Analýza kvality vína pomocou logistickej regresie

1. Úvod

Cieľom projektu bolo klasifikovať kvalitu vína ako **nadpriemernú** alebo **podpriemernú** na základe fyzikálno-chemických vlastností. Pracovalo sa s **datasetom Wine Quality** , ktorý obsahuje údaje ako napr. obsah alkoholu, kyslosti, cukru, pH a ďalšie premenné.

2. Príprava dát

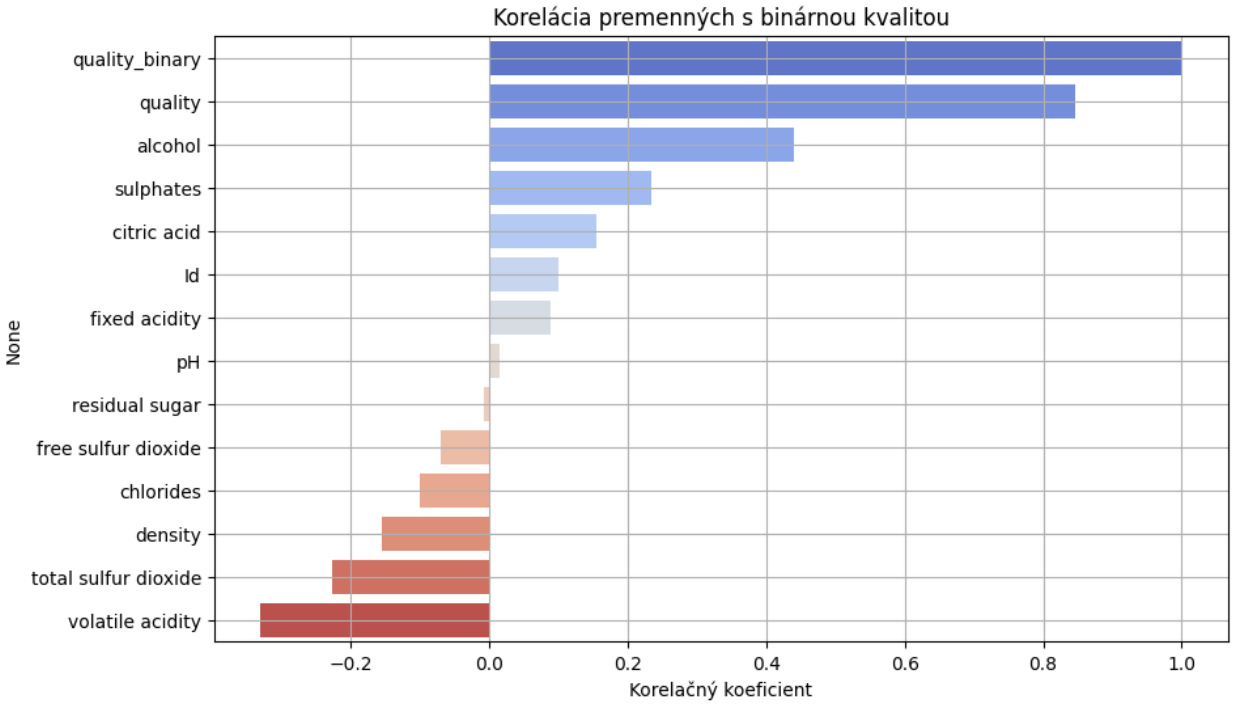
- **Celkový počet vzoriek:** 1599
- Pôvodná premenná quality bola nahradená binárnou premennou quality_binary:
    - 1 = nadpriemerná kvalita (vyššia ako priemer)
    - 0 = podpriemerná kvalita

Dáta boli rozdelené na trénovaciu (70 %) a testovaciu (30 %) množinu. Normalizácia vstupných premenných prebehla pomocou štandardnej škály.

