Природно-математички факултет Крагујевац

A blue and silver logo

Description automatically generated

Увод у науку о подацима

Семинарски рад

Чланови тима:

Милош Милетић 66/2020

Емилија Јевремовић 47/2020

Професор:

др Бранко Арсић

Садржај

[1 Увод 3](#_Toc178772347)

[2 Представљање проблема 4](#_Toc178772348)

[3 Припрема података 6](#_Toc178772349)

[3.1 Колона *Distance\_from\_Home* 11](#_Toc178772350)

[3.2 Колона *Parental\_Education\_Level* 13](#_Toc178772351)

[3.3 Колона *Teacher\_Quality* 14](#_Toc178772352)

[4 Опис података 16](#_Toc178772353)

[4.1 Описи нумеричких колона 16](#_Toc178772354)

[4.1.1 Нумеричка колона *Hours\_Studied* 16](#_Toc178772355)

[4.1.2 Нумеричка колона *Attendance* 19](#_Toc178772356)

[4.1.3 Нумеричка колона *Sleep\_Hours* 20](#_Toc178772357)

[4.1.4 Нумеричка колона *Previous\_Scores* 22](#_Toc178772358)

[4.1.5 Нумеричка колона *Tutoring\_Sessions* 24](#_Toc178772359)

[4.1.6 Нумеричка колона *Physical\_Activity* 27](#_Toc178772360)

[4.1.7 Нумеричка колона *Exam\_Score* 28](#_Toc178772361)

[4.2 Описи категоријских колона 30](#_Toc178772362)

[4.2.1 Категоријска колона *Parental\_Involvement* 30](#_Toc178772363)

[4.2.2 Категоријска колона *Access\_to\_Resources* 31](#_Toc178772364)

[4.2.3 Категоријска колона *Extracurricular\_Activities* 32](#_Toc178772365)

[4.2.4 Категоријска колона *Motivation\_Level* 32](#_Toc178772366)

[4.2.5 Категоријска колона *Internet\_Access* 33](#_Toc178772367)

[4.2.6 Категоријска колона *Family\_Income* 34](#_Toc178772368)

[4.2.7 Категоријска колона *Teacher\_Quality* 35](#_Toc178772369)

[4.2.8 Категоријска колона *School\_Type* 36](#_Toc178772370)

[4.2.9 Категоријска колона *Peer\_Influence* 37](#_Toc178772371)

[4.2.10 Категоријска колона *Learning\_Disabilities* 38](#_Toc178772372)

[4.2.11 Категоријска колона *Parental\_Education\_Level* 39](#_Toc178772373)

[4.2.12 Категоријска колона *Distance\_from\_Home* 40](#_Toc178772374)

[4.2.13 Категоријска колона *Gender* 41](#_Toc178772375)

[5 Анализа 43](#_Toc178772376)

[5.1 Анализа између предиктора и одговора 43](#_Toc178772377)

[5.1.1 Утицај колоне *Hours\_Studied* 43](#_Toc178772378)

[5.1.2 Утицај колоне *Attendance* 45](#_Toc178772379)

[5.1.3 Утицај колоне *Access\_to\_Resource* 46](#_Toc178772380)

[5.1.4 Утицај колоне *Sleep\_Hours* 47](#_Toc178772381)

[5.1.5 Утицај колоне *Motivaton\_Level* 48](#_Toc178772382)

[5.1.6 Утицај колоне *Internet\_Access* 49](#_Toc178772383)

[5.1.7 Утицај колоне *Tutoring\_Sessions* 50](#_Toc178772384)

[5.2 Мултиваријантна анализа 54](#_Toc178772385)

[5.2.1 Повезаност између *Hours\_Studied* и *Attendance* 54](#_Toc178772386)

[5.2.2 Повезаност између *Hours\_Studied* и *Tutoring\_Categories* 55](#_Toc178772387)

[5.2.3 Повезаност између *Internet\_Access* и *Access\_to\_Resources* 56](#_Toc178772388)

[5.2.4 Повезаност између *Parental\_Education\_Level* и *Family\_Income* 56](#_Toc178772389)

[5.2.5 Повезаност између *Tutoring\_Categories* и *Internet\_Access* 57](#_Toc178772390)

[5.2.6 Повезаност између *Teacher\_Quality, Tutoring\_Categories* и *Gender* 58](#_Toc178772391)

[5.2.7 Повезаност између *Physical\_Activity* и *Gender* 59](#_Toc178772392)

[5.2.8 Повезаност између *School\_Type* и *Distance\_From\_Home* 60](#_Toc178772393)

[6 Креирање модела 62](#_Toc178772394)

[6.1 Линеарна регресија (*Linear Regression*) 66](#_Toc178772395)

[6.2 Стабло одлучивања (*Decision Tree*) 71](#_Toc178772396)

[6.3 *Random Forest* 77](#_Toc178772397)

[7 Закључак 78](#_Toc178772398)

[8 Литература 79](#_Toc178772399)

# Увод

У овом раду анализираћемо скуп података под називом „[*Student Performance Dataset*](https://www.kaggle.com/datasets/lainguyn123/student-performance-factors)“, помоћу кога ћемо имати увид у факторе који утичу на успех студената на испитима. Овај скуп података обухвата информације о навикама у учењу као што су: присуство предавањима, време које студенти проводе учећи, као и демографски подаци студената и њихових породица. Како бисмо боље разумели образовне процесе, као и који фактори највише утичу на постигнута академска достигнућа студената, извршићемо детаљну анализу ових података.

Циљ овог рада је да кроз анализу података откријемо који фактори највише утичу на успех студената. Посебну пажњу усмерићемо на њихове навике у учењу и ниво мотивације, јер се на прву лопту чини да ови аспекти имају директан утицај на резултате на испитима. Надамо се да ће резултати ове анализе помоћи студентима да постигну боље резултате и напредак.

Поред тога, бавићемо се и изградњом предиктивних модела који ће моћи да предвиде академски успех на основу унетих параметара. На тај начин, могу се идентификовати потенцијални проблеми код студената и омогућити да их на време увиде како би се побољшала њихова будућа постигнућа.

# Представљање проблема

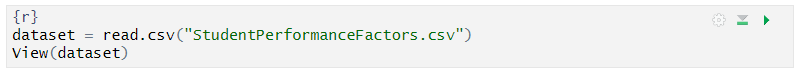
У овом раду користи се ***Student Performance Dataset***, који садржи податке о студентима и различитим факторима који утичу на њихов академски успех. *Dataset,* као што смо већ поменули, обухвата информације о навикама учења, присуству на часовима, ангажовању родитеља и демографским карактеристикама, а циљ је анализирати како ови фактори доприносе постигнућима на испитима (*Exam\_Score*). Проблем којим ћемо се прво бавити је идентификација кључних фактора који највише утичу на успех студената и како се ти подаци могу искористити за предвиђање будућих резултата.

За почетак ћемо учитати потребне библиотеке за рад на пројекту.

A white rectangular object with a black border

Description automatically generated

Затим ћемо импортовати *dataset*



Представићемо *dataset* и објаснити сваку колону. Функцијом *str* проверавамо каква је структура датих колона. У примеру можемо видети да имамо 7 колона које су нумеричког типа (*int*) и 13 колона које су знаковног типа (*chr*).

A screenshot of a computer

Description automatically generated

Као и на слици изнад, функцијом *dim* можемо видети димензије скупа података. Види се да наш скуп података садржи 6607 редова и 20 колона.

