Klasyfikacja cen telefonów komórkowych na podstawie cech technicznych

Weronika Kłujszo (223599), Michał Korzeniewski (223399),
Miłosz Malinowski (223391), Piotr Misiejuk (223302)

10 czerwca 2025

Spis treści

1 Streszczenie

Projekt dotyczy problemu klasyfikacji telefonów komórkowych do jednej z czterech klas cenowych na podstawie ich cech technicznych. Celem było zbudowanie modeli, które potrafią skutecznie przewidzieć przybliżoną cenę urządzenia bez znajomości jego rzeczywistej wartości rynkowej. W tym celu wykorzystano publiczny zbiór danych zawierający informacje o około 2000 telefonów, obejmujący zarówno zmienne ilościowe (np. pamięć RAM, rozdzielczość ekranu, liczba rdzeni procesora), jak i kategoryczne (np. obsługa 4G, ekran dotykowy).

W ramach projektu przeprowadzono analizę danych, ich wstępne przetworzenie i wizualizację, a następnie oceniono skuteczność wybranych algorytmów klasyfikacyjnych. Wyniki pokazują, że na podstawie odpowiednio dobranych parametrów technicznych możliwe jest z dużą dokładnością przewidywanie przynależności telefonu do określonego przedziału cenowego.

2 Wprowadzenie

Telefony komórkowe stanowią jedno z najpopularniejszych i najszybciej rozwijających się typów urządzeń codziennego użytku. Współczesny konsument ma do wyboru dziesiątki modeli różniących się parametrami technicznymi, jakością wykonania i – co istotne – ceną. Analiza tych różnic może pomóc w przewidywaniu, do jakiego przedziału cenowego należy konkretny model, co ma zastosowanie m.in. w marketingu, handlu internetowym czy planowaniu produkcji, ale również może być przydatne dla klientów próbujących określić jak poszczególne cechy telefonów wpływają na ich widełki cenowe. W niniejszym projekcie dokonano klasyfikacji telefonów komórkowych na cztery grupy cenowe (od najtańszych do najdroższych) na podstawie ich parametrów technicznych. Wykorzystano w tym celu publiczny zbiór danych Mobile Price Classification dostępny na platformie Kaggle, który zawiera dane o ponad 2000 modeli telefonów. Przeprowadzono analizę zmiennych, przygotowanie danych (w tym transformacje), a następnie ocenę skuteczności zastosowanych modeli klasyfikacyjnych.

3 Cel i zakres badania

Celem projektu jest podzielenie telefonów na cztery klasy, od najtańsztch do najdroższych, na podstawie ich parametrów technicznych takich jak posiadanie ekranu dotykowego, ilość pamięci wewnętrznej czy długość i szerokość, za pomocą metod klasyfikacji. Zakres badania obejmuje:

- wczytanie i przygotowanie danych,
- skalowanie danych przy użyciu StandardScaler oraz MinMaxScaler,
- budowę i ocenę modeli klasyfikacyjnych (Logistic Regression, LDA, Random Forest, model hybrydowy),
- podsumowanie wyników i zweryfikowanie ich względem rzeczywistych cen telefonów.

4 Przegląd literatury

Klasyfikacja cen produktów na podstawie cech technicznych to temat szeroko analizowany w literaturze związanej z uczeniem maszynowym. W szczególności, problem klasyfikacji telefonów komórkowych stanowi typowy przykład zadania klasyfikacji wieloklasowej z wykorzystaniem cech tabelarycznych.

W pracy [1] autorzy porównali różne algorytmy klasyfikacyjne w kontekście predykcji cen telefonów ze zbioru danych Kaggle Mobile Price Classification. Wykazano, że najlepsze wyniki osiągają modele zespołowe, takie jak Random Forest, które pozwalają uzyskać dokładność sięgającą 96%.

Podobne wnioski przedstawiono w [2], gdzie przeanalizowano skuteczność pięciu algorytmów (KNN, regresja logistyczna, drzewo decyzyjne, Random Forest oraz SVM) w klasyfikacji cen telefonów. W pracy tej podkreślono, że metody zespołowe przewyższają klasyfikatory proste, zwłaszcza w rozróżnianiu klas pośrednich.

Metody hybrydowe (ang. ensemble learning) są coraz częściej wykorzystywane w zadaniach klasyfikacji. W pracy [3] zaprezentowano skuteczność klasyfikatorów opartych na głosowaniu (ang. Voting Classifier) w przewidywaniu ryzyka finansowego. Wykazano, że połączenie kilku modeli bazowych, takich jak Random Forest, LDA czy XGBoost, prowadzi do większej stabilności i poprawy ogólnej skuteczności klasyfikacji.

Klasyczne metody, takie jak regresja logistyczna [4] czy analiza dyskryminacyjna LDA [5], wciąż pozostają wartościowym punktem odniesienia w analizie danych tabelarycznych ze względu na ich interpretowalność i prostotę.

Warto również odnotować, że techniki ensemble zostały szeroko opisane w literaturze, np. w przeglądowej pracy [6], gdzie podkreślono, że łączenie modeli prowadzi do zwiększenia dokładności i odporności na przeuczenie.

Podsumowując, dotychczasowe badania wskazują, że metody zespołowe, a w szczególności Random Forest i hybrydowe VotingClassifier, są obecnie jednymi z najbardziej efektywnych podejść w zadaniach klasyfikacji cen telefonów na podstawie cech technicznych.

5 Słowa kluczowe

POL: klasyfikacja cen telefonów, analiza danych, parametry techniczne, predykcja klasy, dokładność modelu

ENG: mobile price classification, data analysis, technical parameters, class prediction, model accuracy

6 Zmienne wybrane do analizy oraz ich wizualizacja

Zbiór danych zawiera 20 zmiennych, będących cechami telefonu komórkowego. Zmienne zostaly podzielone na ilościowe oraz kategoryczne.

- **Ilościowe:**
  - **battery_power** – pojemność baterii w mAh,
  - **clock_speed** – częstotliwość zegara procesora w GHz,
  - **fc** – rozdzielczość przedniego aparatu w MP,
  - **int_memory** – pamięć wewnętrzna w GB,
  - **m_dep** – głębokość telefonu w cm,
  - **mobile_wt** – masa telefonu w gramach,
  - **n_cores** – liczba rdzeni procesora,

  - **pc** – rozdzielczość tylnego aparatu w MP,
  - **px_height** – wysokość ekranu w pikselach,
  - **px_width** – szerokość ekranu w pikselach,
  - **ram** – pamięć RAM w MB,
  - **sc_h** – wysokość ekranu w cm,
  - **sc_w** – szerokość ekranu w cm,
  - **talk_time** – czas rozmów w godzinach.
- **Kategoryczne:**
  - **three_g** – obsługa 3G,
  - **four_g** – obsługa 4G,
  - **blue** – obsługa Bluetooth,
  - **dual_sim** – obsługa dwóch kart SIM,
  - **touch_screen** – ekran dotykowy,
  - **wifi** – obsługa WiFi.
- **Zmienna docelowa:**
  - **price_range** – zmienna kategoryczna przyjmująca wartości od 0 do 3, oznaczająca klasę cenową telefonu:
    * 0 – klasa **Low Cost** (najtańsze telefony),
    * 1 – klasa **Medium Cost**,
    * 2 – klasa **High Cost**,
    * 3 – klasa **Very High Cost** (najdroższe telefony).

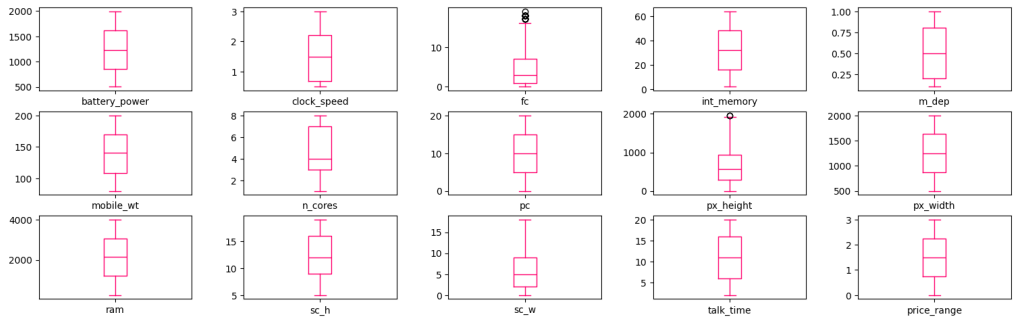
    Zmienna ta została wykorzystana jako etykieta klas w zadaniu klasyfikacyjnym. Jest zbalansowana — każda z klas zawiera około 500 obserwacji.

