VYSOKÁ ŠKOLA POLYTECHNICKÁ JIHLAVA

Aplikovaná informatika

DATOVÁ VĚDA: PREDIKCE KVALITY VÍNA

Bakalářská práce

Autor práce: Pavel Plíhal

Vedoucí práce: doc. Dr. Ing. Jan Voráček, CSc.

Jihlava 2026

Vysoká škola polytechnická Jihlava

Tolstého 16, 586 01 Jihlava

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Autor práce:	**Pavel Plíhal**
Studijní program:	Aplikovaná informatika
Garant studijního programu:	Ing. Lenka Kuklišová Pavelková, Ph.D.
Název práce:	**Datová věda: predikce kvality vína**
Vedoucí práce:	doc. Dr. Ing. Jan Voráček, CSc.
Cíl práce:	Cílem práce je aplikace úplného životního cyklu datové vědy na problematiku predikce vnímané kvality vína. Půjde při ní zejména o nalezení nejvýznamnějších degustačních a chemických analýz, predikujících senzoricky hodnocené vlastnosti analyzovaných vzorků. Naplnění cíle zahrnuje identifikaci problému, rešerši souvisejících zdrojů, návrh a optimalizaci vhodných experimentů, čištění a konsolidaci relevantních dat, jejich explorativní analýzy a induktivní modelování, srozumitelnou vizualizaci výsledků a diskusi možných řešení definovaného problému na základě shromážděných, vyhodnocených a řádně diskutovaných podkladů. Základní datový korpus bude převzat z veřejných zdrojů a student ho může dle vlastních úvah dále modifikovat. Experimenty budou realizovány a distribuovány formou interaktivního online notebooku, využívajícího problémově orientované knihovny jazyka Python. Při vývoji finálního řešení bude využita agilní metodika Rapid Application Development s podporou umělé inteligence.

Abstrakt

Bakalářská práce se zabývá predikcí kvality vína na základě fyzikálně-chemických parametrů s využitím metod strojového učení. Cílem bylo navrhnout systematický analytický postup zahrnující přípravu dat, explorativní analýzu a porovnání vybraných modelů. Byly aplikovány klasifikační i regresní přístupy a analyzován vliv struktury datasetu, zejména nevyváženosti tříd a překryvu mezi kategoriemi kvality. Výsledky ukazují, že kvalitu vína lze predikovat, avšak přesnost modelů je limitována charakterem dat. Nejlepších výsledků dosahují nelineární modely, zejména Random Forest, díky schopnosti zachytit komplexní vztahy mezi proměnnými, zatímco lineární modely se ukázaly jako méně vhodné. Experimenty dále ukázaly, že vyvažování tříd a redukce atributů nepředstavují univerzální řešení a mohou vést ke zhoršení výkonu modelu.

Klíčová slova

predikce kvality vína, strojové učení, klasifikace, regrese, nevyváženost tříd, výběr atributů, Random Forest

Abstract

This bachelor's thesis focuses on the prediction of wine quality based on physicochemical parameters using machine learning methods. The aim was to design a systematic analytical workflow including data preprocessing, exploratory data analysis, and comparison of selected models. Both classification and regression approaches were applied, and the impact of dataset structure, particularly class imbalance and overlap between quality categories, was analyzed. The results show that wine quality can be predicted; however, model performance is limited by the nature of the data. The best results were achieved by nonlinear models, especially Random Forest, due to their ability to capture complex relationships between variables, while linear models proved less suitable. Additional experiments showed that class balancing and feature reduction are not universal solutions and may even lead to a decrease in model performance.

Keywords

wine quality prediction; machine learning; classification; regression; class imbalance; feature selection; Random Forest

Prohlašuji, že předložená bakalářská práce je původní a zpracoval jsem ji samostatně. Prohlašuji, že citace použitých pramenů je úplná, že jsem v práci neporušil autorská práva (ve smyslu zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů, v platném znění, dále též „**AZ**“).

Byl jsem seznámen s tím, že na mou bakalářskou práci se plně vztahuje **AZ**, zejména § 60 (školní dílo).

Podle § 47b zákona o vysokých školách souhlasím se zveřejněním své práce podle Směrnice pro vedení, vypracování a zveřejňování závěrečných prací na VŠPJ, a to bez ohledu na výsledek obhajoby.

Beru na vědomí, že VŠPJ má právo na uzavření licenční smlouvy o užití mé bakalářské práce a prohlašuji, že **s o u h l a s í m** s případným užitím mé bakalářské práce (prodej, zapůjčení apod.).

Jsem si vědom toho, že užít své bakalářské práce či poskytnout licenci k jejímu využití mohu jen se souhlasem VŠPJ, která má právo ode mě požadovat přiměřený příspěvek na úhradu nákladů, vynaložených vysokou školou na vytvoření díla (až do jejich skutečné výše), z výdělku dosaženého v souvislosti s užitím díla či poskytnutím licence.

V Jihlavě dne 10. dubna 2026

…….

Podpis studenta

Poděkování

Rád bych poděkoval vedoucímu své bakalářské práce za odborné vedení, cenné rady a připomínky, které významně přispěly ke zpracování této práce. Poděkování patří také mé rodině za podporu během celého studia.

Obsah

Seznam obrázků

Seznam tabulek

Seznam zkratek

AI	Artificial Intelligence
AUC	Area Under Curve
CRISP-DM	Cross-Industry Standard Process for Data Mining
EDA	Exploratory Data Analysis
FN	False Negative
FP	False Positive
FPR	False Positive Rate
KNN	K-Nearest Neighbours
LDA	Linear Discriminant Analysis
MAE	Mean absolute error
PCA	Principal Component Analysis
PCE	Principal Component Analysis
Q-Q graf	Quantile-Quantil plot
R^2	Koeficient Determinace
RAD	Rapid Application Development
RBF kernel	Radial Basis Function kernel
RFECV	Recursive Feature Elimination with Cross-Validation
RMSE	Root Mean Squared Error
SVM	Support Vector Machines
t-DSNE	t-Distributed Stochastic Neighbor Embedding
TN	True Negative
TP	True Positive
VIF	Variance Inflation Factor

Úvod

V posledních letech dochází k dynamickému rozvoji datové vědy, která se stala významným nástrojem pro analýzu, modelování a predikci jevů v mnoha oblastech lidské činnosti. V oblasti potravinářství a enologie nachází datová věda uplatnění zejména při zkoumání faktorů ovlivňujících kvalitu vína a při snaze vztahy popsat či předpovědět pomocí vhodných modelů. Predikce kvality vína představuje prakticky významný problém, neboť propojuje chemické analýzy s lidským senzorickým hodnocením, které je ze své podstaty částečně subjektivní.

Práce se zabývá aplikací metod datové vědy na problematiku predikce vnímané kvality vína. Zaměřuje se na identifikaci významných fyzikálně-chemických vlastností, které ovlivňují výsledné hodnocení, a na návrh a vyhodnocení modelů strojového učení schopných tuto kvalitu predikovat. Současně je cílem práce aplikovat kompletní životní cyklus datové vědy, zahrnující přípravu dat, explorativní analýzu, modelování, evaluaci a interpretaci výsledků.

Pro analýzu byl využit dataset pocházející z veřejně dostupných zdrojů, obsahující fyzikálně-chemické vlastnosti vín a jejich senzorické hodnocení. Data byla podrobena procesu čištění, transformace a explorativní analýzy s cílem odhalit jejich strukturu, identifikovat vztahy mezi proměnnými a připravit je pro následné modelování.

V metodické části byly aplikovány vybrané metody strojového učení implementované v jazyce Python s využitím knihoven pandas, scikit-learn a dalších nástrojů pro analýzu a vizualizaci dat. Experimenty byly realizovány v interaktivním prostředí notebooku, které umožňuje přehlednou prezentaci analytického postupu i dosažených výsledků. Součástí práce je také experimentální ověření vlivu vybraných faktorů, jako je nevyváženost dat nebo redukce vstupních atributů, na výkon modelů.

Přínosem práce je praktická demonstrace kompletního analytického procesu datové vědy aplikovaného na reálný problém predikce kvality vína. Práce zároveň přispívá k lepšímu pochopení vztahů mezi fyzikálně-chemickými vlastnostmi vína a jeho výsledným hodnocením a poukazuje na omezení i možnosti využití metod strojového učení.

1 Teoretická část

Teoretická část vymezuje odborný rámec práce a představuje klíčové koncepty, na nichž jsou založené následující části. Zaměřuje se na charakteristiku zkoumaného problému, metodické přístupy využívané v datové vědě, základní principy prediktivního modelování a na přehled vybraných studií zabývajících se predikcí kvality vína. Současně shrnuje hlavní techniky, postupy a faktory, které ovlivňují jak samotnou kvalitu vína, tak i výslednou úspěšnost predikčních modelů.

1.1 Charakteristika a kontext zkoumaného problému

Predikce kvality vína představuje mezioborový problém na pomezí potravinářství, chemické analýzy a datové vědy. Tradiční hodnocení kvality vína je zpravidla založeno na senzorickém posuzování prováděném vyškolenými hodnotiteli. Daný postup je však časově náročný, finančně zatěžující a do určité míry také subjektivní. Alternativní možnost proto představuje využití laboratorně měřených fyzikálně-chemických parametrů, na jejichž základě lze pomocí statistických a prediktivních metod odhadovat výsledné hodnocení kvality.

Zájem o tuto problematiku v posledních letech významně vzrostl také díky dostupnosti veřejných datových sad, které obsahují údaje o červených a bílých vínech spolu s vybranými fyzikálně-chemickými charakteristikami, například kyselostí, obsahem zbytkového cukru, alkoholu, chloridů nebo oxidu siřičitého. Proměnné jsou doplněny hodnotou senzoricky určené kvality. Dataset je často využíván jako referenční úloha při studiu metod strojového učení, a nejen díky své dostupnosti, ale také kvůli relativně komplexním a částečně nelineárním vztahům mezi vstupními proměnnými a cílovou proměnnou (Cortez a kol., 2009).

Z pohledu datové vědy je hlavní výzkumný problém spojen s otázkou, zda lze kvalitu vína spolehlivě predikovat na základě měřitelných fyzikálně-chemických atributů a které modelovací přístupy poskytují nejlepší výsledky. Významnou součástí problému je také vhodné předzpracování dat, volba odpovídajících validačních postupů a interpretace výsledků modelů. V praktickém kontextu je totiž žádoucí nejen dosáhnout co nejvyšší predikční přesnosti, ale zároveň identifikovat atributy, které mají na výsledné hodnocení kvality vína největší vliv.

1.2 Cíl a výzkumné otázky

Cílem práce je analyzovat možnosti predikce kvality vína na základě fyzikálně-chemických parametrů získaných laboratorním měřením. Práce se zaměřuje na vytvoření systematického analytického postupu zahrnujícího přípravu a předzpracování dat, výběr vhodných modelovacích přístupů strojového učení, jejich vyhodnocení pomocí odpovídajících metrik a interpretaci získaných výsledků. Součástí práce je také identifikace atributů, které mají největší vliv na výsledné hodnocení kvality vína.

Za účelem naplnění uvedeného cíle jsou formulovány následující výzkumné otázky:

• Lze kvalitu vína spolehlivě predikovat na základě laboratorně měřených fyzikálně-chemických parametrů?

• Které modely strojového učení dosahují v dané úloze nejlepších výsledků a za jakých podmínek?

• Jak významně přispívají jednotlivé fyzikálně-chemické vlastnosti k predikci výsledného hodnocení kvality vína?

• Jak ovlivňuje nerovnoměrné zastoupení jednotlivých tříd kvality výkon klasifikačních modelů?

• Zlepší se stabilita a přesnost klasifikačního modelu při omezení úlohy na nejčastější třídy kvality?

• Jaký dopad má redukce vstupních atributů na výkon regresních modelů?

1.3 Datová věda a její charakteristika

Datová věda představuje mezioborovou disciplínu, která kombinuje metody statistiky, informatiky a matematického modelování s cílem získávat nové poznatky z dat. Zaměřuje se na systematickou práci s daty v celém jejich životním cyklu, od jejich sběru a správy přes analýzu a modelování až po interpretaci výsledků a jejich praktické využití. Datová věda tak patří mezi klíčové oblasti moderních analytických disciplín a nachází uplatnění v řadě oborů, například ve zdravotnictví, financích, průmyslu, marketingu nebo potravinářství (Zhang, 2023).

Typickým rysem datové vědy je propojení různých analytických přístupů, které umožňují efektivně pracovat s daty různého typu, rozsahu a kvality. Mezi základní nástroje datové vědy patří statistická analýza, metody strojového učení, vizualizace dat a databázové technologie. V praktických aplikacích se často využívají programovací jazyky, jako jsou Python nebo R, které poskytují rozsáhlé knihovny pro práci s daty, jejich analýzu a modelování. Současná podoba datové vědy vznikla jako reakce na rychlý nárůst objemu dostupných dat a na potřebu systematických metod, které umožňují data analyzovat a interpretovat. Významnou roli přitom sehrál rozvoj výpočetních technologií a algoritmů strojového učení, které umožňují automatizované zpracování rozsáhlých datových souborů. Datová věda tak představuje propojení teoretických přístupů a praktických nástrojů, jejichž cílem je transformovat surová data na informace s reálnou užitnou hodnotou (Shahnawaz a Kumar, 2025).

V kontextu práce slouží datová věda jako metodický rámec pro analýzu fyzikálně-chemických parametrů vína a pro tvorbu prediktivních modelů, které umožňují odhadovat jeho kvalitu. Metody datové vědy tak umožňují nejen identifikovat vztahy mezi jednotlivými proměnnými, ale také vyhodnotit spolehlivost modelů a interpretovat faktory, které mají na výsledné hodnocení kvality vína největší vliv.

1.4 Životní cyklus (RAD + CRISP-DM)

Životní cyklus datově orientovaného projektu představuje strukturovaný proces, který umožňuje systematické zpracování dat a postupné vytváření analytického řešení. Proces zpravidla začíná definicí problému a pokračuje přes analýzu dat, modelování, vyhodnocení výsledků až po jejich interpretaci a případné praktické využití. Jednotlivé kroky na sebe logicky navazují a zároveň umožňují průběžné ověřování správnosti zvoleného postupu.

V oblasti datové vědy existuje několik metodických rámců, které proces popisují. Mezi nejrozšířenější patří model Cross-Industry Standard Process for Data Mining (CRISP-DM), přístup Rapid Application Developmen (RAD) a rozšíření RAD o umělou inteligenci. Oba přístupy zdůrazňují iterativní charakter práce s daty a možnost průběžně upravovat analytický postup podle získaných výsledků (Chapman a kol., 2000).

1.4.1 CRISP-DM

CRISP-DM patří mezi nejrozšířenější metodické rámce pro řízení projektů zaměřených na datovou analýzu a data mining. Jeho popularita vychází především z univerzálnosti, přehledné struktury a z provázanosti jednotlivých fází. Model je rozdělen do šesti základních etap, mezi nimiž se projekt může iterativně vracet v závislosti na průběžných výsledcích analýzy (Casonatto, Souza, Mariano, 2024).

První fáze je zaměřena na porozumění problému a definování hlavních cílů projektu. Zahrnuje identifikaci potřeb zadavatele a stanovení kritérií úspěchu. Správné vymezení problému je zásadní, protože určuje směr celého analytického procesu. Druhá fáze, označovaná jako porozumění datům, zahrnuje sběr dostupných dat, jejich základní prozkoumání a kontrolu kvality. Jsou identifikovány problémy, jako jsou chybějící hodnoty, odlehlá pozorování nebo nesoulad mezi strukturou dat a cíli projektu. Třetí fáze se věnuje přípravě dat, která bývá v praxi časově nejnáročnější částí celého projektu. Zahrnuje čištění dat, transformace proměnných, škálování nebo výběr relevantních atributů. Kvalita provedení má zásadní vliv na výsledný výkon modelů. Ve čtvrté fázi dochází k modelování. Analytik vybírá vhodné statistické a strojově-učící metody odpovídající povaze problému, například regresi, klasifikaci nebo shlukování. Součástí je také trénování modelů, ladění jejich parametrů a optimalizace výkonu. Pátá fáze představuje vyhodnocení modelů, kdy jsou výsledky analyzovány pomocí vhodných metrik a posuzuje se, zda model splňuje původně stanovené cíle. V případě nedostatečných výsledků může být nutné vrátit se k předchozím krokům a upravit například přípravu dat nebo výběr modelu. Poslední fáze zahrnuje nasazení výsledného řešení, které může mít podobu implementace modelu do aplikace, vytvoření analytických nástrojů nebo prezentace získaných poznatků. I po nasazení může projekt pokračovat dalšími iteracemi, a model dále rozvíjet nebo přizpůsobit novým datům (Staneviciene a kol., 2024).

CRISP-DM klade důraz na flexibilitu a opakovatelnost analytického procesu. Iterativní charakter modelu umožňuje průběžně upravovat analytický postup podle nově získaných poznatků, což z něj činí vhodný rámec pro řešení komplexních datových úloh (Chapman a kol., 2000).

1.4.2 Rapid Application Development

Metodika RAD představuje přístup k vývoji informačních systémů, který zdůrazňuje rychlost vývoje, flexibilitu a průběžnou zpětnou vazbu od uživatelů. Na rozdíl od tradičních sekvenčních metodik, které předpokládají lineární postup jednotlivých fází vývoje, RAD pracuje s krátkými vývojovými cykly a iterativním prototypováním. Základním principem metodiky je vytváření prototypů, tedy zjednodušených verzí budoucího systému. Prototypy jsou následně postupně upravovány na základě zpětné vazby uživatelů, což umožňuje průběžně zpřesňovat funkcionalitu i samotné cíle projektu. RAD umožňuje snížit riziko, že výsledné řešení nebude odpovídat skutečným potřebám uživatelů, a zároveň zkracuje dobu mezi návrhem a vznikem prvních

funkčních výsledků. Iterativní charakter RAD je velmi vhodný také pro projekty v oblasti datové vědy. Analytik může průběžně experimentovat s různými modely strojového učení, testovat jejich parametry, porovnávat alternativní algoritmy a vyhodnocovat dopady různých přístupů k předzpracování dat. Krátké vývojové cykly tak umožňují rychle reagovat na nové poznatky získané během analýzy dat (Mackay, 1999).

V současné době dochází k významnému propojení metodiky Rapid Application Development s oblastí artificial intelligence (AI), což dále posiluje její význam v moderním vývoji softwaru a datových aplikací. Nástroje založené na umělé inteligenci, zejména generativní modely, umožňují automatizaci řady činností v rámci vývoje aplikací, jako je generování zdrojového kódu, návrh uživatelského rozhraní nebo testování. Následkem dochází k dalšímu zrychlení vývojového cyklu, který je pro RAD klíčový (Hartman a kol., 2025).

Z hlediska vývoje datově orientovaných aplikací a systémů strojového učení přináší propojení RAD a AI také nový výzkumný směr označovaný jako „application-driven AI“ (RAD-AI). Klade důraz na vývoj AI systémů, které by měly být úzce spojeny s konkrétní aplikační doménou a potřebami uživatelů. Klíčovou roli zde hraje iterativní testování, zpětná vazba od uživatelů a postupné zlepšování modelů v reálném prostředí. RAD-AI přístup navrhuje tři hlavní fáze: první představuje tvorbu interdisciplinárních týmů a analýzu potřeb uživatelů, fáze druhá návrh metod a metrik přizpůsobených konkrétnímu kontextu a třetí se zabývá ověřováním funkčnosti a dlouhodobé udržitelnosti řešení v praxi. Cyklický proces je plně kompatibilní s principy RAD, zejména s důrazem na iterativní vývoj a průběžnou validaci řešení. Integrace umělé inteligence do vývojového procesu tak nejen urychluje samotný vývoj aplikací, ale zároveň umožňuje vytvářet sofistikovanější a adaptivnější systémy. V kontextu datové vědy, znamená například rychlejší experimentování s modely, automatizaci ladění hyperparametrů nebo efektivnější práci s daty. Kombinace RAD a AI tak představuje moderní přístup, který reflektuje současné požadavky na flexibilitu, rychlost a schopnost reagovat na dynamicky se měnící prostředí (Hartman a kol., 2025).

1.5 Techniky a nástroje

Datová věda využívá široké spektrum technik a nástrojů, které umožňují efektivní práci s daty v celém jejich životním cyklu. Techniky zahrnují postupy pro sběr, přípravu, analýzu, modelování i interpretaci dat a společně tvoří základ analytických procesů. Volba konkrétních metod závisí na charakteru řešeného problému, typu dostupných dat i požadavcích na přesnost, interpretovatelnost a reprodukovatelnost výsledků (Shahnawaz a Kumar, 2025).

V rámci datově orientovaných úloh hraje důležitou roli zejména předzpracování dat, explorativní analýza dat, analýza vztahů mezi proměnnými, výběr a transformace atributů, tvorba prediktivních modelů a jejich následné vyhodnocení. V případě predikce kvality vína, zmíněné techniky umožňují systematicky zkoumat vztahy mezi fyzikálně-chemickými parametry a výsledným hodnocením kvality, identifikovat důležité atributy a porovnávat výkon různých modelovacích přístupů.

Vedle samotných analytických metod je významná také volba softwarového prostředí. V oblasti datové vědy patří mezi nejrozšířenější programovací jazyky Python a R. Python je široce využíván díky rozsáhlému ekosystému knihoven, například NumPy, pandas, scikit-learn nebo Matplotlib,

které poskytují nástroje pro práci s daty, statistickou analýzu, vizualizaci i modelování. V experimentální praxi se často používají interaktivní prostředí typu Jupyter Notebook nebo cloudové platformy, jako je Kaggle, které podporují iterativní vývoj a reprodukovatelnost analytických postupů (Kroese a kol., 2024).

1.5.1 Kvalita dat a její vliv na volbu analytických technik

Kvalita dat představuje jeden z rozhodujících faktorů ovlivňujících výsledky datově orientovaných analýz i stabilitu prediktivních modelů. V odborné literatuře jsou mezi základní dimenze kvality dat obvykle řazeny úplnost, přesnost, konzistence, aktuálnost a relevance. Dimenze určují, do jaké míry jsou data vhodná pro konkrétní analytický úkol, a současně ovlivňují výběr vhodných technik pro jejich zpracování (Batini a kol., 2009).

Při nízké kvalitě dat, například v případě vysokého počtu chybějících hodnot, nekonzistentních formátů nebo nepřesných záznamů, dochází ke snížení spolehlivosti výstupů i jejich interpretovatelnosti. Nekvalitní data mohou vést k chybným závěrům, nevhodné volbě modelů a ke zhoršení schopnosti modelů generalizovat na nová data. Naopak kvalitní datový základ umožňuje efektivnější využití i pokročilejších analytických metod a zvyšuje důvěryhodnost získaných výsledků (Gong a kol., 2023).

Kvalita dat má přímý dopad také na výkon konkrétních analytických technik. Některé modely jsou citlivé na šum, odlehlé hodnoty nebo chybějící data více než jiné. Proto je nezbytné věnovat dostatečnou pozornost předzpracování dat ještě před samotnou fází modelování. Právě systematické posouzení kvality dat a odpovídající volba analytických metod představují důležitý předpoklad úspěšného řešení datově orientovaných úloh (Wang a kol., 2023).

1.5.2 Standardizace a normalizace dat

Při práci s numerickými daty je často nutné provést jejich transformaci tak, aby byly jednotlivé proměnné vzájemně srovnatelné z hlediska měřítka. Mezi nejčastěji používané transformační techniky patří standardizace a normalizace. Standardizace převádí hodnoty proměnné na hodnoty s nulovým průměrem a jednotkovou směrodatnou odchylkou. Transformace je definována vztahem

$$z = \frac{(x - \mu)}{\sigma} \tag{1}$$

kde x je původní hodnota, μ průměr proměnné a σ směrodatná odchylka. Normalizace zase převádí hodnoty do pevně daného intervalu, nejčastěji ⟨0,1⟩, podle vztahu

$$x' = \frac{(x - x_{min})}{(x_{max} - x_{min})} \tag{2}$$

Kde x je původní hodnota proměnné, x' je normalizovaná hodnota, x_{min} je nejmenší hodnota proměnné v datasetu a x_{max} je největší hodnota proměnné v datasetu. Rozdíl mezi přístupy spočívá v tom, že standardizace zachovává tvar rozdělení dat, zatímco normalizace omezuje rozsah hodnot na pevný interval. V případě přítomnosti odlehlých hodnot může normalizace způsobovat, že většina dat bude soustředěna do úzkého intervalu hodnot (Brijith, 2023).

Volba vhodné transformace závisí na charakteru dat a na použitém modelu. Standardizace je často preferována u metod citlivých na měřítko vstupních proměnných, například u algoritmů založených na vzdálenosti (K-Nearest Neighbours), metod analýzy hlavních komponent (Principal Component Analysis, PCA), Support Vector Machines (SVM) nebo lineárních modelů s regularizací.

1.5.3 Explorativní analýza dat a vztahy mezi proměnnými

Explorativní analýza dat (Exploratory Data Analysis, EDA) představuje důležitou fázi analytického procesu, jejímž cílem je získat základní i hlubší porozumění datové sadě ještě před samotným modelováním. EDA umožňuje identifikovat strukturu dat, rozdělení jednotlivých proměnných, jejich variabilitu, přítomnost odlehlých hodnot, případné nesrovnalosti v datech a vztahy mezi vstupními atributy i mezi atributy a cílovou proměnnou (Downey, 2025; Victoria a kol., 2023).

Základním krokem explorativní analýzy je popisná statistika. Poskytuje kvantitativní přehled o jednotlivých proměnných a umožňuje rychle posoudit jejich rozsah, střední hodnotu a míru variability. Mezi nejčastěji používané ukazatele patří počet pozorování, aritmetický průměr, medián, minimum, maximum, směrodatná odchylka a kvartily.

Důležitou součástí EDA je také analýza rozdělení proměnných. K danému účelu se využívají zejména histogramy, boxploty, houslové grafy a bodové grafy. Histogramy umožňují vizuálně posoudit tvar rozdělení proměnných, například zda je rozdělení symetrické, pravostranně nebo levostranně zešikmené, případně zda obsahuje více vrcholů. Boxploty poskytují stručné grafické shrnutí rozdělení, včetně mediánu, kvartilů a možných odlehlých hodnot. Bodové grafy pak umožňují posuzovat vztahy mezi dvojicemi proměnných a odhalovat případné lineární či nelineární závislosti. V rámci explorativní analýzy dat je důležité posoudit také tvar rozdělení jednotlivých proměnných, zejména s ohledem na předpoklady některých statistických a modelovacích metod. Proto se využívají jak vizuální, tak statistické přístupy. Posouzení rozdělení proměnných má přímý vliv na volbu transformačních technik a modelů. V případě odchylek od normálního rozdělení je vhodné uvažovat o použití transformací nebo modelů, které nejsou na normalitu citlivé (Komorowski a kol., 2016).

Mezi vizuální metody patří zejména quantile-quantile plots (Q-Q grafy), které porovnávají empirické kvantily dat s kvantily teoretického normálního rozdělení. Pokud se body v grafu nacházejí přibližně na přímce, lze usuzovat na shodu s normálním rozdělením. Odchylky od přímky signalizují asymetrii, šikmost nebo přítomnost odlehlých hodnot. Pro formální testování normality rozdělení se používají statistické testy, mezi které patří například D'Agostino K^2 test a Anderson–Darling test. D'Agostino K^2 test kombinuje informace o šikmosti a špičatosti rozdělení a ověřuje, zda se významně liší od normálního rozdělení. Anderson–Darling test klade větší důraz na rozdíly v krajních částech rozdělení, a je proto citlivější na odchylky v ocasech rozdělení. Výsledkem testů je obvykle p-hodnota, na jejímž základě lze rozhodnout, zda lze nulovou hypotézu o normalitě rozdělení zamítnout. Pokud je p-hodnota menší než zvolená hladina významnosti, obvykle 0,05, je normalita rozdělení zamítnuta (Ghasemi a Zahediasl, 2012).

V rámci explorativní analýzy je významná také korelační analýza, která slouží ke zkoumání síly a směru vztahu mezi dvojicemi numerických proměnných. Nejčastěji se používá Pearsonův korelační koeficient. Hodnota korelačního koeficientu se pohybuje v intervalu od -1 do 1.

Hodnoty blízké 1 značí silnou kladnou lineární závislost, hodnoty blízké -1 silnou zápornou lineární závislost a hodnoty blízké 0 naznačují absenci lineární závislosti. Korelační analýza je užitečná nejen pro interpretaci dat, ale také pro odhalení redundantních proměnných a pro metodické zdůvodnění dalších kroků, například redukce dimenzionality nebo analýzy multikolinearity (Fu, Tang a Zhao, 2025).

Na korelační analýzu navazuje problematika multikolinearity. Multikolinearita nastává tehdy, když je některá vstupní proměnná silně lineárně závislá na ostatních vstupních proměnných. Jev může komplikovat interpretaci, zejména modelů, které pracují s explicitními koeficienty, a může vést ke zvýšení nestability odhadů parametrů. Pro kvantifikaci multikolinearity se používá ukazatel Variance Inflation Factor (VIF), který lze vyjádřit vztahem

$$VIF_i = \frac{1}{\left(1 - R_i^2\right)} \tag{3}$$

Kde koeficient determinace (R_i^2) představuje pomocné regrese, v níž je i-tá proměnná vysvětlována všemi ostatními vstupními proměnnými. Vyšší hodnota VIF určuje silnější závislost dané proměnné na ostatních proměnných v modelu. Za orientační pravidlo se obvykle považuje, že hodnota VIF blízká 1 značí téměř nezávislou proměnnou, hodnoty menší než 5 ukazují nízkou multikolinearitu, hodnoty mezi 5 a 10 naznačují střední závislost a hodnoty vyšší než 10 mohou signalizovat silnou redundanci informace. VIF analýzu lze využít především jako doplňující nástroj ke korelační analýze a k metodickému zdůvodnění použití redukce dimenzionality. Výsledky korelační analýzy a VIF jsou využívány nejen pro interpretaci vztahů mezi proměnnými, ale také jako podklad pro rozhodování o případné redukci dimenzionality a výběru vhodných modelovacích metod (Kim, 2019; Kuo, Xu a Friesen, 2025).

Po identifikaci možných problémů s multikolinearitou pomocí ukazatele VIF je vhodné zvážit techniky, které umožňují redukovat rozměr dat a odstranit redundantní informaci mezi jednotlivými proměnnými. Jednou z nejčastěji používaných metod je analýza hlavních komponent (Principal Component Analysis, PCA). PCA převádí původní sadu vzájemně korelovaných proměnných do nového prostoru tvořeného menším počtem vzájemně nekorelovaných komponent, přičemž každá nově vzniklá komponenta zachycuje určitou část variability původních dat. Komponenty vznikají jako lineární kombinace původních proměnných, přičemž váhy jednotlivých proměnných jsou voleny tak, aby komponenta zachytila maximální možnou variabilitu. Obecný tvar j-té hlavní komponenty lze vyjádřit vztahem:

$$z_j = w_{j1} * x_1 + w_{j2} * x_2 + \cdots + w_{jp} * x_p \tag{4}$$

kde z_j je j-tá hlavní komponenta, x_1 až x_p jsou původní proměnné a w_{j1} až w_{jp} jsou váhy (loadingy), které určují příspěvek jednotlivých proměnných do komponenty. První hlavní komponenta zachycuje největší možnou část variability dat, druhá komponenta zachycuje další největší část variability při podmínce nekorelovanosti s první komponentou a další komponenty postupně popisují zbývající část variability. Význam jednotlivých komponent se typicky hodnotí pomocí podílu vysvětlené variance, který je definován jako:

$$podíl\ vysvětlené\ variace_j = \frac{\lambda j}{\sum_{k=1}^{p} \lambda_k} \tag{5}$$

kde λ_j je vlastní číslo odpovídající j-té komponentě a λ_k označuje vlastní čísla všech komponent (pro $k = 1, \ldots, p$), jejichž součet představuje celkovou variabilitu datového souboru. V praxi je často sledována také kumulativní vysvětlená variance, která ukazuje, jaká část celkové

variability dat je zachycena prvním, druhým, třetím a dalším počtem komponent. Ukazatel umožňuje rozhodnout, kolik komponent je potřeba pro zachycení podstatné části informace, například 90 % variability. Důležitou podmínkou správného použití PCA je standardizace dat. PCA je citlivá na rozdílné měřítko proměnných, a pokud by standardizace nebyla provedena, mohly by proměnné s větší variabilitou nepřiměřeně dominovat výsledným komponentám. V rámci interpretace PCA se často používají dva typy grafů. Score plot zobrazuje jednotlivé vzorky promítnuté do prostoru prvních hlavních komponent a umožňuje odhalovat přirozenou strukturu dat, skupiny podobných pozorování nebo případné odlehlé body. Loading plot naopak zobrazuje váhy původních proměnných v jednotlivých komponentách a pomáhá vysvětlit, které fyzikálně-chemické charakteristiky se nejvíce podílejí na utváření hlavních směrů variability v datech. Slouží zejména k identifikaci hlavních směrů variability v datech a k lepšímu pochopení struktury datové sady (Jewsbury a Johnson, 2025).

Na výsledky PCA může navazovat shluková analýza, jejímž cílem je identifikovat skupiny vzorků s podobnými charakteristikami. Jedním z nejpoužívanějších algoritmů je K-means, který patří mezi metody učení bez učitele. Algoritmus data rozdělí do předem zadaného počtu shluků tak, aby si byly vzorky uvnitř jednoho shluku co nejpodobnější a zároveň aby se jednotlivé shluky mezi sebou co nejvíce lišily. Algoritmus minimalizuje tzv. vnitroshlukovou variabilitu, vyjádřenou jako součet čtverců vzdáleností jednotlivých bodů od centroidu svého shluku:

$$J = \sum_{k=1}^{K} \sum_{x \in C_k} \left| |x - \mu_k| \right|^2 \tag{6}$$

kde K je počet shluků, C_k je k-tý shluk a μ_k jeho centroid. Při použití algoritmu K-means je klíčové určit vhodný počet shluků. K určení se často používá Elbow metoda. Jejím principem je postupně testovat různé hodnoty počtu shluků K a pro každou z nich vypočítat hodnotu vnitro shlukové variability. Optimální počet shluků se pak hledá v bodě, kde dochází ke zpomalení poklesu chyby, tedy v místě připomínajícím loket. Elbow metoda tak představuje kompromis mezi jednoduchostí modelu a schopností zachytit přirozenou strukturu dat (Umargono, Suseno a Gunawan, 2020).

