Podatkovno rudarjenje in odkrivanje zakonitosti v podatkovnih bazah

Ljupčo Todorovski

Univerza v Ljubljani, Fakulteta za upravo
Institut Jožef Stefan, Odsek za tehnologije znanja (E-8)

April 2020

Pregled predavanja

Podatkovno rudarjenje (*data mining*)

- Definicije in povezani pojmi
- Proces podatkovnega rudarjenja
- Opis posameznih faz procesa

Primeri in težave podatkovnega rudarjenja.

Definicije

Friedman (ESL)

Računalniško-podprta, raziskovalna analiza podatkov v (običajno) velikih, kompleksnih podatkovnih množicah.

Witten (Weka)

- Pristop k reševanju problemov z analizo podatkov shranjenih v podatkovnih bazah.
- Odkrivanje implicitnih, predhodno neznanih in potencialno uporabnih informacij iz podatkov.

Aggarwal

Proučevanje zbiranja, čiščenja, obdelave, analize in pridobivanja koristnega znanja iz podatkov.

Podatkovno rudarjenje in strojno učenje

Strojno učenje

Proučevanje algoritmov, ki izboljšujejo svojo uspešnost iz izkušenj.

- Učenje je izboljševanje uspešnosti v nekem okolju skozi pridobivanje znanja, ki je rezultat izkušenj v tem okolju.

Povezava s podatkovnim rudarjenjem

Strojno učenje je bistveni korak podatkovnega rudarjenja, saj algoritmi strojnega učenja omogočajo odkrivanje vzorcev in modelov iz podatkov.

Razlika od podatkovnega rudarjenja

Fokus na algoritmih za učenje in ne na odkrivanju vzorcev in modelov.

Odkrivanje zakonitosti v bazah podatkov

KDD: Knowledge Discovery in Databases

Definicija

Množica tehnologij za odkrivanje ne-trivialnih, implicitnih, predhodno neznanih in potencialno koristnih informacij iz podatkov shranjenih v podatkovnih bazah.

Odkrite informacije so običajno vzorci in modeli, ki omogočajo

1. boljše razumevanje podatkov,
2. napovedovanje bodočega obnašanja sistemov in
3. boljše odločanje.

Povezava s podatkovnim rudarjenjem

- Zajema celotni podatkovni cikel in ne le odkrivanje vzorcev in modelov
- Številni avtorji uporabljata pojma kot sinonima