A white rectangular object with a black border

Description automatically generated

Опис колона које садржи скуп података *Student Performance Factors*

* ***Hours\_Studied*** – Број сати које је студент провео учећи недељно
* ***Attendance*** – Проценат долазака на часове
* ***Parental\_Involvement*** – Ниво укључености родитеља у образовање студента (Low, Medium, High)
* ***Access\_To\_Resources*** – Доступност образовних ресурса (Low, Medium, High)
* ***Extracurricular\_Activities*** - Учешће у ваннаставним активностима (Yes, No)
* ***Sleep\_Hours*** – Просечан број сати сна по ноћи
* ***Previous\_Scores*** – Оцене са претходних испита
* ***Motivation\_Level*** – Ниво мотивације студента (Low, Medium, High)
* ***Internet\_Access*** – Доступност приступа интернету (Yes, No)
* ***Tutoring\_Sessions*** – Број посећених допунских часова месечно
* ***Family\_Income*** – Ниво породичног прихода (Low, Medium, High)
* ***Teacher\_Quality*** – Квалитет наставника (Low, Medium, High)
* ***School\_Type*** – Врста школе коју похађа (Public, Private)
* ***Peer\_Influence*** – Утицај вршњака на академски успех (Positive, Neutral, Negative)
* ***Physical\_Activity*** – Просечан број сати физичке активности недељно
* ***Learning\_Disabilities*** – Присуство сметњи у учењу (Yes, No)
* ***Parental\_Education\_Level*** – Највиши ниво образовања родитеља (High School, College, Postgraduate)
* ***Distance\_from\_Home*** – Удаљеност од куће до школе (Near, Moderate, Far)
* ***Gender*** – Пол студента (Male, Female)
* ***Exam\_Score*** – Поени са завршног испита

# Припрема података

Сада ћемо да проверимо колико наша база података има *NA* вредности. Након провере морамо утврдити да ли ћемо избацити неку од колона или покушати да попунимо неким приближним вредностима. Функција *summary* нам омогућава да видимо додатне информације о свакој колони као што су минималне и максималне вредности, просечна вредност, медијана као и први и трећи квартал за нумеричке колоне, док за категоријске колоне приказује учесталост сваке категорије.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Такође, функција *summary* може да нам покаже да ли у одређеној колони имамо присутне *NA* вредности. У нашем примеру можемо да видимо да се ни за једну колону не приказују информације о *NA* вредностима.

Након дубље анализе базе података дошли смо до закључка да *NA* вредности нису забележене у свом подразумеваном облику, већ је на тим местима остављено празно поље. Сада ћемо да прикажемо колоне које садрже празна поља као и њихов број.

A screenshot of a computer program

Description automatically generated

Прво смо уз помоћ функције *sapply* пролазили по колонама и за сваку колону рачунали колико има поља са празним вредностима. Онда смо приказали само оне колоне које имају више од 0 празних поља и дошли до закључка да имамо 3 колоне са *NA* вредностима. То су колоне: *Teacher\_Quality, Parental\_Education\_Level* и *Distance\_from\_Home.*

Уместо празних поља у бази података, сада ћемо убацити да пише *NA* у тим пољима како бисмо касније могли да користимо додатне фунцкције које раде са тим вредностима и тиме олакшали рад.

A screenshot of a computer program

Description automatically generated

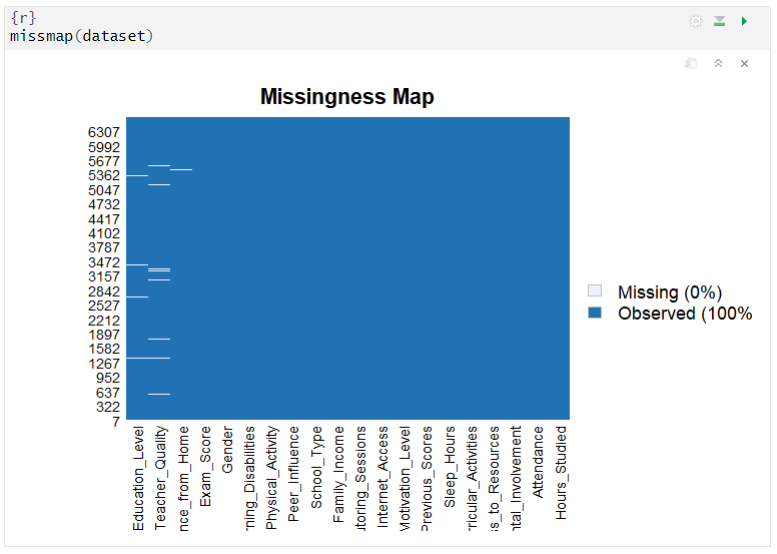
У овом примеру смо у свако празно поље додали *NA* вредност. Такође смо приказали све колоне и број *NA* вредности у свим колонама. Након тога ћемо да процентуално прикажемо колико има недостајућих вредности у колонама и на основу тога донети закључак шта да радимо са њима.

A screenshot of a computer program

Description automatically generated

У примеру видимо да у колонама које имају недостајуће вредности постоји око 1% *NA* вредности за ту колону.

Функција *missmap* у пакету *Amelia* се користи за визуелни приказ *NA* вредности у подацима. Ту можемо преко графика видети како изгледа расподела *NA* вредности по колонама.



Пошто имамо укупно мање од 1% *NA* вредности, *Missing* је 0%.

Сада ћемо да проверимо да ли постоје неки редови у којима се налази више од једне *NA* вредности.

A screenshot of a computer

Description automatically generated

Пошто редови са више недостајућих вредности могу указивати на озбиљнији проблем у подацима, брисање тих редова би била добра опција. Такође број тих редова је веома мали (6/6607) тако да их хоћемо обрисати.

A screen shot of a computer code

Description automatically generated

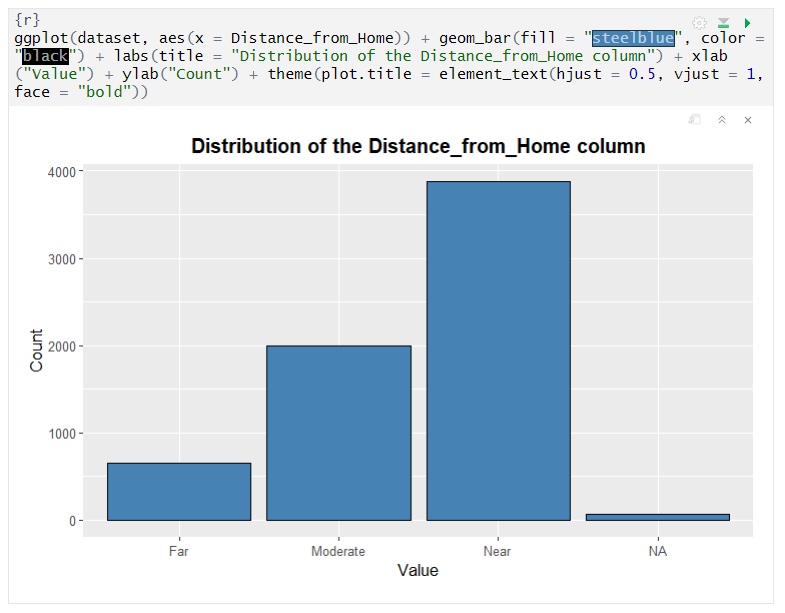
Остале недостајуће вредности планирамо да попунимо са вредношћу која се најчешће појављује у тој колони. То је честа пракса када је број недостајућих података низак и пошто је наш *dataset* мали, не желимо да избацимо податке који могу садржати друге корисне информације.

