**NOSQL BDA – projektni zadatak**

**Zadatak je potrebno riješiti u Jupyter notebooku i predati dva dana prije regularnog ispitnog roka kao link na servis tipa Dropbox na mail :** [**ispiti@goranklepac.com**](mailto:ispiti@goranklepac.com)**, u subjectu navesti NOSQL BDA Ime Prezime .**

1. Kontrola kvalitete podataka

* Prepoznavanje nedostajućih vrijednosti
* Rekodiranje nedostajućih vrijednosti npr .null. -> „missing“
* Prepoznavanje uniformnih ili nelogičnih distribucija
* Provjeriti / redizajnirati /odrediti ciljnu varijablu i njenu kvalitetu
* Objasniti / komentirati nalaze

2. Statističke analize uzorka podataka

* Za svaku kategoričku varijablu prikazati distribucija frekvencije u grafu
* Za svaku kontinuiranu varijablu izračunati srednju vrijednost, z vrijednosti, zaobljenost distribucije, nagnutost distribucije
* Izračunati korelacije svake varijable sa ciljnom varijablom
* Izračunati korelacije svake varijable izražene kroz Z vrijednost i kategorizirane u razrede <-3 , (-3 do 3),>3 s ciljnom varijablom
* Objasniti što ste spoznali vezano uz promatrani skup podataka temeljem analiza iz točke 2
* Objasniti / komentirati nalaze
* Sve mora biti vidljivo u Jupyter Notebooku, kod i rezultati, kao i komentari

3. IV, WoE izračuni

* Izračunati IV, WoE **za svaku varijablu** i prikazati u Jupyter Notebooku
* Izdvojiti varijable sa najjačim IV-om i prikazati grafički WoE graf za sve najjače varijable
* Grupirati razrede po kriteriju WoE za najjače varijable
* Objasniti što ste spoznali temeljem ove analize (profili, uzroci, uzročno posljedične veze)
* Predložite kako biste riješili problem vezano uz vaš skup podataka temeljem IV, WoE analize (što bi trebalo mijenjati, raditi drugačije, gdje je problem uzrok problema, koji su rizični profili … )
* Sve mora biti vidljivo u Jupyter Notebooku, kod i rezultati, kao i komentari
* Objasniti / komentirati nalaze