Wydział Informatyki i Telekomunikacji

Sztuczna Inteligencja i Inżynieria Wiedzy

Lista nr 4

Jakub Krupiński
255356

1. Wprowadzenie

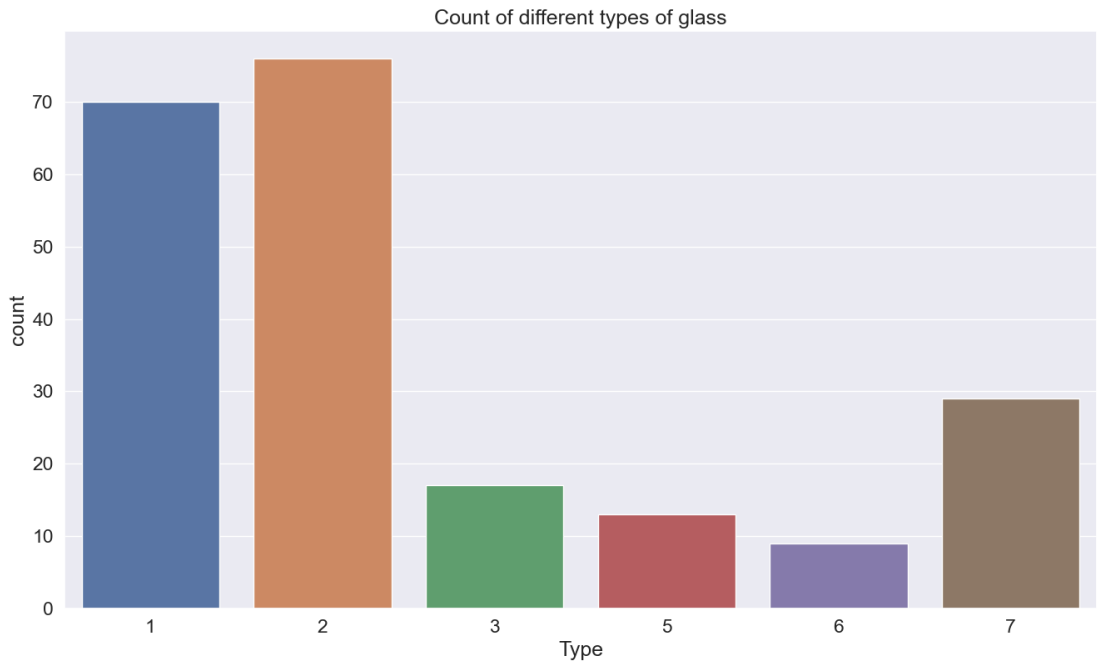
W wykonywanym ćwiczeniu przeprowadzam analizę zestawu danych dotyczącego klasyfikacji szkła. Źródłem danych jest plik ‘glass.data’. W ramach ćwiczenia wykonałem:

a. Odczytanie danych z pliku
b. Wstępną eksplorację danych
c. Podział zestawu danych na zestaw uczący metodami: PCA, selekcji cech, normalizacji, standaryzacji, dyskretyzacji i skalowania min-max, a także bez procesowania
d. Testowanie klasyfikatorów na podstawie metod: SVC, drzewa decyzyjnego i naiwnego klasyfikatora Bayesa
e. Ocenę klasyfikacji

