David Mašek (masd04)
Viktor Sívek (sivv01)

Předpověď srdečního selhání – dataset

Porozumění problematice

Kardiovaskulární onemocnění ročně zabíjí přibližně 17 milionů lidí na celém světě a projevují se především jako infarkty a srdeční selhání. Srdeční selhání nastává, když srdce nemůže pumpovat dostatek krve, aby uspokojilo potřeby těla. Dostupné elektronické lékařské záznamy pacientů kvantifikují symptomy, tělesné rysy a hodnoty klinických laboratorních testů, které lze použít k provedení biostatistické analýzy zaměřené na zvýraznění vzory a korelace, které by jinak lékaři nezjistili. Zejména strojové učení dokáže předvídat přežití pacientů z jejich dat a dokáže rozlišit nejdůležitější vlastnosti mezi těmi, které jsou obsaženy v jejich lékařských záznamech.[1]

Porozumění datům:

Heart Failure Prediction Dataset zpracovává tělesné rysy a hodnoty klinických laboratorních testů u zkoumaných subjektů. Dataset obsahuje 918 záznamů(řádků, neboli zkoumaných subjektů) a 12 atributů (sloupců, neboli zkoumaných hodnot). Záznamy jsou kombinací několika datasetů z různých zdrojů.[9]

Age - věk pacienta

Datový typ: **numeric**

Jednotky: **roky**

Rozsah hodnot: **28 - 77**

Průměrný věk: **54**

Nejpočetnější skupina: **54 - 56 let** (92 výskytů).

Sex - pohlaví pacienta

Datový typ: **binary**

Hodnoty: **M / F** (male / female)

David Mašek (masd04)
Viktor Sívek (sivv01)

Počet M: **725**

Počet F: **193**

ChestPainType - typ bolesti na hrudi

Bolesti na hrudi může způsobovat typická bolest srdce (TA) způsobená sníženým průtokem krve do srdečních svalů.[2] Když člověk zaživá bolesti na hrudi, které nesplňují kritéria pro typickou bolest srdce, je známá jako atypická bolest na hrudi (ATA). Nesrdeční bolest na hrudi (NAP) je termín, který se používá k popisu bolesti na hrudi, která není způsobena srdečním onemocněním nebo infarktem. [3] [4]

Datový typ: **nominal**

Hodnoty (počet): **ASY: Asymptomatic**(496) **/ ATA: Atypical Angina**(173) **/ NAP: Non-Anginal Pain**(203) **/ TA: Typical Angina**(46)

Nejpočetnější: **ASY**

RestingBP - krevní tlak v klidovém stádiu

Předpokládáme, že v případě uvedených hodnot se jedná o tlak systolický, zachycující nejvyšší naměřenou hodnotu v daném místě. U dospělého zdravého jedince by systolický tlak neměl přesáhnout hodnotu 140 mmHg.

V datech se u jedné hodnoty vyskytovala 0. Tu jsme se rozhodli nahradit mediánem nenulových hodnot, tedy hodnotou 130. [5], [6]

Datový typ: **numeric**

Jednotky: **mmHg**

Rozsah hodnot: **80 - 200**

Průměr: **132.54**

Nejpočetnější skupina: **120 - 125 mmHg** (158 výskytů).

Cholesterol - množství cholesterolu v krvi

Cholesterol je částice tuku v krvi, která hraje důležitou roli v řadě tělesných procesů. Zvýšená hladina cholesterolu v krvi je způsobena buď zvýšenou konzumací především živočišných tuků

nebo zvýšenou syntézou cholesterolu v játrech.[7] V tomto sloupci se u 172 subjektů vyskytovala hodnota 0. Tu jsme se rozhodli nahradit mediánem všech zbývajících nenulových hodnot, který činí 237.

Datový typ: **numeric**

Jednotky: **mmHg**

Rozsah hodnot: **85 - 603**

Průměr: **244,7**

Nejpočetnější skupina: **120 - 125 mmHg** (158 výskytů).

FastingBS - hladina cukru v krvi nalačno

Hladina cukru v krvi nalačno 99 mg/dl nebo nižší je normální, 100 až 125 mg/dl znamená, že máte prediabetes, a 126 mg/dl nebo vyšší znamená, že máte cukrovku.[8]

Datový typ: **binary**

Hodnoty: **1** - jestli hladina cukru v krvi > 120 mg/dl, **0** - jinak nula

Počet 1: **704**

Počet 0: **214**

RestingECG - výsledky klidového kardiogramu

Hypertrofie levé komory **(LVH)** - Neboli zmohutnění její svaloviny bývá často nalezeno u chronických kardiaků. Na EKG se vyskytují hluboké negativní kmity QRS ve V1 a V2 a vysoké pozitivní kmity QRS ve V5 a V6. Zátěž levé komory může býti spojena i s výskytem depresí ST a negativních vln T ve V5 a V6.

Výkyv v intervalu ST-T **(ST) -** Při výskytu výchylky intervalu ST od izoterické křivky při depresi/elevaci ve výši 0.05 mV a více. **(Normal)** - Výsledek měření EKG v normě. [10] [11] [12] [13]

Datový typ: **nominal**

Hodnoty (počet): **LVH**(188) / **Normal**(552) / **ST**(178)

Nejpočetnější: **Normal**

MaxHR - maximální dosažená tepová frekvence

Datový typ: **numeric**

Jednotky: **bpm**

Rozsah hodnot: **60 - 202**

Průměr: **136.81**

Nejpočetnější skupina: **140 - 145 bpm** (84 výskytů).

ExerciseAngina - bolest na hrudi při námaze

Datový typ: **binary**

Hodnoty: **Y / N** (Yes / No)

Počet Y: **371**

Počet N: **547**

Oldpeak

Porovnání ST deprese při zátěži a v klidovém režimu. [14]

Datový typ: **numeric**

Rozsah hodnot: (**-2.6**) **- 6.2**

Průměr: 1.32

Nejpočetnější skupina: **0 - 0.5** (426 výskytů).

David Mašek (masd04)
Viktor Sívek (sivv01)

ST_Slope

Míra spádu vlny ST při při maximální míře zátěže.

Downsloping **(Down)**- klesající. Flat **(Flat)** - rovná. Upsloping **(Up)** - rostoucí. [14]

Datový typ: **nominal**

Hodnoty: (počet): **Down**(63) / **Flat**(460) / **Up**(395)

Nejpočetnější: **Flat**

HeartDisease - výskyt srdeční choroby

Datový typ: **binary**

Hodnoty: **1 / 0** (Ano / Ne)

Počet Ano: **508**

Počet Ne: **410**

Specifikace zadání

V naší práci se budeme snažit nalézt odpovědi na následující dílčí otázky s využitím analytických metod.

