Предикција оцене парфема

Aутори пројекта: Нина Кузминац Е2 82/2024 и Вукашин Докмановић Е2 79/2024.

## Циљ

Циљ овог пројекта је развој система који ће бити у стању да предвиди оцене парфема на основу различитих улазних параметара попут нота парфема, акорда, бренда, пола за који је парфем намењен, годишње доба за које је погодан и историјски подаци о корисничким оценама.

## Мотивација

У свету парфема где је конкуренција огромна, прецизна предвиђања могу помоћи произвођачима у развоју производа који ће имати већи комерцијални успех, а трговцима у оптимизацији залиха и побољшању продаје. Купци се често ослањају на оцене како би одлучили који парфем да купе, предвиђање оцене нових парфема пре него што изађу на тржиште могло би значајно помоћи купцима приликом одабира парфема.

## Релевантна литература

### 3.1. Price Prediction of Seasonal Items Using Machine Learning and Statistical Methods[[pdf](https://d1wqtxts1xzle7.cloudfront.net/95971407/pdf-libre.pdf?1671383352=&response-content-disposition=inline%3B+filename%3DPrice_Prediction_of_Seasonal_Items_Using.pdf&Expires=1732273859&Signature=HwOL8Lsg0CwVyhSx20tr1gC2vMqTwZZIvETTWSKYORUI9O0DwZijs33kkxSToe9OYGshvnZfdXZvj1EULHVD56Mg5J828KkLIA8VmU7SRtHas6JFVvOujo191SXQOtDrSh0MLBCWwIhfh128o8UGT8kbCrUlhibLeBo6KRLFzAxQBfMKfX1Es4WpxC-WlFeUs~pvmfruxsAz-nF2HnpQ0et1Q0yojTZugPxdp4cRy-lcm2Bsu8LM50OjgGz~HpRZttTPbLGy9YHBqT6nkpzZuVVg2f0L~~8dPjp~X45iVIAKFHBmj0Q3t8mucj8Hg-QIdbxCoz9E5S1729G~XvJbJw__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA)]

**Циљ рада**:употреба и упоређивање статистичких модела и модела машинског учења за предвиђање цена сезонских добара, попут божићних поклона, зарад повећања профита трговцима.

**Опис методологије**:коришћене су методологије *Linear Regression, Ridge Regression, SVМ, Random Forest Regression* и *ARIMA model.* Улаз у систем су подаци из скупа података као што су идентификатор поклона, тип поклона, индустрија којој поклон припада итд. док излаз представљају цене добијене предвиђањем помоћу раније наведених методологија.

**Опис скупа података**:свака торка унутар овог скупа се састоји из следећих обележја: идентификатора поклона, типа поклона, индустрије којој поклон припада, датум када је стигао поклон на стање, цене поклона итд. Скуп података се налази на [линку](https://www.kaggle.com/datasets/shashwatwork/christmas-gift-price-prediction).

**Начин евалуације**:у раду су коришћене метрике *MAE*, *RMSE*, *R-Squared* и *MAPE.*

**Резултати**:*Random forest* модел је постигао најбоље резултате на свим евалуационим метрикама (*MAE – 20.46, RMSE – 31.31, MAPE – 1.63* и *R² - 61%*). *ARIMA* модел је био други по успешности, док последње место заузима *Ridge regression* модел.

**Закључак**:овај рад се показао као релевантан због саме примене методологије *Random forest,* која је постигла најбоље резултате приликом предикције цене у односу на остале примењене методологије. Спрам овог успеха, одлучили смо се да користимо методологију *Random forest* приликом израде пројекта. Такође користићемо *MAE* јер се показала као веродостојна метрика приликом евалуације.

# 3.2. Stock Closing Price Prediction using Machine Learning Techniques[[pdf](https://www.sciencedirect.com/science/article/pii/S1877050920307924)]

**Циљ рада**: предвиђање затварајуће цене акција компаније користећи методологије машинског учења. Рад тежи дапобољша тачност предвиђања анализирајући финансијске податке, уз генерисање нових података за побољшање предикције.

**Методологија**: у раду су коришћене *АNN* и *Random forest* методологије*.* Улазне вредности чине финансијски подаци, као што су отворена, висока, ниска и затварајућа цена, као и променљиве које су накнадно направљене. На основу ових улаза, излаз модела је предвиђена затварајућа цена акција за наредни дан.

