Анастасија Терзић (ИН44/2021)

Хелена Ђорђевић (ИН42/2021)

Марина Кхан (ИН49/2021)

Предикција телесне тежине на основу етничких и демографских фактора  
(Здравље и телесна тежина)

Извештај за практично истраживање

# Увод

Подаци постали су тема од посебног интереса због њиховог високог, али често и

скривеног потенцијала. Њихов значај доказан је у готово свим аспектима живота, јер

они дају могућност проучавања прошлости и предвиђања будућности. Ово је тема о

којој се мало говори, а анализа података може бити значајнија него што просечан човек

може замислити. Покушаћемо то да вам дочарамо кроз анализу података из конкретно

једне сфере живота- Здравље и његов утицај на квалитет живота. Овај семинарски рад

намењен је истраживању и примењивању стечених знања на курсу Методе и технике

науке о подацима на Факултету техничких наука у Новом Саду. На истраживању радиле

су студенткиње III године смера Информациони инжењеринг. Истраживање вршено је

у програмском језику Python(Jupiter), који је са својим библиотекама феноменалан алат

за анализу и обраду података. Подаци над којима је анализа вршена у даљем контексту

интерпретираћемо као “dataset”. Пре почетка анализе едуковале смо се на тему свакодневних

навика у различитим демографским окружењима (конкретно навике са кључних „Location“ из

обрађиваног dataset-а), и сходно томе у даљем контексту биће спомињате свакодневне навике

као карактеристика демографских фактора.

## Предмет истраживања

Истраживање базирано је на локацији, полу, годинама,

нивоу прихода, нивоу самопуздања, статусу телесне тежине као кључним ставкама из

dataset-а. Ове ставке даће нам до знања, на основу анализе истих, битност и њихов

утицај на квалитет живота.

## Циљеви истраживања

Развити предиктивни модел који ефикасно предвиђа параметаре који се тичу здравља (телесна

тежина, самопоуздање) је примарни циљ истраживања, до ког долазимо применом стечених

знања на курсу. Поред формалног циља који је неопходан за израду овог рада, наш етички циљ

биће да Вам дочарамо значај и подстакнемо Вас да се више бавите физичком активношћу

иводите рачуна о животном стилу као начину живота.

## Задаци истраживања

Прикупљање и анализа података из dataset-а представља неизоставни почетак сваког

истраживања. Затим следи визуализација података као важан фактор приказа односа

података. Неизоставан део пројекта представља примена метода мултиваријабилне

регресије за предвиђање циљних параметара где ћемо применити најкомплексније

градиво и тиме јасно ставити до знања циљеве истраживања. Анализа корелација између

различитих фактора следи након, ради прецизније предикције. Затим применићемо и стечено

знање из области неуронских мрежа и покушати истренирати исту тако да нам даје поуздану

предикцију. Евалуација и оптимизација модела предвиђања заједно са дискусијом чиниће

закључак.

## Очекивани резултати истраживања

Очекујемо да ће резултати истраживања омогућити боље разумевање и значај свакодневних

навика као кључном фактору за срећан, здрав и испуњен живот. Такође, истраживање

ће пружити препоруке за унапређење животних навика и здравља популације. Препоруке

попут усвајања животних навика које су карактеристичне за одређену демографску групу.

Очекујемо да ћемо на основу стечених знања без проблема моћи да направимо предиктивни

модел који нам поуздано предвиђа и показује утицај између података примењивајући стечено

знање.

# Методологија

## Коришћени подаци

Подаци коришћени у овом истраживању су из Behavioral Risk Factor Surveillance System

([https://data.cdc.gov/Behavioral-Risk-Factors/BRFSS-Table-of-Overweight-and-Obesity-BMI- /fqb7-mgjf/about\_data](https://data.cdc.gov/Behavioral-Risk-Factors/BRFSS-Table-of-Overweight-and-Obesity-BMI-%20/fqb7-mgjf/about_data)). Они обухватају информације које утичу на квалитет живота.

## Претходна истраживања других особа над коришћеним подацима

На основу наших истраживања нисмо пронашле нити једно opensource истраживање

рађено на обрађиваном dataset-у. Сходно етичким принципима у илегална истраживања

нисмо залазиле.

