Maša Grubor

grubor.masa@gmail.com

# 

GitHub Repository:

https://github.com/masagrubor/Employee-Retention-ANN.git

predviđanje zadržavanja zaposlenih

Korišćenjem veštačke neuronske mreže (ANN)

Contents

[0](#_Toc189500583)

[1. UVOD 2](#_Toc189500584)

[2. PODACI I OBRADA 3](#_Toc189500585)

[2.1 Opis skupa podataka: 3](#_Toc189500586)

[2.2 Struktura skupa podataka 3](#_Toc189500587)

[2.2.1 Analiza distribucije klasa: 4](#_Toc189500588)

[2.2.2 Analiza korelacije karakteristika 5](#_Toc189500589)

[2.3 Čišćenje i transformacija podataka 6](#_Toc189500590)

[3. IMPLEMENTACIJA 7](#_Toc189500591)

[3.1 Korišćene tehnologije i biblioteke 7](#_Toc189500592)

[4. Arhitektura ANN modela 8](#_Toc189500593)

[5. Treniranje i validacija 9](#_Toc189500594)

[5.1 Parametri treninga 9](#_Toc189500595)

[5.2 Procena modela 9](#_Toc189500596)

[6. REZULTATI I EVALUACIJA 10](#_Toc189500597)

[6.1 Poređenje performansi 10](#_Toc189500598)

[6.2 Analiza grafika 11](#_Toc189500599)

[Confusion matrix: 11](#_Toc189500600)

[Training Loss Curve: 13](#_Toc189500601)

[6.3 Ključni uvidi 13](#_Toc189500602)

[7. Zaključak i budući rad 14](#_Toc189500603)

[7.1 Rezime 14](#_Toc189500604)

[7.2 Budući napredak 14](#_Toc189500605)

# UVOD

Zadržavanje zaposlenih je ključni aspekt upravljanja ljudskim resursima, jer visoke stope osipanja mogu dovesti do pada morala, oslabljenja timova i privremenog pada produktivnosti. Svrha ovog modela veštačke neuronske mreže je da predvidi nameru zaposlenog da napusti organizaciju na osnovu niza faktora kao što su zadovoljstvo poslom, plata, opterećenje i mogućnost razvoja karijere.

Identifikovanjem radnika sa visokim rizikom od odlaska, odeljenja za ljudske resurse mogu delovati i preduzeti prokativne strategije kao što su poboljšanje radnog okruženja i kreiranje opcija za razvoj karijere. Ovaj prisup omogućava timovima koji upravljaju radnom snagom da se pripreme za odlazak zaposlenog ili intervenišu kako bi ih ohrabrili da ostanu.

# PODACI I OBRADA

## Opis skupa podataka:

Skup podataka korišćen za treniranje modela je potiče iz **IBM HR Analytics Employee Attrition & Performance** dataset-a dostupnog na Kaggle-u. Sastoji se od istorijskih zapisa zaposlenih, koji pokrivaju više aspekata angažovanja zaposlenih, zadovoljstva poslom, plate i indikatore performansi.

* Ukupna evidencija: 1470 zaposlenih
* Ukupno karakteristika: 31 (nakon obrade)
* Ciljna promenljiva: Odlazak (DA/NE)

Ključni atributi:

* Demografske informacije: starost, pol, udaljenost od kuće
* Atributi u vezi sa poslom: uloga, odeljenje, godine u kompaniji, mesečni prihod, nivo posla
* Faktori radnog okruženja: status prekovremenog rada, ukupno radno vreme, ravnoteža između posla i privatnog života, poslovna putovanja
* Pokazatelji učinka i napretka: nivo zadovoljstva poslom, ocena učinka, završeni sati obuke, status unapređenja, godine od poslednjeg unapređenja

## Struktura skupa podataka

Skup podataka se sastoji od 1470 zapisa sa sledećom raspodelom zaposlenih koji su ostali u odnosu na one koji su otišli:

* Zadržano zaposlenih: 83,88% (1233 zaposlenih)
* Zaposleni koji su otišli: 16,12% (237 zaposlenih)

Ova neravnoteža je uzeta u obzir prilikom treniranja modela kako bi se izbegla pristrasnost prema većinskoj klasi.

