



UNIVERZITET U NOVOM SADU
FAKULTET TEHNIČKIH NAUKA
U NOVOM SADU




Јелена Милијевић

Предикција добитника филмске награде Оскар употребом машинског учења

МАСТЕР РАД
- Мастер академске студије -

Нови Сад, 2024.

	УНИВЕРЗИТЕТ У НОВОМ САДУ ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА 21000 НОВИ САД, Трг Доситеја Обрадовића 6	Датум:
	ЗАДАТАК ЗА ИЗРАДУ МАСТЕР РАДА	Лист:
		1/1

(Податке уноси предметни наставник - ментор)

Врста студија:	Мастер академске студије
Студијски програм:	Рачунарство и аутоматика
Руководилац студијског програма:	проф. др Мирна Капетина

Студент:	Јелена Милијевћ	Број	E2 91/2023
Област:	Електротехничко и рачунарско инжењерство		
Ментор:	Др Александар Ковачевић, ред. проф		
НА ОСНОВУ ПОДНЕТЕ ПРИЈАВЕ, ПРИЛОЖЕНЕ ДОКУМЕНТАЦИЈЕ И ОДРЕДБИ СТАТУТА ФАКУЛТЕТА ИЗДАЈЕ СЕ ЗАДАТАК ЗА ДИПЛОМСКИ РАД, СА СЛЕДЕЋИМ ЕЛЕМЕНТИМА: - проблем – тема рада; - начин решавања проблема и начин практичне провере резултата рада, ако је таква провера неопходна; - литература			

НАСЛОВ МАСТЕР РАДА:

Предикција добитника филмске награде Оскар употребом машинског учења

ТЕКСТ ЗАДАТКА:

1. Анализирати стање у области. 2. Применити технике истраживања и експлоративне анализе података на решавање проблема класификације добитника Оскара 3. Описати скупове података, који ће бити коришћен приликом анализе. 4. Дефинисати и имплементирати моделе за класификацију: Логистичка регресија, SVM, Random Forest, Bagging, XGBoost, Неуронска мрежа 5. Тестирати и евалуирати коришћене моделе

Руководилац студијског програма:	Ментор рада:

Примерак за: <input type="checkbox"/> - Студента; <input type="checkbox"/> - Ментора
--

КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА

Редни број, РБР :	
Идентификациони број, ИБР :	
Тип документације, ТД :	монографска публикација
Тип записа, ТЗ :	текстуални штампани документ
Врста рада, ВР :	мастер рад
Аутор, АУ :	Јелена Милијевић
Ментор, МН :	др Александар Ковачевић, ред. проф.
Наслов рада, НР :	Предикција добитника филмске награде Оскар употребом машинског учења
Језик публикације, ЈП :	српски
Језик извода, ЈИ :	српски / енглески
Земља публикавања, ЗП :	Србија
Уже географско подручје, УГП :	Војводина
Година, ГО :	2024
Издавач, ИЗ :	ауторски репринт
Место и адреса, МА :	Нови Сад, Факултет техничких наука, Трг Доситеја Обрадовића 6
Физички опис рада, ФО :	10 / 75/ 0 / 14/ 10/ 12/ 0
Научна област, НО :	Рачунарство и аутоматика
Научна дисциплина, НД :	Машинско учење
Предметна одредница / кључне речи, ПО :	Логистичка перспектива, SVM, Random Forest, Bagging, XGBoost, Neuronska mreža
УДК	
Чува се, ЧУ :	Библиотека Факултета техничких наука, Трг Доситеја Обрадовића 6, Нови Сад
Важна напомена, ВН :	
Извод, ИЗ :	У раду је вршена предикција добитника награде Оскар за најбољи филм и за глумачко остварење. Истраживање ове теме проистиче из њеног значаја за filmsku industriju. Korišćeni algoritmi su: Logistička Regresija, SVM, Random Forest, Bagging, XGBoost i Неуронска Мрежа. За евалуацију модела коришћена је 10-унакрсна валидација. У првој предикцији најбоље се показао Random Forest са тањчошћу 91,59%, а у другој XGBoost са 84,39%.
Датум прихватања теме, ДП :	
Датум одбране, ДО :	
Чланови комисије, КО :	

председник	Др Јелена Сливка, ванр. проф
члан	Др Лидија Крстановић, ванр. проф.
ментор	Др Александар Ковачевић, ред. проф
Потпис ментора	

KEY WORDS DOCUMENTATION

Accession number, ANO :	
Identification number, INO :	
Document type, DT :	monographic publication
Type of record, TR :	textual material
Contents code, CC :	master thesis
Author, AU :	Jelena Milijević
Mentor, MN :	Aleksandar Kovačević, full professor, PhD
Title, TI :	Using machine learning to predict Oscar winners
Language of text, LT :	Serbian
Language of abstract, LA :	Serbian / English
Country of publication, CP :	Serbia
Locality of publication, LP :	Vojvodina
Publication year, PY :	2024
Publisher, PB :	author's reprint
Publication place, PP :	Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6
Physical description, PD :	10 / 75/ 0 / 14/ 10/ 12/ 0
Scientific field, SF :	Electrical and computer engineering
Scientific discipline, SD :	Machine Learning
Subject / Keywords, S/KW :	Logistic Regression, SVM, Random Forest, Bagging, XGBoost, Neural Network
UDC	
Holding data, HD :	Library of the Faculty of Technical Sciences, Trg Dositeja Obradovića 6, Novi Sad
Note, N :	
Abstract, AB :	The study focused on predicting the winners of the Oscar award for Best Picture and for acting achievements. This research is motivated by its significance to the film industry. The algorithms used include: Logistic Regression, SVM, Random Forest, Bagging, XGBoost, and Neural Networks. Model evaluation was conducted using 10-fold cross-validation. In the first prediction, Random Forest demonstrated the best performance with an accuracy of 91.59%, while in the second prediction, XGBoost achieved an accuracy of 84.39%.
Accepted by sci. Board on, ASB :	
Defended on, DE :	
Defense board, DB :	
president	Jelena Slivka, associate professor, PhD

member	Lidija Krstanović, associate professor, PhD
mentor	Aleksandar Kovačević, full professor., PhD
Mentor's signature	

SADRŽAJ

KЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА	4
KEY WORDS DOCUMENTATION	7
UVOD	11
PREGLED STANJA U OBLASTI	14
TEORIJSKI POJMOVI I DEFINICIJE	18
3.1. Logistička Regresija.....	18
3.2. Support vector machine (SVM)	19
3.3. Random Forest i Bagging	20
3.4. XGBoost.....	22
3.5. Neuronska mreža.....	23
3.6. One-Hot Encoding i TF-IDF	26
METODOLOGIJA.....	29
4.1. Struktura sistema.....	29
4. 1.1. Eksplorativna analiza podataka (EDA).....	30
4.1.2. Modeli za klasifikaciju	31
EKSPERIMENTI.....	35
5.1. Skupovi podataka	35
5.1.1 Skup podataka nominovanih filmova	35
5.1.2. Skup podataka nominovanih glumaca	37
5.2. Eksperiment 1 – EDA	39
5.2.1. EDA za set nominovanih filmova	39
5.2.2. EDA za set nominovanih glumaca	43
5.3. Eksperiment 2 – Logistička Regresija.....	47
5.4. Eksperiment 3 – SVM	48
5.5. Eksperiment 4 – Random Forest	48
5.6. Eksperiment 5 – Random Forest sa Bagging-om.....	49
5.7. Eksperiment 6 – XGBoost	49
5.8. Eksperiment 7- Neuronska mreža	50
5.9. Evaluacija.....	50
REZULTATI.....	52
6.1. Rezultati eksperimenta 2	52
6.2. Rezultati eksperimenta 3	53
6.3. Rezultati eksperimenta 4	54
6.4. Rezultati eksperimenta 5	55
6.5. Rezultati eksperimenta 6	56
6.6. Rezultati eksperimenta 7	57
6.7. Uporedna analiza rezultata modela	57

DISKUSIJA	61
ZAKLJUČAK.....	69
LITERATURA	72
BIOGRAFIJA	75

UVOD

U dinamičnom svetu filmske industrije, Oskar nagrada predstavlja vrhunac priznanja za izuzetna dostignuća u kinematografiji. Predviđanje pobjednika za nagradu Oskar predstavlja izazov koji intrigira ne samo filmsku industriju, već i akademsku zajednicu. Oskar je jedan od najprestižnijih događaja u globalnoj filmskoj industriji, čiji dobitnici postaju sinonim za vrhunska dostignuća u filmskom stvaralaštvu.

Razlog za istraživanje ove teme leži u njenoj relevantnosti za filmsku industriju, ali i u potencijalnim doprinosima koji bi mogli proizaći iz preciznog predviđanja pobjednika. Efikasno predviđanje pobjednika Oscara može imati značajan uticaj na filmsku industriju, unapređujući strategije produkcije, marketinga i distribucije, te zadovoljavajući očekivanja publike i stvaralaca. Ovo istraživanje predstavlja korak napred ka sistematičnijem razumevanju faktora koji doprinose uspehu u filmskoj industriji primenom mašinskog učenja. Objašnjenje rezultata naglašava da precizno predviđanje pobjednika Oscara može značajno smanjiti rizik i troškove produkcije, omogućavajući filmskim studijima da usmere resurse na projekte sa većom šansom za uspeh.

Ovaj rad istražuje mogućnost primene metodologije mašinskog učenja u predviđanju pobjednika Oscara u kategoriji za najbolji film, kao i za glumačko ostvarenje. Izazovi pri realizaciji prediktivnog modela za predviđanje ovih kategorija su značajni i zahtevaju pažljivu analizu. Kompleksnost podataka predstavlja jedan od glavnih izazova. Podaci o filmovima i glumcima mogu obuhvatiti širok spektar atributa, uključujući žanr, kritike publike, režiju, kao i informacije o prethodnim nagradama i nominacijama. Ova raznovrsnost podataka zahteva detaljnu analizu i odabir relevantnih atributa koji će se koristiti u modelovanju. Neujednačenost podataka može biti problematična. Informacije o filmovima mogu biti neujednačene u pogledu kvaliteta i dostupnosti. Dok neki filmovi imaju detaljne informacije o svim atributima, drugi mogu biti slabo dokumentovani ili im nedostaju važne informacije. Ovo otežava proces modelovanja, naročito kada se koriste algoritmi mašinskog učenja koji zahtevaju potpune podatke.

U ovom radu je rađeno prikupljanje podataka, kao i eksplorativna analiza tih podataka. Analizirajući obimne podatke o filmovima, uključujući žanr, režiju, kao i prethodne nominacije i osvojene nagrade, primenjeni su različiti algoritmi: Logistička regresija, Random Forest, XGBoost, Veštačka neuronska mreža, Support Vector Machine i Random Forest sa Bagging-om. Ovaj ceo postupak je urađen i za predviđanje dobitnika nagrade za glumu. Takođe su primenjeni isti ovi algoritmi na

podatke kao što su godina rođenja glumca, starost, prethodne nominacije i osvajanja ove i drugih nagrada.

Kako bi se modeli evaluirali korišćena je 10-ostruka unakrsna validacija. Takođe za svaki model je računata tačnost, preciznost, odziv, *F1* mera i matrica konfuzije. Za svaki algoritam rađeno je prikaz shap vrednosti koje govore koji podaci najviše utiču na predikciju pobednika.

Ovaj rad je strukturisan u više celina. Drugo poglavlje predstavlja pregled srodnih radova koji se bave sličnom temom. Treće poglavlje obrađuje neophodne teorijske pojmove i definicije, koje su ključne za razumevanje korišćenih algoritama za predviđanje pobednika nagrade Oskar. U četvrtom poglavlju je izložena korišćena metodologija. Peto poglavlje je posvećeno eksperimentima sprovedenim na izabranim skupovima podataka i evaluaciji sistema. Šesto poglavlje pruža detaljnu analizu dobijenih rezultata. Sedmo poglavlje uključuje diskusiju o postignutim rezultatima i poređenje sa drugim radovima na ovu temu. Osmo poglavlje sadrži zaključak, koji rezimira uvide stečene tokom rada i predlaže moguće pravce za buduća istraživanja i razvoj rešenja. Deveto poglavlje, Literatura, predstavlja spisak svih korišćenih izvora, navedenih redosledom kako su citirani u tekstu. Rad se završava odeljkom sa podacima o kandidatu, koji uključuje kratku biografiju.

PREGLED STANJA U OBLASTI

U ovom poglavlju će se razmatrati rešenja relevantna za problem predikcije pobjednika nagrade Oskar.

Jedan od prvih radova vezanih za predikciju uspeha filma bio je rad iz 2000. godine. Ovaj rad [1] nudi detaljnu analizu i metodologiju za predikciju filmskih zarada, ističući važnost različitih faktora pre i nakon izlaska filma, uključujući i uticaj Oskar nominacija. Analiza se bazira na 311 filмова koji su pušteni u SAD tokom 1998. godine. Autori su koristili regresione modele za predikciju logaritmovanih vrednosti domaćih zarada filмова. Modeli su kombinacija kategorijskih i kontinuiranih prediktora (analiza kovarijanse). Model nakon prvog vikenda pokazao se znatno preciznijim, posebno za filmove puštene na više od 10 ekrana sa merom R^2 od 96.6%.

Iain Pardoe [2] se bavio statističkom analizom predikcije dobitnika Oscara u četiri glavne kategorije (najbolji film, režija, glumac i glumica u glavnoj ulozi) od 1938. do 2004. godine. Za predikciju je korišćen Multinomial Logit Model (takođe poznat kao McFaddenov uslovni logit model). Analiza obuhvata Oskare od 1938. do 2004. godine, ukupno 268 dobitnika u četiri kategorije. Model je tačno predvideo 186 od 268 dobitnika u glavnim kategorijama, što odgovara ukupnoj tačnosti od 69%. Predikcije su se poboljšale s vremenom, sa 81% tačnosti u poslednjih 30 godina (1975–2004). Uticaj varijabli poput nominacija za druge Oskare i pobjede na Zlatnim globusima postao je važniji tokom vremena, dok su prethodne pobjede imale sve manji uticaj na pobjedu u budućnosti, posebno kod glumica.

U sledećem radu Iain Pardoe i Dean K. Simonton [3] se fokusiraju na korišćenje modela diskretnog izbora za predikciju pobjednika Oscara u četiri glavne kategorije. Analizirani su podaci o nominacijama i pobjednicima Oscara od 1938. do 2006. godine, pokrivajući četiri glavne kategorije. Koriste se MNL(Multinomial Logit Model) modeli za predikciju pobjednika. Ovaj model predviđa verovatnoću izbora između diskretnih alternativa (nominovani filmovi/pojedinci). Takođe su ispitani ML(Mixed Logit Model) modeli, koji omogućavaju varijabilnost parametara između različitih izbora. Ovi modeli uklanjaju nezavisnost irelevantnih alternativa (IIA) i mogu bolje približiti stvarni proces izbora. MNL model je tačno predvideo 190 od 276 pobjednika (69%) u glavnim kategorijama od 1938. do 2006. Godine.

Rad [4] predstavlja jedan od prvih pokušaja korišćenja mašinskog učenja za predikciju dobitnika Oscara. Predikcija je takođe obuhvatala četiri glavne kategorije. Metodologije koje su korišćene za predikciju su:

SVM, Logistic Regression, Random Forest, Gaussian Naive Bayes, Multinomial Naive Bayes. Najbolji model za nagradu najbolja glumica pokazao se SVM, za nagradu najbolji glumac Multinomial Naive Bayes, a za nagrade najbolji film i najbolji režiser Logistička regresija.

Rad [5] koristi mašinsko učenje za predviđanje popularnosti filmova. Primenjene metodologije u ovom radu su: Logistic Regression, Simple Logistics, J48, Naive Bayes, Multilayer Perceptron Neural Network, PART. Metode su evaluirane koristeći 10-ostruku unakrsnu validaciju. Najveća preciznost je postignuta sa Simple Logistic i Logistic Regression od 84,34% i 84,15% respektivno. Implementirani klasifikator neuronske mreže koji je Multilayer Perceptron daje tačnost od 79,07%, a rezultati stabla odlučivanja J48 daju tačnost 82,42%. Studija je ustanovila da su najbolji rezultati postignuti upotrebom logističke regresije. Takođe, parametri metascor i broj glasova po filmu su se pokazali kao atributi koji imaju najveći uticaj na predikciju popularnosti filma.

