Klasyfikacja cen telefonów komórkowych na podstawie cech technicznych

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

3 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: regresji logistycznej, LDA, lasu losowego oraz modelu hybrydowego. 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 z wykorzystaniem cech technicznych znajduje szerokie zastosowanie w analizie danych i uczeniu maszynowym. Wśród podobnych badań można wymienić:

W pracy [5] autorzy przeanalizowali dane techniczne telefonów komórkowych ze zbioru Kaggle i ocenili skuteczność różnych klasyfikatorów, takich jak regresja logistyczna, drzewa decyzyjne oraz Random Forest. Wykazano, że modele zespołowe, w tym Random Forest, charakteryzują się wysoką dokładnością, sięgającą nawet 96

W badaniu [6] przedstawiono porównanie pięciu algorytmów: KNN, regresji logistycznej, drzewa decyzyjnego, Random Forest oraz SVM w kontekście klasyfikacji cen telefonów. Wnioski wskazują na dużą przydatność modeli opartych na głosowaniu oraz wysoką skuteczność predykcyjną RF i SVM.

W artykule [3] opisano zastosowanie klasyfikatora hybrydowego (VotingClassifier) w zadaniu predykcji ryzyka finansowego. Pokazano, że połączenie kilku modeli bazowych (np. Random Forest, LDA, XGBoost) prowadzi do większej stabilności i dokładności. Podejście to znajduje zastosowanie również w klasyfikacji wieloklasowej, takiej jak klasyfikacja telefonów wg ceny.

Z kolei w publikacjach [4] oraz [1] szczegółowo omówiono teoretyczne podstawy regresji logistycznej i analizy dyskryminacyjnej, które są często stosowane w klasyfikacji danych tablicowych. Ich interpretowalność i prostota czynią je wartościowym punktem odniesienia w analizach tego typu.

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.

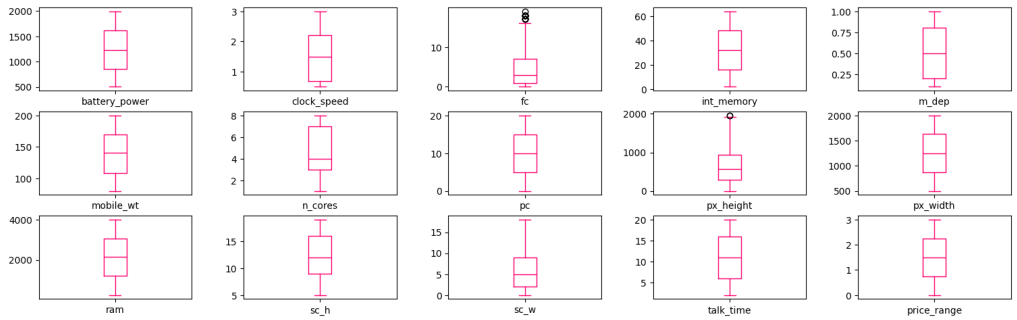
- **Kategoryczne:**
  - **three_g** – obsługa 3G,
  - **four_g** – obsługa 4G,
  - **blue** – obsługa Bluetooth,
  - **dual_sim** – obsługa dwóch kart SIM,
  - **touch_screen** – ekran dotykowy,
  - **wifi** – obsługa WiFi.

Na potrzeby projektu wprowadzimy własną zmienną `price_range`, pełniącą rolę zmiennej objaśnianej w procesie klasyfikacji. Będzie to nasza zmienna docelowa, kategoryczna wieloklasowa, podzielona na cztery klasy cenowe (0 – najtańsze, 3 – najdroższe).

