FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Zadání bakalářské práce

Název:	Automatická explorační analýza dat pro binární klasifikaci pomocí knihovny pandas profiling
Student:	Jan Čáp
Vedoucí:	Ing. Daniel Vašata, Ph.D.
Studijní program:	Informatika
Obor / specializace:	Znalostní inženýrství
Katedra:	Katedra aplikované matematiky
Platnost zadání:	do konce letního semestru 2023/2024

Pokyny pro vypracování

Typickým úvodním krokem při seznamování se s novým datasetem je zevrubná explorace. Je nutné získat představu o rozdělení jednotlivých veličin, podílu chybějících hodnot, potenciálních odlehlých či přetížených hodnotách, korelovanosti veličin apod. Zpravidla to znamená provést celou řadu repetitivních úkonů. Řadu z nich lze nicméně poměrně dobře automatizovat. Slibným nástrojem na tomto poli je knihovna Pandas Profiling pro Python. S její pomocí lze velmi snadno vytvořit report shrnující vše výše uvedené.

Často, když je na základě datasetu prováděna binární klasifikace, by bylo velmi užitečné sledovat popsané statistiky zvlášť na každé ze subpopulací. Zjištěné rozdíly by uživateli pomohly identifikovat, na které z veličin by se měl při statistickém modelování zaměřit přednostně. To bohužel knihovna Pandas Profiling neumožňuje.

Cílem práce je navrhnout a implementovat rozšíření zmíněné knihovny, která by zlepšila její použitelnost na analýzu datasetů při binární klasifikaci. Po provedení rešerše vhodných metod student:
1. Pro jednotlivé typy vstupních veličin (numerické, kategoriální, binární) navrhne vhodné popisné statistiky a grafy, srovnávající jejich rozdělení v subpopulacích dle cílové proměnné.
2. Bude se zabývat přítomností chybějících pozorování a jejich případným vztahem k cílové proměnné.

Elektronicky schválil/a Ing. Magda Friedjungová, Ph.D. dne 14. února 2023 v Praze.

3. Na základě jednotlivých veličin se pokusí identifikovat subpopulace s výrazně odlišnou prevalencí cílové proměnné.
4. Při zkoumání veličin navrhne vhodnou transformaci vzhledem k zamýšlenému užití pro statistické modelování.

Získané informace bude student jednak prezentovat uživateli způsobem obvyklým ve zvoleném frameworku. Jednak na jejich základě postaví referenční model, který poskytne uživateli základní představu o predikovatelnosti cílové proměnné a přínosu jednotlivých veličin. Použitelnost a přínosnost svého rozšíření demonstruje na datasetu titanik a případně na dalším vhodném datasetu.

Bakalářská práce

SUPERVISED PANDAS PROFILING

Jan Čáp

Fakulta informačních technologií
Katedra teoretické informatiky
Vedoucí: Ing. Daniel Vašata, Ph.D.
11. května 2023

Odkaz na tuto práci: Čáp Jan. *Supervised Pandas Profiling*. Bakalářská práce. České vysoké učení technické v Praze, Fakulta informačních technologií, 2023.

Obsah

Seznam obrázků

Seznam tabulek

Seznam výpisů kódu

Především děkuji Mgr. Dominiku Matulovi za jeho drahocenný čas, trpělivost a cenné rady k tématu práce. Dále bych rád poděkoval Ing. Danielovi Vašatovi, Ph.D. za vedení práce na Fakultě informačních technologií ČVUT v Praze.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací. Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů. V souladu s ust. § 2373 odst. 2 zákona č. 89/2012 Sb., občanský zákoník, ve znění pozdějších předpisů, tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, jež jsou její součástí či přílohou a veškeré jejich dokumentace (dále souhrnně jen „Dílo“), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli způsobem, který nesnižuje hodnotu Díla a za jakýmkoli účelem (včetně užití k výdělečným účelům). Toto oprávnění je časově, teritoriálně i množstevně neomezené.

V Praze dne 11. května 2023 .....

Abstrakt

Práce se zabývá automatickou explorací dat s binární klasifikací. Je provedena rešerše již existujících řešení pro automatickou exploraci dat. Dále jsou prozkoumány statistické testy a metody vhodné pro testování závislosti dvou proměnných. Jsou zde také prozkoumány vhodné možnosti vizualizací rozložení dat. V další části je navrženo rozšíření do knihovny *Pandas Profiling*, která byla vybrána v rešerši. Rozšíření se specializuje na binární klasifikaci. Rozšíření obsahuje grafy a statistiky reprezentující závislost sloupců na cílové proměnné, vizualizaci závislostí chybějících hodnot na cílové proměnné, navržené transformace sloupců a trénování výchozího modelu pro klasifikaci cílové proměnné.

Na základě návrhu bylo implementováno rozšíření knihovny *Pandas Profiling*, které urychlí exploraci dat s binární klasifikací.

Klíčová slova rozšíření knihovny, knihovna Pandas Profiling, automatická explorace dat, binární klasifikace, datascience komunita, Python

Abstract

This work deals with automatic data exploration with binary classification. A search of already existing solutions for automatic data exploration is performed. Furthermore, statistical tests and methods suitable for testing the dependence of two variables are investigated. Suitable options for data distribution visualizations are also explored. In the next section, an extension to the *Pandas Profiling* library selected in the search is proposed. The extension specializes in binary classification. The extension includes graphs and statistics representing the dependency of columns on the target variable, visualization of the dependency of missing values on the target variable, proposed column transformations and training of the default model for target variable classification.

Based on the design, an extension to the *Pandas Profiling* library was implemented to speed up data exploration with binary classification.

Keywords library extension, Pandas Profiling, automatic exploration, binary classification, datascience community, Python

Seznam zkratek

CLI command-line interface. 26, 27, 36

CSV comma-separated values. 27

DAG directed acyclic graph. 35

EDA Exploratory Data Analysis. 6

HTML Hypertext Markup Language. 1, 4, 6, 9, 26, 37

IDF inverse document frequency. 32

Q-Q plot quantile-quantile plot. 4

TF term frequency. 32

TF-IDF term frequency–inverse document frequency. 32, 37

URL Uniform Resource Locator. 25

YAML Yet Another Markup Language. 27, 36

Úvod

Nutným úvodním krokem při seznamování se s novou datovou sadou je explorace dat. Je nutné získat představu o rozdělení jednotlivých veličin, podílu chybějících hodnot, potenciálních odlehlých či přetížených hodnotách, korelovanosti veličin apod. Zpravidla to znamená provést celou řadu repetitivních úkonů pro každou veličinu zvlášť. Spoustu z těchto úkonů lze nicméně jednoduše automatizovat. Slibným nástrojem na tomto poli je knihovna *Pandas Profiling* [1] pro programovací jazyk *Python* [2]. S její pomocí lze velmi snadno vytvořit report shrnující vše výše uvedené. Pomocí knihovny je možné vytvořit Hypertext Markup Language (HTML) report s popisnými statistikami a vizualizacemi pro jednotlivé sloupce. V mnoha projektech se objevují data s binární klasifikací, které by při exploraci potřebovali jiné statistické metody, než data bez binární klasifikace. *Pandas Profiling* se specializuje na obecnou exploraci dat a vytváření reportů nad binární klasifikací momentálně nepodporuje. Příkladem binární klasifikace jsou například data z bankovního sektoru, kde je třeba odhadnout, zda zákazník splatí, či nesplatí úvěr, pokud mu bude poskytnut. Dalším příkladem binární klasifikace může být detekce podvodných transakcí platební kartou. Nebo například předpověď, zda bude mít letadlo zpoždění. Pro datové sady s binární klasifikací by bylo velmi užitečné sledovat popsané statistiky zvlášť na každé ze subpopulací. Zjištěné rozdíly by pak uživateli pomohly identifikovat, na které z veličin by se měl při statistickém modelování zaměřit přednostně. S motivací automatizovat exploraci dat s binární klasifikací vzniklo téma této bakalářské práce.

Jedním z cílů teoretické části práce je vytvořit rešerši již existujících řešení pro automatickou exploraci dat. Dále je třeba zjistit, která z prozkoumaných řešení umí pracovat s binární klasifikací. Dalším cílem je prozkoumání statistických metod pro vysvětlení a zobrazení závislosti dvou proměnných. Z těchto metod jsou vybrány metody, které danou závislost nejlépe vystihují a následně jsou použity při implementaci.

Cílem praktické části práce je vytvořit rozšíření do veřejné knihovny *Pandas Profiling*, které přidá funkcionalitu pro analýzu datových sad při binární klasifikaci. Rozšíření se skládá ze čtyř částí. V první části jsou pro jednotlivé typy vstupních veličin (numerická, kategoriální, textová) navrženy a implementovány popisné statistiky a grafy srovnávající jejich rozdělení v subpopulacích dle cílové proměnné. Druhá část se zabývá problematikou chybějících hodnot a jejich případný vztah k cílové proměnné. Jsou implementovány vhodné statistické testy pro hledání závislosti mezi chybějícími hodnotami a cílovou proměnou. Třetí část je zaměřena na transformace proměnných. Jsou navrženy transformace pro numerické, kategorické a textové veličiny, které by mohly mít pozitivní vliv na kvalitu dat. Čtvrtá část praktické části je zaměřena na vytvoření prediktivního modelu pro klasifikaci cílové proměnné. Model musí být dostatečně robustní, aby bez problému zpracoval všechny typy proměnných. Model bude ohodnocen pomocí standardních metrik: *accuracy*, *precision*, *recall* a *f1-score*.

Bakalářské práce se skládá ze čtyř kapitol. První kapitola je zaměřena na rešerši existujících řešení pro automatickou exploraci dat. Ve druhé kapitole jsou prozkoumány statistické metody,

které se zabývají porovnáním a testováním dvou subpopulací. Statistické metody jsou prozkoumány pro numerické, kategorické a binární veličiny. Část druhé kapitoly je dále věnována vizualizacím. U jednotlivých vizualizací jsou porovnány jejich výhody a nevýhody. Třetí kapitola je zaměřena na návrh úprav knihovny. Je zde vysvětleno, jaké statistické testy a metody byly použity a co vše je v knihovně změněno. Čtvrtá kapitola je věnována praktické části bakalářské práce, do níž patří samotná implementace rozšíření.

Kapitola 1
Rešerše

Tato kapitola je zaměřena na rešerši existujících řešení pro automatickou exploraci dat. V datascience komunitě jsou nejčastěji používány programovací jazyky Python [2] a R [3]. Z tohoto důvodu bude rešerše omezena pouze na knihovny napsané v jazyce Python. Pro ukázky je zvolena datová sada Titanik. Datovou sadu je možné stáhnout ze zdroje [4].

1.1 Pandas

Pandas [5] je nejrozšířenější knihovnou pro práci s daty v jazyce Python. Umožňuje uložit datovou sadu do objektu `DataFrame()` a dále s ní provádět transformace, filtrování atd. Objekt typu `DataFrame()` má implementovanou metodu `describe()`, která vygeneruje tabulku se základními informacemi o jednotlivých proměnných, jako je počet hodnot, průměrná hodnota, směrodatná odchylka, minimální a maximální hodnota a hodnoty pro jednotlivé kvartily. Metoda pracuje pouze s číselnými veličinami, textové a kategoriální veličiny do tabulky nezahrnuje. Metoda nemá možnost vytvořit tabulku nad binární klasifikací. Jedinou možností je provést `describe()` zvlášť pro obě subpopulace dle zadané binární klasifikace a vzniklé tabulky následně manuálně porovnat. Další nevýhodou metody `describe()` je komplikovaná orientace v tabulce při vyšším počtu proměnných. Příklad použití metody na ukázkové sadě se nachází na obrázku 1.1.

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Obrázek 1.1 Ukázka metody *pandas* describe na datové sadě Titanik.

1.2 D-tale

D-tale [6] je knihovnou pro exploraci dat s jednoduchým uživatelským rozhraním v Pythonu. Pro vytvoření reportu stačí spustit jednořádkový příkaz `dtale.show(df)` s datovou sadou jako parametrem. Tento příkaz zobrazí tabulku, která se nachází na obrázku 1.2. Tabulka na první pohled připomíná tabulku *Microsoft Excel*. Lze v ní barevně označit sloupce, připnout sloupce důležité a schovat sloupce nepotřebné, nahradit hodnoty ve sloupci jinými, změnit typ sloupce, smazat nebo zobrazit duplicity v datech, vizualizovat data ve sloupci pomocí box plotu, histogramu, či Quantile-quantile plot (Q-Q plot). Knihovna umí pro každý graf, který vykreslí, vrátit zdrojový kód. Jednou z dalších možností je sekce *describe*, která umožňuje vykreslit komplexnější grafy. Nachází se zde i možnost vybrat pro histogram sloupec cílové proměnné, podle kterého je graf rozdělen. *D-Tale* je nástroj usnadňující manuální exploraci dat. Nejedná se tedy o automatickou exploraci dat. Demo knihovny je k dispozici v [7].

D-TALE Actions Visualize Highlight Settings

12 / 891	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Pa
0	1	0	3	Braund, Mr. Owen Harris	male	22.00	1	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	
2	3	1	3	Heikkinen, Miss. Laina	female	26.00	0	
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	
4	5	0	3	Allen, Mr. William Henry	male	35.00	0	
5	6	0	3	Moran, Mr. James	male	nan	0	
6	7	0	1	McCarthy, Mr. Timothy J	male	54.00	0	
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.00	3	
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0	
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1	
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.00	1	
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.00	0	
12	13	0	3	Saundercock, Mr. William Henry	male	20.00	0	
13	14	0	3	Andersson, Mr. Anders Johan	male	39.00	1	
14	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.00	0	
15	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55.00	0	

Obrázek 1.2 Ukázka knihovny *D-tale* na datové sadě Titanik.

1.3 Pandas Profiling

Knihovna *Pandas Profiling* [1], která byla na začátku ledna roku 2023 přejmenována na *ydata profiling*[1], patří mezi nejrozšířenější knihovny pro automatickou exploraci dat v datascience komunitě. Z důvodu popularity této knihovny a z důvodu, že tato knihovna byla následně vybrána pro implementaci rozšíření v rámci této práce, je jí v rešerši věnována větší pozornost než předchozím knihovnám. *Pandas Profiling* umožňuje vytvořit HTML report. Report vytvořený pomocí *Pandas Profiling* knihovny má následující části: přehled, proměnné, interakce proměnných, korelace, chybějící hodnoty a vzorek dat.

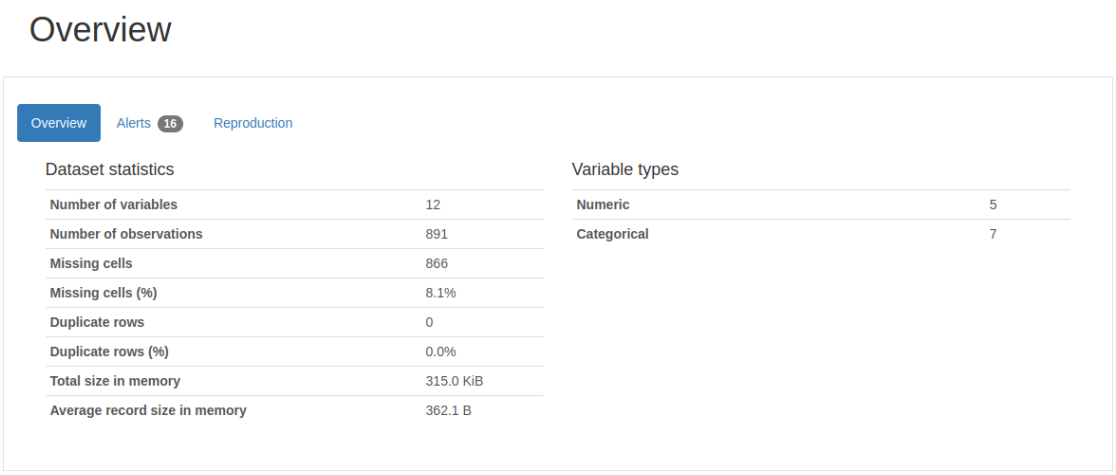
První sekce reportu je pojmenována Přehled. Obsahuje tři podsekce: Přehled (Overview), Upozornění (Alerts) a Informace o reportu (Reproduction). Ukázka této sekce reportu se nachází na obrázku 1.3.

V podsekci *Přehled* jsou zobrazeny statistiky datového souboru, mezi které patří: počet sloupců, počet řádků, počet chybějících hodnot a počet duplicitních hodnot. Dále se zde nachází přehled datových typů vyskytujících se v datové sadě a jejich zastoupení.

[1] Dále v textu bude použit pouze název *Pandas Profiling*.

V podsekci *Upozornění* se nachází přehled důležitých vlastností jednotlivých sloupců. Vlastnosti, které *Pandas Profiling* zahrnuje mezi upozornění jsou: vysoká korelace mezi dvěma či více sloupci, vysoká kardinalita dat, počet chybějících hodnot, rovnoměrné rozdělení dat a upozornění na výskyt nulových hodnot. Pomocí této sekce je možné získat rychlý přehled o potenciálně důležitých či problematických sloupcích v datech.

V podsekci *Informace o reportu* se nachází datum a čas vygenerování reportu, informace o tom, jak dlouho generování trvalo a verze použité knihovny *Pandas Profiling*. V této sekci se také nachází možnost stažení konfiguračního souboru, který byl použit při vytvoření reportu.