Sledeći rad [6] takođe koristi tehnike mašinskog učenja za predviđanje uspešnosti filma. Metodologije korišćene u ovom radu su: Adaptive Tree Boosting, Gradient Tree Boosting, Linear Discriminant, Logistic Regression, Neural Network, Random Forest i Support Vector Classifier. Kako bi se izračunala tačnost modela korišćen je Average Percent Hit Rate (APHR). Najbolje performanse daje algoritam Gradient Tree Boosting, a najgore Support Vector Classifier.

U radu [7] se vrši evaluacija performansi klasifikacionih tehnika mašinskog učenja za predviđanje uspešnosti filma. Metodologije primenjene u ovom radu za rešavanje problema su: Logistic Regression, Support Vector Machine, Random Forest, Gaussian Naive Bayes, AdaBoost, Stochastic Gradient Descent, Multilayer Perceptron Neural Network. Skup podataka je sadržao 755 filmova između 2012. i 2015. godine. Logistic Regression i Gaussian Naive Bayes su efikasni za male skupove podataka ali nisu adekvatni za kompleksne podatke, gde su se druge metode poput SVM pokazale boljim. Od ukupno 755 filmova, model MLP-a tačno predviđa 442 filma. MLP postiže tačnost od 58.53% za tačno predviđanje i 89.67% za predviđanje sa jednim odstupanjem. Za evaluaciju rešenja je korišćena 10-ostruka unakrsna validacija. Pored unakrsne validacije upotrebljene su i metrike preciznost, odziv i F1 ocena.

U radu [8] se vrši predviđanje pobednika nagrade Oskar za najbolji film na osnovu odabranih karakteristika. Korišćena metodologija je Binary Logistic Regression. Od 41 dobitnika nagrade za najbolji film model je tačno predvideo njih 27 (65,9% uspešnosti), a od 200 gubitnika nagrada

model je tačno predvideo 190 njih (95% uspešnosti). Ukupno ovaj model ima 90% uspešnosti.

Tema rada [9] je da razume trendove u dodeli Oskara i da pronade korelacije između pobjedničkog entiteta i različitih varijabli. Podaci koji su analizirani se odnose na dodelu Oskara od 2000. do 2019. godine. Studija koristi metode kao što su ANOVA analiza i binarna logistička regresija sa drugim statističkim testovima za izračunavanje varijabli koje su najefikasnije u poređenju i razumevanju razlika između pobjednika i nominovanih u svakoj kategoriji. Ustanovljeno je da ogroman broj filmova nominovanih za Oskara u periodu 2000.-2019. pripadaju žanru drama. Još jedan od izvedenih zaključaka jeste da filmovi koji duže traju imaju veću šansu da osvoje Oskara.

U radu [10] vršena je predikcija ekonomskog uspeha filma korišćenjem metodologija: Random Forest, SVM i Multilayer Perceptron Neural Network. Autori su koristili skup podataka koji obuhvata 3167 filmova iz perioda između 1980. i 2019. godine. Vršili su dve klasifikacije, da li će film povratiti potrošeni budžet i kojoj profitnoj klasi film pripada, gde su filmovi klasifikovani u šest različitih kategorija na osnovu iznosa profita. Autori su koristili 10-ostruku unakrsnu validaciju za sve eksperimente, kao i metrike poput preciznosti i F1 ocene za evaluaciju tačnosti modela. Random Forest je pokazao najbolje performanse u većini evaluiranih scenarija, sa tačnošću od 96.7% u predviđanju da li će film povratiti budžet. Za Profitne klase, najbolja tačnost iznosila je 50% za skup podataka sa širokom distribucijom filmova. SVM za predikciju vezanu za budžet je dao 94.79% tačnosti, a za profitne klase 45.25%. Dok je neuronska mreža za budžet dala 94.1%, a za profitne klase 42.63%.

TEORIJSKI POJMOVI I DEFINICIJE

U ovom poglavlju će biti opisani algoritmi mašinskog učenja koji su upotrebljeni za predikciju dobitnika nagrade Oskar. U potpoglavlju 3.1 opisana je Logistička Regresija, u potpoglavlju 3.2 opisan je SVM algoritam. Zatim u potpoglavlju 3.3 opisan je Random Forest i Bagging. U potpoglavlju 3.4 opisan je XGBoost, a u potpoglavlju 3.5 veštačka neuronska mreža. U zadnjem potpoglavlju su objašnjeni One-Hot Encoding i TF-IDF pošto se koriste prilikom pretvaranja tekstualnih podataka u numerički format.

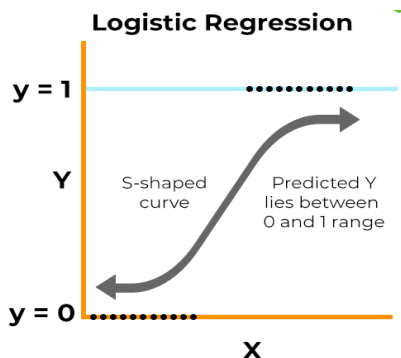
3.1. Logistička Regresija

Logistička regresija je nadgledani algoritam mašinskog učenja koji ispunjava zadatke binarne klasifikacije predviđanjem ishoda. Model daje binarni ishod ograničen na dva moguća ishoda: da/ne, 0/1 ili tačno/netačno.

Prvo se izračunava linearna kombinacija nezavisnih promenljivih (karakteristika) pomoći formule (1), gde su b_1, \dots, b_n koeficijenti regresije, a x_1, x_2, \dots, x_n nezavisne promenljive.

$$z = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n \quad (1)$$

Logistička regresija koristi logističku funkciju (takođe poznatu i kao sigmoidna funkcija) da pretvori izlaz linearne funkcije u verovatnoću. Sigmoidna funkcija se odnosi na krivu u obliku slova S koja pretvara bilo koju stvarnu vrednost u opseg između 0 i 1.



Slika 3.1.1 Prikaz logističke regresije.

Izvor: <https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/>

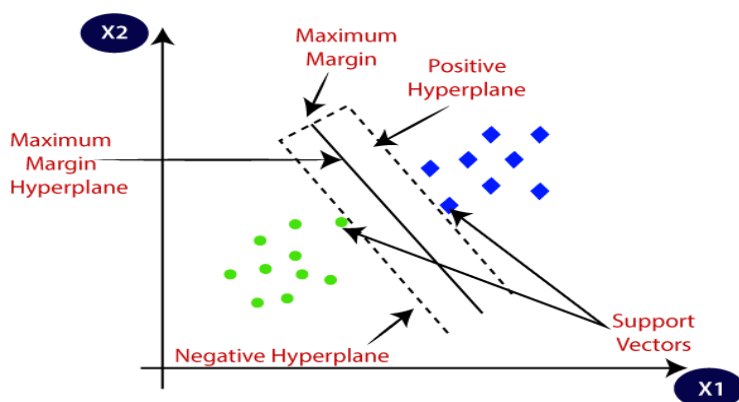
Formula (2) je sigmoidna funkcija koja se nazvana aktivaciona funkcija za logističku regresiju, gde je p verovatnoća da primer pripada pozitivnoj klasi (npr. "1" ili "da"). Nakon izračunavanja verovatnoće, koristi se prag (obično 0.5) da bi se odredila klasifikacija. Ako je $p \geq 0.5$, primer se klasifikuje kao pripadnik pozitivne klase, a ako $p < 0.5$, primer se klasifikuje kao pripadnik negativne klase.

$$p = \frac{1}{1 + e^{-z}} \quad (2)$$

3.2. Support vector machine (SVM)

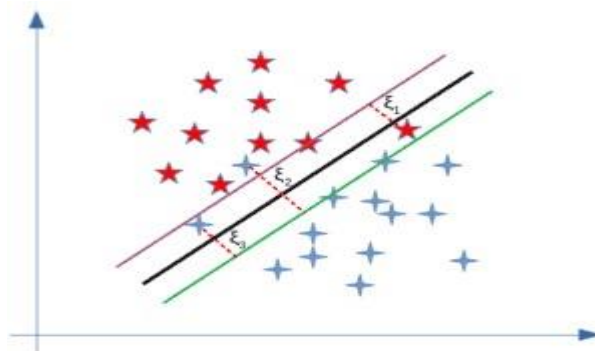
SVM je jedan od najpopularnijih algoritama za nadgledano učenje, koje se koristi za probleme klasifikacije i regresije. Međutim, prvenstveno se koristi za probleme klasifikacije u mašinskom učenju.

Osnovna ideja SVM-a je pronaći hiper-ravan (ili više njih) koji najbolje razdvaja podatke u različite klase. U dvodimenzionalnom prostoru, to je prava linija, u trodimenzionalnom je ravan, dok u višim dimenzijama postaje hiper-ravan. SVM bira ekstremne tačke/vektore koji pomažu u kreiranju hiper-ravni. Ovi ekstremni slučajevi se nazivaju vektori podrške, stoga se algoritam naziva mašina vektora podrške. Takođe, ti ekstremni slučajevi se koriste za izračunavanje margine. Margina je rastojanje između hiper-ravni i najbližih podataka iz svake klase (support vektora). SVM algoritam teži maksimizaciji ove margine, tj. da pronade hiper-ravan koja ima najveću moguću marginu između klasa.



Slika 3.2.1. Primer SVM algoritma. Izvor: <https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>

SVM algoritam podrazumevano implementira tvrdu marginu (Hard Margin), što znači da pokušava da nađe hiper-ravan koji striktno razdvaja sve podatke. Ovo je dobro kada su podaci linearno razdvojivi, međutim može biti problematično ako postoje šumovi ili podaci koji se preklapaju. U ovakvim slučajevima koristi se meka margina (Soft Margin). Ona dopušta određeni stepen greške (preklapanje podataka) kako bi model bio otporniji na šum. Soft Margin uvodi regularizacioni parametar C koji balansira između maksimizacije margine i minimizacije greške.



Slika 3.2.1. Primer SVM algoritma. Izvor:
<https://www.baeldung.com/cs/svm-hard-margin-vs-soft-margin>

Može da rešava i linearne i nelinearne probleme. Ako podaci mogu biti jasno razdvojeni pravom linijom (u 2D), odnosno hiper-ravni (u višedimenzionalnom prostoru), koristi se linearni SVM. Ako podaci nisu linearno razdvojivi, SVM može koristiti kernel funkcije kako bi transformisao podatke u viši-dimenzionalni prostor gde se mogu linearno razdvojiti. Najčešće korišćeni kerneli su *linear*, *rbf*, *poli* i drugi.

3.3. Random Forest i Bagging

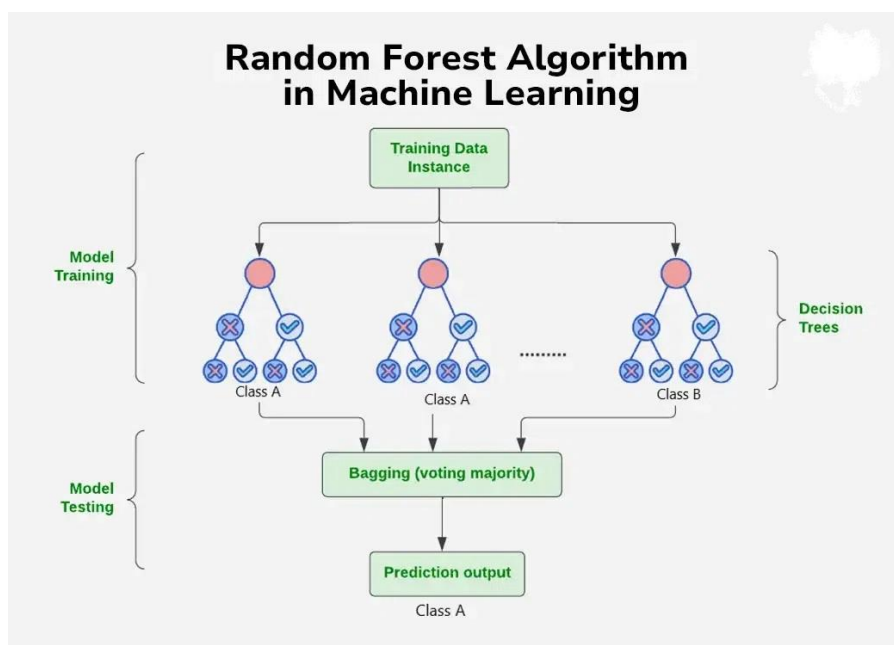
Random Forest je popularan i snažan ansambl algoritam za klasifikaciju, regresiju, i druge zadatke, koji koristi više odluka stabala za donošenje konačne odluke. Random forest algoritmi imaju tri glavna hiper-parametra koja moraju biti podešena pre treniranja. To uključuje veličinu čvora, broj stabala i broj uzoraka karakteristika.

Odluka stablo (Decision Tree) je osnovni gradivni element Random Forest-a. To je model koji donosi odluke na osnovu postavljenih uslova u čvorovima stabla, koji vode do konačnog ishoda u listovima stabla. Svako stablo u šumi se trenira na različitim podskupovima podataka. Tokom treniranja, samo nasumično odabrani podskupovi karakteristika se koriste

za podelu u svakom čvoru stabla, što dodatno smanjuje korelaciju između stabala i smanjuje preprilagođavanje (overfitting).

Ansambl metoda kombinuje predikcije više modela kako bi se poboljšale performanse u poređenju sa pojedinačnim modelima. Random Forest koristi tehniku poznatu kao *bagging* (Bootstrap Aggregating), gde više odluka stabala radi zajedno. Bagging je metoda ansambl tehnike koja kombinuje više modela kako bi smanjila varijansu i poboljšala preciznost. Ključni koraci bagginga su:

1. **Bootstrap uzorkovanje:** Kreiranje više različitih podskupova podataka (uzoraka sa zamenom) iz originalnog skupa podataka.
2. **Treniranje nezavisnih modela:** Za svaki bootstrap uzorak trenira se poseban model (u Random Forestu to su odluka stabla).
3. **Agregacija rezultata:** Predikcije svih modela se kombinuju:
 - **Za klasifikaciju:** Većinsko glasanje (majority voting).
 - **Za regresiju:** Usrednjavanje (averaging) rezultata.

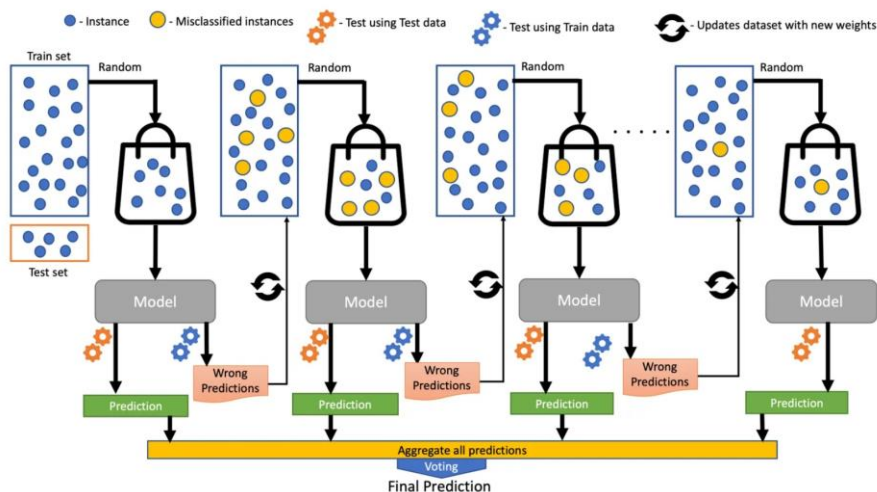


Slika 3.3.1. Struktura Random Forest algoritma. Izvor: <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>

3.4. XGBoost

XGBoost je optimizovana implementacija gradient boosting algoritma, dizajnirana da bude izuzetno efikasna, fleksibilna i prenosiva. Gradient Boosting je popularan algoritam za boosting. Kod gradient boostinga, svaki prediktor ispravlja grešku svog prethodnika.

XGBoost je implementacija Gradient Boosted stabala odluke. U ovom algoritmu, stabla odluke se kreiraju u sekvencijalnom obliku. Težine igraju važnu ulogu u XGBoost-u. One se dodeljuju svim nezavisnim promenljivama koje se zatim ubacuju u stablo odluke koje predviđa rezultate. Težina promenljivih koje su pogrešno predviđene od strane stabla se povećava i te promenljive se zatim ubacuju u drugo stablo odluke. Ovi pojedinačni klasifikatori/prediktori se zatim kombinuju da daju jači i precizniji model. Može se koristiti za regresione, klasifikacione, rangirajuće i prediktivne probleme koje definiše korisnik.



