Зборник радова Факултета техничких наука, Нови Сад

UDK: 004.633
DOI: https://doi.org/10.24867/28BE27Zerajic

ПРЕДИКЦИЈА КАШЊЕЊА АВИОНСКИХ ЛЕТОВА КОРИШЋЕЊЕМ АЛГОРИТАМА МАШИНСКОГ УЧЕЊА

PREDICTION OF FLIGHT DELAYS USING MACHINE LEARNING ALGORITHMS

Катарина Жерајић, Јелена Сливка, *Факултет техничких наука, Нови Сад*

Област – РАЧУНАРСТВО И АУТОМАТИКА

Кратак садржај – *У овом раду се истражује проблем кашњења авионских летова. У развијенијим државама где кашњења авионских летова могу да представљају значајан финансијски губитак, постоје институције које се баве праћењем и анализом овог проблема. У циљу одређивања фактора који утичу на кашњење авионских летова, обучени су модели машинског учења. За предикцију су коришћени су подаци о летовима, авионима, аеродромима и временским условима у време лета. Кашњења су подељена у три класе: занемарљиво кашњење (до 15 минута), мало кашњење (између 15 и 60 минута) и велико кашњење (преко 60 минута).*

Кључне речи: *предвиђање кашњења летова, експлоративна анализа података, модел система, KNN, SVM.*

Abstract – *This paper tackles the problem of flight delays. In more developed countries, where flight delays can lead to significant financial loss, institutions are founded to monitor and analyze this problem. This paper analyses the factors that influence flight delays by training machine learning models on data on flights, planes, airports, and weather conditions at the time of flight. Flight delays are divided into three classes: negligible delays (up to 15 minutes), small delays (between 15 and 60 minutes), and long delays (over 60 minutes).*

Keywords: *flight delay prediction, exploratory data analysis, system model, KNN, SVM.*

1. УВОД

У данашње време, кашњење авионских летова је постало честа појава, што може изазвати низ проблема за путнике и за авио-компаније.

За авио-компаније, кашњење летова може утицати на пад репутације уколико нису у могућности да адекватно управљају ситуацијом и пруже алтернативне аранжмане за путнике који су погођени кашњењем лета.

Уколико би имали приступ поузданим информацијама о потенцијалним проблемима који би могли изазвати кашњење летова, авио-компаније би могле да правовремено осмисле алтернативе за своје путнике, што би повећало њихово задовољство и позитивно утицало на углед авио-компаније.

Предикција кашњења летова би била од велике користи и за путнике, омогућавајући им да се припреме за могуће изазове и да испланирају алтернативе у случају да дође до кашњења летова.

С друге стране, ова врста предикције може имати и шире бенефите за индустрију, помажући у бољем управљању капацитетима и ресурсима на аеродромима, као и у побољшању целокупног транспортног система.

Циљ овог рада је да се, применом машинског учења, утврди веза између података о авионским летовима, временским условима и карактеристикама авиона са могућим кашњењем лета. У том циљу су прикупљени јавно доступни подаци о авионима, аеродромима и летовима, као и подаци о временским условима. Да би се утврдили фактори који имају највећи утицај на кашњење авиона, извршена је експлоративна анализа података. На основу резултата експлоративне анализе изабран је скуп обележја за тренирање модела за предикцију.

У наставку рада биће детаљно објашњени различити аспекти решаваног проблема. У поглављу 2 се налази осврт на радове који се баве сличном тематиком. Поглавље 3 садржи опис скупа података, анализе спроведене над добијеним подацима и описује алгоритме коришћене за предикцију кашњења авионских летова. Поглавље 4 садржи дискусију на тему резултата добијених на основу формираних модела. На крају, поглавље 5 закључује овај рад.

2. ПРЕТХОДНА РЈЕШЕЊА

Претрагом радова на тему предикције кашњења авионских летова, пронађено је неколико радова са различитим приступом у решавању овог проблема.

