234053
*Numer indeksu*
Paweł Galewicz
*Imię i nazwisko*

234067
*Numer indeksu*
Bartosz Jurczewski
*Imię i nazwisko*

234102
*Numer indeksu*
Zbigniew Nowacki
*Imię i nazwisko*

234106
*Numer indeksu*
Karol Podlewski
*Imię i nazwisko*

234128
*Numer indeksu*
Piotr Wardęcki
*Imię i nazwisko*

**Kierunek** Informatyka Stosowana
**Stopień** II
**Specjalizacja** Data Science
**Semestr** 1

**Data oddania** 6 maja 2020

## Metody uczenia maszynowego

# Problem set 3

# Spis treści

# 1 Cel

Zadanie polegało na ocenie jakości modeli, które przygotowane były w poprzednich zadaniach – klasyfikacji oraz analizy skupień. Do oceny tej wykonane zostały różne testy z wykorzystaniem odpowiednich miar ewaluacyjnych oraz metod walidacji.

# 2 Opis implementacji

Algorytmy oraz przygotowane do nich testy zostały zaimplementowane za pomocą języka Python w wersji 3.8.2. Wykorzystano w nim biblioteki NumPy, Sklearn i Pandas.

# 3 Opis zbiorów danych

## 3.1 Klasyfikacja

Bazowaliśmy na trzech zestawach danych, przeznaczonych do klasyfikacji binarnej:

A. League of Legends Diamond Ranked Games (10 min) – zbiór zawierający dane dotyczące meczy rankingowych na wysokim poziomie w grze League of Legends, składający się z 38 cech, po 19 dla każdej drużyny. Zebrane wartości to informacje o między innymi zgromadzonym złocie, zniszczonych wieżach, zgładzonych potworach czy przeciwnikach na przestrzeni pierwszych 10 minut mapy. Do cech tych przypisana jest informacja o zwycięstwie danej drużyny.

B. Rain in Australia – historia danych pogodowych z 10 lat (data, lokalizacja, temperatury, opady, wiatr, ciśnienie, wilgotność, nasłonecznienie itp). Na podstawie 23 cech należy określić, czy kolejnego dnia wystąpią opady.

C. Pima Indians Diabetes Database – zbiór danych zawierający 8 niepowiązanych ze sobą pomiarów diagnostycznych (m. in. ilość ciąż, poziom glukozy, ciśnienie, bmi, wiek) na podstawie których dokonywana jest ocena, czy kobieta choruje na cukrzycę.

## 3.2 Klasteryzacja

Bazowaliśmy na trzech zestawach danych:

A. Mall Customer Segmentation - Zawiera informacje zbierane od klientów centrum handlowego uczestniczących w programie lojalnościowym. Są to między innymi dane o wieku, płci, rocznym przychodzie. Każdy klient ma też wyliczoną ocenę wydatków w sklepie w przedziale od 1 do 100.

B. Wine quality selection – Zestaw danych przedstawiający współczynniki wina wraz z przyporządkowaną ich jakością.

C. Credit Card Dataset – Zbiór opisujący zachowania aktywnych klientów bankowych, agregowanych przez 6miesięcy.

# 4 Ewaluacja modeli klasyfikacyjnych

Model klasyfikacji przygotowany został w ramach zadania 1. Zawiera on implementacje następujących metod:

1. Algorytm drzew decyzyjnych
2. Naiwny klasyfikator Bayesa
3. Maszyna wektorów nośnych
4. Klasyfikator k-najbliższych sąsiadów
5. Algorytm sztucznych sieci neuronowych

Na każdej z metod przeprowadzone zostały testy, które miały na celu sprawdzenie jak dobrze spełniają one swoje zadanie. Za każdym razem wykorzystano 75% danych jako część treningową.

## 4.1 Wartości podstawowych metryk ewaluacyjnych

Pierwszy test zakładał stworzenie macierzy pomyłek dla każdej metody klasyfikacji i na jej podstawie wyliczenie podstawowych metryk ewaluacyjnych:

- Dokładność – Określa jak duży odsetek obserwacji został prawidłowo zaklasyfikowany.
- Precyzja – Oznacza jak dużo zaklasyfikowanych do danej klasy obserwacji rzeczywiście do niej należy.
- Czułość – Stosunek liczby obserwacji oznaczonych jako *true positive* do sumy *true positive* i *false negative*. Jej wartość interpretować można jako zdolność do prawidłowego zakwalifikowania obserwacji do odpowiedniej klasy.
- Specyficzność – Stosunek liczby obserwacji *true negative* do sumy *true negative* i *false positive*. Określa jak dobrze model rozpoznaje, że dana obserwacja nie należy do danej klasy.

**Zbiór A**

Tabela 1: Macierz pomyłek dla algorytmu drzew decyzyjnych, dla zbioru A

| | | **Klasa predykowana** | |
|---|---|---|---|
| | | **Pozytywna** | **Negatywna** |
| **Klasa rzeczywista** | **Pozytywna** | 785 | 461 |
| | **Negatywna** | 457 | 766 |

Tabela 2: Macierz pomyłek dla naiwnego klasyfikatora Bayesa, dla zbioru A

| | | **Klasa predykowana** | |
|---|---|---|---|
| | | **Pozytywna** | **Negatywna** |
| **Klasa rzeczywista** | **Pozytywna** | 879 | 338 |
| | **Negatywna** | 331 | 921 |

Tabela 3: Macierz pomyłek dla maszyny wektorów nośnych, dla zbioru A

| | | **Klasa predykowana** | |
|---|---|---|---|
| | | **Pozytywna** | **Negatywna** |
| **Klasa rzeczywista** | **Pozytywna** | 889 | 378 |
| | **Negatywna** | 304 | 898 |

Tabela 4: Macierz pomyłek dla klasyfikatora k-nn, dla zbioru A

| | | **Klasa predykowana** | |
|---|---|---|---|
| | | **Pozytywna** | **Negatywna** |
| **Klasa rzeczywista** | **Pozytywna** | 832 | 411 |
| | **Negatywna** | 373 | 853 |

Tabela 5: Macierz pomyłek dla sztucznej sieci neuronowej, dla zbioru A

| | | **Klasa predykowana** | |
|---|---|---|---|
| | | **Pozytywna** | **Negatywna** |
| **Klasa rzeczywista** | **Pozytywna** | 1207 | 22 |
| | **Negatywna** | 1079 | 161 |

