WORD PROJEKT 1

Meno študenta: Samuel Weis  
Trieda: 3. ročník  
Dátum: 16.9 2025

Úvod

Analyzovaný dataset obsahuje 303 záznamy pacientov so 14 atribútmi popisujúcimi rôzne klinické parametre, ako sú vek, pohlavie, typ bolesti na hrudníku (cp), krvný tlak v pokoji (trestbps), hladina cholesterolu (chol) a ďalšie. Primárnym cieľom tohto datasetu je predpovedať prítomnosť srdcového ochorenia (cieľový atribút 'target', kde 0 znamená absenciu a 1 prítomnosť ochorenia). Data slúžia ako štandardný benchmark pre binárnu klasifikáciu v medicíne.

Postup

Data boli načítané a následne bola vykonaná exploratívna analýza. Boli vytvorené kľúčové vizualizácie na pochopenie distribúcie atribútov a ich vzťahu k cieľovej premennej. Všetky grafy boli vygenerované pomocou knižnice matplotlib v jazyku Python.

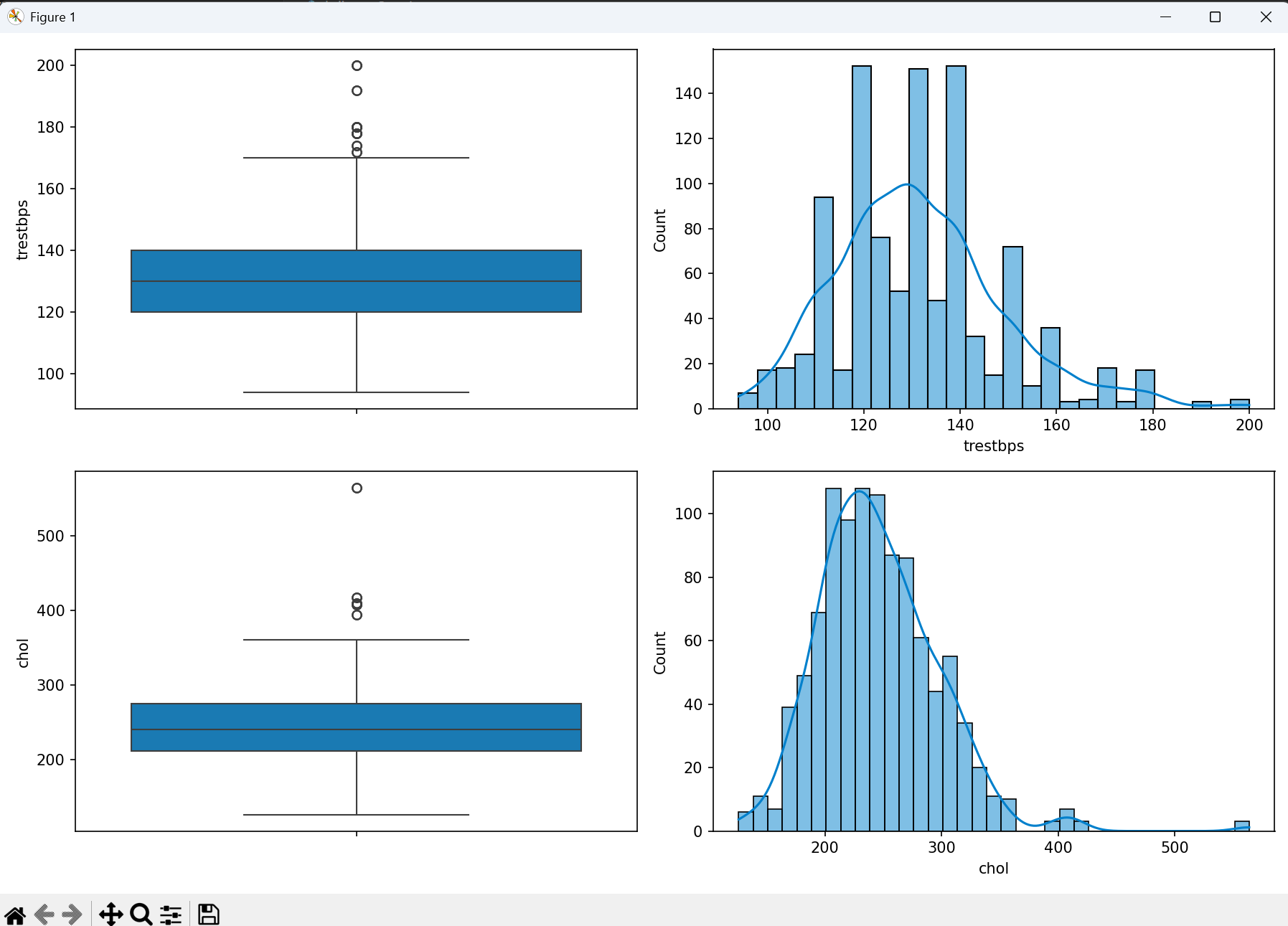
Odľahlé hodnoty boli identifikované metódou IQR. Capovanie ich obmedzilo na horný limit **170 pre trestbps** a **371 pre chol**.

