ПУШКИЦЕ
ФОН андерграунд

Inteligentni sistemi – Januar 2018

Uputstva

- Projekat je potrebno kreirati na desktopu i dati mu naziv u format: 52-Prezime Ime-Indeks
- Dozvoljeno je korišćenje skripte R-Cheetsheets.pdf i fajla Evaluationmetrics.R koji su raspoloživi u folderu sa tekstom zadatka
- Nije dozvoljeno korišćenje Interneta, niti korišćenje pomoćnih materijala u elektronskom, papirnom, niti bilo kom drugom obliku

Zadatak: Da li krenuti autoputem 407?

U fajlu "travel-times.csv" nalaze se podaci koje je vozač XY neko vreme sakupljao preko svog GPS uređaja vozeći se svojim autom od kuće na posao ili s posla kući. Značenja pojedinih varijabli su:

- **Date**: datum kada je obavljena vožnja
- **StartTime**: vreme kada je vozač započeo vožnju
- **DayOfWeek**: dan u nedelji kada je vožnja obavljena
- **GoingTo**: smer kretanja (Work – na posao, Home – kući)
- **Distance**: pređeni put u kilometrima
- **MaxSpeed**: najveća brzina tokom vožnje
- **AvgSpeed**: srednja brzina tokom cele vožnje
- **AvgMovingSpeed**: srednja brzina zabeležena samo dok se auto kretao (bez čekanja tokom zastoja, na semaforima, itd.)
- **FuelEconomy**: gruba procena potrošnje goriva u [litara/100km] tokom vožnje
- **TotalTime**: trajanje cele vožnje u minutima
- **MovingTime**: vreme tokom kojeg se auto kretao (to jest, ne uzimajući u obzir zastoje, nesreće na putevima, kraća zaustavljanja i sl.)
- **Congestion407**: opterećenost (zagušenost) autoputa sa oznakom 407 u vremenskom periodu kada je vožnja obavljena (niže vrednosti označavaju manju opterećenost)
- **Comments**: komentari vozača

Svaka vožnja obavlja se makar delimično preko autoputa sa oznakom 407 (autoput 407). Vozač ima dve alternative – da vozi sve vreme autoputem 407, ili da na nekim deonicama vozi sporednim putevima. Zadatak je primenom KNN algoritma predvideti koju alternative je vozač odabrao.

Potrebno je:

1. Proveriti da li neka od varijabli ima nedostajuće vrednosti (NAs) ili vrednosti "" ili "-" i ukoliko je to slučaj zameniti takve vrednosti adekvatnijim

2. Kreirati izlaznu varijablu Take407All sa dve moguće vrednosti: "Yes", što označava da sve deonice u vožnji treba preći autoputem I "No", što označava da neke deonice treba preći sporednim putevima. Vrednost "Yes" se postavlja ako je vrednost varijable Congestion407 manja od vrednosti na 60. percentilu i pri tome vozač nije uneo nikakav komentar za tu vožnju, a vrednost "No" u ostalim slučajevima . Uzeti vrednost "Yes" za pozitivnu klasu.

3. Primenom kros validacije sa 10 iteracija (10-fold cross-validation) odrediti najbolju vrednost za parameter K.

4. Kreirati klasifikacioni model koristeći KNN algoritam, a na osnovu izabrane vrednosti za K.

5. Za kreirani model:

- Kreirati i interpretirati matricu konfuzije
- Objasniti šta predstavljaju četiri metrike koje se najčešće koriste za procenu klasifikatora
- Izračunati i protumačiti vrednosti evalucionih metrika.