3. Exploračná analýza – vizualizácie

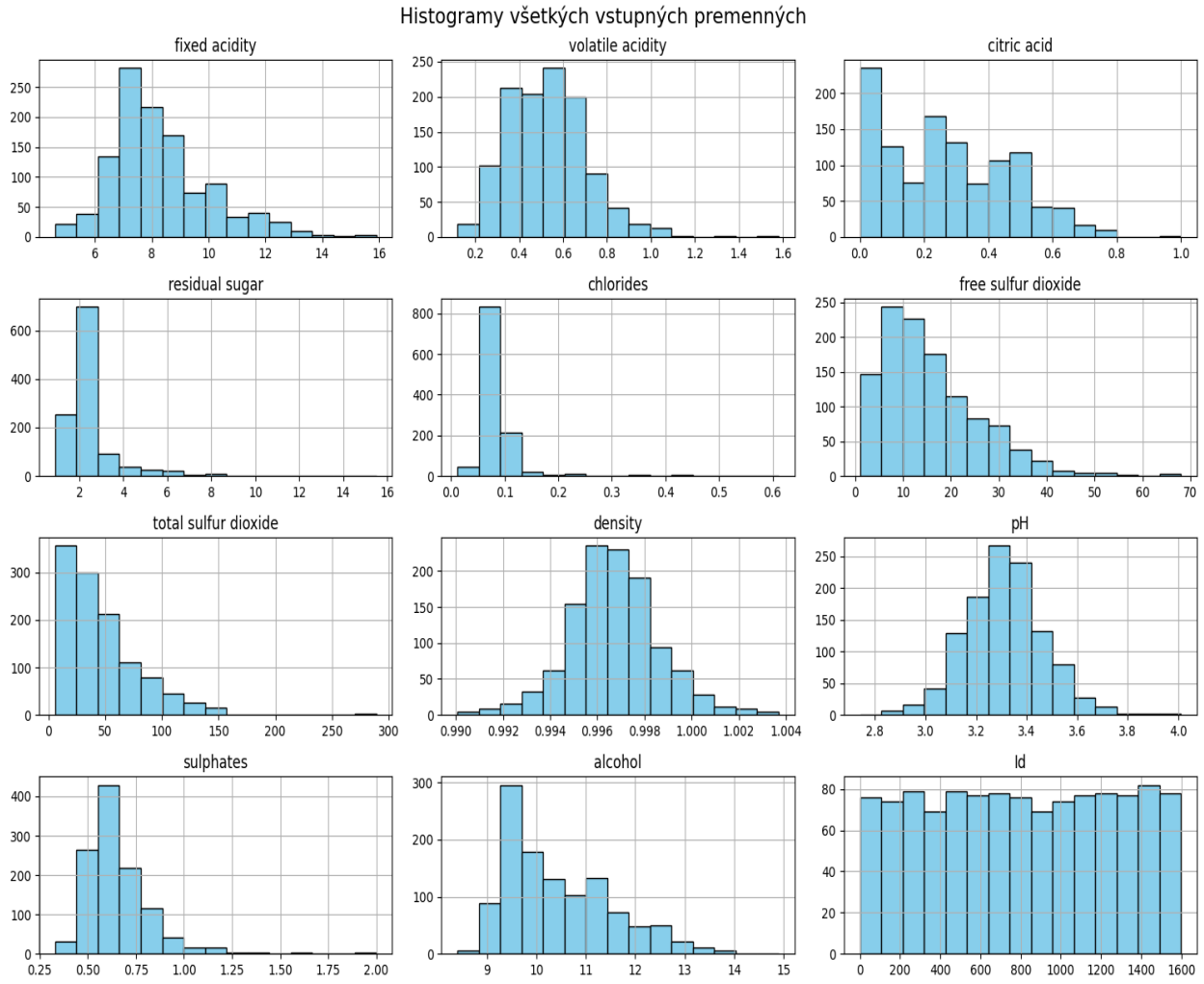
Korelačná matica (heatmapa)

Korelačná matica zobrazuje vzájomné vzťahy medzi číselnými premennými. Silné pozitívne alebo negatívne korelácie môžu naznačovať multikolinearitu, čo je dôležité pri výbere vstupov pre model. Napríklad premenné alcohol a density môžu byť nepriamo úmerné.



