**## OPIS**

Popis skupova podataka za predmet PUI

**## Sadržaj**

**## Skupovi podataka**

* bikesharing\_daily - bikesharing\_daily.csv: Skup u kojem se predviđa dnevni broj iznajmljenih bicikala (regresijski problem)
  + OPIS: Podaci predstavljaju dnevne podatke o iznajmljivanju bicikala u određenom vremenskom periodu. Cilj analize je izrada modela predikcije broja bicikala koji će biti iznajmljeni na temelju različitih značajki poput temperature, vlažnosti, brzine vjetra, dana u tjednu i vremenskih uvjeta.
  + Problematika za studente informatike: Primjena strojnog učenja (regresijski modeli, neuronske mreže) kako bi se predvidio broj bicikala koji će biti iznajmljeni na temelju vremenskih uvjeta i radnih dana. Identifikacija relevantnih značajki.
  + Problematika za studente ekonomije: Kako vremenski uvjeti i sezona utječu na potražnju za biciklima. Kako se mijenja potražnja za biciklima tijekom radnih dana u usporedbi s vikendima i praznicima. Kako prilagoditi marketinšku strategiju da se poboljša iznajmljivanje tijekom dana slabe potražnje. Može li se optimizirati prihod na temelju dinamičkog određivanja cijena (viša cijena kad je visoka potražnja).
* credit\_data - credit\_data.csv: Skup podataka u kojem se klasificira kreditni rizik korisnika bankarskih usluga
  + OPIS: Studenti informatike i ekonomije surađuju na analizi podataka o kreditnom riziku kako bi razvili model koji može predvidjeti vjerojatnost da će klijent imati "dobar" ili "loš" kreditni rejting. Raspolažete tabličnim skupom podataka koji sadrži informacije o klijentima banke, uključujući: spol, dob, zanimanje, stambeni status, štedni i tekući račun, kredit, itd.
  + Problematika za studente informatike: Primjena strojnog učenja za predviđanje kreditnog rizika klijenta (npr. logistička regresija, odluke stabla, random forest, SVM, neuronske mreže).
  + Problematika za studente ekonomije: Identifikacija ključnih faktora koji utječu na kreditni rizik. Razmatranje prednosti i ograničenja predikcijskog modela u stvarnim uvjetima bankarskog poslovanja. Na temelju analize predložiti kako banke mogu poboljšati procjenu kreditnog rizika. Predložiti strategije za smanjenje loših kredita. Prilagoditi ponudu kredita različitim skupinama korisnika (segmentacija).
* housing - housing.csv: Skup podataka za predviđanje cijene stana
  + OPIS: Skup podataka koji sadrži informacije o nekretninama. Cilj je razviti model za predviđanje cijenu nekretnine na temelju različitih faktora poput površine, broja soba, dodatnih pogodnosti i lokacije.
  + Problematika za studente informatike: Razvoj modela strojnog učenja (linearna regresija, odluke stabla, random forest, XGBoost itd.). Pravilno odabiranje značajnih varijabli koje najviše utječu na cijenu nekretnine.
  + Problematika za studente ekonomije: Identifikacija ključnih ekonomskih faktora koji najviše utječu na formiranje cijena - inflacija, kamatne stope, ponuda i potražnje. Analiza kako predloženi prediktivni model može pomoći investitorima i kupcima u donošenju odluka. Na temelju analize prediktivnog modela predložiti strategije za bolje ulaganje u nekretnine. Studenti mogu na temelju predviđanja cijena nekretnina razviti plan investiranja.
* traffic - traffic.csv Skup podataka za predviđanje prometa
  + OPIS: Prometna gužva je značajan problem u urbanim sredinama, a na nju utječu različiti faktori poput vremenskih uvjeta, kvalitete infrastrukture i gospodarskih aktivnosti. Cilj je razvili prediktivni model koji može predvidjeti prometnu gužvu na temelju različitih faktora poput vremenskih uvjeta, kvalitete zraka i doba dana.