Tabela 6: Wartości metryk ewaluacyjnych dla zbioru A

| **Algorytm** | **Precyzja** | **Dokładność** | **Czułość** | **Specyficzność** |
|---|---|---|---|---|
| Drzew decyzyjnych | 0.632 | 0.628 | 0.63 | 0.626 |
| Naiwny klasyfikator Bayesa | 0.726 | 0.729 | 0.722 | 0.736 |
| Maszyna wektorów nośnych | 0.745 | 0.724 | 0.702 | 0.747 |
| Klasyfikator k-nn | 0.69 | 0.682 | 0.669 | 0.696 |
| Sztucznych sieci neuronowych | 0.528 | 0.554 | 0.982 | 0.13 |

**Zbiór B**

Tabela 7: Macierz pomyłek dla algorytmu drzew decyzyjnych, dla zbioru B

| | | **Klasa predykowana** | |
|---|---|---|---|
| | | **Pozytywna** | **Negatywna** |
| **Klasa rzeczywista** | **Pozytywna** | 3115 | 0 |
| | **Negatywna** | 0 | 10989 |

Tabela 8: Macierz pomyłek dla naiwnego klasyfikatora Bayesa, dla zbioru B

| | | **Klasa predykowana** | |
|---|---|---|---|
| | | **Pozytywna** | **Negatywna** |
| **Klasa rzeczywista** | **Pozytywna** | 3036 | 1 |
| | **Negatywna** | 812 | 10255 |

Tabela 9: Macierz pomyłek dla maszyny wektorów nośnych, dla zbioru B

| | | **Klasa predykowana** | |
|---|---|---|---|
| | | **Pozytywna** | **Negatywna** |
| **Klasa rzeczywista** | **Pozytywna** | 485 | 2644 |
| | **Negatywna** | 1 | 10974 |

Tabela 10: Macierz pomyłek dla klasyfikatora k-nn, dla zbioru B

| | | **Klasa predykowana** | |
|---|---|---|---|
| | | **Pozytywna** | **Negatywna** |
| **Klasa rzeczywista** | **Pozytywna** | 1589 | 1497 |
| | **Negatywna** | 405 | 10613 |

Tabela 11: Macierz pomyłek dla sztucznej sieci neuronowej, dla zbioru B

| | | **Klasa predykowana** | |
|---|---|---|---|
| | | **Pozytywna** | **Negatywna** |
| **Klasa rzeczywista** | **Pozytywna** | 3108 | 6 |
| | **Negatywna** | 147 | 10843 |

Tabela 12: Wartości metryk ewaluacyjnych dla zbioru B

| **Algorytm** | **Precyzja** | **Dokładność** | **Czułość** | **Specyficzność** |
|---|---|---|---|---|
| Drzew decyzyjnych | 1.0 | 1.0 | 1.0 | 1.0 |
| Naiwny klasyfikator Bayesa | 0.789 | 0.942 | 1.0 | 0.927 |
| Maszyna wektorów nośnych | 0.998 | 0.812 | 0.155 | 1.0 |
| Klasyfikator k-nn | 0.797 | 0.865 | 0.515 | 0.963 |
| Sztucznych sieci neuronowych | 0.955 | 0.989 | 0.998 | 0.987 |

**Zbiór C**

Tabela 13: Macierz pomyłek dla algorytmu drzew decyzyjnych, dla zbioru C

| | | **Klasa predykowana** | |
|---|---|---|---|
| | | **Pozytywna** | **Negatywna** |
| **Klasa rzeczywista** | **Pozytywna** | 40 | 30 |
| | **Negatywna** | 24 | 97 |

Tabela 14: Macierz pomyłek dla naiwnego klasyfikatora Bayesa, dla zbioru C

| | | **Klasa predykowana** | |
|---|---|---|---|
| | | **Pozytywna** | **Negatywna** |
| **Klasa rzeczywista** | **Pozytywna** | 36 | 33 |
| | **Negatywna** | 20 | 102 |

Tabela 15: Macierz pomyłek dla maszyny wektorów nośnych, dla zbioru C

| | | **Klasa predykowana** | |
|---|---|---|---|
| | | **Pozytywna** | **Negatywna** |
| **Klasa rzeczywista** | **Pozytywna** | 36 | 29 |
| | **Negatywna** | 10 | 116 |

Tabela 16: Macierz pomyłek dla klasyfikatora k-nn, dla zbioru C

| | | **Klasa predykowana** | |
|---|---|---|---|
| | | **Pozytywna** | **Negatywna** |
| **Klasa rzeczywista** | **Pozytywna** | 39 | 29 |
| | **Negatywna** | 23 | 100 |

Tabela 17: Macierz pomyłek dla sztucznej sieci neuronowej, dla zbioru C

| | | **Klasa predykowana** | |
|---|---|---|---|
| | | **Pozytywna** | **Negatywna** |
| **Klasa rzeczywista** | **Pozytywna** | 41 | 14 |
| | **Negatywna** | 40 | 96 |

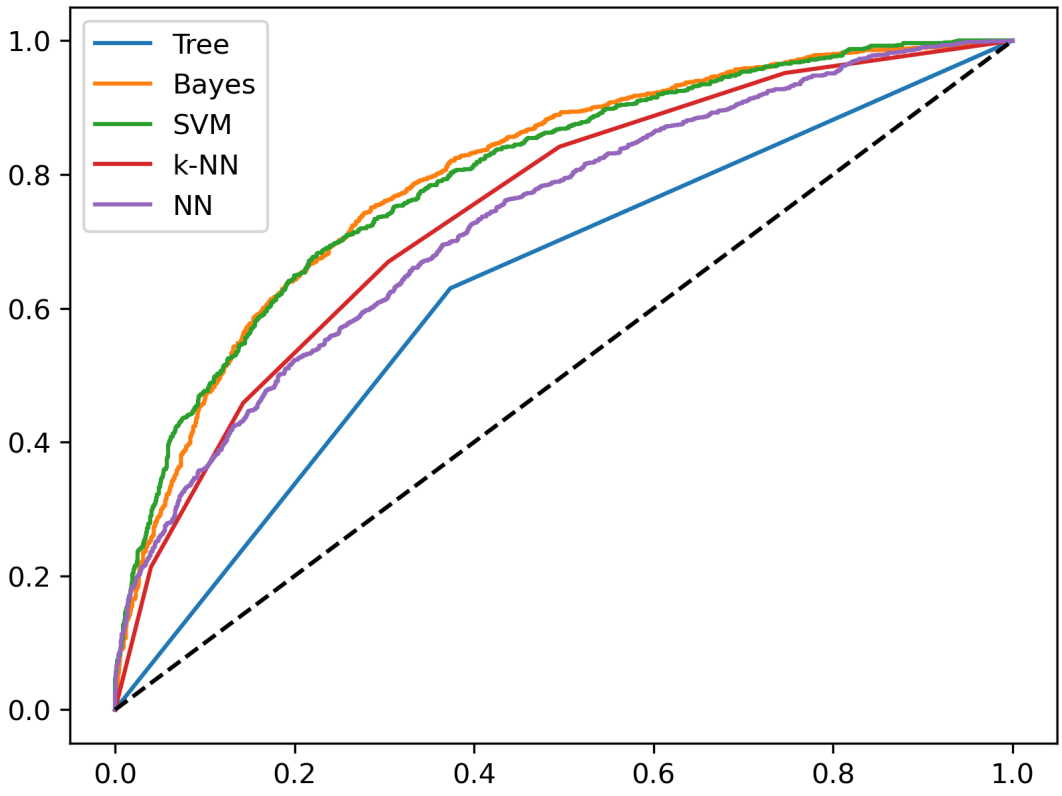
Tabela 18: Wartości metryk ewaluacyjnych dla zbioru C