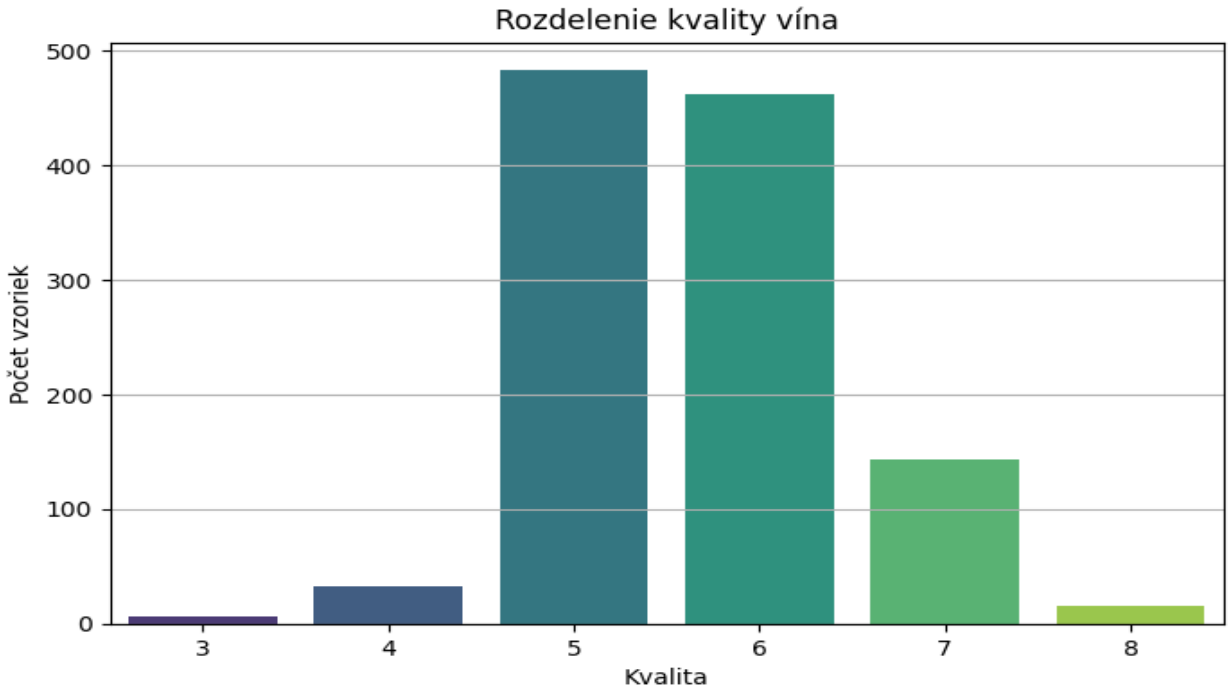
Histogramy pre všetky vstupné premenné

Histogramy znázorňujú rozdelenie jednotlivých číselných premenných v datasete vína. Viaceré premenné, ako napríklad alkohol alebo citric acid, majú mierne šikmé rozdelenie, čo môže ovplyvniť výkon modelov. Tieto vizualizácie pomáhajú identifikovať prípadné odľahlé hodnoty, asimetriu a rozsah dát.



Histogram: kvalita vína

Histogram zobrazuje frekvenciu jednotlivých hodnotení kvality vína v datasete. Väčšina vín má hodnotenie v rozsahu 5 až 7, pričom najčastejšie sa vyskytuje kvalita 6. Rozdelenie je mierne šikmé doľava, čo znamená, že vyššie hodnotenia (8, 9) sú zriedkavé. Táto distribúcia naznačuje, že dataset je mierne nevyvážený, čo môže ovplyvniť výkonnosť klasifikačných modelov.