Obrázek 1.3 Ukázka knihovny *Pandas Profiling* – Přehled

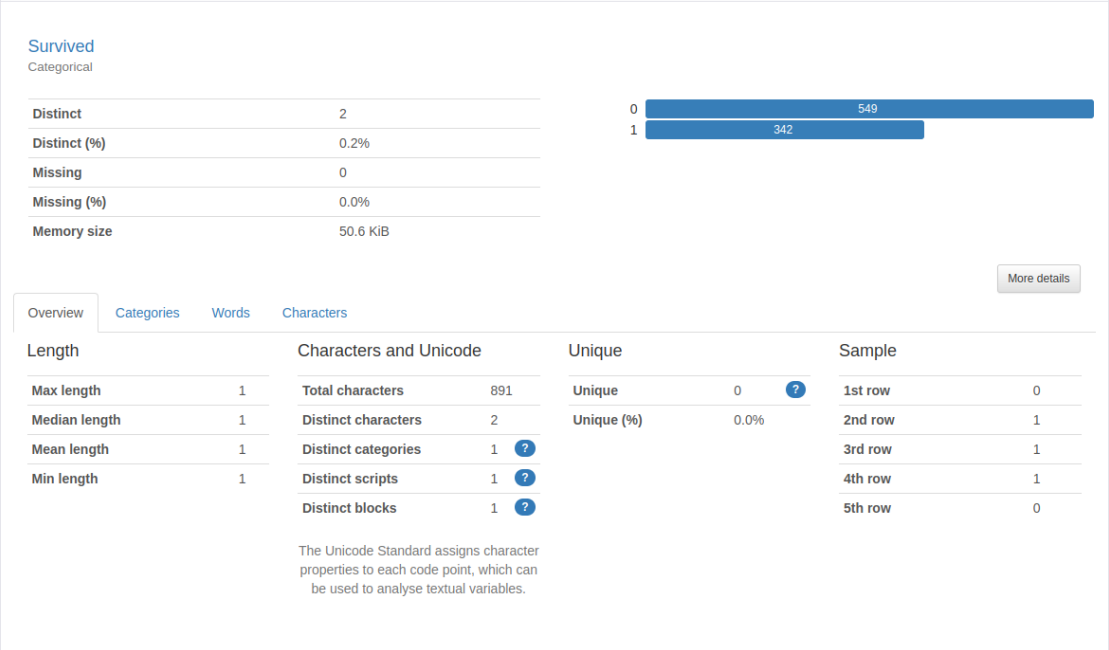
Druhá sekce reportu je věnována jednotlivým proměnným (sloupcům z datové sady). Pro každou proměnnou je zobrazen její datový typ a přehled vybraných popisných statistik a vizualizace dat. Dále je možné zobrazit podrobnější informace o proměnné. Informace a vizualizace proměnné se liší dle datového typu proměnné. Pozornost je věnována pouze numerickým a kategoriálním veličinám. Textové veličiny knihovna *Pandas Profiling* v době psaní práce nepodporuje a ostatní podporované datové typy jsou speciální případy kategoriálních veličin.

U kategoriálních veličin je zobrazen počet unikátních hodnot a počet chybějících hodnot. K vizualizaci dat je použit barplot se zastoupením jednotlivých tříd. V podrobnějších informacích jsou zobrazeny statistiky kategorií, jako je průměrná, minimální a maximální délka názvu kategorie, celkový počet znaků a zastoupení jednotlivých kategorií v datech. V podrobnějších informacích se také nachází informace o zastoupení jednotlivých slov a znaků.

U numerických veličin jsou zobrazeny následující popisné statistiky: počet unikátních hodnot a počet chybějících hodnot, průměrná, minimální a maximální hodnota, počet nulových hodnot a zastoupení negativních hodnot v datech. K vizualizaci numerických veličin je použit histogram. V podrobnějších informacích se nacházejí hodnoty pro 5. a 95. percentil, hodnoty prvního a třetího kvartilu, mediánová hodnota, rozsah hodnot ve sloupci, mezikvartilové rozpětí (rozdíl mezi třetím a prvním kvantilem), hodnoty pro standardní odchylku a rozptyl a součet všech hodnot. Dále se zde nachází sekce pro hodnoty s největším zastoupením a sekce pro extrémní hodnoty s přehledem několika nejvyšších a několika nejnižších hodnot. Ukázka numerické proměnné v *Pandas Profiling* reportu se nachází na obrázku 1.4.

Ve třetí sekci reportu se nacházejí interakce mezi jednotlivými proměnnými. Tato sekce obsahuje korelační diagram (scatter plot) pro zobrazení závislostí mezi číselnými proměnnými. V diagramu chybí kategoriální veličiny.

Čtvrtá sekce se věnuje korelacím. *Pandas Profiling* pracuje s různými metodami pro výpočet korelace. Mezi podporované metody patří *Spearmanova*, *Pearsonova*, *Kendallova* a *Cramerova*



Obrázek 1.4 Ukázka knihovny *Pandas Profiling* – Popis proměnné přežití v ukázkové datové sadě.

korelace. Při vytváření reportu je možné nastavit, které korelace budou vygenerovány a které nikoliv.

Pátá sekce se věnuje chybějícím hodnotám v datech. Obsahuje přehled o tom, kolik validních hodnot se nachází v jednotlivých sloupcích, a také korelační matice vypovídající o tom, jak moc na sobě chybějící hodnoty u jednotlivých proměnných závisí.

Poslední sekce reportu obsahuje vzorek dat, kde je zobrazeno N prvních a N posledních řádků z datové sady.

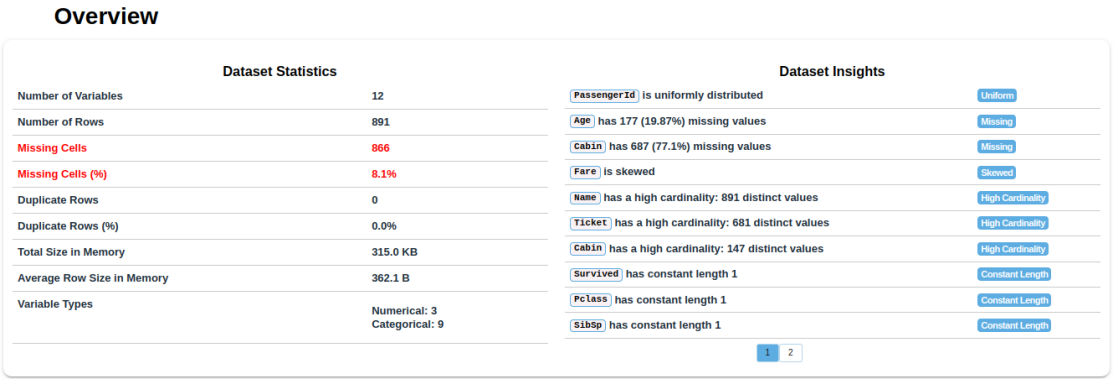
Jedna z výhod knihovny *Pandas Profiling* je také možnost přizpůsobit si vygenerovaný report pomocí konfiguračního souboru. Je možné upravit vzhled reportu, ale také nastavit, které sekce reportu budou zobrazeny a které ne.

Ukázka vytvořeného reportu pomocí knihovny *Pandas Profiling* na datové sadě Titanik se nachází zde [8].

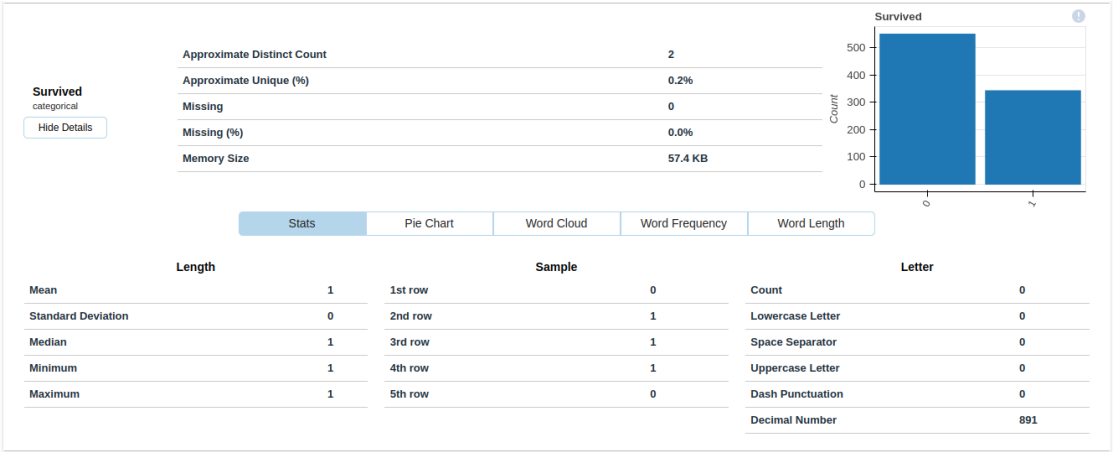
1.4 DataPrep

Knihovna *DataPrep* [9] obsahuje tři moduly. Pro automatickou exploraci dat je relevantní pouze modul Exploratory Data Analysis (EDA). Tento modul obsahuje funkce pro generování reportu a jednotlivých grafů. Pomocí funkce `create_report()` lze vytvořit HTML report. Report vygenerovaný pomocí *DataPrep* knihovny je podobný reportu z knihovny *Pandas Profiling*. Skládá se dokonce ze stejných sekcí: přehled, proměnné, interakce, korelace a chybějící hodnoty. Rozdíly mezi *Pandas Profiling* reportem a *DataPrep* reportem jsou z pohledu uživatele minimální. Je poznat, že se knihovny vzájemně inspirovaly řešením některých vizualizací (porovnání přehledu reportu *DataPrep* 1.5 a *Pandas Profilingu* 1.3.) Sekce proměnných je také podobná, (viz porovnání sekce proměnné v reportu *Dataprep* 1.6 a sekce proměnné v reportu *Pandas Profiling* 1.4) .

Výhodou oproti knihovně *Pandas Profiling*, je možnost extrahovat jednotlivé části reportu pomocí následujících funkcí: Funkce `plot()` umožňuje více funkcionalit, dle vstupních parametrů. V případě jednoho vstupního parametru (datová sada) vygeneruje vizualizaci distribucí pro jednotlivé proměnné. Pro dva vstupní parametry (datová sada a název sloupce) vygeneruje



Obrázek 1.5 Ukázka knihovny *DataPrep* – Přehled



Obrázek 1.6 Ukázka knihovny *DataPrep* – Popis proměnné přežití v ukázkové datové sadě.

podrobnější vizualizace a popisné statistiky pro specifikovaný sloupec. Pro tři vstupní parametry (datová sada, první sloupec, druhý sloupec) vygeneruje grafy pro vizualizaci vztahů těchto dvou sloupců. Funkce `plot_correlation()` umožňuje vygenerovat korelační matice a podporuje *Pearsonovu*, *Spearmanovu* a *Kendallovu* korelaci. Další funkcí v tomto modulu je `plot_missing()`, která se zabývá chybějícími hodnotami. S její pomocí lze zobrazit, v jakých proměnných chybějí hodnoty, jestli spolu chybějící hodnoty nějakým způsobem souvisí, a také to, jak by datovou sadu ovlivnilo, kdyby byly odstraněny všechny řádky s chybějící hodnotou ve specifikované proměnné. Funkce `plot_diff()` umožňuje zobrazit rozdíl mezi dvěma datovými sadami. Nic z těchto funkcionalit knihovna *Pandas Profiling* neumožňuje.

Ukázka použití knihovny je k dispozici zde [10].

1.5 Sweetviz

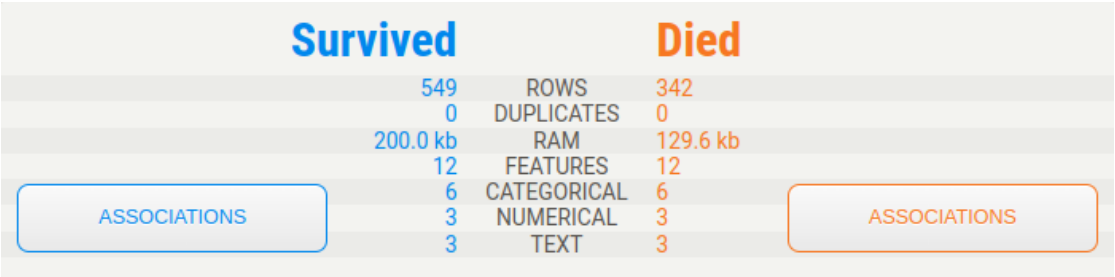
Knihovna *Sweetviz* [11] je další z možností, pro generování automatického reportu pro exploraci dat. Design vygenerovaného reportu je poměrně nepřehledný. Množství informací, které report obsahuje je oproti předchozím knihovnám značně omezené. Mezi výhody knihovny *Sweetviz* patří podpora datového typu pro textové proměnné. Předchozí knihovny totiž textové proměnné klasifikují jako kategoriální. Další výhodou této knihovny je možnost vytvořit report nad dvěma datovými sadami a vzájemně je porovnat. Při rozdělení ukázkové datové sady Titanik dle příznaku

přežití do dvou datových sad, lze vytvořit report s oddělenými vizualizacemi dle přežití. *Sweetviz* report se skládá ze dvou částí. První část obsahuje přehled a druhá se věnuje jednotlivým proměnným.

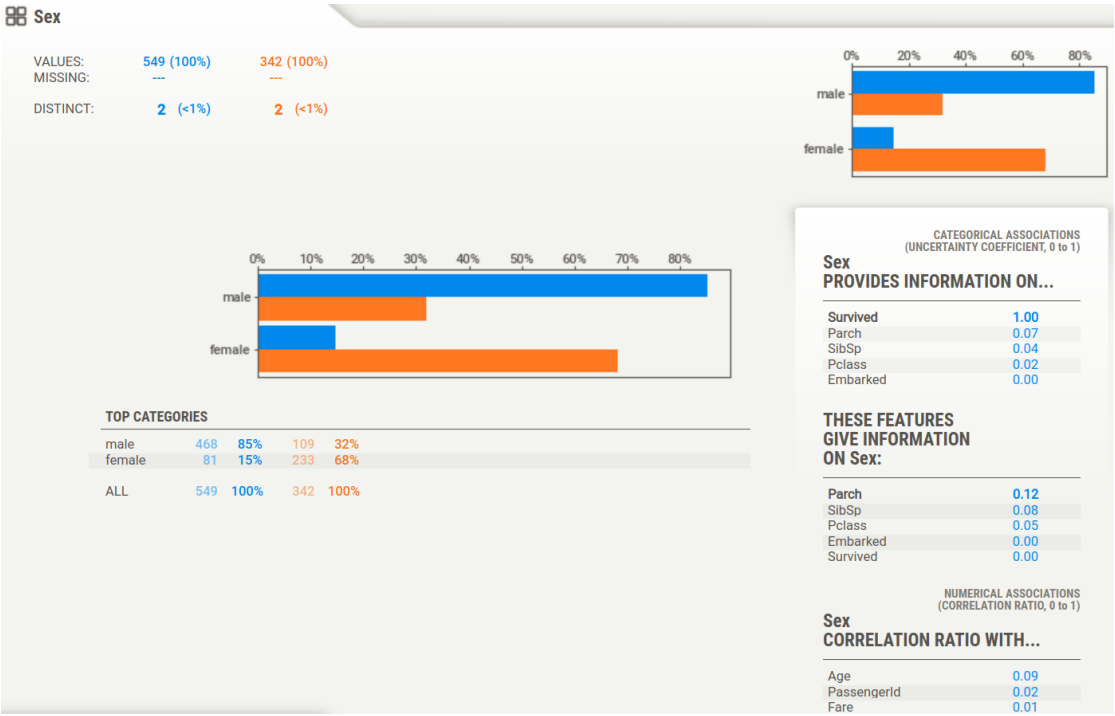
V části *přehled* je zobrazen počet řádků a sloupců v jednotlivých datových sadách, datové typy a jejich zastoupení. Ukázka části *přehled* v reportu vygenerovaném pomocí *Sweetviz* knihovny se nachází na obrázku 1.7.

Druhá sekce zobrazuje pro jednotlivé proměnné vizualizaci rozložení dat, počty hodnot a informace o závislosti s ostatními proměnnými. Ukázka proměnné pohlaví se nachází na obrázku 1.8.

Demo knihovny *Sweetviz* je dostupné zde [12].



Obrázek 1.7 Ukázka knihovny *Sweetviz* – Přehled



Obrázek 1.8 Ukázka knihovny *Sweetviz* – Popis proměnné pohlaví v ukázkové datové sadě.

1.6 Výběr vhodné knihovny pro implementaci rozšíření

Cílem této práce je usnadnit exploraci dat s binární klasifikací. Proto bude vybrána jedna z představených knihoven, do které bude implementováno rozšíření se specializací na binární klasifikaci. Knihovna *D-tale* není nástrojem pro automatickou exploraci dat, jedná se o nástroj usnadňující manuální exploraci dat. Z tohoto důvodu není pro implnementaci vhodným kandidátem. Knihovna *Sweetviz* podporuje generování reportu ze dvou různých datových sad nebo subsekcí jedné datové sady. Toho by bylo možné využít při implementaci rozšíření se specializací na binární klasifikaci. Knihovna *Sweetviz* však zaostává za knihovnami *Pandas Profiling* a *DataPrep* v množství informací, které jsou v reportu obsaženy. Není také tolik rozšířená, jako zmíněné knihovny. Z těchto důvodů není vhodným kandidátem pro implementaci rozšíření. Knihovna *Pandas Profiling* a knihovna *DataPrep* jsou komplexními knihovnami, které generují report ve formátu HTML. Report obsahuje spoustu důležitých statistik a grafů k jednotlivým proměnným. Rozšíření se specializací na binární klasifikací by bylo možné implementovat u obou z nich. Z důvodu rozšířenosti v datascience komunitě byla nakonec pro implementaci rozšíření vybrána knihovna *Pandas Profiling*.

Kapitola 2
Teoretická část

Tato kapitola je zaměřena na statistické testy a jiné metody pro odhalení závislosti dvou veličin. Dále se věnuje vizualizacím. Definice bez citace jsou obecně známými metodami a mohou být nalezeny například v knize [13].