7 Wstępna analiza danych

7.1 Statystyki opisowe

W celu lepszego zrozumienia zbioru danych przeprowadzono analizę podstawowych statystyk opisowych. Dane obejmują zarówno zmienne ilościowe, jak

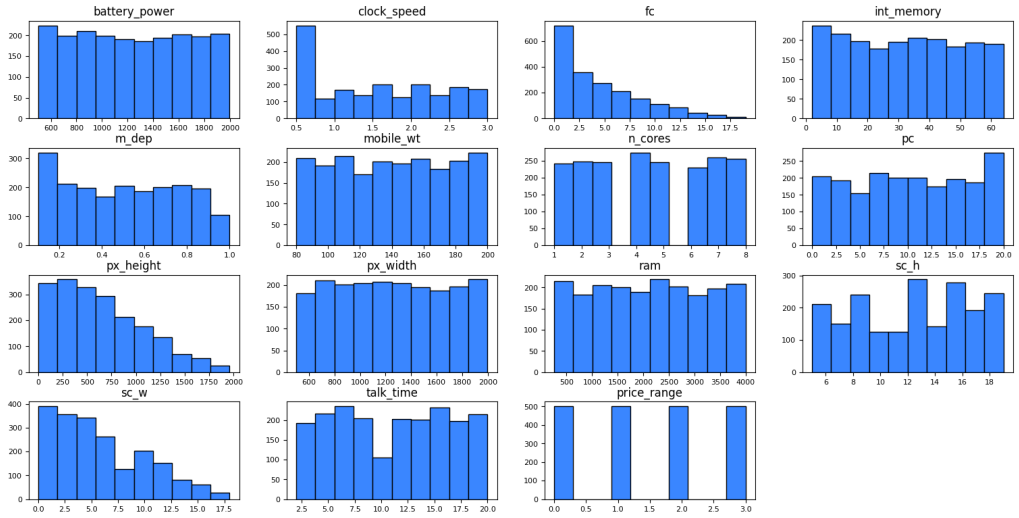
i kategoryczne. Poniżej zaprezentowano rozkłady i wykresy pudełkowe dla zmiennych ilościowych.



Rys. 1. Wykresy pudełkowe dla zmiennych ilościowych

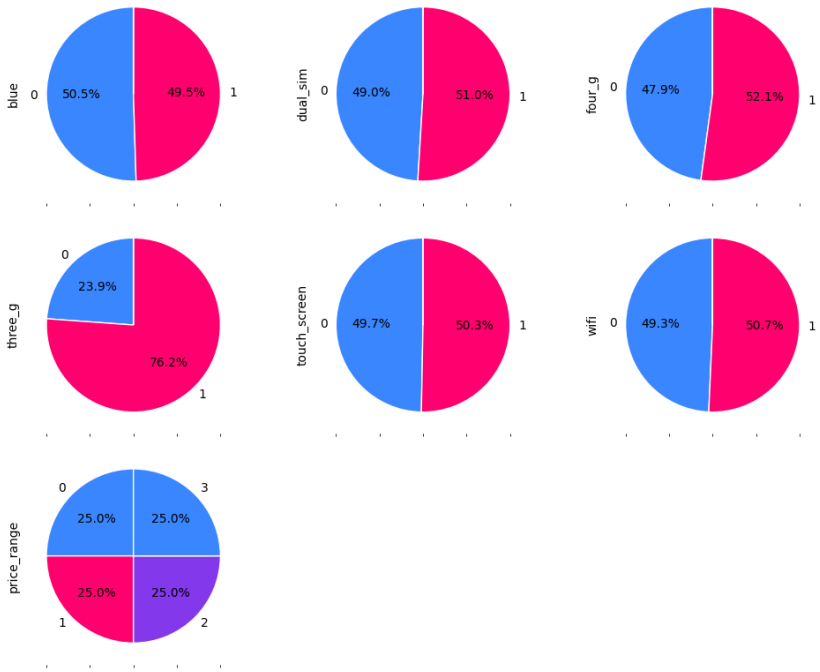
W większości zmiennych obserwujemy stosunkowo symetryczne rozkłady, choć dla niektórych, jak np. `fc` (przedni aparat), występują wartości odstające. Wartości takie mogą świadczyć o nielicznych modelach o bardzo wysokich parametrach w danej kategorii.

Na poniższych histogramach przedstawiono rozkłady zmiennych ilościowych. Widać, że wiele z nich (jak `px_height`, `fc`, `m_dep`) ma rozkłady skośne lub nieregularne, co może mieć wpływ na skuteczność wybranych algorytmów klasyfikacji.

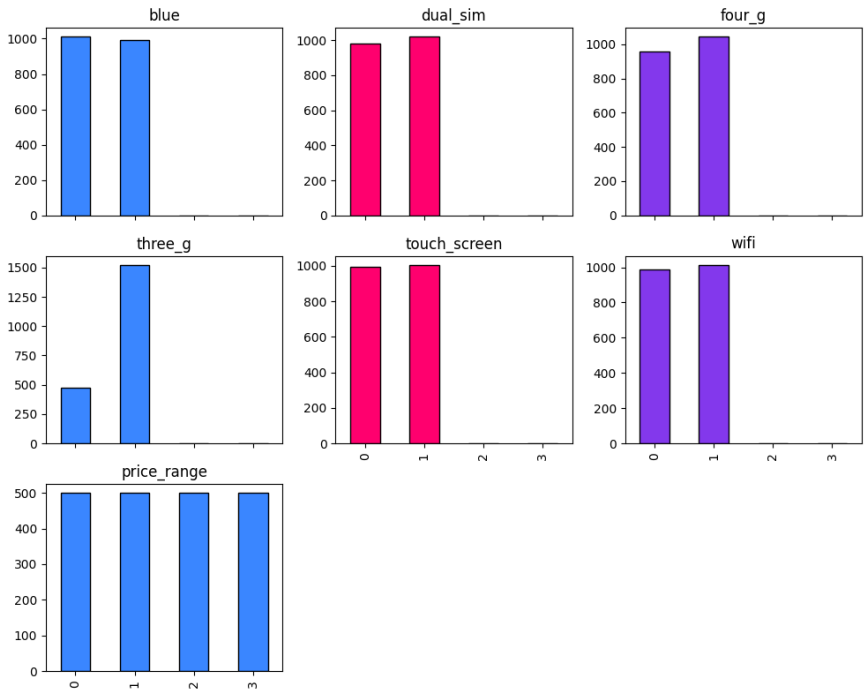


Rys. 2. Histogramy zmiennych ilościowych

W zbiorze występuje sześć zmiennych binarnych opisujących dostępność wybranych funkcji telefonu (np. Bluetooth, ekran dotykowy). Na poniższych wykresach kołowych i słupkowych przedstawiono ich proporcje.



Rys. 3. Udział klas dla zmiennych binarnych (wykresy kołowe)



Rys. 4. Liczność klas dla zmiennych binarnych (wykresy słupkowe)

Zmienna `three_g` dominuje pod względem udziału klasy pozytywnej (76% telefonów obsługuje 3G), co może być istotnym predyktorem w klasyfikacji cen. Pozostałe cechy binarne mają rozkłady bardziej zrównoważone.

Aby uzyskać ogólny obraz danych, obliczono podstawowe statystyki opisowe dla zmiennych ciągłych. Dla każdej zmiennej obliczono:

- średnią,
- medianę,
- maksimum,
- minimum,
- odchylenie standardowe,
- skośność.

Zmienna	Średnia	Mediana	Minimum	Maksimum	Odchylenie Standardowe	Skośność
battery_power	1238.52	1226.00	501.00	1998.00	439.42	0.03
clock_speed	1.52	1.50	0.50	3.00	0.82	0.18
fc	4.31	3.00	0.00	19.00	4.34	1.02
int_memory	32.05	32.00	2.00	64.00	18.15	0.06
m_dep	0.50	0.50	0.10	1.00	0.29	0.09
mobile_wt	140.25	141.00	80.00	200.00	35.40	0.01
n_cores	4.52	4.00	1.00	8.00	2.29	0.00
pc	9.92	10.00	0.00	20.00	6.06	0.02
px_height	645.11	564.00	0.00	1960.00	443.78	0.67
px_width	1251.52	1247.00	500.00	1998.00	432.20	0.01
ram	2124.21	2146.50	256.00	3998.00	1084.73	0.01
sc_h	12.31	12.00	5.00	19.00	4.21	-0.10
sc_w	5.77	5.00	0.00	18.00	4.36	0.63
talk_time	11.01	11.00	2.00	20.00	5.46	0.01

Tabela 1: Statystyki opisowe dla zmiennych ilościowych

7.2 Brakujące oraz odstające dane

W analizowanym zbiorze danych nie stwierdzono brakujących wartości.

Wartości odstające zostały wyznaczone zgodnie z regułą 1.5 IQR, gdzie wartości poniżej Q1-1.5*IQR bądź powyżej Q3+1.5*IQR zostały uznane za odstające.

Oznaczenia:

Q1 - pierwszy kwartyl,

Q3 - trzeci kwartyl,

IQR - odstęp ćwiartkowy.