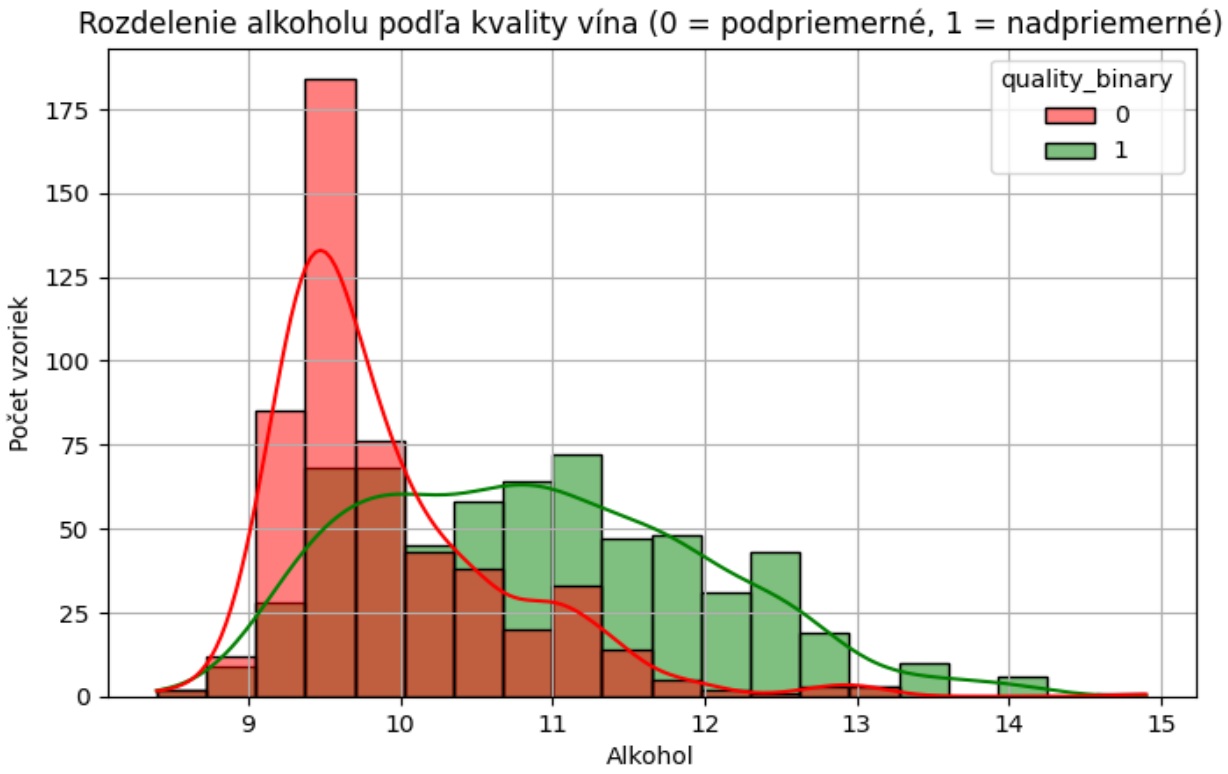


Histogram s rozdelením podľa triedy: pod vs. nad priemerom

Histogramy zobrazujú rozdelenie hodnôt vybraných vstupných premenných (napr. alcohol, volatile acidity, sulphates...) oddelene pre triedy **podpriemerná kvalita (0)** a **nadpriemerná kvalita (1)**.

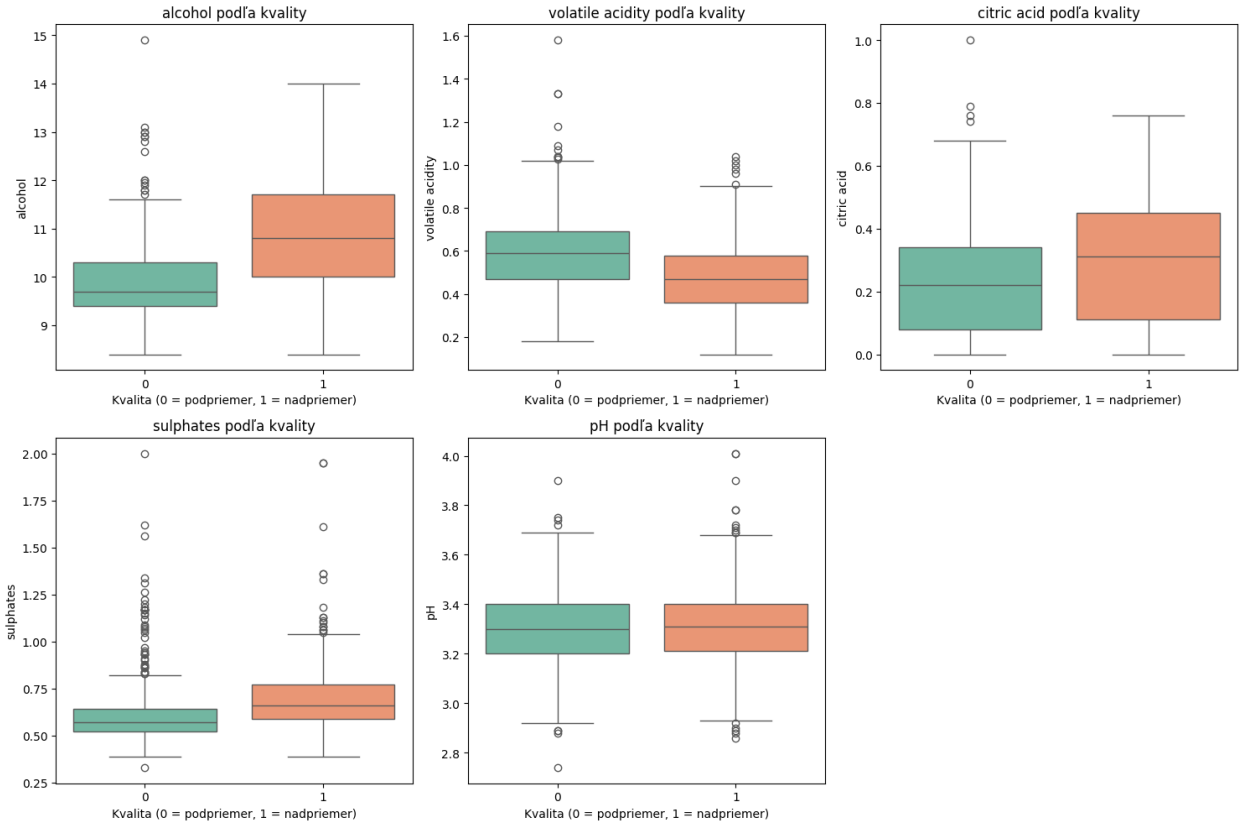
Z grafov je možné pozorovať, že:

