Na začiatku projektu bolo nutné zadefinovať knižnice, ktoré budú využívané pri exploratívnej analýze.

Obrázok, na ktorom je text

Automaticky generovaný popis

Databázu údajov sme vyhľadali na stránke www.kaggle.com v kategórii klasifikácia dát. Po analýze databáz, sme došli k záveru, že dáta vhodné pre našu databázu budú z oblasti medicíny (presnejšie dáta o mŕtvici). Tieto dáta sme si stiahli a následne ich importovali do premennej ako dataframe (dátový rámec)



Po načítaní dát, z súboru formátu csv, sme si overili, že sa nám dáta správne načítali.  
Zobrazili sme prvých a posledných päť záznamov z databázy.  
Vidíme, že záznamov našej databáze je 5110.

Obrázok, na ktorom je stôl

Automaticky generovaný popis

V databáze mohol nastať prípad chýbajúceho záznamu a v tomto prípade, je nutné všetky záznamy (riadky) odstrániť. Robíme to z dôvodu skvalitňovania dát a vyššej dôveryhodnosti údajov, s ktorými je v tomto projekte pracované. Pri výstup sme si všimli, že z 5110 dát sa nachádza nedostatok záznamov pri premennej (stĺpci) bmi. Po preskúmaní dát, sme si uvedomili nadbytočnosť údaju ever-married nakoľko tento údaj sme nepovažovali za potencionálne ovplyvnenie príčiny mŕtvice.

Obrázok, na ktorom je text

Automaticky generovaný popis

Údaj o manželstve bol odstránený a následne sme vypísali dátový rámec ponížený o tento záznam.

Obrázok, na ktorom je text

Automaticky generovaný popis

Následne bol analyzovaný počet záznamov v jednotlivých premenných, kde bola chýbajúca hodnota. Po vypísaní presných hodnôt sme zistili, že pri bmi je chýbajúcich 201 záznamov. Pre úplnosť a vyššiu kvalitu výpočtov, sme tieto údaje odstránili.

Obrázok, na ktorom je stôl

Automaticky generovaný popis Obrázok, na ktorom je stôl

Automaticky generovaný popis

Po vymazaní prázdnych záznamov sme si vypísali aktuálny popis údajov, ktoré sa v dátovom rámci nachádzajú. Vidíme maticu, v ktorej sa nachádzajú údaje, z ktorých vieme vyčítať napríklad : maximálny vek pri zázname (82), priemerná úroveň cukru (105.3), stredné hodnoty pri bmi (28.1), počet záznamov (4909) a mnoho iných použiteľných údajov

Obrázok, na ktorom je stôl

Automaticky generovaný popis

Po analyzovaní dát, sme uznali za vhodné rozdeliť si premenné na číselné a slovné. Vďaka tomuto rozdeleniu vieme následne presnejšie pracovať s jednotlivými záznamami. Pomôže nám to aj jednoduchšie zobraziť a zorientovať sa v jednotlivých hodnotách.  
Môžeme si všimnúť rozdiel, kde číselné premenné sú vo väčšom množstve a vypísané boli pod seba – slovné boli len štyri a zobrazilo ich v jednom riadku.