W przypadku danych ilościowych, w kolumnach, gdzie wartości odstające stanowiły mniej niż 5% wszystkich danych, zostały one usunięte i nie były brane pod uwagę w dalszej analizie. Przypadki, gdzie wartości odstające stanowiły więcej niż 5% wszystkich danych nie wystąpiły.

Zmienna	**Ilość wartości skrajnych**	**Procent wartości skrajnych**
battery_power	0.00	0.00
blue	0.00	0.00
clock_speed	0.00	0.00
dual_sim	0.00	0.00
fc	18.00	0.01
four_g	0.00	0.00
int_memory	0.00	0.00
m_dep	0.00	0.00
mobile_wt	0.00	0.00
n_cores	0.00	0.00
pc	0.00	0.00
px_height	2.00	0.00
px_width	0.00	0.00
ram	0.00	0.00
sc_h	0.00	0.00
sc_w	0.00	0.00
talk_time	0.00	0.00
three_g	0.00	0.00
touch_screen	0.00	0.00
wifi	0.00	0.00
price_range	0.00	0.00

Tabela 2: Statystyki wartości odstających

7.3 Skalowanie danych

W celu zapewnienia spójności w zakresie wartości cech ilościowych przeprowadzono ich skalowanie, co umożliwia prawidłowe działanie wielu algorytmów klasyfikacyjnych. Przeskalowanie zmiennych redukuje wpływ różnych jednostek miar i ułatwia porównywalność.

7.3.1 Standaryzacja

Standaryzacja (inaczej normalizacja z wykorzystaniem średniej i odchylenia standardowego) polega na przekształceniu danych w taki sposób, że każda cecha ma średnią wartość 0 oraz odchylenie standardowe 1. Dzięki temu zmienne z różnych zakresów wartości stają się porównywalne i nie mają wpływu na wyniki analiz statystycznych.

Matematycznie:

$$x' = \frac{x - \mu}{\sigma}$$

gdzie μ to średnia, a σ odchylenie standardowe.

7.3.2 Normalizacja Min-Max

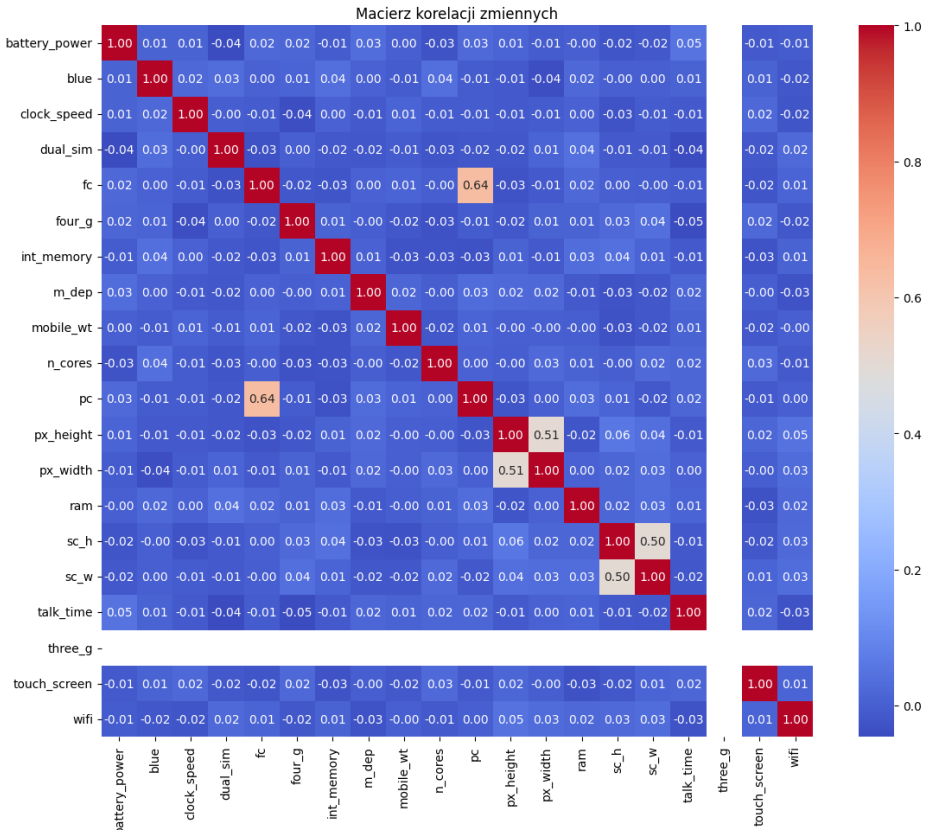
Alternatywnie stosuje się skalowanie Min-Max:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

które przekształca dane do przedziału [0, 1], zachowując proporcje między wartościami.

7.3.3 Macierz korelacji zmiennych

W celu sprawdzenia współzależności między zmiennymi objaśniającymi obliczono macierz korelacji Pearsona. Macierz pozwala zidentyfikować silnie skorelowane cechy, które mogą wprowadzać redundancję do modelu.



Rys. 5. Macierz korelacji zmiennych ilościowych

Jak widać, większość zmiennych nie wykazuje silnej korelacji. Najwyższe zależności zaobserwowano między:

- `px_width` a `px_height` – korelacja $\approx 0{,}51$,
- `fc` a `pc` – korelacja $\approx 0{,}64$,
- `sc_h` a `sc_w` – korelacja $\approx 0{,}50$.

Ponieważ wartości te nie przekraczają 0,8, nie zdecydowano się na eliminację żadnej z cech na podstawie korelacji.

8 Omówienie metod klasyfikacji

- **Regresja logistyczna** – klasyczny liniowy model klasyfikacyjny,

- **LDA (Linear Discriminant Analysis)** – metoda liniowa oparta na analizie wariancji klas,
- **Random Forest** – metoda klasyfikacji bazująca na zespole drzew decyzyjnych,
- **Model hybrydowy (ensemble)** – połączenie kilku modeli bazowych w celu uzyskania lepszej dokładności, np. poprzez głosowanie większościowe.

8.1 Regresja logistyczna

Regresja logistyczna to klasyczny model liniowy używany do klasyfikacji, który modeluje logarytmiczne prawdopodobieństwo przynależności do danej klasy. W przypadku klasyfikacji wieloklasowej używa się funkcji softmax:

$$P(y = k \mid x) = \frac{e^{\beta_k^T x}}{\sum_{j=1}^{K} e^{\beta_j^T x}} \tag{1}$$

gdzie:

- x – wektor cech,
- β_k – wektor współczynników regresji dla klasy k,
- K – liczba klas.

Regresja logistyczna została po raz pierwszy wprowadzona przez Davida Coxa w 1958 roku jako metoda modelowania zależności pomiędzy zmiennymi objaśniającymi a prawdopodobieństwem zajścia danego zdarzenia [7]. Metoda ta jest szeroko stosowana w analizie danych i statystyce, zwłaszcza w badaniach medycznych, marketingu i klasyfikacji wieloklasowej [4].

Model regresji logistycznej został użyty jako punkt odniesienia dla bardziej złożonych algorytmów. Z uwagi na swoją prostotę i interpretowalność, dobrze sprawdza się przy liniowo separowalnych danych. Model został dopasowany do danych po uprzednim skalowaniu cech ilościowych.

Model został wytrenowany przy użyciu solvera `liblinear`, który dobrze sprawdza się przy małych i średnich zbiorach danych. Zastosowano również standaryzację cech w celu zapewnienia porównywalnej skali zmiennych wejściowych.

Ewaluacja modelu obejmowała:

- walidację krzyżową z 5 podziałami,
- Leave-One-Out Cross Validation (LOOCV),
- metryki: dokładność, f1-score (macro), ROC AUC (One-vs-Rest),
- analizę ważności cech przy użyciu wykresu SHAP.

Wyniki dla standrayzacji:

- Dokładność (test set): 77.21%,
- F1-score (macro): 0,7499,
- ROC AUC (One-vs-Rest): 0,9142,
- Walidacja krzyżowa (5-krotna): średnia 76.96% $\pm$ 2.10%,
- Walidacja Leave-One-Out: średnia dokładność 78.29%.

Wyniki dla MinMax:

- Dokładność (test set): **76.64%**,
- F1-score (macro): **0,7428**,
- ROC AUC (One-vs-Rest): **0,9128**,
- Walidacja krzyżowa (5-krotna): średnia **76.64% $\pm$ 2.07%**,
- Walidacja Leave-One-Out: średnia dokładność **77.57%**.

Zastosowanie skalowania MinMax skutkowało bardzo podobnymi wynikami co przy standaryzacji – dokładność oraz F1-score różniły się o mniej niż 1%. Oznacza to, że model regresji logistycznej jest stabilny względem rodzaju skalowania, co wynika z jego liniowej natury.

Model poradził sobie dobrze, szczególnie przy rozpoznawaniu klas skrajnych. Trudności pojawiały się przy rozróżnianiu klas środkowych (średnia cena, wysoka cena), co wynikać może z ich podobieństw cechowych.

