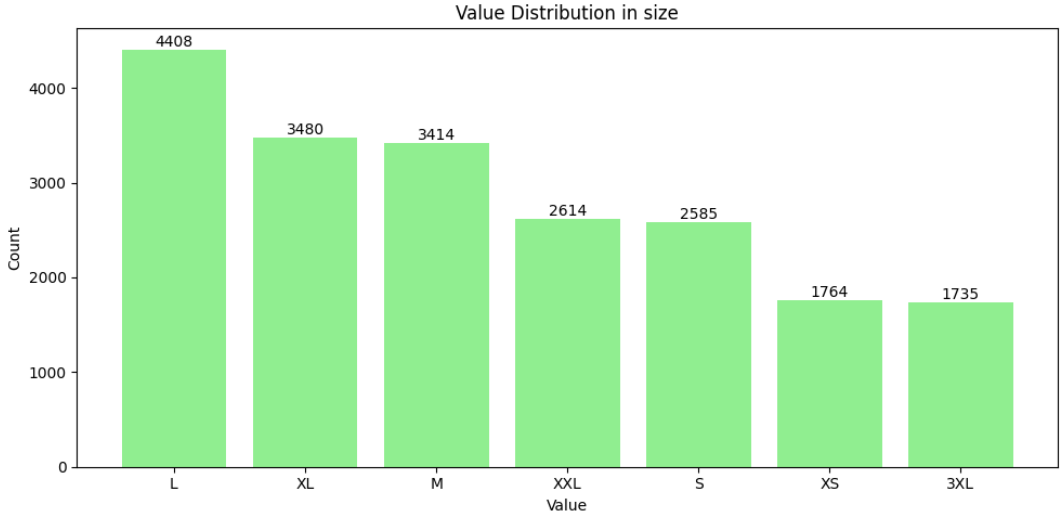
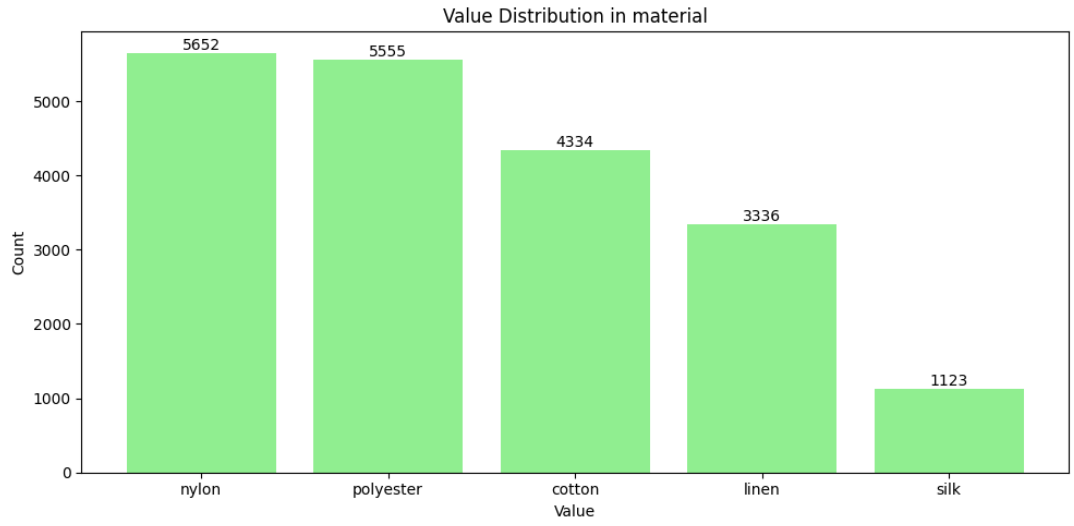
