

Sklop 1: Vizualno iskanje soodvisnosti v podatkih

Uvod

Primerjavo dveh kvantitativnih vrednosti je mogoče izvesti z vizualno primerjavo z uporabo raztresnega grafa* (*ang. scatter plot*). Na tem grafu predstavimo eno spremenljivko na osi x in drugo na osi y. Območji spremenljivk morata biti ustrezno skalirani. Na grafu lahko opazujemo trende, odvisnosti, vzorce in ekstremne odmike.

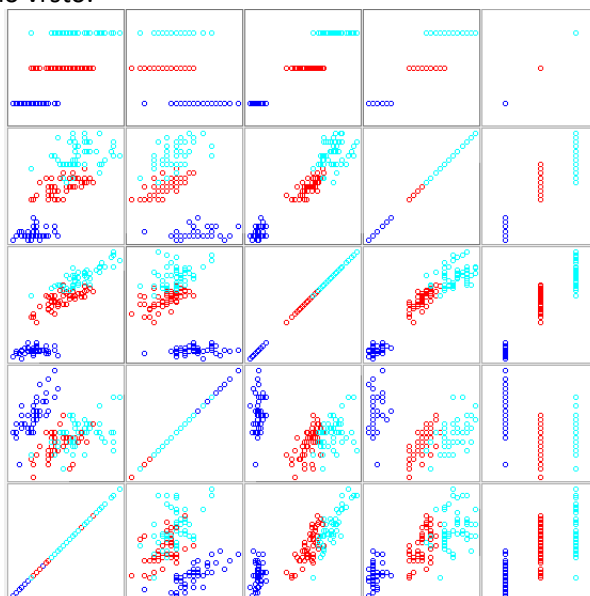
V primeru več značilk narišemo matriko raztresnih grafov in hitro opazimo soodvisnosti značilk.

Naloga 1

Odprite datoteko s podatki in v zavihku Starost Višina in narišite raztresni graf, na katerem primerjamo starost in višino otrok.

Naloga 2

V letu 1936 je bil napravljen nabor podatkov cvetlic perunik treh vrst: Iris Setosa, Iris Versicolor in Iris Virginica. V okviru nabora podatkov so izmerjene dolžine in širine cvetnih (petal) in venčnih (sepal) listov. Izmerjene meritve predstavljajo značilke, na podlagi katerih lahko posamezno cvetlico klasificiramo v posamezno vrsto.



-Narišite matriko raztresnih grafov podatkov na zavihku Iris in ugotovite kateri dve značilki sta med sabo najbolj soodvisni in kateri dve najmanj.

-Na raztresnem grafu najbolj soodvisnih značilk ugotovite, če lahko napravite klasifikator, ki bo posamezne vrste cvetlic učinkovito ločil med sabo.

Najbolj soodvisni značilki: _____

Par značilk, ki bi ga lahko uporabil za klasifikator: _____

*Raztresni graf se prevaja v slovenščino tudi kot razsevni graf

Vir

-Ian H. Witten, Eibe Frank, Mark A. Hall. Data Mining. Practical Machine Learning. Tools and Techniques. Third Edition

- [http://en.wikipedia.org/wiki/Features_\(pattern_recognition\)](http://en.wikipedia.org/wiki/Features_(pattern_recognition))

- http://en.wikipedia.org/wiki/Iris_flower_data_set

-Population Reference Bureau: <http://www.prb.org/>

Vaja pripravljena v okviru KCUI delavnice na UL-FE