Lineární diskriminační analýza (Linear Discriminant Analysis, LDA) slouží k nalezení lineární transformace vstupních proměnných umožňující co nejlepší oddělení jednotlivých tříd. Na rozdíl od PCA, která pracuje bez znalosti cílové proměnné, je LDA metodou učení s učitelem. Jejím cílem je nalézt takovou lineární transformaci vstupních proměnných, která co nejlépe oddělí jednotlivé třídy. Zatímco PCA maximalizuje celkovou variabilitu dat, LDA maximalizuje poměr variability mezi třídami ku variabilitě uvnitř tříd. Metodu lze vyjádřit jako maximalizaci kritéria

$$J(w) = \frac{(\, w^T * S_B * w \,)}{(\, w^T * S_W * w \,)} \tag{7}$$

kde w představuje hledaný transformační vektor, S_B je matice mezitřídního rozptylu a S_W je matice vnitrotřídního rozptylu. LDA je využívána zejména pro analýzu separability tříd a pro posouzení významu jednotlivých atributů ve vztahu k cílové proměnné. (Qu a Pei, 2024).

Mezi metody výběru příznaků patří také Recursive Feature Elimination with Cross-Validation (RFECV). Jedná se o metodu automatizovaného výběru příznaků, která postupně odstraňuje méně důležité proměnné a současně sleduje, jak se redukce promítá do výkonnosti modelu. Princip RFECV lze popsat následovně: model je nejprve natrénován na všech dostupných

příznacích, poté jsou proměnné seřazeny podle jejich významu, nejméně důležité proměnné jsou odstraněny a výkon modelu je po každém kroku vyhodnocen pomocí křížové validace. Jako optimální je zvolen počet příznaků, který poskytuje nejlepší generalizační výkon. Metoda zároveň doplňuje předchozí analýzy založené na korelacích, PCA, VIF a LDA, protože poskytuje pohled na důležitost proměnných přímo z hlediska predikčního výkonu modelu (Awad a Fraihat, 2023).

1.5.4 Klasifikační a regresní modely v učení s učitelem

Učení s učitelem (supervised learning) představuje základní paradigma strojového učení, při němž jsou modely trénovány na datech obsahujících vstupní proměnné a odpovídající cílovou proměnnou. Cílem je naučit model vztah mezi vstupními atributy a výstupní proměnnou tak, aby bylo možné s dostatečnou přesností predikovat hodnotu cílové proměnné i pro nové, dosud nepozorované případy (Alnuaimi a Albaldawi, 2024).

V rámci učení s učitelem lze rozlišovat dvě hlavní skupiny úloh. Regresní modely slouží k predikci spojitých numerických hodnot, zatímco klasifikační modely přiřazují pozorování do jedné z předem definovaných kategorií. V případě predikce kvality vína lze problém interpretovat oběma způsoby. Na jedné straně je možné kvalitu chápat jako číselně vyjádřenou hodnotu a řešit úlohu regresně, na straně druhé má hodnocení kvality diskrétní charakter a lze jej formulovat jako úlohu vícetřídní klasifikace.

Lineární modely patří mezi základní metody strojového učení. Jejich princip spočívá ve vytvoření lineární kombinace vstupních proměnných, která aproximuje cílovou proměnnou. V regresní úloze je typickým zástupcem lineární regrese. Model předpokládá lineární vztah mezi vstupními atributy a cílovou proměnnou. V klasifikační úloze se používá logistická regrese, která vychází ze stejného lineárního principu. Výstup modelu je však transformován pomocí logistické nebo softmax funkce na pravděpodobnost příslušnosti k jednotlivým třídám. Výsledná klasifikace je určena třídou s nejvyšší pravděpodobností. Výhodou lineárních modelů je jejich jednoduchost, nízká výpočetní náročnost a dobrá interpretovatelnost. Omezením je jejich schopnost zachytit pouze lineární vztahy mezi proměnnými. V důsledku zmíněných vlastností mohou lineární modely dosahovat nižší přesnosti v úlohách, kde jsou vztahy mezi proměnnými složitější nebo nelineární (James a kol., 2013).

Rozhodovací stromy představují nelineární modely, které vytvářejí hierarchickou strukturu rozhodovacích pravidel. Algoritmus postupně rozděluje datový prostor podle hodnot jednotlivých atributů tak, aby maximalizoval homogenitu vzniklých podskupin. V klasifikační úloze strom přiřazuje každému listovému uzlu určitou třídu, která odpovídá nejčastější třídě trénovacích dat v daném uzlu. V regresní variantě modelu je výsledná hodnota určena například průměrem cílových hodnot vzorků nacházejících se v příslušném listovém uzlu. Rozhodovací stromy mají výhodu v dobré interpretovatelnosti a schopnosti zachytit nelineární vztahy mezi proměnnými. Nevýhodou může být jejich vyšší variabilita a náchylnost k přeučení, zejména pokud není omezena hloubka stromu nebo další parametry regulující jeho složitost. Vlastnost může vést k nestabilním výsledkům při změně trénovacích dat, což je částečně řešeno použitím ensemble metod (Mienye a Jere, 2024).

Support Vector Machines (SVM) představují algoritmus založený na hledání optimální rozhodovací hranice mezi třídami. Model se snaží nalézt hyperrovinu, která maximalizuje

vzdálenost mezi jednotlivými třídami, tzv. margin. V klasifikační úloze SVM vytváří rozhodovací hranici oddělující jednotlivé třídy. V regresní variantě, označované jako Support Vector Regression (SVR), model hledá funkci, která aproximuje vztah mezi vstupními proměnnými a cílovou hodnotou při současném omezení velikosti chyby. Pro modelování nelineárních vztahů se často využívají kernelové funkce, například radiální bazická funkce (Radial Basis Function kernel, RBF kernel), která umožňuje transformovat data do prostoru s vyšší dimenzionalitou. SVM i SVR jsou známé svou schopností modelovat komplexní vztahy mezi proměnnými, avšak jejich nevýhodou může být citlivost na nastavení hyperparametrů a vyšší výpočetní náročnost. V praxi tak mohou poskytovat velmi přesné výsledky, avšak za cenu složitějšího ladění modelu a delšího trénování (Wang, Li a Wang, 2023; Ajin, Segoni a Fanti, 2024).

Algoritmus K-Nearest Neighbours (KNN) patří mezi tzv. instance-based metody. Na rozdíl od většiny ostatních algoritmů nevytváří explicitní model během trénování, ale uchovává trénovací data a využívá je přímo při predikci. Při klasifikaci je novému vzorku přiřazena třída, která se nejčastěji vyskytuje mezi jeho $\boldsymbol{k}$ nejbližšími sousedy. V regresní variantě algoritmu je výsledná hodnota určena průměrem cílových hodnot sousedů. Výhodou KNN je jeho jednoduchost a schopnost zachytit lokální struktury v datech. Nevýhodou je citlivost na škálování proměnných a vyšší výpočetní náročnost při predikci, protože je nutné opakovaně vyhledávat nejbližší sousedy v trénovacích datech. KNN je rovněž citlivý na volbu parametru $\boldsymbol{k}$, který významně ovlivňuje výslednou kvalitu predikce (Halder a kol., 2024).

Jednotlivé uvedené modely se tedy liší nejen svou konstrukcí, ale i vhodností pro různé typy dat a úloh. Lineární modely vynikají interpretovatelností a jednoduchostí, avšak jejich schopnost zachytit komplexní vztahy je omezená. Naopak nelineární modely, jako jsou rozhodovací stromy, SVM nebo KNN, umožňují modelovat složitější závislosti, avšak často za cenu vyšší výpočetní náročnosti nebo citlivosti na nastavení parametrů. Mezi nelineárními modely existují významné rozdíly: KNN je založen na lokální podobnosti dat a dobře zachycuje lokální struktury, avšak je citlivý na škálování a volbu parametru $\boldsymbol{k}$. SVM naopak hledá globální rozhodovací hranici a pomocí kernelů dokáže modelovat i složité vztahy, ale vyžaduje pečlivé ladění hyperparametrů.

1.5.5 Ensemble metody

Ensemble metody představují pokročilé přístupy strojového učení, jejichž cílem je zvýšit přesnost a stabilitu predikčních modelů prostřednictvím kombinace více dílčích modelů. Základní myšlenkou metod je, že více modelů, které se doplňují ve svých silných a slabých stránkách, může společně dosahovat lepších výsledků než jednotlivý model (Mienye a Sun, 2022).

Významnou roli v oblasti ensemble metod hrají rozhodovací stromy, které jsou často využívány jako základní stavební prvky ensemble přístupů, například v metodách typu Random Forest nebo boosting. Modely se vyznačují relativně vysokou variabilitou, což umožňuje jejich efektivní kombinaci vedoucí ke stabilnějším a přesnějším výsledkům (Blockeel a kol., 2023).

Jedním ze základních a nejrozšířenějších ensemble algoritmů je Random Forest. Metoda je založena na kombinaci velkého počtu rozhodovacích stromů, přičemž každý strom je trénován na náhodném podvzorku dat (bagging) a současně využívá náhodně vybranou podmnožinu vstupních atributů. Zmíněným způsobem dochází ke snížení korelace mezi jednotlivými stromy a ke zlepšení generalizační schopnosti modelu. V klasifikační úloze je výsledná predikce určena

hlasováním jednotlivých stromů, zatímco v regresní úloze je výsledná hodnota získána průměrováním jejich predikcí. Random Forest tak dokáže efektivně snížit varianci modelu a současně zachovat schopnost zachytit nelineární vztahy v datech. Mezi jeho hlavní výhody patří dobrá predikční výkonnost, odolnost vůči přeučení a možnost odhadovat důležitost jednotlivých proměnných. Nevýhodou může být nižší interpretovatelnost ve srovnání s jednoduššími modely. Díky zmíněným vlastnostem bývá Random Forest často považován za silný základní model (baseline), vůči němuž jsou porovnávány pokročilejší ensemble přístupy (Salman, Kalakech a Steiti, 2024).

Na princip Random Forestu navazují další ensemble přístupy, které kombinují více modelů různými způsoby. Jedním z nejjednodušších způsobů kombinace modelů je tzv. voting. Princip metody spočívá v tom, že více modelů nezávisle vytváří predikce, které jsou následně agregovány do jednoho výsledku. V klasifikační úloze může být výsledná třída určena například většinovým hlasováním jednotlivých modelů. V regresní úloze se místo hlasování používá průměrování predikovaných hodnot. Výhodou voting metod je jejich jednoduchost a schopnost zlepšit stabilitu predikce bez nutnosti složitého ladění. Nevýhodou je, že všechny modely mají stejnou váhu a metoda nebere v úvahu jejich individuální kvalitu. Rozšířením voting je tzv. soft voting, které se používá v klasifikačních úlohách. Na rozdíl od hard voting, kde modely hlasují pouze pro konkrétní třídu, soft voting pracuje s pravděpodobnostmi jednotlivých tříd. Každý model poskytne pravděpodobnostní odhad příslušnosti k jednotlivým třídám a výsledná predikce je určena na základě průměru pravděpodobností. Soft voting umožňuje lépe využít informace o nejistotě modelů a často vede k vyšší přesnosti než jednoduché hlasování. Účinnost metody však závisí na kvalitě jednotlivých modelů, přičemž kombinace slabších nebo silně korelovaných modelů nemusí vést k výraznému zlepšení výkonu (Manconi a kol., 2022).

Dalším pokročilým ensemble přístupem je stacking, který kombinuje více modelů pomocí tzv. metamodelu. Základní modely nejprve vytvářejí své predikce, které jsou následně použity jako vstup pro další model, označovaný jako metamodel. Metamodel se učí, jak nejlépe kombinovat výstupy jednotlivých modelů, aby minimalizoval výslednou chybu. V klasifikační úloze jsou jako vstupy pro metamodel často používány pravděpodobnosti jednotlivých tříd, zatímco v regresní úloze se využívají přímo predikované hodnoty jednotlivých modelů. Stacking tak umožňuje efektivně využít silné stránky jednotlivých modelů a současně potlačit jejich slabiny. Výhodou stackingu je jeho schopnost dosahovat vysoké přesnosti díky inteligentní kombinaci modelů. Nevýhodou je vyšší složitost, výpočetní náročnost a riziko přeučení, pokud není správně nastavena validace modelu. Přestože stacking často dosahuje velmi dobrých výsledků, jeho přínos není zaručen a závisí na vhodné kombinaci základních modelů a správném nastavení validačního schématu (Khoshkroodi, Sani a Aajami, 2024).

Ensemble metody obecně představují efektivní způsob, jak zlepšit výkon prediktivních modelů, zejména při práci s komplexními datovými sadami. Kombinace různých modelovacích přístupů umožňuje zachytit jak lineární, tak nelineární vztahy v datech a zvyšuje robustnost výsledného modelu (Kunapuli, 2023).

1.5.6 Evaluace modelů

Vyhodnocení modelů představuje nezbytnou součást analytického procesu, protože umožňuje posoudit, nakolik jsou vytvořené modely schopné přesně a spolehlivě predikovat cílovou

proměnnou. Volba evaluačních metrik závisí na typu řešené úlohy i na charakteru cílové proměnné. Správně zvolená evaluace je klíčová nejen pro porovnání modelů mezi sebou, ale také pro pochopení jejich silných a slabých stránek a identifikaci jejich případných nedostatků (Miller a kol., 2024).

V případě klasifikačních úloh je cílem modelu správně přiřadit jednotlivá pozorování do jedné z předem definovaných tříd. Základním nástrojem pro vyhodnocení klasifikace je matice záměn (confusion matrix), která přehledně zachycuje počty správných a nesprávných klasifikací (Stracke a kol., 2025).

Matice záměn obsahuje čtyři základní hodnoty. True Positive (TP) představuje počet správně klasifikovaných pozitivních případů, True Negative (TN) počet správně klasifikovaných negativních případů. False Positive (FP) označuje případy, které model nesprávně klasifikoval jako pozitivní, a False Negative (FN) případy, které byly nesprávně klasifikovány jako negativní. Na základě vyjmenovaných hodnot lze definovat řadu evaluačních metrik. Accuracy (přesnost) vyjadřuje podíl správně klasifikovaných vzorků na celkovém počtu pozorování:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

Jedná se o nejjednodušší metriku, která poskytuje celkový přehled o úspěšnosti modelu. Nevýhodou je, že může být zavádějící u nevyvážených dat, kdy model dosahuje vysoké přesnosti i při špatné klasifikaci méně zastoupených tříd. Precision (přesnost pozitivní predikce) vyjadřuje podíl správně klasifikovaných pozitivních případů ze všech případů, které model označil jako pozitivní:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{9}$$

Metrika je důležitá zejména v situacích, kdy je nežádoucí vysoký počet falešně pozitivních výsledků. Recall (citlivost) vyjadřuje podíl správně klasifikovaných pozitivních případů ze všech skutečně pozitivních případů:

$$\text{Recall} = \frac{TP}{TP+FN} \tag{10}$$

Recall je důležitý v případech, kdy je klíčové zachytit co nejvíce pozitivních případů, například v medicínských aplikacích. F1-score představuje harmonický průměr precision a recall:

$$\text{F1} = 2 * \frac{Precision*Recall}{Precision+Recall} \tag{11}$$

F1-score vyvažuje obě předchozí metriky a je vhodná zejména v případě nevyvážených dat, kde samotná accuracy neposkytuje dostatečně vypovídající informace. Dalším důležitým nástrojem je Receiver Operating Characteristic (ROC) křivka, která zobrazuje vztah mezi True Positive Rate (recall) a False Positive Rate (FPR):

$$\text{FPR} = \frac{FP}{FP+TN} \tag{12}$$

Plocha pod ROC křivkou (Area Under Curve, AUC) vyjadřuje schopnost modelu rozlišovat mezi třídami. Vyšší hodnota AUC indikuje lepší klasifikační schopnost modelu (Vujović, 2021).

V regresních úlohách je cílem modelu predikovat spojitou numerickou hodnotu. Vyhodnocení modelu je založeno na porovnání skutečných a predikovaných hodnot.

Mean absolute error (MAE) vyjadřuje průměrnou absolutní chybu mezi skutečnými a predikovanými hodnotami:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{13}$$

kde n označuje počet pozorování, y_i skutečnou hodnotu cílové proměnné a $\hat{y}_i$ odpovídající hodnotu predikovanou modelem. MAE udává průměrnou velikost chyby v původních jednotkách a je relativně robustní vůči odlehlým hodnotám (Rauf a kol., 2024).

Root Mean Squared Error (RMSE) je definováno jako druhá odmocnina průměrné kvadratické chyby:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{14}$$

Na rozdíl od MAE penalizuje větší chyby výrazně více, protože chyby jsou umocněny na druhou. RMSE je proto citlivější na odlehlé hodnoty a je vhodné jej použít v situacích, kdy jsou velké chyby nežádoucí. Koeficient determinace R^2 vyjadřuje, jak dobře model vysvětluje variabilitu dat:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{15}$$

kde y_i představuje skutečné hodnoty, $\hat{y}_i$ predikované hodnoty, $\bar{y}$ průměr skutečných hodnot a n počet pozorování. Čitatel vyjadřuje součet kvadratických chyb modelu, zatímco jmenovatel představuje celkovou variabilitu dat. Hodnota R^2se pohybuje v intervalu ⟨-∞, 1⟩. Hodnota blízká 1 znamená, že model velmi dobře vysvětluje variabilitu dat, zatímco nízké nebo záporné hodnoty indikují, že model nedokáže data dostatečně popsat (Cenita, Asuncion a Victoriano, 2023).

Správně zvolená kombinace evaluačních metrik umožňuje komplexní posouzení kvality modelu. V klasifikačních úlohách je vhodné sledovat více metrik současně, zejména v případě nevyvážených dat. V regresních úlohách je vhodné kombinovat metriky poskytující snadnou interpretaci (MAE) s metrikami citlivými na větší chyby (RMSE) a s ukazateli vysvětlené variability (R^2).

V rámci hodnocení prediktivních modelů je vhodné doplnit základní evaluační metriky o informaci o jejich variabilitě. Jednorázová hodnota metriky, například accuracy, F1-score nebo RMSE, poskytuje pouze omezený pohled na výkonnost modelu, protože nezohledňuje variabilitu výsledků způsobenou rozdělením dat nebo náhodností procesu učení (Stracke a kol., 2025).

Pro robustnější posouzení výkonu modelů se proto využívají intervaly spolehlivosti (confidence intervals, CI), které vyjadřují rozsah hodnot, v němž se s danou pravděpodobností nachází skutečná hodnota sledované metriky (Botella a Sánchez-Meca, 2024). V kontextu strojového učení umožňují intervaly spolehlivosti kvantifikovat nejistotu odhadu výkonu modelu a zachytit jeho citlivost na změny v trénovacích datech (Heskes, 1996).

Při použití křížové validace je model opakovaně trénován a vyhodnocován na různých podmnožinách dat, čímž vzniká množina hodnot dané metriky. Z množiny lze následně odvodit interval spolehlivosti pomocí percentilové metody. Konkrétně je interval určen jako rozmezí mezi 2,5. a 97,5. percentilem, což odpovídá přibližně 95 % intervalu spolehlivosti.

Statistický přístup umožňuje lépe posoudit stabilitu modelu. Model s úzkým intervalem spolehlivosti lze považovat za stabilnější než model s širokým intervalem, i pokud mají oba modely podobnou střední hodnotu metriky.

Pro přehledné porovnání více modelů lze intervaly spolehlivosti vizualizovat pomocí tzv. forest plotu, který umožňuje současně zobrazit střední hodnotu metriky i její variabilitu. Způsob vizualizace usnadňuje identifikaci modelů, které dosahují nejen vysoké přesnosti, ale také stabilních výsledků (Botella a Sánchez-Meca, 2024).

Evaluace modelů tak představuje klíčový krok při výběru nejvhodnějšího modelu a interpretaci výsledků analytického procesu.

1.5.7 Validační strategie modelů

Validační strategie představují klíčovou součást procesu tvorby a vyhodnocení prediktivních modelů. Jejich cílem je zajistit, aby byl odhad výkonnosti modelu co nejvíce realistický a aby model nebyl hodnocen na datech, která již „viděl“ během trénování. Správně zvolená validační strategie umožňuje posoudit schopnost modelu generalizovat na nová, dosud nepozorovaná data, což je zásadní pro jeho praktické využití. Základním přístupem k validaci modelu je rozdělení datové sady na trénovací a testovací část (tzv. train/test split). Trénovací data slouží k učení modelu, zatímco testovací data jsou využita výhradně pro jeho finální vyhodnocení. Postup umožňuje získat nezávislý odhad výkonnosti modelu. Nevýhodou jednoduchého rozdělení dat je však závislost výsledků na konkrétním náhodném rozdělení, které může vést k vyšší variabilitě odhadu. Proto se v praxi často využívá křížová validace (cross-validation), která poskytuje robustnější odhad výkonnosti modelu. Principem křížové validace je rozdělení dat do několika částí (tzv. foldů). Model je následně opakovaně trénován na většině dat a testován na zbývající části. Postup se opakuje tak, aby každá část dat byla jednou použita jako validační. Výsledná metrika je pak obvykle určena jako průměr přes všechny iterace. Nejčastěji používanou variantou je k-fold cross-validation, kde je datová sada rozdělena do $\boldsymbol{k}$ stejně velkých částí (Lopez a kol., 2025).

V kontextu klasifikačních úloh je důležitým rozšířením křížové validace tzv. stratifikovaná křížová validace (Stratified K-Fold). Která zajišťuje, že rozdělení tříd cílové proměnné je ve všech foldech přibližně stejné jako v původní datové sadě. Stratifikace je zvláště důležitá v případě nerovnoměrně zastoupených tříd, protože zabraňuje situaci, kdy by některé třídy byly v jednotlivých částech dat nedostatečně reprezentovány. Důsledkem zvyšuje stabilitu i spolehlivost odhadu výkonnosti modelu (Mahesh a kol., 2023).

Důležitým principem při validaci modelů je oddělení procesu ladění hyperparametrů od finálního testování. Hyperparametry modelu jsou obvykle optimalizovány pomocí validačních dat, například v rámci křížové validace. Testovací data by však měla být použita až na samém konci procesu, a pouze jednou pro finální vyhodnocení modelu. Pokud by byla testovací data použita během ladění modelu, došlo by k jejich „úniku“ (data leakage), což by vedlo k nadhodnocení výkonnosti modelu a zkreslení výsledků (Stracke a kol., 2025).

Správně zvolená validační strategie tak představuje nezbytný předpoklad pro objektivní porovnání modelů a pro zajištění jejich schopnosti generalizace. V kombinaci s vhodně

zvolenými evaluačními metrikami umožňuje spolehlivě posoudit kvalitu modelu a jeho použitelnost v praxi (Lopez a kol., 2025).

1.5.8 Přeučení, podučení a learning curves

Při tvorbě prediktivních modelů je důležité nejen dosáhnout vysoké přesnosti na trénovacích datech, ale především zajistit schopnost modelu generalizovat na nová, dosud nepozorovaná data. V tomto kontextu se často setkáváme se dvěma základními problémy, označovanými jako přeučení (overfitting) a podučení (underfitting). K přeučení dochází v situaci, kdy model příliš přesně zachytí strukturu trénovacích dat, včetně náhodného šumu a odlehlých hodnot. Model dosahuje velmi dobrých výsledků na trénovacích datech, avšak jeho výkon na validačních nebo testovacích datech je výrazně horší. Přeučení je typické pro příliš komplexní modely, které mají vysokou kapacitu a dokážou se přizpůsobit i nevýznamným detailům v datech. Naopak podučení nastává tehdy, když model není schopen zachytit ani základní strukturu dat. V takovém případě dosahuje model nízké přesnosti jak na trénovacích, tak i na validačních datech. Podučení bývá způsobeno příliš jednoduchým modelem nebo nevhodnou volbou vstupních atributů. Uvedené problémy úzce souvisejí s tzv. kompromisem mezi bias a variance (bias-variance tradeoff). Bias vyjadřuje systematickou chybu modelu, tedy míru, do jaké model nedokáže zachytit skutečný vztah mezi proměnnými. Variance naopak vyjadřuje citlivost modelu na změny v trénovacích datech. Model s vysokým biasem bývá příliš jednoduchý a vede k podučení, zatímco model s vysokou variancí bývá příliš komplexní a náchylný k přeučení. Cílem modelování je nalézt takovou úroveň složitosti modelu, která minimalizuje obě složky chyby a vede k dobré generalizaci (Aliferis a Simon, 2024).

Pro identifikaci přeučení a podučení je běžné porovnávat výkon modelu na trénovacích a validačních datech. Pokud model dosahuje výrazně lepších výsledků na trénovacích datech než na validačních, jedná se zpravidla o přeučení. Naopak pokud jsou výsledky špatné na obou typech dat, může se jednat o podučení. Porovnání poskytuje důležitou informaci o tom, zda je model vhodně nastaven z hlediska své složitosti. Užitečným nástrojem pro analýzu chování modelu jsou tzv. learning curves (učící křivky). Které zobrazují závislost výkonnosti modelu na velikosti trénovací množiny dat. Obvykle se zobrazují dvě křivky – výkon modelu na trénovacích datech a výkon na validačních datech. V případě přeučení se typicky projevuje výrazný rozdíl mezi oběma křivkami, kdy model dosahuje velmi dobrých výsledků na trénovacích datech, ale podstatně horších na validačních. Naopak při podučení se obě křivky nacházejí na nízké úrovni a jejich hodnoty jsou si blízké. Learning curves tak umožňují nejen diagnostikovat problém, ale také naznačit možné řešení, například úpravu složitosti modelu, změnu množství trénovacích dat nebo volbu jiného modelovacího přístupu (Zaloumis, Rajasekhar a Simpson, 2025).

Analýza přeučení a podučení spolu s využitím learning curves představuje důležitý nástroj pro ladění modelů a pro pochopení jejich chování. Přístupy umožňují lépe interpretovat výsledky modelování a přispívají k tvorbě robustních a spolehlivých prediktivních modelů.

1.5.9 Reprodukovatelnost a použité prostředí

Reprodukovatelnost analytických postupů představuje jeden ze základních principů moderní datové vědy a výpočetního výzkumu. Označuje schopnost nezávislého výzkumníka znovu

dosáhnout shodných nebo velmi podobných výsledků při použití stejných dat, metod a analytických kroků. Transparentnost analytického procesu je s reprodukovatelností úzce spjata, protože umožňuje dohledat jednotlivé kroky analýzy, ověřit jejich správnost a celý postup opakovat (Sandve, 2013).

Reprodukovatelnost vyžaduje systematickou dokumentaci pracovního postupu, uchování vstupních dat, zaznamenání použitých parametrů a verzování kódu. Jako například platforma Kaggle, která umožňuje uchovávat analytický proces v podobě interaktivního notebooku kombinujícího textový komentář, výpočetní kroky i vizualizační výstupy. Platforma současně podporuje verzování a historii úprav, čímž přispívá k transparentnosti a opakovatelnosti provedených experimentů (Coursera, 2025).

1.5.10 Implementace datové vědy

Na základě výše uvedených teoretických východisek je v následujícím přehledu stručně shrnuto, jaké fáze datově orientovaného postupu byly v praktické části práce implementovány a jaké konkrétní metody, techniky a nástroje byly v jednotlivých fázích využity.

1. Formulace problému,
    - vymezení cíle práce,
    - definice úlohy,
        - klasifikace,
        - regrese.
2. Získání dat,
    - Zdroj dat – Wine Quality Data Set (Red & White Wine) (Ruthgn, 2024),
    - popis datasetu,
        - počet záznamů a atributů - df.shape,
        - význam jednotlivých proměnných,
        - cílová proměnná (quality).
3. Příprava dat,
    - transformace proměnných,
        - standardizace - StandardScaler(),
        - normalizace - MinMaxScaler(),
        - OneHotEncoder - OneHotEncoder(),
        - Transformace.
4. Explorativní analýza dat,
    - Jednorozměrná analýza (1-D),
        - popisná statistika - data.describe(),
        - histogramy,
        - Q-Q grafy,
    - Dvourozměrná analýza (2-D),
        - korelační matice - sns.heatmap(corr_matrix),
    - Více-rozměrná analýza (n-D),
        - VIF – variance_inflation_factor(),

- Redukce dimenzionality a výběr příznaků,
  - PCA,
    - pca.explained_variance_ratio_.cumsum(),
    - PCA(n_components=2),
    - Loading komponent,
    - Elbow metoda,
  - LDA,
    - Počet LDA komponent,
    - projekce dat,
    - význam příznaků,
    - Distribuce významnosti,
    - Kumulativní významnost,
  - RFECV – RFECV(),
- Shlukování,
  - K-means – Kmeans().

5. Modelování,
   - rozdělení dat,
     - train/test split,
     - KFold – Kfold(),
     - StratifiedKFold – StratifiedKFold(),
   - klasifikační modely,
     - Logistic Regression – LogisticRegression(),
     - Decision Tree Classifier – DecisionTreeClassifier(),
     - SVC – SVC(),
     - KNN – KneighborsClassifier(),
   - regresní modely,
     - Linear Regression - LinearRegression(),
     - Decision Tree Regressor – DecisionTreeRegressor(),
     - SVR – SVR(),
     - KNN Regressor – KneighborsRegressor(),
   - ensemble metody,
     - Random Forest Classifier – RandomForestClassifier(),
     - Soft Voting Classifier – VotingClassifier(voting="soft"),
       - Logistic Regression,
       - Decision Tree Classifier,
       - SVM,
       - KNN,
     - Stacking Classifier – StackingClassifier(),
       - Logistic Regression,
       - Decision Tree Classifier,
       - SVM,
       - KNN,
     - Random Forest Regressor – RandomForestRegressor,

        - Voting Regressor – VotingRegressor(),
            - LinearRegression,
            - Decision Tree Regressor,
            - SVR,
            - KNN Regressor,
        - Stacking Regressor – StackingRegressor(),
            - LinearRegression,
            - Decision Tree Regressor,
            - SVR,
            - KNN Regressor.
6. Vyhodnocení,
    - klasifikační metriky,
        - confidence intervals,
        - Accuracy – accuracy_score(),
        - Recall - recall_score(),
        - F1-score - f1_score(),
        - ROC,
        - Forest plot,
        - Learning Curve,
    - regresní metriky,
        - confidence intervals,
        - RMSE - mean_squared_error(),
        - MAE - mean_absolute_error(),
        - R^2 - r2_score(),
        - Forest plot.

1.6 Přehled dosavadních studií a přístupů k predikci kvality vína

Predikce kvality vína na základě fyzikálně-chemických parametrů patří mezi často řešené úlohy v oblasti datové vědy a strojového učení, a jak ve vzdělávacím, tak i aplikačním kontextu. Wine Quality Data Set, dostupný například na platformě Kaggle, patří mezi nejpoužívanější datasety a obsahuje několik tisíc vzorků vína včetně jejich chemických vlastností a senzorického hodnocení kvality. Úlohu lze formulovat jak jako regresní problém, tak jako vícetřídní klasifikaci (Ruthgn, 2024).

V dostupných projektech a studiích se opakovaně objevuje podobný analytický postup, který zahrnuje explorativní analýzu dat (EDA), předzpracování dat a následné modelování pomocí různých algoritmů strojového učení. Například projekt Red Wine Quality Prediction – EDA & Classification se zaměřuje na detailní explorativní analýzu dat a porovnání základních klasifikačních modelů. Zvláštní důraz je kladen na vizualizaci dat a identifikaci nerovnoměrného rozdělení cílové proměnné, které má významný vliv na následné modelování (Monfared, 2024).

Další přístup je prezentován v projektu Wine Quality Prediction – Comparing Top ML Models, kde jsou systematicky porovnávány různé algoritmy strojového učení, včetně metod založených na ensemble přístupech, jako jsou Random Forest nebo Gradient Boosting. Výsledky přístupů

ukazují, že nelineární modely, zejména ensemble metody, dosahují v úloze zpravidla lepších výsledků než jednodušší lineární modely (Yasser, 2022).

V některých projektech je úloha dále zjednodušována, například převodem původní vícetřídní klasifikace na binární nebo omezený počet tříd. Přístup lze nalézt například v notebooku Wine Quality Prediction, kde je kvalita vína rozdělena na dvě skupiny. Zjednodušení může vést ke zvýšení dosažené přesnosti, avšak za cenu ztráty detailní informace o jednotlivých úrovních kvality (Elshamy, 2023).

Souhrnně lze konstatovat, že dosavadní přístupy k predikci kvality vína vykazují několik společných rysů: využití nelineárních modelů, důraz na explorativní analýzu dat a opakující se problém nevyváženosti tříd. Poznatky poskytují základ pro návrh experimentální části práce a umožňují následné srovnání dosažených výsledků s existujícími řešeními.

1.7 Faktory ovlivňující kvalitu vína

Predikce kvality vína na základě fyzikálně-chemických parametrů patří mezi často řešené úlohy datové vědy. Wine Quality Data Set, dostupný například na Kaggle, obsahuje tisíce vzorků s chemickými vlastnostmi a senzorickým hodnocením a umožňuje formulaci úlohy jako regresi i vícetřídní klasifikaci (Ruthgn, 2024).

Mezi významné faktory ovlivňující kvalitu vína patří zejména obsah alkoholu, koncentrace kyselin, obsah oxidu siřičitého, chloridy a další chemické složky. Proměnné mají přímý vliv na senzorické vlastnosti vína, jako je chuťový profil, aroma a celková vyváženost. Empirické studie ukazují, že obsah alkoholu bývá pozitivně korelován s hodnocením kvality, zatímco vyšší hodnoty těkavé kyselosti mohou indikovat nežádoucí chemické procesy a negativně ovlivňovat senzorický dojem. Podobně zvýšené koncentrace chloridů či oxidu siřičitého mohou mít negativní vliv na výslednou kvalitu vína (Liu, 2024).

Další důležitou charakteristikou je rovnováha mezi jednotlivými chemickými složkami. Například kyselost ovlivňuje svěžest a stabilitu vína, zatímco hodnota pH souvisí s jeho chemickou stabilitou a mikrobiologickou odolností. I když některé proměnné, jako například zbytkový cukr, nemusí být ve všech studiích identifikovány jako klíčové prediktory kvality, přispívají k celkovému senzorickému vnímání vína a mohou hrát roli v kombinaci s dalšími faktory (Jiang, Liu a Yang, 2023).

Kvalitu vína dále ovlivňují technologické faktory spojené s procesem výroby, jako je způsob a délka fermentace, řízení teploty, použití specifických kvasinek či podmínky zrání. Význam faktorů potvrzují i studie zaměřené na vliv technologických podmínek, například způsob sklizně, teplotní režim či doba zpracování hroznů, které mohou ovlivnit výsledné chemické složení vína. Proměnné se obvykle v dostupných datových sadách nevyskytují, nicméně v praxi významně ovlivňují výsledné fyzikálně-chemické vlastnosti vína, a i jeho kvalitu (Catania, Vallone a Pipitone, 2009).

1.7.1 Vymezení pojmu kvalita vína

Kvalita vína představuje komplexní a vícerozměrný pojem, který nelze redukovat na jedinou veličinu. V odborné literatuře je chápána jako kombinace senzorických vlastností a fyzikálně-

chemického složení, přičemž tyto složky jsou vzájemně úzce propojené. Chemické charakteristiky, jako obsah alkoholu, kyselost, pH nebo koncentrace oxidu siřičitého, ovlivňují senzorický profil vína, zatímco výsledné hodnocení je určováno lidským vnímáním. Senzorická kvalita vychází z hodnocení organoleptických vlastností (aroma, chuť, barva a celkový dojem) a i přes standardizované postupy zůstává do určité míry subjektivní, což se promítá i do dat využívaných pro analytické a prediktivní účely (Basalekou a kol., 2023).