- **Nadpriemerné vína** majú spravidla **vyšší obsah alkoholu** a často aj vyššie hodnoty **sulphates**.
- **Podpriemerné vína** majú zvyčajne **vyššiu volatilnú kyslosť (volatile acidity)**, čo môže negatívne vplývať na ich hodnotenie.



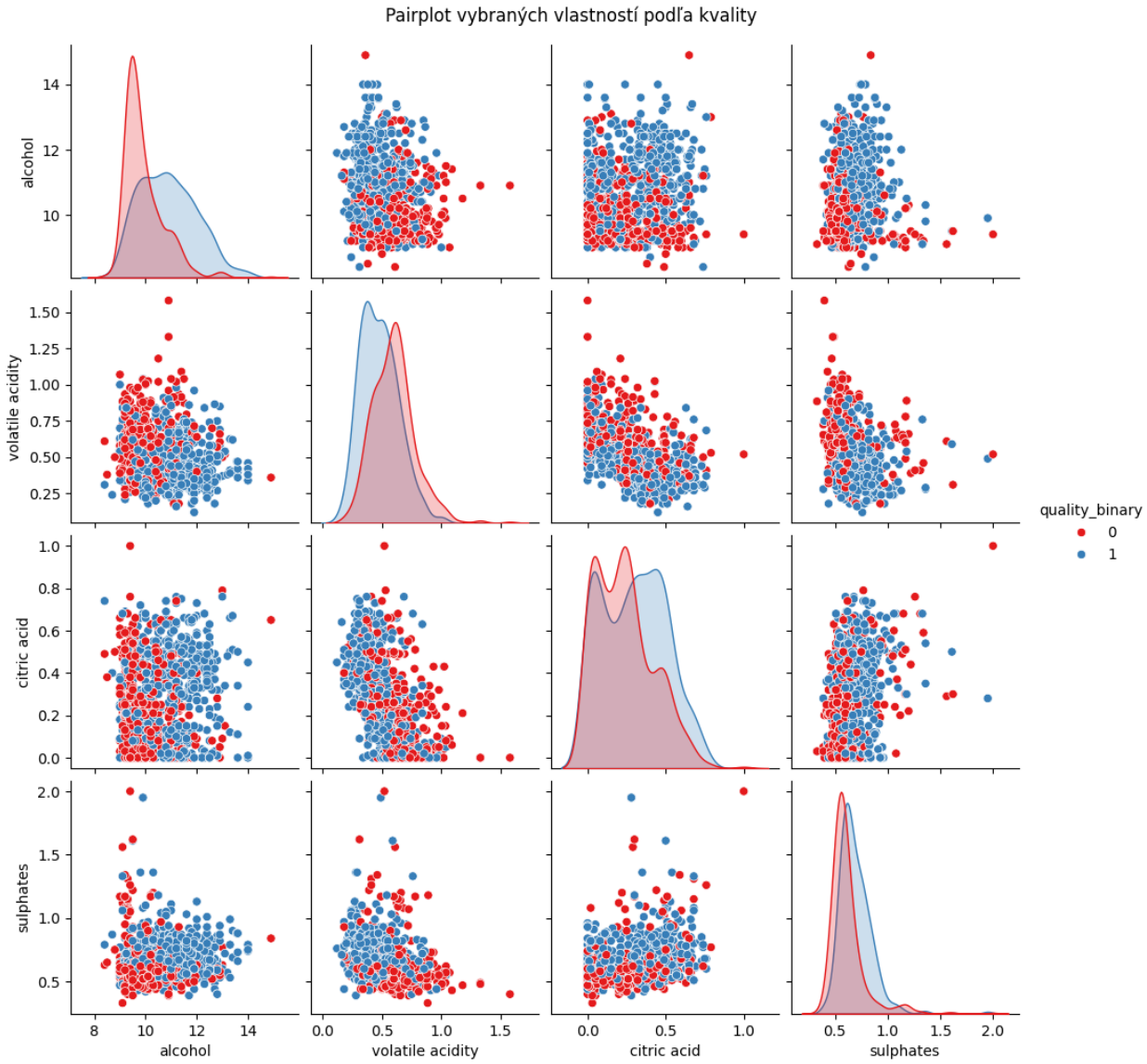
Boxploty – porovnanie vlastností podľa kvality (binárne)

Boxploty ukazujú rozloženie hodnôt jednotlivých vlastností v závislosti od kvality vína. V prípade alkoholu je zrejmé, že vyššia kvalita vína je spojená s vyšším obsahom alkoholu. Tento vzťah naznačuje, že alkohol je významným prediktorom kvality.



