

ИНСТИТУТ ЗА МАТЕМАТИКУ И ИНФОРМАТИКУ

ПРИРОДНО-МАТЕМАТИЧКИ ФАКУЛТЕТ

УНИВЕРЗИТЕТ У КРАГУЈЕВЦУ

СЕМИНАРСКИ РАД

ПРЕДСТАВЉАЊЕ И ТУМАЧЕЊЕ

СКУПА ПОДАТАКА „DIABETES 012 – HEALTH INDICATORS (BRFSS 2015)2

|  |  |
| --- | --- |
| Ментор | Студенти |
| др Бранко Арсић | Магдалена Стамeновић 83/2021 |
|  | Алекса Радивојевић 76/2021 |

Садржај

[Увод 2](#_Toc649752087)

[Упознавање података 3](#_Toc288029448)

[Учитавање скупа података 4](#_Toc617614332)

[Информисање о скупу података 4](#_Toc1283620340)

[Припрема података за ЕДА 11](#_Toc1586491563)

[Експлоративна анализа података (ЕДА) 12](#_Toc1306102819)

[Униваријантна анализа 13](#_Toc1886888039)

[Анализа 13](#_Toc1500578169)

[Табела закључака 34](#_Toc1622960085)

[Биваријантна анализа 36](#_Toc1895836705)

[Имплементација метода статистичких тестова 37](#_Toc228377885)

[Однос са циљаном карактеристиком Diabetes\_012 38](#_Toc1850236602)

[Однос карактеристика које нису циљане 80](#_Toc508720879)

[Табеле закључака 99](#_Toc1995625750)

[Чишћење података 106](#_Toc638077986)

[Трансформација података 106](#_Toc192538983)

[Инжењеринг карактеристика (Feature Engeenering) 107](#_Toc490980569)

[Инжењеринг карактеристике PhysHlthCat 107](#_Toc289595314)

[Инжењеринг карактеристике MentHlthCat 108](#_Toc1347834997)

[Инжењеринг карактеристике ЕducationCat 108](#_Toc698263461)

[Инжењеринг карактеристике IncomeCat 109](#_Toc191843794)

[Инжењеринг карактеристике AgeCat 110](#_Toc976037883)

[Инжењеринг карактеристике CardioRiskScore 112](#_Toc1239395130)

[Инжењеринг карактеристике LifestyleRiskScore 114](#_Toc270028341)

[Инжењеринг карактеристике HealthScore 116](#_Toc1632581963)

[Инжењеринг карактеристике DietScore 118](#_Toc1965047748)

[Инжењеринг карактеристике SocioEconomicStatus 120](#_Toc1120416968)

[Биваријантна анализа нових карактеристика 123](#_Toc284131280)

[Однос са циљаном карактеристиком Diabetes\_012 123](#_Toc68114605)

[PhysHlthCat vs Diabetes\_012 123](#_Toc1282600922)

[MentHlthCat vs Diabetes\_012 124](#_Toc87777966)

[EducationCat vs Diabetes\_012 126](#_Toc323133834)

[IncomeCat vs Diabetes\_012 128](#_Toc653856206)

[AgeCat vs Diabetes\_012 130](#_Toc753615939)

[CardioRiskScore vs Diabetes\_012 132](#_Toc1003977851)

[LifestyleRiskScore vs Diabetes\_012 134](#_Toc1683971659)

[HealthScore vs Diabetes\_012 135](#_Toc584053498)

[SocioEconomicStatus vs Diabetes\_012 137](#_Toc178683585)

[Однос карактеристикама које нису циљане 139](#_Toc1305896224)

[IncomeCat VS EducationCat 139](#_Toc1179468315)

[MentHlthCat VS Stroke 141](#_Toc2064342004)

[PhysHlthCat vs DiffWalk 143](#_Toc1823889300)

[AgeCat vs HealthScore 145](#_Toc842095191)

[Табела закључака 147](#_Toc463031592)

[Селекција предиктора (Feature Engeenering) 148](#_Toc574550208)

[Моделовање 151](#_Toc5585208)

[Дефинисање тренинг и тест скупа 151](#_Toc987186077)

[Балансирање података циљане карактеристике Diabetes\_012 153](#_Toc1168154826)

[Регуларизација и РСА 154](#_Toc464060662)

[Тренирање модела 155](#_Toc833554591)

[Избор модела 155](#_Toc1612001202)

[Унакрсна валидација (k-fold) 156](#_Toc878198444)

[Финално тренирање 160](#_Toc1109989355)

[Метрике модела 162](#_Toc1057047015)

[Референце 165](#_Toc179839559)

# Увод

Семинарски радом анализирали смо скуп података „[Diabetes 012 – Health Indicators (BRFSS 2015)](https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?resource=download)“ који представља велики статистички скуп здравствених параметара прикупљених у оквиру једног од највећих истраживачких програма о здравственим навикама и стању становништва у Сједињеним Америчким Државама, Behavioral Risk Factor Surveillance System (BRFSS).

У следећем поглављу ће детаљно бити описан садржај скупа, али ево кратак резиме. Подаци из овог скупа обухватају преко 253.000 опсервација, а свака од њих садржи информације о здравственим навикама, присуству хроничних болести, физичкој активности, исхрани, демографским обележјима и општем здравственом стању.

Мотивација за избор овог скупа података заснива се на чињеници да дијабетес представља један од најраспрострањенијих глобалних здравствених проблема данашњице. Према подацима Међународне дијабетес федерације, у 2024. години у Србији је процењено 781.500 одраслих особа ( старости 20-79 година) болује од дијабетеса. Истраживање спроведено у периоду од 2007. до 2017. године показује учесталост дијабетеса типа 1 код младих особа ( <19 година) са просеком од 11,82 оболелих на 100.000 становника, док је просек за узраст до 14 година просек 14,28 оболелих на 100.000 становника. Забрињавајући фактор овог истраживања је повећање учесталости обољевања код деце за 5.9% по години, при чему је најризичнији период живота од 5 до 14 година. Сматрамо да би примена метода науке о подацима на приложеном скупу података, у циљу разумевања фактора ризика и ране предикције дијабетеса допринела смањивању наведених стопа.

Циљеви истраживања:

1. Иницијална припрема и чишћење података, укључујући испитивање структуре скупа, анализу нетипичних вредности, трансформацију променљивих.
2. Детаљна униваријанта, биваријантна и описна анализа података у сврху детекције међусобних утицаја главних индикатора на појаву дијабетеса.
3. Одабир и формирање модела машинског учења, а затим испитивање поузданости и робусности на изабраном скупу података.

# Упознавање података

## Учитавање скупа података

На сајту су доступна три фајла (три скупа података). Разлика између ова три скупа је у расподели и категоријама у карактеристици „Diabetes\_012“. Фајл који смо ми изабрали за рад је „diabetes\_012\_health\_indicators\_BRFSS2015.csv“, смештен у истом фолдеру као и „R“ скрипта под називом „diabetes\_dataset.csv“. Прво смо поставили радни директоријум као директоријум у ком се налази „R“ скрипта.

|  |
| --- |
|  |
| radni\_direktorijum <- dirname(rstudioapi::getActiveDocumentContext()$path)  setwd(radni\_direktorijum)  cat("Radni direktorijum je postavljen na:", getwd(), "\n") |
|  |
| Radni direktorijum je postavljen na: E:/FAX/UNP/Семинарски рад |

Затим смо учитали скуп података у радно окружење под променљивом data.

|  |
| --- |
|  |
| data <- read.csv("diabetes\_dataset.csv")  cat("Skup podataka mozete pogledati u promenljivoj data koja se automatski otvorila u radnom okruzenju", View(data)) |
|  |
| Skup podataka mozete pogledati u promenljivoj data koja se automatski otvorila u radnom okruzenju |

## Информисање о скупу података

На почетку анализе детаљно смо се информисали о структури скупа података и значењу свакe карактеристике (feature-а). Број карактеристика (feature-а) и број опсервација у скупу података добијени су коришћењем R функције dim(), након чега су вредности исписане командом cat(). На овај начин утврђено је да скуп садржи 22 карактеристикe (feature-а) и 253680 опсервација, што је довољно за формирање доброг предиктивног модела.

|  |
| --- |
|  |
| dimenzije\_skupa = dim(data)  cat("Broj featura skupa: ", dimenzije\_skupa[2], "\n", "Broj opservacija: ", dimenzije\_skupa[1]) |
|  |
| Broj featura skupa: 22  Broj opservacija: 253680 |

Значење карактеристика (feature-а) смо приказали у следећој табели. На основу доменског знања смо их поделили у категорије како би приликом анализе података лакше испитивали синергијски утицај на предиктивну карактеристику.

|  |  |
| --- | --- |
| Diabetes\_012 | Означава да ли испитаник има дијабетес, предијабетес или га нема уопште. Ова карактеристика се узима као предиктивна. |
| Параметри здравља | |
| HighBP | Да ли испитаник има висок крвни притисак (хипертензију) |
| HighChol | Да ли испитаник има висок холестерол. |
| CholCheck | Да ли је испитаник радио проверу холестерола последњих 5 година |
| BMI | Индекс телесне масе, који је прорачун тежине и висине. |
| GenHlth | Како је испитаник оценио генерално своје здравље о д1 до 5 (1 – одлично, 2 – врло добро, 3 – добро, 4 – задовољавајуће, 5 - лоше ) |
| Параметри стања | |
| HeartDiseaseorAttack | Да ли испитаник има хронично срчано обољење или је имао срчани удар |
| Stroke | Да ли је испитаник имао мождани удар |
| MentHlth | Број дана у последњих 30 дана када је испитаник имао менталне проблеме или лоше ментално здравље попут стреса, регулације емоција, депресије. |
| PhysHlth | Број дана у последњих 30 дана када је испитаник имао физичке проблеме или лоше физички осећао попут болести, повреда. |
| DiffWalk | Да ли испитаник има потешкоће у кретању или пењању уз степенице. |
| Параметри исхране | |
| Fruits | Да ли испитаник једе воће једном или више пута у току дана |
| Veggies | Да ли испитаник једе поврће једном или више пута у току дана |
| Параметри животних навика |  |
| Smoker | Да ли је испитаник пушач (сматра се да јесте ако је попушио барем 100 цигара у животу) |
| HvyAlcoholConsump | Да ли је испитаник зависник од алкохола (више од 14 пића недељно за мушкарце, односно 7 пића за жене) |
| PhysActivity | Да ли је испитаник имао рекреативну физичку активност последњих 30 дана |
| Социјално-економски параметри | |
| AnyHealthcare | Да ли испитаник има некакву врсту здравствене неге укључујући здравствено осигурање или било какво здравствено покриће |
| NoDocbcCost | Да ли је у последњих 12 месеци испитаник имао потребу за доктором али није могао да приушти због трошкова |
| Income | Колике приходе има испитаник од 1 до 8 (1 - мање од 10.000 долара, 5 - мање од 35.000 долара, 8 - 75.000 долара или више) |
| Демографски параметри | |
| Sex | Ког је пола испитаник |
| Age | Колико испитаник има година |
| Education | Који степен едукације има испитаник од 1 до 6 (1 - Никада није ишао у школу или само у вртић 2 - основна школа, 3 – трогодишња средња школа, 4 - четворогодишња средња школа, 5 - факултет 3 године, 6 - факултет 4 године или више |

Након што смо стекли увид у значење и контекст сваке појединачне карактеристике (feature-а), приступили смо анализи структуре скупа података. Испитивање структуре података представља један од кључних корака у процесу припреме података, јер нам омогућава да проверимо да ли је скуп података исправно учитан, те да проценимо да ли су све променљиве погодне за даљу статистичку анализу и изградњу модела машинског учења.

Основни циљ ове фазе је да стекнемо следеће увиде:

* Формат и типови података, као и идентификација потенцијалних категоријских променљивих.
* Провера валидности и квалитета података како би установили потенцијалне грешке у уносу, неправилни опсези, неочекиване вредности.
* Информисање о основној расподели и варијабилности вредности.

За иницијално испитивање структуре скупа података користили смо функцију str(), која пружа сажет али информативан приказ типа објекта и преглед првих неколико вредности за сваку променљиву.

|  |
| --- |
|  |
| str(data) |
|  |
| 'data.frame': 253680 obs. of 22 variables:  $ Diabetes\_012 : num 0 0 0 0 0 0 0 0 2 0 ...  $ HighBP : num 1 0 1 1 1 1 1 1 1 0 ...  $ HighChol : num 1 0 1 0 1 1 0 1 1 0 ...  $ CholCheck : num 1 0 1 1 1 1 1 1 1 1 ...  $ BMI : num 40 25 28 27 24 25 30 25 30 24 ...  $ Smoker : num 1 1 0 0 0 1 1 1 1 0 ...  $ Stroke : num 0 0 0 0 0 0 0 0 0 0 ...  $ HeartDiseaseorAttack: num 0 0 0 0 0 0 0 0 1 0 ...  $ PhysActivity : num 0 1 0 1 1 1 0 1 0 0 ...  $ Fruits : num 0 0 1 1 1 1 0 0 1 0 ...  $ Veggies : num 1 0 0 1 1 1 0 1 1 1 ...  $ HvyAlcoholConsump : num 0 0 0 0 0 0 0 0 0 0 ...  $ AnyHealthcare : num 1 0 1 1 1 1 1 1 1 1 ...  $ NoDocbcCost : num 0 1 1 0 0 0 0 0 0 0 ...  $ GenHlth : num 5 3 5 2 2 2 3 3 5 2 ...  $ MentHlth : num 18 0 30 0 3 0 0 0 30 0 ...  $ PhysHlth : num 15 0 30 0 0 2 14 0 30 0 ...  $ DiffWalk : num 1 0 1 0 0 0 0 1 1 0 ...  $ Sex : num 0 0 0 0 0 1 0 0 0 1 ...  $ Age : num 9 7 9 11 11 10 9 11 9 8 ...  $ Education : num 4 6 4 3 5 6 6 4 5 4 ...  $ Income : num 3 1 8 6 4 8 7 4 1 3 ... |

Увидели смо да су све карактеристике учитане као нумеричке, али на основу информација о значењу можемо закључити да нису по природи све нумеричке. HighBP, HighChol, CholCheck, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, DiffWalk, Sex оне представљају да/не вредности и модели их третирају као категорисјке променљиве. Diabetes\_012 такође може бити категоријска променљива, па је проблем који решавамо мултифункционалан, а не бинарни. Поред њих установили увидели смо ординалне скале у карактеристикама GenHlth, Education, Income. Док су карактеристике BMI, MentHlth, PhysHlth, Age континуалне карактеристике. Трасформацију типова ћемо обрадити у [поглављу припрема за експлоративну анализу](#_Припрема_података_за).

Други увид је потврда документације да нема недостајућих вредности, па је следећи корак провера валидности и квалитета података како би установили потенцијалне грешке у уносу, неправилни опсези, неочекиване вредности. За ту сврху користили смо функцију описне статистике summary(), која уједно даје и информације о основној расподели.

|  |
| --- |
|  |
| summary(data) |
|  |
|  |
|  |

Увиди на основу описне статистике summary() који ће нам помоћи у креирању предиктивних модела су приказани у табели испод.

|  |  |  |
| --- | --- | --- |
| Diabetes\_012 | Максимум нам потврђује категоријску класификацију од три категорије. Вредност медијане, првог и трећег квантила је 0, то значи да је више од 75% опсервација нема дијабетес, док средња вредност говори да 29,69% има неки облик дијабетеса. Закључили смо да расподела класа није у равнотежи што је проблематично моделима. У наредним поглављима биће објашњено решење проблема. | |
| Параметри здравља | | |
| HighBP | Средња вредност нам говори да 42,9% опсервација има повишен притисак. Овај податак нам говори и о високој варијанси (Var=0.429⋅(1−0.429)=0.429⋅0.571=0.245 приближно једнако максимално могућој 0.25) | |
| HighChol | Иста ситуација као код HighBP. | |
| CholCheck | Укупно 96,27% испитаника је проверило ниво холестерола у последњих пет година, што указује на веома малу варијабилност ове променљиве. Ипак, посматрано из угла доменског знања, ова карактеристика је значајна јер омогућава процену валидности податка о високом холестеролу. Наиме, пошто вредност холестерола има значајан утицај на модел (што се види из варијансе ове променљиве), поставља се питање да ли су подаци поуздани у случајевима када испитаник није вршио проверу у последњих пет година.  Стога је неопходно испитати бројност таквих “непотврђених” уноса. Уколико их има релативно мало, могли би се искључити да се не би нарушила робусност модела. Међутим, ако су заступљени у већем броју, бољи приступ је додавање нове променљиве која указује на валидност уноса о холестеролу, уместо њиховог уклањања. | |
| BMI | На основу доменског знања, претпостављамо да има аутлајера или грешка у уносу, јер минимална вредност је екстремно низак индекс, као и максимална вредност као екстремно висок. У даљем раду смо детаљније испитали. Већина испитаника је преко нормалне тежине, а како је средња вредност већа од медијане то нам говори да постоје веома високе вредности. | |
| GenHlth | На основу медијане закључујемо да више од половине испитаника сматра да је њихово здравље врло добро или боље, а по трећем квантилу са категоријом добро достиже и 75%. Да би расподела била равномернија разматраћемо о прегруписању постојећих категорија. | |
| Параметри стања | | |
| HeartDiseaseorAttack | На основу средње вредности закључујемо да 9,4% опсервација има хронично срчано обољење или је имао срчани удар. На основу доменског знања закључили смо да је однос реалан и смислен. | |
| Stroke | Променљива Stroke показује да је мождани удар изузетно ретка појава у анализираном узорку, будући да је само 4,06% испитаника пријавило да је икада имало мождани удар. Јасно је видљиво кроз податак да су први квартил, медијана и трећи квартил једнаки нули па је варијабилност мала. Овим се очекује да ће њен индивидуални утицај на модел бити ограничен, променљива и даље има потенцијалну релевантност због добро познате повезаности можданог удара са метаболичким поремећајима, укључујући дијабетес. Стога ћемо у наставку рада посебно анализирати синергијски утицај ове променљиве у комбинацији са другим здравственим индикаторима. | |
| MentHlth | На основу медијане закључујемо да је најмање пола опсервација рекло да ниједан дан није имало стрес последњих месец дана. У просеку имали су око 3 дана са менталним потешкоћама што је на основу опсега до 30 тежња података ка нижим вредностима. Вредности су у опсегу по документацији (минимум и максимум). Закључујемо да треба графички испитати дистрибуцију па груписати вредности по категоријама у уоченим опсезима. | |
| PhysHlth | Подаци нам говоре слично као и код карактеристике MentHlth. У овом случају просек је око 4 дана са физичким потешкоћама, али исто имамо дистрибуцију података ка нижим вредностима опсега. Закључак је испитати графички детаљније. | |
| DiffWalk | На основу анализе закључујемо да постоје опсервације са и без потешкоћа. Расподела тих вредности није равномерна. У просеку 16,82% опсервација има потешкоће. Ово не чини њену варијансу велико, али на основу доменског знања закључујемо да је повезана са другим здравственим карактеристикама. У даљем раду ћемо испитивати њен синергијски утицај. | |
| Параметри исхране | |
| Fruits | 63.4% опсервација једе воће једном или више пута у току дана, што чини високу варијансу карактеристике. | |
| Veggies | 81.1% опсервација једе поврће једном или више пута у току дана. Висок проценат нам говори о мањој варијанси. | |
| Параметри животних навика | | |
| Smoker | Средња вредност нам говори да је 43,3% опсервација пушач, што чини добру варијабилност. | |
| HvyAlcoholConsump | 5.6% опсервација спада у алкохоличаре, што чини веома малу варијансу и неравномерну расподелу. | |
| PhysActivity | Видимо да је 75,6% опсервација физички активно. На основу доменског знања овај податак нам се учинио сумњивим па смо га у даљем раду детаљније анализирали у биваријантној анализи. | |
| Социјално-економски параметри | | |
| AnyHealthcare | 95% испитаника има неки вид здравствене заштите. Висок проценат је и знак мале варијабилности, па предпостављамо да овај параметар неће бити од велико значајан за модел. | |
| NoDocbcCost | 8,41% опсервација у последњих 12 месеци имао је потребу за доктором али није могао да приушти због трошкова. | |
| Income | Као ординална категорија варијанса зависи од расподела. Видимо теденцију као вишим приходима, тако да смо узели у даље испитивање, како би равномерно биле заступљене групе. | |
| Демографски параметри | | |
| Sex | Видимо равномерну расподелу, однос 44% мукараца и 56% жена. А самим тим по формули даје високу варијансу. | |
| Age | Медијана и средња вредност су приближно једнаке па можемо закључити симетричну расподелу. Распон вредности је од 1 до 13, па на основу доменског знања закључујемо да се ради о старосним групама. За формирање група у расподели која би користила моделу користили смо вредности квантила. Биће обајшњено у поглављу трансформације података. | |
| Education | Већина опсервација је концентрисана у мањем опсегу виших категорија образовања (1st Qu. = 4 (завршену четворогодишњу средњу школу), а 50% има већу од 5 (трогодишњи факултет). Закључили смо да можемо прегруписати постојеће категорије како би добили равномернију расподелу. | |

Након детаљног упознавања са значењем променљивих и анализе њихових расподела, као и формулисаних почетних закључака, приступили смо процесу припреме података за експлоративну анализу података (EDA).

# Припрема података за ЕДА

У прошлој фази установили смо да у скупу података нема недостајућих вредности. С обзиром на то, у овој припремној фази фокус је на дефинисању исправних типова карактеристика и факторизацији категоријских променљивих.

Факторизација је поступак претварања променљивих које представљају категорије у посебан тип података факторе. Фактор је структура података која садржи ограничен број категорија и користи се када променљива не представља непрекидну нумеричку вредност, већ одређене класе или редослед класа. Иако су све променљиве у оригиналном скупу података биле учитане као нумеричке, током прегледа структуре података (str()) и описних статистика (summary()) установљено је да велики број њих садржи вредности у опсегу 0/1 или у малом дискретном скупу бројева (0,1,2,…). Факторизација је спроведена како би се ове вредности правилно третирале у статистичкој анализи и визуелизацији, јер би њихово задржавање као numeric довело до погрешног тумачења.