Z hlediska datové analýzy a prediktivního modelování je kvalita vína zpravidla reprezentována numerickým senzorickým hodnocením, které slouží jako cílová proměnná. Hodnocení bývá vyjádřeno pomocí bodové škály, typicky v rozsahu 0–10, přičemž má ordinální charakter, jelikož jednotlivé hodnoty vyjadřují především pořadí kvality, nikoli přesně definované intervaly mezi stupni. Rozdíly mezi sousedními hodnotami proto nemusí odpovídat stejnému kvalitativnímu posunu, což má významné důsledky pro volbu vhodných analytických metod a přístupů při modelování kvality vína. Chemické parametry tak představují objektivní vstupy, které umožňují kvantifikovat vztah mezi složením vína a jeho výsledným hodnocením pomocí statistických metod a technik strojového učení (Hopfer a kol., 2015).

1.7.2 Variabilita senzorického hodnocení a její důsledky

Senzorické hodnocení vína představuje důležitý nástroj pro posouzení jeho kvality, avšak výsledky hodnocení jsou zatíženy přirozenou variabilitou mezi jednotlivými hodnotiteli. Studie zaměřené na spolehlivost senzorických testů poukazují na skutečnost, že i při použití standardizovaných metodik mohou být výsledky ovlivněny jak interními, tak externími faktory. Variabilita se projevuje nejen mezi jednotlivými členy hodnotitelského panelu, ale také v rámci opakovaných hodnocení stejného hodnotitele za rozdílných podmínek. Mezi významné zdroje variability patří individuální rozdíly ve vnímání senzorických podnětů, úroveň zkušeností a odborných znalostí hodnotitelů. Zkušenější degustátoři mohou být citlivější na jemné rozdíly v aroma nebo chuti, zatímco méně zkušení hodnotitelé mohou posuzovat vzorky odlišným způsobem. Rozdíly se následně promítají do bodového hodnocení a zvyšují rozptyl výsledků v rámci hodnotitelského panelu (MAURYA, Neelesh a Kumar, 2024).

Další významnou roli hrají psychologické a fyziologické faktory, které mohou ovlivňovat senzorické vnímání jednotlivců. Mezi faktory patří například věk, zdravotní stav, individuální citlivost smyslů, míra únavy, očekávání hodnotitele nebo prostředí, ve kterém degustace probíhá. Vlivy mohou způsobit, že stejný vzorek vína je různými hodnotiteli vnímán odlišně, případně že stejný hodnotitel poskytne rozdílné hodnocení v závislosti na aktuálních podmínkách testování. Psychologické a fyziologické faktory tak představují další zdroj variability senzorických dat (Alpeza, Buljević a Budimir, 2025).

Uvedené aspekty mají zásadní důsledky pro analytické zpracování výsledků senzorického hodnocení. Z pohledu datové vědy a prediktivního modelování je nutné si uvědomit, že senzorické skóre nepředstavuje absolutní a objektivní měřítko kvality, ale výsledek lidského hodnocení zatížený nejistotou. Nejistota se v datech projevuje jako šum, který může ovlivňovat stabilitu modelů a jejich schopnost generalizace. V důsledku mohou prediktivní modely zachycovat nejen skutečné vztahy mezi fyzikálně-chemickými parametry a kvalitou vína, ale i náhodnou variabilitu vyplývající ze subjektivního hodnocení kvality jednotlivými degustátory. (Jiang, Liu a Yang, 2023).

2 Metodika

Kapitola popisuje metodický postup použitý při řešení úlohy predikce kvality vína pomocí metod datové vědy a strojového učení. Cílem metodické části je systematicky představit jednotlivé kroky zpracování dat, od jejich načtení a základní analýzy až po vytvoření a vyhodnocení predikčních modelů.

Analytický postup vychází z běžného workflow datové vědy, které zahrnuje fáze přípravy dat, explorační analýzy, návrhu modelů a jejich následného vyhodnocení. Nejprve je provedena kontrola kvality dat a jejich základní statistická analýza. Následuje explorační analýza dat, jejímž cílem je identifikovat strukturu dat, vztahy mezi jednotlivými proměnnými a potenciálně významné atributy ovlivňující kvalitu vína. Součástí je také využití metod redukce dimenzionality a analýzy struktury dat.

Na základě získaných poznatků jsou následně vytvořeny predikční modely využívající metody klasifikace a regrese. Modely jsou trénovány na trénovací množině dat a jejich výkon je vyhodnocen pomocí vhodných metrik a validačních strategií, zejména křížové validace. Součástí práce je rovněž návrh doplňkových experimentů, které analyzují vliv vybraných faktorů, například četnosti tříd nebo redukce vstupních atributů, na výkon vytvořených modelů.

Z důvodu přehlednosti jsou v práci prezentovány pouze klíčové výsledky jednotlivých modelů a experimentů. Podrobné výstupy, doplňkové grafy, průběžné testy a kompletní implementace jsou uvedeny v přiloženém notebooku (viz Příloha A.1).

Je rovněž nutné poznamenat, že dosažené výsledky se mohou při jednotlivých spuštěních notebooku mírně lišit. Variabilitu způsobuje především stochastická povaha některých použitých metod. Odchylky se mohou projevit jak ve výsledných metrikách, tak i v dílčích výstupech modelů. Významnější rozdíly lze pozorovat také v měření časové náročnosti, které je ovlivněno využitím cloudového výpočetního prostředí platformy Kaggle, kde se dostupné výpočetní prostředky mohou v čase lišit v závislosti na aktuálním zatížení infrastruktury. Prezentované výsledky odpovídají stavu získanému v době zpracování této práce.

2.1 Uživatelské rozhraní a interaktivní prvky notebooku

Součástí implementace je rozšíření interaktivního notebooku o uživatelské funkce, které umožňují praktickou aplikaci navržených modelů a experimentální práci s daty. Prvky byly realizovány s využitím knihovny ipywidgets a slouží k demonstraci využití modelů na nových datech i k analýze jejich chování při změně parametrů (viz Příloha A.1).

První funkce umožňuje zadání vlastních hodnot fyzikálně-chemických vlastností vína a následnou predikci jeho kvality pomocí natrénovaných modelů. Výsledky jsou prezentovány ve formě přehledných tabulek, které obsahují jak predikovanou hodnotu (v případě regrese), tak i pravděpodobnostní rozdělení tříd (v případě klasifikace).

Další funkcionalitou je možnost hromadné predikce kvality vína na základě uživatelem nahraného datasetu. Notebook umožňuje načtení dat ve formátu CSV nebo Excel a následné zpracování pomocí nejlepšího vybraného modelu. Výstupem je rozšířený dataset doplněný o predikce kvality, který lze dále exportovat a využít pro další analýzu.

Notebook dále obsahuje interaktivní nástroj pro experimentální úpravu hyperparametrů modelů. Uživatel může měnit vybrané parametry, model dočasně přetrénovat a porovnat jeho výkon s původní variantou. Funkce umožňuje lépe pochopit vliv jednotlivých parametrů na chování modelu a jeho predikční schopnost.

Interaktivní prvky rozšiřují analytickou část práce o praktickou rovinu a umožňují nejen reprodukci experimentů, ale i jejich další rozšiřování a aplikaci na nová data (viz Příloha A.1).

2.2 Metodologický rámec práce

Práce využívá přístup datové vědy, který kombinuje statistické metody, explorační analýzu dat a metody strojového učení. Analytický proces byl realizován iterativním způsobem, kdy jednotlivé kroky analýzy byly postupně zpřesňovány na základě průběžných výsledků a konzultací s vedoucím práce.

Vedoucí práce v rámci projektu vystupoval v roli zadavatele, který průběžně poskytoval metodické připomínky a doporučení k dalšímu směrování analýzy. Na základě konzultací docházelo k postupným úpravám analytického postupu, například v oblasti struktury notebooku, způsobu vizualizace dat nebo výběru vhodných metod modelování. Proces byl doplněn o využití nástrojů umělé inteligence, které sloužily jako podpůrný prostředek při návrhu a implementaci jednotlivých kroků. Zatímco nástroje umělé inteligence poskytovaly alternativní návrhy a urychlovaly iterace, klíčová rozhodnutí o dalším postupu byla činěna na základě konzultací s vedoucím práce a vlastního porozumění řešené problematice.

Pro realizaci analytické části byl využit programovací jazyk Python a prostředí Jupyter Notebook (Kaggle notebook, viz Příloha A.1), které umožňuje kombinovat zdrojový kód, vizualizace a komentovaný text v jednom dokumentu. Formát byl zvolen s ohledem na transparentnost analytického postupu a možnost snadné reprodukovatelnosti jednotlivých kroků analýzy.

Důležitým aspektem práce je také reprodukovatelnost provedeného výzkumu. Veškeré kroky analytického procesu, včetně přípravy dat, explorační analýzy, modelování a vyhodnocení, jsou zdokumentovány ve veřejně dostupném Jupyter Notebooku, jehož odkaz je součástí přílohy A.1.

Notebook obsahuje nejen implementaci jednotlivých metod, ale také detailní popisy použitých technik, parametrů modelů a nastavení validačních postupů. Aby bylo možné celý analytický proces zpětně ověřit, reprodukovat a případně dále rozšiřovat.

2.2.1 Datově-vědní přístup k řešení problému

Řešený problém predikce kvality vína byl formulován jako úloha datové vědy, která zahrnuje analýzu struktury dat a následné vytvoření predikčních modelů. Analytický proces proto zahrnoval několik na sebe navazujících fází, zejména přípravu dat, explorační analýzu dat, návrh a trénování modelů strojového učení a jejich následné vyhodnocení.

Důležitou součástí je také iterativní charakter analýzy. Jednotlivé kroky nebyly realizovány lineárně, ale byly postupně zpřesňovány na základě průběžných výsledků. Například poznatky

získané během explorační analýzy dat vedly k úpravám výběru atributů nebo ke změnám v konfiguraci některých modelů.

Při implementaci analytických postupů byla využita také podpora nástrojů založených na umělé inteligenci, které byly použity zejména při návrhu některých částí zdrojového kódu. Vygenerovaný kód však nebyl používán přímo, ale byl vždy analyzován, upravován a přizpůsobován konkrétním potřebám projektu.

2.2.2 Metodologický rámec RAD-AI / CRISP-DM

Celý analytický proces lze popsat pomocí metodologického rámce používaného v projektu datové vědy, zejména metodiky CRISP-DM a iterativních přístupů typu RAD.

Metodiky zdůrazňují iterativní charakter analytického procesu, kdy jednotlivé fáze analýzy na sebe navazují, ale zároveň se mohou opakovaně vracet a upravovat na základě nových poznatků. Typický proces zahrnuje fáze porozumění problému, porozumění datům, přípravy dat, modelování a vyhodnocení modelů.

V rámci práce byly jednotlivé fáze metodiky realizovány v následující podobě: nejprve byla provedena analýza dostupného datasetu a jeho základních vlastností, následně byla realizována explorační analýza dat a analýza jejich struktury. Na základě získaných poznatků byly vytvořeny klasifikační a regresní modely pro predikci kvality vína a jejich výkon byl vyhodnocen pomocí vhodných metrik.

2.2.3 Struktura analytického workflow

Použitý analytický workflow byl realizován v prostředí Jupyter Notebook, které umožňuje kombinovat programový kód, vizualizace a textové komentáře. Notebook zajišťuje přehlednost analytického procesu a umožňuje snadnou reprodukci jednotlivých kroků analýzy (viz Příloha A.1).

Notebook byl strukturován do několika logických částí. Nejprve byla provedena kontrola kvality dat a základní statistická analýza datasetu. Následovala explorační analýza dat, která zahrnovala analýzu distribuce proměnných, korelační analýzu a metody redukce dimenzionality, jako je analýza hlavních komponent.

Další část byla věnována vytvoření a vyhodnocení predikčních modelů. V práci byly využity jak klasifikační, tak regresní přístupy, které umožňují modelovat kvalitu vína různými způsoby. Součástí práce jsou také doplňkové experimenty, které analyzují vliv různých faktorů na výkon modelů, například vliv četnosti jednotlivých tříd nebo vliv redukce vstupních atributů.

Součástí využitého prostředí Kaggle je také jeho komunitní charakter, který umožňuje sdílení datasetů, analytických notebooků a výsledků mezi jednotlivými uživateli. Otevřenost podporuje transparentnost a reprodukovatelnost datově-vědních projektů a umožňuje porovnání různých přístupů k řešení stejného problému. V rámci práce byl využit Wine Quality Data Set (Ruthgn, 2024) veřejně dostupný na platformě Kaggle (viz Příloha A.1).

Platforma Kaggle byla využita primárně jako prostředí pro realizaci vlastního analytického řešení. Postup analýzy, návrh modelů i jejich implementace byly vytvořeny samostatně bez přímého přebírání existujících řešení z komunitních notebooků. Komunitní obsah platformy byl vnímán

spíše jako potenciální zdroj inspirace a bude využit zejména pro následné porovnání dosažených výsledků s alternativními přístupy.

2.2.4 Využití nástrojů umělé inteligence v rámci řešení práce

V rámci řešení byly využity nástroje založené na umělé inteligenci jako podpůrný prostředek při realizaci jednotlivých kroků analytického procesu. Nástroje nebyly využívány jako autonomní řešení, ale jako interaktivní nástroj podporující návrh, implementaci a interpretaci jednotlivých částí analýzy. Jejich využití tak odpovídá moderním přístupům propojujícím metodiky typu Rapid Application Development s umělou inteligencí, kde AI slouží jako prostředek pro urychlení iterací a podporu experimentování v rámci vývoje aplikací a datových řešení.

Použití nástrojů umělé inteligence se promítlo především do oblasti implementace analytického workflow. V rámci programování byly zmíněné nástroje využívány pro návrh struktury kódu, generování jednotlivých částí implementace a návrh vizualizací. Požadavky na výstup byly vždy explicitně specifikovány, například ve formě požadované struktury grafů, tabulek nebo způsobu ukládání dat. Vygenerovaný kód však nebyl přebírán bez úprav, ale byl podroben detailní kontrole a následné modifikaci tak, aby odpovídal konkrétním potřebám projektu a charakteru analyzovaného datasetu.

Další oblastí využití byla podpora při řešení chyb a ladění kódu. Nástroje umělé inteligence byly využívány pro analýzu chybových hlášení a návrh možných způsobů jejich odstranění. Navržená řešení byla následně ověřována v rámci implementace a upravována podle konkrétní situace. Přístup umožnil efektivněji identifikovat příčiny problémů, avšak konečné rozhodnutí o způsobu jejich řešení bylo vždy založeno na vlastní analýze a porozumění implementovanému řešení.

V oblasti návrhu modelů strojového učení sloužily nástroje umělé inteligence jako podpůrný prostředek při výběru vhodných algoritmů a návrhu experimentů. Směr řešení, včetně volby typu úlohy (klasifikace či regrese), výběru vstupních atributů a způsobu validace modelů, byl stanoven na základě vlastního návrhu a průběžných konzultací s vedoucím práce. Nástroje umělé inteligence zde plnily roli pomocného nástroje pro generování alternativních přístupů a jejich následné porovnání.

Specifickou roli sehrály zmíněné nástroje také při interpretaci výsledků. Vlastní interpretace výsledků byla vždy formulována na základě analýzy výstupů modelů a vizualizací. Následně byla interpretace konfrontována s návrhy generovanými nástroji umělé inteligence, což umožnilo ověřit správnost závěrů a případně rozšířit pohled na analyzovaný problém. Princip lze chápat jako formu interaktivní konzultace, která podporuje kritické myšlení a validaci získaných poznatků.

Nástroje umělé inteligence byly dále využity při tvorbě textových částí notebooku, zejména pro formulaci některých vysvětlujících pasáží a zpřesnění odborné terminologie. Výsledné texty však byly vždy upraveny a přizpůsobeny kontextu práce tak, aby odpovídaly požadavkům na odborný styl a přesnost.

Z metodologického hlediska lze využití umělé inteligence v práci interpretovat jako praktickou realizaci principů propojení RAD a AI, někdy označovaných jako „application-driven AI“. Je kladen důraz na úzkou vazbu mezi návrhem řešení a konkrétní aplikační doménou, přičemž nástroje

umělé inteligence slouží k podpoře iterativního vývoje, rychlého prototypování a průběžné validace navrhovaných řešení.

Klíčovým aspektem využití AI nástrojů byla důsledná validace všech generovaných výstupů. Výstupy nástrojů umělé inteligence nebyly přebírány bez kritického posouzení, ale byly systematicky ověřovány z hlediska správnosti, konzistence a vhodnosti pro daný problém. Kontrola zahrnovala jak syntaktickou a logickou správnost kódu, tak i věcnou správnost interpretace výsledků a volby metod. Kontrola je nezbytná, protože nástroje umělé inteligence mohou generovat nepřesné nebo nevhodné návrhy, zejména pokud neberou v úvahu specifika konkrétního datasetu.

Zároveň je nutné zdůraznit, že nástroje umělé inteligence nemají přímý přístup k analyzovaným datům ani k plnému kontextu řešeného problému. Konečná rozhodnutí byla proto vždy činěna na základě vlastního porozumění problému, znalosti použitých metod a výsledků experimentů.

Celkově lze konstatovat, že využití nástrojů umělé inteligence přispělo k efektivnějšímu průběhu řešení práce, zejména z hlediska rychlosti iterací, návrhu alternativních řešení a podpory analytického myšlení. Současně však vyžadovalo aktivní a kritický přístup, který zajistil, že výsledné řešení odpovídá odborným požadavkům a není pouze nekritickým převzetím generovaných výstupů.

2.3 Data a jejich charakteristika

Analyzovaný dataset obsahuje celkem 6 497 záznamů a 13 atributů. Dataset zahrnuje fyzikálně-chemické vlastnosti jednotlivých vzorků vína a jejich senzorické hodnocení kvality. Každý řádek datasetu představuje jeden vzorek vína popsaný souborem chemických charakteristik. Většina atributů je numerického typu (float64) a reprezentuje měřené chemické vlastnosti vína. Výjimku tvoří atribut type, který je kategoriální proměnnou rozlišující typ vína (např. bílé nebo červené), a atribut quality, který představuje výsledné hodnocení kvality vína (Cortez a kol., 2009).

Pro účely práce byl využit veřejně dostupný dataset Wine Quality Data Set (Red & White Wine), dostupný na platformě Kaggle. Dataset představuje kombinovanou verzi původních dat publikovaných v rámci výzkumu (Cortez a kol., 2009), který se zabýval predikcí kvality portugalských vín typu Vinho Verde na základě fyzikálně-chemických charakteristik. Původní data pocházejí z repozitáře UCI Machine Learning Repository, kde jsou zveřejněna ve formě dvou samostatných datasetů pro červené a bílé víno (Cortez a kol., 2009).

V rámci implementace byl v interaktivním notebooku vytvořen také prvek umožňující načtení alternativního zdrojového datasetu. Uživatel tak může pracovat nejen s výchozí datovou sadou, ale také s vlastními daty ve stejném formátu, aniž by musel zasáhnout do samotného kroku.

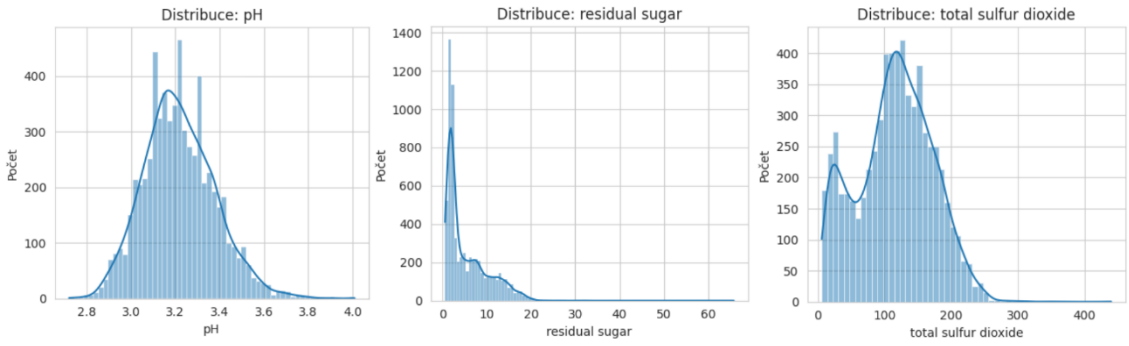
2.3.1 Popis atributů datasetu

Dataset obsahuje několik atributů popisujících fyzikálně-chemické vlastnosti jednotlivých vzorků vína. Atributy představují měřené chemické charakteristiky, které mohou mít vliv na výslednou kvalitu vína.

Konkrétně dataset obsahuje následující atributy:

- type – typ vína (bílé nebo červené),
- fixed acidity – koncentrace fixních kyselin,
- volatile acidity – koncentrace těkavých kyselin,
- citric acid – obsah kyseliny citronové,
- residual sugar – množství zbytkového cukru,
- chlorides – koncentrace chloridů,
- free sulfur dioxide – množství volného oxidu siřičitého,
- total sulfur dioxide – celkové množství oxidu siřičitého,
- density – hustota vína,
- pH – kyselost vína,
- sulphates – obsah síranů,
- alcohol – obsah alkoholu,
- quality – senzorické hodnocení kvality vína.

Pro lepší pochopení charakteru jednotlivých proměnných byla analyzována jejich distribuce pomocí histogramů.



Obr. 1: Vybrané distribuce vstupních proměnných
Zdroj: Cortez a kol. (2009), vlastní zpracování

Z důvodu přehlednosti jsou v práci uvedeny pouze vybrané reprezentativní histogramy. Kompletní sada grafů je dostupná v Příloze A.1.

Z analyzovaných histogramů je patrné, že jednotlivé proměnné vykazují různé typy rozdělení. U některých proměnných lze pozorovat rozdělení blízké normálnímu (např. pH), zatímco jiné proměnné vykazují výraznou pravostrannou šikmost (např. residual sugar). Další proměnné mají složitější průběh rozdělení s širším rozptylem hodnot (např. total sulfur dioxide).

Uvedené příklady slouží jako reprezentativní ukázky, přičemž obdobné charakteristiky lze pozorovat i u dalších atributů datasetu.

U některých proměnných jsou patrné také dlouhé ocasy rozdělení a extrémní hodnoty, což naznačuje přítomnost odlehlých pozorování.

Charakteristiky naznačují, že:

- vztahy mezi proměnnými nemusí být čistě lineární,
- některé modely mohou být citlivé na odlehlé hodnoty,
- a může být vhodné zvážit transformace dat nebo použití robustnějších modelů.

2.4 Příprava a čištění dat

Před samotnou explorační analýzou a modelováním bylo nutné provést přípravu a kontrolu kvality datasetu. Cílem fáze bylo ověřit integritu dat, identifikovat případné problémy v datasetu a připravit data ve formě vhodné pro další analytické kroky.

Proces přípravy dat zahrnoval zejména kontrolu výskytu chybějících hodnot, identifikaci duplicitních záznamů, ověření datových typů jednotlivých atributů a základní validační kontroly vybraných proměnných. Součástí fáze byla také analýza rozdělení numerických proměnných a demonstrace vlivu různých metod škálování dat.

2.4.1 Kontrola kvality dat

Prvním krokem přípravy dat byla kontrola základní kvality datasetu. Bylo ověřeno, zda dataset obsahuje chybějící hodnoty, nekonzistentní datové typy nebo jiné strukturální problémy, které by mohly negativně ovlivnit další analýzu. Detailní implementace jednotlivých kontrol je uvedena v Příloze A.1.

Kontrola datasetu ukázala, že žádný z atributů neobsahuje chybějící hodnoty. Datové typy jednotlivých proměnných odpovídají očekávání – většina vstupních proměnných je reprezentována jako numerické hodnoty typu float64, zatímco cílová proměnná quality je celočíselného typu int64. Atribut type představuje kategoriální proměnnou rozlišující typ vína.

Součástí kontroly byla také identifikace duplicitních záznamů. V datasetu bylo nalezeno 1 177 duplicitních řádků. U chemického složení vína však může být výskyt identických záznamů přirozený, například v případě opakovaných laboratorních měření nebo identických vzorků vína. Proto nebyly duplicitní záznamy odstraněny, aby nedošlo k narušení původního rozdělení datasetu.

Dále byly provedeny základní validační kontroly očekávaných rozsahů vybraných proměnných, zejména atributů alcohol, pH a density. Kontroly neodhalily žádné hodnoty mimo očekávaný rozsah.

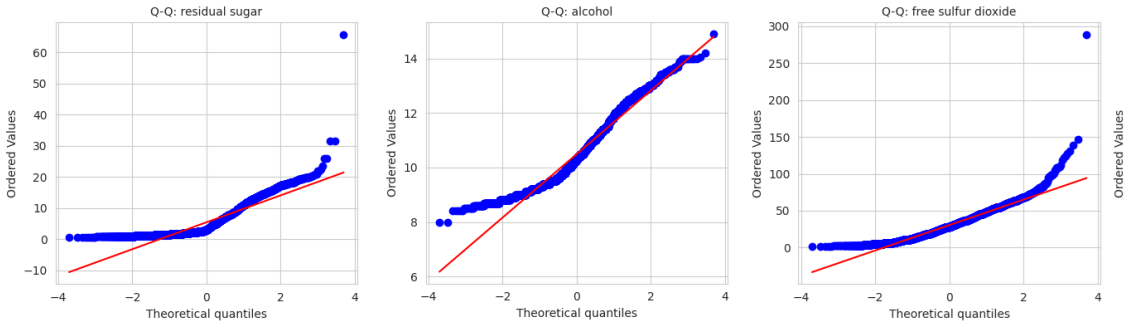
Byla rovněž ověřena konzistence cílové proměnné quality, která odpovídá známé škále hodnocení kvality vína v intervalu 3 až 9.

2.4.2 Analýza rozdělení dat a testy normality

Pro numerické proměnné byly provedeny statistické testy normality, konkrétně D'Agostino K^2 test a Anderson–Darling test. Testy umožňují posoudit, zda lze rozdělení dat považovat za přibližně normální.

Výsledky obou testů ukázaly velmi malé p-hodnoty u všech analyzovaných proměnných, což vede k zamítnutí hypotézy normálního rozdělení při hladině významnosti $\alpha = 0{,}05$.

Je však důležité poznamenat, že u relativně velkých datasetů jsou testy normality velmi citlivé a mohou indikovat statisticky významné odchylky i v případě poměrně malých rozdílů od ideálního normálního rozdělení. Následně byla statistická analýza doplněna také vizuální kontrolou pomocí Q–Q grafů.



Obr. 2: Vybrané Q-Q grafy pro posouzení normality rozdělení
Zdroj: Cortez a kol. (2009), vlastní zpracování

Z důvodu přehlednosti jsou v práci uvedeny pouze vybrané reprezentativní Q–Q grafy. Kompletní sada grafů je dostupná v Příloze A.1.

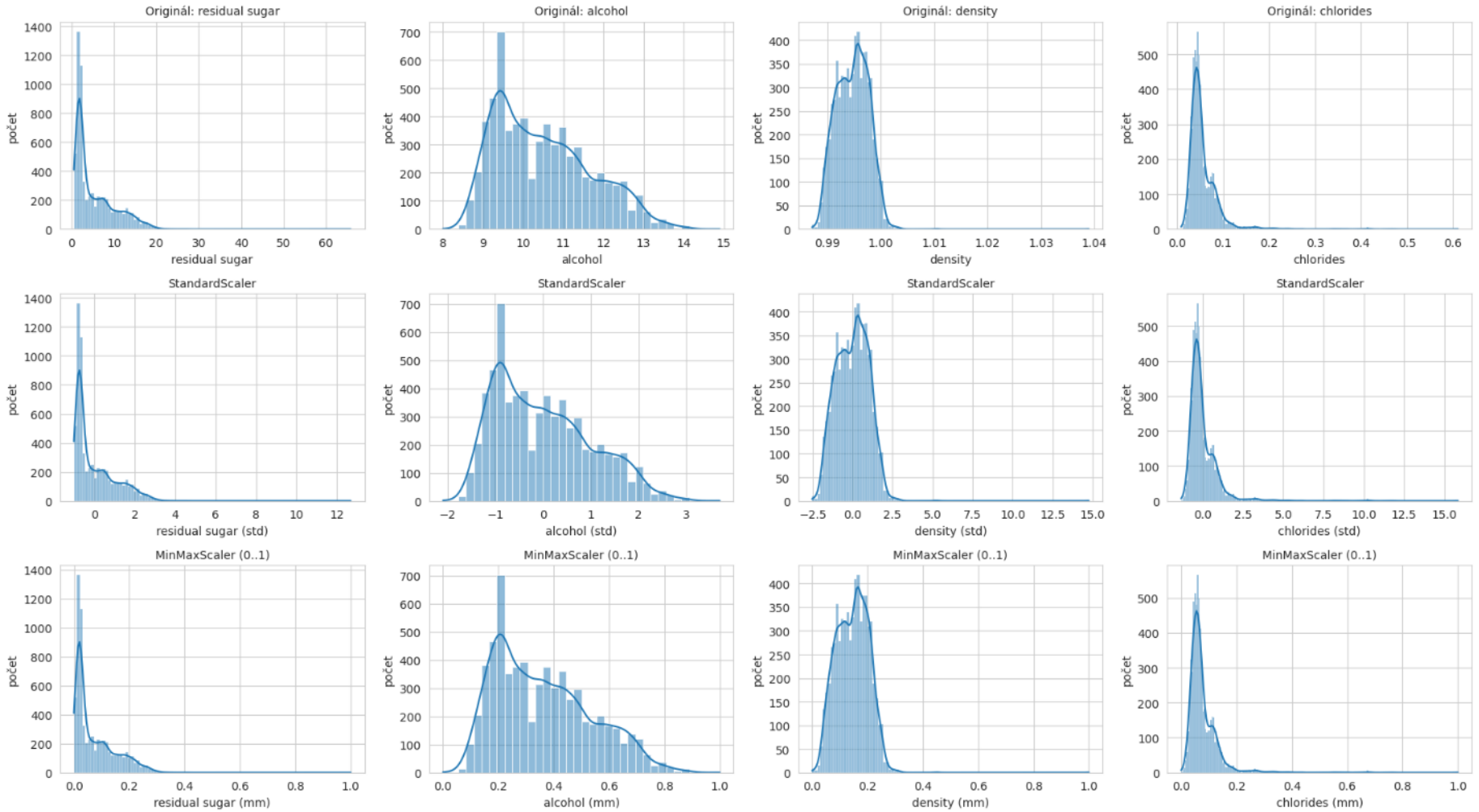
Z analyzovaných grafů je patrné, že řada proměnných vykazuje asymetrické rozdělení, často s pravostrannou šikmostí (např. residual sugar). U proměnných jsou zároveň patrné výrazné odchylky v horních kvantilech, což naznačuje přítomnost extrémních hodnot. Naopak některé proměnné vykazují rozdělení relativně blízké normálnímu (např. alcohol), přesto však dochází k určitým odchylkám zejména ve vyšších kvantilech. U vybraných atributů (např. free sulfur dioxide) lze pozorovat odchylky především v horní části rozdělení, což poukazuje na existenci dlouhých ocasů distribuce.

Uvedené příklady slouží jako reprezentativní ukázky, přičemž obdobné charakteristiky lze pozorovat i u dalších atributů datasetu. Z metodického hlediska, výsledky naznačují, že dataset obsahuje asymetrická rozdělení a přítomnost odlehlých hodnot, což vede k závěru, že při následném modelování může být vhodné využít robustnější nebo nelineární metody.

2.4.3 Normalizace a standardizace dat

Dalším krokem přípravy dat byla analýza vlivu dvou běžně používaných transformačních metod – standardizace a normalizace – na rozdělení vstupních proměnných. Metody slouží k úpravě měřítka dat a jsou důležité zejména pro algoritmy citlivé na rozsah hodnot jednotlivých atributů.

Standardizace transformuje data tak, aby měla nulový průměr a jednotkovou směrodatnou odchylku, zatímco normalizace převádí hodnoty do předem definovaného intervalu, nejčastěji $\langle 0, 1 \rangle$.



Obr. 3: Porovnání rozdělení vybraných proměnných před a po aplikaci škálovacích metod

Zdroj: Cortez a kol. (2009), vlastní zpracování

Z grafu je patrné, že standardizace zachovává tvar původního rozdělení dat, přičemž pouze mění jejich měřítko. Naopak normalizace převádí hodnoty do omezeného intervalu, což může vést ke ztrátě rozlišovací schopnosti u proměnných s výraznou šikmostí nebo přítomností extrémních hodnot. Dané případy způsobují, že většina hodnot je soustředěna do úzkého rozsahu.

Z metodického hlediska je standardizace vhodná zejména pro algoritmy citlivé na měřítko dat, například:

- PCA (Principal Component Analysis),
- K-Nearest Neighbours,
- Support Vector Machines,
- lineární modely s regularizací.

Na základě analýzy je v dalších částech práce preferována standardizace, která poskytuje stabilnější transformaci dat pro následné analytické kroky.

Na závěr fáze byla vytvořena pracovní kopie datasetu o rozměru 6 497 × 13, která je dále využívána v následujících částech analýzy. Detailní implementace transformačních kroků je uvedena v Příloze A.1.

2.5 Explorační analýza dat (EDA)

Explorační analýza dat (EDA) představuje klíčovou fázi datově-analytického procesu, jejímž cílem je získat přehled o vlastnostech datasetu, identifikovat strukturu dat, vztahy mezi proměnnými a potenciální problémy, které mohou ovlivnit následné modelování.

Na začátku EDA jsou analyzovány základní statistické charakteristiky datasetu, distribuce jednotlivých proměnných a vztahy mezi fyzikálně-chemickými vlastnostmi vína a jeho kvalitou. Analýza je doplněna vizualizacemi ve formě histogramů, korelačních matic a bodových grafů.