Obrázok, na ktorom je text

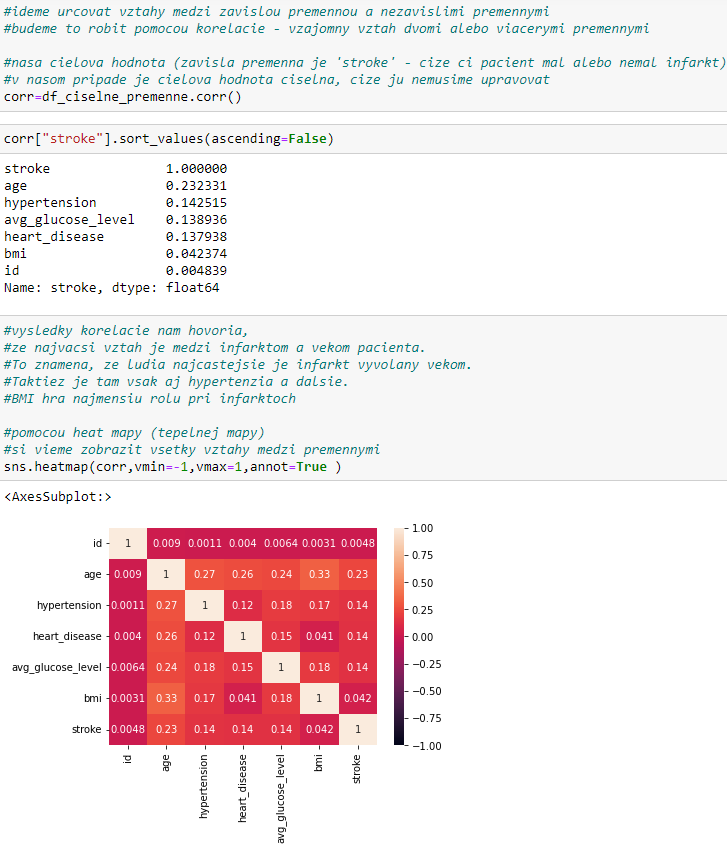
Automaticky generovaný popis

Z polí obsahujúce slovné a číselné premenné, vytvoríme dva dátové rámce, ktoré budú obsahovať ich hodnoty. Následne si ich vypíšeme, aby sme potvrdili pravdivosť úkonu.  
Môžeme si všimnúť, že jednotlivé záznamy nie sú v číselnom poradí (odstránili sme prázdne hodnoty z bmi) a oba dátové rámce majú rozdielne hodnoty stĺpcov.

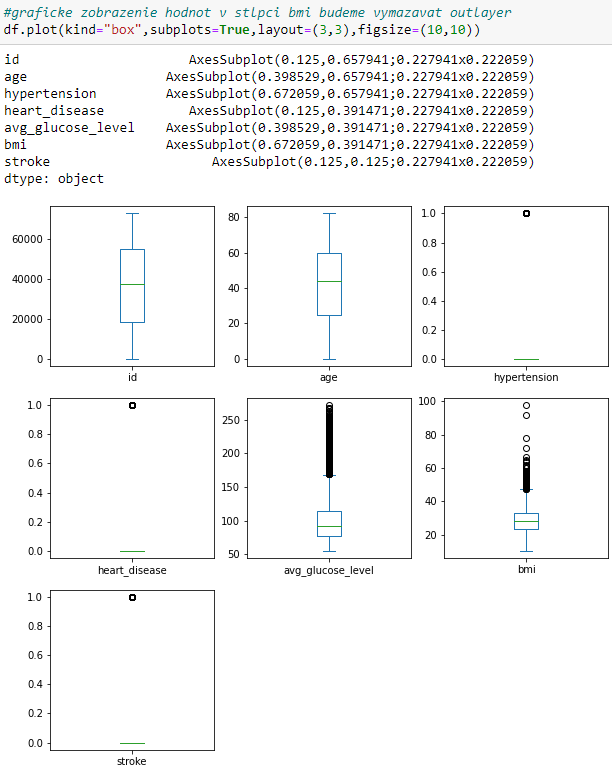
Obrázok, na ktorom je stôl

Automaticky generovaný popis

Po rozčlenení premenných a naplnení dátového rámcu vytvoríme korelačnú maticu. Táto matica nám vraví o vzájomnom vzťahu resp. ovplyvňovaní alebo dopadu hodnôt, jednotlivých premenných medzi sebou. Po vypočítaní hodnôt korelácie, sme jednotlivé premenné zoradili zostupne zamietnutím podmienky ascending (vzostupne = nepravda). Získané údaje sme vypísali pomocou nástroja na zobrazovanie – heatmap “tepelná mapa“. V korelačnej matici je na diagonále vždy číslo 1, čo je maximálna kladná korelácia aká sa dá všeobecne dosiahnuť.



Následne keď sme zanalyzovali koreláciu, môžeme pristúpiť k zobrazeniu grafov pre jednotlivé premenné. Grafy nám zobrazujú hodnoty min, 25%, 50%, 75%, max a outliery. Jednotlivé prázdne krúžky predstavujú hodnoty mimo hraníc grafu (kvarlitov) - outlayeri.  
Boxplot je teda rozdelený na štyri kvartily po 25% so strednou hodnotou predstavujúcu čiaru v strede. Tento graf nám zachytáva všetky hodnoty.