| Algorytm | Precyzja | Dokładność | Czułość | Specyficzność |
|---|---|---|---|---|
| Drzew decyzyjnych | 0.625 | 0.717 | 0.571 | 0.802 |
| Naiwny klasyfikator Bayesa | 0.643 | 0.723 | 0.522 | 0.836 |
| Maszyna wektorów nośnych | 0.783 | 0.796 | 0.554 | 0.921 |
| Klasyfikator k-nn | 0.629 | 0.728 | 0.574 | 0.813 |
| Sztucznych sieci neuronowych | 0.506 | 0.717 | 0.745 | 0.706 |

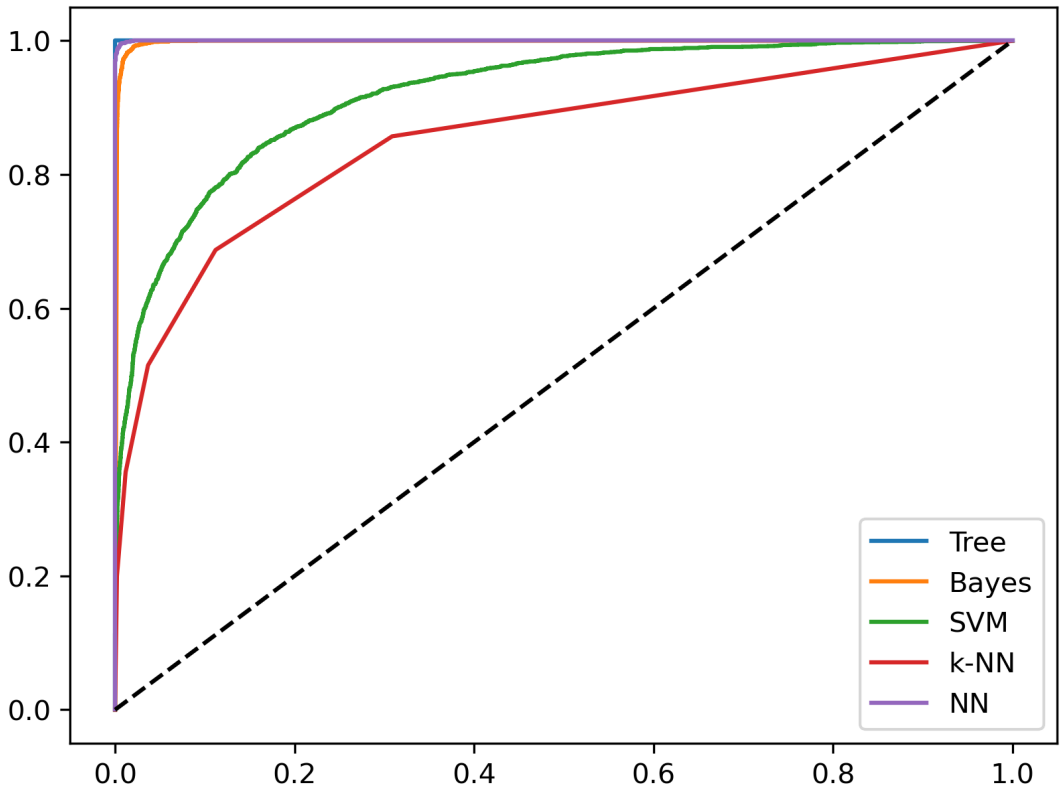
## 4.2 Krzywe ROC

Drugim testem było wyznaczenie krzywych ROC dla każdej zaimplementowanej metody. Pokazuje ona zależność między czułością, a dopełnieniem specyficzności (1 - specificity) dla różnych progów alfa. Pole pod stworzonym wykresem pozwala wyliczyć jakość klasyfikatora.

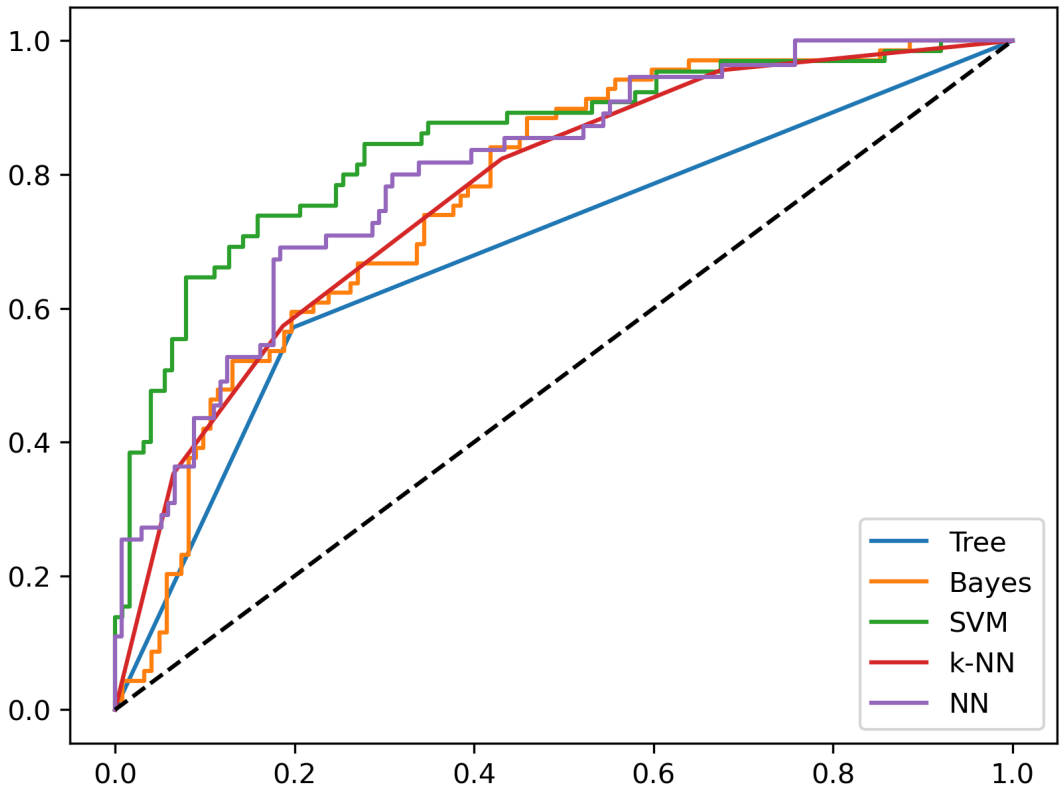
W celu jaśniejszego porównania wszystkie krzywe zostały przedstawione na jednym, zbiorczym wykresie.