1. Které pohlaví je náchylnější na srdeční choroby?
2. Měli bychom se obávat o své srdce při výskytu bolesti na hrudi při námaze?
3. Která věková kategorie jsou nejnáchylnější k srdečním chorobám?
4. Má hladina cholesterolu vliv na srdeční choroby?
5. Je srdce s vysokou tepovou frekvencí odolnější vůči chorobám?
6. Můžeme se bez bolestí na hrudi cítit v bezpečí?
7. S kterou hodnotou ST_Slope je nejvyšší pravděpodobnost srdeční choroby?
8. Při kterých predispozicích je subjekt nejnáchylnější k srdečnímu onemocnění?
9. Při kterých je náchylný nejméně?

Odpovědi na dílčí otázky jsou zodpovězeny u grafů zobrazujících příslušná data.

David Mašek (masd04)
Viktor Sívek (sivv01)

Vizualizace atributů

Jednotlivé atributy jsme vizualizovali pomocí Pythonu. Do Pythonu jsme naimportovali ‘pandas’ a další potřebné softwarové knihovny, které využijeme při naší analýze.

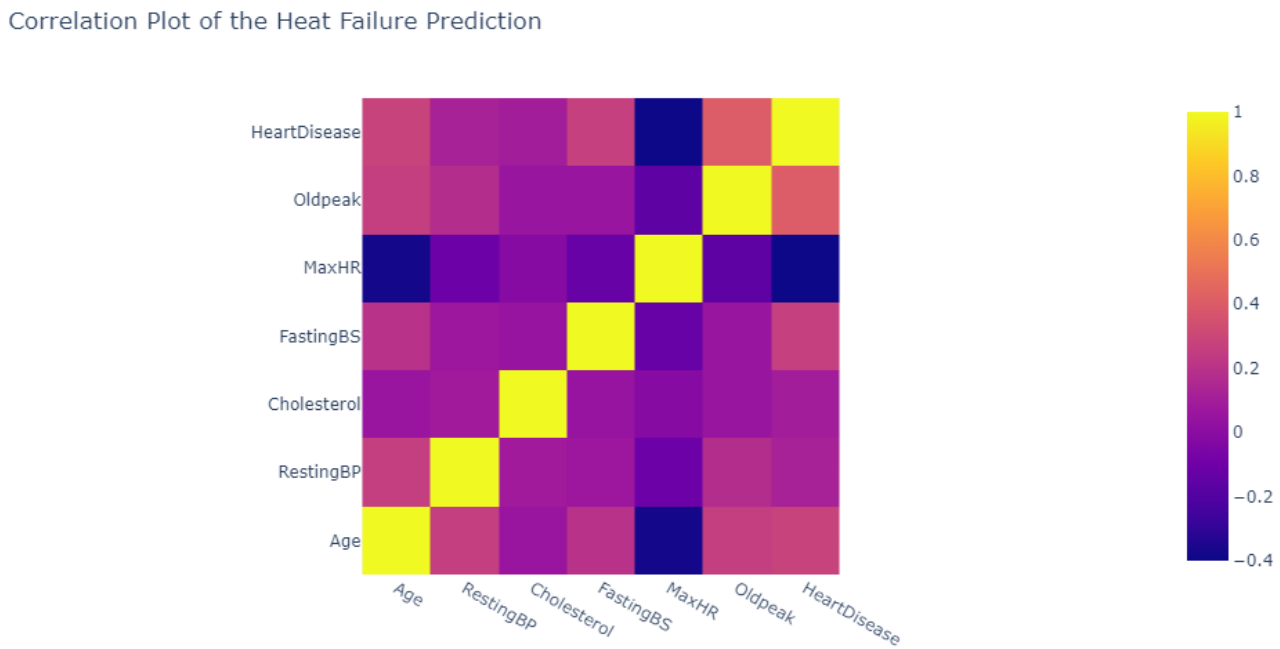
Z důvodu nepovedené implementace programu Lotylda jsme pro další analýzu dat využili program BigML a jeho funkci pro hledání asociací.

Pro zopakování jsme si zobrazili tabulku s prvními pěti hodnotami datasetu.

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

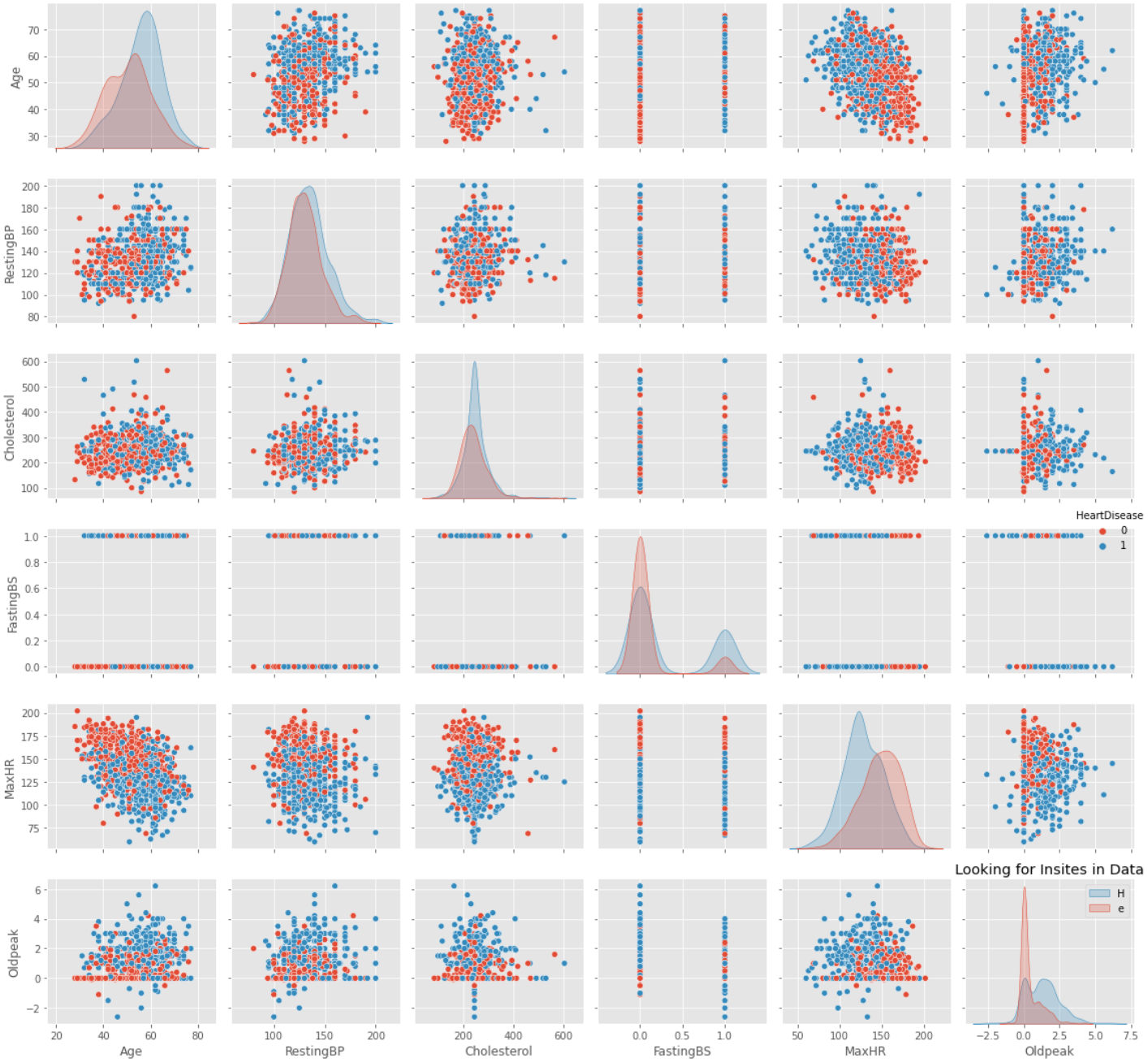
Následně jsme přešli k vizualizaci korelací mezi jednotlivými atributy.