## Методе истраживања

* Прикупљање података (одабир валидног dataset-а)
* Обрада података
  + - * Кодирање категоричких вредности
      * Нормализација неопходних података
      * Обрада недостајућих вредности
* Мултиваријабилна регресија са применом L1, L2 и Elastic Net норме.
* Корелациона анализа између атрибута.
* Тренинг, оптимизација и евалуација модела.
* Креирање неуронске мреже
* Тренирање, тестирање и оптимизација неуронске мреже

# Резултати

Пре него што пређемо на резултате скренућемо пажњу на величине обрађиваних скупова.

Наиме, поделили смо dataset на тренинг, тест и валидациони скуп. Тренинг скуп

Обухвата 70% dataset-а (45215 редова), док тестни и валидациони скуп обухватају по 15%

дataset-а (9690 редова).

## Приказ резултата

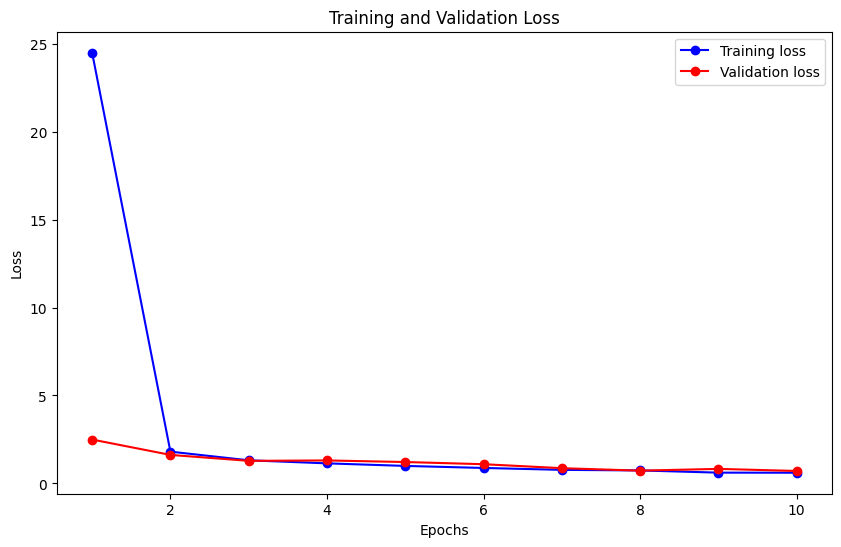
Резултати добијени имплементацијом мултиваријабилне регресије на физичку активност:

* Lasso Regression MSE(L1) ≈ 0.00408
* Ridge Regression RMSE (L2) ≈0.0036
* Elastic Net Regression ≈0.004077

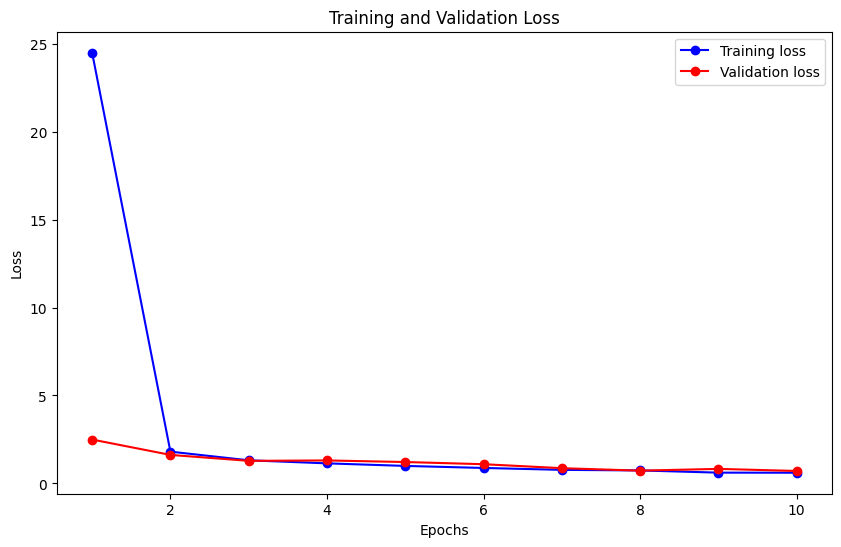
Резултати добијени тренирањем неуронске мреже:

* Test Loss ≈ 0.725
* Test MAE ≈ 0.605

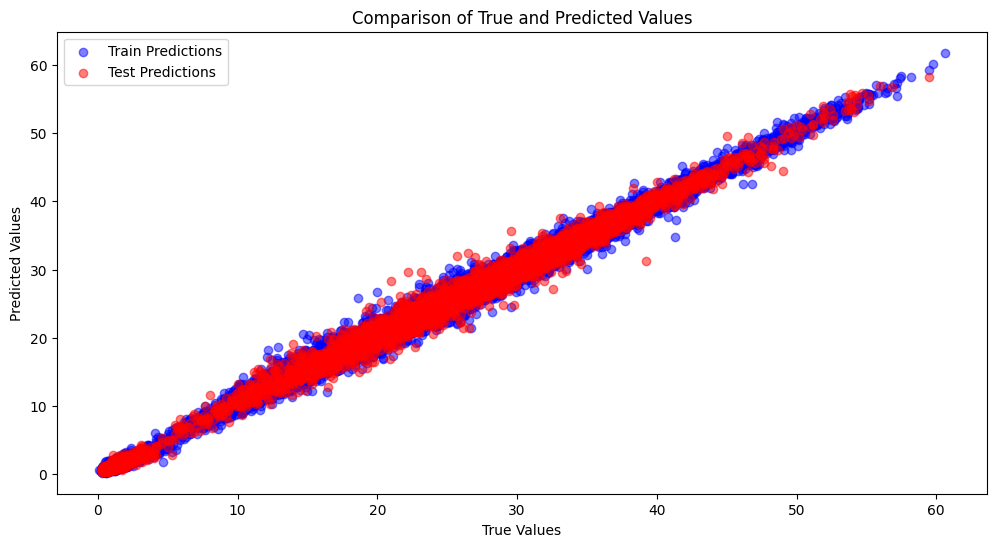
Граф 1:



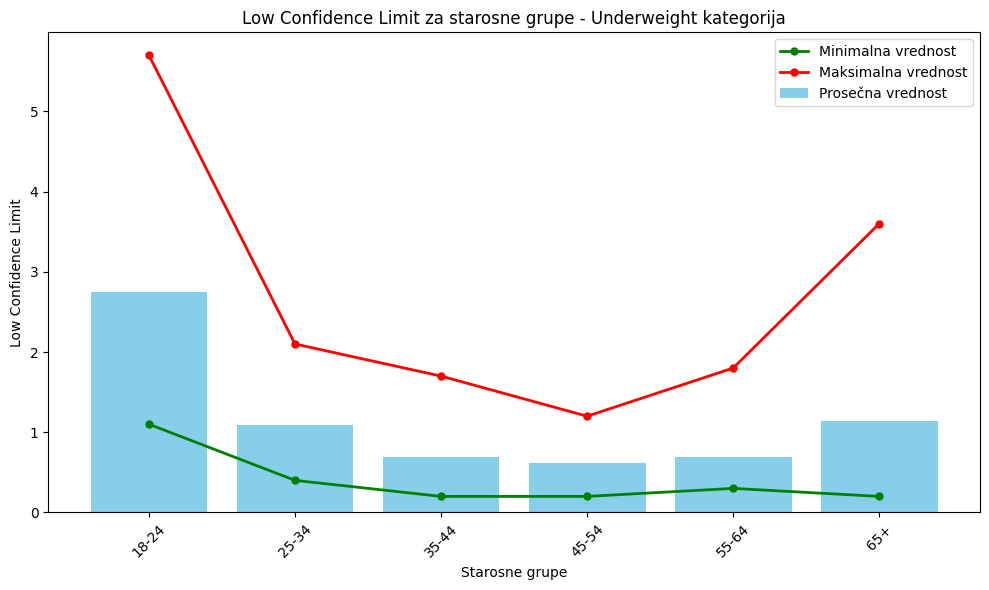
Граф 2:



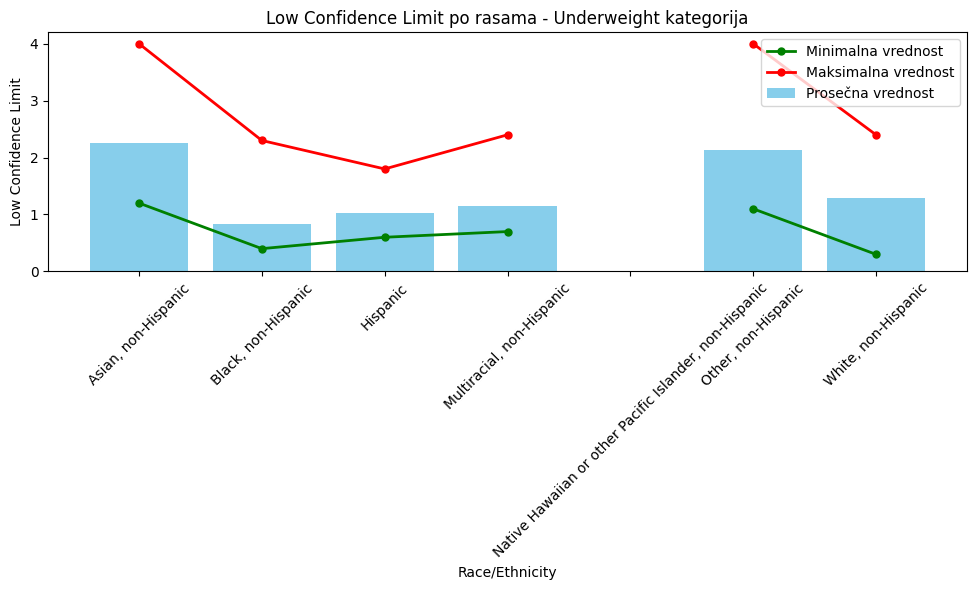
Граф 3:



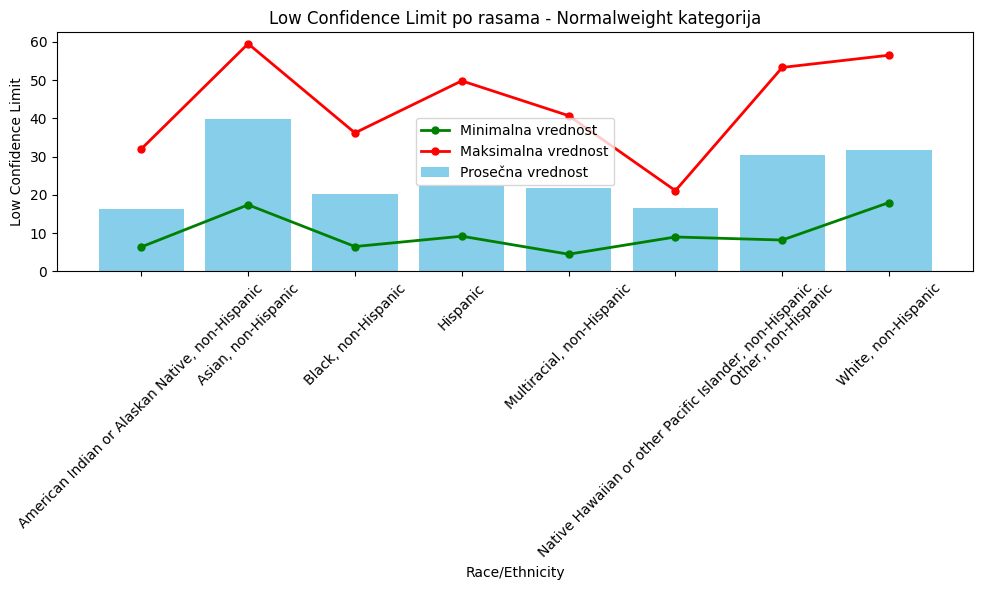
Граф 4:



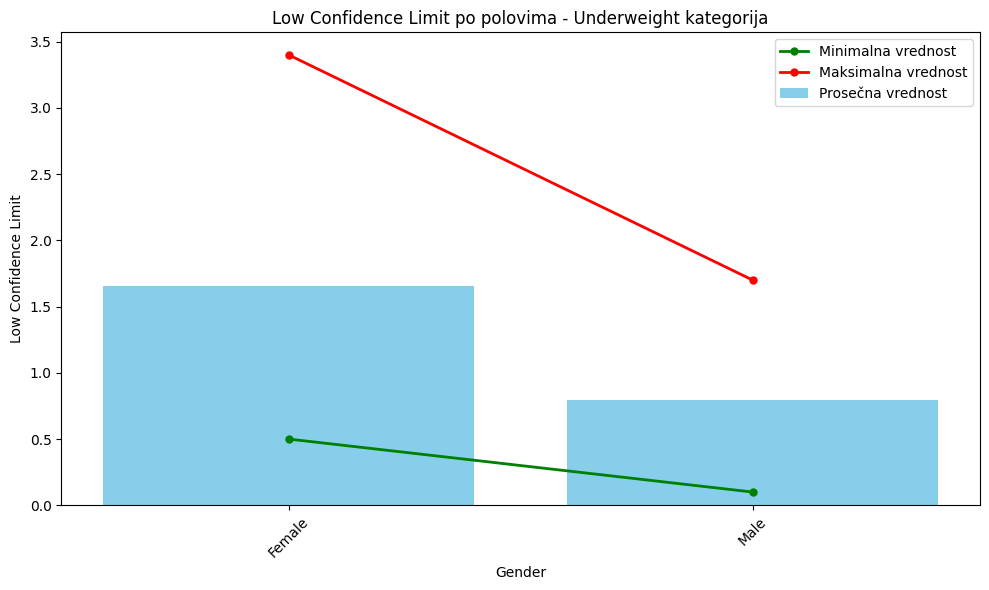
Граф 5:



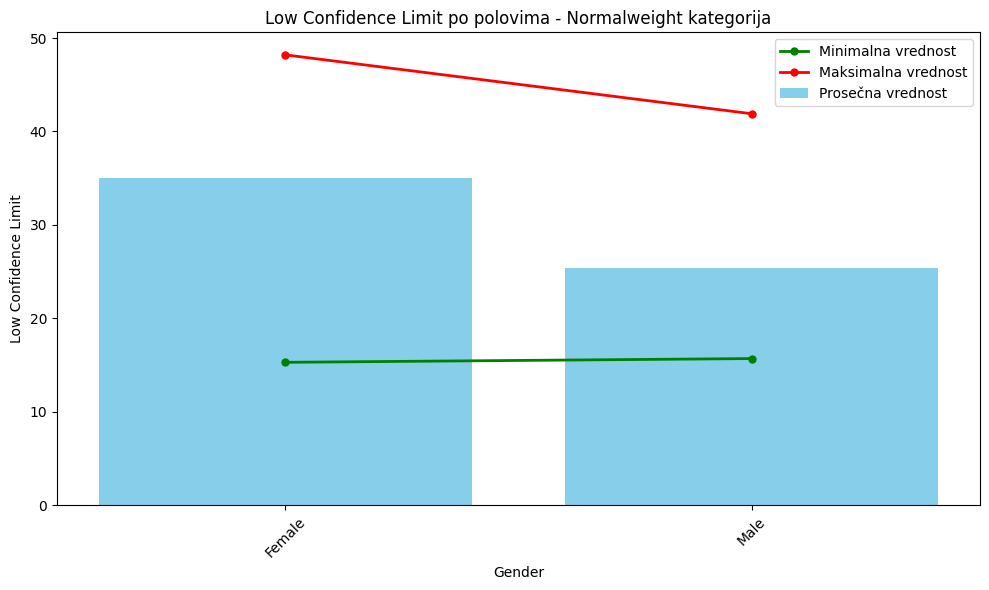
Граф 6:



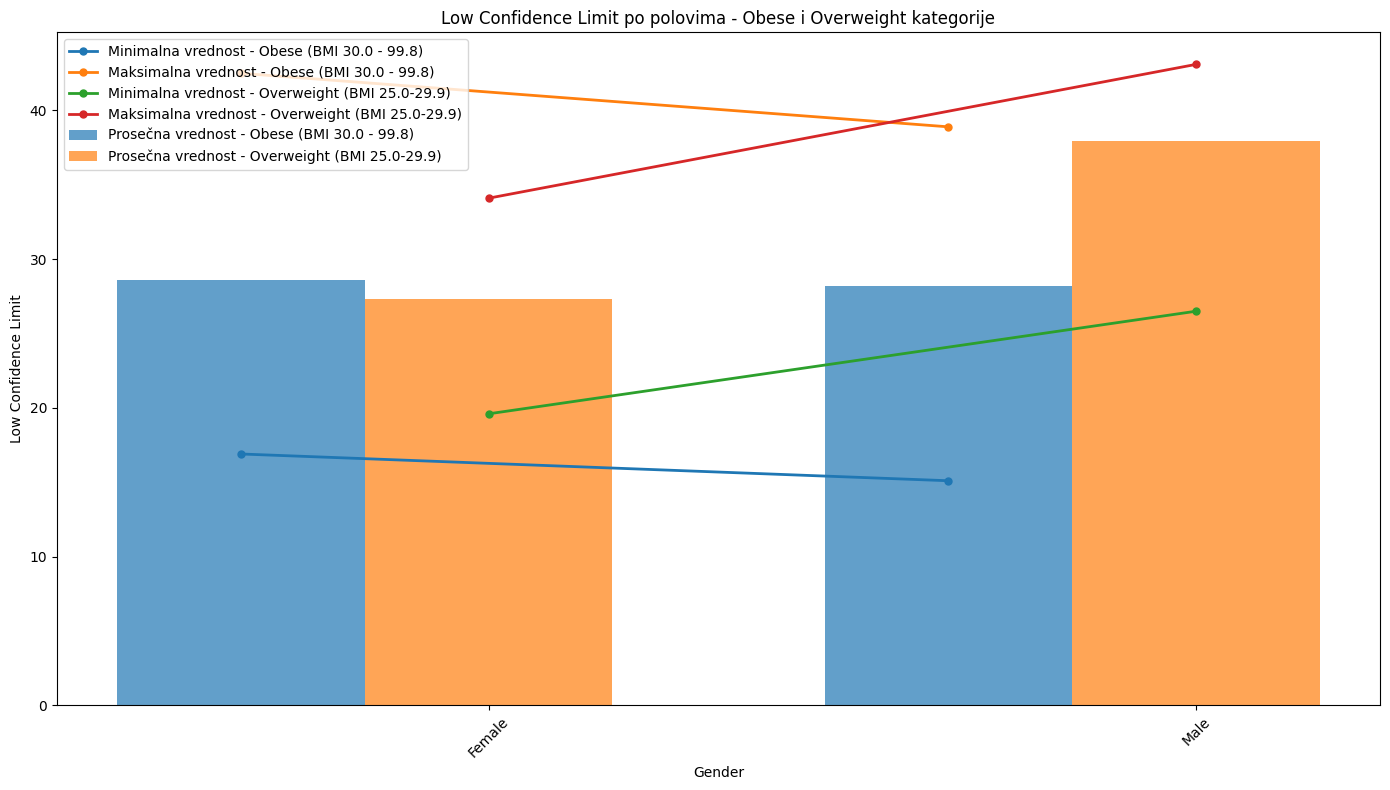
Граф 7:



Граф 8:



Граф 9:



## Тумачење резултата

Вредности резултата добијених мултиваријабилном регресијом можемо рећи да су добре с

