Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra informačních technologií

Bakalářská práce

Čištění dat pomocí strojového učení

Alisher Baibussinov

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

Provozně ekonomická fakulta

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Alisher Baibussinov

Informatika

Název práce

Čištění dat pomocí strojového učení

Název anglicky

Data cleaning using machine learning

Cíle práce

Hlavním cílem práce je analyzovat možnosti čistění dat pomocí strojového učení a zhodnotit vhodnost zvoleného řešení na vybrané datové sadě.
Dílčí cíle práce jsou:
- na základě studia odborných informačních zdrojů analyzovat dostupná řešení pro čištění dat
- analyzovat možnosti využití strojového učení jako nástroje pro čištění dat
- navrhnout a implementovat konkrétní způsob řešení na testovací sadě dat a zhodnotit vhodnost zvoleného postupu

Metodika

Metodika teoretické části práce je založena na studiu a analýze dostupných literárních a online informačních zdrojů v oblasti čištění dat a strojového učení. Metodika praktické části spočívá ve vymezení běžných chyb v datech a způsobech jejich odstranění. Budou analyzována možnosti využití strojového učení pro čištění dat. Bude navržen postup řešení a bude implementován v praxi za použití jazyka Python. Bude provedena validace modelu a změřena jeho úspěšnost při práci se zvolenou datovou sadou a následně určena vhodnost zvoleného postupu. Na základě sjednocení poznatků teoretické a praktické části budou formulovány závěry práce.

Doporučený rozsah práce

40-50

Klíčová slova

čištění dat, zpracování dat, strojové učení, Python, příprava dat

---

Doporučené zdroje informací

Aurélien, Géron. Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems. Beijing: O'Reilly, 2022. ISBN-10: 1492032646.

Bruce, Peter; Bruce, Andrew; Gedeck, Peter. Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python. Sebastopol, CA: O'Reilly, 2020. ISBN-10: 149207294X.

Chu, Xu; Ilyas, Ihab F. Data Cleaning. New York, NY: Association for Computing Machinery, 2019. ISBN-10: 1450371531.

McKinney, Wes. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. Sebastopol, CA: O'Reilly Media, Inc., 2018. ISBN-10: 1491957662.

Předběžný termín obhajoby

2022/23 LS – PEF

Vedoucí práce

Ing. Jan Pavlík

Garantující pracoviště

Katedra informačních technologií

Elektronicky schváleno dne 19. 6. 2022

doc. Ing. Jiří Vaněk, Ph.D.

Vedoucí katedry

Elektronicky schváleno dne 27. 10. 2022

doc. Ing. Tomáš Šubrt, Ph.D.

Děkan

V Praze dne 25. 08. 2023

Čestné prohlášení

Prohlašuji, že svou bakalářskou práci "Čištění dat pomocí strojového učení" jsem vypracoval samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou citovány v práci a uvedeny v seznamu použitých zdrojů na konci práce. Jako autor uvedené bakalářské práce dále prohlašuji, že jsem v souvislosti s jejím vytvořením neporušil autorská práva třetích osob.

V Praze dne 15.03.2024 _____

Poděkování

Rád bych touto cestou poděkoval svému vedoucímu Ing. Janu Pavlíkovi, Ph.D., který mi po celou dobu psaní této práce poskytoval dokonalé vedení a podporu.

Čištění dat pomocí strojového učení

Abstrakt

Tato práce zkoumá použitelnost a efektivitu využití algoritmů strojového učení pro úlohy čištění dat. Prostřednictvím komplexního přehledu existující literatury a odborných zdrojů jsou analyzována různá řešení čištění dat, zejména s důrazem na použitelnost přístupů strojového učení. Teoretická východiska zahrnují témata, jako jsou hodnocení kvality dat, průzkumná analýza dat, zpracování chybějících dat, metodiky detekce odlehlých hodnot a techniky výběru rysů. V praktické části této práce je navrženo a implementováno konkrétní řešení založené na strojovém učení pro vícerozměrnou detekci odlehlých hodnot. Jako primární algoritmus pro detekci odlehlých hodnot je použit algoritmus Support Vector Machines (SVM), a to díky jeho schopnosti zpracovávat komplexní a vícerozměrná data. Pro vyhodnocení vhodnosti a výkonnosti navrženého přístupu je opatrně vybrán testovací datový soubor.

Klíčová slova: čištění dat, zpracování dat, příprava dat, strojové učení, detekce odlehlých hodnot, Support Vector Machines, kvalita dat, průzkumná analýza dat, chybějící data, výběr prediktorů, python.

Data cleaning with machine learning

Abstract

This paper investigates the applicability and effectiveness of using machine learning algorithms for data cleaning tasks. Through a comprehensive review of existing literature and scholar sources, various data cleaning solutions are analysed, with particular emphasis on the applicability of machine learning approaches. The theoretical background includes topics such as data quality assessment, exploratory data analysis, missing data handling, outlier detection methodologies, and feature selection techniques. In the practical part of this work, a specific machine learning based solution for multivariate outlier detection is proposed and implemented. Support Vector Machines (SVM) algorithm is used as the primary algorithm for outlier detection due to its ability to handle complex and high dimensional data. A test dataset is carefully selected to evaluate the suitability and performance of the proposed approach.

Keywords: data cleaning, data processing, data preparation, machine learning, outlier detection, Support Vector Machines, data quality, exploratory data analysis, missing data, feature selection, python.

Obsah

1 Úvod ... 10

2 Cíl práce a metodika ... 11

2.1 Cíl ... 11

2.2 Metodika ... 11

3 Teoretická východiska ... 12

3.1 Kvalita dat ... 12

3.1.1 Typy problémů ... 12

3.1.2 Dimenze kvality dat ... 12

3.1.3 Strategie zlepšování kvality dat ... 13

3.1.4 Metody hodnoceni kvality dat ... 14

3.2 Průzkum dat ... 14

3.2.1 Úrovně organizace dat ... 15

3.2.2 Úrovně měření dat ... 16

3.2.3 Vizualizační techniky ... 17

3.3 Chybějící data ... 20

3.3.1 Vzory chybějících dat ... 20

3.3.2 Mechanismy chybění ... 20

3.3.3 Metody odstraňování ... 21

3.3.4 Často používané metody imputace ... 22

3.3.5 Metody strojového učení ... 23

3.4 Odlehlé hodnoty ... 24

3.4.1 Typy odlehlých hodnot ... 24

3.4.2 Typy metod detekce odlehlých hodnot ... 25

3.4.3 Analýza hlavních komponent ... 26

3.4.4 DBSCAN ... 27

3.5 Výběr prediktorů ... 28

3.5.1 Metody filtrování ... 28

3.5.2 Obalové metody ... 30

4 Vlastní práce ... 32

4.1 Popis datové sady ... 32

4.2 Explorační analýza dat ... 32

4.2.1 Jednorozměrné metody ... 33

4.2.2 Dvourozměrné metody ... 34

4.3 Příprava dat ... 36

4.3.1 Čištění dat ... 36

4.3.2 Výběr prediktorů a deklarace cílové proměnné ... 37

4.3.3 Rozdělení dat do trénovací a testovací sady 38
4.4 Trénování modelů SVM 38
4.4.1 Standardní hyperparametry 39
4.4.2 Optimalizace hyperparametrů 39
5 Výsledky a diskuse 41
5.1 Matice záměn 41
5.1.1 Správnost 42
5.1.2 Přesnost 43
5.1.3 Citlivost 43
5.1.4 Určitost 44
6 Závěr 45
7 Seznam použitých zdrojů 46
8 Přílohy 52
8.1 Obrázky 52

1 Úvod

V posledních desetiletích společnosti shromažďují obrovské množství dat z širokého spektra zdrojů, aby mohly provádět lepší a informovanější analýzy. Velké technologické společnosti například využívají složité algoritmy k vytěžování a analýze uživatelských dat za účelem poskytování personalizovaných reklam. Finanční instituce a pojišťovny shromažďují a zkoumají svá data, aby mohly lépe spravovat finanční prostředky a odhalovat podvody. Světová ekonomika 21. století se spoléhá na to, že data shromážděná od spotřebitelů budou vzkvétat.

Kvalita dat však může výrazně ovlivnit přesnost a efektivitu rozhodování. Nezpracovaná data, známá také jako primární data, obsahují mnoho chyb, jako jsou chybějící hodnoty, odlehlé hodnoty, duplikáty, typografické chyby atd. a mohou potenciálně vnést do analýzy nepřesnosti a vést k nesprávným rozhodnutím. V peněžním vyjádření může podle průzkumu "Gartner Quality Market Survey" stát špatná kvalita dat organizace až 15 milionů dolarů ročně (1). Podle společnosti IBM by mohla stát americkou ekonomiku až 3,1 bilionu dolarů ročně. Byla také označena za hlavní příčinu neúspěchu 40 % nových obchodních iniciativ (2).

Klasický přístup k čištění dat je manuální proces, který vyžaduje značné lidské úsilí a je časově náročný. Podle průzkumu "2021 State of Data Science" společnosti Anaconda respondenti uvedli, že tráví *"39 % svého času přípravou dat a jejich čištěním, což je více než čas strávený trénováním modelů, výběrem modelů a nasazením modelů dohromady"* (3). S rozvojem strojového učení se staly možnými automatizované techniky čištění dat, které mohou snížit zátěž lidí a zlepšit přesnost a kvalitu dat.

2 Cíl práce a metodika

2.1 Cíl

Hlavním cílem práce je analyzovat možnosti čistění dat pomocí strojového učení a zhodnotit vhodnost zvoleného řešení na vybrané datové sadě.
Dílčí cíle práce jsou:

- na základě studia odborných informačních zdrojů analyzovat dostupná řešení pro čištění dat.
- analyzovat možnosti využití strojového učení jako nástroje pro čištění dat.
- navrhnout a implementovat konkrétní způsob řešení na testovací sadě dat a zhodnotit vhodnost zvoleného postupu.

2.2 Metodika

Metodika teoretické části práce je založena na studiu a analýze dostupných literárních a online informačních zdrojů v oblasti čištění dat a strojového učení. Metodika praktické části spočívá ve vymezení běžných chyb v datech a způsobech jejich odstranění. Budou analyzována možnosti využití strojového učení pro čištění dat. Bude navržen postup řešení a bude implementován v praxi za použití jazyka Python. Bude provedena validace modelu a změřena jeho úspěšnost při práci se zvolenou datovou sadou a následně určena vhodnost zvoleného postupu. Na základě sjednocení poznatků teoretické a praktické části budou formulovány závěry práce.

3 Teoretická východiska

3.1 Kvalita dat

„Kvalita dat je často definována jako "vhodnost pro použití", tj. hodnocení toho, do jaké míry určitá data vyhovují účelu uživatele.“ (4)

Před použitím datové sady pro jakoukoli aplikaci je nezbytné jí porozumět. Nedodržení tohoto požadavku může vést k chybným analýzám a nedůvěryhodným závěrům. Posuzování kvality dat napříč navrženými metrikami pro řešení problémů s kvalitou snižuje úsilí potřebné pro iterativní ladění potrubí strojového učení s cílem zlepšit výkonnost modelu (5).

Problémy s kvalitou dat se vyskytují v jednotlivých datových souborech a databázích v důsledku chyb v pravopisu při zadávání dat, chybějících informací nebo jiných neplatných údajů. Pokud je třeba integrovat více zdrojů dat, např. v datových skladech nebo globálních webových informačních systémech, potřeba čištění dat se výrazně zvyšuje (6).

3.1.1 Typy problémů

Rozlišuje se mezi problémy s **jedním nebo více zdroji** a mezi problémy souvisejícími se **schématy nebo instancemi**. Problémy na úrovni schématu se projevují i v instancích. Lze je řešit na úrovni schématu s lepším návrhem. Problémy na úrovni instancí se však týkají chyb a nekonzistencí ve skutečném obsahu dat, které nejsou na úrovni schématu viditelné. Na ně se primárně zaměřuje čištění dat. Kvalita dat zdroje závisí na tom, do jaké míry se řídí schématem a integritními omezeními, která kontrolují povolené hodnoty dat. U zdrojů bez schématu, jako jsou například soubory, existuje jen málo omezení pro to, jaká data lze zadávat a ukládat, což vede k vysoké pravděpodobnosti výskytu chyb a nekonzistencí (6).

3.1.2 Dimenze kvality dat

Dimenze kvality dat poskytuje způsob měření a řízení kvality dat a informací. Nejčastěji zmiňovanými dimenzemi kvality dat jsou úplnost, přesnost, konzistence, aktuálnost a jedinečnost, přičemž definice se u výzkumníků liší.

- **Úplnost** vyjadřuje, zda jsou v souboru dat obsaženy všechny požadované údaje a zda splňují cíle daného projektu. Pokud údaje v souboru dat chybí, může to vést k zavádějícím trendům a zešikmit výsledky analýzy.