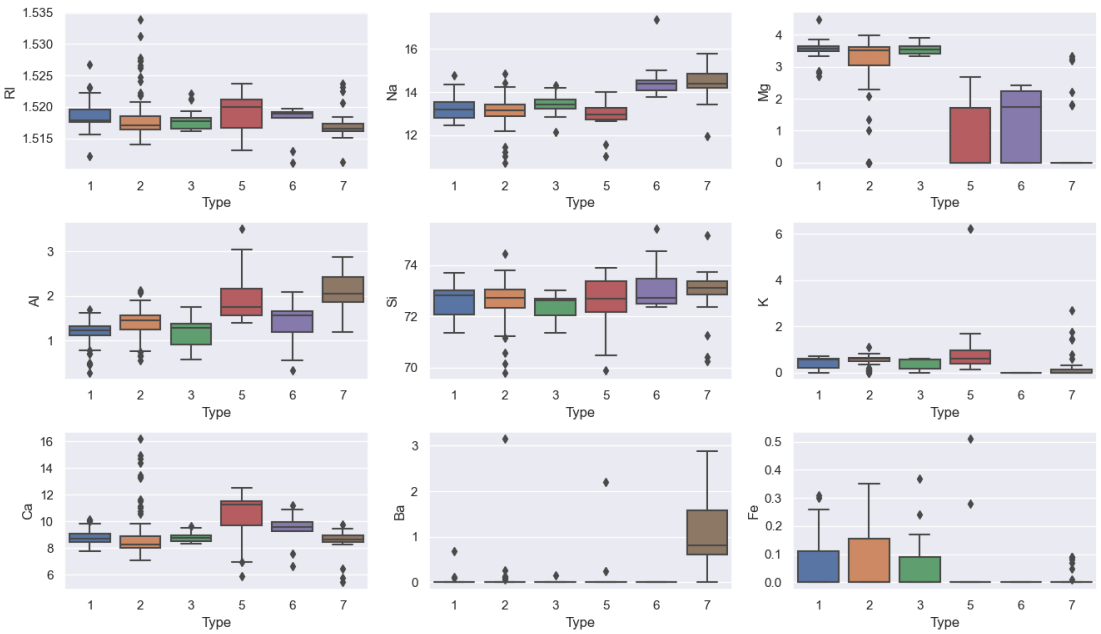
2. Eksploracja danych

Badany dataset dotyczy klasyfikacji rodzajów szkła. Posiada 214 rekordów, 10 atrybutów (w tym indeks) i atrybut klasy. Wyróżnione jest też 7 różnych rodzajów szkła (typ 4 nie znajduje się w datasecie).

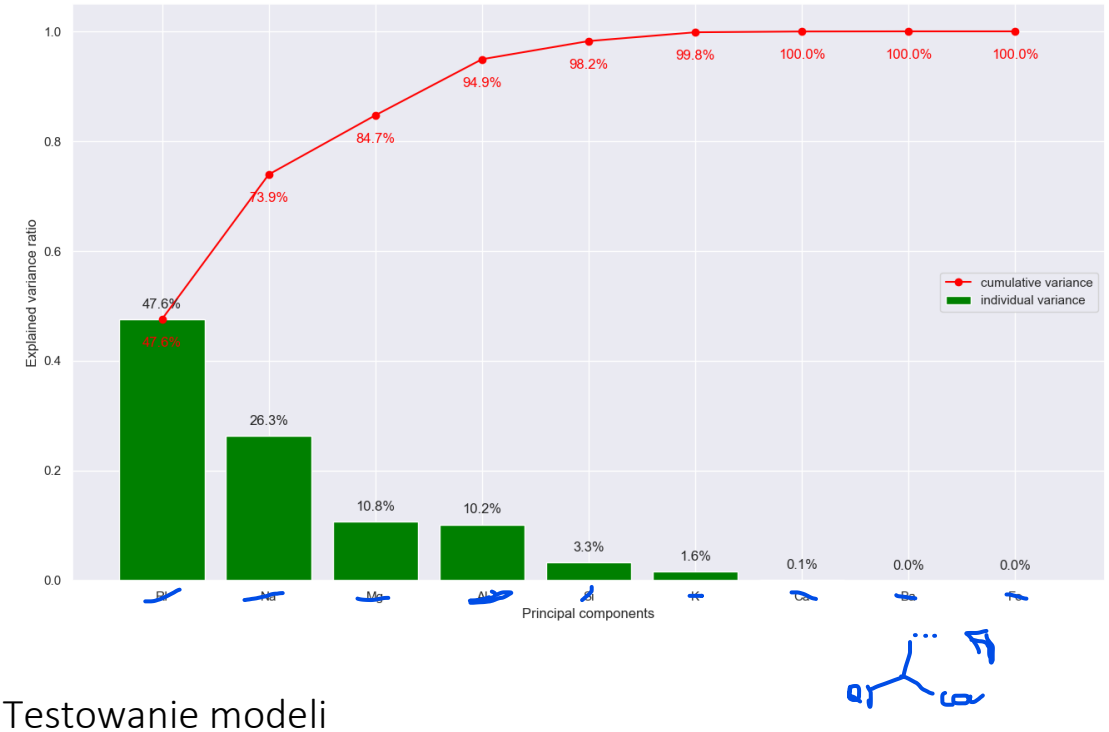
Pierwszą rzeczą którą wykonałem w ramach eksploracji było sprawdzenie tego jak różnorodny jest badany w ramach tego ćwiczenia dataset - po załączonym wykresie widać od razu, że pierwszy i drugi typ szkła powtarzają się zdecydowanie częściej niż pozostałe



