**Данило Дамјановић (ИН27/2021)**

Предикција перформанса играча у њиховој првој сезони у НБА лиги на основу Драфт Комбајн мерења.  
(ЛМН021)

Извештај за практично истраживање

# Увод

## Предмет истраживања

Ово истраживање се бави развојем и применом различитих класичних модела машинског учења за предвиђање прилагођене ефикасности по минуту (Adjusted EPM) и укупне ефикасности (EFF) кошаркаша. Модели машинског учења као што су линеарна регресија, модел случајних шума и XGBoost примењени су на податке са „Драфт комбајн”-а и статистике новајлија у циљу идентификације најбољег модела за предвиђање перформанси играча.

## Циљеви истраживања

Циљ овог истраживања је детаљно упоређивање перформанси различитих модела машинског учења у предвиђању прилагођене ефикасности по минуту и укупне ефикасности кошаркаша. Специфични циљеви укључују:

* Анализу и примену различитих модела машинског учења на доступним подацима.
* Упоређивање резултата добијених коришћењем различитих модела у предвиђању EFF и AEPM.
* Идентификацију модела који најбоље предвиђају перформансе играча путем метрика као што су просечна квадратна грешка (MSE) и R2.
* Завршна процена могућности адекватног предвиђања перформанса играча на основу његових физичких карактеристика са мерења на „Драфт комбајн”-у.

## Задаци истраживања

У циљу испуњења наведених циљева, истраживање ће обухватити следеће задатке:

* Прикупљање и припрема података са „Драфт комбајн”'-а и статистике новајлија, укључујући чишћење података и обједињавање у јединствен скуп података.
* Примена техника предобраде података као што су стандардизација нумеричких вредности и енкодирање категоријалних података.
* Развој и тренирање различитих модела машинског учења, укључујући линеарну регресију, рандом форест и XGBoost.
* Евалуација перформанси модела коришћењем метрика као што су MSE и R2, како би се упоредили резултати и идентификовали најбољи модели за предвиђање ефикасности (EFF) и AEPM.

## Очекивани резултати истраживања

Очекује се да ће ово истраживање омогућити дубљи увид у перформансе различитих модела машинског учења у предвиђању EFF и AEPM кошаркаша. Резултати ће показати који модели најбоље предвиђају перформансе играча и који фактори највише утичу на та предвиђања. Ово истраживање ће такође пружити корисне информације о примени различитих техника машинског учења и њиховој ефикасности у анализи атлетских података, што ће допринети бољем разумевању и унапређењу аналитичких метода у овој области.

# Методологија

## Коришћени подаци

У овом истраживању коришћени су подаци са драфт-комбајна и статистика са драфт-комбајна за новајлије. Подаци су прикупљени из следећих извора:

### Подаци са НБА „Драфт комбајн”- а (draft\_combine.csv):

Овај скуп података садржи мерења и резултате играча са драфт-комбајна, укључујући висину, тежину, дужину крила, разне физичке тестове (нпр. скок у даљ, спринт на 3/4 терена, итд.), као и перформансе у различитим атлетским тестовима, као и тестовима снаге.

Пример колона укључује: „player\_name”, „yearDraft”, „heightWithoutShoes”, „weight”, „wingspan”, „standingReach”, „laneAgilityTime”, „threeQuarterSprint”, „benchPress”, „verticalLeap”, итд.

### Подаци из прве сезоне новајлија у НБА лиги (rookie\_stats.xlsx):

Oвај скуп података садржи статистике перформанси играча током њихове прве сезоне у НБА. Садржи податке као што су број одиграних утакмица, број поена, број скокова, број асистенција и друге статистички показатељи.

Пример колона укључује: „Name”, „Year Drafted”, „GP”, MIN”, „PTS”, „REB”, „AST”, „STL”, „BLK”, „TOV”, „FG%”, „3P%”, „FT%”, итд.

## Претходна истраживања других особа над коришћеним подацима

Нису пронађена било каква истраживања са целокупним подацима коришћеним у овом истраживању. Анализа оваквих података је самостална идеја аутора.

## Методе истраживања

Истраживање је спроведено применом различитих статистичких метода, као и метода машинског учења ради анализе и предвиђања ефикасности играча. Главни кораци и методе су следећи:

### Припрема података

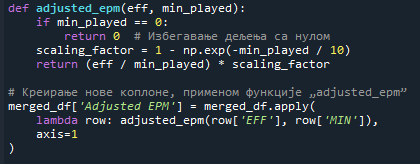
* Избачене су колоне које не доприносе директно анализи или имају превише недостајућих вредности.
* Недостајуће вредности у нумеричким колонама су попуњене просечним вредностима.