2.5.1 Popisná statistika datasetu

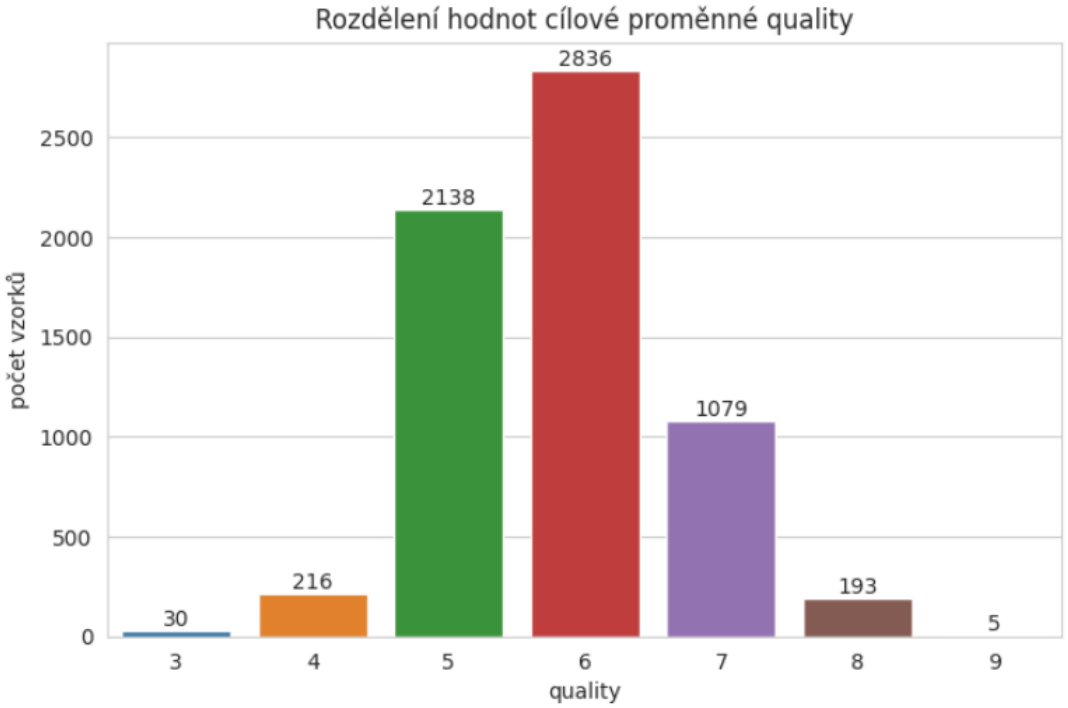
Základní statistické charakteristiky numerických proměnných poskytují přehled o rozsahu hodnot, variabilitě a centrální tendenci jednotlivých atributů datasetu (viz Příloha A.1.).

Všechny proměnné obsahují 6 497 pozorování, což potvrzuje, že dataset neobsahuje chybějící hodnoty. Průměrné hodnoty, mediány a kvartily jednotlivých proměnných poskytují základní informaci o rozložení dat.

Cílová proměnná quality nabývá hodnot v intervalu 3–9, přičemž její základní statistické charakteristiky jsou následující:

- průměr ≈ 5,82,
- medián = 6,
- směrodatná odchylka ≈ 0,87.

Hodnoty naznačují, že většina vzorků vína je hodnocena jako středně kvalitní.



Obr. 4: Rozdělení hodnot cílové proměnné quality

Zdroj: Cortez a kol. (2009), vlastní zpracování

Graf rozdělení cílové proměnné ukazuje výraznou koncentraci pozorování v kategoriích 5 a 6, zatímco extrémní hodnoty (3, 4, 8 a zejména 9) se v datasetu vyskytují výrazně méně často.

Rozdělení znamená, že dataset vykazuje nerovnováhu tříd, což může negativně ovlivnit výkon klasifikačních modelů, zejména při predikci méně zastoupených kategorií.

2.5.2 Analýza distribuce proměnných

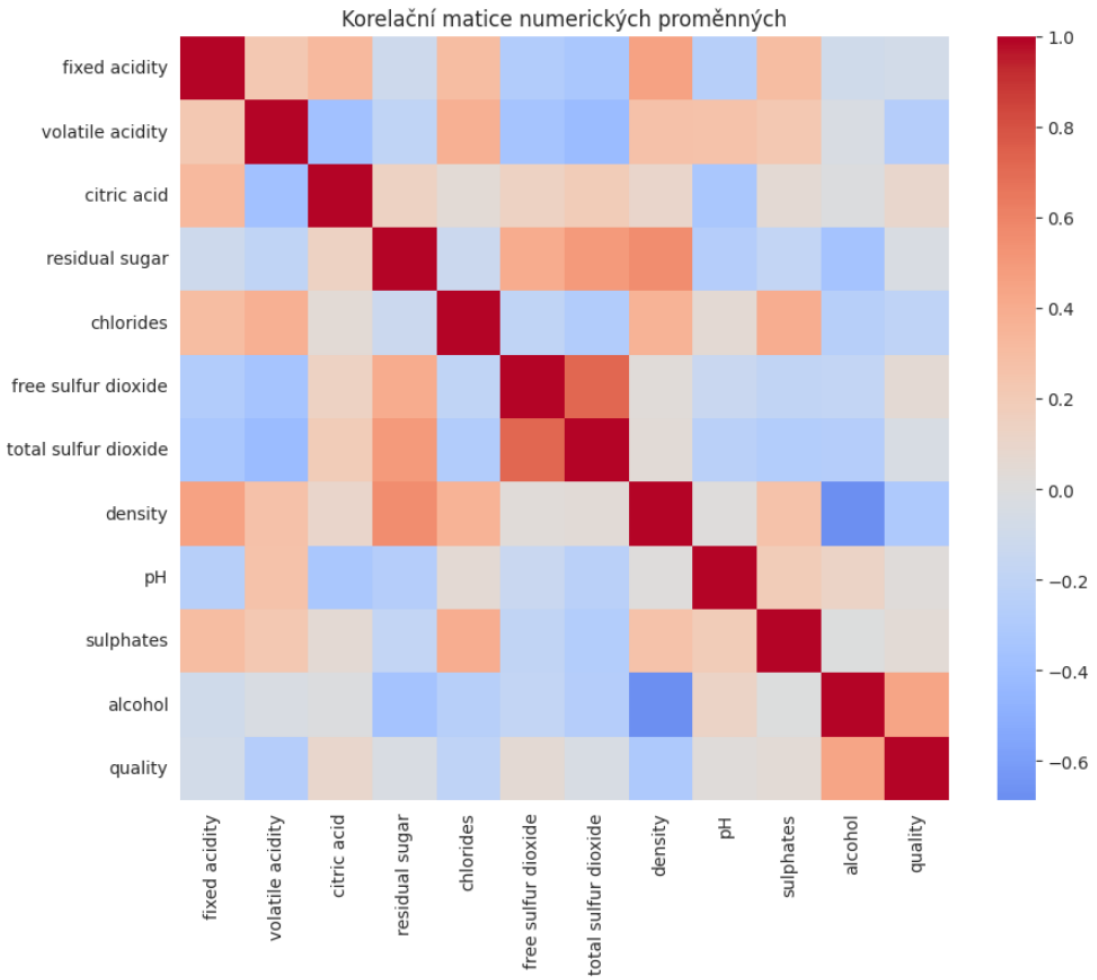
Distribuce jednotlivých fyzikálně-chemických proměnných byla analyzována pomocí histogramů a hustotních funkcí, které jsou uvedeny v předchozí části práce.

Analýza ukazuje, že většina proměnných nevykazuje ideální normální rozdělení. U řady atributů lze pozorovat pravostrannou šikmost a přítomnost extrémních hodnot, zatímco jiné proměnné jsou koncentrovány v užším intervalu hodnot.

Rozdíly v rozdělení a měřítku jednotlivých proměnných mají významný dopad na volbu metod předzpracování dat a následné modelování. Zejména představují jeden z důvodů pro použití standardizace dat, která umožňuje sjednotit měřítko vstupních atributů.

2.5.3 Vztahy mezi proměnnými

Pro analýzu vztahů mezi proměnnými byla vytvořena korelační matice numerických atributů.



Obr. 5: Korelační matice numerických proměnných
Zdroj: Cortez a kol. (2009), vlastní zpracování

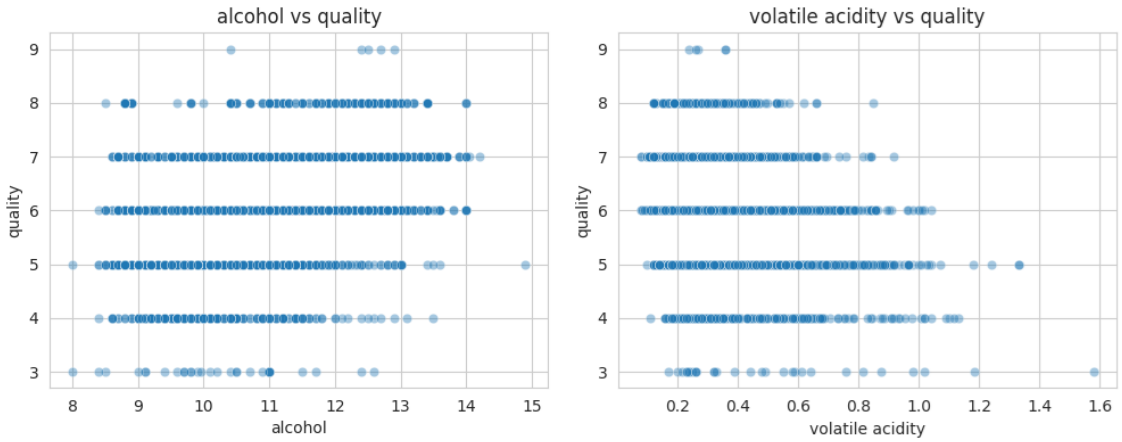
Z korelační matice je patrné, že ve vztahu k cílové proměnné quality vykazuje nejvýraznější pozitivní korelaci proměnná alcohol. Korelace naznačuje, že vyšší obsah alkoholu bývá v datasetu častěji spojen s vyšším hodnocením kvality vína.

Naopak negativní korelace s kvalitou je patrná zejména u proměnných volatile acidity a density, což znamená, že vyšší hodnoty atributů jsou častěji spojeny s nižším hodnocením kvality.

Korelační matice dále ukazuje výraznou pozitivní závislost mezi proměnnými free sulfur dioxide a total sulfur dioxide, což odpovídá jejich fyzikální interpretaci, kdy celkové množství zahrnuje i volnou frakci.

Silná korelace mezi některými vstupními proměnnými může naznačovat přítomnost multikolinearity, která může být problematická zejména pro lineární modely. Problém lze řešit například pomocí metod redukce dimenzionality.

Pro doplnění korelační analýzy byly vytvořeny scatter ploty zobrazující vztah mezi vybranými proměnnými a cílovou proměnnou quality.



Obr. 6: Vybrané vztahy mezi proměnnými a kvalitou vína

Zdroj: Cortez a kol. (2009), vlastní zpracování

Z vizualizací je patrné, že proměnná alcohol vykazuje zřetelný pozitivní trend vůči kvalitě, zatímco volatile acidity má spíše negativní vztah. U ostatních proměnných je vztah ke kvalitě méně jednoznačný a dochází k výraznému překrývání jednotlivých tříd.

Výsledky naznačují, že kvalitu vína nelze spolehlivě vysvětlit pouze jednou proměnnou, ale spíše kombinací více fyzikálně-chemických vlastností.

Kompletní sada vizualizací je dostupná v Příloze A.1.

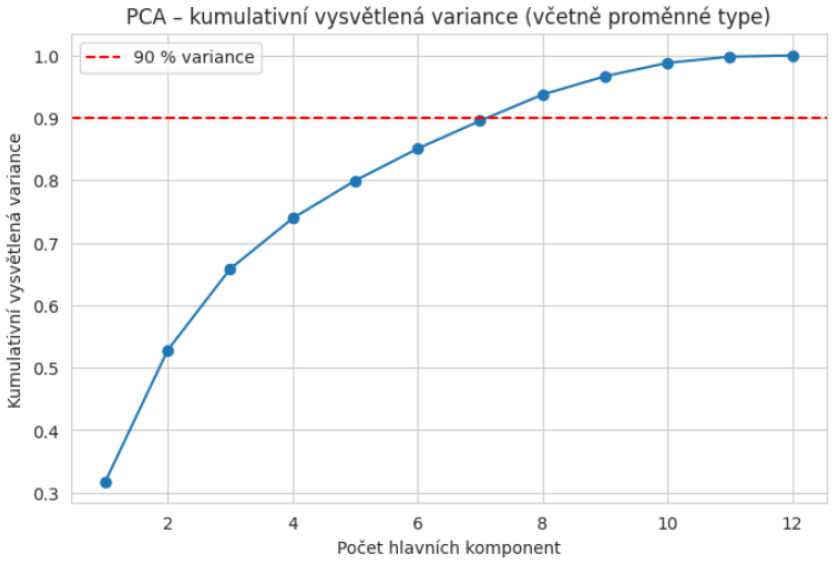
2.5.4 Analýza hlavních komponent (PCA)

Analýza hlavních komponent (Principal Component Analysis, PCA) byla v práci použita jako metoda pro redukci dimenzionality a současně jako nástroj pro analýzu vnitřní struktury dat. Cílem nebylo nahradit původní proměnné při predikčním modelování, ale zejména zjistit, do jaké míry lze variabilitu datasetu popsat menším počtem latentních faktorů a zda se v prostoru hlavních komponent projeví určité přirozené struktury mezi vzorky vín.

Před samotnou aplikací PCA byla data odpovídajícím způsobem připravena. Do analýzy byly zahrnuty numerické fyzikálně-chemické atributy vína a současně byla do binární podoby převedena také proměnná type, aby bylo možné zachytit i vliv typu vína na výslednou strukturu dat. Naopak cílová proměnná quality nebyla do PCA zahrnuta, protože PCA představuje metodu učení bez učitele a zahrnutí cílové proměnné by vedlo k nežádoucímu zkreslení interpretace výsledků.

Všechny vstupní proměnné byly před výpočtem PCA standardizovány, aby měly nulový průměr a jednotkovou směrodatnou odchylku. Krok je pro PCA zásadní, protože jednotlivé atributy mají v původním datasetu výrazně odlišná měřítka a bez standardizace by výsledné komponenty byly nepřiměřeně ovlivněny proměnnými s vyšší variabilitou.

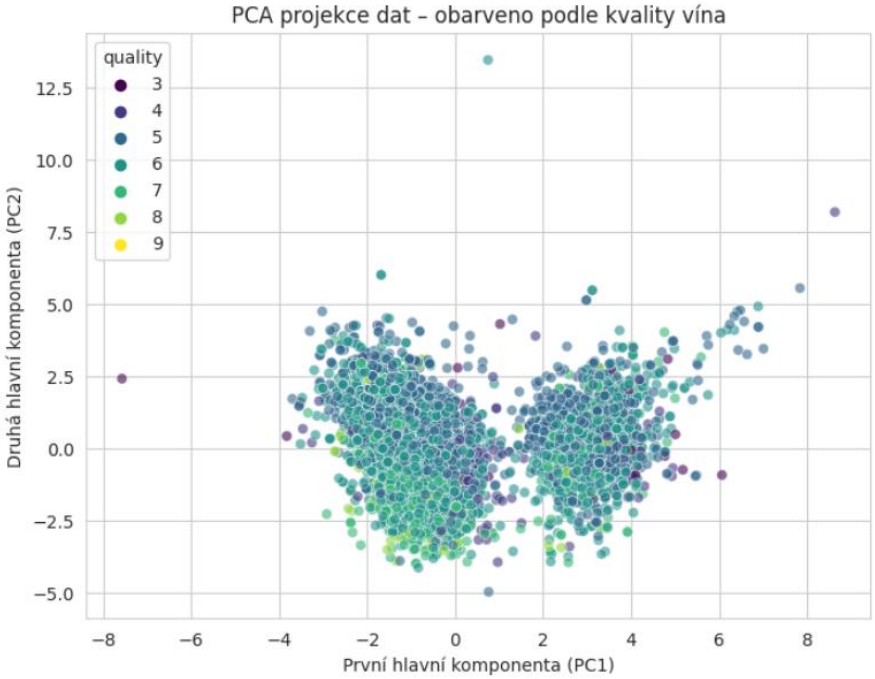
Graf kumulativní vysvětlené variance ukazuje, jaká část celkové variability datasetu je vysvětlena postupným přidáváním hlavních komponent. Z výsledků vyplývá, že přibližně 90 % celkové variability je zachyceno prvními sedmi až osmi komponentami. Konkrétně první komponenta vysvětluje přibližně 31,7 % variability, druhá komponenta dalších 21,1 % a třetí přibližně 13,0 %. Prvních osm komponent pak dohromady vysvětluje přibližně 93,7 % variability datasetu.



Obr. 7: PCA – kumulativní vysvětlená variace
Zdroj: Cortez a kol. (2009), vlastní zpracování

Výsledek ukazuje, že původní vícerozměrný dataset lze do značné míry popsat menším počtem nových proměnných, aniž by došlo k výrazné ztrátě informace. A současně potvrzuje, že mezi původními proměnnými existují korelované vztahy a určitá redundance, kterou PCA shrnuje do menšího počtu latentních faktorů.

Projekce vzorků vín do prostoru prvních dvou hlavních komponent ukazuje, jak jsou jednotlivé vzorky rozmístěny v prostoru definovaném dominantními osami variability datasetu. Jednotlivé body představují konkrétní vzorky vín a jejich barva odpovídá hodnotě cílové proměnné quality.

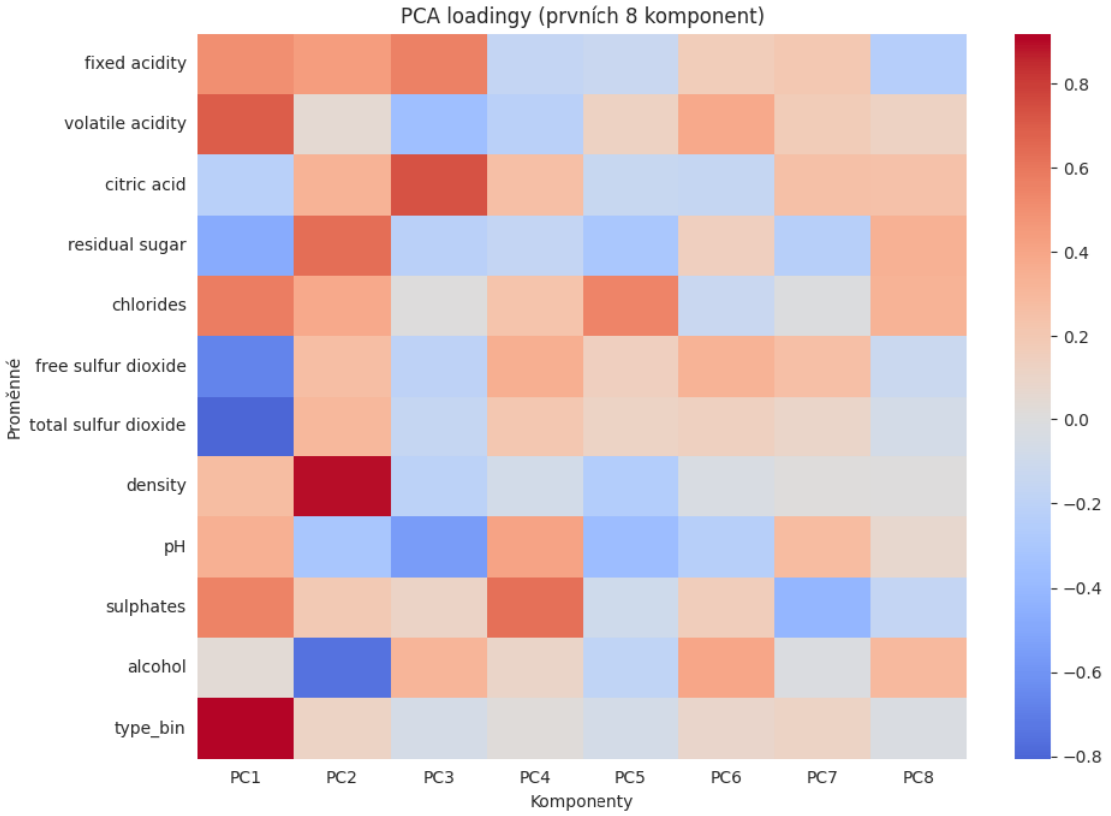


Obr. 8: PCA projekce dat
Zdroj: Cortez a kol. (2009), vlastní zpracování

Z grafu je patrné, že jednotlivé úrovně kvality se v prostoru prvních dvou hlavních komponent výrazně překrývají. Mezi sousedními třídami kvality neexistují ostré hranice a vína s podobným chemickým profilem mohou mít rozdílné hodnocení kvality. Výsledek naznačuje, že kvalita vína není jednoznačně určena pouze měřenými fyzikálně-chemickými atributy a že vztah mezi chemickým složením a výsledným hodnocením kvality je spíše komplexní.

Zmíněný poznatek je důležitý i pro další modelování. Pokud by byla kvalita silně lineárně rozdělitelná na základě několika dominantních os chemické variability, bylo by možné očekávat výraznější oddělení tříd již v prostoru prvních komponent. Jelikož k takovému oddělení nedochází, lze předpokládat, že pro modelování kvality budou vhodnější metody schopné zachytit složitější a potenciálně nelineární vztahy.

Současně je ale v grafu patrná určitá vnitřní struktura dat, která naznačuje, že dataset není homogenní a že obsahuje přirozené skupiny vzorků s podobným chemickým profilem. Poznatek tvoří přirozený přechod ke shlukové analýze.



Obr. 9: Loading proměnných

Zdroj: Cortez a kol. (2009), vlastní zpracování

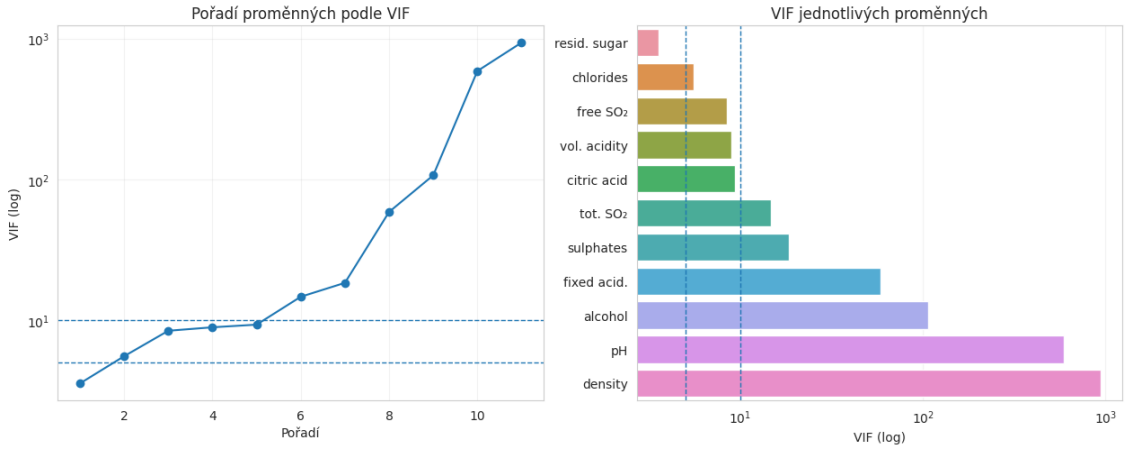
Z grafu vyplývá že první komponenta vysvětluje přibližně 31,7 % variability a představuje dominantní osu datasetu. Nejsilnější loading má proměnná type_bin, dále se výrazně uplatňují proměnné total sulfur dioxide, free sulfur dioxide, volatile acidity, chlorides a sulphates. Zjištění nám udává, že nejvýznamnějším zdrojem variability v datasetu není přímo samotná kvalita vína, ale spíše globální rozdíl mezi chemickými profily bílých a červených vín, případně s nimi spojené technologické charakteristiky. Výsledek je logický, protože spojení červených a bílých vín do jednoho datasetu přirozeně přináší vyšší globální heterogenitu.

Druhá komponenta vysvětluje přibližně 21,1 % variability. Nejvýrazněji se na ní podílejí proměnné density, alcohol, residual sugar, fixed acidity a částečně také chlorides. Daná kombinace dobře odpovídá chemicko-technologickému vztahu mezi stupněm prokvašení vína, obsahem alkoholu, hustotou a množstvím zbytkového cukru. Vyšší obsah alkoholu bývá typicky spojen s nižší hustotou a nižším množstvím zbytkového cukru. Druhou komponentu lze proto interpretovat jako osu související s fermentačním a extraktivním charakterem vína.

Třetí komponenta vysvětluje přibližně 13 % variability a je tvořena zejména proměnnými citric acid, fixed acidity a záporně orientovanou proměnnou pH. Komponenta tak odpovídá především kyselinové rovnováze vína, tedy vztahu mezi celkovou kyselostí a hodnotou pH. Z hlediska interpretace je výsledek velmi smysluplný, protože kyseliny a pH představují úzce související chemické charakteristiky, které významně ovlivňují senzorický profil i stabilitu vína.

Další komponenty (PC4 až PC8) zachycují jemnější rozdíly v chemickém složení vín a jejich podíl na celkové variabilitě je nižší. Komponenty již nepředstavují dominantní globální strukturu dat, ale doplňují detailnější chemickou charakteristiku vzorků.

Výsledky PCA úzce souvisejí také s předchozí korelační analýzou a s přítomností multikolinearity mezi vstupními atributy. Zjištění potvrzuje i VIF analýza, která ukazuje velmi vysoké hodnoty u některých proměnných, zejména density, pH, alcohol a fixed acidity.



Obr. 10: VIF proměných

Zdroj: Cortez a kol. (2009), vlastní zpracování

To znamená, že část informací obsažených v datasetu je mezi proměnnými do určité míry redundantní. PCA redundanci přirozeně redukuje způsobem, že vytváří nové, vzájemně ortogonální komponenty, které shrnují korelované informace do menšího počtu latentních proměnných.

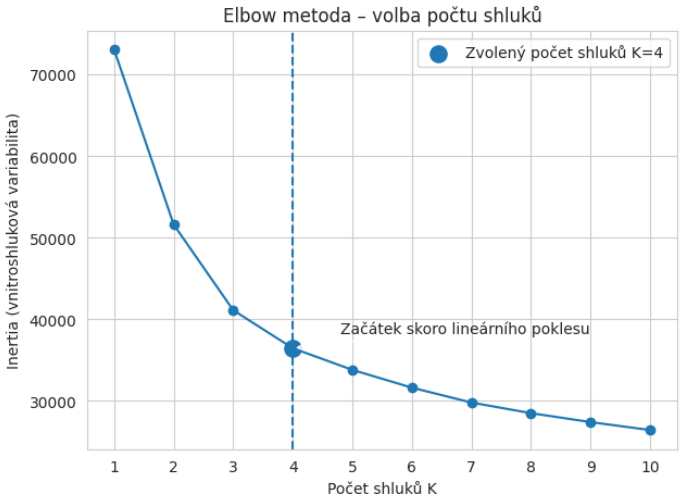
Výsledky PCA nebyly přímo využity pro redukci vstupních atributů v predikčních modelech, ale sloužily především k analýze struktury dat a interpretaci vztahů mezi proměnnými.

2.5.5 Shluková analýza (clustering)

Shluková analýza byla použita s cílem identifikovat přirozené skupiny vín na základě jejich chemického profilu. Na rozdíl od klasifikačních metod nevyužívá předem známé třídy kvality, ale snaží se nalézt podobnosti mezi vzorky čistě na základě vstupních atributů.

Shlukování bylo provedeno pomocí algoritmu K-means v prostoru hlavních komponent PCA. Postup byl zvolen proto, že PCA prostor omezuje vliv multikolinearity a částečně redukuje šum v datech, což může vést ke stabilnější a interpretovatelnější shlukové struktuře.

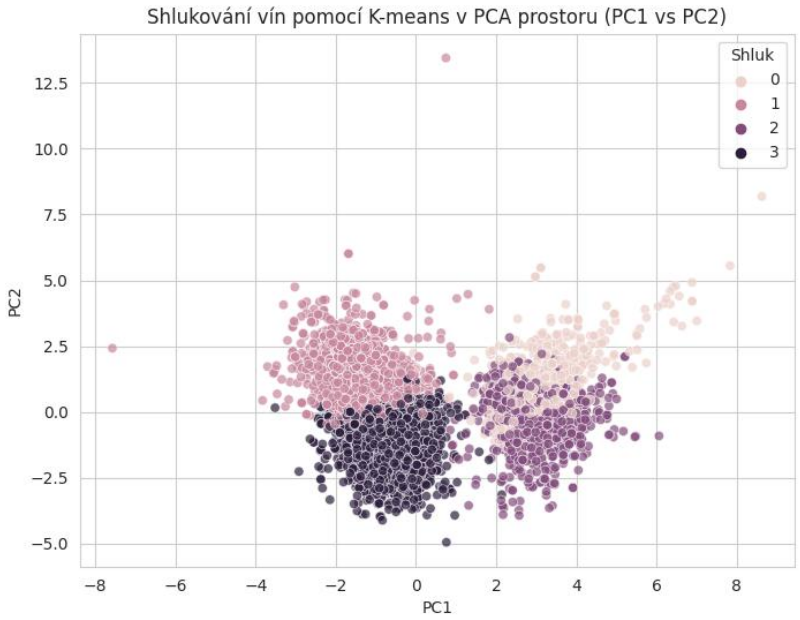
Pro odhad vhodného počtu shluků byla použita elbow metoda, která sleduje změnu vnitroshlukové variability v závislosti na počtu shluků K. Z grafu je patrný výrazný pokles variability mezi hodnotami $K = 1$ až $K = 3$, zatímco od hodnoty přibližně $K = 4$ se pokles začíná zpomalovat a další navyšování počtu shluků již nepřináší tak výrazné zlepšení.



Obr. 11: Elbow graf

Zdroj: Cortez a kol. (2009), vlastní zpracování

Na tomto základě byla zvolena hodnota K = 4, která představuje rozumný kompromis mezi zachycením hlavní struktury dat a zachováním interpretovatelnosti výsledků.



Obr. 12: Shlukování vín pomocí K-means v PCA prostoru (PC1 vs PC2)

Zdroj: Cortez a kol. (2009), vlastní zpracování

Projekce shluků do prostoru prvních dvou hlavních komponent ukazuje, že identifikované skupiny mají v PCA prostoru smysluplnou strukturu a nejsou rozmístěny zcela náhodně. Přestože

mezi některými shluky dochází k částečnému překryvu, lze pozorovat relativně odlišné oblasti odpovídající různým chemickým profilům vín.

Projekce potvrzuje, že dataset obsahuje několik základních typů chemického složení, které nejsou totožné s kategoriemi kvality, ale představují spíše objektivní chemické profily vzorků.

Shluková analýza rozdělila dataset na čtyři skupiny o nerovnoměrné velikosti:

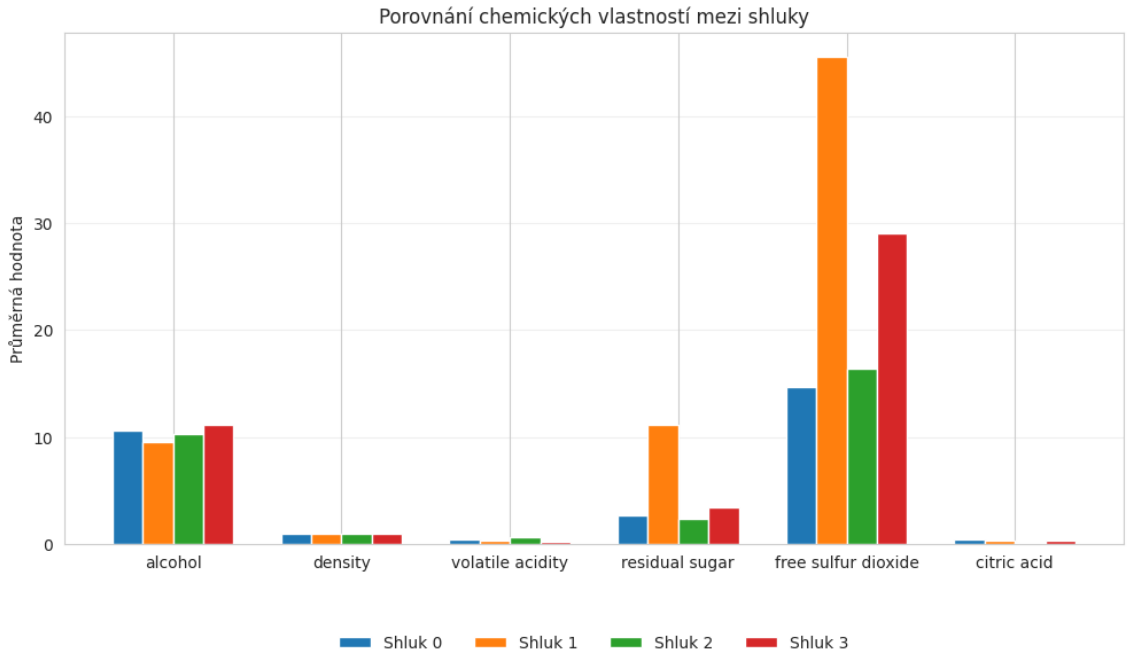
- shluk 0: 657 vzorků,
- shluk 1: 1 892 vzorků,
- shluk 2: 953 vzorků,
- shluk 3: 2 995 vzorků.

Nerovnoměrné zastoupení naznačuje existenci dominantních chemických profilů spolu s menšími specializovanými skupinami.

Průměrná kvalita v jednotlivých shlucích je:

- shluk 0: 5,87,
- shluk 1: 5,62,
- shluk 2: 5,45,
- shluk 3: 6,05.

Výsledky ukazují, že kvalita sama o sobě nevytváří samostatné shluky, ale že se liší uvnitř chemicky definovaných skupin. Nejvyšší průměrná kvalita byla pozorována ve shluku 3, zatímco nejnižší ve shluku 2.



Obr. 13: Porovnání chemických vlastností mezi shluky

Zdroj: Cortez a kol. (2009), vlastní zpracování

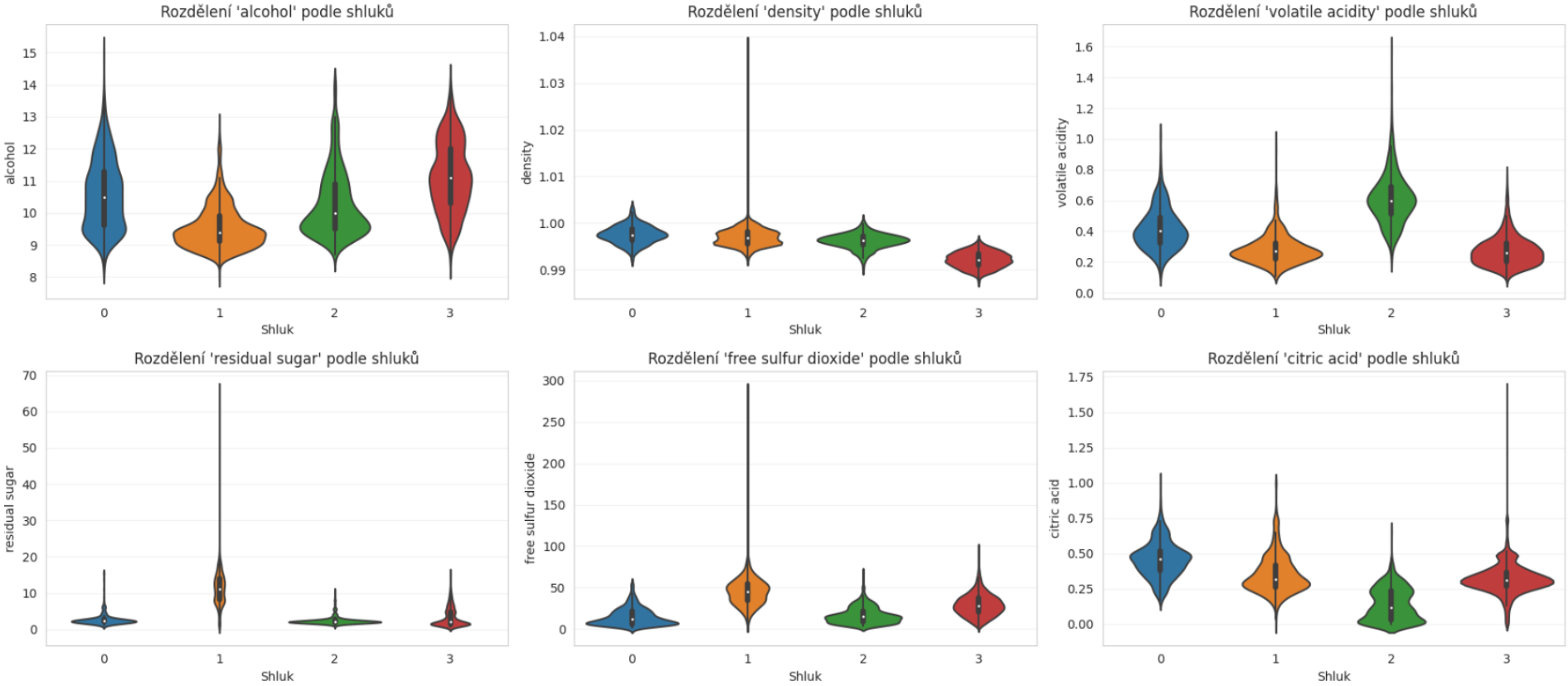
Pro interpretaci shluků byly porovnány vybrané chemické atributy: alcohol, density, volatile acidity, residual sugar, free sulfur dioxide a citric acid. Kombinace průměrných hodnot a violin plotů umožňuje posoudit nejen rozdíly v průměrech, ale také variabilitu a tvar rozdělení proměnných v jednotlivých skupinách.