**Опис скупа података**: подаци су прикупљени са *Yahoo Finance* и обухватају десетогодишњи период од 2009. до 2019. године за пет компанија: *Nike, Goldman Sachs, Johnson & Johnson, Pfizer* и *JP Morgan Chase & Co*. Подаци укључују дневне вредности цена (отварајућа, висока, ниска, затварајућа) и обим трговања.

**Начин евалуације**:коришћене метрике за евалуацију перформанси модела су *RMSE*, *MAPE* и *MBE.*

**Резултати**:*ANN* је показала боље резултате у предвиђању затварајућих цена акција у поређењу са *Random forest.* Употребивши компаративну анализу метрика попут *RMSE, MAPE, MBE,* *ANN* је имала ниже вредности грешака у поређењу са *Random forest*,

**Закључак**: Овај рад је релевантан за нас јер је упоредио две различите методологије за предикцију затварајуће цене акција, где се показало да је *ANN* дала боље резултате у поређењу са *Random Forest* методом, због тога ћемо користити *АNN* као бољу методологију за наше потребе предикције. Сматрамо да je метрикa *RMSE* корисна и користићемо је приликом евалуације нашег решења.

# 3.3. Movie Rating Prediction using Ensemble Learning Algorithms[[pdf](https://www.semanticscholar.org/paper/Movie-Rating-Prediction-using-Ensemble-Learning-Mhowwala-Sulthana/78c92259268f32fd06ff25b7fba3817642626c2f?p2df)]

**Циљ рада**: циљ овог рада је да се прикупе подаци о филмовима са *IMDb*-а и података са сајтова као што су *YouTube* и *Wikipedia,* а потом упореде перформансе два алгоритма за машинско учење приликом предвиђања оцене филма.

**Методологија**: коришћене су методологије *Random forest* и *XGBoost*. На основу метаподатака о филмовима и података о филмовима са друштвених мрежа, уз помоћ методологија, оцена филма је била предвиђана.

**Опис скупа података**: подаци везани за филмове су добијени помоћу скрејпинга *IMDb* сајта и уз помоћ коришћења *TMDb* и *OMDb API-*ја. Добављени су подаци попут наслова, продуцената, глумаца, оцена итд. Да би се разумело заинтересовање за филм пре премијере, добављали су се подаци преко *Wikimedia REST API*-ја. Преко *YouTube API*-јасу, за трејлер сваког филма, добављени лајкови, дељења и коментари.

**Начин евалуације**:евалуација перформанси модела спроведена је коришћењем десетоструке унакрсне валидације. Као метрике коришћени су *МSE*, *МАЕ* и *RMSE*. Мерења су поновљена за два скупа података: један који укључује само метаподатке филмова и други који укључује и податке са друштвених мрежа.

**Резултати**:*XGBoost* јепоказао боље резулате на оба скупа података од *Random forest* методе и остварио је *RMSE* од чак 0.30, док је *Random forest* остварио *RMSE* од 0.40 .

**Закључак**: сматрамо да је овај рад релевантан за наш пројекат јер користи *XGBoost* који је дао добре резултате приликом предикције.

# 3.4. Social success of perfumes [[pdf](https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0218664&type=printable)]

**Циљ рада**: утврдити какав утицај специфичне комбинације нота имају на успех формула и мириса парфема. Истраживање се такође фокусира на препознавање конкретних нота које могу значајно да унапреде целокупну композицију неког парфема.

**Методологија**:коришћена је формула инспирисана Бајесовом статистиком за комбиновање просечне оцене са бројем рецензија. Такође, примењена је мрежна анализа која обухвата испитивање веза између парфема и нота, као и утврђивање утицаја појединих нота на успешност композиција. За процену значаја акорда и нота за успех парфема, коришћена је Коенова мера величине ефекта.

**Опис скупа података**:у доступним подацима налази се 10,599 парфема, од којих сваки садржи комбинације између 990 различитих нота. За сваки парфем могу се пронаћи информације о просечној оцени коју су корисници дали, као и о укупном броју гласова. Поред тога, за одређене парфеме је наведена и цена.

**Начин евалуације**: коришћене су статистичке анализе, укључујући пермутацијске тестове (10,000 симулација) за утврђивање значајности разлика у величини ефекта. З-резултати су коришћени за одређивање да ли су одређени акорди или ноте презаступљени или недовољно заступљени у односу на случајни модел. Такође, вршено је упоређивање резултата мрежне анализе са стварним бројем рецензија и оценама парфема.