7 Wstępna analiza danych

7.1 Statystyki opisowe

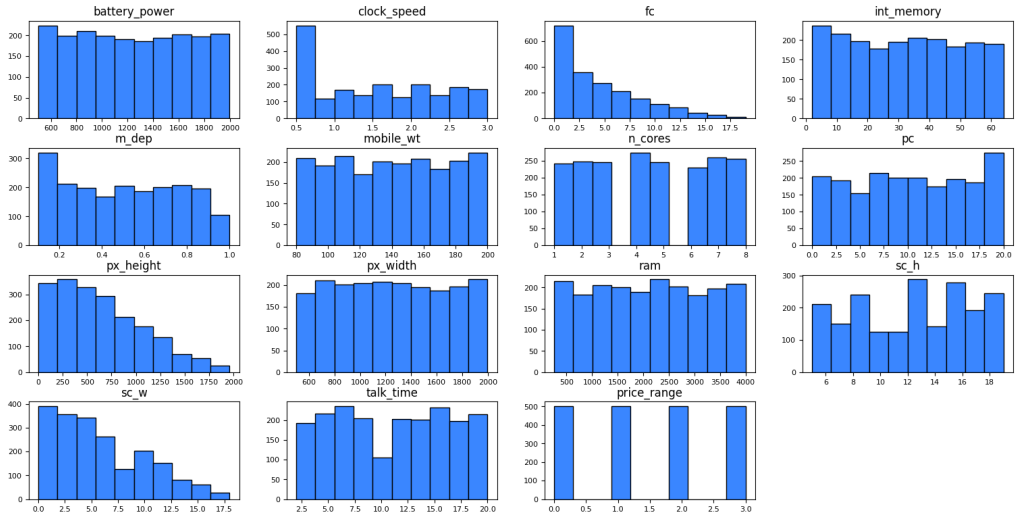
W celu lepszego zrozumienia zbioru danych przeprowadzono analizę podstawowych statystyk opisowych. Dane obejmują zarówno zmienne ilościowe, jak i kategoryczne. Poniżej zaprezentowano rozkłady i wykresy pudełkowe dla zmiennych ilościowych.



Rys. 1. Wykresy pudełkowe dla zmiennych ilościowych

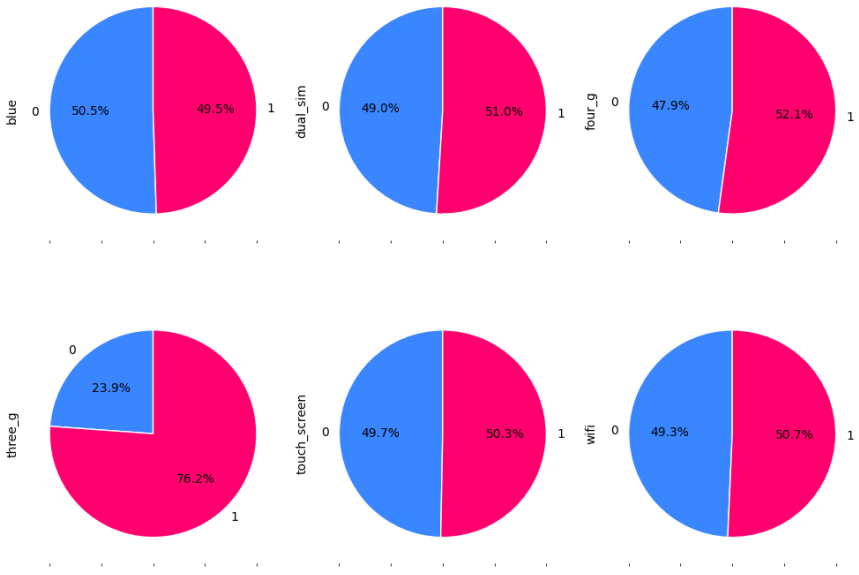
W większości zmiennych obserwujemy stosunkowo symetryczne rozkłady, choć dla niektórych, jak np. `fc` (przedni aparat), występują wartości odstające. Wartości takie mogą świadczyć o nielicznych modelach o bardzo wysokich parametrach w danej kategorii.

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.

