Podstawy uczenia maszynowego

11.04.2024

Laboratorium 5

Klasteryzacja

Łukasz Stępień, Kacper Fus

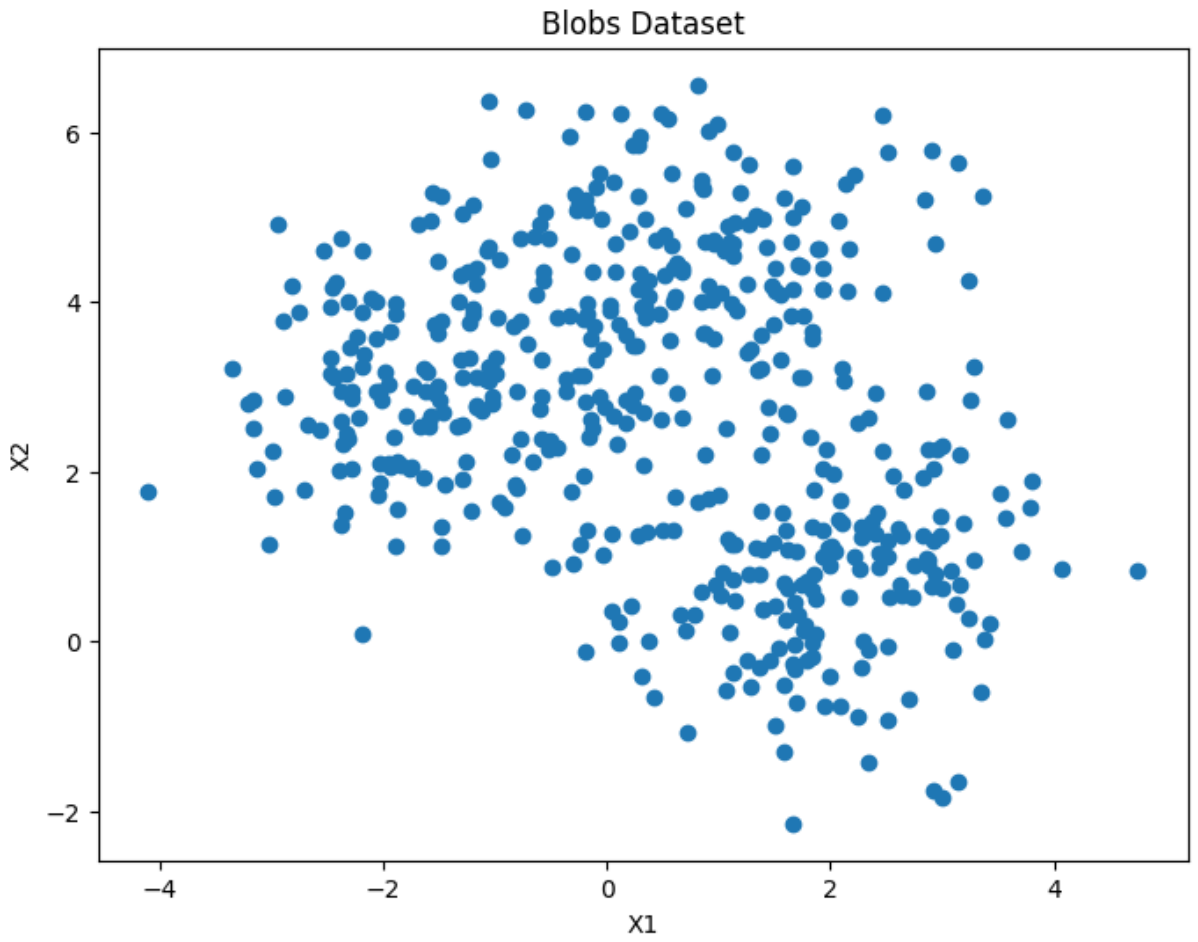
1. Cel zadania

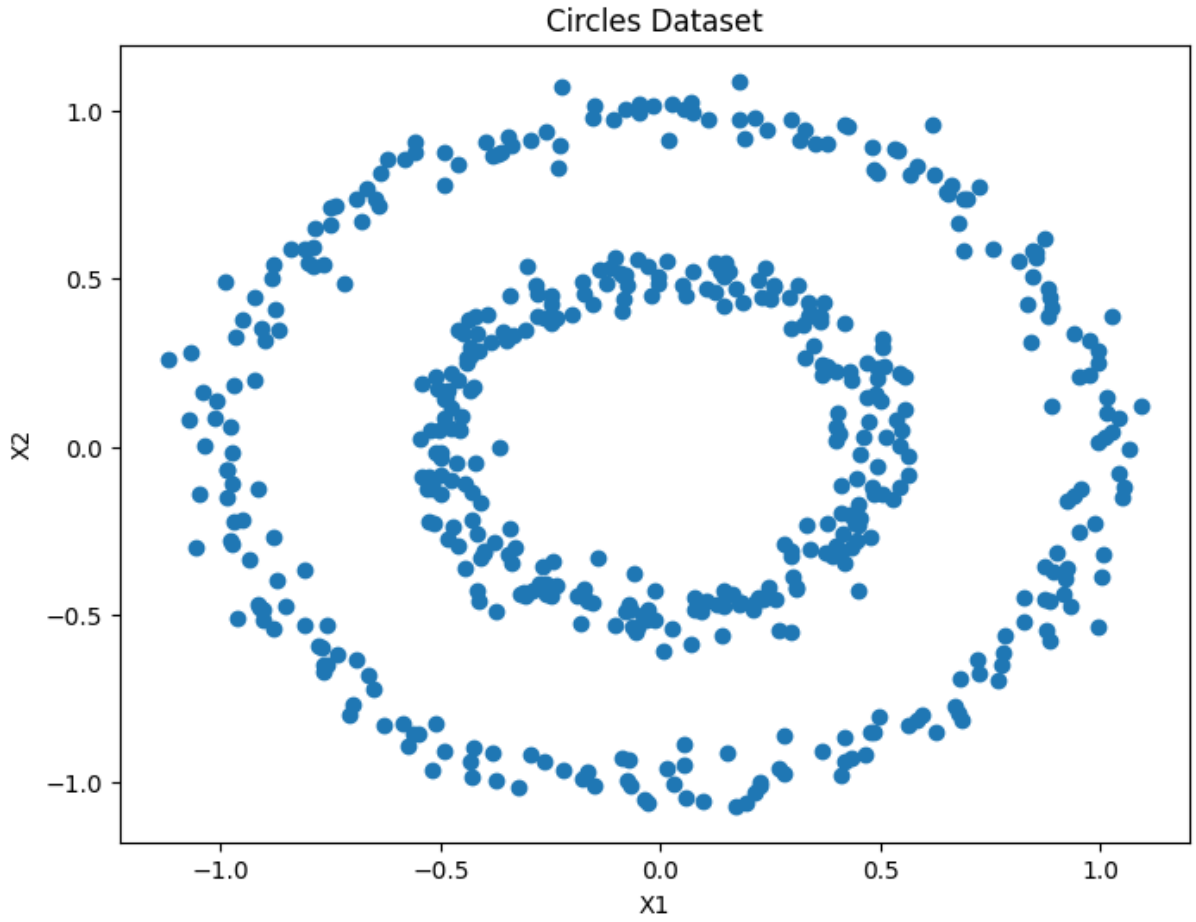
Celem zadania było zapoznanie się z trzema popularnymi algorytmami klasteryzacji: k-means, DBSCAN i klasteryzacją spektralną oraz zbadanie ich skuteczności na różnych zbiorach danych.