Pairplot – vzťahy medzi viacerými premennými + farebne podľa kvality

Párové grafy zobrazujú vzťahy medzi vybranými dvojicami premenných. Umožňujú vizuálne identifikovať zhluky alebo vzory, ktoré by mohli naznačovať rozdiely medzi triedami kvality.



4. Model

Použitý model: **Logistická regresia**

- Cieľ: predikovať pravdepodobnosť, že víno bude nadpriemernej kvality.
- Vstupné premenné: všetky numerické vlastnosti vína okrem samotnej kvality.

5. Výsledky modelu

Presnosť:

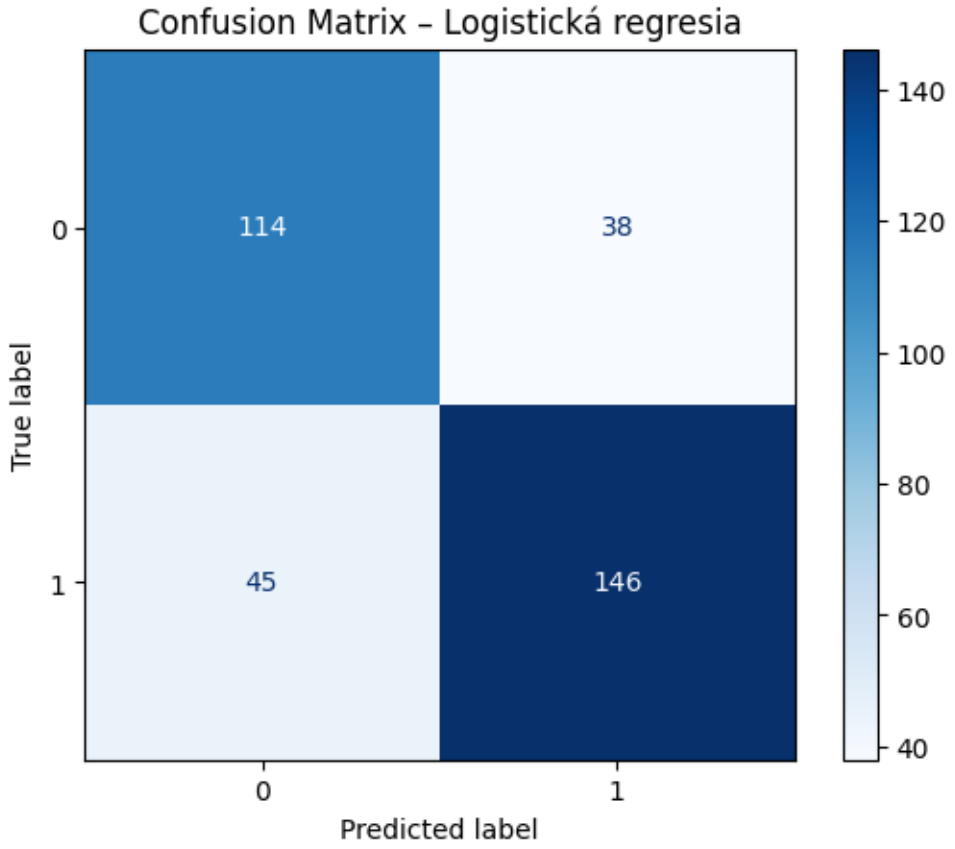
- **Accuracy:** 75.8 %

Confusion matrix:

[[114 38] <- 0 **(podpriemerné)**

 [45 146]] <- 1 **(nadpriemerné)**

Zobrazuje počet správne a nesprávne klasifikovaných vzoriek. Väčšina vín bola správne rozpoznaná ako podpriemerné (0) alebo nadpriemerné (1), s vyváženým počtom chýb medzi triedami.

