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

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.

2. 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.

```
Afghanistan,5,1,648,16,10,2,0,3,5,1,1,0,1,1,1,0,green,0,0,0,0,1,0,0,1,0,0,black,green
Albania,3,1,29,3,6,6,0,0,3,1,0,0,1,0,1,0,red,0,0,0,0,1,0,0,0,1,0,red,red
Algeria,4,1,2388,20,8,2,2,0,3,1,1,0,0,1,0,0,green,0,0,0,0,1,1,0,0,0,0,green,white
American-Samoa,6,3,0,0,1,1,0,0,5,1,0,1,1,1,0,1,blue,0,0,0,0,0,0,1,1,1,0,blue,red
Andorra,3,1,0,0,6,0,3,0,3,1,0,1,1,0,0,0,gold,0,0,0,0,0,0,0,0,0,0,blue,red
Angola,4,2,1247,7,10,5,0,2,3,1,0,0,1,0,1,0,red,0,0,0,0,1,0,0,1,0,0,red,black
```

Rysunek 1Wycinek 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 km^2
5. Populacja zaokrąglana do milionów
6. Język główny lub przynależność do rodziny języków:
    1) Angielski,
    2) Hiszpański,
    3) Francuski,
    4) Niemiecki,
    5) Słowiański,
    6) Inny Indo-Europejski,
    7) Chiński,
    8) Arabski,
    9) Japoński/Turecki/Fiński/Węgierski,
    10) Inne
7. **Religia**:
    1) Katolicka,
    2) Inne chrześcijańskie,
    3) Muzułmańska,
    4) Buddyzm,
    5) Hinduizm,
    6) Etniczna,
    7) Marksistowska,
    8) Inne

Kolejne atrybuty opisują flagę:

8. Liczba pionowych pasków
9. Liczba poziomych pasków
10. Liczba różnych kolorów

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

11. Czerwony
12. Zielony
13. Niebieski
14. Złoty lub żółty
15. Biały
16. Czarny
17. Pomarańczowy lub brązowy
18. Główny odcień: kolor dominujący (Remis rozwiązywany poprzez próbę pobrania koloru na górze flagi, następnie centralnego, następnie skrajnie z lewej)

Liczba symboli:

19. Liczba okręgów
20. Liczba pionowych krzyży
21. Liczba diagonalnych krzyży
22. Liczba kwartałów (wydzielonych sekcji)
23. Liczba słońc lub symboli gwiazd

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

24. Obecność półksiężyców
25. Obecność trójkątów
26. Obecność innych symboli nieożywionych (np. łódka)
27. Obecność symboli ożywionych (np. zwierzę, ludzka ręka)
28. Obecność tekstu

Kolory w rogach:

29. Kolor w lewym górnym rogu
30. 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

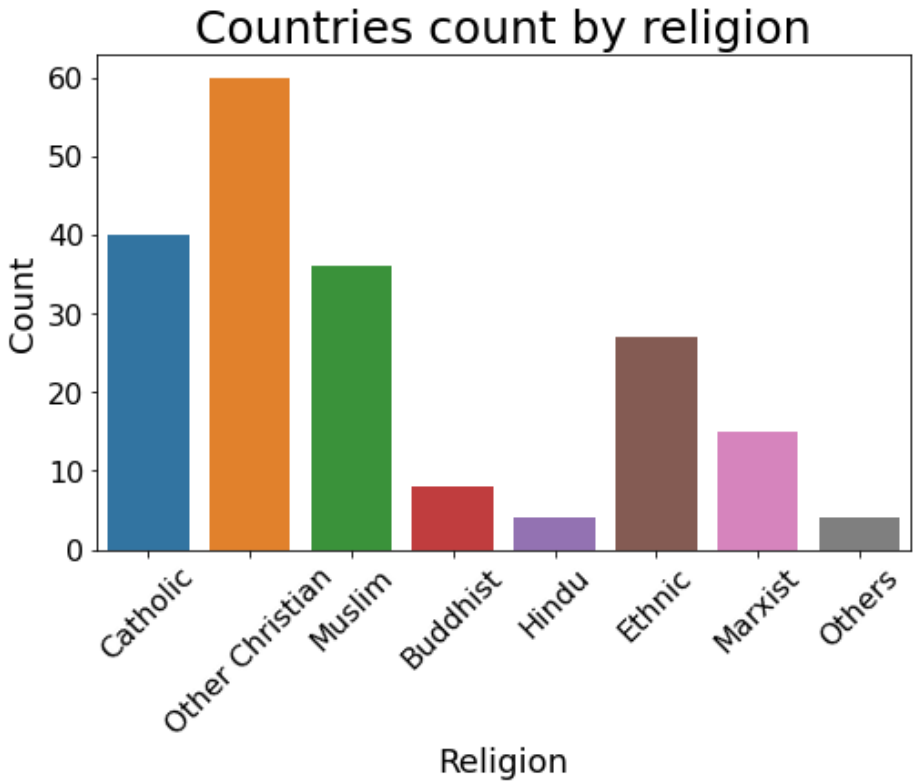
Po dokonaniu zamiany tabela z danymi prezentuje się w następujący sposób:

	Landmass	Zone	Area	Population	Language	Religion	Bars	Stripes	Colours	Red	...	Saltires	Quarters	Sunstars	Crescent	Triangle	Icon
Afghanistan	5	1	648	16	10	2	0	3	5	1	...	0	0	1	0	0	1
Albania	3	1	29	3	6	6	0	0	3	1	...	0	0	1	0	0	0
Algeria	4	1	2388	20	8	2	2	0	3	1	...	0	0	1	1	0	0
American-Samoa	6	3	0	0	1	1	0	0	5	1	...	0	0	0	0	1	1
Andorra	3	1	0	0	6	0	3	0	3	1	...	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Western-Samoa	6	3	3	0	1	1	0	0	3	1	...	0	1	5	0	0	0
Yugoslavia	3	1	256	22	6	6	0	3	4	1	...	0	0	1	0	0	0
Zaire	4	2	905	28	10	5	0	0	4	1	...	0	0	0	0	0	1
Zambia	4	2	753	6	10	5	3	0	4	1	...	0	0	0	0	0	0
Zimbabwe	4	2	391	8	10	5	0	7	5	1	...	0	0	1	0	1	1

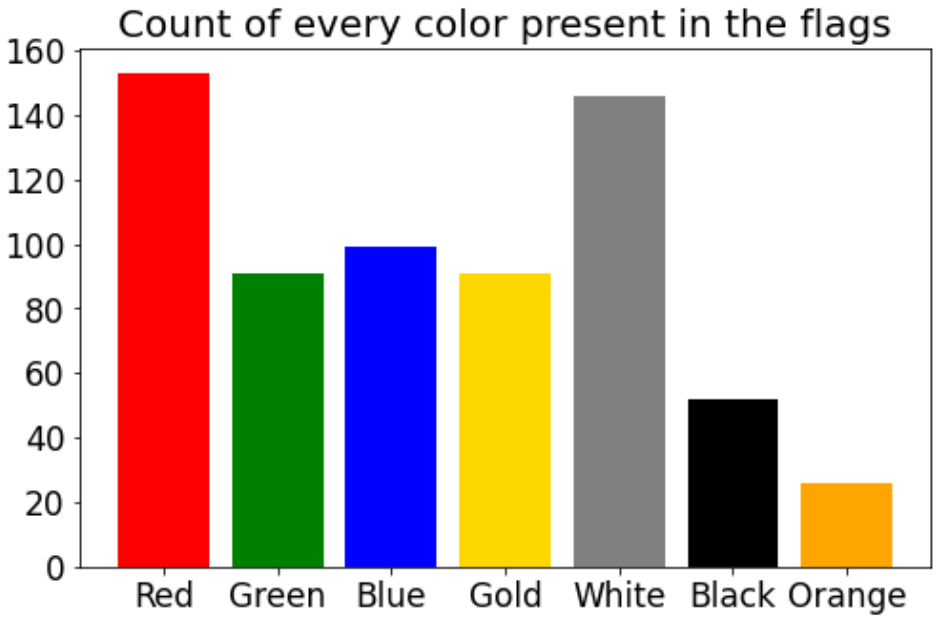
Rysunek 2 Przedstawienie danych

3. Eksploracja danych

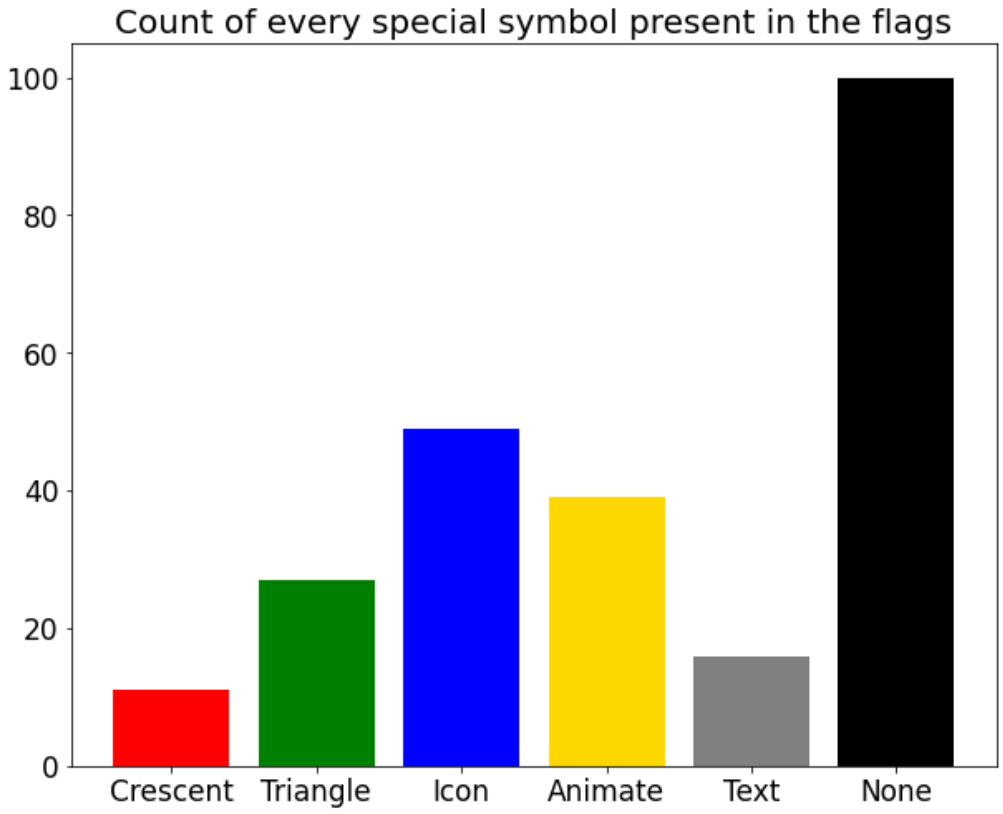
3.1. Wizualizacja zbioru



Wykres 1 Liczba państw i dominujących w nich religii



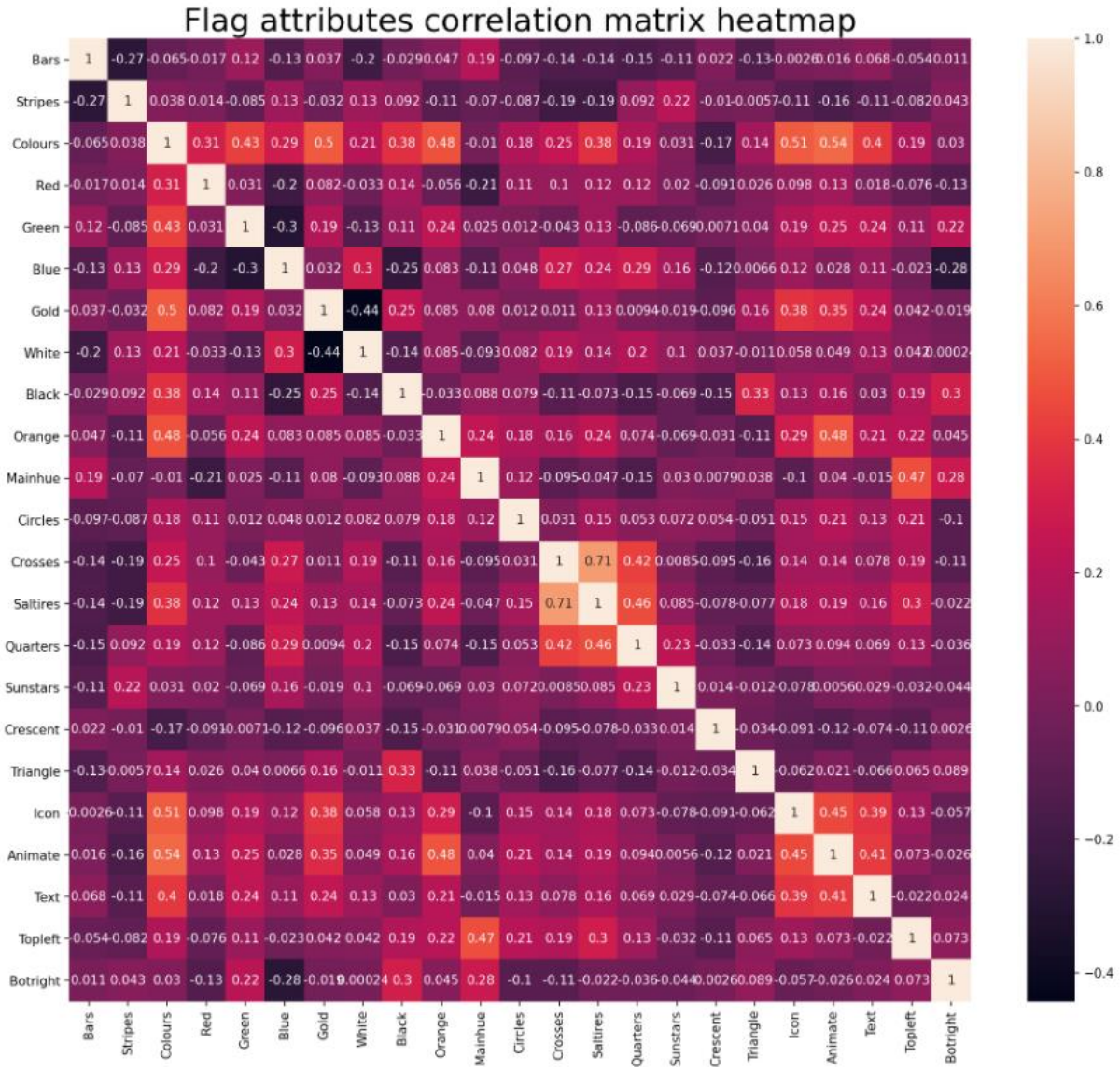
Wykres 2 Liczba kolorów obecnych na flagach