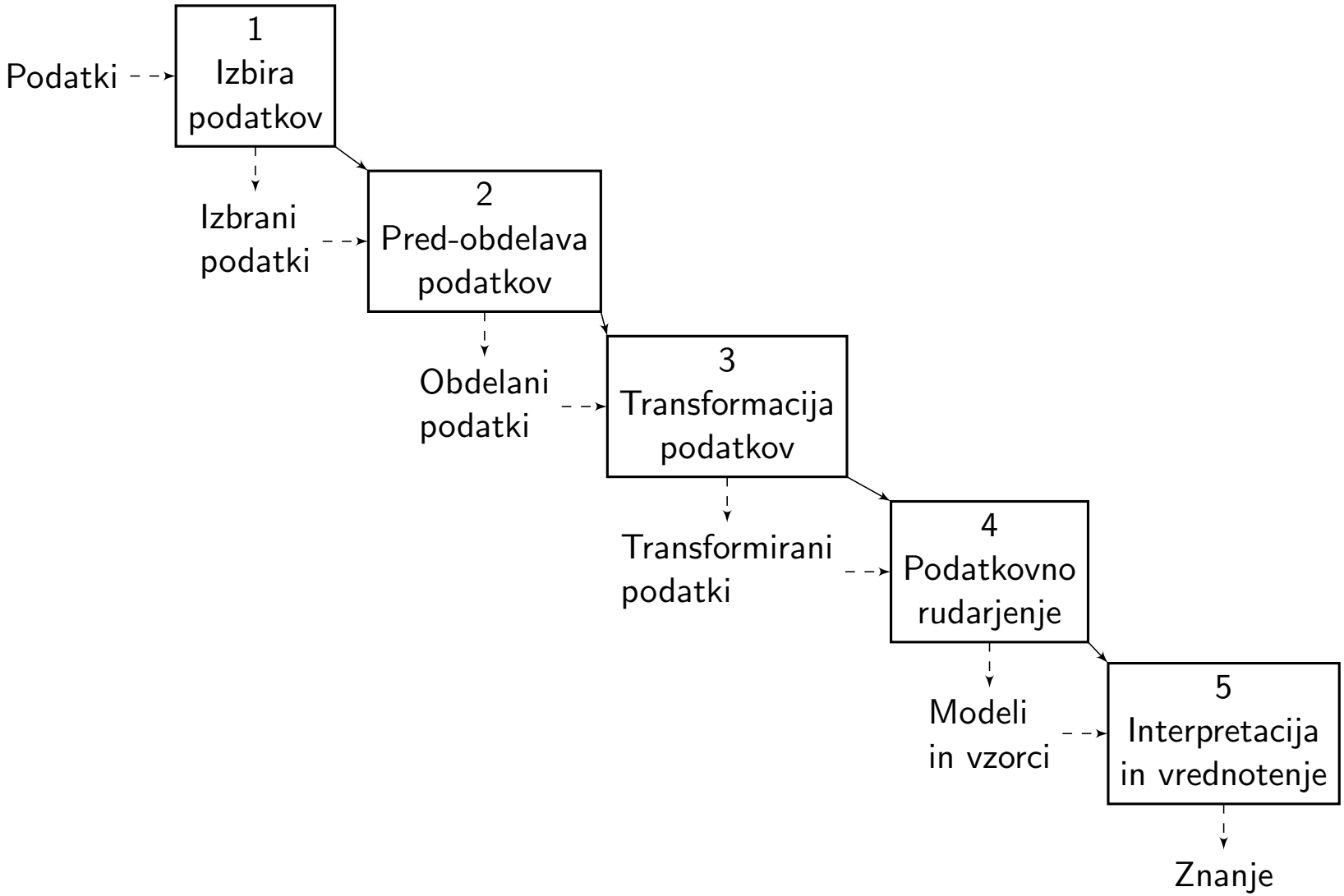
Proces odkrivanja zakonitosti v bazah podatkov

KDD je kompleksen proces izbire, priprave, obdelave in rudarjenja podatkov ter vrednotenja in uporabe odkritih modelov in vzorcev.

Koraki procesa *KDD*

1. *Selection*: izbira podatkov
2. *Preprocessing*: pred-obdelava podatkov
3. *Transformation*: transformacija podatkov
4. *Data mining*: podatkovno rudarjenje, odkrivanje vzorcev in modelov
5. *Interpretation and evaluation*: interpretacija in vrednotenje vzorcev in modelov ter njihova pretvorba v znanje

Koraki procesa *KDD*



1 Izbira podatkov

Ključno vprašanje

Kateri podatki so relevantni za problem, ki ga rešujem?

Tudi zbiranje podatkov

- Posebna strojna oprema: senzorji
- Posebna programska oprema: pajki
- Ročno: uporabniške ankete

Pomembna vloga področnih ekspertov.

2 Pred-obdelava podatkov

Ključna naloga

Priprava podatkov v obliko, ki je ustrezna za analizo.

Običajna oblika je tabela

- Vrstice: primeri
- Stolpci: spremenljivke, tudi značilke (*features*) ali atributi

Običajne težave

- Integracija različnih podatkovnih virov
- Ne-strukturirani ali pol-strukturirani podatkovni tipi, predavanje *Obravnava različnih tipov podatkov in vložitve*

3 Transformacija podatkov

Tri ključne naloge

Čiščenje podatkov

- Manjkajoči in napačni podatki
- Predavanje *Obravnava nepopolnih podatkov*

Izbira in transformacija ciljnih in napovednih spremenljivk

Predavanje *Izbira in konstrukcija napovednih spremenljivk*

Vzorčenje in obteževanje primerov

- Izbira primerov za učne in testne podatkovne množice
- Obteževanje primerov pri neenakomernih porazdelitvah
- Predavanje *Neenakomerna porazdelitev vrednosti ciljne spremenljivke*

4 Podatkovno rudarjenje

Ključni nalogi

- Učenje napovednih modelov iz podatkov
- Odkrivanje vzorcev v podatkih

Uporaba metod strojnega učenja

Kreativna uporaba in kombiniranje metod za učenje napovednih modelov in odkrivanje vzorcev iz prvega dela semestra.