### Analiza distribucije klasa:

A graph of a class distribution

Description automatically generated

Grafikon distribucije po klasama naglašava značajnu klasnu neravnozežu, sa visokim udelom radnika koji ostaju, a relativno malim procentom odustajanja. Taj disbalans je rešen tokom treniranja modela korišćenjem stratifikovanog uzorkovanja i tehnika ponderisanja, kako bi se izbegla pristrasnost modela ka predviđanju odlaska zaposlenih, a ne zadržavanja.

### Analiza korelacije karakteristika

A screenshot of a graph

Description automatically generated

Data Heapmap-a za korelaciju karakteristika otkriva sledeće informacije o međuzavisnosti karakteristika:

* Mesečna plata, nivo posla i godine rada imaju srednji nivo korelacije, pri čemu iskusni radnici mahom imaju visoku platu, odnosno seniorske pozicije.
* Zadovoljstvo poslom, status prekovremenog rada i ravnoteža između posla I života imaju jake veze sa odlascima, što ih čini veoma korisnim prilikom definisanja strategija zadržavanja.
* I broj godina od poslednjeg unapređenja I broj godina na datoj poziciji imaju slične odnose I može se videti da će radnici na sličnoj ulozi tokom dužeg vremena imati veliku verovatnoću odlaska.

Ova zapažanja su uzeta u obzir pri odabiru karakteristika za treniranje modela. Karakteristike sa visokim stepenom korelacije sa fluktuacijom radnika su prioritizovane, a karakteristike sa niskim efektom su eliminisane u pokušaju da se maksimiziraju performanse modela.

## Čišćenje i transformacija podataka

* Kategoričke varijable su bile one-hot encode-ovane (npr. uloga, odeljenje…) da bi se omogućilo da ih ANN obradi.
* Normalizacija: kontinuiranje numeričke varijable kao što su plata i godine u kompaniji su skalirane korišćenjem Min-Max skaliranja kako bi se sve karakteristike dovele u uporediv opseg, poboljšavajući stabilnost modela.
* Odstupanja u numeričkim karakteristikama su identifikovana i obrađena korišćenjem metode Interquartile range (IQR) kako bi se osiguralo da ekstremne vrednosti ne utiču negativno na treniranje modela.
* Podaci su podeljeni tako da se 80% skupa koristi za treniranje a 20% za testiranje, održavajući ravnotežu klase kroz stratifikovano uzorkovanje. Ovo je obezbedilo da obe klase budu proporcijalno zastupljene u skupovima, sprečavajući bilo kakvu pristrasnost u predviđanjima modela.

A red and blue squares with numbers

Description automatically generated

Ova vizualizacija prikazuje kako one-hot encoding konvertuje kategoričke karakteristike u numerički format unutar skupa podataka. Svaki red odgovara kodiranju pojedinačne karakteristike, dok svaka kolona predstavlja jednog zaposlenog. Boje omogućavaju jasan prikaz prisustva (1, prikazano crvenom) i odsustva (0, prikazano plavom) određenih kategoričkih atributa za svakog zaposlenog.

Na primer, zaposleni može pripadati samo jednoj radnoj ulozi (npr. JobRole\_Manager, JobRole\_Engineer, JobRole\_Analyst), jednom odeljenju (npr. Department\_HR, Department\_IT, Department\_Sales) i imati jedan status prekovremenog rada (npr. OverTime\_Yes, OverTime\_No). Heatmap-a obezbeđuje da ne postoje konfliktna kodiranja – svaka grupa ima samo jednu vrednost “1” po zaposlenom, čime se osigurava pravilna struktura podataka za model.