Rysunek 1: Krzywe ROC dla zbioru A



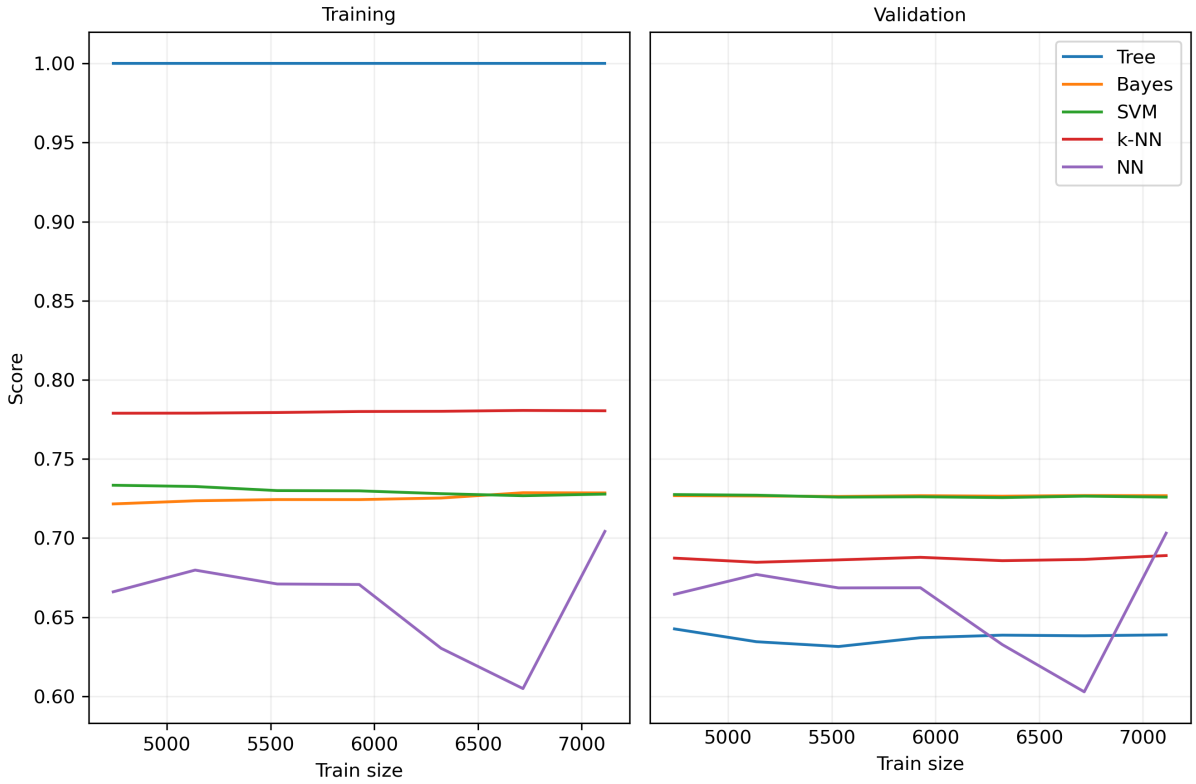

Rysunek 2: Krzywe ROC dla zbioru B



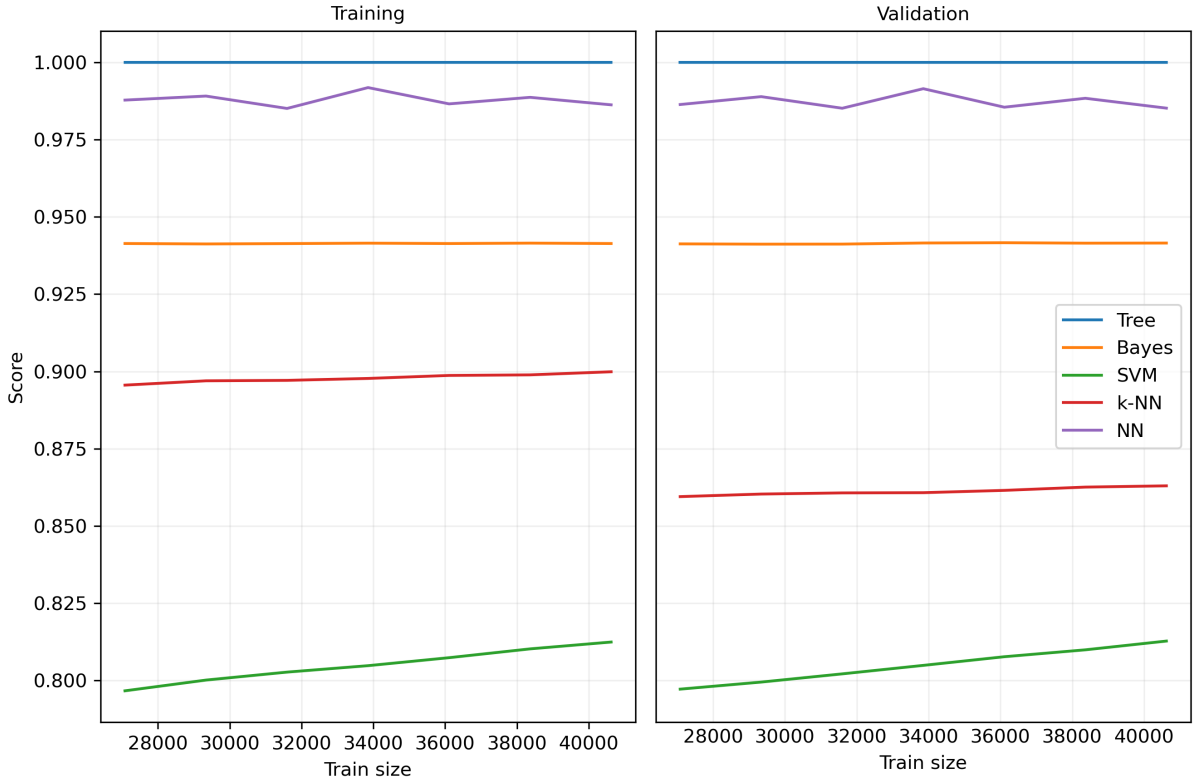

Rysunek 3: Krzywe ROC dla zbioru C

### 4.3 Krzywe uczenia się

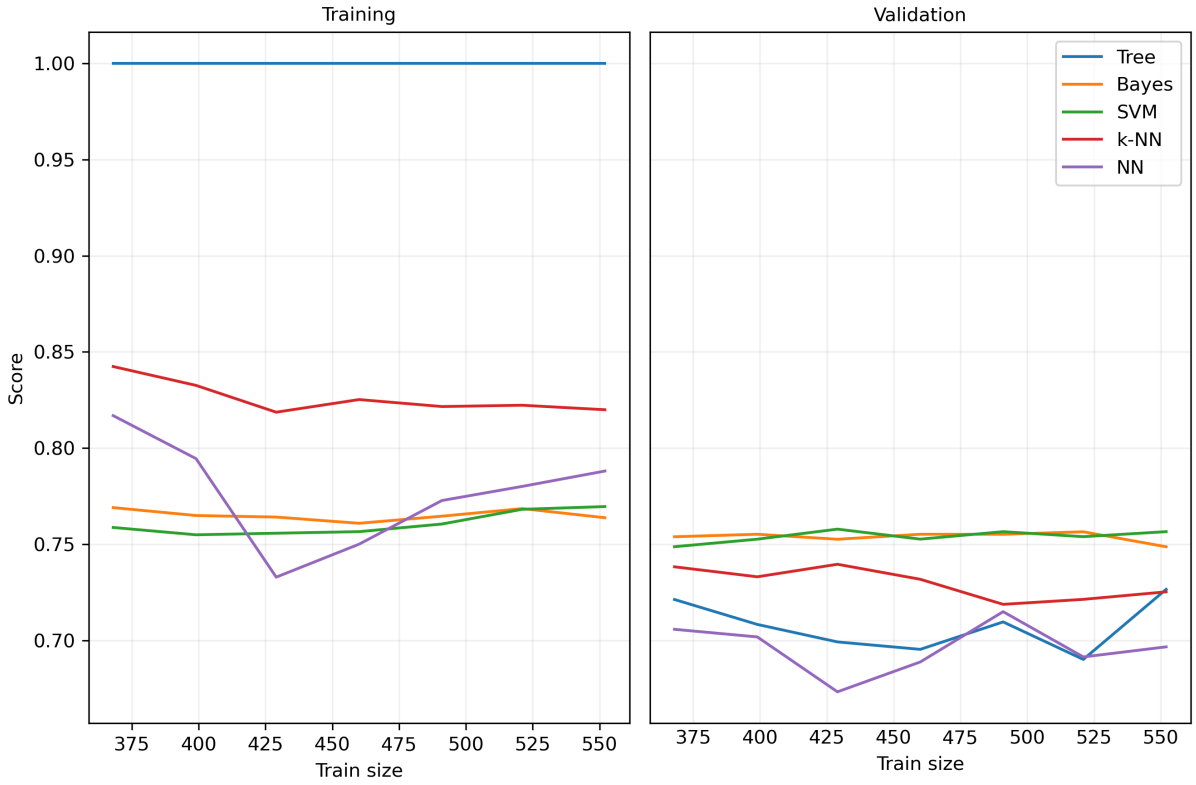
Kolejną częścią ewaluacji modelu było wyznaczenie krzywych uczenia się metod klasyfikacji. Wykres ten prezentuje proces zdobywania informacji przez klasyfikator w czasie. Dzięki temu określić można szybkość, z jaką dana metoda ”uczy się”. Dla każdego zestawu danych wygenerowane zostały wykresy porównujące wyniki na zbiorze treningowym oraz testowym. Każdy z klasyfikatorów był trenowany kolejno na 60, 65, 70, 75, 80, 85 i 90 procentach całego zbioru danych.