5 Interpretacija in vrednotenje

Ključna vprašanja

- Kakšen je pomen modelov v kontekstu reševanja problema?
- Kakšen je pomen vzorcev v istem kontekstu?
- Kakšen je prispevek modelov in vzorcev k znanju s področja?

Pomembna vloga domenskih ekspertov.

Naloga

Faza 2: Pred-obdelava podatkov

Običajni problem razvrščanja, učna množica

- Primeri so besedila
- Ciljna spremenljivka: razred (oznaka) besedila
- Vhodne spremenljivke: opis besedila

Težava

- Kako dobiti vektorsko predstavitev besedila?
- Kako definirati vhodne spremenljivke?

Faza 3: Transformacija podatkov

Preprosta vektorska predstavitev besedila

- Vhodne spremenljivke ustrezajo besedam iz slovarja
- Vsaka spremenljivka je Boolova
- Vrednost beleži prisotnost besede v besedilu

Težava

- Ogromno število vhodnih spremenljivk
- Prekletstvo večdimenzionalnosti

Faza 1: Izbira podatkov

Vrednosti ciljne spremenljivke

- Vsi primeri v učni množici morajo biti že razvrščeni
- Običajno človek mora določiti vrednost ciljne spremenljivke
- Časovno zahtevno opravilo, veliko ročnega dela

Težavi

- Kolikšno je minimalno število označenih primerov?
- Kaj če je v enem razredu zelo malo primerov?

Prebeg kot problem dvojiškega razvrščanja

Primeri

Stranke storitvenega podjetja (npr. telekomunikacijskega operaterja)

Spremenljivke

- Ciljna spremenljivka: ali je stranka prebegnila?
- Opisne spremenljivke: podatki o stranki

Težave

Faza 1: Izbira podatkov

Če učna množica zajame vse stranke, bo delež prebeglih majhen (pod 1%).

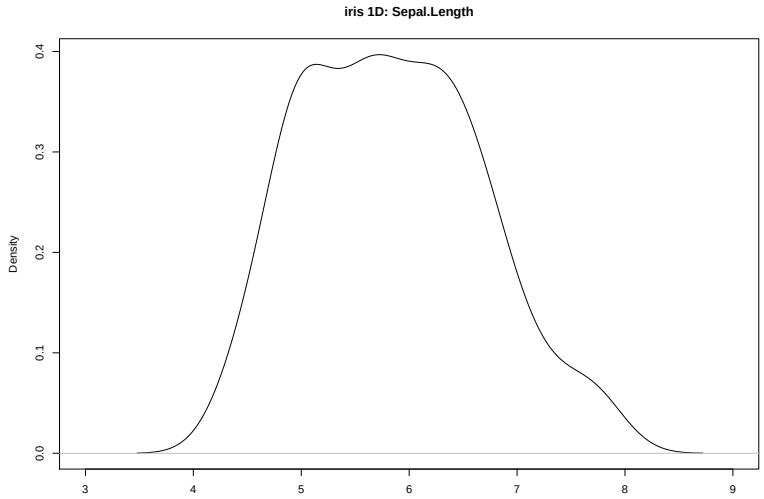
Faza 2: Pred-obdelava podatkov

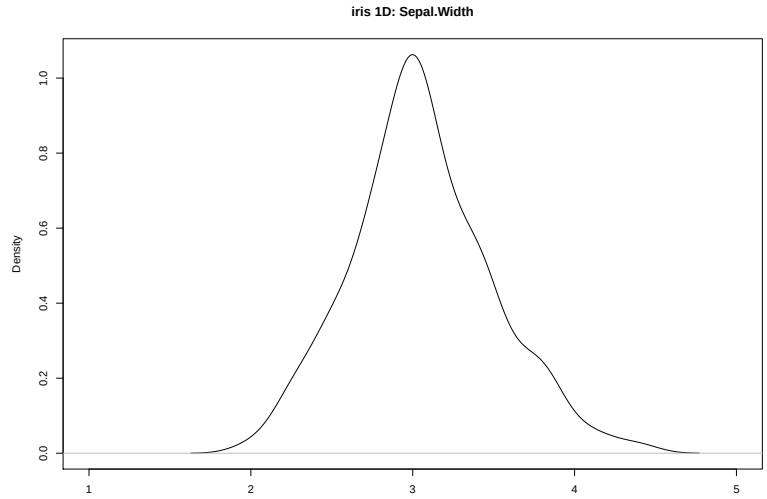
- Podatki vsebujejo tudi omrežje klicev stranke
- Kako to omrežje spraviti v vektorsko obliko opisnih spremenljivk?
- Manjkajoči podatki iz naslova anonimnih uporabnikov