Следећим Р кодом реализовали смо [закључке](#УвидиОтиповимКарактеристика) о факторизацији:

|  |
| --- |
| Факторизација бинарних карактеристика у категорије не/да. |
| daNe\_kategorije = c("HighBP", "HighChol", "CholCheck", "Smoker", "Stroke",  "HeartDiseaseorAttack", "PhysActivity", "Fruits", "Veggies",  "HvyAlcoholConsump", "AnyHealthcare", "NoDocbcCost", "DiffWalk")  data[daNe\_kategorije] = lapply(data[daNe\_kategorije], factor, levels = c(0,1), labels = c("Ne","Da")) |
| Факторизација бинарне карактеристике Sex у категорије: женско / мушко. |
| data$Sex = factor(data$Sex, levels = c(0,1), labels = c("zensko","musko")) |
| Факторизација карактеристике Diabetes\_012 у категорије: нема дијабетес / предијабетес / дијабетес. |
| nivoi\_Diabetes\_012 = sort(unique(data$Diabetes\_012))  data$Diabetes\_012 = factor(data$Diabetes\_012,  levels = nivoi\_Diabetes\_012,  labels = c("nema dijabetes", "predijabetes", "dijabetes")) |
| Конверзија променљивих GenHlth, Education и Income у ординалне факторе, односно категорије са дефинисаним редоследом. |
| ordinalne\_kategorije = c("GenHlth", "Education", "Income")  nivoi\_GenHlth = sort(unique(data$GenHlth), decreasing = TRUE)  data$GenHlth = factor(  data$GenHlth,  levels = nivoi\_GenHlth,  labels = c("Odlično", "Vrlo dobro", "Dobro", "Zadovoljavajuće", "Loše"),  ordered = TRUE  )  nivoi\_Income = sort(unique(data$Income))  data$Income = factor(data$Income, levels = nivoi\_Income,  ordered = TRUE,  labels = c("<10.000$",  "10.000-14.999$",  "15.000-19.999$",  "20.000-24.999$",  "25.000-34.999$",  "35.000-49.999$",  "50.000-74.999$",  ">=75.000$"))  nivoi\_Education = sort(unique(data$Education))  data$Education<- factor(data$Education, levels = nivoi\_Education,  ordered = TRUE,  labels = c("bez skole/isao u vrtic",  "osnovna skola",  "3-god srednja",  "4-god srednja",  "3-god fakultet",  "4-god fakultet")) |

Након тога смо функцијом str() поново добили увид у типове карактеристика.

|  |
| --- |
| > str(data)  'data.frame': 253680 obs. of 22 variables:  $ Diabetes\_012 : Factor w/ 3 levels "nema dijabetes",..: 1 1 1 1 1 1 1 1 3 1 ...  $ HighBP : Factor w/ 2 levels "Ne","Da": 2 1 2 2 2 2 2 2 2 1 ...  $ HighChol : Factor w/ 2 levels "Ne","Da": 2 1 2 1 2 2 1 2 2 1 ...  $ CholCheck : Factor w/ 2 levels "Ne","Da": 2 1 2 2 2 2 2 2 2 2 ...  $ BMI : num 40 25 28 27 24 25 30 25 30 24 ...  $ Smoker : Factor w/ 2 levels "Ne","Da": 2 2 1 1 1 2 2 2 2 1 ...  $ Stroke : Factor w/ 2 levels "Ne","Da": 1 1 1 1 1 1 1 1 1 1 ...  $ HeartDiseaseorAttack: Factor w/ 2 levels "Ne","Da": 1 1 1 1 1 1 1 1 2 1 ...  $ PhysActivity : Factor w/ 2 levels "Ne","Da": 1 2 1 2 2 2 1 2 1 1 ...  $ Fruits : Factor w/ 2 levels "Ne","Da": 1 1 2 2 2 2 1 1 2 1 ...  $ Veggies : Factor w/ 2 levels "Ne","Da": 2 1 1 2 2 2 1 2 2 2 ...  $ HvyAlcoholConsump : Factor w/ 2 levels "Ne","Da": 1 1 1 1 1 1 1 1 1 1 ...  $ AnyHealthcare : Factor w/ 2 levels "Ne","Da": 2 1 2 2 2 2 2 2 2 2 ...  $ NoDocbcCost : Factor w/ 2 levels "Ne","Da": 1 2 2 1 1 1 1 1 1 1 ...  $ GenHlth : Ord.factor w/ 5 levels "1"<"2"<"3"<"4"<..: 5 3 5 2 2 2 3 3 5 2 ...  $ MentHlth : num 18 0 30 0 3 0 0 0 30 0 ...  $ PhysHlth : num 15 0 30 0 0 2 14 0 30 0 ...  $ DiffWalk : Factor w/ 2 levels "Ne","Da": 2 1 2 1 1 1 1 2 2 1 ...  $ Sex : Factor w/ 2 levels "zensko","musko": 1 1 1 1 1 2 1 1 1 2 ...  $ Age : num 9 7 9 11 11 10 9 11 9 8 ...  $ Education : Ord.factor w/ 6 levels "bez skole/isao u vrtic"<..: 4 6 4 3 5 6 6 4 5 4 ...  $ Income : Ord.factor w/ 8 levels "<10.000$"<"10.000-14.999$"<..: 3 1 8 6 4 8 7 4 1 3 ... |

Сада можемо прећи на детаљнију анализу и припрему података за моделе.

# Експлоративна анализа података (ЕДА)

У оквиру експлораторне анализе података детаљније смо испитали све увиде добијене у почетној фази информисања о подацима. На основу резултата ЕДА фазе идентификовани су кључни обрасци, потенцијалне неправилности и потребе за додатним трансформацијама. Сви закључци из ове фазе директно су примењени у наредном кораку припреме и трансформације података. На крају поглавља је приказана табела са коначним трансформацијама које треба извршити за сваку карактеристику и табела са међусобним синергијама.

Процес експлоративне анализе смо спровели у два степена:

* Униваријантна анализа,
* Биваријантна анализа,

## Униваријантна анализа

Униваријанта анализа подразумева анализу једне карактеристике (feature-а). Радили смо по систему претпоставки из фазе информисања и доказивали те претпоставке, на крају формирајући закључак.

### Анализа

Претпоставка бр. 1: У карактеристици (feature-у) Diabetes\_012 расподела класа није у равнотежи.

|  |
| --- |
| ggplot(data, aes(x = Diabetes\_012, fill = Diabetes\_012)) +  geom\_bar() +  labs(title = "Distribucija kategorija Diabetes\_012", x = "Kategorije", y = "Broj opservacija") +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |

На основу графика потврђујемо да класе у карактеристици (feature-у) Diabetes\_012 нису балансирани тј. класе нису равнимерно заступљене. Класа „нема дијабетес“ је доминантна, са нешто више од 200 000 опсервација од укупно 253 680, што може изазвати склонност модела да увек предвиђа ову класу. Због тога је потребно применити методе балансирања података, као што су oversampling, undersampling, како би модел боље препознавао и мање заступљене класе. Перформансе модела треба мерити помоћу F1-score и recall по класи, а не само на основу тачности (accuracy), како би се осигурало да све класе буду адекватно предвиђене.

Претпоставка бр. 2: Вредност HighBP за 42,9% опсервација је позитивна тј. да имају повишен крвни притисак.

|  |
| --- |
| ggplot(data, aes(x = HighBP , fill = HighBP )) +  geom\_bar() +  labs(title = "Distribucija kategorija HighBP", subtitle = "Da li ispitanik ima visok pritisak (hipertenziju)?", x = "Kategorije", y = "Broj opservacija") +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |

Графиком потврђујемо претпоставку да је вредност HighBP позитивна код 42,9% опсервација, односно да значајан део испитаника има повишен крвни притисак. Оваква расподела HighBP чини је веома информативном тј. варијанса је висока. Варијанса представља меру распршености података око средње вредности, рачуна се по формули

Максимална варијанса код бинарних променљивих (што је и HighBP ) износи 0.25 и јавља се када су класе подједнако заступљене (p = 0.5). Што је варијанса ближа овој вредности, то је већа информативност променљиве, јер се категорије значајно разликују. Уколико је варијанса ниска то значи да је једна категорија доминантна и да модел на основу промељиве не може ефикасно да учи.

|  |
| --- |
| > varijansa\_highBP = var(as.numeric(data$HighBP))  > cat("Varijansa HighBP: ", varijansa\_highBP, "od maksimmalne moguce 0.25\n")  Varijansa HighBP: 0.2449601 od maksimmalne moguce 0.25 |

У случају променљиве HighBP, добијена варијанса 0.2449601 што је веома близу максималној вредности за бинарне податке.

Претпоставка бр. 3: Вредност HighChol за 42,4% опсервација је позитивна тј. да имају повишен холестерол.

|  |
| --- |
| ggplot(data, aes(x = HighChol , fill = HighChol )) +  geom\_bar() +  labs(title = "Distribucija kategorija HighChol", subtitle = "Da li ispitanik ima visok holesterol?", x = "Kategorije", y = "Broj opservacija") +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |
|  |

Графиком потврђујемо претпоставку. Оваква расподела HighChol чини је веома информативном тј. варијанса је висока. Што потврђује следећи код:

|  |
| --- |
| > cat("Varijansa HighChol: ", varijansa\_highChol, "od maksimmalne moguce 0.25\n")  Varijansa HighChol: 0.2442433 od maksimalne moguće 0.25 |

Претпоставка бр. 4: Укупно 96,27% испитаника је проверило ниво холестерола у последњих пет година, што указује на веома малу варијабилност ове променљиве.

Прво смо визуелно представили дистрибуцију ове карактеристике.

|  |
| --- |
| ggplot(data, aes(x = CholCheck , fill = CholCheck )) +  geom\_bar() +  labs(title = "Distribucija kategorija CholCheck", subtitle = "Da li je ispitanik proverio holesterol poslednjih 5 godina?", x = "Kategorije", y = "Broj opservacija") +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |

Графиком потврђујемо претпоставку. Оваква расподела CholCheck чини је неинформативном тј. варијанса је ниска. Што потврђује следећи код:

|  |
| --- |
| > cat("Varijansa CholCheck: ", varijansa\_CholCheck, "od maksimmalne moguce 0.25\n")  Varijansa CholCheck: 0.03593707 od maksimmalne moguce 0.25 |

Претпоставка бр. 5: У карактеристици БМИ има аутлајера или грешка у уносу, јер је минимална вредност екстремно низак индекс, као и максимална вредност као екстремно висок.

Прво смо визуелно представили дистрибуцију ове карактеристике.

|  |
| --- |
| ggplot(data, aes(x = BMI)) +  geom\_histogram(bins = 50, fill = "#7C4B73FF", color="#381A61FF") +  labs(title = "Distribucija BMI", x = "BMI", y = "Broj opservacija") |
|  |

Преко приказане дистрибуције видимо реп ка већим вредностима БМИ. Свака вредност која је значајно удаљена од већине података представља аутлајер, то не значи нужно грешку, може бити екстрем. Како су на графику тешко уочљиви аутлајери, користили смо прецизнији график боксплот.

|  |
| --- |
|  |

Тачке које се налазе изван кутије представљају екстремне вредности (аутлајере) према IQR методу. Ови аутлајери указују на испитанике са значајно високим или ниским BMI вредностима, које могу бити природни, али и потенцијално резултат грешке у уносу. Број таквих података добићемо коришћењем IQR методе.

|  |
| --- |
| > donja\_granica <- quantile(data$BMI, 0.25) - 1.5 \* IQR(data$BMI)  > gornja\_granica <- quantile(data$BMI, 0.75) + 1.5 \* IQR(data$BMI)  >  > autlajeri\_statisticki <- data %>% filter(BMI < donja\_granica | BMI > gornja\_granica)  > cat("Po IQR metodu statističkih autlajera ima " , nrow(autlajeri\_statisticki), ".Što je", nrow(autlajeri\_statisticki)/nrow(data)\*100,"% ukupnog broja opservacija\n")  Po IQR metodu statističkih autlajera ima 9847. Što je 3.881662 % ukupnog broja opservacija |

На основу IQR методе, идентификовано је 9 847 аутлајера у подацима за променљиву BMI, што чини приближно 3,88 % укупног броја опсервација. Ове вредности могу представљати реалне, али ретке случајеве као и потенцијалне грешке у уносу.

Да би тачно дефинисали прво смо се водили доменским знањем. По таблицама вредности између 12 -14 или 60 - 70 могу бити екстремне, али потенцијално реалне. Тако да смо све вредности преко окарактерисали као доменске аутјлајере.

|  |
| --- |
|  |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | |  | | --- | | > autlajeri\_domenski <- data %>% filter(BMI < 12 | BMI > 70)  > cat("Po domenskom znanju autlajera ima " , nrow(autlajeri\_domenski), ".Što je", nrow(autlajeri\_domenski)/nrow(data)\*100,"% ukupnog broja opservacija\n")  Po domenskom znanju autlajera ima 584 .Što je 0.2302113 % ukupnog broja opservacija | |  | | |  | | --- | |  | | | |

Закључак је да 0.2302113% представља мали проценат од укупног броја података одлучили смо да доменске аутлајере не користимо у моделима.

Претпоставка бр. 6: GenHelt нема равномерну расподелу класа, 75% испитаника сматра своје здравље 1 – одлично, 2 – врло добро, 3 – добро, чиме се остале класе свести у једну.

|  |
| --- |
| ggplot(data, aes(x = GenHlth , fill = GenHlth )) +  geom\_bar() +  labs(title = "Distribucija kategorija GenHlth", subtitle = "Kako ispitanik ocenjuje svoje opste zdravlje?", x = "Kategorije", y = "Broj opservacija") +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |

На основу графика потврђујемо неуравнотежену расподелу. Испитали смо и тачну процентуалну расподелу сваке класе.

|  |
| --- |
| data %>% group\_by(GenHlth) %>%  summarise(  Broj\_opservacija = n(),  Procenat\_opservacija = round(n() / nrow(data) \* 100, 2)  ) |
|  |

На основу процентуалне расподеле класа закључујемо да класе задовољавајуће и лоше спојимо под једну класу која означава лошије здравствено стање.

Претпоставка бр. 7: Средња вредност HeartDiseaseorAttack карактеристике показује да 9,4% опсервација има хронично срчано обољење или је имао срчани удар. Што чини промељиву неуравнотеженом.

На основу доменског знања закључили смо да је овака однос у уззорку реалан и смислен. Како би добили бољи увид у расподелу података користићемо стубичасти график.

|  |
| --- |
| ggplot(data, aes(x = HeartDiseaseorAttack , fill = HeartDiseaseorAttack )) +  geom\_bar() +  labs(title = "Distribucija kategorija HeartDiseaseorAttack ", subtitle = "Da li je ispitanik imao srčani udar ili ima bolest srca?", x = "Kategorije", y = "Broj opservacija") +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |

Графиком потврђујемо претпоставку. Оваква расподела HeartDiseaseorAttack чини је неинформативном тј. варијанса је ниска. Што потврђује следећи код:

|  |
| --- |
| > varijansa\_HeartDiseaseorAttack = var(as.numeric(data$HeartDiseaseorAttack ))  > cat("Varijansa HeartDiseaseorAttack : ", varijansa\_HeartDiseaseorAttack , "od maksimmalne moguce 0.25\n")  Varijansa HeartDiseaseorAttack : 0.085315 od maksimmalne moguce 0.25 |

Претпоставка бр. 8: Променљива Stroke показује да је мождани удар изузетно ретка појава.

|  |
| --- |
| ggplot(data, aes(x = Stroke , fill = Stroke )) +  geom\_bar() +  labs(title = "Distribucija kategorija Stroke ", subtitle = "Da li je ispitanik imao moždani udar?", x = "Kategorije", y = "Broj opservacija") +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |

Графиком потврђујемо претпоставку. Оваква расподела Stroke чини је неинформативном тј. варијанса је ниска. Што потврђује следећи код:

|  |
| --- |
| > varijansa\_Stroke = var(as.numeric(data$Stroke ))  > cat("Varijansa Stroke : ", varijansa\_Stroke , "od maksimmalne moguce 0.25\n")  Varijansa Stroke : 0.03892496 od maksimmalne moguce 0.25 |

Претпоставка бр. 9: MentHlth указује да већина испитаника није имала менталне потешкоће током већег дела последњих 30 дана, при чему је расподела броја дана са стресом концентрисана ка нижим вредностима.

|  |
| --- |
| ggplot(data, aes(x = MentHlth)) +  geom\_histogram(bins = 50, fill = "#7C4B73FF", color="#381A61FF") +  labs(title = "Distribucija MentHlth", x = "MentHlth", y = "Broj opservacija",  subtitle = "Koliko dana u poslednjih mesec dana je ispitanik imao stres?") |
|  |

На основу графика видимо неравномерну расподелу одговора. Закључујемо да треба свести одговоре на класе “да” и “не” (имали су потешкоће са менталним здрављем / нису имали потешкоће са менталним здрављем).

Претпоставка бр. 10: PhysHlth указује да већина испитаника није имала физичке потешкоће током већег дела последњих 30 дана, при чему је расподела броја дана са физичким потешкоћима концентрисана ка нижим вредностима.

|  |
| --- |
| ggplot(data, aes(x = PhysHlth)) +  geom\_histogram(bins = 50, fill = "#7C4B73FF", color="#381A61FF") +  labs(title = "Distribucija PhysHlth", x = "PhysHlth", y = "Broj opservacija",  subtitle = "Koliko dana u poslednjih mesec dana je ispitanik imao fizicke poteskoce?") |
|  |

На основу графика видимо неравномерну расподелу одговора. Закључујемо да треба свести одговоре на класе “да” и “не” (имали су потешкоће са физичким здрављем / нису имали потешкоће са физичким здрављем).

Претпоставка бр. 11: 63.4% опсервација једе воће једном или више пута у току дана, што чини високу варијансу карактеристике Fruits.

|  |
| --- |
| ggplot(data, aes(x = Fruits , fill = Fruits )) +  geom\_bar() +  labs(title = "Distribucija kategorija Fruits", subtitle = "Da li ispitanik jede jednu porciju ili više voća u toku dana?", x = "Kategorije", y = "Broj opservacija") +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |

Графиком потврђујемо претпоставку. Oваква расподела Fruits чини је веома информативном тј. варијанса је висока. Како је Fruits бинарна категоријска карактеристика ово је чини веома утицајном за модел тј. варијанса је висока. Што потврђује следећи код:

|  |
| --- |
| > varijansa\_fruits = var(as.numeric(data$Fruits))  > cat("Varijansa Fruits: ", varijansa\_fruits, "od maksimmalne moguce 0.25\n")  Varijansa Fruits: 0.2319763 od maksimmalne moguce 0.25 |

Претпоставка бр. 12: Расподела категорија није равномерна у промељивој DiffWalk.

|  |
| --- |
| ggplot(data, aes(x = DiffWalk , fill = DiffWalk )) +  geom\_bar() +  labs(title = "Distribucija kategorija DiffWalk ", subtitle = "Da li je ispitanik ima poteškoće u kretanju ili penjanju uz stepenice?", x = "Kategorije", y = "Broj opservacija") +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |

Овaкав визуелни однос не чини њену варијансу великом, али на основу доменског знања закључујемо да је повезана са другим здравственим карактеристикама. У даљем раду ћемо испитивати њен синергијски утицај. Потврда ниже варијабилности:

|  |
| --- |
| > varijansa\_DiffWalk = var(as.numeric(data$DiffWalk ))  > cat("Varijansa DiffWalk : ", varijansa\_DiffWalk , "od maksimmalne moguce 0.25\n")  Varijansa DiffWalk : 0.1399251 od maksimmalne moguce 0.25 |

Претпоставка бр. 13: 81.1% опсервација једе поврће једном или више пута у току дана. Висок проценат нам говори о мањој варијанси Veggies.

|  |
| --- |
| ggplot(data, aes(x = Veggies , fill = Veggies )) +  geom\_bar() +  labs(title = "Distribucija kategorija Veggies", subtitle = "Da li ispitanik jede jednu porciju ili više povrća u toku dana?", x = "Kategorije", y = "Broj opservacija") +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |

Графиком потврђујемо претпоставку. Оваква расподела Veggies чини је неинформативном тј. варијанса је ниска. Што потврђује следећи код:

|  |
| --- |
| > varijansa\_veggies = var(as.numeric(data$Veggies))  > cat("Varijansa Veggies: ", varijansa\_veggies, "od maksimmalne moguce 0.25\n")  Varijansa Veggies: 0.1530182 od maksimmalne moguce 0.25 |

Претпоставка бр. 14: 43,3% опсервација пушач, што је чини утицајном на модел тј. висок проценат нам говори о високој варијанси Smoker.

|  |
| --- |
| ggplot(data, aes(x = Smoker , fill = Smoker )) +  geom\_bar() +  labs(title = "Distribucija kategorija Smoker", subtitle = "Da li је ispitanik konzumirao minimalno 100 cigareta u toku zivota?", x = "Kategorije", y = "Broj opservacija") +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |

Графиком потврђујемо претпоставку. Oваква расподела Smoker чини је веома информативном тј. варијанса је висока. Што потврђује следећи код:

|  |
| --- |
| > varijansa\_smoker = var(as.numeric(data$Smoker))  > cat("Varijansa Smoker: ", varijansa\_smoker, "od maksimmalne moguce 0.25\n")  Varijansa Smoker: 0.2467712 od maksimmalne moguce 0.25 |

Претпоставка бр. 15: 5.6% опсервација спада у алкохоличаре. Расподела је неравномерна тј. орјентисана ка једној класи, па карактеристика HvyAlcoholConsump није утицајна за модел тј. нема високу варијансу.

|  |
| --- |
| ggplot(data, aes(x = HvyAlcoholConsump , fill = HvyAlcoholConsump )) +  geom\_bar() +  labs(title = "Distribucija kategorija HvyAlcoholConsump", subtitle = "Da li је ispitanik konzumirao 7 ili više pića u toku nedelje?", x = "Kategorije", y = "Broj opservacija") +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |

Графиком потврђујемо претпоставку. Oваква расподела HvyAlcoholConsump чини је неинформативном тј. варијанса је ниска. Како је бинарна категоријска карактеристика ово је чини неутицајном за модел тј. варијанса је мала. Што потврђује следећи код:

|  |
| --- |
| > varijansa\_hvyAlcoholConsump = var(as.numeric(data$HvyAlcoholConsump))  > cat("Varijansa HvyAlcoholConsump: ", varijansa\_hvyAlcoholConsump, "od maksimmalne moguce 0.25\n")  Varijansa HvyAlcoholConsump: 0.05303891 od maksimmalne moguce 0.25 |

Претпоставка бр. 16: Већина опсервација је имало рекреативну физичку активност послењих 30 дана (PhysActivity).

|  |
| --- |
| ggplot(data, aes(x = PhysActivity , fill = PhysActivity )) +  geom\_bar() +  labs(title = "Distribucija kategorija PhysActivity ", subtitle = "Da li je ispitanik imao rekreativnu fizičku aktivnost poslednjih 30 dana?", x = "Kategorije", y = "Broj opservacija") +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |

Графиком потврђујемо претпоставку. Oваква расподела PhysActivity чини је неинформативном тј. варијанса је ниска. Што потврђује следећи код:

|  |
| --- |
| > varijansa\_PhysActivity = var(as.numeric(data$PhysActivity ))  > cat("Varijansa PhysActivity : ", varijansa\_PhysActivity , "od maksimmalne moguce 0.25\n")  Varijansa PhysActivity : 0.1841861 od maksimmalne moguce 0.25 |

Претпоставка бр. 17: 95% испитаника има неки вид здравствене заштите. Висок проценат је и знак мале варијабилности тиме AnyHealthcare карактеристика неће бити од велико значајан за модел.

|  |
| --- |
| ggplot(data, aes(x = AnyHealthcare , fill = AnyHealthcare )) +  geom\_bar() +  labs(title = "Distribucija kategorija AnyHealthcare", subtitle = "Da li ispitanik ima neki vid zdravstvene zastite?", x = "Kategorije", y = "Broj opservacija") +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |

Графиком потврђујемо претпоставку. Oваква расподела AnyHealthcare чини је неинформативном тј. варијанса је ниска. Што потврђује следећи код:

|  |
| --- |
| > varijansa\_аnyHealthcare = var(as.numeric(data$AnyHealthcare))  > cat("Varijansa AnyHealthcare: ", varijansa\_аnyHealthcare, "od maksimmalne moguce 0.25\n")  Varijansa AnyHealthcare: 0.04655182 od maksimmalne moguce 0.25 |

Претпоставка бр. 18: 8,41% опсервација у последњих 12 месеци имао је потребу за доктором али није могао да приушти због трошкова. Ова статистика показује NoDocbcCost као ниско варијабилну карактеристику.

|  |
| --- |
| ggplot(data, aes(x = NoDocbcCost , fill = NoDocbcCost )) +  geom\_bar() +  labs(title = "Distribucija kategorija NoDocbcCost", subtitle = "Da li ispitanik imao potrebu za doktorom, a nije moga to da priušti?", x = "Kategorije", y = "Broj opservacija") +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |

Графиком потврђујемо претпоставку. Oваква расподела NoDocbcCost чини је неинформативном тј. варијанса је ниска. Што потврђује следећи код:

|  |
| --- |
| > varijansa\_noDocbcCost = var(as.numeric(data$NoDocbcCost))  > cat("Varijansa NoDocbcCost: ", varijansa\_noDocbcCost, "od maksimmalne moguce 0.25\n")  Varijansa NoDocbcCost: 0.07709147 od maksimmalne moguce 0.25 |

Претпоставка бр. 19: Income као ординална категорија има теденцију као вишим категоријама прихода, уз асиметричну расподелу категорија.

|  |
| --- |
| Дефинисали смо вектор боја како би покрили све категорије на графику. |
| colorS <- c("#88A0DCFF", "#381A61FF", "#7C4B73FF", "#ED968CFF", "#AB3329FF", "#E78429FF", "#F9D14AFF","#73652DFF") |
| График приказа расподеле променљиве Icome |
| ggplot(data, aes(x = Income, fill = Income)) +  geom\_bar() +  labs(  title = "Distribucija kategorija Income",  y = "Broj opservacija",  ) +  geom\_text(stat = "count", aes(label = ..count..), vjust = -0.5) +  scale\_fill\_manual(values = colorS)+  theme(legend.position = "bottom",  axis.text.x = element\_blank()) |
|  |

На основу графика видимо тенденцију података ка вишим категоријама прихода (90385 опсервација има 75000+ приходе). Да би нам однос био јаснији, испитали смо процентуалну заступљеност категорија помоћу функције prop.table(). Пошто ова функција ради над табелама броја појављивања категорија, ординалну променљиву *Income* смо проследили у виду табеле.

|  |
| --- |
| > prop.table(table(data$Income))\*100  <10.000$ 10.000-14.999$ 15.000-19.999$ 20.000-24.999$ 25.000-34.999$ 35.000-49.999$  3.867471 4.644828 6.304793 7.937165 10.203012 14.376380  50.000-74.999$ >=75.000$  17.036818 35.629533 |

Највећи удео испитаника припада највишој приходној категорији (≥75.000$), која обухвата око 35,6% укупних опсервација. Уколико се посматрају заједно категорије са приходима већим од 50.000$, може се закључити да више од половине испитаника (преко 52%) остварује високе приходе (>= 50.000$).