Slika 3.4.1. Primer boosting algoritma. Izvor: <https://medium.com/sfu-csmp/xgboost-a-deep-dive-into-boosting-f06c9c41349>

Definiše se funkciju gubitka koja odgovara problemu (npr. log loss za binarnu klasifikaciju, mean squared error za regresiju). Model počinje sa jednostavnom inicijalnom predikcijom, obično srednjom vrednošću ili medianom ciljne varijable. Svaki novi model u nizu trenira se da bi korigovao greške koje je napravio prethodni model. To se postiže dodavanjem sledećeg modela koji predviđa negativni gradient funkcije

gubitka. Gradient se izračunava za svaku instancu kao parcijalni derivat funkcije gubitka sa obzirom na trenutnu predikciju (3).

$$g_i^{(m)} = \frac{\partial L(y_i, \hat{y}_i^{(m-1)})}{\partial \hat{y}_i} \quad (3)$$

Ovaj gradient $g_i^{(m)}$ predstavlja koliko bi se funkcija gubitka promenila ako bi se trenutna predikcija $\hat{y}_i^{(m-1)}$ promenila. Trenira se novo stablo koje se fokusira na minimizaciju grešaka (gradijenta) prethodnog modela. Nakon što se stablo trenira, predikcije se ažuriraju tako što se dodaje nova predikcija sa težinom koja je izračunata tokom treniranja stabla (4),

$$\hat{y}_i^{(m)} = \hat{y}_i^{(m-1)} + \nu * f_m(x_i) \quad (4)$$

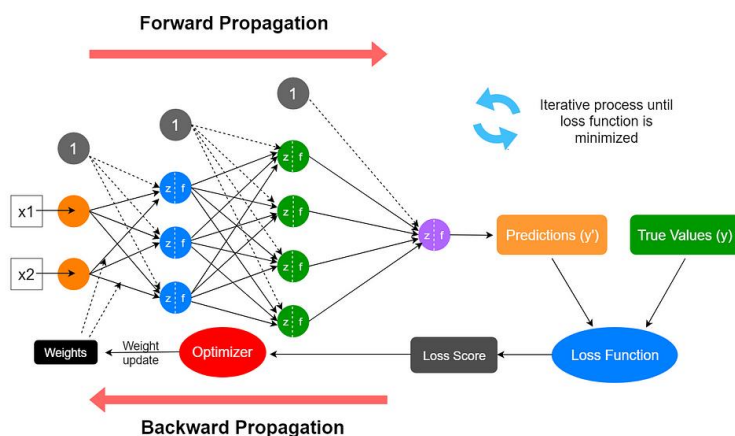
gde je ν faktor učenja (learning rate), a $f_m(x_i)$ predikcija novog stabla. Ovaj proces se ponavlja za unapred definisani broj iteracija (broj stabala) ili dok se ne postigne zadovoljavajuće poboljšanje performansi. Konačna predikcija modela se dobija kao zbir predikcija svih stabala u modelu. Svako stablo doprinosi sa svojom predikcijom umanjeno za faktor učenja. XGBoost koristi regularizaciju (L1 i L2) da bi smanjio složenost modela i sprečio preprilagođavanje. Regularizacija se dodaje u funkciju gubitka i koristi se tokom treniranja stabala.

3.5. Neuronska mreža

Neuronske mreže su osnovni deo mnogih savremenih metoda mašinskog učenja i dubokog učenja. Osnovni gradivni blokovi neuronskih mreža su neuroni. Svaki neuron prima ulaze, obrađuje ih i šalje rezultat napolje. Takođe bitna stavka kod neuronskih mreža su aktivacione funkcije. To je funkcija koja određuje izlaz neurona na osnovu njegovog ulaza. Uobičajene aktivacione funkcije uključuju:

- **Sigmoidna funkcija:** $\sigma(x) = \frac{1}{1+e^{-x}}$
- **ReLU (Rectified Linear Unit):** $ReLU(x) = \max(0, x)$
- **Tanh:** $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

Što se tiče strukture neuronske mreže ona se sastoji iz ulaznog sloja, skrivenih slojeva i izlaznog sloja. Ulazni sloj je sloj neurona koji prima ulaze (feature-e) iz podataka. Skriveni slojevi predstavljaju jedan ili više slojeva neurona između ulaznog i izlaznog sloja. Ovi slojevi uče i obrađuju podatke. Izlazni sloj je sloj neurona koji proizvodi konačne predikcije ili rezultate mreže.



Slika 3.5.1. Struktura neuronske mreže. Izvor: <https://medium.com/data-science-365/overview-of-a-neural-networks-learning-process-61690a502fa>

Proces obuke neuronske mreže obuhvata nekoliko ključnih koraka. Vršiti se postavljanje težina, gdje na početku, težine svih veza između neurona se obično inicijalizuju nasumično ili pomoću nekih heurističkih metoda (npr. Xavier inicijalizacija). Zatim se definišu aktivacione funkcije za svaki sloj (npr. ReLU, Sigmoid, Tanh). Ulazni podaci se šalju kroz mrežu. Svaki neuron izračunava svoje aktivacije na osnovu ulaza i težina. Ovo se vrši pomoću funkcije aktivacije. Formula za izračunavanje aktivacije neurona je:

$$a_j = \phi \left(\sum_i \omega_{ij} x_i + b_j \right) \quad (5)$$

gde a_j aktivacija neurona j , w_{ij} težina između neurona i i neurona j , x_i ulaz i , b_j bias za neuron j , a ϕ aktivaciona funkcija.

Takođe se računa funkcija gubitka (Loss Function), koja predstavlja meru razlike između predikcija modela i stvarnih vrednosti (ciljnih vrednosti). Primeri funkcije gubitka su Mean Squared Error za regresiju i

Cross-Entropy Loss za klasifikaciju. Formula za funkciju gubitka (npr. Cross-Entropy Loss) je:

$$L(y, \hat{y}) = - \sum_{c=1}^C y_c * \log(\hat{y}_c) \quad (6)$$

gde je y_c stvarna oznaka za klasu c (1 ako je ta klasa stvarna, u suprotnom 0), \hat{y}_c verovatnoća koju model dodeljuje klasi c , a C broj klasa.

Zatim se vrši računanje gradijenta. U ovom koraku koristi se lančano pravilo za izračunavanje gradijenata funkcije gubitka u odnosu na težine mreže. To znači da se izračunavaju parcijalni derivati funkcije gubitka u odnosu na svaku težinu u mreži. Ovi gradijenti pokazuju koliko bi se funkcija gubitka promenila ako bi se težine promenile, što omogućava modelu da se prilagodi i poboljša svoje performanse tokom obuke. Formula za gradijent težine je:

$$\frac{\partial L}{\partial \omega_{ij}} = \frac{\partial L}{\partial a_j} * \frac{\partial a_j}{\partial z_j} * \frac{\partial z_j}{\partial \omega_{ij}} \quad (7)$$

gde je z_j linearna kombinacija ulaza za neuron j , a a_j aktivacija neurona j . Zatim se vrši ažuriranje težina. Koriste se algoritmi optimizacije kao što su Stohastički Gradient Descent (SGD), Adam ili drugi za ažuriranje težina na osnovu izračunatih gradijenata. Formula za ažuriranje težine je:

$$\omega_{ij} \leftarrow \omega_{ij} - \eta * \frac{\partial L}{\partial \omega_{ij}} \quad (8)$$

gde je η brzina učenja (learning rate), a $\frac{\partial L}{\partial \omega_{ij}}$ gradijent funkcije gubitka u odnosu na težinu ω_{ij} . Ceo proces se ponavlja za više epoha. Svaka epoha prolazi kroz sve trening podatke.

Na osnovu arhitekture neuronske mreže se mogu podeliti na:

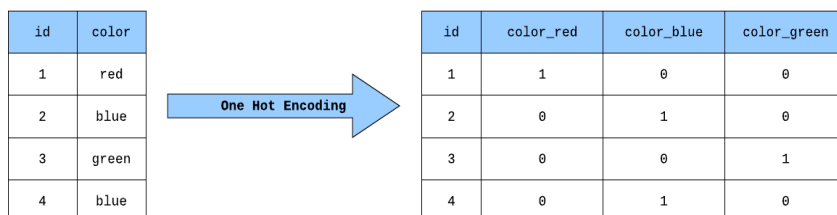
- **Jednoslojna Neuronska Mreža (Single-Layer Perceptron):** Najjednostavnija forma neuronske mreže sa jednim slojem neurona (samo izlazni sloj).
- **Višeslojna Neuronska Mreža (Multi-Layer Perceptron, MLP):** Mreža sa jednim ili više skrivenih slojeva između ulaznog i izlaznog sloja.

- **Konvolucione Neuronske Mreže (Convolutional Neural Networks, CNNs):** Koriste se za obradu slika i uključuju konvolucione slojeve za ekstrakciju karakteristika.
- **Rekurentne Neuronske Mreže (Recurrent Neural Networks, RNNs):** Koriste se za sekvencijalne podatke (npr. tekst) i imaju petlje koje omogućavaju pamćenje prethodnih informacija.
- **LSTM (Long Short-Term Memory) i GRU (Gated Recurrent Units):** Varijante RNN-a koje se koriste za bolje pamćenje dugoročnih zavisnosti.

3.6. One-Hot Encoding i TF-IDF

One-Hot Encoding i TF-IDF su tehnike za pretvaranje kategorijskih i tekstualnih podataka u numerički format, ali se koriste u različitim kontekstima i imaju različite ciljeve.

One-Hot Encoding je tehnika za pretvaranje kategorijskih (diskretnih) varijabli u numerički format. Svaka kategorija se pretvara u binarni vektor sa samo jednim aktivnim (1) bitom, dok su svi ostali bitovi nulti (0). Ovo omogućava modelima mašinskog učenja da rade sa kategorijskim podacima.



Slika 3.6.1. Primer One-Hot Encoding-a. Izvor:
<https://towardsdatascience.com/building-a-one-hot-encoding-layer-with-tensorflow-f907d686bf39>

TF-IDF (Term Frequency Inverse Document Frequency) je tehnika za pretvaranje tekstualnih podataka u numerički format, koja meri značaj svake reči u dokumentu u odnosu na sve druge dokumente u korpusu. Kombinuje dva aspekta:

- **Term Frequency (TF):** Učestalost reči u dokumentu.
- **Inverse Document Frequency (IDF):** Koliko je reč retka u celokupnom korpusu dokumenata.

TF-IDF vrednost se dobija kombinovanjem TF i IDF vrednosti kako bi se dobila ukupna važnost reči u dokumentu.

TF se izračunava kao:

$$TF(t, d) = \frac{\text{Broj pojavljivanja reči } t \text{ u dokumentu } d}{\text{Ukupan broj reči u dokumentu } d}$$

- t je reč (term)
- d je dokument

IDF se izračunava kao:

$$IDF(t, D) = \log \left(\frac{N}{\text{Broj dokumenata koji sadrže reč } t} \right)$$

- N je ukupan broj dokumenata u korpusu.
- Broj dokumenata koji sadrže reč t je broj dokumenata u kojima se reč t pojavljuje.

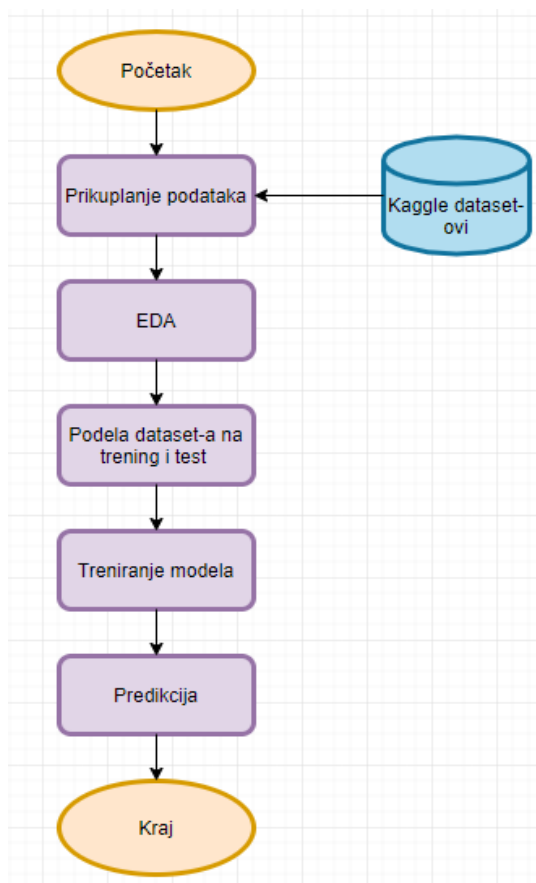
TF-IDF se izračunava kao:

$$TF - IDF(t, d, D) = TF(t, d) \times IDF(t, D)$$

METODOLOGIJA

U ovom poglavlju je detaljno opisana struktura sistema korišćenog za predikciju, koja je identična za predviđanje dobitnika Oscara za najbolji film, kao i za predikciju dobitnika Oscara za glumačko ostvarenje. Najpre je predstavljena celokupna arhitektura sistema, a zatim su detaljno objašnjeni svi njegovi ključni elementi i njihove uloge u procesu predikcije.

4.1. Struktura sistema



Slika 4.1.1.1. Šematski prikaz stukture sistema za predikciju Dobitnika Oscara

Sam sistem započinje prikupljanjem podataka sa različitih izvora i sajtova, čime se formira osnovni skup podataka za analizu. Detaljan opis ovih skupova podataka biće pružen u narednom poglavlju, gde će biti objašnjeni njihovi izvori, struktura i relevantnost za istraživanje. Kreiraju se dva skupa podataka jedan za predikciju dobitnika nagrade Oskar za najbolji film i drugi za predikciju dobitnika nagrade Oskar za glumu.

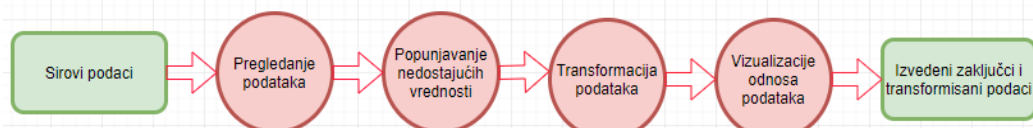
Nakon što su podaci prikupljeni, sprovedena je eksplorativna analiza podataka (EDA) kako bi se identifikovali ključni obrasci, odnosi i potencijalni problemi u podacima, uključujući nedostajuće vrednosti. Ovaj proces je omogućio dubinsko razumevanje strukture podataka i bio je ključan za njihovu dalju obradu.

Sledeći korak bio je čišćenje i sređivanje podataka, nakon čega je izvršena podela skupa podataka na trening i test skup u odnosu 80:20. Budući da se radi o nebalansiranom skupu podataka, gde postoji značajno veći broj nominovanih u odnosu na one koji su osvojili Oscara, posebna pažnja posvećena je stratifikaciji tokom podele. Stratifikacija je osigurala da svaki podskup zadrži proporcionalni odnos klasa, čime je povećana pouzdanost evaluacije modela.

Trening modela je sproveden koristeći šest različitih algoritama: Logistička Regresija, SVM, Random Forest, Random Forest uz Bagging, XGBoost i neuronska mreža. Svaki od ovih modela je obučen na trening skupu, a zatim evaluiran koristeći 10-ostruku unakrsnu validaciju, što je omogućilo robustan uvid u performanse modela. Za svaki model izračunate su ključne metrike, uključujući tačnost, preciznost, odziv, F1 meru i matricu konfuzije. Ove metrike pružaju sveobuhvatan pregled sposobnosti modela da pravilno klasifikuje primere i omogućavaju upoređivanje njihovih performansi u kontekstu konkretne aplikacije.

4. 1.1. Eksplorativna analiza podataka (EDA)

Eksplorativna analiza koja je vršena u ovom sistemu se sastoji iz nekoliko faza koje su objašnjene u nastavku.