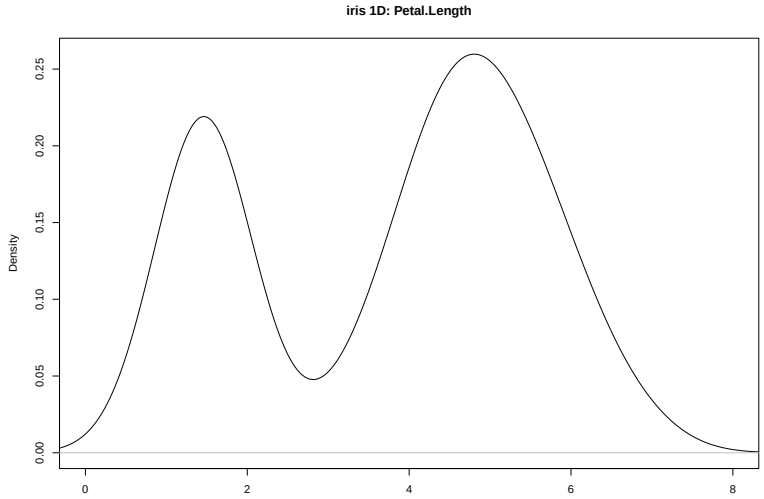
Fazi 4 in 5: Vrednotenje modelov

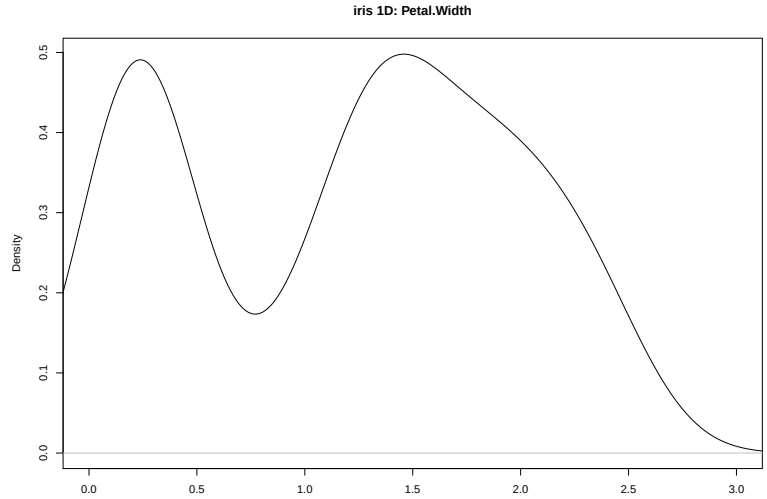
Kako vrednotiti točnost modela, če je malo pozitivnih primerov?

