Uczenie Maszynowe I SI – sprawozdanie z zadania 3

Michał Świątczak

1. Wstęp

Celem ćwiczenia było wybranie zestawu danych klasyfikacyjnych, zastosowanie dwóch modeli uczenia maszynowego (w tym przypadku – drzewo decyzyjne oraz las losowy) o różnych stopniach złożoności oraz porównanie ich efektywności razem z analizą zestawu danych.

Dla przejrzystości oraz wygody uruchomienia kodu, zastosowano format Jupyter Notebook, ze środowiskiem przygotowanym szczególnie pod ML & Data Science w Anacondzie.

2. Przebieg analizy

Ze strony kaggle.com pobrano zestaw treningowy Mobile Price Classification (https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification), który opisuje telefony za pomocą 20 parametrów, oraz dodatkowa kolumna zawierająca klasyfikację cenową (0,1,2 i 3, gdzie 0 to najtańszy zakres). Zestaw danych zawiera 2000 próbek.

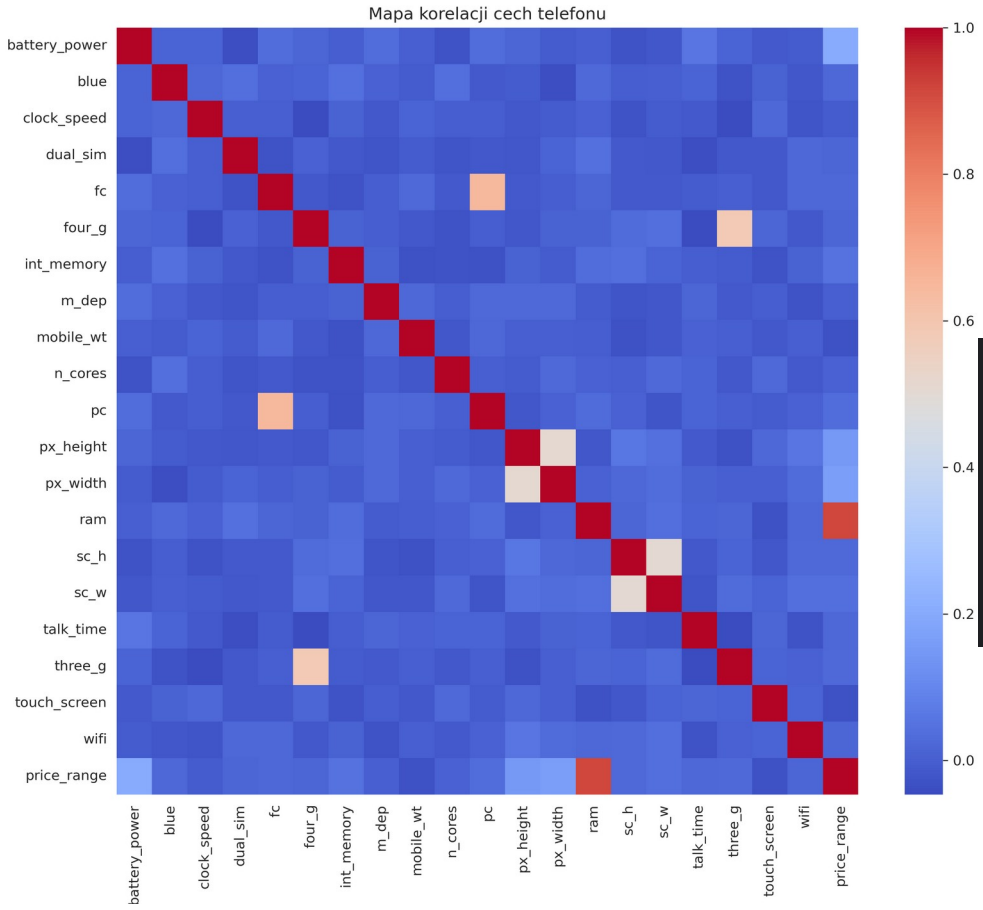
Pierwszym krokiem było pobranie oraz załadowanie zestawu do pythona za pomocą biblioteki pandas. Warto zauważyć, że plik .zip z zestawem, posiada dwa wpisy = test oraz train. Zestaw test.csv służy tylko i wyłącznie do walidacji oraz porównania modelu z innymi na Kaggle.