- **Přesnost** je míra bezchybnosti, spolehlivosti a správnosti údajů. Jinými slovy, do jaké míry údaje přesně odrážejí popisovaný objekt.
- **Aktuálnost** popisuje skutečnost, že je možné mít současná data, která jsou neúčinná, protože jsou pro konkrétní použití opožděná.
- **Konzistence** je vnímána jako integritní omezení a je považována za klíčový problém kvality dat. Stejná data by měla být konzistentní napříč různými úložišti, softwarovými balíčky a formáty souborů.
- **Jedinečnost** popisuje, že prvky nebo objekty by měly být v určité datové sadě zastoupeny pouze jednou. Duplicitní data nejenže zešikmuje výsledky, ale mohou také prodloužit dobu zpracování a zvětšit úložný prostor (7; 8).

3.1.3 Strategie zlepšování kvality dat

Procesně řízená strategie je strategie, která přepracovává proces tvorby nebo úpravy dat s cílem zlepšit jejich kvalitu. Odstraňuje příčiny problémů s kvalitou a dosahuje lepších výsledků v čase. **Datově řízená** je strategie zlepšování kvality dat přímou úpravou jejich hodnoty. Je nákladnější než procesně řízená a krátkodobě účinnější (9; 7).

Neomezený seznam technik zlepšování, které se používají v rámci datově řízené strategie, je následující:

- **Získávání dat** – vylepšuje data tím, že shromažďuje kvalitnější data, která nahrazují hodnoty způsobující problémy.
- **Standardizace (nebo normalizace)** – nahrazuje nebo doplňuje nestandardní hodnoty dat odpovídajícími hodnotami, které jsou v souladu se standardem (přezdívky jsou nahrazeny odpovídajícími názvy).
- **Propojení záznamů** – identifikuje reprezentace dat ve dvou (nebo více) tabulkách, které mohou odkazovat na stejný objekt reálného světa.
- **Integrace dat a schémat** – definuje jednotný pohled na data poskytovaná heterogenními zdroji dat. Hlavním účelem integrace je umožnit uživateli přístup k datům uloženým v heterogenních zdrojích dat prostřednictvím jednotného pohledu na tato data.
- **Důvěryhodnost zdroje** – vybírá zdroje dat na základě kvality.
- **Lokalizace a oprava chyb** – identifikují a odstraňují chyby v kvalitě tím, že odhalují záznamy, které nesplňují daný soubor norem kvality (10).

Procesně řízené strategie se vyznačují dvěma hlavními technikami:

- **Přepracování procesů** – odstraňuje příčiny nízké kvality dat a přidává nový proces pro generování vysoce kvalitních dat.
- **Kontrola procesu** – kontroluje a řídí proces získávaní dat (9; 7).

3.1.4 Metody hodnoceni kvality dat

Na základě konkrétní oblasti existují tři různé přístupy k hodnocení dimenzí kvality dat:

- **Empirický přístup** je obecné označení pro jakoukoli výzkumnou metodu, která vyvozuje závěry z pozorovatelných důkazů. Používá se k určení dimenzí pro hodnocení vhodnosti dat pro použití ke konkrétnímu účelu. Příkladem jsou metody, jako je akční výzkum, případová studie, statistická analýza a ekonometrie (11).
- **Teoretický přístup** určuje dimenze na základě zavedené teorie a může poskytnout úplný soubor dimenzí kvality dat. Například Wand a Wang (12) tvrdili, že data by měla být ve vyčerpávajícím mapování s reálným světem. Tj. stav reálného světa lze mapovat na více než jeden stav v informačním systému, ale stav v informačním systému nemůže reprezentovat dva nebo více stavů v reálném světě. Proto jsou údaje považovány za neúplné, pokud žádný stav v informačním systému neodpovídá stavu v reálném světě. Podobně jsou údaje považovány za nejednoznačné, pokud stav v informačním systému odpovídá dvěma nebo více stavům reálného světa. Údaje jsou nesmyslné, pokud stav v informačním systému neodpovídá žádnému stavu reálného světa (13).
- **Intuitivní přístup** je zvolen na základě osobních zkušeností a znalostí výzkumníka o tom, jaké vlastnosti jsou v konkrétním rámci důležité. Smysl použití této metody spočívá v tom, že umožňuje jednotlivci vybrat dimenze důležité pro konkrétní cíle daného rámce (7).

3.2 Průzkum dat

Průzkum dat neboli **explorační analýza dat** (exploratory data analysis, EDA) je proces popisu dat pomocí statistických a vizualizačních technik s cílem zdůraznit důležité aspekty těchto dat pro další analýzu. Jedná se o prozkoumání datového souboru z mnoha hledisek, jeho popis a shrnutí, aniž by byly učiněny jakékoli předpoklady o jeho obsahu. Je to významný krok, který je třeba učinit před zahájením jakéhokoli statistického modelování

nebo strojového učení, aby bylo zajištěno, že data jsou skutečně vhodná a že v nich nejsou zjevné chyby. Mělo by to být součástí projektů datové vědy v každé organizaci (14).

Přestože někdy mají výzkumníci tendenci věnovat více času návrhu architektury modelu a ladění parametrů, neměli bychom ignorovat význam průzkumu dat. Například společnost Amazon vytvořila v roce 2018 nástroj pro nábor zaměstnanců s umělou inteligencí, který sloužil k prověřování životopisů. Společnost použila k trénování modelu životopisy uchazečů za posledních 10 let. Doporučení modelu se však silně přikláněla k mužům, a dokonce penalizovala životopisy, které obsahovaly slova související s ženami. Důvodem takového zkreslení je nevyvážený počet mužských a ženských uchazečů za posledních 10 let. Ve skutečnosti, pokud by byl správně proveden krok průzkumu dat, bylo by možné nevyváženost tříd řešit a před nasazením modelu provést vhodné techniky (15).

3.2.1 Úrovně organizace dat

Pro analýzu a další využití dat ve strojovém učení je důležité pochopit, že existují tři běžné stupně organizace dat: strukturovaná, semi-strukturovaná a nestrukturovaná. Jejich sběr a škálování probíhají různými způsoby a každý z nich se nachází v jiném typu databáze (16).

- **Nestrukturovaná data** zahrnují různé druhy obsahu, jako jsou dokumenty, videa, audio soubory, příspěvky na sociálních sítích a e-maily. Tyto typy dat je těžké standardizovat a kategorizovat. Nestrukturovaná data se často skládají spíše z kolekcí dat než z jasného datového prvku – například z dokumentu s tisíci slovy, které se týkají několika témat. V současné době tvoří 90 % dat nestrukturovaná data (16; 17).
- **Strukturovaná data** jsou obvykle uložena v tabulkovém formátu a nacházejí se v relačních databázích. Snadno se spravují a lze v nich dobře vyhledávat, a to jak prostřednictvím dotazů generovaných člověkem, tak i automatizovanou analýzou pomocí tradičních statistických metod a algoritmů strojového učení. Strukturovaná data se používají téměř ve všech odvětvích. Mezi běžné příklady aplikací, které se spoléhají na strukturovaná data, patří řízení vztahů se zákazníky, fakturační systémy, databáze produktů a seznamy kontaktů (16; 17; 18).
- **Semi-strukturovaná data** (např. JSON, CSV, XML) jsou "mostem" mezi strukturovanými a nestrukturovanými daty. Nemají předem definovaný datový model, nicméně data mají určitou strukturu díky přítomnosti metadat, sémantických prvků a organizačních vlastností, které umožňují jejich analýzu (19; 20).

Vzhledem k tomu, že většina dat existuje ve volném formátu, je třeba použít techniky předběžné analýzy, tzv. **preprocessing**, aby bylo možné alespoň na část dat aplikovat strukturu pro další analýzu. To lze provést použitím nových charakteristik, které popisují data: počet slov nebo frází, existence určitých speciálních znaků, relativní délka textu atd (21).

3.2.2 Úrovně měření dat

Ze statistického hlediska existuje více typů proměnných než jen prosté rozlišení mezi numerickými a kategoriálními daty. Ve skutečnosti existují čtyři tzv. úrovně měření dat, které určují, co proměnná skutečně znamená a jaké matematické operace na ni lze aplikovat. A typ měření dat proměnných může ovlivnit způsob, jakým modely strojového učení s těmito daty zacházejí a jak se z nich učí (22).

První úroveň měření je **nominální úroveň měření**. V této úrovni měření se čísla v proměnné používají pouze ke klasifikaci údajů (23). Jedinou informací, kterou nominální data nesou, je skupina, do které pozorování patří. Příkladem může být barva. Žlutá barva by mohla být zakódována jako 1 a modrá jako 2, ale tato čísla by neměla žádnou konkrétní hodnotu ani význam. A takové kódování by automaticky neznamenalo, že zelená je rovna 1,5. Pokud jde o matematické operace, s tím se toho moc dělat nedá. Je možné pouze vypočítat modus nominální proměnné. Jiné míry, jako je průměr nebo medián, nemají smysl (22).

Druhou úrovní měření je **ordinální úroveň měření**. Tato úroveň měření zobrazuje určitý uspořádaný vztah mezi pozorováními proměnné (23), neposkytuje však relativní rozdíly mezi pozorováními, což znamená, že je není možné sčítat nebo odečítat a získat tak skutečný význam (21). Jedním z příkladů je úroveň vzdělání: základní, střední, bakalářské, magisterské atd. Pořadí nám umožňuje vypočítat medián. Výpočet průměru však u ordinálních dat nemá smysl (22).

Třetí úroveň měření je **intervalová úroveň měření**. Intervalová úroveň měření nejen klasifikuje a uspořádává měření, ale také určuje, že vzdálenosti mezi jednotlivými intervaly na stupnici jsou ekvivalentní podél stupnice od nízkého intervalu po vysoký interval (23). Dobrým příkladem je teplota měřená ve stupních Celsia: rozdíl mezi 1 stupněm a 5 stupni je stejný jako mezi 20 a 24 stupni. Údaje na této úrovni nemají přirozený počáteční bod ani

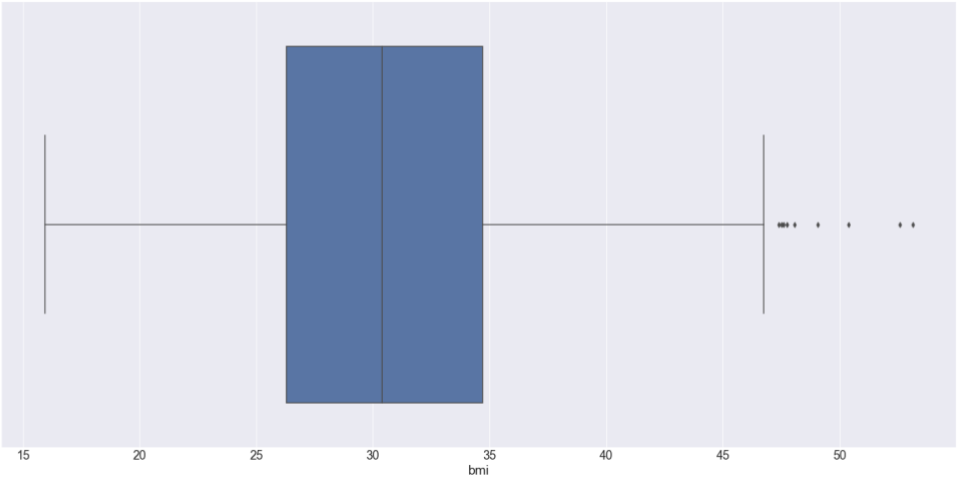
přirozenou nulu, což znamená, že být na nule stupňů Celsia neznamená, že neexistuje žádná teplota (21). V případě proměnných intervalového typu má kromě modusu a mediánu smysl také výpočet aritmetického průměru. Na intervalová data je také možné aplikovat lineární transformace (22).

Čtvrtou úrovní měření je **úroveň poměrového měření**. V této úrovni měření mohou mít pozorování kromě stejných intervalů také nulovou hodnotu, která umožňuje vypočítat poměry mezi dvěma datovými body (22). Příkladem je cena, délka, hmotnost, nějaké množství nebo teplota měřená v Kelvinech. Protože u tohoto typu dat je možné brát poměry, je také možné použít škálovací transformace, například násobení (23).

