# Remzi Meri? Ceylan

May 16, 2018

## 1 Veri Madencilii Ödevi - Bahar 2018

## 2 Remzi Meriç Ceylan

---

### 2.1 1.Özet

Projenin temel ksmlar, veri tanmlama, algoritma tantm ve modellemesi yapld,. Veri katagorik olduu için gruplama yapmakta zorlanld ve parametre seçimi, yorumlama yaplamad.

### 2.2 2.Giri

```
In [1]: import numpy as np
        import pandas as pd
        from sklearn.cluster import KMeans, DBSCAN, AgglomerativeClustering
        from IPython.display import Math, Latex
        import matplotlib.pyplot as plt

        from IPython.display import display
        pd.options.display.max_columns = None
```

```
In [2]: flags = pd.read_csv("flags.csv", header=None)
        flags.columns = list(range(1,31))
        flags.index = flags[1]
        del flags[1]
        flags.head(10)
```

Out[2]:

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | \ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Afghanistan | 5 | 1 | 648 | 16 | 10 | 2 | 0 | 3 | 5 | 1 | 1 | 0 | 1 | 1 | 1 | |
| Albania | 3 | 1 | 29 | 3 | 6 | 6 | 0 | 0 | 3 | 1 | 0 | 0 | 1 | 0 | 1 | |
| Algeria | 4 | 1 | 2388 | 20 | 8 | 2 | 2 | 0 | 3 | 1 | 1 | 0 | 0 | 1 | 0 | |
| American-Samoa | 6 | 3 | 0 | 0 | 1 | 1 | 0 | 0 | 5 | 1 | 0 | 1 | 1 | 1 | 0 | |
| Andorra | 3 | 1 | 0 | 0 | 6 | 0 | 3 | 0 | 3 | 1 | 0 | 1 | 1 | 0 | 0 | |
| Angola | 4 | 2 | 1247 | 7 | 10 | 5 | 0 | 2 | 3 | 1 | 0 | 0 | 1 | 0 | 1 | |
| Anguilla | 1 | 4 | 0 | 0 | 1 | 1 | 0 | 1 | 3 | 0 | 0 | 1 | 0 | 1 | 0 | |

```
Antigua-Barbuda   1   4     0   0   1   1   0   1   5   1   0   1   1   1   1
Argentina         2   3  2777  28   2   0   0   3   2   0   0   1   0   1   0
Argentine         2   3  2777  28   2   0   0   3   3   0   0   1   1   1   0

                 17     18  19  20  21  22  23  24  25  26  27  28     29  \
1
Afghanistan       0  green   0   0   0   0   1   0   0   1   0   0  black
Albania           0    red   0   0   0   0   1   0   0   0   1   0    red
Algeria           0  green   0   0   0   0   1   1   0   0   0   0  green
American-Samoa    1   blue   0   0   0   0   0   0   1   1   1   0   blue
Andorra           0   gold   0   0   0   0   0   0   0   0   0   0   blue
Angola            0    red   0   0   0   0   1   0   0   1   0   0    red
Anguilla          1  white   0   0   0   0   0   0   0   0   1   0  white
Antigua-Barbuda   0    red   0   0   0   0   1   0   1   0   0   0  black
Argentina         0   blue   0   0   0   0   0   0   0   0   0   0   blue
Argentine         0   blue   0   0   0   0   1   0   0   0   0   0   blue

                    30
1
Afghanistan      green
Albania            red
Algeria          white
American-Samoa     red
Andorra            red
Angola           black
Anguilla          blue
Antigua-Barbuda    red
Argentina         blue
Argentine         blue
```

In [3]:
```
def enumerate_dataframe(df):
    for column in df:
        if type(df[column][0]) == np.int64:
            pass
        else:
            print("Column Name: ",df[column].name)
            print(df[column].unique())
            key=[]
            value=[]
            for v, k in enumerate(df[column].unique()):
                key.append(k)
                value.append(v)
                print(k,"\t", v)
            print("\n")
            df[column] = df[column].map(dict(zip(key, value)))
    return df
```

In [4]:
```
flags = enumerate_dataframe(flags)
```

```
Column Name:  18
['green' 'red' 'blue' 'gold' 'white' 'orange' 'black' 'brown']
green          0
red          1
blue          2
gold          3
white          4
orange          5
black          6
brown          7


Column Name:  29
['black' 'red' 'green' 'blue' 'white' 'orange' 'gold']
black          0
red          1
green          2
blue          3
white          4
orange          5
gold          6


Column Name:  30
['green' 'red' 'white' 'black' 'blue' 'gold' 'orange' 'brown']
green          0
red          1
white          2
black          3
blue          4
gold          5
orange          6
brown          7
```

```
In [5]: flags.columns = [
        "landmass",
        "zone",
        "area_km2",
        "population",
        "language",
        "religion",
        "bars_count",
        "stripes_count",
        "colours_count",
        "red_exist",
```

```
        "green_exist",
        "blue_exist",
        "gold_yellow_exist",
        "white_exist",
        "black_exist",
        "orange_brown_exist",
        "mainhue_predominant_colour",
        "circles_count",
        "crosses_count",
        "saltires_count",
        "quarters_count",
        "sunstars_count",
        "crescent_exist",
        "triangle_exist",
        "icon_exist",
        "animate_exist",
        "text_exist",
        "topleft_colour",
        "botleft_colour"
        ]

        flags.index.name = "country_name"
```

