K-means clustering Project Report

Jihoon Bae

Language used : Python

Original:



K = 3:



Only expresses the basic shades.

K = 4:



Sky and land are divided out of gray shades

K = 5:



Sea and cloud are expressed with light blue

K = 6:



Details added for mountain

K = 7:



Shade added for mountain.

K = 8:



Shade added for wave.

K = 9:



Dark part of the sky is expressed with dark blue.

K = 10:



Can’t really tell the big difference. Mountain color is bit changed.

I will be using at least 5 clusters. Because that’s the point where we can identify sky, land, sea, and wave with minimum colors. All ks after 5 was almost about details. If I don’t necessarily need the detail, I would use 5 clusters.

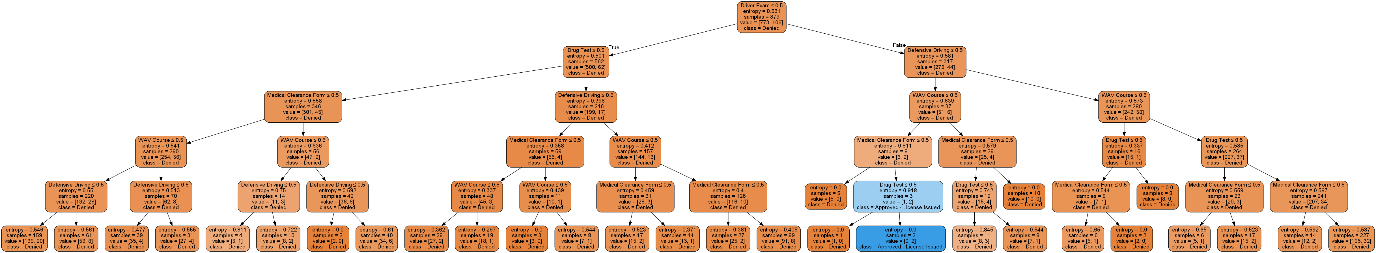
**Extra**

I first tried the driver license approval based on driver’s

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Drug Test, | WAV Course, | Defensive Driving, | Driver Exam, | Medical Clearance Form |

But the result was so random because there was so many people approved while they shouldn’t be.

Data was from <https://data.world/city-of-ny/dpec-ucu7/workspace/file?filename=tlc-new-driver-application-status-1.csv>



The tree looked like this.

I tried all over because the result was a mess.

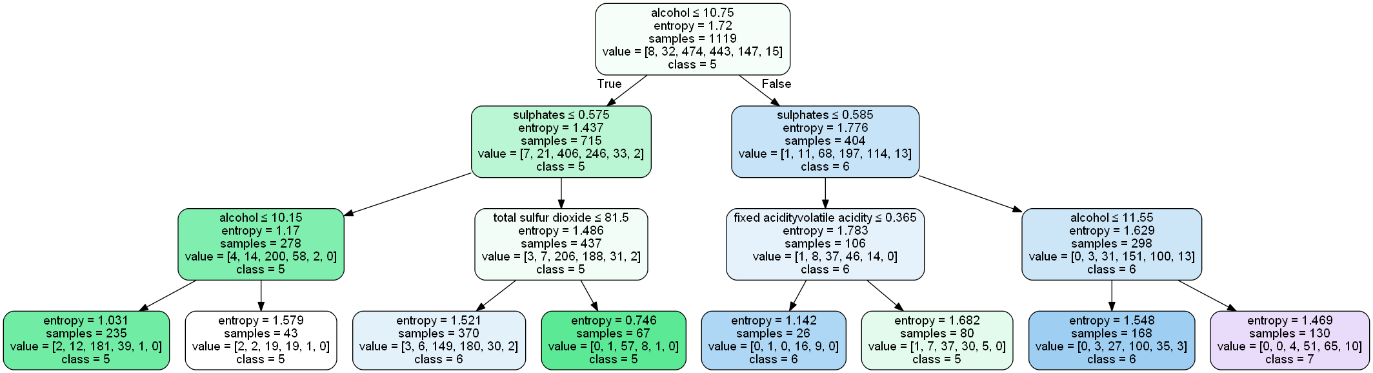
Then I changed the topic to red wine quality classification.

Data is from <https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009/data>

Quality label is classified base on

 ['fixed acidity volatile acidity','citric acid','residual sugar','chlorides','free sulfur dioxide','total sulfur dioxide','density','pH','sulphates','alcohol']

I chose the data set because I learned a lesson if I choose some data that has no strict standard for the label, I won’t get fine tree.



This is the tree. Label was ranged from 3 to 8 but the tree gets too huge to go through all. So I only went through 3rd depth.