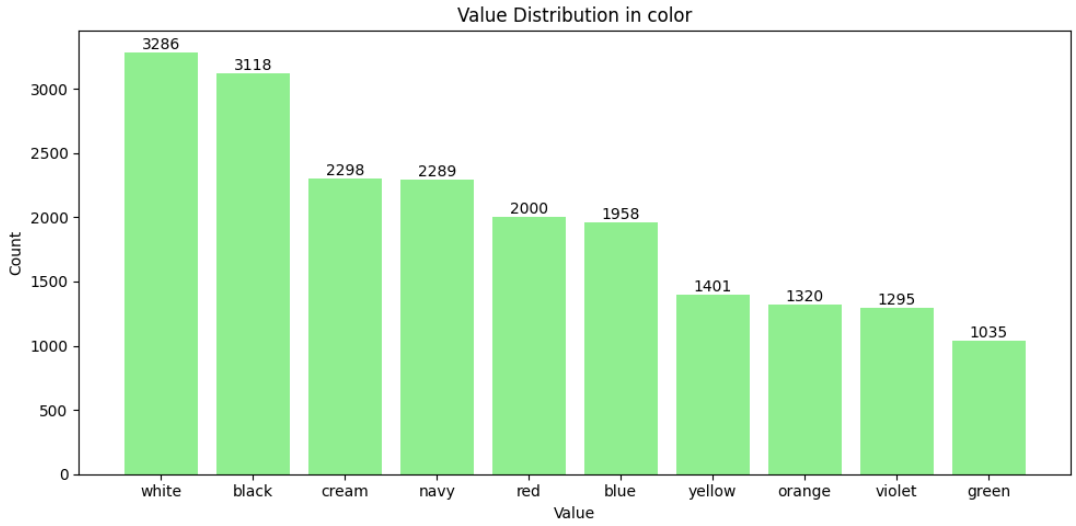
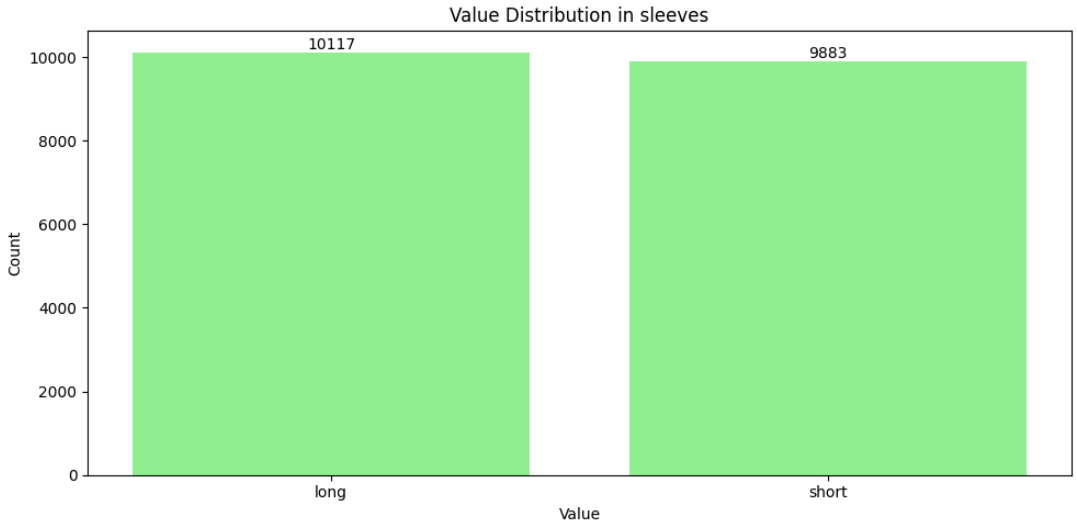
1. Eksploracja danych

