**SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU**

**FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA OSIJEK**

Sveučilišni diplomski studij

**PREDVIĐANJE PLAĆA U IT SEKTORU**

**Raspoznavanje uzoraka i strojno učenje**

**Domagoj Vučetić**

**Osijek, 2023.**

**SADRŽAJ**

[**1.** **UVOD** 3](#_Toc138623423)

[**2.** **STROJNO UČENJE** 4](#_Toc138623424)

[**2.1** **Regresija** 5](#_Toc138623425)

[**3.** **MODELI** 6](#_Toc138623426)

[**3.1** **Linearna regresija** 6](#_Toc138623427)

[**3.2** **Drveće odlučivanja (engl. *Decision Tree*)** 7](#_Toc138623428)

[**3.3** **Regresija nasumične šume (engl. *Random forest)*** 8](#_Toc138623429)

[**4.** **REZULTATI** 9](#_Toc138623430)

[**4.1** **Linearna regresija** 10](#_Toc138623431)

[**4.2** **Drveće odlučivanja (engl. *Decision Tree*)** 11](#_Toc138623432)

[**4.3** **Regresija nasumične šume (engl. *Random forest)*** 13](#_Toc138623433)

[**5.** **ZAKLJUČAK** 15](#_Toc138623434)

# **UVOD**

U današnjem digitalnom dobu, IT sektor je postao ključni igrač u svijetu poslovanja. Sve veći broj ljudi privlači mogućnost zapošljavanja u ovoj industriji, jer se smatra da IT poslovi nude visoke plaće i napredne karijerne mogućnosti. Međutim, mnogi se pitaju koliko je točno moguće predvidjeti plaće u IT sektoru.

Cilj ovog seminara je istražiti predikciju plaća u IT sektoru kroz primjenu različitih modela strojnog učenja. Korišten je programski jezik Python u razvojnom okruženju Anaconda.

# **STROJNO UČENJE**

Strojno učenje je područje umjetne inteligencije koje se bavi razvojem algoritama i tehnika kojima se računalima omogućuje da uče i donose zaključke iz podataka, bez eksplicitnog programiranja. Ova grana računalnih znanosti ima ključnu ulogu u raznim industrijama, uključujući IT sektor, gdje se primjenjuje za razumijevanje, analizu i predviđanje kompleksnih fenomena, uključujući plaće.

Strojno učenje koristi matematičke i statističke modele kako bi računalima omogućilo učenje iz podataka. Podaci se koriste za treniranje modela, što znači da algoritmi analiziraju uzorke i zakonitosti u podacima kako bi stekli znanje o određenom problemu ili fenomenu. Na temelju tog znanja, modeli mogu donositi predikcije ili donositi odluke.

Postoje različite tehnike strojnog učenja, uključujući nadzirano učenje, nenadzirano učenje i pojačano učenje. U kontekstu predikcije plaća u IT sektoru, nadzirano učenje odnosno regresija može biti relevantna. Ova tehnika koristi skup podataka koji sadrži informacije o IT profesionalcima, uključujući godine iskustva, godine u poduzeću i ostale faktore koji mogu utjecati na visinu plaća. Na temelju tih podataka, algoritmi strojnog učenja mogu naučiti kako ti faktori utječu na plaće i mogu se koristiti za predviđanje plaća za nove primjere.

## **Regresija**

Regresija je tehnika u strojnom učenju koja se koristi za predviđanje numeričkih vrijednosti na temelju podataka. Cilj regresije je pronaći matematički model koji opisuje odnos između ulaznih varijabli i ciljne varijable kako bi se moglo predvidjeti ili procijeniti vrijednost ciljne varijable za nove primjere.

U regresiji, ulazni podaci se nazivaju nezavisne varijable ili značajke, dok se ciljna varijabla koju želimo predvidjeti naziva ovisna varijabla. Na temelju dostupnih podataka, regresijski modeli pronalaze matematičku funkciju koja najbolje opisuje vezu između ulaznih varijabli i ciljne varijable.

Postoji nekoliko različitih pristupa regresiji, kao što su linearna regresija, regresija temeljena na stablima odlučivanja, regresija podržana vektorima i mnogi drugi. Svaki od ovih pristupa ima svoje karakteristike i prednosti, ovisno o prirodi podataka i očekivanoj složenosti odnosa između varijabli.