3.2.3 Vizualizační techniky

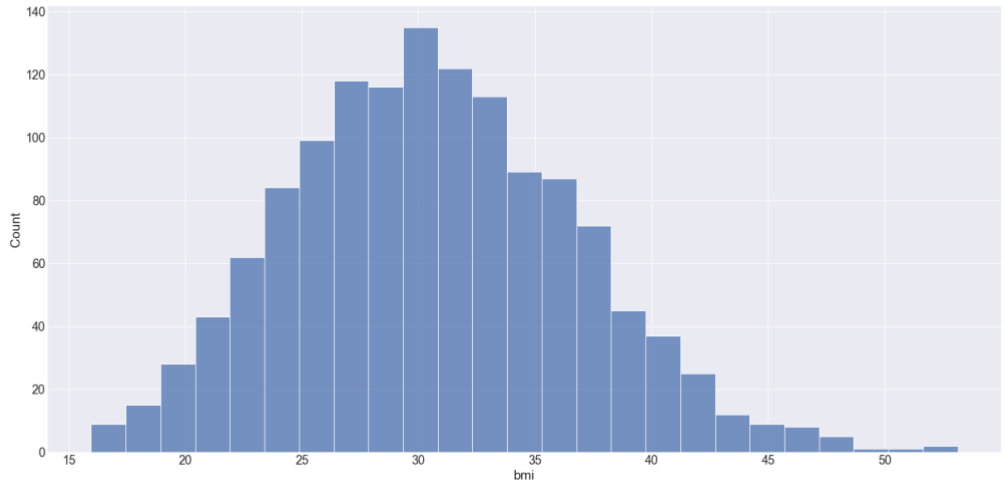
Jako příklad pro vizualizaci byl vybrán datový soubor "Medical cost", který byl představen v knize "Machine Learning with R" od Bretta Lantze (24). Jakékoli charakteristiky dat však nebudou uvedeny, protože cílem následujících vizualizací je pouze demonstrace často používaných metod.

Krabicové grafy (viz obrázek 1), které zavedl Tukey (25), jsou založeny na percentilech a umožňují rychlou vizualizaci rozložení dat (26).



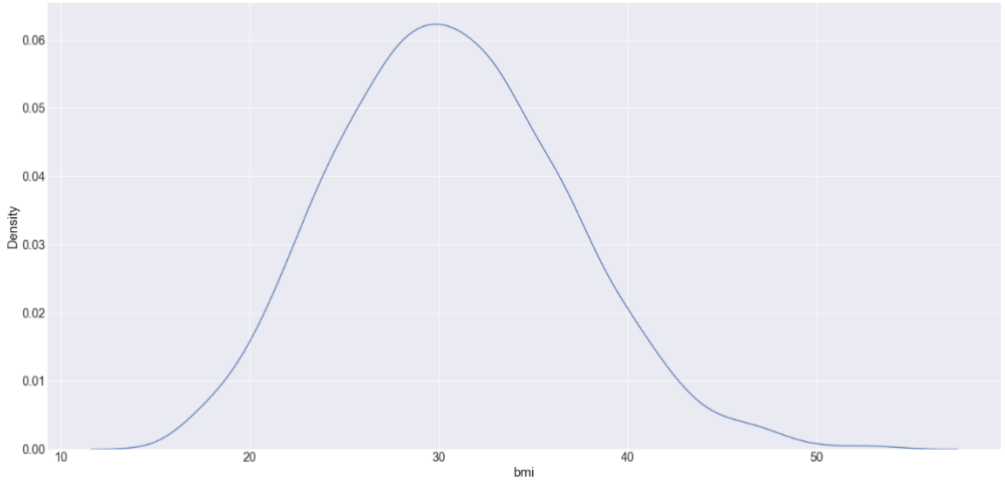
Obrázek 1 - Ukázka krabicového grafu (vlastní zpracování)

Histogram (viz obrázek 2) je způsob vizualizace tabulky četností, kde na ose x jsou obdélníkové pruhy a na ose y jsou počty hodnot (26).



Obrázek 2 - Ukázka histogramu (vlastní zpracování)

S histogramem souvisí **graf hustoty** (viz obrázek 3), který zobrazuje rozložení hodnot dat jako spojitou čáru. Graf hustoty lze považovat za vyhlazený histogram, i když se obvykle počítá přímo z dat pomocí jádrového odhadu hustoty (26).

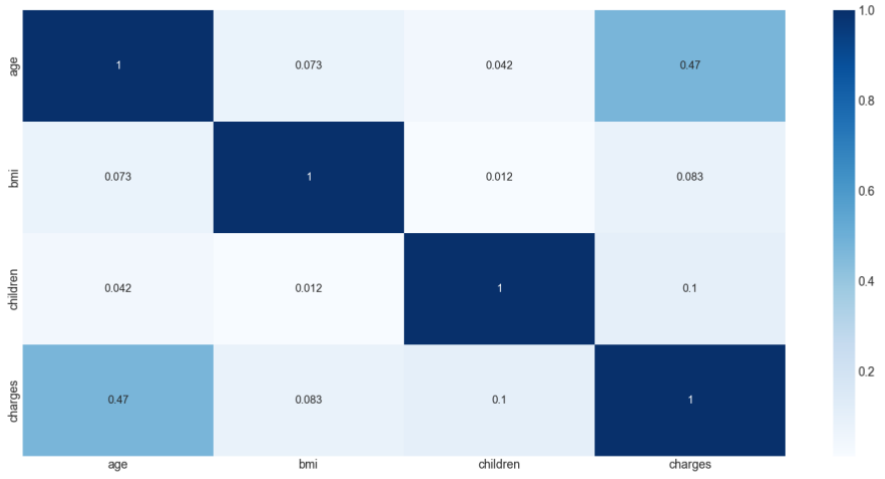


Obrázek 3 - Ukázka grafu hustoty (vlastní zpracování)

Standardním způsobem vizualizace vztahu mezi dvěma měřenými datovými proměnnými je **korelační diagram** (viz obrázek 4). Osa x představuje jednu proměnnou a osa y druhou, a každý bod na grafu je záznam (26).

Obrázek 4 - Ukázka korelačního diagramu (vlastní zpracování)

Korelační matice je statistická technika používaná k vyhodnocení vztahu mezi dvěma proměnnými v souboru dat. Matice je tabulka, v níž každá buňka obsahuje korelační koeficient. Nejčastěji se používá při sestavování regresních modelů. Korelační matice se vizualizuje pomocí **teplotní mapy** (viz obrázek 5) (27).



Obrázek 5 - Ukázka teplotní mapy (vlastní zpracování)

3.3 Chybějící data

Strojové učení se stalo základním pilířem při analýze a získávání informací z dat a často se vyskytuje problém chybějících hodnot. Chybějící hodnoty vznikají v důsledku různých faktorů, jako je selhání systému při sběru dat nebo lidská chyba při předběžném zpracování dat. Přesto je důležité se s chybějícími hodnotami před analýzou dat vypořádat, protože ignorování nebo vynechání chybějících hodnot může vést ke zkreslené nebo nesprávné analýze. (28) Problém chybějících hodnot je obvykle běžný ve všech oblastech, které pracují s daty, a způsobuje různé problémy, jako je zhoršení výkonnosti, problémy s analýzou dat a zkreslené výsledky způsobené rozdíly v chybějících a úplných hodnotách (29).

3.3.1 Vzory chybějících dat

Pro lepší pochopení problému chybějících dat a pro výběr vhodné strategie imputace vytvořili statistici dvoudílnou taxonomii nejčastějších problémů s chybějícími daty. První dimenzí je vzor chybějících dat, který definuje konkrétní rozložení chybějících pozorování napříč datovými případy a proměnnými, které tvoří datový soubor. V literatuře však neexistuje standardní seznam vzorů chybějících dat (30).

- **Jednorozměrné** – chybí pouze jedna proměnná. Tento vzorec je ve většině oborů vzácný a vzniká v experimentálních studiích (31).
- **Monotónní** – proměnné v datech mohou být uspořádány, tento vzor je obvykle spojen s dlouhotrvajícími studiemi, kde členové vypadnou a už se nevrátí. S monotónním vzorem dat se pracuje snadněji, protože vzorce mezi chybějícími hodnotami jsou snadno pozorovatelné (32; 33).
- **Nemonotónní** – v tomto případě se jedná o vzor chybějících údajů, kdy chybějící hodnota jedné proměnné neovlivňuje chybějící hodnoty jiných proměnných (34).

3.3.2 Mechanismy chybění

Mechanismy, které vedou k chybějícím hodnotám v datech, většinou ovlivňují některé předpoklady podporující většinu metod zpracování chybějících dat, a proto se v literatuře chybějící data definují podle těchto mechanismů:

- **Zcela náhodně chybějící data** (Missing completely at random, MCAR) - jedná se o případ, kdy chybějící data nejsou závislá na pozorovaných a nepozorovaných měřeních (28).

- **Náhodně chybějící data** (Missing at random, MAR) - pravděpodobnost chybějící hodnoty v MAR souvisí pouze s pozorovatelnými daty. S náhodně chybějícími hodnotami (MAR) se nejčastěji setkáváme v souborech dat zdravotnických studií. V rámci tohoto mechanismu lze chybějící hodnoty zpracovat pomocí pozorovaných prediktorů (35).
- **Ne náhodně chybějící data** (Missing not at random, MNAR) - jedná se o případy, kdy chybějící údaje nejsou ani MCAR, ani MAR. Chybějící údaje závisí rovnoměrně na chybějících i pozorovaných hodnotách. Při této metodě je zpracování chybějících hodnot obvykle nemožné, protože závisí na nepozorovaných datech. Tento mechanismus se většinou používá v různých oblastech převážně v oblasti (bio)medicíny, ale uplatňuje se také v souborech psychologických a vzdělávacích dat (28).

Podle Grahama (36) je většinou nemožné jednoznačně rozdělit chybějící data do těchto tří mechanismů, protože představa, že chybějící data zcela nesouvisí s jinými nechybějícími proměnnými, je velmi náročná, protože tak či onak chybějící hodnoty souvisejí s nechybějícími proměnnými. Mnoho výzkumníků však uvádí, že nejjednodušší je doplnit všechny chybějící údaje do určité míry jako MAR, protože MAR se nachází uprostřed tohoto kontinua.

3.3.3 Metody odstraňování

Při tomto přístupu se odstraní všechny záznamy s chybějícími hodnotami. Odstranění je považováno za nejjednodušší přístup, protože není třeba se pokoušet odhadnout hodnotu. Odstraňování má však slabinu, protože vnáší do analýzy zkreslení, zejména pokud chybějící údaje nejsou rozděleny náhodně. Proces odstraňování lze provést dvěma způsoby, párovým nebo seznamovým odstraňováním (37).

- Při **seznamovém odstraňování** se odstraní každý záznam, který obsahuje jednu nebo více chybějících hodnot. Za předpokladu, že data nejsou MCAR, však seznamové odstraňování vede ke zkreslení. Zatímco pokud jsou vzorky dat dostatečně velké a předpoklad MCAR je splněn, pak může být odstraňování podle seznamu rozumným přístupem. Pokud nejsou vzorky dat dostatečně velké nebo není splněn předpoklad MCAR, pak odstraňování podle seznamu není nejvhodnějším

přístupem. Odstraňování podle seznamu může také vést ke ztrátě některých důležitých informací, zejména pokud je počet vyřazených případů vysoký (38).

- Pro zmírnění ztráty informací při odstraňování podle seznamu lze použít **párové odstraňování**. Párové mazání se totiž provádí tak, že snižuje ztráty, ke kterým by mohlo dojít při odstraňování podle seznamu. To se provádí tak, že se hodnoty vylučují pouze tehdy, když existuje určitý datový bod potřebný k ověření. Je také známo, že u dat MCAR nebo MAR dává nízké výsledky zkreslení (39).

3.3.4 Často používané metody imputace

Jednoduchá imputace zahrnuje nahrazení chybějících hodnot pomocí kvantitativních nebo kvalitativních charakteristik všech nechybějících hodnot. Při jednoduché imputaci se chybějící údaje zpracovávají různými metodami, jako je modus, průměr nebo medián dostupných hodnot. Ve většině studií se používají metody jednoduché imputace, protože jsou snadné a lze je použít jako nenáročnou referenční techniku. Jednoduché metody imputace však mohou u souborů dat s vysokou dimenzí vést ke zkreslení nebo nereálným výsledkům. Také se zdá, že s nově vznikající generací velkých dat má tato metoda špatné výsledky, a proto je nevhodná pro implementaci na takové datové soubory (40; 28).

Regresní imputace je jednou z preferovaných statistických metod pro zpracování chybějících hodnot. Tato metoda se také označuje jako imputace podmíněných průměrů, kdy jsou chybějící hodnoty nahrazeny předpovídanou hodnotou vytvořenou na základě regresního modelu, pokud data jsou MAR. Celkový proces regrese je dvoufázová metoda: v prvním kroku se použijí všechna úplná pozorování k sestavení regresního modelu, ve druhém kroku se na základě sestaveného regresního modelu imputují chybějící údaje (41).

Imputace Hot-Deck zvládá chybějící hodnoty tak, že chybějící hodnoty porovná s jinými hodnotami v souboru dat u několika dalších klíčových proměnných, které mají úplné hodnoty. Metoda má různé variace, ale ta, která umožňuje přirozenou variabilitu chybějících údajů, vybírá soubor všech případů. Tento soubor se nazývá dárcovský soubor, který je totožný s případy s chybějícími údaji u mnoha proměnných, a z tohoto souboru se náhodně vybere jeden případ. Chybějící hodnota je pak nahrazena údaji z náhodně vybraných případů (42).