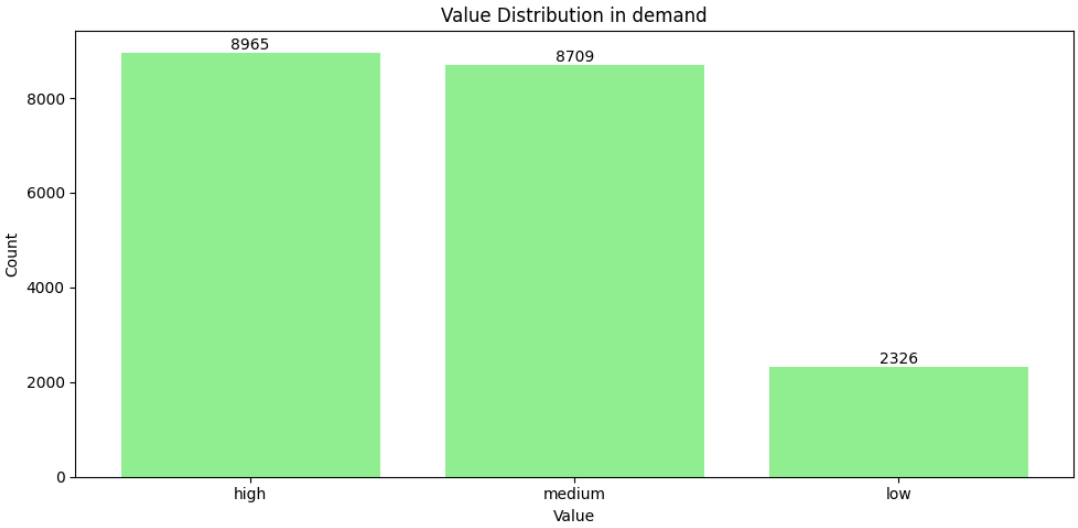
Rozkład wartości



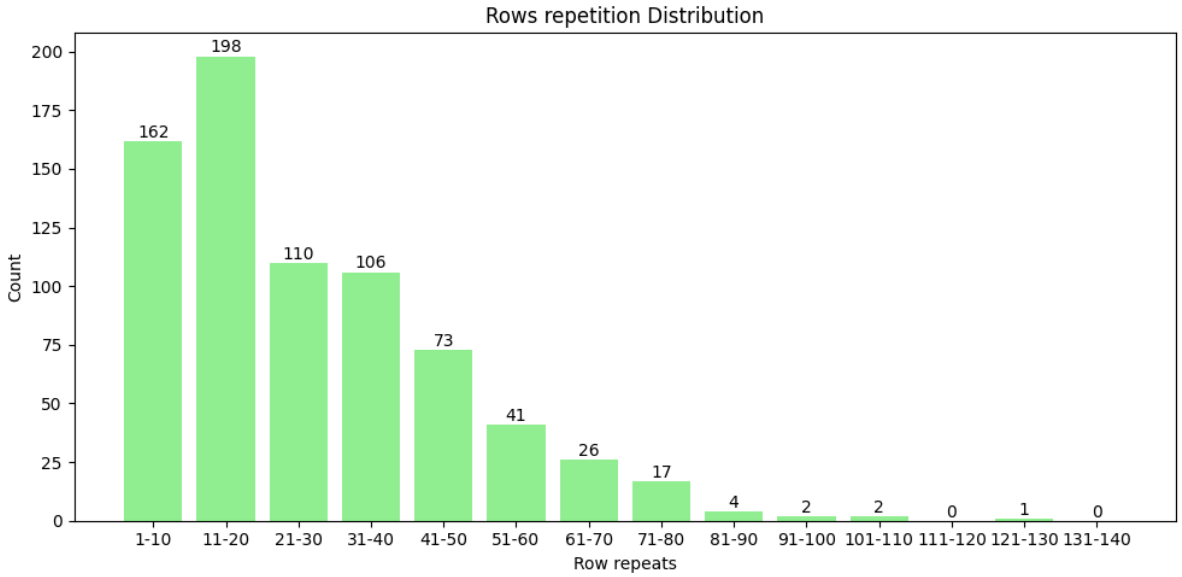




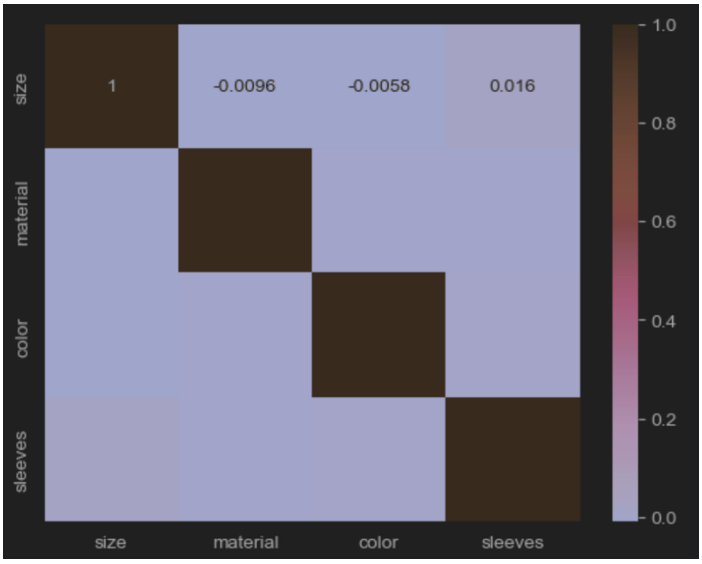




Powtarzające się wiersze



Macierz korelacji



Uwagi dotyczące zbioru danych

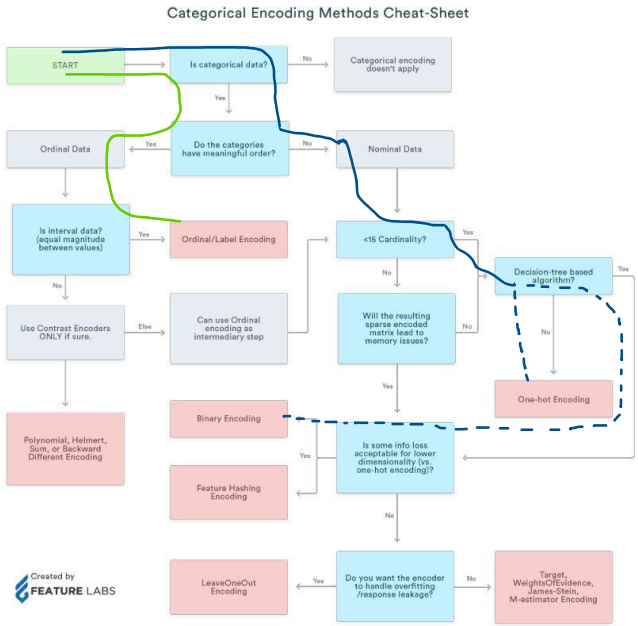
- Dane są poprawne
  - brak niezupełnych rekordów
  - brak niepoprawnie wprowadzonych nazw (wartości)
  - wiersz nagłówkowy
- Rozkład danych jest różnorodny
- Wiele wierszy się powtarza
- **Wyłącznie dane kategoryczne (2 kolumny porządkowe, 2 nominalne)**
- **Kolumna etykiet (demand – low, medium, high)**
- Rozkład etykiet nie jest zbalansowany – znacznie mniej demand=low
- **Cechy nie są skorelowane – niewymagana selekcja cech**

2. Przygotowanie danych

size	material	color	sleeves	demand
S	nylon	white	long	medium
XL	polyester	cream	short	high
S	silk	blue	short	medium
M	cotton	black	short	medium

Podgląd danych

Dane kategoryczne -> numeryczne



https://www.kdnuggets.com/2021/05/deal-with-categorical-data-machine-learning.html

Dane porządkowe:

size	**sleeves**
0 – XS	0 – short
1 – S	1 – long
…	
6 – 3XL	

size	material	color	sleeves	demand
1	nylon	white	1	medium
4	polyester	cream	0	high
1	silk	blue	0	medium
2	cotton	black	0	medium