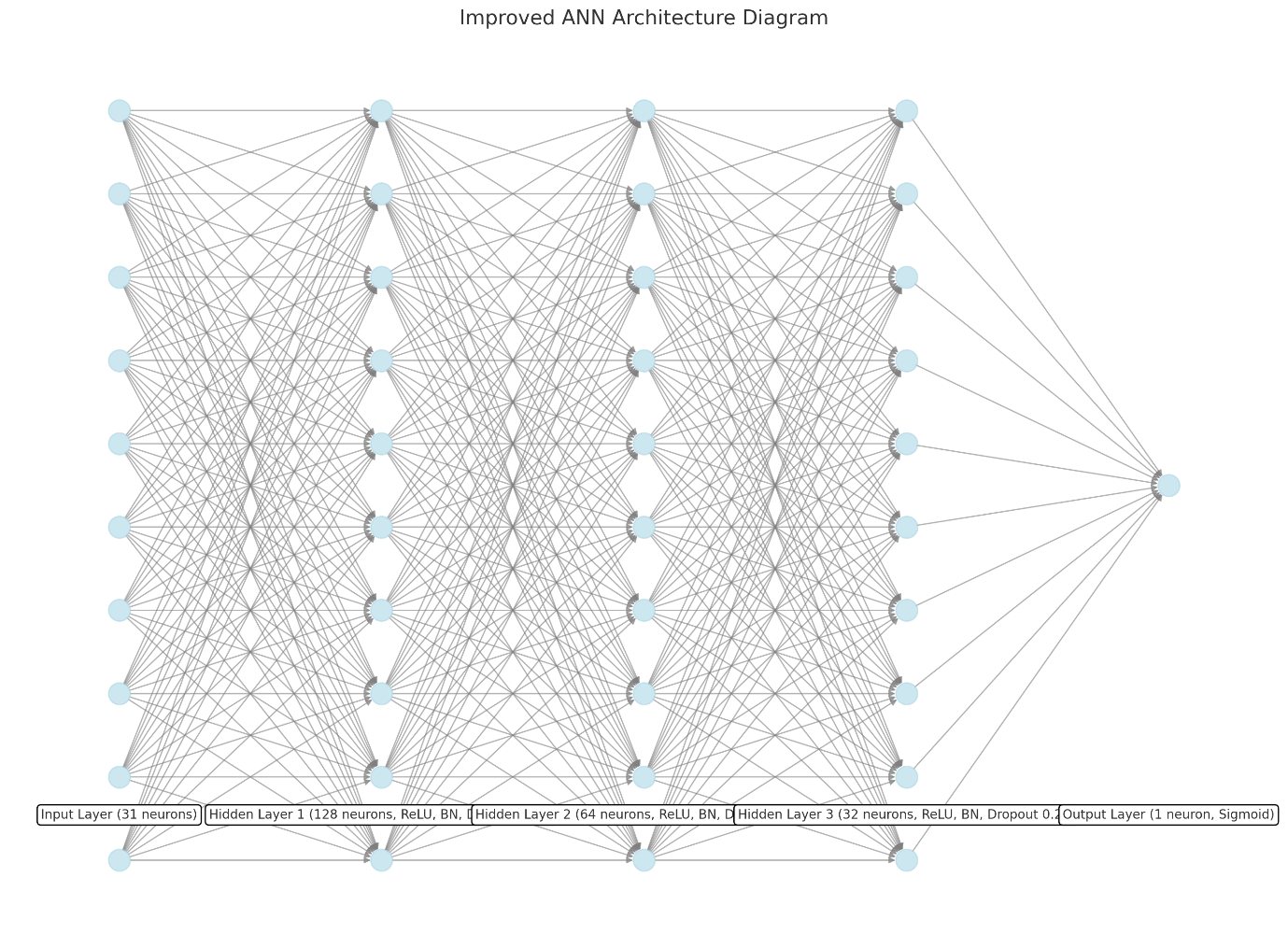
# IMPLEMENTACIJA

Model je implementiran u Python-u, koristeći različite biblioteke za treniranje, testiranje i predprocesiranje podataka.