8.2 Analiza dyskryminacyjna liniowa (LDA)

LDA zakłada normalność rozkładu danych w obrębie klas oraz równość macierzy kowariancji. Model ten sprawdza się przy małej liczbie obserwacji i prostych granicach decyzyjnych.

Funkcja dyskryminacyjna dla klasy k ma postać:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k \tag{2}$$

gdzie:

- x – wektor cech,
- μ_k – średni wektor klasy k,
- Σ – wspólna macierz kowariancji,
- π_k – aprioryczne prawdopodobieństwo klasy k.

Obiekt x przypisywany jest do klasy o największej wartości $\delta_k(x)$.

Analiza dyskryminacyjna liniowa (LDA) została wprowadzona przez Ronalda Fishera w 1936 roku jako metoda do rozróżniania klas na podstawie zmiennych ilościowych [5]. LDA jest klasycznym narzędziem statystycznym, powszechnie stosowanym w biomedycynie, systemach rozpoznawania obrazów i klasyfikacji tekstów [10].

Model LDA został wykorzystany do klasyfikacji danych ze względu na jego zdolność do modelowania zależności między klasami przy założeniu normalności danych oraz równości macierzy kowariancji.

W analizie nie przyjęto dodatkowego uproszczenia cech jako niezależnych — zastosowano pełne (zwykłe) LDA, a nie wersję naiwną (z diagonalną macierzą kowariancji). Pomimo umiarkowanych współzależności między kilkoma parami zmiennych (korelacje ≈ 0.5 – 0.64), uznano, że są one akceptowalne w kontekście założeń modelu.

W celu oceny skuteczności modelu zastosowano dwie metody walidacji:

- **Stratyfikowana walidacja krzyżowa (Stratified K-Fold)** z 5 podziałami, pozwalająca na stabilną ocenę dokładności klasyfikatora przy zachowaniu proporcji klas w każdym foldzie.
- **Leave-One-Out (LOO)** – ekstremalna wersja walidacji krzyżowej, gdzie każdy przykład jest używany jako zbiór testowy dokładnie raz.

Wyniki dla standrayzacji:

- Dokładność (test set): 94.91%,
- F1-score (macro): 0,9481,
- ROC AUC (One-vs-Rest): 0,9958,
- Walidacja krzyżowa (5-krotna): 94.29% ± 0.77%,
- Walidacja Leave-One-Out: 94.29%.

Wyniki dla MinMax:

- Dokładność (test set): **94.23%**,
- F1-score (macro): **0,9423**,
- ROC AUC (One-vs-Rest): **0,9956**,
- Walidacja krzyżowa (5-krotna): średnia **94.23% ± 0.74%**,
- Walidacja Leave-One-Out: średnia dokładność **94.57%**.

Wyniki uzyskane po MinMaxScalingu były niemal identyczne jak przy standaryzacji, z dokładnością oraz F1-score na poziomie 94%. Wskazuje to, że LDA – jako metoda oparta na kowariancji – również dobrze radzi sobie przy obu podejściach do skalowania danych.

Model osiągnął średnią dokładność na poziomie około 79% w walidacji krzyżowej oraz porównywalne wyniki w LOO. Na podstawie macierzy pomyłek i wskaźników takich jak F1-score oraz ROC AUC można stwierdzić, że LDA najlepiej klasyfikuje klasy skrajne (0 i 3), natomiast częściej myli klasy środkowe.

8.3 Random Forest

Random Forest to popularna metoda zespołowa (ang. ensemble learning), która wykorzystuje wiele drzew decyzyjnych w celu poprawy stabilności i dokładności predykcji. Każde drzewo jest trenowane na innym losowym podzbiorze danych (z tzw. bootstrapowaniem), a w trakcie jego budowy wybierana jest losowa podgrupa cech przy każdym rozgałęzieniu (ang. feature bagging).

- Każde drzewo głosuje niezależnie na jedną z klas,
- Ostateczna decyzja modelu to klasa wybrana przez większość drzew (głosowanie większościowe).

Zaletą lasu losowego jest jego odporność na nadmierne dopasowanie (overfitting) oraz możliwość uchwycenia nieliniowych zależności między zmiennymi. Dodatkowo metoda ta umożliwia ocenę znaczenia poszczególnych cech wejściowych (np. za pomocą wskaźnika Gini lub metryk opartych na głębokości podziału w drzewach).

Metoda Random Forest została zaproponowana przez Leo Breimana w 2001 roku [8] jako rozszerzenie klasycznych drzew decyzyjnych, oparte na technikach baggingu i losowego wyboru cech. Random Forest wykazuje bardzo dobrą odporność na nadmierne dopasowanie (overfitting) i znajduje szerokie zastosowanie w różnorodnych zadaniach klasyfikacji i regresji.

Model ten szczególnie dobrze radzi sobie w zadaniach klasyfikacji wieloklasowej, gdzie zależności pomiędzy cechami są złożone i trudne do uchwycenia liniowymi metodami.

$$\hat{y} = \text{mode}\left\{h_1(x), h_2(x), \ldots, h_T(x)\right\} \tag{3}$$

gdzie $h_t(x)$ to predykcja t-tego drzewa decyzyjnego, a T to całkowita liczba drzew w lesie.

W naszym przypadku zastosowano las losowy o 1000 drzewach i maksymalnej głębokości 5. Model został wytrenowany na 70% danych, a pozostałe 30% przeznaczono na testy.

Wyniki dla standaryzacji:

- Dokładność: 82.08%,
- F1-score (macro): 0,8126,
- ROC AUC (One-vs-Rest): 0,9613.

Wyniki dla MinMax

- Dokładność (test set): **82.69%**,
- F1-score (macro): **0,8191**,
- ROC AUC (One-vs-Rest): **0,9637**.

Dla Random Forest różnica między skalowaniem Standard a MinMax była minimalna – dokładność wzrosła nieznacznie z 82.08% do 82.69%, a F1-score również uległ niewielkiej poprawie. Ponieważ RF nie wymaga skalowania, różnice wynikają jedynie z drobnych efektów numerycznych lub przypadkowego podziału danych.

Model bardzo dobrze rozróżniał wszystkie klasy, zwłaszcza klasy 0 i 3. W analizie cech najważniejsze były: `ram`, `battery_power`, `px_width`, `px_height`. Najlepsze pojedyncze drzewo w lesie zostało wybrane na podstawie jego skuteczności na zbiorze testowym. Zostało ono zaprezentowane graficznie w formie drzewa decyzyjnego (Rys. 6).

8.4 Model hybrydowy z klasyfikatorem VotingClassifier

Model hybrydowy korzysta z metody głosowania większościowego wśród różnych klasyfikatorów bazowych. Ostateczna predykcja $\hat{y}$ dla przykładu x jest dana wzorem:

$$\hat{y} = \text{mode}\left\{f_1(x), f_2(x), \ldots, f_K(x)\right\} \tag{4}$$

gdzie $f_k(x)$ oznacza wynik k-tego klasyfikatora bazowego, a K to liczba modeli składowych.

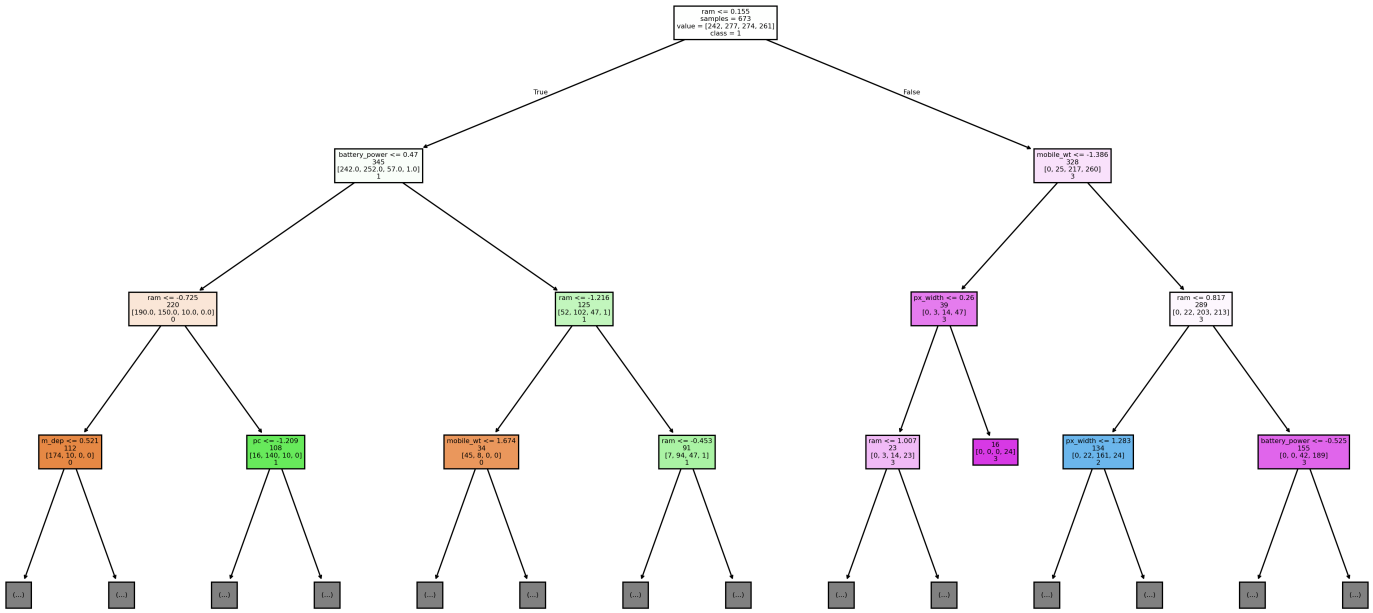
8.4.1 Wprowadzenie

Model hybrydowy wykorzystuje klasyfikator typu `VotingClassifier` z biblioteki `scikit-learn`, który łączy predykcje wielu różnych modeli bazowych, aby uzyskać bardziej stabilne i dokładne wyniki klasyfikacji. W tym przypadku zastosowano głosowanie miękkie (soft voting), które polega na obliczeniu średniej ważonej prawdopodobieństw klas przewidywanych przez poszczególne modele bazowe.