Světlé odstíny značí pozitivní korelace, tmavé ty negativní. Z grafu je patrné, že srdeční choroba má vysokou negativní korelaci s hodnotami maximální tepové frekvence. S rostoucí maximální tepovou frekvencí se tedy pravděpodobnost výskytu srdeční choroby snižuje. Naopak u hodnot Oldpeak a krevního cukru je korelace se srdeční chorobou pozitivní. Srdeční choroba je tedy s růstem hodnot těchto atributů pravděpodobnější.



David Mašek (masd04)
Viktor Sívek (sivv01)

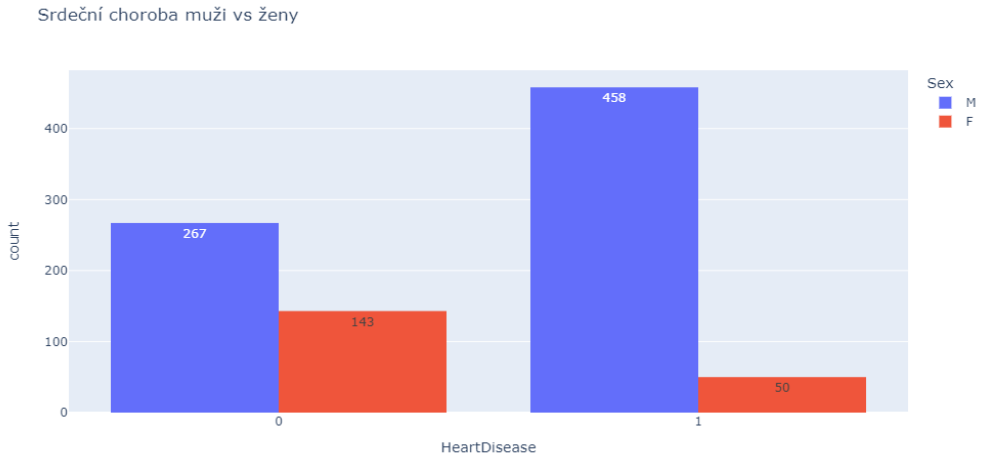
Dále jsme vytvořili grafy pro korelace mezi jednotlivými atributy pro lepší porozumění jejich vzájemným závislostem. U těchto grafů sledujeme rozptyl a směr korelace teček, přičemž modrá barva značí výskyt srdeční choroby a barva oranžová její absenci.



David Mašek (masd04)
Viktor Sívek (sivv01)

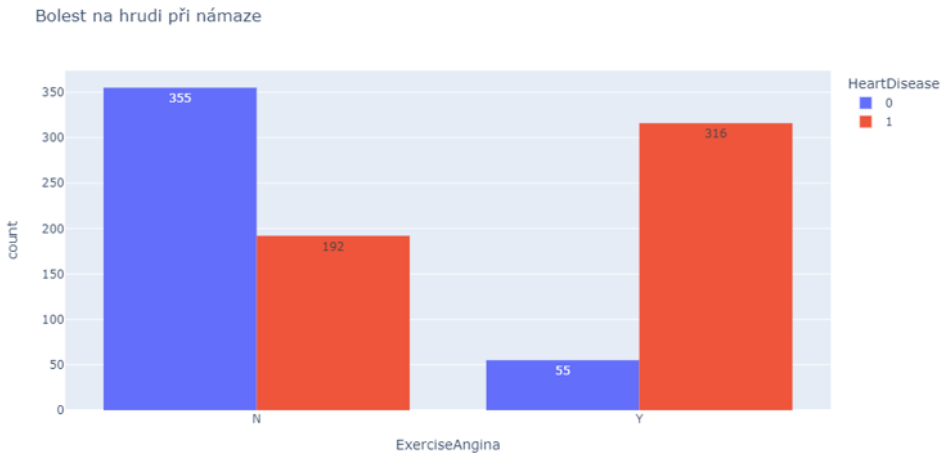
První si vizualizujeme jaké je rozdělení mužů a žen v datasetu a jak se u jednotlivých pohlaví vyskytuje srdeční choroba. Jak můžeme vidět v datasetu se vyskytuje převážně mužské pohlaví. Srdeční choroba se vyskytuje u 458 (63.17%) mužů a pouze u 50 (25.9%) žen z datasetu.

1) Odpověď na první otázku tudíž zní, že muži jsou náchylnější na výskyt srdečních chorob.



V následujícím grafu vidíme porovnání četnosti výskytu srdeční choroby u osob s a bez bolesti na hrudi při námaze. Z grafu je patrné, že v případě osob bez bolestí na hrudi při námaze se choroby srdce vyskytují výrazně méně (pouze u 35% subjektů). U sledovaných osob potýkajících se s bolestmi na hrudi při námaze nedošlo k výskytu srdeční choroby pouze u 55 (14.82%). V případě výskytu těchto bolestí je tedy poměrně vysoké riziko výskytu srdeční choroby. Nicméně v opačném případě, kdy se bolesti nevyskytují, rozhodně nemáme jistotu, že se nevyskytne choroba. Dle níže zobrazených výsledků je u osoby bez výskytu bolesti na hrudi při námaze „pouze“ 65% šance, že dožije bez srdeční choroby.