Рад [1] се бави предикцијом одлагања авионских летова на територији САД-а на основу података о авиону, укрцавању путника и терета у комбинацији са подацима о лету. Модели машинског учења који су коришћени за предикцију су: *Decision Tree, Random Forest, Extra Trees, Bagging, Gradient Boosting* и *XGBoost*. За сваки модел су израчунате тачност, одзив и Ф-мера. *XGBoost* класификатор је дао најбоље резултате, праћен *Random Forest* алгоритмом. На основу најбоље остварених резултата одлучено је да се у овом раду користе *XGBoost* класификатор и *Random Forest* алгоритам. Поред тога, користиће се методе евалуације које су коришћене у раду [1].

НАПОМЕНА:
Овај рад проистекао је из мастер рада чији ментор је била др Јелена Сливка, ванр. проф.

Рад [2] је анализирао кашњења летова Делта авио-компаније. Коришћени су подаци о летовима, који укључују термин лета, аеродром поласка, аеродром доласка, пријеме доласка, разлика између постигнутог и жељеног времена доласка, број летова и удаљеност између аеродрома. Аутори рада [2] су експериментисали са следећим моделима машинског учења: *Gradient Boosting Classifier, Decision Tree, Naive Bayes, SVM, Random Forest, Extra Trees, Bagging, logistic regression*. За евалуацију модела су користили прецизност, тачност, одзив и специфичност. Најбољи резултати остварени су употребом *Gradient boosting* класификатора, што иде у прилог одлуци да се и у овом раду користи *XGBoost*.

Рад [3] је истраживао утицај временских услова на кашњење авионских летова. У ту сврху су аутори користили податке о летовима, временским условима и гужвама на терминалу. Експериментисали су са следећим моделима машинског учења: *Decision Tree, neural networks, SVM*. Решење је евалуирано поделом на три подскупа на основу параметара временских услова. Утврђено је да нема велике разлике у резултатима који су постигнути применом горе наведених алгоритама. С обзиром да су резултати за сва три алгоритма слични, одлучено је да *SVM* буде један од алгоритама за предикцију коришћених у овом раду.

3. МЕТОДОЛОГИЈА

У овом поглављу је представљена имплементација система за предикцију кашњења авионских летова.

Као улаз у систем коришћена је комбинација више скупова података. Међу њима су подаци о авионима, аеродромима и летовима, као и подаци о временским условима. Циљна варијабла, односно излаз из система је класа кашњења којој неки лет припада. Скуп података описан је у поглављу 3.1.

Након прикупљања података, извршена је експлоративна анализа како би се одредила адекватна обележја за предикцију кашњења лета. Поступак експлоративне анализе је описан у поглављу 3.2.

Над креираним скупом података су тренирани модели машинског учења приказани у поглављу 3.3.

3.1. Скупови података

Скуп података који је коришћен у овом раду је произведен од стране Завода за статистику саобраћаја (енг. *Bureau of Transportation Statistics*) [4] у Сједињеним Америчким Државама. Овај скуп садржи три различита скупа података о комерцијалним летовима из 2015. године, организованих у CSV датотеке. Први скуп података садржи информације о авио-компанијама, други скуп података садржи информације о аеродромима у Сједињеним Америчким Државама, а трећи и најзначајнији скуп података садржи информације о летовима који су организовани у току 2015. године. Овај сет садржи информације о више од пет милиона летова, а сваки лет је описан 31 атрибутом.

На основу радова анализираних у поглављу 2, изведен је закључак да би старост и тип авиона могли да имају утицај на предикцију кашњења летова. Стога су прикупљени подаци са сајта о авионима [5] који на основу вредности TAIL_NUMBER обележја, враћа податке о карактеристикама авиона. Ови подаци подразумевају информације о типу и старости авиона.

Подаци о временским условима за полазни и долазни аеродром из 2015. године су прикупљени позивањем *API*-ја сајта о временским условима [6].