Slika 4.1.1.1. Šematski prikaz eksplorativne analize podataka

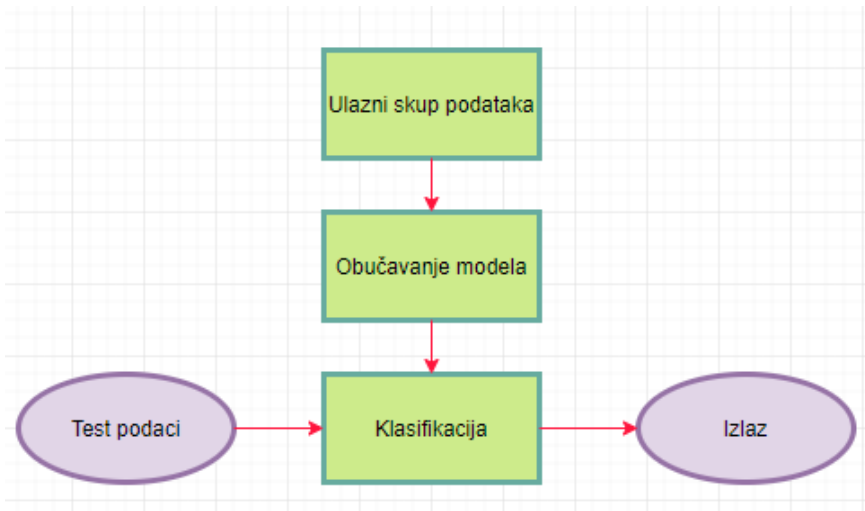
Prva faza obrade podataka obuhvatala je detaljnu analizu, pri čemu je posebna pažnja posvećena upravljanju nedostajućim vrednostima. Nedostajuće vrednosti su u najvećem broju slučajeva popunjavane računanjem srednje vrednosti ili unošenjem najčešće prisutne vrednosti u skupu podataka.

Kako bi se smanjio uticaj ekstremnih vrednosti i postigla bolja normalizacija podataka, primenjene su transformacije poput korenovanja i logaritmovanja na odabranim numeričkim atributima. Takođe, za pretvaranje tekstualnih podataka u numeričke vrednosti korišćene su metode One-Hot Encoding i TF-IDF, što je omogućilo efikasniju obradu i analizu podataka.

Na kraju ovog procesa, urađene su vizualizacije koje prikazuju odnose između različitih aspekata podataka. Ove grafičke reprezentacije pružaju dubok uvid u međusobne veze među promenljivama i omogućavaju jasnije sagledavanje obrazaca koji su prisutni u skupu podataka. Detaljan prikaz i analiza ovih vizualizacija biće predstavljeni u narednom poglavlju.

4.1.2. Modeli za klasifikaciju

U okviru ovog sistema korišćeni su različiti modeli za klasifikaciju, uključujući Logističku Regresiju, SVM, Random Forest, Random Forest uz primenu Bagging-a, XGBoost, i neuronsku mrežu. Izbor određenih algoritama, kao što su Logistička Regresija, SVM, Random Forest i neuronska mreža, bio je motivisan njihovim dokazanim performansama u prethodnim istraživanjima na slične teme. Ovi modeli su se pokazali kao pouzdani alati za rešavanje problema klasifikacije, pa su zbog toga uključeni u analizu.



Slika 4.1.2.1. Šematski prikaz klasifikacije

Pored ovih već poznatih algoritama, uvedeni su i novi modeli poput XGBoost i Bagging-a, kako bi se izvršila detaljna uporedna analiza njihovih performansi u poređenju sa tradicionalnijim pristupima. Ovaj pristup omogućava dublje razumevanje prednosti i slabosti različitih algoritama, čime se doprinosi izboru optimalnog modela za konkretan zadatak klasifikacije u ovom sistemu. Ulaz u svaki model je obrađen skup podataka sa numeričkim vrednostima, a izlaz je klasifikacija koja označava da li je osvojen Oskar ili ne.

Detalji svakog modela kao i njihovi hiper-parametri biće objašnjeni u narednom poglavlju. Sada će biti samo objašnjena struktura neuronskih mreža, korišćenih za predikciju dobitnika Oscara za najbolji film i predikciju dobitnika oscara za glumu. Obe neuronske mreže pripadaju Osnovnim Densnim Neuronskim Mrežama (Fully Connected Neural Networks, FCNN) odnosno Višeslojnim Neuronskim Mrežama (Multi-Layer Perceptron, MLP).

Prva neuronska mreža koja se koristi za predviđanje dobitnika nagrade Oskar za najbolji film se sastoji iz tri sloja. Prvi sloj se sastoji od 128 neurona gde je aktivaciona funkcija ReLU (Rectified Linear Unit). Drugi sloj ima 64 neurona i takođe koristi ReLU aktivaciju. Poslednji sloj sadrži samo jedan neuron sa sigmoidnom aktivacionom funkcijom, koja je idealna za binarne klasifikacione zadatke. Ova arhitektura je odabrana jer pruža dobar balans između složenosti i kapaciteta modela. Jednostavna struktura sa dva skrivena sloja i

progresivnim smanjenjem broja neurona omogućava efikasno učenje i generalizaciju bez prekomernog prenaučavanja, posebno na manjem skupu podataka. Model se optimizuje pomoću funkcije gubitka `binary_crossentropy`. Ova funkcija gubitka je standard za binarne klasifikacione zadatke, jer meri razliku između predikovanih verovatnoća i stvarnih binarnih oznaka (0 ili 1). Optimizacija se vrši tako da se minimizira vrednost ove funkcije, čime model postepeno poboljšava svoje predikcije tokom treniranja. Korišćenje *adam* optimizatora omogućava brz i efikasan proces treniranja.

Druga neuronska mreža koja se koristi za predikciju dobitnika nagrade Oskar za glumu sastoji se od 4 ključna sloja. Prvi sloj je sloj sa 512 neurona, koji koristi ReLU (Rectified Linear Unit) aktivaciju. Drugi sloj je sloj sa 256 neurona, koji takođe koristi ReLU aktivaciju i L2 regularizaciju sa parametrom 0.001. Treći sloj je sloj sa 128 neurona, koji koristi ReLU aktivaciju i L2 regularizaciju sa parametrom 0.001. Četvrti sloj je sloj sa jednim neuronom i sigmoid aktivacijom. Između svaka dva sloja se nalazi Dropout sloj sa stopom od 50%. L2 regularizacija pomaže u smanjenju overfittinga dodavanjem kazne za velike vrednosti težina, čime se model održava generalizovanim. Dropout slojevi, takođe pomažu u smanjenju overfittinga time što nasumično isključuju deo neurona tokom treniranja. Za funkciju gubitka takođe je korišćen `binary_crossentropy` i za optimizator *adam* kao i u prošloj neuronskoj mreži. Ova neuronska mreža je složenija u poređenju sa prethodnom zbog većeg skupa podataka koji se koristi za obuku. Veći obim podataka omogućava modelu da uči složenije obrasce i karakteristike, što opravdava upotrebu složenije arhitekture sa više slojeva. Ova dodatna složenost pomaže u efikasnijem uočavanju i modelovanju detaljnih informacija, što može poboljšati performanse modela u predviđanju.

EKSPERIMENTI

U ovom poglavlju biće predstavljene svi eksperimenti sprovedeni u okviru ovog istraživanja. Prvo će biti opisani korišćeni skupovi podataka, a zatim će detaljno biti obrađen svaki model koji je upotrebljen za predikciju dobitnika Oscara za najbolji film, kao i za predikciju dobitnika Oscara za glumačko ostvarenje. Na kraju, biće opisan postupak evaluacije koji je korišćen za procenu performansi modela.

5.1. Skupovi podataka

Ovde će biti objašnjena dva skupa podataka, jedan koji je korišćen za predikciju dobitnika Oscara za najbolji film i drugi koji je korišćen za predikciju dobitnika Oscara za glumu.

5.1.1 Skup podataka nominovanih filmova

Prilikom prikupljanja podataka, za ovaj skup podataka izabrani su filmovi koji su bili nominovani od 1961. do 2021. godine. Ovo je urađeno zbog specifičnosti aktuelnih standarda Akademije Filmske Umetnosti i Nauka prilikom odabira pobednika, koji uključuju i nedostupnost određenih atributa koji su postali relevantni tokom godina, a nisu bili prisutni na starijim ceremonijama. Jedan od primera ovog problema je uključivanje informacija o nominacijama i osvojenim nagradama na filmskim festivalima koji prethode dodeli Oscara, a koji su se počeli pojavljivati kasnije u istoriji ceremonije.

Ukupno su preuzeta tri seta podataka o nominovanim filmovima, koji su kasnije spojeni u jedan set. Setovi su organizovani u CSV datoteke. Najkompletniji preuzeti set podataka obuhvata filmska ostvarenja nominovana za najbolji film u periodu od 1961. do 2019. godine [10]. Ovaj set je poslužio kao osnova za formiranje konačnog skupa podataka zbog raznolikosti atributa kojima raspolaže. Nedostajući podaci za ceremoniju iz 2021. godine pronađeni su u okviru drugog seta podataka [10]. Spajanjem dva seta dobijen je novi set podataka od ukupno 342 uzorka koji sadrži informacije o nazivu filma, kategoriji Oscara za koji je film nominovan, nazivu nominovanog filma, godini u kojoj se održala ceremonija, ocenama koje je film postigao, ukupnom broju nominacija za nagradu Oskar, datum prve projekcije filma, nominacije i osvojene nagrade na drugim filmskim festivalima, žanr i MPPA (Motion Picture Production Association) oznake. Poslednji atribut pruža informacije o tome za koje uzraste je film prikladan. MPPA oznaka može biti važna za distribuciju i marketinške svrhe, ali nije

presudna za osvajanje nagrade koja se dodeljuje na osnovu kvaliteta filma i stoga je ovaj atribut uklonjen iz inicijalnog seta podataka.

Iz trećeg preuzetog skupa podataka [11], upotrebljena je duration kolona koja sadrži informaciju o trajanju filma izraženo u minutima. Jedini vid modifikacije ovog skupa podataka je preimenovanje labela original_title u Film kako bi se moglo uspešno izvršiti spajanje sa prethodnim setom podataka po nazivu filma. Nakon izvršenog spajanja kreiran je inicijalni set podataka od 342 filma, koji će se u nastavku referencirati kao set nominovanih filmova.

Set nominovanih filmova sadrži sledeće labelle:

- Binarna vrednost da li je film osvojio Oskara (vrednost 1) ili nije (vrednost 0), koja za potrebe ovog rada predstavlja ciljni atribut
- Godina u kojoj je film bio nominovan za Oskara, odnosno godina održavanja ceremonije
- Deo godine odnosno kvartal u kom je film prvi put prikazan publici, predstavljen kroz četiri atributa (Q1, Q2, Q3, Q4) sa binarnom vrednošću (1 – film je izašao u datom kvartalu i 0 – nije izašao u datom kvartalu)
- Ocena filma sa sajta IMDB (Internet Movie Database), predstavljena kao decimalni broj sa vrednošću u rasponu od 0.0 do 10.0
- Ocena publike preuzeta sa sajta RottenTomatoes, predstavljen kao celobrojni broj sa vrednošću u rasponu od 0 do 100
- Ocena kritičara preuzeta sa sajta RottenTomatoes (celobrojna vrednost od 0 do 100)
- Ukupan broj nominacija za Oskara u svim kategorijama
- Oskar nominacija za Najboljeg režisera, predstavljena kao binarna vrednost (1 – osvojeno i 0 – nije osvojeno)
- Podatak da li je film osvojio DGA (Director Guild of America) nagradu
- Podatak da li je film nominovan i da li je osvojio nagradu BAFTA (British Academy of Film and Television Arts) za najbolji film
- Podatak da li je film nominovan i da li je osvojio nagradu Golden Globe za najbolju dramu
- Podatak da li je film nominovan i da li je osvojio nagradu Golden Globe za najbolju komediju
- Podatak da li je film nominovan i da li je osvojio Critics choice nagradu za najbolji film

- Podatak da li je film nominovan i da li je osvojio SAGA (Screen Actors Guild Awards) nagradu za najbolju glumačku postavu
- Podatak da li je film nominovan i da li je osvojio PGA (Producers Guild Award) nagradu
- Žanr filma, koji može biti akcija, biografija, kriminalistički, komedija, drama, horor, fantazija, naučna fantastika, misterija, mjuzikl, romantični, istorijski, sa ratnom tematikom, triler, avanturistički, porodični, sportski i/ili western, pri čemu se jedan film može svrstati u više žanrova
- Trajanje filma u minutima
- Naziv filma

5.1.2. Skup podataka nominovanih glumaca

Prilikom prikupljanja ovog skupa podataka uzeti su podaci o nominovanim glumcima i glumicama u periodu od 1961. do 2021. godine. Ovaj skup podataka sadrži podatke o glumcima i glumicama nominovanih za Oskara u glavnoj ulozi kao i o glumcima i glumicama nominovanih u sporednoj ulozi.

Ukupno su preuzeta dva seta podataka sa podacima o nominovanim glumcima i glumicama, koji su kasnije spojeni u jedan set. Setovi su organizovani u CSV datoteke. Najkompletniji preuzeti set podataka obuhvata glumce i glumice nominovanih u periodu od 1961. do 2019. godine [10]. Ovaj set je poslužio kao osnova za formiranje konačnog skupa podataka zbog raznolikosti atributa kojima raspolaže. Nedostajući podaci za ceremoniju iz 2021. godine pronađeni su u okviru drugog seta podataka [10]. Spajanjem dva seta dobijen je novi set podataka od ukupno 1185 uzorka. Kao i u prošlom setu podataka izbačena je MPAA oznaka. Ovaj skup podataka će se u nastavku referencirati kao set nominovanih glumaca.

Set nominovanih glumaca sadrži sledeće labele:

- Binarna vrednost da li je glumac/glumica osvojio/la Oskara (vrednost 1) ili nije (vrednost 0), koja za potrebe ovog rada predstavlja ciljni atribut
- Kategorija gde su vrednosti: Actor, Actress, Supporting Actress, Supporting Actor
- Naziv filma za kojeg su glumci nominovani za Oskara
- Ime i prezime glumca/glumice
- Godina održavanja ceremonije Oskara

- Deo godine odnosno kvartal u kom je film prvi put prikazan publici, predstavljen kroz četiri atributa (Q1, Q2, Q3, Q4) sa binarnom vrednošću (1 – film je izašao u datom kvartalu i 0 – nije izašao u datom kvartalu)
- Ocena filma sa sajta IMDB (Internet Movie Database), predstavljena kao decimalni broj sa vrednošću u rasponu od 0.0 do 10.0
- Ocena publike preuzeta sa sajta RottenTomatoes, predstavljena kao celobrojni broj sa vrednošću u rasponu od 0 do 100
- Ocena kritičara preuzeta sa sajta RottenTomatoes (celobrojna vrednost od 0 do 100)
- Ukupan broj nominacija za Oskara u svim kategorijama za taj film
- Podatak da li je glumac/glumica nominovan/a i da li je osvojio/la nagradu BAFTA (British Academy of Film and Television Arts)
- Podatak da li je glumac/glumica nominovan/a i da li je osvojio/la nagradu Golden Globe za ulogu u drami
- Podatak da li je glumac/glumica nominovan/a i da li je osvojio/la nagradu Golden Globe za ulogu u komediji
- Podatak da li glumac/glumica ima prethodnih nominacija za Oskara i koliko
- Podatak da li glumac/glumica ima prethodnih osvajanja nagrade Oskar i koliko
- Godina rođenja glumca/glumice
- Godine glumca/glumice
- Kvartal u kom starosnom dobu je glumac/glumica osvojio/la Oskara gde su vrednosti: 0-25, 25-35, 35-45, 45-55, 55-65, 65-75, 75+
- Podatak kom žanru pripada film u kojem su glumac/glumica nominovan/a, film može da pripada više žanrova
- Podatak da li je glumac/glumica nominovan/a i da li je osvojio/la Critics choice nagradu
- Podatak da li je glumac/glumica nominovan/a i da li je osvojio SAGA (Screen Actors Guild Awards) nagradu
- Pol

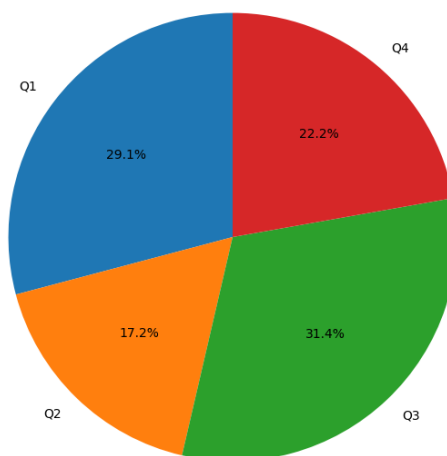
5.2. Eksperiment 1 – EDA

Ovde će biti detaljno objašnjena eksplorativna analiza koja je vršena za oba skupa podataka. Prvo će biti objašnjena za set nominovanih filmova, a onda za set nominovanih glumaca.