Wykres 3 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.

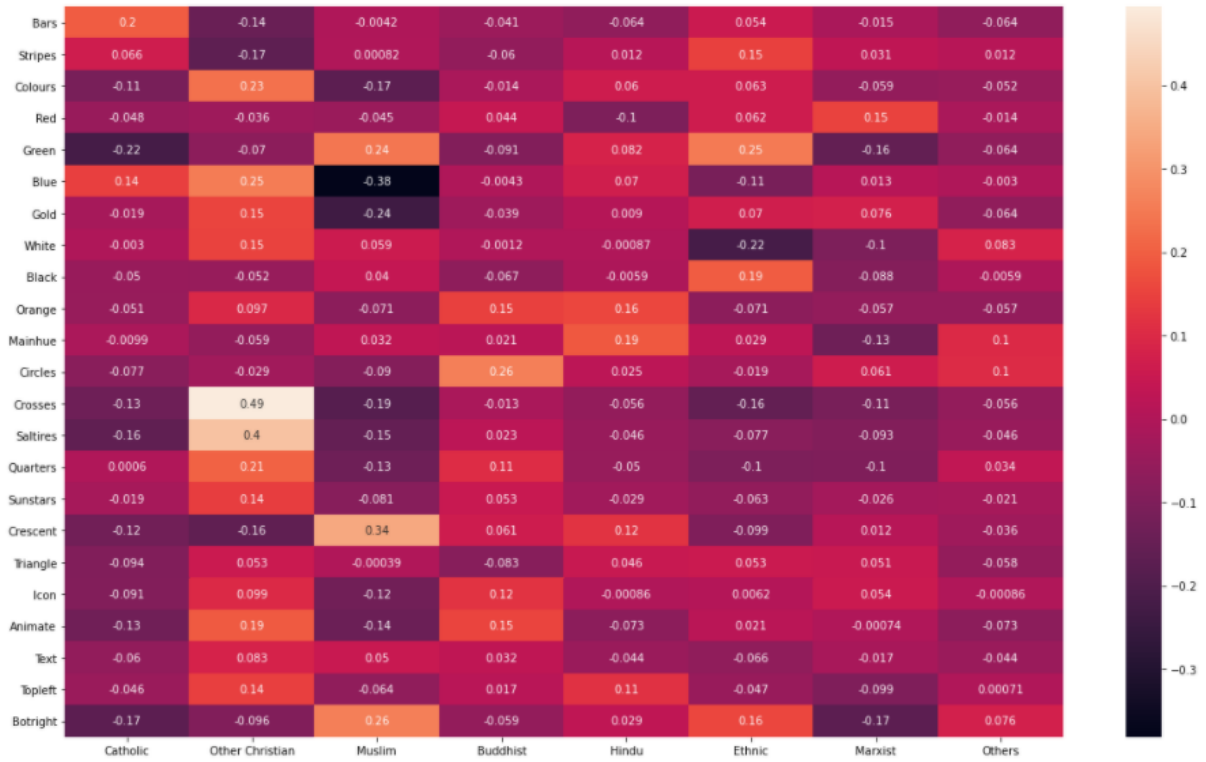
3.2. Korelacja danych



Rysunek 3 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 4 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.