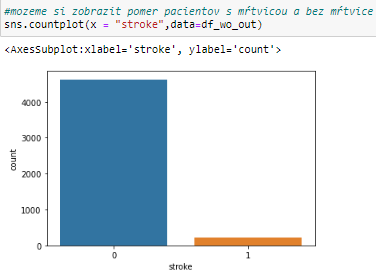
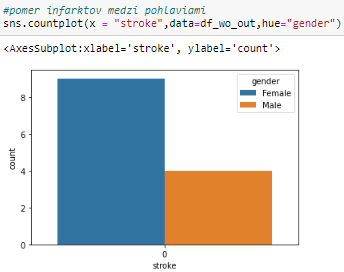


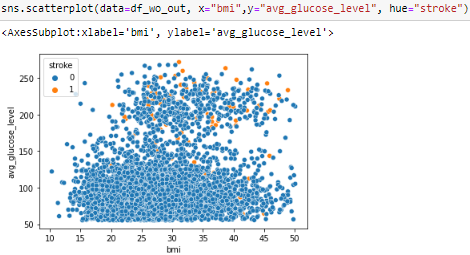
Na vyššie uvedených grafoch môžeme vidieť veľké rozpätie hodnôt v outlieroch pri bmi. Aby sme dosiahli vyššiu kvalitu údajov odstránime outlieri, ktoré sa nachádzajú mimo hlavú skupinu outlierov. Odstránene realizujeme znížením resp. vymazaním top 1.5% hodnôt z bmi.

Obrázok, na ktorom je text

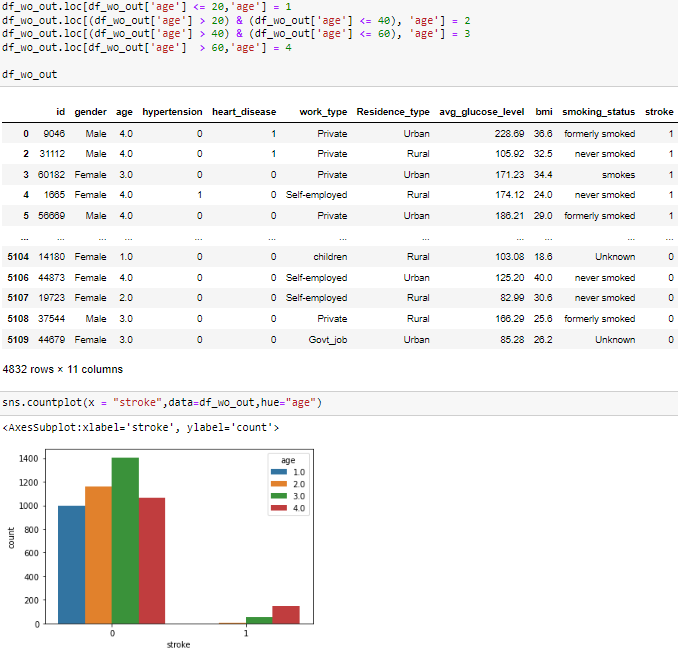
Automaticky generovaný popis

Po odstránení hodnôt, sme pristúpili k vykresleniu grafov pre : početnosť mŕtvic a bez ochorenia, pomer mŕtvic v závislosti s pohlaviami medzi mužmi a ženami a jednotlivé mŕtvice pri porovnaní v prípade bmi s priemernou úrovňou cukru v tele.



Posledným grafom pred prechodom na trénovaciu a testovaciu časť, bolo zobrazenie počtu participantov v prieskume. Jednotlivé vekové skupiny sme rozdelili do skupín po intervaloch dvadsiatich rokov. Následne sme tieto intervaly vpísali do dátového rámcu a zobrazili graf, ktorý porovnáva vekové skupiny s počtom mŕtvic resp. zdravých ľudí.



Po ukončení exploratívnej analýzy sme začali s procesom testovania a trénovania dát.  
Prvým krokom bolo zadefinovanie premenných X a y.   
X je premenná obsahujúca údaje, na základe ktorých sa budú dáta trénovať a testovať.  
y je nám značí o výsledku, či pacient mal alebo nemal mŕtvicu.  
Z knižnice sklearn použijeme knižnicu train\_test\_split, vďaka ktorej sa nám dáta rozdelia na testovacie a trénovacie X. Preddefinovaný pomer sme nechali na hodnote 75/25 resp. 75% dát bude pridelených do trénovacej časti a 25% dát bude priradených testovacej časti.  
Veľkosť vidíme na základe príkazu na vypísanie vzniknutých premenných.  
Obrázok, na ktorom je text

Automaticky generovaný popis

Zobrazíme trénovaciu a testovaciu časť. Vidíme, že dáta sú v defaultnom pomere 75/25.

Obrázok, na ktorom je stôl

Automaticky generovaný popis Obrázok, na ktorom je stôl

Automaticky generovaný popis

Attribute Information

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST\_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

<https://holypython.com/svm/support-vector-machine-step-by-step/>

<https://data36.com/random-forest-in-python/>