5.2.1. EDA za set nominovanih filmova

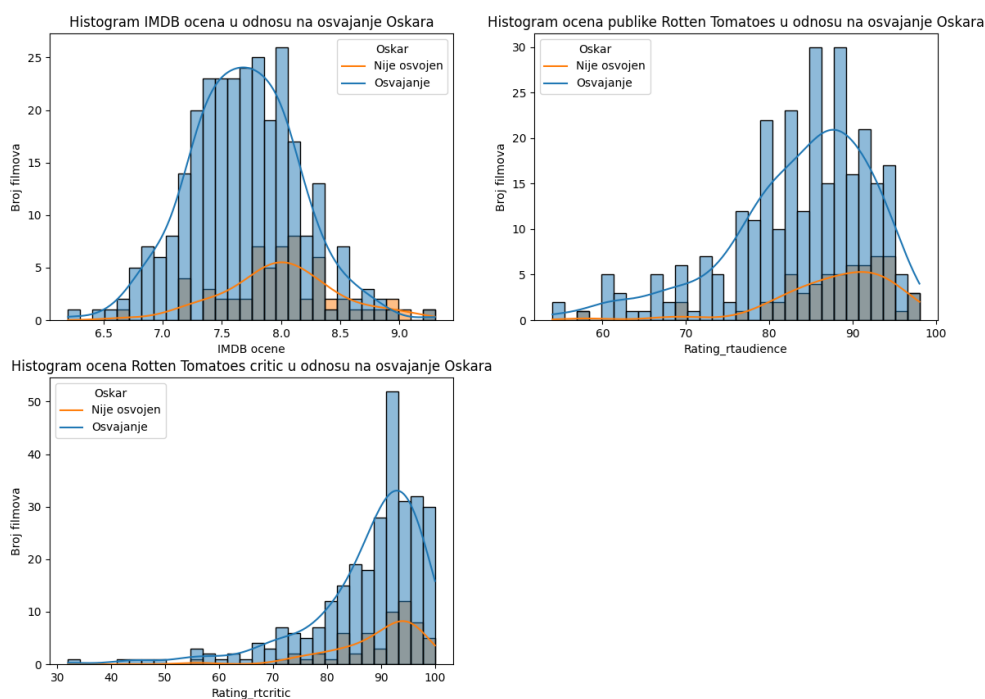
U fazi pripreme podataka, prvo je izvršeno popunjavanje nedostajućih vrednosti u koloni koja se odnosi na trajanje filma, koristeći srednju vrednost kao zamenu za prazne unose. Kako bi se smanjio uticaj ekstremnih vrednosti i doprinelo normalizaciji podataka, primenjene su transformacije korenovanja i logaritmovanja na odabrane numeričke attribute, uključujući dužinu trajanja filma, kao i ocene publike i kritičara. Pored toga, da bi se tekstualni naziv filma pretvorio u numerički format, korišćen je TF-IDF algoritam. Ovaj postupak je omogućio da se značaj svake reči u naslovu filma kvantifikuje, čime je poboljšana mogućnost analize tekstualnih podataka.

Na kraju procesa analize, izvedene su različite vizualizacije podataka koje su pružile uvid u prisutne trendove među vrednostima specifičnih obeležja u kontekstu predikcije pobednika. Prva vizualizacija otkriva da je najveći broj filmova koji su osvojili Oscara objavljen u trećoj četvrtini godine. Takođe, zanimljivo je da značajan broj filmova koji su osvojili Oscara pripada prvoj četvrtini godine, što je možda iznenađujuće i pruža dodatne uvide u obrasce izlaska filmova i njihovih uspeha.



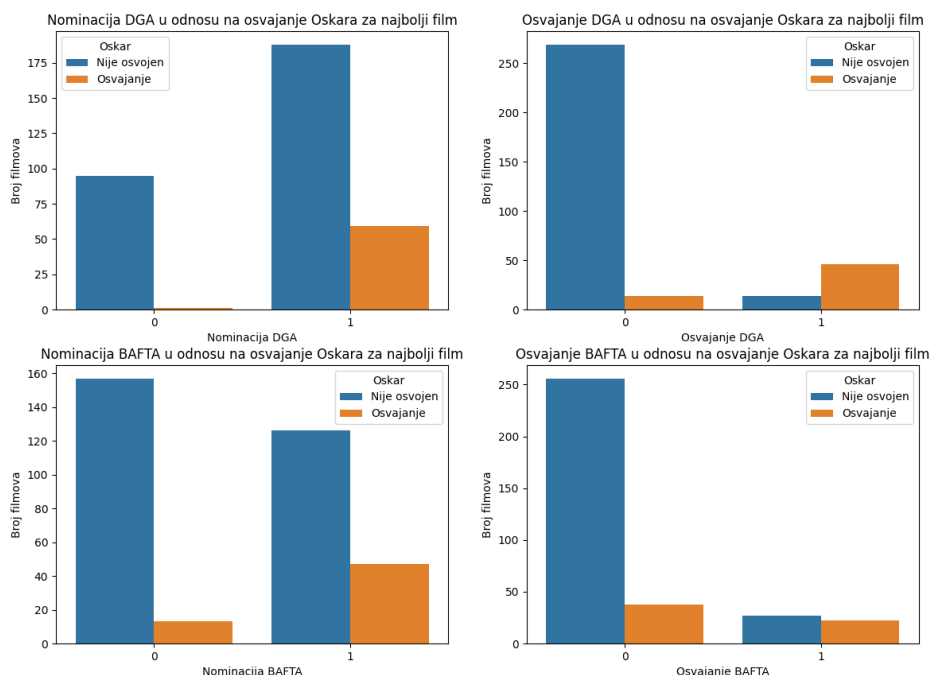
Grafik 5.2.1.1. Odnos pobedničkih filmova po četvrtinama godine

Na sledećem grafikonu su prikazani histogrami sa različitim filmskim ocenama u odnosu na osvajanje Oscara. Ocene koje su prikazane su IMDB ocena, ocena publike Rotten Tomatoes i ocena kritičara Rotten Tomatoes. Na sva tri histograma se može videti da filmovi sa višim ocenama imaju veću šansu da osvoje Oscara.



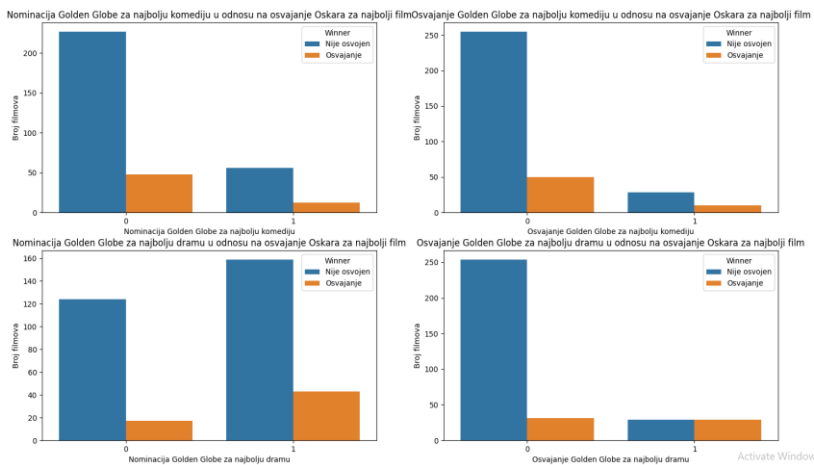
Grafik 5.2.1.2 Prikaz različitih ocena u odnosu na osvajanje Oscara

Na sledećem grafikonu su prikazane četiri odvojene analize koje upoređuju rezultate nominacija i osvajanja nagrada DGA (Directors Guild of America) i BAFTA (British Academy of Film and Television Arts) sa osvajanjima Oscara za najbolji film. Iz ovog grafika se može zaključiti da nominacija za DGA i BAFTA povećava verovatnoću da film osvoji Oscara, ali nije garant uspeha. Takođe i osvajanje DGA ili BAFTA značajno povećava šansu da film osvoji Oscara, ali i dalje ne garantuje uspeh.



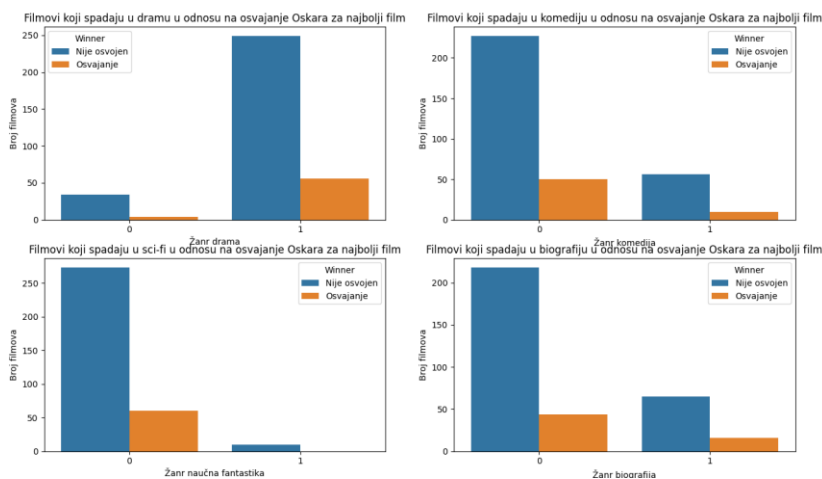
Grafik 5.2.1.3. Prikaz nominacija i osvajanja nagrada DGA i BAFTA u odnosu na osvajanje nagrade Oskar

Naredni grafikon prikazuje analizu povezanosti između nominacija i osvajanja Zlatnog Globusa u kategorijama za najbolju komediju i dramu, i njihovog uticaja na osvajanje Oscara za najbolji film. Nominacija za Zlatni Globus u obe kategorije (komedija i drama) povećava šansu za osvajanje Oscara, ali nije presudna. Osvajanje Zlatnog globusa u obe kategorije takođe povećava šanse za Oscara, ali postoji značajan broj filmova koji su osvojili Zlatni globus, a nisu osvojili Oscara. Takođe, veće šanse da osvoji Oscara ima film koji je nominovan i koji je osvojio Zlatni Globus za dramu, nego za komediju.



Grafik 5.2.1.4. Prikaz nominacija i osvajanja nagrade Zlatno globus u odnosu na osvajanje nagrade Oskar

Sledeći grafikon prikazuje odnose između žanra filmova i njihovog uspeha u osvajanju Oscara za najbolji film. Odabrani žanrovi za prikaz su: drama, komedija, biografija i naučna fantastika. Grafikoni pokazuju da su drame i biografski filmovi generalno uspešniji u osvajanju Oscara za najbolji film, dok su komedije i naučna fantastika žanrovi sa manjim šansama za osvajanje ove nagrade. Drama se ističe kao žanr sa najvećim brojem osvojenih Oscara u poređenju sa ostalima.

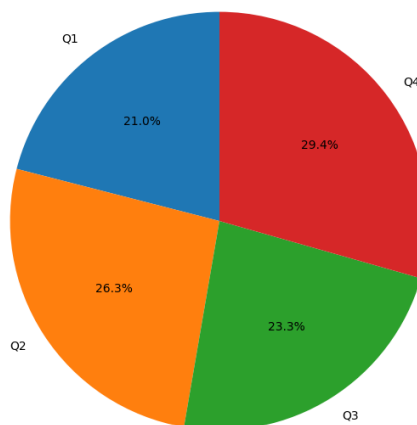


Grafik 5.2.1.4. Prikaz žanra kojem film pripada u odnosu na osvajanje nagrade Oskar

5.2.2. EDA za set nominovanih glumaca

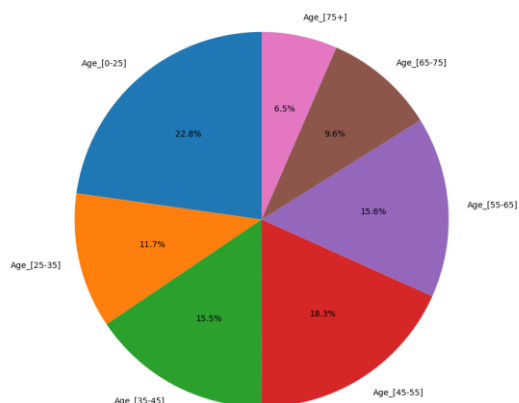
U fazi pripreme podataka, nedostajuće vrednosti u ključnim kolonama, kao što su godina rođenja glumca/glumice i njegova/njena starosna dob, popunjene su srednjom vrednošću. Nakon popunjavanja, nad numeričkim atributima, posebno onima koji predstavljaju ocene publike i kritičara, izvršene su transformacije kako bi se stabilizovala varijansa i postigla normalna distribucija. Konkretno, primenjeno je korenovanje i logaritmovanje ovih atributa. Za obradu tekstualnih podataka, nazivi filmova, koji su prvobitno bili u obliku običnog teksta, pretvoreni su u numerički format korišćenjem TF-IDF algoritma. Ova transformacija omogućila je modelu da kvantifikuje važnost svake reči unutar naslova, olakšavajući analizu. Pored toga, kolona sa kategorijom, kao i kolona koja sadrži imena i prezimena, konvertovane su u numerički format pomoću one-hot encoding tehnike. Ova tehnika je kreirala binarne kolone za svaku kategoriju ili jedinstveno ime, čime je omogućeno modelu da efikasno obradi ove promenljive. Kombinacija ovih tehnika obrade podataka obezbedila je sveobuhvatan i robusan pristup pripremi podataka za dalju analizu i modelovanje.

Na samom karaju analize podataka urađene su različite vizualizacije kako bi se uvideli prisutni trendovi među vrednostima specifičnih obeležja u kontekstu predikcije pobednika. Prvi grafikon otkriva da je najveći broj filmova za koje su glumac/glumica osvojili Oskara objavljen u poslednjoj četvrtini godine. Takođe, zanimljivo je da je relativno jednaka raspodela po četvrtinama godine u kojima je izašao film u odnosu da li je glumac/glumica osvojio/la Oskara za taj film.



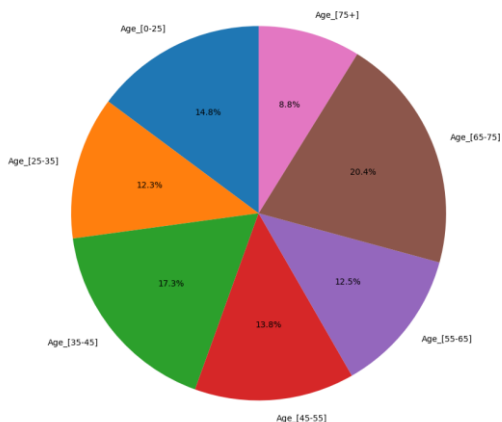
Grafik 5.2.2.1. Prikaz odnosa filmova po četvrtinama godine za koje je glumac osvojio Oskara

Naredni grafikon pokazuje u kom starosnom dobu je glumac osvojio Oscara. Može se uočiti da najveći broj glumaca koji su osvojili Oscara pripada starosnoj dobi ispod 25 godina. Međutim ima veliki procenat glumaca koji su osvojili Oscara a pripadaju starosnoj dobi 35-45, 45-55 i 65-75.