  + Problematika za studente informatike: Razvoj modela strojnog učenja (regresija, neuronske mreže, random forest itd.) za predikciju prometa. Pravilno odabiranje značajnih varijabli koje najviše utječu na promet.
  + Problematika za studente ekonomije: Istraživanje kako vremenske uvjeti, kvaliteta zraka i doba dana utječu na promet. Analiza kako predloženi prediktivni model može pomoći u urbanističkom planiranju i optimizaciji prometa. Razviti smjernice za gradske vlasti, prijevozne kompanije i ekonomske subjekte. Razviti strategiju prometnih ograničenja s obzirom na vremenske uvjete.
* online\_news\_pop - online\_news\_pop.csv: Skup podataka za predviđanja dijeljenja vijesti
  + OPIS: Popularnost online članaka igra ključnu ulogu u digitalnom marketingu i medijskim strategijama. Faktori poput sadržaja, strukture teksta, korištenih multimedijskih elemenata i vremena objave mogu značajno utjecati na broj dijeljenja članka na društvenim mrežama.
  + Problematika za studente informatike: Razvoj modela strojnog učenja (regresija, neuronske mreže, random forest itd.) za predikciju popularnosti članka.
  + Problematika za studente ekonomije: Istraživanje utjecaja različitih tipova sadržaja na popularnost članaka. Analiza utjecaja dana objave na dijeljenje sadržaja (radni dani vs. vikend). Razmatranje uloge sentimenta i strukture članka u digitalnom marketingu. Ispitivanje mogućnosti implementacije rezultata u realnim poslovnim situacijama. Studenti mogu predlagati strategije za povećanje popularnosti članaka ili razviti smjernice za kreiranje sadržaja koji će imati veću vidljivost i doseg. Studenti mogu produbiti razumijevanje kako varijable poput strukture teksta, multimedijskih elemenata i vremena objave utječu na popularnost članaka te s time utjecati na prinose novinskih agencija. Tu mogu implementirati i analizu dugoročnih trendova u digitalnom marketingu i promjena u korisničkom ponašanju.
* flight\_satisfaction - flight\_satisfaction.csv: Skup za predvižanje zadovoljstva putnika letom
  + OPIS: Zračni promet igra ključnu ulogu u globalnoj povezanosti, a zadovoljstvo putnika izravno utječe na reputaciju i profitabilnost aviokompanija. Cilj ovog interdisciplinarnog projekta je analizirati podatke o zadovoljstvu putnika te razviti model koji može predvidjeti zadovoljstvo putnika na temelju različitih faktora putovanja.
  + Problematika za studente informatike: Razvoj modela strojnog učenja (logistička regresija, Random Forest, neuronske mreže itd.) za predviđanje zadovoljstva putnika. Identifikacija ključnih varijabli koje najviše utječu na zadovoljstvo putnika.
  + Problematika za studente ekonomije: Istraživanje kako različiti aspekti usluge i kašnjenja utječu na percepciju putnika. Analiza potencijalnih troškova i koristi povezanih s poboljšanjem ključnih faktora zadovoljstva. Analiza uloge lojalnosti kupaca u ocjenama zadovoljstva. Studenti mogu razmatrati kako predloženi predikcijski model može pomoći aviokompanijama u poboljšanju usluga te pružanju poslovnih preporuka za poboljšanje korisničkog iskustva.
* heat\_load - heat\_load.csv: Skup za predviđanje potrebe grijanja -  Npr. za prodaju grijačih jedinica
  + OPIS: Podaci koji su u skupu odnose se na karakteristike zgrade ili objekta, specifično u kontekstu energetske efikasnosti. Predviđa se toplinski učinak grijanja, odnosno opterećenja za grijanje.
  + Problematika za studente informatike: : Primjena tehnika strojnog učenja kako bi se analizirali podaci i predvidjela potrošnja energije i efikasnost grijanja za zgrade.