Kolejną przebadaną rzeczą było sprawdzenie tego jak rozkładają się wartości różnych parametrów w datasecie – widać tu, że zarówno same wartości jak i same zakresy potrafią znacznie się różnić pomiędzy typami szkła, co może stanowić pewną poszlakę w kontekście przyszłego przetwarzania tych danych – można tu zobaczyć np., że wartości takich parametrów jak RI, SI czy Al są dość równomiernie rozłożone, zarówno pod względem wartosci jak ich zakresów, podczas gdy wartości np. Ba są bardzo nierówne i ciężko będzie sprawić, by faktycznie były użyteczne dla moich modeli.



Ostatni wykres bada proporcje wariacji dla każdego komponentu i ich dokładny wpływ – widać tu, że wpływ ostatnich trzech parametrów: Ca, Ba i Fe jest praktycznie niezauważalny – co zostanie później wykorzystane przy tworzenie podziału danych, np. dla PCA.

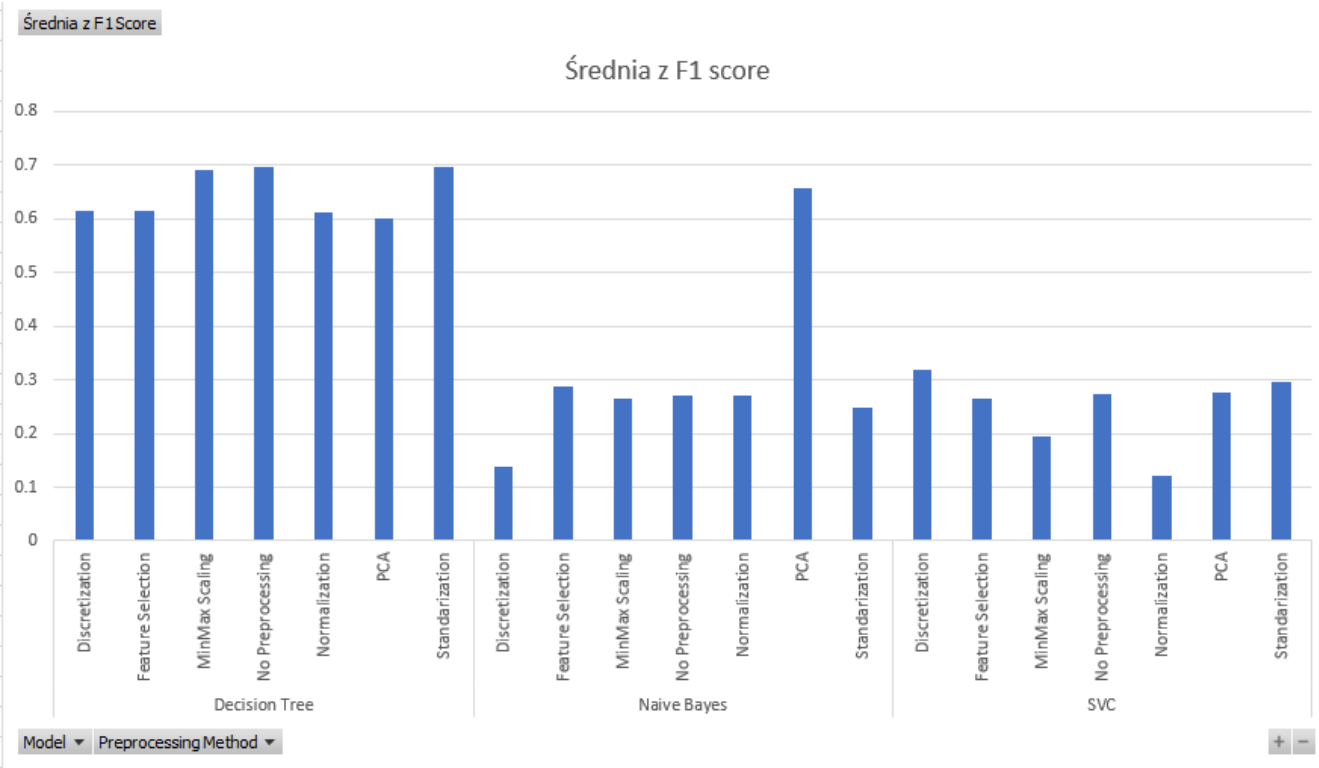


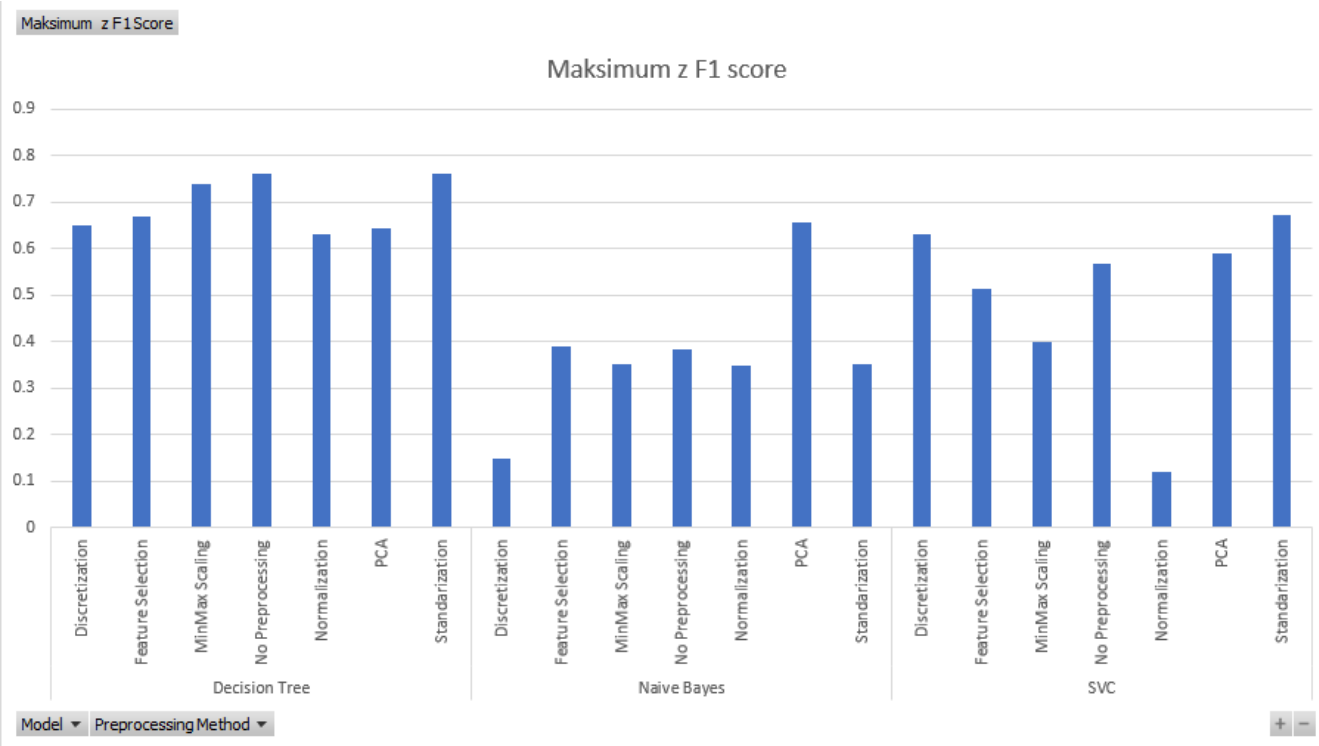