Ordinal Encoding

Dane nominalne (*material, color*):

	material_cotton	material_linen	material_nylon	material_polyester	material_silk
0	False	False	True	False	False
1	False	False	False	True	False
2	False	False	False	False	True
3	True	False	False	False	False

One Hot Encoding

Ostatecznie 18 kolumn:

- size
- sleeves
- material x 5
- color x 10
- demand

Podział na zbiór treningowy/testowy

```
X = df.drop(['demand'], axis=1)
y = df['demand']
```
Executed at 2024.05.28 12:23:36 in 3ms

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```
Executed at 2024.05.28 12:23:40 in 206ms

Normalizacja

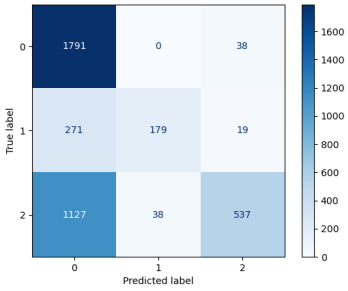
0	1	2	3	4	5	6	7	8	9	10	11	12
0.500000	0.500000	0.000000	0.000000	0.500000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00
0.942809	0.000000	0.000000	0.000000	0.000000	0.235702	0.00000	0.000000	0.000000	0.235702	0.000000	0.000000	0.00
0.577350	0.000000	0.000000	0.000000	0.000000	0.000000	0.57735	0.000000	0.577350	0.000000	0.000000	0.000000	0.00
0.816497	0.000000	0.408248	0.000000	0.000000	0.000000	0.00000	0.408248	0.000000	0.000000	0.000000	0.000000	0.00
0.917663	0.229416	0.000000	0.000000	0.000000	0.229416	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.22

PCA

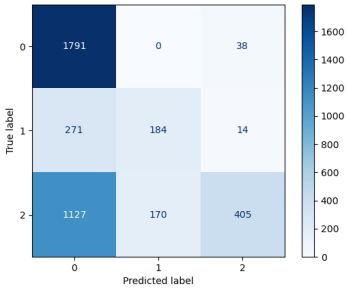
	0	1	2	3	4	5	6	7	8	9	10	
0	2.000403	0.726973	-0.468815	-0.369243	-0.105877	0.791994	0.413158	-0.000296	-0.061324	-0.004380	-0.063410	
1	-0.992858	-0.697322	0.503819	-0.383502	-0.144909	-0.060208	-0.388289	-0.726619	-0.456330	-0.062721	-0.153721	
2	2.004616	-0.026895	0.495441	0.097837	0.099931	-0.066958	-0.234689	0.037397	0.436093	0.769139	-0.316638	
3	1.004463	-0.033628	0.473888	0.804143	-0.378696	-0.633894	0.647671	0.009328	-0.072693	-0.007788	-0.066027	-(
4	-0.999216	-0.683171	-0.488959	-0.423377	-0.133428	-0.017667	-0.133393	-0.015690	0.082359	0.013225	0.448197	-(
5	3.006031	-0.700640	0.480334	-0.414750	-0.131398	-0.621643	0.630344	-0.015716	-0.082809	-0.024598	-0.071135	-(