Przeprowadzono wstępne sprawdzenie czystości oraz ciągłości danych, za pomocą funkcji df.info().

Zestaw nie posiada żadnych null-values, więc nie ma potrzeby na modyfikację zestawu.

Sprawdzono również ilości wpisów dla konkretnych kategorii cenowych. Zestaw ten jest idealnie rozłożony, ponieważ każda z klas posiada 500 wpisów. Przed rozpoczeciem trenowania modelu, utworzono macierz korelacji parametrów telefonów, aby zobaczyć które parametry są najbardziej skorelowane z kategorią cenową. Następnie podzielono dane na testowe (20%, 400 elementów) oraz treningowe (80%, 1600 elementów). Do tych danych dopasowano dwa modele za pomoca funkcji model.fit(x_train,y_train) – **model drzewa deycyzjnego** oraz **model lasu losowego.** Porównano wartości przewidziane z wartościami testowymi, oraz wyznaczono dokładność dla obydwu modeli. Utworzono macierze pomyłek, oraz wykres waznosci cechy od ceny (na podstawie waznosci w lesie losowym, **las_model.feature_importances_**)

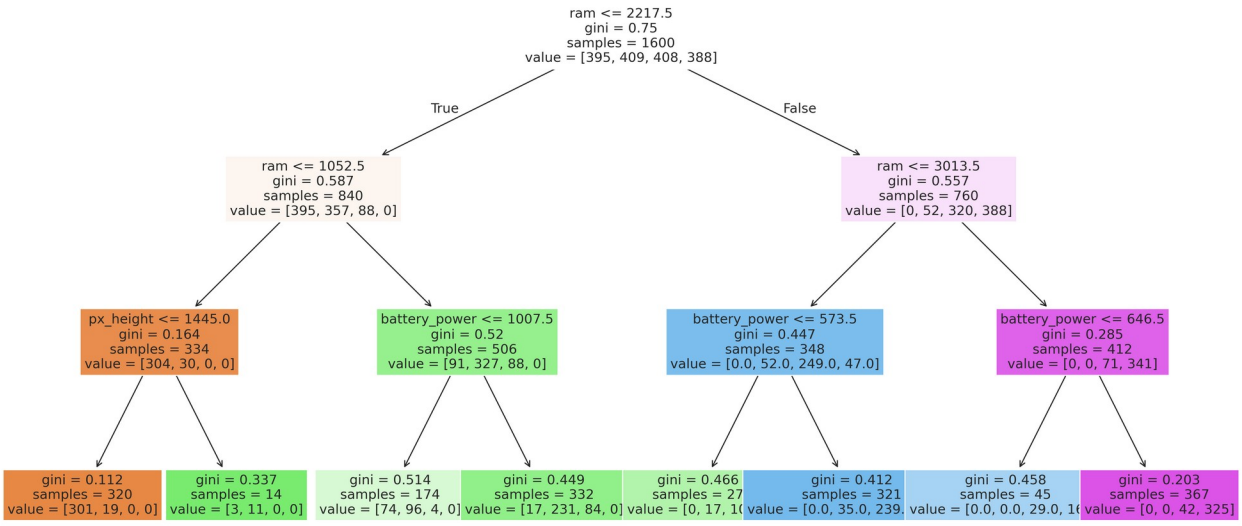
3. Wizualizacje i interpretacja



price_range	1.000000
ram	0.917046
battery_power	0.200723
px_width	0.165818
px_height	0.148858
int_memory	0.044435
sc_w	0.038711