### Формулација проблема

* Дефинисани су два главна циља предикције: „EFF” (ефикасност играча) и „Adjusted EPM” (прилагођена ефикасност по минуту).
* Ефикасност играча је позната метрика у свету кошарке. Спада у основне кошаркашке метрике и често се користи за квантификацију перформанса играча.
* Прилагођена ефикасност по минуту („Adjusted EPM”) је метрика коју је извео аутор за потребе овог истраживања. Циљ ове метрике је да се пружи прецизнија процена учинка играча, узимајући у обзир број одиграних минута. На тај начин се избегавају предрасуде у анализи играча који су одиграли мањи број минута у поређењу са онима који су на паркету провели више минута. Рачуна се као производ фактора скалиранја и количника ефикасности и одиграних минута. Фактор скалирања је:



Део кода који представља његово израчунавање:



### Машинско учење:

### Линеарна регресија: Коришћена је за иницијалну базну линију предикције.

### „RandomForestRegressor”: Примењен је са идејом побољшања тачности предвиђања коришћењем ансамбл методе.

### „XGBRegressor”: Примењен је градијент бустинг модел у циљу побољшања тачности предвиђања сложенијим ансамбл моделом.

### Евалуација модела

* Модели су тренирани и тестирани на одвојеним скуповима података.
* За евалуацију модела коришћене су метрике: „Mean Squared Error (MSE) ” и „R^2 Score”.
* Перформансе модела су анализиране кроз дијаграме распршења стварних и предвиђених вредности.

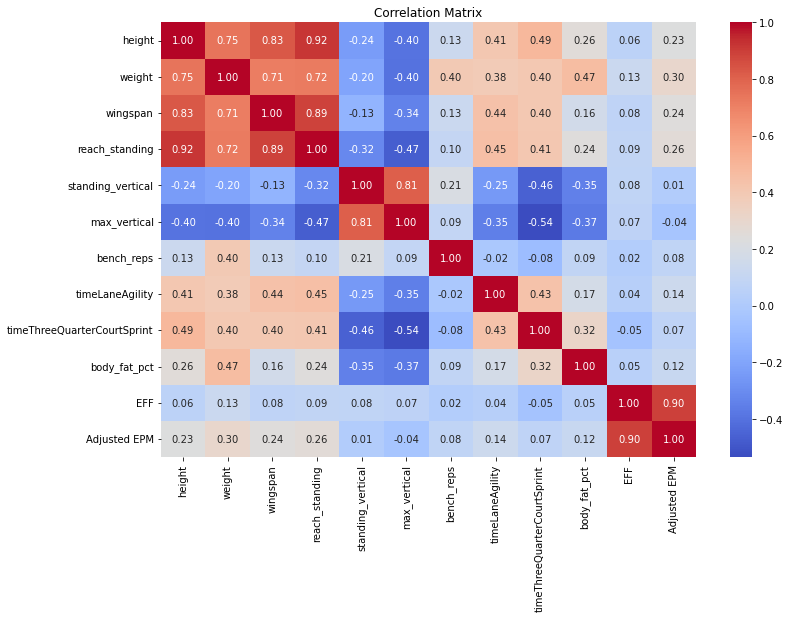
### Анализа значајних карактеристика

* Извршена анализа значајних карактеристика у циљу идентификовација најважнијих карактеристика које доприносе предикцији перформанса играча.