Součástí rozšíření knihovny *Pandas Profiling* bude srovnání subpopulací proměnných na základě cílové proměnné. Jinými slovy, každá proměnná bude rozdělena do dvou populací dle cílové proměnné. Aby bylo možné určit, zda se tyto dvě populace významně liší, je potřeba vybrat a použít statistické testy. Představením vybraných statistických testů se zabývá následující kapitola.

2.1 Statistické testy

„Statistické testy představují způsob, jak matematicky určit, zda se dva soubory dat od sebe významně liší. K tomu statistické testy používají několik statistických měr, jako je průměr, směrodatná odchylka a variační koeficient. Po výpočtu statistických měr je statistický test porovnán se souborem předem stanovených kritérií. Pokud data splňují kritéria, statistický test dojde k závěru, že mezi oběma soubory dat existuje významný rozdíl. V závislosti na typu analyzovaných údajů lze použít různé statistické testy. Mezi nejběžnější statistické testy však patří t-testy, chí-kvadrát testy a testy ANOVA.“ [14]

2.1.1 T-test

Studentův t-test je jedním z nejznámějších statistických testů. T-test má tři varianty: *jednovýběrový*, *dvouvýběrový* a *párový*. T-testy pracují s náhodným výběrem $\mathrm{N}(\mu, \sigma^2)$. Hodnoty náhodného výběru jsou označeny $X_1, X_2, ..., X_n$. Střední hodnota a rozptyl jsou odhadnuty pomocí výběrového průměru $\bar{X}_n$ a výběrového rozptylu s_n^2:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

2.1.1.1 Jednovýběrový t-test

Jednovýběrový t-test slouží k ověření hypotézy o střední hodnotě daného náhodného výběru. Předpokládá, že data mají normální rozdělení a že rozptyl rozdělení není znám. Při testování

hypotézy o střední hodnotě je nulová hypotéza $H_0 : \mu = \mu_0$, kde μ_0 je konkrétní hodnota. Alternativa H_A má tři různé varianty, dle zadání úlohy.

Nulová hypotéze $H_0 : \mu = \mu_0$ je zamítnuta ve prospěch oboustranné alternativa $H_A : \mu \neq \mu_0$, pokud testovaná hodnota μ_0 neleží v oboustranném konfidenčním intervalu, který je definován následovně:

$$(\bar{X}_n - t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}}, \quad \bar{X}_n + t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}}),$$

kde α je hladina významnosti testu.

Nulová hypotéze $H_0 : \mu = \mu_0$ je zamítnuta ve prospěch jednostranné alternativa $H_A : \mu > \mu_0$, pokud testovaná hodnota μ_0 neleží v horním jednostranném konfidenčním intervalu:

$$(\bar{X}_n - t_{\alpha,n-1}\frac{s_n}{\sqrt{n}}, \quad +\infty),$$

kde α je hladina významnosti testu.

Nulová hypotéze $H_0 : \mu = \mu_0$ je zamítnuta ve prospěch oboustranné alternativa $H_A : \mu < \mu_0$, pokud testovaná hodnota μ_0 neleží v dolním jednostranném konfidenčním intervalu:

$$(-\infty, \quad \bar{X}_n + t_{\alpha,n-1}\frac{s_n}{\sqrt{n}}),$$

kde α je hladina významnosti testu.

2.1.1.2 Dvouvýběrový t-test

Dvouvýběrový t-test slouží k porovnání dvou nezávislých náhodných výběrů. Předpokládá, že veličiny jsou nezávislé a normálně rozdělené. $X_1, X_2, \ldots, X_{n_1}$ je náhodný výběr z $X_i \sim \mathrm{N}(\mu_1, \sigma_1^2)$ a $Y_1, Y_2, \ldots, Y_{n_2}$ je náhodný výběr z $Y_i \sim \mathrm{N}(\mu_2, \sigma_2^2)$. Při testování závislosti, či nezávislosti středních hodnot je nulová hypotéza $H_0 : \mu_1 = \mu_2$ a alternativa $H_A : \mu_1 \neq \mu_2$. Dvouvýběrový t-test má dvě varianty dle toho, zda mají oba výběry shodný rozptyl ($\sigma_1^2 = \sigma_2^2$), či nikoliv.

Při stejném rozptylu náhodných výběrů $\sigma_1^2 = \sigma_2^2$ je testovaná statistika T vypočítána pomocí následujícího vzorce:

$$T = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2}}{\sqrt{\frac{(n_1-1)s_X^2 + (n_2-1)s_Y^2}{n_1+n_2-2}}}\sqrt{\frac{n_1 n_2}{n_1+n_2}},$$

kde s_X^2 a s_Y^2 jsou výběrové rozptyly X_i a Y_i. Kritická hodnota testované statistiky T pro vyvrácení nulové hypotézy ve prospěch alternativy je

$$|T| > t_{\alpha/2,n_1+n_2-2}.$$

Při různém rozptylu náhodných výběrů $\sigma_1^2 \neq \sigma_2^2$ je testovaná statistika T vypočítána pomocí vzorce

$$T = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2}}{s_d},$$

kde s_X^2 a s_Y^2 jsou výběrové rozptyly X_i a Y_i a kde

$$s_d = \sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}.$$

Kritická hodnota testované statistiky T pro vyvrácení nulové hypotézy ve prospěch alternativy je

$$|T| > t_{\alpha/2,n_d},$$

kde

$$n_d = \frac{s_d^4}{\frac{1}{n_1-1}(\frac{s_X^2}{n_1})^2 + \frac{1}{n_2-1}(\frac{s_Y^2}{n_2})^2},$$

kde s_X^2 a s_Y^2 jsou výběrové rozptyly X_i a Y_i.

2.1.1.3 Párový t-test

Párový t-test pracuje s náhodným výběrem párů $(X_1, Y_1), \ldots, (X_n, Y_n)$, přičemž veličiny uvnitř párů mohou být závislé. Párový t-test předpokládá normální rozdělení obou veličin, které lze definovat následovně: $X_i \sim \mathrm{N}(\mu_1, \sigma_1^2)$ a $Y_i \sim \mathrm{N}(\mu_2, \sigma_2^2)$. Testovaná hypotéza je definována $H_0 : \mu_1 = \mu_2$. Vytvořením $Z_i = Y_i - X_i$ vznikne nové rozdělení, se střední hodnotou $\mu_{diff} = \mu_2 - \mu_1$, které má také normální rozdělení. Test shody středních hodnot lze převést na jednovýběrový t-test s oboustrannou alternativou, kde nulová hypotéze je $H_0 : \mu_{diff} = 0$ a alternativa $H_A : \mu_1 \neq \mu_2$.

„Korelační t-test neboli párový t-test je závislý typ testu a provádí se v případě, že se vzorky skládají ze shodných párů podobných jednotek nebo v případech opakovaných měření. Například se mohou vyskytnout případy, kdy jsou stejní pacienti opakovaně testováni před a po absolvování určité léčby. Každý pacient je použit jako kontrolní vzorek proti sobě samému.“ [15]

2.1.2 Chí-kvadrát test nezávislosti

Testem, umožňující testování závislost či nezávislost dvou diskrétních veličin, je chí-kvadrát test nezávislosti. *„Chí-kvadrát (X^2) test nezávislosti je typem Pearsonova chí-kvadrát testu. Pearsonovy chí-kvadrát testy jsou neparametrické testy pro kategoriální proměnné. Používají se k určení, zda se data významně liší od toho, co bylo očekáváno.“* [16] Chí-kvadrát test nezávislosti vychází z pozorovaných četností, což jsou počty pozorování v kontingenční tabulce. Test porovnává pozorované četnosti s očekávanými četnostmi. Pokud jsou proměnné závislé, budou pozorované a očekávané četnosti podobné. Tabulka očekávaných četností je definována následovně:

$$E_{ij} = \frac{R_i \cdot C_j}{N},$$

kde R_i je součet hodnot i-tého řádku, C_j součet hodnot j-tého sloupce a N je celkový součet hodnot všech pozorování. Dále je spočítána testovací statistika s označením X^2 pomocí následujícího vzorce:

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

kde O_{ij} je reálná hodnota i-tého řádku j-tého sloupce v tabulce četností a E_{ij} je očekávaná hodnota i-tého řádku j-tého sloupce. Kritická hodnota testované statistiky X^2 pro vyvrácení nulové hypotézy ve prospěch alternativy je

$$X^2 \geq \chi^2_{\alpha,(r-1)(c-1)},$$

kde r je počet řádků a c je počet sloupců v kontingenční tabulce. [16]

Pro příklad použití chí-kvadrát testu nezávislosti slouží tabulka 2.1. Test nezávislosti je proveden na hladině významnosti $\alpha = 0{,}05$. Tabulka očekávaných četností pro tento příklad je zobrazena v tabulce 2.2. Dále je třeba spočítat testovací statistiku:

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 2{,}78 + 5 + 2{,}78 + 5 = 15{,}56.$$

Nyní stačí porovnat statistiku s chí-kvadrát rozdělením. Stupně volnosti chí-kvadrát rozdělení jsou rovny $(r-1)(c-1) = (2-1) \cdot (2-1) = 1$.

$$15{,}56 = X^2 \geq \chi^2_{\alpha;(r-1)(c-1)} = \chi^2_{0{,}05;1} = 3{,}841,$$

proto je nulová hypotéza H_0 na hladině důvěryhodnosti 95 % zamítnuta ve prospěch alternativy H_A.

Tabulka 2.1 Tabulka četností atletů/kuřáků.

Atlet/Kuřák	Ne	Ano	Součet
Ano	14	4	18
Ne	0	10	10
Součet	14	14	28

Tabulka 2.2 Tabulka očekávaných četností pro tabulku 2.1.

Atlet/Kuřák	Ne	Ano	Součet
Ano	$(18 \cdot 14)/28 = 9$	$(18 \cdot 14)/28 = 9$	18
Ne	$(10 \cdot 14)/28 = 5$	$(10 \cdot 14)/28 = 5$	10
Součet	14	14	28

2.1.3 Fisherův exaktní test

Fisherův exaktní test patří mezi exaktní testy pro testování závislosti dvou binárních veličin. *„Fisherův přesný test je vhodné použít zejména v případě malých počtů. Chí-kvadrát test je v podstatě aproximací výsledků exaktního testu, takže chybné výsledky by mohly být potenciálně získány z malého počtu pozorování.“* [17] Pro testování závislosti pomocí Fisherova exaktního testu jsou definovány dvě hypotézy: nulová hypotéza H_0, dle které jsou veličiny nezávislé, a alternativa H_A, dle které jsou veličiny závislé. Dle nulové hypotézy H_0 pochází vstupní tabulka z hypergeometrického rozdělení s následujícími parametry: celkový počet pozorování $M = a + b + c + d$, počet pozorování prvního typu $n = a + b$ a počet vybraných prvků $N = a + c$, pro vstupní tabulku ve formátu $[[a, b], [c, d]]$. Toto rozdělení podporuje tabulky, pro které platí:

$$\max(0, a - d) \leq x \leq a + \min(b, c),$$

kde x lze interpretovat jako levý horní prvek tabulky. Tabulky v rozdělení mají tvar zobrazený v tabulce 2.3. [18]

Tabulka 2.3 Tvar tabulek hypergeometrického rozdělení.

	Ne	Ano
Ano	x	$n - x$
Ne	$N - x$	$M - (n + N) + x$

Pravděpodobnost získání jedné tabulky je dána vzorcem:

$$p_{a,b,c,d} = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}.$$

K vyhodnocení testu na určité hladině důvěryhodnosti je třeba sečíst všechny pravděpodobnosti výskytu tabulek, které mají nižší nebo rovnou pravděpodobnost výskytu jako zdrojová tabulka.

Pro ukázkový příklad použití Fisherova exaktního testu je použita tabulka 2.1 s označením buněk podle tabulky 2.4. Pravděpodobnost získání takové tabulky za předpokladu nezávislosti kouření a skutečnosti, že zkoumaným je atlet, je dána vzorcem:

$$p_{14,4,0,10} = 0{,}000076.$$

Dále dle definice musí platit následující vzorec:

$$\max(0, a - d) = 4 <= x <= 14 = a + \min(b, c),$$

Tabulka 2.4 Označení buněk tabulky četností 2.1.

Atlet/Kuřák	Ne	Ano
Ano	a	b
Ne	c	d

kde x reprezentuje horní levý roh tabulky četností. V příkladovém hypergeometrickém rozdělení tedy existuje 11 možností rozdělení dat. Pravděpodobnosti těchto jedenácti možných rozdělení jsou vypočítány dle vzorce a vycházejí následovně:

$$[7{,}62776506 \cdot 10^{-5}; 2{,}13577422 \cdot 10^{-3}; 2{,}08237986 \cdot 10^{-2}; 9{,}51945080 \cdot 10^{-2};$$
$$2{,}29061785 \cdot 10^{-1}; 3{,}05415713 \cdot 10^{-1}; 2{,}29061785 \cdot 10^{-1}; 9{,}51945080 \cdot 10^{-2};$$
$$2{,}08237986 \cdot 10^{-2}; 2{,}13577422 \cdot 10^{-3}; 7{,}62776506 \cdot 10^{-5}].$$

Pravděpodobnosti, které jsou nižší nebo rovny pravděpodobnosti zdrojové tabulky jsou pouze dvě: 0,000076 a 0,000076. P-hodnota oboustranného Fisherova exaktního testu je tedy rovna hodnotě 0,00015. Platí tedy:

$$p_{\text{val}} < 1 - 0{,}95.$$

Z tohoto důvodu je nulová hypotéza H_0 zamítnuta ve prospěch alternativy, a to na hladině důvěryhodnosti 95 %.

2.2 Statistické metody pro zkoumání subpopulací

Statistické testy neumí odhalit všechny odlišnosti dvou veličin. T-test umožňuje testovat závislost či nezávislost průměrných hodnot dvou veličin, což ale není jediná vlastnost, ve které se mohou proměnné lišit. Pomocí chí-kvadrátu je možné určit, zda jsou diskrétní data závislá, není ale možné říci, jaké hodnoty v diskrétní proměnné mají nejvíce vychýlené počty výskytů. Z tohoto důvodu budou prozkoumány další statistické metody.

2.2.1 Relativní riziko

Relativní riziko je poměrem pravděpodobnosti příslušnosti k jedné skupině oproti pravděpodobnosti příslušnosti k druhé skupině [19].

Vzorec pro relativní riziko je následující:

$$\text{RR} = \frac{p_1}{p_2},$$

kde p_1 je pravděpodobnost výskytu v první skupině a p_2 je pravděpodobnost výskytu ve druhé skupině.

Následuje příklad s daty z tabulky 2.5 a označením buněk podle tabulky 2.4, kde bude spočítán relativní risk kouření u ne-atletů oproti atletům. Vzorec pro tento příklad vypadá následovně:

$$\text{RR} = \frac{p_1}{p_2} = \frac{\frac{10}{10+10}}{\frac{1}{9+1}} = 5,$$

kde p_1 je pravděpodobnost kouření u ne-atletů a p_2 je pravděpodobnost kouření u atletů. Je tedy 5-krát pravděpodobnější, že kouří ne-atlet, než že kouří atlet.

Tabulka 2.5 Tabulka četností pro ukázky relativního rizika a poměru šancí.

Atlet/Kuřák	Ne	Ano
Ano	9	1
Ne	10	10

2.2.2 Poměr šancí

Poměr šancí je poměrem šance příslušnosti k jedné skupině oproti šanci příslušnosti k druhé skupině [20]. Podobá se relativnímu riziku s jediným rozdílem, že namísto pravděpodobností pracuje se šancemi. Vzorec pro *poměr šancí* vypadá následovně:

$$\mathrm{OR} = \frac{\mathrm{odds}_1}{\mathrm{odds}_2},$$

kde $odds_1$ je šance příslušnosti k první skupině a $odds_2$ je šance příslušnosti k druhé skupině.

Následuje příklad s daty z tabulky 2.5 a označením buněk podle tabulky 2.4, kde bude spočítán *poměr šancí* kouření u ne-atletů oproti atletům. Šance kouření u atletů je $1 : 9$, zatímco šance kouření u ne-atletů je $1 : 1$.

$$\mathrm{OR} = \frac{\mathrm{odds}_1}{\mathrm{odds}_2} = \frac{\frac{d}{c}}{\frac{b}{a}} = \frac{ad}{bc} = \frac{9 \cdot 10}{1 \cdot 10} = 9,$$

kde $odds_1$ je šance příslušnosti k první skupině (kouřící ne-atleti) a $odds_2$ je šance příslušnosti k druhé skupině (kouřící atleti). Poměr šancí je 9 (šance, že ne-atlet kouří je 9 krát vyšší než šance, že kouří atlet).

Definiční obor *poměru šancí* je zdola omezen nulou, zatímco nahoře je neomezen, má tedy šikmé rozdělení. Možnost, jak jej upravit, je použití logaritmu. Logaritmický *poměr šancí* má přibližně normální rozdělení [21]. Další výhodou logaritmického *poměru šancí* je, že při prohození skupin je výsledná hodnota stejná s opačným znaménkem:

$$\log_2 \frac{1}{0{,}5} = 1,$$

$$\log_2 \frac{0{,}5}{1} = -1.$$

Konfidenční interval pro *poměr šancí* lze spočítat pomocí následujícího vzorce [20]:

$$\mathrm{CI}_{0,95} = e^{\ln \mathrm{OR} \pm 1{,}96\sqrt{1/a+1/b+1/c+1/d}}.$$

Poměr šancí a relativní riziko se používají v lékařství. Pomocí obou metod je možné vyhodnotit například kvalitu léčby.