3. Klasyfikacja + ocena

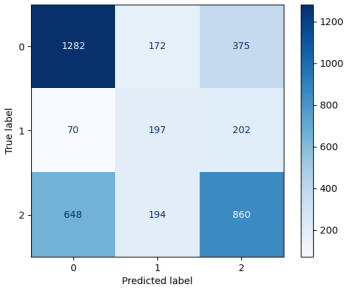
Naiwny Bayes



Brak normalizacji/PCA



Normalizacja

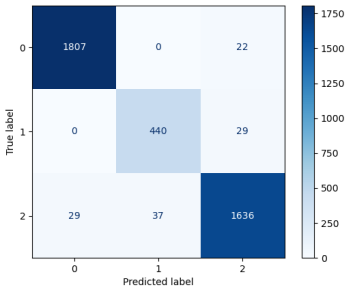


PCA

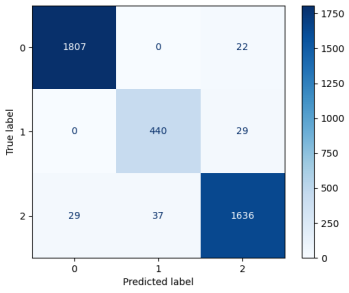
Avg()	Brak normalizacji / PCA	Normalizacja	PCA
Accuracy	0,63	0,59	0,58
Recall	0,56	0,54	0,54
Precision	0,76	0,66	0,53
F1-score	0,57	0,51	0,53

- Najlepsze metryki bez dodatkowego preprocessingu
- Najwyższa metryka – precyzja – zaledwie 76% prawidłowych pozytywnie sprognozowanych

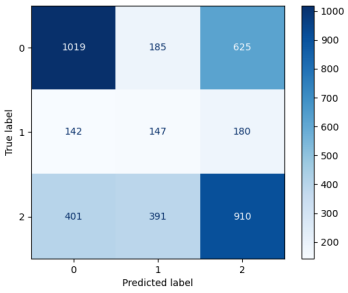
Drzewo decyzyjne



Brak normalizacji/PCA



Normalizacja

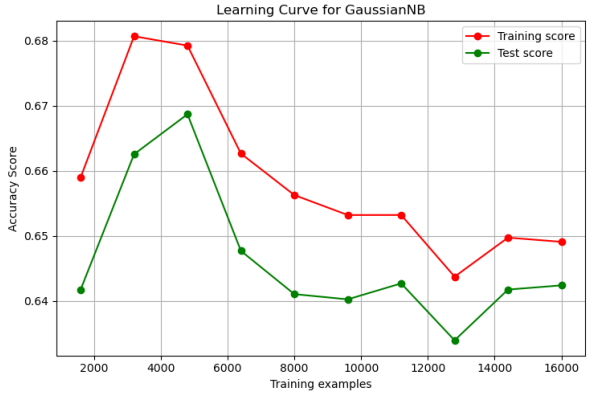


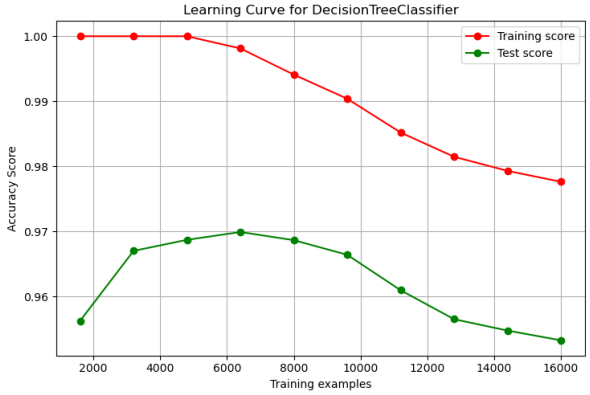
PCA

Avg()	Brak normalizacji / PCA	Normalizacja	PCA
Accuracy	0,97	0,97	0,54
Recall	0,96	0,96	0,47
Precision	0,96	0,96	0,48
F1-score	0,96	0,96	0,4t

- Najlepsze metryki bez dodatkowego preprocessingu
- Wszystkie metryki blisko 100%
- Znaczna przewaga nad Bayesem

Krzywa uczenia





- Naiwny Bayes lepiej generalizuje, natomiast Drzewo Decyzyjne stosuje overfitting
- Drzewo Decyzyjne uczy się bardziej stabilnie