In [6]:
```
        """
        landmass                  -> 1=N.America, 2=S.America, 3=Europe, 4=Africa, 5=Asia, 6=Oc
        zone                      -> 1=NE, 2=SE, 3=SW, 4=NW
        language                  -> 1=English, 2=Spanish, 3=French, 4=German, 5=Slavic, 6=Othe
                                  9=Japanese/Turkish/Finnish/Magyar, 10=Others
        religion                  -> 0=Catholic, 1=Other Christian, 2=Muslim, 3=Buddhist, 4=Hin
        mainhue_predominant_colour -> green=0, red=1, blue=2, gold=3, white=4, orange=5, black=6
        topleft colour            -> black=0, red=1, green=2, blue=3, white=4, orange=5, gold=6
        botleft_colour            -> green=0, red=1, white=2, black=3, blue=4, gold=5, orange=6
        """
        flags.head(10)
```

Out[6]:

| country_name | landmass | zone | area_km2 | population | language | religion |
|---|---|---|---|---|---|---|
| Afghanistan | 5 | 1 | 648 | 16 | 10 | 2 |
| Albania | 3 | 1 | 29 | 3 | 6 | 6 |
| Algeria | 4 | 1 | 2388 | 20 | 8 | 2 |
| American-Samoa | 6 | 3 | 0 | 0 | 1 | 1 |
| Andorra | 3 | 1 | 0 | 0 | 6 | 0 |
| Angola | 4 | 2 | 1247 | 7 | 10 | 5 |
| Anguilla | 1 | 4 | 0 | 0 | 1 | 1 |
| Antigua-Barbuda | 1 | 4 | 0 | 0 | 1 | 1 |
| Argentina | 2 | 3 | 2777 | 28 | 2 | 0 |
| Argentine | 2 | 3 | 2777 | 28 | 2 | 0 |

| country_name | bars_count | stripes_count | colours_count | red_exist | \ |
|---|---|---|---|---|---|
| Afghanistan | 0 | 3 | 5 | 1 | |
| Albania | 0 | 0 | 3 | 1 | |
| Algeria | 2 | 0 | 3 | 1 | |
| American-Samoa | 0 | 0 | 5 | 1 | |
| Andorra | 3 | 0 | 3 | 1 | |
| Angola | 0 | 2 | 3 | 1 | |
| Anguilla | 0 | 1 | 3 | 0 | |
| Antigua-Barbuda | 0 | 1 | 5 | 1 | |
| Argentina | 0 | 3 | 2 | 0 | |
| Argentine | 0 | 3 | 3 | 0 | |

| country_name | green_exist | blue_exist | gold_yellow_exist | white_exist | \ |
|---|---|---|---|---|---|
| Afghanistan | 1 | 0 | 1 | 1 | |
| Albania | 0 | 0 | 1 | 0 | |
| Algeria | 1 | 0 | 0 | 1 | |
| American-Samoa | 0 | 1 | 1 | 1 | |
| Andorra | 0 | 1 | 1 | 0 | |
| Angola | 0 | 0 | 1 | 0 | |
| Anguilla | 0 | 1 | 0 | 1 | |
| Antigua-Barbuda | 0 | 1 | 1 | 1 | |
| Argentina | 0 | 1 | 0 | 1 | |
| Argentine | 0 | 1 | 1 | 1 | |

| country_name | black_exist | orange_brown_exist | mainhue_predominant_colour | \ |
|---|---|---|---|---|
| Afghanistan | 1 | 0 | 0 | |
| Albania | 1 | 0 | 1 | |
| Algeria | 0 | 0 | 0 | |
| American-Samoa | 0 | 1 | 2 | |
| Andorra | 0 | 0 | 3 | |
| Angola | 1 | 0 | 1 | |
| Anguilla | 0 | 1 | 4 | |
| Antigua-Barbuda | 1 | 0 | 1 | |
| Argentina | 0 | 0 | 2 | |
| Argentine | 0 | 0 | 2 | |