Vizualizacija posameznih spremenljivk

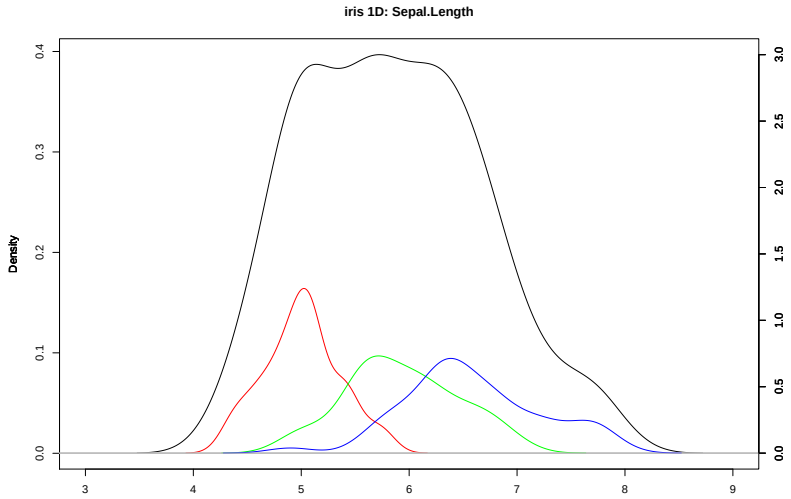


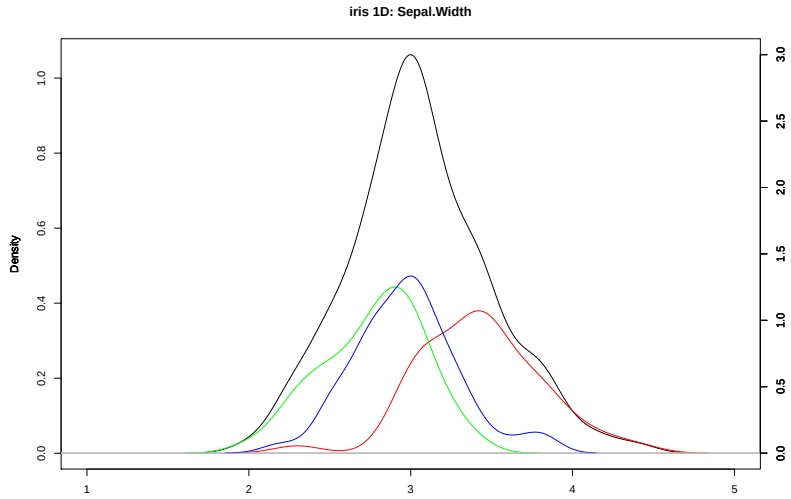


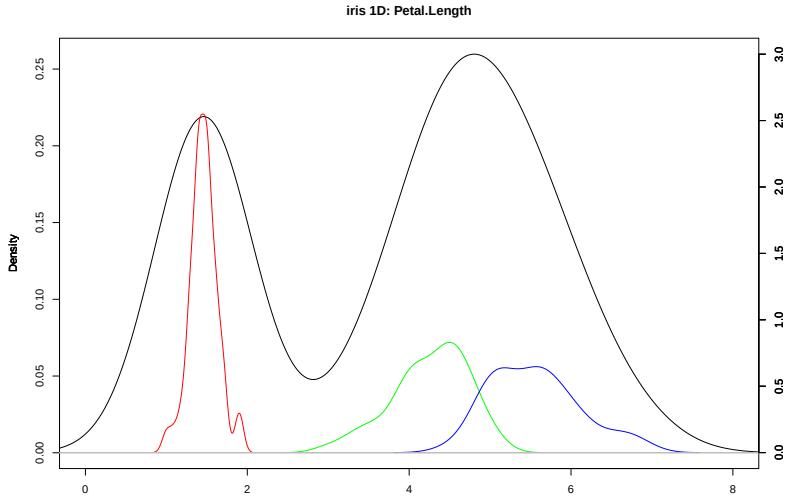


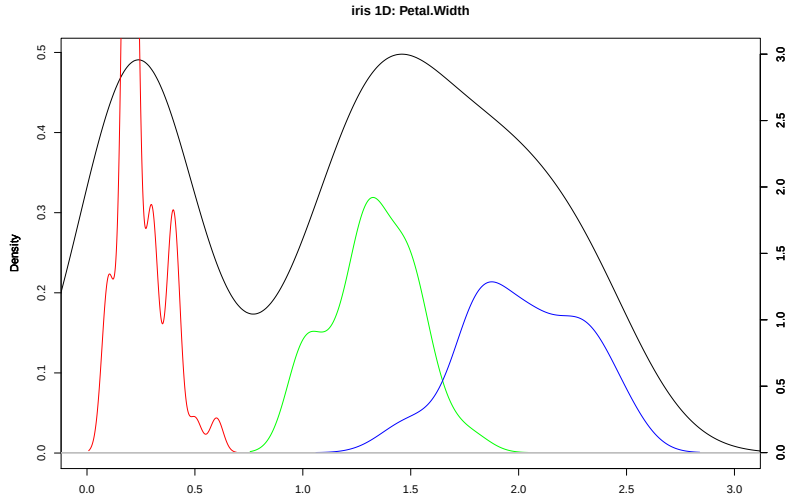


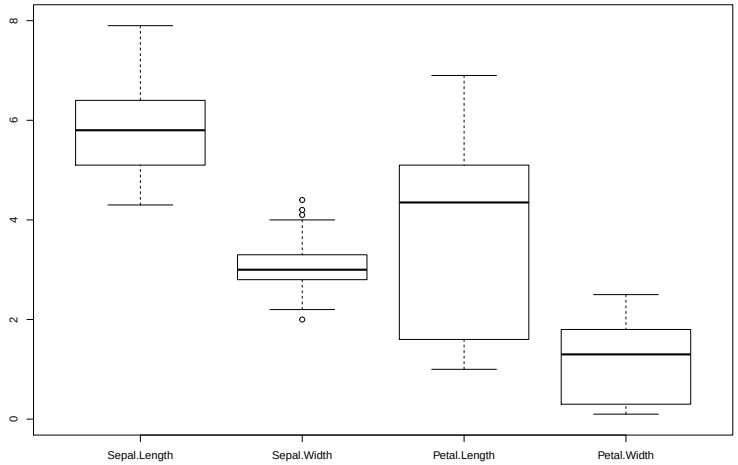
Vizualizacija posameznih spremenljivk in razreda