Vícenásobná imputace doplňuje chybějící hodnoty generováním čísel odvozených z rozdělení a vztahů mezi sledovanými proměnnými v souboru dat. Vícenásobná imputace se liší od metod jednoduché imputace, protože chybějící údaje se doplňují mnohokrát a pro každou chybějící hodnotu se odhaduje mnoho různých hodnot. Použití více důvěryhodných hodnot poskytuje kvantifikaci nejistoty při odhadu toho, jaké by mohly být chybějící hodnoty, čímž se zamezí vytváření falešných přesností (43).

3.3.5 Metody strojového učení

Algoritmus k-nejbližších sousedů (k-nearest neighbours, KNN) funguje tak, že klasifikuje nejbližší sousedy chybějících hodnot a používá tyto sousedy pro imputaci pomocí míry vzdálenosti mezi instancemi. Pro imputaci KNN lze použít několik měr vzdálenosti, například Minkowského vzdálenost, Manhattanovu vzdálenost, kosinovou vzdálenost, Jaccardovu vzdálenost, Hammingovu vzdálenost a euklidovskou vzdálenost, avšak uvádí se, že euklidovská vzdálenost poskytuje efektivitu a produktivitu, a proto je nejpoužívanější mírou vzdálenosti. Technika imputace KNN je flexibilní jak pro diskrétní, tak pro spojitá data. Imputace KNN má však nevýhody, jako je nízká přesnost při imputování proměnných a zavádí falešné asociace tam, kde neexistují. Další slabinou imputace KNN je, že prohledává celý soubor dat, a tím zvyšuje výpočetní čas (44; 45; 46).

Rozhodovací strom je algoritmus strojového učení, který znázorňuje všechny možné výsledky a cesty vedoucí k daným výsledkům ve formě stromové struktury. Imputace chybějících hodnot pomocí této metody se provádí tak, že se sestaví rozhodovací stromy pro pozorování chybějících hodnot každé proměnné a poté se doplní chybějící hodnoty každé chybějící proměnné pomocí odpovídajícího stromu. Predikce chybějících hodnot se pak zobrazí v koncovém uzlu. Tento algoritmus navíc dokáže pracovat s číselnými i kategorickými proměnnými, identifikovat většinu proměnných a zbytek eliminovat. Rozhodovací stromy však mohou vytvářet složitý strom, který bývá časově náročný, ale má nízké zkreslení (47).

Ansámblové metody jsou strategie, které vytvářejí více modelů a poté je kombinují, aby vytvořily jeden lepší výsledek. Tato metoda obvykle poskytuje přesnější výsledky, než by poskytl jediný model. Tak tomu bylo i v soutěžích strojového učení, kde vítězné modely používaly techniky ensemble. Studie potvrdily, že ansámblové algoritmy zpracování

chybějících dat překonávají algoritmy strojového učení s jedním základním modelem. Existuje několik ansámblových strategií, které se používají, a patří mezi ně mimo jiné Bagging, Boosting a Stacking (48).

3.4 Odlehlé hodnoty

Odlehlé hodnoty jsou extrémní nebo netypické hodnoty, které mohou snížit a zkreslit informace v souboru dat. Problém, jak se vypořádat s odlehlými hodnotami, je již dlouho předmětem zájmu. Některé odlehlé hodnoty ve statistických průzkumech obsahují nějakou chybu, která vyžaduje opravu. Jiné nemusí zahrnovat chybu, ale představují tendenci odlišnou od většiny, přičemž mají v souboru dat velkou návrhovou váhu. Většina běžných metod detekce odlehlých hodnot v oblasti oficiální statistiky jsou jednorozměrné metody uplatňované především při hledání chybných pozorování, aby bylo možné je opravit a vytvořit zcela validní soubory dat. Typickým příkladem je kontrola rozsahu pro určení horní a dolní hranice pro "normální" (tj. nikoliv odlehlá) data, stejně jako metoda kvartilů. Tyto jednorozměrné metody však nemohou odhalit vícerozměrné odlehlé hodnoty, tj. odlehlé hodnoty zahrnující různé vztahy mezi proměnnými. Ve vícerozměrných případech se oproti vícerozměrným metodám často používají korelační diagramy nebo jiné vizualizační techniky, a to z důvodu jejich výpočetní náročnosti a doby zpracování nebo obtížnosti spojené s kontrolou zjištěných vícerozměrných odlehlých hodnot (49).

3.4.1 Typy odlehlých hodnot

Ve statistice a datové vědě existují tři obecně uznávané kategorie, do kterých spadají všechny odlehlé hodnoty:

- **Globální odlehlé hodnoty** (bodové anomálie) – datový bod je považován za globální odlehlou hodnotu, pokud se jeho hodnota výrazně odchyluje od celého souboru dat, ve kterém se nachází.
- **Kontextové** (podmíněné) **odlehlé hodnoty** – datový bod je považován za kontextovou odlehlou hodnotu, pokud se jeho hodnota výrazně odchyluje od ostatních datových bodů ve stejném kontextu. To znamená, že stejná hodnota nemusí být považována za odlehlou, pokud se vyskytla v jiném kontextu. Například u dat časových řad je "kontext" téměř vždy dočasný, protože data časových řad jsou záznamy určité veličiny v čase.

- **Kolektivní odlehlé hodnoty** – hodnoty jednotlivých datových bodů v rámci souboru dat jsou považovány za anomální, pokud se tyto hodnoty jako soubor významně odchylují od celého souboru dat, ale hodnoty jednotlivých datových bodů samy o sobě nejsou anomální ani v kontextovém, ani v globálním smyslu (50).

Odlehlé hodnoty mohou být také **jednorozměrné** nebo **vícerozměrné**. Vícerozměrné odlehlé hodnoty jsou pozorování, která neodpovídají korelační struktuře souboru dat. Zatímco tedy jednorozměrná detekce odlehlých hodnot se provádí nezávisle na každé proměnné, vícerozměrné metody zkoumají vztah několika proměnných. Vícerozměrná odlehlá hodnota je pozorování s charakteristikami odlišnými od vícerozměrného rozdělení většiny pozorování. Detekce vícerozměrných odlehlých hodnot je mnohem obtížnější úkol než detekce jednorozměrných odlehlých hodnot, protože existuje několik směrů, ve kterých může být bod odlehlý (51).

3.4.2 Typy metod detekce odlehlých hodnot

Techniky detekce odlehlých hodnot **založené na statistice** předpokládají, že normální datové body se objevují v oblastech stochastického modelu s vysokou pravděpodobností, zatímco odlehlé hodnoty se vyskytují v oblastech stochastického modelu s nízkou pravděpodobností. Existují dvě běžně používané kategorie přístupů pro detekci odlehlých hodnot na základě statistiky. První kategorie je založena na metodách **testování hypotéz**, jako je Grubbsův test a Tietjen-Mooreův test. Obvykle se při nich na základě pozorovaných datových bodů vypočítá testovací statistika, která slouží k určení, zda je třeba zamítnout nulovou hypotézu (v souboru dat se nevyskytuje žádná odlehlá hodnota). Druhá kategorie technik detekce odlehlých hodnot založených na statistice se zaměřuje na **určení rozdělení nebo odvození funkce hustoty pravděpodobnosti** na základě pozorovaných dat. Datové body, které mají podle hustoty pravděpodobnosti nízkou pravděpodobnost, jsou prohlášeny za odlehlé hodnoty (52).

Techniky detekce odlehlých hodnot založené na **vzdálenosti** často definují vzdálenost mezi datovými body, která se používá pro definování normálního chování. Metody detekce odlehlých hodnot založené na vzdálenosti lze dále rozdělit na globální nebo lokální metody v závislosti na referenčním souboru použitém při určování, zda je bod odlehlou hodnotou. **Globální metoda** detekce odlehlých hodnot založená na vzdálenosti určuje, zda je bod

odlehlou hodnotou, na základě vzdálenosti mezi tímto datovým bodem a všemi ostatními datovými body v souboru dat. Naproti tomu **lokální metoda** při určování odlehlých hodnot zohledňuje vzdálenost mezi bodem a body v jeho okolí (52).

Techniky detekce odlehlých hodnot založené na **modelu** se nejprve naučí klasifikační model ze sady označených datových bodů a poté použijí vycvičený klasifikátor na testovací datový bod, aby určily, zda se jedná o odlehlou hodnotu. Přístupy založené na modelu předpokládají, že klasifikátor lze naučit tak, aby pomocí daného prostoru příznaků rozlišoval mezi normálními a anomálními datovými body. Datové body označují jako odlehlé, pokud je žádný z naučených modelů neklasifikuje jako normální body. Na základě značek, které jsou k dispozici pro trénování klasifikátoru, lze přístupy založené na modelech dále rozdělit do dvou podkategorií: techniky založené na modelech více tříd a techniky založené na modelech jedné třídy. **Techniky založené na modelu více tříd** předpokládají, že trénovací datové body obsahují označené instance patřící do více normálních tříd. Naproti tomu **techniky založené na modelu jedné třídy** předpokládají, že všechny body trénovacích dat patří do jedné normální třídy (52).

3.4.3 Analýza hlavních komponent

Analýza hlavních komponent (Principal Component Analysis, PCA) je velmi oblíbená metoda redukce dimenze. Snaží se vysvětlit kovariantní strukturu dat pomocí malého počtu komponent. Tyto komponenty jsou lineárními kombinacemi původních proměnných a často umožňují interpretaci a lepší pochopení různých zdrojů variability (53). Protože komponenty přímo odhalují variační strukturu dat, mohou mít často větší vypovídací hodnotu než původní funkce. Existuje řada běžných způsobů použití PCA:

- **Redukce dimenze** – pokud jsou znaky vysoce redundantní (konkrétně multikolineární), PCA rozdělí redundanci na jednu nebo více komponent s téměř nulovým rozptylem, které pak mohou být vypuštěny, protože obsahují jen málo informací nebo neobsahují žádné.
- **Detekce anomálií** – neobvyklé odchylky, které nejsou patrné z původních znaků, se často projeví ve složkách s nízkým rozptylem. Tyto složky mohou mít vysokou vypovídací hodnotu v úloze detekce anomálií nebo odlehlých hodnot.

- **Redukce šumu** – soubor údajů ze senzorů má často společný šum na pozadí. PCA někdy dokáže shromáždit (informativní) signál do menšího počtu prvků, zatímco šum nechá, čímž zvýší poměr signálu k šumu.
- **Dekorrelace** – některé algoritmy strojového učení mají problémy s vysoce korelovanými znaky. PCA transformuje korelované znaky na nekorelované komponenty, s nimiž může algoritmus snadněji pracovat (54).

Pokud je k analýze dat zvolena metoda PCA, je součástí procesu také kontrola, zda lze data skutečně analyzovat pomocí PCA. Předpoklady jsou následující:

- **Předpoklad #1** - více proměnných, které by měly být měřeny na spojité úrovni (i když se velmi často používají ordinální proměnné).
- **Předpoklad #2** - mezi všemi proměnnými musí existovat lineární vztah. Důvodem tohoto předpokladu je, že PCA je založena na Pearsonových korelačních koeficientech.
- **Předpoklad #3** - měla by existovat přiměřenost výběru, což jednoduše znamená, že aby PCA poskytla spolehlivý výsledek, je zapotřebí dostatečně velký vzorek. Obecně se jako minimální velikost vzorku doporučuje minimálně 150 případů nebo 5 až 10 případů na proměnnou.
- **Předpoklad #4** - data by měla být vhodná pro redukci. Efektivně by měly existovat přiměřené korelace mezi proměnnými, aby bylo možné proměnné redukovat na menší počet složek.
- **Předpoklad #5** - neměly by se vyskytovat žádné významné odlehlé hodnoty. Odlehlé hodnoty jsou důležité, protože mohou mít neúměrný vliv na výsledky (55).

3.4.4 DBSCAN

Prostorové shlukování aplikací se šumem na základě hustoty (density-based spatial clustering of applications with noise, DBSCAN) je algoritmus, který může také definovat anomálie v datových souborech. Vyžaduje dva uživatelem definované parametry, kterými jsou vzdálenost sousedství **epsilon** a **minimální počet bodů**. Pro daný bod se body ve vzdálenosti epsilon nazývají sousedé tohoto bodu. Pokud je počet sousedních bodů bodu větší než minimální počet bodů, nazývá se tato skupina bodů **shluk**. DBSCAN označuje datové body jako jádrové body, hraniční body a odlehlé (anomální) body. **Jádrové body** jsou ty, které mají alespoň minimální počet bodů ve vzdálenosti epsilon. **Hraniční body** lze

definovat jako body, které nejsou jádrovými body, ale jsou sousedy jádrových bodů. **Odlehlé body** jsou ty, které nejsou ani jádrovými, ani hraničními body (56).