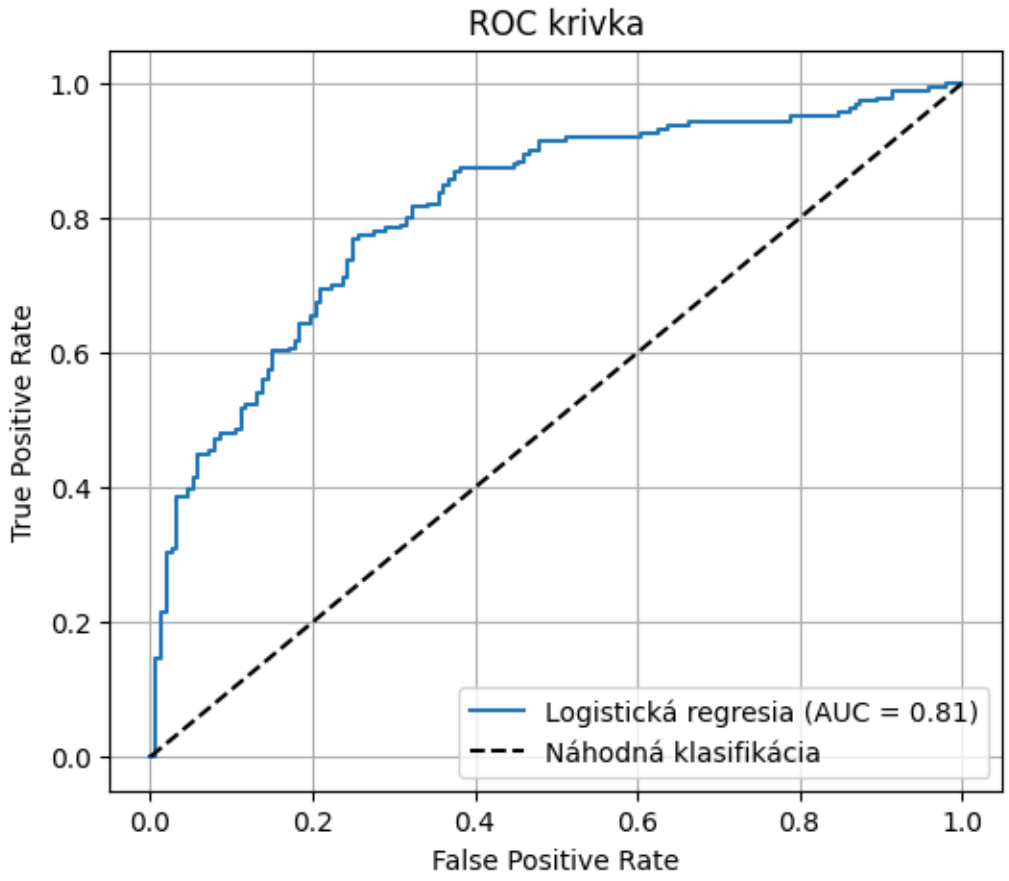


Klasifikačná správa:

Trieda	**Presnosť (Precision)**	**Recall**	**F1-score**	**Podpora (Support)**
0 – podpriemerná	0.72	0.75	0.73	152
1 – nadpriemerná	0.79	0.76	0.78	191
Priemer	**0.76**	**0.76**	**0.76**	**343**