  + Problematika za studente ekonomije: Analiza troškova i koristi energetske efikasnosti, npr. kako promjena relativne kompaktnosti ili povećanje staklenih površina može utjecati na ukupne troškove grijanja i hlađenja. Studenti mogu koristiti analizu povrata na investiciju (ROI) za procjenu koliko će ulaganja u energetsku učinkovitost (npr. smanjenje površine staklenih površina, poboljšanje relativne kompaktnosti objekta) donijeti uštede. Određivanje koje vrste građevinskih karakteristika mogu povećati tržišnu vrijednost objekta, na temelju energetske efikasnosti, mogu biti korisni u analizama investicija i ekonomskih modela. U mnogim zemljama postoje regulative koje zahtijevaju od investitora da smanje energetsku potrošnju, kao i poticaji za gradnju energetski učinkovitih zgrada (npr. zelene zgrade).
* cooling\_load - heat\_load.csv: Skup za predviđanje potrebe hlađenja - Npr. za prodaju rasladnih uređaja
  + OPIS: Cilj ovog zadatka je da studenti ekonomije i informatike surađuju na analizi i optimizaciji energetskih potreba zgrade na temelju podataka koji uključuju karakteristike objekta, poput relativne kompaktnosti, površine, površine zidova i krova, visine, orijentacije, staklenih površina, i toplinskog opterećenja za hlađenje.
  + Problematika za studente informatike: Razvoj modela za predviđanje opterećenja za hlađenje: Studenti informatike će koristiti analizu podataka kako bi razvili model koji predviđa opterećenje za hlađenje (Cooling Load) na temelju varijabli poput relativne kompaktnosti, orijentacije, površine staklenih površina, i drugih karakteristika zgrade.
  + Problematika za studente ekonomije: Procjena ekonomskih učinaka energetske učinkovitosti - analiza kako bi promjena staklenih površina ili relativne kompaktnosti zgrade mogla smanjiti troškove hlađenja. Studenti ekonomije će razviti model za izračunavanje povrata na investiciju (ROI) za poboljšanje energetske učinkovitosti, kao što je smanjenje površine staklenih površina ili bolje orijentacije zgrade, s ciljem smanjenja troškova hlađenja. Procjena uštede za investitore koji odluče uložiti u energetski učinkovitiju gradnju.
* cement - cement.csv. Skup podataka za predviđanje tlačne čvrstoće betona (concrete copressive strength - MPa)
  + OPIS: Cilj ovog zadatka je da studenti ekonomije i informatike surađuju u analizi i optimizaciji sastava betona i njegovih svojstava, s naglaskom na predviđanje kompresivne čvrstoće betona na temelju različitih parametara.
  + Problematika za studente informatike: Studenti će raditi s podacima koji uključuju sastav betona (cement, šljaku iz visokih peći, leteći pepeo, vodu, superplastifikator, koarse agregate, fine agregate) te starost betona, i pomoću tih podataka razviti model koji može predvidjeti kompresivnu čvrstoću betona u megapaskalima (MPa).
  + Problematika za studente ekonomije: Studenti će istražiti cijene svake komponente betonske smjese. Na temelju tih cijena, razviti će troškovni model koji prikazuje koliko košta proizvodnja betonske smjese za različite omjere sastava i za različite vrste primjene betona (npr. Beton za temelje, zidove, pilone, mostove, nadvožnjake s jedne strane ili ukrasne stupove, ograde, kipove i skulpture kod druge krajnosti). Ako se betonske smjese optimiziraju za bolje performanse i dugotrajniju izdržljivost, studenti mogu procijeniti dugoročne uštede u gradnji, jer betoni s višom kompresivnom čvrstoćom mogu zahtijevati manje održavanja i biti ekonomski isplativiji na duže staze.
