# **Uputa za prvi dio izrade samostalnog praktičnog rada**

U prvom dijelu izrade rada trebate na temelju odabranog skupa podataka za prezentaciju pripremiti sljedeće:

1. **Razumijevanje domene** (istraživanje, pregled literature):
   * odaberite dva do pet relevantna znanstvena ili stručna članka na temu vašeg skupa podataka koristeći izvore poput [Google Scholar](https://scholar.google.com/) ili [ScienceDirect](https://www.sciencedirect.com/).
   * Zabilježite:
     + koje metode su primijenjene;
     + koje su evaluacijske metrike korištene i kakvi su rezultati dobiveni u tim istraživanjima;
     + kakva je kvaliteta razvijenih modela i koje smjernice autori preporučuju za buduća istraživanja.
2. **Razumijevanje podataka**
   * opišite skup podataka uključujući informacije o tipovima varijabli te osnovni opis varijabli;
   * napravite deskriptivnu statistiku za svaki tip varijable: za numeričke varijable uključite srednju vrijednost, medijan, standardnu devijaciju, donji i gornji kvartil te minimum i maksimum, a za kvalitativne varijable mod i frekvenciju pojavljivanja;
   * grafički prikažite prikladnim grafičkim dijagramima distribuciju varijabli;
   * prikažite i opišite matricom korelacije i/ili dijagramom raspršenja odnos između numeričkih varijabli;
   * provjerite i zabilježite prisutnost nedostajućih vrijednosti, dupliciranih zapisa i ekstremnih vrijednosti te navedite kako ćete s njima postupiti u pripremi podataka.
3. **Odabir metoda**
   * odaberite dvije od dolje navedenih metoda te ih kratko opišite uz objašnjenje koje varijable ćete koristiti za primjenu odgovarajućih testova;
     + Metode:
       - hi-kvadrat test;
       - t-test i odgovarajući neparametarski test ukoliko nisu ispunjeni uvjeti za primjenu t-testa;
       - ANOVA i odgovarajući neparametarski test ukoliko nisu ispunjeni uvjeti za njezinu primjenu te primjena post-hoc testa.
   * na temelju razumijevanja domene odaberite uz obrazloženje zavisnu varijablu na način da ćete izabrati jednu od sljedećih:
     + numeričku varijablu za primjenu linearne regresije;
     + kvalitativnu varijablu za primjenu logističke regresije.
       - *Napomena: Ako nema varijable za primjenu logističke regresije, tada zavisnu numeričku varijablu transformirajte na smislen način tako da se može koristiti za primjenu logističke regresije*.

Analiza skupa podataka radi se u RStudiju.

## **Skupovi podataka**

1. Kvaliteta vina (Wine Quality)

<https://archive.ics.uci.edu/dataset/186/wine+quality>

1. Procjena automobila (Car Evaluation)

<https://archive.ics.uci.edu/dataset/19/car+evaluation>

1. Podaci o kreditu u Njemačkoj (German Credit Data)

<https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data>

1. Potrošnja goriva automobila (Auto MPG)

<https://archive.ics.uci.edu/dataset/9/auto+mpg>

1. Iznajmljivanje bicikala (Bike Sharing)

<https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset>

* + odabrati skup podataka *day.csv*

1. Procjena razine pretilosti na temelju prehrambenih navika i tjelesnog stanja (Estimation of Obesity Levels Based On Eating Habits and Physical Condition)

<https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition>

1. Abalone (školjka Morsko uho)

<https://archive.ics.uci.edu/dataset/1/abalone>

1. Čvrstoća betona (Concrete Compressive Strength)

<https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength>

1. Odljev korisnika od pružatelja telekomunikacijskih usluga (Telecom Churn Dataset)

<https://www.kaggle.com/datasets/arashnic/telecom-churn-dataset/data>

* + napraviti podskup od 1500 slučajno odabranih opservacija iz danog skupa podataka

1. Zdravstveni pokazatelji dijabetesa (CDC Diabetes Health Indicators)

<https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators>

* + iz skupa *diabetes \_ binary \_ 5050split \_ health \_ indicators \_ BRFSS2015.csv* napraviti podskup od 1500 opservacija izabranih na slučajan način tako da 50% ispitanika nema dijabetes (0) i 50% ispitanika ima predijabetes ili dijabetes (1)

1. Predikcija odustajanja učenika od školovanja i akademski uspjeh (Predict Students' Dropout and Academic Success)

<https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success>

* + napraviti podskup od 1500 slučajno odabranih opservacija iz danog skupa podataka tako da je jednak omjer klasa varijable “*Target*” kao i u originalnom skupu podataka

1. Neispunjenje obveza klijenata kreditnih kartica (Default of Credit Card Clients)

<https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients>

* + napraviti podskup od 1500 opservacija izabranih na slučajan način

1. Šumski požari (Forest Fires)

<https://archive.ics.uci.edu/dataset/162/forest+fires>

1. Klinički zapisi o srčanom zastoju (Heart Failure Clinical Records)

<https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records>