Predlog projekta: Regresija za određivanje cijene medicinskih troškova pacijenta

Student: Anja Vujačić RA 209/2021

# Regresija za određivanje cijene medicinskih troškova pacijenta

## 1. Definicija problema:

Razmatra se problem procjene medicinskih troškova na osnovu podataka o pacijentima. Cilj je razvijanje regresionog modela koji može precizno predvidjeti troškove liječenja za svakog pacijenta.

## 2. Skup podataka:

- **Starost (*age*)**: Starost glavnog osiguranika, izražena u godinama.
- **Pol (*sex*)**: Pol osiguranika, moguće vrijednosti su "ženski" ili "muški".
- **Indeks tjelesne mase (*BMI*)**: Indeks tjelesne mase osiguranika, koji pruža uvid u tjelesnu težinu u odnosu na visinu. Izračunava se kao odnos težine (u kilogramima) i kvadrata visine (u metrima), sa idealnim vrijednostima između 18,5 i 24,9. Ova mjera omogućava procjenu uticaja tjelesne mase na medicinske troškove.
- **Djeca (*children*)**: Broj djece koja su pokrivena zdravstvenim osiguranjem ili broj osoba koje su zavisne od osiguranika.
- **Pušač (*smoker*)**: Informacija o tome da li osiguranik konzumira cigarete ili ne.
- **Region (*region*)**: Lokacija prebivališta osiguranika u Sjedinjenim Američkim Državama, moguće vrijednosti su "sjeveroistok", "jugoistok", "jugozapad" i "sjeverozapad".
- **Troškovi (*charges*)**: Individualni medicinski troškovi koje fakturiše zdravstveno osiguranje. Ova vrijednost predstavlja ključni izlazni podatak koji se predviđa regresionim modelom.
- **Medical ID**: broj kartona
- Broj instanci prije uklanjanja nedostajućih vrijednosti je 1338.

Razvija se regresioni model koji predviđa medicinske troškove (označene kao "*charges*"), kontinuiranu numeričku vrijednost, koristeći ostale karakteristike pacijenata kao ulazne podatke.

## 3. Način pretprocesiranja podataka:

U procesu pretprocesiranja podataka, prvo se nailazi na **problem nedostajućih vrijednosti**. Analizom skupa podataka primjećuje se to da je većina informacija u koloni "Medical ID" nedostajuća, te se ova kolona eliminiše iz daljeg istraživanja. Takođe, primjećuje se da nedostajuće vrijednosti u koloni "charges" predstavljaju izazov, jer je to ključna promjenljiva čiju vrijednost treba predvidjeti. Iz tog razloga eliminišu se redovi gdje u pomenutoj koloni nedostaju vrijednosti.

Što se tiče kolone "BMI" koja ima nedostajuće vrijednosti, primjenjuje se pristup popunjavanja tih nedostajućih vrijednosti. Konkretno, koristiće se prosječna vrijednost za odgovarajuće godine i pol

kao zamjena za ove nedostajuće vrijednosti, čime se obezbeđuje kontinuitet analize i izbjegava gubitak podataka.

Nakon rješavanja problema sa nedostajućim vrijednostima, slijedi **transformacija labela**. Za kolone "*sex*" i "*smoker*", koje su kategoričke, primenjuje se *label encoding* kako bi se numerički predstavile ove kategorije. S druge strane, za kolonu "*region*" koristi se *one hot encoding* kako bi se efikasno reprezentovali različiti regioni, bez uvođenja redosljeda ili hijerarhije.

## 4. Metodologija

**Prikupljanje podataka**: Preuzimanje sa Kaggle platforme obezbjeđuje detaljan skup podataka o troškovima zdravstvenih osiguranja.

**Link do podataka** je ovdje.

**Pretprocesiranje podataka**: Ovaj korak obuhvata eliminaciju nedostajućih vrijednosti, uključujući i uklanjanje cijelih kolona ili redova gdje podaci nedostaju, kao i popunjavanje nedostajućih vrijednosti za kolonu "BMI" prosječnom vrijednošću za odgovarajuće godine i pol.

**Skaliranje podataka**: Nakon pretprocesiranja, primenjuje se skaliranje podataka kako bi se osiguralo da su sve karakteristike podataka u odgovarajućem opsegu. Koristi se MinMaxScaler i StandardScaler, a kasnije se bira optimalni pristup.

**Podjela podataka na skupove**: Podaci se dijele na trening i test skupove, gdje će test skup sadržati 20% podataka, osiguravajući da model bude evaluiran na nezavisnom skupu podataka.

**Izbor modela**: Za rješavanje problema procjene medicinskih troškova koriste se različiti regresioni modeli: Linear Regression, Lasso i Ridge Regression (derivirani iz Linear Regression), SVM Regression, Decision Tree Regression, kao i neuronska mreža za regresioni problem. Ovaj pristup omogućava analizu i upoređivanje performansi različitih modela.

**Podešavanje hiperparametara izabranog modela**: Za svaki odabrani model, vrši se podešavanje hiperparametara kako bi se postigle optimalne performanse. To uključuje korišćenje metoda kao što su *grid search* ili *random search*.

**Evaluacija greške**: Za evaluaciju performansi modela koriste se metrike kao što su Mean Absolute Error (MAE), Mean Squared Error (MSE) i Root Mean Squared Error (RMSE). Ove metrike omogućavaju kvantitativni opis toga koliko dobro model procjenjuje medicinske troškove.

**Potencijalno mijenjanje parametara i izbor drugačijeg modela**: Na osnovu rezultata evaluacije greške, razmotriće se mogućnost promjene parametara ili izbor drugačijeg modela kako bi se poboljšale performanse sistema. Ovaj iterativni proces omogućava kontinuirano poboljšavanje modela i prilagođavanje specifičnostima podataka.

## 5. Način evaluacije

Nakon podjele skupa podataka na trening i test skupove u odnosu 80:20, planiram da koristim nekoliko metrika kako bih evaluirala performanse svog modela. Specifično, koristiću pet standardnih metrika za regresione probleme: Mean Absolute Error (MAE), Mean Squared Error (MSE) i Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE),Coefficient of Variation (CV). Ove metrike omogućavaju kvantitativno ocjenjivanje razlike između stvarnih i predviđenih vrijednosti medicinskih troškova.

## 6. Tehnologija

Tehnološka infrastruktura ovog projekta oslanja se na Python programski jezik. Za implementaciju algoritama mašinskog učenja, koristi se biblioteka Scikit-learn, koja pruža bogat set algoritama i alata za rad sa podacima. Za implementaciju neuronskih mreža, kao što je TensorFlow, projekat se oslanja takođe na Python ekosistem.

# Relavantna literatura


- Deep Learning, MIT Press essential knowledge series. Author, John D. Kelleher. Publisher, MIT Press, 2019. ISBN, 0262354896, 9780262354899. Length, 272 pages.
- Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition by Aurélien Géron. Released September 2019. Publisher(s): O'Reilly Media, Inc. ISBN: 9781492032649