4. Wybór klasyfikatora

4.1. Sposób wyboru klasyfikatora

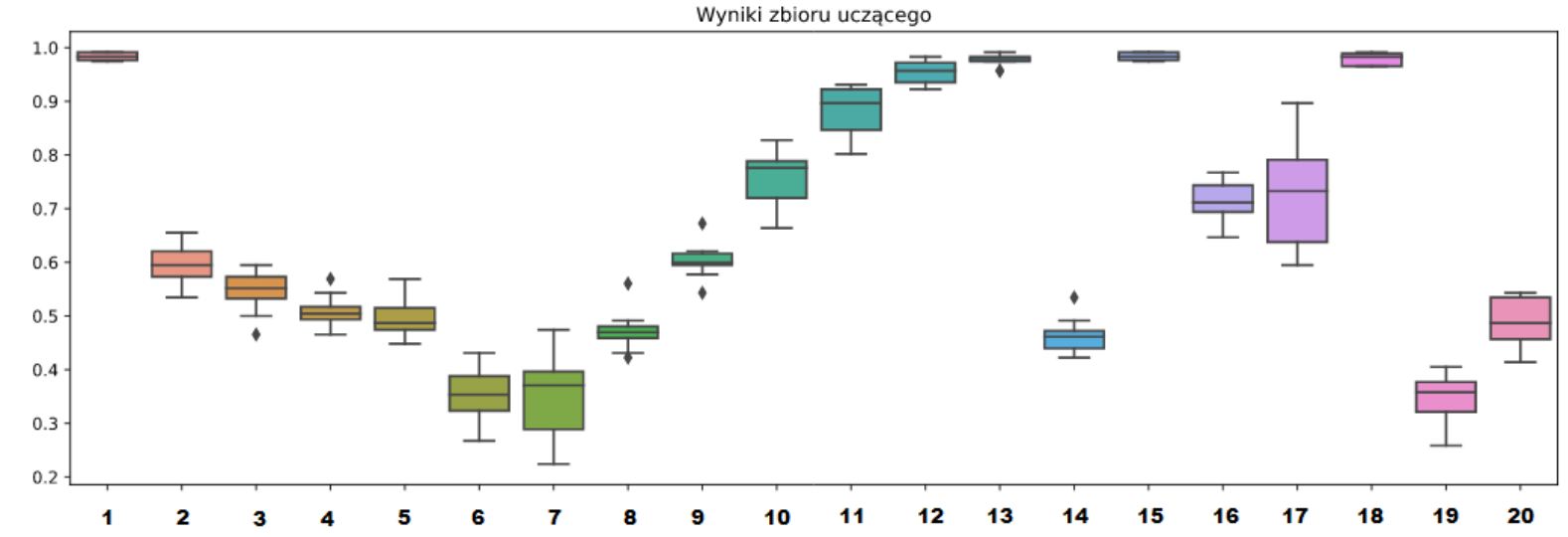
Ostatecznie klasyfikatory wybraliśmy na podstawie otrzymanych wyników. Każda klasyfikacja zbioru uczącego i testowego była oceniana w skali 0-1, gdzie 1 oznaczało pełne dopasowanie nowej klasyfikacji do pierwotnego podziału.