## Колона *Distance\_from\_Home*

Сада ћемо вредности које су недостајуће да попунимо са вредношћу која се највише појављује у тој колони. Прво ћемо за колону *Distance\_from\_Home* да пронађемо ту вредност и након тога ћемо да је убацимо у *NA* поља. Користићемо библиотеку *tidyverse* и унутар ње библиотеку *ggplot* преко које ћемо графички да представимо расподелу колоне *Distance\_from\_Home.*

A white rectangular object with a black border

Description automatically generated



На графику можемо da видimo да се највише појављује вредност *Near* у колони *Distance\_from\_Home. NA* вредности ћемо да попунимо том вредношћу јер је најзаступљенија. Та промена неће да утиче толико на податке јер имамо веома мало недостајућих података, а пуно *Near* вредности.

A screen shot of a computer

Description automatically generated

Када смо попунили недостајуће вредности, расподела колоне *Distance\_from\_Home* изгледа као што видимо на слици изнад. Број вредности у *Near* категорији се повећао за број недостајућих вредности.

## Колона *Parental\_Education\_Level*

Сада ћемо да поновимо поступак попуњавања недостајућих вредности и у колони *Parental\_Education\_Level.* Наћи ћемо најучесталију вредност и након тога приказати графички расподелу колоне и видећемо како ћемо да решимо проблем *NA* вредности у тој колони.

A screenshot of a computer

Description automatically generated

A graph of a bar chart

Description automatically generated with medium confidence

Видимо да се у овој колони највише понавља вредност *High School,* тако да ћемо наше недостајуће вредности у овој колони заменити том вредношћу. Због малог броја недостајућих вредности, ова промена неће значајно утицати на податке у нашем *dataset-у.*

A screenshot of a computer program

Description automatically generated

Када смо попунили недостајуће вредности у колони *Parental\_Education\_Level,* расподела колоне изгледа као на слици изнад. Број вредности у *High School* категорији се повећао за број недостајућих вредности.

## Колона *Teacher\_Quality*

Остало нам је да решимо проблем недостајућих вредности у колони *Teacher\_Quality.* Видећемо графички и бројчано каква је раподела ове колоне и на основу тога ћемо одлучити шта даље да радимо са недостајућим вредностима.

A white rectangular object with a black border

Description automatically generated

A graph of a number of blue rectangular objects

Description automatically generated with medium confidence

Можемо видети да има доста више вредности *Medium* у колони од осталиих тако да ћемо поља са недостајућим вредностима да попунимо овом вредношћу. Тиме нећемо толико да утичемо на податке јер имамо мало недостајућих вредности у односу на остатак података.

A screen shot of a computer

Description automatically generated

Када смо попунили недостајуће податке одређеном вредношћу, расподела колоне *Teacher\_Quality* изгледа као на слици изнад. Број вредности у *Medium* категорији се повећао за број недостајућих вредности.

# Опис података

У овом поглављу ћемо се детаљније посветити анализи структуре података, представићемо атрибуте *dataset-a*, њихове типове и распон вредности, како бисмо боље разумели наш скуп података и касније извршили предикцију.

## Описи нумеричких колона

Нумеричке колоне у *dataset-u* ћемо анализирати како би се стекло боље разумевање расподеле вредности и идентификовале евентуалне аномалије у подацима. За сваку колону ћемо испитати које су минималне и максималне вредности, квартили (25%, 50% и 75%), као и средња вредност и стандардна девијација.

### Нумеричка колона *Hours\_Studied*

*Hours\_Studied* представља број сати које је студент провео учећи недељно.

Функција *summary* нам приказује основне статистике о подацима из колоне *Hours Studied* као што су: минимална вредност, максимална вредност, просечна вредност, медијана (вредност која дели податке на два једнака дела, где је 50% података испод те вредности), први квантил (25% података испод те вредности) и трећи квантил (75% података испод те вредности).

A screenshot of a computer

Description automatically generated

Ови подаци су релативно равномерно распоређени, али максимална вредност (44) изгледа као да одступа, што би могло бити вредно додатне анализе, као потенцијални *outlier*.

A graph of a column

Description automatically generated with medium confidence

На основу хистограма можемо да видимо да је већина студената учила између 15 и 25 сати недељно, што значи да су подаци сконцентрисани у средњем делу. Мали број података одскаче од уобичајених вредности. Сада ћемо преко *boxplot-a* да прикажемо ове податке како бимсо утврдили постојање *oulier-a.*

A screen shot of a graph

Description automatically generated

Постоје *outlier-i* када је *Hours Studied* мањи од 4 и већи од 36. Међутим, уопште није немогуће да неки студенти проводе толико времена учећи и вероватно ће те вредности имати утицај на резултате тестова. Сматрамо да ће нам ове вредности бити од помоћи у креирању модела.

Користећи Z-score методу можемо идентификовати тачне вредности *outlier-а*.

A screenshot of a computer code

Description automatically generated

Резултати су показали следеће вредности: 43, 1, 38, 2, 39, 44, као и неколико других поновљених вредности. Ове вредности су значајно одвојене од просека, што указује на то да представљају екстремне случајеве у овом скупу података.

### Нумеричка колона *Attendance*

*Attendance* представља проценат долазака на часове.

Помовићемо поступак као и са претходном колоном и позвати функцију *summary* за колону *Attendance.*

A screenshot of a computer

Description automatically generated

Према овим подацима можемо да закључимо да је присуство студената углавном високо, са минималном вредношћу од 60% и максималном од 100%. Медијана је 80%, што значи да половина студената има присуство веће од 80%. Просечна вредност је такође близу медијане, 79.98%, што указује на релативно симетричну дистрибуцију, што можемо да докажемо и помоћу функције *skewness* која израчунава асиметрију (секвенс) дистрибуције. Секвенс близу 0 (између -0.5 и 0.5) указује на симетричност.

A close-up of a white background

Description automatically generated

С обзиром на то да је *skewness* практично нула, можемо закључити да подаци о присуству студената немају екстремне вредности и да су равномерно распоређени око медијане. Ипак ћемо се уверити у то и преко графика, на пример преко *boxplot-а.*

A graph with blue squares

Description automatically generated

Овде заиста можемо да се уверимо да нема вредности које одскачу.

### Нумеричка колона *Sleep\_Hours*

*Sleep\_Hours* представља просечан број сати сна по ноћи.

Помовићемо поступак као и са претходном колоном и позвати функцију *summary* за колону *Sleep\_Hours.*

A screenshot of a computer

Description automatically generated

Просек броја сати спавања је ~7 сати, што указује да већина има уобичајену количину сна, поготово за студенте. Слично као и за колону *Attendance,* делује да је симетрична расподела података. Проверићемо то и пркео *histogram-a.*

A graph of a number of blue bars

Description automatically generated with medium confidence

Већина података се концентрише око 6-8 сати, што представља здрав опсег сна. Студенти који спавају 4 сата или 10 сати могу се сматрати изолованим случајевима. Ове вредности нису превише честе, али су у границама реалног понашања. Касније ћемо видети да ли и какав утицај сан може да има на перформансе на испитима.

Следећи корак јесте рачунање стандардне девијације која је важна статистичка мера и помаже нам да разумемо расподелу података у односу на средњу вредност.

A white rectangular object with a black border

Description automatically generated

Стандардна девијација износи 1.468 сати. Ова вредност указује на умерену варијабилност у расподели података. У овом случају, студенти спавају у широком распону сати, што може указивати на различите навике спавања. Вредности које значајно одступају од овог просека могу бити предмет даљег истраживања, с тога ћемо користити *boxplot*.

A graph of a number of hours

Description automatically generated with medium confidence

Са овог графика можемо закључити да нема вредности које одскачу.

### Нумеричка колона *Previous\_Scores*

*Previous\_Scores* представља оцене са претходних испита.