„Interpretace poměru šancí může být neintuitivní, ale téměř ve všech reálných případech je nepravděpodobné, že by interpretace poměru šancí jako relativního rizika změnila kvalitativní hodnocení výsledků studie. Poměr šancí bude při interpretaci jako relativní riziko vždy nadhodnocený a míra nadhodnocení se bude zvyšovat jak s rostoucím počátečním rizikem, tak s rostoucí velikostí případného účinku léčby. Neexistuje však žádný bod, v němž by míra nadhodnocení mohla vést ke kvalitativně odlišnému hodnocení studie.“ [22]

2.3 Vyhlazovací metody

Některé modely a statistické metody jsou náchylné na chybějící a zašuměná data. Pomocí následujících metod je možné data vyhladit a do nezastoupených skupin přidat umělá pozorování.

2.3.1 Laplaceovo vyhlazování

Laplaceovo vyhlazování je jedna z nejjednodušších vyhlazovacích metod. Metoda přidá do každé skupiny v datech α umělých pozorování. V případě, že by nějaká skupina nebyla v datech zastoupena, po použití Laplaceova vyhlazování bude v této skupině α umělých pozorování.

„Laplaceovo vyhlazování je vyhlazovací metoda, pomocí které lze řešit problém nulové pravděpodobnosti v algoritmu Naïve Bayes. Použití vyšších hodnot alfa posune pravděpodobnost k hodnotě 0,5, tj. pravděpodobnost slova rovná 0,5 pro pozitivní i negativní recenze. Protože z toho nezískáme mnoho informací, není to výhodné. Proto je vhodnější použít hodnotu alfa=1.“ [23]

2.3.2 Beta rozdělení

Beta rozdělení je definováno pomocí dvou parametrů α a β [24]. Průměrná hodnota rozdělení $\mathrm{E}X$ a rozptyl $\mathrm{var}X$ jsou definovány následovně:

$$\mathrm{E}X = \frac{\alpha}{\alpha+\beta} \qquad \mathrm{var}X = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Beta rozdělení je možné použít k vyhlazení pozorování na základě očekávaného průměru (prioru). Nejprve je třeba zvolit parametry α_0 a β_0 tak, aby průměrná hodnota rozdělení $\mathrm{Beta}(\alpha_0;\beta_0)$ byla rovna prioru. Pomocí velikosti parametrů α_0 a β_0 je možné určit sílu prioru oproti pozorovaným datům. Následně je vytvořeno rozdělení:

$$\mathrm{Beta}(\alpha_0+\alpha;\beta_0+\beta),$$

kde α reprezentuje pozitivní počet výskytů v pozorování a β reprezentuje negativní počet výskytů v pozorování. Výsledná hodnota pravděpodobnosti je reprezentována střední hodnotou rozdělení, která je z důvodu použití parametrů α_0 a β_0 posunuta směrem k prioru. [25]

Následuje příklad beta vyhlazování na datech z Titaniku. Úkolem je predikovat pravděpodobnost přežití skupiny 2 mužů, z nichž 1 přežil a 1 nepřežil a skupiny 200 mužů z nichž 100 přežilo a 100 nepřežilo. Bez použití vyhlazování by obě skupiny dopadly stejně s pravděpodobností přežití 0,5. S použitím beta rozdělení je brán v potaz prior (pravděpodobnost přežití v populaci). V celé populaci, která je k dispozici bylo 577 mužů z nichž 468 nepřežilo a 109 přežilo. Pravděpodobnost přežití muže je tedy rovna $\frac{109}{109+468} = 0{,}1889$. To lze aproximovat beta rozdělením například s parametry $\alpha = 10$ a $\beta = 43$. $\frac{10}{10+43} = 0{,}1887$. Beta rozdělení s těmito parametry je zobrazeno na obrázku 2.1.

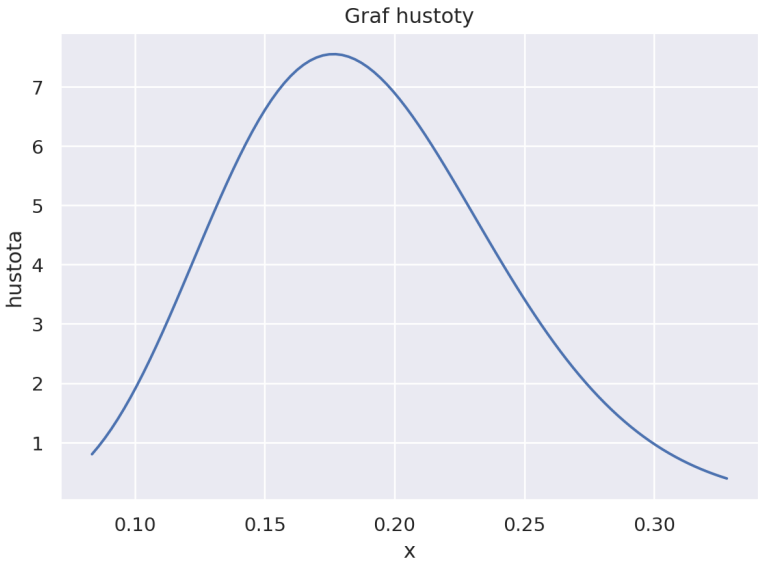
V grafu 2.2 jsou zobrazeny distribuční funkce tří beta rozdělení. První beta rozdělení reprezentuje výchozí rozdělení (prior). Druhá distribuční funkce beta rozdělení reprezentuje skupinu dvou mužů s beta vyhlazením a třetí beta rozdělení reprezentuje skupinu 200 mužů se stejným beta vyhlazením. Vzorec pro skupinu dvou mužů je

$$\mathrm{Beta}(\alpha_0+1;\beta_0+1) = \mathrm{Beta}(11;44)$$

a vzorec pro skupinu 200 mužů je

$$\mathrm{Beta}(\alpha_0+100;\beta_0+100) = \mathrm{Beta}(110;143).$$

Pravděpodobnost je následně vypočítána jako průměrná hodnota rozdělení. Z grafu vyplývá, že pro skupinu s málo záznamy hraje hlavní roli očekávaný výsledek (prior). Naopak u dat s dostatkem záznamů se výsledek blíží nevyhlazenému, je ovšem mírně posunut k prioru. Pro skupinu dvou mužů, ze které jeden přežil a jeden nepřežil je pravděpodobnost přežití po vyhlazení rovna 0,2. Pro skupinu 200 mužů, z nichž 100 přežilo a 100 nepřežilo je pravděpodobnost přežití po vyhlazení rovna 0,43.



Obrázek 2.1 Beta rozdělení s nastavením parametrů $\alpha = 10$ a $\beta = 43$

2.4 Vizualizace dat

"Vizualizace dat je částečně umění a částečně věda. Výzvou je správně zvládnout umění, aniž by se věda zvrhla, a naopak. Vizualizace dat musí především přesně zprostředkovat data. Nesmí zavádět ani zkreslovat. Pokud je jedno číslo dvakrát větší než druhé, ale ve vizualizaci vypadají přibližně stejně, pak je vizualizace špatná. Zároveň by vizualizace dat měla být esteticky příjemná. Dobré vizuální prezentace obvykle umocňují sdělení vizualizace. Pokud obrázek obsahuje rušivé barvy, nevyvážené vizuální prvky nebo jiné prvky, které odvádějí pozornost, pak bude pro diváka obtížnější obrázek prohlédnout a správně interpretovat." [26]

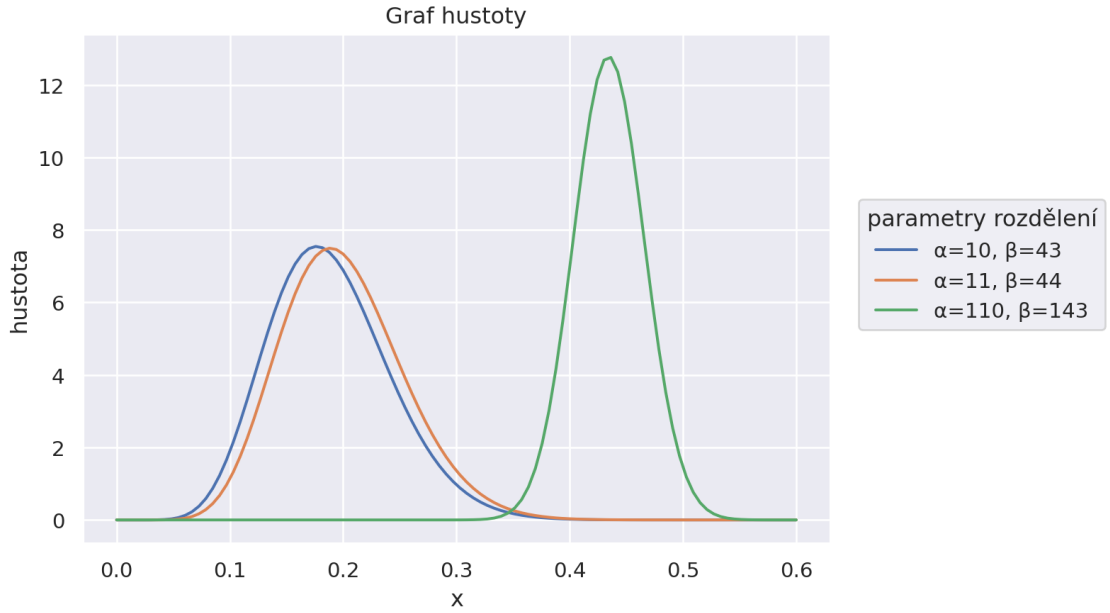
Pro jednotlivé datové typy budou podrobněji prozkoumány možnosti vizualizace. Pro každý datový typ se může hodit jiný typ grafu. Zkoumání vhodných grafů bude zaměřeno na následující veličiny: *číselné*, *kategoriální* a *textové*.

2.4.1 Číselné veličiny

Standardem pro vizualizaci distribuce číselných veličin je histogram. Histogram se špatně zvoleným počtem sloupců může data zkreslit. Pokud je zvolen nízký počet sloupců, může se z dat ztratit informace. Naopak, pokud je počet sloupců příliš vysoký, histogram bude zašumělý a bude těžké najít trend. Na obrázku 2.3 jsou zobrazeny 4 vizualizace stejných dat s různými počty sloupců. V prvním grafu (s nejvíce sloupci) je spoustu šumu a v posledním grafu (s nejméně sloupci) se ztratila informace o vyšším počtu lidí s věkem blížícím se k nule. [26]

V případě numerické proměnné s binární klasifikací je potřeba vykreslit dvě distribuce. Zvlášť pro data s kladným výsledkem a zvlášť pro data se záporným výsledkem cílové proměnné. Nabízí se několik variant grafu. První z možností je vykreslit sloupce pro jednu skupinu a nad nimi sloupce pro druhou skupinu s různými barvami (viz obrázek 2.4). U tohoto způsobu je přehledně vidět pouze distribuce první skupiny. Distribuce druhé skupiny je z grafu obtížně rozpoznatelná. Problém je také s porovnáním histogramů. Je velmi obtížné porovnat histogramy, které jsou postaveny „na sobě". Další problém tohoto grafu je, že není jasné, kde začínají sloupce histogramu druhé skupiny. Mohou začínat na ose x, nebo na místě, kde končí histogram první skupiny.

Druhá možnost je dát histogramy přes sebe, aby oba začínali od nuly na ose x a vyřešit tím



Obrázek 2.2 Beta rozdělení pro vyhlazená data.

problém porovnatelnosti z předchozího způsobu vizualizace. Upravením transparentnosti jsou v grafu 2.5 vidět oba histogramy. Problém tohoto způsobu je, že v grafu vzniká více než dvě barvy. Na první pohled není jisté, zda se v grafu vyskytují pouze dva histogramy, nebo více.

Třetí možností je použít histogram, který zobrazí sloupce jednotlivých distribucí vedle sebe (viz obrázek 2.6). Problém tohoto grafu spočívá v posunutí histogramu první skupiny mírně doleva a histogramu druhé skupiny mírně doprava. Pro nižší počet sloupců toto není problém, s vyšším počtem sloupců se však graf stává nepřehledný.

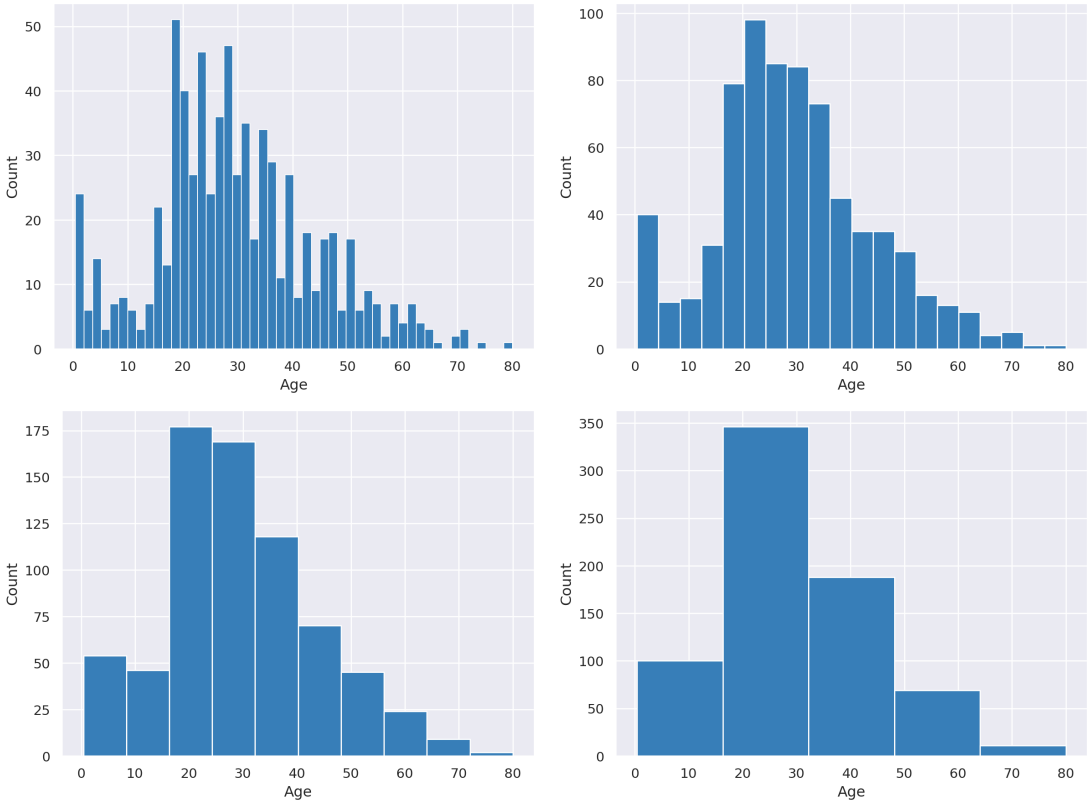
Další možnost je použít motýlí graf, kterému se také říká pyramidový graf. Tento graf vykreslí dva histogramy proti sobě v opačném směru (viz obrázek 2.7).

2.4.2 Kategoriální veličiny

Pro vizualizaci kategoriálních veličin se nejčastěji používá barplot. Pro odlišení číselných a kategoriálních vizualizací budou u kategoriálních vizualizací prohozeny osy x a y. Ukázka barplotu se nachází na obrázku 2.8. Graf je dobře čitelný a data pro kladnou cílovou proměnnou jsou dobře porovnatelné s daty pro zápornou cílovou proměnnou.

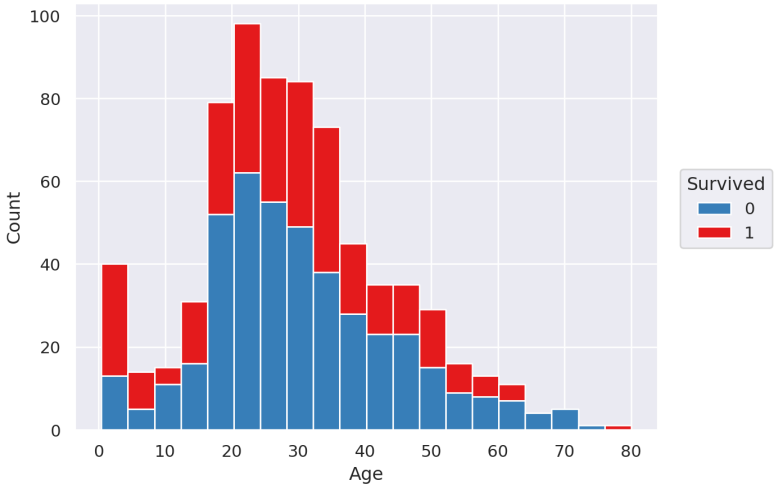
2.4.3 Textové veličiny

Pro vizualizaci textových veličin se nehodí žádný z předchozích grafů. Při použití kategoriální vizualizace by vzniklo spoustu kategorií s nízkým počtem výskytů. Jako ideální možnost pro vizualizaci textových dat se nabízí wordcloud. Wordcloud je typ grafu, ve kterém jsou slova zobrazena s velikostí dle počtu výskytů. Slova s vysokým počtem výskytů jsou zobrazena větším fontem, slova s nižším počtem výskytů menším fontem. Vizualizovat wordcloud s binární klasifikací je možné pomocí barvy. Každé slovo je obarveno dle počtu výskytů v jednotlivých skupinách. Například slovo, které se vyskytuje pouze v řádcích s pozitivní vysvětlovanou proměnou se zobrazí stejnou barvou, kterou je reprezentován pozitivní výsledek cílové proměnné. Slova s výskytem v obou skupinách vysvětlované proměnné jsou obarveny pomocí barevného gradi-

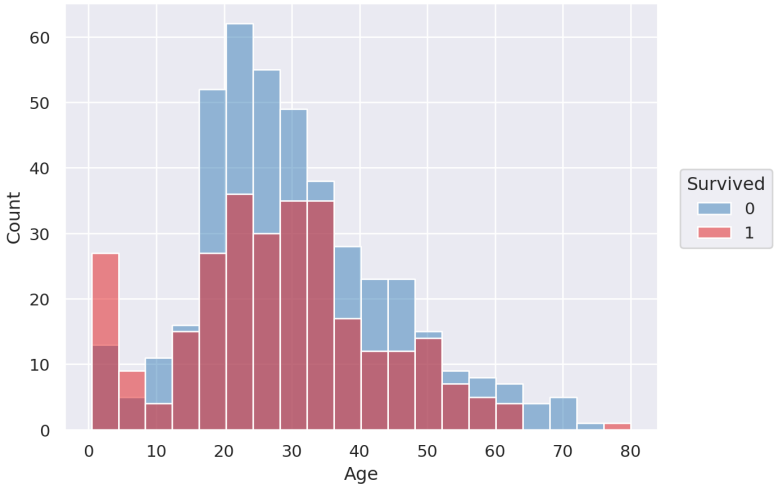


Obrázek 2.3 Histogramy s různým počtem sloupců. Všechny histogramy jsou vizualizace stejných dat.