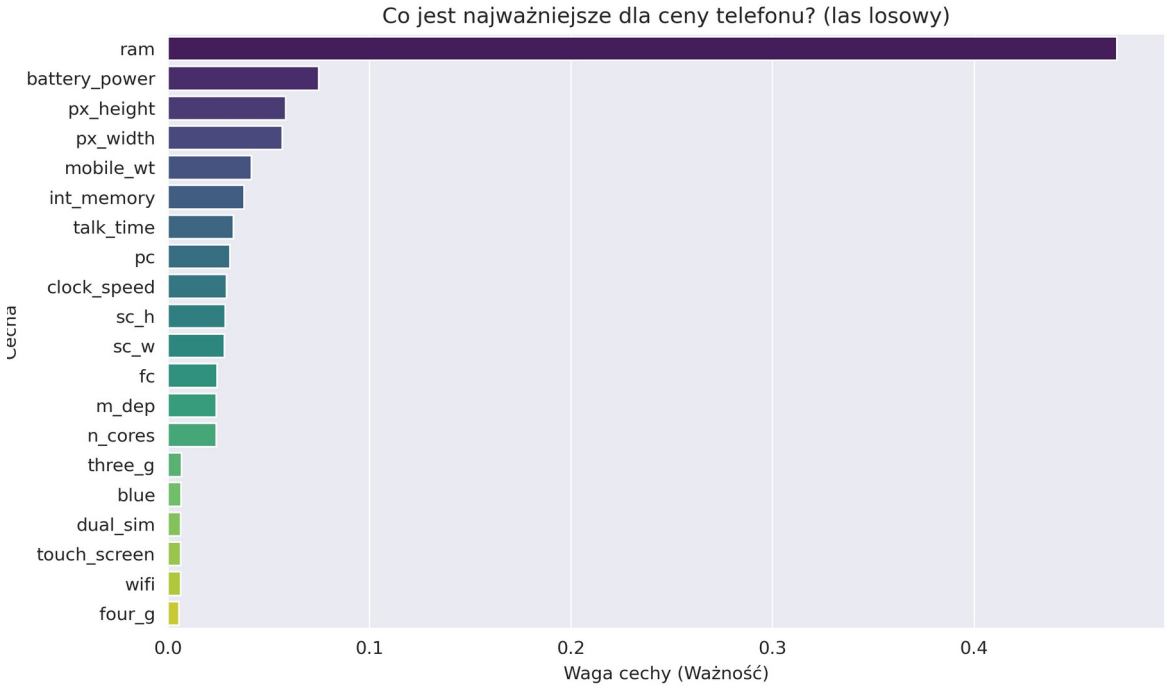
Na podstawie mapy korelacji można zauważyć, że najwiekszą korelację z ceną ma parametr **pamięci ram,** następnie **moc baterii** i **rozdzielczosc ekranu**.

a) Drzewo Decyzyjne (wizualizacja uproszczona, dla głębokości = 3)