Поновићемо поступак као и са претходном колоном и позвати функцију *summary* за колону *Previous\_Scores.*

A screenshot of a computer

Description automatically generated

Укупно, расподела *Previous\_Scores* изгледа као да је уравнотежена, са већином студената у распону од 63 до 88, а просек и медијана указују на то да је већина студената остварила добар резултат. Ипак, постоје и студенти са значајно нижим резултатима, најнижи резултат је 50.00, што може указивати на потребу за допунском наставом.

A graph of blue lines

Description automatically generated with medium confidence

Са графика изнад можемо да закључимо да су вредности приближно равномерно распоређене, свака вредност има сличну фреквенцију. Уочљива разлика је једино за вредности које су екстремно ниске или екстремно високе (за вредности 50 и 100), али нема доминантних вредности или великих одступања у појединим сегментима. Пошто је распон вредности велики вероватно је стандардна девијација висока.

A white rectangular object with black text

Description automatically generated

Као што смо и претпоставили вредност је висока, чак и у односу на вредност стандардне девијације за колону *Sleep\_Hours* која је износила ~1.4*.*

Пошто су све вредности једнако заступљене, мислимо да нема екстремних вредности. Али ћемо свакако приказати то и на *boxplot-u*.

A screenshot of a graph

Description automatically generated

Како смо и очекивали, нема *oulier-a.*

### Нумеричка колона *Tutoring\_Sessions*

*Tutoring\_Sessions* представља број посећених допунских часова месечно.

Поновићемо поступак као и са претходном колоном и позвати функцију *summary* за колону *Tutoring\_Sessions.*

A screenshot of a computer

Description automatically generated

Дистрибуција података показује да већина студената има мали број долазака на допунске часове, са 25% оних који немају ниједан долазак. Само неколико студената има већи број долазака, што указује на неуједначеност у коришћењу туторства.

Израчунаћемо *skewness* да бисмо видели да ли су подаци у колони *Tutoring\_Sessions* симетрични.

A screen shot of a computer

Description automatically generated

Пошто резултат ове функције мора да буде између -0.5 и 0.5 да би се подаци сматрали симетричним, у овом примеру можемо видети да подаци нису симетрични јер је резултат функције 0.81. Сада ћемо преко *histogram-a* визуелно приказати расподелу података у колони *Tutoring\_Sessions.*

A graph on a screen

Description automatically generated

Са слике се јасно види да постоји солидан број студената (чак 25%) који нису ишли на допунске часове код тутора. Такође се види да је мали број студената ишло на више од 3 часа месечно. Остали број студената је углавном имао 1 до 2 часа месечно. Ова колона би била погодна за *Feature Engineering,* који ћемо одрадити касније.

A graph on a screen

Description automatically generated

Преко *boxplot-a* јасно видимо које вредности одскачу од просека (4, 5, 6, 7, 8).

### Нумеричка колона *Physical\_Activity*

*Physical\_Activity* представља просечан број сати физичке активности недељно.

Поновићемо поступак као и са претходном колоном и позвати функцију *summary* за колону *Physical\_Activity.*

A screenshot of a computer

Description automatically generated

Можемо да видимо да је минимална вредност је 0, што значи да неки људи уопште немају физичку активност, али медијана (3 сата) указује да је пола студената активно најмање 3 сата недељно. Средња вредност (2.968) је слична медијани, што значи да су подаци релативно симетрични и да не постоје екстремни *oulier-i* који би значајно утицали на просечне вредности.

A graph of a bar chart

Description automatically generated with medium confidence

Већина података је сконцентрисана на око 3 сата физичке активности недељно. Има студената који или немају никакву или имају високу физичку активност, али је број таквих студената знатно мањи у односу на остатак. Како су подаци врло симетрично распоређени, нема потребе да приказујемо *boxplot* у потрази за *outlier-ima*.

### Нумеричка колона *Exam\_Score*

*Exam\_Score* представља оцену завршног испита.

Поновићемо поступак као и са претходном колоном и позвати функцију *summary* за колону *Exam\_Score.*

A close up of a number

Description automatically generated

Медијана износи 67 и близу је средње вредности од 67.24, што указује на то да су резултати распоређени релативно симетрично. Први квартил и трећи квартил показују да је половина студената постигла резултате између 65 и 69. Дакле, већина студената је постигла резултате близу просека, а мали број студената је постигао резултате који су бољи од просека.

A graph of a graph

Description automatically generated

Као што смо и претпоставили, на основу функције *summary,* мали број студената је постигао резултате боље од 75.

## Описи категоријских колона

### Категоријска колона *Parental\_Involvement*

*Parental\_Involvement* представља ниво укључености родитеља у образовање студента *(Low, Medium, High).* Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

A white box with black text

Description automatically generated

A screenshot of a computer screen

Description automatically generated

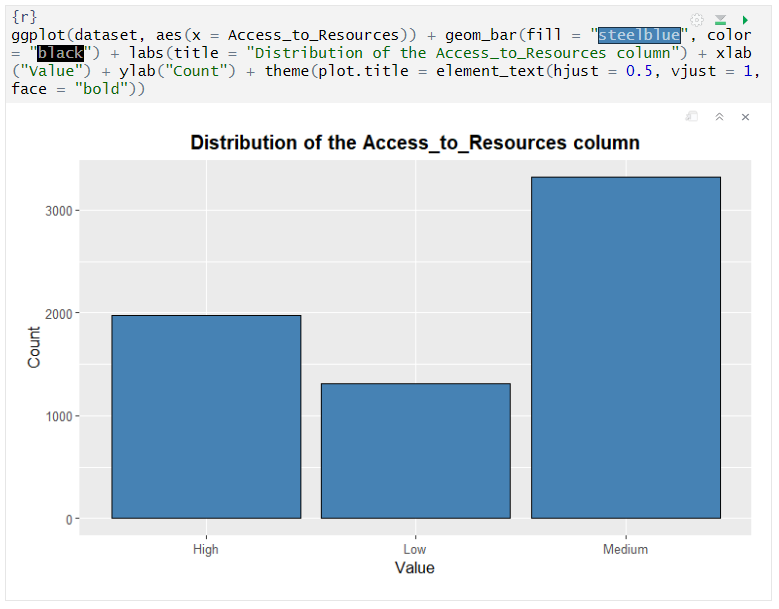
Са графика можемо видети да је укљученост родитеља у образовање студената најчешће средње (*Medium*). Касније ћемо видети да ли то има неки утицај на колону која се предвиђа.

### Категоријска колона *Access\_to\_Resources*

*Access\_To\_Resources* представља доступност образовних ресурса *(Low, Medium, High).* Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

A screenshot of a computer

Description automatically generated



Са графика можемо видети да је доступност образовних ресусра за студенте углавном средње (*Medium*).

### Категоријска колона *Extracurricular\_Activities*

*Extracurricular\_Activities* представља учешће у ваннаставним активностима *(Yes, No).* Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

A close-up of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

Са графика можемо видети да је више студената имало ваннаставне активности.

### Категоријска колона *Motivation\_Level*

*Motivation\_Level* представља ниво мотивације студента *(Low, Medium, High).* Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

A white rectangular object with a black border

Description automatically generated

A screen shot of a graph

Description automatically generated

Са графика можемо видети да је мотивација студената углавном средње вредности (*Medium*).

### Категоријска колона *Internet\_Access*

*Internet\_Access* представља доступност приступа интернету *(Yes, No).* Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

A white rectangular object with a black border

Description automatically generated

A screen shot of a computer

Description automatically generated

Као што је и очекивано, велики број студената има приступ интернету наспрам веома малог броја оних који немају. Питамо се колико ће значајан фактор ово бити касније.