**Резултати**: истакнуто је да презаступљене ноте и акорди нису нужно најуспешнији, односно, њихов утицај на број рецензија и оцену неког парфема је мали, док недовољно заступљене комбинације могу имати велики потенцијал јер имају већи утицај на успех парфема и требало би их даље истраживати. Ноте са највећим потенцијалом за побољшање других композиција укључују мошус, ванилу, цветне ноте и дрвене ноте. Такође, није пронађена јака корелација између цене парфема или старости парфема и његове популарности.

**Закључак**: показатељи да акорди и ноте имају значајнији утицај на успех, оцену и број рецензија парфема од осталих атрибута попут цене или године издавања парфема, су нама релевантни јер знамо на које атрибуте треба да обратимо више пажње приликом добављања података и предикције оцене.

## Скуп података

Скуп података ће бити ручно направљен употребом скрејпинга сајтова попут: [јасмина](https://www.jasmin.rs/), [сефоре,](https://www.sephora.com/) [метрополитена](https://metropoliten.rs/) и [фрагрантике](https://www.fragrantica.com/). Скрејповаће се оквирно 1900 парфема. Атрибути који ће бити добављени за сваки парфем су: назив бренда, назив парфема, акорди, ноте, пол за који је намењен, трајност, јачина трага парфема, оцена и број гласова, мера намењености парфема за одређено годишње доба, мера намењености парфема за одређено доба дана, име и презиме креатора парфема, опис парфема и *URL* фотографије парфема. На основу обележја парфема биће предвиђена оцена парфема, која има вредност у опсегу од један до пет.

## Предложена методологија

Након прикупљања података вршиће се претпроцесирање у виду *One-Hot Encoding-*а категоричких вредности и нота за улаз у *Random Forest* и *XGBoost*, као и прављење *Embedding-*анота за улаз у *ANN*. Улаз у алгоритме ће бити одабрани подаци из претходно формираног скупа, као што су бренд, пол, ноте, акорди, годишње доба, оцена и други, док ће излаз представљати предвиђена оцена парфема на основу његових карактеристика. Употребићемо алгоритме као што су *Random forest*, *XGBoost* и *ANN*. Након тога ће бити извршена евалуација и поређење. Такође, извршићемо експлоративну анализу података (*EDA*), где ћемо обрадити следећа питања:

* *What are the top 10 least prevalent accords?*
* *What are the top 10 most prevalent accords?*
* *What are the top 10 least prevalent notes?*
* *What are the top 10 most prevalent notes?*
* *What note is the most prevalent in the top 10 rated fragrances that have at least 5000 reviews?*
* *What accord is the most prevalent in the top 10 rated fragrances that have at least 5000 reviews?*
* *What note is the most prevalent in the worst 10 rated fragrances that have at least 5000 reviews?*
* *What accord is the most prevalent in the worst 10 rated fragrances that have at least 5000 reviews?*
* *What 3-note combination is the most common amongst the fragrances rated 4.50 and higher, that also have at least 5000 reviews?*
* *Who are the top 5 perfumers based on the rating of their fragrances?*
* *What is the percentage of perfume houses that participated in creating the top 20 rated perfumes that have more than 10,000 reviews?*
* *Which season is the most recommended for fragrances rated 4.50 and above with at least 5000 reviews?*
* *What are the most common accords and notes associated with fragrances that are recommended for winter season?*
* *What are the most common accords and notes associated with fragrances that are recommended for summer season?*
* *What percentage of fragrances rated below 3.00 are recommended for summer season?*
* *What are the most prevalent notes in the top 5 rated fragrances with at least 5000 reviews that are recommended for night time?*

## Метод евалуације

За евалуацију предикције оцене парфема користићемо метрике *МАЕ* и *RMSE*, где ћемо рачунати одступање предвиђене оцене парфема од оригиналне оцене парфема. Коришћењем метрика *MAE* и *RMSE* покрићемо општи учинак модела за већину случајева (*MAE*) и осетљивост на велике грешке (*RMSE*) модела. Овим приступом ћемо добити шири увид у перформансе модела, што ће нам помоћи да боље разумемо његове слабости и могућности за унапређење. Размеру величине тренинг и тест скупа ћемо иницијално проценити на 80:20.