|  |
| --- |
|  |
| **Grupowanie państw na podstawie wyglądu ich flag** |
|  |
| **Podstawy reprezentacji i analizy danych – PROJEKT Grupa 16** |

January 24, 2021

Wykonali:

Gulewski Rafał, Nowakowski Marek

Spis treści

[1. Wstęp 2](#_Toc62394380)

[2. Zbiór danych 2](#_Toc62394381)

[3. Eksploracja danych 2](#_Toc62394382)

[4. Wybór klasyfikatora 2](#_Toc62394383)

[5. Klasyfikacja 2](#_Toc62394384)

[6. Podsumowanie 2](#_Toc62394385)

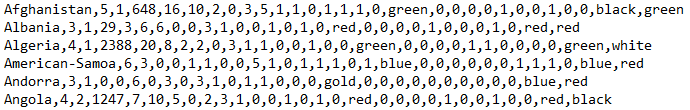
# Wstęp

Celem projektu jest stworzenie modelu klasyfikującego, przyporządkowującego każdemu krajowi religię, na podstawie wyglądu jego flagi. Wykorzystując informacje o flagach takie jak kolory, znajdujące się na fladze kształty, lub symbole, model nauczy się klasyfikacji religii na zbiorze treningowym, by następnie zmierzyć się z modelem testowym. Konieczna będzie analiza zbioru ze względu na występowanie danych atrybutów oraz ich korelacje, odpowiedni podział na zbiór uczący i weryfikacyjny i porównanie wyników różnych metod grupowania. Wynikiem klasyfikacji będzie zbiór państw rozdzielony na poszczególne grupy, w zależności przynależności do danej religii.

# Zbiór danych

# 2.1. Informacje o zbiorze danych

W projekcie został wykorzystany zbiór danych na temat flag i państw ze zbioru uczenia maszynowego „UCI Machine Learning Repository”. Stworzony został na podstawie książki „Collins Gem Guide to Flags”. Zbiór ten powstał z myślą o klasyfikacji. Składa się z 194 obiektów opisanych przez 30 atrybutów. W zbiorze nie brakuje żadnych danych. Dane są z 1990 r., co powoduje pewne nieaktualności w pojedynczych przypadkach takich jak np. zmieniona trzykrotnie od tego czasu flaga Kongo, lub obecność już nieistniejącej Czechosłowacji. Dane w surowej postaci są plikiem tekstowym, gdzie każdy obiekt zaczyna się od nowej linii a atrybuty oddzielone są przecinkami.



Rysunek Wycinek surowego pliku z danymi

# 2.2. Informacje o atrybutach

**Atrybuty to kolejno:**

1. **Nazwa państwa**
2. Położenie na kontynencie:
   1. Ameryka północna,
   2. Ameryka południowa,
   3. Europa,
   4. Afryka,
   5. Azja,
   6. Oceania
3. Strefa geograficzna oparta na Greenwich i równiku:
   1. Północny wschód,
   2. Południowy wschód,
   3. Południowy zachód,
   4. Północny zachód
4. Powierzchnia w tysiącach
5. Populacja zaokrąglana do milionów
6. Język główny lub przynależność do rodziny języków:
7. Angielski,
8. Hiszpański,
9. Francuski,
10. Niemiecki,
11. Słowiański,
12. Inny Indo-Europejski,
13. Chiński,
14. Arabski,
15. Japoński/Turecki/Fiński/Węgierski,
16. Inne
17. **Religia**:
    1. Katolicka,
    2. Inne chrześcijańskie,
    3. Muzułmańska,
    4. Buddyzm,
    5. Hinduizm,
    6. Etniczna,
    7. Marksistowska,
    8. Inne

Kolejne atrybuty opisują flagę:

1. Liczba pionowych pasków
2. Liczba poziomych pasków
3. Liczba różnych kolorów

Obecność kolorów (1 – obecny na fladze, 0 – brak danego koloru na fladze):

1. Czerwony
2. Zielony
3. Niebieski
4. Złoty lub żółty
5. Biały
6. Czarny
7. Pomarańczowy lub brązowy
8. Główny odcień: kolor dominujący (Remis rozwiązywany poprzez próbę pobrania koloru na górze flagi, następnie centralnego, następnie skrajnie z lewej)

Liczba symboli:

1. Liczba okręgów
2. Liczba pionowych krzyży
3. Liczba diagonalnych krzyży
4. Liczba kwartałów (wydzielonych sekcji)
5. Liczba słońc lub symboli gwiazd

Obecność symboli (1 – obecny na fladze, 0 – brak):

1. Obecność półksiężyców
2. Obecność trójkątów
3. Obecność innych symboli nieożywionych (np. łódka)
4. Obecność symboli ożywionych (np. zwierzę, ludzka ręka)
5. Obecność tekstu

Kolory w rogach:

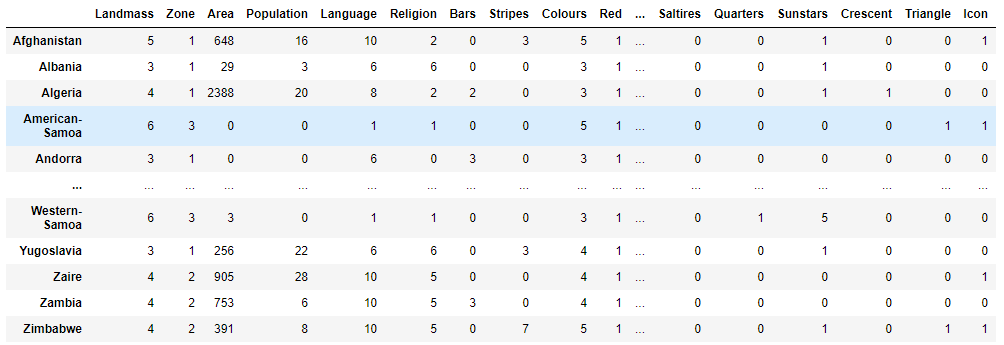
1. Kolor w lewym górnym rogu
2. Kolor w lewym dolnym rogu

# 2.3. Przygotowanie danych

Dane nie wymagają dużej obróbki, po wczytaniu ich do ramki danych, zamieniamy dane tekstowe na liczbowe – są to kolory, zapisane w pliku nazwami angielskimi. Przyporządkowujemy kolorom odpowiednie numery:

1. Czerwony
2. Niebieski
3. Zielony
4. Biały
5. Złoty
6. Czarny
7. Pomarańczowy
8. Brązowy

Po dokonaniu zamiany tabela z danymi prezentuje się w następujący sposób:



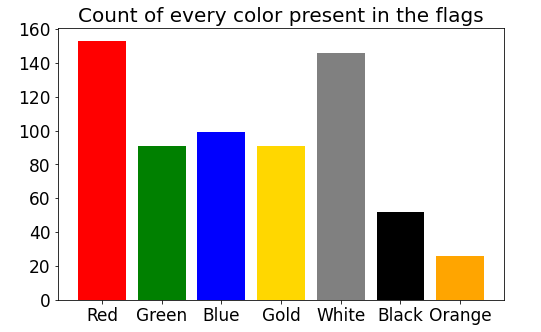
Rysunek 2 Przedstawienie danych

# Eksploracja danych

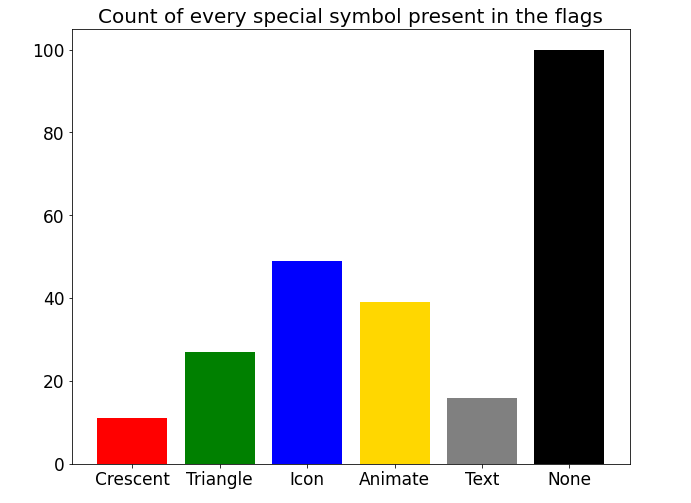
# 3.1. Wizualizacja zbioru

# 

Wykres Liczba państw i dominujących w nich religii



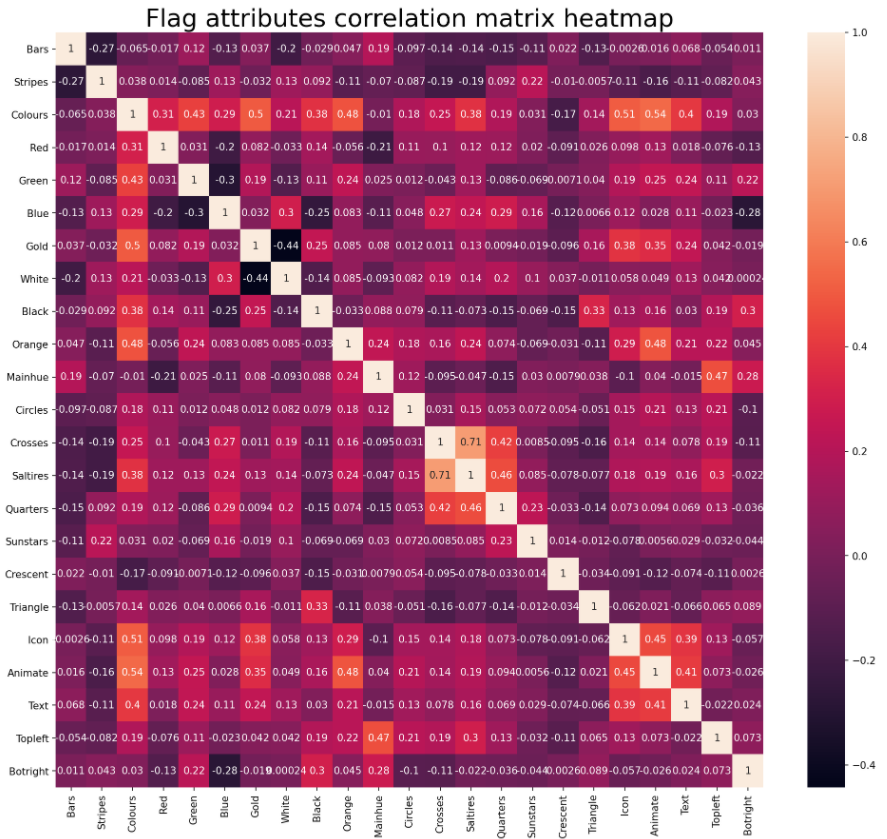
Wykres Liczba kolorów obecnych na flagach