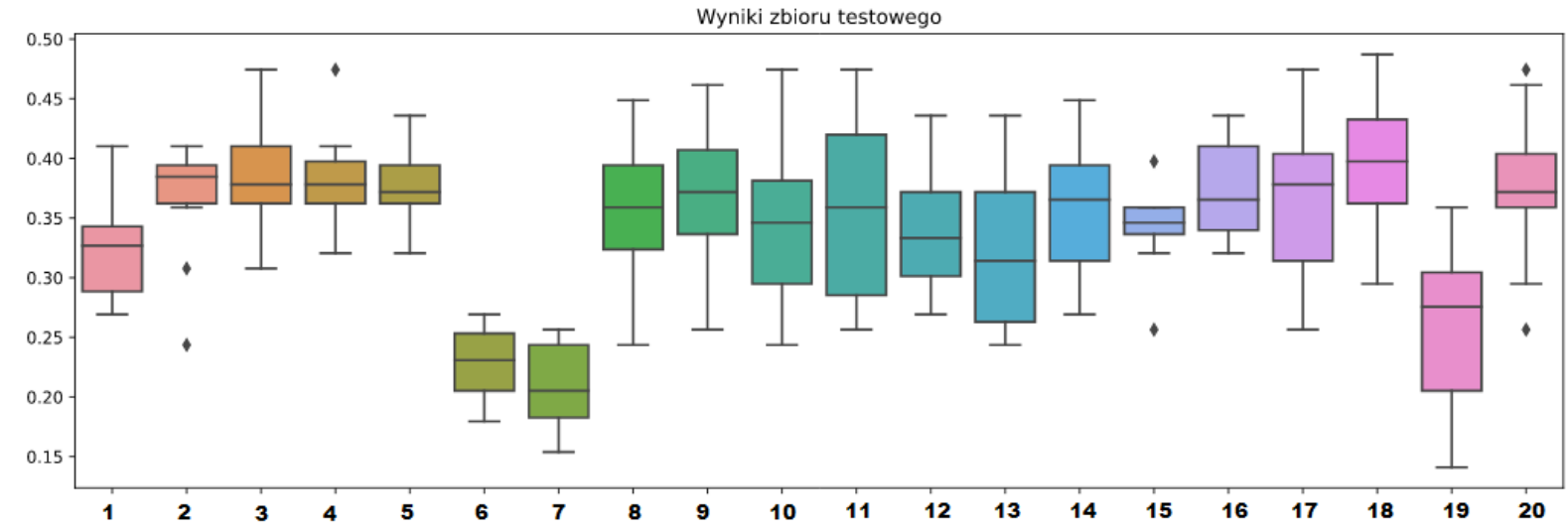
4.2. Spis klasyfikatorów

Porównane klasyfikatory:

- k-najbliższych sąsiadów (dla k = 1, 3, 5, 7, 9) [numery 1-5]
- najbliższego prototypu [numer 6]
- naiwny Bayesa [numer 7]
- Drzewo decyzyjne (o głębokości = 3, 5, 7, 9, 11, 13) [numery 8-13]
- SVC [numery 14-15]
- Losowego lasu [numer 16]
- Spadku gradientu [numer 17]
- Typu MLP [numer 18]
- AdaBoost [numer 19]
- Proces Gaussowski [numer 20]

4.3. Porównanie wyników klasyfikacji





Powyższe wykresy przedstawiają wynik klasyfikacji dla 10 różnych podziałów zbioru na uczący i testowy. Jak możemy zauważyć większość klasyfikatorów grupuje dane z poprawnością na poziomie 35%. Najlepiej działającym klasyfikatorem (spośród sprawdzanych dla zbioru) okazał się **klasyfikator typu MLP**. Uzyskał on średnio najlepszy wynik klasyfikacji zbioru testowego (ok. 40%) oraz jeden z najlepszych wyników klasyfikacji zbioru uczącego (ok. 97%). Podobny wynik dla zbioru testowego uzyskał klasyfikator 9 - najbliższych sąsiadów, jednak dla zbioru uczącego uzyskał wynik 2 razy gorszy. Natomiast klasyfikatory które równie dobrze przeprowadziły klasyfikacje zbioru uczącego m.in. klasyfikator SVC, drzewo decyzyjne o głębokości 13, uzyskały gorsze wyniki dla zbioru testowego.

Klasyfikatory zostały również sprawdzone dla podziału zbioru bez elementu losowości. Wyniki prezentują się następująco:

Klasyfikator	Wynik dla zbioru uczącego	Wynik dla zbioru testowego
1 - najbliższego sąsiada	0,974	0,307
3 – najbliższych sąsiadów	0,525	0,307
5 – najbliższych sąsiadów	0,551	0,346
7 – najbliższych sąsiadów	0,474	0,346
9 – najbliższych sąsiadów	0,456	0,371
Najbliższego prototypu	0,318	0,166
Naiwny Bayesa	0,293	0,153
Drzewo decyzyjne (głębokość 3)	0,465	0,423
Drzewo decyzyjne (głębokość 5)	0,612	0,371
Drzewo decyzyjne (głębokość 7)	0,741	0,423
Drzewo decyzyjne (głębokość 9)	0,862	0,423
Drzewo decyzyjne (głębokość 11)	0,922	0,371
Drzewo decyzyjne (głębokość 13)	0,974	0,397
SVC	0,456	0,384
SVC	0,974	0,333
RandomForest	0,793	0,384
GradientBoosting	0,629	0,282
MLP	**0,974**	**0,474**
AdaBoost	0,448	0,397
GaussianProcess	0,491	0,333

Dla podziału zbiorów bez elementu losowości podziału również najlepsze wyniki otrzymał klasyfikator typu MLP, dlatego główną klasyfikację przeprowadziliśmy z jego użyciem.