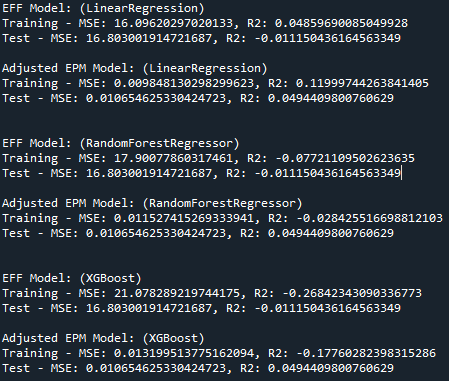
# Резултати

### Приказ резултата

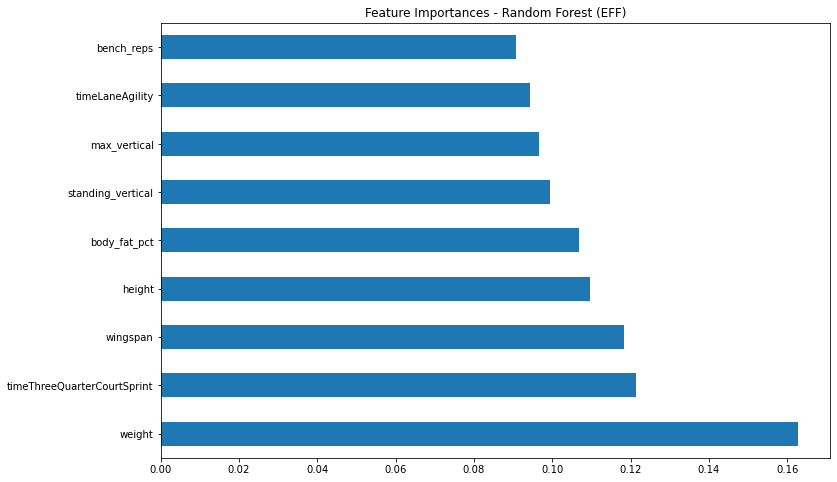
Матрица корелације преосталих колoна након избацивања непотребних изгледа овако:

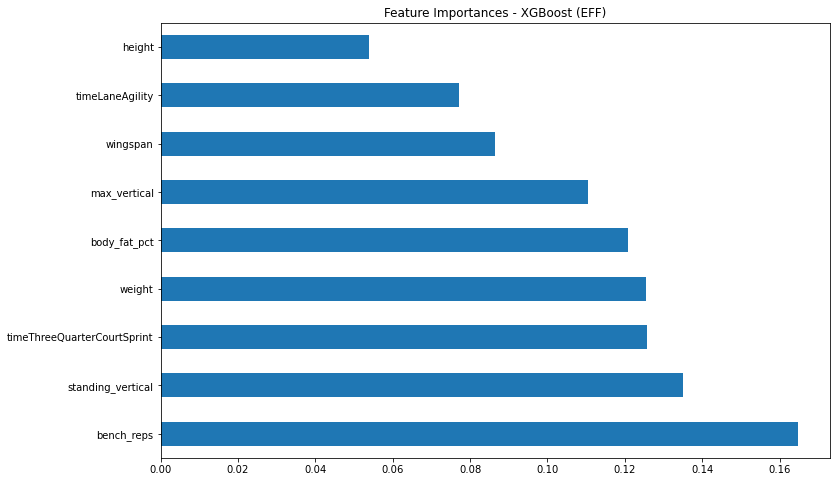


Резултати тест статистика изгледају овако



Анализа значајних карактеристика је следећа:





## Тумачење резултата

Матрица корелације показује да су у корелацији већој од 0.9 само висина и максималан дохват. Из тог разлога је избачена колона максималан дохват. Анализа значајних карактеристика показује да су све физичке карактеристике врло мало значајне код предикције перформанса играча. То потврђују и резултати тестирања модела:

### Линеарна регресија

* EFF Model: Овај модел показује ниску вредност R2 за тренинг податке (0.0486) и негативну вредност R2 за тест податке. То указује на то да модел не може добро да предвиђа ефикасност играча.
* **Adjusted EPM Model:** Модел показује бољи R2 за тренинг податке (0.1200) и тест податке (0.0494), што указује на бољу, али и даље релативно слабу предиктивну способност.

3.2.2 Случајне шуме

* EFF Model: Овај модел има још лошије перформансе са негативним R2 вредностима (-0.0772 за тренинг податке и -0.0111 за тест податке). То значи да модел значајно оверфитује и не може добро да се искористи за овај проблем.
* Adjusted EPM Model: Перформансе овог модела су лоше, са негативним R2 податке за тренинг податке (-0.0284) и слабим позитивним R2  за тест податке (0.0494).

3.2.3 XGBoost:

* EFF Model: Перформансе су још лошије са негативним R2 (-0.2684 за тренинг податке и -0.0111 за тест податке). Овај модел се чини превише сложеним за овај задатак.
* Adjusted EPM Model: Слично као Рандом Форест, овај модел има негативне вредности R2 за тренинг податке (-0.1776) и нешто бољу, али и даље ниску R2 вредност за тест податке (0.0494).

MSE вредност је приближна за све моделе (око 16.8 за EFF и 0.01 за АEPM ) на тестном скупу података. Код тренинг скупа линерна регресија има најнижу вредност MSE што говори да она се она најбоље прилагођава подацима.

