10 KI8 KI6 KI4 KI2 KI0 BA25 BA23 BA21 BA19 BA17 BA15 BA13 BA11 BA9 BA7 BA5 BA3 BA1 SA26 SA24 SA22 SA20 SA18 SA16 SA14 SA12 SA10 SA8 SA6 SA4 SA1 SA2 MA19 MA0 NZ8 SI8 SI9 MA20 MA1 NZ9 SI10 TO21 TO11 TO1 FI18 FI8 JA25 JA15 JA5 BR22 BR12 BR2 KI11 KI12 MA21 MA2 NZ10 SI11 SI12']

Output is analyzed in the conclusion.

***Part 2:***

***-Data Entry:*** The data entry process was identical to part 1, the only difference was I removed the island code to see how that would effect the clusters.

***-Running the program:*** I ran it the same as part 1.

***-Results:*** In contrast to part 1, I was shocked to see the difference between the two runs. After adjusting the dataset, this run also took 322 iterations, but I was left with only two clusters:

['MA26', 'MA25 MA24 MA23 MA22 MA21 MA20 MA19 MA18 MA17 MA16 MA15 MA14 MA13 MA12 MA11 MA10 MA9 MA8 MA7 MA6 MA5 MA4 MA3 MA2 MA1 MA0 NZ26 NZ25 NZ24 NZ23 NZ22 NZ21 NZ20 NZ19 NZ18 NZ17 NZ16 NZ15 NZ14 NZ13 NZ12 NZ11 NZ10 NZ9 NZ8 NZ7 NZ6 NZ5 NZ4 NZ3 NZ2 NZ1 SI26NZ0 SI25 SI24 SI23 SI22 SI21 SI20 SI19 SI18 SI17 SI16 SI15 SI14 SI13 SI12 SI11 SI10 SI9 SI8 SI7 SI6 SI5 SI4 SI3 SI2 SI1 SI0 TO26 TO25 TO24 TO23 TO22 TO21 TO20 TO19 TO18 TO17 TO16 TO15 TO14 TO13 TO12 TO11 TO10 TO9 TO8 TO7 TO6 TO5 TO4 TO3 TO2 TO1 TO0 FI26 FI25 FI24 FI23 FI22 FI21 FI20 FI19 FI18 FI17 FI16 FI15 FI14 FI13 FI12 FI11 FI10 FI9 FI8 FI7 FI6 FI5 FI4 FI3 FI2 FI1 FI0 JA26 JA25 JA24 JA23 JA22 JA21 JA20 JA19 JA18 JA17 JA16 JA15 JA14 JA13 JA12 JA11 JA10 JA9 JA8 JA7 JA6 JA5 JA4 JA3 JA2 JA1 JA0 BR26 BR25 BR24 BR23 BR22 BR21 BR20 BR19 BR18 BR17 BR16 BR15 BR14 BR13 BR12 BR11 BR10 BR9 BR8 BR7 BR6 BR5 BR4 BR3 BR2 BR1 BR0 KI26 KI25 KI24 KI23 KI22 KI21 KI20 KI19 KI18 KI17 KI16 KI15 KI14 KI13 KI12 KI11 KI10 KI9 KI8 KI7 KI6 KI5 KI4 KI3 KI2 KI1 KI0 BA26 BA25 BA24 BA23 BA22 BA21 BA20 BA19 BA18 BA17 BA16 BA15 BA14 BA13 BA12 BA11 BA10 BA9 BA8 BA7 BA6 BA5 BA4 BA3 BA2 BA1 BA0 SA26 SA25 SA24 SA23 SA22 SA21 SA20 SA19 SA18 SA17 SA16 SA15 SA14 SA13 SA12 SA11 SA10 SA9 SA8 SA7 SA6 SA5 SA4 SA3 SA2 SA0 PR26 PR25 PR24 PR23 PR22 PR21 PR20 PR19 PR18 PR17 PR16 PR15 PR14 PR13 PR12 PR11 PR10 PR9 PR8 PR7 PR6 PR5 PR4 PR3 PR2 PR1 PR0 MI26 MI25 MI24 MI23 MI22 MI21 MI20 MI19 MI18 MI17 MI16 MI15 MI14 MI13 MI12 MI11 MI10 MI9 MI8 MI7 MI6 MI5 MI4 MI3 MI2 MI0 MI1 SA1']