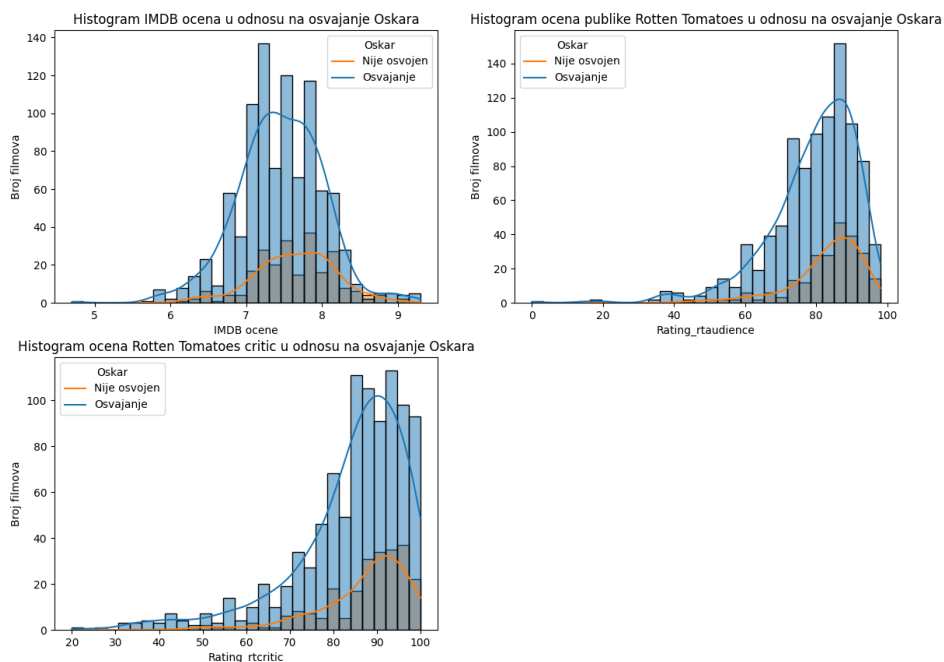
Grafik 5.2.2.2. Prikaz starosnih doba u kojima su glumaci osvojili Oscara

Sledeći grafikon pokazuje starosnu dob u kojoj je glumica osvojila Oscara. Može se uočiti da najveći broj glumica koje su osvojile Oscara pripada starosnom dobu 65-75. Takođe, značajan procenat glumica koje su osvojile Oscara pripada starosnom dobu 35-45.



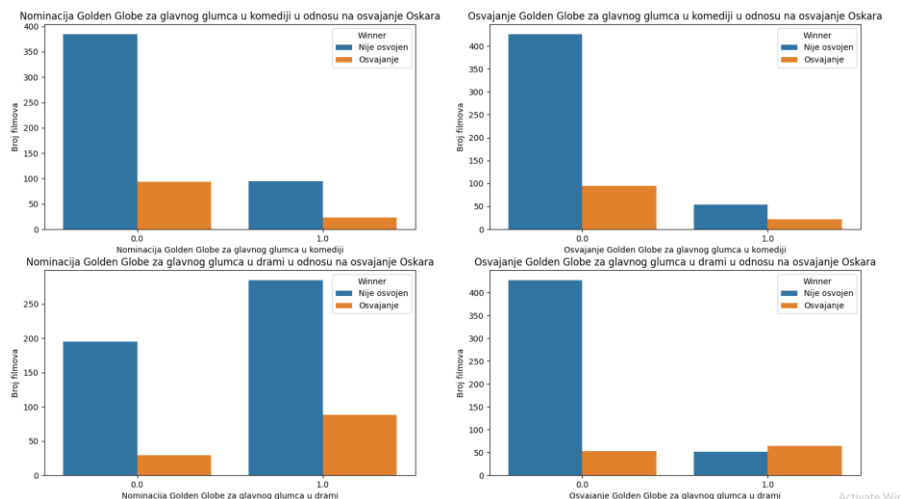
Grafik 5.2.2.3. Prikaz starosnih doba u kojima su glumice osvojile Oscara

Na sledećem grafikonu su prikazani histogrami sa različitim filmskim ocenama u odnosu na osvajanje Oscara za glumu. Ocene koje su prikazane su IMDB ocena, ocena publike Rotten Tomatoes i ocena kritičara Rotten Tomatoes. Na sva tri histograma se može videti da filmovi sa višim ocenama donose veću šansu glumcima koji u njima glume da osvoje Oscara.



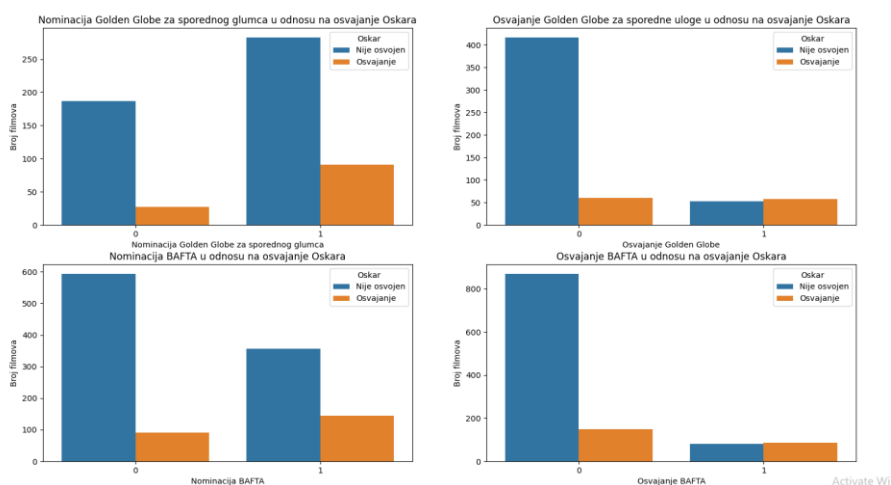
Grafik 5.2.2.4 Prikaz različitih ocena u odnosu na osvajanje Oscara

Na sledećem grafikonu su prikazane četiri odvojene analize koje upoređuju nominacije i osvajanja nagrade Zlatni Globus u glavnoj ulozi za dramu i za komediju sa osvajanjima Oscara za glumu. Iz ovog grafika se može zaključiti da nominacija i osvajanje Zlatnog Globusa za glavnu ulogu u komediji ima manju verovatnoću da će glumac osvojiti Oscara, dok nominacija i osvajanje Zlatnog Globusa za glavnu ulogu u drami povećava verovatnoću da glumac osvoji Oscara, ali nije garant uspeha.



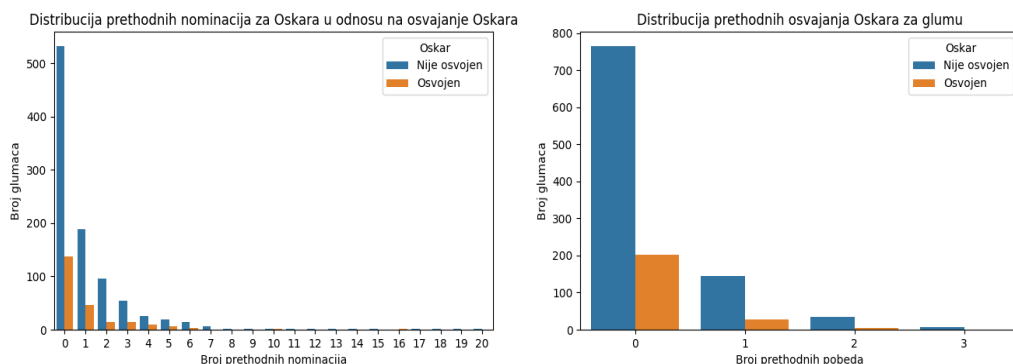
Grafik 5.2.2.5. Prikaz nominacija i osvajanja nagrade Zlatni Globus u odnosu na osvajanje nagrade Oskar

Sledeći grafikon prikazuje odnos između nominacija i osvajanja Zlatnog Globusa za sporednu ulogu, kao i BAFTA nagrada, u poređenju sa osvajanjem Oscara. Iz njega se može zaključiti da nominacije za Zlatni Globus i BAFTA nagradu povećavaju verovatnoću da glumac osvoji Oscara, ali to ne garantuje uspeh. Slično tome, osvajanje Zlatnog Globusa ili BAFTA doprinosi većoj šansi da glumac osvoji Oscara, ali ni tada uspeh nije zagarantovan.



Grafik 5.2.2.6. Prikaz nominacija i osvajanje Zlatnog Globusa i BAFTA u odnosu na osvajanje Oscara

Na narednom grafikonu se može videti odnos prethodnih nominacija i osvajanja Oscara. Iako prethodne nominacije i pobeđe mogu povećati šanse glumca da ponovo osvoji Oscara, ovi grafici pokazuju da većina pobednika zapravo nije imala prethodne nominacije ili pobeđe. To sugerise da, iako iskustvo u vidu prethodnih nominacija i pobeđa može igrati neku ulogu, ono nije presudno.



Grafik 5.2.2.7. Prikaz prethodnih nominacija i osvajanja Oscara

5.3. Eksperiment 2 – Logistička Regresija

Ovde će biti opisana dva modela Logističke regresije. Prvo će biti opisan onaj koji je korišćen za predikciju dobitnika Oscara za najbolji film, a zatim drugi koji je korišćen za predikciju dobitnika Oscara za glumu.

Hiper-parametri u prvom modelu Logističke regresije su se podešavali ručno kako bi se dobilo optimalno rešenje. Ovaj model logističke regresije je konfigurisan sa podešavanjima koja omogućavaju visoku preciznost i kontrolu nad regularizacijom i optimizacijom. Korišćenje `max_iter=2000` osigurava da model ima dovoljno iteracija za konvergenciju. Parametar `C=0.1` dodaje jaku L2 regularizaciju kako bi se smanjila prenaučenosť. Korišćenje `random_state=26` omogućava reproduktivnost rezultata, dok `penalty='l2'` i `solver='liblinear'` definišu način na koji se model regularizuje i optimizuje. `fit_intercept=False` implicira da se slobodan član ne koristi, dok `class_weight='balanced'` pomaže u rešavanju problema nerazmernih klasa.

U drugom modelu Logističke regresije takođe su se hiper-parametri podešavali ručno. Ovaj model logističke regresije je podešen za dugotrajnu obuku sa `max_iter=8000`, što omogućava modelu da prođe kroz mnogo iteracija, posebno u situacijama kada je potrebno dodatno usavršavanje. Korišćenje `C=100.0` implicira slabiju regularizaciju, što omogućava modelu da bolje prati podatke. Algoritam `lbfgs` je izabran za optimizaciju, što je efikasno za složene modele sa velikim brojem karakteristika. Postavljanje `fit_intercept=False` ukazuje na to da slobodan član nije uključen u model, dok `class_weight='balanced'` pomaže u rešavanju problema nerazmernih klasa.

5.4. Ekperiment 3 – SVM

Ovde će takođe biti opisana dva modela kao i u prošlom eksperimentu. Za optimizaciju parametara oba modela korišćen je Grid Search, koji omogućava sistematsko pretraživanje različitih kombinacija hiper-parametara kako bi se pronašli najbolji. Grid Search-a je ispitivao različite vrednosti za parametre `C`, `gamma` i `kernel`. Za prvi model najbolje su se pokazali parametri `C=1`, `gamma=0.1` i `kernel='sigmoid'`. Dok za drugi model najbolje su se pokazali parametri `C=1`, `gamma=0.01` i `kernel='sigmoid'`.

5.5. Esperiment 4 – Random Forest

Ovde su se hiper-parametri za oba modela podešavala ručno.

U prvom modelu `RandomForestClassifier` je konfigurisan sa 100 stabala, koristi Gini impurity kao kriterijum, i postavljen je na maksimalnu dubinu od 1, što rezultira veoma plitkim stablima. Minimalan broj uzoraka potreban za podelu čvora je 2, a minimalan broj uzoraka u listu je 1. Model koristi sve dostupne karakteristike za podelu i ne koristi bootstrap uzorkovanje, što znači da koristi ceo set podataka za obuku svakog stabla. Parametar `random_state` je postavljen na 42 kako bi se omogućila reprodukcija rezultata.

U drugom modelu `RandomForestClassifier` koristi 300 stabala i koristi entropiju kao kriterijum za podelu. Stabla su ograničena na maksimalnu dubinu od 4 nivoa, a za podelu čvorova je potrebno najmanje 4 uzorka. Svaki list mora sadržati najmanje 2 uzorka. Model koristi 60% dostupnih karakteristika za podelu i koristi bootstrap uzorkovanje za obuku stabala. Parametar `random_state` je postavljen na 42 za reprodukciju rezultata, dok je `class_weight='balanced'` korišćen za balansiranje težine klasa.

5.6. Eksperiment 5 – Random Forest sa Bagging-om

Ovde su takođe hiper-parametri podešavani ručno u oba modela.

U prvom modelu `RandomForestClassifier` koristi samo 10 stabala sa plitkom dubinom (maksimalno 1 nivo), Gini impurity kao kriterijum, i omogućava svaki čvor da se deli sa minimalno 2 uzorka, dok svaki list može imati samo 1 uzorak. Model koristi sve karakteristike za podelu i koristi bootstrap uzorkovanje za obuku stabala. Ovaj plitki `RandomForestClassifier` je zatim korišćen kao bazni model u `BaggingClassifier`-u sa 10 instanci, gde se rezultati svakog modela kombinuju kako bi se poboljšala ukupna stabilnost i preciznost modela. `BaggingClassifier` koristi `random_state` 42 za omogućavanje reproduktivnosti rezultata.

U drugom modelu `BalancedRandomForestClassifier` koristi 200 stabala sa plitkom dubinom (maksimalno 1 nivo) i entropiju kao kriterijum za merenje nečistoće. Model koristi 60% od ukupnog broja karakteristika za podelu i koristi bootstrap uzorkovanje za obuku stabala, sa automatskim balansiranjem klasa. Ovaj model se zatim koristi kao bazni model u `BaggingClassifier`-u, sa 10 instanci, gde se rezultati svakog modela kombinuju kako bi se poboljšala ukupna stabilnost i preciznost modela. `BaggingClassifier` koristi `random_state` 42 za omogućavanje reproduktivnosti rezultata.

5.7. Eksperiment 6 – XGBoost

Ovde će biti predstavljena dva modela gde su hiper-parametri takođe podešavani ručno.

U prvom modelu vrednost za hiper-parametar `n_estimators` je 100. Korišćenjem 100 stabala, model pokušava da postigne dobar balans između kompleksnosti i vremena obuke. Parametar `learning_rate=0.1` kontroliše brzinu kojom se model uči iz podataka. Maksimalna dubina svakog stabla u modelu je 2, dok je `min_child_weight=1`. Parametri `subsample=0.8` i `colsample_bytree=0.8` pomažu u smanjenju prekomernog prilagođavanja. Regularizacija je blago primenjena kroz `gamma=0.01`, dok su `reg_alpha=0` i `reg_lambda=1` postavljeni za osnovnu primenu regularizacije. Parametar `random_state` je 42.

Drugi model takođe koristi `n_estimators=100` i `learning_rate=0.1`. Parametri poput `max_depth=4` i `min_child_weight=2` postavljaju model na srednju složenost, dok `subsample=0.8` i `colsample_bytree=0.8` pomažu u smanjenju prekomernog prilagođavanja. Regularizacija je primenjena kroz `gamma=0.1`,

`reg_alpha=1`, i `reg_lambda=2`, čime se pomaže u održavanju modela jednostavnim i smanjenju prekomernog prilagođavanja.

5.8. Eksperiment 7- Neuronska mreža

Ovde su takođe trenirana dva modela čija je detaljna struktura već objašnjena u prethodnom poglavlju. Ova dva modela se razlikuju po svojoj složenosti i broju slojeva, u zavisnosti od obima podataka na kojima se obučavaju.

Prvi model je dizajniran da bude jednostavniji, jer se obučava na manjem skupu podataka. Sastoji se od tri sloja, što omogućava efikasno treniranje i generalizaciju na manjem skupu podataka. Drugi model je, s druge strane, složeniji, jer se obučava na većem skupu podataka, što zahteva dodatne slojeve kako bi se uhvatili složeniji obrasci u podacima. Ovaj model sadrži četiri ključna sloja, a između svakog para slojeva umetnuti su Dropout slojevi. Dropout slojevi imaju važnu ulogu u smanjenju overfitting-a, jer nasumično isključuju određeni procenat neurona tokom treninga, što pomaže u sprečavanju modela da postane previše prilagođen specifičnom skupu podataka. Obe neuronske mreže koriste za optimizator `adam`, a za funkciju gubitka `binary_crossentropy`.

Što se tiče parametara treninga, prvi model je treniran tokom 30 epoha, sa `batch_size` od 15. Ovaj manji broj epoha i veličina batch-a odgovaraju jednostavnijoj arhitekturi i manjem skupu podataka, omogućavajući bržu obuku bez gubitka preciznosti. Drugi model, zbog svoje složenosti i većeg skupa podataka, zahtevao je obuku tokom 100 epoha, sa `batch_size` od 64. Ovi parametri omogućili su modelu da detaljno nauči složenije obrasce iz podataka, uz istovremeno smanjenje rizika od overfitting-a zahvaljujući Dropout slojevima.