Shluk 3 vykazuje nejvyšší průměrný obsah alkoholu, nejnižší hustotu a relativně nižší volatilní kyselost. Profil odpovídá vínům s vyšším stupněm prokvašení. Současně jde o shluk s nejvyšší průměrnou kvalitou, což naznačuje, že kombinace chemických vlastností je v rámci analyzovaného datasetu spojena s příznivějším hodnocením.

Shluk 1 se vyznačuje výrazně nejvyšší hodnotou residual sugar a současně nejvyšším obsahem free sulfur dioxide. Současně má nižší obsah alkoholu. Kombinace odpovídá technologicky odlišnému profilu vín, pravděpodobně s vyšší sladkostí nebo nižším stupněm prokvašení a intenzivnější stabilizací.

Shluk 2 se od ostatních odlišuje nejvyšší hodnotou volatile acidity a současně nízkým obsahem citric acid. Profil odpovídá chemicky méně vyváženým vínům a současně jde o shluk s nejnižší průměrnou kvalitou. Interpretace je konzistentní i s předchozí korelační analýzou, podle které vyšší volatilní kyselost souvisí s nižším hodnocením kvality.

Shluk 0 představuje skupinu s méně extrémním profilem a se středními hodnotami většiny analyzovaných atributů. Ve srovnání s ostatními shluky působí jako přechodová skupina mezi dominantnějšími chemickými typy vín. Průměrná kvalita je zde mírně nad průměrem celého datasetu.



Obr. 14: Violin grafy vybraných proměnných

Zdroj: Cortez a kol. (2009), vlastní zpracování

Violin ploty ukazují, že rozdíly mezi shluky nejsou dány pouze rozdílnými průměrnými hodnotami, ale také variabilitou, asymetrií rozdělení a přítomností extrémních hodnot. To doplňuje předchozí analýzu založenou na průměrných hodnotách a umožňuje detailnější pochopení vnitřní struktury jednotlivých shluků.

Zvlášť výrazné je například u:

- residual sugar, kde shluk 1 vykazuje nejen vysoký průměr, ale i široké rozdělení hodnot, což naznačuje značnou heterogenitu skupiny,
- free sulfur dioxide, kde je ve shluku 1 patrný dlouhý ocas směrem k vyšším hodnotám, což potvrzuje přítomnost extrémních pozorování,
- volatile acidity, kde je shluk 2 zřetelně posunut k vyšším hodnotám, což odpovídá jeho nižší průměrné kvalitě,
- alcohol a density, kde se mezi shluky projevuje očekávaný protichůdný vztah, přičemž vztah je patrný nejen na úrovni průměrů, ale i v celém rozdělení hodnot.

2.5.6 Lineární diskriminační analýza (LDA)

Lineární diskriminační analýza (Linear Discriminant Analysis, LDA) byla použita jako metoda učení s učitelem explorační analýzy dat, jejímž cílem bylo posoudit, do jaké míry lze jednotlivé úrovně kvality vína odlišit na základě fyzikálně-chemických vlastností. Na rozdíl od předchozích metod, jako byla PCA nebo shluková analýza, LDA při konstrukci projekčního prostoru využívá informaci o třídách, tedy v tomto případě o hodnotách cílové proměnné quality.

Cílem analýzy nebylo vytvořit finální klasifikační model, ale především:

- zjistit, zda jsou jednotlivé třídy kvality alespoň částečně lineárně separovatelné,
- identifikovat proměnné, které nejvíce přispívají k rozlišení kvality,
- porovnat, nakolik je struktura tříd kvality v souladu s chemickou strukturou datasetu zjištěnou v předchozí PCA a shlukové analýze.

Pokud by chemické vlastnosti vína plně určovaly jeho výslednou kvalitu, bylo by možné očekávat, že LDA nalezne diskriminační prostor, ve kterém budou jednotlivé třídy kvality relativně dobře odděleny. Naopak výrazné překrývání tříd by naznačovalo, že hodnocení kvality je výsledkem složitější kombinace faktorů a může obsahovat i určitý prvek subjektivity.

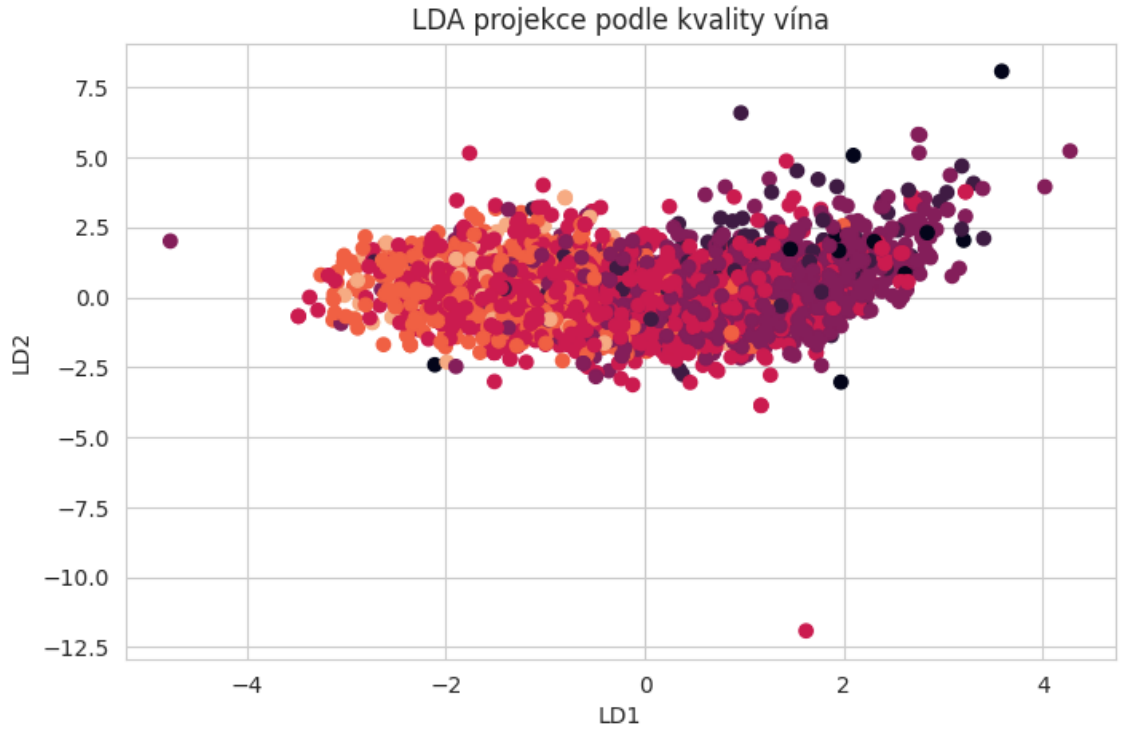
Pro lineární diskriminační analýzu byly jako vstupní proměnné použity numerické fyzikálně-chemické vlastnosti vína, zatímco proměnná quality byla použita jako cílová třída.

Před samotnou aplikací LDA byly provedeny následující kroky:

- výběr numerických vstupních proměnných,
- vyloučení kategoriálních atributů,
- standardizace všech vstupních atributů.

Standardizace byla provedena z důvodu rozdílného měřítka jednotlivých proměnných. Aby proměnné s větším rozsahem nebo vyšší variabilitou nepřiměřeně neovlivňovaly výslednou diskriminační projekci.

Vzhledem k tomu, že dataset obsahuje 7 tříd kvality, může LDA vytvořit maximálně 6 diskriminačních komponent. Vzniklo celkem 6 diskriminačních os (LD1–LD6).



Obr. 15: LDA projekce podle kvality vína

Zdroj: Cortez a kol. (2009), vlastní zpracování

Graf zobrazuje projekci jednotlivých vzorků vín do prostoru prvních dvou diskriminačních komponent LD1 a LD2, které zachycují nejvýznamnější část diskriminační informace mezi třídami kvality.

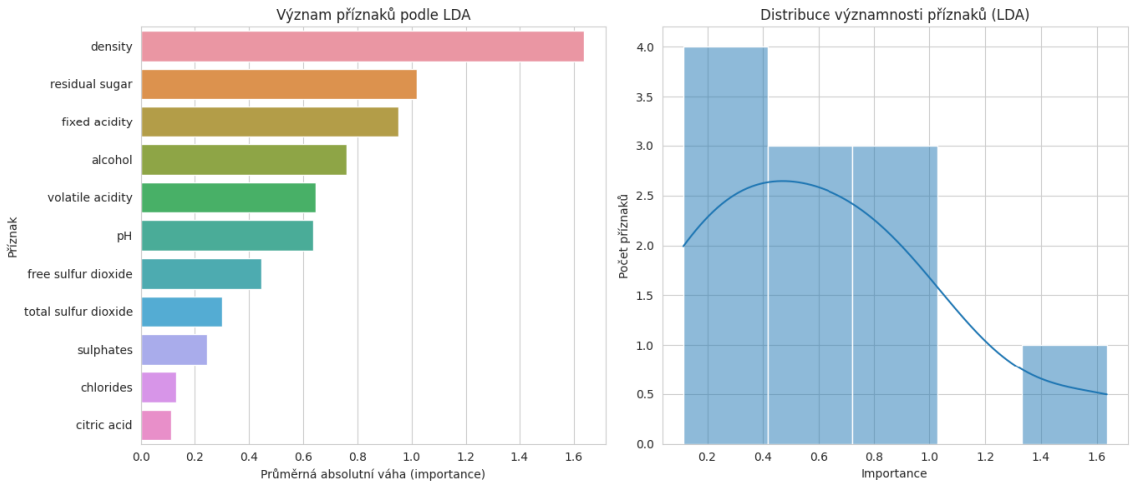
Z výsledného rozložení bodů lze pozorovat několik důležitých skutečností. Především je patrné, že podél první diskriminační osy LD1 dochází k určitému uspořádání tříd, tedy že některé úrovně kvality mají tendenci obsazovat odlišné oblasti projekčního prostoru. Nejde však o ostré rozdělení, ale spíše o plynulý přechod mezi sousedními třídami.

Současně je z grafu zřejmé, že mezi jednotlivými třídami existuje výrazný překryv. Vína různých kvalit se často nacházejí ve stejných nebo velmi blízkých oblastech diskriminačního prostoru. To znamená, že kvalita vína není na základě dostupných chemických proměnných jednoduše lineárně separovatelná.

Výsledek je v souladu s předchozí PCA analýzou, kde se také ukázalo, že chemická struktura dat nevede k ostrému oddělení jednotlivých úrovní kvality. LDA sice díky znalosti tříd zlepšuje orientaci prostoru směrem k separaci tříd, avšak ani v tomto případě nevznikají jasně oddělené skupiny.

Z interpretačního hlediska to naznačuje, že:

- chemické vlastnosti vín skutečně nesou relevantní informaci o kvalitě,
- informace však není dostatečná pro jednoduché lineární rozlišení všech tříd,
- kvalita vína je pravděpodobně výsledkem komplexní kombinace více chemických faktorů a částečně i senzorického hodnocení.



Obr. 16: Grafy distribuce a významu příznaků

Zdroj: Cortez a kol. (2009), vlastní zpracování

Vedle projekce dat umožňuje LDA také analyzovat, které proměnné nejvíce přispívají k rozlišení jednotlivých tříd kvality. Každá diskriminační komponenta je lineární kombinací původních vstupních proměnných a velikost jejích koeficientů lze využít jako orientační ukazatel důležitosti jednotlivých příznaků.

Významnost proměnných byla vyjádřena pomocí průměrné absolutní hodnoty koeficientů napříč diskriminačními komponentami. Takto získané pořadí ukazuje, které fyzikálně-chemické vlastnosti mají největší schopnost odlišovat jednotlivé úrovně kvality vína.

Nejvyšší diskriminační význam podle LDA analýzy vykazují proměnné:

- density,
- residual sugar,
- fixed acidity,
- alcohol,
- volatile acidity,
- pH.

Naopak nižší diskriminační význam vykazují například:

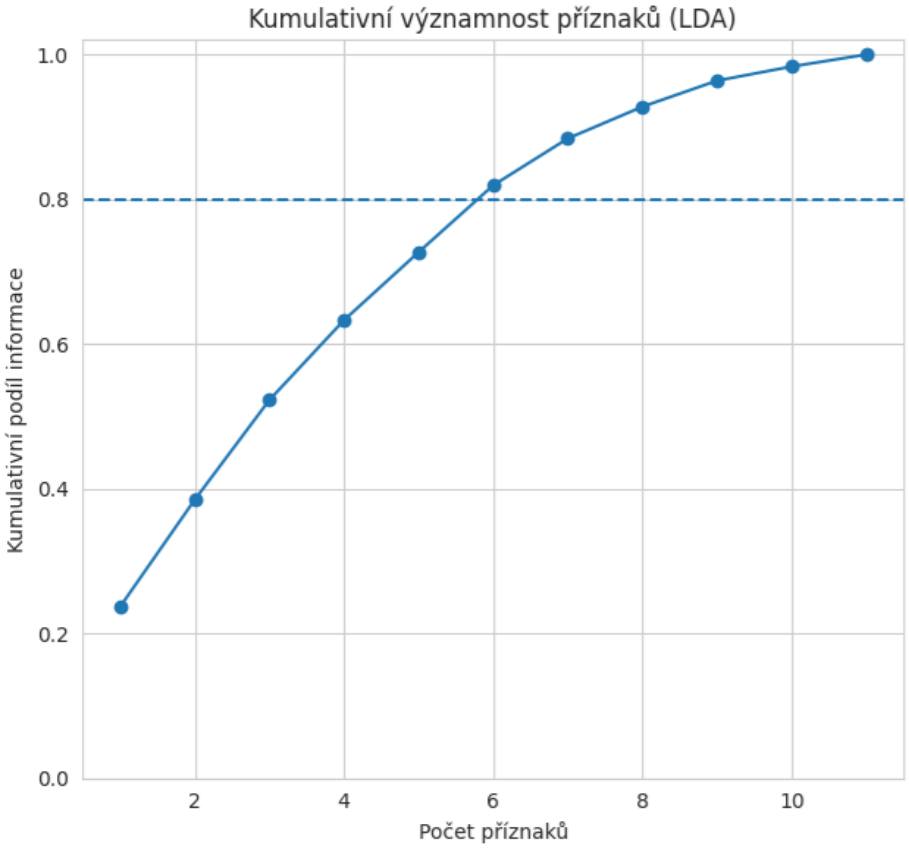
- citric acid,
- chlorides,
- sulphates.

Distribuce významnosti ukazuje, že většina proměnných má střední diskriminační schopnost, zatímco pouze několik málo atributů výrazně dominuje.

Výsledky jsou do značné míry konzistentní s předchozími kroky explorační analýzy. Už korelační analýza i PCA naznačily význam proměnných souvisejících s fermentací, hustotou, alkoholem a kyselostí. LDA nyní ukazuje, že stejné proměnné hrají důležitou roli také při samotném rozlišování tříd kvality.

Zvláště zajímavé je postavení proměnných density a alcohol. Atributy se opakovaně ukazují jako důležité jak pro celkovou strukturu dat, tak i pro odlišení úrovní kvality. To naznačuje,

že procesy související s fermentací a koncentrací rozpuštěných látek patří mezi klíčové chemické faktory spojené s hodnocením vína.



Obr. 17: Kumulativní významnost příznaků

Zdroj: Cortez a kol. (2009), vlastní zpracování

Graf kumulativní významnosti ukazuje, že přibližně prvních šest příznaků nese více než 80 % celkové diskriminační informace. To znamená, že většina informace potřebné k rozlišení kvality vína je soustředěna v relativně malé podmnožině vstupních atributů.

Z praktického hlediska je to důležitý poznatek, protože naznačuje možnost částečné redukce dimenzionality bez zásadní ztráty informace. Současně to zvyšuje interpretovatelnost dalších modelů, protože ukazuje, že ne všechny proměnné mají pro rozlišení kvality stejný význam.

Výsledky LDA dobře navazují na předchozí explorační analýzu. Již korelační analýza i PCA naznačily význam proměnných souvisejících s hustotou, alkoholem a obecně fermentačním profilem vína. Shluková analýza zase ukázala, že přirozené skupiny vín se liší právě v oblastech jako obsah alkoholu, hustota, obsah zbytkového cukru nebo volatilní kyselost. LDA nyní potvrzuje, že stejné nebo velmi podobné proměnné patří také mezi klíčové faktory při rozlišování úrovní kvality vína.

Konzistence mezi různými analytickými postupy zvyšuje důvěryhodnost výsledků. Zatímco PCA a clustering popisovaly obecnou chemickou strukturu dat, LDA ukazuje, jak se chemická struktura promítá do cílové proměnné kvality.

Výsledky LDA lze chápat také jako formu orientačního výběru příznaků. Smyslem analýzy není stanovit finální množinu atributů pro predikční modely, ale identifikovat proměnné s vysokou diskriminační schopností. Poznatky mohou sloužit jako podklad pro následné

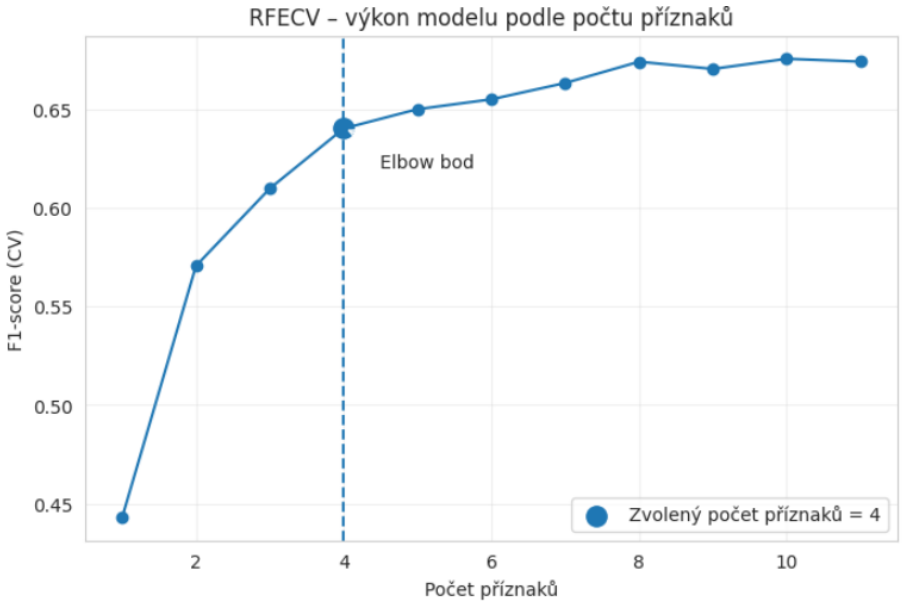
experimenty s redukovanými sadami vstupních atributů. Je však nutné zdůraznit, že významnost odvozená z LDA odráží především lineární vztahy, a proto nemusí plně vystihovat důležitost proměnných pro nelineární modely.

2.5.7 Výběr příznaků pomocí RFECV a RFE

V návaznosti na výsledky lineární diskriminační analýzy byl proveden doplňkový výběr příznaků pomocí metod Recursive Feature Elimination (RFE) a Recursive Feature Elimination with Cross-Validation (RFECV). Cílem kroku bylo identifikovat menší podmnožinu vstupních atributů, která si zachová většinu predikční informace.

Na rozdíl od LDA, která hodnotí diskriminační význam proměnných z hlediska lineární separace tříd, RFECV posuzuje význam příznaků na základě skutečného výkonu predikčního modelu. Jako základní estimator byl vybrán model Random Forest, který je vhodný pro zachycení nelineárních vztahů a je relativně robustní vůči multikolinearitě mezi vstupními proměnnými.

Metoda RFECV iterativně odstraňuje méně důležité příznaky a po každém kroku vyhodnocuje výkon modelu pomocí křížové validace. Výsledkem je závislost validačního skóre na počtu použitých atributů.



Obr. 18: RFECV – výkon modelu podle počtu příznaků

Zdroj: Cortez a kol. (2009), vlastní zpracování

Z grafu je patrné, že validační výkon modelu roste výrazně při zahrnutí prvních několika příznaků, zatímco při dalším navyšování jejich počtu se tempo zlepšení postupně zpomaluje. Nejvyšší výkon je dosažen přibližně při 10 příznacích, avšak již kolem hodnoty 4 příznaků je patrný bod zpomalení růstu. Bod lze interpretovat jako praktický kompromis mezi přesností modelu a jeho složitostí.

Na základě pozorování byla jako doplňková redukovaná varianta sledována sada čtyř příznaků, konkrétně volatile acidity, total sulfur dioxide, density a alcohol. Atributy reprezentují klíčové chemické a technologické charakteristiky vína, zejména kyselinový profil, obsah sirných sloučenin, hustotu a fermentační proces.

Výsledky RFECV a RFE jsou konzistentní s předchozí LDA analýzou, protože i zde se mezi nejvýznamnějšími proměnnými objevují atributy související s hustotou, alkoholem a kyselinovým profilem vína. Shoda napříč různými metodami zvyšuje důvěryhodnost závěrů explorační analýzy.

Výsledky analýzy budou dále využity v experimentální části práce pro porovnání výkonu modelů při použití plné a redukované množiny vstupních atributů.

2.6 Příprava dat pro modelování

Po dokončení explorační analýzy dat následovala příprava datasetu pro úlohy učení s učitelem. Data byla upravena tak, aby bylo možné korektně trénovat a vyhodnocovat jak regresní modely, tak multitřídní klasifikační modely pro predikci kvality vína na základě jeho fyzikálně-chemických vlastností.

Cílem bylo zajistit konzistentní předzpracování dat, zabránit úniku informace z testovací množiny (data leakage) a umožnit férové porovnání jednotlivých modelů i jejich variant. Klíčovým krokem bylo oddělení dat používaných pro trénování a ladění modelů od dat určených výhradně pro finální vyhodnocení.

V rámci práce byla jako výchozí použita plná sada vstupních proměnných, která zahrnuje všechny dostupné fyzikálně-chemické atributy vína a proměnnou typu vína. Redukovaná sada příznaků identifikovaná v předchozí explorační analýze byla využívána pouze doplňkově v experimentální části práce a nepředstavuje hlavní variantu modelování.

Cílová proměnná quality byla zpracována dvěma způsoby. V regresní úloze byla interpretována jako spojitá numerická hodnota, zatímco v klasifikační úloze byla chápána jako diskrétní třída odpovídající jednotlivým úrovním kvality. Přístup umožňuje porovnat dva odlišné modelovací přístupy aplikované na stejný problém.

2.6.1 Rozdělení datasetu na trénovací a testovací množinu

Prvním krokem přípravy dat pro modelování bylo rozdělení datasetu na trénovací a testovací množinu. Smyslem rozdělení je oddělit data, která slouží k učení a ladění modelů, od dat, která jsou vyhrazena pouze pro jejich závěrečné nezávislé vyhodnocení.

Rozdělení datasetu bylo provedeno samostatně pro regresní a klasifikační variantu úlohy. U regresní úlohy bylo použito běžné náhodné rozdělení dat. U klasifikační úlohy bylo navíc využito stratifikované dělení podle cílové proměnné quality, aby bylo zachováno podobné zastoupení jednotlivých tříd kvality v trénovací i testovací množině.

Krok je zvláště důležitý, protože cílová proměnná quality není v datasetu zastoupena rovnoměrně a některé třídy se vyskytují výrazně méně často než jiné. Stratifikace proto pomáhá zajistit, že testovací i trénovací část budou z hlediska rozdělení tříd vzájemně srovnatelné.

V rámci implementace byla v přiloženém Jupyter Notebooku k dispozici také možnost interaktivní volby velikosti testovací množiny, která umožňovala ověřit stabilitu výsledků při různých poměrech trénovacích a testovacích dat. V práci byl jako výchozí použit standardní poměr 80:20, tedy 80 % dat pro trénování a 20 % pro testování.

Z metodického hlediska je důležité, že testovací množina nebyla v žádném kroku používána pro ladění hyperparametrů ani pro výběr modelu. Sloužila výhradně pro finální vyhodnocení již natrénovaných a naladěných modelů. A bylo zabráněno zkreslení výsledků a nadhodnocení skutečné generalizační schopnosti modelů.

V rámci implementace byl v přiloženém Jupyter Notebooku vytvořen interaktivní prvek umožňující dynamickou volbu velikosti testovací množiny. Uživatel tak může měnit poměr mezi trénovacími a testovacími daty a sledovat, jak změna poměru ovlivňuje výkon modelů (viz Příloha A.1).

2.6.2 Vytvoření pipeline pro předzpracování dat

Pro zajištění konzistentního a bezpečného předzpracování dat byla vytvořena preprocessing pipeline založená na objektu ColumnTransformer. Přístup umožňuje odděleně zpracovávat numerické a kategoriální proměnné a zároveň zajistit jednotnou aplikaci transformací napříč všemi modely.

Vstupní atributy byly rozděleny na dvě skupiny:

- numerické proměnné reprezentující fyzikálně-chemická měření vína,
- kategoriální proměnné, konkrétně atribut type rozlišující červené a bílé víno.

Numerické proměnné byly transformovány pomocí StandardScaler, který převádí data na nulový průměr a jednotkovou směrodatnou odchylku. Transformace je důležitá zejména pro algoritmy citlivé na měřítko dat, například lineární modely, Support Vector Machines nebo metody založené na vzdálenostech, jako je K-Nearest Neighbours.

Kategoriální proměnné byly zpracovány pomocí OneHotEncoder, který převádí kategorické hodnoty do binární reprezentace.

Z metodického hlediska pipeline významně omezuje riziko data leakage, protože všechny transformační kroky jsou uvažovány pouze na trénovacích datech a následně aplikovány na testovací množinu. Čímž je zajištěno, že testovací data zůstávají nezávislá a výsledky modelů odpovídají realistickému scénáři jejich nasazení.

Pipeline zároveň zvyšuje reprodukovatelnost celého workflow, protože propojuje předzpracování a modelování do jednoho konzistentního řetězce. Přístup je klíčový zejména při porovnávání více modelů, jelikož zajišťuje, že všechny modely pracují se stejně připravenými daty. Pipeline byla následně integrována do jednotlivých modelů prostřednictvím objektu Pipeline z knihovny scikit-learn.

2.6.3 Strategie validace modelů

Pro ladění hyperparametrů a průběžné vyhodnocování modelů byla použita křížová validace, aplikovaná výhradně na trénovací množinu. Umožňuje efektivnější využití dostupných dat a poskytuje robustnější odhad výkonnosti modelu než jednorázové rozdělení na trénovací a validační část.

Pro ladění hyperparametrů byly použity dvě validační strategie v závislosti na charakteru úlohy:

- pro regresi metoda KFold,
- pro klasifikaci metoda StratifiedKFold, která zachovává poměr tříd v jednotlivých validačních foldách.

V obou případech byla validace nastavena jako 4-fold cross-validation se zapnutým mícháním dat (shuffle=True) a pevně nastaveným náhodným seedem. Použití pevného seed zajišťuje reprodukovatelnost experimentů a umožňuje opakované spuštění analýzy se stejným rozdělením dat.

Použití stratifikované křížové validace u klasifikační úlohy je metodicky důležité vzhledem k nerovnoměrnému zastoupení tříd cílové proměnné quality. Stratifikace zajišťuje, že každá validační folda obsahuje podobné rozdělení tříd jako celá trénovací množina, čímž se snižuje riziko zkreslení validačních výsledků.

Celkový workflow modelování byl nastaven následovně:

1. dataset byl rozdělen na trénovací a testovací část,
2. preprocessing pipeline byla učena pouze na trénovacích datech,
3. na trénovací množině probíhalo ladění modelů pomocí křížové validace,
4. testovací množina byla použita výhradně pro finální vyhodnocení modelu.

Stratifikace zajišťuje korektní oddělení fází výběru modelu a závěrečného testování a minimalizuje riziko nadhodnocení dosažených výsledků.

2.7 Klasifikační modely

Klasifikační část byla navržena tak, aby umožňovala férové porovnání různých modelovacích přístupů, od jednodušších lineárních modelů až po nelineární a ensemble metody. Současně byl kladen důraz na metodicky korektní workflow, které odděluje fázi benchmarku, ladění hyperparametrů a finálního testování na nezávislé testovací množině.

2.7.1 Formulace klasifikační úlohy

Klasifikační úloha je v práci formulována jako problém multitřídní klasifikace, ve kterém vstup tvoří fyzikálně-chemické vlastnosti vína a výstupem je třída odpovídající hodnotě proměnné quality. Jednotlivé třídy odpovídají ordinalním hodnotám kvality, které se v datasetu vyskytují na škále od 3 do 9.

Z metodického hlediska jde o klasifikaci, která má několik specifik:

- cílová proměnná je diskrétní, ale současně ordinální,
- jednotlivé třídy nejsou v datasetu zastoupeny rovnoměrně,
- sousední třídy si jsou z hlediska chemického složení často velmi blízké, což bylo potvrzeno také lineární diskriminační analýzou (LDA), která ukázala výrazné překrývání tříd v diskriminačním prostoru
- problém nelze předpokládat jako jednoduše lineárně rozdělitelný.

Klasifikační přístup je vhodný zejména tehdy, když má být výstup interpretován jako konkrétní kategorie kvality a kdy je cílem vytvořit model, který bude schopen víno přiřadit do odpovídající

kvalitativní třídy. Oproti regresnímu přístupu tak poskytuje přímo rozhodovací výstup ve formě diskrétního zařazení.

Formulace úlohy zároveň dobře odpovídá praktickému použití v doméně hodnocení vína, kde může být zajímavější určit, do které úrovně kvality daný vzorek patří¨, než pouze odhadnout desetinnou numerickou hodnotu.

2.7.2 Implementační a metodické principy klasifikační analýzy

Klasifikační analýza byla realizována v prostředí Python s využitím knihovny scikit-learn. Implementace byla navržena tak, aby byla reprodukovatelná, metodicky konzistentní a odolná vůči běžným chybám při experimentování.

Každý klasifikátor byl implementován ve formě pipeline, která spojuje předzpracování dat a model do jednoho celku. Přístup zajišťuje, že všechny transformace jsou učeny pouze na trénovacích datech.

Součástí workflow bylo také využití Out-of-Fold (OOF) predikcí při benchmarku modelů. Každá predikce použitá pro výpočet validačních metrik byla získána modelem, který daný vzorek neviděl při trénování, což poskytuje realističtější odhad výkonu modelu na neznámých datech.

2.7.3 Použité metriky klasifikace

Pro hodnocení klasifikačních modelů bylo využito více metrik, protože v úloze vícetřídní klasifikace s nerovnoměrným zastoupením jednotlivých tříd by použití jediné metriky mohlo vést ke zkreslené interpretaci výsledků. Jednotlivé metriky totiž zachycují různé aspekty výkonu modelu, a jejich kombinace proto umožňuje komplexnější a metodicky vhodnější posouzení kvality klasifikace.

Základní metrikou byla accuracy, která vyjadřuje podíl správně klasifikovaných vzorků na celkovém počtu pozorování. Výhodou metriky je její jednoduchost, intuitivní interpretace a snadné porovnání mezi jednotlivými modely. V případě nevyvážených tříd však může být její vypovídací schopnost omezená, protože i model s relativně vysokou hodnotou accuracy může selhávat při predikci méně četných kategorií.

Dále byly využity metriky precision, recall a F1-score, a to ve variantě weighted average, která při výpočtu zohledňuje četnost jednotlivých tříd v datasetu. Volba je vhodná zejména tehdy, když některé kategorie kvality obsahují výrazně více vzorků než jiné. Zvláště weighted F1-score zohledňuje jak přesnost, tak úplnost klasifikace napříč všemi kategoriemi.

Další použitou metrikou byla ROC-AUC ve variantě multiclass One-vs-Rest (OvR). V tomto přístupu je každá třída postupně považována za pozitivní, zatímco všechny ostatní třídy tvoří společně negativní skupinu. Pro každou takto definovanou binární úlohu je následně vypočtena ROC křivka a odpovídající plocha pod křivkou. Výsledná hodnota poskytuje doplňující informaci o schopnosti modelu rozlišovat jednotlivé třídy na základě jejich predikčních skóre nebo pravděpodobností, nikoli pouze na základě finálního přiřazení třídy. V multitřídní úloze tak ROC-AUC rozšiřuje interpretaci výsledků o pohled na separační schopnost modelu.

Vedle bodových odhadů metrik bylo v rámci rozšířeného statistického hodnocení využito také opakované validační schéma Repeated Stratified K-Fold, které umožnilo sledovat variabilitu

výsledků při různých rozděleních trénovacích dat. Pro jednotlivé metriky tak byly kromě průměrných hodnot stanoveny také 95% intervaly spolehlivosti, což přispívá k robustnějšímu porovnání modelů a umožňuje posoudit nejen jejich průměrný výkon, ale i stabilitu.

Kromě numerických metrik byla jako diagnostický nástroj využívána také confusion matrix. Umožňuje podrobněji analyzovat, mezi kterými třídami dochází k nejčastějším záměnám, a doplňuje tak čistě numerické hodnocení o kvalitativní pohled na charakter chyb modelu. V kontextu datasetu vín je očekávatelné, že k nejčastějším záměnám dochází mezi sousedními třídami kvality, například mezi hodnotami 5 a 6 nebo 6 a 7, zatímco okrajové a málo zastoupené třídy bývají z hlediska klasifikace problematičtější. Confusion matrix proto slouží jako důležitý prostředek pro interpretaci praktického chování modelu nad rámec souhrnných metrik.

