


Prognozowanie satysfakcji i klasyfikacja klientów linii lotniczych



Zbiór danych

- **Rozmiar zbioru danych:** 129 487 ankiet przeprowadzonych na pasażerach linii lotniczych.
 - **Zakres danych:** 25 zmiennych opisujących różne aspekty podróży i charakterystyk pasażerów.
 - **Przykładowe zmienne:**
 - Typ podróży (służbowa, prywatna).
 - Płeć pasażera.
 - Metoda rezerwacji (online, biuro podróży).
 - Wiek pasażera.
-



Opis problemu do rozwiązania

- **Opracowanie modelu drzewa decyzyjnego do przewidywania satysfakcji pasażerów.**

Stworzenie i wdrożenie modelu, który pozwoli na precyzyjne prognozowanie poziomu zadowolenia pasażerów na podstawie dostępnych danych.

- **Zastosowanie algorytmu K-means w celu identyfikacji segmentów pasażerów oraz analizy ich zachowań.**

Grupowanie pasażerów na podstawie podobieństw w ich cechach i zachowaniach, aby lepiej zrozumieć różnorodne potrzeby i oczekiwania.

- **Analiza wyników i wyciągnięcie praktycznych wniosków.**

Interpretacja uzyskanych rezultatów w celu sformułowania rekomendacji i strategii mających na celu poprawę satysfakcji pasażerów oraz optymalizację usług.

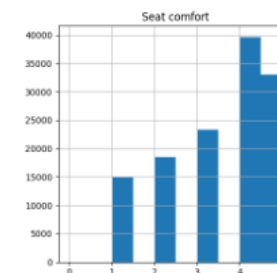
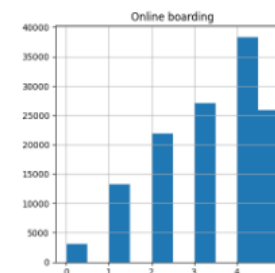
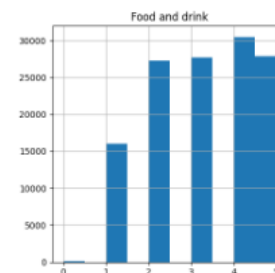
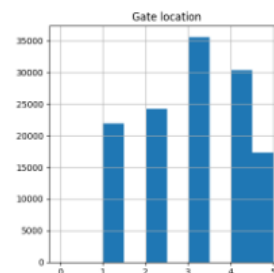
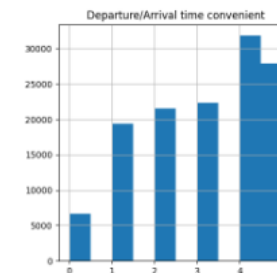
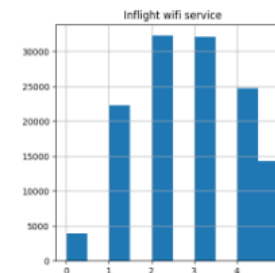
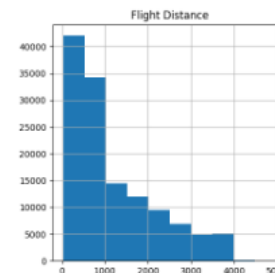
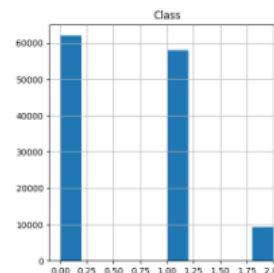
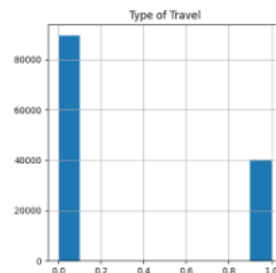
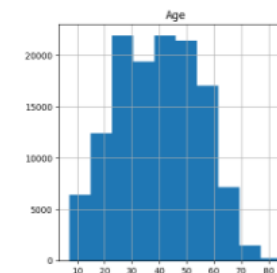
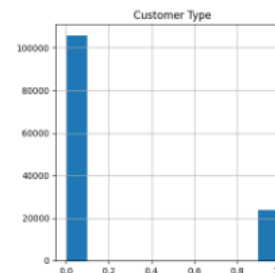
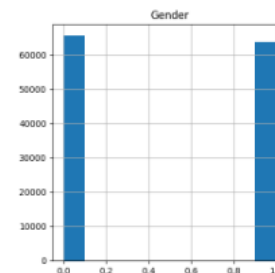
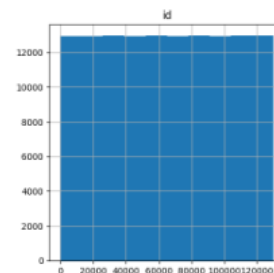
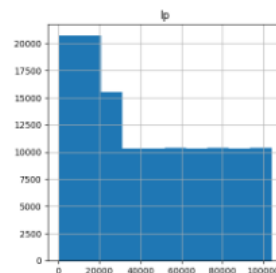
Praca z danymi