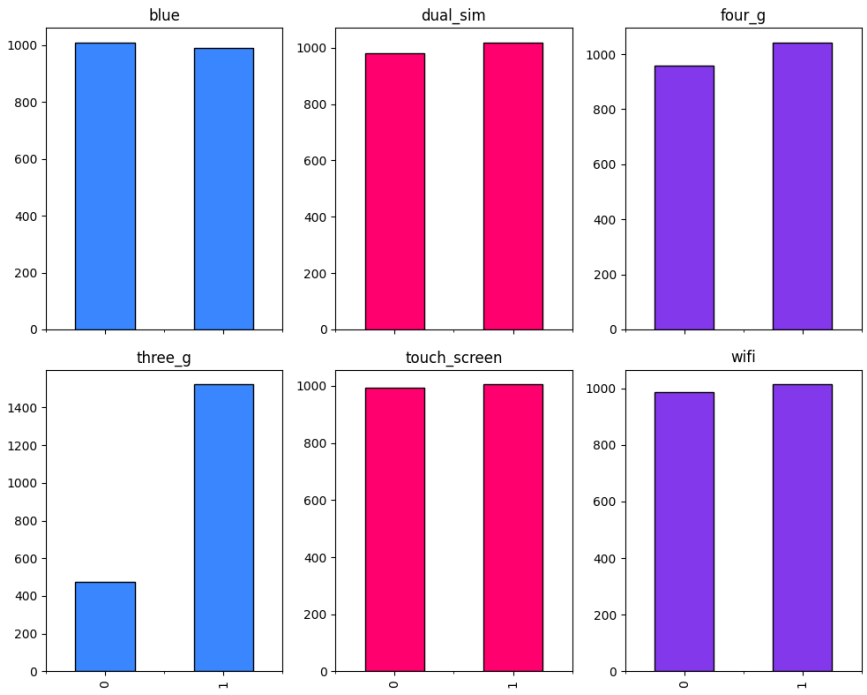


Rys. 2. Histogramy zmiennych ilościowych

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.

7.2 Brakujące oraz odstające dane

W analizowanym zbiorze danych nie stwierdzono brakujących wartości. Jednak niektóre zmienne, takie jak `fc` (przedni aparat), zawierają wartości odstające, co może sugerować obecność modeli flagowych o wyjątkowo wysokich parametrach. Wartości te pozostawiono bez modyfikacji, gdyż są one rzeczywiste i mogą zawierać istotne informacje klasyfikacyjne.

7.3 Skalowanie danych

W celu zapewnienia spójności w zakresie wartości cech ilościowych przeprowadzono ich skalowanie, co umożliwia prawidłowe działanie wielu algoryt-

mów 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

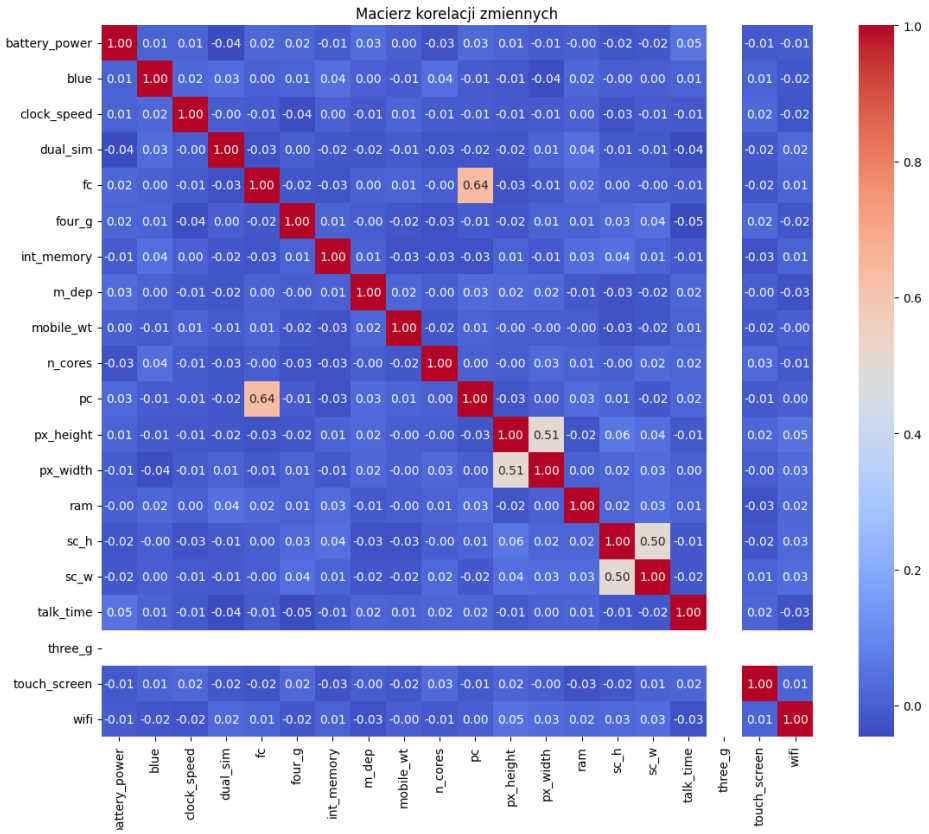
Alternatywnie stosuje się skalowanie Min-Max:

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

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.