Wykres Liczba symboli obecnych na flagach

Z powyższych wykresów dowiadujemy się wielu ciekawych informacji na temat zbioru danych. Większość flag nie zawiera żadnych symboli i składa się tylko i wyłącznie z kolorów. Najpopularniejszymi symbolami są przedmioty nieożywione oraz ludzie, o dziwo wśród flag znajduje się sporo trójkątów, a tekst i krzyże występują relatywnie rzadko. Wśród kolorów dominuje czerwony oraz biały, dosyć rzadko pojawia się pomarańczowy. Nie jest zaskakujące że wśród religii dominuje chrześcijaństwo, zaraz za nim znajduje się katolicyzm, popularnością również cieszy się muzułmanizm. Z racji że dane pochodzą z 1990 roku możemy również zauważyć jak w wielu państwach dominuje marksizm.

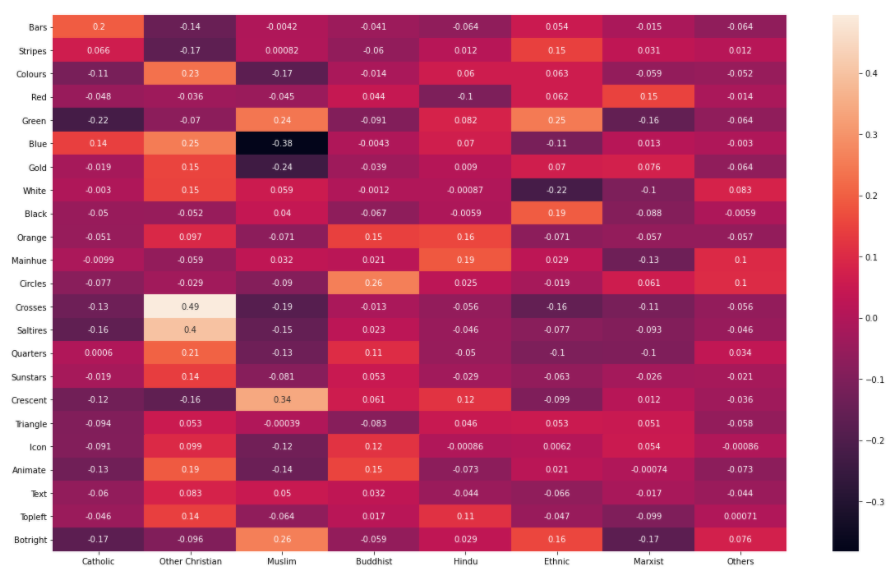
# 2.3. Korelacja danych



Rysunek Mapa ciepła korelacji atrybutów flag

Jak można zauważyć z mapy ciepła korelacji atrybutów dotyczących flagi, najbardziej zależnymi od siebie atrybutami są krzyże poziome i poprzeczne. Reszta atrybutów zachowuje stosunkowo niski współczynnik korelacji z wyłączeniem takich przypadków jak np. liczba kolorów i obecność symboli ożywionych na fladze lub ikon. Najmniej skorelowane są wystąpienia par kolorów biały i złoty, czy niebieski i zielony.

Bardziej interesującą nas korelacją jest zależność atrybutów od danej religii.



Rysunek Mapa ciepła korelacji atrybutów flag i religii

Choć współczynniki korelacji są dosyć niskie, możemy zauważyć ciekawe zależności takie jak obecność krzyża na fladze jest wysoko skorelowana z religiami chrześcijańskimi, lub brak kolorów niebieskiego i złotego w flagach krajów muzułmańskich. W flagach muzułmańskich często pojawia się jednak półksiężyc i dominuje kolor zielony. W krajach gdzie dominuje buddyzm często pojawiają się okręgi, a w krajach gdzie panuje ideologia marksistowska najczęstszym kolorem jest czerwony.

# Wybór klasyfikatora

//Wyniki poszczególnych klasyfikatorów dla losowych zestawów danych (podział na uczony/testowy) ze zbioru

# Klasyfikacja

//Wyniki klasyfikacji: macierz pomyłem, wykresy przynależności religijnej w danych grupach , wizualna prezentacja

# Podsumowanie

//wiadomo wnioski wyciągnięte z projekty czy można tak pogrupować i ogólnie takie tam