## Korišćene tehnologije i biblioteke

* Google Colab – Cloud okruženje za izvršavanje Python notebook-a
* Python 3.9+
* Pandas, NumPy – Manipulacija i obrada podataka
* Scikit-Learn – Predprocesiranje podataka, konvencionalni algoritmi (Decision Tree, Logistic Regression)
* PyTorch – Kreiranje, treniranje i testiranje veštačke neuronske mreže (ANN)
* Matplotlib, Seaborn – Vizualizacija podataka, uključujući konfuzionu matricu i precision-recall krivu

# Arhitektura ANN modela



Odabrani ANN model prati arhitekturu neuronske mreže širenja unapred sa više skrivenih slojeva da bi se opazili složeni odnosi unutar podataka.

* Ulazni sloj: 31 neuron (po jedan za svaku karakteristiku)
* Skriveni slojevi:
  + Prvi skriveni sloj: 128 neurona, aktivacija ReLu, batch normalizacija
  + Drugi skriveni sloj: 64 neurona, aktivacija ReLu, batch normalizacija
  + Treći skriveni sloj: 32 neurona, aktivacija ReLu, batch normalizacija
* Dropout sloj: nanesen na sve skrivene slojeve kako bi se sprečilo overfit-ovanje modela
* Izlazni sloj: 1 neuron sa sigmoidnom funkcijom aktivacije za binarnu klasifikaciju (ostao = 0, otišao = 1)

# Treniranje i validacija

## Parametri treninga

* Funkcija gubitka: Focal Loss (ponderisana funkcija gubitka za upravljanje neravnotežom klase)
* Optimizator: Adam sa stopom učenja 0,0015
* Veličina serije: 32
* Broj iteracija: 50
* Planer brzine učenja: StepLR sa veličinom koraka 5 i ɣ = 0,97

## Procena modela

Pokazatelji učinka:

* Tačnost (accuracy)
* Preciznost, opoziv, F1-score (precision, recall, F1-score)
* AUC-ROC kriva

Strategija validacije:

* K-fold unakrsna validacija
* Analiza konfuzione matrice (confusion matrix) – za procenu lažno pozitivnih i lažno negativnih rezultata

# REZULTATI I EVALUACIJA

## Poređenje performansi

Da bismo potvrdili performanse ANN modela, uporedili smo ga sa klasifikatorom stabla odlučivanja i logističkom regresijom, što su tradicionalne tehnike mašinskog učenja.

|  |  |  |  |
| --- | --- | --- | --- |
| MODEL | ACCURACY | PRECISION | RECALL |
| ANN | |  | | --- | | **88.1%** | | |  | | --- | | **66.67%** | | |  | | --- | | **51.06%** | |
| Decision Tree | |  | | --- | | **83.33%** | | |  | | --- | | **44.44%** | | |  | | --- | | **17.02%** | |
| Logistic Regression | **87.41%** | **72.73%** | **34.04%** |

ANN u poređenju sa stablom odlučivanja

* Veća tačnost sa ANN: Sa 88,1% tračnosti, ANN model je nadmašio 83,33% tačnosti stabla odlučivanja, što dokazuje da ANN najbolje generalizuje neviđene podatke o fluktuaciji zaposlenih.
* Značajan pad preciznosti kod stabla odlučivanja: Model stabla odlučivanja ima preciznost od 44,44% što znači da je za mnoge radnike predviđeno da će napustiti posao a zapravo su ostali. ANN model ga značajno nadmašuje u pogledu preciznosti, sa 66,67%
* Dramatičan pad opoziva kod stabla odlučivanja: U prognoziranju fluktuacije zaposlenih, visok opoziv je od velikog značaja, jer neuspeh u preciznom prepoznavanju stvarnih slulaja napuštanja može imati visoku cenu. Stablo odlučivanja je uspelo da postigne samo 17,02% opoziva, što znači da nije uspelo da prepozna 83% stvarnih slučajeva napuštanja zaposlenih. Recall za ANN, s druge strane bio je 51,06% što ga čini mnogo pouzdanijim modelom za prognozu fluktuacije.