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:

- Dokładność (test set): 82%,
- F1-score (macro): 0,8201,
- ROC AUC (One-vs-Rest): 0,9394,
- Walidacja krzyżowa (5-krotna): średnia 82,18% ± 1,15%,
- Walidacja Leave-One-Out: średnia dokładność 82,16%.

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 $\delta_k(x)$.

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 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:

- Dokładność (test set): 79%,
- F1-score (macro): 0,7910,
- ROC AUC (One-vs-Rest): 0,9300,
- Walidacja krzyżowa (5-krotna): 79,45% $\pm$ 1,58%,
- Walidacja Leave-One-Out: 79,35%.

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).

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

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

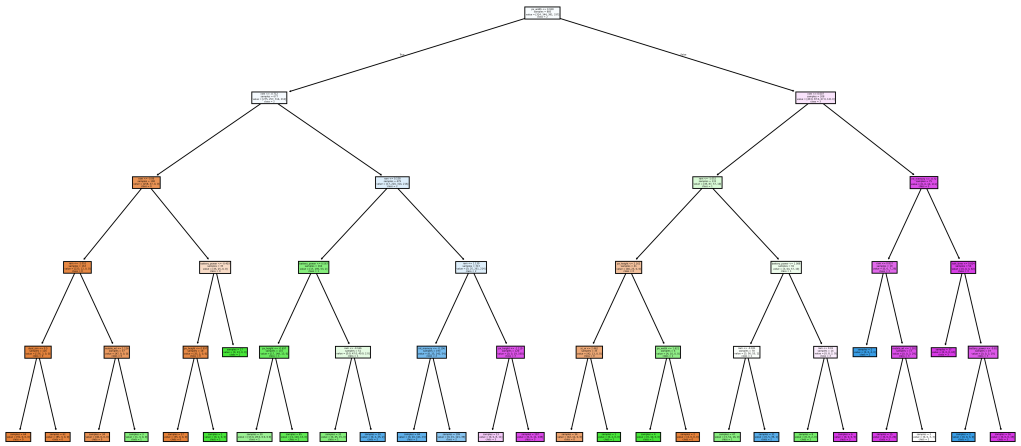
gdzie $h_t(x)$ to predykcja t-tego drzewa decyzyjnego, a T to całkowita liczba drzew w lesie.

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.

Uzyskane wyniki:

- Dokładność: 91%,
- F1-score (macro): 0,9097,
- ROC AUC (One-vs-Rest): 0,9846.

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).



Rys. 6. Przykładowe drzewo decyzyjne z modelu Random Forest

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.

8.4.2 Skład modeli bazowych

- **RandomForestClassifier** – zespół drzew decyzyjnych,

- **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:

- Dokładność (test set): 89%,
- F1-score (macro): 0,8893,
- ROC AUC (One-vs-Rest): 0,9780,
- Walidacja krzyżowa (5-krotna): 88,78% $\pm$ 0,86%.