### Категоријска колона *Family\_Income*

*Family\_Income* представља ниво породичног прихода *(Low, Medium, High).* Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

A white rectangular object with a black border

Description automatically generated

A screenshot of a graph

Description automatically generated

Са графика можемо закључити да је у односу на *Low* и *Medium, High* вредност у колони *Family\_Income* је доста мања од остале две категорије, што значи да постоји мањи број студената који имају висок породични приход.

### Категоријска колона *Teacher\_Quality*

*Teacher\_Quality* представља квалитет наставника *(Low, Medium, High).* Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

A white rectangular object with a black border

Description automatically generated

A screen shot of a graph

Description automatically generated

Видимо да највише има професора који имају средњу оцену (*Medium*), док је број оних који су лоше оцењени (*Low*) најмањи, што је добро. Видећемо колики утицај ово има на допунску наставу и крајњи резултат на испиту.

### Категоријска колона *School\_Type*

*School\_Type* представља врсту школе коју студент похађа *(Public, Private).* Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

A white rectangular object with a black border

Description automatically generated

A screenshot of a computer screen

Description automatically generated

Можемо видети да је већи број оних студената који похађају државну школу.

### Категоријска колона *Peer\_Influence*

*Peer\_Influence* представља утицај вршњака на академски успех студента *(Positive, Neutral, Negative).* Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

A white rectangular object with a black border

Description automatically generated

A screenshot of a graph

Description automatically generated

Можемо приметити да је најмањи број утицаја негативан (*Negative*), док су неутрални (*Neutral*) и позитивни (*Positive*) утицаји скоро једнаки. Вероватно је да студенти на које колеге лоше утичу имају лошије резултате.

### Категоријска колона *Learning\_Disabilities*

*Learning\_Disabilities* представља присуство сметњи у учењу *(Yes, No).* Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

A white box with black text

Description automatically generated

A screen shot of a graph

Description automatically generated

Можемо приметити да постоје студенти који имају сметње у учењу, касније ћемо видети колики утицај оне имају на њихов крајни резултат на тестовима.

### Категоријска колона *Parental\_Education\_Level*

*Parental\_Education\_Level* представља највиши ниво образовања родитеља *(High School, College, Postgraduate).* Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

A white box with black text

Description automatically generated

A screen shot of a graph

Description automatically generated

Примећујемо да највећи број студената има родитеље чији је највиши ниво едукације средња школа (*High School*), док је најмање оних који су наставили школовање након факултета на мастер или докторским студијама (*Postgraduate*).

### Категоријска колона *Distance\_from\_Home*

*Distance\_from\_Home* представља удаљеност од куће до школе *(Near, Moderate, Far).* Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

A screenshot of a computer

Description automatically generated

A graph with blue squares

Description automatically generated

Примећујемо да највећи број студената живи близу школе (*Near*), а да је број оних који живе далеко јако мали (*Far*).

### Категоријска колона *Gender*

*Gender* представља пол студента *(Male, Female).* Приказаћемо табелу учестаности да бисмо видели колико имамо категорија за ову колону.

A white rectangular object with a black border

Description automatically generated

A screen shot of a computer

Description automatically generated

Евидентно је да је већи број студената мушког пола.

# Анализа

У овом поглављу истражићемо различите факторе који могу имати утицаја на постигнуте резултате студената на испитима, како бисмо идентификовали кључне варијабле које највише доприносе академском успеху. Планирана анализа ће обухватати и нумеричке и категоријске факторе, како бисмо стекли дубље разумевање корелација и могућности предикције.

## Анализа између предиктора и одговора

У овом делу ћемо да анализирамо оне колоне за које сматрамо да имају утицај на предикцију колоне *Exam\_Score.* Колоне које ћемо обрађивати у даљем раду смо изабрали на основу доменског знања, док оне за оне које нису изабране сматрамо да немају утицај на излазну колону.

### Утицај колоне *Hours\_Studied*

Први фактор који анализирамо је број сати проведених у учењу недељно (*Hours\_Studied*) у односу на оцену на завршном испиту (*Exam\_Score*). Очекивано је да студенти који више времена проводе учећи остварују боље резултате.

A screen shot of a graph

Description automatically generated

Са графика се може приметити позитиван однос између ове две променљиве, у већини случајева, како се повећава број сати у учењу, тако се повећава и резултат на испиту. Иако тренд није строго линеаран, делује да већи број сати у учењу води ка бољим резултатима, али подаци су прилично раштркани, што значи да број сати не гарантује директно већи успех. Такође, можемо да приметимо да има студената који проводе мање сати у учењу, а ипак постижу високе оцене, као и оних који више сати уче, а ипак немају сјајне резултате. Ово су потенцијални *outlier-i*.

A computer screen shot of a computer code

Description automatically generated

A graph with a line going up

Description automatically generated

График приказује позитивну повезаност између броја сати проведених у учењу и резултата на испиту, више сати учења доводи до бољих оцена, али са нешто већим интервалом поверења, значи да постоји већа варијабилност у резултатима код студената који уче више од 35 сати недељно или код оних који уче мање од 10 сати недељно. Постоји могућност да су они који нису учили довољно, а остварили су веома добре резултате, или екстремно надарени или су варали на испитима.

Издвојићемо случајеве када студенти имају више од 80 поена.

A screenshot of a computer

Description automatically generated

Анализирали смо овај подскуп у односу на целокупан *dataset* и не постоје неки значајни патерни или навике које студенти који постижу високе резултате имају у односу на остатак. Све вредности и медијане су јако сличне са минималним разликама.

Свакако ћемо приказати подскуп на графику.

A screen shot of a graph

Description automatically generated

Примећујемо да ретко који студент који има добре резултате учи мање од 10 сати недељно, за студенте који ипак уче мање од 10 сати се поставља питање да ли су можда варали на тестовима, јер остали атрибути попут мотивације, присуства и претходних резултата немају одређен патерн, што се види на наредној слици.

A screenshot of a computer

Description automatically generated

Већина студената је сконцентрисана између 15 и 25 сати, па чак и ту има доста варијације у резултатима. Закључујемо да време проведено у учењу није фактор који самостално може да утиче на поене.

### Утицај колоне *Attendance*

Овај график приказује однос између присуства на часовима (*Attendance*) и постигнуте оцене на завршном испиту (*Exam\_Score*). Очекује се да студенти са већим присуством имају боље резултате на испиту.

A graph with blue dots

Description automatically generated

График нам показује да што више студенти присуствују часовима, углавном добијају боље оцене, али то није увек строго правило. Они који имају више од 90% присуства често постижу боље резултате у односу на оне са мањим присуством који остају у опсегу од 60 до 70 поена. Наравно постоје и изузеци, неки студенти са мањим присуством постижу високе оцене. Дакле, присуство је важно, али није једини фактор који утиче на успех.

A computer screen shot of a blue box

Description automatically generated

A graph with a line drawn on it

Description automatically generated

График показује позитивну повезаност између присуства студената и њихових резултата на испиту, у глобалу можемо рећи да веће присуство води ка бољим оценама.

### Утицај колоне *Access\_to\_Resource*

Сада ћемо анализирати какав утицај има доступност образовних ресурса студента на његове резултате на испиту.

A screenshot of a computer

Description automatically generated

Закључујемо да нема значајних разлика у распону оцена између различитих нивоа доступности образовних ресурса. За све нивое *(High, Low, Medium)*, већина студената добија оцене у распону од 60 до 80 поена, без јасног тренда који би указивао на то да већа доступност ресурса доводи до бољих резултата. Постоји неколико *outlier-а* са високим оценама изнад 90, али они су присутни код свих категорија, што значи да овај фактор сам по себи можда није пресудан.