5.9. Evaluacija

Nakon sređivanja podataka podaci su podeljeni na trening i test skup u odnosu 80:20. Takođe pošto se radi o nebalansiranom skupu podataka, jer ima mnogo više nominovanih od onih koji su osvojili Oskara, pri podeli podataka vođeno je računa o stratifikaciji. Kako bi se modeli evaluirali korišćena je 10-ostruka unakrsna validacija i to je uzeta u obzir tačnost. Takođe za svaki model je računata preciznost, tačnost, odziv, F1 mera i matrica konfuzije. Koršćena je 10-ostruka unakrsna validacija što znači da je trening skup podeljen na 10 delova (foldova). U svakom ciklusu, model se trenira na 9 foldova, a testira na preostalom jednom. Ovaj proces se ponavlja 10 puta, pri čemu se svaki fold koristi jednom kao test skup.

REZULTATI

U ovom poglavlju biće dati rezultati svih eksperimenata pomenutih u prethodnom poglavlju. Prvo će biti prikazani rezultati po eksperimentima, a onda prikaz rezultata svih eksperimenata zajedno.

6.1. Rezultati eksperimenta 2

Ovde će biti prikazani rezultati dva modela Logističke Regresije. Prvi koji vrši predikciju dobitnika Oscara za najbolji film i drugi koji vrši predikciju dobitnika Oscara za glumu. U prvoj tabeli je prikazana tačnost algoritma po svakom fold-u i ukupna tačnost. Dok u drugoj tabeli se nalazi tačnost, preciznost, odziv i mera F1.

Fold	Model 1	Model 2
1	92,86%	85,26%
2	96,43%	81,05%
3	85,71%	78,73%
4	85,71%	86,31%
5	88,89%	82,1%
6	85,18%	82,1%
7	81,48%	79,89%
8	77,78%	88,21%
9	100%	87,25%
10	85,18%	86,6%
Ukupno	87,92%	83,75%

Tabela 6.1.1. Prikaz rezultata 10-ostruke unakrsne validacije za logističku regresiju

	Model 1	Model 2
Tačnost	92, 75%	86,5%
Preciznost	76,92%	66,67%
Odziv	83,33%	63,82%
F1	80%	65,22%

Tabela 6.1.2. Prikaz metrika za logističku regresiju

6.2. Rezultati eksperimenta 3

Ovde su takođe prikazani rezultati za oba SVM modela u dve tabele kao i za prošli eksperiment.

Fold	Model 1	Model 2
1	80%	83,19%
2	84,28%	82,35%
3	91,42%	84,03%
4	79,41%	86,55%
5	91,18%	84,03%
6	85,29%	79,66%
7	73,53%	83,05%
8	88,23%	82,2%
9	94,12%	87,28%
10	97,06%	85,59%
Ukupno	87,45%	83,54%

Tabela 6.1.1. Prikaz rezultata 10-ostruke unakrsne validacije za SVM

	Model 1	Model 2
Tačnost	88,41%	84,81%
Preciznost	70%	63,33%
Odziv	58,33%	40,42%
F1	63,63%	49,35%

Tabela 6.2.2. Prikaz metrika za SVM

6.3. Rezultati eksperimenta 4

Ovde su prikazani rezultati za dva Random Forest modela, takođe su prikazani u dve tabele.

Fold	Model 1	Model 2
1	92,86%	87,37%
2	100%	77,89%
3	85,71%	78,95%
4	92,86%	87,37%
5	92,59%	84,21%
6	88,89%	83,16%
7	96,3%	81,05%
8	81,48%	82,1%
9	96,3%	85,1%
10	88,89%	80,85%
Ukupno	91,59%	82,81%

Tabela 6.3.1. Prikaz rezultata 10-ostruke unakrsne validacije za Random Forest

	Model 1	Model 2
Tačnost	92,75%	84,39%
Preciznost	88,89%	59,61%
Odziv	66,67%	65,96%
F1	76,19%	62,63%

Tabela 6.2.2. Prikaz metrika za Random Forest

6.4. Rezultati eksperimenta 5

Ovde su prikazani rezultati dva modela Random Forest sa Bagging-om, isto u dve tabele.

Fold	Model 1	Model 2
1	92,86%	80%
2	92,86%	77,89%
3	96,43%	87,36%
4	96,43%	80%
5	85,18%	84,21%
6	85,18%	82,1%
7	96,30%	77,89%
8	85,18%	87,37%
9	100%	79,78%
10	85,18%	82,98%
Ukupno	91,56%	81,96%

Tabela 6.4.1. Prikaz rezultata 10-ostruke unakrsne validacije za Random Forest sa Bagging-om

	Model 1	Model 2
Tačnost	92,75%	82,43%
Preciznost	88,89%	58,34%
Odziv	66,67%	61,57%
F1	76,19%	48,57% %

Tabela 6.4.2. Prikaz metrika za Random Forest sa Bagging-om

6.5. Rezultati eksperimenta 6

Ovde su prikazani rezultati dva XGBoost modela, takođe u dve tabele.

Fold	Model 1	Model 2
1	92,86%	86,31%
2	96,43%	81,05%
3	85,71%	80%
4	92,86%	86,31%
5	92,59%	84,21%
6	92,59%	85,26%
7	88,89%	80%
8	81,48%	85,26%
9	96,29%	87,23%
10	85,18%	85,1%
Ukupno	90,49%	84,39%

Tabela 6.5.1. Prikaz rezultata 10-ostruke unakrsne validacije za XGBoost

	Model 1	Model 2
Tačnost	91,3%	86,07%
Preciznost	100%	71,87%
Odziv	50%	51,06%
F1	66,67%	58,54%

Tabela 6.5.2. Prikaz metrika za XGBoost

6.6. Rezultati eksperimenta 7

Ovde su prikazani rezultati dva modela neuronske mreže, takođe u dve tabele.

Fold	Model 1	Model 2
1	88,57%	84,03%
2	97,14%	78,89%
3	88,57%	82,35%
4	97,06%	80,67%
5	97,06%	75,63%
6	95,29%	81,36%
7	88,24%	74,58%
8	82,35%	81,36%
9	91,18%	79,66%
10	82,35%	76,27%
Ukupno	89,78%	81,16%

Tabela 6.6.1. Prikaz rezultata 10-ostruke unakrsne validacije za neuronske mreže

	Model 1	Model 2
Tačnost	89,85%	82,15%
Preciznost	72,73%	52,36%
Odziv	66,67%	57,44%
F1	69,56%	51,92%

Tabela 6.6.2. Prikaz metrika za neuronske mreže

6.7. Uporedna analiza rezultata modela

U ovom potpoglavlju pružen je sveobuhvatan pregled rezultata svih korišćenih modela, uz zaključke o tome koji su se modeli pokazali kao najbolji, a koji kao najgori u predikciji dobitnika Oscara za najbolji film i glumu.

Prvo su prikazani rezultati za sve modele koji vrše predikciju dobitnika nagrade Oskar za najbolji film. Nakon toga su prikazani rezultati modela koji vrše predikciju dobitnika nagrade Oskar za glumu.

Modeli	10-ostruka unakrsna validacija
Logistic Reggresion	87,92%
SVM	87,45%
Random Forest	91,59%
RF sa Bagging-om	91,56%
XGBoost	90,49%
NN	89,78%

Tabela 6.7.1. Prikaz 10-ostuke unakrsne validacije za modele vrše predikciju za najbolji film

U analizi predikcije dobitnika nagrade Oskar za najbolji film, najbolji rezultati su postignuti korišćenjem modela Random Forest i Random Forest sa Bagging-om. Ova dva modela su pokazala gotovo identične performanse, što ukazuje na to da dodatna složenost koju donosi Bagging nije značajno unapredila već odlične rezultate koje pruža Random Forest.

Međutim, većina modela se suočava s izazovima kada je reč o tačnoj klasifikaciji pozitivnih instanci, odnosno filmova koji su zaista osvojili nagradu. U ovom kontekstu, Logistička regresija se ističe kao model koji najbolje klasifikuje pozitivne instance, pružajući značajno bolje rezultate u poređenju sa ostalim modelima.

Sa druge strane, algoritam XGBoost, iako efikasan u tačnoj klasifikaciji svih negativnih instanci (filmova koji nisu osvojili nagradu), pokazuje slabost kada je reč o klasifikaciji pozitivnih instanci. Ovo znači da, iako XGBoost precizno prepoznaje filmove koji nisu osvojili nagradu, nije pouzdan u predviđanju stvarnih dobitnika.

Kada je reč o ukupnim rezultatima, SVM algoritam se pokazao kao najmanje uspešan. Njegove performanse su značajno slabije u poređenju sa ostalim modelima, što ga čini najmanje pogodnim za ovaj konkretan zadatak predikcije Oskara.

Modeli	10-ostruka unakrsna validacija
Logistic Reggresion	83,75%
SVM	83,54%
Random Forest	82,81%
RF sa Bagging-om	81,96%
XGBoost	84,39%
NN	81,16%

Tabela 6.7.2. Prikaz 10-unakrsne validacije za modele vrše predikciju za glumu

U analizi predikcije dobitnika nagrade Oskar za glumu, algoritam XGBoost se istakao kao najuspešniji model, postigavši najbolje ukupne rezultate u predviđanju pobjednika. Uz XGBoost, algoritmi SVM i Logistička regresija također su pokazali vrlo dobre performanse, demonstrirajući visoku tačnost i pouzdanost.

Međutim, slično kao i u prethodnoj analizi predikcije za najbolji film, većina modela se suočava sa izazovima u tačnom predviđanju pozitivnih instanci, odnosno slučajeva u kojima su glumci zaista osvojili nagradu. U ovom kontekstu, Random Forest se izdvojio kao model koji najbolje klasifikuje pozitivne instance, omogućavajući precizno predviđanje dobitnika. Logistička regresija je također postigla dobre rezultate u ovoj oblasti, odmah nakon Random Forest-a.

Sa druge strane, algoritam SVM se pokazao najpouzdanijim u klasifikaciji negativnih instanci, odnosno u tačnom prepoznavanju glumaca koji nisu osvojili nagradu. Ovo znači da, iako SVM nije najbolji u predviđanju pobjednika, on izuzetno dobro razlikuje slučajeve kada nagrada nije osvojena.

Nasuprot ovim uspešnim modelima, neuronska mreža se pokazala kao najlošiji model za predikciju dobitnika Oscara za glumu. Njene performanse su bile znatno ispod očekivanja, što ukazuje na to da je ova metoda možda previše složena ili neprilagođena specifičnostima podataka u ovom slučaju. Pored toga, iako je Random Forest sa Bagging-om pružio određeni nivo preciznosti, njegove performanse nisu bile značajno bolje od običnog Random Forest modela i nisu bile dovoljno dobre da se izdvoji kao efikasan pristup za ovu vrstu predikcije.

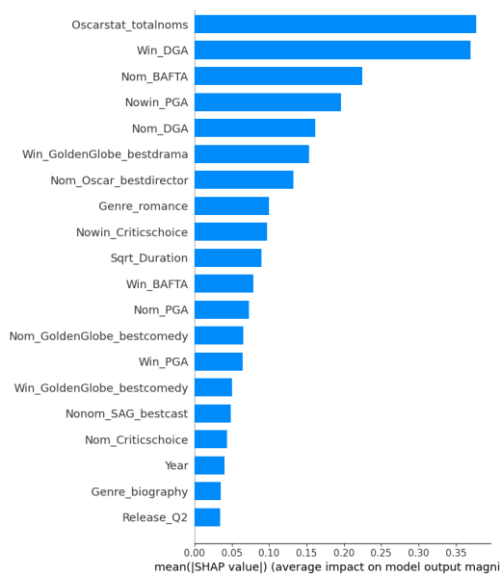
DISKUSIJA

U ovom poglavlju pružena je detaljna analiza rezultata, zajedno sa ključnim zaključcima izvedenim iz istraživanja. Nakon toga, sledi deo u kojem se vrše poređenja sa prethodnim studijama, čime se ističu određene prednosti i mane ovog rada.

Analizom dobijenih rezultata utvrđeno je da najbolje performanse za predikciju za najbolji film imaju Random Forest i Random Forest sa Bagging-om, takođe i Logistička regresija, a najgore SVM model.

Eksplorativnom analizom podataka utvrđeno je da filmovi koji su osvojili Oskara često imaju i nominaciju za najboljeg režisera. Takođe kao što je i očekivano da najviše nominovanih filmova za nagradu Oskar i onih koji su osvojili tu nagradu pripadaju žanru drama. Što se tiče odnosa nagrade Oskar sa drugim nagradama najviše ima poklapanja sa PGA i DGA nagradama, odnosno filmovi koji su osvojili ove nagrade češće osvajaju nagradu Oskar. Još jedno od zanimljivih zapažanja jeste da najviše filmova koji su nominovani za nagradu Oskar su izašli u poslednjoj četvrtini godine što je i očekivano, a oni koji su je zapravo osvojili izašli su u trećoj i prvoj četvrtini godine.

Dodatnom analizom pomoću SHAP vrednosti utvrđeno je koja obeležja imaju najveći uticaj na predikciju, uglavnom su to nominacije filma za neku nagradu ili da li je osvojena neka nagrada. Kod Logističke regresije prva tri parametra koja utiču na predikciju su ukupan broj nominacija za Oskara, osvojena DGA nagrada i nominacija za BAFTA nagradu. Dok kod XGBoost algoritma su osvojena DGA nagrada, ocena Rotten Tomatoes kritičara i nominacija za BAFTA nagradu. Uglavnom kod svih algoritama prva tri parametra koja utiču na predikciju dobijenih SHAP vrednostima su slični, obično se dva poklapaju a treći se razlikuje i redosled im je drugačiji. U nastavku je dat prikaz SHAP vrednosti za Logističku regresiju (grafik 7.1.) kao i prikaz SHAP vrednosti za XGBoost algoritam (grafik 7.2.).



Grafik 7.1. Prikaz SHAP vednosti za logističku regresiju

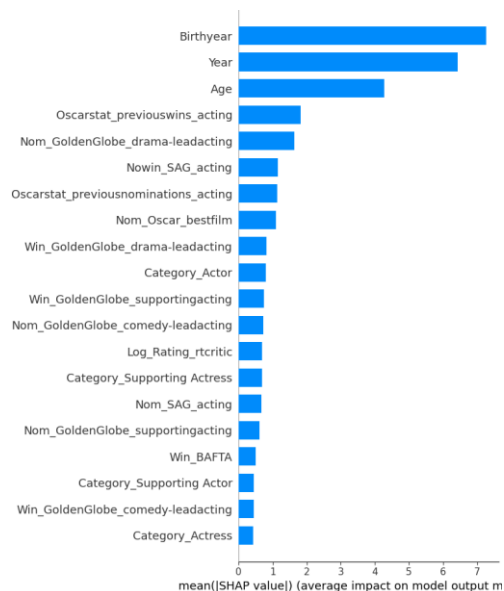


Grafik 7.2. Prikaz SHAP vednosti za XGBoost

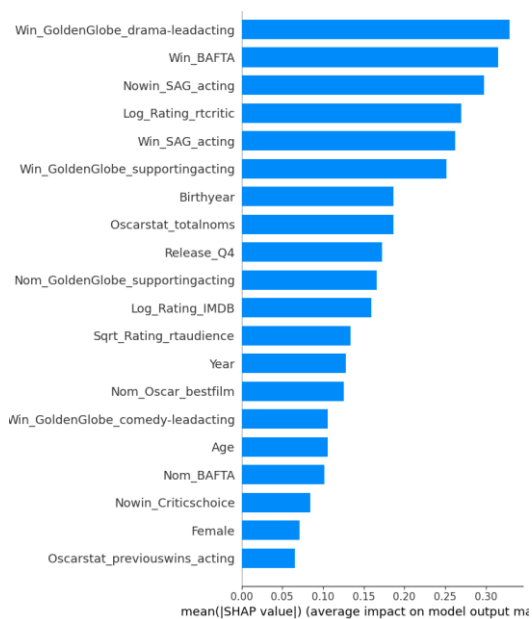
Analizom dobijenih rezultata utvrđeno je da najbolje performanse za predikciju za glumu imaju XGBoost i Logistička regresija, takođe i SVM, a najgore neuronska mreža.