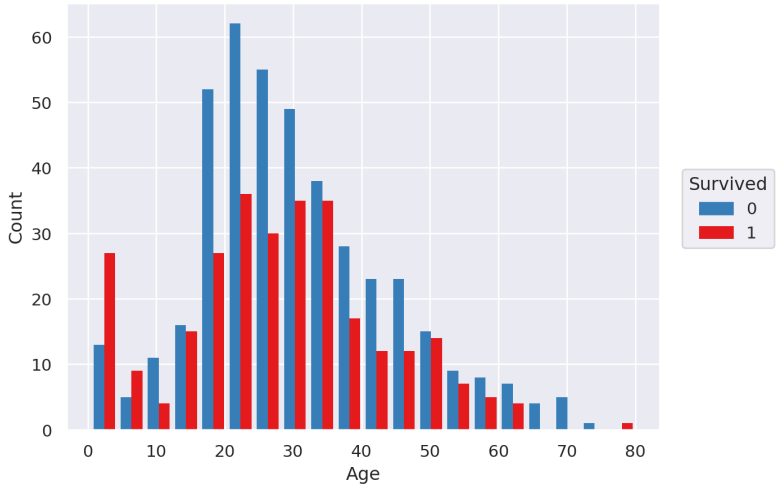
entu mezi barvami, které reprezentují pozitivní a negativní výsledny cílové proměnné. Ukázka wordcloudu se nachází na obrázku 2.9.



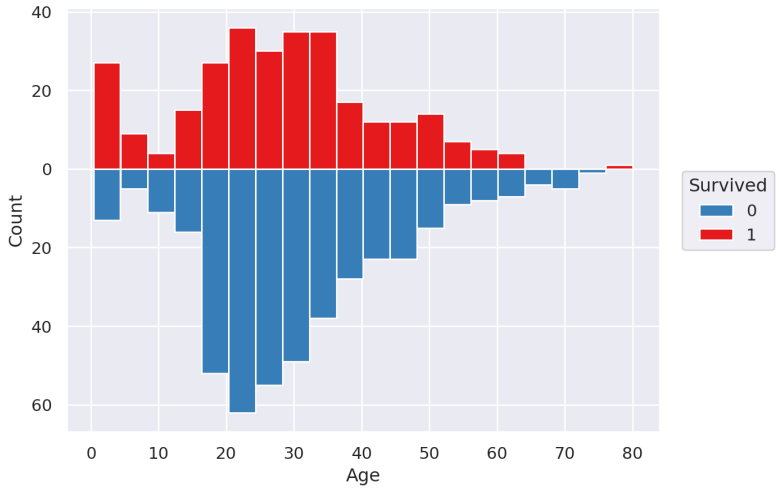
Obrázek 2.4 Histogram věku cestujících na Titaniku dle přežití. Histogramy se nepřekrývají a červený histogram je zobrazen nad modrým histogramem.



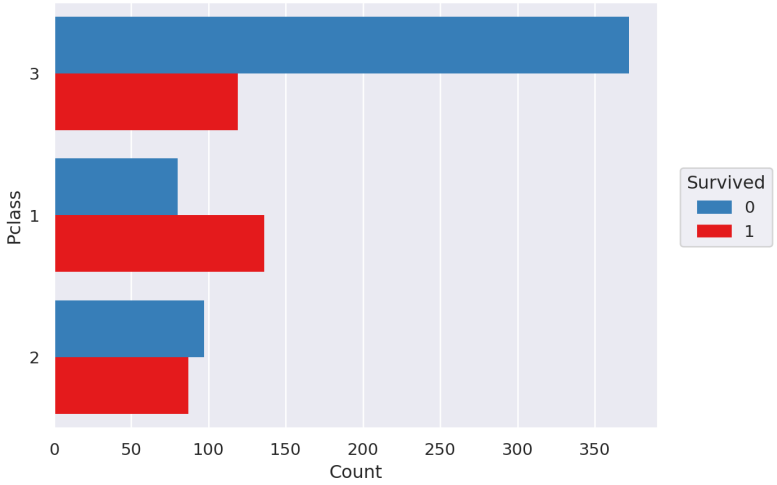
Obrázek 2.5 Histogram věku cestujících na Titaniku dle přežití. Histogramy se překrývají.



Obrázek 2.6 Histogram věku cestujících na Titaniku dle přežití. Histogramy jsou zobrazeny vedle sebe.



Obrázek 2.7 Histogram věku cestujících na Titaniku dle přežití. Motýlí graf.



Obrázek 2.8 Barplot zakoupené třídy na Titaniku dle přežití.

Obrázek 2.9 Wordcloud proměnné jméno z datové sady Titanik. Barva je zvolena dle přežití.

Kapitola 3
Návrh úprav knihovny

Tato kapitola je věnována návrhu úprav knihovny Pandas Profiling pro zlepšení automatické explorace dat při binární klasifikaci.

3.1 Datové typy

Knihovna *Pandas Profiling* podporovala v době psaní práce následující datové typy:

Číselné: celočíselné a desetinné veličiny

Kategoriální: kategoriální a textové veličiny

Datum a čas: data ve formátu datetime

Binární: binární veličiny

Při vytváření reportu je možné navíc zapnout rozšířené datové typy, které se v základním nastavení řadí do kategoriálních proměnných. Mezi tyto typy patří:

Cesta: veličiny obsahující cesty k souborům

Soubor: veličiny obsahující soubory

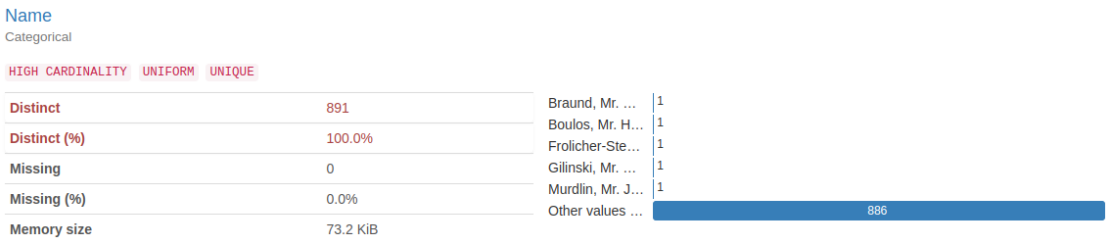
Obrázek: veličiny obsahující obrázky

URL: veličiny obsahující Uniform Resource Locator (URL) adresy

Typ pro cesty k souborům přináší navíc informace o souborových typech a složkách, typ pro soubory přidává navíc informace o velikosti souborů a datum vytvoření a upravení souborů, a typ pro obrázky přidává statistiky pro rozměry obrázků. Z hlediska testování závislosti mezi rozšířenými datovými typy a cílovou proměnou je možné použít stejný přístup jako u kategoriálních veličin, proto jim v této práci nebude věnována speciální pozornost. Práce se také nebude explicitně zabývat binárními veličinami, jedná se totiž o speciální případ kategorických veličin.

Použití kategoriálního datového typu pro textové veličiny (jako je například jméno a příjmení) není nejvhodnější. U takových dat je velmi pravděpodobné, že budou obsahovat spoustu unikátních hodnot. Vizualizace textové veličiny, klasifikované jako kategoriální, je zobrazena na obrázku 3.1. Tato vizualizace nepřináší žádnou informaci.

Ze sloupce se jménem by bylo možné získat více informací. Například počet lidí se stejným křestním jménem, nebo příjmením. Z tohoto důvodu bude přidán nový datový typ pro textové veličiny.



Obrázek 3.1 Textová veličina klasifikována jako kategoriální veličina.

Vytvořením nového datového typu pro textové veličiny vzniká problém s vizualizací. Mezi obvyklé vizualizace rozdělení patří barplot a histogram. Ani jeden z těchto přístupů však není vhodný pro textové veličiny, neboť se v datech nachází velké množství málo zastoupených kategorií. Pro vizualizaci textových veličin se nabízí wordcloud, navržený v kapitole 2.4.3, s jehož pomocí lze zobrazit zastoupení jednotlivých slov v datech. Při binární klasifikaci by pro uživatele mohla být důležitá informace o výskytu jednotlivých slov dle cílové proměnné, čehož je možné docílit pomocí barvy slov.

Vztah datového typu pro textové veličiny a datového typu pro kategoriální veličiny bude implementován pomocí dvou limitních hodnot, dle kterých bude rozhodnuto, zda se jedná o textovou, či kategoriální veličinu. První limitní hodnota bude udávat maximální absolutní počet unikátních hodnot ve sloupci a druhá limitní hodnota bude reprezentovat relativní maximální počet unikátních hodnot. Pokud nebude ani jedna limitní hodnota překročena, budou data klasifikována jako kategoriální. V základním nastavení bude první limitní hodnota nastavena na 50 (aby byla textová proměnná převedena na kategoriální, nesmí obsahovat více než 50 unikátních hodnot) a druhá na 0,5 (veličina nesmí obsahovat více než polovinu unikátních hodnot). Nastavení těchto limitních hodnot bude možné změnit v konfiguračním souboru.

3.2 Uživatelské rozhraní knihovny

V knihovně *Pandas Profiling* jsou implementovány dva přístupy použití. První přístup použití knihovny je přímo pomocí programovacího jazyku *Python*, pomocí třídy `ProfileReport()`. Druhý přístup je pomocí příkazové řádky Command-line interface (CLI), který na pozadí používá již zmíněnou třídu `ProfileReport()`. Konstruktor třídy `ProfileReport()` obsahuje celou řadu nepovinných parametrů. Mezi důležité parametry patří `config_file`, pomocí kterého lze specifikovat cestu ke konfiguračnímu souboru. Konfigurační soubor obsahuje veškeré nastavení reportu, s jehož pomocí je možné přizpůsobit report dle potřeb uživatele. Parametr `minimal` umožňuje nastavit konfiguraci reportu tak, aby bylo možné pracovat s velkými datovými sadami tím, že vypne některé výpočetně náročné operace.

Ukázka použití knihovny *Pandas Profiling* přímo pomocí programovacího jazyku *Python* se nachází v ukázkovém kódu 1, kde je znázorněno, jak vytvořit HTML report.

```
import pandas as pd
from pandas_profiling import ProfileReport

df = pd.read_csv('data.csv')
profile = ProfileReport(df, title='Title of report')
profile.to_file('out_file_name.html')
```

Výpis kódu 1 Použití knihovny *Pandas Profiling* pomocí jazyka *Python*.

Druhý přístup je pomocí příkazové řádky CLI. Příklad použití příkazové řádky je znázorněn v kódu 2. Příkaz má jeden povinný parametr `input_file`, který představuje cestu k souboru s daty ve formátu Comma-separated values (CSV). Nepovinné parametry jsou vybrané parametry z konstruktoru `ProfileReport()` nebo z konfiguračního souboru. Za zmínku stojí např. `--infer_dtypes`, pomocí kterého je možné zapnout automatickou klasifikaci datových typů sloupců.

```
pandas_profiling [-h] [--version] [-s] [-m] [-e] [--pool_size POOL_SIZE]
↪  [--title TITLE] [--infer_dtypes] [--no-infer_dtypes] [--config_file
↪  CONFIG_FILE] input_file [output_file]
```

Výpis kódu 2 Použití knihovny *Pandas Profiling* pomocí příkazové řádky.

Do konstruktoru `ProfileReport()` budou přidány dva nové parametry: `target_col` pro volbu sloupce s cílovou proměnnou a `target_positive_values` pro volbu pozitivních hodnot v cílové proměnné. V případě vynechání druhého parametru budou pozitivní hodnoty odhadnuty výchozími pozitivními hodnotami z konfiguračního souboru. Do CLI rozhraní příkazové řádky budou přidány dva nepovinné parametry: `--target_col` a `--target_positive_values`, které reprezentují výše zmíněné parametry v `ProfileReport()` konstruktoru.

Konfigurační soubor vytvářeného reportu je ve formátu Yet Another Markup Language (YAML) a obsahuje mnoho nastavení, která ovlivňují výsledný report.

3.3 Popisné statistiky a grafy

Jedním z hlavních cílů práce je nalezení závislostí mezi jednotlivými proměnnými a cílovou proměnnou a výsledky přehledně reprezentovat. Hlavními nástroji pro efektivní reprezentaci výsledků budou statistické testy, vhodně zvolené vizualizace a příslušné popisné statistiky. Praktickým cílem práce je vytvořit rozšíření do knihovny *Pandas Profiling*, je tedy důležité, aby se jednalo o standardní statistické testy a standardní způsoby vizualizací pro jednoduchost použití novými uživateli knihovny *Pandas Profiling*. Je též třeba, aby vizualizace podporovaly zobrazení dvou populací dle cílové proměnné a aby byly tyto dvě populace jednoduše porovnatelné. Budou přiřazeny barvy pro jednotlivé kategorie binární cílové proměnné. Tyto barvy budou použity při vizualizaci subpopulací jednotlivých proměnných dle cílové proměnné. Barvy lze upravit v konfiguračním souboru.

3.3.1 Číselné veličiny

Pro testování závislosti číselných veličin na cílové proměnné byl vybrán dvouvýběrový studentův t-test, který byl představen v kapitole 2.1.1. Studentův t-test byl vybrán proto, že se jedná o všeobecně známý statistický test a jeho výsledky jsou jednoduše interpretovatelné. V základním nastavení bude test proveden na hladině významnosti 5 %, avšak toto nastavení lze změnit v konfiguračním souboru. Nastavení hladiny významnosti se vtahuje také na kategoriální veličiny. Dojde-li k zamítnutí nulové hypotézy (tj. střední hodnoty se napříč populacemi liší), bude tato informace reprezentována v sekci *Upozornění*. Informace o upozorněních se nacházejí v kapitole 3.6. Uživatel tak bude upozorněn na proměnné, ve kterých se subpopulace dle cílové proměnné významně liší, což může urychlit práci s velkými datovými sadami s vysokým počtem sloupců.

Bude také upravena vizualizace dat tak, aby bylo jednoduché rozpoznat rozdíly v jednotlivých subpopulacích. V době psaní práce byl pro vizualizaci použit histogram, pomocí kterého lze zobrazit pouze jednu distribuci. K vizualizaci dat dle cílové proměnné bude použit histogram, který zobrazí sloupce jednotlivých distribucí vedle sebe. Histogram byl představený v kapitole 2.4.1.

Tento typ vizualizace byl vybrán z důvodu, že se jedná o běžné zobrazení dvou histogramů v jednom grafu a je jednoduché porovnat mezi sebou zobrazené grafy.

3.3.2 Kategoriální veličiny

Pro testování závislosti kategoriálních veličin na cílové proměnné byl vybrán chí-kvadrát test nezávislosti představený v kapitole 2.1.2. Jedná se o standard pro testování nezávislosti dvou kategoriálních veličin. Stejně jako v případě číselných veličin, dojde-li k zamítnutí nulové hypotézy (tj. pokud na zvolené hladině významnosti bude zamítnuta hypotéza nezávislosti), bude tato informace reprezentována v sekci *Upozornění*, viz kapitola 3.6.

Dále bude změněna vizualizace kategoriálních veličin. V době psaní této práce byla pro vizualizaci kategoriálních veličin použita tabulka počtů obsahující barplot, která byla implementována přímo v knihovně *Pandas Profiling*. Z důvodu náročnosti upravení stávajícího přístupu tak, aby bylo možné zobrazit dvě subpopulace, byla zvolena nová vizualizace pomocí barplotu, který byl představen v kapitole 2.4.2. Jedná se o stejný způsob vizualizace jako u číselných veličin, s tím rozdílem, že osy x a y budou pro snadné rozpoznání datových typů na základě vizualizace prohozeny.

3.3.3 Textové veličiny

Textové veličiny byly v době psaní práce klasifikovány jako kategoriální veličiny, a proto budou obsahovat většinu popisných statistik z datového typu pro kategoriální veličiny. U textových veličin nebude použit žádný test pro testování závislosti na cílové proměnné. V textových veličinách bývají zpravidla všechny hodnoty unikátní, proto použití chí-kvadrát testu nedává smysl.

K vizualizaci textových veličin bude použit wordcloud, který byl představen v kapitole 2.4.3. Text bude převeden na malá písmena, rozdělen na jednotlivá slova a následně bude vytvořena tabulka četností. Díky této vizualizaci bude možné zjistit, jaká slova se v textových datech vyskytují nejčastěji. Dle barvy slov bude také možné určit, se kterou hodnotou cílové proměnné jsou slova více spjatá. Barva bude získána pomocí barevného gradientu mezi barvami přiřazeným kategoriím cílové proměnné. Pro data zobrazená v tabulce 3.1 bude výsledná tabulka četností vypadat následovně (viz tabulka 3.2) a vizualizace pomocí wordcloud je zobrazena na obrázku 3.2.