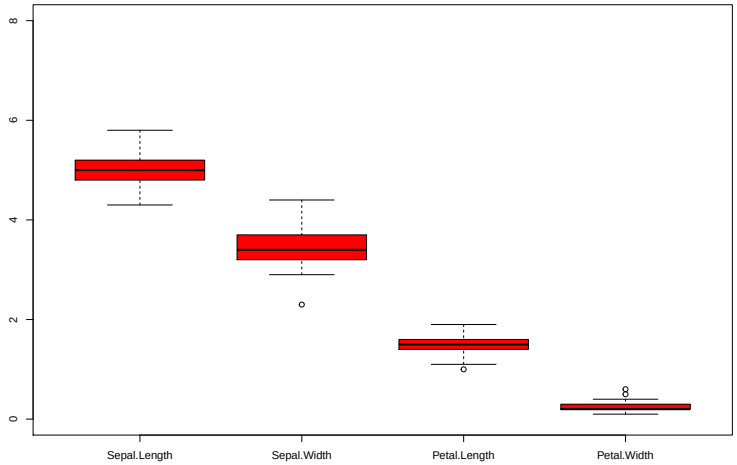


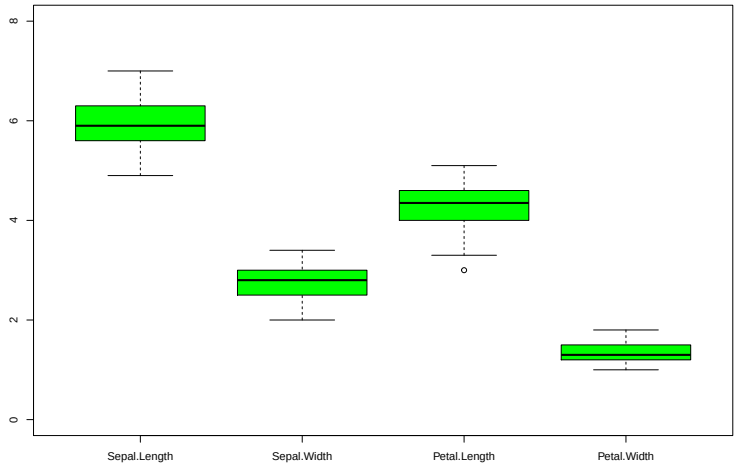


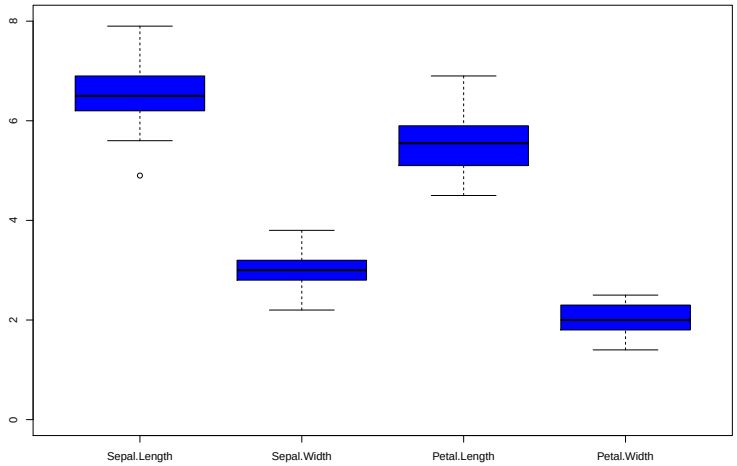






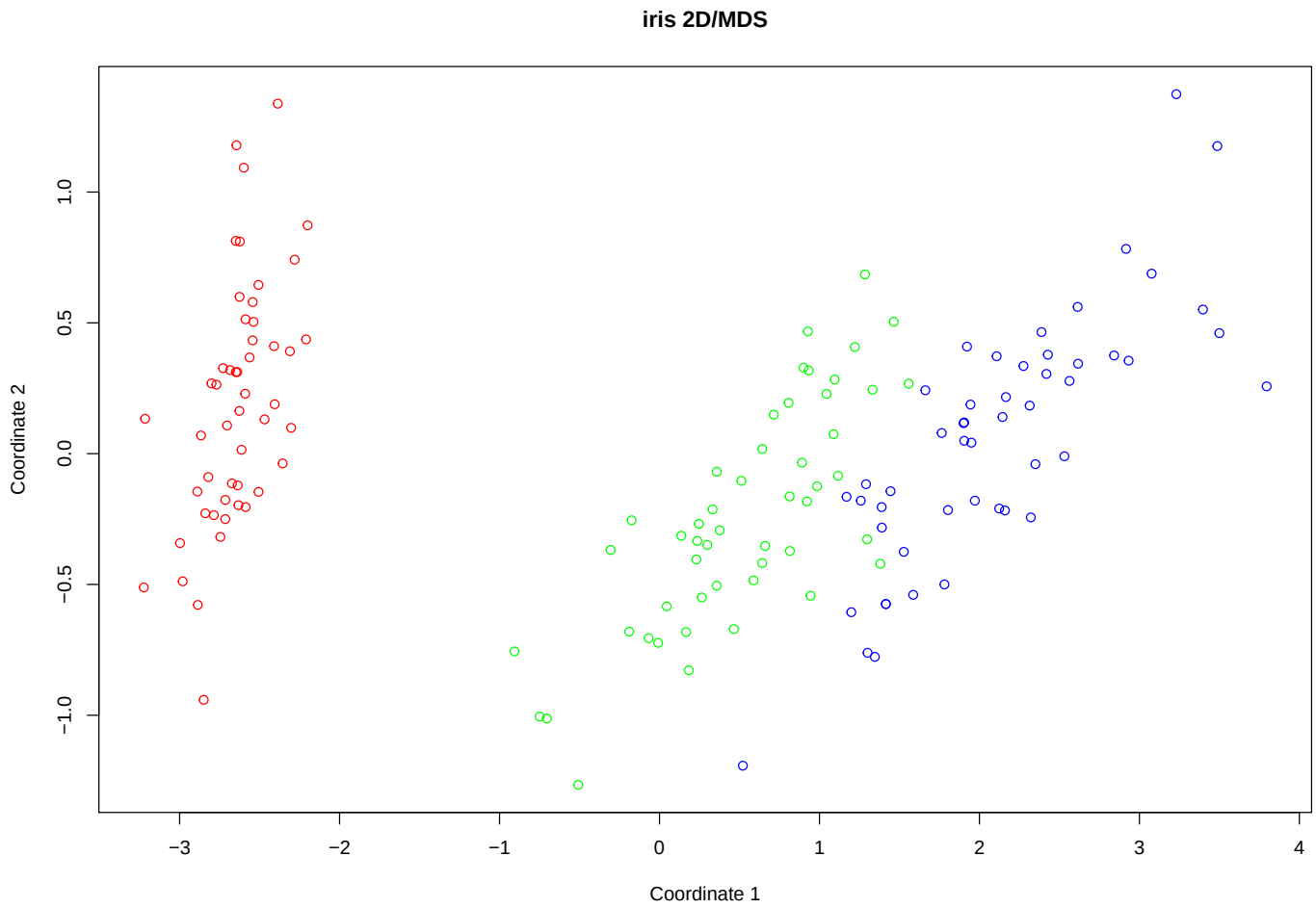






Vizualizacija posameznih spremenljivk: škatle

Več-dimenzionalna vizualizacija: MDS



Pregled predavanj do konca semestra

Rešitve omenjenih težav

1. Obravnava besedil: vreče besed in vpetja
2. Zmanjševanje razsežnosti: izbira in tvorba spremenljivk
3. Zapolnjevanje manjkajočih vrednosti
4. Neenakomerna porazdelitev vrednosti ciljne spremenljivke