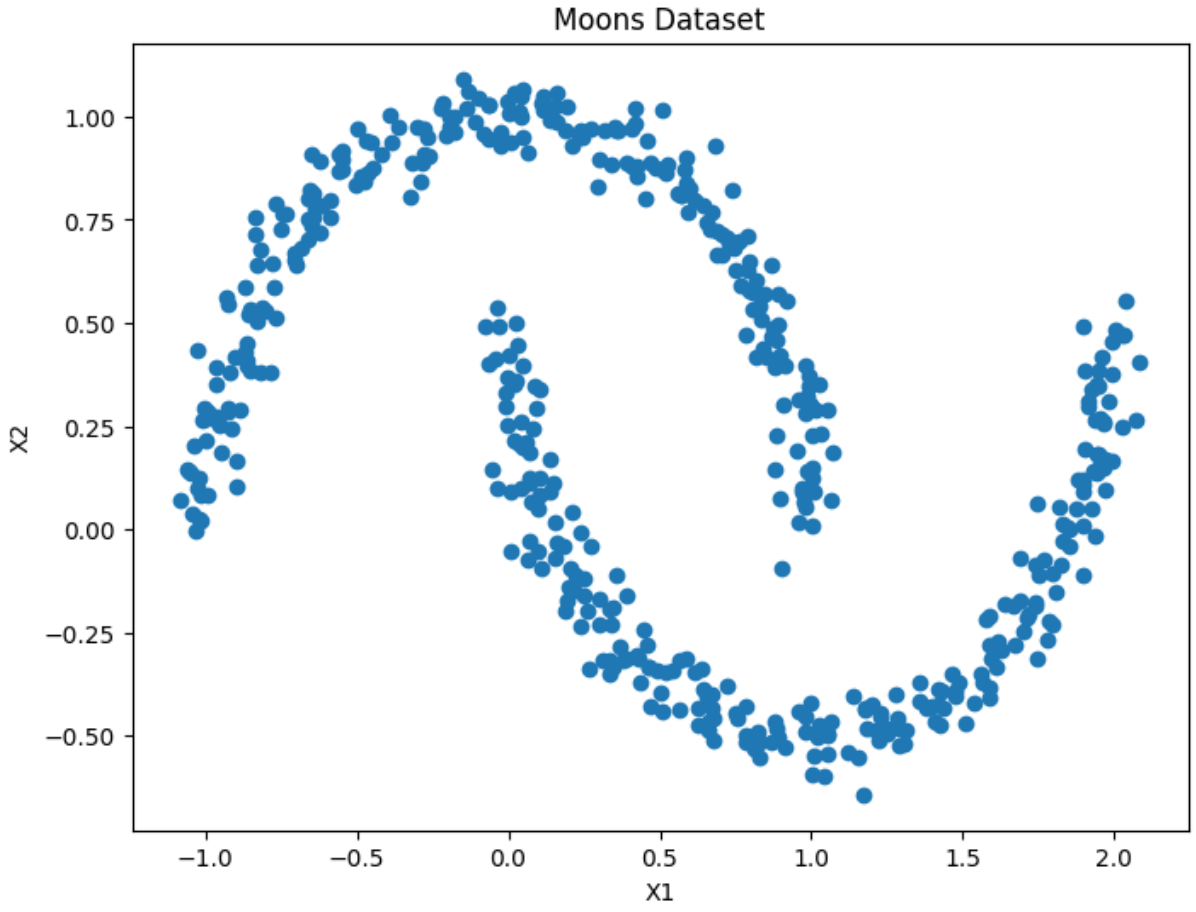
2. Implementacja

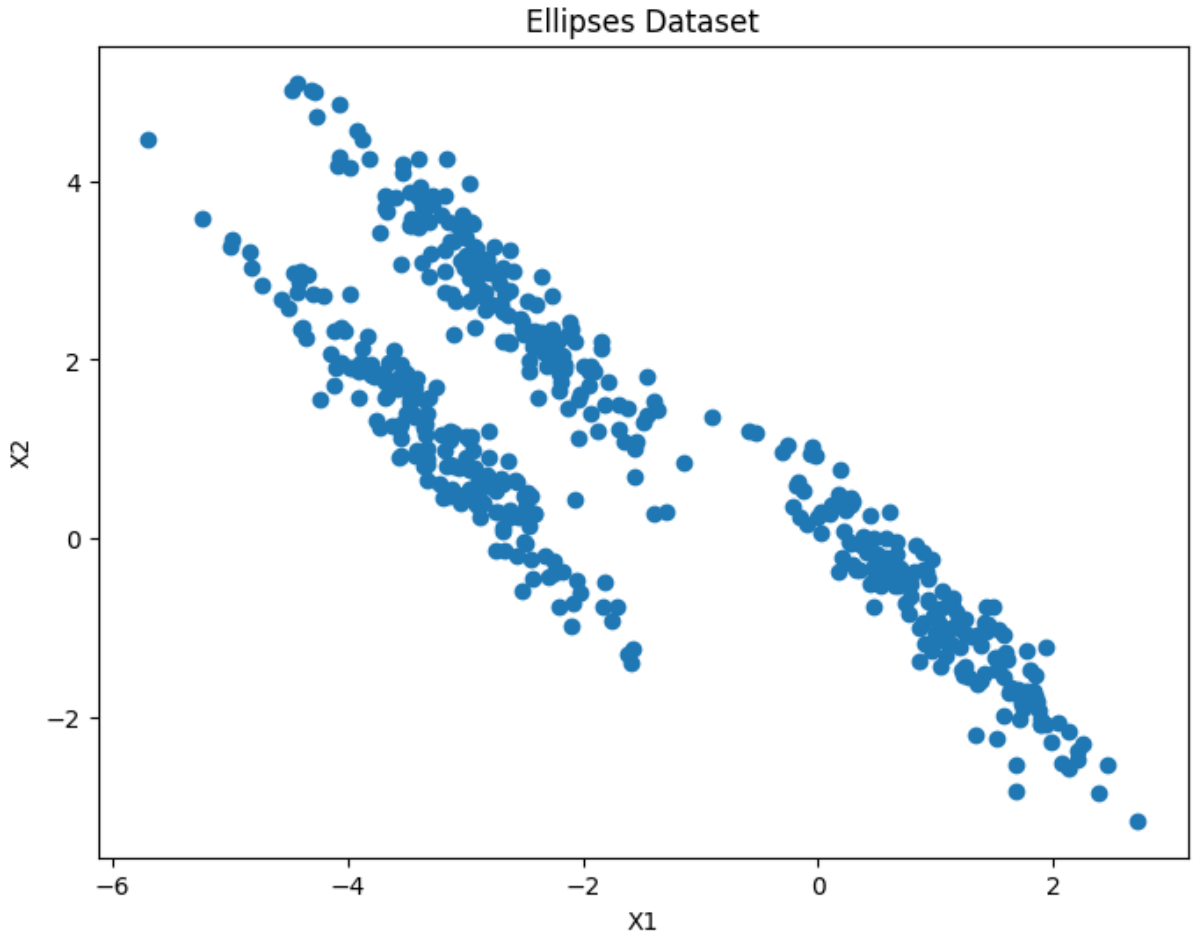
Wykorzystano syntetycznie generowane zbiory „Blobs", „Circles", „Moons" oraz „Ellipses". Następnie zbadano jak algorytmy k-means, DBSCAN oraz klasteryzacja spektralna radzą sobie z tymi zbiorami, ze względu na różne parametry. Następnie wykorzystano zbiór banknotes, który zawiera informacje o cechach banknotów oraz ich statusie jako oryginalne lub sfałszowane. Zadanie polegało na zastosowaniu algorytmów k-means i DBSCAN do pogrupowania banknotów w klastry.