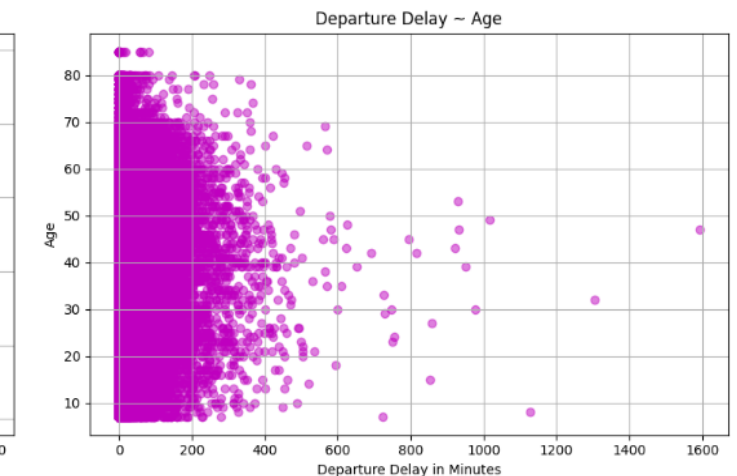
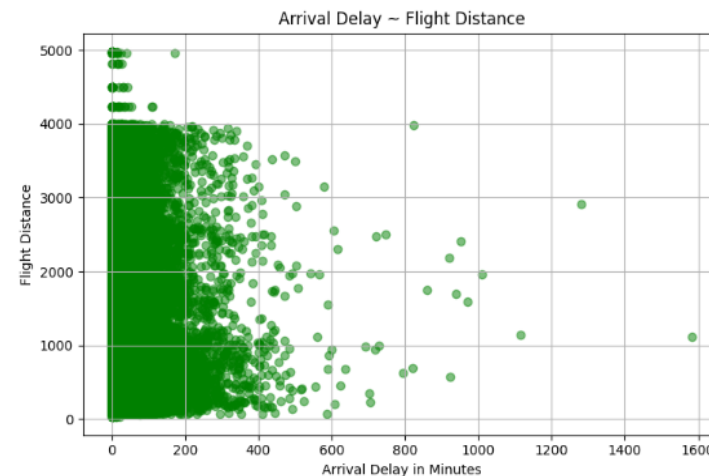
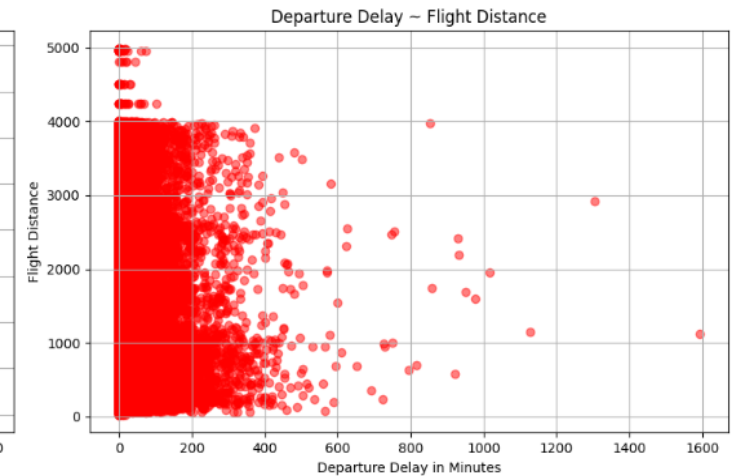
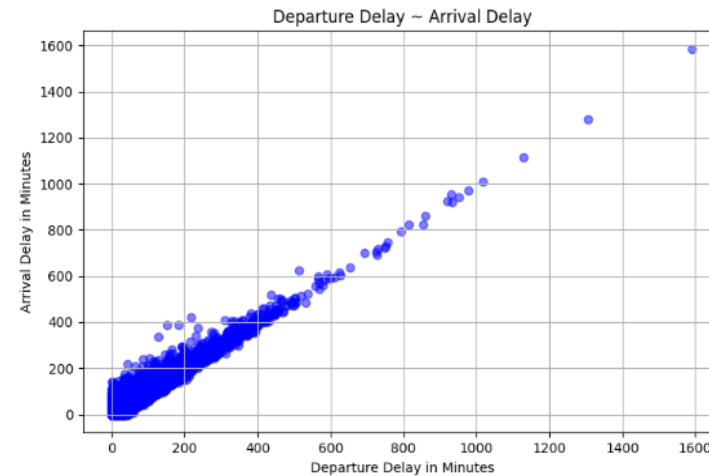
- **Wczytanie danych:** Import danych z pliku źródłowego.
- **Sprawdzenie brakujących wartości:** Identyfikacja i uzupełnienie braków.
- **Kodowanie zmiennych kategorycznych:** Zamiana na wartości liczbowe.
- **Statystyki opisowe:** Wyliczenie średniej, mediany, wariancji, oraz dodanie dominanty.