Са аспекта доменског знања, приходи су повезани са бројним аспектима живота појединца, укључујући квалитет живота, приступ здравственој заштити и здраве животне навике. Стога се може претпоставити да променљива *Income* има потенцијалну индиректну повезаност са појавом дијабетеса, због чега је њено даље разматрање и трансформација категорија у мањи број група оправдана у наставку рада.

На основу дистрибуције категорија по количини примања јасно је да постоји асиметрија међу њима,зато ћемо у даљим корацима направити нову варијаблу на основу ове и доменског знања о категоријама становништва по основу годишњих примања где ће ових 8 категорија бити барем преполовљено,што ће допринети умањењу асиметрије међу њима.

Претпоставка бр. 20: Категорија која одређује пол испитаника (Sex) има равномерну расподелу што чини утицајном за модел.

|  |
| --- |
| ggplot(data, aes(x = Sex , fill = Sex )) +  geom\_bar() +  labs(title = "Distribucija kategorija Sex", subtitle = "Kol pola je ispitanik?", x = "Kategorije", y = "Broj opservacija") +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |

Графиком потврђујемо претпоставку. Oваква расподела Sex чини је веома информативном тј. варијанса је висока. Што потврђује следећи код:

|  |
| --- |
| > varijansa\_sex = var(as.numeric(data$Sex))  > cat("Varijansa Sex: ", varijansa\_sex, "od maksimmalne moguce 0.25\n")  Varijansa Sex: 0.2464419 od maksimmalne moguce 0.25 |

Претпоставка бр. 21: Расподела категорије Age је симетрична, са распоном вредности од 1 до 13.

|  |
| --- |
| ggplot(data, aes(x = Age )) +  geom\_histogram(bins = 50, fill = "#7C4B73FF", color="#381A61FF") +  labs(title = "Distribucija Age ", x = "Age ", y = "Broj opservacija",  subtitle = "Starost ispitanika") |
|  |

Age је нумеричка променљива (показано у делу информисања о скупу података), па смо за приказ дистрибуције исте користили хистограм. На основу овог графика уочавамо симетрију у расподели, чиме потврђујемо претпоставку.

Као што је речено у поглављу информисања о структури података вредности су целобројне у распону од 1 до 13, на основу доменског знања закључујемо да се не ради о тачном броју година већ о старосној групи. Све категорије су довољно заступљене што чини расподелу стабилном.

Обзиром на наведено, а у циљу примене у моделу, потребно је извршити трансформацију ове променљиве формирањем ординалних категорија које ће омогућити уравнотеженију и интерпретабилнију расподелу. Пре категоризације, у биваријантној анализи се могу додатно испитати интервали, како би се потенцијално смањио број група и поједноставила интерпретација модела.

На основу хистограма види се да ће постојати доста категорија,тј 13, што је много, а притом иако је дистрибуција међу вредностима задовољавајућа, потребно је напоменути да би по модел било боље дистрибуцију мало побољшати,а притом је потребно и смањити број категорија, што ће се према даљим истраживањима и утврдити на колико категорија свести ову варијаблу, а да се притом побољша дистрибуција.

Претпоставка бр. 22: Већина вредности у категорији Education има тенденцију ка вишим категоријама образовања, што чини расподелу неуравнотеженом и губи значајност за модел.

|  |
| --- |
| ggplot(data, aes(x = Education, fill = Education)) +  geom\_bar() +  labs(  title = "Distribucija kategorija Education",  y = "Broj opservacija",  ) +  geom\_text(stat = "count", aes(label = ..count..), vjust = -0.5) +  scale\_fill\_manual(values = colors)+  theme(legend.position = "bottom",  axis.text.x = element\_blank()) |
|  |

Графиком смо потврдили претпоставку. Закључили смо да можемо прегруписати постојеће категорије како би добили равномернију расподелу.

### Табела закључака

|  |  |
| --- | --- |
| Diabetes\_012 | Потребно применити методе балансирања података, као што су oversampling (SMOTE), undersampling. |
| Параметри здравља | |
| HighBP | Варијанса износи 0,2449 од максимални 0,25 што је чини веома информативном. |
| HighChol | Варијанса износи 0,2424 од максимални 0,25 што је чини веома информативном. |
| CholCheck | Варијанса износи 0.03593707 од максимални 0,25 што је не чини информативном. |
| BMI | Уочавају се екстремне вредности од којих су <12 и >70 окарактерисани као доменски аулајери и чине 0.2302113% опсервација. Утицај аутлејера треба испитати биваријантном анализом. |
| GenHlth | На основу процентуалне расподеле класа закључујемо да класе задовољавајуће и лоше спојимо под једну класу која означава лошије здравствено стање. |
| Параметри стања | |
| HeartDiseaseorAttack | На основу средње вредности закључујемо да 9,4% опсервација има хронично срчано обољење или је имао срчани удар. На основу доменског знања закључили смо да је однос реалан и смислен. Варијанса износи 0.085315 од максимални 0,25 што је не чини информативном. |
| Stroke | Варијанса износи 0.03892496 од максимални 0,25 што је не чини информативном. Променљива и даље има потенцијалну релевантност због добро познате повезаности можданог удара са метаболичким поремећајима, укључујући дијабетес. |
| MentHlth | Највећа концентација одговора је за вредност 0 тј, да испитаник није имао стрес у последњих 30 дана, стварајући неуравнотежену расподелу. Закључујемо да треба испитати потенцијалне категорије. |
| PhysHlth | Највећа концентација одговора је за вредност 0 тј, да испитаник није имао физичких потешкоћа у последњих 30 дана, стварајући неуравнотежену расподелу. Закључујемо да треба испитати потенцијалне категорије. |
| DiffWalk | Варијанса износи 0.03892496 од максимални 0,25 што је не чини информативном. У даљем раду ћемо испитивати њен синергијски утицај. |
| Параметри исхране | |
| Fruits | Варијанса износи 0.2319763 од максимални 0,25 што је чини информативном. |
| Veggies | Варијанса износи 0.1530182 од максимални 0,25 што је не чини информативном. |
| Параметри животних навика |  |
| Smoker | Варијанса износи 0.2467712 од максимални 0,25 што је чини утицајном. |
| HvyAlcoholConsump | Варијанса износи 0.05303891 од максимални 0,25 што је не чини информативном. |
| PhysActivity | Велика већина опсервација је имало рекреативну физички активност последњих 30 дана. Варијанса износи 0.1841861 од максимални 0,25 што је не чини много информативном. На основу доменског знања сматрамо да ова карактеристика има везе са потешкоћава у креању DiffWalk. |
| Социјално-економски параметри | |
| AnyHealthcare | Варијанса износи 0.04655182 од максимални 0,25 што је не чини утицајном. |
| NoDocbcCost | Варијанса износи 0.07709147 од максимални 0,25 што је не чини информативном. |
| Income | Уочава се јасна тенденција ка вишим приходима (преко 52%) остварује високе приходе. Са аспекта доменског знања, приходи су повезани са бројним аспектима живота појединца, укључујући квалитет живота, приступ здравственој заштити и здраве животне навике, због чега је њено даље разматрање и трансформација категорија у мањи број група оправдана у наставку рада. |
| Демографски параметри | |
| Sex | Варијанса износи 0.2464419 од максимални 0,25 што је чини информативном. |
| Age | Расподела категорије Age је симетрична, са распоном вредности од 1 до 13 на основу доменског знања закључујемо да се не ради о тачном броју година већ о старосној групи. Потребна је категоризација. |
| Education | Већина вредности у категорији Education има тенденцију ка вишим категоријама образовања, што чини расподелу неуравнотеженом и губи значајност за модел, с тога прегруписавање је оправдано. |

## Биваријантна анализа

Биваријантна анализа обухвата визуелну и статистичку анализу две карактеристике у циљу добијања закључака и њиховом односу и зависностима. Испитивали смо постојање образаца, повезаности, али и јачине те повезаности.

У зависности од типа карактеристика које се међусобно анализирају, визуелно биваријантном анализом обухватили смо:

* однос категоријске и нумеричке преко бокс плот дијаграма и density дијаграма,
* однос категоријске и категоријске карактеристике преко стубичастог дијаграма и топлотне мапе.

Добијене претпоставке смо тестирали статистички, опет у зависности од типа типа карактеристика које се међусобно анализирају:

* однос категоријске (са више од две категорије) и нумеричке преко анова теста и post-hoc анализе ми смо изабрали Tukey HSD тест,
* однос категоријске и категоријске карактеристике преко χ² теста и Крамеровог коефицијента.

Статистички тестови имплементирани су у функцијама anova\_test, tukey\_fun, chi\_sq\_test, cramer\_v.

### Имплементација метода статистичких тестова

#### АНОВА тест (Analysis of Variance)

АНОВА тест се користи за испитивање да ли постоји статистички значајна разлика у средњим вредностима нумеричке променљиве између три или више категорија категоријске променљиве. Уколико бар једна категорија задовољава разлику тест показује статистички значај, али не показује између којих категорија постоји разлика. Зато се приступа post-hoc анализама. Ако у резултатима добијемо да је „p-value“ < 0.05, бар једна група се значајно разликује од осталих.

Имплементација:

|  |
| --- |
| anova\_test <- function(numericka\_var, grupna\_var) {  model <- aov(numericka\_var ~ as.factor(grupna\_var))  return(summary(model))  } |

#### Tukey HSD тест (Honestly Significant Difference)

Tukey HSD тест представља накнадни тест који се примењује након статистички значајног АНОВА теста. Обим тестом се омогућава поређење свих парова категорија и утврђивање између који парова постоји статистички значајна разлика у средњим вредностима. Тај резултат можемо видети у табели у првој колони која нам говори за колико јединица је разлика између категорија.

Имплементација:

|  |
| --- |
| tukey\_fun <- function(numericka\_var, grupna\_var) {  model <- aov(numericka\_var ~ as.factor(grupna\_var))  rezultat <- TukeyHSD(model)  return(rezultat)  } |

#### χ² тест независности

χ² тест независности се користи за испитивање да ли постоји статистички значајна повезаност између две категоријске променљиве. Тестира се хипотеза да су посматране променљиве независне једна од друге, на основу упоређивања очекиваних и посматраних фреквенција. Ако у резултатима добијемо да је „p-value“ < 0.05, постоји статистички значајна повезаност, а све преко се сматра независним карактеристикама.

Имплементација:

|  |
| --- |
| chi\_sq\_test <- function(var1, var2) {  tabela <- table(var1, var2)  rezultat <- chisq.test(tabela)  return(rezultat)  } |

#### Крамеров коефицијент

Крамеров коефицијент представља меру јачине повезаности између две категоријске променљиве и користи се након χ² теста, у случају када је утврђена статистички значајна повезаност. Вредности коефицијента су од 0 до 1, где тежња ка 1 указује на јаку повезаност.

Имплементација:

|  |
| --- |
| cramer\_v <- function(var1, var2) {  tabela <- table(var1, var2)  chi2 <- chisq.test(tabela)$statistic  n <- sum(tabela)  phi2 <- chi2 / n  r <- nrow(tabela)  k <- ncol(tabela)  v <- sqrt(phi2 / min(r - 1, k - 1))  return(as.numeric(v))  } |

#### Пирсонов коефицијент корелације

Коефицијент нам квантификује линеарност између две нумеричке варијабле, реалном вредношћу која се креће између -1 и 1. У случају да је на коефицијент на -1 имамо негативну линеарност међу варијаблама,како се креће ка 0 она је све слабија, на 0 ће бити непостојећа, док ће како се креће ка 1 постајати све јача и биће позитивна линеарност.

Имплементација:

|  |
| --- |
| pearson\_funkcija <- function(x, y) {  if(length(x) != length(y)) stop("Vektori moraju biti iste dužine")  mean\_x <- mean(x)  mean\_y <- mean(y)  brojilac <- sum((x - mean\_x) \* (y - mean\_y))  imenilac <- sqrt(sum((x - mean\_x)^2) \* sum((y - mean\_y)^2))  return(brojilac / imenilac)  } |

### Однос са циљаном карактеристиком Diabetes\_012

У овом поглављу испитана је повезаност сваке карактеристике са циљном променљивом Diabetes\_012. Резултати ових анализа пружају увид у то које карактеристике имају самостални утицај на стање дијабетеса.

#### HighBP vs Diabetes\_012

Анализирамо учесталост испитаника са високим крвним притиском (HighBP) у односу на резултате дијабетичког стања (Diabetes\_012). Тип карактеристике HighBP је бинарна категоријска променљива са категоријама Да (има повишен крвни притисак), Не (нема повишен крвни притисак). Дистрибуцију повишеног крвног притиска унутар сваке класе дијабетеса приказали смо стубичастим дијаграмом и топлотним дијаграмом.

|  |  |
| --- | --- |
| ggplot(data, aes(x = Diabetes\_012, fill = HighBP)) +  geom\_bar(position = "dodge") +  labs(  title = "Distribucija kategorija HighBP po kategorijama Diabetes\_012",  y = "Broj opservacija",  fill = "Povišen krvni pritisak"  ) + scale\_fill\_paletteer\_d("MetBrewer::Archambault") +  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  )  data %>%  count(Diabetes\_012, HighBP) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = HighBP, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija HighBP po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "HighBP",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) | |
|  |  |

Са стубичастог графика визуелно је јасно да код особа без дијабетеса доминирају испитаници који немају повишен крвни притисак, док код особа са неким стањем дијабетеса доминирају они са повишеним крвним притиском. Ако обратимо пажњу на градијент топлотног дијаграма, уочавамо да има променљивог односа у распореду који потврђује образац уочен на стубичастом дијаграму. Указујући да је повишен крвни притисак значајно повезан са статусом дијабетеса и има аналитичку вредност у односу на циљану променљиву Diabetes\_012.

Претпоставку ћемо испитати статистички преко Крамеровог коефицијента и χ² теста, наравно обе методе смо већ предефинисали и имплементирали па их онда једноставно позивамо.

|  |
| --- |
| chi\_sq\_test(data$HighBP, data$Diabetes\_012)  cramer\_v(data$HighBP, data$Diabetes\_012) |
|  |

χ² тест независности показује статистички значајну повезаност између променљивих Diabetes\_012 и HighBP χ² = 18795, a p-value < 0,00000000000000022 што је доста испод границе статистичке значајности од 0,05. Јачина ове повезаности износи 0,27 (Крамеров коефицијент) што је слаба јачина везе али веома близу границе са умереном. Овај самостални утицај нећемо разматрати, али због постојања статистички значајне везе (χ² тест) узећемо у разматрање мултиваријантне анализе.

#### HighChol vs Diabetes\_012

Анализирамо учесталост испитаника са високим холестеролом (HighChol) у односу на резултате дијабетичког стања (Diabetes\_012). Тип карактеристике HighChol је бинарна категоријска променљива са категоријама Да (има повишен холестерол), Не (нема повишен холестерол). Дистрибуцију повишеног холестерола унутар сваке класе дијабетеса приказали смо стубичастим дијаграмом i топлотним дијаграмом.

|  |  |
| --- | --- |
| ggplot(data, aes(x = Diabetes\_012, fill = HighChol)) +  geom\_bar(position = "dodge") +  labs(  title = "Distribucija kategorija HighChol po kategorijama Diabetes\_012",  y = "Broj opservacija",  fill = "Povišen holesterol"  ) + scale\_fill\_paletteer\_d("MetBrewer::Archambault") +  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  )+  theme(legend.position="bottom")  data %>%  count(Diabetes\_012, HighChol) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = HighChol, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija HighChol po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "HighChol",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) | |
|  |  |

Са стубичастог графика визуелно је јасно да код особа без дијабетеса доминирају испитаници који немају повишен холестерол (132 673 испитаника), док код особа са неким стањем дијабетеса доминирају они са повишеним холестеролом. Ако обратимо пажњу на градијент топлотног дијаграма, уочавамо да има променљивог односа у распореду који потврђује образац уочен на стубичастом дијаграму. Указујући да је повишен крвни притисак значајно повезан са статусом дијабетеса и има аналитичку вредност у односу на циљану променљиву Diabetes\_012.

Претпоставку ћемо испитати статистички преко Крамеровог коефицијента и χ² теста, наравно обе методе смо већ предефинисали и имплементирали па их онда једноставно позивамо.

|  |
| --- |
| chi\_sq\_test(data$HighChol, data$Diabetes\_012)  cramer\_v(data$HighChol, data$Diabetes\_012) |
|  |

χ² тест независности показује статистички значајну повезаност између променљивих Diabetes\_012 и HighChol χ² = 11259, a p-value < 0,00000000000000022 што је доста испод границе статистичке значајности од 0,05. Јачина ове повезаности износи 0,21 што је слаба повезаност. Овај самостални утицај нећемо разматрати, али због постојања статистички значајне везе (χ² тест) узећемо у разматрање мултиваријантне анализе.

#### CholCheck vs Diabetes\_012

Анализирамо учесталост испитаника са провером холестерола у последњих 5 година (CholCheck) у односу на резултате дијабетичког стања (Diabetes\_012). Тип карактеристике CholCheck је бинарна категоријска променљива са категоријама Да (има проверен холестерол), Не (нема проверен холестерол). Дистрибуцију провере холестерола унутар сваке класе дијабетеса приказали смо стубичастим дијаграмом i топлотним дијаграмом.

Претпоставку ћемо испитати статистички преко Крамеровог коефицијента и χ² теста, наравно обе методе смо већ предефинисали и имплементирали па их онда једноставно позивамо.

|  |  |
| --- | --- |
| ggplot(data, aes(x = Diabetes\_012, fill = CholCheck)) +  geom\_bar(position = "dodge") +  labs(  title = "Distribucija kategorija CholCheck po kategorijama Diabetes\_012",  y = "Broj opservacija",  fill = "Proveren holesterol poslednjih 5 godina"  ) + scale\_fill\_paletteer\_d("MetBrewer::Archambault") +  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  ) +  theme(legend.position="bottom")  data %>%  count(Diabetes\_012, CholCheck) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = CholCheck, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija CholCheck po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "CholCheck",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) | |
|  |  |

Са стубичастог графика визуелно је јасно да код особа без дијабетеса доминирају испитаници који су проверили холестерол последњих 5 година, код особа са неким стањем дијабетеса такође доминирају они који су проверили холестерол последњих 5 година. Градијент на топлотним дијаграму показује променљив однос, али распоред не открива јасан образац међу класама дијабетеса. То указује да провера холестерола није самостално повезана са статусом дијабетеса и нема аналитичку вредност у односу на циљану променљиву Diabetes\_012.

Претпоставку ћемо испитати статистички преко Крамеровог коефицијента и χ² теста, наравно обе методе смо већ предефинисали и имплементирали па их онда једноставно позивамо.

|  |
| --- |
| chi\_sq\_test(data$CholCheck, data$Diabetes\_012)  cramer\_v(data$CholCheck, data$Diabetes\_012) |
|  |

χ² тест независности показује статистички значајну повезаност између променљивих Diabetes\_012 i CholCheck χ² = 1173, a p-value < 0,00000000000000022 што је доста испод границе статистичке значајности од 0,05. Међутим, графички преглед не открива јасан образац, а Крамеров коефицијент V = 0.068 указује на веома слабу повезаност. Стога се може закључити да променљива *CholCheck* нема практичан самостални ефекат на статус дијабетеса. За мултиваријантну анализу, ова променљива ће бити разматрана само ако други фактори указују на могућу интеракцију или контролни ефекат.

#### BMI vs Diabetes\_012

Анализирамо расподелу вредности индекса телесне масе (BMI) у односу на резултате дијабетичког стања (Diabetes\_012). Променљива BMI је нумеричка карактеристика, која описује однос телесне масе и висине испитаника. Да бисмо испитали везу између BMI и дијабетичког статуса, анализирали смо дистрибуцију BMI вредности унутар сваке класе дијабетеса, користећи боксплот дијаграм, ради поређења медијана, интерквартилних опсега и присуства екстремних вредности и density дијаграм, ради увида у облик расподеле BMI у свакој групи.

|  |  |
| --- | --- |
| vrenostiBoxPlot <- data%>%  group\_by(Diabetes\_012)%>%  summarise(  Q1 = quantile(BMI, 0.25),  median = median(BMI),  Q3 = quantile(BMI, 0.75),  )%>%  pivot\_longer(  cols = c(Q1, median, Q3)  )  ggplot(data, aes(x = Diabetes\_012, y = BMI, fill = Diabetes\_012))+  geom\_boxplot()+  geom\_point(  data = vrenostiBoxPlot,  aes(x = Diabetes\_012, y = value),  shape = 22,  size = 5,  fill = "white",  color = "#E78429FF"  )+  geom\_text(  data = vrenostiBoxPlot,  aes(x = Diabetes\_012, y = value, label = round(value, 1)),  size = 3  )+  labs(  title = "Raspodela BMI po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "BMI"  )+  scale\_fill\_paletteer\_d("MetBrewer::Archambault")+  theme(legend.position = "none")  ggplot(data, aes(x = BMI, fill = Diabetes\_012)) +  geom\_density(alpha = 0.6) +  labs(  title = "Gustina raspodele BMI po kategorijama Diabetes\_012",  x = "BMI",  y = "Gustina",  fill = "Diabetes\_012"  ) +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") | |
|  |  |

На основу бокс плота уочавамо јасан тренд раста БМИ, код испитаника без дијабетеса прва половина података се креће око 24 до 27, код предијабетеса од 26 до 30, а са дијабетесом још више од 27 до 31. Ово указује на позитивну повезаност између БМИ и дијабетеса. Код дијабетичара је расподела шире растегнута што указује на већу хетерогеност телесне масе.