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

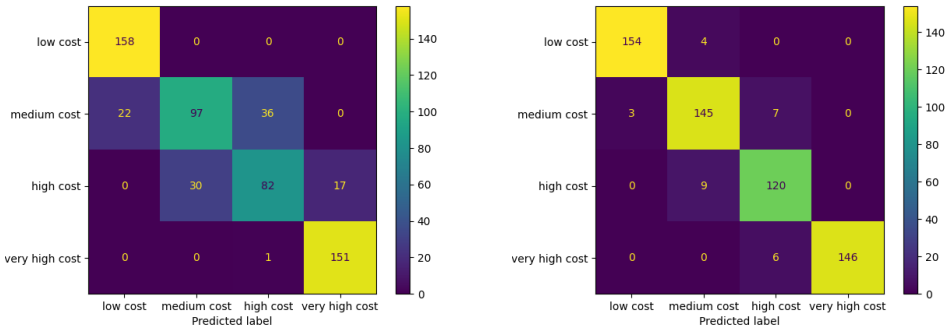
9 Rezultaty oraz omówienie wyników

Tabela 1: Porównanie wyników modeli klasyfikacyjnych

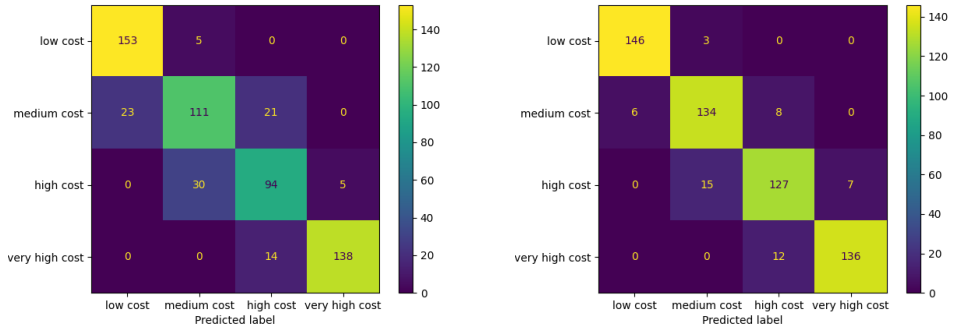
Model	**Dokładność (%)**	**F1-score (macro)**	**ROC AUC (ovr)**
Regresja logistyczna	82	0,8201	0,9394
LDA	79	0,7910	0,9300
Random Forest	91	0,9097	0,9846
Model hybrydowy	89	0,8893	0,9780

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

9.1 Macierze pomyłek



Rys. 7. Macierze pomyłek dla regresji logistycznej oraz LDA



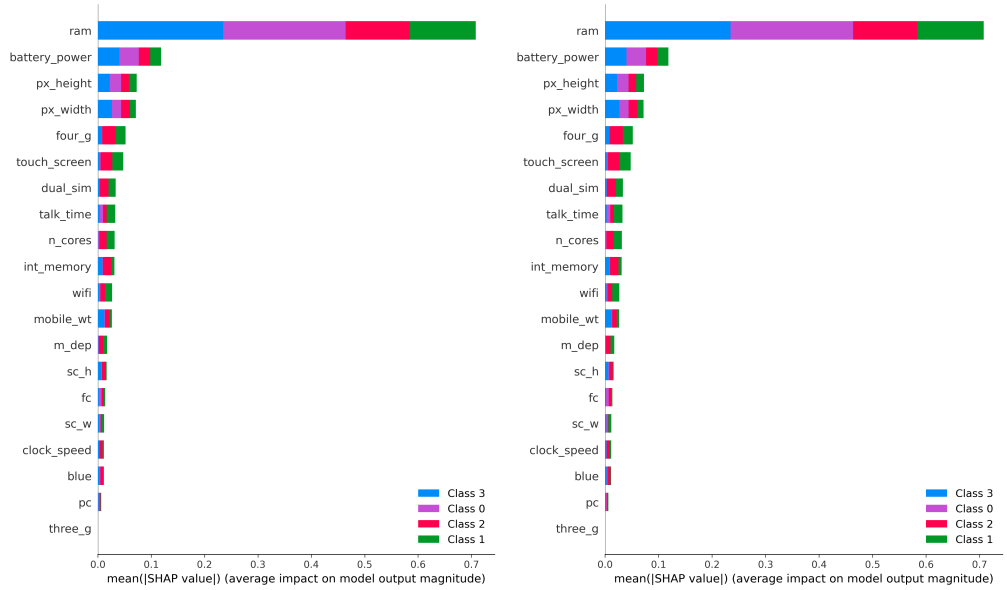
Rys. 8. Macierze pomyłek dla Random Forest oraz modelu hybrydowego