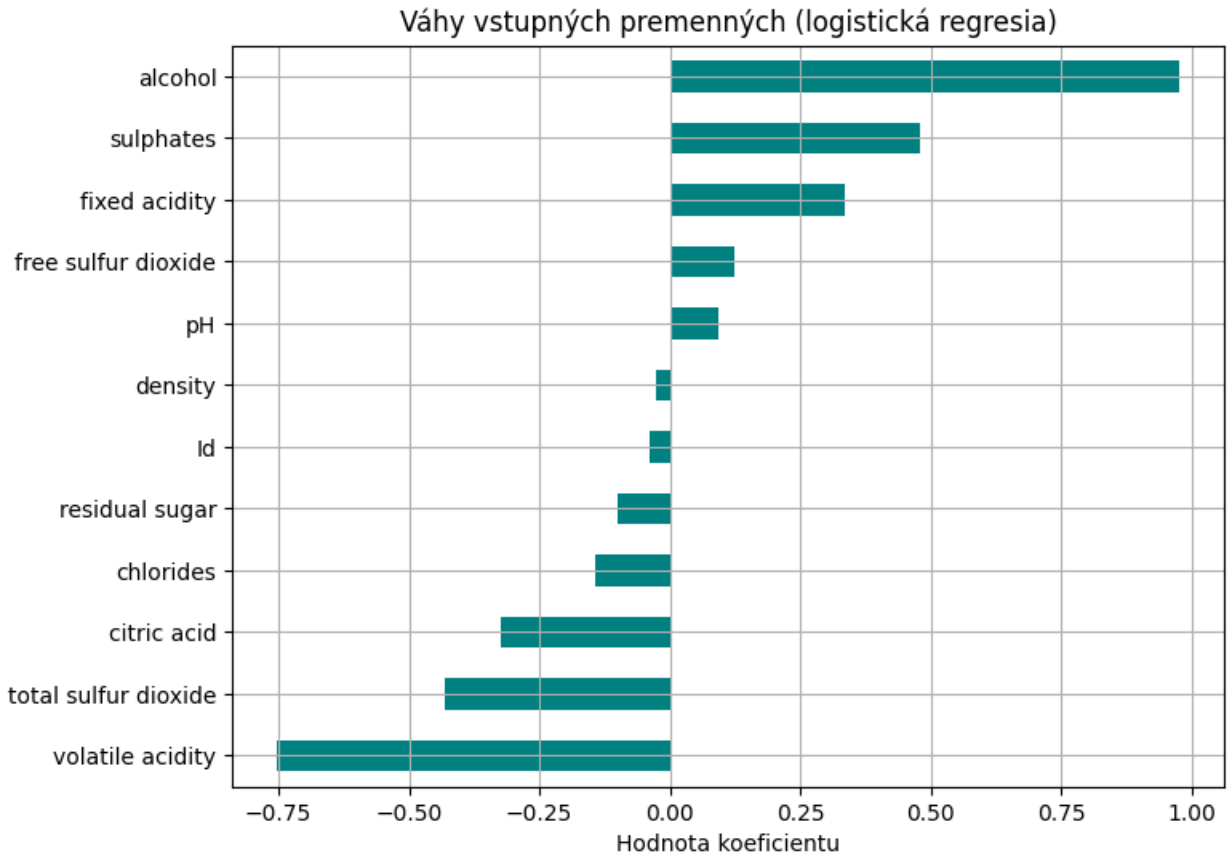
ROC krivka – Výkonnosť klasifikátora

ROC krivka znázorňuje vzťah medzi mierou False Positive Rate a True Positive Rate pri rôznych prahoch rozhodovania. Model dosiahol AUC skóre približne 0.84, čo indikuje veľmi dobrú schopnosť rozlišovať medzi triedami.



Váhy vstupných premenných – Interpretácia modelu

Graf zobrazuje koeficienty logistickej regresie pre jednotlivé vstupné premenné. Premenné s vyššou absolútnou hodnotou koeficientu majú väčší vplyv na predikciu. Napríklad, vyšší obsah alkoholu zvyšuje pravdepodobnosť nadpriemernej kvality vína, zatiaľ čo vyššia volatilná kyslosť túto pravdepodobnosť znižuje.



6. Zhrnutie a odporúčania

V tomto projekte sme sa zamerali na predikciu kvality vína pomocou logistickej regresie. Kvalita bola rozdelená na dve triedy – *podpriemerná* a *nadpriemerná* – podľa priemernej hodnoty kvality v datasete. Po predspracovaní dát a základnej exploračnej analýze sme vytvorili klasifikačný model.

Model dosiahol nasledovné výsledky:

- **Presnosť:** 75.8 %
- **F1-skóre:** 0.76 (priemer)
- **ROC AUC:** ~0.84
- **Matica zámien:** vyvážené chyby medzi triedami

Na základe klasifikačnej správy a ROC krivky možno povedať, že model má solídnu schopnosť rozlišovať medzi vínami rôznej kvality. Významný pozitívny vplyv na predikciu má najmä **alkohol**, zatiaľ čo **volatile acidity** má negatívny dopad.

7. Odporúčania

- **Rozšíriť model o ďalšie algoritmy**: napr. rozhodovacie stromy, Random forest alebo gradient boosting pre porovnanie výkonu.
- **Vyskúšať vyváženie tried** (napr. pomocou SMOTE alebo class weight), ak by sa vyskytla nevyváženosť medzi triedami.
- **Zvážiť štandardizáciu premenných**, najmä pri použití algoritmov citlivých na mierku.
- **Zvýšiť interpretovateľnosť** modelu pomocou metód ako SHAP alebo LIME.
- **Získať viac údajov** alebo doplniť dataset o ďalšie premenné (napr. krajina pôvodu, odroda hrozna), čo môže zvýšiť predikčnú schopnosť.