Ključni korak u regresiji je treniranje modela na dostupnim podacima. To uključuje pronalaženje parametara modela koji minimiziraju grešku predikcije između stvarnih vrijednosti i predviđenih vrijednosti.

Jednom kada je model treniran, može se koristiti za predviđanje vrijednosti za nove ulazne primjere. To omogućuje regresijskim modelima da se primijene u raznim područjima, kao što su financije, ekonomija, medicina, industrija, i mnogi drugi, gdje je predviđanje numeričkih vrijednosti ključno.

# **MODELI**

Modeli za regresiju su alati strojnog učenja koji se koriste za predviđanje kontinuiranih numeričkih vrijednosti. Regresijski modeli se koriste za analizu veza između ulaznih varijabli i ciljne varijable te za stvaranje funkcionalne veze koja može predvidjeti vrijednost ciljne varijable na temelju vrijednosti ulaza. U ovom primjeru koriste se 3 modela, linearna regresija, regresija drveća odlučivanja i

## **Linearna regresija**

Linearna regresija je jedan od osnovnih algoritama strojnog učenja koji se koristi za predviđanje numeričkih vrijednosti na temelju linearnog modela. Ideja linearne regresije je pronaći najbolju linearnu vezu između nezavisnih varijabli (ulaza) i zavisne varijable (cilja).

Linearna regresija pretpostavlja da postoji linearna veza između ulaznih varijabli i ciljne varijable. Osnovna pretpostavka je da se ciljna varijabla može opisati kao linearna kombinacija ulaznih varijabli, uz neke pretpostavke o greškama.

Matematički, linearna regresija se može predstaviti kao:

y = w0 + w1x1 + w2x2 + ... + wn\*xn + ε

gdje je y ciljna varijabla koju želimo predvidjeti, x1, x2, ..., xn su ulazne varijable, w0, w1, w2, ..., wn su koeficijenti (težine) koji se prilagođavaju modelu, ε je greška koja predstavlja nesavršeno poklapanje između stvarnih i predviđenih vrijednosti.

Cilj linearnog regresijskog modela je pronaći optimalne vrijednosti koeficijenata w0, w1, w2, ..., wn kako bi se minimizirala greška između stvarnih i predviđenih vrijednosti.

Prednosti linearne regresije uključuju jednostavnost interpretacije rezultata, brzu implementaciju i dobre performanse kada je linearna veza između varijabli dovoljno jasna. Međutim, linearna regresija ima ograničenja i ne može uhvatiti složene nelinearne veze između varijabli.

## **Drveće odlučivanja (engl. *Decision Tree*)**

Regresija drveća odlučivanja je metoda strojnog učenja koja se koristi za predviđanje kontinuiranih numeričkih vrijednosti. Temelji se na konceptu drveća odlučivanja, ali se primjenjuje za regresijske probleme umjesto klasifikacije.

Drvo odlučivanja je hijerarhijska struktura koja se sastoji od čvorova i grana. Svaki čvor predstavlja testiranje određene značajke, a grane predstavljaju moguće rezultate tog testa. Na temelju odgovora na test, algoritam odlučuje koji čvor treba biti slijedeći u stablu, sve dok se ne dosegne krajnji čvor koji predstavlja predviđenu vrijednost.

Regresija drveća odlučivanja radi na sličan način, ali umjesto da svaki čvor predstavlja klasifikacijski rezultat, on predstavlja numeričku vrijednost. Algoritam gradi stablo tako da optimizira odabir testova koji maksimiziraju smanjenje greške predviđanja.

Prilikom gradnje stabla, algoritam koristi različite mjere kao što su srednja kvadratna greška ili srednja apsolutna greška kako bi odabrao najbolju značajku i vrijednost za testiranje na svakom čvoru. Cilj je minimizirati grešku predviđanja na temelju dostupnih podataka.

Regresija drveća odlučivanja ima nekoliko prednosti. Prvo, intuitivno je razumljiva jer se rezultati mogu prikazati u obliku stabla odlučivanja. Također, može se nositi s različitim tipovima podataka, uključujući kategoričke i numeričke varijable. Osim toga, drveće odlučivanja može rukovati i s velikim brojem značajki te je otporno na prekomjerno prilagođavanje (engl. *overfitting*) podacima.