Eksplorativnom analizom podataka utvrđeno je da veću šansu imaju glumci da osvoje Oskara ako je film u kojem su glumili nominovan za Oskara. Takođe kao što je i očekivano da najviše nominovanih glumaca za nagradu Oskar i onih koji su osvojili tu nagradu su glumili u filmovima koji pripadaju žanru drama. Što se tiče odnosa nagrade Oskar sa drugim nagradama, veću verovatnoću da osvoje Oskara imaju glumci ako su osvojili Zlatni globus za dramu nego za komediju. Takođe je veća verovatnoća da će glumac osvojiti Oskara ako je osvojio i BAFTA nagradu. Još jedno od zanimljivih zapažanja jeste da najviše glumaca koji su nominovani za Oskara su glumili u filmovima koji su izašli u poslednjoj četvrtini godine što je i očekivano, a oni koji su je zapravo osvojili glumili su u filmovima koji su izašli u drugoj i poslednjoj četvrtini godine. Takođe, jedno od zanimljivih zapažanja jeste da najviše glumaca koji su osvojili Oskara pripadaju starosnom dobu ispod 25 godina i 45-55 godina. Dok glumice koje su osvojile Oskara pripadaju najviše starosnom dobu 35-45 i 65-75 godina.

Dodatnom analizom pomoću SHAP vrednosti utvrđeno je koja obeležja imaju najveći uticaj na predikciju. Uglavnom su to nominacije glumca za neku nagradu ili da li je osvojena neka nagrada ili godište glumca. Kod Logističke regresije prva tri parametra koja utiču na predikciju su godina rođenja glumca, godina izlaska filma za koji su nominovani i same godine glumca. Dok kod XGBoost algoritma su osvojena nagrada Zlatni Globus, BAFTA i SAG nagrada. Uglavnom kod svih ostalih algoritama prva tri parametra koja utiču na predikciju dobijenih SHAP vrednostima su slični, uglavnom su to neka osvajanja ili nominacije nekih nagrada. U nastavku je dat prikaz SHAP vrednosti za Logističku regresiju (grafik 7.3.) kao i prikaz SHAP vrednosti za XGBoost algoritam (grafik 7.4.).



Grafik 7.3. Prikaz SHAP vednosti za logističku regresiju



Grafik 7.4. Prikaz SHAP vednosti za XGBoost

U radu [4] korišćeni su različiti algoritmi, uključujući SVM, Logističku regresiju i Random Forest, pri čemu je Logistička Regresija dala najbolje rezultate za predikciju dobitnika nagrade Oskar za najbolji film. Kada je reč o predikciji nagrada za glumu, algoritam SVM se pokazao najuspešnijim za kategoriju najbolje glumice, dok je Multinomial Naive Bayes postigao najbolje rezultate u kategoriji najboljeg glumca. U ovom radu takođe su primenjeni algoritmi SVM, Logistička regresija i Random Forest. Za predikciju dobitnika nagrade Oskar za najbolji film, Random Forest je ostvario rezultate jednake onima koje je postigla Logistička Regresija. Međutim, za predikciju Oskara u kategorijama za glumu, SVM se nije pokazao kao najbolji izbor. Što se tiče skupova podataka, rad [4] i ovaj rad koriste neke zajedničke atribute, poput godine izlaska filma, ocena publike i kritičara na Rotten Tomatoes, žanra, trajanja filma i IMDB ocene. Međutim, ovaj rad dodatno uključuje podatke vezane za nominacije i osvajanja drugih prestižnih nagrada, kao što su BAFTA, Zlatni globus, DGA, PGA, SAGA, i druge, čime se proširuje analiza i omogućava sveobuhvatnija predikcija.

U radu [7] za predikciju uspešnosti filmova korišćeni su algoritmi kao što su Logistička regresija, SVM, Random Forest, i Multilayer Perceptron (MLP) neuronska mreža, koji su takođe primenjeni i u ovom istraživanju. Međutim, skup podataka korišćen u radu [7] značajno se razlikuje od skupova podataka koji su korišćeni ovde, što otežava direktna poređenja rezultata. Logistička regresija se u radu [7] pokazala kao efikasna na manjim skupovima podataka, dok je SVM bio uspešniji kada su u pitanju kompleksniji podaci. U ovom istraživanju, Logistička regresija se pokazala kao vrlo pouzdana ne samo za predikciju dobitnika Oskara za najbolji film, koji se oslanja na manji skup podataka, već i za predikciju dobitnika Oskara za glumu, gde je skup podataka veći i složeniji. Sa druge strane, SVM algoritam, kao i u radu [7], nije bio toliko efikasan na manjem skupu podataka za najbolji film, ali je bio jedan od uspešnijih modela za predikciju dobitnika Oskara za glumu, gde je skup podataka složeniji. U radu [7], MLP neuronska mreža sa tri skrivena sloja postigla je tačnost od 58,53% za tačne predikcije i 89,67% za predikcije sa jednim odstupanjem. U ovom radu, prva neuronska mreža sa dva skrivena sloja postigla je tačnost od 89,78%, dok je druga, koja ima tri skrivena sloja sa Dropout slojem između svakog para slojeva, postigla tačnost od 81,16%.

U radu [8], koji se bavio predviđanjem dobitnika Oskara za najbolji film, korišćena je Logistička regresija, kao i u ovom radu. Njihov model postigao je tačnost od 90%, dok je u ovom istraživanju

postignuta nešto viša tačnost od 92,75%. Iako se skupovi podataka razlikuju, oba rada su koristila neke iste parametre, kao što su žanr filma, datum objavljivanja, trajanje filma i ukupan broj nominacija za Oskara. Oba istraživanja su došla do sličnih zaključaka. Jedan od ključnih zaključaka je da većina filmova koji osvoje Oskara za najbolji film pripadaju žanru drama, što ukazuje na značaj ovog žanra u predviđanju pobjednika. Takođe, oba rada su istakla važnost nominacije Oskara u kategoriji najbolji režiser, koja značajno povećava šanse da film osvoji Oskara za najbolji film. Ovi rezultati naglašavaju kako određeni faktori, poput žanra i režiserske nominacije, mogu biti ključni pokazatelji uspeha filma na dodeli Oskara.

Rad [9] bavi se analizom trendova u dodeli Oskara, s ciljem da identifikuje korelacije između pobjedničkih filmova i različitih varijabli. U tom istraživanju, analizirani su podaci o dodeli Oskara za period od 2000. do 2019. godine u svim kategorijama, dok ovaj rad obuhvata širi vremenski okvir, od 1961. do 2021. godine i samo kategorije glume i najbolji film. Iako su korišćene različite metodologije i skupovi podataka, oba istraživanja su došla do sličnih zaključaka. Kao što je već pomenuto u radu [8], zaključeno je da većina filmova nominovanih za Oskara, posebno u kategoriji za najbolji film, pripada žanru drama. Ova korelacija između žanra i uspeha na dodeli Oskara ukazuje na dominantnu ulogu drame u filmskoj industriji kada su u pitanju prestižne nagrade. Takođe, oba rada su primetila da filmovi sa dužim trajanjem imaju veću verovatnoću da osvoje Oskara u kategoriji za najbolji film. Ova zapažanja sugerišu da publika i glasači možda vrednuju složenije, temeljno razvijene priče koje se obično nalaze u dužim filmovima, što doprinosi njihovom uspehu na dodeli nagrada.

U radu [10] fokus je bio na predikciji ekonomske uspešnosti filma, što predstavlja značajnu razliku u odnosu na temu ovog rada, gde se predviđa osvajanje Oskara za najbolji film. Iako oba rada koriste slične tehnike mašinskog učenja, kao što su SVM, Random Forest, i Multilayer Perceptron (MLP) neuronska mreža, postoji nekoliko ključnih razlika koje otežavaju direktno poređenje. Prvo, podaci korišćeni u radu [10] obuhvataju filmove iz perioda od 1980. do 2019. godine, dok u ovom istraživanju podaci pokrivaju širi vremenski period, od 1961. do 2021. godine. Iako oba skupa podataka sadrže zajedničke karakteristike kao što su žanr filma,

trajanje filma, i datum izlaska, razlikuju se u mnogim drugim aspektima. Na primer, podaci u ovom radu fokusiraju se na faktore koji su relevantni za predviđanje nagrada, dok je u radu [10] cilj bio predviđanje profita filma, što uključuje dodatne ekonomske faktore poput budžeta i prihoda. Uprkos tome što oba rada koriste iste algoritme mašinskog učenja, rezultati se razlikuju. U radu [10], Random Forest je postigao najbolje rezultate u predikciji ekonomske uspešnosti filma, dok su se SVM i MLP pokazali manje efikasnim, pri čemu je neuronska mreža bila najmanje uspešna. U ovom radu, iako je Random Forest ponovo dominirao u tačnosti predikcija, neuronska mreža je nadmašila SVM model, što ukazuje na različite performanse algoritama u zavisnosti od prirode problema i podataka. Zbog različitih ciljeva istraživanja i razlika u korišćenim skupovima podataka, poređenje ova dva rada mora biti oprezno. Dok je u oba slučaja Random Forest bio najuspešniji model, različite performanse SVM-a i MLP-a sugerišu da efikasnost algoritma zavisi od specifičnosti problema koji se rešava.

ZAKLJUČAK

U ovom radu predstavljen je sistem za predikciju dobitnika nagrade Oskar u kategorijama za najbolji film i za glumačko ostvarenje. Motivacija za istraživanje ove teme proističe iz njenog značaja za filmsku industriju, ali i iz potencijalnih koristi koje može doneti precizno predviđanje pobjednika Oscara. Pouzdane prognoze ishoda ove prestižne nagrade mogu značajno uticati na filmsku industriju, unapređujući strategije u produkciji, marketingu i distribuciji, dok istovremeno ispunjavaju očekivanja publike i filmskih stvaralaca.

U radu je sprovedena eksplorativna analiza oba skupa podataka, što je omogućilo sticanje uvida u tendencije i zavisnosti između određenih karakteristika i osvajanja nagrade Oskar. Jedan od ključnih zaključaka jeste da prisustvo nominacija i osvajanja drugih prestižnih filmskih nagrada, poput BAFTA, SAGA, DGA, PGA i Zlatnog globusa, značajno povećava šanse za osvajanjem Oscara, što ih čini jednim od najvažnijih faktora. Takođe, primećena je snažna povezanost između nominacije za najboljeg režisera i osvajanja Oscara za najbolji film. Utvrđeno je i da filmovi iz žanra drame, kao i glumci koji u njima glume, imaju veću verovatnoću da osvoje Oskara. Međutim, prethodne nominacije ili osvajanja Oscara za glumu nisu pokazale značajan uticaj na mogućnost ponovnog osvajanja ove nagrade.

Nakon eksplorativne analize podataka, vršena su treniranja modela. Modeli koji su korišćeni u obe predikcije su: Logistička Regresija, SVM, Random Forest, Random Forest sa Baggingom, XGBoost, Neuronska mreža. Kako bi se modeli evaluirali korišćena je 10-ostruka unakrsna validacija i to je uzeta u obzir tačnost. Takođe za svaki model je računata preciznost, tačnost, odziv, F1 mera i matrica konfuzije. Najbolje rezultate u predikciji za najbolji film daju Random Forest i Random Forest sa Baggingom, a najlošije daje SVM. Dok za predikciju dobitnika Oscara za glumu najbolje rezultate daju XGBoost i Logistička Regresija, a najgore Neuronska mreža.

Kako bi se poboljšalo rešenje, postoji više stvari koje se mogu razmotriti. Postoje beskonačne mogućnosti varijabli za merenje uspeha filma, bilo da se radi o predviđanjima za Oscara ili finansijskim rezultatima. Potrebno je neprestano raditi na identifikaciji novih faktora, ali i na tome da podaci postanu dostupniji nego što su trenutno. Ovim pristupom može se postići efikasan napredak u istraživanju filmske industrije.

Važno je napomenuti da su oba skupa podataka u ovoj analizi bila ograničena na period od 1961. do 2021. godine. U budućnosti, moglo bi se razmotriti proširenje skupa podataka uključivanjem informacija o nominacijama i pobednicima Oscara koji su se dogodili pre 1961. godine ili posle 2021. godine. Takođe moguće je razmotriti proširenje oba skupa podataka sa nekim novim parametrima. Što se tiče skupa podataka za najbolji film može se dodati opis filma ili neke informacije koje se tiču finansijskog aspekta filma. Što se tiče skupa podataka za predikciju dobitnika Oscara za glumu, pored proširenja skupa podataka, može se razmotriti i njegova dodatna segmentacija. Trenutno, ovaj skup podataka obuhvata informacije za najbolje glumce, najbolje glumice, najbolje sporedne glumce i najbolje sporedne glumice. Razdvajanje skupa podataka na više specifičnih skupova moglo bi omogućiti precizniju analizu i poboljšanje modela za svaku pojedinačnu kategoriju.

Pored proširenja skupova podataka moguće je unaprediti rešenje boljim izborom hiper-parametara, korišćenjem različitih pristupa za njihovu optimizaciju. U ovom radu podešavanje hiper-parametara se vršilo ručno ili korišćenjem GridSearch-a kod nekih algoritama. Mogu se razmotriti sledeći pristupi: Gradient optimizacija, Bayesian optimizacija, Random Search ili neki od evolucionih algoritama. Korišćenjem ovih naprednijih pristupa moguće je postići bolju optimizaciju modela i poboljšanje njegovih performansi.

Takođe, može se razmotriti uvođenje novih modela za obučavanje koji su se koristili u prethodnim istraživanjima. Neki od modela koji su korišćeni u drugim radovima i koji bi mogli doprineti unapređenju performansi uključuju: Naive Bayes, AdaBoost, Adaptive Tree Boosting, Gradient Tree Boosting. Uključivanje ovih modela može proširiti spektar tehnika koje se koriste za analizu i potencijalno poboljšati rezultate.

LITERATURA

- [1] Jeffrey S. Simonoff, Ilana R. Sparrow, Predicting movie grosses: Winners and losers, blockbusters and sleepers, 2000.
- [2] Iain Pardoe, "Just how predictable are the Oscars?", 2005.
- [3] Iain Pardoe, Dean K. Simonton, Applying discrete choice models to predict Academy Award winners, 2008.
- [4] Predicting the 85th Academy Awards: Stephen Barber, Kasey Le, Sean O'Donnell December 13, 2012.
- [5] Prediction of Movies popularity Using Machine Learning Techniques: Muhammad Hassan Latif, Hammad Afzal, National University of Sceinces and technology, Pakistan, August 2016.
- [6] Predicting movie success with machine learning techniques: Kyuhan Lee, Jinsoo Park, Iljoo Kim, Youngseok Choi, 2016.
- [7] Performance evaluation of seven machine learning classification techniques for movie box office success prediction: Nahid Quader, Md. Osman Gani, Dipankar Chaki , December 2017.
- [8] Predicting the "Best Picture" Oscar Award Winner: Paul Ables, 2018.
- [9] The statistics of the Oscars, Florida Southern College, Jacqueline R. Carlton, 2019.
- [10] Revisiting predictions of movie economic success: random Forest applied to profits: Thaís Luiza Donega e Souza, & Marislei Nishijima, Ricardo Pires, Mart 2023.
- [11] [Online].Available;<https://www.kaggle.com/datasets/matevaradi/oscar-prediction-dataset>
- [12] [Online].Available: <https://zenodo.org/records/4244691>
- [13] Logistička Regresija, https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
- [14] SVM, <https://scikit-learn.org/stable/modules/svm.html>
- [15] Random Forest, <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>
- [16] Bagging i Bosdting, <https://www.geeksforgeeks.org/bagging-vs-boosting-in-machine-learning/>

- [17] XGBoost, <https://www.geeksforgeeks.org/xgboost/>
- [18] Neuronska mreža, https://scikit-learn.org/stable/modules/neural_networks_supervised.html
- [19] TF-IDF, https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
- [20] Eksplorativna analiza podataka, <https://medium.com/@azisamcalvin/python-for-data-science-implementing-exploratory-data-analysis-eda-and-k-means-clustering-bcf1d24adc12>

BIOGRAFIJA

Jelena Milijević je rođena 05.08.1998. u Beogradu. Živi u Rumi gde je stekla svoje osnovno i srednje obrazovanje. Školske 2019/20 godine se upisuje na Fakultet tehničkih nauka na studijski program Računarstvo i automatika. Godine 2023. završava osnovne akademske i upisuje master studije. Položila je sve ispite predviđene planom i programom i stekla uslov za odbranu završnog rada.