На другом графику можемо видети да крива за „нема дијабетес” је концентрисана око нижих BMI вредности. Предијабетес и дијабетес је померен удесно у односу на здраве. Иако је помак вредности приметан графици се у великој мери преклапају у опсегу вредности, дакле нема јасног раздвајања, што БМИ чини визуелно слабим предиктором. Тумачење је приказано визуелно на следећем графику:

|  |
| --- |
|  |
| ggplot(data, aes(x = BMI, fill = Diabetes\_012)) +  geom\_density(alpha = 0.6) +  labs(  title = "Gustina raspodele BMI po kategorijama Diabetes\_012",  x = "BMI",  y = "Gustina",  fill = "Diabetes\_012"  ) +  scale\_fill\_paletteer\_d("MetBrewer::Archambault")+  geom\_vline(xintercept = c(25, 30), linetype = "dashed", color = "black", size = 1) |

Ако испратимо криву групе „нема дијабетес“ видимо доста осцилација што указује на неравномерну расподелу, дакле у овој групи постоје подгрупе испитаника са различитим БМИ. Предијабетес и дијабетес имају један доминантан врх.

Као што је утврђено у униваријантној анализи, статистички аутлајери су присутни у расподели променљиве BMI. На боксплот дијаграму уочава се да су аутлајери заступљени у свим категоријама дијабетеса. Како би се проценио њихов утицај на категорије циљане променљиве, спроведена је даља анализа. Поред оригиналне визуелне репрезентације расподеле, анализирана је и расподела скупа података без доменских и додатно без статистичких аутлајера, дефинисаних у оквиру униваријантне анализе.

Формирани скупови потребни за визуелну репрезентацију:

|  |
| --- |
| BMI\_bezdomeskihAutlajera = data %>%  filter(BMI >= 12 & BMI <= 70)  BMI\_bezStatistickihAutlajera = data %>%  filter(BMI >= donja\_granica & BMI <= gornja\_granica) |
| donja\_granica и gornja\_granica дефинисане су у униваријантној анализи следећим кодом: |

За анализу утицаја доменских аутлејера формирали смо збирни скуп да подацима који имају аутлејере, како би омогућили facet wrap:

|  |
| --- |
| BMI\_bezdomenskih\_sabrani <- bind\_rows(  data %>%  select(BMI, Diabetes\_012) %>%  mutate(Skup = "originalni skup"),    BMI\_bezdomeskihAutlajera %>%  select(BMI, Diabetes\_012) %>%  mutate(Skup = "bez domenskih autlajera"),  ) |

Затим смо нацртали графике са исписаним вредностима медијана, првог и трећег кватила како би јасније уочили да ли је дошло до утицаја аутлајера:

|  |
| --- |
| statistickeVrednostiBezDomenskih = BMI\_bezdomenskih\_sabrani %>%  group\_by(Skup, Diabetes\_012) %>%  summarise(  Q1 = quantile(BMI, 0.25),  Median = quantile(BMI, 0.5),  Q3 = quantile(BMI, 0.75)  )  ggplot(BMI\_bezdomenskih\_sabrani, aes(x = Diabetes\_012, y = BMI, fill = Diabetes\_012)) +  geom\_boxplot(outlier.alpha = 0.3) +  facet\_wrap(~ Skup)+  geom\_text(  data = statistickeVrednostiBezDomenskih,  aes(x = Diabetes\_012, y = Q1, label = round(Q1, 1)),  vjust = 1.5, color = "blue", size = 3,  ) +  geom\_text(  data = statistickeVrednostiBezDomenskih,  aes(x = Diabetes\_012, y = Median, label = round(Median, 1)),  vjust = -0.5, color = "red", size = 3  ) +  geom\_text(  data = statistickeVrednostiBezDomenskih,  aes(x = Diabetes\_012, y = Q3, label = round(Q3, 1)),  vjust = -1.5, color = "darkgreen", size = 3  ) +  labs(  title = "Uporedjivanje uticaja domenskih autlejera karakteristike BMI na Diabetes\_012",  y = "BMI",  x = "Dijabetes"  )+  theme(legend.position = "none")+  scale\_fill\_paletteer\_d("MetBrewer::Archambault") |
|  |

Као што видимо по ивицама кутија, али и по нумеричким показатељима може се уочити да доменски аутлајери немају никакви утицај на категорије дијабетеса. Додатно са доменског аспекта не представљају реалне вредности па нису репрезентативане за већину популације, зато закључујемо да их у поглављу чишћења података уклонимо.

За анализу утицаја статистичких аутлејера формирали смо збирни скуп са подацима који имају аутлејере, како би омогућили facet wrap:

|  |
| --- |
| BMI\_bezstatistickih\_sabrani <- bind\_rows(  data %>%  select(BMI, Diabetes\_012) %>%  mutate(Skup = "originalni skup"),    BMI\_bezStatistickihAutlajera %>%  select(BMI, Diabetes\_012) %>%  mutate(Skup = "bez statistickih autlajera"),  ) |

Затим смо нацртали графике са исписаним вредностима медијана, првог и трећег кватила како би јасније уочили да ли је дошло до утицаја аутлајера:

|  |
| --- |
| statistickeVrednostiBezStatistickih <- BMI\_bezstatistickih\_sabrani %>%  group\_by(Skup, Diabetes\_012) %>%  summarise(  Q1 = quantile(BMI, 0.25),  Median = quantile(BMI, 0.5),  Q3 = quantile(BMI, 0.75)  )  ggplot(BMI\_bezstatistickih\_sabrani, aes(x = Diabetes\_012, y = BMI, fill = Diabetes\_012)) +  geom\_boxplot(outlier.alpha = 0.3) +  facet\_wrap(~ Skup)+  geom\_text(  data = statistickeVrednostiBezStatistickih,  aes(x = Diabetes\_012, y = Q1, label = round(Q1, 1)),  vjust = 1.5, color = "blue", size = 3,  ) +  geom\_text(  data = statistickeVrednostiBezStatistickih,  aes(x = Diabetes\_012, y = Median, label = round(Median, 1)),  vjust = -0.5, color = "red", size = 3  ) +  geom\_text(  data = statistickeVrednostiBezStatistickih,  aes(x = Diabetes\_012, y = Q3, label = round(Q3, 1)),  vjust = -1.5, color = "darkgreen", size = 3  ) +  labs(  title = "Uporedjivanje uticaja statistickih autlejera karakteristike BMI na Diabetes\_012",  y = "BMI",  x = "Dijabetes"  )+  theme(legend.position = "none")+  scale\_fill\_paletteer\_d("MetBrewer::Archambault") |
|  |

Са упоредних дијаграма може се уочити да уклањање статистичких аутлајера не доводи до значајних промена медијана и интерквартилних распона, при чему је однос између вредности BMI и категорија дијабетеса задржан. Статистички аутлајери, иако екстремни са статистичког становишта, представљају реалне вредности гојазности у доменском смислу. С обзиром на то да њихово уклањање не утиче на анализирани однос, донета је одлука да се задрже у даљој анализи.

На основу претходних тумачења претпостављамо да је БМИ има статистички значајан, али нејасан и самосталан ефекат, што ћемо статистички потврдити са анова тесту. Анова је већ имплементирана фукција раније у коду, па смо је само позвали.

|  |
| --- |
| anova\_test(data$BMI, data$Diabetes\_012) |
|  |

F value = 6768, Pr(>F) < 2e-16 означава да је п вредност практично 0 тј. постоји статистички значајна разлика између просечног БМИ у бар две категорије Diabetes\_012. Уз то стоји и \*\*\* што нам потврђује визуелну претпоставку да BMI има утицај на категорије дијабетеса.

Ипак, сам анова тест не указује које тачно групе се разликују. Због тога смо користили post-hoc тест (Tukey HSD), који нам омогућава да за све парове утврдимо између којих група постоји статистички значајна разлика. Методу за тест смо већ предефинисали и имплементирали па је онда једноставно позивамо:

|  |
| --- |
| tukey\_fun(data$BMI, data$Diabetes\_012) |
|  |

На основу резулатата видимо предијабетес има значајно виши BMI од здравих (diff = 2.98), дијабетес има још већи BMI у поређењу са здравима (diff = 4.20) и између предијабетеса и дијабетеса постоји значајна разлика (diff = 1.22), иако мања него у односу на здраве.

#### Smoker vs Diabetes\_012

Анализирамо учесталост испитаника који су пушачи (Smoker) у односу на резултате дијабетичког стања (Diabetes\_012). Тип карактеристике Smoker је бинарна категоријска променљива са категоријама Да (да испитаник је пушач), Не (испитаник није пушач). Дистрибуцију зависности пушења унутар сваке класе дијабетеса приказали смо стубичастим дијаграмом i топлотним дијаграмом.

|  |  |
| --- | --- |
| ggplot(data, aes(x = Diabetes\_012, fill = Smoker)) +  geom\_bar(position = "dodge") +  labs(  title = "Distribucija kategorija Smoker po kategorijama Diabetes\_012",  y = "Broj opservacija",  fill = "Pušač"  ) + scale\_fill\_paletteer\_d("MetBrewer::Archambault")+  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  )+  theme(legend.position = "bottom")  data %>%  count(Diabetes\_012, Smoker) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = Smoker, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija Smoker po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "Smoker",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) | |
|  |  |

Са стубичастог графика визуелно је јасно да код особа без дијабетеса доминирају не пушачи, док код особа са дијабетесом има више пушача. Дакле образац постоји, међутим ако обратимо пажњу на градијент топлотног дијаграма, уочавамо да је јако слаб за категорије предијабетеса и дијабетеса. Претпостављамо да пушење није значајно повезано са статусом дијабетеса и нема снажну или самосталну аналитичку вредност у односу на циљану променљиву Diabetes\_012.

Претпоставку ћемо испитати статистички преко Крамеровог коефицијента и χ² теста, наравно обе методе смо већ предефинисали и имплементирали па их онда једноставно позивамо

|  |
| --- |
| chi\_sq\_test(data$Smoker, data$Diabetes\_012)  cramer\_v(data$Smoker, data$Diabetes\_012) |
|  |

χ² тест независности показује статистички значајну повезаност између променљивих Diabetes\_012 i χ² = 1010, a p-value < 0,00000000000000022 што је доста испод границе статистичке значајности од 0,05. Међутим, графички преглед показује слабу повезаност (слаб градијент), што потврђује Крамеров коефицијент V = 0.068. Стога се може закључити да променљива Smoker нема практичан самостални ефекат на статус дијабетеса. За мултиваријантну анализу, ова променљива ће бити разматрана само ако други фактори указују на могућу интеракцију или контролни ефекат.

#### Stroke vs Diabetes\_012

Анализирамо појаву можданог удара (Stroke) у односу на резултате дијабетичког стања (Diabetes\_012). Тип карактеристике Stroke је бинарна категоријска променљива са категоријама Да (да испитаник је имао мождани удар), Не (испитаник није имао мождани удар). Дистрибуцију зависности појаве можданог удара унутар сваке класе дијабетеса приказали смо стубичастим дијаграмом i топлотним дијаграмом.

|  |  |
| --- | --- |
| ggplot(data, aes(x = Diabetes\_012, fill = Stroke)) +  geom\_bar(position = "dodge") +  labs(  title = "Distribucija kategorija Stroke po kategorijama Diabetes\_012",  y = "Broj opservacija",  fill = "Moždani udar"  ) + scale\_fill\_paletteer\_d("MetBrewer::Archambault")+  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  )+  theme(legend.position = "bottom")  data %>%  count(Diabetes\_012, Stroke) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = Stroke, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija Stroke po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "Stroke",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) | |
|  |  |

Са стубичастог графика визуелно је јасно да код особа без дијабетеса доминирају испитаници који нису имали мождани удар, код особа са неким стањем дијабетеса такође доминирају они који нису имали мождани удар. Градијент на топлотним дијаграму показује променљив однос, али распоред не открива јасан образац међу класама дијабетеса. То указује да мождани удар није самостално повезана са статусом дијабетеса и нема аналитичку вредност у односу на циљану променљиву Diabetes\_012.

Претпоставку ћемо испитати статистички преко Крамеровог коефицијента и χ² теста, наравно обе методе смо већ предефинисали и имплементирали па их онда једноставно позивамо.

|  |
| --- |
| chi\_sq\_test(data$ Stroke, data$Diabetes\_012)  cramer\_v(data$ Stroke, data$Diabetes\_012) |
|  |

χ² тест независности показује статистички значајну повезаност између променљивих Diabetes\_012 i Stroke χ² = 1173, a p-value < 0,00000000000000022 што је доста испод границе статистичке значајности од 0,05. Међутим, графички преглед не открива јасан образац, а Крамеров коефицијент V = 0.107 указује на веома слабу повезаност. Стога се може закључити да променљива *Stroke* нема практичан самостални ефекат на статус дијабетеса. За мултиваријантну анализу, ова променљива ће бити разматрана само ако други фактори указују на могућу интеракцију или контролни ефекат.

#### HeartDiseaseorAttack vs Diabetes\_012

Анализирамо појаву срчаног удара или хипертензије (HeartDiseaseorAttack) у односу на резултате дијабетичког стања (Diabetes\_012). Тип карактеристике HeartDiseaseorAttack је бинарна категоријска променљива са категоријама Да (да испитаник је имао срчани удар или хипертензију), Не (испитаник није имао срчани удар или хипертензију). Дистрибуцију зависности појаве срчаног удара или хипертензије унутар сваке класе дијабетеса приказали смо стубичастим дијаграмом i топлотним дијаграмом.

|  |  |
| --- | --- |
| ggplot(data, aes(x = Diabetes\_012, fill = HeartDiseaseorAttack)) +  geom\_bar(position = "dodge") +  labs(  title = "Distribucija kategorija HeartDiseaseorAttack po kategorijama Diabetes\_012",  y = "Broj opservacija",  fill = "Bolest srca/Srčani udar"  ) + scale\_fill\_paletteer\_d("MetBrewer::Archambault")+  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  )+  theme(legend.position = "bottom")  data %>%  count(Diabetes\_012, HeartDiseaseorAttack) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = HeartDiseaseorAttack, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija HeartDiseaseorAttack po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "HeartDiseaseorAttack",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) | |
|  |  |

Са стубичастог графика визуелно је јасно да код особа без дијабетеса доминирају испитаници који нису имали срчани удар или хипертензију, код особа са неким стањем дијабетеса такође доминирају они који нису имали срчани удар или хипертензију. Градијент на топлотним дијаграму показује променљив однос, али распоред не открива јасан образац међу класама дијабетеса. То указује да срчани удар или хипертензија није самостално повезана са статусом дијабетеса и нема аналитичку вредност у односу на циљану променљиву Diabetes\_012.

Претпоставку ћемо испитати статистички преко Крамеровог коефицијента и χ² теста, наравно обе методе смо већ предефинисали и имплементирали па их онда једноставно позивамо.

|  |
| --- |
| chi\_sq\_test(data$HeartDiseaseorAttack , data$Diabetes\_012)  cramer\_v(data$HeartDiseaseorAttack , data$Diabetes\_012) |
|  |

χ² тест независности показује статистички значајну повезаност између променљивих Diabetes\_012 i HeartDiseaseorAttack χ² = 8244,9, a p-value < 0,00000000000000022 што је доста испод границе статистичке значајности од 0,05. Међутим, графички преглед не открива јасан образац, а Крамеров коефицијент V = 0,18 указује на слабу повезаност. Стога се може закључити да променљива HeartDiseaseorAttack нема практичан самостални ефекат на статус дијабетеса. За мултиваријантну анализу, ова променљива ће бити разматрана само ако други фактори указују на могућу интеракцију или контролни ефекат.

#### PhysActivity vs Diabetes\_012

Анализирамо физичку активност (PhysActivity) у односу на резултате дијабетичког стања (Diabetes\_012). Тип карактеристике PhysActivity је бинарна категоријска променљива са категоријама Да (да испитаник је физички активан), Не (испитаник није физички активан). Дистрибуцију зависности физичке активност унутар сваке класе дијабетеса приказали смо стубичастим дијаграмом i топлотним дијаграмом.

|  |  |
| --- | --- |
| ggplot(data, aes(x = Diabetes\_012, fill = PhysActivity)) +  geom\_bar(position = "dodge") +  labs(  title = "Distribucija kategorija PhysActivity po kategorijama Diabetes\_012",  y = "Broj opservacija",  fill = "Fizička aktivnost"  ) + scale\_fill\_paletteer\_d("MetBrewer::Archambault")+  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  )+  theme(legend.position = "bottom")  data %>%  count(Diabetes\_012, PhysActivity) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = PhysActivity, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija PhysActivity po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "PhysActivity",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) | |
|  |  |

Са стубичастог графика визуелно је јасно да код особа без дијабетеса доминирају испитаници који имају физичку активност, код особа са неким стањем дијабетеса такође доминирају они који имају физичку активност. Градијент на топлотним дијаграму показује променљив однос, али распоред не открива јасан образац међу класама дијабетеса. То указује да физичка активност није самостално повезана са статусом дијабетеса и нема аналитичку вредност у односу на циљану променљиву Diabetes\_012. Податак о физичкој активности не указује на то да ли су испитаници били активни пре развоја дијабетеса. Испитаници су се изјашњавали о физичкој активности последњих 30 дана, а временски податак о дијагнози дијабетеса немамо. Могуће је да се физичка активност наставила или започела након дијагнозе, али из овог сета података то се не може закључити.

Претпоставку ћемо испитати статистички преко Крамеровог коефицијента и χ² теста, наравно обе методе смо већ предефинисали и имплементирали па их онда једноставно позивамо.

|  |
| --- |
| chi\_sq\_test(data$PhysActivity , data$Diabetes\_012)  cramer\_v(data$PhysActivity , data$Diabetes\_012) |
|  |

χ² тест независности показује статистички значајну повезаност између променљивих Diabetes\_012 i PhysActivity χ² = 3789, a p-value < 0,00000000000000022 што је доста испод границе статистичке значајности од 0,05. Међутим, графички преглед не открива јасан образац, а Крамеров коефицијент V = 0.122 указује на слабу повезаност. Стога се може закључити да променљива PhysActivity нема практичан самостални ефекат на статус дијабетеса. За мултиваријантну анализу, ова променљива ће бити разматрана само ако други фактори указују на могућу интеракцију или контролни ефекат.

#### Fruits vs Diabetes\_012

Анализирамо учесталост испитаника који су конзумирају воће (Fruits) у односу на резултате дијабетичког стања (Diabetes\_012). Тип карактеристике Fruits је бинарна категоријска променљива са категоријама Да (испитаник конзумира воће једном или више пута у току дана), Не (испитаник не конзумира воће једном или више пута у току дана). Дистрибуцију конзумације воћа унутар сваке класе дијабетеса приказали смо стубичастим дијаграмом i топлотним дијаграмом.

|  |  |
| --- | --- |
| ggplot(data, aes(x = Diabetes\_012, fill = Fruits)) +  geom\_bar(position = "dodge") +  labs(  title = "Distribucija kategorija Fruits po kategorijama Diabetes\_012",  y = "Broj opservacija",  fill = "Konzumacija voća"  ) + scale\_fill\_paletteer\_d("MetBrewer::Archambault")+  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  )+  theme(legend.position = "bottom")  data %>%  count(Diabetes\_012, Fruits) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = Fruits, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija Fruits po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "Fruits",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) | |
|  |  |

Са стубичастог графика визуелно је јасно да код особа без дијабетеса доминирају испитаници који конзумирају воће, код особа са неким стањем дијабетеса такође има више оних који конзумирају воће. Градијент на топлотним дијаграму показује променљив однос, али распоред не открива јасан образац међу класама дијабетеса. То указује да конзумација воћа није самостално повезана са статусом дијабетеса и нема аналитичку вредност у односу на циљану променљиву Diabetes\_012.

Претпоставку ћемо испитати статистички преко Крамеровог коефицијента и χ² теста, наравно обе методе смо већ предефинисали и имплементирали па их онда једноставно позивамо.

|  |
| --- |
| chi\_sq\_test(data$Fruits, data$Diabetes\_012)  cramer\_v(data$ Fruits, data$Diabetes\_012) |
|  |

χ² тест независности показује статистички значајну повезаност између променљивих Diabetes\_012 i Fruits χ² = 454,35, a p-value < 0,00000000000000022 што је доста испод границе статистичке значајности од 0,05. Међутим, графички преглед не открива јасан образац, а Крамеров коефицијент V = 0.042 указује на веома слабу повезаност. Стога се може закључити да променљива Fruits нема практичан самостални ефекат на статус дијабетеса. За мултиваријантну анализу, ова променљива ће бити разматрана само ако други фактори указују на могућу интеракцију или контролни ефекат.

#### Veggies vs Diabetes\_012

Анализирамо учесталост испитаника који су конзумирају поврће (Veggies) у односу на резултате дијабетичког стања (Diabetes\_012). Тип карактеристике Veggies је бинарна категоријска променљива са категоријама Да (испитаник конзумира поврће једном или више пута у току дана ), Не (испитаник не конзумира поврће једном или више пута у току дана). Дистрибуцију конзумације воћа унутар сваке класе дијабетеса приказали смо стубичастим дијаграмом i топлотним дијаграмом.

|  |  |
| --- | --- |
| ggplot(data, aes(x = Diabetes\_012, fill = Veggies)) +  geom\_bar(position = "dodge") +  labs(  title = "Distribucija kategorija Veggies po kategorijama Diabetes\_012",  y = "Broj opservacija",  fill = "Konzumacija povrća"  ) + scale\_fill\_paletteer\_d("MetBrewer::Archambault")+  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  )+  theme(legend.position = "bottom")  data %>%  count(Diabetes\_012, Veggies) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = Veggies, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija Veggies po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "Veggies",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) | |
|  |  |

Са стубичастог графика визуелно је јасно да код особа без дијабетеса доминирају испитаници који конзумирају поврће, код особа са неким стањем дијабетеса такође има више оних који конзумирају поврће. Градијент на топлотним дијаграму показује променљив однос, али распоред не открива јасан образац међу класама дијабетеса. То указује да конзумација поврћа није самостално повезана са статусом дијабетеса и нема аналитичку вредност у односу на циљану променљиву Diabetes\_012.

Претпоставку ћемо испитати статистички преко Крамеровог коефицијента и χ² теста, наравно обе методе смо већ предефинисали и имплементирали па их онда једноставно позивамо.

|  |
| --- |
| chi\_sq\_test(data$Veggies, data$Diabetes\_012)  cramer\_v(data$Veggies, data$Diabetes\_012) |
|  |

χ² тест независности показује статистички значајну повезаност између променљивих Diabetes\_012 i Veggies χ² = 893,84, a p-value < 0,00000000000000022 што је доста испод границе статистичке значајности од 0,05. Међутим, графички преглед не открива јасан образац, а Крамеров коефицијент V = 0.059 указује на веома слабу повезаност. Стога се може закључити да променљива Veggies нема практичан самостални ефекат на статус дијабетеса. За мултиваријантну анализу, ова променљива ће бити разматрана само ако други фактори указују на могућу интеракцију или контролни ефекат.