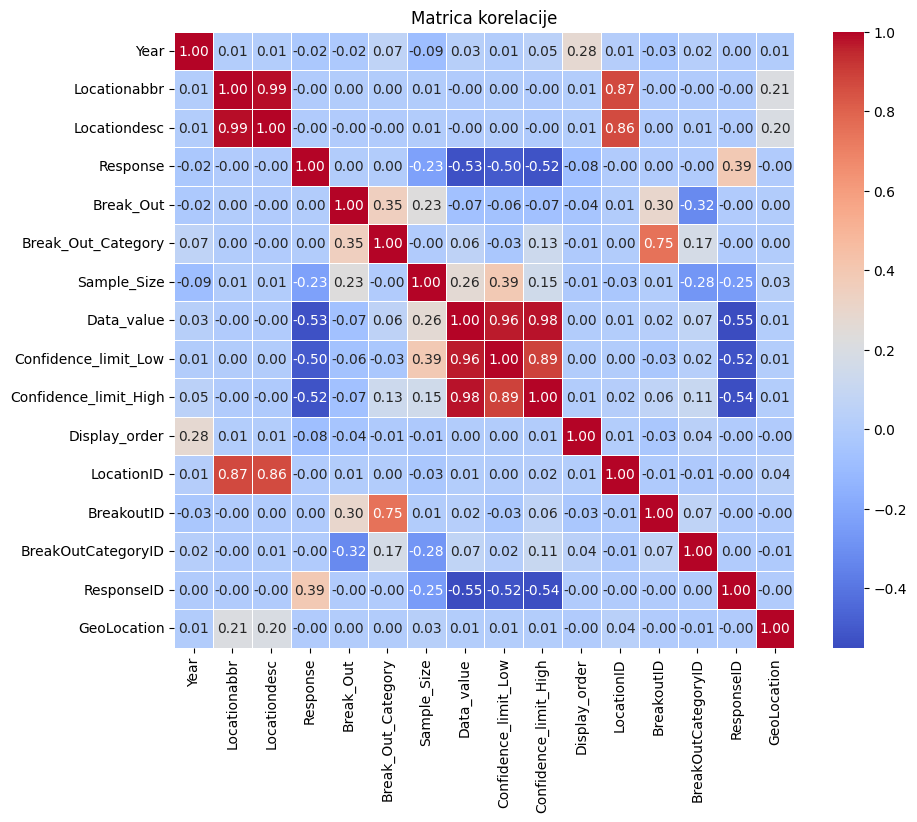
обзиром да смо добиле изузетно мале вредности, што би значило да је модел добар. Тачније

речено предвиђамо са веома великом прецижношћу. За постизање боље прецизности

користили смо матрицу корелације да би видели која обележја су у јакој корелацији и на која

треба да обратимо пажњу и избацимо. Обележја су у јакој корелацији у колико је број што

ближи 1. Добијена матрица корелације изгледа овако:



Вредности резултата добијених тренирањем неуронске мреже такође нам говоре да смо добро

истренирали исту и да можемо са сигурношћу тврдити да су предвиђени резултати врло слични

тачним (реалним) вредностима.

Граф 1 и Граф 2:

Претстављају нам „криве“ где плава крива означава тренинг а црвена валидационе податке.

Поклапање ове две криве сигнализира нам успешност у предвиђању. С обзиром да се ове две

криве крећу истим током и имају сличне вредности то је добар знак за даље коментарисање

добијених резултата (не узимајући у обзир почетак плаве криве која нам претставља изузетке

и специјалне случајеве).

Граф 3:

На овом графу јасно можемо видети поклапање предвиђених и стварних вредности. Што нам

је дало потвду да смо добро одрадиле главни део задатка.

Граф 4:

Овај граф нам представља ниво мањка самопоуздања по старосним групама за људе који су

потхрањени. Оно што је на овом графу јасно приказано јесте заправо та разлика код популације

између 18 и 24 године. Запрепашћујуће је порасла средња вредност, што би значило да већина

те популације има ниско самопуздање у поређењу са осталим добима. Претпостављамо да на

то највећи утицај имају трендови и друштвене мреже које намећу стандарде што са психолошке

стране није ни мало здраво, а поготово не за тако младу популацију.

Граф 5:

На овом графу имамо представу такође мањка самопоуздања за потхрањену популацију по

расама. Одступање можемо видети код азијата и „native Hawaiian“ што нас је поприлично

изненадило с обзиром на познати хавајски концепт 'Pono'

(<https://www.unchainyourbrain.org/2019/11/14/why-clarifying-your-values-boosts-self-esteem-the-hawaiian-concept-of-pono/> ).

Граф 6:

Овај граф односи се на исте ствари као претходни с тим што посматрамо популацију за коју

важи да је нормалне тежине. Овде је средња вредност за мањак самопоуздања код америпчке

популације највећа без обзира што су у питању људи нормалне тежине. Нажалост, сва

истраживања говоре да су за то криве друштвене мреже и наметнути нереални стандарди.