3. Testowanie modeli

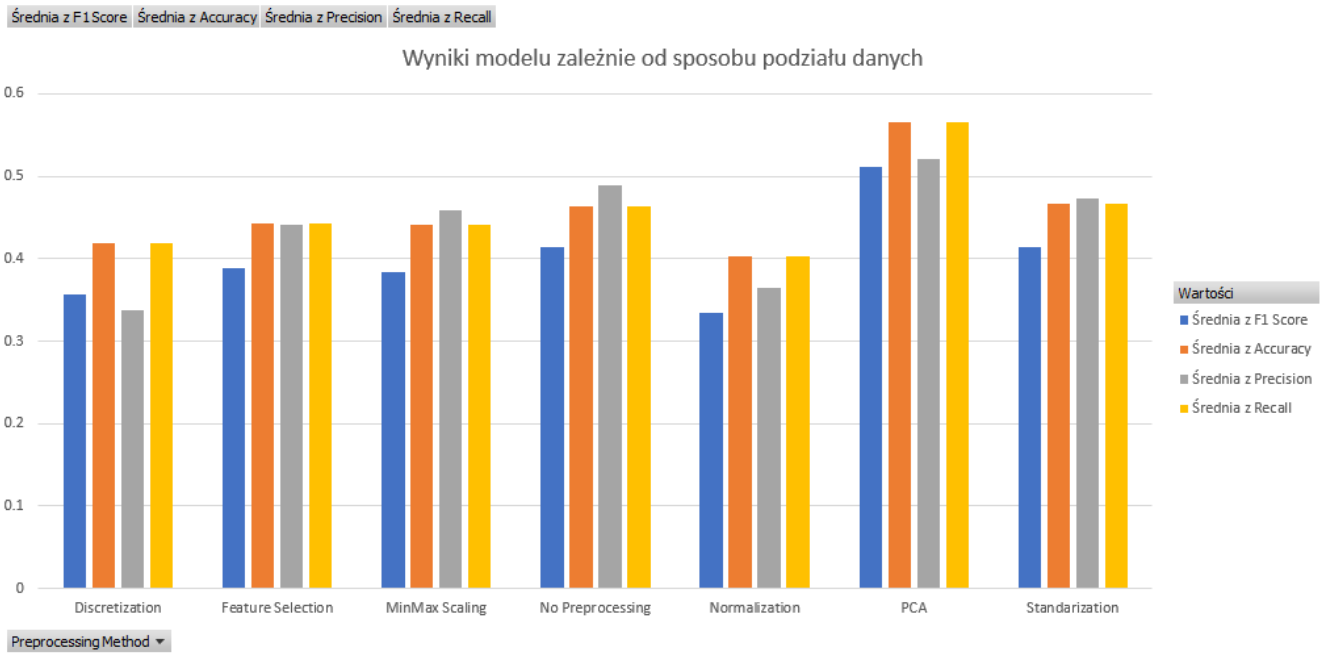
Modele te są testowane dla trzech modeli: SVC, DecisionTree oraz przez klasyfikator Bayesa. Każdy z nich testowany jest dla każdej z metod przygotowania danych, dla każdego z pięciu zestawów hiperparametrów. Wyniki każdego z modeli są oceniane przez: accuracy score, precision score, recall score i f1 score, a nastepnie sa zapisywane do pliku .csv za pomocą biblioteki pandas.

