VJEŽBA 4: EKSPLORATIVNA ANALIZA PODATAKA

I. Cilj vježbe:

Upoznati se sa statističkom analizom te metodom vizualizacije skupa izmjerenih podataka

II. Opis vježbe:

Prije postupka modeliranja na temelju skupa izmjerenih podataka, moguće je vršiti eksplorativnu analizu nad skupom podataka. Eksplorativna analiza podataka je postupak analize skupa podataka pri čemu se dobiju sažete glavne (statističke) karakteristike skupa često vizualno odnosno grafičko prikazane. Na taj je način moguće dobiti više uvida u informativnost te međusobne odnose izmjerenih podataka prije nego što se krene u postupak modeliranja.
U ovoj će se vježbi naučiti neki od postupaka eksplorativne analize podataka u Python-u. Vršit će se statističke kalkulacije pomoću *statistics* Python biblioteke te grafička analiza raznih grafova dobivenih pomoću biblioteka poput *matplotlib-a*.

III. Rad na vježbi:

Preuzeti datoteku *irisdata.mat* s Loomen stranice predmeta (https://loomen.carnet.hr/course/view.php?id=6207). U njoj se opisuju biometrijske značajke cvjetova (duljina i širina čašice odnosno latice) triju vrsta perunike odnosno irisa (setosa – 1, versicolor – 2 i virginica -3). Detaljan opis može se pronaći u datoteci *irisdata.txt*.

Postupak	PYTHON biblioteka	Funkcija
Određivanje mjere centralne tendencije	`statistics, numpy`	`mean, median, mode, percentile`
Određivanje mjere rasipanja	`statstics`	`variance,stddev`
Crtanje histograma	`matplotlib`	`hist`
Određivanje empirijske kumulativne funkcije distribucije	`numpy`	`cumsum, histogram`
Box-plot	`matplotlib`	`boxplot`
Dijagram raspršenja ili rasipanja (scatter plot)	`mlxtend.plotting, matplotlib`	`scatter hist, scatter`

IV. Zadatak:

Preuzeti datoteku *bodydata.mat*. Datoteka sadrži mjerenja različitih antropometrijskih značajki poput visine, težine, promjera i opsega različitih dijelova tijela fizički aktivnih osoba. Ukupno su na raspolaganju mjerenja izvršena na 507 osoba (od toga je 247 muškaraca, a 260 žena) pri čemu je mjereno 17 različitih značajki. Detaljan opis može se pronaći u datoteci *bodydata.txt*.

1. Odrediti srednje vrijednosti, median i standardnu devijaciju pojedinačno za sve značajke grupirane po spolu. Komentirati dobivene vrijednosti.
2. Pomoću histograma usporediti sve značajke grupirane po spolu tako da se na istoj slici prikazuju dva histograma određene značajke za oba spola. Komentirati dobivene slike.
3. Odrediti i nacrtati empirijsku kumulativnu funkciju distribucije za sve značajke tako da se na istoj slici uspoređuje značajka po spolovima. Naznačiti na svim slikama 1. i 3. kvartil. Komentirati dobivene slike.
4. Nacrtati *box-plot* za sve značajke po spolovima. Komentirati dobivene slike.
5. Nacrtajte 2D dijagrame rasipanja (s histogramima pri rubovima dijagrama) za parove značajki grupirane po spolu:
   - težina – razmak između ramena;
   - težina – opseg prsa;
   - težina – opseg struka;
   - težina – opseg struka oko pupka;
   - visina – razmak između ramena;
   - visina – opseg prsa;
   - visina – opseg struka;
   - visina – opseg struka oko pupka;

   Komentirati dobivene slike.