Након добављања података, следећи корак је чишћење података, што подразумева следеће фазе:

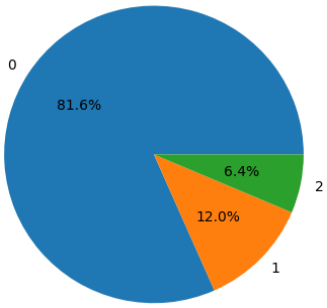
- **Уклањање недостајућих вредности.** Формирани скуп података о летовима има преко пет милиона редова, па је одлучено је да редови са недостајућим вредностима циљног обележја буду уклоњени.
- **Смањивање скупа података.** Одлучено је да се скуп података смањи са пет милиона на око 10 000 узорака због спорости прикупљања података о авионима и временским условима. При смањењу скупа података је коришћена стратегија узимања сваког к-тог узорка. Да би добијени скуп података био права слика већег скупа од којег је добијен, било је потребно сортирати полазни скуп по вредностима циљног обележја. Након сортирања скупа података, да би се добило 10 000 узорака, узет је сваки 500-ти узорак из сортираног скупа података.
- **Категоризација циљног обележја.** Након смањивања скупа података, примјећен је велики распон вредности циљног обележја (од -80 до 2000). Извршена је категоризација циљног обележја у три категорије:
    - 0 категорија - занемарљиво кашњење (мање од 15 минута)
    - 1 категорија - мало кашњење (од 15 до 60 минута)
    - 2 категорија - велико кашњење (више од 60 минута)
- **Уклањање аутлајера.** Након категоризације циљног обележја, примјећено је да постоје обележја чије вредности доста одскачу од осталих вредности. На основу графика вредности обележја је утврђено за која обележја да се уклоне вредности које одскачу (енг. *outliers*).

3.2. Експлоративна анализа и обрада података

Циљ експлоративне анализе података је да се разумеју и истраже подаци како би се открили кључне карактеристике и трендови у скупу података.

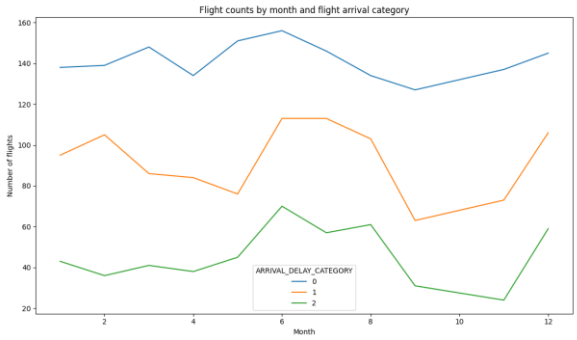
Први корак у фази експлоративне анализе био је утврђивање односа између категорија кашњења летова. На слици 1 се налази процентуална расподела летова према категоријама кашњења.

На слици 1 види се да највећи проценат скупа података чине летови који су имали кашњење мање од 15 минута. Ово значи да је скуп података неизбалансиран, што може негативно утицати на методе класификације.



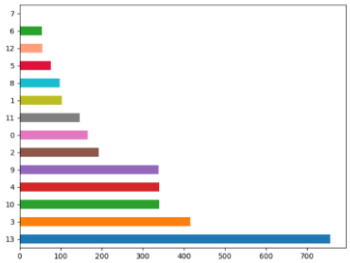
Слика 1. *Расподела кашњења према категоријама*

На слици 2 приказан је график на коме се види зависност између месеца у години и броја кашњења летова у свакој од категорија кашњења. Можемо закључити да је у летњим месецима (6 - 9 месеца) највећи број кашњења по свим категоријама, па је из тог разлога одлучено да обележје MONTH буде укључено у предикцију.