### Утицај колоне *Sleep\_Hours*

Сада ћемо анализирати колону *Sleep\_Hours* и након анализе колоне видети да ли она има неки утицај на излазну колону *Exam\_Score.*

A screenshot of a graph

Description automatically generated

Видимо да већина студената, без обзира на број сати сна, има оцене концентрисане око 70 поена, што значи да нема значајних разлика у оценама у зависности од броја сати сна. Постоје неки *outlier-и* са изузетно високим резултатима, али они су равномерно распоређени кроз различите категорије, па бисмо рекли да количина сна можда нема велики утицај на резултате на испиту.

### Утицај колоне *Motivaton\_Level*

Анализираћемо колону *Motivation\_Level* у односу на излазну колону и покушаћемо да донесемо закључак да ли ова колона има утицај на предикцију излазне колоне.

A screenshot of a computer screen

Description automatically generated

Можемо да видимо да постоји врло мала разлика у медијанама између различитих нивоа мотивације, па вероватно мотивација не игра значајну улогу у просечним резултатима. Међутим, студенти са високим нивоом мотивације имају нешто шири распон оцена, укључујући неколико изузетно високих резултата, док студенти са ниском мотивацијом такође имају појединачне *outlier-е* са високим оценама, али су у већини ближе просеку.

### Утицај колоне *Internet\_Access*

Сада ћемо анализирати какав утицај има приступ интернету на студентове резултате на испиту.

A screenshot of a computer screen

Description automatically generated

Видљиво је да студенти који имају приступ интернету имају нешто шири распон оцена, укључујући неколико изузетно високих резултата, док студенти без приступа интернету углавном постижу резултате ближе медијану око 70 поена. Иако постоје неки *outlier-и*, изгледа да приступ интернету нема утицај какав смо очекивали.

### Утицај колоне *Tutoring\_Sessions*

Сада ћемо видети какав утицај има колона *Tutoring\_Sessions* на колону која се предвиђа.

A screen shot of a graph

Description automatically generated

Како се број сесија повећава, видимо благ пораст у медијани оцена, а посебно код студената који су имали 5 или више сесија. Постоје и неки outlier-и, али генерално, чини се да већи број сесија доводи до бољих резултата. Ова колона би могла бити корисна за *feature engineering*, јер видимо потенцијални позитиван утицај на резултате на испиту.

Одлучили смо да колону *Tutoring\_Sessions* поделимо у 3 категорије: *None* за оне студенте који уопште нису имали допунске часове, *Between 1 and 3* за оне који су имали између 1 и 3 часа месечно и *More than 3* за оне који су имали више од 3.

A screenshot of a computer

Description automatically generated

На графику испод ћемо приказати расподелу новокреиране категоријске променљиве *Tutoring\_Categories*.

A graph with blue squares

Description automatically generated

Видимо да већина студената припада категорији *Between 1 and 3*, затим иду они који немају никакве допунске часове, а само мали број студената има већи број часова месечно.

Сада ћемо анализирати какав утицај допунски часови имају на крајњи резултат на испиту.

A screenshot of a computer screen

Description automatically generated

Овај график показује да студенти без допунских часова имају медијану око 70, али има неколико изузетака који постижу веома високе резултате. Слично томе, студенти који имају између 1 и 3 часа имају сличну медијану и распон резултата. Док студенти са више од 3 часа имају већу медијану и мањи распон поена, што указује на то да већи број сесија можда има позитиван утицај на успех, иако је узорак мањи и садржи неколико екстремних вредности.

Обрадили смо колоне за које смо сматрали да имају утицај на крајње резултате на испиту, остале колоне смо такође испитивали помоћу графика, али смо закључили да њихов утицај на резултате испита није значајан. Такође, на основу доменског знања, смо одлучили да неке од тих колона избацимо из даље анализе, јер њихово укључивање не би побољшало тачност предикције.

На пример за колону *Peer\_Influence* смо претпостављали да ће имати већи утицај на крајни резултат, поготово за студенте на које колеге утичу лоше, међутим када смо направили график, видели смо да то баш и није тако.

A screen shot of a graph

Description automatically generated

Просечни резултати су веома слични, све три категорије имају сличан распон у резултатима, чак имају у сличан број изузетака. Нема очигледног обрасца или утицаја на поене на испиту.

Слично важи и за колону *Learning\_Disabilities.* Очекивали смо да постоји видљива разлика у резултатима између студената који имају потешкоћа у учењу и оних који немају. Међутим већ на наредном графику се види да та претпоставка не важи. Медијана јесте нижа за студенте са потешкоћама, али и они могу да постижу врло добре резултате. Горња граница јесте нижа (89 поена) у односу на студенте без потешкоћа, али не постоји довољно јака веза која би указивала да потешкоће у учењу имају значајан утицај на крајњи исход.

A screenshot of a computer screen

Description automatically generated

Колоне попут пола студента, типа факултета и укључења родитеља у образовање студента смо заобишли у обради на основу домеског знања. Сматрамо да пол нема утицаја на крајњи резултат, као ни ниво укључености родитеља у образовање. Из свакодневног живота можемо закључити да ови фактори немају утицај на успех студента.

## Мултиваријантна анализа

До сада смо се фокусирали на појединачне факторе који могу утицати на резултате студената на испиту, тако што смо их анализирали изоловано. У стварности академски успех зависи од комбинације више фактора. Зато је неопходно применити мултиваријантну анализу, како бисмо истражили међусобне односе између различитих променљивих и њихов заједнички утицај на резултат. Овај приступ ће нам омогућити да направимо бољи модел за предвиђање крајњих резултата на испиту.

### Повезаност између *Hours\_Studied* и *Attendance*

На основу анализа везаних за утицај различитих колона на излаз, закључили смо да би најбоље било да кренемо са комбинацијом колона *Hours\_Studied* и *Attendancе.* Зато што су ове две променљиве показале највећи индивидуални утицај на резултате на испиту.

Наредни график приказује однос између броја сати проведених у учењу и резултата на испиту, а боја представља присуство.

A screen shot of a graph

Description automatically generated

Можемо да видимо да и време проведено у учењу и присуство јесу важни фактори утицаја на резултате испита. Веће присуство изгледа да надокнађује мањи број сати учења, док нижи ниво присуства захтева више сати учења да би се постигли бољи резултати.

Представићемо овај график на другачији начин и пошто је тешко протумачити присуство и број сати проведених у учењу, поделићемо ове вредности у интервале и покушати да одатле нешто закључимо.

A screenshot of a computer

Description automatically generated

Видимо да веће присуство у комбинацији са више сати учења даје боље резултате, сада ћемо то представити на графику. Одлучили смо се за *heatmap* график.

A screenshot of a graph

Description automatically generated

Најбољи резултати се постижу када студенти уче између 40 и 60 сати недељно и имају присуство веће од 80% (просек је 78 поена). Такође, добре резултате (70-75) могу да постигну студенти који уче мање од 40 сати недељно, али имају веће присуство, или обрнуто, уколико је мање присуство (60-80) надокнади се са више сати учења. Такође примећујемо да не постоје студенти који уче више од 40 сати, а да имају долазност мању од 60%.

### Повезаност између *Hours\_Studied* и *Tutoring\_Categories*

Сада ћемо истражити какав утицај заједно имају време проведено у учењу и број допунских часова на резултате испита.

Прво ћемо одредити просек освојених поена за комбинације различитих интервала за време проведено у учењу и број допунских часова.

A screenshot of a computer

Description automatically generated

Примећујемо да студенти који уопште не иду на допунску наставу имају боље резултате уз више учења. Генерално, са порастом сати учења у комбинацији са допунском наставом се виде бољи резултати.