2.1 Wizualizacja zborów




Circles Dataset
1.0
0.5
0.0
−0.5
−1.0
X2
−1.0
−0.5
0.0
0.5
1.0
X1


Moons Dataset
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
X2
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
X1



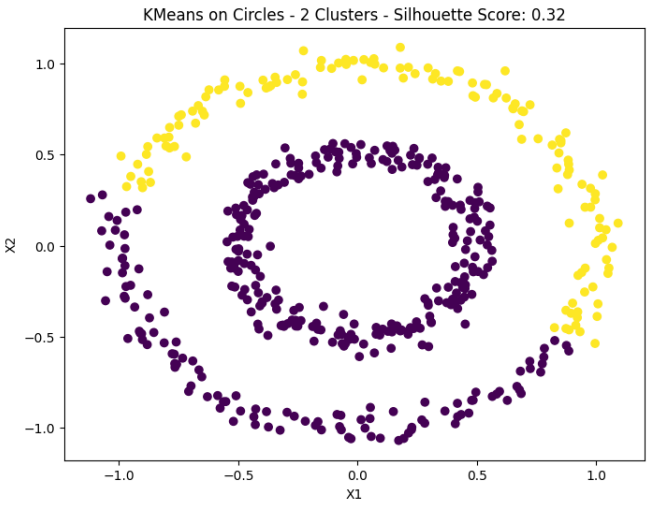
2.2. Wyniki testów

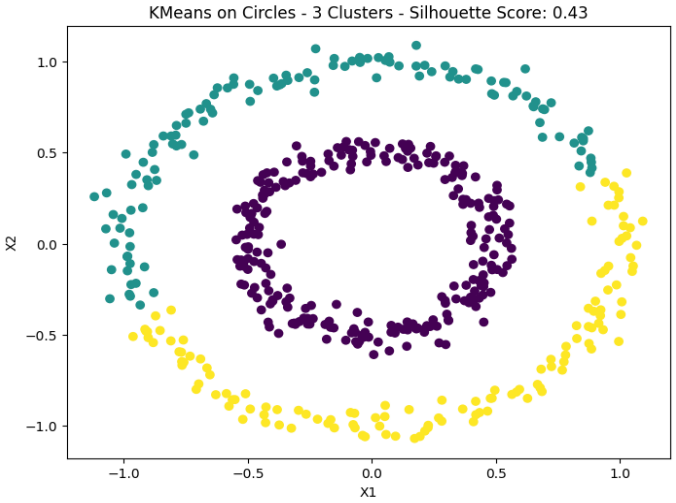
```
Accuracy for KMeans:
Circles Dataset:
N Clusters: 2, Accuracy: 0.50
N Clusters: 3, Accuracy: 0.34
N Clusters: 4, Accuracy: 0.24

Moons Dataset:
N Clusters: 2, Accuracy: 0.75
N Clusters: 3, Accuracy: 0.56
N Clusters: 4, Accuracy: 0.45

Ellipses Dataset:
N Clusters: 2, Accuracy: 0.32
N Clusters: 3, Accuracy: 0.11
N Clusters: 4, Accuracy: 0.07

Blobs Dataset:
N Clusters: 2, Accuracy: 0.03
N Clusters: 3, Accuracy: 0.05
N Clusters: 4, Accuracy: 0.18
```