Celkově zvolená sada metrik umožnila hodnotit klasifikační modely z více hledisek — z pohledu celkové úspěšnosti, rovnováhy mezi přesností a úplností klasifikace, schopnosti rozlišovat jednotlivé třídy i stability výsledků napříč validačními běhy. Přístup poskytuje metodicky vhodný základ pro férové porovnání modelů v úloze predikce kvality vína.

2.7.4 Přehled použitých klasifikačních algoritmů

Pro klasifikační úlohu byla zvolena sada algoritmů reprezentujících různé modelovací přístupy. Cílem bylo porovnat lineární, nelineární i vzdálenostní metody a získat tak širší pohled, jaký typ modelu je pro danou úlohu nejvhodnější.

Logistická regrese ve variantě multinomial byla použita jako lineární baseline model pro multitřídní klasifikaci. Slouží především jako referenční přístup, vůči němuž lze porovnávat složitější algoritmy. Její výhodou je dobrá interpretovatelnost a stabilita, zatímco omezením je schopnost zachytit pouze lineární vztahy mezi proměnnými.

Rozhodovací strom byl zařazen jako zástupce nelineárních modelů schopných zachytit interakce mezi proměnnými. Současně poskytuje relativně dobře interpretovatelnou strukturu rozhodování. Jeho nevýhodou je vyšší náchylnost k přeučení, zejména pokud není vhodně regulována jeho složitost.

Support Vector Machine (SVM) s RBF jádrem byl zvolen jako reprezentant modelů schopných vytvářet komplexní nelineární rozhodovací hranice. V explorativní analýze nebyla zjištěna jednoznačná lineární separace tříd, proto představuje model vhodného kandidáta pro zachycení složitějších vztahů v datech. Jeho výkon je však citlivý na volbu hyperparametrů a na správné škálování vstupních proměnných.

K-Nearest Neighbours (KNN) byl zařazen jako jednoduchý model založený na podobnosti mezi vzorky, který nevyžaduje explicitní učení parametrů. V kontextu práce slouží jako kontrast k modelům založeným na učení parametrických vztahů a umožňuje posoudit, do jaké míry lze kvalitu vína odhadovat na základě lokální struktury dat.

2.7.5 Ensemble metody pro klasifikaci

Vedle samostatných klasifikátorů byly analyzovány také ensemble metody, jejichž cílem je kombinovat více modelů a využít jejich komplementární vlastnosti.

Random Forest byl vybrán jako reprezentant stromového ensemble přístupu vhodného pro tabulková data. V kontextu úlohy představuje důležitý kontrast k lineárním modelům i metodám založeným na vzdálenosti, protože umožňuje zachytit nelineární vztahy a interakce mezi fyzikálně-chemickými atributy bez nutnosti jejich explicitního modelování. Současně lze očekávat, že bude relativně odolný vůči šumu v datech i vůči částečné multikolinearitě mezi proměnnými.

Soft Voting bylo zařazeno jako jednoduchý způsob kombinace více odlišných klasifikátorů. Smyslem přístupu je ověřit, zda může pouhé průměrování pravděpodobnostních výstupů několika modelů vést ke stabilnějším výsledkům než použití jednotlivých modelů samostatně. Metoda je vhodná zejména v situaci, kdy jednotlivé klasifikátory zachycují různé aspekty struktury dat a dopouštějí se odlišných typů chyb. Soft Voting byl realizován kombinací následujících základních modelů: logistické regrese, rozhodovacího stromu, SVM s RBF jádrem a metody KNN.

Stacking představuje pokročilejší ensemble strategii, která byla do práce zařazena s cílem ověřit, zda lze kombinaci základních modelů využít ještě efektivněji než u prostého průměrování. V práci byl stacking implementován tak, aby meta-model pracoval nad predikovanými pravděpodobnostmi základních klasifikátorů. Tento přístup umožňuje využít nejen samotné finální predikce, ale také míru jistoty jednotlivých modelů při rozhodování o třídě, a může tak lépe zachytit jejich vzájemnou komplementaritu. Stejně jako u metody Soft Voting tvořily základní vrstvu modely logistické regrese, rozhodovací strom, SVM s RBF jádrem a KNN. Jako meta-model byla zvolena logistická regrese, která se na základě výstupů těchto modelů učí optimálně kombinovat jejich predikce.

Ensemble metody byly do analýzy zařazeny proto, že úloha predikce kvality vína vykazuje znaky problému, u něhož nelze předpokládat dominanci jediného jednoduchého modelovacího principu. Překrývání sousedních tříd, nevyváženost kategorií i přítomnost nelineárních vztahů naznačují, že kombinace více klasifikátorů může představovat vhodný způsob, jak zvýšit robustnost modelu i jeho schopnost generalizace.

2.7.6 Strategie benchmarku, ladění a finálního testování

Vyhodnocení klasifikačních modelů bylo rozděleno do tří postupných fází tak, aby nedocházelo k optimistickému zkreslení výsledků.

V první fázi byly všechny kandidátní modely porovnány pomocí stratifikované křížové validace pouze na trénovací množině. Benchmark sloužil k férovému porovnání modelů za stejných podmínek a k identifikaci nejslibnějších kandidátů.

Metriky byly počítány z validačních predikcí, tedy z predikcí na foldách, které model neviděl při trénování. Takto získané hodnoty představují konzervativnější a metodicky správnější odhad výkonu než vyhodnocení přímo na datech použitých pro trénování.

Ve druhé fázi bylo provedeno ladění hyperparametrů pomocí GridSearchCV, a to pouze u nejlepších modelů vybraných v benchmarkové části. Daný postup byl zvolen z důvodu omezení výpočetní náročnosti a současně pro snížení rizika experimentálního overfittingu, který by mohl vzniknout při plošném ladění všech modelů bez předchozí selekce.

GridSearchCV využívala stejnou stratifikovanou validační strategii jako benchmark, čímž byla zachována metodická konzistence porovnání.

Ve třetí a poslední fázi byly vybrané modely vyhodnoceny na nezávislé testovací množině, která nebyla použita ani při benchmarku, ani při ladění hyperparametrů. Fáze poskytuje nejrealističtější odhad generalizační schopnosti výsledných modelů a zajišťuje, že výběr modelu i ladění hyperparametrů probíhaly výhradně na trénovacích datech, zatímco testovací množina sloužila pouze k závěrečnému reportování výkonu.

2.7.7 Omezení klasifikační úlohy a interpretace varování

Při aplikaci stratifikované křížové validace se v průběhu experimentů objevovalo varování knihovny scikit-learn informující o tom, že nejméně zastoupená třída v cílové proměnné obsahuje pouze 3 vzorky, což je méně než počet validačních foldů.

Varování je důsledkem výrazné nevyváženosti tříd cílové proměnné quality, zejména u extrémních hodnot kvality, které se v datasetu vyskytují jen velmi vzácně. V situaci nelze zaručit, že každá validační folda bude obsahovat reprezentativní zastoupení všech tříd.

Varování neznamená chybu implementace, ale upozorňuje na inherentní vlastnost datasetu a omezení klasifikační úlohy. Jak již bylo naznačeno v explorační analýze, cílová proměnná quality vykazuje nevyvážené zastoupení tříd, zejména u extrémních hodnot kvality, což komplikuje jejich stabilní učení i validaci.

Proto je nutné výsledky interpretovat nejen pomocí globálních metrik, ale také s využitím confusion matrix a třídově citlivějších pohledů na výkon modelu. Poznatek zároveň potvrzuje, že největší obtíž klasifikační úlohy spočívá v predikci málo zastoupených extrémních tříd kvality.

2.8 Regresní modely

Vedle klasifikačního přístupu byla v práci řešena také úloha regrese, ve které je cílová proměnná quality chápána jako numerická veličina. Cílem regresního modelování je odhadnout hodnotu kvality vína na základě jeho fyzikálně-chemických vlastností.

Regresní přístup byl do práce zařazen jako doplňkový pohled ke klasifikaci, který umožňuje detailněji analyzovat vztah mezi vstupními proměnnými a výslednou kvalitou.

Stejně jako v klasifikační části byl i zde kladen důraz na metodicky korektní workflow zahrnující oddělení benchmarku modelů, ladění hyperparametrů a finálního testování na nezávislé testovací množině.

2.8.1 Formulace regresní úlohy

V regresní variantě úlohy je proměnná quality interpretována jako numerická veličina, kterou se model snaží co nejpřesněji predikovat.

Přestože je proměnná v datasetu zaznamenána v podobě celočíselných hodnot, regresní přístup je metodicky opodstatněný, protože umožňuje modelovat kvalitu jako spojitou škálu a současně přesně kvantifikovat velikost predikční chyby.

Zařazení regrese vedle klasifikace umožňuje porovnat dva odlišné přístupy k řešení úlohy a posoudit, která interpretace cílové proměnné vede k lepším výsledkům.

2.8.2 Implementační a metodické principy regresní analýzy

Regresní analýza byla implementována v prostředí Python s využitím knihovny scikit-learn. Z metodického hlediska vychází část ze stejného workflow jako klasifikační analýza, zejména pokud jde o použití pipeline, oddělení trénovacích a testovacích dat a aplikaci křížové validace.

Stejně jako v klasifikační části byly jednotlivé modely realizovány pomocí pipeline, která propojuje preprocessing a vlastní regresní algoritmus. Přístup zajišťuje konzistentní zpracování dat a minimalizuje riziko data leakage, protože všechny transformační kroky jsou učeny pouze na trénovacích datech.

Hlavní rozdíl oproti klasifikační úloze spočívá ve formulaci cílové proměnné a ve způsobu vyhodnocení modelů. Zatímco klasifikace pracuje s diskrétními třídami, regresní modely predikují spojitou hodnotu, což se promítá zejména do volby metrik a interpretace výsledků.

Podobně jako v klasifikační části byly i zde využity Out-of-Fold (OOF) predikce pro získání realističtějšího odhadu výkonu modelů na validačních datech. OOF umožňuje analyzovat chování modelů bez nutnosti vytvářet samostatnou validační množinu mimo křížovou validaci.

2.8.3 Použité metriky regrese

Pro vyhodnocení regresních modelů byly použity metriky RMSE, MAE a R^2, které společně umožňují hodnotit výkon modelů z více perspektiv.

Jako hlavní optimalizační metrika byla zvolena RMSE, která klade větší důraz na větší chyby. V kontextu úlohy je důležité minimalizovat výrazné odchylky mezi predikovanou a skutečnou kvalitou, a proto byla RMSE použita jako hlavní kritérium při benchmarku i při ladění hyperparametrů.

MAE byla využita jako doplňková metrika poskytující intuitivní interpretaci průměrné chyby v jednotkách cílové proměnné. Slouží zejména k posouzení typické velikosti odchylky modelu a umožňuje lépe interpretovat výsledky z praktického hlediska.

R^2 bylo použito jako doplňující metrika zaměřená na celkovou vysvětlovací schopnost modelu. Umožňuje posoudit, do jaké míry model zachycuje variabilitu cílové proměnné, a doplňuje tak pohled založený na velikosti chyb.

V rámci evaluace modelů nebylo hodnocení založeno pouze na jedné hodnotě metriky, ale také na analýze její variability. Pro vyhodnocení byla použita metoda opakované křížové validace (Repeated K-Fold cross-validation), která umožňuje získat více nezávislých odhadů výkonu modelu při různých náhodných rozděleních dat.

Pro každý model byly metriky RMSE, MAE a R^2 počítány v jednotlivých validačních bězích. Na základě hodnot byly následně určeny agregované statistiky, konkrétně průměrná hodnota, medián, směrodatná odchylka a 95% interval spolehlivosti stanovený pomocí percentilů.

2.8.4 Přehled použitých regresních algoritmů

Pro regresní úlohu byla zvolena sada modelů odpovídající algoritmům použitým v klasifikační části, přičemž byly využity jejich regresní varianty. Přístup umožňuje přímé porovnání chování stejných modelovacích principů při odlišné interpretaci cílové proměnné.

Zatímco v klasifikační úloze modely přiřazují vzorky do diskrétních tříd kvality, jejich regresní varianty umožňují odhadovat spojitou hodnotu kvality a poskytují detailnější informaci o velikosti odchylky mezi predikcí a skutečností.

Lineární regrese představuje regresní analogii logistické regrese a slouží jako základní baseline model. Umožňuje posoudit, do jaké míry lze kvalitu vína aproximovat lineárním vztahem a jak velké jsou odchylky zjednodušeného přístupu.

Rozhodovací strom (regresní varianta) navazuje na klasifikační strom a umožňuje modelovat nelineární vztahy mezi proměnnými. V regresní úloze navíc poskytuje informaci o lokálních hodnotách predikce, což umožňuje detailněji analyzovat rozdíly mezi jednotlivými oblastmi datového prostoru.

Support Vector Regression (SVR) představuje regresní rozšíření modelu SVM. Umožňuje aproximovat složitější nelineární vztahy mezi vstupními atributy a cílovou proměnnou a současně minimalizovat velikost predikční chyby v kontinuálním prostoru.

K-Nearest Neighbours Regressor (KNN) vychází z klasifikační varianty založené na podobnosti vzorků. V regresním případě poskytuje predikci ve formě průměru hodnot sousedních vzorků, což umožňuje plynulejší modelování lokální struktury dat bez nutnosti explicitního učení globálního modelu.

2.8.5 Ensemble metody pro regresi

Stejně jako u klasifikace byly i v regresní úloze použity ensemble metody, které kombinují více modelů. V tomto případě však kombinují spojité predikce jednotlivých regresorů, což umožňuje nejen zvýšit stabilitu modelu, ale také zpřesnit samotnou numerickou predikci.

Random Forest Regressor představuje regresní variantu stromového ensemble modelu. V porovnání s klasifikační verzí umožňuje nejen rozhodovat o třídě, ale i poskytovat stabilní odhad spojité hodnoty kvality.

Voting Regressor kombinuje predikce více modelů pomocí průměrování. Oproti klasifikačnímu soft voting zde dochází k přímé agregaci numerických odhadů, což umožňuje snížit variabilitu jednotlivých modelů a zpřesnit výslednou predikci. Stejně jako v případě klasifikačních ensemble metod se varianta skládá ze stejných typů modelů, avšak v jejich regresní podobě, konkrétně lineární regrese, rozhodovacího stromu, SVR a KNN.

Stacking Regressor navazuje na klasifikační variantu stackingu, avšak pracuje se spojitými výstupy základních modelů. Meta-model se tak učí optimálně kombinovat numerické predikce, což může vést k přesnějšímu zachycení vztahu mezi vstupními proměnnými a kvalitou vína. Obdobně jako u klasifikačního stackingu jsou i zde použity stejné typy základních modelů v regresní variantě, přičemž roli meta-modelu plní regulovaná lineární regrese (Ridge), která se učí kombinovat jejich výstupy.

2.8.6 Strategie benchmarku, ladění a finálního testování

Strategie vyhodnocení regresních modelů vychází ze stejného principu jako v klasifikační části (viz kapitola 2.7.6) a je založena na třech navazujících fázích: benchmark modelů, ladění hyperparametrů a finální testování.

V první fázi byly jednotlivé regresní modely porovnány pomocí křížové validace na trénovací množině s cílem identifikovat nejslibnější kandidáty. Ve druhé fázi bylo provedeno ladění hyperparametrů pomocí GridSearchCV, přičemž jako optimalizační kritérium byla použita metrika RMSE.

Ve třetí fázi byly vybrané modely vyhodnoceny na nezávislé testovací množině, která nebyla použita ani při benchmarku, ani při ladění. Postup zajišťuje metodicky korektní oddělení jednotlivých fází a poskytuje realistický odhad generalizační schopnosti modelů.

2.8.7 Diagnostické grafy regresních modelů

Kromě numerických metrik byly v regresní části využity také diagnostické grafy, které slouží jako doplňkový nástroj pro analýzu chování modelů a charakteru jejich chyb.

Pro jednotlivé modely byly vytvářeny tři základní typy grafů: graf predikovaných a skutečných hodnot, graf reziduí vůči predikcím a histogram rozdělení reziduí. Grafy umožňují posoudit nejen celkovou přesnost modelu, ale také odhalit případné systematické odchylky, nelinearity nebo nerovnoměrnou variabilitu chyb.

Z metodického hlediska představují diagnostické grafy důležitý doplněk číselných metrik, protože umožňují identifikovat vlastnosti modelu, které se v agregovaných ukazatelích nemusí plně projevit.

2.9 Návrh experimentů

Navržené experimenty slouží k doplnění základního modelování o hlubší analýzu chování modelů v různých podmínkách. Jejich cílem je ověřit stabilitu a robustnost dosažených výsledků a identifikovat limity použitých přístupů. Experimenty se zaměřují zejména na vliv struktury dat, volby příznaků a rozdílných modelovacích strategií na výslednou predikční výkonnost.

Experimenty přímo navazují na poznatky získané v rámci explorační analýzy dat, zejména na zjištěnou nevyváženost tříd, překryv tříd kvality a význam jednotlivých vstupních atributů.

Detailní implementace všech uvedených experimentů, včetně definice pipeline, výběru atributů a nastavení modelů, je dostupná v přiloženém notebooku (viz Příloha A.1).

2.9.1 Analýza výkonů modelů podle četnosti vzorků

Experiment byl navržen s cílem analyzovat, jak se prediktivní modely chovají v různých částech rozdělení cílové proměnné quality. Standardní globální metriky poskytují pouze celkový přehled o výkonnosti modelu, avšak neumožňují posoudit, zda model dosahuje srovnatelné kvality predikce napříč celým spektrem hodnot, nebo zda je jeho výkon soustředěn především do nejčetnějších oblastí dat.

Motivací experimentu je skutečnost, že analyzovaný dataset vykazuje výraznou nerovnováhu tříd, kdy většina vzorků spadá do středních hodnot kvality, zatímco extrémní hodnoty jsou zastoupeny pouze omezeně. Současná struktura může vést k situaci, kdy model dosahuje dobrých globálních výsledků, ale jeho výkon je výrazně odlišný v méně četných částech datového rozdělení.

Experiment byl proveden samostatně pro klasifikační i regresní úlohu a v obou případech byly analyzovány finálně vybrané a naladěné modely. Vyhodnocení bylo realizováno výhradně na odložené testovací množině, aby byla zachována metodická korektnost a nedošlo k ovlivnění výsledků opakovaným využíváním trénovacích dat.

Pro účely analýzy byla testovací množina rozdělena do tří skupin podle skutečné hodnoty cílové proměnné quality:

- četné třídy (5–7),
- méně četné spodní třídy (3–4),
- méně četné horní třídy (8–9).

Rozdělení reflektuje přirozenou strukturu datasetu, kde střední hodnoty reprezentují dominantní část dat, zatímco okrajové hodnoty odpovídají méně zastoupeným extrémům kvality.

V klasifikační části experimentu byly modely vyhodnoceny v rámci jednotlivých skupin pomocí metrik Accuracy a F1-score (weighted), které umožňují posoudit nejen celkovou úspěšnost klasifikace, ale i schopnost modelu pracovat s nevyváženým rozdělením tříd.

V regresní části byly modely hodnoceny pomocí metrik RMSE, MAE a R^2, které poskytují doplňující pohled na velikost predikční chyby i schopnost modelu vysvětlit variabilitu dat v jednotlivých částech rozdělení.

2.9.2 Vyvážený tří-třídní klasifikační model pro kvality 5, 6 a 7

Experiment byl navržen s cílem ověřit, zda umělé vyvážení tříd v nejčetnější části datasetu může vést ke stabilnějšímu a férovějšímu klasifikačnímu modelu. Motivací experimentu byla skutečnost, že cílová proměnná quality vykazuje výrazně nevyvážené rozdělení, přičemž největší počet vzorků připadá na střední hodnoty kvality, zejména na třídy 5, 6 a 7. Současná struktura dat může vést k preferenci dominantních tříd a ke slabšímu výkonu modelů na méně zastoupených kategoriích.

Experiment se proto zaměřuje výhradně na tři nejčastější třídy kvality, tedy 5, 6 a 7. Třídy současně reprezentují přirozeně interpretovatelnou střední oblast kvality vín a umožňují sledovat vliv vyvážení tříd bez zahrnutí extrémních hodnot kvality. Původní multiclass úloha je tak v tomto experimentu zjednodušena na tří-třídní klasifikační problém.

V první fázi byla z původního datasetu vybrána pouze pozorování s hodnotami quality $\in \{5, 6, 7\}$. Následně byla pro vytvoření vyváženého datasetu použita metoda downsamplingu, při níž byly početnější třídy náhodně redukovány na velikost nejmenší z uvažovaných tříd. Výsledkem vznikla datová sada s rovnoměrným zastoupením všech tří kategorií.

Vyvážený dataset byl dále rozdělen stratifikovaným způsobem na trénovací a testovací množinu. Stratifikace byla použita s cílem zachovat stejné zastoupení tříd v obou částech datasetu a zajistit férové vyhodnocení modelů na vyvážené testovací množině.

Na rozdíl od původní verze experimentu nebyla použita samostatná validační množina. Ladění hyperparametrů nového modelu probíhalo výhradně na trénovací množině pomocí stratifikované křížové validace. Testovací množina byla následně využita pouze pro finální vyhodnocení modelů. Postup odpovídá metodice použité v hlavní části práce a zajišťuje konzistentní experimentální nastavení.

V experimentu byly porovnány dva přístupy. Prvním byl původní nejlepší naladěný klasifikační model z hlavní části práce, který byl natrénován na kompletním původním datasetu a následně vyhodnocen na nově vytvořené testovací množině obsahující pouze třídy 5, 6 a 7. Druhým byl nový model stejného typu, který byl znovu natrénován a naladěn pouze na vyváženém datasetu obsahujícím tři třídy.

Je však nutné zdůraznit, že se nejedná o zcela symetrické porovnání dvou modelů na identickém problému. Původní model byl učen na širším spektru tříd, zatímco nový model řeší zúžený tří-třídní problém. Výsledky proto slouží především k analýze dopadu vyvážení tříd a redukce dat, nikoli jako přímé srovnání dvou ekvivalentních modelů.

Srovnání umožňuje současně posoudit dopad redukce trénovacích dat. Experiment tak sleduje, zda případný přínos vyvážení tříd převáží nad ztrátou informace způsobenou downsamplingem.

Pro vyhodnocení obou modelů byly použity metriky Accuracy, F1-score weighted a F1-score macro. Zatímco Accuracy a weighted F1-score umožňují posoudit celkovou klasifikační úspěšnost, F1-score macro slouží jako doplňující ukazatel rovnoměrnosti výkonu napříč jednotlivými třídami.

2.9.3 Analýza vlivu redukce vstupních atributů na výkon modelu

Cílem experimentu bylo ověřit, jaký dopad má redukce počtu vstupních atributů na výkon prediktivních modelů kvality vína. V předchozích částech práce byly na základě explorační analýzy dat a lineární diskriminační analýzy identifikovány proměnné, které se jeví jako významné z hlediska vztahu ke kvalitě vína. Nabízí se proto otázka, zda je možné využít pouze omezenou množinu atributů a zároveň zachovat přijatelnou úroveň predikční přesnosti.

Experiment porovnává dvě varianty vstupních dat:

- FULL varianta, která využívá kompletní sadu vstupních atributů použitých v hlavním modelování,
- EDA varianta, která využívá redukovanou množinu čtyř atributů vybraných na základě předchozí analýzy (density, residual sugar, fixed acidity, alcohol).

Atributy reprezentují důležité chemické a technologické charakteristiky vína a byly vybrány na základě jejich významnosti v předchozích analytických krocích.

Aby bylo možné izolovat vliv samotné redukce atributů, byl ve všech případech zachován stejný modelový přístup i stejná testovací množina. Jedinou měněnou veličinou byl počet vstupních atributů.

Pro obě úlohy byly vytvořeny dvě varianty pipeline:

- FULL pipeline – s kompletní množinou vstupních atributů,
- EDA pipeline – pouze s redukovanou množinou čtyř atributů.

Vyhodnocení obou variant probíhalo na stejné odložené testovací množině, aby bylo zajištěno férové porovnání.

Experiment byl proveden pro regresní i klasifikační variantu úlohy. V obou případech byla predikční výkonnost hodnocena pomocí odpovídajících metrik – u regrese pomocí RMSE, MAE a R^2, u klasifikace pomocí Accuracy a F1-score ve variantách weighted a macro.

Současně byla sledována také výpočetní náročnost jednotlivých variant, konkrétně čas tréninku modelu a čas predikce.

3 Výsledky analýz

Zde jsou prezentovány výsledky klasifikačních a regresních modelů pro predikci kvality vína. Navazuje se zde na metodickou část, ve které byly popsány použité algoritmy, validační strategie a metriky hodnocení. Cílem kapitoly je shrnout dosažené výsledky jednotlivých modelů a porovnat jejich výkonnost na trénovacích i testovacích datech, včetně výsledků navržených experimentů.

3.1 Výsledky klasifikačních modelů

Výkon klasifikačních modelů byl vyhodnocen pomocí Out-of-Fold (OOF) predikcí, které poskytují realistický odhad validační schopnosti modelů na trénovacích datech. Pro hodnocení byly použity metriky accuracy, precision, recall, F1 skóre a průměrná hodnota AUC. Souhrn výsledků jednotlivých modelů je uveden v tabulce.

Tab. 1: Výsledky klasifikačních modelů pomocí OOF

Model	*Accuracy*	*Precision*	*Recall*	*F1*	*Mean AUC*
Random Forest Classifier	0.667116	0.67779	0.667116	0.65291	0.839189
Decision Tree Classifier	0.569367	0.567554	0.569367	0.568374	0.616876
SVM (RBF Kernel)	0.565711	0.56452	0.565711	0.528337	0.795392
KNN Classifier	0.547624	0.525042	0.547624	0.525678	0.688892
Logistic Regression (Multinomial)	0.544353	0.523052	0.544353	0.512253	0.767709

Zdroj: vlastní zpracování

Detailní diagnostické grafy pro jednotlivé modely, včetně confusion matrix, ROC křivek, Precision–Recall křivek a learning curves, jsou dostupné v přiloženém notebooku (viz Příloha A.1).

Výsledky ensemble přístupů, které kombinují více klasifikačních modelů, jsou uvedeny v následující tabulce.

Tab. 2: Výsledky klasifikačních ensamble pomocí OOF

Model	*Accuracy*	*Precision*	*Recall*	*F1*	*Mean AUC*
Ensemble – Soft Voting	0.607658	0.596233	0.607658	0.590824	0.818206
Ensemble – Stacking	0.605542	0.598363	0.605542	0.589421	0.810037

Zdroj: vlastní zpracování

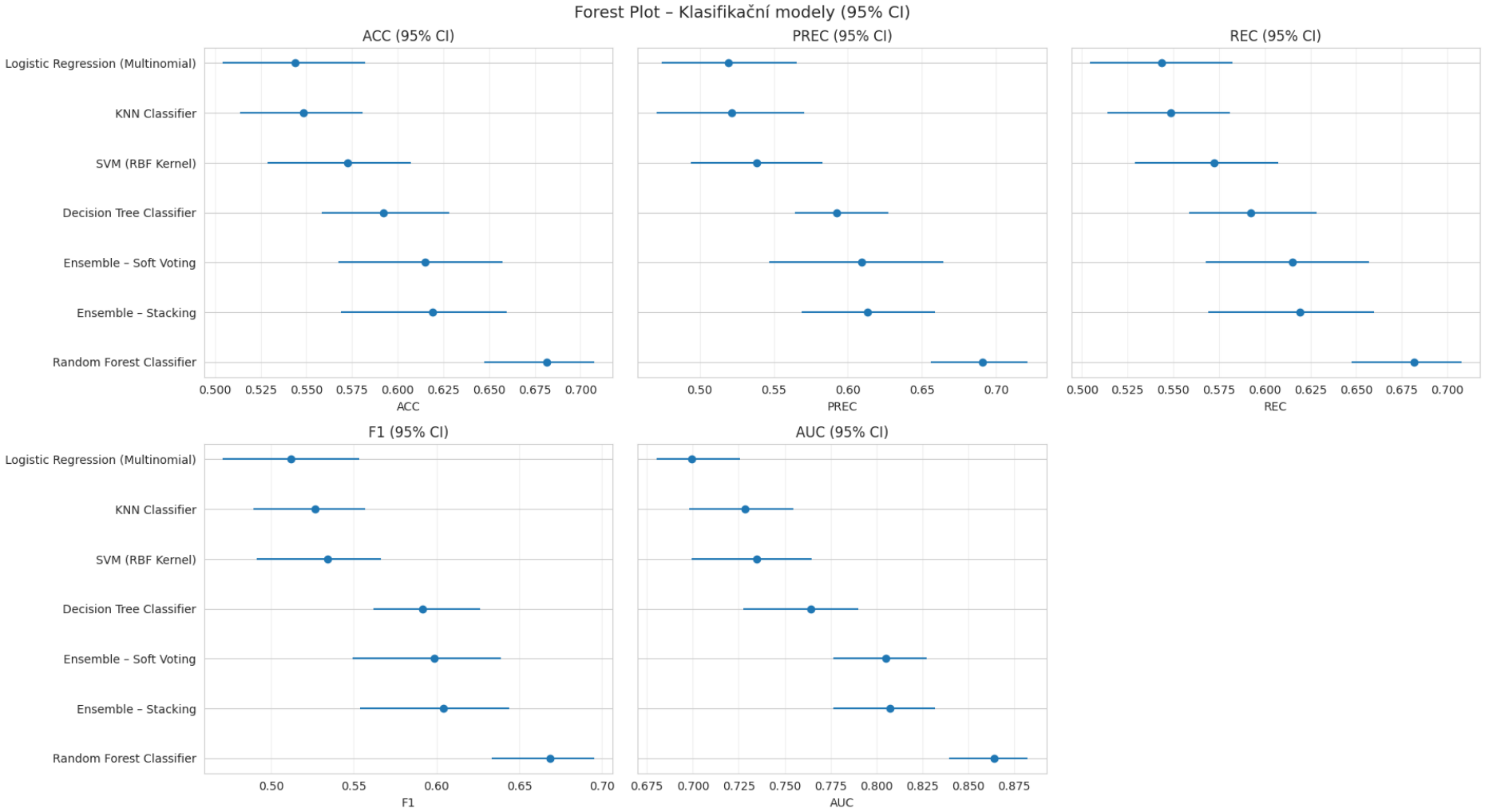
Na základě předchozích experimentů byly vybrány finální modely, které byly následně optimalizovány a vyhodnoceny na testovací množině. Souhrn dosažených metrik je uveden v tabulce.

Tab. 3: Výsledky klasifikačních modelů po ladění

Model	*Accuracy*	*Precision*	*Recall*	*F1*	*Mean AUC*
Random Forest Classifier (tuned)	0.692308	0.694318	0.692308	0.678486	0.870719
Ensemble – Soft Voting	0.639231	0.625927	0.639231	0.619872	0.847211

Zdroj: vlastní zpracování

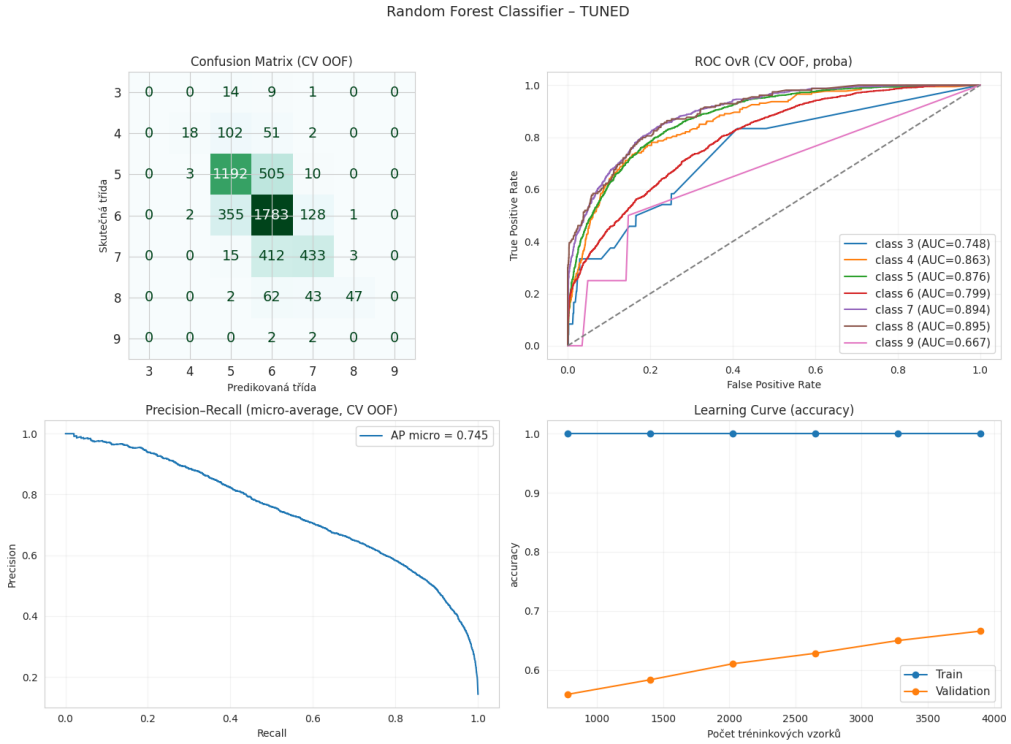
Pro doplnění výsledků bylo provedeno také statistické hodnocení pomocí Repeated Stratified K-Fold. Výsledky jsou prezentovány formou forest plotů, které zobrazují průměrné hodnoty jednotlivých metrik a jejich 95% intervaly spolehlivosti napříč validačními běhy.



Obr. 19: Forest Plot – Klasifikační modely

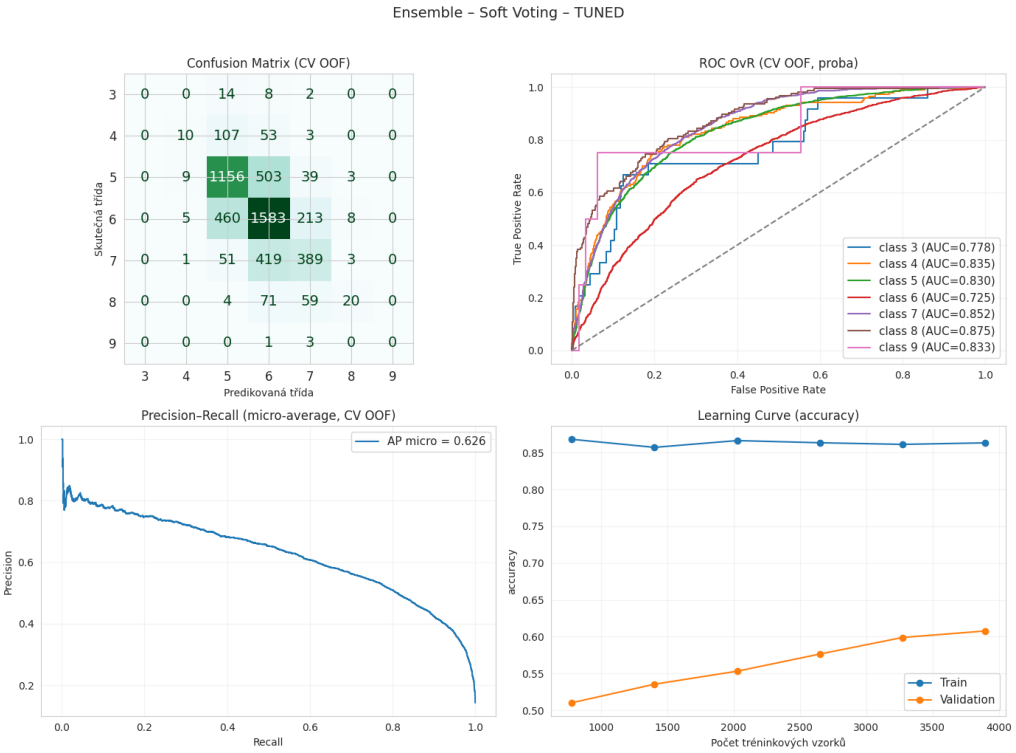
Zdroj: Vlastní zpracován

Pro detailnější posouzení výkonu finálních modelů jsou na obrázcích zobrazeny evaluační grafy nejlepších modelů, přičemž kompletní sada grafů je dostupná v přiloženém notebooku (viz Příloha A.1).