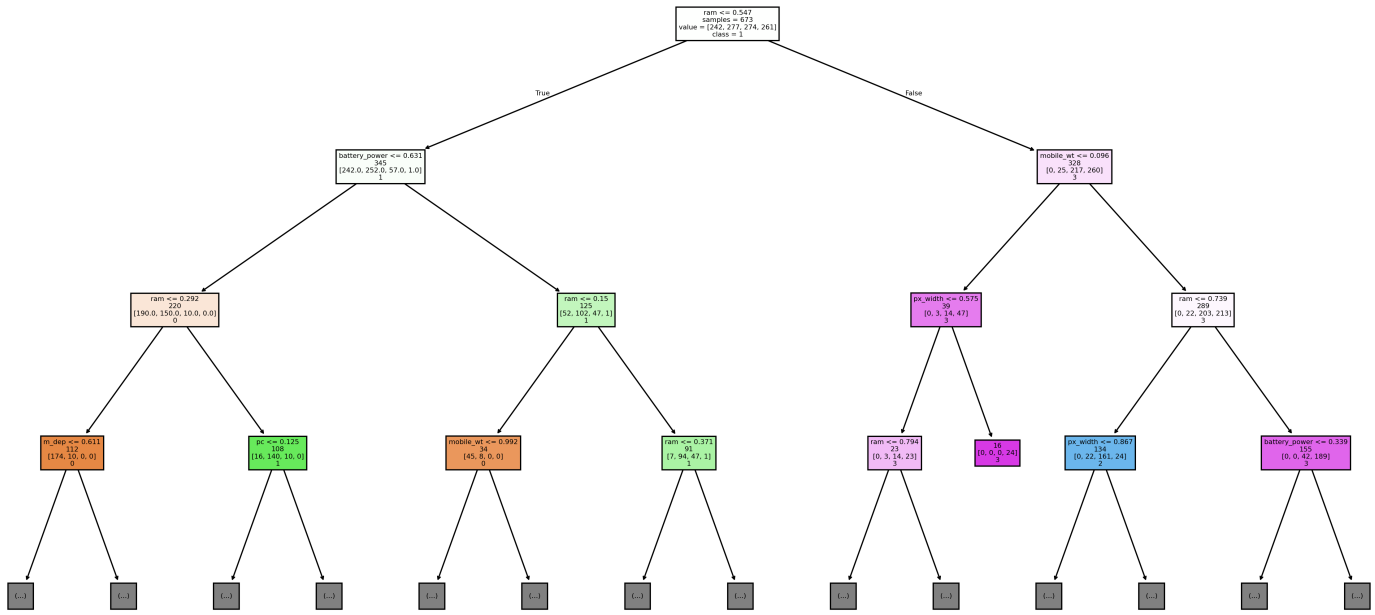
Modele hybrydowe oparte na głosowaniu wielu klasyfikatorów (ang. ensemble voting) są znaną i skuteczną techniką wykorzystywaną w uczeniu maszynowym. VotingClassifier jest implementacją tej idei w bibliotece `scikit-learn` [9]. Liczne badania wykazują, że łączenie wyników różnych modeli poprawia stabilność i dokładność predykcji [3].

8.4.2 Skład modeli bazowych

- **RandomForestClassifier** – zespół drzew decyzyjnych,



Rys. 6. Przykładowe drzewo decyzyjne z modelu Random Forest, kolory odpowiadają poszczególnym klasom: pomarańczowy - 0, zielony - 1, niebieski - 2, fioletowy - 3



Rys. 7. Przykładowe drzewo decyzyjne z modelu Random Forest po skalowaniu MinMax, kolory odpowiadają poszczególnym klasom:
pomarańczowy - 0, zielony - 1, niebieski - 2, fioletowy - 3

- **KNeighborsClassifier** – metoda najbliższych sąsiadów,
- **LinearDiscriminantAnalysis** – klasyfikator dyskryminacyjny,
- **XGBClassifier** – boostingowy klasyfikator o wysokiej skuteczności.

8.4.3 Przetwarzanie danych wejściowych

Dane numeryczne są standaryzowane (`StandardScaler`), cechy kategoryczne pozostawiono bez zmian, a transformacje zastosowano za pomocą `ColumnTransformer`.

8.4.4 Podział danych i ocena

Dane podzielono z zachowaniem rozkładu klas (`stratify`). Do oceny wykorzystano 5-krotną walidację krzyżową (`StratifiedKFold`) oraz metryki:

- dokładność,
- F1-score,
- ROC AUC.

Wyniki dla standryzacji:

- Dokładność (test set): 91.15%,
- F1-score (macro): 0,9113,
- ROC AUC (One-vs-Rest): 0,9871,
- Walidacja krzyżowa (5-krotna): 91.83% $\pm$ 1.35%.

Wyniki dla MinMax:

- Dokładność (test set): **91.15%**,
- F1-score (macro): **0,9107**,
- ROC AUC (One-vs-Rest): **0,9873**,
- Walidacja krzyżowa (5-krotna): średnia **91.71% $\pm$ 1.35%**.

Dla modelu hybrydowego VotingClassifier wyniki były niemal identyczne niezależnie od zastosowanego skalowania – dokładność i ROC AUC różniły się o mniej niż 0.2 punktu procentowego. Oznacza to, że agregacja wyników z różnych modeli zapewnia odporność na drobne różnice wynikające z transformacji danych.

Model hybrydowy łączył zalety klasyfikatorów bazowych, osiągając wyniki bliskie Random Forest. Szczególnie dobrze radził sobie w rozróżnianiu klas środkowych, gdzie modele liniowe miały trudności.

9 Przykład użycia modeli na stworzonych sztucznie obserwacjach

W celu zilustrowania działania modeli, utworzono cztery przykładowe sztuczne obserwacje reprezentujące hipotetyczne telefony o określonych parametrach technicznych. Dla każdej z tych obserwacji przewidziano klasę cenową za pomocą każdego z modeli.

Przykład	Regresja Logistyczna	LDA	Random Forest	Model Hybrydowy
Telefon A	0	0	0	0
Telefon B	1	1	1	1
Telefon C	2	2	3	2
Telefon D	3	3	3	3

Tabela 3: Przykład predykcji modeli na sztucznych danych

Parametry techniczne dla sztucznych telefonów przedstawiono w tabeli poniżej.

Telefon	ram	battery_power	px_width	px_height	n_cores	fc	pc	int_memory	four_g	touch_screen	wifi	dual_sim	blue	three_g
Telefon A	512	900	720	1280	2	2	5	8	0	0	1	1	1	1
Telefon B	2048	1500	1080	1920	4	5	8	32	1	1	1	1	1	1
Telefon C	3072	1800	1440	2560	8	8	13	64	1	0	1	1	1	1
Telefon D	4096	2200	2160	3840	8	16	20	128	1	1	1	1	1	1

Tabela 4: Parametry techniczne stworzonych sztucznie telefonów

Opis przypadków:

- **Telefon A** — tani telefon budżetowy, niska pamięć RAM, mała bateria, brak 4G, brak ekranu dotykowego.

- **Telefon B** — telefon klasy średniej, umiarkowane parametry, pełne wsparcie 4G i funkcji multimedialnych.
- **Telefon C** — dawny flagowiec, mocne parametry techniczne, ale bez ekranu dotykowego.
- **Telefon D** — nowoczesny flagowiec z najwyższymi parametrami.

Wszystkie modele poprawnie rozpoznały przypadki skrajne (Telefon A i Telefon D). Warto zauważyć, że różnice w przypisaniu klasy dla Telefonu C (przypadek, w którym Random Forest przewidział klasę 3 zamiast 2) mogą wynikać z naturalnej niestabilności modeli opartych na drzewach decyzyjnych, szczególnie w przypadku predykcji pojedynczych obserwacji o cechach nietypowych lub znajdujących się na granicy między klasami.