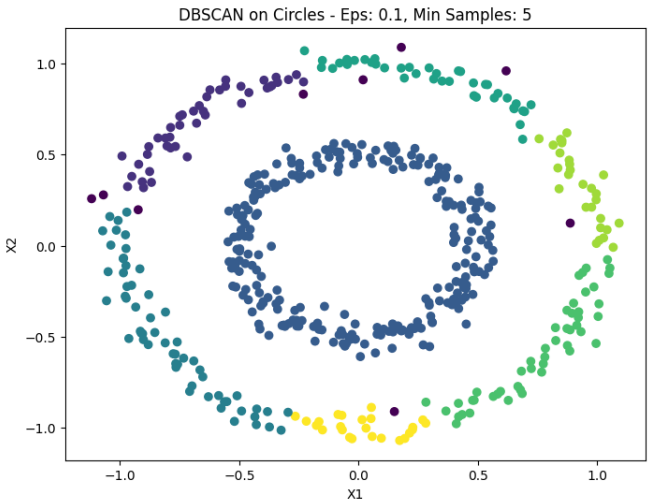
```
Accuracy for DBSCAN:
Circles Dataset:
Eps: 0.1, Min Samples: 5, Accuracy: 0.59
Eps: 0.1, Min Samples: 10, Accuracy: 0.00
Eps: 0.1, Min Samples: 15, Accuracy: 0.04
Eps: 0.2, Min Samples: 5, Accuracy: 1.00
Eps: 0.2, Min Samples: 10, Accuracy: 1.00
Eps: 0.2, Min Samples: 15, Accuracy: 0.77

Moons Dataset:
Eps: 0.1, Min Samples: 5, Accuracy: 0.00
Eps: 0.1, Min Samples: 10, Accuracy: 0.23
Eps: 0.1, Min Samples: 15, Accuracy: 0.09
Eps: 0.2, Min Samples: 5, Accuracy: 0.00
Eps: 0.2, Min Samples: 10, Accuracy: 0.00
Eps: 0.2, Min Samples: 15, Accuracy: 0.00
Eps: 0.3, Min Samples: 5, Accuracy: 0.00
Eps: 0.3, Min Samples: 10, Accuracy: 0.00
Eps: 0.3, Min Samples: 15, Accuracy: 0.00

Ellipses Dataset:
Eps: 0.1, Min Samples: 5, Accuracy: 0.05
Eps: 0.1, Min Samples: 10, Accuracy: 0.24
Eps: 0.1, Min Samples: 15, Accuracy: 0.10
Eps: 0.2, Min Samples: 5, Accuracy: 0.06
Eps: 0.2, Min Samples: 10, Accuracy: 0.31
Eps: 0.2, Min Samples: 15, Accuracy: 0.27
Eps: 0.3, Min Samples: 5, Accuracy: 0.25

Blobs Dataset:
Eps: 0.1, Min Samples: 5, Accuracy: 0.00
Eps: 0.1, Min Samples: 10, Accuracy: 0.03
Eps: 0.1, Min Samples: 15, Accuracy: 0.21
Eps: 0.2, Min Samples: 5, Accuracy: 0.03
```