#### HvyAlcoholConsump vs Diabetes\_012

У овој анализи ће се посматрати колика је учестаност људи који често конзумирају, односно ретко конзумирају алкохол а да притом имају неки вид дијабетеса. То радимо анализом мултикласне променљиве Diabetes\_012, која нам говори да ли испитаник има дијабетес/предијабетес или га нема, и бинарне категорисјке променљиве HvyAlcoholConump која нам говори да ли испитаник има тенденције да често конзумира алкохол, односно да ли конзумира више од 14 пића на недељном нивоу за мушкарце и 7 за жене,наравно ова промељива је факторизована на једноставније „Da i ne“. Одговор о пропорционалности класа ћемо добити употребом сложеног стубичног дијаграма добијеног из датог кода:

|  |
| --- |
| ggplot(data, aes(x = Diabetes\_012, fill = HvyAlcoholConsump)) +  geom\_bar(position = "dodge") +  geom\_text(stat = "count",  aes(label = ..count..),  vjust = -0.5,  position = position\_dodge(width = 0.9),  size = 3) +  scale\_fill\_manual(values = colors)+  labs(title = "Distribucija HvyAlcoholConsump u odnosu na Diabetes\_012",  x = NULL,  y = "Broj ispitanika",  fill ="Konzumira dosta alkohola") |
|  |

На основу датог дијаграма оно што се да видети јесте да немамо нарочиту пропорцију узорака за групе људи који су већи конзументи и оних који то нису, што је и очекивано јер је у униваријантној анализи доказано да од укупног броја испитаника, преко 90% њих није тежак конзумент алкохола. Међутим оно што је евидентно јесте да за све класе дијабетеса конзуматори и неконзуматори алкохола имају пропорционално исти однос, што би значило да нема веће зависности.

Са друге стране да бисмо утврдили да ли постоји корелација између варијабли употребићемо Топлотну мапу (heat map) на основу приложеног кода:

|  |
| --- |
| data %>%  count(Diabetes\_012, HvyAlcoholConsump) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = HvyAlcoholConsump, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija HvyAlcoholConsump po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "HvyAlcoholConsump",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) |
|  |

Иако је пропорција међу групама лоша, визуалним посматрањем топлотне мапе видимо да нема неке релације између оболелих који конзумирају и не конзумирају алкохол, што ће рећи да, барем по топлотној мапи, нема релације између датих варијабли.

Међутим, иако нам график говори да нема неке нарочите релације међу датим варијаблама, сама визуелна инспекција нам није довољан доказ, па како бисмо потврдили оно што нам график говори, употребићемо Крамеров „V“ коефицијент и χ² тест, наравно обе методе смо већ предефинисали и имплементирали па их онда једноставно позивамо:

|  |
| --- |
| chi\_sq\_test(data$HvyAlcoholConsump, data$Diabetes\_012)  cramer\_v(data$HvyAlcoholConsump, data$Diabetes\_012) |

χ² тест je израчунао χ² = 850.32, df = 2, p-value < 2.2e-16, што би рекло да по вредност „p” са математичке стране постоји статистичка повезаност између употребе алкохола и дијабетеса, X-squared има релативно нижу вредност, што ће рећи да су ове две варијабле независне. Са друге стране иако нам је „chi“ тест дао одговор да употреба алкохола има повезаност са дијабетесом, Крамеров коефицијент нам је дао релативно ниску вредност од 0.0579, што ће рећи да у пракси нема готово никакав утицај на сам исход да ли особа има дијабетес или не, па се слободно може рећи да је ова варијабла занемарива при превиђању типа дијабетеса код особе.

#### AnyHealthcare vs Diabetes\_012

У овој анализи ће се посматрати колика је учестаност људи који имају било какав вид здраствене неге, односно немају, а да притом имају неки вид дијабетеса. То радимо анализом мултикласне променљиве Diabetes\_012, која нам говори да ли испитаник има дијабетес/предијабетес или га нема, и бинарне категорисјке променљиве AnyHealthcare која нам говори да ли испитаник има неки вид здравствене неге, наравно ова промељива је факторизована на једноставније „Da” и „Не“, као што је случај код анализе са HvyAlcoholConsump . Одговор ћемо добити употребом сложеног стубичног дијаграма добијеног из датог кода:

|  |
| --- |
| ggplot(data, aes(x = Diabetes\_012, fill = AnyHealthcare)) +  geom\_bar(position = "dodge") +  geom\_text(stat = "count",  aes(label = ..count..),  vjust = -0.5,  position = position\_dodge(width = 0.9),  size = 3) +  scale\_fill\_manual(values = colors) +  labs(title = "Distribucija klasa AnyHealthcare u odnosu na klase Diabetes\_012",  x = NULL,  y = "Broj opservacija",  fill = "Ima z. negu") |
|  |

Према визуелној инспекцији добијеног графика, опет видимо да нам је број испитаника према приступу здравственој нези непропорционалан, исто као што је било при анализи са алкохолом, са тим што у овом случају 95% испитаника има приступ неком виду здравствене неге. На основу података рекло се види да готово све опсервације имају здравствену заштиту, што и не чуди с обзиром да је испитивање вршено у САД,једној од најразвијенијих држава света. Оно што је такође уочљиво јесте да пропорционално постоје исти трендови при посматрању да ли различите класе дијабетеса имају приступ здравственој заштити или не.

Да бисмо визуелно утврдили да ли постоји корелација опет ћемо се послужити топлотном мапом:

|  |
| --- |
| data %>%  count(Diabetes\_012, AnyHealthcare) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = AnyHealthcare, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija AnyHealthcare po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "AnyHealthcare",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) |
|  |

Иако је пропорција међу чланова група лоша, визуалном инспекцијом топлотне мапе видимо да је повезаност између оних који имају приступ здравственој нези са и без дијабетеса непостојећа у односу на оне који немају здравствену негу, што ће рећи да нам визуелна инспекција даје основну хипотезу да релације између датих променљивих нема.

Међутим, као и код конзумирања алкохола, преглед графика нам није довољан како бисмо могли да потврдимо хипотезу о неповезаности између приступу здравственој нези и дијабетесу, те нам је потребно да опет израчунамо Крамеров „V” коефицијент, као и χ² тест, који нам оквирно дају колико је јака релација између две променљиве, као и то да ли уопште постоји и зато ћемо их сада израчунати:

|  |
| --- |
| chi\_sq\_test(data$AnyHealthcare, data$Diabetes\_012)  cramer\_v(data$AnyHealthcare, data$Diabetes\_012) |

Напокон, када извршимо дате функције за χ² тест смо добили следећи резултат:

χ² тест = 69.078, df = 2, p-value = 9.998e-16. За дати резултат на основу параметара p који је практично 0, може се закључити да постоји статистичка повезаност између двеју варијабли, а вредност χ² ,која је релативно ниска у односу на скуп података нам говори да је одступање независности од приступа здравственој нези од тога да ли неко има дијабетес занемарива. Стога нам је потребан Крамеров коефицијент да нам каже да ли приступ здравственој нези уопште има било какав утицај на то да ли неко има дијабетес или не. За Крамеров коефицијент смо добили вредност од 0.01650162, која нам говори да у пракси и за модел који би предвидео да ли неко има дијабетес или не, приступ здравственој нези нема готово никакав утицај.

#### NoDocbcCost vs Diabetes\_012

Следећа анализа нам треба дати одговор на постојање повезаности између појаве дијабетеса и немогућности одласка лекару услед недостатка финансијских средстава. NoDocBcCost нам говори да ли је испитаник у претходних месец дана имао средства да оде у посету лекару. Као и код претходне две анализе и ова варијабла има два нивоа „Да” и „Не“. Одговор о пропорционалности класа добити употребом сложеног стубастог дијаграма:

|  |
| --- |
| ggplot(data, aes(x = Diabetes\_012, fill = NoDocbcCost)) +  geom\_bar(position = "dodge") +  geom\_text(stat = "count",  aes(label = ..count..),  vjust = -0.5,  position = position\_dodge(width = 0.9),  size = 3) +  scale\_fill\_manual(values = colors) +  labs(title = "Distribucija klasa NoDocbcCost u odnosu na tipove Diabetes\_012",  x = NULL,  y = "Broj opservacija",  fill = "Bez mogućnosti odlaska") |
|  |

Као и код претходне две анализе, и у овој је број испитаника према могућности одласка лекару непропорционалан, свега 8.7% није имало могућност одласка лекару, што чини ову варијаблу пристрасном. Исто се може рећи да пропорционално имамо готово идентичне односе између оних који су имали могућности да оду лекари, и оних који то нису имали, без обзира да ли је у питању неко ко има дијабетес или не.

Визуално утврђујемо постојаност повезаности између датих променљивих употребом топлотне мапе:

|  |
| --- |
| data %>%  count(Diabetes\_012, NoDocbcCost) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = NoDocbcCost, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija NoDocbcCost po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "NoDocbcCost",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) |
|  |

Визуалном инспекцијом дате топлотне мапе јасно је уочљиво да не постоји нарочита повезаност између класа, иако је пропорциоанлност броја испитаника у односу на класе дијабетеса лоша, видимо да за сваку класу дијабетеса имамо практччно исти однос. Свакако радимо Крамеров коефицијент и χ² тест како бисмо одбранили тезу о неповезаности.

|  |
| --- |
| chi\_sq\_test(data$NoDocbcCost, data$Diabetes\_012)  cramer\_v(data$NoDocbcCost, data$Diabetes\_012)  chi\_sq\_test(data$NoDocbcCost, data$Diabetes\_012)  X-squared = 396.08, df = 2, p-value < 2.2e-16  > cramer\_v(data$NoDocbcCost, data$Diabetes\_012)  [1] 0.03951385 |

Извршавањем датих функција, поново добијамо сличну ситуацију, χ² тест нам кроз параметре “p” које је готово 0 и χ² које је око 396 нам говори да постоји статистичка повезаност која до душе је међу класама које су независне. Вредност Крамеровог коефицијента која износи приближних 0.0395 нам говори и да је ова повезаност занемарива, тј. да је утицај ове варијабле над будућом предвидивом варијаблом Diabetes\_012 заправо занемарива, и неће представљати битан фактор за предикцију модела.

#### GenHlth vs Diabetes\_012

У овој анализи ћемо се бавити повезаношћу оцене самих испитаника у вези са њиховим здрављем на скали од 1 до 5 у односу на предиктивну варијаблу Diabetes\_012. Ради се о категоричкој променљивој тако да се опет служимо стубастим графиком као и топлотном мапом за визуелизацију:

|  |
| --- |
| ggplot(data, aes(x = Diabetes\_012, fill = GenHlth)) +  geom\_bar(position = "dodge") +  geom\_text(stat = "count",  aes(label = ..count..),  vjust = -0.5,  position = position\_dodge(width = 0.9),  size = 3) +  scale\_fill\_manual(values = colors) +  labs(title = "Distribucija GenHlth u odnosu na tipove Diabetes\_012",  x = NULL,  y = "Broj opservacija",  fill = "Ocena") |
|  |

У односу на претходне 3 анализе, овде јасно видимо да постоје разлике у дистрибуцији између класа. Док код опсервација у којима испитаници нису имали дијабетес превладавају оцене о општем здрављу које су биле већином Добре, Врло добре и Одлично, то није нарочит случај код друга два стања, односно где су испитаници имали дијагностикован предијабетес/дијабетес. Наравно овде се може уочити и да имамо највише опсервација код којих испитаници немају дијабетес, али се такође може рећи да је број оцена које говоре да испитанику није добро пропорционално највећи управо код испитаника који имају дијабетес, што ће рећи да је ова варијабла потенцијални предиктор за наш модел, али остаје да се види са још неким мерилима о којима више у наставку.

Како имамо доста нивоа оцена здравља,било би пожељно урадити и box plot који ће нам боље дати увид у дистрибуцију оцена међу типовима дијабетеса:

|  |
| --- |
| ggplot(data, aes(x = Diabetes\_012, y = as.numeric(GenHlth), fill = Diabetes\_012)) +  geom\_boxplot() +  scale\_fill\_manual(values=colorS)+  labs(title = "Distribucija ocena zdravlja po tipovima dijabetesa",  y = "Ocena zdravlja (1-Odlično, 5-Loše)",  x = "Dijabetes status") |
|  |

На основу box plota јасно је да постоје разлике у дистрибуцији оцена здравља на основу типова дијабетеса, тј. евидентно је да како се дијабетес појављује,тако је дистрибуција оцена на све нижем нивоу у односу на оне који немају дијабетес.

Како бисмо утврдили да ли има релације визуално, служимо се топлотном мапом.

|  |
| --- |
| data %>%  count(Diabetes\_012, GenHlth) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = GenHlth, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija GenHlth po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "GenHlth",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) |
|  |

Као што се дало предочити са стубичастим графиком, овде постоје већи помаци у повезаности у односу на претходне 3 анализе. Евидентно је да пропорционално, иако имамо мање опсервација са дијабетесом него оних без, ми можемо утврдити да пропорционално највише резултата где је некоме лоше има управо у опсервацијама са дијабетесом, у односу на онима који се генерално осећају здраво, а којих највише пропорционално има без дијабетеса.

Међутим као и пре, није нам довољан само визуални приказ да би утврдили да ли је нешто повезано или не, потребно је употребити и математичко-статистичке методе како бисмо дошли до мерила да ли је нешто повезано, зато поново користимо χ² тест, као и “V” коефицијент.

|  |
| --- |
| chi\_sq\_test(data$GenHlth, data$Diabetes\_012)  X-squared = 24248, df = 8, p-value < 2.2e-16  cramer\_v(data$GenHlth, data$Diabetes\_012)  0.2186154 |

Ове бројке нам говоре да је статистичка повезаност итекако постојећа, то нам говори вредност „p“, која је готово 0, међутим оно што је интересантније, јесте да је вредност χ² поприлично висока, што нам говори да статистички то није случајно, и да постоје међузависности варијабли. Наравно како би потврдили утицај ту је Крамеров коефицијент који је овог пута нешто виши, односно већи је од 0.1, што је релативна вредност да ли нешто има неки утицај или не, стога можемо донети закључак да ће ова варијабла утицати на то да ли неко има дијабетес или не, али више ће се показати у мултиваријатној анализи.

#### MentHlth vs Diabetes\_012

У овој анализи ће се говорити о повезаности између менталног здравља испитаника са и без дијабетеса. MentHlth нам говори о менталним проблемима испитаника, односно да ли је испитаник у претходних 30 дана имао неки психички проблем и када. Потребно је утврдити да ли појава менталних проблема прати дијабетес. За разлику од претходне 4 анализе, у овој учествује нумеричка варијабла,те нам је потребно употреби други тип графика, односно „Box plot” и „Density plot”

|  |
| --- |
| ggplot(data, aes(x = as.factor(Diabetes\_012), y = MentHlth, fill = as.factor(Diabetes\_012))) +  geom\_boxplot(alpha = 0.7) +  scale\_fill\_manual(values=colors) +  labs(title = "Distribucija MentHlth po tipovima Diabetes\_012", fill = "Dijabetes") |
|  |

Шта нам говори дати „Box plot“? Па као прво видимо да су нам за све три групе медијане готово на 0 што нам говори да преко половине испитаника није осећало било какве менталне потешкоће. Наравно за све три групе можемо видети и солидан број изузетака (outliera), што значи да постоје екстремне вредности за све 3 групе испитаника, што значи да екстремни случајеви не зависе од саме групе. Са друге стране види се да су интерквартилни опсези група предијабетес и дијабетес нешто већи од групе која нема дијабетес, што значи да постоји потенцијална повезаност ових двеју варијабли, али наравно ово је само претпоставка, остаје се утврдити другим методама.

Хистограм ћемо употребити да би видело колика је тачна учестаност испитаника који нису осетили било какве менталне потешкоће у периоду од месец дана:

|  |
| --- |
| ggplot(data, aes(x = MentHlth, fill = Diabetes\_012)) +  geom\_histogram(binwidth = 1, position = "dodge", alpha = 0.8) +  scale\_fill\_manual(values =colors) +  labs(title = "Histogram MentHlth u odnosu na Diabetes\_012",  y = "Broj opservacija",  fill = "Dijabetes") |
|  |

Дакле наше сумње су се обистиниле, велика већина опсервација из скупа података није имала никакве менталне потешкоће, и то су готово све опсервације без дијабетеса. Једина већа група, мада занемарива у односу на прво поменуту су људи који нису имали проблема али имају дијабетес, што нам даје слабе назнаке да овде постоји повезаност. Наравно пре свега ћемо употребити мапу густине како бисмо добили мало бољу представу о дистрибуцији типова дијабетеса према броју дана са менталним потешкоћама.

|  |
| --- |
| ggplot(data, aes(x = MentHlth, color = Diabetes\_012, fill = Diabetes\_012)) +  geom\_density(alpha = 0.2) +  scale\_fill\_manual(values=colorS)+  labs(title = "Gustina MentHlth po tipu dijabetesa") |
|  |

На основу мапе густине јасно је да већина опсервација тежи ка 0,небитно од типа дијабетеса,мада оно што је добро истаћи јесте да ако се погледају појаве репова на одређеним данима, евидентно је да не постоје неке значајне разлике у промени густине оних који имају дијабетес/предијабетес односно немају га.

Ову асиметричност карактеристике *MentHlth* смо увидели и у униваријантној анализи, а сада у биваријантној потврдили. Када узмемо у обзир додатне закључке идентичне расподеле по групама дијабетеса. Разлике између група се пре свега виде у учесталости већих вредности, а не у главној тенденцији ка 0, што додатно указује да променљива *MentHlth* нема снажну дискриминаторну моћ када се посматра у свом континуалном облику. На основу хистограма предложени интервал група је:

|  |  |
| --- | --- |
| Категорија | Опис |
| 0 | нема проблема |
| 1-5 | благи проблеми |
| 6-15 | умерени проблеми |
| 16-30 | тешки проблеми |

Претпотавку да *MentHlth* у свом нумеричком облику није утицајна испитаћемо статички. Како се ради о нумеричкој варијабли овде ћемо користити “anova test” и “tukey test”. Путем анове дознајемо да ли је просек лоших дана исти у свим групама, док је „Tukey” ту да нам каже које су то групе које садрже разлику.

|  |
| --- |
| anova\_test(data$MentHlth,data$Diabetes\_012)  Df Sum Sq Mean Sq F value Pr(>F)  as.factor(grupna\_var) 2 78369 39185 717.1 <2e-16 \*\*\*  Residuals 253677 13861367 55  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  > tukey\_fun(data$MentHlth,data$Diabetes\_012)  Tukey multiple comparisons of means  95% family-wise confidence level  Fit: aov(formula = numericka\_var ~ as.factor(grupna\_var))  $`as.factor(grupna\_var)`  diff lwr upr p adj  predijabetes-nema dijabetes 1.585503 1.3281776 1.8428284 0.000000  dijabetes-nema dijabetes 1.517402 1.4179230 1.6168810 0.000000  dijabetes-predijabetes -0.068101 -0.3388471 0.2026451 0.825752 |

Из добијених резултата урађених тестова може се закључити следеће:

Код анова теста значајна нам је вредност F и p, која је релативно висока, говори да постоји разлика у просеку менталног здравља међу испитаницима који имају, односно немају некакав тип дијабетеса. Међутим да би се открило код којих класа испитаника постоји пораст у просеку дана са менталним потешкоћама нам је потребна употреба „Tukey” теста. Његови резултати нам говоре да у просеку људи без дијабетеса имају за око 1.5 дана мање дана са менталним потешкоћама него они који имају дијабетес/предијабетес, док је однос између дијабетеса/предијабетеса у основи занемарив јер му је разлика око 0.5. Иако делује да постоји статистичка значајност на ментално здравље буде предиктор, у пракси при предикцији он нема високи потенцијал да буде предиктор, тј. занемарив је.

#### PhysHlth vs Diabetes\_012

У овој анализи ћемо испитати да ли варијабла PhysHlth има икакву повезаност са варијаблом Diabetes\_012. У основи прво поменута варијабла нам говори да ли је испитаник имао икаквих физичких потешкоћа у претходних месец дана, слично као и код варијабле MntHlth. Структурно гледано и ова варијабла је нумеричка, па ћемо се опет служити бокс плотом и хистограмом:

|  |
| --- |
| ggplot(data, aes(x = Diabetes\_012, y = PhysHlth, fill = Diabetes\_012)) +  geom\_boxplot(alpha = 0.7) +  scale\_fill\_manual(values=colors) +  labs(title = "Distribucija PhysHlth po tipovima Diabetes\_012",  fill = "Dijabetes") |
|  |

Овде се може видети да постоји одређени растући тренд, док је кутија за нема дијабетес најмања, готово равномерно кутија расте са сваком следећом класом типа дијабетеса. Посебно је интересантно то што се медијана коначно одлепљује од нуле код опсервација са дијабетесом, што ће рећи да постоји осетно више физичких тегоба код људи са дијабетесом него код оних вез њега. Такође је лако уочљиво да изузетака код опсервација са дијабетесом нема, што ће рећи да постоји известан образац тегоба код људи који имају дијабетес. Ово нам даје индицију да ће се ова варијабла наћи међу предикторима у нашем моделу, али наравно још је рано о томе говорити, па ћемо се послужити још неким техникама како бисмо то утврдили.

|  |
| --- |
| ggplot(data, aes(x = PhysHlth, fill = Diabetes\_012)) +  geom\_histogram(binwidth = 1, position = "dodge", alpha = 0.8) +  scale\_fill\_manual(values =colors) +  labs(title = "Histogram PhysHlth u odnosu na Diabetes\_012",  y = "Broj opservacija",  fill = "Dijabetes") |
|  |

Хистограм нам пак даје мало обесхрабрујућу слику јер нам говори да у скупу података највише има испитаника без физичких потешкоћа, и притом најбројнији скуп опсервација је управо онај где испитаник нема дијабетес. Наравно пре свега ћемо употребити мапу густине како бисмо добили мало бољу представу о дистрибуцији типова дијабетеса према броју дана са физичким потешкоћама.

|  |
| --- |
| ggplot(data, aes(x = PhysHlth, color = Diabetes\_012, fill = Diabetes\_012)) +  geom\_density(alpha = 0.2) +  scale\_fill\_manual(values=colorS)+  labs(title = "Gustina PhysHlth po tipu dijabetesa") |
|  |

На основу мапе густине јасно је да не постоји нарочитих промена према данима са физичким потешкоћама осим када је у питању свих 30 дана у месецу када је јасно да густина све 3 групе одједном расте, односно постоји већа шанса да ће испитаник пријавити да је имао физичке потешкоће цео месец у односу на мањи број дана у истом. Наравно и даље је евидентна асиметрија у броју опсервација са и без икаквих потешкоћа где се види да већина опсервација невезано за класу припада без икаквих потешкоћа у току месеца.

