**Dokumentacja zadania 4: Analiza klasyfikacji jakości wina z użyciem SVM**

**Artem Kukushkin 317140**

**Cel projektu**

Celem było porównanie skuteczności klasyfikatora SVM z modelem bazowym (regresją logistyczną) w binarnej klasyfikacji jakości czerwonego wina na podstawie fizykochemicznych cech. Skupiono się nie tylko na dokładności, ale również na jakości przewidywań i równowadze między klasami.

**Przygotowanie danych**

Zmienna docelowa (quality) została przekształcona w zmienną binarną:

**Złe wina** (ocena < 6) jako klasa 0.

**Dobre wina** (ocena ≥ 6) jako klasa 1.

Po podziale danych na zbiór treningowy i testowy (70% treningowy; 30% testowy) dokonano standaryzacji cech.

**Trening i optymalizacja SVM**

Zastosowano przeszukiwanie siatki hiperparametrów (GridSearchCV), aby dobrać najlepsze parametry SVM. Najlepiej sprawdziła się funkcja jądra RBF z parametrami:

C = 1 — kompromis między marginesem a błędami klasyfikacji,

gamma = 0.1 — zrównoważony wpływ pojedynczych punktów na granicę decyzyjną.

**Analiza wyników**

**A screenshot of a computer

AI-generated content may be incorrect.**

**SVM (jądro RBF)**

**Dokładność: 75,6%** — oznacza, że 3 na 4 próbki zostały poprawnie sklasyfikowane. W kontekście rzeczywistej aplikacji (np. automatyczna selekcja jakości win) to dobry wynik.

**F1-score dla klasy 1 (dobre wina): 0.77** — model dobrze radzi sobie z wykrywaniem wysokiej jakości win, co może być cenniejsze z biznesowego punktu widzenia (np. promocja jakościowych produktów).

**Równowaga klas**: F1 dla klasy 0 wynosi 0.74 — model nie faworyzuje jednej klasy, co jest istotne w przypadku niezbalansowanych danych.

**Macierz pomyłek** ujawnia, że pomyłki rozkładają się równomiernie:

48 błędnych predykcji dobrych win wśród złych (false positives),

69 błędnych predykcji złych win wśród dobrych (false negatives).

Interpretacja: SVM zachowuje dobrą równowagę między precyzją a czułością. Fałszywe alarmy (false positives) są nieco mniej liczne niż w modelu bazowym, co może być istotne, jeśli błędne zakwalifikowanie wina jako "dobrego" niesie za sobą konsekwencje (np. wysoka cena).

**Regresja logistyczna (model bazowy)**

**Dokładność: 73,5%** — o 2% niższa niż SVM, co potwierdza, że model jest mniej dopasowany do danych.

**Więcej błędów klasyfikacji klasy 0 (złe wina)** — 56 przypadków zostało błędnie uznanych za dobre (więcej niż w SVM).

**Precyzja klasy 0: 0.69** vs **0.71 w SVM** — model częściej "myli się na plus", co może prowadzić do błędnych decyzji biznesowych (np. wprowadzenie słabego wina na rynek premium).

**Czułość klasy 1 (dobre wina): 0.73**, vs **0.74 w SVM** — różnica niewielka, ale przy dużych zbiorach może być istotna.

**Interpretacja**: Regresja logistyczna gorzej radzi sobie z identyfikacją złych win — wrażliwa na liniowość granicy decyzyjnej, nie modeluje dobrze nieliniowych zależności w danych. W przypadku bardziej złożonych zbiorów może zaniżać wyniki klasyfikacji.

**Wnioski końcowe**

**SVM z jądrem RBF lepiej modeluje nieliniowe zależności** w danych, co przekłada się na wyższą dokładność, lepszy F1-score i mniejszą liczbę błędnych klasyfikacji.

**Równowaga między klasami** w wynikach SVM świadczy o solidnym dopasowaniu modelu do problemu — nie ma faworyzowania klasy dominującej.

**Model bazowy jest prostszy, ale mniej elastyczny** — może być szybszy w treningu, ale jego zastosowanie ogranicza się do prostszych zależności.