5. Klasyfikacja

5.1. Macierz pomyłek

Macierz pomyłek dla klasyfikatora typu MLP:

- Zbiór uczący

grupa	1	2	3	4	5	6	7	8
1 -katolicka	15	0	0	0	0	0	0	0
2 - Inne chrześcijańskie	2	31	0	0	0	0	0	0
3 - Muzułmańska	0	0	23	0	0	0	0	0
4 - Buddyzm	0	0	0	7	0	0	0	0
5 - Hinduizm	0	0	0	0	4	0	0	0
6 - Etniczna	1	0	0	0	0	19	0	0
7 - Marksistowska	0	0	1	0	0	0	10	0
8 - inne	0	0	0	0	0	0	0	3

- Zbiór testowy

grupa	1	2	3	4	5	6	7	8
1 -katolicka	8	8	4	0		1	4	
2 - Inne chrześcijańskie	3	19	1	1		1	1	
3 - Muzułmańska	0	4	8	0		1	0	
4 - Buddyzm	0	1	0	0		0	0	
5 - Hinduizm								
6 - Etniczna	0	1	1	1		4	0	
7 - Marksistowska	0	2	0	0		0	2	
8 - inne	0	0	1	0		0	0	

Z powyższych tabel łatwo możemy wywnioskować, że zdecydowanie łatwiej spośród wszystkich wyznań, na podstawie podziału flag możemy stwierdzić, że dane państwo jest katolickie, bądź innej religii chrześcijańskiej.

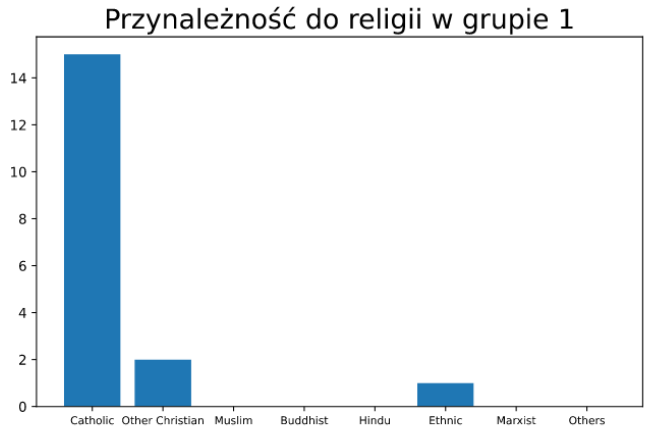
Skuteczność klasyfikacji – zbiór uczący: **96.55%**
Skuteczność klasyfikacji – zbiór testowy: **52.56%**

Z macierzy dla zbioru testowego możemy też dowiedzieć się następujących informacji:

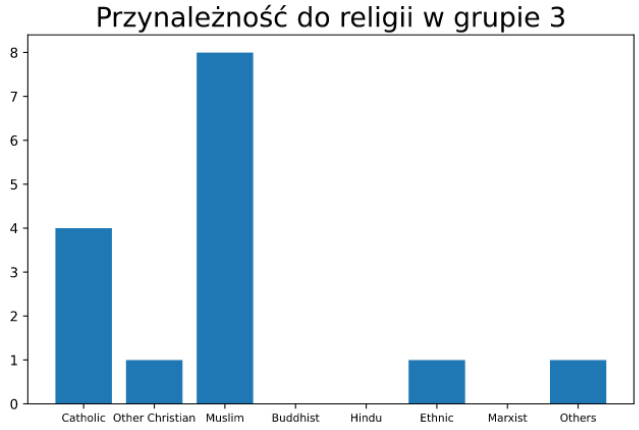
- Kraje chrześcijańskie – mają największy współczynnik czułości
- Kraje katolickie – mają największy współczynnik precyzji
- Dla Buddyzmu, Marksizmu i „innych” nie dopasowano poprawnie żadnego państwa
- Zbiór testowy nie zawierał kraju hinduskiego