2) Při bolestech na hrudi bychom se o své srdce obavát rozhodně měli, nicméně ani bez výskytu tohoto typu bolesti nemáme absenci srdeční choroby zaručenu.

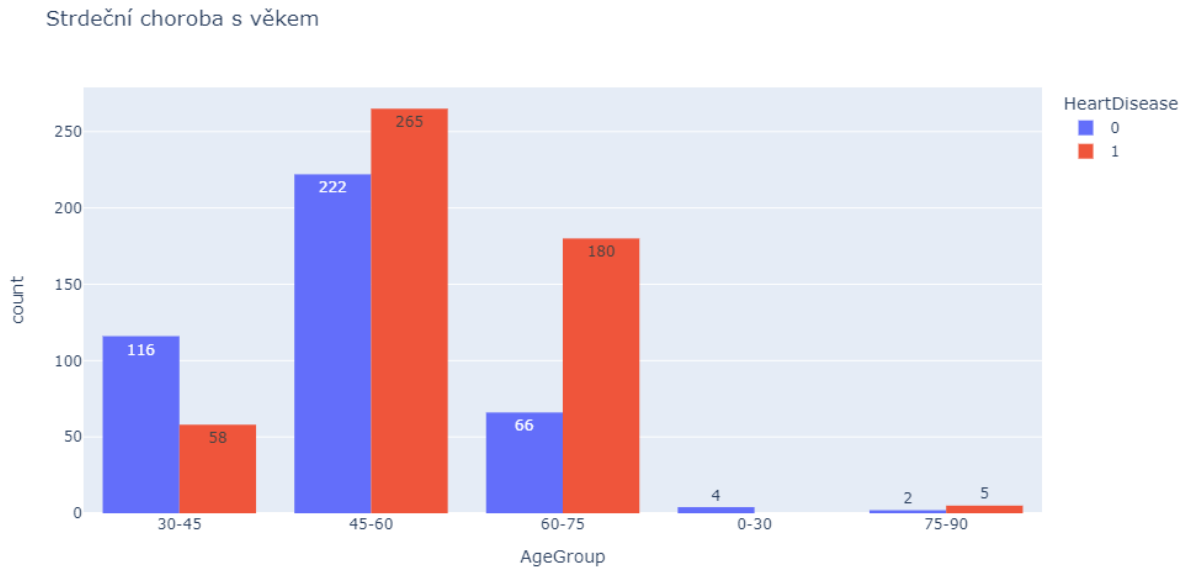


ExerciseAngina = Y	HeartDisease = 1	40.4140%	34.4230%	85.1750%	12.0590%	1.5392

V dalším grafu vidíme porovnání výskytu srdečních chorob u věkových skupin rozdělených do intervalů podle světové zdravotnické organizace (WHO). [15]

Z grafu je patrný nárůst četnosti srdečních chorob s rostoucím věkem, přičemž s největší pravděpodobností se choroba vyskytne mezi 60-75 lety. Při takto nízkém vzorku testovaných subjektů nelze považovat tento údaj za pravdivý, nicméně u testovaných subjektů mladších 30 let je 0% výskyt srdečních chorob.

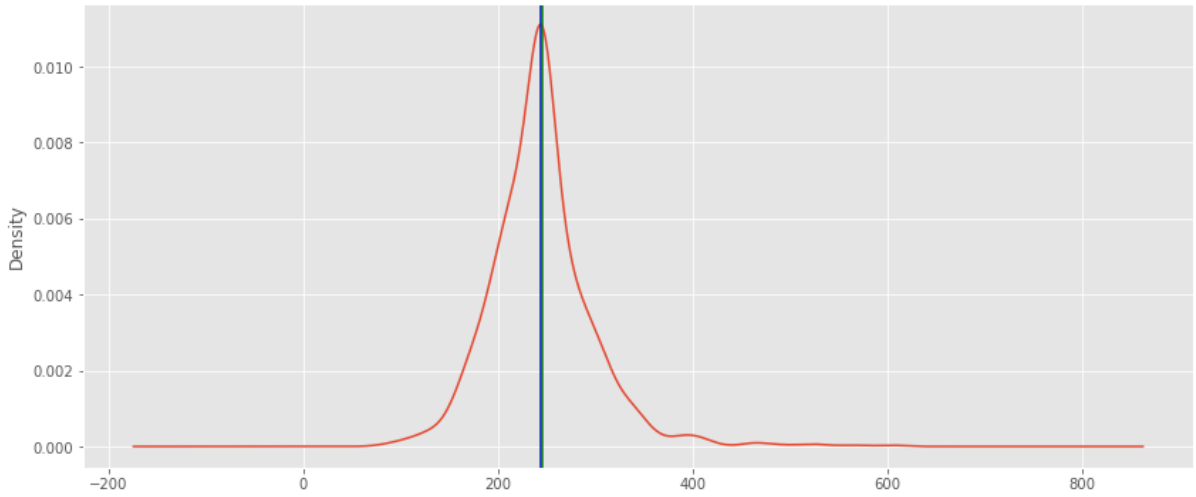
3) Nejnáchylnější věkovou kategorií jsou lidé mezi 60 a 75 lety života s pravděpodobností výskytu srdeční choroby 73.2%