# Закључак

Предвиђање прилагођене ефикасности по минуту (Adjusted EPM) даје нешто боље резултате од предвиђања обичне ефикасности на свим имплементираним моделима. Оба предвиђања нису задовољавајућа и сугеришу да постоји минимална побезаност између физичких карактеристика и перформанса играча, те да овај вид предвиђања није адекватан. Перформанс играча и играчки квалитет зависи од мноштва фактора који не обухватају само његове физичке карактеристика са мерења. Предвиђање перформанса на основу физичких карактеристика је веома незахвално и не даје добре резултате. Тестови на свим имплементираним моделима то и покаѕују. Закључак је да није захвално предвиђати перформансе кошаркаша на основу његових физичких карактеристика, макар не у НБА лиги.

## Анализа испуњења циљева истраживања

Циљеви овог истраживања су били:

4.1.1 Aнализa и примена различитих модела машинског учења на податацима са „Драфт комбајн”-а и статистиком новајлија у њиховох првој сезони у НБА лиги.

Будући да су модели успешно имплементирани и примењени, као и да је извршена адекватна анализа, закључак је да је овај циљ испуњем.

4.1.2. Упоређивање резултата добијених коришћењем различитих модела у предвиђању EFF и AEPM и идентификација модела који најбоље предвиђају перформанс играча на основу метрика као што су просечна квадратна грешка и R2.

Извршено је упоређивање резултата добијених коришћењем различитих модела и закључак је да модел линеране регресије има нешто боље перформансе од осталих модела због бољих вредности на тестовима. Овај циљ је, такође, испуњен.

### 4.1.3. Завршна процена могућности адекватног предвиђања перформанса играча на основу његових физичких карактеристика са мерења на „Драфт комбајн”-у.

Овај циљ је, такође испуњен (погледати почетак поглавља Заључак).

## Анализа остварења очекиваних резултата истраживања

Ово истраживање је омогућило дубљи увид у перформансе различитих модела машинског учења у предвиђању ефикасности и AEPM кошаркаша у првој сезони у НБА лиги. Резултати су показали који модел најбоље предвиђа перформансе играча, али су показали и да ни једно предвишање није довољно добро за унапређење и боље разумевање аналитичких метода у овој области. Очекивани резултати су делимично остварени.

## Могућности за примену истраживања у пракси

Ово истраживање није дало корисне резултате и као такво нема већу могућност примене

у пракси.

## Идеје за побољшање и разраду истраживања

Увоћење нових и компликованијих модела не делује као адекватно решење за побољшање овог истраживања. Што је компликованији модел био тестиран то су добијени резултати бивали лошији. Мало је вероватно да ће било какво побољшање модела битно утицати на резултате. Друга опција је промена података. У колико би се тако нешто радило, дошло би до промене целокупног концепта и идеје овог истраживања јер су у њега укључени сви подаци са „Драфт комбајн”-а. Закључак је да не постоје адекватне идеје за побољшаење и разраду истраживања. У колико би тако нешто постојало, сигурно је да би се нашло у саставу овог истраживања.

# Литература

1. Литература са предавања и вежби на предмету МИТНОП
2. Hagan MT, Demuth HB, Beale MH, De Jesús O. Neural Network Design. 2nd edition. eBook. Internet: https://hagan.okstate.edu/nnd.html
3. Keras. Keras: Deep learning for humans. Internet: <https://keras.io/>
4. Keras. Getting started with Keras. Internet: https://keras.io/getting\_started/
5. Keras. Keras 3 API documentation. Internet: https://keras.io/api/ keras-team. keras. Internet: <https://github.com/keras-team/keras/>
6. Matplotlib. Matplotlib: Visualization with Python. Internet: https://matplotlib.org/
7. Matplotlib. Using Matplotlib — Matplotlib 3.8.3 documentation. Internet: <https://matplotlib.org/stable/users/index>
8. Rougier NP. Scientific Visualization: Python + Matplotlib. 2021. Internet: <https://www.labri.fr/perso/nrougier/scientificvisualization.html>
9. pandas. pandas - Python data analysis library. Internet: <https://pandas.pydata.org/>
10. pandas. pandas documentation — pandas 2.2.1 documentation. Internet: <https://pandas.pydata.org/docs/>
11. NumPy. NumPy. Internet: <https://numpy.org/>
12. NumPy. NumPy documentation — NumPy v1.26 Manual. Internet: <https://numpy.org/doc/stable/index.html>
13. Scikit-learn. https://scikit-learn.org/