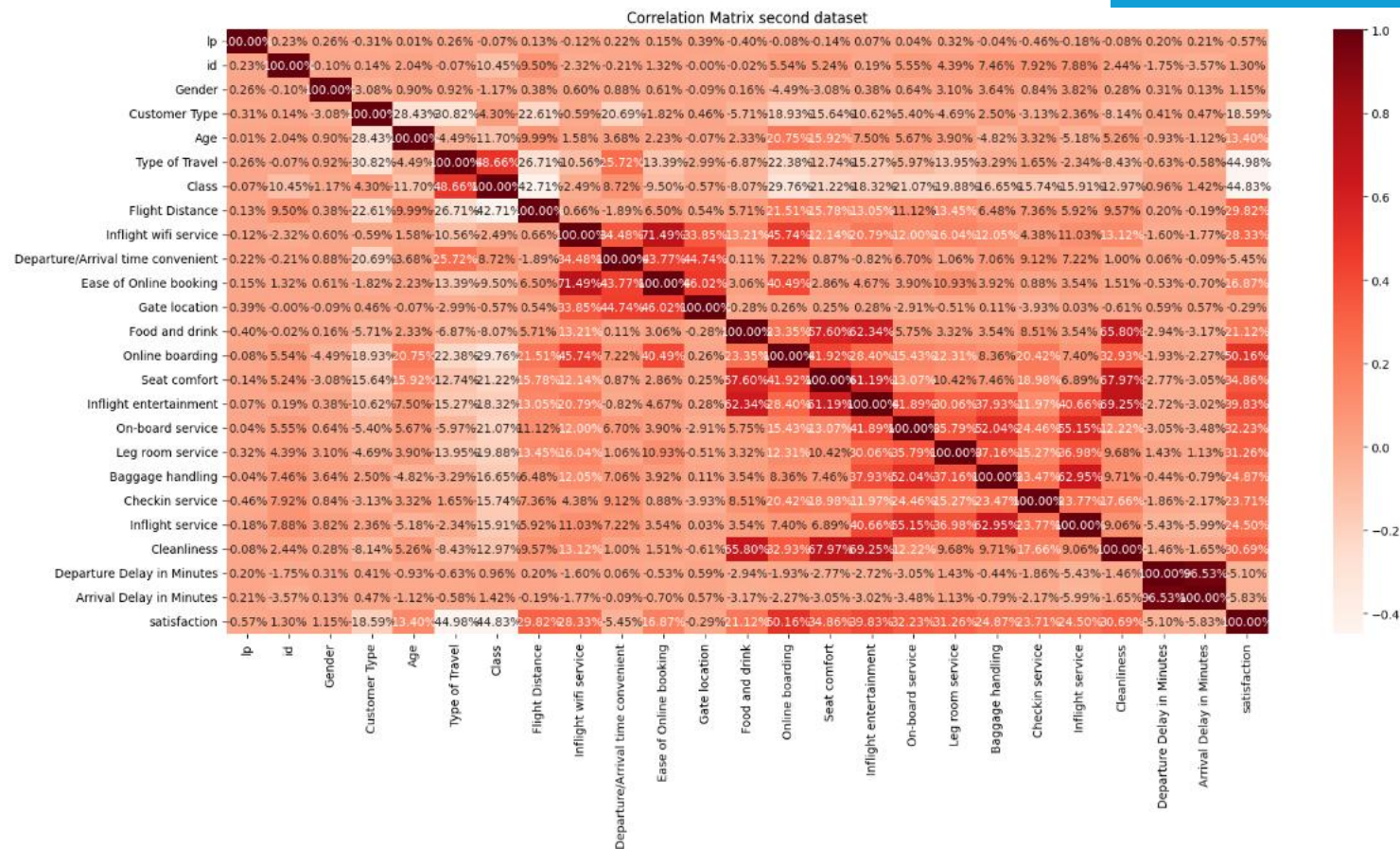
Prezentowanie histogramów




Tworzenie wykresów
rozrzutu dla zależności
między poszczególnymi
zmiennymi



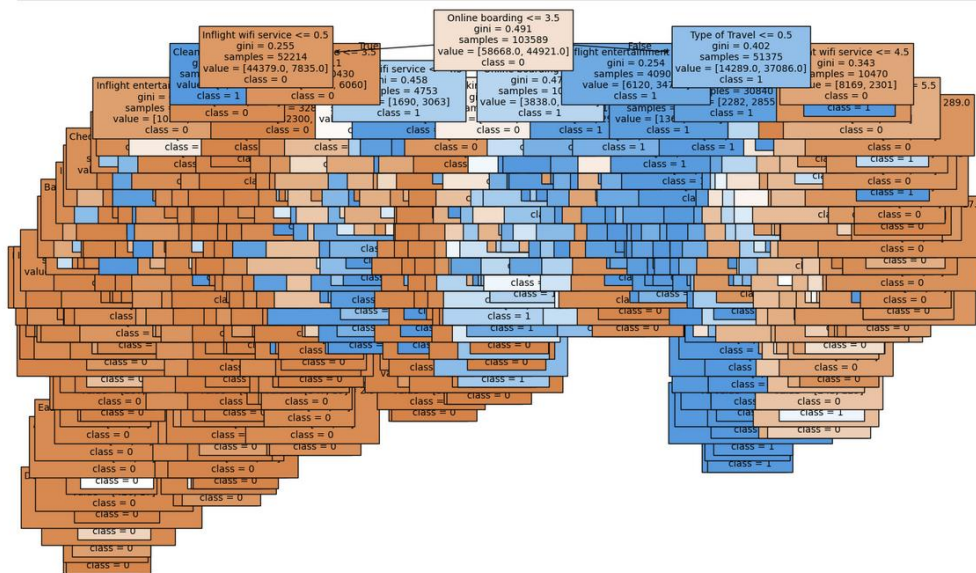
Tworzenie wykresu z macierzą korelacji



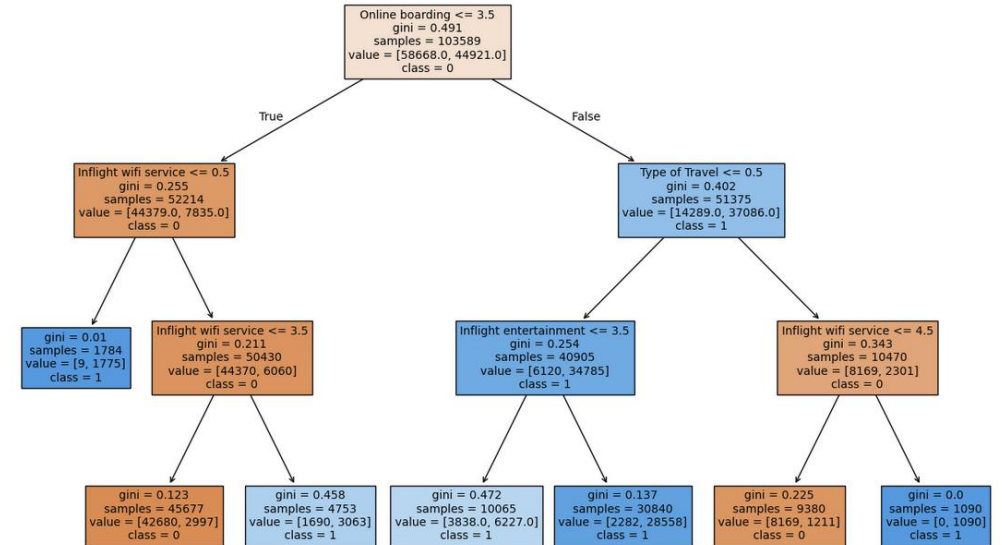
- 
- Użyta funkcja podziału drzewa: GINI
 - Accuracy na zbiorze treningowym: 95,6%.
 - Accuracy na zbiorze testowym: 94,9%
 - min_samples_split: 200
 - Przy zmianie wartości min_samples_split z domyślnego 2 na 200 accuracy zbioru trenningowe spadła ze 100% (overfitting) na 95,6%, accuracy zbioru treningowego wzrosło z 94.65 % na 94.9%
-

Wpływ zmianny ccp_alpha (Cost Complexity Pruning Alpha)