Rysunek 4: Zestawienie krzywych dla wszystkich klasyfikatorów, dla zbioru A




Rysunek 5: Zestawienie krzywych dla wszystkich klasyfikatorów, dla zbioru B




Rysunek 6: Zestawienie krzywych dla wszystkich klasyfikatorów, dla zbioru C

# 5 Ewaluacja modeli analizy skupień

Zadanie 2 wymagało stworzenia modelu analizy skupień opartego o następujące metody:

1. Algorytm EM,
2. Algorytm k-średnich,
3. Algorytm hierarchiczne aglomeracyjnego,
4. Metoda gęstościowej DBSCAN,
5. Metoda optics.

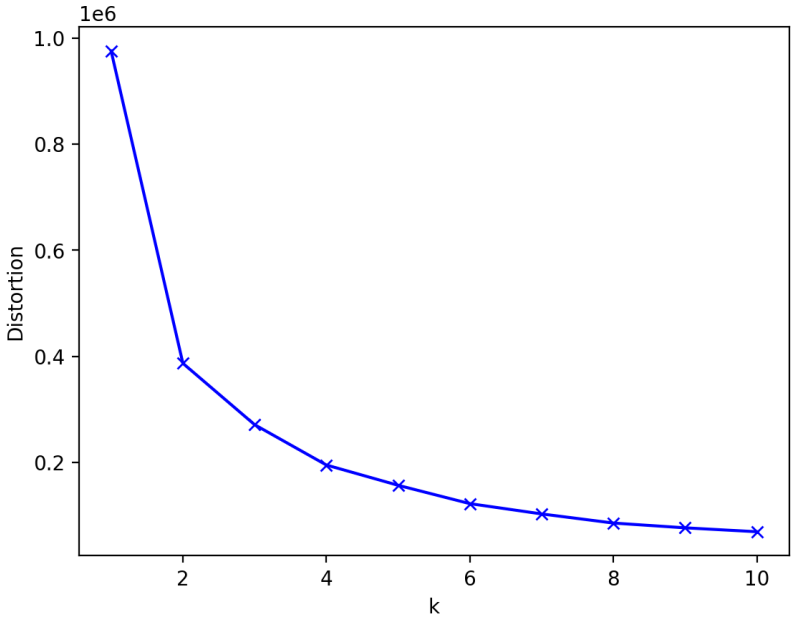
Ocenie poddane zostały wyznaczone w ramach danych metod skupienia.