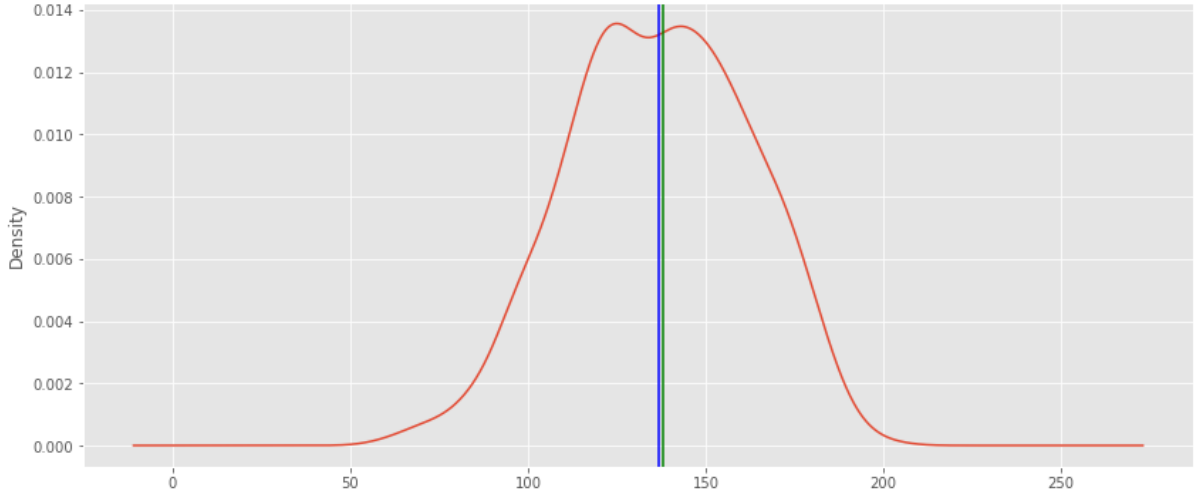
Age <= 44	HeartDisease = 0	19.3900%	13.0720%	67.4160%	4.4120%	1.5095

Pro lepší porozumění atributům cholesterol a Max HR si tyto atributy zobrazíme v grafu spolu s jejich průměrem a mediánem pro lepší představu o rozložení hodnot těchto atributů. Jak můžeme vidět hladina cholesterolu u subjektů se pohybuje přibližně převážně mezi 200-300 mmHg. A maximální tepová frekvence se pohybuje převážně mezi 100-175 bpm.

Cholesterol:



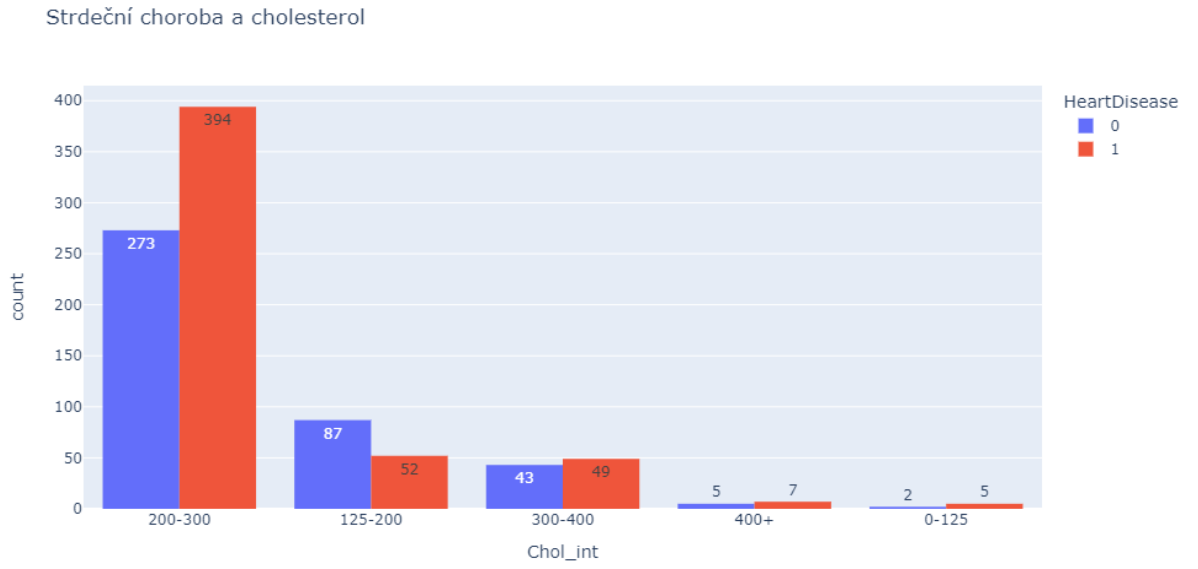
Max HR:



Dle stránky medicalnewstoday.com je zdravá hodnota cholesterolu u žen i mužů starších 20 let mezi 125 – 200 mg/dl. [16]

Subjektů se “zdravou” hodnotou cholesterolu máme v datasetu 139 a dle grafu je zřejmé, že subjekty s hodnotou v tomto intervalu mají opravdu nejnižší pravděpodobnost výskytu srdeční choroby. Při hodnotě cholesterolu mezi 200 - 300 mg/dl se riziko srdečního onemocnění zvyšuje a dle našich dat je při této hodnotě 59.1% pravděpodobnost srdečního onemocnění. U hodnot mezi 300 a 400 mg/dl je pravděpodobnost výskytu choroby 53.3%. Počet měřených subjektů je však poměrně nízký. Při hodnotách vyšších zůstává výskyt choroby pravděpodobnější, nicméně takto vysoká hodnota se u testovaných subjektů vyskytuje v příliš nízkém počtu.

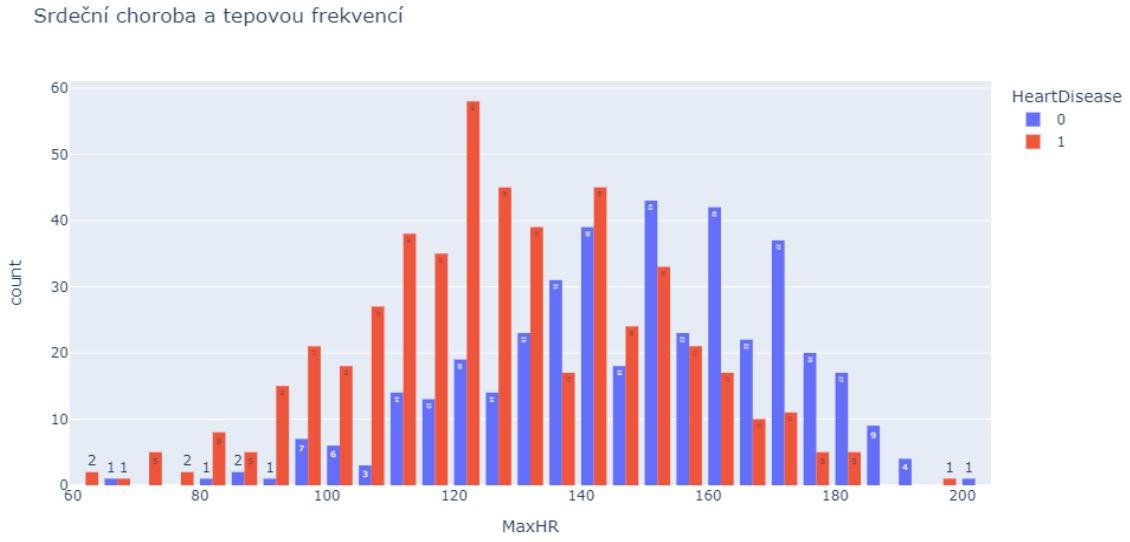
4) Při nedodržení zdravé hladiny cholesterolu v krvi se opravdu stavíme do rizika výskytu srdeční choroby.



