ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
Fakulta Elektrotechnická

Zaklady Datovych Analyz

Semestralni prace studenta

Obor: SIT - BI

Ivan Pasportnikov

2024

Contents

1 Úvod **2**
1.1 Formulace problému 2

2 Vstupní data **2**

3 Otazky **3**

4 Hypotézy **3**

5 Postup **4**

6 Závěr **8**

1 Úvod

1.1 Formulace problému

Srdeční selhání (SS) je hlavní příčinou úmrtí číslo 1 na celém světě, přičemž ročně způsobí přibližně 17,9 milionu úmrtí, což představuje 31% všech úmrtí na světě. Více než 3 z 5 úmrtí na srdeční selhání jsou spojeny s průměrnými nástroji a přístupy, a více než třetina těchto úmrtí nastává neočekávaně u lidí ve věku nad 70 let. Nedostatečná přesnost je rozšířeným problémem, způsobeným srdečními selháními, a tento dataset obsahuje 11 znaků, které lze použít k předpovídání možného srdečního onemocnění.

Lidé se srdečním selháním nebo s vysokým srdečním rizikem (kvůli nedostatku několika rizikových faktorů, jako jsou genetika, dieta, předchozí nemoci nebo stávající zdravotní stav) potřebují včasné vyhodnocení a léčbu, ve které může analyza a model strojového učení poskytnout značnou pomoc.

Info je z Wikipedia
World Health Organization

2 Vstupní data

Analýza lékařského datasetu
V mém projektu se chystám prozkoumat dataset dat věnovaný

srdečnímu selhání pacientů.

```
Dataset obsahuje věk pacienta, pohlaví pacienta, typ bolesti v hrudi,
krevní tlak v klidu, hladina cholesterolu, hladina cukru v krvi nalačno,
výsledky elektrokardiogramu v klidu, maximální dosažená srdeční frekvence,
ExerciseAngina, ST deprese vyvolaná cvičením v porovnání s klidem,
the slope of the peak exercise ST segment, výsledný stav srdce.
```

Dataset Link **https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/code?datasetId=1582403&searchQuery=py**

3 Otazky

- **Jaké jsou hlavní rizikové faktory srdečního selhání?**
- **Jak často srdeční selhání probíhá asymptomatické?**

World Health Organization
"The effects of behavioural risk factors may show up in individuals as raised blood pressure, raised blood glucose..."
"Often, there are no symptoms of the underlying disease of the blood vessels."

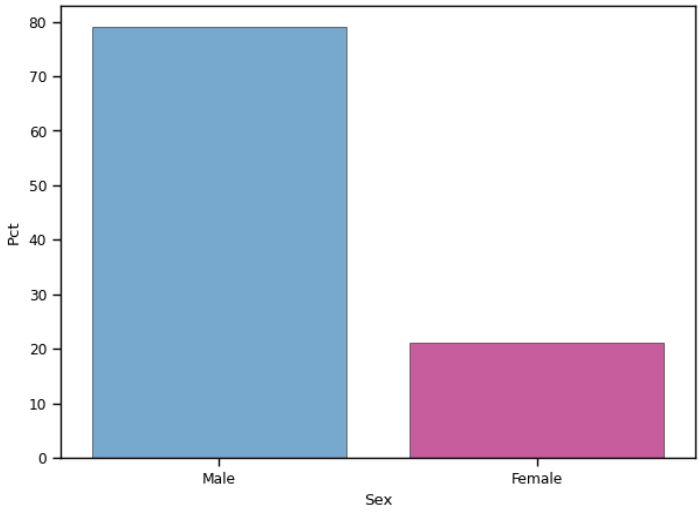
4 Hypotézy

- **Rozdíly mezi muži a ženami ve výskytu srdečního selhání:**
  - **Hypotéza: Existují statisticky významné rozdíly ve výskytu srdečního selhání mezi muži a ženami. Muži trpí srdečním selháním častěji než ženy.**
  - **Ověření: Srovnáme procentuální zastoupení srdečního selhání u mužů a žen a analyzujeme rozdíly v klinických charakteristikách.**
- **Literatura**
  - **BMJ Global Health**
  - **National Library of Medicine**