## 5.1 Analiza wyznaczonych skupień

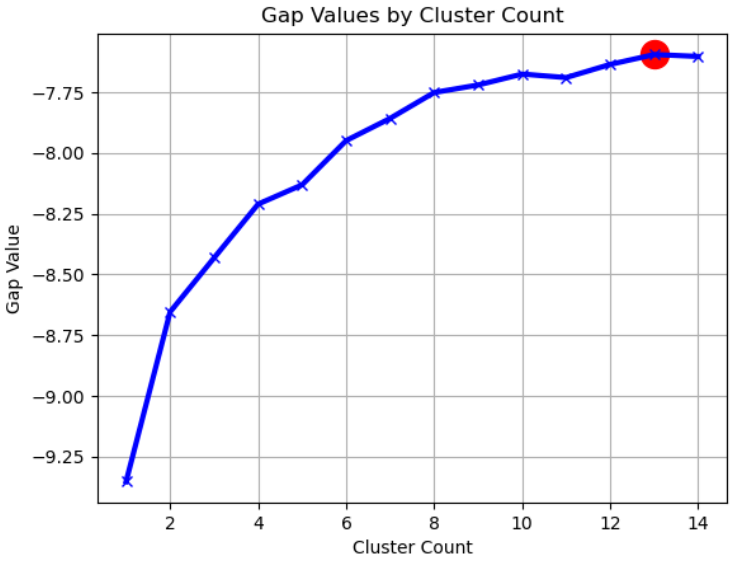
Ewaluacja modelu analizy skupień zakładała po pierwsze sprawdzenie, wyznaczonych różnymi metodami, liczby skupień przyjmowanych w algorytmach. Do tego celu został wykorzystana metoda łokcia która polega na wykreśleniu zmiany w funkcji liczby skupień i wybraniu krzywej jako liczby używanych skupień oraz metoda statystyki luk która określa liczbę klastrów w zbiorze danych. Następnie wygenerowane skupienia zostały poddane analizie jakości. Ze względu na specyfikę zbiorów danych - brak etykiet - wykorzystane zostały następujące miary wewnętrzne:

- Silhouette - jest miarą podobieństwa obiektu do własnego skupienia w porównaniu do innych skupień. Metoda silhouette ma przedział od -1 do +1, gdzie wysoka wartość wskazuje, że obiekt jest dobrze dopasowany do własnego klastra i słabo dopasowany do sąsiednich klastrów. Jeśli wiele punktów ma niską lub ujemną wartość, konfiguracja grupowania może mieć za dużo lub za mało klastrów.
- Calinski-Harabasz - metoda ta używa grupowania k-średnich, aby uzyskać wyniki dla różnych wartości k. Przebiegi k-średnich są losowo inicjowane i dlatego muszą być uruchamiane wiele razy, aby zapewnić optymalne grupowanie.
- Davies-Bouldin - jest to metoda w której sprawdzanie poprawności grupowania odbywa się przy użyciu ilości cech charakterystycznych dla zestawu danych.