Međutim, regresija drveća odlučivanja također može imati neke nedostatke. Stabla odlučivanja mogu biti sklonija pretjeranoj složenosti. Također, osjetljiva su na male varijacije u podacima, što može rezultirati nestabilnim stablima. Da bi se riješili ovi problemi, koriste se tehnike kao što su podrezivanje stabala i korištenje ansambl metoda poput Random Forest-a.

## **Regresija nasumične šume (engl. *Random forest)***

Random Forest je metoda strojnog učenja koja koristi ansambl tehniku za rješavanje problema klasifikacije i regresije. To je kombinacija više drveća odlučivanja koja zajedno donose odluke o klasifikaciji ili predviđanju.

Ideja iza Random Foresta je da se konstruiraju mnoge slučajne podskupine od originalnog skupa podataka. Zatim se za svaki podskup konstruira stablo odlučivanja. Ključna razlika u odnosu na obično stablo odlučivanja je da se prilikom konstrukcije svakog čvora odabire samo podskup slučajno odabranih značajki iz cjelokupnog skupa značajki.

Nakon što su sva stabla izgrađena, klasifikacija ili predviđanje se vrši putem glasanja (za klasifikaciju) ili srednje vrijednosti (za regresiju) svih stabala. Svako stablo ima jednaku težinu u donošenju odluke.

Random Forest ima nekoliko prednosti. Prvo, može se nositi s visoko dimenzionalnim skupovima podataka koji sadrže veliki broj značajki. Također, Random Forest je otporan na prekomjerno prilagođavanje (overfitting) zbog korištenja slučajnih podskupova značajki i bootstrap uzoraka. Osim toga, može se koristiti za procjenu važnosti značajki, što pomaže u razumijevanju utjecaja pojedinih značajki na rezultate.

Važno je napomenuti da Random Forest ima i neke ograničenja. Interpretacija rezultata može biti izazovna zbog kombinacije više stabala. Također, izgradnja i predviđanje s više stabala može biti vremenski zahtjevno, posebno za velike skupove podataka.

Random Forest je popularna metoda strojnog učenja zbog svoje sposobnosti da pruži dobre rezultate u mnogim problemima klasifikacije i regresije.

# **REZULTATI**

Modeli se evaluiraju kako bi se dobila objektivna procjena njihove performanse i kvalitete. Evaluacija modela omogućuje nam da kvantificiramo koliko dobro model radi na predviđanju novih podataka.

Preciznost regresijskih modela se obično mjeri pomoću metrika koje ocjenjuju koliko dobro model predviđa numeričke vrijednosti. Nekoliko uobičajenih metrika koje se koriste za procjenu preciznosti regresijskih modela su:

Srednja kvadratna pogreška (engl. *Mean Squared Error*, MSE) je popularna metrika koja mjeri prosječnu kvadratnu razliku između stvarnih i predviđenih vrijednosti. Manje vrijednosti MSE ukazuju na veću preciznost modela, a optimalna vrijednost je 0.

Korijen srednje kvadratne pogreške (RMSE) je samo korijen iz MSE i koristi se kako bi se dobila mjera pogreške koja je izražena u istim jedinicama kao i ciljna varijabla. RMSE je korisna jer daje intuitivno razumljive rezultate. Izračunava se kao korijen iz MSE

Koeficijent determinacije R2 je metrika koja ocjenjuje koliko dobro model objašnjava varijabilnost ciljne varijable. Vrijednosti R2 se kreću od 0 do 1, gdje veće vrijednosti ukazuju na veću preciznost. Vrijednost 1 označava da model savršeno objašnjava varijabilnost, dok vrijednost 0 znači da model nema prediktivnu snagu.

Korišteni skup podatak ima preko 60000 tisuća uzoraka iz kojeg su uzete značajke „totalyearlycompensation“, „yearsofexperience“ i „yearsatcompany“. Značajke „yearsofexperience“ i „yearsatcompany“ uzete su kao ulazne veličine na osnovu kojih se predviđala izlazna veličina odnosno „totalyearlycompensation“.

## **Linearna regresija**

Evaluacija modela linearne regresije za trening set:

Mean Squared Error (MSE): 12431667945.91025

Root Mean Squared Error (RMSE): 111497.38986142343

R2: 0.19646852415996974

Evaluacija modela linearne regresije za test set:

Mean Squared Error (MSE): 12800081723.978876

Root Mean Squared Error (RMSE): 113137.44616164392

R2: 0.19430085412362907

Na slici 1 možemo vidjeti ravninu koja najbolje odgovara trening podatcima, dok na slici 2 vidimo ravninu koja najbolje odgovara test podatcima.