Слика 2. *Расподела кашњења по категорији у односу на месец у години*

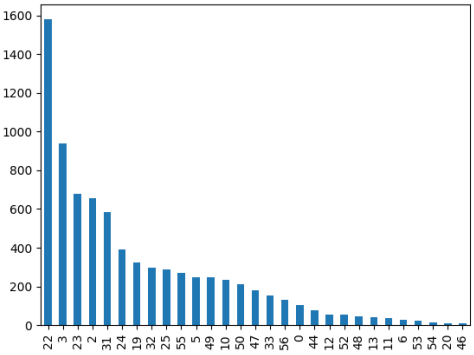
На слици 3 приказан је утицај авио-компаније на кашњење летова. На графику се види да једна компанија има убедљиво највише кашњења, самим тим је одлучено да се обележје AIRLINE узме у обзир приликом предикције кашњења авиона.



Слика 3. *Расподела кашњења према авио-компанији*

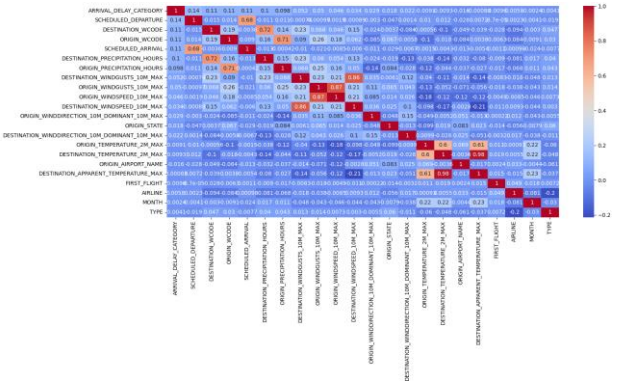
На слици 4 се може видети број летова који касне за сваки тип авиона. Такође је видљиво да се неколико типова авиона издвајају по броју кашњења. Због тога је обележје TYPE узето као једно од обележја за тренирање модела.

Као последњи корак у експлоративној анализи, се посматра зависност обележја, како би се утврдило која обележја се могу искористити за обучавање модела. Зависност обележја се може утврдити анализом матрице корелације почетног скупа обележја.



Слика 4. *Расподела кашњења према типу авиона*

На слици 5 је приказана матрица корелације за обележја која су изабрана за предикцију на основу експлоративне анализе и анализе нумеричких вредности корелација са циљним обележјем.



Слика 5. *Матрица корелације одабраног скупа обележја*

3.3. Коришћени модели машинског учења

Ово поглавље дискутује неколико модела машинског учења који су коришћени за предикцију кашњења у авионском саобраћају:

- К-најближих комшија (енг. *K-nearest neighbour - KNN*)
- Метод случајне шуме (енг. *Random forest - RF*)
- *XGBoost* класификатор (*XGB*)
- Машине потпорних вектора (енг. *Support Vector Machine - SVM*)

KNN (K-Nearest Neighbors) модел се користи за класификацију и регресију. Основна идеја модела је да, ако два објекта имају сличне карактеристике, они припадају истој категорији.

Random Forest је модел машинског учења који се састоји од више стабала одлучивања која се генеришу случајним избором узорка података и случајним избором карактеристика. Свако стабло даје своје предвиђање, а коначно предвиђање се добија агрегацијом предвиђања појединачних стабала одлуке.

XGBoost класификатор је базиран на идеји градијентног *boosting*-а, који је поступак тренирања више слабијих модела како би се постигла боља тачност предвиђања.

SVM има за основну идеју проналажење хипер-равни која најбоље раздваја различите класе података. Хипер-раван се бира тако да буде максимално удаљена од најближих тачака различитих класа.

4. РЕЗУЛТАТИ И ДИСКУСИЈА

Први корак у тренирању сваког од модела је подјела података на тренинг, валидациони и тест скуп. Скупови су подељени у односу 80% тренинг скуп, 10% валидациони скуп и 10% тест скуп.

Након оптимизације параметара, која је вршена на валидационом скупу података, извршено је обучавање модела помоћу тренинг скупа података. За евалуацију су коришћене следеће метрике: Ф1-мјера, одзив, тачност и прецизност. У табели 1 приказани су резултати предикције кашњења авионских летова над тест скупом података.