A screenshot of a graph

Description automatically generated

Најбоље резултате постижу студенти који су учили више од 40 сати недељно уз допунске часове (1-3), примећујемо да не постоје студенти који или нису ишли или су ишли на више од 3 допунска часа, а да су притом учили више од 40 сати недељно. Због тога не можемо за категорију од 40+ сати да видимо утицај допунске наставе. За остале интервале (0-20 и 20-40 сати недељно) можемо да приметимо да допунска настава има позитиван утицај на поене на испиту.

### Повезаност између *Internet\_Access* и *Access\_to\_Resources*

Занимало нас је да ли комбинација приступа интернету и осталим образовним ресурсима може да има значајнији утицај на резултате испита.



Једино што се истиче на овом графику је то да када студенти немају приступ интернету и имају низак ниво доступности образовним ресурсима, медијана је уочљиво нижа од осталих комбинација и нема вредности које одступају. Највећи резултат који је постигнут је ~70 поена.

### Повезаност између *Parental\_Education\_Level* и *Family\_Income*

Претпоставили смо да образовање родитеља може утицати на економски статус породице. Виши ниво образовања често је повезан са бољим радним позицијама и већим приходима, па смо желели да испитамо да ли ове две варијабле имају неку корелацију.

A graph of a graph of a parent

Description automatically generated with medium confidence

Родитељи са средњошколским образовањем углавном припадају нижим и средњим приходима, иако категорија *Postgraduate* имаја нешто већи број у категорији средњих и високих прихода у поређењу са осталим групама, образовање не гарантује висок приход. Да бисмо мало боље анализирали и упоредили различите категорије, сагледаћемо процентуалну расподелу.

A screen shot of a graph

Description automatically generated

Са овог графика се заиста може уочити да ниво образовања нема устицај на приход. Разлике су минималне.

### Повезаност између *Tutoring\_Categories* и *Internet\_Access*

Раније смо истражили заједнички утицај интернета и приступа образовним материјалима, сада ћемо видети да ли приступ интернету може имати утицај на то да ли студенти имају потребу за допунском наставом.

A screenshot of a computer screen

Description automatically generated

Приметили смо да приступ интернету има значај у случају када студенти немају ниједан допунски час, иако је медијана слична распон поена за оне који имају интернет је доста већи и има више изузетака који постижу јако добре резултате. Слично важи и за категорију студената који имају између 1 и 3 допунска часа, где се и на графику очигледно види да имају већу медијану они који имају приступ интернету, а имају и више изузетака.

### Повезаност између *Teacher\_Quality, Tutoring\_Categories* и *Gender*

Замисао је да квалитет професора може да има значајан утицај на број допунских часова, односно, уколико студент сматра да је квалитет професора лош то би значило да му је потребно више допунских часова. И на крају ћемо видети да ли и пол има додатни утицај.

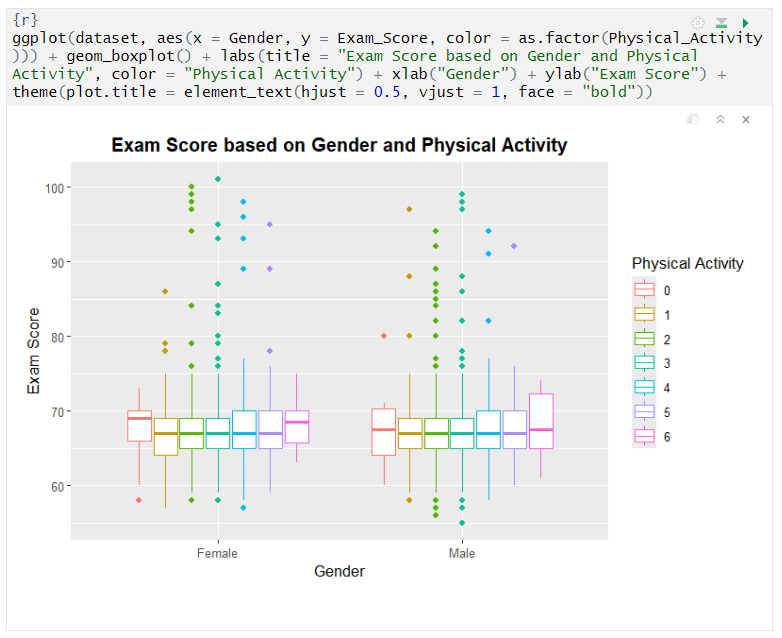
A screenshot of a computer screen

Description automatically generated

Прво примећујемо да пол нема никакав утицај на број допунских часова, јер је распоред врло сличан. Такође видимо да иако неки студенти оцењују професоре као лоше, не похађају више допунских часова у односу на оне који су их ставили у категорију средњег квалитета. Сматрамо да је повезаност минимална до непостојећа.

### Повезаност између *Physical\_Activity* и *Gender*

Физичка активност може да утиче на концентрацију и опште ментално здравље, што на крају може утицати и на резултате тестова. Видећемо да ли је то истина и да ли пол има везе са тим.



Једина приметна разлика се види код студенткиња које немају физичку активност, оне имају бољу медијану у односу на остале студенткиње. Следеће су студенткиње које имају највишу физичку активност, њихова медијана је већа од остатка групе. Код мушкараца постоји сличан патерн али га је теже уочити, што значи да физичка активност нема велики утицај на њихов успех.

### Повезаност између *School\_Type* и *Distance\_From\_Home*

Занима нас да ли удаљеност од куће има утицај на тип школе који студенти уписују, да ли већа удаљеност значи да се студенти пре одлуче за приватни или јавни факултет. Видећемо да ли заједно имају утицај на резултате испита.



Највећа разлика се види када је удаљеност велика, студенти на приватним факултетима имају нешто лошији просек у односу на студенте са државних факултета. Док, када је у питању мала удаљеност од куће, мало бољи успех имају студенти који су на приватним факултетима. Ове разлике нису много значајне у односу на неке претходне примере, и не указују на јасан тренд који има утицај на успех на испиту.

# Креирање модела

Након што смо истражили везе између самих предиктора, као и између предиктора и одговора започећемо креирање модела.

Прво ћемо приказати матрицу корелације како бисмо имали бољи увид у корелацију између нумеричких колона.

A screenshot of a computer

Description automatically generated

Примећујемо да *Exam\_Score* има најбољу корелацију са *Attendance,* а потом са *Hours\_Studied.* Нешто мању корелацију има са *Previous\_Scores* и *Tutoring\_Sessions*.

Затим ћемо трансформисати све категоријске колоне у фактор варијабле.

A computer screen shot of a computer code

Description automatically generated

Потом се уверавамо да су колоне заправо конвертоване.

A screenshot of a computer

Description automatically generated

Наредни корак јесте да наш *dataset* поделимо да тренинг и тест скупове, у односу 75:25.

A screenshot of a computer code

Description automatically generated

Приказаћемо расподелу резултата испита у тренинг и тест скупу понаособ.

A graph of a graph

Description automatically generated

A graph of a graph

Description automatically generated with medium confidence

Помоћу графика боље видимо расподеле резултата испита унутар оба скупа и можемо да се уверимо да су слично подељени и да је све спремно за даљу анализу.

## Линеарна регресија (*Linear Regression*)

Користићемо линеарну регресију како бисмо анализирали однос између предиктора и одговора. Прво ћемо у модел убацити предикторе за које смо се уверили да имају највећи утицај.

A screenshot of a computer

Description automatically generated

Тачност модела је ~51% што и није толико добар резултат, али је F-статистика > 1 и p-вредност < 0.5 из чега видимо да постоји повезаност између предиктора и одговора. Покушаћемо да убацимо и колоне *Previous\_Scores* и *Tutoring\_Sessions* јер су оне следеће дале најбоље резултате и надамо се да ће се тачност модела повећати.