4. Hiperparametry

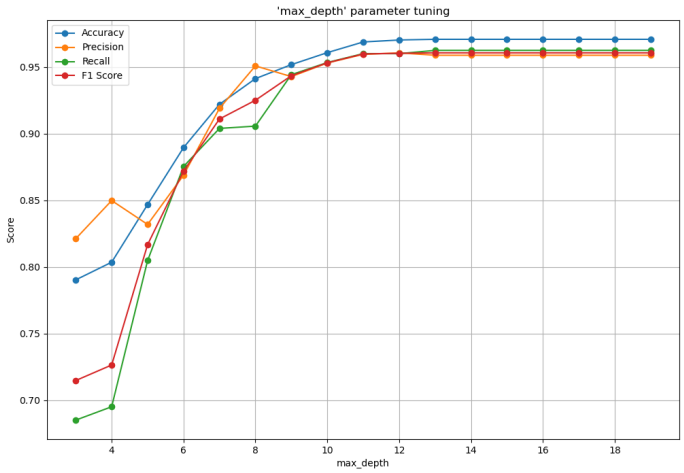
Wybrane parametry w Drzewie Decyzyjnym:

- *max_depth* – maksymalna głębokość drzewa
  min_samples_leaf – minimalna liczba rekordów znajdujących się w węźle drzewa
- *max_leaf_nodes* – maksymalna liczba liści

Domyślny zestaw hiperparametrów

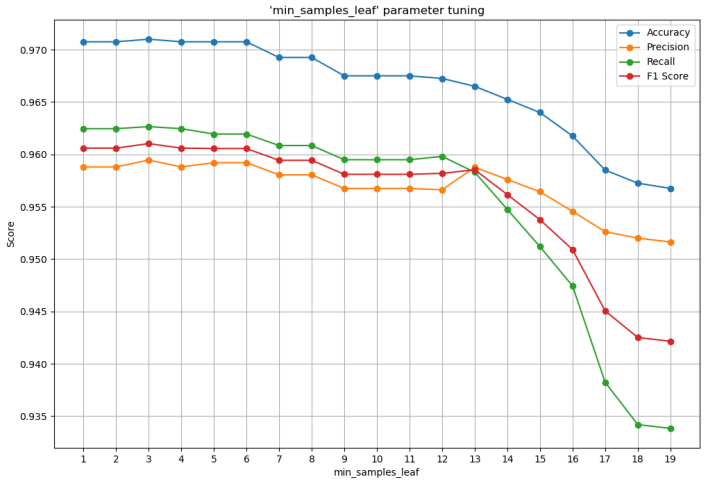
- max_depth = None
- min_samples_leaf = 1
- max_leaf_nodes = None

max_depth



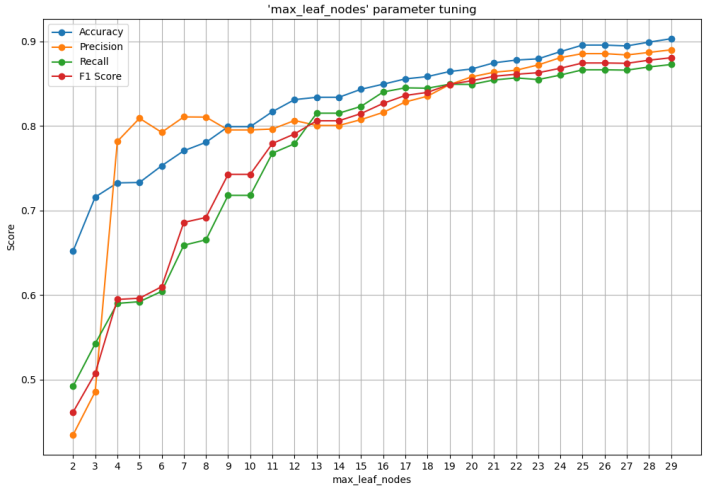
Brak zmian dla max_depth

min_samples_leaf



min_samples_leaf = 3

max_leaf_nodes



max_leaf_nodes = None

Końcowy zestaw hiperparametrów

- max_depth = None
- min_samples_leaf = 3
- max_leaf_nodes = None