ANN u poređenju sa logističkom regresijom

* Veća tačnost sa ANN: Oba modela postižu dobre rezultate ali ANN ima blagu prednost, što ukazuje na bolju generalizaciju za nove slučajeve.
* Povećana preciznost kod logističke regresije: Logistička regresija postiže preciznost od 72,73% u poređenju sa 66,67% kod ANN-a. To znači da logistička regresija pravi manje netačnih pozitivnih predikcija – kada predvidi da će zaposleni napustiti posao, češće je u pravu nego ANN.
* Veći opoziv kod ANN-a: Ovo pokazuje da ANN identifikuje veći procenat stvarnih odlazaka zaposlenih.

Šta ove razlike znače?

* Logistička regresija je konzervativnija u predviđanju odlazaka zaposlenih, što je čini specifičnijom, ali manje efikasnom u prepoznavanju stvarnih slučajeva odlska. Iako je to manje rizično za HR odluke, može dovesti do propuštanja zaposlenih koji zaista planiraju da odu.
* ANN bolje balansira tačnost i recall, što ga čini jačim alatom za proaktivne strategije zadržavanja zaposlenih.
* U stvarnim HR aplikacijama, neuspeh u identifikaciji zaposlenih koji planiraju da odu može biti skupo plaćen, zbog čega je ANN preferirana alternativa.

## Analiza grafika

### Confusion matrix:

A graph with blue squares

Description automatically generated

Ova analiza pokazuje da model najpreciznije određuje zaposlene koji ostaju, ali da mu je potrebno poboljšanje u tačnom identifikovanju onih koji će napustiti posao. Lažno pozitivni slučajevi ukazuju na to da model nekad ne uspeva da prepozna zaposlene koji izaista odlaze. Lažno negativni slučajevi sugerišu na to da model precenjuje rizik od odlaska zaposlenih, što može dovesti do nepotrebnih HR intervencija.

Precision-Recall Curve:

A graph showing a line graph

Description automatically generated with medium confidence

Precision-Recall kriva pokazuje kako model balansira preciznost i recall pri različitim pragovima klasifikacije.

* Pri niskim vrednostima recall-a, preciznost je veoma visoka (~1.0), što znači da je model veoma siguran u svojim predikcijama, ali prepoznaje samo mali broj stvarnih slučajeva odlazaka zaposlenih.
* Kako recall raste, preciznost počinje da oscilira pre nego što postepeno opada, što ukazuje na kompromis između otkiravanje većeg broja zaposlenih koji će otići i održavanja pouzdanosti predikcija.
* Trend opadanja preciznosti sugeriše da, iako model generalno dobro radi, potrebno je dodatno podešavanje kako bi se bolje balansirali lažno pozitivni i lažno negativni rezultati.

Komplenzacija između recall-a i preciznosti naglašava da povećanje recall-a dolazi po cenu preciznosti, što ističe važnost pažljivog prilagođavanja praga klasifikacije i poboljšanja karakteristika modela. U HR aplikacijama ključno je balansirati visok recall i minimizirati nepotrebne lažno pozitivne slučajeve kako bi se izbegle preterane reakcije na zaposlene koji zapravo nisu u riziku od odlaska.

### Training Loss Curve:

A graph with a line

Description automatically generated

Pokazuje postebeni pad gubitka, što ukazuje na to da model zaista uči i da nije prenaučen. Glatkoća krive osigurava da su parametri, kao što su stopa učenja i veličina serije, pravilno podešeni.

## Ključni uvidi

* Zaposleni sa niskim zadovoljsvom poslom i čestim prekovremenim radom su česće odlazili.
* Plate i mogućnosti napredovanja u karijeri značajno su uticale na stope zadržavanja.
* ANN je uspešno uhvatio nelinearne zavisnosti između više funkcija.

# Zaključak i budući rad

## Rezime

Primena ANN-a u predviđanju zadržavanja zaposlenih pokazala je prednost u performansama u odnosu na tradicionalne algoritme u mašinskom učenju. Model je precizno identifikovao faktore koji utiču na odluke zaposlenih, a HR može te informacije da iskoristi u svojim strategijama za zadržavanje radne snage.

## Budući napredak

* Prilagođavanje pragova odlučivanja kako bi se minimizirale greške u pogrešnoj klasifikaciji.
* Primena modela u HR kontrolnoj tabli za predikciju u realnom vremenu i podršku pri donošenju odluka.