***ID3:***

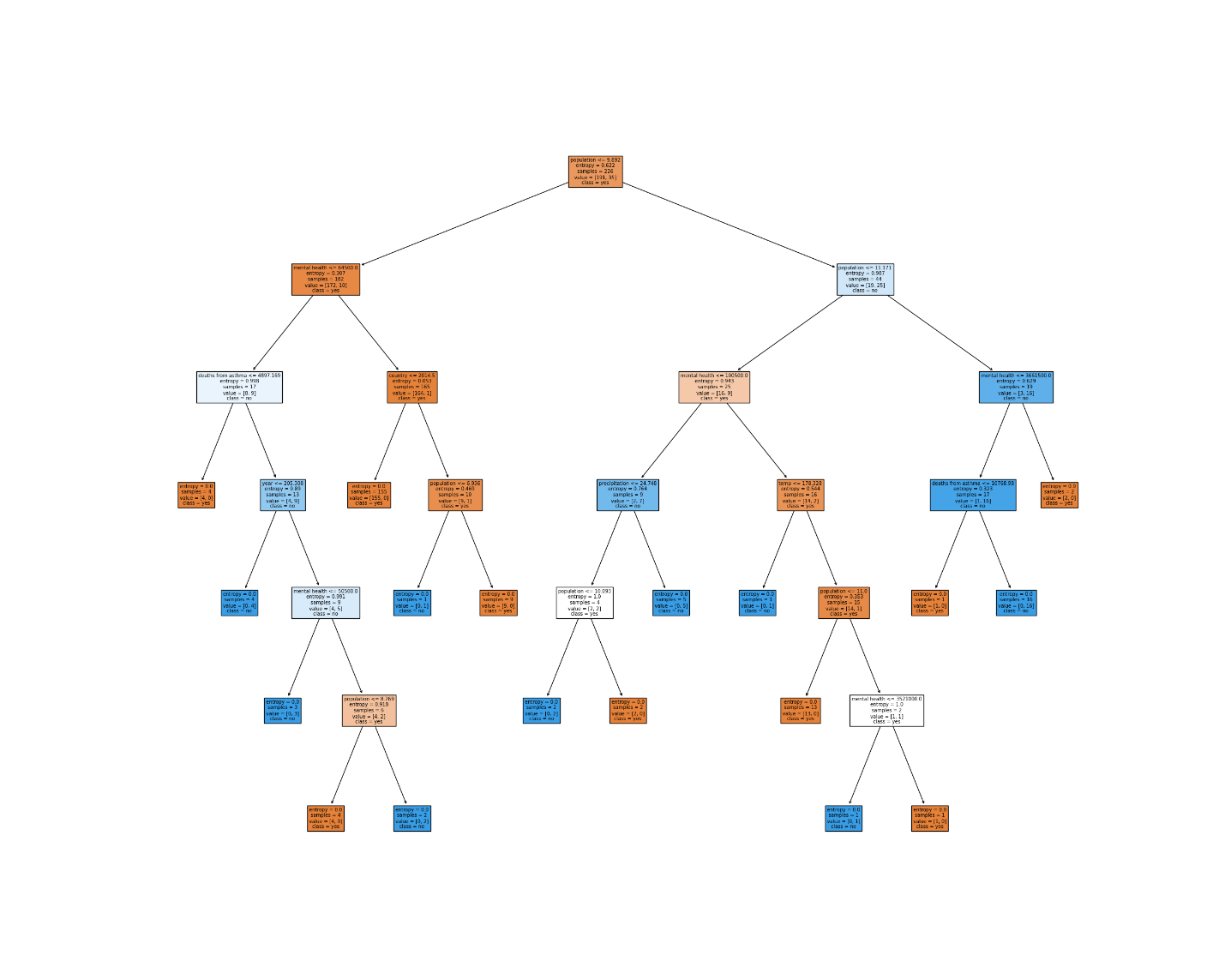
***-Program:*** The program I chose for this algorithm was ‘classification2.py’, provided by the professor and found on Blackboard. I chose to use this algorithm because I liked the functionality that it enabled; it output a picture of the tree based on the data. I really enjoyed this feature because it allows me to visually see what the tree is doing. I chose this over ‘classification.py’, also provided by the professor, because ‘classification2.py’ handled continuous attributes better, as all of my attributes were continuous.