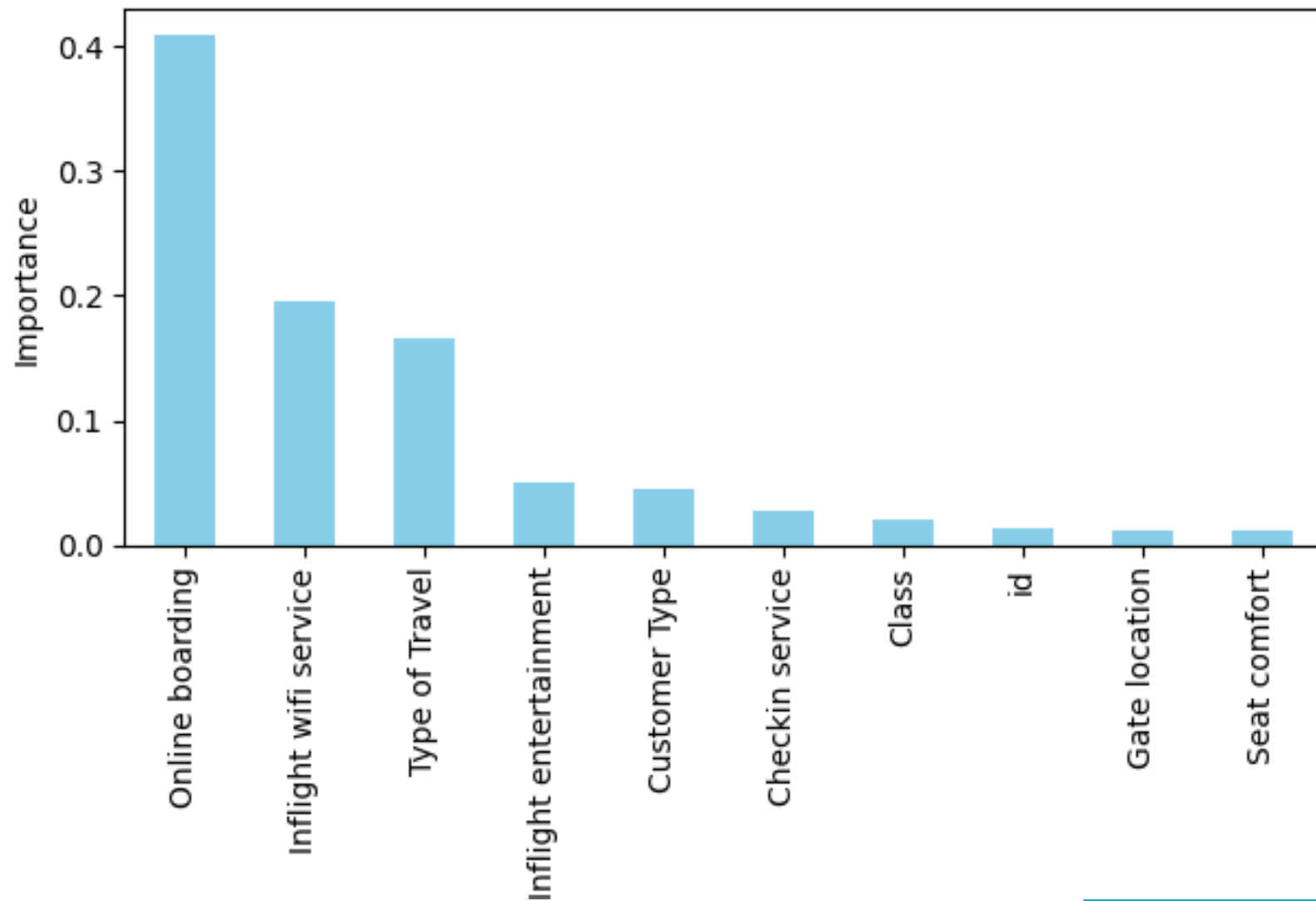
ccp_alpha=0



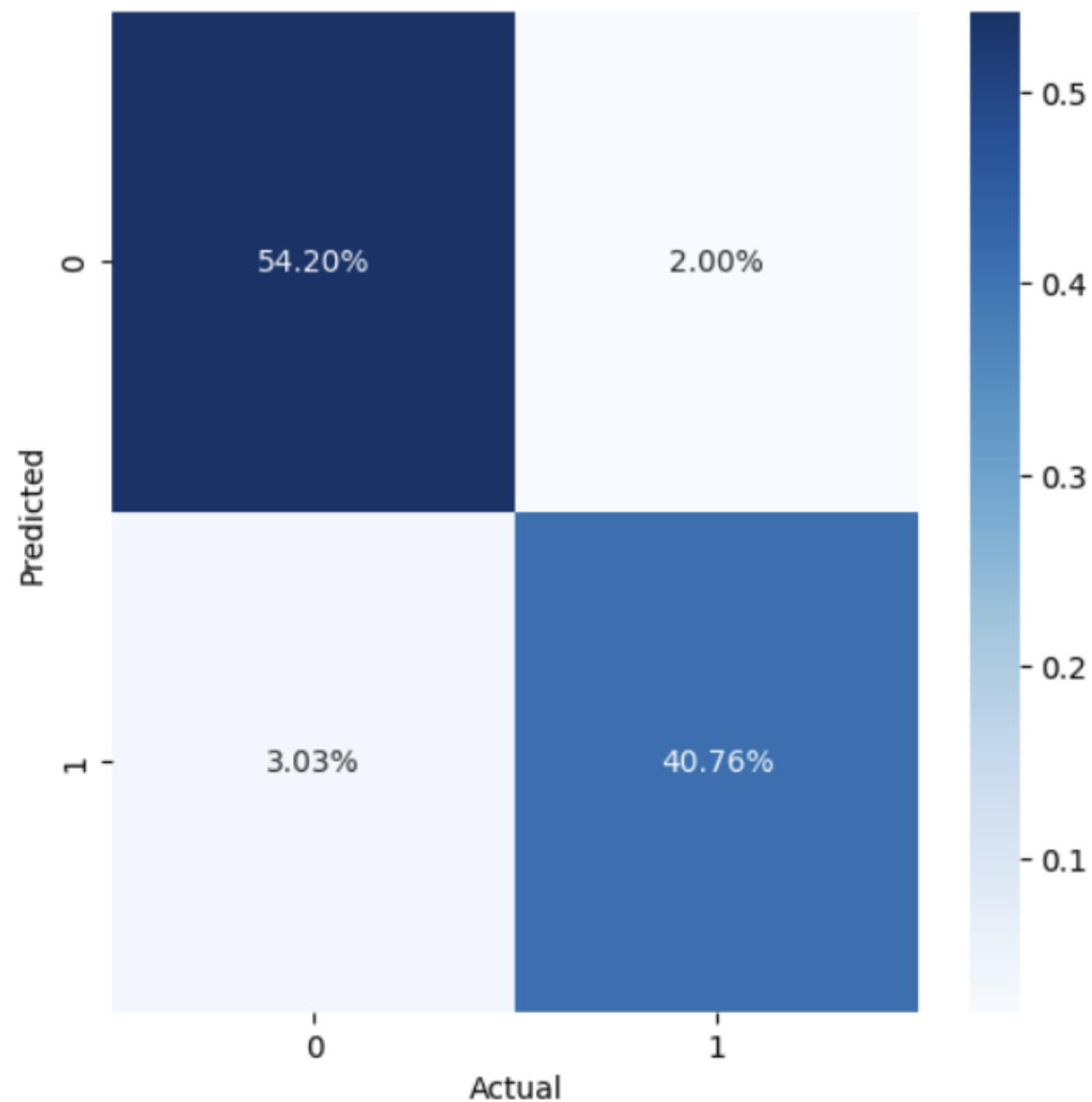
ccp_alpha=0.01



Istotność zmiennych



Confusion matrix drzewa decyzyjnego

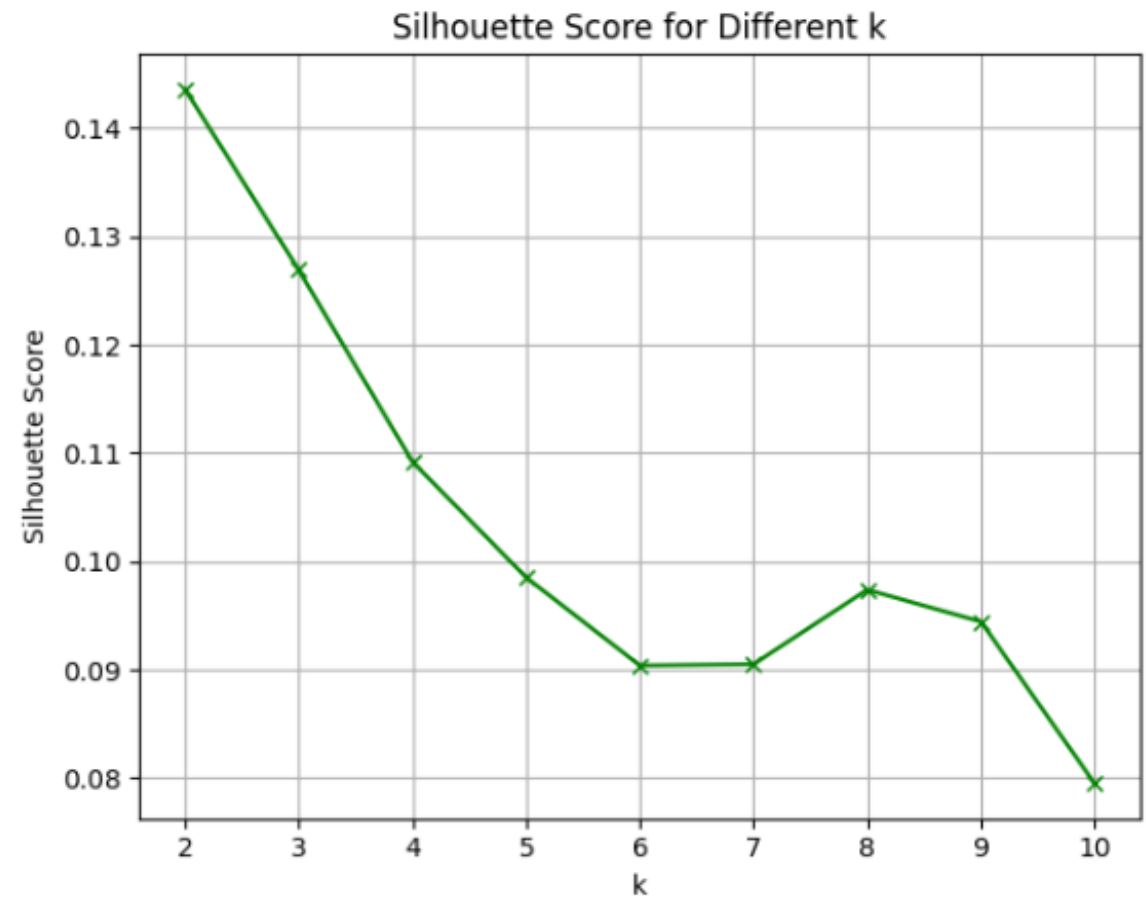