Такву асиметричност смо увидели и у униваријантној анализи. Разлике између група се показује у промени структуре и већем броју дана са потешкоћама, што додатно указује да променљива PhysHlth имати већу значајност у категоријском облику. На основу хистограма предложени интервал група је:

|  |  |
| --- | --- |
| Категорија | Опис |
| 0 | нема проблема |
| 1-5 | благи проблеми |
| 6-15 | умерени проблеми |
| 16-30 | тешки проблеми |

Наравно остаје да се даље утврђује употребом математичко-статистичких метода да ли ова варијабла има предикаторски потенцијал као нумеричка, те из тог разлога користимо анова и “Tukey” тест.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | | anova\_test(data$PhysHlth,data$Diabetes\_012)  Df Sum Sq Mean Sq F value Pr(>F)  as.factor(grupna\_var) 2 600673 300337 4079 <2e-16 \*\*\*  Residuals 253677 18679609 74  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  > tukey\_fun(data$PhysHlth,data$Diabetes\_012)  Tukey multiple comparisons of means  95% family-wise confidence level  Fit: aov(formula = numericka\_var ~ as.factor(grupna\_var))  $`as.factor(grupna\_var)`  diff lwr upr p adj  predijabetes-nema dijabetes 2.765889 2.467170 3.064609 0  dijabetes-nema dijabetes 4.372063 4.256581 4.487544 0  dijabetes-predijabetes 1.606174 1.291875 1.920473 0 | |  | | |  | | --- | |  | | |

Дакле након извршених тестова добијамо јаснију слику. Анова тест нам је дао поприличну високу вредност параметра F=4079,што ће рећи да међу групама итекако постоји разлика међу групама дијабетичара и њихових физичких потешкоћа, чиме параметар “p” потврђује нам на значајности ове разлике. Наравно да би видели међу којим класама разлика постоји користиће нам други обављени тест, односно „Tukey”, који нам говори да је од особе која нема дијабетес до особе која га има, просечан број дана са физичким потешкоћама у просеку износи 4.37 дана, наравно ни остале разлике нису занемариве, што није био случај код менталних потешкоћа, па тако разлика између особа са дијабетесом и предијабетесом износи 1.6 дана, док код предијабетеса и оних који га немају износи 2.8 дана, што нису занемариви бројеви, те је безбедно рећи да физичке потешкоће имају утицаја на то да ли неко има дијабетес или не.   
  
Ово иде у прилог о одлуци прегрупације, која ће очувати значајност, а истовремено редуковати асиметричност.

#### DiffWalk vs Diabetes\_012

Садашња анализа ће нам дати повезаност између варијабле DiffWalk и Diabetes\_012. Ова варијабла нам говори да ли испитаник има потешкоће у кретању, те је ово класна варијабла са два нивоа, односно „Da“ и „Ne”, наравно како то бива са класним варијаблама употребљавамо стубичасти дијаграм и топлотну мапу ради добијања хипотезе:

|  |
| --- |
| ggplot(data, aes(x = Diabetes\_012, fill =DiffWalk )) +  geom\_bar(position = "dodge") +  geom\_text(stat = "count",  aes(label = ..count..),  vjust = -0.5,  position = position\_dodge(width = 0.9),  size = 3) +  scale\_fill\_manual(values = colors) +  labs(title = "Distribucija DiffWalk u odnosu na tipove Diabetes\_012",  x = NULL,  y = "Broj opservacija",  fill = "Ima poteškoća") |
|  |

Према графику може се уочити, да нам пропорционалност потешкоћа и није баш најбоља, али је лако уочљиво да нам број људи са потешкоћама пропорционално расте са појавом дијабетеса, што би рекло да постоји повезаност, мада је потребно то додатно утврдити.

|  |
| --- |
| data %>%  count(Diabetes\_012, DiffWalk) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = DiffWalk, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija DiffWalk po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "DiffWalk",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) |
|  |

Топлотна мапа нам засигурно даје ширу слику. Дефинитивно број потешкоћа пропорционално расте у односу на класе дијабетеса. Са тим у вези валидно је дати претпоставку да постоји повезаност између потешкоћа у кретању са стањима дијабетеса, наравно ово је потребно додатно математички доказати.

|  |
| --- |
| chi\_sq\_test(data$DiffWalk, data$Diabetes\_012)  data: tabela  X-squared = 12777, df = 2, p-value < 2.2e-16  cramer\_v(data$DiffWalk, data$Diabetes\_012)  0.2244245 |

Дефинитивно се може рећи да имамо повезаност сада, χ² тест нам говори да постоји статистичка повезаност између датих варијабли, то нам говори параметар “p” који је готово једнак 0, са друге стране χ² тест вредност која је поприлично висока, нам говори да постоји јака корелација међу датим варијаблама. Са друге стране како бисмо добили тачну јачину повезаности. Зато имамо Крамеров коефицијент који је у овом случају висок, односно 0.2244, што је у односу на релативну вредност већи, па се дефинитивно може потврди да овде постоји повезаност,али да је боље употребити је у мултиваријантној анализи него је држати под самостални предиктор.

#### Sex vs Diabetes\_012

Сада ћемо да истражујемо повезаност појаве дијабетеса и полова. Пол је у овом случају категоријска променљива са две вредности, односно „zensko“ и „musko“, те и овде употребљавамо методе као и код категоријских променљивих. Прво ћемо све визуализовати:

|  |
| --- |
| ggplot(data, aes(x = Diabetes\_012, fill =Sex )) +  geom\_bar(position = "dodge") +  geom\_text(stat = "count",  aes(label = ..count..),  vjust = -0.5,  position = position\_dodge(width = 0.9),  size = 3) +  scale\_fill\_manual(values = colors) +  labs(title = "Distribucija polova u odnosu na tipove Diabetes\_012",  y = "Broj opservacija",  fill = "Pol") |
|  |

Овде увиђамо да не постоји нарочито одступање по половима у односу на појаву дијабетеса, што би рекло да вероватно не постоји нарочита повезаност. Наравно сам преглед стубичастог графика нам не може засигурно рећи о постојању повезаности датих варијабли, зао ћемо извршити додатне анализе.

|  |
| --- |
| data %>%  count(Diabetes\_012, Sex) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = Sex, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija Sex po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "Sex",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) |
|  |

Овде је уочљиво да нема нарочитих промена у односу између полова у односу на тип дијабетеса, што би нам у основи могло дати хипотезу да не постоји нарочита повезаност између пола испитаника и дијабетеса, али да бисмо то потврдили радимо тестове.

|  |
| --- |
| chi\_sq\_test(data$Diabetes\_012,data$Sex)  X-squared =250.85, df = 2, p-value < 2.2e-16  cramer\_v(data$Diabetes\_012,data$Sex)  0.03144593 |

Иако је вредност параметра p у χ² тесту говори да постоји статистичка значајност, али да је одступање од независности занемариво што нам говори χ² вредност. Наравно саму јачину утицаја утврђујемо Крамеровим коефицијентом који нам говори да је ова повезаност занемарива са вредношћу од 0.031,па је безбедно рећи да дата варијабла нема предиктивну вредност, бар не самостално.

#### Age vs Diabetes\_012

У овој анализи ћемо се бавити повезаношћу између старости испитаника и појаве дијабетеса. Старосна доб је нумеричка вредност, и представља редни број старосне групе којој испитаник припада (закључено у униваријантној анализи). У овом случају се служимо хистограмом и бар плотом да бисмо извукли претпоставке о зависности:

|  |
| --- |
| ggplot(data, aes(x = Diabetes\_012, y = Age, fill = Diabetes\_012)) +  geom\_boxplot(alpha = 0.7) +  scale\_fill\_manual(values=colors) +  labs(title = "Distribucija godina po tipovima Diabetes\_012",  fill = "Dijabetes") |
|  |

На основу бокс плота видимо да је појава предијабетеса и дијабетеса учесталија код испитаника који су у просеку старији од испитаника из групе који немају дијабетес. Наравно иако постоји јасан растући тренд код година у односу на појаву дијабетеса, постоји неколико изузетака у групи са дијабетесом, што је са доменског знања очекивано услед гојазности (висок БМИ) или генетике испитаника.

|  |
| --- |
| ggplot(data, aes(x = Age, fill = Diabetes\_012)) +  geom\_histogram(binwidth = 1, position = "dodge", alpha = 0.8) +  scale\_fill\_manual(values =colors) +  labs(title = "Histogram godina u odnosu na Diabetes\_012",  y = "Broj opservacija",  fill = "Dijabetes") |
|  |

На основу прегледа хистограма увидено је да број оболелих од дијабетеса креће експоненцијално да расте од старосне доби 7, што ће рећи да постоји хипотеза која нам тврди да су старосна доб и појава дијабетеса повезани, зато ћемо урадити тестове како бисмо ово потврдили.

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | | anova\_test(data$Age,data$Diabetes\_012)  Df Sum Sq Mean Sq F value Pr(>F)  as.factor(grupna\_var) 2 82130 41065 4560 <2e-16 \*\*\*  Residuals 253677 2284255 9  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  tukey\_fun(data$Age,data$Diabetes\_012)  Tukey multiple comparisons of means  95% family-wise confidence level  Fit: aov(formula = numericka\_var ~ as.factor(grupna\_var))  $`as.factor(grupna\_var)`  diff lwr upr p adj  predijabetes-nema dijabetes 1.2967924 1.1923320 1.401253 0  dijabetes-nema dijabetes 1.5924939 1.5521107 1.632877 0  dijabetes-predijabetes 0.2957015 0.1857929 0.405610 0 | |  | |  | |

На основу анова теста евидентно је да постоји разлика у старосној доби између оних који имају неки облик дијабетеса и оних који га немају. То је евидентно услед високе вредност F=4560, а додатно на значајности ове разлике нам потврђује параметар „p”=2e-16. Tukey нам није дао нарочиту разлику међу класама па се може слободно рећи да је варијабала Age у слабој релацији са Diabetes\_012.

#### Education vs Diabetes\_012

Сада ћемо анализирати повезаност појаве дијабетеса са нивоом образовања испитаника. У овом случају ради се о категоричној варијабли па ћемо употребити стубасти дијаграм и топлотну мапу ради утврђивања основне сумње да ли повезаност постоји, као и саму дистрибуцију међу класама:

|  |
| --- |
| ggplot(data, aes(x = Diabetes\_012, fill =Education )) +  geom\_bar(position = "dodge") +  geom\_text(stat = "count",  aes(label = ..count..),  vjust = -0.5,  position = position\_dodge(width = 0.9),  size = 3) +  scale\_fill\_manual(values = colorS) +  labs(title = "Distribucija klasa Education u odnosu na tipove Diabetes\_012",  y = "Broj opservacija",  fill = "Nivo obrazovanja") |
|  |

Иако имамо доста мање опсервација где испитаник има неку врсту дијабетеса, јасно је да са пропорционалне стране постоји утицај образовања на појаву дијабетеса, то је нарочито уочљиво у класи где испитаници имају дијабетес, јер је евидентно да постоји разлика у пропорцији расподеле типова образовања у односу на то да ли испитаник има дијабетес или не, наравно прецизније утврђујемо са топлотном мапом.

|  |
| --- |
| data %>%  count(Diabetes\_012, Education) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = Education, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija Education po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "Education",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) |
|  |

На основу топлотне мапе уочљиво је да постоји повезаност нивоа образовања са дијабетесом, услед чињенице да постоји другачији поредак између класа дијабетеса, а у односу на ниво образовања. Нарочито уочљиво код класе дијабетес, где се јасно види да је поредак нивоа образовања другачији у односу на испитанике без дијабетеса и њиховог нивоа образовања. Сада имамо довољно доказа да је хипотеза о повезаности нивоа образовања постојана са класом дијабетеса.

Наравно то све доказујемо статистички.

|  |
| --- |
| chi\_sq\_test(data$Diabetes\_012,data$Education)  X-squared = 4560.6, df = 10, p-value < 2.2e-16  cramer\_v(data$Diabetes\_012,data$Education)  0.09481014 |

На основу χ² теста рекло би се да постоји статистичка повезаност и да она одступа од независности, јер χ² има солидну вредност. Међутим јачина утицаја нивоа образовања на појаву дијабетеса се показала слабом са ниском вредношћу Крамеровог коефицијента од свега 0.0948, па се опет може речи да неће имати неки значајнији утицај на модел, мада се може показати значајнија у мултиваријантној анализи.

Ако поново погледамо графике изнад можемо да уочимо да у свакој групи дијабетеса има неуравнотежену расподелу тј. тенденцију ка вишим групама образовања. Овај увид смо имали и униваријантној анализи, тако да је прегрупација оправдана. Оно што је више уочљивије на топлотној мапи како да формирамо нове категорије. Видимо је да у све три категорије дијабетеса категорије образовања „без школе/вртић“ и „основна школа“ имају приближан однос (јако слаб уочљив градијент). Додатно гледајући са доменског аспекта могу се сврстати у једну групу као основно образовање. Само ова промена нам неће изменити расподелу значајно, зато ћемо испитати процентуално

|  |
| --- |
| prop.table(table(data$Education,data$Diabetes\_012), margin = 2)\*100 |
|  |

Увидом и тренд података у процентима уз ограничење доменског знања категорије образовања су прегруписане у четири нивоа:

|  |  |
| --- | --- |
| Нове категорије | Опсег старих категорија |
| Ниско образовање | Без школе/вртић |
| Основно образовање | Основна школа |
| Средње образовање | 3-годишња и 4-годишња средња школа |
| Високо образовање | Високо образовање |

#### Income vs Diabetes\_012

Сада вршимо анализу над висином примања испитаника у односу на појаву неког типа дијабетеса. Income je категоричка варијабла те ћемо се као и до сада послужити стубастим графиком и топлотном мапом за визуелну анализу дистрибуције и потенцијалне повезаности:

|  |
| --- |
| ggplot(data, aes(x = Diabetes\_012, fill =Income )) +  geom\_bar(position = "dodge") +  geom\_text(stat = "count",  aes(label = ..count..),  vjust = -0.5,  position = position\_dodge(width = 0.9),  size = 3) +  scale\_fill\_manual(values = colors) +  labs(title = "Distribucija nivoa primanja u odnosu na tipove Diabetes\_012",  y = "Broj opservacija",  fill = "Primanja") |
|  |

Као што је већ случај, број испитаника који немају дијабетес је драстично већи од оних који имају предијабетес/дијабетес. Међутим на први поглед, свака од класа дијабетеса има сличну расподелу нивоа примања, па би се на први поглед рекло да нема повезаности, наравно ово се мора утврдити.

|  |
| --- |
| data %>%  count(Diabetes\_012, Income) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = Income, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija Income po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "Income",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) |
|  |

Када погледамо топлотну мапу за појаву дијабетеса и ниво примања, добијамо другачију слику, наиме уочљиво је да је однос између оних који имају дијабетес а који имају висока примања (преко 75000 долара) са људима који имају ниска примања (мање од 10000 долара) знатно мањи, око 1:3, у односу на однос људи који немају дијабетес, где је однос људи са највишим примањима у односу на оне са најнижим, око 1:12, као што је то случај и са највишим примањима у односу на средња примања, итд. У сваком случају јасно је да су односи примања по класама дијабетеса другачији, па се ипак може сумњати да постоји одређена повезаност двеју варијабли.

Наравно ово ћемо све потврдити тестовима.

|  |
| --- |
| chi\_sq\_test(data$Diabetes\_012,data$Income)  X-squared = 7816.5, df = 14, p-value < 2.2e-16  cramer\_v(data$Diabetes\_012,data$Income)  0.1241215 |

Тест χ² нам говори да статистичка повезаност постоји вредношћу параметра „p“ која је готово једнака 0 и која је мања од 0.05, а параметар χ² нам говори да та повезаност има значаја својом релативно високом вредношћу која износи 7816. Са друге стране Крамеров индекс нам говори да иако постоји статистичка повезаност која има значаја, њен утицај на Diabetes\_012 није нарочито јак и може се показати као бољи предиктор у мултиваријантној анализи.

У униваријантној анализи се примећена неуравнотежена расподела са теденцијом ка вишим вредностима примања, што смо биваријантном анализом и потврдили. Донет је закључак о другачијој подели категорија. Са стубичастог дијаграма увиђамо да сваке две суседне категорије од најмање до претпоследње, имају веома малу разлику посебно за категорије „предијабетес“ и „дијабетес“. Додатно на топлотној мапи видимо градијенте за исто, који су скоро неприметни. Ово нам говори да нема наглог скока примања у односу на циљ тј. понашају се слично за исту категорију дијабетеса. Ако претпоставимо да статистички ово јесу добри индикатори за спајање, увиђамо да ће и даље бити неуравнотежености у расподели. Како би боље увидели индикаторе погледајмо процентуални удео:

|  |
| --- |
| tabla\_Incom\_Diabetes\_broj <- table(data$Income, data$Diabetes\_012)  prop.table(tabla\_Incom\_Diabetes\_broj, margin = 2) \* 100 |
|  |

### Однос карактеристика које нису циљане

Поред испитивања односа између појединачних карактеристика и циљне променљиве, у овом поглављу анализирани су и односи између карактеристика које нису циљане. Резултат ове анализе идентификација међусобне повезаности предикторских променљивих, потенцијалне интеракције, као и могуће редундантности у подацима. Разумевање ових односа представља основу за даљу мултиваријантну анализу.

#### Income VS Education

У овој анализи ћемо утврђивати у каквој су релацији варијабле за ниво образовања и примања испитаника. Како су обе категоријске променљиве, тако ћемо користити стубичасти график и топлотну мапу за визуелизацију:

|  |
| --- |
| ggplot(data, aes(x = Education, fill =Income )) +  geom\_bar(position = "dodge")+  scale\_fill\_manual(values = colors) +  labs(title = "Distribucija nivoa primanja u odnosu na nivo obrazovanje",  y = "Broj opservacija",  fill = "Primanja") |
|  |

На основу стубичастог дијаграма евидентно је да једна група одскаче по бројности, а то су опсервације са четворо-годишњим факултетом, међутим ако погледамо трендове распоређивања јасно је да само особе са било каквим факултетом прате растући тренд од најнижег до највишег нивоа примања, док је слика нешто другачији код особа са средњом школом, a без школе је очигледно број опсервација занемарив.

|  |
| --- |
| data %>%  count(Education, Income) %>%  group\_by(Education) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Education, y = Income, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija visine primanja u odnosu na tipove nivo obrazovanja",  x = "Education",  y = "Income",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) |
|  |

На основу топлотне мапе видимо да постоји јасна разлика у поретку примања између људи који су ишли на факултет, и оних који нису. На основу тога се може поставити основна тврдња да постоји одређена повезаност. Наравно да бисмо уопште могли рећи да она постоји потребно је да урадимо тестове.

|  |
| --- |
| chi\_sq\_test(data$Education,data$Income)  X-squared = 60337, df = 35, p-value < 2.2e-16  cramer\_v(data$Education,data$Income)  0.2181037 |

Као што се види у резултатима, очигледно је да статистичка повезаност и значајно одступање од независности варијабли постоји, то нам говори мала вредност параметра “p”, али и висока χ² вредност, што нам даје јасан увид да имамо јаку повезаност. Наравно Крамеров коефицијент нам говори да иако повезаност постоји, она јесте утицајна међу датим варијабла, али не јако, већ у неком умереном интезитету.

#### HighChol VS CholCheck

У следећој анализи ћемо утврђивати да ли постоји веза између појаве холестерола и провере холестерола у последњих 5 година код испитаника. Ради се о категоријским променљивима, па на основу тога поново користим топлотну мапу и стубичасти график да бисмо визуелно испитали релацију.

|  |
| --- |
| ggplot(data, aes(x = CholCheck, fill =HighChol )) +  geom\_bar(position = "dodge")+  scale\_fill\_manual(values = colors) +  labs(title = "Distribucija HighChol u odnosu na CholCheck",  y = "Broj opservacija",  fill = "Visok holesterol") |
|  |

На основу графика јасно да је јако мало опсервација у којима испитаник није испитивао свој холестерол у претходник 5 година, свакако јасно је да у оба случаја однос људи са холестеролом је готово исти. Међутим однос је боље потврдити топлотном мапом.

|  |
| --- |
| data %>%  count(CholCheck,HighChol) %>%  group\_by(CholCheck) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = CholCheck, y = HighChol, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija HighChol u odnosu na CholCheck",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) |
|  |

На основу топлотне мапе, евидентно је да је удео опсервације у којима испитаници имају холестерол у односи на оне који га немају пропорционално већи код оних који су холестерол контролисали, што нам може дати основну сумњу да нека повезаност постоји, наравно да бисмо ово утврдили користићемо тестове.

|  |
| --- |
| chi\_sq\_test(data$CholCheck,data$HighChol)  X-squared = 1859.7, df = 1, p-value < 2.2e-16  cramer\_v(data$CholCheck,data$HighChol)  0.08562119 |

На основу тестова χ² нам говори да постоји статистичка повезаност, мада и није нарочита, односно одступања од независности двеју варијабли је релативно слаба, наравно и међусобни утицај варијабли је слаб, односно Крамер има релативно слабу вредности, испод 0.1, тако да знамо да не постоји нарочита повезаност између контролисања холестерола у последњих 5 година и његове појаве код опсервације.

#### DiffWalk vs PhysHlth

У овој анализи утврђујемо да ли постоји повезаност између физичких потешкоћа које је испитаник осећао кроз последњих 30 дана и потешкоћа са кретањем. У овом случају DiffWalk нам је категоријска променљива, док је PhysHlth нумеричка. Зато у овом случају користимо хистограм и боксплот за анализу:

|  |
| --- |
| ggplot(data, aes(x = DiffWalk, y = PhysHlth, fill = DiffWalk)) +  geom\_boxplot(alpha = 0.7) +  scale\_fill\_manual(values=colors) +  labs(title = "Distribucija PhysHlth u odnosu na DiffWalk",  fill = "Poteškoće u kretanju") |
|  |

Дати дијаграм нам даје поприличну очиту слику, а то је да они који немају проблема у кретању у просеку нису имали никакве потешкоће, тј. преко 50% испитаника их није имало, што се и види по медијани која је на 0 за оне без потешкоћа у кретању, мада има изузетака. Са друге стране они који су имали потешкоћа у кретању обично су осећали и тегобе, са просеком од око 10 дана. Наравно повезаност нећемо само графиком утврдити, потребно је осмотрити још неке анализе.

|  |
| --- |
| ggplot(data, aes(x = PhysHlth, fill = DiffWalk)) +  geom\_histogram(binwidth = 1, position = "dodge", alpha = 0.8) +  scale\_fill\_manual(values =colors) +  labs(title = "Histogram dana teskoba u odnosu na poteškoće u kretanju",  y = "Broj opservacija",  fill = "Ima poteškoće") |
|  |

Хистограм нам превасходно говори да у скупу опсервација највише има оних који нису имали никакве тескобе као ни проблема са кретањем, са већину дана евидентан је тренд у ком је већи број испитаника који су имали потешкоће за тај број дана а да су притом имали и потешкоћа у кретању.

Наравно како бисмо јасније одредили повезаност и потврдили потенцијалну повезаност из графика радимо тестове:

|  |
| --- |
| anova\_test(data$PhysHlth,data$DiffWalk)  Df Sum Sq Mean Sq  as.factor(grupna\_var) 1 4412919 4412919  Residuals 253678 14867364 59  F value Pr(>F)  as.factor(grupna\_var) 75296 <2e-16 \*\*\*  Residuals  ---  Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  > tukey\_fun(data$PhysHlth,data$DiffWalk)  diff lwr upr p adj  Da-Ne 11.14995 11.07031 11.22959 0 |

Резултати тестова нам дефинитивно потврђују повезаност потешкоћа у кретању са осећајем физичких потешкоћа. Прво нам АНОВА тест даје неуобичајену велику вредност за F што ће рећи да итекако постоје разлике у вредностима потешкоћа у односу на оне који имају потешкоће у кретању и оне који их немају, статистичку повезаност нам потврђује и изузетно мала вредност параметра „p“ која је готово равна 0.