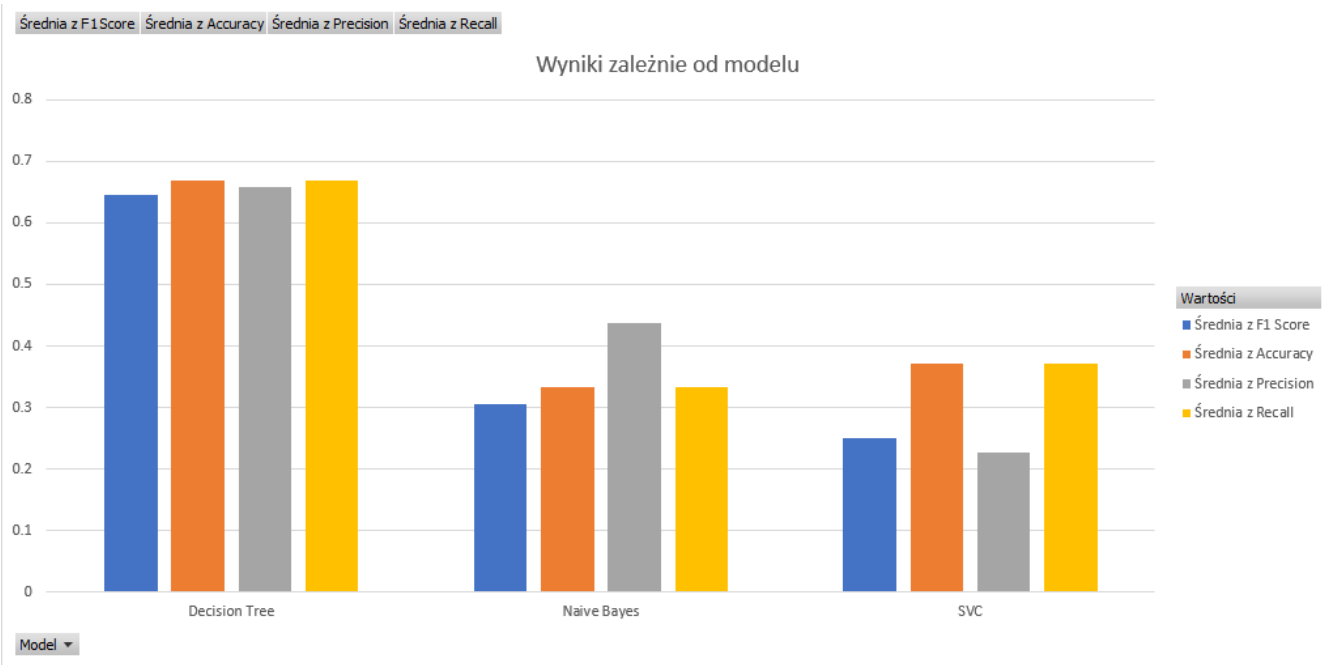
Feature selection oraz PCA wywoływane sa odpowiednio dla k=6 oraz n_bins=6. Każdy model w moim programie wywoływany jest dla ‘losowego’ seedu 1, aby zapewnić powtarzalność otrzymywanych wyników.