3.5 Výběr prediktorů

Výběr prediktorů (feature selection) se zaměřuje především na odstranění neinformativních nebo nadbytečných prediktorů z modelu. Mnoho modelů, zejména těch založených na regresních sklonech a průsečících, odhaduje parametry pro každý člen modelu. Z tohoto důvodu může přítomnost neinformativních proměnných zvýšit nejistotu predikcí a snížit celkovou účinnost modelu. Některé modely jsou přirozeně odolné vůči neinformativním prediktorům. Například modely založené na stromech a pravidlech, MARS a lasso ze své podstaty provádějí výběr příznaků. Pokud například při konstrukci stromu není prediktor použit v žádném rozdělení, je predikční rovnice funkčně nezávislá na prediktoru (57).

Nejjednodušší formou výběru prvků by bylo odstranění prvků s velmi nízkým rozptylem. Pokud mají rysy velmi nízký rozptyl (tj. velmi blízký 0), jsou blízké konstantě, a tudíž nepřidávají do žádného modelu vůbec žádnou hodnotu (58).

3.5.1 Metody filtrování

Filtrovací metody vyhodnocují prediktory před trénováním modelu a na základě tohoto vyhodnocení je do modelu vložena podmnožina prediktorů. Většina těchto technik je jednorozměrná, což znamená, že hodnotí každý prediktor samostatně. V tomto případě existence korelovaných prediktorů umožňuje vybrat důležité, ale nadbytečné prediktory. Zřejmým důsledkem tohoto problému je, že je vybráno příliš mnoho prediktorů a v důsledku toho vznikají problémy s kolinearitou (57).

Jako základ pro výběr příznaků pomocí metod filtru se běžně používají statistické míry korelace mezi vstupními a výstupními proměnnými. Výběr statistických měr je do značné míry závislý na typech dat proměnných (59). Vzájemný vztah, kovariance nebo asociace mezi dvěma nebo více proměnnými se nazývá **korelace**. Nezabývá se ani změnami veličin X a Y jednotlivě, ale měřením současných změn obou proměnných (26).

Pearsonův korelační koeficient je mírou síly lineárního vztahu mezi dvěma proměnnými a označuje se **r**. Může nabývat hodnot od +1 do -1. Hodnota 0 znamená, že mezi oběma proměnnými neexistuje žádný vztah. Hodnota větší než 0 znamená pozitivní asociaci Hodnota menší než 0 znamená negativní asociaci. Prvním a nejdůležitějším krokem před analýzou dat pomocí Pearsonovy korelace je ověřit, zda je vhodné tento statistický test použít:

- **Předpoklad #1** - dvě proměnné by měly být měřeny na spojité úrovni (tzn. jsou měřeny na úrovni intervalu nebo poměru), obě proměnné však nemusí být měřeny na stejné škále.
- **Předpoklad #2** - dvě proměnné představují párová pozorování.
- **Předpoklad #3** - měla by existovat nezávislost případů.
- **Předpoklad #4** - mezi dvěma proměnnými by měl existovat lineární vztah.
- **Předpoklad #5** - teoreticky by obě proměnné měly odpovídat dvourozměrnému normálnímu rozdělení, ačkoli v praxi se často uznává, že postačí prostá jednorozměrná normalita obou proměnných.
- **Předpoklad #6** - měla by existovat homoskedasticita, což znamená, že rozptyly podél přímky nejlepší shody zůstávají podobné.
- **Předpoklad #7** - neměly by existovat žádné jednorozměrné ani vícerozměrné odlehlé hodnoty (60).

Spearmanův korelační koeficient je neparametrická míra síly a směru asociace, která existuje mezi dvěma řadovými proměnnými měřenými alespoň na ordinální úrovni. Označuje se řeckým písmenem **ρ**. Spearmanovu korelaci lze použít, pokud dvě proměnné nejsou normálně rozděleny. Není také příliš citlivá na odlehlé hodnoty. Test se používá buď pro ordinální proměnné, nebo pro spojitá data, která nesplnila předpoklady nutné pro provedení Pearsonovy korelace. Tento koeficient musí rovněž splnit několik předpokladů:

- **Předpoklad #1** - proměnné by měly být měřeny na ordinální, intervalové nebo poměrové úrovni.
- **Předpoklad #2** - proměnné představují párová pozorování.
- **Předpoklad #3** - mezi oběma proměnnými existuje monotónní vztah. Monotónní vztah existuje, když buď proměnné společně zvyšují svou hodnotu, nebo když se hodnota jedné proměnné zvyšuje, hodnota druhé proměnné se snižuje (61).

Jednofaktorová analýza rozptylu (The one-way analysis of variance, ANOVA) se používá ke zjištění, zda existují statisticky významné rozdíly mezi průměry dvou nebo více nezávislých (nesouvisejících) skupin. Pokud mezi skupinami není žádný významný rozdíl, bude výsledek F-hodnoty ANOVA blízký 1, a proto nebude mít přidanou hodnotu k predikcím. Předpokládá hypotézu jako:

- **H0**: Průměry všech skupin jsou stejné.
- **H1**: Alespoň jeden průměr skupin se liší.

Pro použití ANOVA musí být splněno 6 předpokladů:

- **Předpoklad #1** - závislá proměnná by měla být měřena na úrovni intervalu nebo poměru.
- **Předpoklad #2** - nezávislá proměnná by se měla skládat ze dvou nebo více kategoriálních, nezávislých skupin.
- **Předpoklad #3** - měla by existovat nezávislost pozorování.
- **Předpoklad #4** - neměly by existovat žádné významné odlehlé hodnoty.
- **Předpoklad #5** - závislá proměnná by měla být přibližně normálně rozdělena pro každou kategorii nezávislé proměnné.
- **Předpoklad #6** - musí existovat homogenita rozptylů. Lze ji testovat pomocí Leveneova testu rovnosti rozptylů (62).

Chí-kvadrát test nezávislosti, označovaný také jako Pearsonův chí-kvadrát test nebo chí-kvadrát test asociace, se používá ke zjištění, zda existuje vztah mezi dvěma kategoriálními proměnnými. Předpoklady jsou následující:

- **Předpoklad #1** - proměnné by měly být měřeny na ordinální nebo nominální úrovni.
- **Předpoklad #2** - proměnné by se měly skládat ze dvou nebo více kategoriálních, nezávislých skupin (63).

3.5.2 Obalové metody

Na rozdíl od filtračních metod používají **obalové metody** (wrapper methods) výkonnost zvoleného algoritmu jako metriku, která pomáhá při výběru nejlepší podmnožiny znaků. To je hlavní výhoda obalových metod a bylo prokázáno, že vede k vyšší prediktivní výkonnosti než jaké lze dosáhnout pomocí filtračních metod. Vyčerpávající prohledávání celého prostoru možných kombinací znaků je však výpočetně neproveditelné. Proto je třeba

definovat heuristické strategie vyhledávání v prostoru možných podmnožin znaků (např. náhodné, sekvenční vyhledávání, genetický algoritmus atd.). Na základě vygenerovaných podmnožin znaků se pak natrénuje a vyhodnotí konkrétní algoritmus. Porovnají se klasifikační výkony vygenerovaných podmnožin a podmnožina, která vede k nejlepšímu výkonu, se vybere jako optimální podmnožina. Prakticky lze zkombinovat jakoukoli vyhledávací strategii a klasifikační algoritmus a vytvořit tak obalovou metodu (64).

Obalové metody jsou závislé na použitém algoritmu. Proto není zaručeno, že vybrané znaky zůstanou optimální i při použití jiného algoritmu. V některých případech může použití výkonnosti algoritmu jako ukazatele pro výběr znaků vést k vytvoření podmnožiny znaků s dobrou přesností v rámci trénovacího souboru dat, ale se špatnou generalizací na externí soubory dat (tj. s větší náchylností k přetypování) (65).

4 Vlastní práce

Jedním z kritických aspektů procesu čištění dat je rozpoznání a ošetření odlehlých hodnot – datových bodů, které se výrazně odchylují od většiny pozorování. Jak bylo uvedeno v literární rešerši, odlehlé hodnoty mohou vznikat v důsledku různých faktorů, jako jsou chyby měření, zkreslení dat nebo skutečné anomálie v procesu generování dat. Odhalení a řešení těchto odlehlých hodnot je nezbytné pro zachování kvality analytických modelů postavených na konkrétním souboru dat.

Tato studie se zaměřuje konkrétně na úlohu detekce vícerozměrných odlehlých hodnot pomocí algoritmu Support Vector Machines (SVM). Cílem této práce je postavit a vyhodnotit přístup založený na SVM pro identifikaci odlehlých hodnot v synteticky vytvořeném souboru dat (https://zenodo.org/records/1171077).

4.1 Popis datové sady

Soubor dat použitý v této práci je synteticky vytvořený soubor dat přizpůsobený pro hodnocení algoritmů detekce odlehlých hodnot. Obsahuje 1000 pozorování a 10 proměnných.

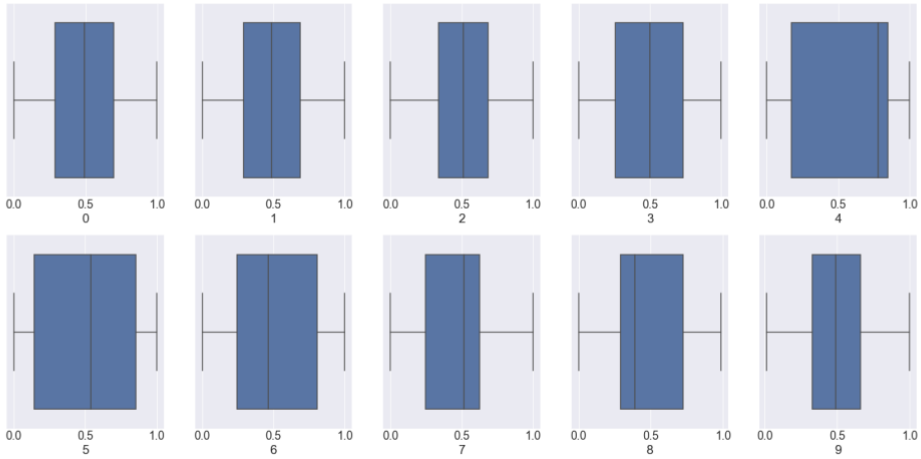
Struktura souboru dat je navržena tak, aby vykazovala různé vzory a rozložení, což umožňuje komplexní hodnocení. Datový soubor navíc obsahuje sloupec "outlier", který označuje, zda je každé pozorování klasifikováno jako odlehlé. Toto označení usnadňuje použití přístupů „učení pod dohledem“ a umožňuje vyhodnotit výkonnost algoritmu.

4.2 Explorační analýza dat

Před zahájením přípravy dat a trénování modelu je nezbytné nejprve pochopit základní strukturu a vlastnosti datové sady. Explorační analýza dat slouží jako nástroj k odhalení poznatků a vzorců skrytých v datech. V této části budou použity grafické techniky, které budou vodítkem pro výběr vhodných metod.

4.2.1 Jednorozměrné metody

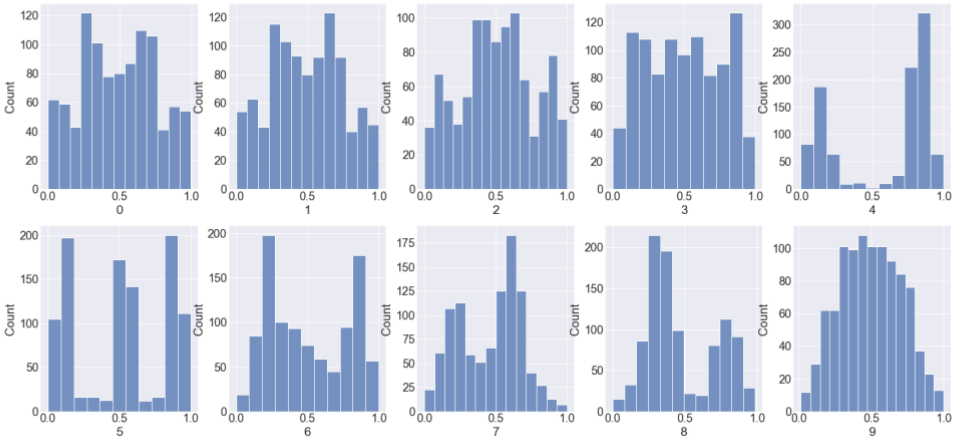
První použitou metodou je **krabicový graf** (viz obrázek 6). Tato metoda poskytuje grafické shrnutí distribuce numerických dat prostřednictvím pěti klíčových statistik: minimum, první kvartil, medián, třetí kvartil a maximum.