10 Rezultaty oraz omówienie wyników

Tabela 5: Porównanie wyników modeli klasyfikacyjnych (dane po standaryzacji)

Model	**Dokładność (%)**	**F1-score (macro)**	**ROC AUC (ovr)**
Regresja logistyczna	77.212	0,7499	0,9142
LDA	94.912	0,9481	0,9958
Random Forest	82.080	0,8126	0,9613
Model hybrydowy	91.151	0,9113	0,9871

Tabela 6: Porównanie wyników modeli klasyfikacyjnych (dane po MinMax)

Model	**Dokładność (%)**	**F1-score (macro)**	**ROC AUC (ovr)**
Regresja logistyczna	76.640	0,7428	0,9128
LDA	94.230	0,9423	0,9956
Random Forest	82.690	0,8191	0,9637
Model hybrydowy	91.150	0,9107	0,9873

Wyniki uzyskane dla danych przeskalowanych metodą MinMaxScaler są bardzo zbliżone do tych osiągniętych po standaryzacji cech (StandardScaler). Najwyższe rezultaty, zarówno pod względem dokładności, F1-score, jak i

ROC AUC, ponownie uzyskały modele LDA oraz hybrydowy, co potwierdza ich wysoką skuteczność i stabilność niezależnie od przyjętego sposobu skalowania danych.

Minimalne różnice w dokładności i F1-score wskazują, że oba podejścia do skalowania są w tym przypadku w pełni akceptowalne. Dla modeli liniowych, a w szczególności regresji logistycznej, zauważono niewielki spadek skuteczności przy zastosowaniu MinMaxScaler w porównaniu ze standaryzacją, co jest zgodne z oczekiwaniami wynikającymi z charakterystyki tych algorytmów.

Warto podkreślić, że dla modeli nieliniowych, takich jak Random Forest czy model hybrydowy oparty na VotingClassifier, wybór metody skalowania miał marginalny wpływ na jakość predykcji. Świadczy to o dużej odporności tych modeli na transformacje skali zmiennych wejściowych i ich zdolności do efektywnego wykorzystywania cech o różnym zakresie wartości.

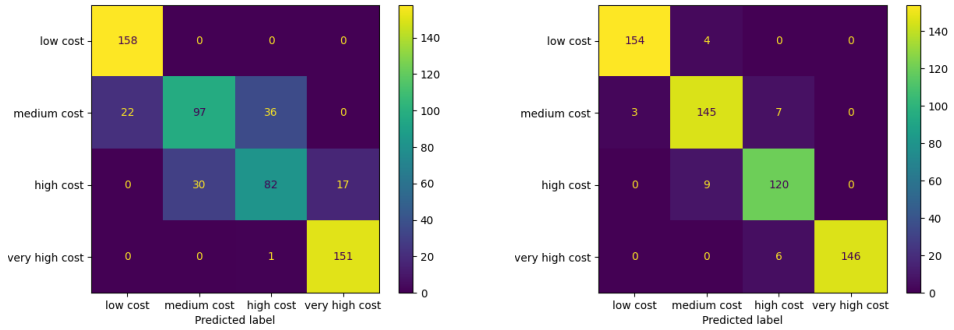
W celu oceny skuteczności modeli klasyfikacyjnych zastosowano macierze pomyłek oraz analizę istotności cech przy pomocy wykresów SHAP.

10.1 Macierze pomyłek

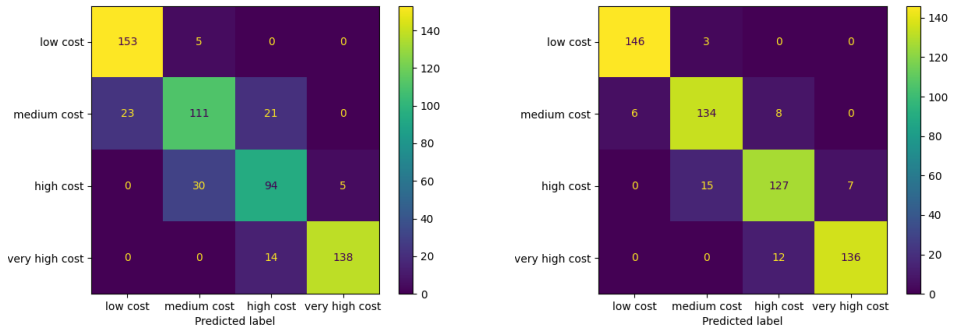
Macierz pomyłek (confusion matrix) pozwala szczegółowo przeanalizować, jak modele klasyfikacyjne radzą sobie z przypisaniem obserwacji do poszczególnych klas. W wierszach macierzy przedstawione są rzeczywiste klasy obserwacji, natomiast w kolumnach — klasy przewidziane przez model.

Idealny model miałby wartości skoncentrowane wyłącznie na przekątnej macierzy (oznaczającej poprawne klasyfikacje). Elementy poza przekątną reprezentują błędne przypisania. Analiza macierzy pomyłek pozwala zidentyfikować, które klasy są najczęściej mylone oraz w jaki sposób modele radzą sobie z rozróżnianiem klas pośrednich.

Poniżej przedstawiono macierze pomyłek dla wszystkich rozważanych modeli.



Rys. 8. Macierze pomyłek dla regresji logistycznej oraz LDA



Rys. 9. Macierze pomyłek dla Random Forest oraz modelu hybrydowego

Najlepsze wyniki uzyskał model hybrydowy oraz LDA, które skutecznie klasyfikowały wszystkie cztery klasy cenowe. Regresja logistyczna i Random Forest częściej myliły klasy sąsiednie (np. średni koszt z wysokim). Warto zauważyć, że:

- osiągnięto bardzo dobre dopasowanie dla klas „low cost” i „very high cost”,
- Model hybrydowy poprawił wyniki lasu losowego i regresji w klasach środkowych.

11 Krzywe ROC

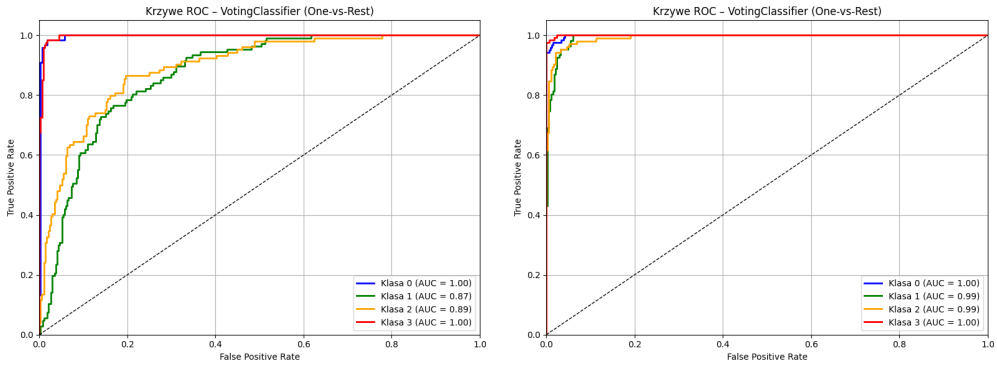
Krzywa ROC (Receiver Operating Characteristic) jest narzędziem pozwalającym ocenić zdolność modelu klasyfikacyjnego do rozróżniania między klasami. W przypadku klasyfikacji wieloklasowej wykorzystano podejście *One-vs-Rest* (OvR), gdzie dla każdej klasy wyliczono osobną krzywą ROC, traktując pozostałe klasy jako jedną wspólną kategorię.

Wysoka wartość AUC (Area Under Curve) świadczy o dobrej jakości modelu — wartość AUC bliska 1 oznacza niemal perfekcyjne rozróżnianie klas, natomiast wartość bliska 0.5 oznacza model losowy.

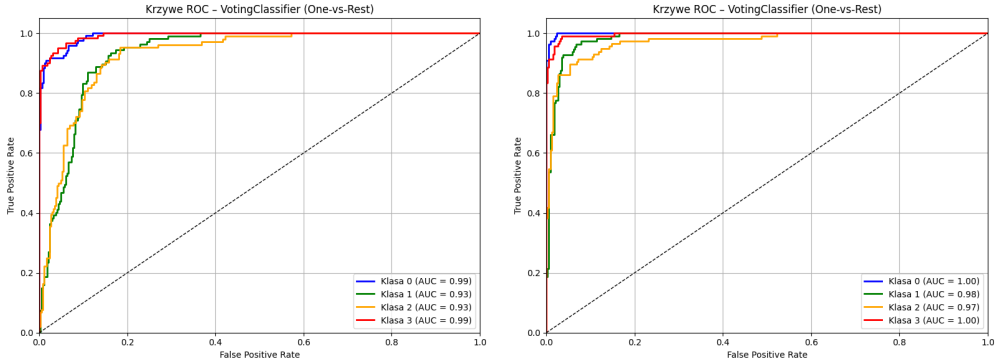
Analizując krzywe ROC dla poszczególnych modeli, możliwe jest uzyskanie dodatkowego wglądu w jakość klasyfikacji, niezależnie od progu decyzyjnego, oraz zrozumienie, w jakim stopniu modele potrafią odróżniać poszczególne klasy cenowe telefonów.