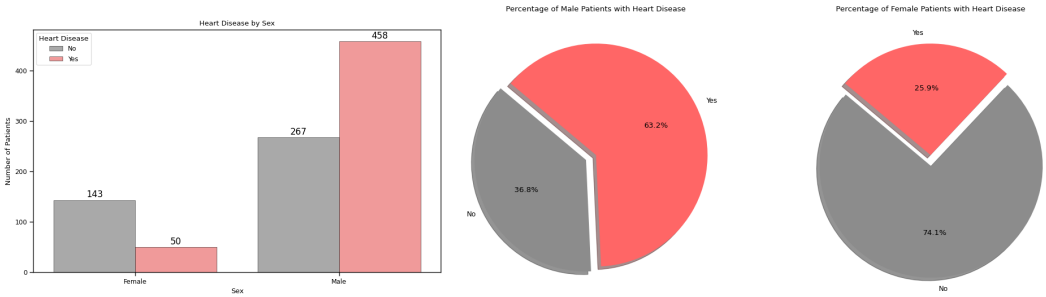
5 Postup

Nejdříve jsem si stáhl data, která se vztahují k tématu.
Popis kroků řešení

1. Formulace problému
2. Import knihoven
3. Načtení a první pohled na dataset
4. Čištění dat
   (a) There are no NaN (Null) values
   (b) There are no Duplicate values in the dataset
5. Data Analysis
   (a) Změna hodnot v kategoriích pro lepší čitelnost
   (b) Vizualizace rozložení pohlaví



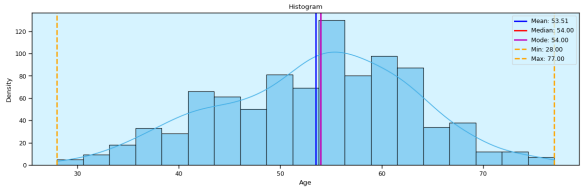
   (c) Analýza srdečního onemocnění podle pohlaví a potvrzení hypotézy



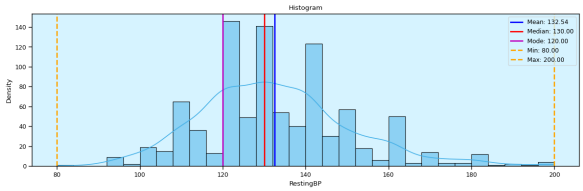
(d) Vizualizace rozložení číselných charakteristik

(e) Vytváření histogramů a koláčových diagramů pro číselné charakteristiky a kategorie bolesti na hrudi.

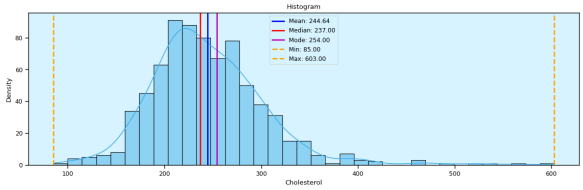
i. hist_plot("Age")



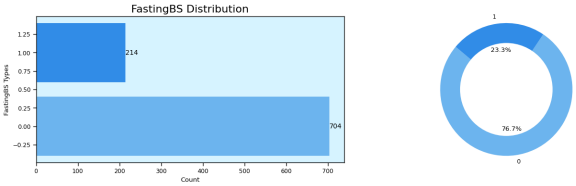
ii. hist_plot("RestingBP")



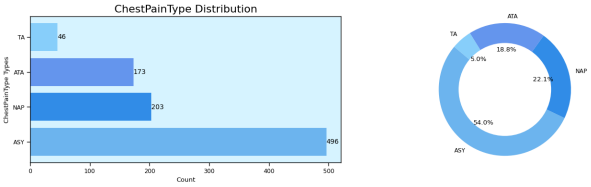
iii. hist_plot("Cholesterol")