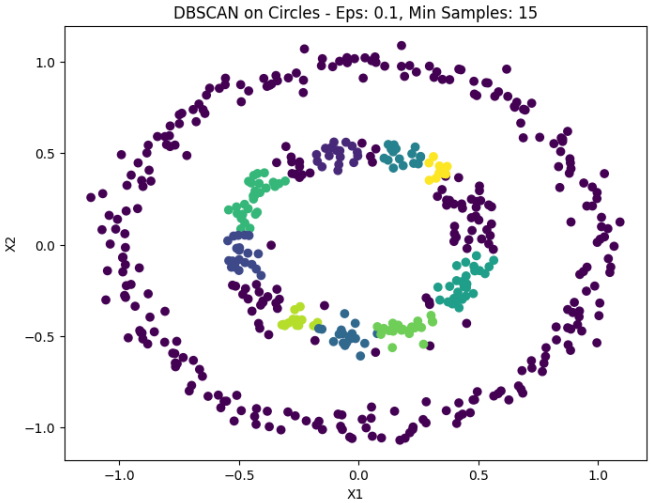
```
Accuracy for Circles Dataset:
N Clusters: 2, Gamma: 0.1, Accuracy: 0.51
N Clusters: 2, Gamma: 1, Accuracy: 0.51
N Clusters: 2, Gamma: 10, Accuracy: 0.49
N Clusters: 3, Gamma: 0.1, Accuracy: 0.32
N Clusters: 3, Gamma: 1, Accuracy: 0.33
N Clusters: 3, Gamma: 10, Accuracy: 0.34
N Clusters: 4, Gamma: 0.1, Accuracy: 0.26
N Clusters: 4, Gamma: 1, Accuracy: 0.25
N Clusters: 4, Gamma: 10, Accuracy: 0.26

Accuracy for Moons Dataset:
N Clusters: 2, Gamma: 0.1, Accuracy: 0.74
N Clusters: 2, Gamma: 1, Accuracy: 0.77
N Clusters: 2, Gamma: 10, Accuracy: 0.01
N Clusters: 3, Gamma: 0.1, Accuracy: 0.12
N Clusters: 3, Gamma: 1, Accuracy: 0.00
N Clusters: 3, Gamma: 10, Accuracy: 0.26
N Clusters: 4, Gamma: 0.1, Accuracy: 0.06
N Clusters: 4, Gamma: 1, Accuracy: 0.45
N Clusters: 4, Gamma: 10, Accuracy: 0.23

Accuracy for Ellipses Dataset:
N Clusters: 2, Gamma: 0.1, Accuracy: 0.33
N Clusters: 2, Gamma: 1, Accuracy: 0.33
N Clusters: 2, Gamma: 10, Accuracy: 0.33
N Clusters: 3, Gamma: 0.1, Accuracy: 0.26
N Clusters: 3, Gamma: 1, Accuracy: 0.06
N Clusters: 3, Gamma: 10, Accuracy: 0.33
N Clusters: 4, Gamma: 0.1, Accuracy: 0.29
N Clusters: 4, Gamma: 1, Accuracy: 0.22
N Clusters: 4, Gamma: 10, Accuracy: 0.00

Accuracy for Blobs Dataset:
N Clusters: 2, Gamma: 0.1, Accuracy: 0.02
N Clusters: 2, Gamma: 1, Accuracy: 0.65
N Clusters: 2, Gamma: 10, Accuracy: 0.65
N Clusters: 3, Gamma: 0.1, Accuracy: 0.32
N Clusters: 3, Gamma: 1, Accuracy: 0.37
N Clusters: 3, Gamma: 10, Accuracy: 0.34
N Clusters: 4, Gamma: 0.1, Accuracy: 0.12
N Clusters: 4, Gamma: 1, Accuracy: 0.01
N Clusters: 4, Gamma: 10, Accuracy: 0.32
```

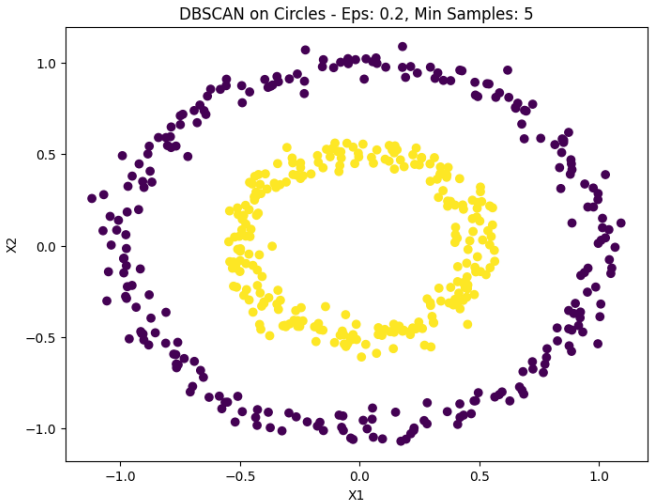
2.3 Wpływ parametrów na kształt klastrów