Tabulka 3.1 Ukázková data pro wordcloud.

Data	Cílová proměnná
Ahoj, ahoj	0
Ahoj světe Ahoj světe	1
Ne	0

Tabulka 3.2 Tabulka četností ukázkových dat pro wordcloud.

Slovo	Počet pozitivních výskytů	Počet negativních výskytů	Barva
ahoj	2	2	0,5
světe	2	0	1
ne	0	1	0

Obrázek 3.2 Wordcloud pro ukázkovou tabulku četností 3.2.

3.4 Identifikace subpopulací s odlišnou proporcí skupin v cílové proměnné

U každé veličiny by bylo vhodné nalézt subpopulace, které se významně liší v zastoupení kategorií cílové veličiny oproti celé populaci. Pro identifikaci takových subpopulací bude použit logaritmický *poměr šancí* se základem dva. *Poměr šancí* byl představen v kapitole 2.2.2. Před výpočtem *poměru šancí* bude provedeno vyhlazení dat. Důvodem pro použití vyhlazení dat je potlačení extrémních hodnot u kategorií s nízkým výskytem. K vyhlazení dat bude použita upravená verze *beta vyhlazování* z kapitoly 2.3.2, kde za prior bude zvolena průměrná hodnota v populaci. Získané hodnoty pomocí logaritmického *poměru šancí* pak budou po vyhlazení mírně posunuty k průměrné hodnotě v populaci. Ukázková data pro tuto kapitolu jsou sloupce pohlaví a přežití v ukázkové datové sadě Titanik (viz tabulka 3.3).

Tabulka 3.3 Kontingenční tabulka pohlaví a přežití v ukázkové datové sadě Titanik.

pohlaví/přežití	Ano	Ne
Muž	109	468
Žena	233	81
celkem	342	549

3.4.1 Použití logaritmu poměru šancí

Poměr šancí je poměr šance výskytu pozitivních pozorování v jedné skupině proti šanci výskytu pozitivních pozorování v druhé skupině. Jako první skupina bude zvolena subpopulace dle analyzované veličiny (v ukázce se jedná o muže, nebo ženy), jako druhá skupina bude zvolena celá populace dané veličiny (v případě ukázky muži a ženy dohromady). Tímto přístupem bude docíleno porovnání jednotlivých subpopulací s průměrem populace dané veličiny.

Příklad použití na datech z tabulky 3.3:

$$\mathrm{OR}_m = \frac{\mathrm{odds}_m}{\mathrm{odds}_{all}} = \frac{\frac{109}{468}}{\frac{342}{549}} = \frac{109 \cdot 549}{342 \cdot 468} = 0{,}37,$$

$$\mathrm{OR}_w = \frac{\mathrm{odds}_w}{\mathrm{odds}_{all}} = \frac{\frac{233}{81}}{\frac{342}{549}} = \frac{233 \cdot 549}{342 \cdot 81} = 4{,}62,$$

kde odds_m je šance přežití u mužů, odds_w je šance přežití u žen, odds_{all} je šance přežití u všech osob, OR_m je poměr šancí pro muže a OR_w je poměr šancí pro ženy. Z těchto výsledků vyplývá, že ženy mají téměř pětkrát vyšší šanci na přežití, než je průměrná šance na přežití v populaci, zatímco muži mají skoro 0,4 krát nižší šanci na přežití, než je průměrná šance na přežití

v populaci. Nad *poměry šancí* je následně provedena logaritmizace, pro normalizaci rozdělení *poměru šancí* (jak bylo představeno v kapitole 2.2.2). Logaritmus *poměru šancí* u mužů vyjde $\log_2(OR_m) = \log_2(0{,}37) = -1{,}43$ a logaritmus *poměru šancí* u žen vyjde $\log_2(OR_m) = \log_2(4{,}62) = 2{,}21$.

3.4.2 Vyhlazení dat

Jak již bylo uvedeno výše, k vyhlazování dat bude použita upravená verze *beta vyhlazování*. Úprava spočívá ve výpočtu parametrů α a β. Parametry α a β budou nastaveny tak, aby platila následující soustava rovnic:

$$\frac{\alpha}{\alpha + \beta} = P(Pos),$$
$$\alpha + \beta = c,$$

kde $P(Pos)$ je pravděpodobnost pozitivního výsledku v binární klasifikaci nad celou populací (prior populace) a c znázorňuje parametr, který je nastaven uživatelem v konfiguračním souboru (výchozí hodnota tohoto parametru je 20). Parametr c reprezentuje počet přidaných umělých pozorování do každé kategorie v populaci.

Příklad použití *beta vyhlazení* na datech z tabulky 3.3 s parametrem $c = 20$:

$$\frac{342}{342 + 549} \approx 0{,}3838,$$
$$\alpha + \beta = 20.$$

Parametry α a β budou vypočítány jako $\alpha = 7{,}676$ a $\beta = 12{,}324$. Do každé subpopulace bude přidáno α pozitivních a β negativních umělých pozorování. To se na logaritmu *poměru šancí* projeví tak, že výsledné hodnoty budou mírně posunuty směrem k nule. Nová kontingenční tabulka po použití *beta vyhlazení* bude vypadat následovně 3.4.

Tabulka 3.4 Kontingenční tabulka pohlaví a přežití v ukázkové datové sadě Titanik s použitím *beta vyhlazení*.

pohlaví/přežití	Ano	Ne
Muž	116,676	480,324
Žena	240,676	93,324

Pokud hodnota logaritmu *poměru šancí* překročí prahovou hodnotu (která je definována v konfiguračním souboru reportu), bude do reportu přidáno upozornění na vychýlenou kategorii v analyzované proměnné. Pokud se v proměnné nachází více kategorií, které překročují prahovou hodnotu, je vygenerováno jedno upozornění pro každou z kategorií. Podrobnější informace o upozorněních se nacházejí v kapitole 3.6.

3.5 Chybějící hodnoty

Do kapitoly chybějící hodnoty v reportu bude přidána záložka pro kontrolu závislosti chybějících hodnot na cílové proměnné. Pro všechny veličiny s chybějícími hodnotami bude vykreslena matice výskytů. V řádcích bude zobrazena cílová proměnná (0/1) a ve sloupcích příznak, zda se jedná o chybějící hodnotu (Ne/Ano). Nad touto maticí bude proveden test nezávislosti. Byl zvolen chí-kvadrát test nezávislosti představený v kapitole 2.1.2. Chí-kvadrát test nezávislosti byl opět zvolen i pro jeho všeobecnou známost. Hladinu významnosti testu bude možné nastavit v konfiguračním souboru reportu (výchozí hodnota bude nastavena na 5 %). Dojde-li k zamítnutí nulové hypotézy (tj. pokud na zvolené hladině významnosti bude zamítnuta hypotéza nezávislosti), bude tato informace reprezentována v sekci *Upozornění* (viz kapitola 3.6).

3.6 Upozornění

Sekce upozornění je důležitou součástí vygenerovaného reportu. Nachází se zde přehled veličin, které jsou nějakým způsobem zajímavé. Knihovna v době psaní práce podporovala následující upozornění:

- CONSTANT: veličiny, u kterých se vyskytuje pouze jedna unikátní hodnota
- DUPLICATES: veličiny obsahující větší výskyt duplicitních hodnot
- EMPTY: veličiny, které neobsahují žádné hodnoty (obsahují pouze chybějící hodnoty)
- HIGH CARDINALITY: kategoriální veličiny s vysokou kardinalitou
- HIGH CORRELATION: veličiny, které korelují s jinými proměnnými; v rámci upozornění jsou vypsány názvy korelovaných proměnných
- INFINITE: číselné veličiny obsahující hodnoty pro nekonečno
- MISSING: veličiny obsahující větší výskyt chybějících hodnot
- REJECTED: veličiny datového typu, který knihovna neumí zpracovat
- SKEWED: číselné veličiny, které jsou zešikmené
- UNIFORM: veličiny s uniformním rozdělením
- UNIQUE: veličiny se všemi unikátními hodnotami
- UNSUPPORTED: veličiny datového typu, který knihovna neumí zpracovat
- ZEROS: veličiny obsahující větší množství nulových hodnot

K již existujícím upozorněním budou přidány upozornění pro implementované statistické testy, detekce subpopulací, které se významně liší v zastoupení kategorií cílové veličiny oproti celé populaci, a pro testování nezávislosti chybějících hodnot a cílové proměnné. Budou přidány následující upozornění:

- DEPENDENT MEAN: číselné veličiny s výrazně odlišnými průměry dle cílové proměnné
- DEPENDENT CATEGORIES: kategoriální veličiny s výrazně odlišným zastoupením dle cílové proměnné
- LOW LOG ODDS RATIO: veličiny, ve kterých se vyskytuje kategorie s významně nižší *poměrem šancí* oproti celé populaci
- HIGH LOG ODDS RATIO: veličiny, ve kterých se vyskytuje kategorie s významně vyšším *poměrem šancí* oproti celé populaci
- MISSING CORRELATED WITH TARGET: veličiny, u kterých test nezávislosti chybějících pozorování a cílové veličiny zamítl nulovou hypotézu

V konfiguračním souboru reportu bude možné nastavit prahové hodnoty pro jednotlivá upozornění. Ve výchozím nastavení budou všechny statistické testy provedeny na hladině důvěryhodnosti 95 %.

3.7 Doporučené transformace

Do reportu bude přidána sekce pro doporučené transformace veličin. Budou definované transformace pro číselné, kategoriální a textové veličiny. Cílem těchto transformací je zlepšit kvalitu dat pro trénování klasifikačního modelu nad cílovou proměnnou.

3.7.1 Číselné veličiny

Pro číselné veličiny budou implementovány tři transformace: *normalizace*, *binning* a *logaritmizace*.

Normalizace má více variant. Mezi nejznámější patří *min-max normalizace*, která od všech dat odečte minimum a vydělí rozdílem maxima a minima. Tím se dostanou všechny hodnoty do intervalu [0,1]. Další ze známých normalizačních metod je *z-skóre normalizace* (též známá pod názvem *standardizace*), která od dat odečte průměrnou hodnotu a vydělí rozptylem. Výsledná veličina bude mít průměrnou hodnotou rovnou 0 a rozptyl rovný 1. Pro implementaci byla vybrána normalizační metoda *z-skóre*.

Binning je metoda pro transformaci spojitých veličin na diskrétní veličiny. *Binning* má dvě varianty: binning stejné šířky (data jsou rozděleny do stejně širokých sloupců) a binning stejné výšky (v každém sloupci se vyskytuje stejný počet hodnot). Binning stejné výšky má obvykle lepší výsledky při použití v regresních modelech a umí se lépe vypořádat s odlehlými hodnotami. Z tohoto důvodu byl vybrán binning stejné výšky. Počet sloupců bude možné změnit v konfiguračním souboru (v základním nastavení bude nastaven na 5).

Logaritmizace může pomoci při trénování modelu, pokud jsou data hodně zešikmená.

3.7.2 Kategoriálníké veličiny

Pro kategoriální veličiny bude implementována pouze jedna transformace a to *one-hot encoding*. *One-hot encoding* je transformace, která vytvoří příznakový sloupec pro každou kategorii v datech. Vytvoří tedy tolik binárních sloupců, kolik je v transformovaném sloupci kategorií.

3.7.3 Textové veličiny

Textové veličiny jsou samy o sobě pro model nepoužitelné. Jako vhodná transformace pro textové veličiny se nabízí metoda Term frequency–inverse document frequency (TF-IDF). TF-IDF se skládá ze dvou částí: *frekvence termínů* Term frequency (TF) a *inverzní frekvence dokumentů* Inverse document frequency (IDF). *Frekvence termínů* TF $\mathrm{tf}(t,d)$ udává relativní četnost termínu t v dokumentu d. *Frekvence termínů* může být definována více způsoby, jeden ze způsobů je

$$\mathrm{tf}(t,d) = \frac{f_{(t,d)}}{\sum_{\forall t' \in d} f(t',d)},$$

kde $f_{(t,d)}$ je počet výskytů termínu t v dokumentu d a $\sum_{\forall t' \in d} f(t',d)$ je součet výskytů všech termínů v dokumentu d. *Inverzní frekvence dokumentů* IDF $\mathrm{idf}(t)$ udává inverzní hodnotu frekvence dokumentu, která měří informační hodnotu termínu t. Jedna z možností pro výpočet inverzní frekvence dokumentu je:

$$\mathrm{idf}(t) = \log(\frac{N}{\mathrm{df}(t)}),$$

kde N je počet dokumentů v souboru dokumentů D a $\mathrm{df}(t)$ je četnost dokumentů d (počet dokumentů d v souboru dokumentů D), ve kterých se vyskytuje termín t. TF-IDF je součin *frekvence termínů* a *inverzní frekvence dokumentů*. [27]

TF-IDF transformace bude použita na tabulce četností slov z kapitoly 3.3.3. Po provedení transformace bude vybráno n nejčetnějších slov (tento limit lze změnit v konfiguračním souboru a v základním nastavení bude nastaven na 50) a ty budou zahrnuty do transformace. Z jednoho sloupce tedy transformací vznikne až n nových číselných sloupců.

3.7.4 Vyhodnocení transformací

Přínosnost transformace bude vyhodnocena přes schopnost lépe predikovat cílovou proměnnou. Budou tedy porovnány výkonnosti modelu s transformovaným sloupcem (jako náhrada původního sloupce) s referenčním modelem, který je představen v kapitole 3.8. Pro porovnání modelů bude použita jedna z následujících metrik: *accuracy*, *precision*, *recall*, nebo *f1-score*. Metriku bude možné zvolit v konfiguračním souboru (v základním nastavení bude zvolena metrika *accuracy*).

Nejprve bude datová sada rozdělena na trénovací a testovací sadu. Transformace bude vždy natrénována pouze na trénovací sadě a následně použita pro transformaci obou datových sad. Až poté je možné trénovat klasifikační model. U navržených transformací bude zobrazen popis transformace, nastavení rozdělení trénovací a testovací sady, nastavení modelu a metriky ohodnocení klasifikačního modelu. Modul transformací dat bude ve výchozím nastavení vypnutý a bude jej možné zapnout pomocí konfiguračního souboru.

3.8 Referenční model

Do reportu bude přidán nový modul pro klasifikační model. Tento modul bude obsahovat referenční klasifikační model natrénovaný nad originálními daty a pokud bude zapnutý modul transformací, bude obsahovat i model natrénovaný nad daty s doporučenými transformacemi. Jako model bude použit *rozhodovací strom s gradientním boostingem*. Jedním z důvodů této volby je robustnost modelu, díky které se model umí vypořádat s různorodými daty. U modelu budou zobrazeny základní metriky: *accuracy*, *precision*, *recall* a *f1-score*. Dále bude zobrazeno nastavení modelu, poměr rozdělení testovacích a trénovacích dat a důležitost jednotlivých veličin pro model. Tento modul bude ve výchozím nastavení vypnutý a bude jej možné zapnout pomocí konfiguračního souboru. Dále bude možné pomocí konfiguračního souboru měnit vybrané hyperparametry modelu.

Kapitola 4
Implementace

Tato kapitola je zaměřena na implementaci rozšíření do knihovny Pandas Profiling. Nachází se zde implementace nového datového typu pro textové veličiny, rozšíření uživatelského rozhraní o definici cílové proměnné, implementace statistických testů k veličinám dle datového typu, úprava vizualizací veličin, implementace testu závislosti chybějících hodnot na cílové proměnné, přidání transformací a implementace klasifikačního modelu.

Při implementaci byl použit programovací jazyk *Python* [2]. Pro práci s daty byla použita knihovna *Pandas* [5]. Základní Statistické testy jsou provedeny s pomocí knihovny *scipy* [28]. Pro vizualizaci dat byla použita knihovna *seaborn* [29].

V implementaci knihovny *Pandas Profiling* byly v době psaní práce hojně používány slovníky. Používání slovníků nepatří mezi nejlepší programovací praktiky. Pro usnadnění práce s knihovnou byly některé slovníky odstraněny a místo nich byly implementovány nové datové třídy `BaseDescription` a `BaseAnalysis`. Třída `BaseDescription` obsahuje všechna data, pro vytvoření reportu. Jedná se o popisné statistiky jednotlivých proměnných, upozornění, korelace, chybějící hodnoty atd. Třída `BaseAnalysis` obsahuje data o reportu samotném, jako je název a datum vytvoření reportu.

4.1 Datové typy

Pro automatické odvozování datových typů je v knihovně *Pandas Profiling* použita knihovna *visions* [30]. Tato knihovna umožňuje vytvářet vlastní datové typy a nastavovat vztahy mezi jednotlivými datovými typy. Vztahy mezi datovými typy tvoří Directed acyclic graph (DAG). Data jsou nejprve klasifikovány do jednoho z datových typů. Následně je prohledáván graf transformací mezi datovými typy, dokud je možné měnit datový typ dat. Datové typy vytvořené s pomocí knihovny *visions* dědí z abstraktní třídy `VisionsBaseType`. Třída má dvě virtuální metody: `contains_op()` a `get_relations()`. První metoda `contains_op()` obsahuje definici pro prvotní klasifikaci do tohoto datového typu. Metoda `get_relations()` definuje transformace mezi datovými typy. Obsahuje všechny typy, ze kterých je možné definovaný datový typ vytvořit.

