Klasyfikacja Jakości Wina – Algorytm k-NN

Tsimafei Kotski 158740
Vasil Kusmartsev 156202
L2

Streszczenie

Zbiór danych zawierał informacje o 1599 winach opisanych 11 atrybutami, zaklasyfikowanych do trzech kategorii jakości: dobra, średnia i zła. Przeprowadzono pełny proces przetwarzania danych, podziału na zbiory treningowy i testowy, normalizację, wybór optymalnej wartości parametru k za pomocą walidacji krzyżowej oraz klasyfikację przy użyciu algorytmu k-NN.

1. Normalizacja danych

Do przetworzenia danych wykorzystano metodę StandardScaler z biblioteki scikit-learn. StandardScaler standaryzuje cechy poprzez przesunięcie ich tak, aby miały średnią równą 0 i odchylenie standardowe równe 1, zgodnie ze wzorem:

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

gdzie:

- z to znormalizowana wartość
- x to oryginalna wartość
- μ to średnia
- σ to odchylenie standardowe

Normalizacja jest krytycznie ważna dla algorytmu k-NN, ponieważ algorytm ten oblicza odległości między obiektami. Cechy z dużymi wartościami mogą całkowicie zdominować cechy z małymi wartościami podczas obliczania odległości euklidesowej. Normalizacja gwarantuje, że wszystkie cechy mają równe znaczenie przy klasyfikacji.

2. Podział Danych na Zbiory Treningowy i Testowy

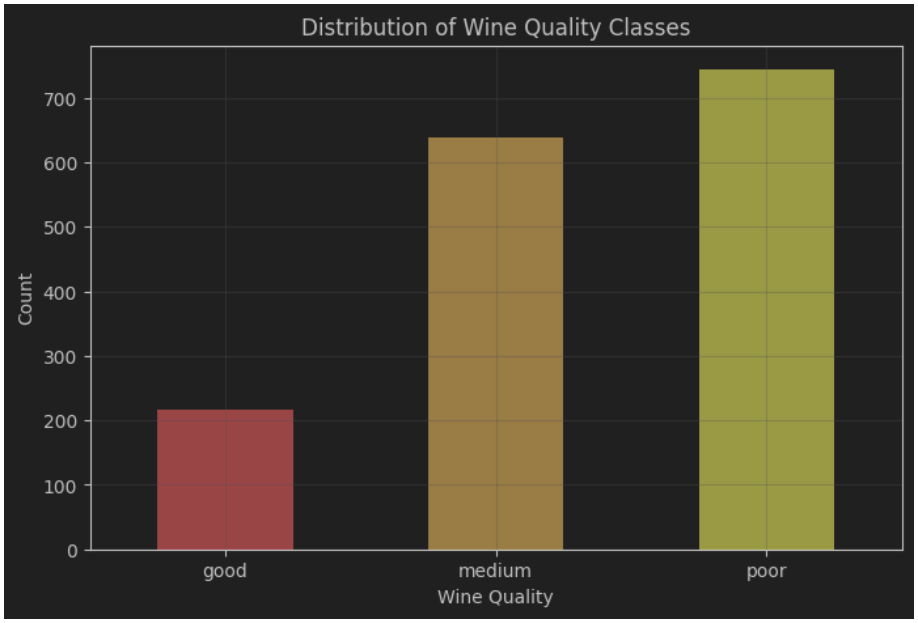
2.1 Parametry Podziału

- Rozmiar zbioru treningowego: 1279 próbek (80%)
- Rozmiar zbioru testowego: 320 próbek (20%)
- Losowe generowanie (random_state): 5

2.2 Rozkład Klas

Zbiór danych charakteryzuje się niezbalansowanym rozkładem klas:

- Wina złej jakości (poor): 744 próbki (46,53%)
- Wina średniej jakości (medium): 638 próbek (39,90%)
- Wina dobrej jakości (good): 217 próbek (13,57%)



3. Wybór Optymalnej Wartości Parametru k

3.1 Walidacja Krzyżowa

Do wyboru optymalnej wartości parametru k zastosowano 5-krotną walidację krzyżową (5-fold cross-validation). Metodyka ta polega na:

1. Podzieleniu zbioru treningowego (1279 próbek) na 5 równych części (foldy)
2. Przeprowadzeniu 5 iteracji, w każdej używając innego foldu jako zestawu walidacyjnego, a pozostałych 4 foldów do trenowania
3. Uśrednieniu wyników z 5 iteracji w celu uzyskania wiarygodnej oceny

3.2 Wyniki dla Poszczególnych Wartości k

Testowano wartości k od 1 do 20. Wyniki 5-krotnej walidacji krzyżowej są następujące:

k	Dokładność CV	k	Dokładność CV
1	0.6881 (68,81%)	11	0.6005 (60,05%)
2	0.5942 (59,42%)	12	0.6052 (60,52%)
3	0.5864 (58,64%)	13	0.6099 (60,99%)
4	0.5755 (57,55%)	14	0.6169 (61,69%)
k	Dokładność CV	k	Dokładność CV
5	0.5911 (59,11%)	15	0.6099 (60,99%)
6	0.5911 (59,11%)	16	0.5919 (59,19%)
7	0.5997 (59,97%)	17	0.6044 (60,44%)
8	0.5997 (59,97%)	18	0.6044 (60,44%)
9	0.6005 (60,05%)	19	0.6114 (61,14%)
10	0.6005 (60,05%)	20	0.6067 (60,67%)

3.3 Analiza Wyników

Wyniki wskazują, że k=1 osiąga najwyższą dokładność walidacji krzyżowej na poziomie 68,81%. Dokładność spada dla k=2 (59,42%), a następnie powoli rośnie dla większych wartości k, osiągając maksimum dla k=14 (61,69%).

Taki wynik sugeruje, że dla tego zbioru danych najbliższy sąsiad (k=1) dostarcza najlepszych prognoz. Możemy wywnioskować, że punkty danych w tej przestrzeni cech są wystarczająco dobrze rozdzielone, a wpływ szumu jest minimalny.

3.4 Wyniki Testowania

- Dokładność walidacji krzyżowej: 0.6881 (68,81%)
- Dokładność na zbiorze testowym: 0.6656 (66,56%)

Różnica między dokładnością walidacji krzyżowej a dokładnością na zbiorze testowym wyniosła 2,25%, co wskazuje na dobrą zdolność uogólniania modelu. Brak znacznej różnicy sugeruje, że model nie doległ overfit-ingowi.