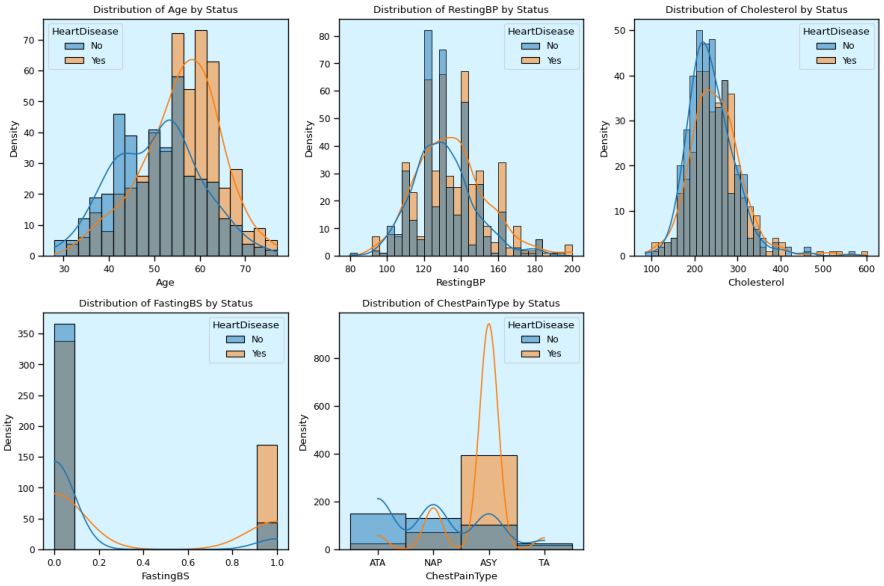
iv. bardonut_chart("FastingBS")



v. bardonut_chart("ChestPainType")



(f) Vizualizuju distribuci vybraných numerických proměnných podle statusu srdečního onemocnění.



Křivky, které jsou zobrazeny na vrcholech jednotlivých grafů, jsou Kernel Density Estimate (KDE) křivky. Tyto křivky poskytují hladký odhad distribuční hustoty dat, což umožňuje lépe vizualizovat, kde se nachází většina hodnot a jaký je tvar distribuce.

i. Distribution of Age by Status:
Modrá křivka (bez srdečního onemocnění) ukazuje, že většina

lidí bez srdečního onemocnění je mladší 50 let. Oranžová křivka (se srdečním onemocněním) ukazuje, že většina lidí se srdečním onemocněním je starší než 50 let. Tento rozdíl potvrzuje závěr, že pravděpodobnost srdečního onemocnění se zvyšuje s věkem.

ii. Distribution of RestingBP by Status:
Modrá křivka (bez srdečního onemocnění) je soustředěna více nalevo, což znamená nižší hodnoty krevního tlaku v klidu. Oranžová křivka (se srdečním onemocněním) je mírně posunuta doprava, což znamená vyšší hodnoty krevního tlaku. Rozložení krevního tlaku v klidu je skoro podobné jak u lidí se srdečním onemocněním, tak u lidí bez něj.

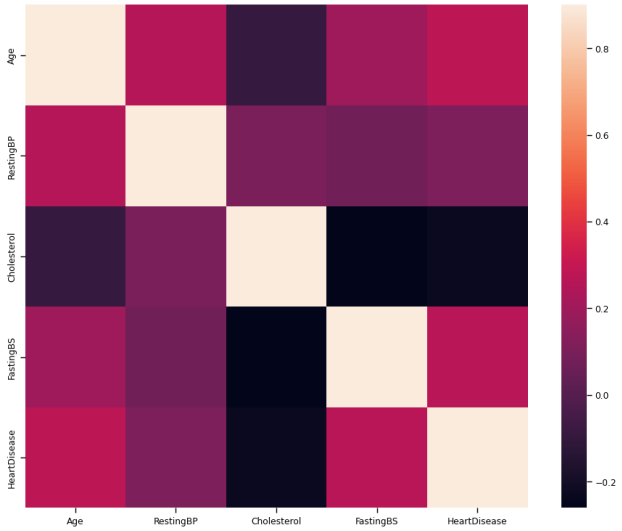
iii. Distribution of Cholesterol by Status:
Modrá křivka ukazuje, že většina lidí bez srdečního onemocnění má hladinu cholesterolu kolem 200 mg/dL. Oranžová křivka je mírně posunuta doprava ale ukazuje širší rozložení hladiny cholesterolu, což naznačuje, že lidé se srdečním onemocněním mají různé hladiny cholesterolu, což podtrhuje potřebu multidimenzionálního přístupu k výzkumu a prevenci srdečních onemocnění.

