Bordianu Ada Maria Alexandra

Diaconu Beatrice

Grupa 234

Proiect probabilitati si statisctica

In programul Rstudio am importat setul de date “train”, folosindu-ne in mare parte de variabilele Age, Survived, Pclass, Fare si Sex.

La primul punct am calculat media pretulului biletului si a numarului de supravietuitori. Am reprezentat grafic prin functia boxplot asocierea dintre Age si Fare. Am aflat varianta lui Fare si am impartit in patru parti egale variabila Pclass cu “quantile”.

La punctul doi, am realizat cu “cor” corelatia dintre Survived si Parch, concluzionand ca este o relatie “slaba”. Asadar, am ales sa facem o regresie simpla intre Survived si Parch, astfel incat sa putem vedea cati oameni au supravietuit in raport cu compartimentul in care au stat. Am vrut sa facem aceasta regresie pentru a vedea daca oamenii care au supravietuit au depins de compartimentul in care se aflau si daca au influentat statistic numarul de supravietuitori. In cazul regresiei multiple, am ales sa continuam principiul si sa facem asocierea dintre survived si parch,fare si pclass. Astfel, am vrut sa vedem cat este infulentat numarul de supravietuitori de catre pretul compartimentului, compartimentul in sine si clasa in care se incadreaza compartimentul. Cand vine vorba de predictie, am ales sa asociem Survived cu Age, astfel incat sa ne orientam cati oameni au supravietuiut in functie de varsta.

Pentru a doua parte a cerintei, am ales sa introducem prin repartitia uniforma variabilele height si weight. Am ales functia runif, deoarece exista o infinitate de cazuri intre doua valori constante, respectiv intre 1 si 2 la inaltime si 30 si 120 la greutate. Asadar, le-am considerat potrivite pentru a realiza o regresie simpla.

La punctul al treilea, repartitia aleasa este chisq.test. Reprezinta un test prin care se face verficarea relatiei dintre doua variabile din train (Survived si Pclass). Am facut distributia pe o variabila predictor, agePred, ideala pentru a vedea daca e apropiata de o distributie normal. In cazul nostru, nu este apropiata deoarece din punctul situat pe grafic la coordonatele (2.30,10) nu mai respecta caracteristicile unei distributii normale.