234 < Cholesterol <= 248	HeartDisease = 1	26.6880%	19.9350%	74.6940%	5.1660%	1.3498

Níže vidíme graf zobrazující hodnoty maximální tepové frekvence v závislosti na výskytu srdeční choroby. Zhruba do 130 bpm je zřetelná poměrně vysoká šance na výskyt srdeční choroby. Při hodnotě vyšší, než 150 bpm se pravděpodobnost výskytu choroby postupně snižuje, počínaje na hodnotě lehce pod 50%.

5) Ano, srdce s nadprůměrnými hodnotami maximální tepové frekvence jsou dle našich dat odolnější vůči srdečním chorobám.

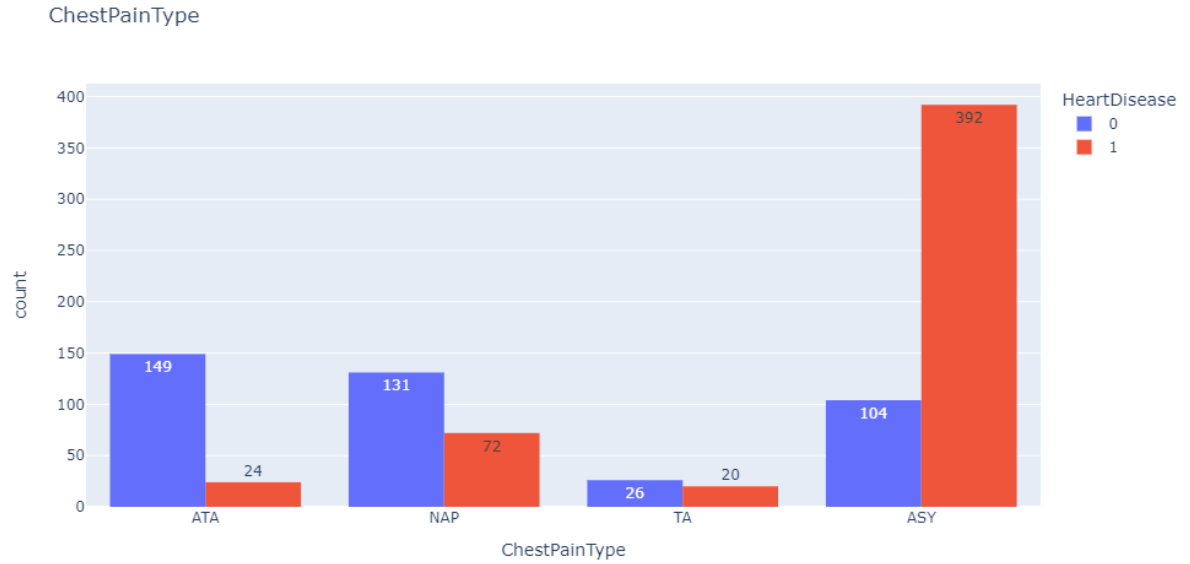


MaxHR <= 114	HeartDisease = 1	19.2810%	15.4680%	80.2260%	4.7990%	1.4498
114 < MaxHR <= 129	HeartDisease = 1	20.0440%	15.0330%	75.0000%	3.9410%	1.3553
MaxHR > 160	HeartDisease = 0	19.1720%	14.4880%	75.5680%	5.9250%	1.6920

David Mašek (masd04)
Viktor Sívek (sivv01)

Následující graf vizualizuje výskyt typů bolesti v závislosti na výskytu srdeční choroby. Zajímavým zjištěním je značně četnější výskyt srdečních chorob u osob bez příznaků bolesti na hrudi a to u 79% případů.

6) Dle našich dat je 79% subjektů bez bolesti na hrudi diagnostikováno se srdečním onemocněním. Bezpečnější tedy teoreticky je bolesti na hrudi pociťovat.

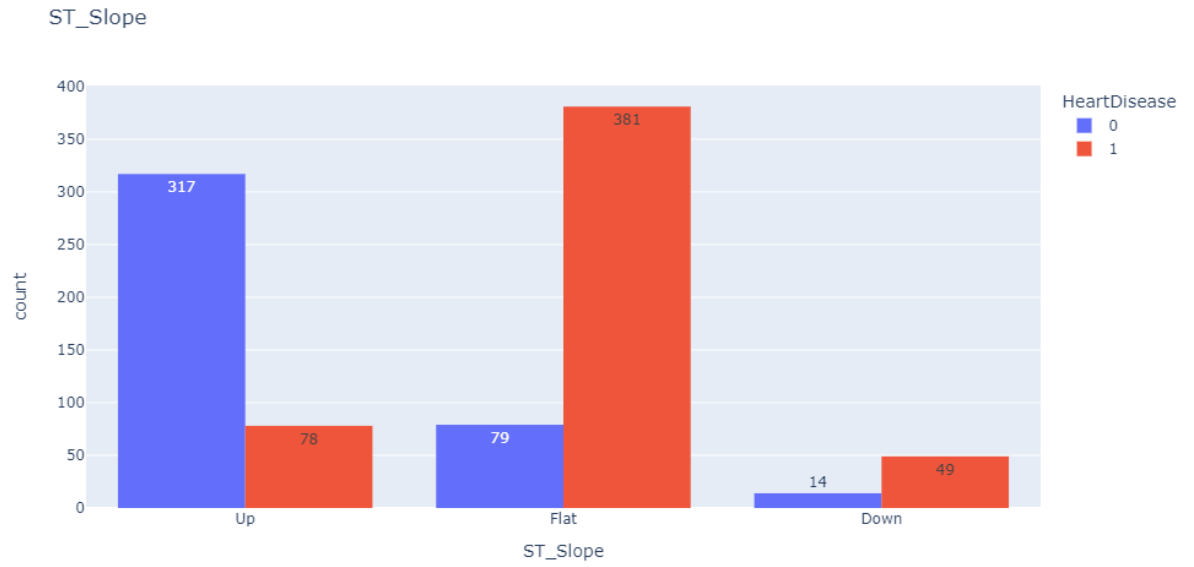


ChestPainType = ASY	HeartDisease = 1	54.0310%	42.7020%	79.0320%	12.8020%	1.4282

V dalším grafu je patrný nejčastější výskyt srdeční choroby u osob s hodnotou Flat u atributu ST_Slope. Nejbezpečnější shledáváme rostoucí křivku ST při maximální zátěži, tedy hodnotu Up. V tomto případě se 80.3% subjektů nepotýká se srdeční chorobou.