Са друге стране „Tukey“ тест нам говори колика је та разлика у вредностима коју нам АНОВА тест говори да постоји, и разлика је велика, у просеку они без проблема у кретању имају 11 дање мања са осећањем физичких тегоба, тако да дефинитивно можемо закључити повезаност ових двеју варијабли.

#### HighBP vs HeartDeseaseAttack

Сада ћемо обавити анализу повезаности појаве високог крвног притиска и хроничних срчаних обољења. Обе од две поменуте варијабле су категоричне, па ћемо употребити топлотну мапу и стибичасти дијаграм ради визуелне анализе повезаности:

|  |
| --- |
| ggplot(data, aes(x = HighBP, fill =HeartDiseaseorAttack )) +  geom\_bar(position = "dodge")+  scale\_fill\_manual(values = colors) +  labs(title = "Distribucija pojave srčanih oboljenja u odnosu na hipertenziju",  y = "Broj opservacija",  fill = "Srčana oboljenja") |
|  |

Иако би нам доменско знање требало потврдити да постоји директна повезаност између срчаних обољења и хипертензије, график нам даје тоталну другу слику, разлог овоме вероватно лежи у томе што је скуп опсервација такав да је пристрастан, односно постоји далеко више оних који немају срчана обољења од оних који их имају, са тим у вези овај график нам не може нешто пуно осим тога рећи, па ћемо сада да пређемо на друге анализе ради даљег утврђивања.

|  |
| --- |
| data %>%  count(HighBP, HeartDiseaseorAttack) %>%  group\_by(HighBP) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = HighBP, y = HeartDiseaseorAttack, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija pojave srčanih oboljenja u odnosu na hipertenziju",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) |
|  |

Како смо већ и сумњали у стубичасти график, сумња је испада исправна у овом случају. Ако погледамо топлотну мапу јасно је да су односи између оних са срчаним обољењима и без, другачији између оних који имају и немају висок крвни притисак, јасно је да је однос код људи са крвним притиском око 1:5, док је код оних без хипертензије тај однос 1:23, па се може узети у обзир претпоставка да овде постоји повезаност, како нам доменско знање и налаже.

Зато ћемо урадити одговарајуће тестове.

|  |
| --- |
| chi\_sq\_test(data$HighBP,data$HeartDiseaseorAttack)  X-squared = 11118, df = 1, p-value < 2.2e-16  cramer\_v(data$HighBP,data$HeartDiseaseorAttack)  0.2093476 |

Као што смо и претпоставили на основу топлотне мапе, имамо повезаност, χ²нам је дао потврду да постоји статистичка повезаност, и да постоји високо одступање од независности међу варијаблама, што нам је рекла висока χ² вредност и мала вредност p параметра. Крамер је ту да нам каже да ли постоји утицај једне варијабле на другу, и евидентно је да постоји, наравно није баш најјача, тј. умерена је, али довољна за узимање у обзир у мултиваријантним анализама.

#### BMI vs HeartDeseaseorAttack

Ова анализа ће нам дати одговор на повезаност индекса телесне тежине и појаве срчаних обољења. Потреба за овом анализом произилази из доменских сазнања. BMI нам представља индекс телесне тежине, нумеричку варијаблу, док је HeartDeseaseorAttack категоријска. Са тим у вези користићемо хистограм и бокс плот за визуалну анализу:

|  |
| --- |
| ggplot(data, aes(x = HeartDiseaseorAttack, y = BMI, fill = HeartDiseaseorAttack)) +  geom\_boxplot(alpha = 0.7) +  scale\_fill\_manual(values=colors) +  labs(title = "Distribucija BMI u odnosu na srčana oboljenja",  fill = "Srčane poteškoće") |
|  |

На основу прегледа дијаграма, не видимо нарочито помераје у медијалним вредностима BMI-a у односу на то да ли неко има срчаних тегоба или не, чак су и кварталне кутије готово идентичне, са благим померајем на горе за опсервације са срчаним тескобама, али опет ништа значајно. Оно што је такође уочљиво јесте да имамо доста аутлајера, односно изузетака. Наравно настављамо даље са анализом.

|  |
| --- |
| ggplot(data, aes(x =BMI, fill = HeartDiseaseorAttack)) +  geom\_histogram(binwidth = 1, position = "dodge", alpha = 0.8) +  scale\_fill\_manual(values =colors) +  labs(title = "Histogram BMI u odnosu na srčane tegobe",  y = "Broj opservacija",  fill = "Ima poteškoće") |
|  |

У хистограму видимо да нам је скуп података неравномеран што се тиче појава срчаних потешкоћа, нешто што је било видљиво и у униваријантној анализи. Поред тога видљиво је да оба стања са и без потешкоћа доживљавају свој врхунац у бројности на готово идентичним вредностима BMI-A, што би рекло да нема нарочитих повезаности између овде две варијабле. Наравно да би утврдили то урадићемо одговарајуће тестове.

|  |
| --- |
| anova\_test(data$BMI,data$HeartDiseaseorAttack)  Df Sum Sq Mean Sq  as.factor(grupna\_var) 1 31010 31010  Residuals 253678 11048380 44  F value Pr(>F)  as.factor(grupna\_var) 712 <2e-16 \*\*\*  Residuals  ---  Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  tukey\_fun(data$BMI,data$HeartDiseaseorAttack)  diff lwr upr p adj  Da-Ne 1.196998 1.109076 1.284921 0 |

Тестови нам потврђују основну сумњу да нема нарочите повезаности, АНОВА нам говори о постојању статистичке значајности малом вредношћу параметра “p” док је F=712 вредност која није нарочито велика, односно релативно је мала. Она нам каже да не постоји нарочита разлика међу класама и просечним вредностима нумеричких варијабли. Што нам „Tukey” и потврђује, износи свега 1.19 за разлику међу класама што је у овом случају занемарио, па се може рећи да иако статистичка повезаност постоји, она је слаба до занемарива.

#### DiffWalk vs PhysActivity

Ову анализу спроводимо услед увиђаја у униваријантној анализи и говори нам о повезаност потешкоћа у кретању са питањем да ли се испитаник бавио физичким активностима у претходних 30 дана. Обе од ове варијабле су категоријске, и обе имају вредност „Da” и „Ne”, па на основу тога знамо да је потребно користити стубичасти дијаграм и топлотну мапу:

|  |
| --- |
| ggplot(data, aes(x = PhysActivity, fill =DiffWalk )) +  geom\_bar(position = "dodge")+  scale\_fill\_manual(values = colors) +  labs(title = "Distribucija DiffWalk u odnosu na PhysActivity",  y = "Broj opservacija",  fill = "Ima poteškoća u hodanju") |
|  |

Иако нам је број опсервација неравномеран по питању физичке активности, може се увидети да су за обе групе број оних који имају потешкоће у кретању идентични, док се за оне који немају потешкоћа ситуација разликује, што је и нормално са обзиром да класе немају исти број записа. Већ овде можемо видети да вероватно постоји повезаност с обзиром да пропорција за оне који имају и немају потешкоћа у ходању није иста за обе класе, али морамо урадити још анализа.

|  |
| --- |
| data %>%  count(PhysActivity, DiffWalk) %>%  group\_by(PhysActivity) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = PhysActivity, y = DiffWalk, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija pojave srčanih oboljenja u odnosu na hipertenziju",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) |
|  |

Топлотна мапа нам потврђује да постоји разлика у односима оних са потешкоћама у шетању у односу на физичку активност. Јасно се види да се односи разликују, јер код оних који су имали рекреативну физичку активност однос оних са потешкоћама у кретању мањи од односа оних који нису имали рекреативне физичке активности, односно 1:8<1:2. Што би рекло да постоји одређена повезаност, коју ћемо сада утврдити употребом тестова.

|  |
| --- |
| chi\_sq\_test(data$PhysActivity,data$DiffWalk)  X-squared = 16259, df = 1, p-value < 2.2e-16  > cramer\_v(data$PhysActivity,data$DiffWalk)  0.2531617 |

На основу урађених тестова добили смо потврду да повезаност итекако постоји, χ² нам кроз параметар те који је близу 0 говори да постоји статистичка повезаност, док је χ² вредност повелика, па би се рекло да постоји зависност између двеју варијабли. Крамеров коефицијент има солидну вредност од 0.25 што нам даје индиције да иако утицај вероватно није велики, довољан је да се узме у разматрање у даљим анализама.

#### MentHlth VS Stroke

Ову анализу радимо више из доменских сазнања, јер се често дешава у пракси да неко са менталним проблемима доживи шлог. Зато ћемо сада упоређивати променљиве MentHlth и Stroke, MentHlth је променљива која нам говори да ли је испитаник имао менталних потешкоћа тј. колико дана у последњих 30 се осећао ментално запослен и она је нумеричка, док нам Stroke показује да ли је испитаник имао мождани удар и она је категоричка, тако да за визуелизацију користимо хистограм и бокс плот:

|  |
| --- |
| ggplot(data, aes(x = Stroke, y = MentHlth, fill = Stroke)) +  geom\_boxplot(alpha = 0.7) +  scale\_fill\_manual(values=colors) +  labs(title = "Distribucija MentHlth u odnosu na Stroke",  fill = "Imao udar") |
|  |

Иако су нам медијане на 0 за оба случаја, односно онде где је испитаник имао и немао шлог практично на 0 што би рекло да преко 50% испитаника није имало менталних потешкоћа, јасно је да је квартална кутија за оне који су имали шлог нешто већа у односу на оне који га нису имали, то сугерише да људи који су га имали имају бар неколика дана више под менталним потешкоћама него што је то случај код оних који га нису имали, постоји доста аутлајера што ће рећи да постоји доста испитаника који имају менталне проблеме без обзира да ли је имао шлог или не.

|  |
| --- |
| ggplot(data, aes(x =MentHlth, fill = Stroke)) +  geom\_histogram(binwidth = 1, position = "dodge", alpha = 0.8) +  scale\_fill\_manual(values =colors) +  labs(title = "Histogram MentHlth u odnosu na Stroke",  y = "Broj opservacija",  fill = "Imao udar") |
|  |

Хистограм нам говори да већина људи из скупа опсервација није имала шлог, и већина није имала менталних потешкоћа, такође се види да је јако мали број оних који су имали шлог у читавом скупу, али то не значи да повезаности нема, како бисмо то утврдили морамо да се позабавимо тестовима који ћемо одрадити.

|  |
| --- |
| anova\_test(data$MentHlth,data$Stroke)  Df Sum Sq Mean Sq  as.factor(grupna\_var) 1 68640 68640  Residuals 253678 13871096 55  F value Pr(>F)  as.factor(grupna\_var) 1255 <2e-16 \*\*\*  Residuals  ---  Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  > tukey\_fun(data$MentHlth,data$Stroke)  Tukey multiple comparisons of means  95% family-wise confidence level  Fit: aov(formula = numericka\_var ~ as.factor(grupna\_var))  $`as.factor(grupna\_var)`  diff lwr upr p adj  Da-Ne 2.636535 2.490686 2.782385 0 |

АНОВА тест нам даје до знања да постоји статистичка повезаност, тиме што је параметар p близу нуле односно 2e-16, F нам говори да постоји одређена разлика између просечних вредности броја дана у класама, како бисмо видели колика је та разлика користимо се “Tukey”-ем који нам говори да је разлика у бројевима дана са менталним потешкоћама између оних којих су имали и нису имали шлог око 2.63 дана, што представља вредност којом се може окарактерисати постојање везе између шлога и менталних проблема,али не толика да је предиктивна.

#### HighChol VS Stroke

На основу доменског знања радимо анализо повезаности високог холестерола и шлога. HighChol нам говори о томе да ли испитаник има висок холестерол и она представља категоричку променљиву, Stroke нам говори да ли је испитаник имао шлог и она је такође категоричка, из тог разлога ћемо употребити субичасти дијаграм и топлотну мапу да утврдимо односе у скупу опсервација:

|  |
| --- |
| ggplot(data, aes(x = Stroke, fill =HighChol )) +  geom\_bar(position = "dodge")+  scale\_fill\_manual(values = colors) +  labs(title = "Distribucija HighChol u odnosu na Stroke",  y = "Broj opservacija",  fill = "Ima holesterol") |
|  |

На основу стубичастих графова иако видимо да се ради о мањку испитаника који су имали шлог, тј. нису већина га није имала, вид се да постоји разлика у односу појаве холестерола међу тим случајевима, наравно боље ћемо ово одредити топлотном мапом.

|  |
| --- |
| data %>%  count(Stroke, HighChol) %>%  group\_by(Stroke) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Stroke, y = HighChol, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija HighChol u odnosu na Stroke",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) |
|  |

Топлотна мапа нам дефинитивно даје потврду да се поретци појаве холестерола разликују по појави шлога, на основу мапе би већ могло да се каже да постоји веза, али сама визуелизација није довољна, па ћемо сада урадити и тестове који то доказују.

|  |
| --- |
| chi\_sq\_test(data$Stroke,data$HighChol)  Pearson's Chi-squared test with Yates'  continuity correction  data: tabela  X-squared = 2175.2, df = 1, p-value < 2.2e-16  > cramer\_v(data$Stroke,data$HighChol)  0.09259986 |

На основу тестова статистичка повезаност постоји, то је јасно на основу “p” параметра из χ² теста, тј. његове ниске вредности, χ² такође нема баш малу вредност, па се може рећи да су варијабле зависне до неког нивоа једна од друге, али не претерано како делује на Топлотној мапи, Крамеров коефицијент такође нема нешто високу вредност, испод 0.1 је, мада није далоко, што значи да су ове две варијабле лабаво повезане, и једино у мултиваријантној анализи би могле да дају нешто јачи резултат.

#### HighChol VS HeartDiseaseorAttack

Анализа се бави провером да ли висок холестерол утиче на појаву срчаних обољења и удара, иако нам доменско каже да то јесте случај, скуп података такође то мора потврдити. Овде се ради о варијабли HighChol која нам говори о томе да ли испитаник има висок холестерол и она представља категоричку променљиву, са друге стране HeartDiseaseorAttack је такође категоричка променљива и она нам говори о томе да ли је испитаник имао срчани удар или напад. Обе имају две категорије односно “Da” и “Ne”, и зато ћемо да их визуализујемо употребом стубичастог графа и топлотне мапе:

|  |
| --- |
| ggplot(data, aes(x = HeartDiseaseorAttack, fill =HighChol )) +  geom\_bar(position = "dodge")+  scale\_fill\_manual(values = colors) +  labs(title = "Distribucija HighChol u odnosu na HeartDisease",  y = "Broj opservacija",  fill = "Ima holesterol") |
|  |

На основу графа види се да имамо дефицит у опсервацијама где је испитаник имао неки проблем са срцем у односу на оне који то нису имали, али је пропорционално број оних који имају холестерол већи у односу на број оних који га нису имали у односу на то да ли је било срчаних проблема или не. Наравно трудимо се да додатним анализама добијемо одговор о повезаности.

|  |
| --- |
| data %>%  count(HeartDiseaseorAttack, HighChol) %>%  group\_by(HeartDiseaseorAttack) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = HeartDiseaseorAttack, y = HighChol, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija HighChol u odnosu na HeartDiseaseorAttack",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) |
|  |

На основу топлотне мапе видимо да постоји разлика у односима да ли неко има холестерол или не у односу на то да ли пати од срчаних обољења. На основу односа виђених на графику може се рећи да повезаности има, али треба узети у обзир пристрасност скупа опсервација па ће нам тестови дати бољи увид:

|  |
| --- |
| chi\_sq\_test(data$HeartDiseaseorAttack,data$HighChol)  Pearson's Chi-squared test with Yates'  continuity correction  data: tabela  X-squared = 8288, df = 1, p-value < 2.2e-16  > cramer\_v(data$HeartDiseaseorAttack,data$HighChol)  0.1807517 |

На основу урађених тестова види се да повезаност појаве холестерола и проблема са срцем постоје. χ² тест је показао вредностима χ²= 8288 и p-value < 2.2e-16 да статистичка повезаност постоји и да постоји одређена међуваријабилна зависност, наравно Крамеров коефицијент нам говори да са вредношћу од 0.1807517 постоји одређени међусобни утицај двеју варијабли али није толики да може бити предикативан, те би било боље ово проверавати у мултиваријатној са још неком варијаблом укљученом.

#### BMI VS Age

У садашњој анализи ћемо проверавити повезаност индекса телесне тежине и старосне доби. У овој анализи радимо са нумеричким варијаблама, те ћемо се послуђити scatter plotom ради визуелне анализе:

|  |
| --- |
| ggplot(data,aes(x=Age,y=BMI))+geom\_point() |
|  |

Дакле на основу дате анализе евидентно је да у свакој од старосних доби имамо доменске изузетке за BMI,односно мањи од 12 и већи од 70, са друге стране, евидентно је и то да за сваку класу највећа концентрација опсервација се налази у опсегу од 20 до 50 за BMI,као и то да је број опсервација за старусну групу 1,10,11,12,13 знатно мања него код осталих група.

|  |
| --- |
| pearson\_funkcija(data$Age,data$BMI)  -0.03661764 |

На основу урађеног теста евидентно је да немамо корелацију између старосне доби и BMI, чак штавише имају тенденцију да иду супротно једна од друге, па се овде може рећи да нема никакве релације.

#### BMI VS PhysHlth

У садашњој анализи ћемо обрађивати корелацију између индекса телесне масе и броја дана са физичким потешкоћама код испитаника, пошто је у доменском знању доказано да постоје извесни физички проблеми код гојазних особа. Код ове анализе имамо посла са бројчаним варијаблама па ћемо користити scatter plot:

|  |
| --- |
| ggplot(data,aes(x=PhysHlth,y=BMI))+geom\_point() |
|  |

На основу scatter-plota евидентно је да имамо мањак у симетрији међу бројевима дана са психичким потешкоћама код испитаника, оно што је до душе уочљиво јесте чињеница да је већина вредности BMI концентрисана управо у распону од 20 до 40. Наравно постоје и изузеци,поготово у доменском смислу који носе вредност преко 70,мада за неке вредности PhysHlth евидентно је да и нема толико изузетака,већ су јасно концентрисани у распону 20-40. Према овом графу рекло би се да постоји комплексна повезаност, мада је ту Пирсонов тест да нам каже више о тој корелацији.

|  |
| --- |
| pearson\_funkcija(data$PhysHlth,data$BMI)  0.1211411 |

На основу Пирсоновог теста видимо да је корелација двеју варијабли слаба, те се може рећи да је ова повезаност употребљива само у корелацији са још неком варијаблом.

#### BMI VS MentHlth

У овој анализи биће упоређивана повезаност међу индексом телесне масе и броја дана са менталним потешкоћама код испитаника. Обе варијабле су нумеричке па ћемо на основу тога употребити scatter plot ради визуализације дистрибуције података:

|  |
| --- |
| ggplot(data,aes(x=MentHlth,y=BMI))+geom\_point() |
|  |

На основу графа видимо да постоји одређени дисбаланс међу бројевима опсервација по данима са менталним проблемима, рекло би се да их највише има код испитаника који их уопште нису имали (0) и оних који су их имали цео месец (30),мада је евидентно да код сваког броја дана једнака сконцентрисаност, од 20 до 40 претежно,мада се појављују доменски изузеци готово сваког дана,односно онде где је BMI преко 70, на основу графика не можемо одредити нарочиту повезаност, али ће нам Пирсонов тест рећи више.

|  |
| --- |
| pearson\_funkcija(data$MentHlth,data$BMI)  0.08531016 |

Пирсонов тест нам даје до знања да корелација изузетно слаба и готово занемарива,па би се рекло да је корелација непостојећа,иако постоје ситни трендови у праћењу међусобно.

#### Age VS PhysHlth

У овој релацији ћемо испитати повезаност старусне доби са физичким потешкоћама код испитаника.Обе од ових варијабли су нумеричке па користимо scatter plot како би визуализовали однос:

|  |
| --- |
| ggplot(data,aes(x=Age,y=PhysHlth))+geom\_point() |
|  |

На основу графа види се да су опсервације равномерно распоређене према старосним групама и да нема одређених коцентрација опсервација по старосним групама, што би рекло да нема нарочитих повезаности, мада нам је потребно употребити Пирсонов тест да потврдимо ову хипотезу.

|  |
| --- |
| pearson\_funkcija(data$Age,data$PhysHlth)  0.09912993 |

На основу резулатата,постоји слаба повезаност,што би значило да ова повезаност може имати предиктиван однос само уколико се употреби са још неком варијаблом.