* ~~bankruptcy - bankruptcy.csv: Skup podataka za predviđanje stečaja poduzeća~~
  + Izbacujemo van jer je prekompleksni dataset
* steel\_industry\_energy\_consuption - steel\_industry\_energy\_consuption.csv: Podaci prikupljeni iz pametne čeličane za predikciju potrošnje energije u industriji čelika
  + OPIS: Predviđanje opterećenja (load prediction) je važna tema u području informatičkih znanosti i ekonomije, posebno u kontekstu energetske industrije. Kada govorimo o predviđanju opterećenja (load forecasting), cilj je razviti modele koji mogu predvidjeti potrošnju električne energije temeljem prošlih podataka. Podaci, kao što su potrošnja električne energije, reaktivna snaga, faktor snage i drugi parametri, igraju ključnu ulogu u optimizaciji energetske proizvodnje i distribucije.
  + Problematika za studente informatike: Za studente informatike, važno je razumjeti kako se koriste različite tehnike za predviđanje opterećenja. Razviti model za predikciju opterećenja.
  + Problematika za studente ekonomije: Prediktivni modeli mogu pomoći u predviđanju opterećenja i identificiranju perioda kada je potrošnja energije najveća (npr. Maximum\_Load) i kada je najmanja (npr. Light\_Load). To omogućuje tvrtkama da bolje planiraju potrošnju energije i pregovaraju o cijenama energije s dobavljačima, smanjujući troškove na temelju predviđenih potreba. Ako se predviđa da će određeno vrijeme biti period visokog opterećenja, prediktivni model može omogućiti optimizaciju korištenja resursa (npr. korištenje jeftinijih izvora energije u određenim vremenskim intervalima ili prebacivanje dijelova proizvodnje na manje opterećenje). Precizno predviđanje tipova opterećenja omogućava čeličani da unaprijedi upravljanje energijom i smanji gubitke. Na primjer, ako se zna da će doći do maksimalnog opterećenja, mogu se poduzeti mjere za smanjenje potrošnje ili optimiziranje proizvodnih procesa kako bi se izbjegla prekomjerna potrošnja energije. orištenjem prediktivnih modela, tvrtka može bolje planirati preventivno održavanje na temelju opterećenja i korištenja opreme. Podaci o CO2 emisijama povezani s razinom opterećenja (kako to implicira jedan od parametara), prediktivni model može pomoći u smanjenju emisija CO2 identificiranjem i optimiziranjem perioda s najvećim emisijama. Ove informacije mogu biti korisne za poboljšanje ekološkog pristupa i smanjenje operativnih troškova povezanih s emisijama. Prediktivni model može pomoći u boljem usklađivanju proizvodnih procesa s različitim razinama opterećenja. Na temelju tih informacija, tvrtka može optimizirati raspored proizvodnje i zaposlenika, smanjujući vrijeme zastoja i neproduktivne periode.
* water\_potability - water\_potability.csv: Skup za predviđanje je li voda za piće ili ne
  + OPIS: Kvaliteta vode ima veliki utjecaj na različite sektore, uključujući poljoprivredu, industriju, te pročišćavanje vode za ljudsku potrošnju.
  + Problematika za studente informatike: Za studente informatike, glavni izazov je razviti modele za predviđanje kvalitete vode na temelju kemijskih karakteristika vode.
  + Problematika za studente ekonomije: Na temelju podataka, moguće je razviti modele koji će predvidjeti promjene u kvaliteti vode u budućnosti. Ovi modeli mogu pomoći u kreiranju ekoloških politika i regulacija, koje osiguravaju da voda ostane sigurna za ljudsku potrošnju. Kvaliteta vode također utječe na poljoprivredu i industrijske procese. Na primjer, voda s visokom razinom tvrdosti može negativno utjecati na proces proizvodnje, dok prisutnost štetnih kemikalija kao što su trihalometani može biti neprihvatljiva za korištenje u industrijskim procesima. : Kvaliteta vode direktno utječe na troškove njenog pročišćavanja. Ako voda ima visoku razinu kloramina, tvrdoće ili zamućenosti, trošak filtracije i tretmana je veći. Stoga, ekonomski model može analizirati kako različite kemijske karakteristike vode povećavaju operativne troškove.
* ~~vrtic - vrtic.csv: Skup podataka za rangiranje prijave za određeni vrtić~~
  + Ovo nema previše smisla, izbacijemo.