### 5.2 Analiza zbioru danych A - ”Mall Customer Segmentation”



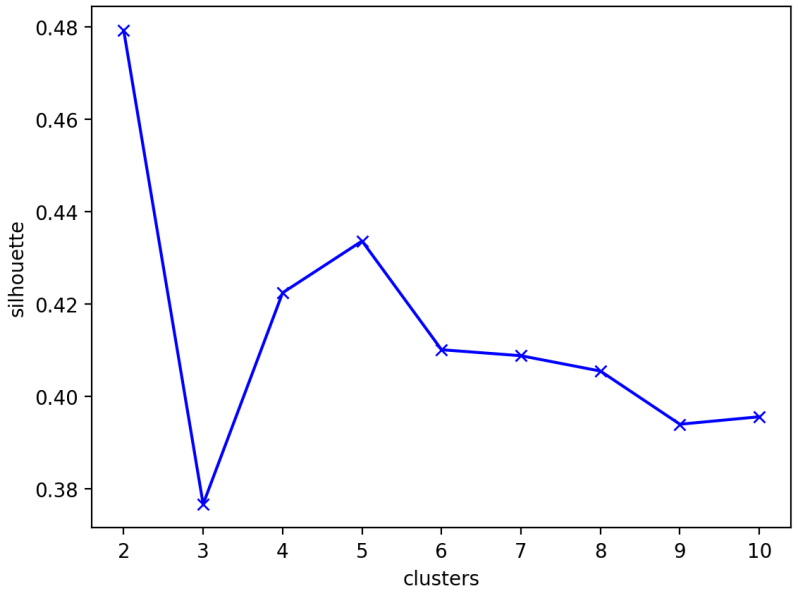

Rysunek 7: Metoda ”łokcia” dla zbioru danych A



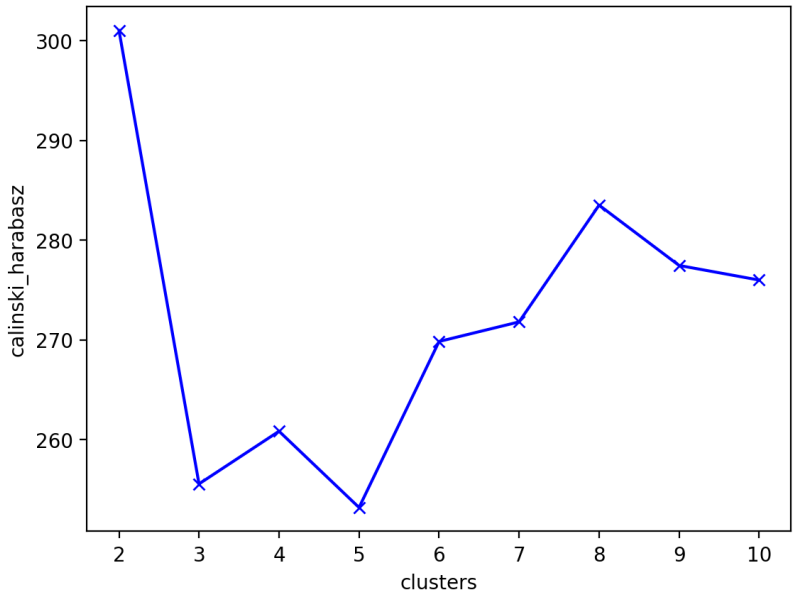

Rysunek 8: Metoda statystyki luk dla zbioru danych A

Tabela 19: Wartości metryk ewaluacyjnych dla zbioru A

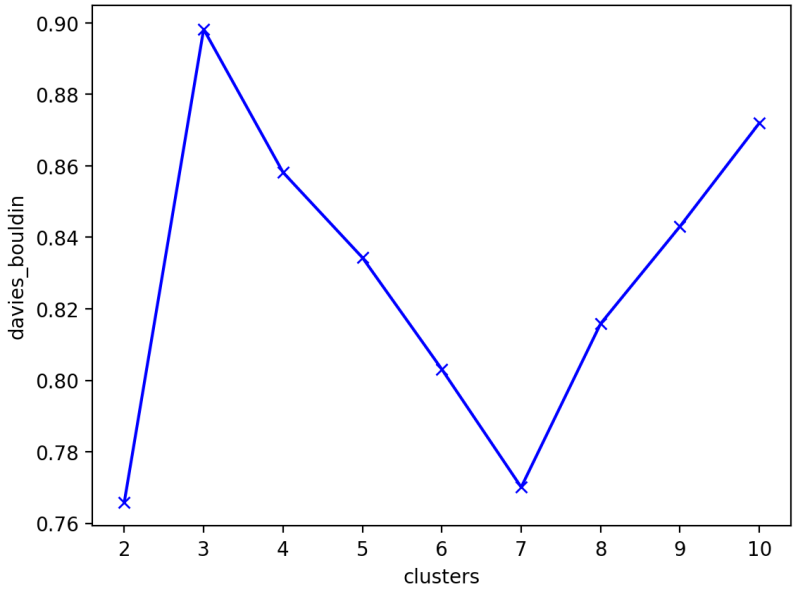
| klaster | Silhouette | Calinski-Harabasz | Davies–Bouldin |
|---|---|---|---|
| 2 | 0.479 | 301.014 | 0.765 |
| 3 | 0.374 | 255.565 | 0.900 |
| 4 | 0.422 | 260.833 | 0.858 |
| 5 | 0.421 | 253.831 | 0.866 |
| 6 | 0.410 | 269.849 | 0.809 |
| 7 | 0.408 | 272.036 | 0.773 |
| 8 | 0.405 | 283.170 | 0.816 |
| 9 | 0.395 | 277.426 | 0.822 |
| 10 | 0.388 | 275.874 | 0.868 |