Poniżej przedstawiono krzywe ROC dla wszystkich rozważanych modeli, osobno dla danych po standaryzacji oraz po skalowaniu MinMax.

11.1 Krzywe ROC – dane po standaryzacji (StandardScaler)

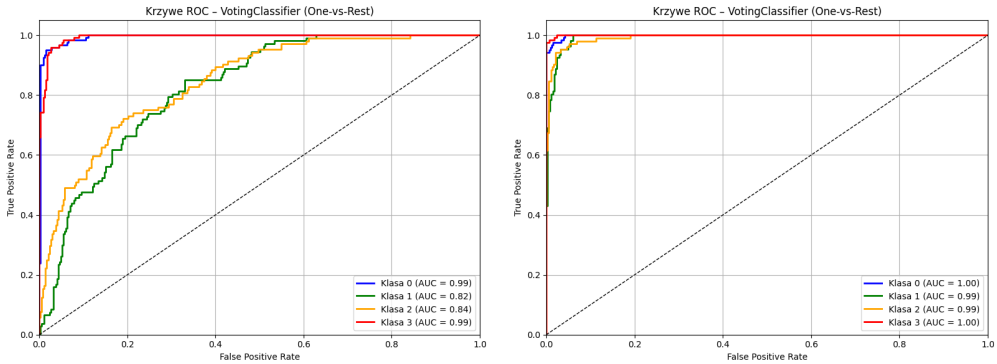


Rys. 10. Krzywe ROC dla regresji logistycznej oraz LDA)

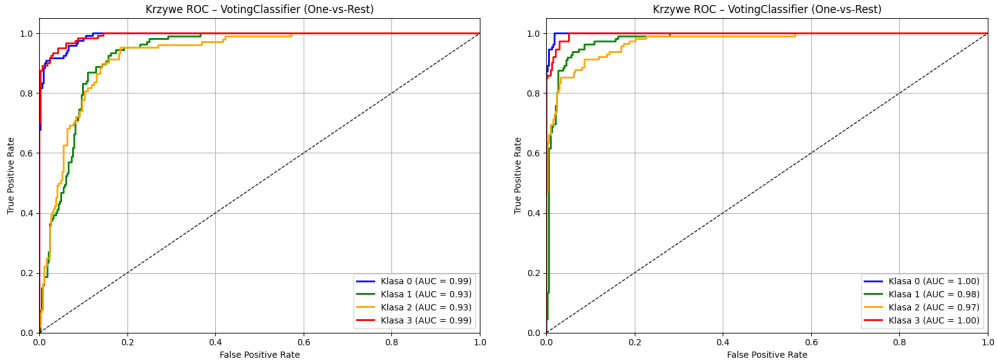


Rys. 11. Krzywe ROC dla Random Forest oraz modelu hybrydowego

11.2 Krzywe ROC – dane po skalowaniu MinMax



Rys. 12. Krzywe ROC dla regresji logistycznej oraz LDA



Rys. 13. Krzywe ROC dla Random Forest oraz modelu hybrydowego

11.3 Porównanie krzywych ROC

Analizując powyższe wykresy można zauważyć, że:

- Modele Random Forest oraz hybrydowy osiągają bardzo wysokie wartości AUC (powyżej 0.93) niezależnie od rodzaju skalowania.
- W przypadku modeli liniowych (regresja logistyczna, LDA), standaryzacja dała lepsze wyniki.
- Najwyższą stabilność i skuteczność na wszystkich klasach wykazało LDA, zaraz za nim model hybrydowy.

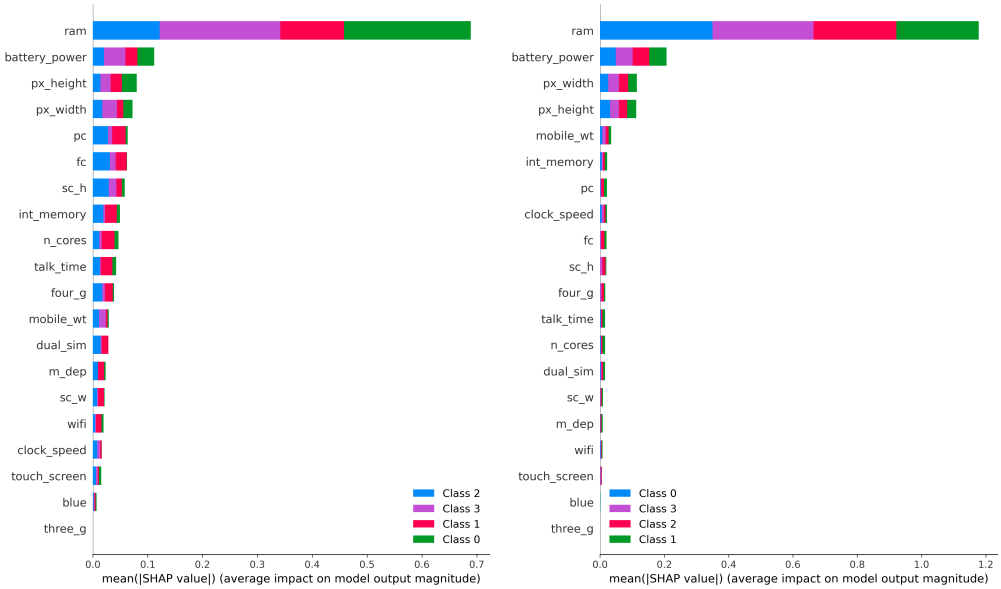
11.4 Analiza istotności cech – wykresy SHAP

Aby zrozumieć, które cechy mają największy wpływ na decyzje modeli klasyfikacyjnych, wykorzystano analizę wartości SHAP (SHapley Additive exPlanations). SHAP opiera się na koncepcji wartości Shapleya z teorii gier kooperacyjnych i umożliwia interpretację wpływu poszczególnych cech na predykcję modelu.

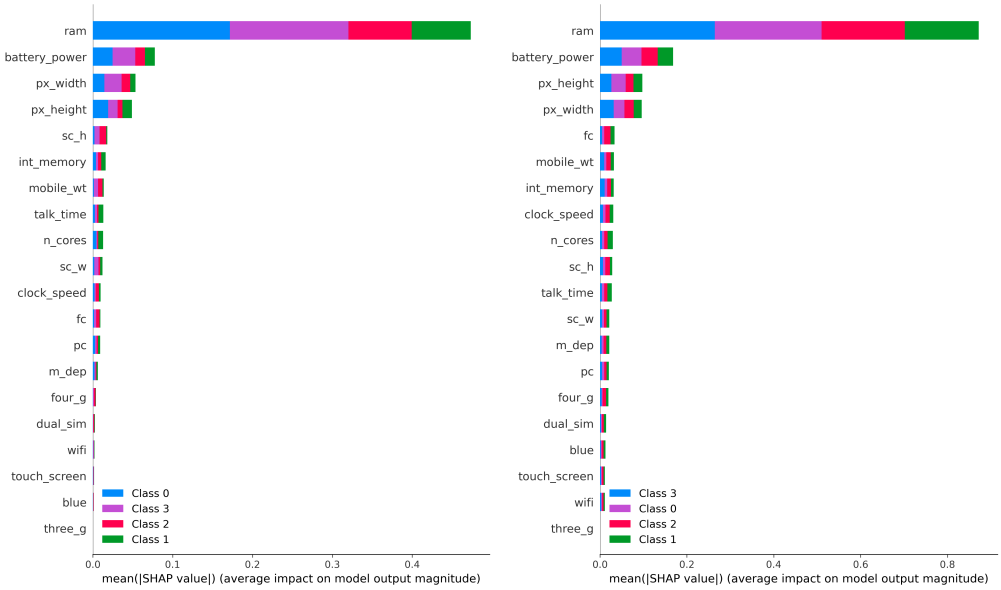
Wartości SHAP pokazują, w jakim stopniu dana cecha przyczynia się do zwiększenia lub zmniejszenia prawdopodobieństwa przypisania obserwacji do określonej klasy. Wysokie wartości dodatnie lub ujemne dla danej cechy oznaczają silny wpływ tej cechy na decyzję modelu.

Poniżej zaprezentowano wykresy SHAP dla wszystkich analizowanych modeli. Przedstawiono dwie wersje: dla danych standaryzowanych oraz dla danych po przeskalowaniu metodą MinMax.