***Part 1:***

***-Data Entry:*** To read in this data, I had to additionally preprocess my dataset to include another column labeled ‘class val’. This column was chosen because I wanted to see if the population from the previous year had increased or decreased based on the population levels of the previous year. To achieve this, I wrote an addition to the program I used to calculate averages in the previous algorithm and added some additional functionality. The program reads in the data in the following format: ‘val, year’. It stores the values and years in 2 parallel arrays, then the program will traverse both arrays, storing two values and finding the difference between them. If the difference is positive, the output is ‘y’, meaning the population increased, and if the output is negative, the output is ‘n’, meaning the population decreased. I then transferred these values to my excel sheet; these values were replaced with a 0 for ‘y’ and a 1 for ‘n’.

***-Running the program:*** I moved all of the newly preprocessed data to another spreadsheet named ‘dataset.xslx’ in the same directory as my ‘classification2.py’, since that program reads data from a spreadsheet, and it calculated the decision tree for me. A fun feature of this program is you’re able to control the size of the testing and training set, so I set the test\_size to be .3, or 30%. It then output the tree to a PNG image titled ‘decision\_tree’.

***-Results:*** The first time I ran this algorithm, the tree was very large; I had input the data incorrectly. After fixing this mistake, the tree output became much smaller and more manageable: (this image is saved as ‘tree1\_noprune.png’)



One limitation of this program is that the output is a bit difficult to read. There are many different branches to this tree and that makes constructing a set of rules for this very difficult. The program does output the accuracy of the rules it generated based on the training set, and for this iteration, the accuracy was:



This accuracy can indicate that the data I pulled together was good and not too overfit; if the accuracy was 100%, that would indicate “too perfect” data, and why bother with a data mining algorithm?

***Part 2:***

***-Data Entry:*** Same as previous.

***-Running the program:*** There is a lovely pruning attribute you can add to the program when you call the method DecisionTreeClassifier(), and it’s ‘max-depth’. For this round, I set ‘max-depth’ to be 5 and tried to see if that would make it a bit easier to come up with some rules.

***-Results:*** (this image is saved as ‘tree1\_prune5.png’)

A picture containing diagram

Description automatically generated

The tree itself is easier to read this time, so let’s generate some rules.

Rules:

*Left side*

IF population <= 9.982 and mental health incidence <= 64500.0 and country <= 1993.5 and year <= 205.308 and mental health incidence <= 50500.0 then N

IF population <= 9.982 and mental health incidence <= 64500.0 and country <= 2014.5 and population <= 6.936, then Y

*Right side*

IF population <= 9.982 and population <= 11.171 and deaths from asthma <= 10390.823 and precipitation <= 24.748 and deaths from asthma <= 9844.444, then Y

IF population <= 9.982 and population <= 11.171 and deaths from asthma <= 10390.823 and population <= 9.895 and population <= 11.0 then Y

IF population <= 9.982 and population <= 11.171 and mental health <= 3661500.0 and deaths from asthma <= 10768.90, then N.

The accuracy for this iteration:



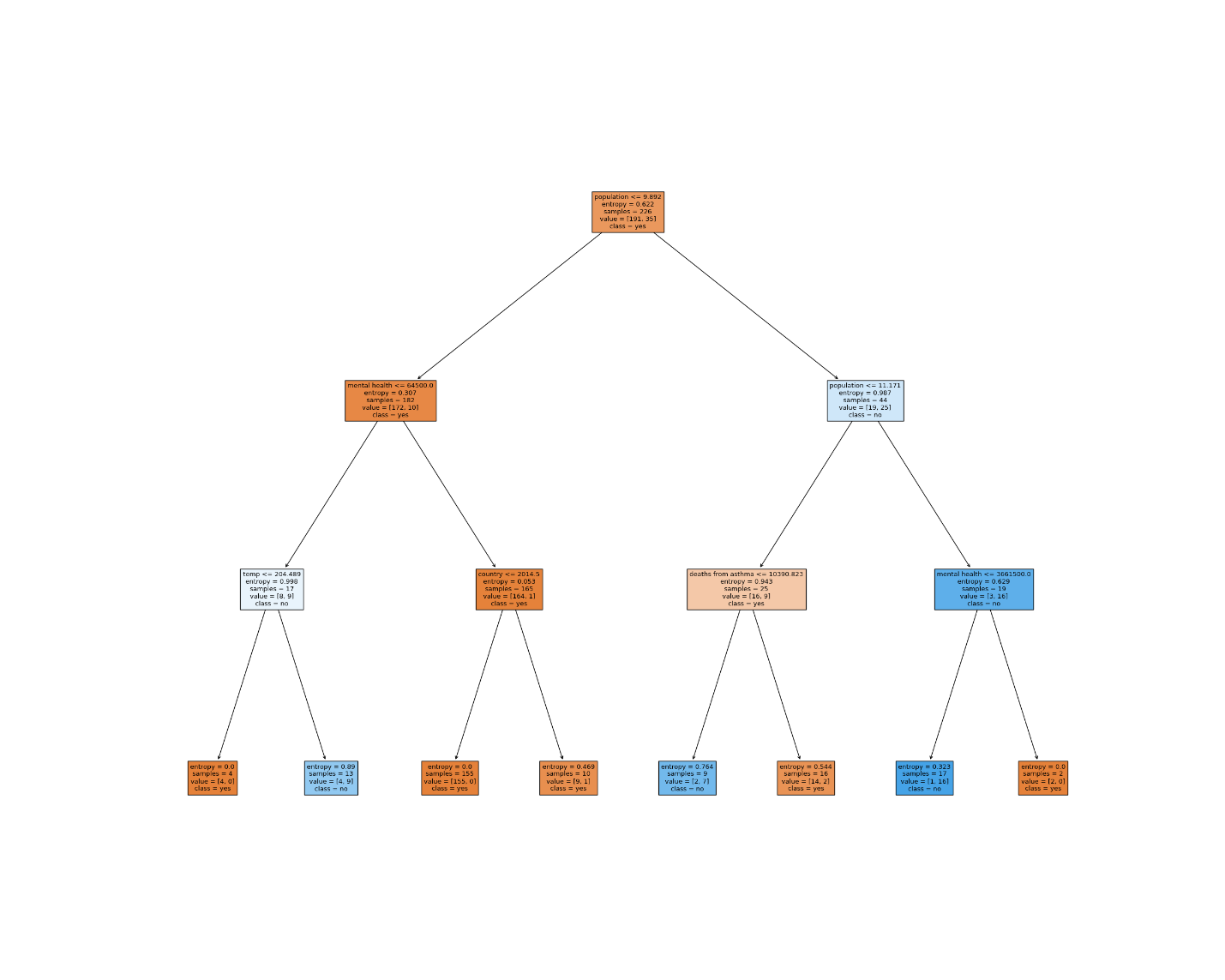
To my surprise, it is lower than last time.

***Part 3:***

***-Data Entry:*** Same as previous.

***-Running the program:*** ‘max-depth’ was set to 3.

***-Results:*** (this image is saved as ‘tree1\_prune3.png’)



This one is much easier to read.

Rules:

*Left side*

IF population <= 9.982 and mental health incidence <= 64500 and temp <= 204.489 then N

IF population <= 9.982 and mental health incidence <= 64500 and country <= 2014.5 then Y

*Right side*

IF population <= 9.982 and population <= 11.171 and deaths from asthma is less than 10390.823, then Y

IF IF population <= 9.982 and population <= 11.171 and mental health incidence <= 3661500.0 then N

Accuracy for this round:



Another surprise, this accuracy is the same as the previous.

***Conclusion-Analysis of output:***

I think my main issues I’m having with understanding my outputs comes down to the attributes that I’ve chosen for this project. I chose attributes that really aren’t too closely related; how are population and asthma prevalence related? I should have been asking those questions in the beginning and chose attributes that made more sense together. I am having trouble understanding why the clustering algorithm put *those* clusters together; I cannot find many similarities. I am thinking maybe the algorithm weighted the temperature more than anything, since some of the clusters seem close together based on their temperature, but I am unsure. For the ID3 results, they are a bit better to understand. I wanted to know what attributes were influencing whether a population was increasing or decreasing, and based on the output from prune = 5, I can have a better understanding. Mental health incidence appeared 4 times in the rules, deaths from asthma appeared 4 times, and population appeared *numerous* times. It can be concluded that the attributes that have an effect on whether the population increases or decreases are mental health indicators, deaths from asthma, and of course, population. I would say that my questions weren’t answered *too* well, but they were answered to an extent.