* cc\_approval\_cleaned - cc\_approval\_cleaned.csv: Skup podataka za predviđanja odobrenja kreditnih kartica
  + OPIS: Ovi podaci se odnose na osobne financije korisnika, gdje se predviđa hoće li korisniku biti odobrena kreditna kartica na temelju različitih demografskih, financijskih i profesionalnih varijabli.
  + Problematika za studente informatike: Za studente informatike, glavni zadatak bi bio razviti model strojnog učenja za predviđanje hoće li korisnik dobiti kreditnu karticu, koristeći navedene varijable.
  + Problematika za studente ekonomije: Analizirati kako visina prihoda i kreditni rejting korisnika utječu na odluku o odobrenju kreditne kartice. Na primjer, veći prihod i bolji kreditni rejting mogu povećati šanse za odobrenje. Analizirati kako etnička pripadnost ili drugi socijalni faktori mogu utjecati na odluku o odobravanju kreditne kartice. Ovo pitanje može uključivati pristranost u sustavu kreditiranja, gdje korisnici iz određenih etničkih ili socijalnih skupina mogu biti diskriminirani prilikom odobrenja. Proučiti kako varijable poput prihoda i zaposlenosti mogu dovesti do ekonomske isključenosti određenih društvenih skupina. Na primjer, korisnici s nižim prihodom ili oni koji nisu zaposleni mogli bi biti manje skloni dobivanju kreditnih kartica, čak i ako ispunjavaju druge uvjete. Studenti mogu izraziti analizu rizika za banke i optimizirati poslovanje banaka – npr. sustav za predviđanje omogućuje automatsko donošenje odluka na temelju unaprijed definiranih pravila – koliko banke mogu uštedjeti s tim sustavom ili kako banke mogu optimizirati ponude za različito segmentirane korisnike.
* hitters - hitters.csv: Skup podataka za predviđanja plaći igrača bejzbola
  + OPIS: Podaci predstavljaju statistiku bejzbolskih igrača iz različitih timova, a cilj je predvidjeti njihovu plaću.
  + Problematika za studente informatike: Cilj je predvidjeti visinu plaće za svakog igrača na temelju njihovih performansi i drugih karakteristika. Ovo je problem regresije jer se radi o predviđanju numeričke vrijednosti.
  + Problematika za studente ekonomije: Ekonomska teorija vs. praktični rezultati: Studenti moraju povezati statističke modele s ekonomskim principima, kao što su ponuda i potražnja za bejzbolskim igračima, utjecaj marketinških prava ili čak utjecaj posjeta utakmicama na plaću igrača. Može se analizirati kako performanse igrača (kao što su broj trčanja i udaraca) donose povrat investicija za timove. Studenti mogu usporediti dvije lige iz podataka te analizirati da li igrači u jednoj ligi imaju tendenciju da zarađuju više nego u drugoj, te koji faktori (kao što su ligaške razlike, tržišni uvjeti i sl.) mogu utjecati na plate u tim ligama. Na temelju statistika o igračima, studenti mogu analizirati da li timovi s višim budžetima ulažu u nekoliko ključnih igrača s velikim plaćama, ili distribuiraju svoja ulaganja u više manjih ugovora. Analiziranje varijable EducationLevel može otkriti koji je obrazovni faktor ključan za uspjeh i visoku platu u bejzbolu, te kako obrazovni sistem može oblikovati mogućnosti za visoke plaće u profesionalnim sportovima. Studenti mogu proučavati kako promjene u sezoni (u smislu broja odigranih utakmica, promjena u pravilu ili ekonomskoj situaciji) utječu na plaću. Iako bi intuitivno veći učinak igrača trebao značiti višu platu, ponekad to nije slučaj. Studenti mogu analizirati kako raspodjela plaća među igračima može biti nejednakija nego što se očekuje, te koje su ekonomske teorije koje to objašnjavaju (npr. teorija tržišta monopola ili teorija jednakosti).