Табела 1 *- Резултати предикције кашњења авионских летова*

Модел	Тачност	Прецизност	Одзив	Ф1 - мјера
XGB	**80%**	65%	**80%**	**72%**
KNN	**80%**	**74%**	**80%**	**72%**
RF	**80%**	65%	**80%**	**72%**
SVM	**80%**	65%	**80%**	**72%**

Као што је видљиво у табели 1, сваки од модела који је коришћен за предикцију дао је сличне резултате - Ф1-мјера износи 72%, док тачност износи 80% за све моделе.

У раду [1] максимална вредност Ф-мере која је постигнута износи 58%, што је горе од резултата добијених у овом раду. У раду [2] Ф-мера није коришћена као метрика евалуације. Умјесто тога, коришћена је *AUC* (*Area Under the Curve*) метрика, где је добијена вредност од 72%.

Ако упоредимо почетни скуп података коришћен у научним радовима [1] и [2] са скупом података коришћеним у овом раду, можемо примјетити да они садрже мање обележја, што је главни разлог слабијих резултата.

У научном раду [3] коришћена је *CSI* (*Critical Success Index*) метрика и постигнута је вредност од 84%.

5. ЗАКЉУЧАК

Циљ овог рада била је анализа кашњења авионских летова. Одрађена је експлоративна анализа да би се утврдило која обележја имају највећи утицај на циљно обележје (класа кашњења). На основу резултата експлоративне анализе изабран је скуп обележја за тренирање модела за предикцију. Резултати који су постигнути приказани су у поглављу 4. Као што је видљиво у табели 1, резултати су слични за све моделе.

Иако су остварени резултати доста добри у поређењу са научним радовима [1], [2] и [3], остављен је простор за унапређење рада. Како 80% скупа података чине летови са занемарљивим кашњењем, потенцијално унапређење би било балансирање скупа података применом техника за *under-* и *over-sampling*.

Такође, контрола ваздушног саобраћаја (*ATC - Air traffic control*) може имати утицаја на кашњење авионских летова, па би значајно унапређење било укључити и информацију о стању ваздушног простора у тренирање модела за предикцију. Осим тога, за предикцију кашњења лета авиона би се могла искористити и информација о претходном лету на који се дати лет конектује, јер постоји могућност ланчане реакције, уколико је претходни лет каснио.

6. ЛИТЕРАТУРА

[1] Mustafa Kurt (2019), MEF university, Flight delay prediction https://openaccess.mef.edu.tr/xmlui/bitstream/handle/20.500.11779/1217/MustafaKurt.pdf

[2] Enwew Chibuike Kenneth (2019), National College of Ireland, A Machine Learning approach predicting flight arrival delay reduction for Delta Airlines https://norma.ncirl.ie/4305/1/chibuikekennethenwere.pdf

[3] Deepak Kulkarni, Yao Wang and Banavar Sridhar (2013), Ames Research Center, Data mining for understanding and improving decision-making affecting ground delay programs https://permanent.access.gpo.gov/gpo52894/20140008891.pdf

[4] Keggle, 2015 Flight Delays and Cancellations, https://www.kaggle.com/datasets/usdot/flight-delays

[5] FlightAware, Flight Tracker/Flight Status https://flightaware.com

[6] Open-Meteo, Free Open-Source Weather API https://open-meteo.com

Кратка биографија:

Катарина Жерајић рођена је 20.10.1999. у Невесињу, БиХ. Године 2022 завршила је основне студије на Факултету техничких наука на коме брани и мастер рад 2024 године из области Електротехнике и рачунарства - Рачунарство и аутоматика.

Школске 2022/23 године је радила као сарадник у настави на Факултету техничких наука. Године 2023 започиње рад као предавач на "ФТН информатици", и као "јава девелопер" у фирми *CAKE*.

контакт: katarinazer6@gmail.com