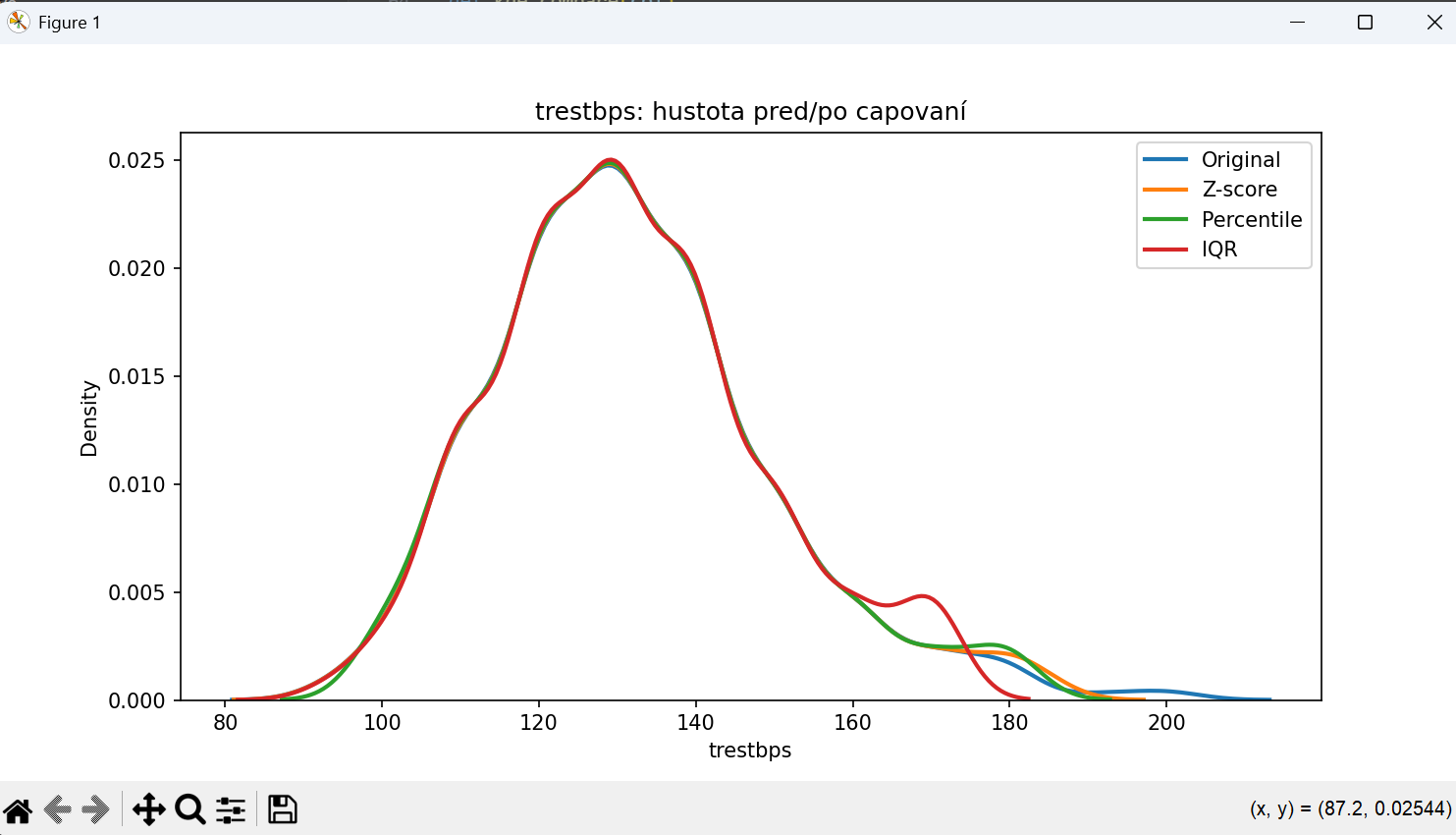
**Zmena mediánu a rozptylu po capovaní:**  
Medián **nezmenil** sa v žiadnom prípade (ostal rovnaký), pretože capovanie ovplyvnilo iba extrémne hodnoty na okrajoch rozdelenia. Rozptyl sa **mierne znížil** – pri trestbps z pôvodných **17,52** na **16,53** a pri chol z **51,59** na **47,75**, čo indikuje menšiu variabilitu dát po odstránení odľahlých hodnôt.

**Ďalšie použitie upravených dát:**  
Upravené dáta by som použil na trénovanie **jednoduchých modelov strojového učenia**, ako je:

**Lineárna regresia** (pre predikciu číselných hodnôt),

**Logistická regresia** alebo **KNN** (pre klasifikáciu ochorení),

**Rozhodovacie stromy** (odolné voči odľahlým hodnotám).



Capovanie znižuje vplyv šumu na model, zlepšuje stabilitu a zvyšuje interpretovateľnosť výsledkov, čo je kľúčové pre klinické rozhodovanie.

V stĺpci trestbps bolo 49 odľahlých hodnôt (4,78 %) a v chol 61 (5,95 %). Najvhodnejšia IQR metóda capovania odstránila extrémne hodnoty (nad 170 pre tlak, nad 371 pre cholesterol) bez zmeny mediánu, ale so znížením rozptylu. Takto upravené dáta sú vhodné pre modely strojového učenia (logistická regresia, KNN), ktoré budú robustnejšie a interpretovateľnejšie.

Kod

<hdisease3.py>

database

[archive\heart.csv](archive/heart.csv)