| country_name | circles_count | crosses_count | saltires_count | quarters_count | \ |
|---|---|---|---|---|---|
| Afghanistan | 0 | 0 | 0 | 0 | |
| Albania | 0 | 0 | 0 | 0 | |
| Algeria | 0 | 0 | 0 | 0 | |
| American-Samoa | 0 | 0 | 0 | 0 | |
| Andorra | 0 | 0 | 0 | 0 | |
| Angola | 0 | 0 | 0 | 0 | |
| Anguilla | 0 | 0 | 0 | 0 | |

```
Antigua-Barbuda              0              0              0              0
Argentina                    0              0              0              0
Argentine                    0              0              0              0

                sunstars_count  crescent_exist  triangle_exist  icon_exist  \
country_name
Afghanistan                  1               0               0           1
Albania                      1               0               0           0
Algeria                      1               1               0           0
American-Samoa               0               0               1           1
Andorra                      0               0               0           0
Angola                       1               0               0           1
Anguilla                     0               0               0           0
Antigua-Barbuda              1               0               1           0
Argentina                    0               0               0           0
Argentine                    1               0               0           0

                animate_exist  text_exist  topleft_colour  botleft_colour
country_name
Afghanistan                 0           0               0               0
Albania                     1           0               1               1
Algeria                     0           0               2               2
American-Samoa              1           0               3               1
Andorra                     0           0               3               1
Angola                      0           0               1               3
Anguilla                    1           0               4               4
Antigua-Barbuda             0           0               0               1
Argentina                   0           0               3               4
Argentine                   0           0               3               4
```

### 2.3 3.Verinin Tanm

Title: Flag database

Source Information -- Creators: Collected primarily from the "Collins Gem Guide to Flags": Collins Publishers (1986). -- Donor: Richard S. Forsyth 8 Grosvenor Avenue Mapperley Park Nottingham NG3 5DX 0602-621676 -- Date: 5/15/1990

Past Usage: -- None known other than what is shown in Forsyth's PC/BEAGLE User's Guide.

Relevant Information: -- This data file contains details of various nations and their flags. In this file the fields are separated by spaces (not commas). With this data you can try things like predicting the religion of a country from its size and the colours in its flag.
-- 10 attributes are numeric-valued. The remainder are either Boolean- or nominal-valued.

Number of Instances: 194

Number of attributes: 30 (overall)

Attribute Information:

1. name Name of the country concerned
2. landmass 1=N.America, 2=S.America, 3=Europe, 4=Africa, 5=Asia, 6=Oceania
3. zone Geographic quadrant, based on Greenwich and the Equator 1=NE, 2=SE, 3=SW, 4=NW

4. area in thousands of square km
5. population in round millions
6. language 1=English, 2=Spanish, 3=French, 4=German, 5=Slavic, 6=Other Indo-European, 7=Chinese, 8=Arabic, 9=Japanese/Turkish/Finnish/Magyar, 10=Others
7. religion 0=Catholic, 1=Other Christian, 2=Muslim, 3=Buddhist, 4=Hindu, 5=Ethnic, 6=Marxist, 7=Others
8. bars Number of vertical bars in the flag
9. stripes Number of horizontal stripes in the flag
10. colours Number of different colours in the flag
11. red 0 if red absent, 1 if red present in the flag
12. green same for green
13. blue same for blue
14. gold same for gold (also yellow)
15. white same for white
16. black same for black
17. orange same for orange (also brown)
18. mainhue predominant colour in the flag (tie-breaks decided by taking the topmost hue, if that fails then the most central hue, and if that fails the leftmost hue) (green=0, red=1, blue=2, gold=3, white=4, orange=5, black=6, brown=7)
19. circles Number of circles in the flag
20. crosses Number of (upright) crosses
21. saltires Number of diagonal crosses
22. quarters Number of quartered sections
23. sunstars Number of sun or star symbols
24. crescent 1 if a crescent moon symbol present, else 0
25. triangle 1 if any triangles present, 0 otherwise
26. icon 1 if an inanimate image present (e.g., a boat), otherwise 0
27. animate 1 if an animate image (e.g., an eagle, a tree, a human hand) present, 0 otherwise
28. text 1 if any letters or writing on the flag (e.g., a motto or slogan), 0 otherwise
29. topleft colour in the top-left corner (moving right to decide tie-breaks) (black=0, red=1, green=2, blue=3, white=4, orange=5, gold=6)
30. botright Colour in the bottom-left corner (moving left to decide tie-breaks) (green=0, red=1, white=2, black=3, blue=4, gold=5, orange=6, brown=7)

Missing values: None

""" 1. country_name 2. landmass 3. zone 4. area_km2 5. population 6. language 7. religion 8. bars_count 9. stripes_count 10. colours_count 11. red_exist 12. green_exist 13. blue_exist 14. gold_yellow_exist 15. white_exist 16. black_exist 17. orange_brown_exist 18. mainhue_predominant_colour 19. circles_count 20. crosses_count 21. saltires_count 22. quarters_count 23. sunstars_count 24. crescent_exist 25. triangle_exist 26. icon_exist 27. animate_exist
28. text_exist 29. topleft_colour
30. botleft_colour """

## 2.4 4.Yöntemler

### 2.4.1 4.1 KMeans

The KMeans algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares. This algorithm