Rysunek 9: Metoda Silhouette dla zbioru danych A

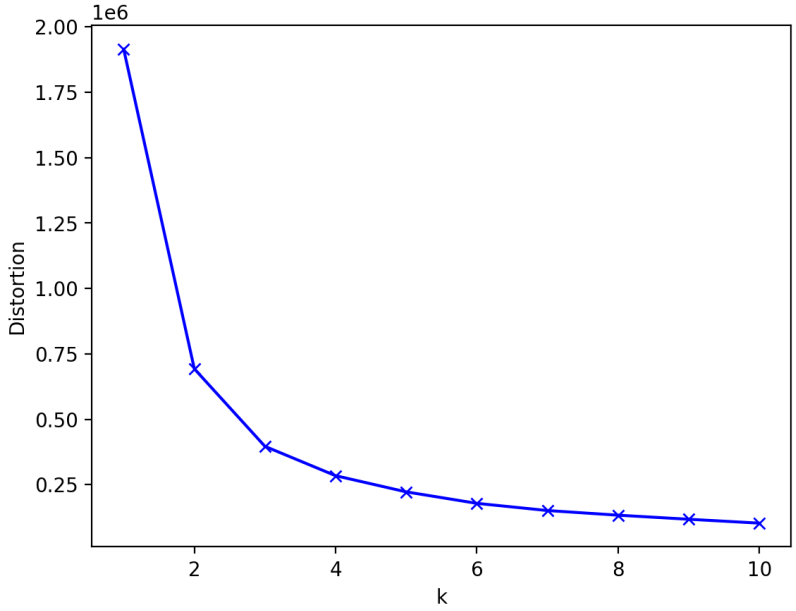



Rysunek 10: Metoda Calinski-Harabasz dla zbioru danych A

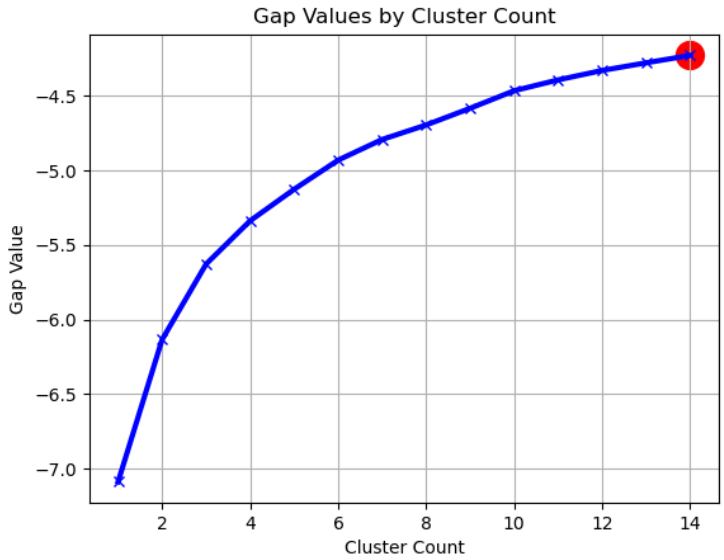



Rysunek 11: Metoda Davies-Bouldin dla zbioru danych A

### 5.3 Analiza zbioru danych B - ”Red Wine Quality”



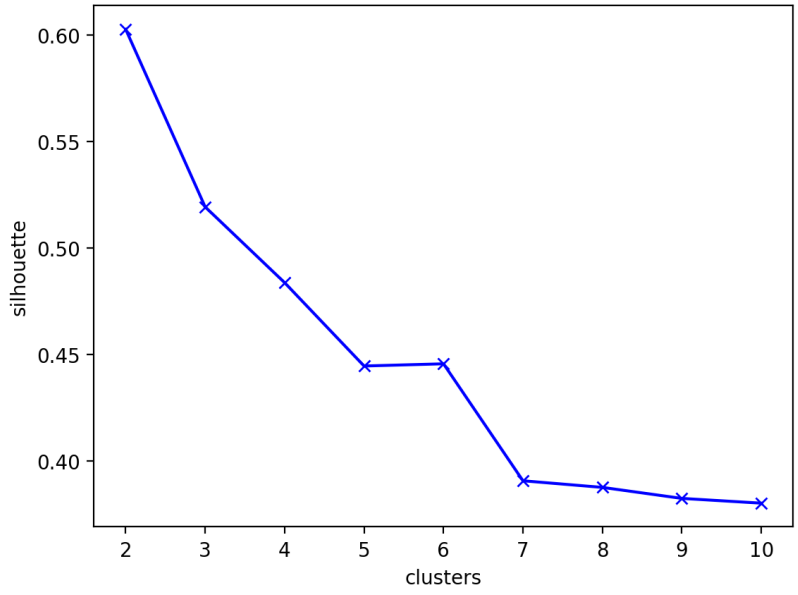

Rysunek 12: Metoda ”łokcia” dla zbioru danych B



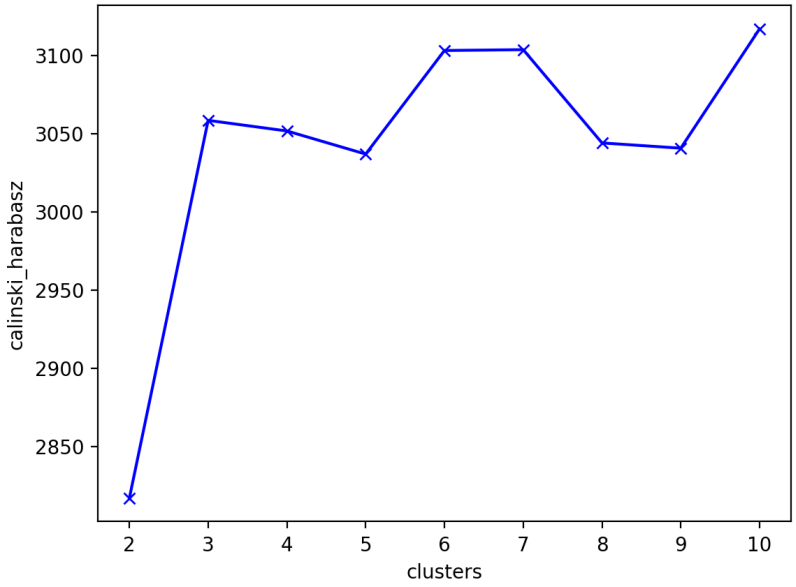

Rysunek 13: Metoda statystyki luk dla zbioru danych B

Tabela 20: Wartości metryk ewaluacyjnych dla zbioru B

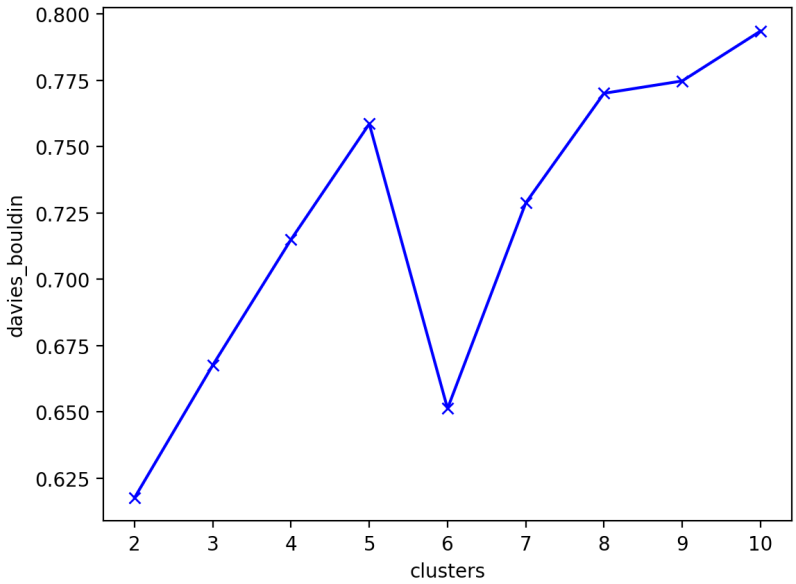
| **klaster** | **Silhouette** | **Calinski-Harabasz** | **Davies–Bouldin** |
|---|---|---|---|
| 2 | 0.602 | 2816.875 | 0.617 |
| 3 | 0.518 | 3058.177 | 0.664 |
| 4 | 0.483 | 3051.568 | 0.715 |
| 5 | 0.444 | 3036.735 | 0.758 |
| 6 | 0.446 | 3103.100 | 0.652 |
| 7 | 0.392 | 3103.871 | 0.734 |
| 8 | 0.393 | 3028.415 | 0.763 |
| 9 | 0.382 | 3040.573 | 0.774 |
| 10 | 0.380 | 3116.814 | 0.793 |




Rysunek 14: Metoda Silhouette dla zbioru danych B




Rysunek 15: Metoda Calinski$_{H}arabaszdlazbiorudanychB$




Rysunek 16: Metoda Davies$_{B}ouldindlazbiorudanychB$

# 6 Wyniki

- Dla największego zbioru danych B wszystkie klasyfikatory odznaczały się zauważalnie wyższą sprawnością, co przekładało się na wyniki uzyskiwane na zbiorze testowym. Algorytm drzew decyzyjnych nie popełnił błędu podczas klasyfikacji danych na tym zbiorze.
- Algorytm sztucznej sieci neuronowej popełnia zdecydowanie więcej błędów pierwszego rodzaju niż inne algorytmy. Jest to widoczne w Czułości, której wartość jest zawsze wysoka dla tego algorytmu. Dla dużego zbioru danych algorytm Bayesa także popełnił więcej błędów pierwszego rodzaju.
- Na krzywych ROC widoczne jest, że algorytm drzew decyzyjnych najlepiej sobie radzi dla dużych zbiorów - jego pole zauważalnie maleje przy zbiorach A oraz C. W przypadku algorytmów maszyny wektorów nośnych oraz k-nn większa liczba przypadków treningowych nie wpływa tak zauważalnie na poprawę pola pod krzywą.
- Na postawie krzywych uczenia można zaobserwować, że zjawisko niedouczenia wystąpiło w przypadku zbioru danych A o małej wielkości przy metodach drzew decyzyjnych oraz k najbliższych sąsiadów; Dla najmniejszego zbioru C wystąpiło również dla metody sieci neuronowej, a także w mniejszym stopniu dla naiwnego Bayesa oraz wektorów; Dla największego zbioru B jedynie metoda k najbliższych sąsiadów miała niższą sprawność dla danych testowych niż treningowych.
- Wszystkie przetestowane metody wyboru optymalnej liczby klastrów wskazały podobne wyniki.
- Dla dużych zbiorów danych najlepszymi metodami do oceny liczby skupień jest metoda łokcia oraz metoda luki.