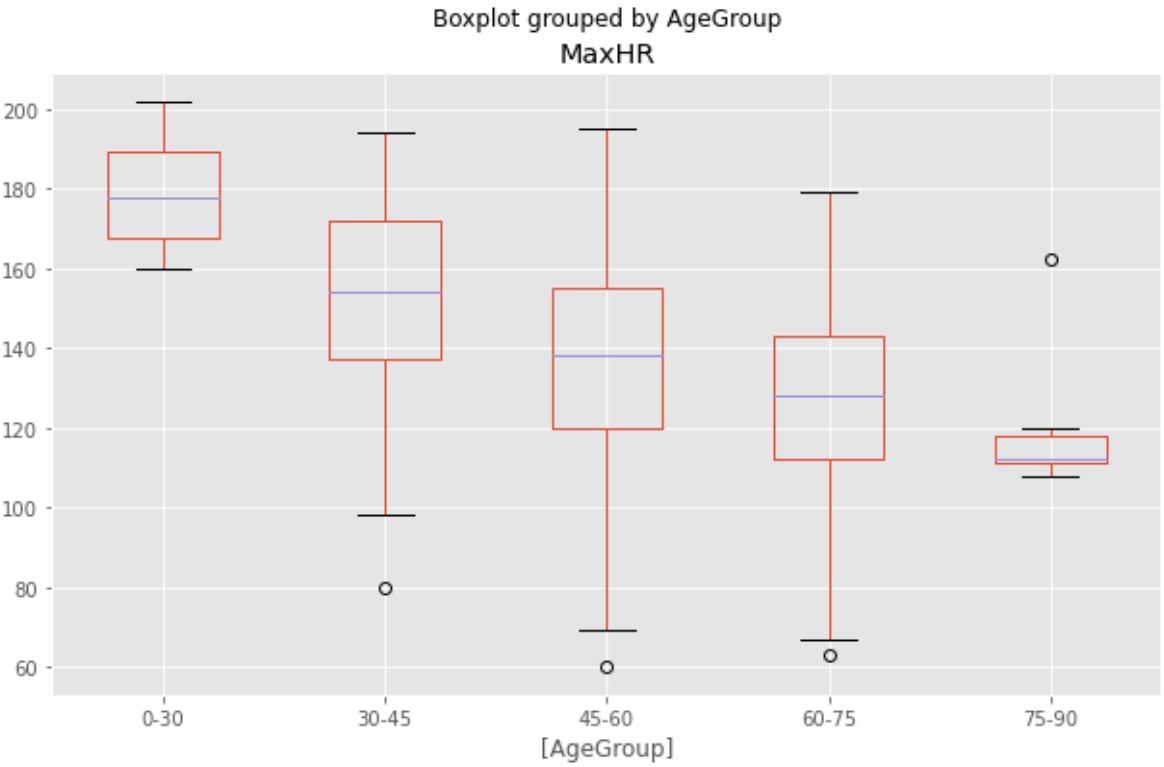
7) Srdeční onemocnění se nejčastěji vyskytují u subjektů s hodnotou ST_Slope = Flat

David Mašek (masd04)
Viktor Sívek (sivv01)

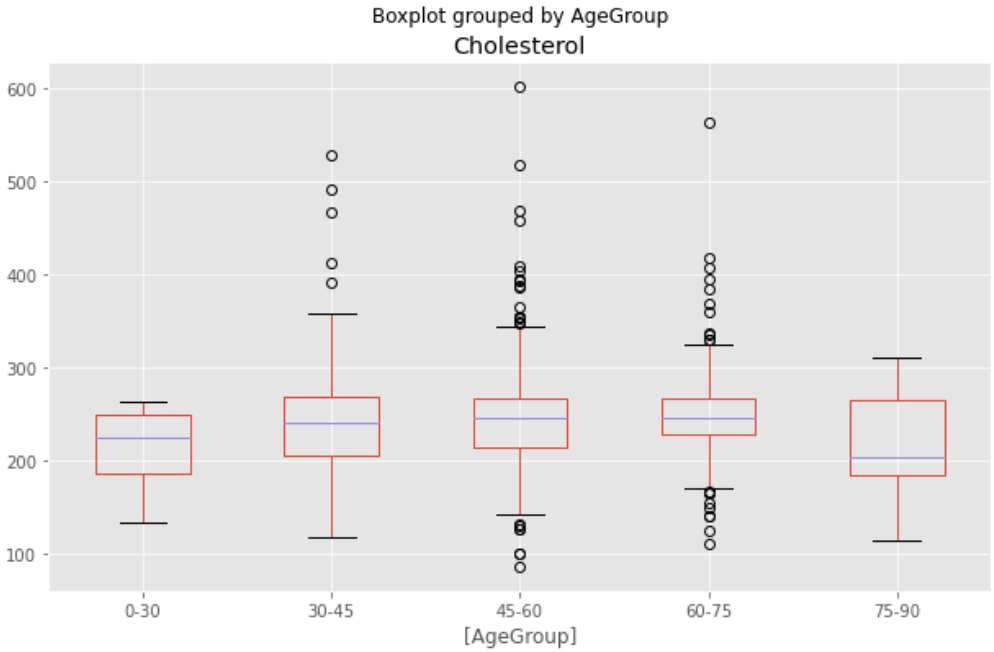


HeartDisease = 1	ST_Slope = Flat	55.3380%	41.5030%	75.0000%	13.7740%	1.4967
ST_Slope = Up	HeartDisease = 0	43.0280%	34.5320%	80.2530%	15.3140%	1.7969

V následujícím grafu vidíme rozsahy hodnot maximální tepové frekvence u předdefinovaných věkových kategorií. U každého z intervalů jsou obdélníkem vyznačeny nejčastěji se vyskytující hodnoty, přičemž modrá přímka značí hodnotu průměrnou. Výše maximálního tepu pochopitelně klesá s rostoucím věkem sledovaných subjektů.



Níže, obdobně jako v předešlém grafu, vidíme rozsahy hodnot naměřeného cholesterolu u věkových skupin. Zde se průměrný cholesterol v porovnání s předešlým grafem s věkem příliš nemění. Nejvíce hodnot vychýlených od průměru najdeme ve intervalu 45 - 60 let věku.