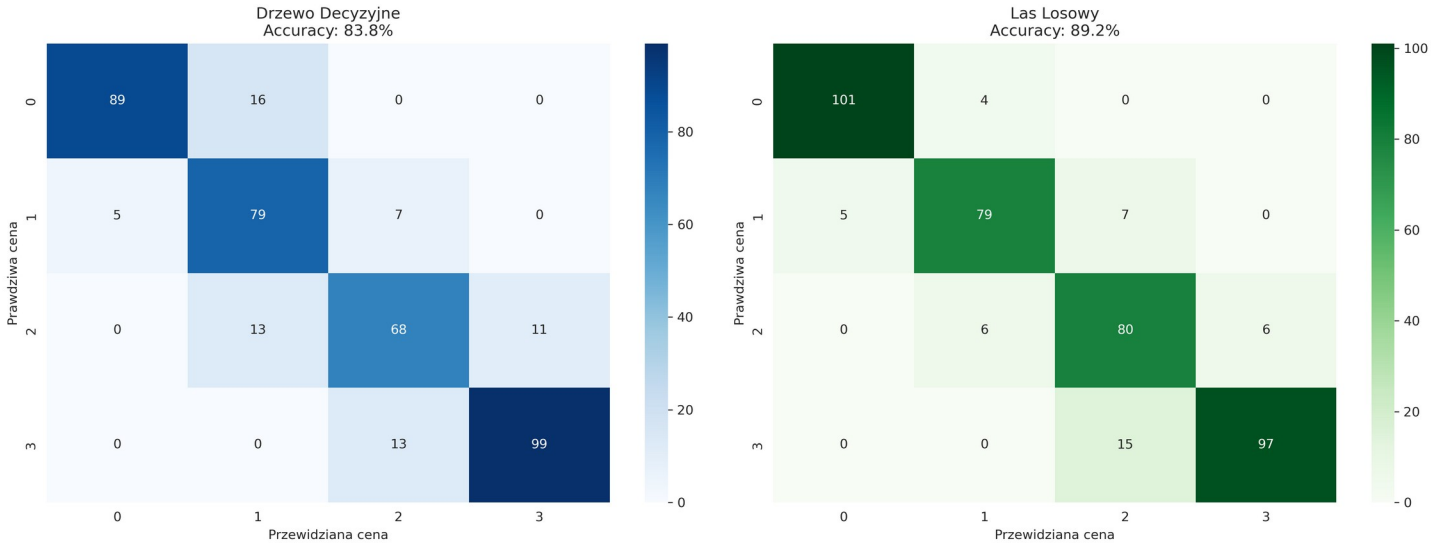


Patrząc na korzeń drzewa, widać, że model również uznał parametr pamięci ram za najważniejszy jeśli chodzi o przewidywanie kategorii cenowej. Zanieczyszczenie Giniego wynosi w korzeniu 0.75, co oznacza duży chaos w danych i zawarcie wszystkiego w jednym węźle. Drzewo dąży do minimalizacji tego parametru, aby uzyskać jednolity zbiór.

W przypadku lasu losowego, niemożliwe jest uzyskanie wykresu wszystkich drzew (n = 100). Utworzono dataframe zaleznosci cech od waznosci wyznaczonej wg modelu, oraz utworzono wykres. Model również słusznie zauważył, że największa korelacja jest pomiędzy pamięcią ram, a ceną.



Po walidacji modelów, utworzono również macierze pomyłek.



4. Wnioski oraz obserwacje

Głębokość	Dokładność
1	54.25
2	75
3	74.4
4	79.5
5	80.6
6	80.5
7	83.5
8	82.75
9	82.5
10	83.25
11	82
12	81.75
13	83.14
14	82.5
…	…

Na początku warto zauważyć, że w przypadku drzewa decyzyjnego, dokładność jest silnie powiązania z głębokością drzewa. Dla każdej głębokości n>14, dokładność osiąga ustaloną wartość 82,5%. Aby upewnić się, że model nie jest przetrenowany, przyjęto wartość n = 7 (acc=83,5%).

Ciekawe jest to, że w wygenerowanej macierzy pomyłek, wartość dokładności jest większa o 0,3 punktów procentowych.

Dokładność lasu losowego jest większa o około 6 punktów procentowych, co daje znaczną poprawę. Model lasu losowego jest diametralnie bardziej precyzyjny w przewidywaniu poprawnej ceny dla wszystkich kategorii, warto zauważyć że niepoprawne przewidywania z macierzy dla drzewa decyzyjnego 0,1 i 1,0 „przeskoczyły” w macierzy lasu do 0,0. Piętą achillesową obydwu modeli jest przewidywanie ceny dla rzeczywistej kategorii cenowej 3, pomyłki 2,3 są podobne w obydwu macierzach. Obydwa modele również dobrze przewidują rzeczywistą kategorię dla 1,1.

Nie sprawdzano zalezności dokładności lasu losowego od ilości drzew. Domyślne n=100.