Sprawozdanie z projektu
Analiza i przewidywanie jakości wina

Autor: Artur Kompała

Wstęp i cel analizy

Celem niniejszego projektu jest zbadanie, czy na podstawie obiektywnych właściwości fizykochemicznych (takich jak kwasowość, zawartość cukru, poziom siarczanów czy alkohol itp) możliwe jest skuteczne przewidzenie oceny jakości wina. Analiza ma również na celu identyfikację cech, które w największym stopniu determinują wysoką ocenę trunku.

Jako zbiór danych wybrano "Wine Quality" pochodzący z UCI Machine Learning Repository. Zbiór ten zawiera dane dotyczące win czerwonych i białych, obejmując 11 zmiennych wejściowych (cechy fizykochemiczne) oraz jedną zmienną wyjściową (ocena jakości w skali punktowej).

Pytanie badawcze:

Czy na podstawie właściwości fizykochemicznych wina jesteśmy w stanie skutecznie przewidzieć, czy wino zostanie ocenione jako wysokiej jakości, oraz które cechy mają na to największy wpływ?

Eksploracyjna analiza danych (EDA)

W pierwszej fazie projektu przeprowadzono wstępną analizę danych w celu zrozumienia ich struktury i jakości.

- **Weryfikacja jakości danych:** Sprawdzono zbiór pod kątem brakujących wartości. Analiza wykazała, że zbiór jest kompletny i nie wymaga imputacji danych (brak wartości null we wszystkich kolumnach).
- **Analiza rozkładu zmiennej celowej:** Wykres rozkładu jakości wina pokazał, że większość próbek posiada oceny średnie (5-6), a win o ocenach skrajnych (bardzo słabych lub wybitnych) jest znacznie mniej. Wskazuje to na problem niezbalansowania klas, co wzięto pod uwagę przy doborze metryk oceny modelu.
- **Korelacje:** Sporządzono macierz korelacji (heatmap), aby zbadać zależności między cechami. Pozwoliło to na wstępną ocenę, które zmienne mogą być silnie skorelowane z jakością wina (np. alkohol), a które mogą wprowadzać szum informacyjny.

Przygotowanie danych i metodologia

Aby umożliwić algorytmom uczenia maszynowego efektywną pracę, dane poddano wstępnej obróbce:

- Podział danych: Zbiór został podzielony na część treningową i testową, co pozwala na rzetelną ocenę zdolności generalizacji modelu na nowych, niewidzianych wcześniej danych.
- Dobór modeli: Do rozwiązania problemu klasyfikacji wybrano dwa algorytmy w celu ich porównania:
    - Regresja Logistyczna: Jako model bazowy, pozwalający ocenić liniowe zależności.
    - Las Losowy (Random Forest): Jako bardziej zaawansowany model zespołowy, zdolny do wychwytywania nieliniowych zależności i bardziej odporny na dysproporcje w danych.

Każdy z modeli został oceniony przy użyciu metryk Accuracy (dokładność) oraz F1-score. Wybór F1-score był podyktowany koniecznością uwzględnienia balansu między precyzją (precision) a czułością (recall), co jest kluczowe przy nierównomiernym rozkładzie ocen jakości.

Wyniki analizy

Przeprowadzono trenowanie i testowanie obu modeli.

- **Porównanie skuteczności**: Analiza porównawcza wykazała, że model Lasu Losowego osiągnął lepsze wyniki niż Regresja Logistyczna. Uzyskał on wyższą dokładność oraz, co istotne, lepszy wynik F1-score. Oznacza to, że algorytm ten lepiej radzi sobie z klasyfikacją próbek, minimalizując liczbę błędnych przyporządkowań.
- **Ważność cech**: Dzięki zastosowaniu Lasu Losowego możliwe było wyznaczenie ważności poszczególnych zmiennych. Analiza wykazała, że największy wpływ na predykcję wysokiej jakości wina mają:
    - Zawartość alkoholu: (często wyższa w winach lepiej ocenianych).
    - Gęstość (density).
    - Kwasowość.

Wnioski i interpretacja

Na podstawie przeprowadzonego eksperymentu można sformułować następujące odpowiedzi na pytanie badawcze:

- **Możliwość predykcji:** Tak, właściwości fizykochemiczne pozwalają na przewidzenie jakości wina z satysfakcjonującą skutecznością, przy czym modele nieliniowe (takie jak Las Losowy) sprawdzają się w tym zadaniu lepiej.
- **Kluczowe czynniki:** Potwierdzono, że parametry takie jak poziom alkoholu i kwasowość są kluczowymi wyznacznikami jakości w percepcji ekspertów oceniających wina w zbiorze danych.

Zastosowanie biznesowe: Opracowany model mógłby posłużyć winiarniom oraz dystrybutorom do wstępnej selekcji win o potencjale na segment premium ("Quality Control"). Automatyzacja tego procesu na wczesnym etapie, jeszcze przed kosztownym procesem testowania przez sommelierów, mogłaby przynieść oszczędności czasu i zasobów.