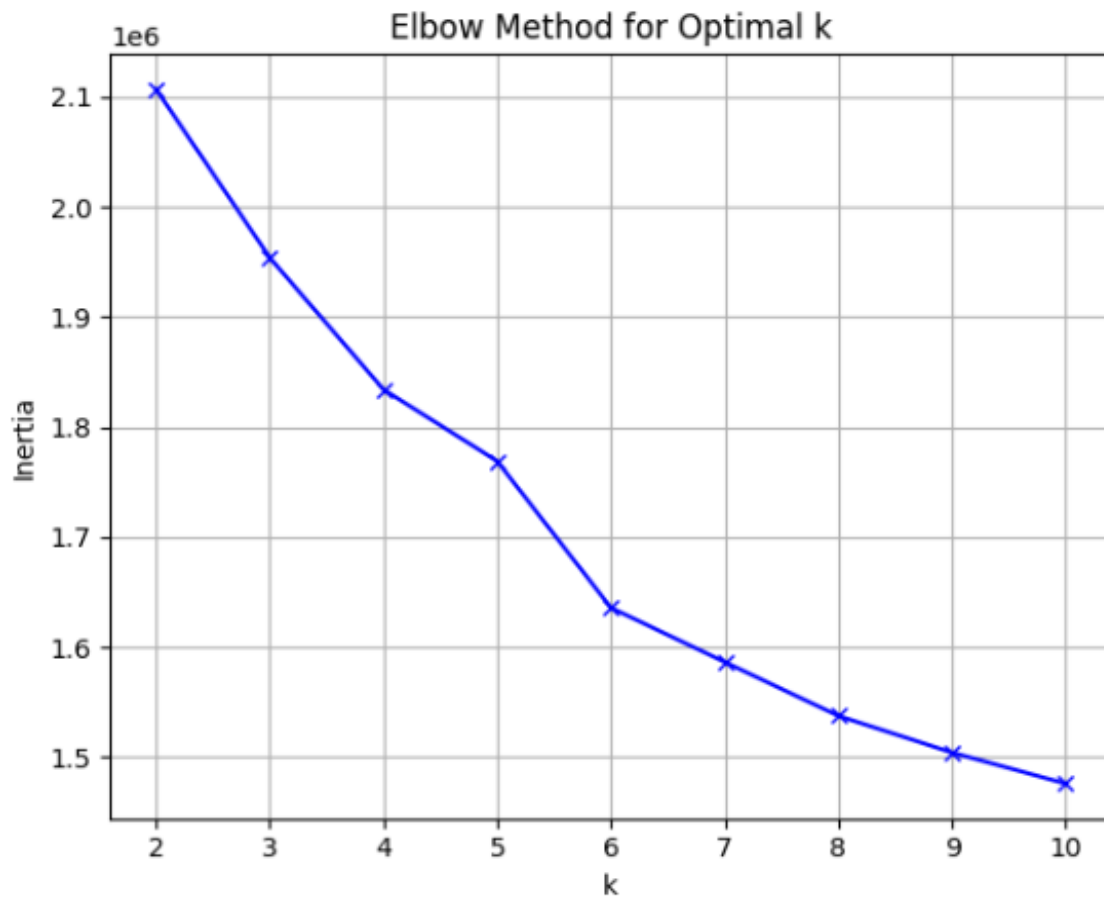




Algorytm centroidów (K-means clustering)

- Ponowne utworzenie zbioru danych.
 - Ustandaryzowanie cech.
 - Wyznaczenie optymalnej liczby klastrów metodą "łokcia".
 - Obliczenie Silhouette Score dla próbki danych (dla całego zbioru zajmowało zbyt dużo czasu)
-

Kreślenie wykresu bezwładności (elbow curve) oraz wyniku profilu (silhouette score)