11.4.1 Dla danych po standaryzacji

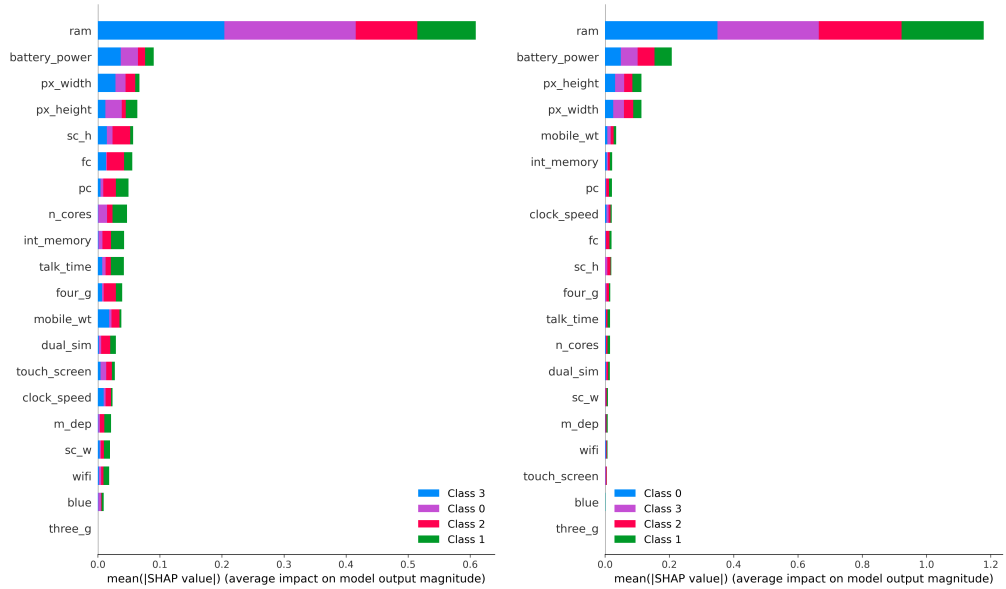


Rys. 14. Wpływ zmiennych na decyzje modeli: regresja logistyczna i LDA

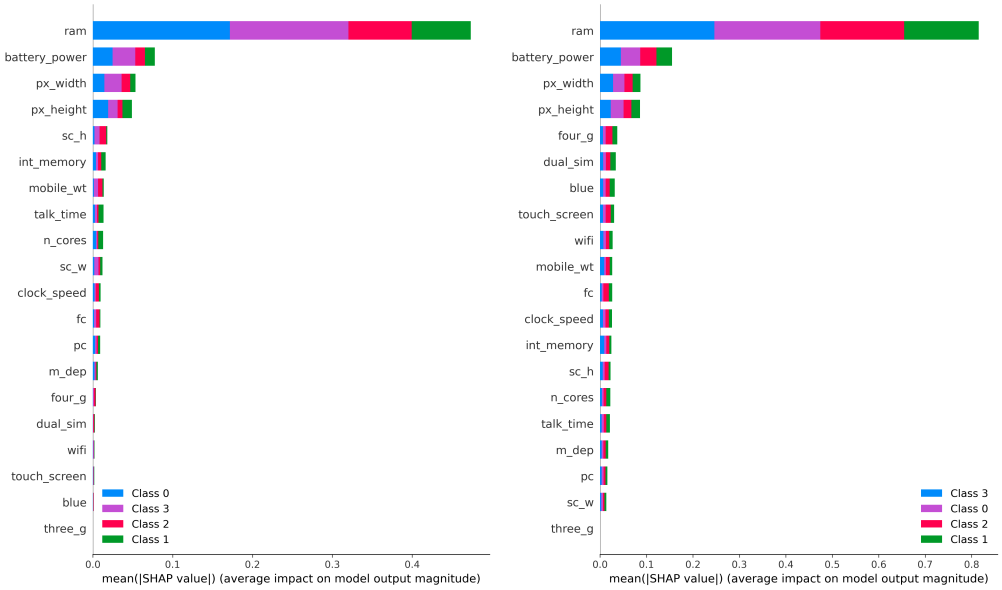


Rys. 15. Wpływ zmiennych na decyzje modeli: Random Forest i model hybrydowy

11.4.2 Dla danych przeskalowanych MinMaxScalerem



Rys. 16. Wpływ zmiennych na decyzje modeli: regresja logistyczna i LDA



Rys. 17. Wpływ zmiennych na decyzje modeli: Random Forest i model hybrydowy

Z analizy SHAP wynika, że najbardziej wpływowymi cechami dla wszystkich modeli były:

- **ram** – najistotniejszy czynnik dla wszystkich klas (im więcej pamięci RAM, tym wyższa klasa),
- **battery_power**, **px_width**, **px_height** – istotne dla modeli drzewiastych (Random Forest, hybrid),
- **four_g**, **touch_screen**, **dual_sim** – miały pewne znaczenie dla klasyfikatorów logistycznych, ale mniejsze ogólne znaczenie.

Wszystkie zastosowane modele osiągnęły stosunkowo wysoką dokładność predykcji. Najlepsze wyniki uzyskano dla modelu hybrydowego, który osiągnął dokładność na poziomie 91%. Regresja logistyczna i LDA miały nieco niższą skuteczność, odpowiednio 82% i 79%.

12 Podsumowanie

Celem projektu była klasyfikacja telefonów komórkowych na podstawie cech technicznych do jednej z czterech klas cenowych, z wykorzystaniem różnych metod klasyfikacyjnych. Przeprowadzono kompleksową analizę danych, obejmującą wstępną eksplorację, przygotowanie danych oraz porównanie skuteczności wybranych modeli.

Porównanie wyników pokazało, że najwyższą skuteczność osiągnął model hybrydowy (VotingClassifier), który uzyskał średnią dokładność powyżej 91%, przewyższając pozostałe modele. Linear Discriminant Analysis (LDA) również wykazała bardzo wysoką skuteczność, co potwierdza, że nawet klasyczne metody statystyczne mogą dobrze sprawdzać się w zadaniach klasyfikacji na danych tabelarycznych. Random Forest pozwolił uchwycić złożone, nieliniowe zależności między cechami, osiągając dobre wyniki, choć wyraźnie ustępujące modelowi hybrydowemu. Regresja logistyczna, mimo prostoty i dobrej interpretowalności, ustępowała pod względem dokładności wszystkim pozostałym modelom.

Przeprowadzona analiza wykazała, że cechy takie jak pamięć RAM, rozdzielczość ekranu oraz pojemność baterii mają kluczowy wpływ na przewidywaną klasę cenową telefonu. Modele hybrydowe i ensemble pozwoliły na lepsze radzenie sobie z trudniejszymi przypadkami, zwłaszcza dla klas pośrednich, co znalazło odzwierciedlenie zarówno w macierzach pomyłek, jak i w wysokich wartościach AUC.

Warto również podkreślić, że wyniki były stabilne niezależnie od zastosowanej metody skalowania cech (Standaryzacja vs. MinMaxScaler), co świadczy o dobrej jakości przygotowania danych.

Podsumowując, możliwe jest skuteczne przewidywanie klasy cenowej telefonu komórkowego na podstawie jego parametrów technicznych. Dobór odpowiedniej metody klasyfikacyjnej ma istotny wpływ na jakość predykcji — szczególnie w kontekście modeli ensemble, które zapewniają wysoką dokładność i stabilność. Wyniki uzyskane w projekcie są zgodne z obserwacjami przedstawionymi w literaturze, a zaproponowane podejście może zostać z powodzeniem wykorzystane w praktycznych zastosowaniach, takich jak rekomendacje produktowe czy automatyczna kategoryzacja urządzeń.

13 Bibliografia

Literatura

[1] Ramireddy, S., Singh, R. (2024). Comparative Evaluation of Machine Learning Models for Mobile Phone Price Prediction. *ResearchGate Preprint*.
https://www.researchgate.net/publication/384970604_Comparative_Evaluation_of_Machine_Learning_Models_for_Mobile_Phone_Price_Prediction_Assessing_Accuracy_Robustness_and_Generalization_Performance

[2] Achemelu, I., Nguyen, L. (2024). Classification of Mobile Price Using Machine Learning. *Proceedings of ICICIS 2024*.
https://ceur-ws.org/Vol-3682/Paper5.pdf

[3] Nguyen, T., Wang, H., Li, Y. (2024). Hybrid Ensemble Classifiers for Risk Prediction. *Expert Systems with Applications*, 235, 121012.
https://doi.org/10.1016/j.eswa.2024.121012

[4] Hosmer, D. W., Lemeshow, S., Sturdivant, R. X. (2013). *Applied Logistic Regression*. Wiley.
https://www.wiley.com/en-us/Applied+Logistic+Regression%2C+3rd+Edition-p-9780470582473

[5] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179–188.
https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

[6] Dietterich, T. G. (2000). Ensemble methods in machine learning. *International Workshop on Multiple Classifier Systems*. Springer.
https://link.springer.com/chapter/10.1007/3-540-45014-9_1

[7] Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2), 215–242.

[8] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
https://link.springer.com/article/10.1023/A:1010933404324

[9] Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
http://www.jmlr.org/papers/v12/pedregosa11a.html

[10] Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
https://hastie.su.domains/ElemStatLearn/