S využitím této knihovny byl přidán nový datový typ pro textové veličiny (představený v kapitole 3.1). Ukázka třídy pro textové veličiny se nachází v kódu 3. Dle `contains_op()` jsou při prvotní klasifikaci do datového typu `Text` přiřazeny proměnné, obsahující pouze hodnoty typu *object*. Dle implementace metody `get_relations()` je možné převést proměnné typu `Unsupported` na datový typ `Text` bez nutnosti úprav.

```
class Text(visions.VisionsBaseType):
    @staticmethod
    def get_relations() -> Sequence[TypeRelation]:
        return [
            IdentityRelation(Unsupported),
        ]

    @staticmethod
    @multimethod
    @series_not_empty
    @series_handle_nulls
    def contains_op(series: pd.Series, state: dict) -> bool:
        return pdt.is_string_dtype(series) and series_is_string(series, state)
```

Výpis kódu 3 Vytvoření nového datového typu pomocí knihovny visions.

4.2 Uživatelské rozhraní knihovny

Binární klasifikace pracuje s binární cílovou proměnnou (cílová proměnná má pouze dvě validní hodnoty). Pro zjednodušení použití rozšíření je přidána definice pozitivních hodnot v cílové proměnné. Ostatní hodnoty jsou automaticky klasifikovány jako negativní. Díky tomuto zobecnění lze pracovat i s cílovou proměnnou, které má více než dvě různé hodnoty.

Do konstruktoru reportu `ProfileReport()` je přidán parametr `target_col` pro sloupec s cílovou proměnnou. Tento parametr je nepovinný a má datový typ `string`. Druhý přidaný parametr je `target_positive_values` pro volbu pozitivních hodnot v cílové proměnné, který je též nepovinný a může obsahovat jednu hodnotu typu `string`, nebo list hodnot typu `string`.

Uživatelské rozhraní pomocí CLI je implementováno s pomocí modulu *argparse* [31], který je součástí *standardních knihoven pro Python* [32]. Zde jsou přidány stejné parametry, jako v kounstruktoru třídy `ProfileReport()`. Ukázka přidání parametrů do CLI pomocí knihovny *argparse* se nachází v kódu 4. Oba parametry jsou typu *string* a jsou nepovinné.

4.2.1 Přidané konfigurace

Do konfiguračního souboru ve formátu YAML byla přidána nastavení, které je možné rozdělit do následujících částí: nastavení cílové proměnné, nastavení veličin, nastavení upozornění, nastavení reportu, nastavení transformací a nastavení výchozího modelu.

4.2.1.1 Cílová proměnná

Nastavení pro cílovou proměnnou obsahuje následující nastavení: název sloupce s cílovou proměnnou, pozitivní hodnoty v cílové proměnné a výchozí pozitivní hodnoty. Pokud nejsou pozitivní hodnoty specifikovány, jsou použity výchozí pozitivní hodnoty.

4.2.1.2 Veličiny

Pro textové veličiny bylo převzato nastavení z kategoriálních veličin. Obsahuje tedy přepínače, zda se v reportu mají zobrazit statistiky pro slova a statistiky pro jednotlivé znaky. Dále zde byly přidány dvě prahové hodnoty pro klasifikaci mezi textovými a kategoriálními veličinami. První prahová hodnota reprezentuje počet unikátních hodnot ve veličině (výchozí hodnota je nastavena na 50). Druhá hodnota reprezentuje počet unikátních hodnot podělený počtem všech hodnot

```
parser = argparse.ArgumentParser(
    description="Profile the variables in a CSV file and generate a HTML
    ↪  report."
)
parser.add_argument(
    "--target_col",
    type=str,
    default=None,
    help="Name of target column.",
)

parser.add_argument(
    "--target_positive_values",
    type=str,
    default=None,
    help="Positive values in target column.",
)
```

Výpis kódu 4 Přidání parametrů do uživatelského rozhraní příkazové řádky knihovny *Pandas Profiling*.

(výchozí hodnota je nastavena na 0,5). Pokud není ani jedna z prahových hodnot překročena, je veličina klasifikována jako kategoriální veličina.

U veličin bylo také přidáno nastavení parametru pro *beta vyhlazování* (viz kapitola 3.4.2), které je v základním nastavení nastaveno na 20.

4.2.1.3 Upozornění

V této sekci konfigurace se nacházejí prahové hodnoty pro upozornění. Nachází se zde nastavení hladiny důvěryhodnosti pro statistické testy (viz kapitola 3.3), nastavení hladiny důvěryhodnosti pro chí-kvadrát test nezávislosti chybějících hodnot na cílové proměnné (viz kapitola 3.5). Pokud je u testu na zvolené hladině zamítnuta nulová hypotéza o nezávislosti, je vytvořeno upozornění. Nakonec se zde nachází prahová hodnota pro logaritmus poměru šancí (viz kapitola 3.4). V případě, že u nějaké kategorie překročí absolutní hodnota logaritmu *poměru šancí* prahovou hodnotu, je vytvořeno upozornění.

4.2.1.4 Report

V této sekci konfiguračního souboru se nachází nastavení reportu. Bylo přidáno nastavení pro vypnutí či zapnutí modulu transformací (viz kapitola 3.7) a modulu referenčního klasifikačního modelu (viz kapitola 3.8). Dále byly přidány příznaky pro zobrazení grafu distribuce a grafu logaritmu *poměru šancí* v přehledové části každé proměnné. Poslední přidané nastavení se týká omezení počtu zobrazených veličin (ve výchozím nastavení je omezení vypnuto). U datových sad s vyšším počtem veličin je vytvořený HTML report příliš velký a není možné s ním jakkoli pracovat. Z tohoto důvodu by mohlo být žádoucí omezit počet zobrazených veličin.

4.2.1.5 Transformace

Sekce transformace v konfiguračním souboru obsahuje nastavení pro navržení transformací, které byly přestaveny v kapitole 3.7. Sekce obsahuje *seed* pro použité transformace, počet sloupců při použití *binningu* u číselných veličin a maximální počet termínů při TF-IDF transformaci u textových veličin.

4.2.1.6 Model

V této sekci se nacházejí konfigurace pro klasifikační model *rozhodovací strom s gradientním boostingem* (viz kapitola 3.8). Nastavení modelu obsahuje relativní velikost testovací množiny (výchozí nastavení je 0,25), podle které jsou data náhodně rozděleny na trénovací a testovací. Data jsou rozděleny pouze jednou. Dále se zde nachází *seed* hodnota pro klasifikační modely, metrika, pomocí které jsou modely vyhodnoceny a porovnávány, a hyperparametry klasifikačního modelu. Mezi ně patří maximální hloubku stromů (výchozí hodnota je 3), maximální počet stromů (výchozí hodnota je nastavena na 10) a maximální počet listů v jednom stromu (hodnota je nastavena na 10). Parametry byly zvoleny takto z důvodu rychlejšího výpočtu.

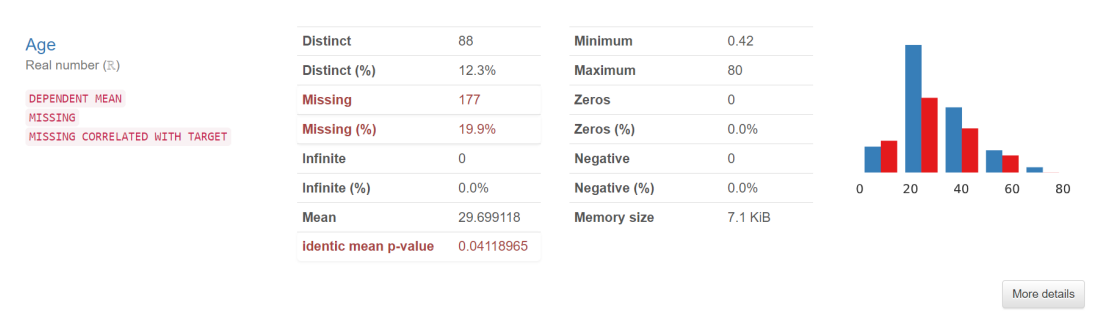
4.3 Popisné statistiky a grafy

U jednotlivých veličin byly implementovány statistické testy, byly upraveny vizualizace distribuce dat a byla přidána vizualizace logaritmu *poměru šancí*.

4.3.1 Číselné veličiny

U číselných veličin byl ke stávající implementaci popisných statistik přidán studentův dvouvýběrový t-test nezávislosti středních hodnot s rozdílnými rozptyly. Při implementaci byla použita funkce `stats.ttest_ind()` z knihovny *scipy*. Ukázka zobrazení numerické veličiny v reportu se nachází na obrázku 4.1. Do první tabulky byl přidán řádek s p-hodnotou pro testování nezávislosti průměrných hodnot dle cílové proměnné. V tomto případě jsou průměrné hodnoty značně odlišné, je tedy přidáno i upozornění na závislost průměrné hodnoty dle cílové proměnné.

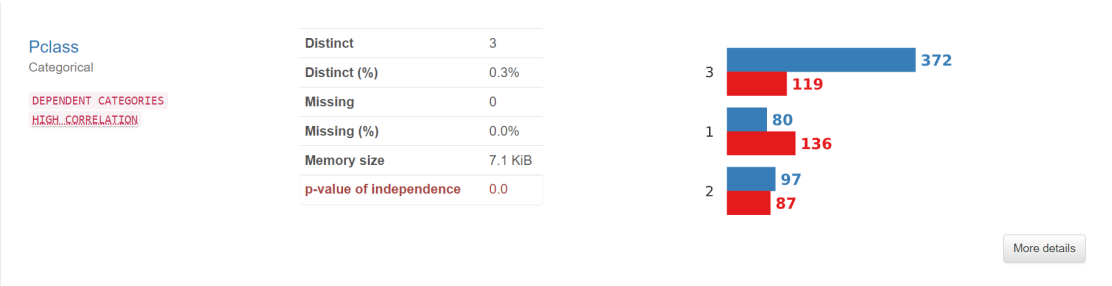
Dále byla upravena vizualizace distribuce dat. K vizualizaci byla použita knihovna *seaborn*.



Obrázek 4.1 Ukázka číselné veličiny z ukázkové datové sady Titanik v *Pandas Profiling* reportu po implementaci rozšíření.

4.3.2 Kategoriální veličiny

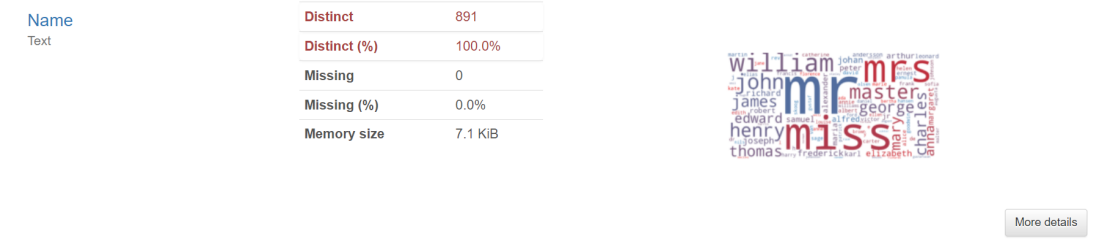
U kategoriálních veličin byl ke stávající implementaci popisných statistik přidán chí-kvadrát testu nezávislosti, který je implementován v knihovně *scipy* v modulu `stats.chi2_contingency()`. Pro implementaci vizualizace byla stejně jako u číselných veličin použita knihovna *seaborn*. Ukázka kategoriální veličiny v reportu je zobrazena na obrázku 4.2. Do první tabulky byl přidán řádek s p-hodnotou pro test nezávislosti kategorií na cílové proměnné. V případě ukázkové proměnné jsou kategorie závislé na cílové proměnné, je tedy vytvořeno upozornění na závislé kategorie na cílové proměnné.



Obrázek 4.2 Ukázka kategoriální veličiny z ukázkové datové sady Titanik v *Pandas Profiling* reportu po implementaci rozšíření.

4.3.3 Textové veličiny

K nově vytvořenému datovému typu pro textové veličiny byly přidány stejné popisné statistiky, které byly implementovány pro kategoriální veličiny. K vizualizaci byla použita knihovna *wordcloud* [33]. Jedná se o nejrozšířenější knihovnu pro tvorbu wordcloud grafů v jazyce *Python*. Ukázka textové veličiny v reportu se nachází na obrázku 4.3. V ukázce lze pomocí obarvení slov (jak bylo diskutováno v kapitole 3.1) zjistit, že muži výrazně častěji nepřežili, naopak ženy častěji přežily.



Obrázek 4.3 Ukázka textové veličiny z ukázkové datové sady Titanik v *Pandas Profiling* reportu po implementaci rozšíření.

4.4 Identifikace subpopulací s odlišnou proporcí skupin v cílové proměnné

Pro identifikaci subpopulací s odlišnou proporcí skupin v cílové proměnné byl použit logaritmus *poměru šancí* se základem dva. Metoda byla implementována dle definice, která byla představena v kapitole 2.2.2, s pomocí knihovny *Pandas*. Před začátkem výpočtu *poměru šancí* je použito *beta vyhlazení* dat. Tím je do každé kategorie přidáno α kladných pozorování a β záporných pozorování. Poté je spočítána *šance* pro celou veličinu a pro každou kategorii ve veličině. Dále je vypočítán *poměr šancí* každé kategorie proti celé populaci. Jako poslední krok je výpočet logaritmu (se základem dva) z vypočítaného *poměru šancí*.

Informace, které je možné získat z grafu distribuce jsou závislé na absolutních počtech. Proto jsou přidány vizualizace logaritmu *poměru šancí*. Logaritmus *poměr šancí* je na absolutních počtech nezávislý. Lze tak lépe porovnat závislosti jednotlivých kategorií na cílové proměnné mezi sebou.

V kódu 5 je ukázka, jak je z kontingenční matice diskrétní veličiny získán logaritmus *poměru šancí*.

```
# count of imaginare observations (= alpha + beta)
imaginary_observations_count = self.config.base.smoothing_parameter
pos_prob = (cont_table[p_target_value].sum()) / (
    cont_table[p_target_value].sum() + cont_table[n_target_value].sum()
)
neg_prob = 1 - pos_prob
# beta smoothing parameters
alpha = pos_prob * imaginary_observations_count
beta = neg_prob * imaginary_observations_count

# odds of categories with smoothing
log_odds[odds] = (cont_table[p_target_value] + alpha) / (
    cont_table[n_target_value] + beta
)

# beta smoothing to whole population
groups = log_odds.shape[0]
population_odds = (cont_table[p_target_value].sum() + groups * alpha) / (
    cont_table[n_target_value].sum() + groups * beta
)

# odds ratio = category odds / population odds
log_odds[odds_ratio] = (
    log_odds[odds] / population_odds
)
log_odds[log_odds_ratio] = np.log2(log_odds[odds_ratio])
```

Výpis kódu 5 Ukázka výpočtu logaritmu *poměru šancí*.

4.4.1 Číselné veličiny

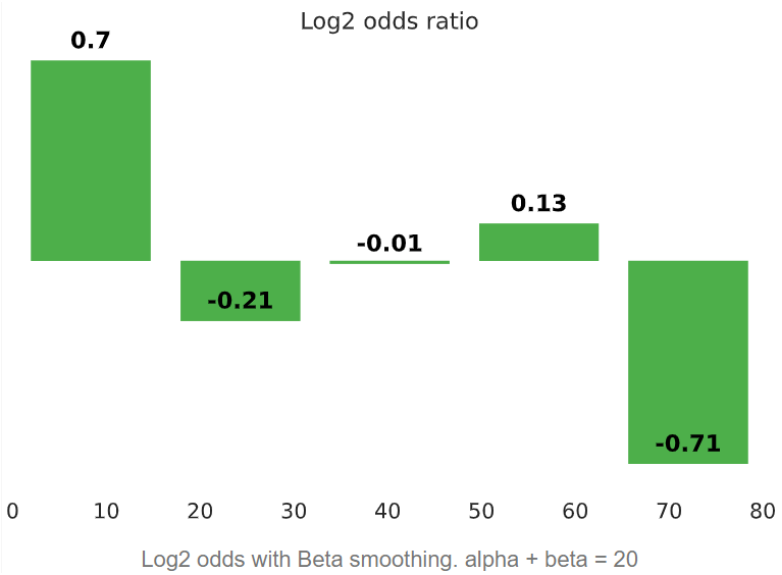
U číselných veličin byl použit binning. Tím vznikla ze spojité veličiny diskrétní veličina, nad kterou lze použít *poměr šancí* pro jednotlivé nově vzniklé třídy. Ukázka logaritmu *poměru šancí* u numerické veličiny věk z ukázkové datové sady Titanik je zobrazena na obrázku 4.4. Z grafu vyplývá, že lidé s věkem v intervalu od cca 0 do cca 10 mají značně vyšší šanci na přežití, než je průměrná šance na přežití v celé populaci. Pro lidi s věkem v intervalu cca 68 a cca 80 je naopak šance na přežití značně snížena oproti průměrné šanci na přežití v populaci.

4.4.2 Kategoriální veličiny

U kategoriálních veličin nebylo před použitím logaritmu *poměru šancí* potřeba provést žádné úpravy. Ukázka logaritmu *poměru šancí* u kategoriální veličiny třída z ukázkové datové sady Titanik je zobrazena na obrázku 4.5. Z grafu vyplývá, že největší šanci na přežití měli lidé ubytovaní v první třídě, dále lidé z druhé třídy a nejhůře dopadli lidé ze třetí třídy.

4.4.3 Textové veličiny

U textových veličin byly provedeny následující transformace. Nejprve jsou všechna data převedena na malá písmena, dále jsou jednotlivé záznamy rozděleny na slova. Poté je vytvořena kontingenční matice s počtem výskytů slova dle cílové proměnné. Pro vizualizaci logaritmu *poměru šancí* byla



Obrázek 4.4 Ukázka logaritmu *poměru šancí* u číselné veličiny v reportu po implementování rozšíření.

upravena tabulka výskytů, implementovaná v knihovně *Pandas Profiling*. Pokud se v textové veličině vyskytuje méně než 10 různých slov, jsou zobrazeny logaritmy *poměrů šancí* všech slov. Pokud se v proměnné vyskytuje více slov, je zobrazeno pět slov s nejvyšším logaritmem *poměru šancí* a pět slov s nejnižším logaritmem *poměru šancí*. Ukázka textové veličiny jméno z datové sady Titanik je zobrazena na obrázku 4.6.