A screenshot of a computer

Description automatically generated

Тачност се повећала на ~56% што је боље, такође видимо да *Adjusted R2*не одступа превише од *Multiple R2* што нам је битно да не бисмо *overfit-*овали податке*,* а и стандардна грешка се смањила.

У модел ћемо додати још две колоне *Parental\_Involvement* и *Access\_to\_Resources*. Како бисмо покушали да повећамо тачност.

A screenshot of a computer

Description automatically generated

Сада тачност износи ~62% што је доста боље у односу на почетни модел. Свих осам предиктора су статистички значајни, са p-вредношћу мањом од 2e-16. Повећањем *Hours\_Studied* и *Tutoring\_Sessions* се повећавају и резултати испита, док негативни коефицијенти за *Parental\_Involvement* и *Access\_to\_Resources* указују на смањење резултата.

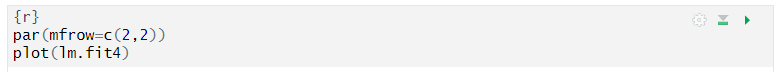
Пробаћемо још мало да побољшамо модел.

A screenshot of a computer

Description automatically generated

Додавањем *Family\_Income* и *Motivation\_Level* добили смо нешто бољи модел са тачношћу од ~64%, што је одлично у односу на почетни модел.

Приказаћемо графички наш модел.



A group of graphs with numbers

Description automatically generated

Помоћу *Residuals vs Fitted* графика проверавамо линеаран однос између предиктора и зависне променљиве. Идеално, резидуали треба да буду равномерно распоређени око хоризонталне линије без јасних образаца. Овде је већина тачака распоређена око линије, али постоје и неке тачке које одступају.

За *Normal Q-Q* график важи да уколико резидуали прате линију онда су нормално распоређени. Видимо да има одступања на крајевима, вероватно због екстремних вредности.

*Scale-Location* график служи за проверу хомоскедастичности. Потребно је да тачке буду једнако распршене око линије. Има доста одступања па можемо да кажемо да варијанса резидуала није константна.

*Residuals vs Leverage* график се користи за откривање *outlier-а*. Као што смо и претпоставили, на основу *Normal Q-Q* графика, имамо неколико екстремних вредности.

## Стабло одлучивања (*Decision Tree*)

Стабло одлучивања омогућава моделирање сложенијих и нелинеарних односа између променљивих, па је корисно ако линеарна регресија не даје довољно прецизне резултате.

A screenshot of a computer

Description automatically generated

Прва подела у стаблу је на основу *Attendance*< 82.5, где се подаци деле на оне са вредностима мањим од 82.5 и оне са вредностима већим или једнаким од 82.5, затим је подела према *Hours\_Studied* итд. На пример, у чвору где је *Attendance* < 69.5 и *Hours\_Studied*< 20.5, просечан резултат испита је ~63.

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A computer screen shot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

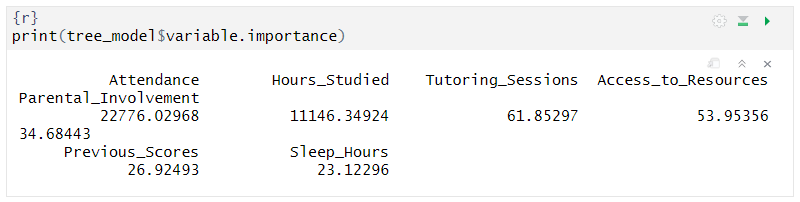
На основу *summary* функције, очиглендно је да су најважнији предиктори *Attendance* (67%) и *Hours\_Studied* (33%). Модел се побољшава кроз неколико фаза цепања, a варијанса (MSE) у резидуалима се смањује, па можемо да кажемо да су ови предиктори релевантни за предвиђање резултата.

Следећи корак је израчунавање метрика тачности модела као што су: *RMSE* (Root Mean Squared Error), *MAE* (Mean Absolute Error), *MSE* (Mean Squared Error), и *R²* (*coefficient of determination*). На основу њих процењујемо колико добро модел предвиђа резултате испита.

A screenshot of a computer program

Description automatically generated

*RMSE* износи ~2.47, што значи да предвиђене вредности у просеку одступају од стварних за око 2.48 поена. *MAE* је ~1.78, значи да је просечно апсолутно одступање предвиђених од стварних вредности релативно мало. *MSE* је 6.12, с обзиром да укључује квадрирање, корисна је да наглси веће грешке. *R²* је ~0.53, што значи да модел објашњава око 53% варијације у резултатима испита.



Видимо да су *Hours\_Studied* и *Attendance* најбитнији предиктори и да додавање нових заправо и не прави разлику јер се стабло одлучивања ослања само на ова два предиктора. Међутим исти резултати се добијају и када су само *Hours\_Studied* и *Attendance* укључени у модел, па га нећемо приказивати.

## *Random Forest*

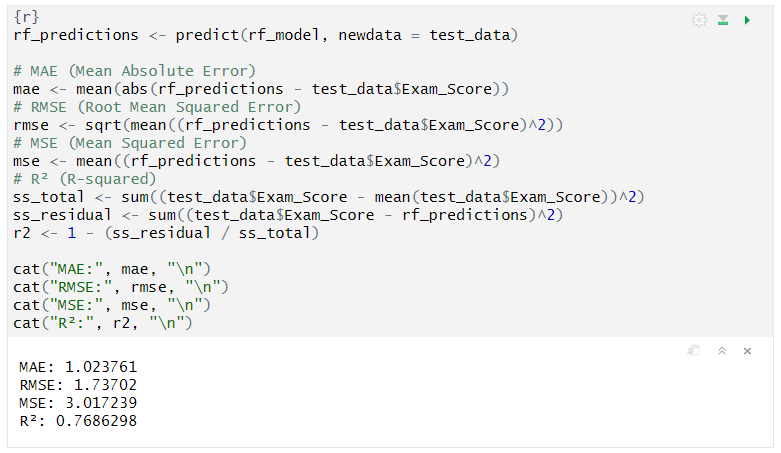
Покушаћемо да направимо и *Random Forest* модел, он прави више стабала одлучивања и комбинује њихове предвиђене вредности. Требало би да повећа стабилност и тачност предвиђања.

A screenshot of a computer program

Description automatically generated

Овде смо креирали *Random Forest* модел са 100 стабала и предвиђамо резултат испита на основу броја сати учења и присуства. Модел објашњава 59.59% варијације у подацима, можемо да приметимо побољшање у односу на претходни модел стабла одлучивања.

Сада ћемо видети да ли су се и метрике тачности побољшале.



Видимо да су и метрике доста боље у осносу на претходне моделе. *MAE*, *RMSE* и *MSE* су се смањиле, па су и просечне грешке у предвиђањима мање. *R²* је повећан, што значи да модел објашњава скоро 77% варијансе у резултатима испита.

# Закључак

Циљ овог рада био је да се анализом података утврди који фактори највише утичу на резултате студената на испитима. Након свих извршених анализа, дошли смо до закључка да су сати учења и присуство на часовима кључни предиктори. Кроз примену различитих модела, најбољи резултати су добијени помоћу Random Forest модела, који је објаснио 76.86% варијације у резултатима испита. Кобминација труда и рада је добитна комбинација за добре резултате на факултету.

# Литература

Увод у науку о подацима – вежбе и предавања

<https://www.r-bloggers.com/2021/04/random-forest-in-r/>

<https://www.kaggle.com/code/mushei/eda-on-student-performance>

<https://www.r-bloggers.com/2021/04/decision-trees-in-r/>

<https://r-graph-gallery.com/>