Model	Params	Accuracy	Precision	Recall	F1 Score
Dummy	strategy=”prior”	0.639	0.408	0.639	0.498
SVM	max_iter=2000, C=1.0, kernel='rbf', degree=3	0.440	0.491	0.440	0.460
Random Forest	max_depth=55, n_jobs=-1, n_estimators=100, criterion='gini'	0.650	0.604	0.650	0.532

Analiza wyników modeli w kontekście EDA

Przeprowadzona wcześniej eksploracyjna analiza danych okazała się pomocna i miała pewien wpływ na skuteczność trenowanych modeli. (nie duży, ale miała)

- **Silna korelacja liczby "loves_count", "reviews" i "rating" z etykietą (LABEL-rating)** – decyzja o ich pozostawieniu jako cech numerycznych została potwierdzona.
- **Wartości tekstowe w kolumnach `review_text` i `review_title`** – zostały uwzględnione przy użyciu `CountVectorizer`,. Zastosowanie ograniczeń `max_features=200`, `min_df=2`, `max_df=0.95` również opierało się na analizie rozkładu słów z EDA.
- **highlight`** zostały zakodowane binarnie na podstawie 20 najczęstszych elementów, aby móc zobaczyć, czy etykiety były powiązane z wysoką oceną
- **Uzupełnienie helpfullness** za pomocą odpowiedniego wzoru umożliwiło pozyskanie autentycznych wartości i ominięcie estymacji
- **Usunięcie kolumn**, które zawierały nieznaczną lub mało ważną wiedzę zmniejszyło liczbę atrybutów i poprawiło wyniki.
- **Utworzenie 4 cech** takich jak 'review_length', 'contains_refund', 'exclamation_count', 'unique_word_count' poszerzyło liczbę atrymutów o bardziej przydatne cechy
- **Naprawiono część kolumn** takich jak highlights, aby zawierały dane w listach, a nie string będący listą
- **Wyczyszczono teskt** za pomocą regexa, stopwords oraz użyto lemmatizera
- **Zastosowanie redukcji wymiarowości (TruncatedSVD)** – pomogło ograniczyć przestrzeń cech po przetworzeniu danych tekstowych i kategorycznych. Było to potrzebne, ponieważ posiadaliśmy dużą liczbę cech wynikającą z one-hot encodingu i wektoryzacji tekstu. Zdecydowanie przyspieszyło to proces uczenia i nieznacznie wpłynęło pozytywnie na wynik

Najlepiej wypadł model Random Forest, co również było przewidywane podczas EDA – cechy były nieliniowe, a ten model dobrze to wykorzystuje. Ponadto działa dobrze przy danych mieszanych: numerycznych, kategorycznych i tekstowych.

Podsumowując, wnioski z EDA były przydatne i znalazły potwierdzenie w skuteczności ostatecznego modelu.