Obrázek 6 - Vizualizace proměnných pomocí krabicových grafů (vlastní zpracování)

Každá proměnná se skládá z hodnot mezi 0 a 1. Proměnné 0, 1, 2, 3 a 9 mají medián blízký 0,5 a přibližně symetrické rozdělení. Nejdůležitější informací, kterou krabicový graf poskytuje, je přítomnost odlehlých hodnot. V těchto datech nejsou žádné jednorozměrné odlehlé hodnoty.

Další metodou, která poskytne doplňující pohled na jednotlivá rozdělení, je **histogram** (viz obrázek 7).

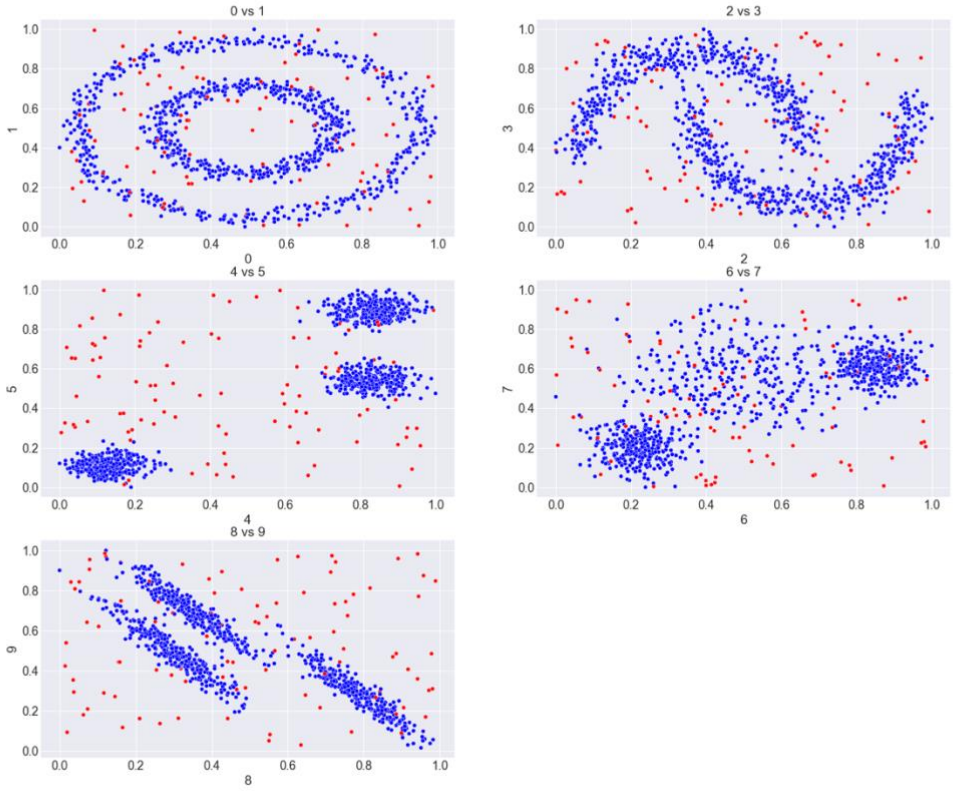


Obrázek 7 – Vizualizace proměnných pomocí histogramů (vlastní zpracování)

Všechny proměnné mají nesymetrické multimodální rozdělení kromě poslední proměnné, která má přibližně normální rozdělení. Tato skutečnost vylučuje použití statistických metod, které předpokládají normalitu dat.

4.2.2 Dvourozměrné metody

Pro pochopení vztahů mezi proměnné byly použity **bodové grafy** s označenými skupinami (modrá – normální hodnoty, červená – odlehlé hodnoty) (viz obrázek 8). Dvojice atributů byly vybrány na základě zdrojového popisu datové sady.



Obrázek 8 - Vizualizace dvojic proměnných pomocí bodových grafů (vlastní zpracování)

Na základě vizualizací jsou v souboru dat shluky tvořené různými kombinacemi proměnných, z nichž každý vykazuje odlišné tvary a vlastnosti:

- **Proměnné 0 a 1**: jsou vytvořeny dva kruhové shluky, které představují symetrické rozložení datových bodů připomínající soustředné kružnice.

- **Proměnné 2 a 3**: objevují se dva shluky ve tvaru půlkruhů, které se vyznačují protáhlým rozložením připomínajícím zakřivení banánu.
- **Proměnné 4 a 5**: jsou patrné tři shluky bodů, které ukazují rozptýlené rozložení datových bodů s různou hustotou.
- **Proměnné 6 a 7**: jsou patrné dva shluky, které mají nerovnoměrná rozložení s různým stupněm rozptylu.
- **Proměnné 8 a 9**: jsou patrné tři shluky s anizotropními tvary, které se vyznačují rozložením protáhlým podél určitých os.

V neposlední řadě je třeba ověřit korelaci mezi proměnnými. Cílem je zajistit, aby mezi nezávislými proměnnými neexistovala multikolinearita. Multikolinearitou se rozumí přítomnost vysokých korelací mezi predikčními proměnnými v regresním modelu, což může představovat problém pro odhad a interpretaci koeficientů. Přestože SVM je především klasifikační algoritmus a neodhaduje koeficienty jako regresní modely, může multikolinearita negativně ovlivnit výkonnost modelu. Pokud jsou predikční proměnné vysoce korelované, může být model příliš citlivý na malé odchylky v trénovacích datech, což vede ke špatnému výkonu na dosud neznámých datech.

Vzhledem k tomu, že data nemají normální rozdělení a mezi proměnnými neexistují silné lineární vztahy, není možné použít Pearsonův korelační koeficient. Proto bude místo toho použit **Spearmanův korelační koeficient**. Vypočtená korelační matice je vizualizována pomocí **teplotní mapy** (viz obrázek 9).

Obrázek 9 - Vizualizace korelační matice daného souboru dat (vlastní zpracování)

- Proměnné 2 a 3, 6 a 9, 8 a 9 - mají zápornou mírnou korelaci.
- Proměnné 4 a 5, 6 a 7, 7 a 8 - mají mírnou pozitivní korelaci.
- Proměnné 6 a 8 - mají vysokou pozitivní korelaci.

Odstranění vysoce korelovaných proměnných může být v určitých případech správným přístupem, avšak v daném problému byly odlehlé hodnoty zjištěny na základě konkrétních dvojic atributů, takže odstranění proměnných pouze na základě jejich korelace nemusí být vhodné, protože by mohlo potenciálně změnit základní strukturu dat a ovlivnit interpretaci odlehlých hodnot.

4.3 Příprava dat

Fáze přípravy dat je klíčovým krokem v procesu strojového učení, kdy se nezpracovaná data transformují a zpracovávají tak, aby byla vhodná pro trénování modelů a predikce.

4.3.1 Čištění dat

Přestože cílem této práce je detekce vícerozměrných odlehlých hodnot, je třeba data vyčistit, aby algoritmus strojového učení správně pracoval. Funguje zde koncept "garbage in, garbage out" (smetí dovnitř, smetí ven), podle kterého chybné vstupy vedou k chybným výstupům.

Jak bylo uvedeno v literární rešerši, čištění dat zahrnuje identifikaci a zpracování chybějících hodnot, odlehlých hodnot a chyb v souboru dat.

Předložený soubor dat neobsahuje žádné chybějící (viz obrázek 10) ani jednorozměrné odlehlé hodnoty (viz obrázek 6).

```
df.isna().sum()
```

```
0          0
1          0
2          0
3          0
4          0
5          0
6          0
7          0
8          0
9          0
outlier    0
dtype: int64
```

Obrázek 10 - Ukázka kódu pro kontrolu chybějících hodnot (vlastní zpracování)

4.3.2 Výběr prediktorů a deklarace cílové proměnné

Pro experimentální účely budou trénovány modely s různými sadami proměnných, aby bylo možné vyhodnotit odolnost modelů vůči změnám ve výběru prediktorů (viz obrázek 11).

```
# feature vector without target variable
X1 = df.drop(['outlier'], axis=1)

# feature vector without target variable and variable "6"
X2 = df.drop(['outlier', '6'], axis=1)

# isolation of target variable
y = df['target_class']
```

Obrázek 11 - Ukázka kódu pro výběr prediktorů a deklaraci cílové proměnné (vlastní zpracování)

- První soubor (**X1**) se skládá ze všech proměnných kromě cílové proměnné ("outlier").
- Druhý soubor (**X2**) se skládá ze všech proměnných kromě cílové proměnné a proměnné "6", která je předmětem vysoké korelace s jinou prediktivní proměnnou.
- **y** představuje izolovanou cílovou proměnnou.

4.3.3 Rozdělení dat do trénovací a testovací sady

Hlavním účelem rozdělení na trénovací a testovací sady je vyhodnotit výkonnost modelu strojového učení na datech, se kterými se model během trénování nesetkal. Trénováním modelu na jedné podmnožině dat a jeho testováním na jiné podmnožině je možné posoudit, jak dobře model zobecňuje na nová, dosud neznámá data.

Pro zajištění reprodukovatelnosti výsledků je parametr **random_state** nastaven na hodnotu 0. Tím je zajištěno, že při každém spuštění kódu bude použito stejné náhodné rozdělení, což usnadní konzistenci procesu vyhodnocování. Navíc, soubor dat je rozdělen na trénovací a testovací množinu s použitím poměru 7:3 (viz obrázek 12).

```
from sklearn.model_selection import train_test_split

# train-test split based on feature selection
X1_train, X1_test, y1_train, y1_test = train_test_split(X1, y, test_size = 0.3, random_state = 0)
X2_train, X2_test, y2_train, y2_test = train_test_split(X2, y, test_size = 0.3, random_state = 0)
```

Obrázek 12 - Ukázka kódu pro rozdělení train-test (vlastní zpracování)

4.4 Trénování modelů SVM

Support Vector Machine (zkráceně SVM) je jedním z výkonných algoritmů strojového učení pro účely klasifikace, regrese a detekce odlehlých hodnot. Klasifikátor SVM vytváří model, který přiřazuje datové body k jedné z daných kategorií. Lze jej tedy považovat za nepravděpodobnostní binární lineární klasifikátor.

Kromě provádění lineární klasifikace může SVM efektivně provádět nelineární klasifikaci pomocí triku s jádrem, který implicitně zobrazuje vstupy do prostorů rysů s vyšší dimenzí, kde jsou data oddělitelná. Jinými slovy převádí nelineární oddělitelné problémy na lineární oddělitelné problémy tím, že k nim přidává další dimenze.

V kontextu SVM existují 4 jádra – **lineární jádro**, **polynomiální jádro**, **jádro RBF** a **sigmoidní jádro**.

4.4.1 Standardní hyperparametry

Nejprve bude natrénován SVM s výchozími hyperparametry, který bude sloužit jako referenční model.

Jak bylo uvedeno v kapitole 4.4.2, model bude vycvičen na různých sadách proměnných, aby bylo možné vybrat nejvýkonnější sadu predikčních rysů.

```
# import SVC classifier
from sklearn.svm import SVC


# import metrics to compute accuracy
from sklearn.metrics import accuracy_score


# classifiers with default hyperparameters
svc_default_1 = SVC()
svc_default_2 = SVC()

# fit classifiers to training sets
svc_default_1.fit(X1_train, y1_train)

svc_default_2.fit(X2_train, y2_train)


# make predictions on test sets
y_pred_1=svc_default_1.predict(X1_test)
y_pred_2=svc_default_2.predict(X2_test)

# compute and print accuracy scores
print('Model accuracy score with default hyperparameters for the first feature vector: {0:0.4f}'. format(accuracy_score(y1_test, y_pred_1)))
print('Model accuracy score with default hyperparameters for the second feature vector: {0:0.4f}'. format(accuracy_score(y2_test, y_pred_2)))
```

```
Model accuracy score with default hyperparameters for the first feature vector: 0.9467
Model accuracy score with default hyperparameters for the second feature vector: 0.9333
```

Obrázek 13 - Ukázka kódu pro trénování a vyhodnocení modelu s výchozími hyperparametry (vlastní zpracování)

I když rozdíl ve skóre přesnosti není velký (1,3 %), ukazuje, že kompletní sada rysů vede k lepším předpovědím (viz obrázek 13). Pro další kroky se tedy bude používat kompletní sada rysů.