Obr. 20: Evaluační grafy Random Forest Classifier

Zdroj: Vlastní zpracování



Obr. 21: Evaluační grafy pro Ensamble Soft Voting

Zdroj: vlastní zpracování

3.2 Výsledky regresních modelů

Regresní modely byly hodnoceny pomocí metrik RMSE, MAE a koeficientu determinace R^2. Kromě numerických metrik byly pro vybrané modely použity také diagnostické grafy, konkrétně vztah mezi predikovanými a skutečnými hodnotami, graf reziduí a distribuce reziduí. Detailní grafické výstupy pro všechny modely jsou dostupné v přiloženém notebooku (viz Příloha A.1).

Benchmark regresních modelů byl proveden pomocí cross-validation na trénovací množině. Souhrn dosažených výsledků je uveden v tabulce.

Tab. 4: Výsledky regresních modelů pomocí cross-validation

Model	*RMSE*	*MAE*	*R2*
Random Forest Regressor	0.631579	0.456098	0.484581
KNN Regressor	0.634643	0.429131	0.479532
SVR (RBF Kernel)	0.699921	0.522848	0.367039
Linear Regression	0.746282	0.575851	0.280287
Decision Tree Regressor	0.890025	0.539548	-0.02426

Zdroj: vlastní zpracování

Výsledky ensemble přístupů jsou uvedeny v následující tabulce.

Tab. 5: Výsledky regresních ensamble pomocí cross-validation

Model	*RMSE*	*MAE*	*R2*
Ensemble – Stacking Regressor	0.625073	0.442296	0.495098
Ensemble – Voting Regressor	0.651451	0.480651	0.45158

Zdroj: vlastní zpracování

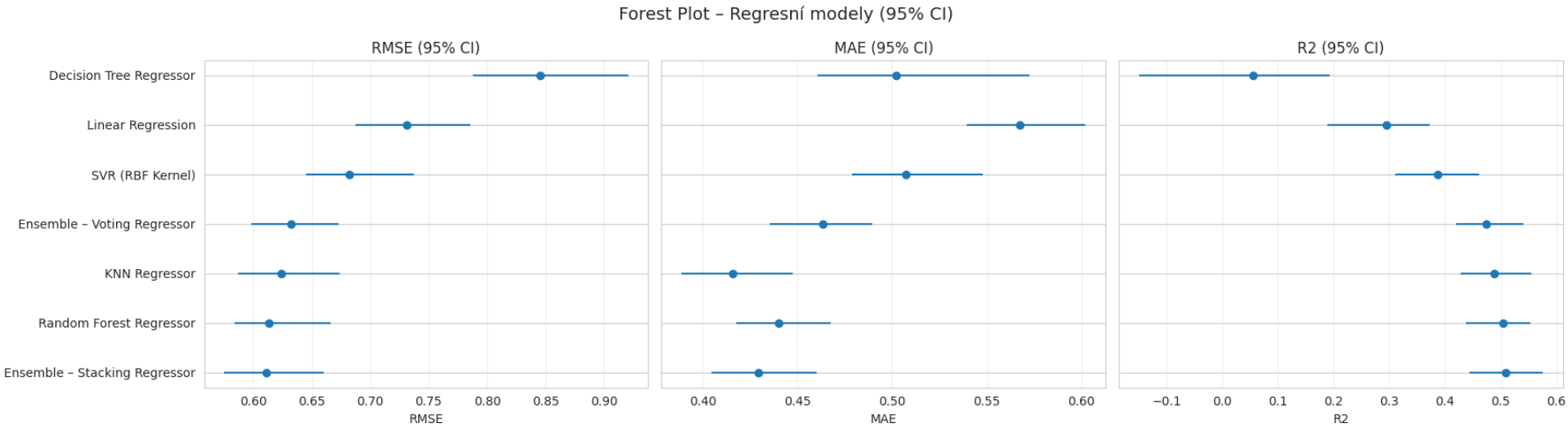
Na základě validačních výsledků byly vybrány dva modely pro finální testování na nezávislé testovací množině. Souhrn dosažených metrik je uveden v tabulce.

Tab. 6: Výsledky regresních modelů po ladění

Model	*RMSE*	*MAE*	*R2*
Random Forest Regressor (tuned)	0.561154	0.402654	0.559465
Ensemble – Stacking Regressor (tuned)	0.562403	0.392214	0.557502

Zdroj: vlastní zpracování

Výsledky statistické evaluace jsou prezentovány pomocí forest plotů, které zobrazují průměrné hodnoty metrik RMSE, MAE a R^2 spolu s jejich 95% intervaly spolehlivosti napříč validačními běhy.



Obr. 22: Grafy Forest Plot – Regresní modely

Zdroj: Vlastní zpracování

3.3 Výsledky analýzy výkonů modelů podle četnosti vzorků

Experiment analyzuje výkon finálních modelů v různých částech rozdělení cílové proměnné quality. Vyhodnocení bylo provedeno na testovací množině, která byla rozdělena do tří skupin podle skutečné hodnoty kvality.

Jednotlivé skupiny byly definovány následovně:

- četné třídy: 5–7,
- méně četné spodní třídy: 3–4,
- méně četné horní třídy: 8–9.

Zastoupení jednotlivých skupin v testovací množině je uvedeno v tabulce.

Tab. 7: Četnosti skupin

Skupina	*Četnost*
Četné (5–7)	1211
Méně četné – spodní (3–4)	49
Méně četné – horní (8–9)	40

Zdroj: vlastní zpracování

Souhrn dosažených výsledků klasifikačních modelů je uveden v následujících tabulkách.

Tab. 8: Výsledky F1 score dle skupin četnosti

F1 score Skupina	*Horní (8–9)*	*Spodní (3–4)*	*Četné (5–7)*
Single (tuned)	0.458824	0.182823	0.728155
Ensemble (tuned)	0.254348	0.076309	0.641455

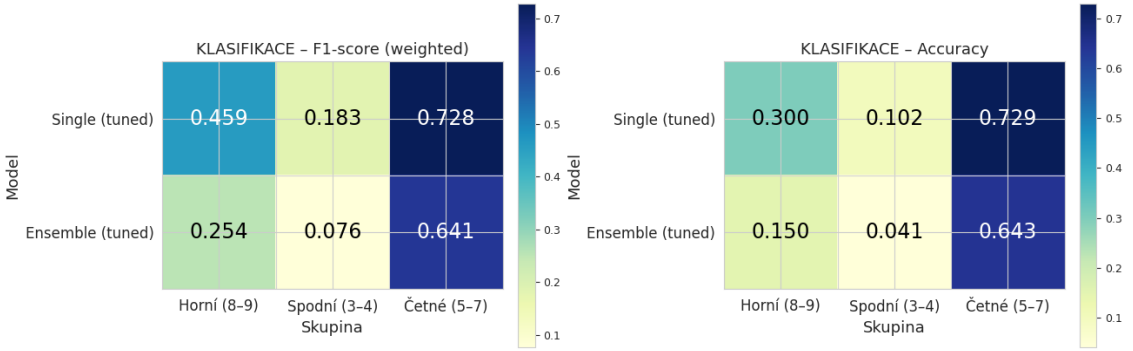
Zdroj: vlastní zpracování

Tab. 9: Výsledky Accurace dle skupin četnosti

Accuracy Skupina	*Horní (8–9)*	*Spodní (3–4)*	*Četné (5–7)*
Single (tuned)	0.3	0.102041	0.729149
Ensemble (tuned)	0.15	0.040816	0.643270

Zdroj: vlastní zpracování

Pro lepší přehlednost jsou výsledky znázorněny také graficky.



Obr. 23: F1 score a Accuracy podle skupin

Zdroj: vlastní zpracování

Souhrn výsledků regresních modelů v jednotlivých skupinách je uveden v následujících tabulkách.

Tab. 10: Výsledky RMSE dle skupin četnosti

RMSE Skupina	*Horní (8–9)*	*Spodní (3–4)*	*Četné (5–7)*
Ensemble (tuned)	1.254939	1.56595	0.49258
Single (tuned)	1.291482	1.600588	0.483419

Zdroj: vlastní zpracování

Tab. 11: Výsledky MAE dle skupin četnosti

MAE Skupina	*Horní (8–9)*	*Spodní (3–4)*	*Četné (5–7)*
Ensemble (tuned)	1.096813	1.405184	0.35965
Single (tuned)	1.13875	1.424796	0.36287

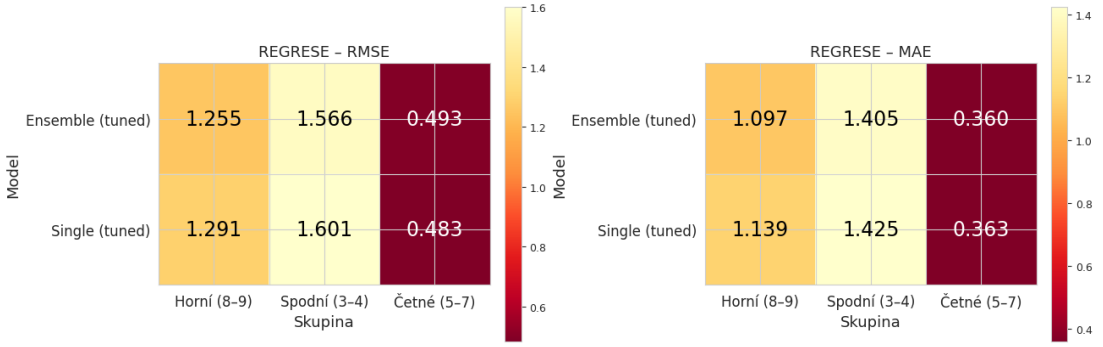
Zdroj: vlastní zpracování

Tab. 12: Výsledky R^2 dle skupin četnosti

R^2 Skupina	*Horní (8–9)*	*Spodní (3–4)*	*Četné (5–7)*
Ensemble (tuned)	-63.6102	-21.8207	0.515839
Single (tuned)	-67.4277	-22.8414	0.533681

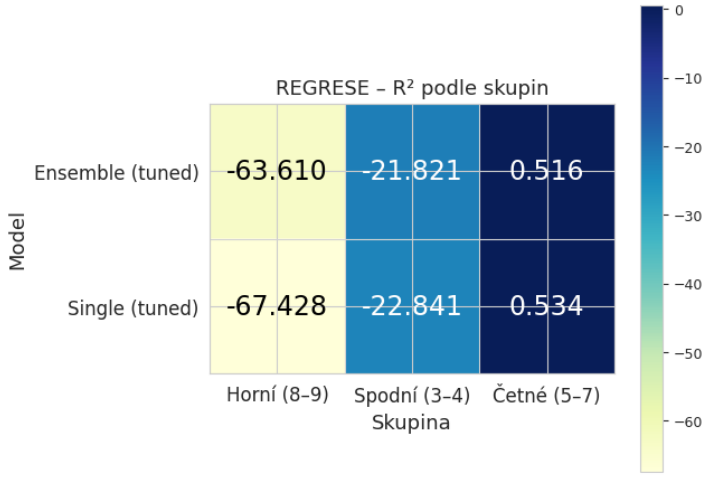
Zdroj: vlastní zpracování

Grafické znázornění výsledků je uvedeno na následujících obrázcích.



Obr. 24: RMSE a MAE podle skupin

Zdroj: vlastní zpracování



Obr. 25: R^2 podle skupin

Zdroj: vlastní zpracování

3.4 Výsledky experimentu s vyvážením tříd

V tomto experimentu byl analyzován vliv vyvážení tříd na výkon klasifikačního modelu v podmnožině dat obsahující pouze třídy kvality 5, 6 a 7.

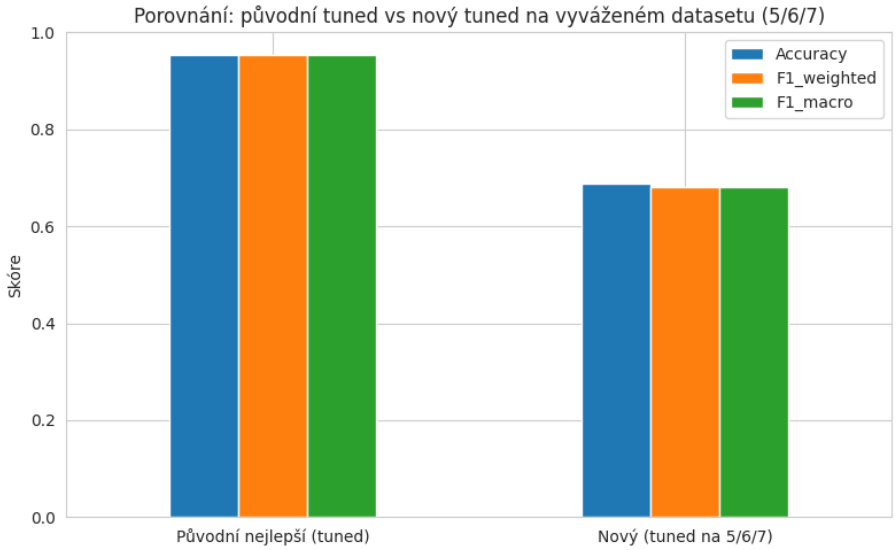
Souhrn dosažených výsledků obou porovnávaných modelů na testovací množině je uveden v tabulce.

Tab. 13: Výsledky vyváženosti tříd

Model	*Accuracy*	*F1 weighted*	*F1 macro*
Původní nejlepší (tuned)	0.953704	0.953763	0.953763
Nový (tuned na 5/6/7)	0.688272	0.681990	0.681990

Zdroj: vlastní zpracování

Pro přehlednější porovnání jsou výsledky znázorněny také graficky.



Obr. 26: Porovnání s vyváženým datasetem

Zdroj: vlastní zpracování

3.5 Výsledky vlivu redukce vstupních atributů na výkon modelu

V tomto experimentu byl analyzován vliv redukce vstupních atributů na výkon prediktivních modelů. Souhrn výsledků regresních a klasifikačních modelů je uveden v následujících tabulkách a grafech.

Tab. 14: Výsledky redukce vstupních atributů pro klasifikaci

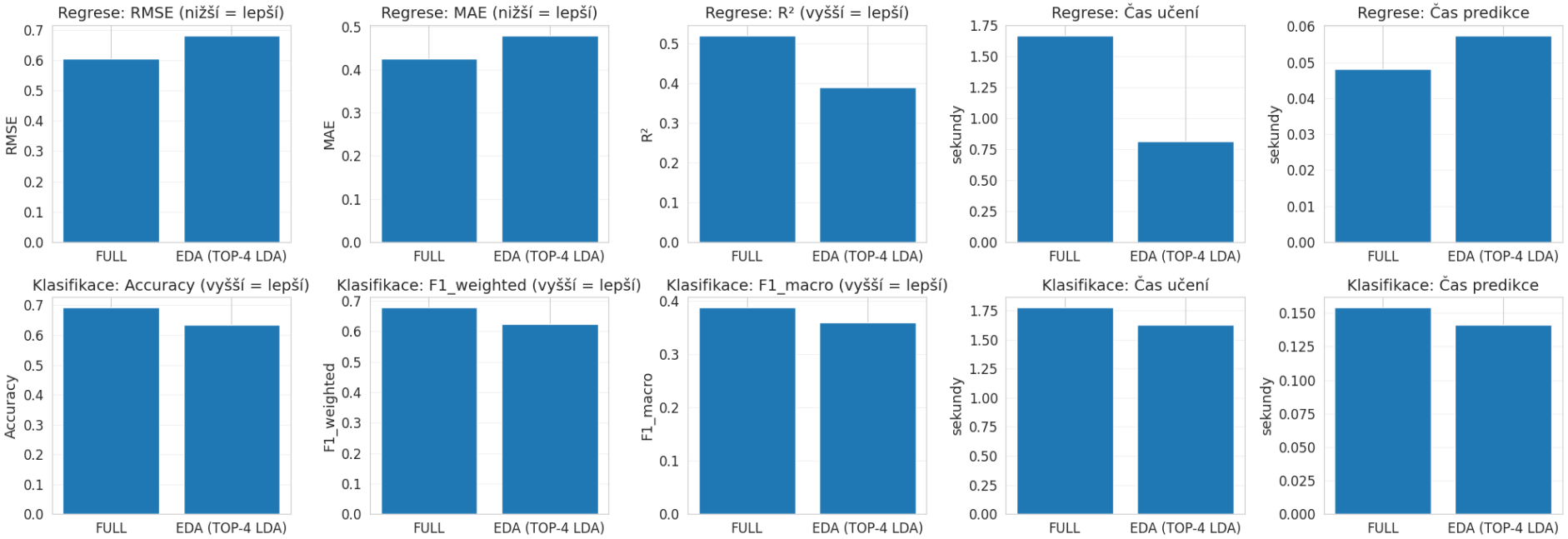
Varianta	*Počet atributů*	*RMSE*	*MAE*	*R2*	*Čas tréninku (s)*	*Čas predikce (s)*
FULL	12	0.604631	0.426769	0.521166	2.551128	0.085709
TOP-4 LDA	4	0.682291	0.479995	0.390261	1.252458	0.094079

Zdroj: vlastní zpracování

Tab. 15: Výsledky redukce vstupních atributů pro regresi

Varianta	*Počet atributů*	*Accuracy*	*F1_weighted*	*F1_macro*	*Čas tréninku (s)*	*Čas predikce (s)*
FULL	12	0.692308	0.678486	0.387877	2.68142	0.279134
TOP-4 LDA	4	0.634615	0.62402	0.360493	2.426747	0.269831

Zdroj: vlastní zpracování



Obr. 27: Grafy porovnání vlivu redukce

Zdroj: vlastní zpracová

4 Diskuse

Zde jsou interpretovány hlavní výsledky dosažené v analytické části práce a jejich význam ve vztahu ke stanovenému cíli a výzkumným otázkám. Pozornost je věnována zejména porovnání klasifikačních a regresních přístupů, vlivu struktury datasetu na výkon modelů a dopadu vybraných experimentálních úprav, jako bylo vyvažování tříd a redukce vstupních atributů. Součástí diskuse je také zasazení dosažených výsledků do kontextu obdobných přístupů publikovaných v komunitních projektech.

4.1 Interpretace hlavních výsledků modelů

Na základě dosažených výsledků lze konstatovat, že kvalitu vína je možné na základě laboratorně měřených fyzikálně-chemických parametrů predikovat, avšak nikoli zcela bez omezení. Nejlepší klasifikační model po ladění, Random Forest Classifier, dosáhl hodnoty Accuracy 0,692, weighted F1-score 0,678 a Mean AUC 0,871, zatímco nejlepší regresní modely dosáhly hodnot RMSE přibližně 0,56 a R^2 přibližně 0,56. Výsledky naznačují, že mezi vstupními atributy a výsledným hodnocením kvality skutečně existuje predikovatelný vztah, současně však potvrzují, že vztah není natolik jednoznačný, aby bylo možné očekávat téměř bezchybnou predikci. Predikce kvality vína je tedy spíše úlohou středně obtížné aproximace než jednoznačného rozhodování.

V klasifikační úloze se jako nejúspěšnější ukázal Random Forest, který překonal logistickou regresi, k-NN i SVM a současně nepodlehl ani porovnání s použitými ensemble přístupy. Výsledek odpovídá charakteru analyzovaných dat. Již v explorativní fázi bylo naznačeno, že vztahy mezi atributy a cílovou proměnnou nejsou čistě lineární a že mezi jednotlivými třídami existuje výrazný překryv. Právě zmíněná kombinace je pro lineární modely problematická. Logistická regrese předpokládá lineárně oddělitelné rozhodovací hranice v prostoru příznaků, a proto v situaci, kdy se sousední třídy kvality částečně prolínají, její výkonnost přirozeně klesá. Na druhou stranu Random Forest dokáže díky soustavě stromových rozdělení zachytit i složitější a nelineární interakce mezi proměnnými, aniž by vyžadoval tak silné strukturální předpoklady o tvaru vztahu mezi vstupem a výstupem. Daná vlastnost je pravděpodobně hlavní důvod, proč právě Random Forest model dosáhl v dané úloze nejlepších výsledků.

Důležité je, že Random Forest je pro využitá data metodicky velmi vhodný. Pracuje dobře i v situaci, kdy jednotlivé proměnné nesou pouze dílčí informaci a až jejich kombinace vytváří užitečný predikční signál. V případě kvality vína je současný scénář realistický: výsledné hodnocení není určováno jedním dominantním parametrem, ale spíše souhrou více fyzikálně-chemických vlastností. Zároveň částečně odpovídá i výzkumné otázce, které modely dosahují nejlepších výsledků a za jakých podmínek. V tomto případě se ukazuje, že za podmínek vícetřídní úlohy, bez zjednodušení cílové proměnné a bez umělého vyvažování tříd, jsou nejvhodnější robustní nelineární stromové modely, především Random Forest.

Naopak slabší výkon logistické regrese, k-NN a v menší míře i SVM lze interpretovat jako důsledek omezení jejich modelovací logiky vůči charakteru dat. U logistické regrese je limitem zejména lineární forma rozhodování. U k-NN může být problémem skutečnost, že při existenci překryvu tříd a vyšší hustoty středních kategorií se lokální sousedství snadno „přelije“ do dominantních tříd, což zhoršuje rozlišování mezi sousedními hodnotami kvality. SVM

s RBF jádrem sice teoreticky umožňuje nelineární separaci, avšak v úloze zřejmě nedokázalo nalézt takovou rozhodovací strukturu, která by překonala stromové modely. Výsledek může souviset jak s překryvem tříd, tak i s tím, že cílová proměnná quality má ordinální charakter a hranice mezi sousedními třídami nejsou ostré, ale spíše plynulé. Ve zmíněné situaci může být i dobře naladěné SVM méně stabilní než baggingový stromový ensemble.

Za pozornost stojí skutečnost, že použité ensemble přístupy nepřekonaly samotný Random Forest. Na první pohled se může zdát paradoxní, že kombinace více modelů vede k horším výsledkům než jeden model. Při bližší interpretaci však získaný výsledek dává smysl. Použitý soft voting i stacking byly sestaveny z modelů Logistic Regression, Decision Tree, SVM a KNN, přičemž mezi základními modely vůbec nefiguroval Random Forest, tedy empiricky nejúspěšnější klasifikátor v celé úloze. Ensemble tak ve skutečnosti nekombinoval nejlepší dostupné prediktory, ale spíše agregoval několik modelů, které byly samy o sobě slabší. Kombinace modelů nemusí získat silnější model, ale pouze stabilizovaný průměr slabších rozhodnutí.

Dalším důvodem může být i omezená diverzita chyb mezi použitými základními modely. Ensemble metody bývají nejpřínosnější tehdy, pokud jednotlivé modely chybují odlišným způsobem. V úloze však lze předpokládat, že většina modelů narážela na stejný problém, totiž na překryv sousedních tříd kvality a dominanci středních kategorií. Pokud základní modely opakovaně selhávají ve stejných oblastech datového prostoru, metamodel stackingu ani průměrování pravděpodobností v soft voting nemají z čeho vytěžit výrazné zlepšení. Výsledek je proto důležitý i z metodického hlediska: samotný fakt, že jde o ensemble metodu, automaticky nezaručuje vyšší výkon. Přínos ensemble závisí na kvalitě i vzájemné komplementaritě použitých základních modelů. V tomto smyslu výsledky práce nepřímo odpovídají i na výzkumnou otázku, za jakých podmínek dosahují modely nejlepších výsledků: nerozhoduje pouze složitost modelu, ale zejména jeho shoda s charakterem dat.

Závěr je dále podpořen analýzou learning curves. Z nich je patrné, že oba modely vykazují známky přeučení, neboť v obou případech existuje rozdíl mezi výkonem na trénovacích a validačních datech. Rozdíl je však výrazně větší u modelu Random Forest, který dosahuje téměř perfektní přesnosti na trénovacích datech, zatímco validační přesnost zůstává podstatně nižší. Model tak vykazuje typické znaky overfittingu, kdy má dostatečnou kapacitu k zapamatování trénovacích dat, avšak jeho generalizační schopnost je omezená. Současně však validační křivka s rostoucím počtem vzorků mírně roste, což naznačuje, že model částečně těží z většího množství dat, ale zároveň naráží na přirozený limit predikovatelnosti dané úlohy.

Použitý ensemble model založený na soft voting nedosahuje vysoké přesnosti ani na trénovacích datech, což ukazuje na jeho omezenou modelovací kapacitu. Přestože i zde existuje rozdíl mezi trénovacím a validačním výkonem, který naznačuje mírné přeučení, dominantním problémem je spíše underfitting. Model tak není schopen dostatečně zachytit komplexní nelineární vztahy v datech, a proto nedosahuje výkonu srovnatelného s Random Forestem.

U regresních modelů jsou rozdíly mezi nejlepšími přístupy podstatně menší než u klasifikace. Random Forest Regressor po ladění dosáhl RMSE 0,561 a R^2 0,559, zatímco tuned Stacking Regressor měl RMSE 0,562 a R^2 0,558. Prakticky tedy nelze hovořit o výrazně lepším modelu; oba přístupy podávají téměř shodný výkon. Výsledek lze interpretovat tak, že regresní formulace úlohy naráží na přirozený strop predikovatelnosti. Cílová proměnná quality je sice zapsána číselně, ale ve své podstatě představuje diskrétní a částečně subjektivní senzorické hodnocení.

Regresní modely se tedy snaží aproximovat veličinu, která není plně spojitá a která navíc může obsahovat šum vyplývající z lidského hodnocení. Složitější regresní model nemusí mít výrazný prostor k překonání silného stromového baseline modelu.

Výrazně slabší výkon lineární regrese oproti stromovým a kernelovým metodám potvrzuje, že ani v regresní formulaci nelze vztah mezi fyzikálně-chemickými vlastnostmi a kvalitou vína chápat jako jednoduchou lineární závislost. Pokud by byl vztah převážně lineární, bylo by možné očekávat menší rozdíl mezi lineární regresí a pokročilejšími modely. Skutečnost, že lineární regrese dosahuje podstatně horších hodnot RMSE i R^2, naznačuje, že rozhodující informace je uložena spíše v nelineárních vazbách a interakcích mezi proměnnými. Zjištění podporuje závěr, že pro daný problém jsou vhodnější nelineární modely než jednoduché lineární přístupy.

Při interpretaci dosažených výsledků je rovněž důležité srovnání s obdobnými projekty z Kaggle. V připravených výpiscích se opakovaně ukazuje, že mezi nejúspěšnější modely patří Random Forest, Gradient Boosting, XGBoost či LightGBM, zatímco logistická regrese a další jednodušší přístupy dosahují slabších výsledků. V tomto ohledu jsou závěry práce s příbuznými experimenty konzistentní: nelineární stromové ensemble metody mají v úloze predikce kvality vína zjevnou výhodu. Současně je však třeba zdůraznit, že některé Kaggle notebooky uvádějí výrazně vyšší hodnoty accuracy či F1 než bylo dosaženo, ale často za odlišných podmínek. V jednom případě byla původní vícetřídní úloha redukována pouze na tři třídy, v jiném dokonce na binární klasifikaci, a v dalším byl použit SMOTE pro vyvážení tříd. Úpravy sice mohou vést ke zlepšení metrik, avšak zároveň mění samotnou obtížnost úlohy a částečně i její interpretaci. Výsledky dosažené v práci jsou proto nižší, ale metodicky realističtější, protože zachovávají původní vícetřídní charakter problému a neodstraňují jeho přirozenou složitost.

Zmíněný aspekt je důležitý i vzhledem k výzkumné otázce, zda lze kvalitu vína spolehlivě predikovat na základě fyzikálně-chemických parametrů. Dosavadní výsledky ukazují, že odpověď zní spíše „ano, ale pouze do určité míry“. Modely jsou schopny zachytit významnou část variability hodnocení, což dokládá jak úspěšnost nejlepších klasifikátorů, tak hodnoty R^2 u regresních modelů. Zároveň však dosažené metriky i pozorovaný překryv tříd naznačují, že samotné laboratorní parametry zřejmě nevysvětlují celé výsledné senzorické hodnocení. Část informace může být vázána na faktory, které v datasetu obsaženy nejsou, případně na samotnou subjektivitu cílového hodnocení. Zmíněné důvody představují důležité omezení celé úlohy i použitých modelovacích metod.

Zatímco otázka vlivu nerovnoměrného zastoupení tříd bude detailněji zodpovězena v navazujících experimentech, již lze naznačit, že dominance středních hodnot quality a obecný překryv sousedních tříd představují jeden z hlavních faktorů omezujících výkonnost klasifikačních modelů. Modely jsou totiž přirozeně tlačeny, aby optimalizovaly výkon na nejčetnějších kategoriích, zatímco rozlišení okrajových tříd je obtížnější. Podobně také otázka významu jednotlivých fyzikálně-chemických vlastností bude přesněji rozvedena v dalších částech práce, nicméně už z porovnání s příbuznými Kaggle studiemi i z vlastních explorativních analýz vyplývá, že vlastnosti jako alcohol nebo volatile acidity patří mezi důležitější prediktory, avšak samy o sobě nestačí k jednoznačnému oddělení všech tříd kvality.

Úloha predikce kvality vína se v daném datasetu ukazuje jako problém s výraznou nelinearitou, částečným překryvem tříd a omezenou separovatelností sousedních kategorií. V takovém prostředí se nejlépe uplatňují robustní stromové ensemble modely, zejména Random Forest,

zatímco jednodušší lineární přístupy narážejí na své strukturální limity. Současně se ukazuje, že ne každá kombinace více modelů automaticky vede ke zlepšení výkonu; pokud ensemble neobsahuje nejsilnější model nebo pokud základní modely chybují podobně, může být jeho přínos omezený.

4.2 Vliv struktury datasetu na výkon modelů

Jedním z klíčových faktorů ovlivňujících výkon prediktivních modelů je samotná struktura datasetu, zejména rozdělení cílové proměnné quality. Jak bylo ukázáno v rámci explorační analýzy dat, dataset je výrazně nevyvážený, přičemž dominantní zastoupení mají střední hodnoty kvality (5–7), zatímco okrajové třídy (3–4 a 8–9) jsou zastoupeny pouze v omezeném počtu případů. Nerovnoměrnost se zásadním způsobem promítá do chování modelů a představuje jeden z hlavních limitujících faktorů celé úlohy.

Výsledky experimentu zaměřeného na výkon modelů podle četnosti vzorků (viz kapitola 3.3.) ukazují, že výkonnost modelů není v rámci datového prostoru homogenní, ale výrazně závisí na hustotě dat v jednotlivých oblastech. Ve skupině četných tříd (5–7), kde je k dispozici dostatečné množství trénovacích dat, dosahují modely relativně vysokých hodnot metrik (např. F1 score ≈ 0,73). V minoritních skupinách dochází k dramatickému poklesu výkonnosti, kdy například ve spodní skupině (3–4) klesá F1 score až k hodnotám okolo 0,18 a accuracy přibližně k 0,10.

Ještě výraznější rozdíly lze pozorovat v regresní úloze. Zatímco ve střední části rozdělení dosahují modely hodnot koeficientu determinace R^2 přibližně 0,53, v minoritních skupinách jsou hodnoty R^2 výrazně záporné (např. až −63 u horních tříd). Záporná hodnota R^2 přitom znamená, že model poskytuje horší predikce než triviální model založený na průměrné hodnotě. V méně četných oblastech modely zcela selhávají a nejsou schopny zachytit vztah mezi vstupními atributy a cílovou proměnnou.

Jev lze vysvětlit kombinací dvou faktorů. Prvním je samotná nerovnováha datasetu, která má za následek, že modely jsou během trénování optimalizovány především na dominantní část datového rozdělení. Druhým faktorem je překryv mezi jednotlivými třídami kvality, který byl pozorován již v rámci PCA a LDA analýzy. V důsledku překryvu nejsou sousední třídy (např. 5 a 6 nebo 6 a 7) jednoznačně oddělitelné, a to zejména v oblastech, kde je k dispozici pouze malé množství trénovacích vzorků. Kombinace nízké četnosti a vysoké míry překryvu tak představuje pro modely výrazně obtížnější problém než samotná nerovnováha dat.

Významným důsledkem je tzv. regrese k průměru, kdy modely mají tendenci predikovat hodnoty blízké nejčetnější části rozdělení. V klasifikační úloze se jev projevuje preferencí středních tříd kvality, zatímco extrémní hodnoty jsou často nesprávně klasifikovány jako běžnější kategorie. V regresní úloze se efekt projevuje „stahováním“ predikcí směrem k průměrné hodnotě quality, což vede k výrazným chybám právě u okrajových hodnot. Mechanismus vysvětluje, proč modely dosahují dobrých globálních metrik, ale současně selhávají v minoritních částech datového prostoru.

Získané výsledky zároveň ukazují, že globální metriky, jako jsou celkové F1 score, accuracy nebo RMSE, mohou poskytovat zkreslený obraz o skutečné výkonnosti modelu. Model, který dosahuje dobrých agregovaných výsledků, může ve skutečnosti fungovat dobře pouze

na dominantní části dat, zatímco v méně zastoupených, ale často interpretačně zajímavějších oblastech selhává. Analýza výkonu podle četnosti vzorků proto představuje důležitý doplněk standardního hodnocení a umožňuje identifikovat slabiny modelu, které by jinak zůstaly skryté.

Podobné chování modelů je popsáno i v komunitních projektech na platformě Kaggle, kde je opakovaně zdůrazňována jak nevyváženost tříd, tak i jejich vzájemný překryv. Faktory jsou uváděny jako hlavní příčina obtížnosti úlohy a limitů dosažitelné přesnosti.