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

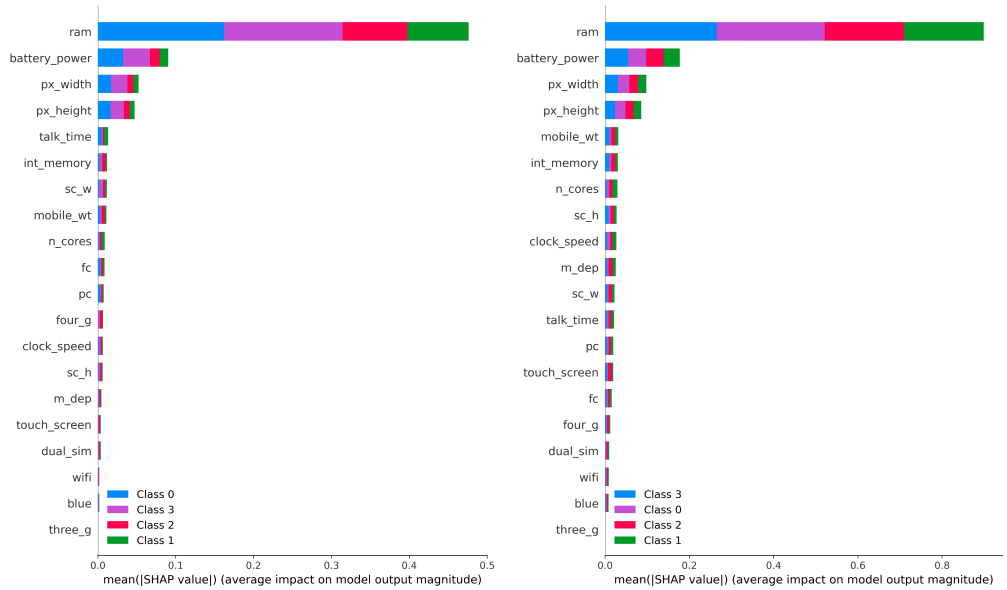
- Random Forest osiągnął bardzo dobre dopasowanie dla klas „low cost" i „very high cost",
- Model hybrydowy poprawił wyniki LDA i regresji w klasach środkowych,
- LDA miało problem z rozróżnianiem „medium cost" i „high cost".

9.2 Analiza istotności cech – wykresy SHAP

Poniższe wykresy SHAP ilustrują średni wpływ poszczególnych cech na predykcję modelu dla każdej z klas.



Rys. 9. Wpływ zmiennych na decyzje modeli: regresja logistyczna i LDA (SHAP)



Rys. 10. Wpływ zmiennych na decyzje modeli: Random Forest i model hybrydowy (SHAP)

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 algorytmu Random Forest, który osiągnął dokładność na poziomie 91%. Regresja logistyczna i LDA miały nieco niższą skuteczność, odpowiednio 82% i 79%. Połączenie modeli w formie głosowania większościowego (ensemble) pozwoliło na uzyskanie stabilnego wyniku na poziomie 89%.

10 Podsumowanie

W przeprowadzonym badaniu porównano skuteczność czterech modeli klasyfikacyjnych w przewidywaniu przedziału cenowego telefonów komórkowych.

Porównanie skuteczności modeli pokazało, że najlepsze wyniki uzyskał algorytm Random Forest z dokładnością 91% oraz F1-score 0,91. Model hybrydowy osiągnął bardzo dobre rezultaty (89% dokładności), łącząc zalety klasyfikatorów bazowych. Regresja logistyczna oraz LDA, mimo niższych wyników (odpowiednio 82% i 79%), sprawdziły się jako proste i interpretowalne modele, szczególnie skuteczne przy klasyfikowaniu klas skrajnych. Modele zespołowe lepiej radziły sobie z rozróżnianiem klas środkowych, co wskazuje na ich większą elastyczność w uchwyceniu złożonych wzorców w danych.

Projekt pokazał, że cechy takie jak pamięć RAM, rozdzielczość ekranu oraz pojemność baterii mają kluczowe znaczenie w przewidywaniu ceny urządzenia.

11 Bibliografia

Literatura

[1] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer, 2009.

[2] G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning*, Springer, 2013.

[3] Nguyen, T., Wang, H., Li, Y. (2024). Hybrid Ensemble Classifiers for Risk Prediction. *Expert Systems with Applications*, 235, 121012.

[4] David W. Hosmer, Stanley Lemeshow, Rodney X. Sturdivant. *Applied Logistic Regression*. Wiley, 2013.

[5] 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`

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