

СТАТИСТИКА ЗА АУТОМАТСКУ АНАЛИЗУ ПОДАТАКА

Упутство за израду семинарских радова

Семинарски рад представља се састоји из презентације и одбране (усмено и на рачунару). Сваки од ова два дела носи максимално 5 поена. Семинарски рад раде по 3 изузетно 2 студента у групи. За презентацију сви студенти који су обрадили исту тему добијају исти број поена, док на одбрани остварени број поена може се разликовати у зависности од знања студената.

Задатак

Све базе података на располагању, заједно с пратећим званичним описом, дате су у .zip фајлу. У табели дат је списак база.

назив базе	кратак опис
tinting ('Car Window Tinting Experiment Data')	ефекти затамњивања стакала код аутомобила
nuclear ('Nuclear Power Station Construction Data')	подаци о изградњи нуклеарне електране
evals ('Professor evaluations and beauty')	вредновање професора
CollegeDistance ('College Distance Data')	истраживање о удаљености школе
HousePrices ('House Prices in the City of Windsor, Canada')	цене кућа у граду Виндзору у Канади
DoctorVisits ('Australian Health Service Utilization Data')	коришћење здравствених услуга у Аустралији
Computers ('Prices of Personal Computers')	цене персоналних рачунара
SaratogaHouses ('Houses in Saratoga County (2006)')	анализа података о кућама у округу Саратога
USSeatBelts ('Effects of Mandatory Seat Belt Laws in the US')	ефекти закона о обавезним сигурносним појасевима

Списак тема

- Ефекти затамњивања стакала код аутомобила (подаци: *tinting*)
- Подаци о изградњи нуклеарне електране (подаци: *nuclear*)
- Истраживање о удаљености школе (подаци: *CollegeDistance*)
- Цене кућа у граду Виндзор, Канада (подаци: *HousePrices*)
- Коришћење здравствених услуга у Аустралији (подаци: *DoctorVisits*)
- Цене персоналних рачунара (подаци: *Computers*)
- Анализа података о кућама у округу Саратога (подаци: *SaratogaHouses*)
- Ефекти закона о обавезним сигурносним појасевима (подаци: *USSeatBelts*)

У семинарском раду треба укратко описати податке (шта представљају), затим одрадити почетну анализу података (дескриптивну статистику). Све базе су реалне, нису специјално сређиване за ову прилику, тако да ћете у њима наћи и променљиве које нећете обрађивати, већ треба направити смислену селекцију. Важно је уочити које су променљиве нумеричке, а које фактори, и променити тип променљиве у бази уколико је то потребно. Уколико има недостајућих података у променљивим од интереса, те елементе узорка треба избацити.

Након тога формирати модел вишеструке линеарне регресије, образложити која променљива се узима као зависна, а које као предиктори. Формирати најпре почетни модел (са свим променљивим), а затим приказати алтернативне моделе добијене селекцијом променљивих и/или трансформацијом. Урадити дијагностику предложеног модела и навести закључке за сваки модел понаособ и одредити који је модел најприкладнији.

Презентација и одбрана

Сви аутори рада имају одбрану у истом термину. Презентација треба да траје максимално 15 минута с равномерном поделом презентовања међу ауторима. Након презентације следе питања која могу бити како у вези рада, тако и градива курса директно везаног за рад.