Граф 7, 8 и 9:

Овде можемо видети представу по половима са све три популације. На којој се види значајан

пораст средње вредности (тј. мањак самопоуздања већи) код жена него код мушкараца у прве

две категорије (граф 7 и 8), док за популацију која има прекомерну тежину и гојазност важи

(граф 9) важи нешто другачије. Јасно на графу можемо видети значајно мање самопоуздање

код мушкараца (средња вредност већа) него код жена прекомерне тежине. Код гојазних људи

ситуација је врло слична, тј. неупоредиво је мала разлика између мушкараца и жена.

# Закључак

Анализа података омогућила нам је да детаљно разумемо и тумачимо утицај анализираних

фактора на здравствене навике и статус телесне тежине. Кроз детаљну примену

мултиваријабилне регресије са различитим техникама, показале су се значајне

корелације између како етничких тако и демографских фактора, и параметра телесне тежине.

Ова сазнања нам омогућавају да идентификујемо кључне области на којима је неопходно

радити и предложимо циљане стратегије (поглавље 4.3) усмерене на унапређење здравствених

навика. Ефикасно одређивање фактора ризика у контексту телесне тежине и демографије може

довести до значајних побољшања у здрављу популације, смањењу процента гојазности и

повећању општег квалитета живота популације.

## Анализа испуњења циљева истраживања

Истраживање је испунило циљеве предвиђања параметара и примене стеченог знања

као формалног дела истраживања. Што се тиче етичког циља, искрено се надамо да смо

Вас подстакле да промените лоше животне навике, а уверили Вас да су ваше добре

навике пут ка успеху, срећном и здравом животу. Надамо се да ћете размислити о усвајању

навика које карактеришу демографске регионе који су се најбоље показали у истраживању.

## Анализа остварења очекиваних резултата истраживања

Очекивани резултати добијени применом коришћених стратегија поклапају се са нашим

почетним претпоставкама везаним за животне навике, такође у тумачењу резултата

(поглавље 3) тврдимо са великом прецизношћу да су предвиђени и очекивани резултати

у снажној повезаности.

## Могућности за примену истраживања у пракси

**Едукација и подизање свести:** Ово истраживање може се користити за развој јавно-

здравствених кампања и активности. Активности могу укључивати радионице,

семинаре и интересантне програме који се фокусирају на значај правилне исхране и

редовне физичке активности.

**Партнерства са приватним сектором (компанијама из уско повезане бранше):** Компаније

које се баве производњом хране ипића, као и фитнес индустрија, могу користити ове податке

за развој и промоцију производа који подржавају здравије животне навике. Ово може

укључивати развој здравих прехрамбених производа или промоције које подстичу

физичку активност међу запосленима. Поред производа компаније могу организовати

бројне догађаје који ће за циљ имати представљање ретких, а корисних информација

везаних за побољшање здравља.

**Иновације у технологији**: Технолошке компаније могу искористити резултате истраживања

за развој апликација и алата за праћење и побољшање здравствених навика. Ове апликације

могу понудити персонализоване препоруке на основу података који су се показали

као кључни у анализи.

Поред ових могућности, такође ово истраживање може бити пример коришћења и

примене корисних алата (поглавље 2.3.) који су данас поред ИТ индустрије неизоставни

у области економије, бизниса, као и у напреднијим криминално-истраживачким

анализама.

## Идеје за побољшање и разраду истраживања

С обзиром на експоненцијални раст могућности које пружа вештачка интелигенција,

без сумње очекујемо и оптимизоване алате за рад са подацима које би применили у

будућим истраживањима. Верујемо да би нам ово помогло да будемо ефикаснији и

ефективнији у раду, као и да тестирамо досадашње стечено знање. Побољшању би допринос

донео и новији dataset који би нам помогао да са још већом сигурношћу можемо предвидети

утицај фактора. У будућим истраживањима, препоручујемо даље испитивање утицаја

културолошких фактора и доступности здравствених ресурса. Овакав приступ не само да ће

помоћи унапређењу здравствених исхода, већ ће и осигурати да

се здравствене предикције спроводе на етички одговоран и социјално осетљив начин.

# Литература

1. Материјали са вежби на курсу *Методе и технике науке о подацима,* Факултет техничких наука у Новом Саду
2. Python 3.13.3 documentation <https://docs.python.org/3/>
3. Материјали са предавања на курсу *Методе и технике науке о подацима,* Факултет техничких наука у Новом Саду
4. Codebasics видео снимци <https://www.youtube.com/@codebasics>