4.4.2 Optimalizace hyperparametrů

Optimalizace hyperparametrů, často prováděná pomocí technik, jako je **GridSearch**, je proces přesnějšího nastavení hyperparametrů modelu strojového učení za účelem optimalizace jeho výkonu. GridSearch zahrnuje následující kroky (viz obrázek 14):

- Definování mřížky hodnot hyperparametrů, ve které se má vyhledávat.
- Trénování modelu pomocí každé kombinace hyperparametrů.
- Vyhodnocení výkonu modelu pomocí křížové validace.
- Výběr kombinace hyperparametrů, která vede k nejlepšímu výkonu.

```
# import GridSearchCV
from sklearn.model_selection import GridSearchCV


# import SVC classifier
from sklearn.svm import SVC


# instantiate classifier with default hyperparameters with kernel=rbf, C=1.0 and gamma=auto
svc = SVC()



# declare parameters for hyperparameter tuning
parameters = [ {'C':[1, 10, 100, 1000], 'kernel':['linear']},
               {'C':[1, 10, 100, 1000], 'kernel':['rbf'], 'gamma':[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]},
               {'C':[1, 10, 100, 1000], 'kernel':['poly'], 'degree': [2,3,4] ,'gamma':[0.01,0.02,0.03,0.04,0.05]}
              ]
grid_search = GridSearchCV(estimator = svc,
                           param_grid = parameters,
                           scoring = 'accuracy',
                           cv = 5,
                           verbose=0)


grid_search.fit(X1_train, y1_train)
```

Obrázek 14 - Ukázka kódu pro implementaci GridSearch (vlastní zpracování)

Na základě vyhledávání GridSearch byly nalezeny následující hyperparametry:

- kernel – RBF.
- C – 1000.
- gama – 0,1.

Tato kombinace hyperparametrů dosáhla přesnosti 0,9714, což je o 2,47 % více než referenční model (viz obrázek 15).

```
# best score achieved during the GridSearchCV
print('GridSearch CV best score : {:.4f}\n\n'.format(grid_search.best_score_))


# print parameters that give the best results
print('Parameters that give the best results :','\n\n', (grid_search.best_params_))
```

```
GridSearch CV best score : 0.9714


Parameters that give the best results :

 {'C': 1000, 'gamma': 0.1, 'kernel': 'rbf'}
```

Obrázek 15 - Ukázka kódu pro přehled nejlepších hyperparametrů a jejich skóre (vlastní zpracování)

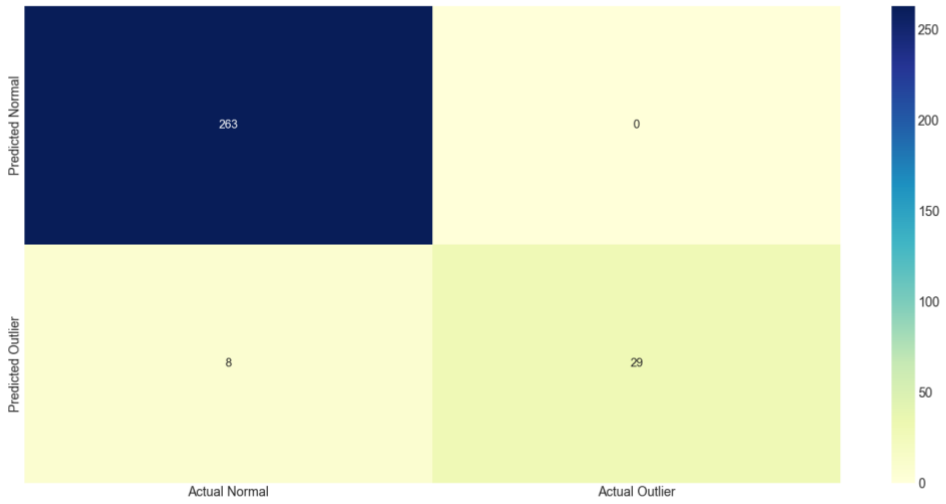
5 Výsledky a diskuse

Na základě výše uvedené analýzy lze konstatovat, že přesnost klasifikačního modelu je přiměřená. Jelikož je však soubor dat nevyvážený (poměr tříd je 10:1), přesnost není vhodným měřítkem pro vyjádření prediktivního výkonu. Proto je třeba prozkoumat alternativní metriky, které poskytnou lepší vodítko při výběru modelů. Předmětem zájmu je zejména základní rozdělení hodnot a typ chyb, kterých se klasifikátor dopouští. Jednou z takových metrik pro analýzu výkonnosti modelu v problému nevyvážených tříd je **matice záměn**.

5.1 Matice záměn

Matice záměny je nástroj pro shrnutí výkonnosti klasifikačního algoritmu. Poskytuje jasný obraz o výkonnosti klasifikačního modelu a typech chyb, které model produkuje. Správné a nesprávné předpovědi jsou rozděleny podle jednotlivých kategorií:

- **Skutečně pozitivní** (True Positive, TP) - počet správně identifikovaných pozitivních případů.
- **Skutečně negativní** (True Negative, TN) - počet správně identifikovaných negativních případů.
- **Falešně pozitivní** (False Positive, FP) - případy, kdy model nesprávně předpovídá pozitivní třídu, zatímco skutečná třída je negativní. Označuje se také jako chyba typu 1.
- **Falešně negativní** (False Negative, FN) - případy, kdy model nesprávně předpovídá negativní třídu, zatímco skutečná třída je pozitivní. Označuje se také jako chyba typu 2.



Obrázek 16 - Vizualizace matice záměn daného souboru dat (vlastní zpracování)

Vytrénovaný model má následující předpovědi (viz obrázek 16):

- správně předpověděl 263 případů jako normální hodnoty.
- správně předpověděl 29 případů jako odlehlé hodnoty.
- neprovedl žádné chybně kladné předpovědi.
- nesprávně předpověděl 8 případů jako odlehlé hodnoty, i když se ve skutečnosti jednalo o normální hodnoty.

Na základě těchto informací lze vypočítat následující metriky, které umožní hlubší pochopení výkonnosti modelu:

- **Accuracy** (správnost)
- **Precision** (přesnost)
- **Sensitivity** (citlivost)
- **Specificity** (určitost)

5.1.1 Správnost

Správnost měří celkovou správnost předpovědí modelu. V tomto kontextu přesnost odráží podíl správně identifikovaných normálních hodnot a správně identifikovaných odlehlých hodnot mezi všemi případy. **Správnost natrénovaného modelu je 97,33 %** (viz obrázek

17). Vysoká správnost naznačuje, že model účinně rozlišuje mezi normálními daty a odlehlými hodnotami.

```
classification_accuracy = (TP + TN) / float(TP + TN + FP + FN)

print('Classification accuracy : {0:0.4f}'.format(classification_accuracy))
```

```
Classification accuracy : 0.9733
```

Obrázek 17 - Ukázka kódu pro výpočet správnosti (vlastní zpracování)

5.1.2 Přesnost

Přesnost měří podíl skutečných normálních hodnot mezi všemi případy předpovězenými modelem jako normální. Jinými slovy, hodnotí schopnost modelu vyhnout se chybné klasifikaci odlehlých hodnot jako normálních dat. **Přesnost natrénovaného modelu je 100 %** (viz obrázek 18). Vysoká přesnost znamená, že model má nízkou míru falešně pozitivních výsledků, což minimalizuje chybnou identifikaci odlehlých hodnot.

```
precision = TP / float(TP + FP)

print('Precision : {0:0.4f}'.format(precision))
```

```
Precision : 1.0000
```

Obrázek 18 - Ukázka kódu pro výpočet přesnosti (vlastní zpracování)

5.1.3 Citlivost

Citlivost měří podíl skutečných normálních hodnot, které byly modelem správně předpovězeny jako normální, mezi všemi skutečnými normálními hodnotami. Odráží schopnost modelu efektivně detekovat normální hodnoty dat při minimalizaci falešně negativních hodnot (odlehlých hodnot chybně klasifikovaných jako normální data). **Citlivost natrénovaného modelu je 97,05 %** (viz obrázek 19).

```
sensitivity = TP / float(TP + FN)

print('Sensitivity : {0:0.4f}'.format(sensitivity))
```

```
Sensitivity : 0.9705
```

Obrázek 19 - Ukázka kódu pro výpočet citlivosti (vlastní zpracování)

5.1.4 Určitost

Určitost měří podíl skutečných odlehlých hodnot, které byly modelem správně předpovězeny jako odlehlé hodnoty, mezi všemi skutečnými odlehlými hodnotami. V tomto kontextu určitost odráží schopnost modelu přesně identifikovat odlehlé hodnoty, aniž by normální hodnoty byly nesprávně klasifikovány jako odlehlé (minimalizuje falešně pozitivní výsledky). **Určitost natrénovaného modelu je 100 %** (viz. obrázek 20).

```
specificity = TN / (TN + FP)

print('Specificity : {0:0.4f}'.format(specificity))
```

```
Specificity : 1.0000
```

Obrázek 20 - Ukázka kódu pro výpočet určitosti (vlastní zpracování)

6 Závěr

Tato práce se zabývala možnostmi čištění dat pomocí technik strojového učení se zaměřením na detekci vícerozměrných odlehlých hodnot. Na základě uvedených cílů tato práce úspěšně dosáhla svého hlavního cíle, kterým byla analýza možností čištění dat pomocí strojového učení a vyhodnocení vhodnosti zvoleného řešení na vybraném souboru dat. Dílčí cíle, mezi něž patřila analýza dostupných řešení pro čištění dat, posouzení potenciálu strojového učení a implementace konkrétního řešení na testovací sadě dat, byly efektivně splněny. Výsledky ukazují, že přístupy strojového učení, zejména detekce odlehlých hodnot založené na SVM, jsou perspektivní pro automatizaci a zvýšení efektivity procesů čištění dat.

Je však důležité si uvědomit zjištěná omezení. Metoda použitá v této studii explicitně neurčuje, které konkrétní atributy přispívají k odlehlosti pozorování. Tento nedostatek informací o konkrétních atributech může představovat problém při pochopení základních příčin odlehlých hodnot. V neposlední řadě je nezbytné si uvědomit, že detekce odlehlých hodnot je ze své podstaty subjektivní a závislá na kontextu. Odlehlé hodnoty mohou představovat nejen potenciální chyby, ale také skutečné a smysluplné datové body, které se odchylují od většiny datového souboru z různých legitimních důvodů, jako jsou vzácné události nebo výjimečné charakteristiky. Výsledky detekce odlehlých hodnot je proto třeba interpretovat opatrně a ověřit je dalšími analýzami a odborným posouzením.

Závěrem lze říci, že ačkoli detekce odlehlých hodnot založená na SVM nabízí perspektivní přístup k čištění dat, je nezbytné si uvědomit a řešit zmíněná omezení. Budoucí výzkumné cesty by se měly zaměřit na vývoj interpretovatelnějších technik detekce odlehlých hodnot a na integraci informací specifických pro jednotlivé atributy s cílem zvýšit účinnost procesů čištění dat.

7 Seznam použitých zdrojů

1. Moore, . How to Create a Business Case for Data Quality Improvement. *Gartner.* [Online] 19 Červen 2018. [Cited: 12 Srpen 2023.] https://www.gartner.com/smarterwithgartner/how-to-create-a-business-case-for-data-quality-improvement#:~:text=Poor%20data%20quality%20destroys%20business,million%20per%20year%20in%20losses..

2. Redman, Thomas C. Bad Data Costs the U.S. $3 Trillion Per Year. *Harvard Business Review.* [Online] 22 Září 2016. [Cited: 16 Srpen 2023.] https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year.

3. Team Anaconda. State of Data Science 2021. *Anaconda.* [Online] 2021. [Cited: 12 Srpen 2023.] https://anaconda.cloud/state-of-data-science-2021.

4. *Data quality assessment from the user's perspective.* Francalanci, Chiara. 2004. stránky 68-73. doi: 10.1145/1012453.1012465.

5. *Overview and Importance of Data Quality for Machine Learning Tasks.* Jain, Abhinav, a další. New York : In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20), 2020. stránky 3561–3562. doi: 10.1145/3394486.3406477.

6. *Data cleaning: Problems and current approaches.* Rahm, Erhard a Hong Hai Do. 4, místo neznámé : IEEE Data Eng. Bull, 2000, Sv. 23, stránky 3-13.

7. *A Short Review of the Literature on Automatic Data Quality.* Chandran, D a Gupta, V. místo neznámé : Journal of Computer and Communications, 2022, Sv. 10, stránky 55-73. doi: 10.4236/jcc.2022.105004.

8. McDonald, Andy. Data quality considerations for Machine Learning Models. *Medium.* [Online] Towards Data Science, 12. Říjen 2022. [Citace: 2023. Červen 20.] https://towardsdatascience.com/data-quality-considerations-for-machine-learning-models-dcbe9cab34cb.

9. *Data quality: A survey of data quality dimensions.* Sidi, F., a další. Kuala Lumpur, Malaysia : International Conference on Information Retrieval & Knowledge Management, 2012. stránky 300-304. doi: 10.1109/InfRKM.2012.6204995.

10. *Methodologies for data quality assessment and improvement.* Batini, Carlo, a další. 3, místo neznámé : ACM Computing Surveys, 2009, Sv. 41. doi: 10.1145/1541880.1541883.

11. *Overview and framework for data and information quality research.* Madnick, Stuart E., a další. 1, místo neznámé : Journal of Data and Information Quality, 2009, Sv. 1, stránky 1-22. doi: 10.1145/1515693.151668.