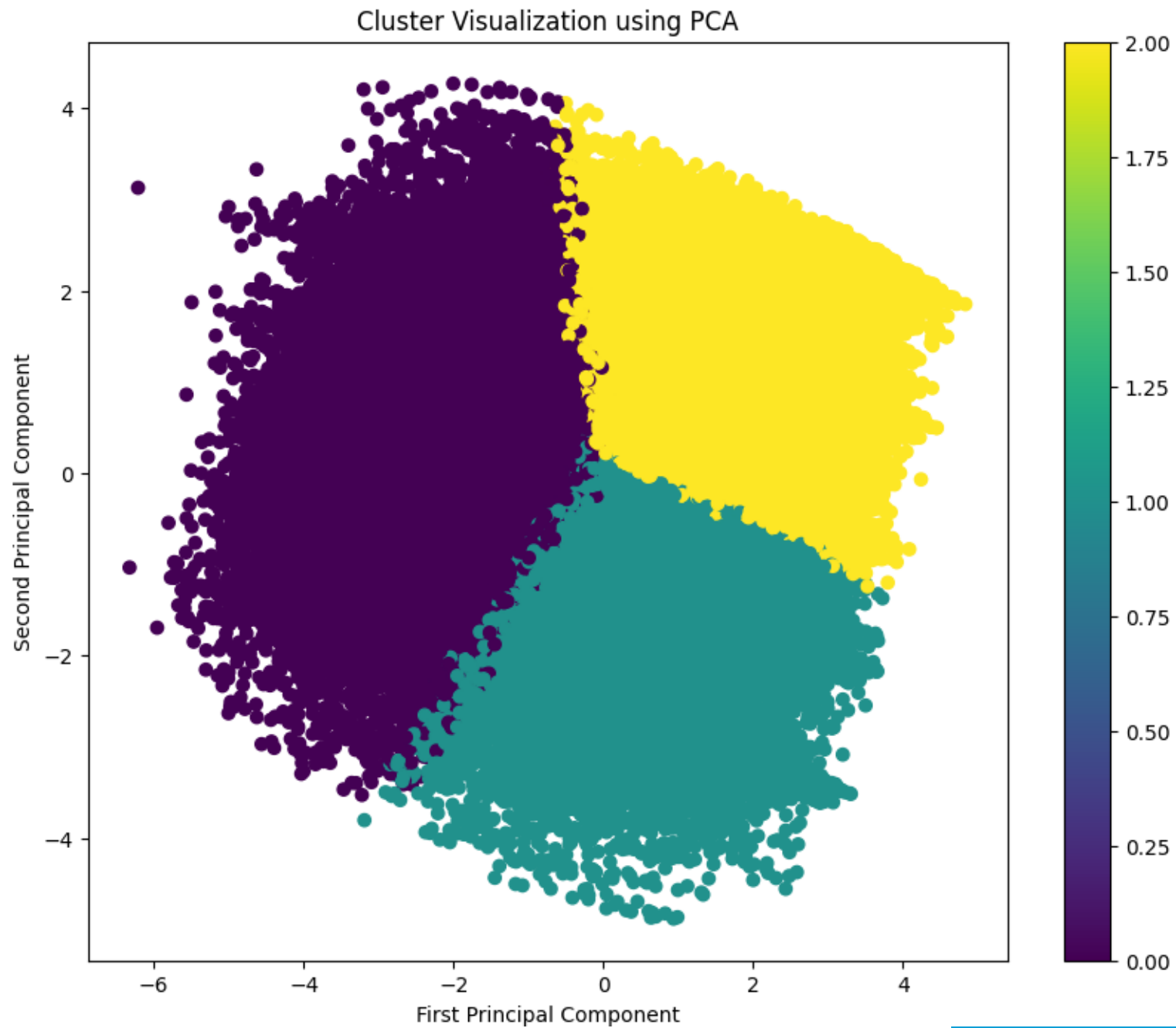


- Inicjalizacja modelu KMeans z wybraną liczbą klastrów
- Analiza klastrów
- Analiza częstotliwości dla kolumn kategorycznych w klastrach





Wizualizacja klastrow





Wnioski

- Jeśli dane są dobrze zorganizowane i mają wyraźnie zdefiniowane klasy, drzewo decyzyjne będzie najodpowiedniejszą metodą.
 - Jeśli dane są bardziej "rozproszone" i potrzebujesz segmentować je w oparciu o podobieństwa, warto zastosować K-means.
-



Skład grupy projektowej oraz podział obowiązków

- Natalia Kwaśniewska - Drzewo decyzyjne, Prezentacja
 - Franciszek Dębicki - Algorytm centroidów, Prezentacja
 - Cyprian Jurkowski – korekty kodu, opracowanie danych , Prezentacja
 - Kamil Kaplita - Algorytm centroidów, Prezentacja
 - Kamil Łempicki - Drzewo decyzyjne, Prezentacja
-