Další asociace nalezeny v BigML:

V programu BigML jsme pomocí funkce asociace nalezli možné hodnoty atributů související s výskytem srdeční choroby u testovaných subjektů.

8) Jako dvě hodnoty atributů nejpravděpodobněji související s výskytem srdečních onemocnění byla shledána asymtomická bolest na hrudi v klidovém režimu v kombinaci s hodnotou Flat u atributu ST_Slope
Další možnou kombinací vedoucí k srdeční chorobě je asymptomatická bolest na hrudi v kombinaci s bolestí na hrudi při zátěži.

ChestPainType = ASY ST_Slope = Flat	HeartDisease = 1	34.6410%	31.4810%	90.8810%	12.3120%	1.6423
ChestPainType = ASY ExerciseAngina = Y	HeartDisease = 1	32.3530%	29.1940%	90.2360%	11.2910%	1.6306

S lehce nižší pravděpodobností vede k srdečnímu onemocnění u mužů hodnota Flat u atributu ST_Slope a také maximální tepová frekvence v intervalu mezi 115 bpm a 129 bpm včetně.

Sex = M ST_Slope = Flat	HeartDisease = 1	41.9390%	37.2550%	88.8310%	14.0470%	1.6053
Sex = M 114 < MaxHR <= 129	HeartDisease = 1	17.2110%	14.4880%	84.1770%	4.9640%	1.5212

9) Naopak mezi kombinace související s absencí srdeční choroby řadíme s nejvyšší pravděpodobností hodnotu Up u atributu ST_Slope společně s atypickou bolestí na hrudi.

ST_Slope = Up ChestPainType = ATA	HeartDisease = 0	15.0330%	14.4880%	96.3770%	7.7740%	2.1579

Dále jsme nalezli kombinace hodnoty Up u atributu ST_Slope s absencí bolesti na hrudi při zátěži a také v kombinaci s normální hodnotou cukru v krvi.

ExerciseAngina = N ST_Slope = Up	HeartDisease = 0	36.7100%	31.6990%	86.3500%	15.3040%	1.9334
FastingBS = 0 ST_Slope = Up	HeartDisease = 0	36.3830%	31.0460%	85.3290%	14.7960%	1.9105

Závěr

Pro práci jsme zvolili postup dle alternativního zadání z důvodu výpadku platformy Lotylda. Vizualizaci dat jsme prováděli pomocí jazyka Python a pro zjištění asociací jsme použili platformu BigML. Původně zvolenou analytickou otázku jsme se rozhodli nahradit dílčími otázkami, které jsme se dle našich nejlepších schopností snažili s co největší přesností zodpovědět.

Obrázky

Obrazový materiál neuvádíme ve zdrojích z důvodu naší vlastní produkce obrázků pomocí jazyka Python.

Citace

[1] CHICCO, Davide a Giuseppe JURMAN, 2020. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making* [online]. **20**(1), 16 [vid. 2022-03-24]. ISSN 1472-6947. Dostupné z: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5

[2] Anon., 2017. Angina. *nhs.uk* [online] [vid. 2022-03-24]. Dostupné z: https://www.nhs.uk/conditions/angina/

[3] Anon., 2020. Atypical Chest Pain. *Premier Pain & Spine* [online]. [vid. 2022-03-24]. Dostupné z: https://www.ppschicago.com/pain-management/chest-pain/atypical-chest-pain/

[4] Anon., [b.r.]. Non-Cardiac Chest Pain. *Cleveland Clinic* [online] [vid. 2022-03-24]. Dostupné z: https://my.clevelandclinic.org/health/diseases/15851-gerd-non-cardiac-chest-pain

[5] SR, MEFANET, síť lékařských fakult ČR a, [b.r.]. *Krevní tlak – WikiSkripta* [online] [vid. 2022-03-23]. Dostupné z: https://www.wikiskripta.eu/w/Krevn%C3%AD_tlak

[6] Anon., [b.r.]. *Systolický krevní tlak – WikiSkripta* [online] [vid. 2022-03-23]. Dostupné z: https://www.wikiskripta.eu/w/Systolick%C3%BD_krevn%C3%AD_tlak

[7] Anon., [b.r.]. *Vysoký cholesterol – příčiny, léčba a prevence* [online] [vid. 2022-03-24]. Dostupné z: https://euc.cz/clanky-a-novinky/clanky/vysoky-cholesterol-priciny-lecba-a-prevence/

[8] CDC, 2022. Diabetes Testing. *Centers for Disease Control and Prevention* [online] [vid. 2022-03-24]. Dostupné z: https://www.cdc.gov/diabetes/basics/getting-tested.html

[9] Anon., [b.r.]. Index of /ml/machine-learning-databases/heart-disease [online] [vid. 2022-03-24]. Dostupné z: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

[10] ŠILAR, Jiří, [b.r.]. DETEKCE PARAMETRŮ POPISUJÍCÍ TVAR T VLNY SIGNÁLU EKG A ZÁKLADNÍ ANALÝZA VLASTNOSTÍ TĚCHTO PARAMETRŮ [online]. 52. Dostupné z: https://is.muni.cz/th/bsk1a/bp_silar.pdf

[11] VESELKA, Josef, [b.r.]. Hypertrofická kardiomyopatie [online]. 77. Dostupné z: https://www.kardio-cz.cz/data/clanek/423/dokumenty/473-veselka-hypertrofickakardiomyopatie.pdf

[12] Anon., 2011. Hypertrofie levé komory - EKG. *Medicína, nemoci, studium na 1. LF UK* [online] [vid. 2022-03-24]. Dostupné z: https://www.stefajir.cz/hypertrofie-leve-komory-ekg

[13] BURNS, Ed a Robert BUTTNER, 2018. Left Ventricular Hypertrophy (LVH). *Life in the Fast Lane • LITFL* [online]. [vid. 2022-03-24]. Dostupné z: https://litfl.com/left-ventricular-hypertrophy-lvh-ecg-library/

[14] CHELLAMMAL, S a R SHARMILA, 2019. Recommendation of Attributes for Heart Disease Prediction using Correlation Measure. **8**(2), 6.

[15] Anon., 2022. *Ontogeneze člověka* [online]. [vid. 2022-05-08]. Dostupné z: https://cs.wikipedia.org/w/index.php?title=Ontogeneze_%C4%8Dlov%C4%9Bka&oldid=21224622

[16] Anon., 2021. *Cholesterol levels by age: Health ranges, what is high, and tips* [online] [vid. 2022-05-08]. Dostupné z: https://www.medicalnewstoday.com/articles/315900