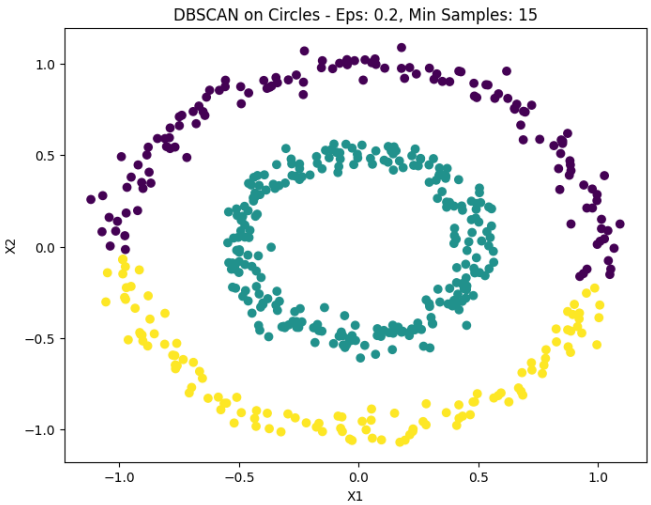


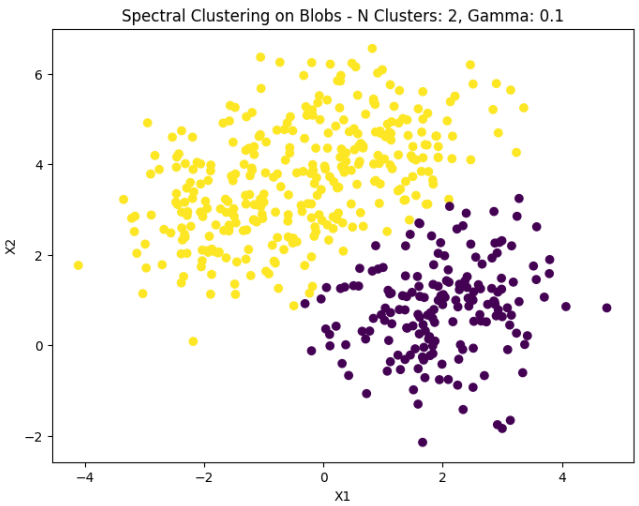


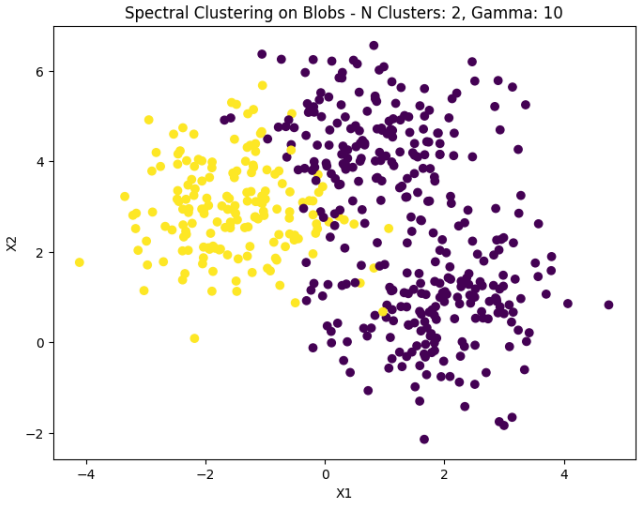


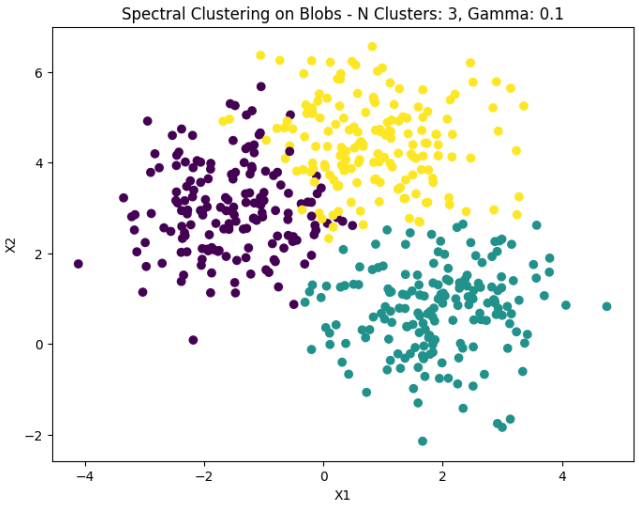


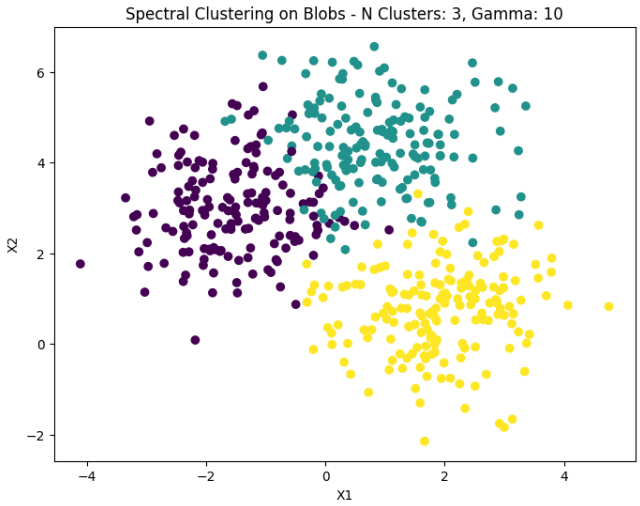












2.4 Wnioski do zadania 1

- K-means
  - Najlepiej sobie poradził ze zbiorem danych Circles oraz Moons. Sugeruje to, że lepiej on radzi sobie ze zbiorami eliptycznymi bądź okrągłymi ,które na dodatek są oddzielone od siebie .
  - Na zaprezentowanych wykresach, widać, że parametr zwiększą liczbę klastrów, na które dzielimy zbiór danych. Każdy ze zbiorów danych, jest równomierny oraz zauważamy granicę między nimi.
- DBSCAN
  - Najlepiej poradził sobie ponownie z Circles, prawdopodobnie dlatego, że reprezentują one zbiór o różnej gęstości,
  - Wpływ tych dwóch parametrów (epsilon i minimalna liczba punktów) na klastry w DBSCAN polega na decydowaniu, jakie punkty będą łączone w klastry na podstawie odległości i gęstości. Umiejętne dobranie tych parametrów sprawi, że uzyskane klastry będą miały równomierne kształty i będą wolne od zakłóceń spowodowanych szumem.

- Klasteryzacja spektralna
    - Poradziła sobie najlepiej z danymi Blobs co może świadczyć o tym, że Klasteryzacja spektralna jest skuteczna w identyfikowaniu klastrów w zbiorach danych, które nie są liniowo separowalne.
    - Zwiększanie liczby klastrów, zwiększa liczbę podzbiorów na jakie dane zostały podzielone, a gamma sprawią, że zbiory stają się mniej równomierne.

2.5 Wyniki doświadczeń do Zadania 2

```
Dokładność przypisania klastrów przez KMeans: 0.96

DBSCAN - Wszystkie banknoty przypisane: False
Homogeneity: 0.0
Completeness: 1.0
V-Measure: 0.0
```

2.6 Wnioski do Zadania 2

- DBSCAN nie przypisał wszystkich banknotów do klastrów. To sugeruje, że niektóre banknoty mogą zostać uznane za punkty odstające lub nieprzypisane do żadnego klastra.
- Homogeniczność mierzy: wartość 0.0 oznacza, że klastry nie są homogeniczne, co sugeruje, że punkty z różnych klas mogą być przypisane do tych samych klastrów.