Slika na kojoj se prikazuje dijagram, crta, dizajn

Opis je automatski generiran

Slika 1. Prikaz podataka za linearnu regresiju na trening podatcima

Slika na kojoj se prikazuje dijagram, crta, dizajn

Opis je automatski generiran

Slika 2. Prikaz podataka za linearnu regresiju na test podatcima

## **Drveće odlučivanja (engl. *Decision Tree*)**

Evaluacija decision tree modela za trening set:

Mean Squared Error (MSE): 11757842166.284641

Root Mean Squared Error (RMSE): 108433.58412542049

R2: 0.24002182895522417

Evaluacija decision tree modela za test set:

Mean Squared Error (MSE): 12925657642.784182

Root Mean Squared Error (RMSE): 113691.0622818882

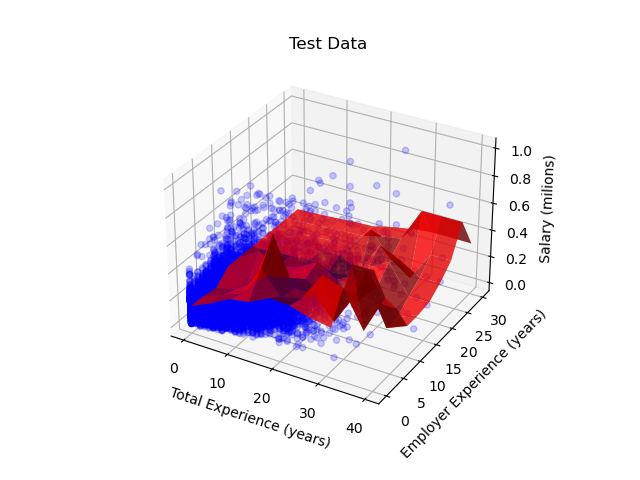
R2: 0.18639649751826948

Na slici 3 možemo vidjeti ravninu koja najbolje odgovara trening podatcima, dok na slici 4 vidimo ravninu koja najbolje odgovara test podatcima.

Slika na kojoj se prikazuje skeč, origami, dijagram

Opis je automatski generiran

Slika 3. Prikaz podataka za decision tree model na trening podatcima



Slika 4. Prikaz podataka za decision tree model na test podatcima

## **Regresija nasumične šume (engl. *Random forest)***

Evaluacija random forest modela za trening set:

Mean Squared Error (MSE): 11778926467.635115

Root Mean Squared Error (RMSE): 108530.76277090803

R2: 0.2386590271288792

Evaluacija random forest modela za test set:

Test Set:

Mean Squared Error (MSE): 12748478050.90051

Root Mean Squared Error (RMSE): 112909.15840134719

R2: 0.19754903926962186

Na slici 5 možemo vidjeti ravninu koja najbolje odgovara trening podatcima, dok na slici 6 vidimo ravninu koja najbolje odgovara test podatcima.

Slika na kojoj se prikazuje dijagram, karta

Opis je automatski generiran

Slika 5. Prikaz podataka za random forest model na trening podatcima

Slika na kojoj se prikazuje dijagram, skeč, crta

Opis je automatski generiran

Slika 6. Prikaz podataka za random forest model na trening podatcima

# **ZAKLJUČAK**

Strojno učenje moćan je alat koji se može koristiti za predviđanje plaća u IT sektoru. Kroz primjenu različitih modela strojnog učenja, kao što su linearna regresija, drveće odlučivanja i regresija nasumične šume, moguće je analizirati uzorke i zakonitosti u podacima te donositi predikcije o visini plaća za nove primjere.

Više od 60000 uzoraka plaća u IT sektoru su istraženi i analizirani kroz primjenu navedenih modela. Evaluacija sva 3 modela pokazuje da nijedan model nije adekvatan za ovaj problem jer imaju velike pogreške.

Jedan od razloga zašto modeli nisu precizni može biti što u skupu podataka ima velikih odstupanja, npr. osoba s 5 godine iskustva i 2 godine u istoj firmi može imati plaću 100000, dok druga osoba koja im 1 godinu iskustva i tek se zaposlio u novoj firmi može imati plaću 200000. Također jedan od problema može biti i ne korištenje adekvatnih značajki za predviđanje.