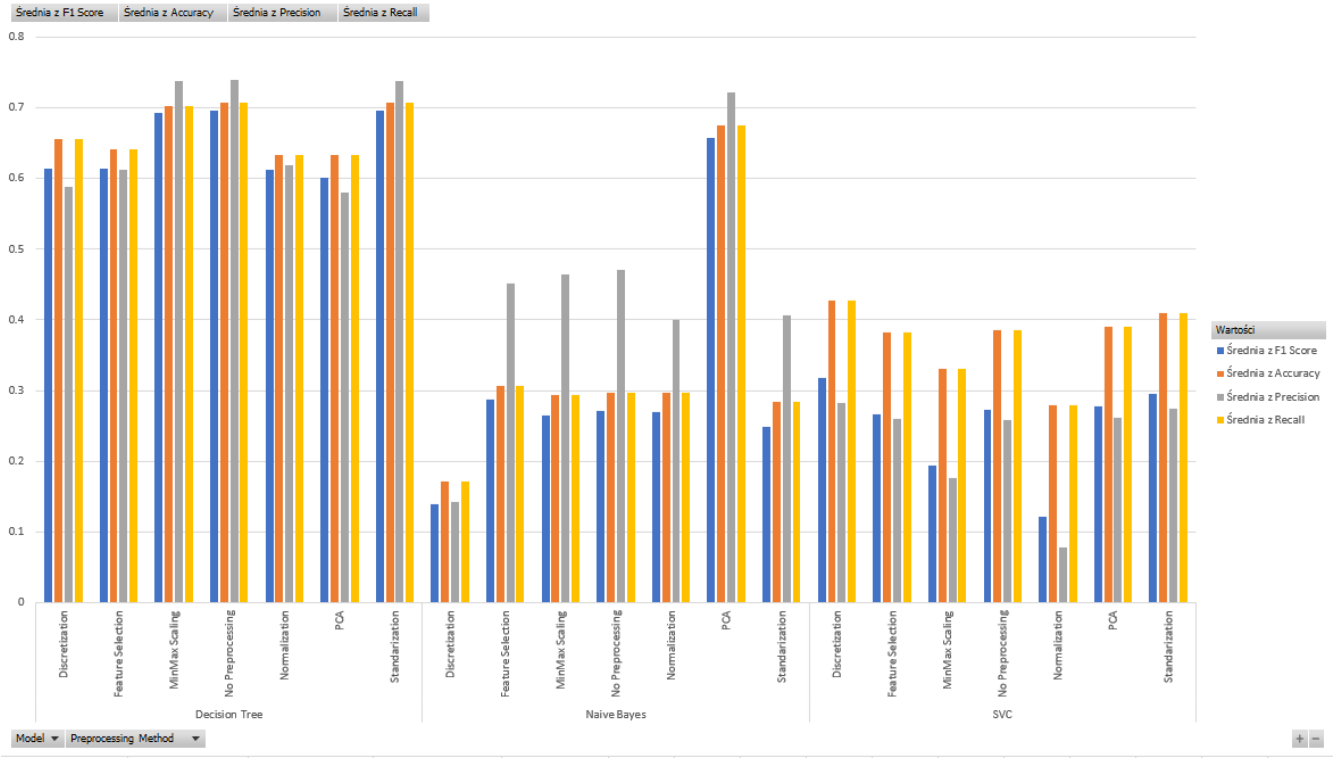
4. Analiza wyników






Średnia z F1 Score
Średnia z Accuracy
Średnia z Precision
Średnia z Recall
Wyniki modelu zależnie od sposobu podziału danych
0.6
0.5
0.4
0.3
0.2
0.1
0
Discretization
Feature Selection
MinMax Scaling
No Preprocessing
Normalization
PCA
Standarization
Wartości
Średnia z F1 Score
Średnia z Accuracy
Średnia z Precision
Średnia z Recall
Preprocessing Method


Średnia z F1 Score
Średnia z Accuracy
Średnia z Precision
Średnia z Recall
Wyniki zależnie od modelu
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Decision Tree
Naive Bayes
SVC
Wartości
Średnia z F1 Score
Średnia z Accuracy
Średnia z Precision
Średnia z Recall
Model



Na podstawie wygenerowanego przeze mnie pliku .csv sporządziłem sobie 5 wykresów, które przedstawiają wyniki modeli w zależności od sposobu przygotowania danych oraz użytego modelu. Co ciekawe – dyskretyzacja bardzo słabo poradziła sobie w przypadku naiwnego klasyfikatora Bayesa, natomiast normalizacja w przypadku SVC. Bez wątpienia jest to po części wynik niepoprawnie dobranych hiperparametrów. Drzewa decyzyjne poradziły sobie najlepiej – wyniki są dla nich najwyższe,a różnice między róznymi rodzajami preprocesingu danych są najmniejsze. Warto zwrócić jednak uwagę na to, że dla drzewa decyzyjnego bardzo dobre wyniki osiągnął model, który pracował na czystych danych – może to wynikać z faktu, że zestaw danych jest tu nieduży, a nieodpowiednie skalibrowanie modeli może doprowadzić do tego, że dojdzie do utraty części informacji – dlatego należałoby poświęcić moim modelom nieco więcej czasu i przeprowadzić eksperymenty na różnych wartościach hiperparametrów, aby upewnić się, że znalezione rozwiązanie na pewno jest optymalne. Jeśli chodzi o wykres średnich wartości wyników F1 w zależności od metody przygotowania danych – sugeruje on, że metoda PCA była tutaj najbardziej skuteczna, należy jednak zauważyć, że owe średnie wartości dotyczą wszystkich trzech klasyfikatorów włącznie – dla klasyfikatora Bayesa PCA poradził sobie wyjątkowo dobrze, przez co ogólna średnia ocena tej metody znacznie wzrosła, mimo że zarówno dla SVC jak i dla drzewa decyzyjnego PCA osiągnął wyniki niewiele różniące się od tych osiągniętych przez inne metody.