- Zupełność mierzy: wartość 1.0 oznacza, że wszystkie punkty oryginalnie należące do jednej klasy zostały przypisane do jednego klastra, ale jakość tych klastrów jest nadal niska.
- V-miara: w tym przypadku, V-miara wynosząca 0.0 sugeruje, że klasteryzacja nie jest zgodna z rzeczywistym podziałem na banknoty oryginalne i sfałszowane.
- K-means, ma wysoką skuteczność w przypisywaniu klastrów, co świadczy, że struktura zbioru była dobrze odzwierciedlona przez algorytm.

- K-means:
    - K-means dobrze radzi sobie z danymi, które mają w miarę jednorodne i wypukłe klastry o podobnej wielkości.
    - Głównymi parametrami k-means są liczba klastrów (k). Wybór odpowiedniej liczby klastrów może być trudny, a niepoprawne ustawienie może prowadzić do złego podziału danych. Wybór punktów startowych może również wpływać na wynik, ale zwykle algorytm jest odporny na różne punkty startowe.

- DBSCAN
    - DBSCAN jest bardziej elastyczny i może radzić sobie z danymi o różnych kształtach klastrów oraz gęstości.
    - Głównymi parametrami DBSCAN są promień epsilon (eps) określający odległość, w której punkty są uznawane za sąsiadów, oraz minimalna liczba punktów (MinPts), które muszą znajdować się w otoczeniu punktu, aby został uznany za centralny. Wybór odpowiednich wartości eps i MinPts może być kluczowy. Zbyt małe eps mogą powodować, że wiele punktów zostanie uznanych za szum, podczas gdy zbyt duże eps mogą prowadzić do połączenia klastrów.

- Klasteryzacja spektralna:
    - Klasteryzacja spektralna może radzić sobie z danymi, które mają nieliniowe kształty klastrów oraz z danymi wysokowymiarowymi.
    - Klasteryzacja spektralna opiera się na wyznaczeniu macierzy podobieństwa lub odległości między punktami, a następnie wyznaczeniu kierunków (wektorów własnych) w tej przestrzeni. Głównym parametrem jest liczba kierunków (k), która określa liczbę klastrów. Wybór odpowiedniej liczby kierunków może być trudny, a niepoprawny wybór może prowadzić do złego podziału danych.