Z pohledu výzkumných otázek lze konstatovat, že nerovnoměrné zastoupení tříd má zásadní vliv na výkon klasifikačních modelů, a to zejména ve smyslu výrazného zhoršení schopnosti modelů rozpoznávat minoritní kategorie. Současně se ukazuje, že samotná velikost datasetu není jediným rozhodujícím faktorem – klíčovou roli hraje také struktura datového prostoru a míra separovatelnosti jednotlivých tříd. Poznatky tvoří důležité východisko pro navazující experimenty, které se zaměřují na úpravu struktury datasetu a její dopad na stabilitu a přesnost modelů.

4.3 Vliv vyvážení tříd na výkon modelu

V rámci experimentu zaměřeného na vyvážení tříd (viz kapitola 3.4.) byl analyzován dopad downsamplingu na výkon klasifikačního modelu v podmnožině dat obsahující pouze třídy kvality 5, 6 a 7. Výsledky ukazují, že model natrénovaný na původním datasetu dosahuje výrazně vyšší výkonnosti než model trénovaný na uměle vyvážených datech. Konkrétně došlo k poklesu accuracy z přibližně 0,96 na 0,68 a k obdobnému snížení hodnot weighted F1-score.

Na první pohled by bylo možné očekávat, že odstranění nerovnováhy tříd povede ke zlepšení schopnosti modelu rozlišovat jednotlivé kategorie. Výsledky experimentu však ukazují opačný efekt, což naznačuje, že samotná nevyváženost datasetu není jediným ani dominantním problémem úlohy. Klíčovou roli zde hraje způsob, jakým bylo vyvážení provedeno.

Hlavním důvodem poklesu výkonnosti modelu je výrazná ztráta informace způsobená downsamplingem. Při vyvážení tříd byly dominantní kategorie redukovány na úroveň nejméně zastoupené třídy, čímž došlo ke snížení počtu trénovacích vzorků přibližně z 6053 na 3237, tedy téměř o polovinu. Krok nevede pouze ke snížení množství dat, ale také ke ztrátě variability a vnitřní struktury datového prostoru. Model tak nemá k dispozici reprezentativní obraz původního rozdělení a jeho schopnost generalizace se výrazně zhoršuje.

Jev lze interpretovat také z pohledu bias–variance tradeoff. Snížením počtu trénovacích vzorků dochází ke zvýšení bias modelu, protože model je nucen učit se z menšího a méně rozmanitého vzorku dat. Současně se snižuje variabilita modelu, avšak za cenu ztráty schopnosti zachytit komplexní vztahy mezi vstupními atributy a cílovou proměnnou. Výsledkem je jednodušší, ale méně přesný model.

Zajímavým zjištěním je, že u obou modelů jsou hodnoty F1-score (weighted) a F1-score (macro) velmi podobné. Jev naznačuje, že modely dosahují relativně vyrovnaného výkonu napříč jednotlivými třídami a nejsou výrazně ovlivněny dominancí některé z kategorií. Lze to přičíst skutečnosti, že experiment je omezen pouze na tři nejčastější třídy kvality (5, 6 a 7), jejichž zastoupení je již v původním datasetu relativně vyrovnané.

Na rozdíl od očekávání se tedy neprojevuje výrazné zlepšení „férovosti“ modelu po aplikaci vyvážení tříd. Hlavním efektem downsamplingu tak není změna rozložení chyb mezi třídami, ale především výrazná ztráta trénovacích dat, která vede k celkovému poklesu predikční výkonnosti modelu.

Důležitým faktorem je také skutečnost, že nevyváženost datasetu v tomto případě pravděpodobně odráží reálné rozdělení kvality vína. Střední hodnoty kvality jsou přirozeně častější než extrémní případy, a model trénovaný na původním datasetu tak lépe odpovídá reálným podmínkám. Umělé vyvážení tříd pomocí downsamplingu sice eliminuje nerovnováhu, ale zároveň narušuje přirozenou strukturu dat a vede k modelu, který se učí na méně realistickém rozdělení.

Srovnání s komunitními projekty na platformě Kaggle ukazuje, že způsob vyvažování dat má zásadní vliv na výsledný výkon modelu. Například v notebooku Wine Quality Prediction – Comparing Top ML Models (2022) je pro řešení nevyváženosti využita metoda SMOTE, která na rozdíl od downsamplingu synteticky generuje nové vzorky minoritních tříd. Použitý přístup vede k rozšíření datasetu a zachování informace, což může vysvětlovat zlepšení metrik zaměřených na méně zastoupené třídy. Naopak použitý downsampling, informaci odstraňuje, což se negativně promítá do celkové výkonnosti modelu.

Z pohledu výzkumných otázek lze konstatovat, že vyvážení tříd může vést ke stabilnějšímu a rovnoměrnějšímu chování modelu, avšak není zaručeno, že povede ke zlepšení jeho celkové přesnosti. V konkrétním případě vedlo vyvážení tříd pomocí downsamplingu k výraznému zhoršení výkonu modelu, což naznačuje, že zachování původní struktury dat může být v některých případech výhodnější než její umělá úprava. Zjištění zároveň ukazuje, že při návrhu modelu je nutné zohlednit konkrétní cíle úlohy a rozhodnout, zda je prioritou maximální přesnost, nebo vyvážený výkon napříč třídami.

4.4 Vliv redukce vstupních atributů na výkon modelu

V experimentu zaměřeném na redukci vstupních atributů (viz kapitola 3.5.) byl analyzován dopad omezení počtu vstupních proměnných na výkon regresních a klasifikačních modelů. Redukce byla provedena výběrem čtyř nejvýznamnějších atributů na základě LDA analýzy, zatímco referenční model využíval plnou množinu dvanácti proměnných.

Výsledky ukazují, že modely využívající plnou množinu atributů dosahují konzistentně lepších výsledků než modely založené pouze na omezené množině proměnných. V regresní úloze došlo při redukci atributů ke zvýšení hodnot RMSE z přibližně 0,60 na 0,68 a současně k poklesu koeficientu determinace R^2 z hodnoty okolo 0,52 na 0,39. Podobný trend lze pozorovat i v klasifikační úloze, kde accuracy klesla z přibližně 0,69 na 0,63 a hodnoty F1-score vykazují obdobné zhoršení.

Výsledky naznačují, že omezení vstupního prostoru vedlo ke ztrátě části informace důležité pro přesnou predikci. Nejedná se však pouze o ztrátu „méně významných“ atributů v tradičním smyslu. Klíčovým zjištěním je, že i proměnné, které nevykazují silný individuální vztah k cílové proměnné, mohou v kombinaci s ostatními atributy přispívat k celkovému výkonu modelu. Modely, zejména nelineární přístupy jako Random Forest, totiž nevyužívají pouze jednotlivé atributy izolovaně, ale dokážou zachytit komplexní interakce mezi více proměnnými současně.

Redukce atributů tak vede nejen ke zjednodušení modelu, ale také k odstranění doplňujících vztahů. Výsledný model je sice méně komplexní, avšak současně méně schopný zachytit jemné struktury v datech. Efekt lze interpretovat i z pohledu bias–variance tradeoff: snížení počtu atributů vede ke zvýšení bias modelu, protože model pracuje s omezeným popisem datového prostoru, a důsledkem ztrácí část predikční schopnosti.

Získané výsledky zároveň ukazují, že význam atributů identifikovaný v rámci explorační analýzy dat nelze přímo interpretovat jako optimální množinu vstupních proměnných pro finální modelování. Zatímco EDA a metody jako LDA hodnotí jednotlivé atributy primárně z hlediska jejich individuální diskriminační schopnosti, samotné modelování využívá širší kontext a kombinace proměnných, které nemusí být zřejmé při izolovaném posuzování. Rozdíl vysvětluje, proč redukce na „nejdůležitější“ atributy nevede ke zlepšení výkonu modelu.

Dalším důležitým faktorem je skutečnost, že použité modely, zejména Random Forest, již implicitně provádějí výběr atributů během trénování. Každý strom v rámci ensemble pracuje pouze s náhodnou podmnožinou proměnných, což přirozeně potlačuje vliv méně informativních atributů. Explicitní redukce vstupního prostoru tak může být v tomto případě redundantní nebo dokonce kontraproduktivní.

Z hlediska výpočetní náročnosti přinesla redukce atributů určité zrychlení, zejména v době tréninku modelu, která se přibližně snížila na polovinu. Přínos však nebyl natolik výrazný, aby kompenzoval ztrátu predikční přesnosti. Doba predikce zůstala prakticky nezměněna, což naznačuje, že počet atributů nemusí být hlavním faktorem ovlivňujícím inferenční čas.

Podobné závěry lze nalézt i v komunitních projektech na platformě Kaggle. Například v notebooku Wine Quality Prediction – Comparing Top ML Models (2022) byly testovány metody výběru atributů, jako je RFE nebo VIF, přičemž nedošlo k významnému zlepšení výkonu modelu a finálně byla ponechána plná množina proměnných. Výsledek podporuje závěr, že dataset již obsahuje relevantní a dostatečně informativní sadu atributů a jejich redukce nemusí být přínosná.

Z pohledu výzkumných otázek lze konstatovat, že redukce vstupních atributů vede ke snížení výkonu modelů a nepředstavuje vhodný způsob, jak zlepšit jejich přesnost. Současně však ukazuje, že mezi přesností modelu a jeho jednoduchostí existuje přirozený kompromis. Zatímco redukovaná množina atributů může vést k jednoduššímu a rychlejšímu modelu, plná množina vstupních proměnných poskytuje vyšší predikční výkon. Kompletní sada atributů se ukazuje jako přesnější, zatímco redukovaná varianta představuje spíše zjednodušený model s omezenou přesností.

Závěr

Cílem práce bylo analyzovat možnosti predikce kvality vína na základě fyzikálně-chemických parametrů a navrhnout systematický analytický postup zahrnující přípravu dat, modelování a vyhodnocení výsledků. Cíl byl naplněn prostřednictvím komplexní analýzy datasetu, aplikace různých přístupů strojového učení a provedení experimentů zaměřených na klíčové faktory ovlivňující výkon modelů.

Výsledky práce ukazují, že kvalitu vína lze na základě dostupných fyzikálně-chemických atributů predikovat, avšak predikce je výrazně limitována strukturou dat. Modely dosahují uspokojivých výsledků především ve střední části rozdělení kvality, zatímco v okrajových třídách jejich výkon výrazně klesá. Jev souvisí nejen s nerovnoměrným zastoupením tříd, ale také s výrazným překryvem mezi sousedními hodnotami kvality, který byl identifikován v rámci PCA a LDA analýzy. V důsledku dochází k efektu „regrese k průměru“, kdy modely mají tendenci predikovat nejčastější hodnoty kvality a selhávají u extrémních případů.

Z hlediska volby modelu bylo prokázáno, že nejlepších výsledků dosahují nelineární přístupy, zejména stromové ensemble modely. Random Forest se ukázal jako nejvýkonnější model jak v klasifikační, tak i regresní úloze, což potvrzuje jeho schopnost zachytit komplexní nelineární vztahy a interakce mezi vstupními atributy. Naopak lineární modely dosahovaly výrazně slabších výsledků, což odpovídá nelineární povaze analyzovaných dat. Zajímavým zjištěním je, že ensemble přístupy založené na kombinaci více modelů nepřinesly zlepšení oproti Random Forest modelu. Výsledek lze vysvětlit omezenou diverzitou použitých modelů a skutečností, že ensemble neobsahoval nejvýkonnější model typu Random Forest.

Významnou roli v predikci kvality vína hraje také kombinace vstupních atributů. Přestože explorační analýza identifikovala některé proměnné jako významnější, experiment zaměřený na redukci atributů ukázal, že omezení modelu pouze na vybrané proměnné vede ke zhoršení výkonu. Výsledek potvrzuje, že modely, zejména nelineární, využívají nejen individuální význam atributů, ale i jejich vzájemné interakce, které nelze jednoduše zachytit při izolovaném posuzování jednotlivých proměnných.

Dalším klíčovým zjištěním je vliv struktury datasetu na výkon modelů. Nerovnoměrné zastoupení tříd má za následek, že modely optimalizují svůj výkon především na dominantních kategoriích, zatímco minoritní třídy jsou predikovány s výrazně nižší přesností. Experimenty dále ukázaly, že vyvažování tříd pomocí downsamplingu nevedlo k významné změně rozložení výkonu mezi třídami, což lze pozorovat z podobnosti hodnot F1-score (weighted) a F1-score (macro) u obou modelů. Hlavním efektem vyvážení tak nebylo zlepšení „férovosti“ modelu, ale především výrazný pokles celkové přesnosti v důsledku ztráty části trénovacích dat.

Na základě zjištění lze konstatovat, že stanovený cíl práce byl splněn a formulované výzkumné otázky byly zodpovězeny. Výsledky ukazují, že fyzikálně-chemické parametry představují významný zdroj informací pro predikci kvality vína, avšak samy o sobě nejsou dostačující pro dosažení vysoké přesnosti. Klíčovým omezením úlohy byla kombinace nevyváženosti dat, překryvu tříd a omezené separovatelnosti jednotlivých kategorií kvality.

Mezi hlavní omezení práce patří použití jednoho konkrétního datasetu, který nemusí plně reprezentovat všechny typy vín, a omezený rozsah vstupních atributů. Dalším omezením

je nevyváženost dat a skutečnost, že nebyly využity všechny pokročilé metody pro její řešení ani rozsáhlá optimalizace hyperparametrů.

Pro další výzkum se nabízí několik směrů. Jedním z nich je využití pokročilejších ensemble metod, zejména boosting přístupů, a aplikace metod pro řešení nevyváženosti dat, například SMOTE. Perspektivní je také využití ordinalních modelů, které lépe odpovídají povaze cílové proměnné, a hlubší analýza interpretovatelnosti modelů pomocí metod, jako jsou SHAP nebo LIME.

Dalším možným směrem je úprava rozdělení vstupních proměnných. Vzhledem k tomu, že některé atributy vykazují výraznou šikmost rozdělení, lze uvažovat o aplikaci transformačních metod, například logaritmické transformace u proměnných s vysokou hodnotou šikmosti (např. větší než 1,5). Úprava by mohla vést ke stabilnějšímu chování regresních modelů a zlepšení jejich predikční schopnosti, což představuje vhodný směr pro další experimentální ověření.

Seznam použité literatury

AJIN, Rajendran Shobha; SEGONI, Samuele a FANTI, Riccardo. Optimization of SVR and CatBoost models using metaheuristic algorithms to assess landslide susceptibility. Online. Scientific Reports. 2024, vol. 14, no. 24851, s. 1-24. ISSN 2045-2322. Dostupné z: https://doi.org/10.1038/s41598-024-72663-x. [cit. 2026-03-28].

ALIFERIS, Constantin a SIMON, Gyorgy. Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI. Online. *Health Informatics*. 2024, vol. 2024, no. 1, s. 477–524. ISSN 2197-3741. Dostupné z: https://doi.org/10.1007/978-3-031-39355-6_10. [cit. 2026-03-22].

ALNUAIMI, Amer F.A.H. a ALBALDAWI, Tasnim H.K. An overview of machine learning classification techniques. Online. BIO Web of Conferences. 2024, vol. 97, no. 2024, s. 1-24. ISSN 2117-4458. Dostupné z: https://doi.org/10.1051/bioconf/20249700133. [cit. 2026-03-27].

ALPEZA, Ivana; BULJEVIĆ, Nina a BUDIMIR, Ivan. The reliability of wine sensory analysis results tested by different correlation models: the example of the impact of SARS-CoV-2 virus infection. Online. *Journal of Central European Agriculture*. 2025,. Roč. 26, č. 3, s. 710-719. ISSN 1332-9049. Dostupné z: https://doi.org/10.5513/JCEA01/26.3.4508. [cit. 2025-12-29].

AWAD, Mohammed a FRAIHAT, Salam. Recursive Feature Elimination with Cross-Validation with Decision Tree: Feature Selection Method for Machine Learning-Based Intrusion Detection Systems. Online. *Journal of Sensor and Actuator Networks*. 2023, vol. 12, no. 5, s. 2-23. ISSN 2224-2708. Dostupné z: https://doi.org/10.3390/jsan12050067. [cit. 2026-03-22].

BASALEKOU, Marianthi; TATARIDIS, Panagiotis; GEORGAKIS, Konstantinos a TSINTONIS, Christos. Measuring Wine Quality and Typicity. Online. Beverages. 2023, vol. 9, no. 2, s. 1-18. ISSN 2306-5710. Dostupné z: https://doi.org/10.3390/beverages9020041. [cit. 2026-03-23].

BATINI, Carlo; CAPPIELLO, Cinzia; FRANCALANCI, Chiara a MAURINO, Andrea. Methodologies for data quality assessment and improvement. Online. ACM Computing Surveys. 2009, vol. 41, no. 3, s. 1-52. ISSN 1557-7341. Dostupné z: https://doi.org/10.1145/1541880.1541883. [cit. 2026-03-22].

BLOCKEEL, Hendrik; DEVOS, Laurens; FRÉNAY, Benoît; NANFACK, Géraldin a NIJSSEN, Siegfried. Decision trees: from efficient prediction to responsible AI. Online. Frontiers in Artificial Intelligence. 2023, vol. 6, no. 1, s. 1-17. ISSN 2624-8212. Dostupné z: https://doi.org/10.3389/frai.2023.1124553. [cit. 2026-03-28].

BOTELLA, Juan a SÁNCHEZ-MECA, Juilo. Meta-analysis: Confidence intervals and Prediction intervals. Online. Anales de Psicología / Annals of Psychology. 2024, vol. 40, no. 2, s. 344-354. ISSN 1695-2294. Dostupné z: https://doi.org/10.6018/analesps.591831. [cit. 2026-03-25].

BRIJITH, Arya. Data Preprocessing for Machine Learning. Online. *International Center for AI and Cyber Security Research and Innovations (CCRI)*. 2023, vol. 3, no. 1, s. 1-4. Dostupné z: https://www.researchgate.net/publication/375003512_Data_Preprocessing_for_Machine_Learning. [cit. 2026-03-22].

CASONATTO, Ricardo Accorsi; SOUZA, Tales De Pádua Grillo a MARIANO, Ari Melo. Quality and Risk Management in Data Mining: A CRISP-DM Perspective. [Online]. Procedia Computer Science. 2024, vol. 242, no. 1, s. 161-168. ISSN 1877-0509. Dostupné z: https://doi.org/10.1016/j.procs.2024.08.257. [cit. 2026-03-22].

CATANIA, Pietro; VALLONE, Mariangela a PIPITONE, Felice. Analysis of the main factors influencing the quality of wine from mechanicallyharvested grapes. Online. *Journal of Agricultural Engineering*. 2009, vol. 40, no. 4, s. 27-30. ISSN 2239-6268. Dostupné z: https://doi.org/10.4081/jae.2009.87. [cit. 2026-03-23].

CENITA, Jonelle Angelo S.; ASUNCION, Paul Richie F. a VICTORIANO, Jayson M. Performance Evaluation of Regression Models in Predicting the Cost of Medical Insurance. Online. *International Journal of Computing Sciences Research*. 2023, roč. 7, č. 2, s. 2052-2065. ISSN 2546-115X. Dostupné z: https://doi.org/10.25147/ijcsr.2017.001.1.146. [cit. 2026-03-23].

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Wine Quality [Dataset]. [Online], UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T.

COURSERA. *What Is Kaggle and What Is It Used For?* Online. 2025. Dostupné z: https://www.coursera.org/articles/kaggle. [cit. 2025-12-29].

DOWNEY, Allen B. Think stats: exploratory data analysis. Third edition. Sebastopol, CA: O'Reilly, 2025. ISBN 978-1-098-19025-5.

ELSHAMY, Abdalla. *WINE QUALITY PREDICTION*. Online. Kaggle. 2023. Dostupné z: https://www.kaggle.com/code/abdallaelshamy/wine-quality-prediction#Decison-Tree. [cit. 2026-03-23].

FU, Haojie; TANG, Shuang a ZHAO, Xudong. Limitations of Correlation Coefficients in Research on Functional Connectomes and Psychological Processes. Online. *Human brain mapping*. 2025, vol. 46, no. 10, s. 1-7. ISSN 1097-0193. Dostupné z: https://doi.org/10.1002/hbm.70287. [cit. 2026-03-22].

GHASEMI, Asghar a ZAHEDIASL, Saleh. Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. [Online] International Journal of Endocrinology Metabolism. 2012, vol. 10, no. 2, s. 486-489. ISSN 1726-9148. Dostupné z: https://doi.org/10.5812/ijem.3505. [cit. 2026-03-22].

GONG, Youdi; LIU, Guangzhen; XUE, Yunzhi; LI, Rui a MENG, Lingzhong. A survey on dataset quality in machine learning. Online. Information and software technology. 2023, vol. 162, no. 1, s. 1-12. ISSN 1873-6025. Dostupné z: https://doi.org/10.1016/j.infsof.2023.107268. [cit. 2026-03-22].

HALDER, Rajib Kumar; UDDIN, Mohammed Nasir; UDDIN, Md. Ashraf; ARYAL, Sunil a KHRAISAT, Ansam. Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. Online. Journal of Big Data. 2024, vol. 11, no. 1. ISSN 2196-1115. Dostupné z: https://doi.org/10.1186/s40537-024-00973-y. [cit. 2026-03-30].

HARTMAN, Sarah; ONG, Cheng Soon; POWLES, Julia a KUHNERT, Petra. Position: We Need Responsible, Application-Driven (RAD) AI Research. Online. 2025, s. 1-12. Dostupné z: https://doi.org/https://doi.org/10.48550/arXiv.2505.04104. [cit. 2026-03-24].

HESKES, Tom. Practical confidence and prediction intervals. Online. Nijmegen, Nizozemsko: RWCP Novel Functions SNN Laboratory; University of Nijmegen Geert Grooteplein 21, 6525 EZ Nijmegen, The Netherlands, 1996. Dostupné z: https://proceedings.neurips.cc/paper_files/paper/1996/file/7940ab47468396569a906f75ff3f20ef-Paper.pdf. [cit. 2026-03-25].

CHAPMAN, Pete; CLINTON, Julian; KERBER, Randy; KHABAZA, Thomas; REINARTZ, Thomas et al. CRISP-DM 1.0. [Online]. SPSS, 2000. Dostupné z: https://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/es/CRISP-DM.pdf. [cit. 2025-11-22].

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. An introduction to statistical learning: with applications in R. Springer texts in statistics. New York: Springer, 2013. ISBN 978-1-4614-7138-7.

JEWSBURY, Paul A. a JOHNSON, Matthew S. Principal component analysis on the covariance matrix for data reduction in large-scale assessments. Online. *Large-scale Assessments in Education*. 2025, vol. 13, no. 30, s. 1-40. ISSN 2196-0739. Dostupné z: https://doi.org/10.1186/s40536-025-00264-9. [cit. 2026-03-22].

JIANG, Xianghui; LIU, Xuanyu; WU, Yutong a YANG, Dehuai. White Wine Quality Prediction and Analysis with Machine Learning Techniques. Online. *Highlights in Science, Engineering and Technology*. 2023, vol. 39, no. 1, s. 321-326. ISSN 2791-0210. Dostupné z: https://doi.org/10.54097/hset.v39i.6548. [cit. 2026-03-23].

KHOSHKROODI, A.; SANI, H. Parvini a AAJAMI, M. Stacking Ensemble-Based Machine Learning Model for Predicting Deterioration Components of Steel W-Section Beams. Online. *Buildings*. 2024, vol. 14, no. 1, s. 1-21. ISSN 2075-5309. Dostupné z: https://doi.org/10.3390/buildings14010240. [cit. 2026-03-22].

KIM, Jong Hae. Multicollinearity and misleading statistical results. Online. *Korean Journal of Anesthesiology*. 2019, vol. 72, no. 6, s. 1-12. ISSN 2005-7563. Dostupné z: https://doi.org/10.4097/kja.19087. [cit. 2026-03-22].

KOMOROWSKI, Matthieu; MARSHALL, Dominic C.; SALCICCIOLI, Justin D. a CRUTAIN, Yves. Exploratory Data Analysis. Online. Secondary Analysis of Electronic Health Records. 2016, s. 185-203. Dostupné z: https://doi.org/10.1007/978-3-319-43742-2_15. [cit. 2026-03-22].

KROESE, Dirk P.; BOTEV, Zdravko I.; TAIMRE, Thomas a VAISMAN, Radislav. Data Science and Machine Learning Mathematical and Statistical Methods. Online. 2024. CRC Press, 2019. ISBN 9781032488684. Dostupné z: https://people.smp.uq.edu.au/DirkKroese/DSML/DSML.pdf. [cit. 2026-03-22].

KUNAPULI, Gautam. Ensemble Methods for Machine Learning. Online. Manning Publications Co., 2023. ISBN 9781617297137. Dostupné z: https://books.google.cz/books?hl=cs&lr=&id=hoK3EAAAQBAJ&oi=fnd&pg=PR11&dq=ensemble+methods&ots=FczVHsZ16c&sig=fwhUqw1wvbnxyKw6rcsJYf2euXQ&redir_esc=y#v=onepage&q=ensemble%20methods&f=false. [cit. 2026-03-28].

KUO, Chih-Ting; XU, Duo a FRIESEN, Rachel. A Brief Review of Unsupervised Machine Learning Algorithms in Astronomy: Dimensionality Reduction and Clustering. Online. *Universe*. 2025, roč. 2025, č. 11, s. 50. ISSN 2218-1997 Dostupné z: https://doi.org/https://doi.org/10.3390/universe11120412. [cit. 2025-12-29].

LIU, Tianze. The Research on Influence Factors of Red Wine Quality. Online. *Transactions on Economics Business and Management Research*. 2024, vol. 10, no. 1, s. 292-297. ISSN 2960-2254. Dostupné z: https://doi.org/10.62051/xxq55t12. [cit. 2026-03-23].

LOPEZ, Eneko; GORLA, Giulia; ETXEBARRIA-ELEZGARAI, Jaione; AMIGO, Jose Manuel a SEIFERT, Andreas. The importance of choosing a proper validation strategy in predictive models. Part 2: Recipes for (avoiding) overfitting-A tutorial. Online. *Analytica Chimica Acta*. 2025, vol. 1384, no. 1, s. 1-15. ISSN 1873-4324. Dostupné z: https://doi.org/10.1016/j.aca.2025.344838. [cit. 2026-03-23].

MACKAY, Hugh. Rapid application development (RAD): an empirical review. Online. European Journal of Information Systems, 1999. Dostupné z: https://www.academia.edu/94828796/Rapid_application_development_RAD_an_empirical_review. [cit. 2025-11-22].

MAHESH, T R; KUMAR, Vinoth; KUMAR, Dhilip; GEMAN, Oana; MARGALA, Martin et al. The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification. Online. *Healthcare Analytics*. 2023, vol. 4, no. 1, s. 1-10. ISSN 2772-4425. Dostupné z: https://doi.org/10.1016/j.health.2023.100247. [cit. 2026-03-23].

MANCONI, Andrea; ARMANO, Giuliano; GNOCCHI, Matteo a MILANESI, Luciano. A Soft-Voting Ensemble Classifier for Detecting Patients Affected by COVID-19. Online. Applied Sciences. 2022, vol. 2022, no. 12, s. 1-23. ISSN 2076-3417. Dostupné z: https://doi.org/10.3390/app12157554. [cit. 2026-03-22].

MAURYA, Neelesh Kumar. Selection and Performance of Sensory Panelists: A Comprehensive Review of Factors Influencing Sensory Evaluation Outcomes. Online. *Nutrition and Food Processing*. 2024, roč. 2024, č. 15, s. 7. Dostupné z: https://doi.org/10.31579/2637-8914/278. [cit. 2025-12-29].

MIENYE, Ibomoiye Domor a JERE, Nobert. A Survey of Decision Trees: Concepts, Algorithms, and Applications. Online. IEEE access. 2024, vol. 12, no. 1, s. 86716-86727. ISSN 2169-3536. Dostupné z: https://doi.org/10.1109/ACCESS.2024.3416838. [cit. 2026-03-28].

MIENYE, Ibomoiye Domor a SUN, Yanxia. A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. Online. *IEEE Access*. 2022, vol. 10, no. 1, s. 1-21. ISSN 2169-3536. Dostupné z: https://doi.org/10.1109/ACCESS.2022.3207287. [cit. 2026-03-22].

MILLER, Catriona; PORTLOCK, Theo; NYAGA, Denis M. a O'SULLIVAN, Justin M. A review of model evaluation metrics for machine learning in genetics and genomics. Online. *Frontiers in Bioinformatics*. 2024, roč. 4. ISSN 2673-7647. Dostupné z: https://doi.org/10.3389/fbinf.2024.1457619. [cit. 2026-03-30].

MONFARED, Melissa. *Red Wine Quality Prediction ◂ EDA & Classification*. Online. Kaggle. 2025. Dostupné z: https://www.kaggle.com/code/melissamonfared/red-wine-quality-prediction-eda-classification?. [cit. 2026-03-23].

QU, Lingxiao a PEI, Yan. A Comprehensive Review on Discriminant Analysis for Addressing Challenges of Class-Level Limitations, Small Sample Size, and Robustness. Online. *Processes*. 2024, vol. 12, no. 7, s. 1-32. ISSN 2227-9717. Dostupné z: https://doi.org/10.3390/pr12071382. [cit. 2026-03-22].

RAUF, Rauf I.; ALRASHEEDI, Masad A.; SADIQ, Rasheedah a ALDAWSARI, Abdulrahman M. A. Evaluating Predictive Accuracy of Regression Models with First-Order Autoregressive Disturbances: A Comparative Approach Using Artificial Neural Networks and Classical Estimators. Online. *Mathematics*. 2024, roč. 12, č. 24, s. 1-23. ISSN 2227-7390. Dostupné z: https://doi.org/10.3390/math12243966. [cit. 2026-03-23].

RUTHGN. Wine Quality Data Set (Red & White Wine). Online. Kaggle. 2024. Dostupné z: https://www.kaggle.com/datasets/ruthgn/wine-quality-data-set-red-white-wine. [cit. 2025-12-09].

SALMAN, Hasan Ahmed; KALAKECH, Ali a STEITI, Amani. Random Forest Algorithm Overview. Online. *Babylonian Journal of Machine Learning*. 2024, vol. 2024, no. 1, s. 69-79. ISSN 3006–5429. Dostupné z: https://doi.org/10.58496/BJML/2024/007. [cit. 2026-03-22].

SANDVE, Geir Kjetil; NEKVIND, Anders; KJETIL, Liland; NEDERBRAGT, Lex; HEDLEY, Simon a kol. Ten Simple Rules for Reproducible Computational Research**.** Online. *PLoS Computational Biology*. 2013, roč. 9, č. 10. s. 1-4. ISSN 1553-7358. Dostupné z: https://doi.org/10.1371/journal.pcbi.1003285. [cit. 2025-12-29].

SHAHNAWAZ, Mohammad a KUMAR, Manish. A Comprehensive Survey on Big Data Analytics: Characteristics, Tools and Techniques. [Online]. *ACM Computing Surveys*. 2025, vol. 57, no. 8, s. 1-33. ISSN 1557-7341. Dostupné z: https://doi.org/10.1145/3718364 [cit. 2026-03-22].

STANEVICIENE, Evelina; GUDONIENE, Daina; PUNYS, Vytenis a KUKSTYS, Arturas. A Case Study on the Data Mining-Based Prediction of Students' Performance for Effective and Sustainable E-Learning. [Online] Sustainability. 2024, vol. 16, no. 23, s. 1-15. ISSN 2071-1050. Dostupné z: https://doi.org/10.3390/su162310442. [cit. 2026-03-22].

STARCKE, Jonathan; SPADAFORA, James; SPADAFORA, Jonathan; SPADAFORA, Phillip a TOMA, Milan. The Effect of Data Leakage and Feature Selection on Machine Learning Performance for Early Parkinson's Disease Detection. Online. Bioengineering. 2025, vol. 12, no. 845, s. 1-21. ISSN 2306-5354. Dostupné z: https://doi.org/10.3390/bioengineering12080845. [cit. 2026-03-23].

Step by guide to Predict Red Wine quality + EDA. Online. 2021. Dostupné z: https://www.kaggle.com/code/vipin20/step-by-guide-to-predict-red-wine-quality-eda?. [cit. 2025-12-13].

UMARGONO, Edy; SUSENO, Jatmiko Endro a GUNAWAN, S.K Vincensius. K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula. Online. *Advances in Social Science, Education and Humanities Research*. 2020, vol. 474, no. 1, s. 121-129. ISSN 2352-5398. Dostupné z: https://doi.org/10.2991/assehr.k.201010.019. [cit. 2026-03-22].

VICTORIA DA POIAN; THEILING, Bethany; CLOUGH, Lily; BRETT A. MCKINNEY; JONATHAN D. MAJOR et al. Exploratory data analysis (EDA) machine learning approaches for ocean world analog mass spectrometry. Online. Frontiers in Astronomy and Space Sciences. 2023, vol. 10. ISSN 2296-987X. Dostupné z: https://doi.org/10.3389/fspas.2023.1134141. [cit. 2026-03-30].

VUJOVIĆ, Željko Đ. Classification Model Evaluation Metrics. Online. *International Journal of Advanced Computer Science and Applications*. 2021, vol. 12, no. 6, s. 599-606. ISSN 2156-5570. Dostupné z: https://doi.org/10.14569/IJACSA.2021.0120670. [cit. 2026-03-23].

WANG, Huajun; LI, Genghui a WANG, Zhenkun. Fast SVM classifier for large-scale classification problems. Online. Information Sciences. 2023, vol. 42, no. 119131. ISSN 0020-0255. Dostupné z: https://doi.org/10.1016/j.ins.2023.119136. [cit. 2026-03-28].

WANG, Jingran; LIU, Yi; LI, Peigong; LIN, Zhenxing; SINDAKIS, Stavros et al. Overview of Data Quality: Examining the Dimensions, Antecedents, and Impacts of Data Quality. Online. *Journal of the Knowledge Economy*. 2023, roč. 2023, č. 1, s. 20. Dostupné z: https://doi.org/10.1007/s13132-022-01096-6. [cit. 2025-12-29].

YASSER, M H. *Wine Quality Prediction -(Comparing Top ML Models)*. Online. Kaggle. 2022. Dostupné z: https://www.kaggle.com/code/yasserh/wine-quality-prediction-comparing-top-ml-models#7.-Gradient-Boosting-Classfier. [cit. 2026-03-23].

ZALOUMIS, Sophie G.; RAJASEKHAR, Megha a SIMPSON, Julie A. How to use learning curves to evaluate the sample size for malaria prediction models developed using machine learning algorithms. Online. Malaria Journal. 2025, vol. 24, no. 242, s. 1-11. ISSN 1475-2875. Dostupné z: https://doi.org/10.1186/s12936-025-05479-3. [cit. 2026-03-23].

ZHANG, Lili. Looking Back to the Future: A Glimpse at Twenty Years of Data Science. [Online] D ata science journal. 2026, vol. 22, no. 1, s. 1-9. ISSN 1683-1470. Dostupné z: https://doi.org/10.5334/dsj-2023-007 [cit. 2026-03-22].

Přílohy A

Příloha A.1 Kaggle notebook

Název: Data Science: Wine Quality Prediction

Dostupné z: https://www.kaggle.com/code/pavelplhal/data-science-wine-quality-prediction