12. *Anchoring data quality dimensions in ontological foundations.* Wand, Yair a Wang, Richard Y. 11, místo neznámé : Communications of the ACM, 1996, Sv. 39, stránky 86-95.

13. *Evolutional Data Quality: A Theory-Specific View.* Liping, Liu a Lauren, Chi. místo neznámé : International Conference on Information Quality, 2002. stránky 292-304.

14. Exploratory Data Analysis and Visualization Techniques in Data Science. *Analytics Vidhya.* [Online] 16. Květen 2021. [Citace: 19. Červen 2023.] https://www.analyticsvidhya.com/blog/2021/08/exploratory-data-analysis-and-visualization-techniques-in-data-science/.

15. Liu, Mingjie. 2 - Data Exploration. *Machine Learning Blog | ML@CMU | Carnegie Mellon University.* [Online] 24. Srpen 2020. https://blog.ml.cmu.edu/2020/08/31/2-data-exploration/.

16. What is structured & Unstructured Data: Examples & differences: Imperva. *Imperva.* [Online] 26. Říjen 2022. [Citace: 12. Srpen 2022.] https://www.imperva.com/learn/data-security/structured-and-unstructured-data/.

17. *Influence of Structured, Semi- Structured, Unstructured data on various data models.* Praveen, Shagufta a Chandra, Umesh. místo neznámé : International Journal of Scientific and Engineering Research, 2020, Sv. 8, stránky 67-69.

18. Crockett, Emma. Structured vs unstructured data: Key differences explained. *Datamation.* [Online] 9. Únor 2023. [Citace: 17. Červenec 2023.] https://www.datamation.com/big-data/structured-vs-unstructured-data/.

19. IBM. Structured vs. unstructured data: What's the difference? *IBM Blog.* [Online] 29. Červen 2021. [Citace: 29. Srpen 2023.] https://www.ibm.com/cloud/blog/structured-vs-unstructured-data.

20. Dhanashree. Everything you need to know about semi-structured data with semi-structured data examples. *Nanonets Intelligent Automation, and Business Process AI Blog.* [Online] 17. Říjen 2022. [Citace: 28. Srpen 2023.]

21. Ozdemir, Sinan. *Principles of data science.* Birmingham : Packt Publishing, 2016. ISBN: 978-1785887918.

22. Oleszak, Michal. Data Measurement Levels. *Medium.* [Online] Towards Data Science, 21. Září 2021. [Citace: 12. Srpen 2023.] https://towardsdatascience.com/data-measurement-

levels-dfa9a4564176#:~:text=The%20four%20data%20measurement%20levels,ordinal%20%2C%20interval%20%2C%20and%20ratio%20.

23. Data levels of measurement. *Statistics Solutions*. [Online] 3. Srpen 2021. [Citace: 12. Červen 2023.] https://www.statisticssolutions.com/dissertation-resources/descriptive-statistics/data-levels-of-measurement/.

24. Lantz, Brett. *Machine learning with r: Expert techniques for Predictive modeling.* místo neznámé : Packt Publishing, 2019.

25. Tukey, John W. *Exploratory Data Analysis.* místo neznámé : Pearson, 1977. ISBN-13 : 978-0201076165.

26. Bruce, Peter, Bruce, Adrew a Gedeck, Peter. *Practical statistics for data scientists 50+ essential concepts using r and python.* Sebastopol : O'Reilly Media, 2020. ISBN: 978-1492072942.

27. Introduction to the correlation matrix. *Built In.* [Online] [Citace: 13. Červen 2023.] https://builtin.com/data-science/correlation-matrix.

28. *A survey on missing data in machine learning.* Emmanuel, T., Maupong, T. a Mpoeleng, D. místo neznámé : J Big Data 8, 2021, Sv. 140. https://doi.org/10.1186/s40537-021-00516-9.

29. *Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry.* Ayilara, O. F., a další. 106, místo neznámé : Health and quality of life outcomes, 2019, Sv. 17(1). https://doi.org/10.1186/s12955-019-1181-2.

30. Berglund, P. a Heeringa, S. G. *Multiple imputation of missing data using SAS.* místo neznámé : Cary: SAS Institute, 2014.

31. *Flexible imputation of missing data.* Demirtas, H. 1, místo neznámé : Journal of Statistical Software, 2018, Sv. 85, stránky 1-5. https://doi.org/10.18637/jss.v085.b04.

32. *Missing Data Imputation Using the Multivariate t Distribution.* Liu, C. místo neznámé : Journal of Multivariate Analysis, Elsevier, 1995, Sv. 53(1), stránky 139-158.

33. *Principled missing data methods for researchers.* Dong, Y. a Peng, CY. J. místo neznámé : SpringerPlus , 2013, Sv. 2. https://doi.org/10.1186/2193-1801-2-222.

34. *Pattern graphs: a graphical approach to nonmonotone missing data.* Chen, Y-C. 2020. https://doi.org/10.48550/arXiv.2004.00744.

35. *A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets.* Gómez-Carracedo, M, a další. místo neznámé : Chemom Intell Lab Syst, 2014, Sv. 134, stránky 23-33.
36. Graham, J W. *Missing Data: Analysis and Design.* New York : Springer, 2012. stránky 47-69. ISBN-13: 978-1461440178.
37. *Statistical Analysis with Missing Data.* Little, R J a Rubin, D B. místo neznámé : John Wiley & Sons, Hoboken, 2019, Sv. 793. https://doi.org/10.1002/9781119482260.
38. *Missing data Part 1: overview, traditional methods.* Williams, R. místo neznámé : Notre Dame: University of Notre Dame, 2015.
39. *Missing data.* Allison, P D. místo neznámé : Thousand Oaks: Sage Publications, 2001, Sv. 136.
40. *SICE: an improved missing data imputation technique.* Khan SI, Hoque ASML. místo neznámé : J Big Data, 2020, Sv. 7(1), stránky 1-21.
41. *Missing data imputation techniques.* Song Q, Shepperd M. místo neznámé : Int J Bus Intell Data Min, 2007, Sv. 2(3), stránky 261-291.
42. *A review of hot deck imputation for survey non-response.* Andridge RR, Little RJ. místo neznámé : Int Stat Rev, 2010, Sv. 78(1), stránky 40-64.
43. *Multiple Imputation: A Flexible Tool for Handling Missing Data.* Li P, Stuart EA, Allison DB. místo neznámé : JAMA, 2015 , Sv. 314(18). doi: 10.1001/jama.2015.15281.
44. *The treatment of missing values and its effect on classifier accuracy. In: Classification, clustering, and data mining applications.* Acuna E, Rodriguez C. New York : Springer, 2004, stránky 639–647. https://doi.org/10.1007/978-3-642-17103-1_60.
45. *kNN-is: an iterative Spark-based design of the k-nearest neighbors classifier for big data.* Maillo J, Ramírez S, Triguero I, Herrera F. místo neznámé : Knowl Based Syst, 2017, Sv. 117, stránky 3-15. https://doi.org/10.1016/j.knosys.2016.06.012.
46. *Missing value imputation in multi attribute data set.* M., Gimpy. místo neznámé : Int J Comput Sci Inf Technol, 2014, Sv. 5(4), stránky 1–7.
47. *Decision forest: twenty years of research.* L., Rokach. místo neznámé : Inf Fusion, 2016, Sv. 27, stránky 111–125. https://doi.org/10.1016/j.inffus.2015.06.005.
48. Zhang C, Ma Y. *Ensemble machine learning: methods and applications.* Boston : Springer, 2012. DOI:10.1007/9781441993267.
49. *Outliers in official statistics.* Wada, K. místo neznámé : Jpn J Stat Data Sci, 2020, Sv. 3, stránky 669–691.

50. Cohen, Ira. Outliers explained: A quick guide to the different types of outliers. *Medium.* [Online] Towards Data Science, 26. Srpen 2021. [Citace: 1. Květen 2023.] https://towardsdatascience.com/outliers-analysis-a-quick-guide-to-the-different-types-of-outliers-e41de37e6bf6.

51. Outlier detection using machine learning. *Charles Holbert.* [Online] 09. Září 2019. [Citace: 11. Duben 2023.] https://www.cfholbert.com/blog/outlier-detection-machine-learning/.

52. Xu, Chu a Ilyas, Ihab F. *Data Cleaning.* místo neznámé : Association for Computing Machinery, 2019.

53. *Robust statistics for outlier.* Rousseeuw, Peter J. a Hubert, Mia. New York : John Wiley & Sons, Inc. WIREs Data Mining Knowl Discov, 2011, Sv. 1, stránky 73-79.

54. Javigallego. Massive PCA + outlier detection tutorial. *Kaggle.* [Online] 06. Červenec 2022. [Citace: 11. Srpen 2023.] https://www.kaggle.com/code/javigallego/massive-pca-outlier-detection-tutorial.

55. Principal components analysis (PCA) using SPSS statistics. *Laerd Statistics.* [Online] [Citace: 29. Duben 2023.] https://statistics.laerd.com/spss-tutorials/principal-components-analysis-pca-using-spss-statistics.php.

56. *Anomaly detection in temperature data using DBSCAN algorithm.* Çelik, M., Dadaşer-Çelik, F. a Dokuz, A. Ş. Istanbul : 2011 International Symposium on Innovations in Intelligent Systems and Applications, 2011, stránky 91-95.

57. Kuhn, Max a Johnson, Kjell. *Applied predictive modeling.* místo neznámé : Springer, 2019. str. 488. ISBN-13: 978-1461468486.

58. Thakur, Abhishek. *Approaching (Almost) Any Machine Learning Problem.* 2020. ISBN-13: 978-9390274437.

59. Brownlee, Jason. How to choose a feature selection method for machine learning. *Machine Learning Mastery.* [Online] 20. Srpen 2020. [Citace: 15. Srpen 2023.] https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/.

60. Pearson product-moment correlation. [Online] Laerd Statistics. [Citace: 15. Srpen 2023.] https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php.

61. Spearman's rank-order correlation. *Laerd Statistics.* [Online] [Citace: 15. Srpen 2023.] https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php.

62. One-way ANOVA in SPSS statistics. *Laerd Statistics.* [Online] [Citace: 15. Srpen 2023.] https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php.
63. Chi-square test for association using SPSS statistics. *Laerd Statistics.* [Online] [Citace: 16. Srpen 2023.] https://statistics.laerd.com/spss-tutorials/chi-square-test-for-association-using-spss-statistics.php.
64. Pudjihartono, Nicholas, a další. A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers.* [Online] 3. Červen 2022. [Citace: 17. Srpen 2023.] https://www.frontiersin.org/articles/10.3389/fbinf.2022.927312/full.
65. *Wrappers for Feature Subset Selection.* Kohavi, R. a John, G. H. místo neznámé : Artif. Intell., 1997, Sv. 97, stránky 273–324.

8 Přílohy

8.1 Obrázky

Obrázek 1 - Ukázka krabicového grafu (vlastní zpracování) 17
Obrázek 2 - Ukázka histogramu (vlastní zpracování) 18
Obrázek 3 - Ukázka grafu hustoty (vlastní zpracování) 18
Obrázek 4 - Ukázka korelačního diagramu (vlastní zpracování).....19
Obrázek 5 - Ukázka teplotní mapy (vlastní zpracování) 19
Obrázek 6 - Vizualizace proměnných pomocí krabicových grafů (vlastní zpracování).....33
Obrázek 7 – Vizualizace proměnných pomocí histogramů (vlastní zpracování) 33
Obrázek 8 - Vizualizace dvojic proměnných pomocí bodových grafů (vlastní zpracování)....34
Obrázek 9 - Vizualizace korelační matice daného souboru dat (vlastní zpracování) 36
Obrázek 10 - Ukázka kódu pro kontrolu chybějících hodnot (vlastní zpracování) 37
Obrázek 11 - Ukázka kódu pro výběr prediktorů a deklaraci cílové proměnné (vlastní zpracování).....37
Obrázek 12 - Ukázka kódu pro rozdělení train-test (vlastní zpracování) 38
Obrázek 13 - Ukázka kódu pro trénování a vyhodnocení modelu s výchozími hyperparametry (vlastní zpracování).....39
Obrázek 14 - Ukázka kódu pro implementaci GridSearch (vlastní zpracování) 40
Obrázek 15 - Ukázka kódu pro přehled nejlepších hyperparametrů a jejich skóre (vlastní zpracování).....40
Obrázek 16 - Vizualizace matice záměn daného souboru dat (vlastní zpracování) 42
Obrázek 17 - Ukázka kódu pro výpočet správnosti (vlastní zpracování) 43
Obrázek 18 - Ukázka kódu pro výpočet přesnosti (vlastní zpracování) 43
Obrázek 19 - Ukázka kódu pro výpočet citlivosti (vlastní zpracování).....43
Obrázek 20 - Ukázka kódu pro výpočet určitosti (vlastní zpracování) 44