4.5 Chybějící hodnoty

Pro reprezentaci závislosti chybějících hodnot na cílové proměnné je použita kontingenční matice. K vykreslení kontingenčních matic je použita knihovna *seaborn*. Ukázka je na obrázku 4.7. Nad touto kontingenční maticí je následně proveden chí-kvadrát test nezávislosti, který je implementován v knihovně *scipy*. Výsledná p-hodnota je následně porovnána s prahovou hodnotou z konfiguračního souboru. V případě, že je prahová hodnota překročena, je vytvořeno upozornění na korelaci chybějících hodnot a hodnot cílové proměnné ve zkoumané veličině. V případě ukázky jsou chybějící data výrazně závislá na cílové proměnné.

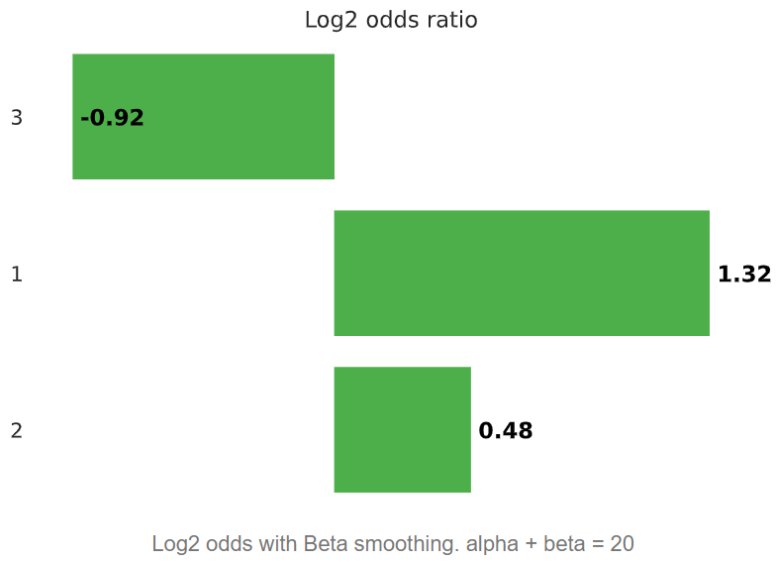
4.6 Doporučené transformace

Všechny transformace byly implementovány pomocí knihovny *scikit learn* [34].

Normalizace byla implementována pomocí třídy `StandardScaler`, která se nachází v modulu *preprocessing* v knihovně *scikit learn*. Tato třída reprezentuje *z-skóre* transformaci.

Binning byl implementován pomocí třídy `KBinsDiscretizer`, která se nachází v modulu *preprocessing* v knihovně *scikit learn*. Pomocí parametru `strategy="quantile"` byl zvolen *binning stejné výšky*. Počet nově vzniklých sloupců je možné nastavit v konfiguračním souboru (výchozí hodnota je 5).

Logaritmizace byla implementována pomocí třídy `FunctionTransformer`, která se nachází v modulu *preprocessing* v knihovně *scikit learn*. Tato třída přijímá jako parametr funkci, kterou následně používá pro transformaci. Jako funkce pro logaritmickou transformaci dat byla zvolena `np.log1p` z knihovny *numpy*, která k hodnotě nejprve přičte konstantu 1 a následně aplikuje přirozený logaritmus.



Obrázek 4.5 Ukázka logaritmu *poměru šancí* u kategoriální veličiny v reportu po implementování rozšíření.

Pro implementaci *one-hot encoding* transformace byla použita třída `OneHotEncoder` z modulu *preprocessing* v knihovně *scikit learn*.

Pro implementaci *tf-idf* transformace byla použita třída `TfidfVectorizer` z modulu *feature_extraction.text* v knihovně *scikit learn*. Ve třídě `TfidfVectorizer` je inverzní četnost dokumentů definována následovně:

$$idf(t) = \log[\frac{1+n}{1+df(t)}] + 1.$$

Pomocí parametru `max_features` je omezen počet použitých termínů. Tento počet je možné nastavit v konfiguračního souboru (výchozí hodnota je nastavena na 50). Termíny jsou seřazeny od nejčetnějších k nejméně četným. Jako dokumenty jsou brány jednotlivé řádky v textové proměnné.

4.7 Referenční model

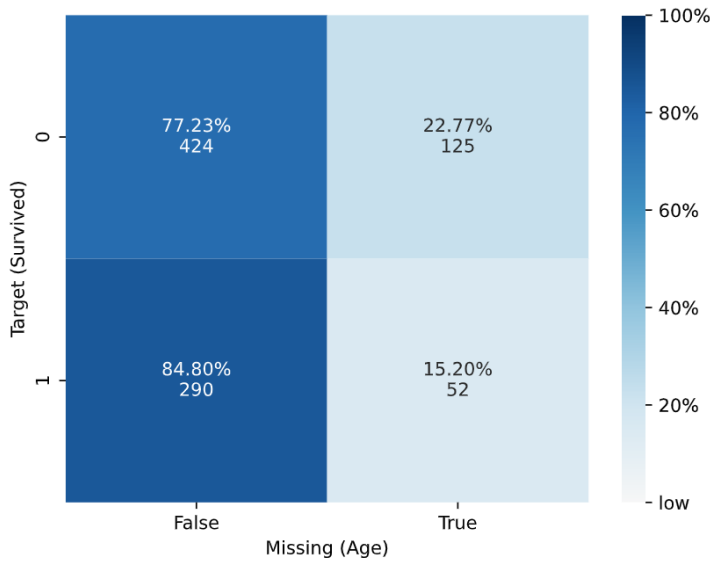
Pro implementaci referenčního modelu byla použita knihovna *LightGBM* [35]. V této knihovně je pomocí třídy `LGBMClassifier` implementován *rozhodovací strom s gradientním boostingem*. Model `LGBMClassifier` umí pracovat s kategoriálními veličinami. To spoustu modelů neumí. Z důvodu lepšího doporučení transformací pro modelování byly z referenčního modelu odstraněny kategoriální a textové veličiny. Tím byla zhoršena kvalita referenčního modelu a je tak dáno více prostoru navrženým transformacím. Nastavení některých parametrů modelu je možné pomocí konfiguračního souboru, jak bylo představeno v kapitole 4.2.1.6.

Vizualizace modelu se skládá z několika částí. První část se zabývá evaluací modelu, která obsahuje hodnoty pro *accuracy*, *precision*, *recall* a *f1-skóre*. Ukázka této části je na obrázku 4.8. Druhá část zobrazuje tabulku očekávaných a predikovaných hodnot. Dále je zobrazeno nastavení modelu, nastavení rozdělení dat na trénovací a testovací a nakonec tabulka s důležitostí jednotlivých sloupců. Ukázka zobrazení nastavení modelu, nastavení rozdělení dat a seznam nejdůležitějších sloupců pro model se nachází na obrázku 4.9.

Log odds words with Beta smoothing. alpha + beta = 20

Value	Positive count	Negative count	Log2 odds ratio
mrs	102	27	1.97
miss	127	55	1.48
elizabeth	13	2	1.11
mary	15	5	0.95
kate	7	0	0.87
...	...	...	
david	0	7	-0.68
john	11	33	-0.74
johan	2	13	-0.78
alfred	1	11	-0.8
mr	84	437	-1.83

Obrázek 4.6 Ukázka logaritmu *poměru šancí* u textové veličiny v reportu po implementování rozšíření.



P-value for the chi-square independence test is 0.0059. Missing values and target variable are related at 95.0% confidence level.

Obrázek 4.7 Ukázka kontingenční matice chybějících hodnot u proměnné věk a cílové proměnné z datové sady Titanik.

Model evaluation

Accuracy (%)	68.2%
Precision (%)	33.0%
Recall (%)	70.7%
F1 score (%)	45.0%

Obrázek 4.8 Ukázka evaluace modelu v reportu po implementování rozšíření.

Model setting

Used model	lightgbm.LGBMClassifier
Boosting type	Gradient Boosting Decision Tree
Maximum tree depth	3
Number of boosted trees	10
Maximum tree leaves	10
Model seed	123456

Train test setting

Train test split policy	random
Test size (%)	25.0%
Train records count	668
Test records count	223

Feature importances

Fare	447.643
Age	135.21
PassengerId	88.149
Pclass	64.6386
SibSp	22.9218
Parch	2.77754

Obrázek 4.9 Ukázka nastavení modelu, nastavení rozdělení dat a seznam nejdůležitějších sloupců v reportu po implementování rozšíření.

Závěr

Práce se zabývala explorací dat s binární klasifikací. Byla provedena rešerše nástrojů pro automatickou exploraci dat. Dále byly prozkoumány statistické testy a metody, které by bylo možné použít pro odhalení datových závislostí na cílové proměnné. Byly též prozkoumány možnosti vizualizací distribuce pro základní datové typy proměnných, jako je *číselný*, *kategoriální* a *textový* datový typ. V další části byly navrženy úpravy knihovny pro automatickou exploraci dat *Pandas Profiling*, které zlepší kvalitu explorace dat s binární klasifikací.

Dále byly implementovány následující úpravy z návrhu: K jednotlivým proměnným byly přidány testy nezávislosti na cílové proměnné a testy subpopulací pomocí logaritmu *relativního rizika*. Byl též přidán test závislosti chybějících hodnot na cílové proměnné. Dále byla knihovna rozšířena o sekci s doporučenými transformacemi jednotlivých sloupců, které zlepší kvalitu dat pro trénování klasifikačního modelu. Jako poslední byla přidána sekce s referenčním modelem, který je natrénován na originálních a na transformovaných datech, a jsou zobrazeny metriky ohodnocení datových modelů.

Výstup práce je přínosem datascience komunitě. Rozšíření implementované v rámci této práce bude přidáno do knihovny *Pandas Profiling*, ze které práce vycházela. Některé části práce, mezi něž patří např. nový datový typ pro textové veličiny, již byly do knihovny přidány.

Bibliografie

1. YDATA LABS INC. *ydata-profiling* [online]. 2023. Ver. 3.6.6 [cit. 2023-04-22]. Dostupné z: `https://github.com/ydataai/ydata-profiling`.
2. PYTHON SOFTWARE FOUNDATION. *Python 3.10.11 documentation* [online]. 2023. [cit. 2023-04-19]. Dostupné z: `https://docs.python.org/3.10/`.
3. THE R FOUNDATION. *The R Project for Statistical Computing* [online]. 2023. [cit. 2023-04-22]. Dostupné z: `https://www.r-project.org/`.
4. DATA SCIENCE DOJO. *Titanic* [online]. 2015. [cit. 2023-05-01]. Dostupné z: `https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv`.
5. PANDAS. *pandas* [soft.]. 2023. [cit. 2023-04-22]. Dostupné z: `https://pandas.pydata.org/`.
6. MAN ALPHA TECHNOLOGY. *D-Tale* [online]. 2023. Ver. 2.14.0 [cit. 2023-04-22]. Dostupné z: `https://github.com/man-group/dtale`.
7. MAN ALPHA TECHNOLOGY. *D-Tale* [online]. 2023. [cit. 2023-04-22]. Dostupné z: `http://alphatechadmin.pythonanywhere.com/dtale/main/1`.
8. YDATA LABS INC. *Titanic Dataset* [online]. 2023. [cit. 2023-04-22]. Dostupné z: `https://ydata-profiling.ydata.ai/examples/master/titanic/titanic_report.html`.
9. SFU DATABASE SYSTEM LAB. *DataPrep* [online]. 2022. Ver. 0.4.5 [cit. 2023-04-22]. Dostupné z: `https://dataprep.ai/`.
10. SFU DATABASE SYSTEM LAB. *DataPrep Demo* [online]. 2022. [cit. 2023-04-22]. Dostupné z: `https://colab.research.google.com/drive/1U_-pAMcne3hK1HbMB3kuEt-093Np_7Uk?usp=sharing#scrollTo=lfdZeenXbLNf`.
11. BERTRAND, Francois. *Sweetviz* [online]. 2022. Ver. 2.1.4 [cit. 2023-04-22]. Dostupné z: `https://github.com/fbdesignpro/sweetviz`.
12. BERTRAND, Francois. *Sweetviz Demo* [online]. 2020. [cit. 2023-04-22]. Dostupné z: `http://cooltiming.com/SWEETVIZ_REPORT.html`.
13. HOGG ROBERT McKean Joseph, Craig Allen. *Introduction to Mathematical Statistics*. 8. vyd. Pearson, 2018. ISBN 978-0134686998.
14. SIRISILLA, Shrutika. *7 Ways to Choose the Right Statistical Test for Your Research Study* [online]. Přel. DEEPL. 2023. [cit. 2023-04-06]. Dostupné z: `https://www.enago.com/academy/right-statistical-test/`.
15. HAYES, A. T-Test: What it is with multiple formulas and when to use them. *Investopedia. Retrieved December*. 2022, roč. 31, s. 2022.

16. TURNEY, Shaun. *Chi-Square Test of Independence* [online]. Přel. DEEPL. 2022. [cit. 2023-04-06]. Dostupné z: https://www.scribbr.com/statistics/chi-square-test-of-independence/.
17. BOWER, Keith M. When to use Fisher's exact test. In: *American Society for Quality, Six Sigma Forum Magazine*. Přel. DEEPL. American Society for Quality Milwaukee, WI, USA, 2003, sv. 2, s. 35–37. Č. 4.
18. SCIPY COMMUNITY. *scipy.stats.fisher_exact* [online]. Přel. DEEPL. 2023. [cit. 2023-04-19]. Dostupné z: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.fisher_exact.html#scipy-stats-fisher-exact.
19. TENNY, Steven; HOFFMAN, Mary R. Relative risk. *Information* [online]. 2017 [cit. 2023-04-27]. Dostupné z: https://europepmc.org/article/nbk/nbk430824. PMID:28613574.
20. SZUMILAS, Magdalena. Explaining odds ratios. *Journal of the Canadian academy of child and adolescent psychiatry* [online]. 2010, roč. 19, č. 3, s. 227 [cit. 2023-07-04]. Dostupné z: https://pubmed.ncbi.nlm.nih.gov/20842279/. PMID:20842279.
21. BLAND, J Martin; ALTMAN, Douglas G. The odds ratio. *Bmj*. 2000, roč. 320, č. 7247, s. 1468.
22. DAVIES, Huw Talfryn Oakley; CROMBIE, Iain Kinloch; TAVAKOLI, Manouche. When can odds ratios mislead? *Bmj*. 1998, roč. 316, č. 7136, s. 989–991.
23. JAYASWAL, Vaibhav. *Laplace smoothing in Naïve Bayes algorithm* [online]. 2020. [cit. 2023-04-13]. Dostupné z: https://towardsdatascience.com/laplace-smoothing-in-na%C3%AFve-bayes-algorithm-9c237a8bdece.
24. JOHNSON, Norman L; KOTZ, Samuel; BALAKRISHNAN, N. Beta distributions. *Continuous univariate distributions. 2nd ed. New York, NY: John Wiley and Sons*. 1994, s. 221–235.
25. ROBINSON, David. *Understanding the beta distribution (using baseball statistics)* [online]. Přel. DEEPL. 2014. [cit. 2023-04-13]. Dostupné z: http://varianceexplained.org/statistics/beta_distribution_and_baseball/.
26. WILKE, Claus O. *Fundamentals of data visualization: a primer on making informative and compelling figures*. 1. vyd. O'Reilly Media, 2019. ISBN 978-1492031086.
27. HAMDAOUI, Y. TF (Term Frequency)-IDF (Inverse Document Frequency) from scratch in python. *Medium. Retrieved July*. 2021, roč. 23, s. 2022.
28. THE SCIPY COMMUNITY. *SciPy documentation* [soft.]. 2023. Ver. 1.10.1 [cit. 2023-04-19]. Dostupné z: https://docs.scipy.org/doc/scipy/.
29. WASKOM, Michael. *seaborn* [soft.]. 2022. Ver. 0.12 [cit. 2023-04-19]. Dostupné z: https://seaborn.pydata.org/.
30. DYLAN PROFILER. *visions* [soft.]. 2021. Ver. 0.7.5 [cit. 2023-04-19]. Dostupné z: https://github.com/dylan-profiler/visions.
31. PYTHON SOFTWARE FOUNDATION. *argparse* [soft.]. 2015. Ver. 1.4.0 [cit. 2023-04-19]. Dostupné z: https://docs.python.org/3/library/argparse.html#module-argparse.
32. PYTHON SOFTWARE FOUNDATION. *The Python Standard Library* [soft.]. 2023. Ver. 3.11 [cit. 2023-05-04]. Dostupné z: https://docs.python.org/3.11/library/index.html.
33. MUELLER, Andreas. *wordcloud* [soft.]. 2023. Ver. 1.9.1.1 [cit. 2023-05-03]. Dostupné z: https://github.com/amueller/word_cloud.
34. COURNAPEAU, David. *scikit-learn* [soft.]. 2023. Ver. 1.2.2 [cit. 2023-05-03]. Dostupné z: https://scikit-learn.org/stable/.
35. MICROSOFT. *LightGBM* [soft.]. 2023. Ver. 3.3.5 [cit. 2023-05-03]. Dostupné z: https://github.com/microsoft/LightGBM.

Obsah přiloženého média

- `README.txt` ..... stručný popis obsahu média a návod na instalaci programu
- `data` ..... složka s ukázkovými datovými sadami
- `out` ..... složka s ukázkovými vygenerovanými reporty
- `supervised-pandas-profiling` ..... balíček knihovny
- `text` ..... text práce
  - `thesis` ..... zdrojová forma práce ve formátu LaTeX
  - `capjan5.pdf` ..... text práce ve formátu PDF