iv. Distribution of FastingBS by Status:
U lidí bez srdečního onemocnění je hladina krevního cukru nalačno (FastingBS) většinou nízká. U lidí se srdečním onemocněním je častěji zvýšená hladina cukru v krvi nalačno. To potvrzuje, že hyperglykémie je významným rizikovým faktorem pro rozvoj srdečních onemocnění.

v. Distribution of ChestPainType by Status:
Modrá křivka ukazuje, že lidé bez srdečního onemocnění mají více typů bolesti na hrudi, jako je asymptomatická (ASY) nebo atypická (ATA). Oranžová křivka ukazuje, že u lidí se srdečním onemocněním je nejběžnější asymptomatický typ (ASY). To znamená, že značná část pacientů se srdečním onemocněním nemusí vykazovat zjevné příznaky.

(g) Vytvoření korelační matici pro vybrané proměnné a vizualizujeme ji pomocí heatmapy

(h) Výpočet a vizualizace korelační matice k identifikaci vztahů mezi číselnými charakteristikami a srdečním onemocněním.



(i) Zde vidíme, že srdeční onemocnění má vysokou negativní korelaci s "Cholesterol" a poněkud negativní-pozitivní korelaci s "RestingBP", kde jako zde pozitivní korelace s "Age" a "FastingBS".

Postup samotné analýzy je popsán v notebooku ZDA_paspoiva.ipynb, který je součástí odevzdané práce.

6 Závěr

Na základě poskytnutých poznámek lze učinit následující závěry:

- Potvrzena hypotéza, že nemocemi srdečního selhání trpí převážně muži. Při procentuálním poměru mužů v tabulce je 79 procent onemocnění u mužů 63,2, zatímco u žen je tato hodnota rovna 25,9
- Byla provedena analýza , aby se zjistilo, jaké faktory ovlivňují srdeční selhání
    1. S věkem se zvyšuje pravděpodobnost onemocnění, zvláště náchylní k onemocnění jsou rizikoví pacienti s věkem 50+
    2. Navzdory určitému nárůstu srdeční frekvence při zvýšení krevního tlaku není tento nárůst významný. To může naznačovat, že ačkoli

je zvýšený krevní tlak jedním z rizikových faktorů, není jediným a rozhodujícím faktorem v tomto souboru dat.

3. Zjištění cholesterolu zdůrazňují potřebu multidimenzionálního přístupu ve výzkumu a prevenci srdečních onemocnění, přičemž uznávají, že kontrola jednoho rizikového faktoru, jako je hladina cholesterolu, by měla být doprovázena kontrolou a dalšími faktory, aby byla maximální účinnost.
4. Dochází k výraznému zvýšení frekvence srdečních onemocnění se zvýšenou hladinou cukru v krvi. To naznačuje, že hyperglykémie (zvýšená hladina cukru) je významným rizikovým faktorem pro rozvoj srdečních onemocnění. Vysoká hladina cukru v krvi může poškodit krevní cévy a přispět k rozvoji aterosklerózy, což zvyšuje riziko srdečních onemocnění.
5. Analýza ukázala, že téměř 50% případů srdečních onemocnění je asymptomatických. To znamená, že značný počet pacientů nemusí pociťovat zjevné příznaky navzdory přítomnosti závažného onemocnění. Asymptomatický průběh srdečních onemocnění je významnou hrozbou, protože takoví pacienti nemusí vyhledat lékařskou pomoc včas, což zvyšuje riziko náhlých a závažných komplikací, jako je infarkt myokardu nebo náhlá srdeční smrt.

- Srdeční onemocnění se obvykle vyvíjejí pod vlivem mnoha faktorů, včetně genetické predispozice, hladiny cholesterolu, životního stylu, přítomnosti dalších onemocnění atd. Je nutné provádět komplexní lékařská vyšetření a zvážit všechny druhy rizikových faktorů, aby byla zajištěna účinná prevence a léčba srdečních onemocnění.

Tato zjištění zdůrazňují potřebu multidimenzionálního přístupu ve výzkumu a prevenci srdečních onemocnění, přičemž uznávají, že kontrola jednoho rizikového faktoru musí být doprovázena kontrolou a dalšími faktory, aby byla maximalizována účinnost.