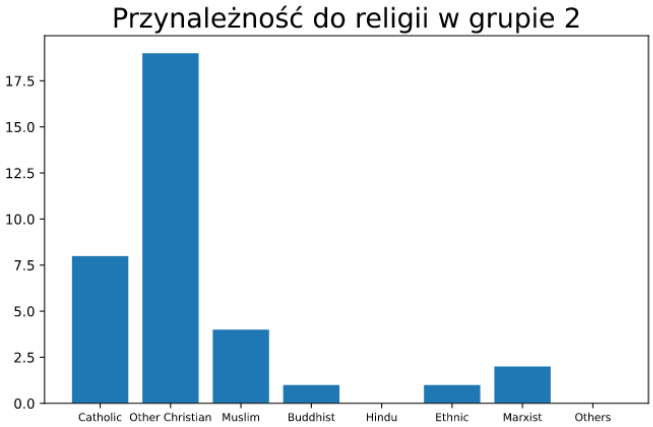
5.2. Wizualizacja klasyfikacji

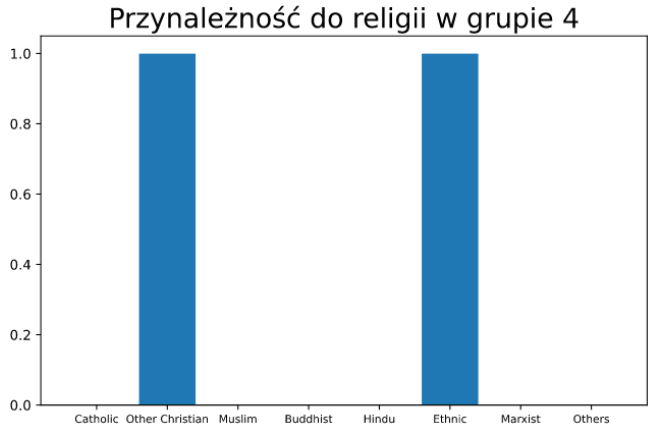
Procentowe występowanie religii w danej grupie (dla zbioru uczącego):

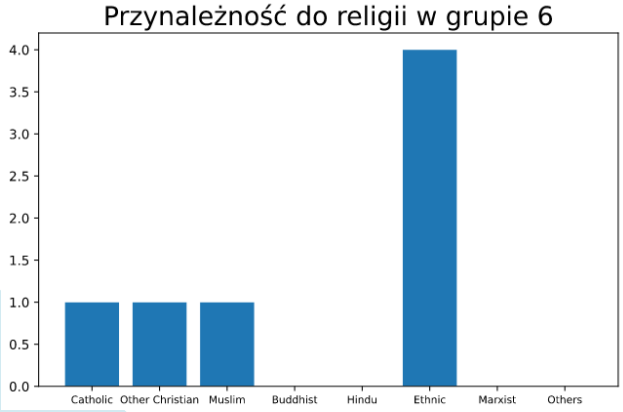


Procentowe występowanie religii w danej grupie (dla zbioru testowego):









ZBIÓR UCZĄCY

Grupa 1:

Grupa 2:

Grupa 3:

ZBIÓR TESTOWY

Grupa 1:

Grupa 2:

Grupa 3:

ZBIÓR UCZĄCY

Grupa 4:

Grupa 5:

Grupa 6:

Grupa 7:

Grupa 8:

ZBIÓR TESTOWY

Grupa 4:

Grupa 5:

Grupa 6:

Grupa 7:

*Ikony flag pochodzą z oddzielnych źródeł i mogą odbiegać od przedstawiających je wartości w zbiorze

Na podstawie analizy otrzymanych grup możemy stwierdzić:

- Kraje chrześcijańskie charakteryzują się krzyżem na swojej fladze
- Wszystkie flagi z charakterystycznymi krzyżami dla Wielkiej Brytanii zostały sklasyfikowane do tej samej grupy

- Dla krajów muzułmańskich charakterystyczne są zielone kolory
- Kraje z grupy pierwszej (katolickie) charakteryzują się barwami: niebieskim, żółtym, białym i składają się z pasków
- W grupie 7 (marksizm), każda flaga posiada czerwoną barwę

Obserwacje wizualnej prezentacji zbioru pokrywają się z założeniami postawionymi wraz z obserwacją korelacji między atrybutami opisowymi, a atrybutem decyzyjnym.

6. Podsumowanie

Podsumowując uzyskane wyniki możemy łatwo dojść do wniosku, że na podstawie klasyfikacji flag nie jesteśmy w stanie poprawnie przydzielić państwu jego religię. W najlepszym uzyskanym przez nas przypadku dopasowanie to wynosiło dla zbioru testowego zaledwie 52%.

Mimo wszystko uważamy, że flagi zostały sklasyfikowane poprawnie, gdyż każda z grup ma charakterystyczne właściwości, które odróżniają ją od pozostałych grup.

Musimy jednak wziąć pod uwagę, że dane na których pracowaliśmy były z 1990 roku, a zarówno wygląd flag państw jak i ich główne wyznanie mogą zmieniać się z czasem dlatego możliwe jest, że aktualnie, bądź w przyszłości podział taki będzie możliwy ze znacznie większą skutecznością.