### Табеле закључака

|  |  |
| --- | --- |
| HighBP vs Diabetes\_012 | Уочено је да не постоји довољно јака веза између високог крвног пристиска и појаве дијабетеса. „chi” тест нам је дао параметре χ² = 18795, a p-value < 0,00000000000000022 , који представљају постојање дате везе између поменутих варијабли, али Крамеров коефицијент (0,27) је готово на самој граници како би HighBP био самостални предиктор, што није задовољавајуће, па је потребно употребити га у мултиваријантној анализи. |
| HighChol vs Diabetes\_012 | Повезаност двеју варијабли постоји али није претерано јака, „chi” тест нам је дао параметре χ² = 11259, a p-value < 0,00000000000000022 који доказују постојање статистичке повезаности и тежњи зависности варијабли,али Крамеров коефицијент у овом случају има оквирну вредност од 0,21 што би рекло да је утицај недовољно јак да би био самостални предиктор, мада ће бити разматран у мултиваријантнон анализи. |
| CholCheck vs Diabetes\_012 | Повезаност двеју варијабли постоји али је релативно слаба, тестови су нам показали резултате који су уопште не могу поставити CholCheck варијаблу као самосталног предиктора, мада су резултати χ² = 1173, a p-value < 0,00000000000000022 такви да говоре о статистичкој значајности, али нам Крамеров коефицијент (0.068) говори да нема неког утицаја,осим ако варијаблу не разматрамо у мултиваријантној анализи. |
| BMI vs Diabetes\_012 | Повезаност BMI са предиктивном варијаблом Diabetes\_012 према АНОВА тесту постоји, тј говори нам да постоје разлика просечних вредности BMI у класама дијабетеса, тј да постоји повезаност између двеју варијабли са резултатом од F value = 6768, Pr(>F) < 2e-16,мада нам “Tukey” говори да однос разлика није исти за све 3 класе дијабетеса,тј највећа разлика је евидентна код људи који имају у немају дијабетес, што је у просеку 4.20, док је за друге односе то нешто мање. Иако знамо да не може бити чист предиктор, BMI ћемо даље испитати у мултиваријантној. |
| GenHlth vs Diabetes\_012 | Анализе и тестови су нам показали да постоји умерена повезаност између GenHlth и Diabetes\_012, тј повезаност је довољна да се даље анализира у мутливаријантној анализи, али није довољна да би GenHlth био самостални предиктор, што нам тестови и доказују. (X-squared = 24248, df = 8, p-value < 2.2e-16) и Крамеров коефицијент који је 0.219,а који нам доказују да постоји извесна повезаност. |
| HeartDiseaseorAttack vs Diabetes\_012 | Анализе и тестови су у овом случају показали да HeartDiseaseorAttack има умерени утицај над Diabetes\_012, тј да он не може бити самостални предиктор, што су тестови потврдили и дали резултате χ² = 8244,9, a p-value < 0,00000000000000022,што нам говори да постоји статистичка значајност и тенденције ка повезаношћу, мада је Крамеров коефицијент релативно низак V = 0,18, тако да ћемо се и овде ослонити на мултиваријантну анализу. |
| Stroke vs Diabetes\_012 | Анализе и тестови су у овом случају показали да Stroke има слабији али не баш безначајан утицај над Diabetes\_012, тј да он не може бити самостални предиктор, што су тестови потврдили и дали резултате χ² = 1173, a p-value < 0,00000000000000022,што нам говори да постоји статистичка значајност и тенденције ка повезаношћу, мада је Крамеров коефицијент слаб V = 0.107, тако да ћемо се овде ослонити на мултиваријантну анализу. |
| MentHlth vs Diabetes\_012 | Проблем са MentHlth варијаблом је та што су опсервације асиметричне, тј много мало људи је имало било какве метналне потешкоће, па ћемо прво морати категорисати ову варијаблу како бисмо је употребили у мултиваријантој анализи, пошто се показало кроз тестове да не постоји готово никаква повезаност међу датим варијаблама, тј не може бити предиктор над Diabetes\_012, тако да једино где бисмо могли да употребимо ову варијаблу јесте у мултиваријантној анализи. |
| PhysHlth vs Diabetes\_012 | Проблем је као и код MentHlth,односно опсервације су тотално асиметричне, па је и овде потребно поделити све по категорији ако би се узимала у обзир кроз мулитваријантну анализу. Разлог зашто не би могла бити самосталан предиктор јесте тај што иако је АНОВА дала добру вреднст F=4079 и 2e-16, тако да нам АНОВА говори да постоји разлика међу класама,мада нам је „Tukey“ дао одговор на то где је разлика, и проблем је тај што разлике нису исте међу свим класама,тако да се ово може употребити само у мултиваријантној. |
| DiffWalk vs Diabetes\_012 | Са становишта анализе и тестова постоји солидна повезаност међу датим варијаблама, тест нам даје резултат X-squared = 12777 p-value < 2.2e-16, што нам говори да постоји статистичка повезаност и висока повезаност класи, мада нам Крамер говори (0.2244245) да немамо нарочити високи утицај, па је боље употребити је у мултиваријантној анализи, него као самостални предиктор. |
| Fruits vs Diabetes\_012 | Са становишта тестова и анализе, променљива Fruits нема готово никакву предиктивну моћ над Diabetes\_012, што нам потврђују и тестови χ² = 454,35, a p-value < 0,00000000000000022, као и низак Крамеров коефицијент V = 0.042,па би се она могла употребити само у мултиваријантној анализи,тј по потреби. |
| Veggies vs Diabetes\_012 | Са становишта тестова и анализе, променљива Veggies нема готово никакву предиктивну моћ над Diabetes\_012, што нам потврђују и тестови Veggies χ²=893,84,a p-value< 2.2e-16, као и низак Крамеров коефицијент V = 0.059,па би се она могла употребити само у мултиваријантној анализи,тј по потреби. |
| Smoker vs Diabetes\_012 | Са становишта тестова и анализе, променљива Smoker нема готово никакву предиктивну моћ над Diabetes\_012, што нам потврђују и тестови χ²=1010,a p-value< 2.2e-16, као и низак Крамеров коефицијент V = 0.068,па би се она могла употребити само у мултиваријантној анализи,тј по потреби. |
| HvyAlcoholConsump vs Diabetes\_012 | Са становишта тестова и анализе, променљива HvyAlcoholConsump нема готово никакву предиктивну моћ над Diabetes\_012, што нам потврђују и тестови χ²=850.32,a p-value< 2.2e-16, као и низак Крамеров коефицијент V = 0.058,па би се она могла употребити само у мултиваријантној анализи,тј по потреби. |
| PhysActivity vs Diabetes\_012 | Овде постоји извесна повезаност, наиме према резултатима тестова и анализи графова види се да је повезаност постојана,мада је резултат тестова (χ² = 3789, p-value< 2.2e) такав да нам говори да та повезаност није нарочито утицајна, и да је варијаблу PhysActivity потребно употребити једино у мултиваријантној анализи,са другим предикторима, јер релативно нема нарочитог утицаја што нам говори ниска Крамерова врередност коефицијента V=0.122 |
| AnyHealthcare vs Diabetes\_012 | Са становишта тестова и анализе, променљива AnyHealthcare нема готово никакву предиктивну моћ над Diabetes\_012, што нам потврђују и тестови χ²=69.078,a p-value< 9.998e-16, као и занемарив Крамеров коефицијент V = 0.016,па би се она могла употребити само у мултиваријантној анализи,тј по потреби. |
| NoDocbcCost vs Diabetes\_012 | Са становишта тестова и анализе, променљива NoDocbcCost нема готово никакву предиктивну моћ над Diabetes\_012, што нам потврђују и тестови χ²=396.08,a p-value< 2.2e-16, као и занемарив Крамеров коефицијент V 0.0395,па би се она могла употребити само у мултиваријантној анализи,тј. по потреби |
| Income vs Diabetes\_012 | На основу тестова види се да постоји умерена повезаност између појаве дијабетеса и нивоа прихода, тест нам је дао резултат χ²=7816.5 и p-value< 2.2e-16, што ће рећи да постоји значајна статистичка повезаност, мада утицај није велики према Крамеровом коефицијенту, тј релативно је низак V = 0.124, тако да ћемо њу више анализирати у склопу мултиваријантне анализе. |
| Sex vs Diabetes\_012 | Са становишта тестова и анализе, променљива за пол испитаника нема готово никакву предиктивну моћ над Diabetes\_012, што нам потврђују и тестови χ²=250.85,a p-value<2.2e-16, као и занемарив Крамеров коефицијент V = 0.031,па би се она могла употребити само у мултиваријантној анализи,тј по потреби. |
| Age vs Diabetes\_012 | Старосна доб је како из доменског знања тако и кроз урађене анализе показала да има релаитвно високу повезаност у односу на појаву дијабетеса, АНОВА тест (F=4560, „p”=2e-16) нам је показао да у односу на класе дијабетеса постоје разлике у старосним добима испитаника, а “Tukey” нам даје увид да су разлике у старосним добима међу класама дијабетеса уочљиве. |
| Education vs Diabetes\_012 | Према урађеној анализи,показало се да постоји слаба повезаност између образовања и дијабетеса, тест нам је показао χ²=4560.6,a p-value<2.2e-16, који нам говоре да постоји статистичка повезаност,али нам Крамер има низак коефицијент V= 0.095, па се нека већа значајност може увидети тек у мултиваријантној анализи. |
| Однос карактеристика које нису циљане |  |
| Income VS Education | На основу анализе утврдило се да постоји извесна повезаност између нивоа образовања и примања испитаника, тест нам је дао поприлично висок резултат χ²=60337,a p-value<2.2e-16,што би рекло да постоји јасна статистичка повезаност, као и јасно одступање од независности међу варијаблама,мада нам Крамеров коефицијент V=0.22 , говори да постоји умерени али не и пресудни утицај међу варијаблама,па је најбоље извршити додатну мултиваријантну анализу. |
| HighChol VS CholCheck | Са становишта тестова и анализе, променљива за CholCheck испитаника нема готово никакву предиктивну моћ над HighChol, што нам потврђују и тестови χ²=1859.7,a p-value<2.2e-16, као и занемарив Крамеров коефицијент V = 0.031,па би се она могла употребити само у мултиваријантној анализи,тј по потреби. |
| HighBP vs HeartDeseaseAttack | У овој анализи, која је урађена на основу доменског знања евидентно је да постоји повезаност, тест нам је дао следећи резултат: χ= 11118,,a p-value<2.2e-16, који нам говоре да постоји статистичка повезаност и да је значајна, док нам Крамер V=0.209 говори да не постоји довољно јак утицај да висок крвни притисак може бити предиктор над срчаним болестима, али у комбинацији са још неким предиктором у мултиваријантној анализи би нам дао прецизнији одговор. |
| BMI vs HeartDeseaseorAttack | У овој анализи се показало да не постоји нарочита повезаност, тј занемарива је, BMI не утиче нарочито на пробеме са срцем и том нам тестови потвђују,АНОВА (712,<2e-16)нам говори да постоји нека разлика у просеку BMI међу испитаницима који имају и немају срчане проблеме,али нам “Tukey” даје занемарујући резултат 1.19,па се ово занемарује. |
| DiffWalk vs PhysActivity | Потешкоће у кретању и обављање физичких активности према тестовима говоре да постоји повезаност, до душе не довољно велика да DiffWalk буде предиктор PhysActivity,тест нам је дао резултат: χ= 16259,,a p-value<2.2e-16, па ће рећи да постоји добра статистичка повезаност , док нам утицај повезаности није довољан да би могао да постоји предикаторски однос, већ се може комбиновати са још предикатора у мултиваријантној (Крамер V=0.25). |
| MentHlth VS Stroke | На основу урађене анализе рекло би се да постоји одређена повезаност између шлога и металних проблема, АНОВА (1255 <2e-16) нам говори о појави разлике у бројевима дана са менталним потешкоћама код оних који су имали шлог и оних који га нису имали,а „Tukey” нам говори да је та разлика око 2.64 дана,што значи да није нарочито предиктиван однос,али у комбинацији са још неким предикатором може добити на значајности. |
| HighChol VS Stroke | На основу урађене анализе висок холестерол нема нарочите везе са појавом шлога код испитаника, иако нам тест χ= = 2175.2,a p-value<2.2e-16 даје резултат који показује статистичку повезаност,утицај је према Крамеровом коефицијенту V=0.093 мали. |
| HighChol VS HeartDiseaseorAttack | На основу урађене анализе висок холестерол је умерено повезан са срчаним проблемима, тест нам показује (χ= = 8288,a p-value<2.2e-16) да постоји извесна статистичка повезаност и није занемарива,али њен утицај (Крамер V=0.18) није довољно јак да би могао постојати предиктиван однос,па је најбоље користити у комбинацији са још неким предикатором |
| BMI vs Age | На основу Пирсоновог теста и дијаграма видимо да нема нарочите повезаности, тј Пирсонов тест нам даје резултат од -0.037, па је јасно да повезаности у овом случају нема. |
| BMI vs PhysHlth | На основу Пиросновог теста и дијаграма се закључује да нема никакве повезаности,Пирсонов тест има слаб резултат од 0.12, што јесте веће од 0.1 али недовољно да се каже да су у чистој релацији,већ се мора употребити у комбинацији са још неким варијаблама. |
| BMI vs MentHlth | Корелација у овом случају је изузетно слаба,Пирсонов тест нам је дао резултат од 0.085 што ће рећи да је повезаност занемарива. |
| Age vs PhysHlth | У овом случају корелација није нарочито постојана,иако је на scatter plot евидентна добра дистрибуција међу старосним групама и данима са физичким потешкоћама. Пирсонов тест нам даје вредност од 0.099,што је приближно релативној вредности 0.1 па се ова повезаност може користити само у комбинацији са још неким варијаблама. |

# Чишћење података

Првобитном униваријантном анализом детектовано је постојање екстремних вредности у карактеристици БМИ. Дефинисане су две врсте екстремних вредности (аутлајера):

* Доменски који обухватају вредности БМИ које нису реалне или су веома ретке са медицинске стране па не представљају репрезентативан узорак (БМИ<12 и БМИ>70) са процентуалним уделом 0.2302113% ,
* Статистички који су детектовани применом IQR методе, са процентним уделом од 3.881662 %.

Даље биваријантном анализом испитан је утицај ових аутлајера на циљану карактеристику Diabetes\_012. Анализа је показала да аутлајери не утичу на циљану променљиву, али како нису репрезентативне, додато удео је <0,5%, донет је закључак да се те опсервације уклоне. За статистичке аутлајере је показало да имају занемарљив утицај на циљану променљиву, али како су доменски реалне вредности и показатељ гојазности одлучено је да буду задржане. Чишћење доменских аутлајера приказано је следећим кодом:

|  |
| --- |
| data\_clean <- data[data$BMI >= 12 & data$BMI <= 70, ] |

Димензионалност скупа пре је била 253 680 опсервација, а сада 253 096.

|  |
| --- |
| print(data.frame(  Skup = c("Originalni podaci", "Nakon čišćenja BMI"),  Broj\_opservacija = c(nrow(data), nrow(data\_clean))  )) |
|  |

# Трансформација података

На основу униваријантне и биваријантне анализе закључено је да карактеристика Age представља старосне групе испитаника, те је стога потребно категорисати је. Пошто садржи велики број категорија (0 до 13) које се тумаче у ординалном поретку, најбоље је користити ординалну факторизацију.

Функцијом јединствености извукли смо све вредности из нумеричке променљиве Age, сортирали са sort, и категоризовали са factor:

|  |
| --- |
| nivoi\_Age = sort(unique(data\_clean$Age))  data\_clean$Age = factor(data\_clean$Age, levels = nivoi\_Age, ordered = TRUE) |

Провера типа:

|  |
| --- |
| > str(data\_clean$Age)  Ord.factor w/ 13 levels "1"<"2"<"3"<"4"<..: 9 7 9 11 11 10 9 11 9 8 ... |

# Инжењеринг карактеристика (Feature Engeenering)

Инжењеринг карактеристика (Feature Engineering) представља процес трансформације постојећих карактеристика или извлачења нових информација кроз комбинације постојећих карактеристика, тако да се резултат бележи као нова карактеристика. Циљ је повећање информативности података, што побољшава предиктивну способност модела и квалитет статистичке анализе.

У нашем раду дефиинисане су нове карактеристике:

* PhysHlthCat
* MentHlthCat
* EducationCat
* IncomeCat
* AgeCat
* CardioRiskScore
* LifestyleRiskScore
* HealthScore
* DietScore
* SocioEconomicStatus

## Инжењеринг карактеристике PhysHlthCat

У униваријантној анализи уочено је да нумеричка карактеристика броја дана са физичким потешкоћама у последњих 30 дана (PhysHlth) има неуравнотежену расподелу са јасно уочљивим категоријама. Даље, биваријантном анализом уочена је подела тих категорија, таква да одржава статистички значај и однос на циљану карактеристику Diabetes\_012. Подела је следећа:

|  |  |
| --- | --- |
| Категорија | Опис |
| 0 | нема проблема |
| 1-5 | благи проблеми |
| 6-15 | умерени проблеми |
| 16-30 | тешки проблеми |

У складу са поделом формирана је карактеристика PhysHlthCat на следећи начин:

|  |
| --- |
| category\_physHlth = c("nema problema", "blagi problemi", "umereni problemi", "teski problemi")  interval\_physHlth = c(0, 5, 15, 30)  data\_clean$PhysHlthCat = NA  data\_clean$PhysHlthCat[data\_clean$PhysHlth == interval\_physHlth[1]] = category\_physHlth[1]  data\_clean$PhysHlthCat[data\_clean$PhysHlth > interval\_physHlth[1] & interval\_physHlth[2] <= 5] = category\_physHlth[2]  data\_clean$PhysHlthCat[data\_clean$PhysHlth > interval\_physHlth[2] & interval\_physHlth[3] <= 15] = category\_physHlth[3]  data\_clean$PhysHlthCat[data\_clean$PhysHlth > interval\_physHlth[3] & interval\_physHlth[4] <= 30] = category\_physHlth[4] |

А затим смо је факторизовали у поретку.

|  |
| --- |
| data\_clean$PhysHlthCat <- factor(data\_clean$PhysHlthCat, levels = category\_physHlth,ordered = TRUE)  > str(data\_clean$PhysHlthCat)  Ord.factor w/ 4 levels "nema problema"<..: 3 1 4 1 1 2 3 1 4 1 … |

## Инжењеринг карактеристике MentHlthCat

У униварјантној анализи уочено је да нумеричка карактеристика броја дана са менталним потешкоћама или стресом у последњих 30 дана (MentHlth) има неуравнотежену расподелу са јасно уочљивим категоријама. Даље, биваријантном анализом уочена је подела тих категорија, таква да одржава статистички значај и однос на циљану карактеристику Diabetes\_012. Подела је следећа:

|  |  |
| --- | --- |
| Категорија | Опис |
| 0 | нема проблема |
| 1-5 | благи проблеми |
| 6-15 | умерени проблеми |
| 16-30 | тешки проблеми |

У складу са поделом формирана је карактеристика MentHlthCat на следећи начин:

|  |
| --- |
| category\_mentHlth = c("nema problema", "blagi problemi", "umereni problemi", "teski problemi")  interval\_mentHlth = c(0, 5, 15, 30)  data\_clean$MentHlthCat = NA  data\_clean$MentHlthCat[data\_clean$MentHlth == interval\_mentHlth[1]] = category\_mentHlth[1]  data\_clean$MentHlthCat[data\_clean$MentHlth > interval\_mentHlth[1] & interval\_mentHlth[2] <= 5] = category\_mentHlth[2]  data\_clean$MentHlthCat[data\_clean$MentHlth > interval\_mentHlth[2] & interval\_mentHlth[3] <= 15] = category\_mentHlth[3]  data\_clean$MentHlthCat[data\_clean$MentHlth > interval\_mentHlth[3] & interval\_mentHlth[4] <= 30] = category\_mentHlth[4] |

А затим смо је факторизовали у поретку.

|  |
| --- |
| data\_clean$MentHlthCat <- factor(data\_clean$MentHlthCat, levels = category\_mentHlth,ordered = TRUE)  > str(data\_clean$MentHlthCat)  Ord.factor w/ 4 levels "nema problema"<..: 4 1 4 1 2 1 1 1 4 1 ... |

## Инжењеринг карактеристике ЕducationCat

У униваријантној анализи уочено је дa расподела карактеристике степена образовања испитаника (Еducation) неуравнотежена па да је потребна другачија структура категорија. Даље, у биваријантној анализи су категорије приказане по категоријама дијабетеса и уочену су границе за спајање категорија. Закључена подела:

|  |  |
| --- | --- |
| Нове категорије | Опсег старих категорија |
| Ниско | Без школе/вртић |
| Основно | Основна школа |
| Средње | 3-годишња и 4-годишња средња школа |
| Високо | Високо образовање |

Формирање нове карактеристике ЕducationCat:

|  |
| --- |
| nivoi\_Education = levels(data\_clean$Education)  data\_clean$EducationCat = factor(data\_clean$Education,  levels = nivoi\_Education,  labels = c("Nisko",  "Osnovno",  "Srednje",  "Srednje",  "Visoko",  "Visoko")  ) |

## Инжењеринг карактеристике IncomeCat

На основу досадашње варијабле Income видимо да постоји много категорија, као и чињеница да број опсервација нису нарочито добро расподељене. На основу свега тога евидентно је да морамо направити категорисјку карактеристику која ће бити заснована на пређашњој Income,која ће бити добро избалансирана и која ће имати мање категорија од садњших 8. Наравно те категорије морају имати смисла, па ћемо се послужити доменским знањем како распоређујемо испитанике по нивоу примања. На основу истраживања за 2015. годину правимо следеће категорије

|  |  |
| --- | --- |
| Нове категорије | Опсег старих категорија |
| Ниска примања | <10.000$,10.000$-14.999$,15.000$-19.999$,20.000$-24.999$ |
| Ниско-средња примања | 25.000$-34.999$,35.000$-49.999$ |
| Средња примања | 50.000$-74999$ |
| Висока примања | 75.000$>= |

Формирање нове карактериситке IncomeCat:

|  |
| --- |
| nivoi\_Income = levels(data\_clean$Income)  data\_clean$IncomeCat = factor(data\_clean$Income,  levels = nivoi\_Income,  labels = c("Niska",  "Niska",  "Niska",  "Niska",  "Nisko-srednja",  "Nisko-srednja",  "Srednja",  "Visoka")) |

На графику видимо дистрибуцију опсервација по класама:

|  |
| --- |
| ggplot(data\_clean, aes(x = IncomeCat, fill = IncomeCat)) +  geom\_bar() +  labs(title = "Distribucija kategorija IncomeCat", x = "Kategorije", y = "Broj opservacija") +  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3 ) +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |

На основу графика видимо да највише има оних са високим примањима,док су нижи и средње нижи у близини један са другим, а средња класа је понајмања,што је по доменском знању и очекивано за САД.

## Инжењеринг карактеристике AgeCat

Као што је већ виђено у униваријантној анализи, са већ постојећу променљиво Age постоји чак 13 категорија, што је много, а приом дистрибуција ње саме се може побољшати. Зато ћемо креирати нову карактеристику AgeCat која ће сажети Аge на свега неколико категорија а да притом начини дистрибуцију категорија бољом. Категорије које креирамо ће бити креиране на основу праћења већ постојећег хистограма тако да имамо следеће категорије:

|  |
| --- |
|  |

На основу хистограма правимо следеће категорије:

|  |  |
| --- | --- |
| Нове категорије | Опсег старих категорија |
| Млади | <=4 |
| Зрели | >=5 && <=8 |
| Старији | >=9 && <=10 |
| Сениори | >10 |

Формирање нове карактериситке AgeCat:

|  |
| --- |
| data\_clean$AgeCat <- factor(data\_clean$Age, levels = 1:13)  levels(data\_clean$AgeCat) <- c(rep("Mladi", 4), rep("Zreli", 4), rep("Stariji", 2), rep("Seniori", 3)) |

На графику видимо дистрибуцију опсервација по класама:

|  |
| --- |
| ggplot(data\_clean, aes(x = AgeCat, fill = AgeCat)) +  geom\_bar() +  labs(title = "Distribucija kategorija AgeCat", x = "Kategorije", y = "Broj opservacija") +  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3 ) +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |

Са графика видимо да највећи удео у опсервацијама имају испитаници који приадају зрелој старсној категорији. Док су сениори и старији прижни,а најмање има у категорији где су млади.

## Инжењеринг карактеристике CardioRiskScore

На основу биваријантне анализе уочили смо следеће јачине повезаности

|  |  |  |
| --- | --- | --- |
| Карактеристике | Крамеров коефицијент | Опис везе |
| HighBP vs Diabetes\_012 | 0,27 | умерена |
| HighChol vs Diabetes\_012 | 0,21 | слаба |
| HeartDiseaseorAttack vs Diabetes\_012 | 0,18 | слаба |
| HighBP vs HeartDeseaseAttack | 0,21 | слаба |
| HighChol vs HeartDiseaseorAttack | 0,18 | слаба |

Дакле, ниједна од ових веза није занемарљива, али ниједна није екстремно јака да би утицала као самостални предиктор. Свака од ових карактеристика (HighBP, HighChol, HeartDisease) мери различити медицински параметар, али заједно чине заједнички здравствени контекст. Зато је одлучено да се на основу њиховог међусобног утицаја формира карактеристика степена кардиолошког ризика (CardioRiskScore).

Карактеристика CardioRiskScore је дефинисана као збир бинарних вредности карактеристика HighBP, HighChol и HeartDiseaseorAttack. Вредности CardioRiskScore крећу се у опсегу од 0 до 3, где већа вредност означава већи број присутних кардиолошких фактора ризика. Изабран је скор приступ уместо бинарне класификације како би се задржала информација о броју присутних фактора ризика.

Формирање је приказано следећим кодом:

|  |
| --- |
| data\_clean$CardioRiskScore = as.numeric(data\_clean$HighBP == "Da") +  as.numeric(data\_clean$HighChol == "Da") +  as.numeric(data\_clean$HeartDiseaseorAttack == "Da") |

Иницијално карактеристике HighBP, HighChol и HeartDiseaseorAttack су бинарне, то значи да категорија „Не“ има вредност 1, а категорија „Да“ има вредност 2. Сабирањем добија се опсег карактеристике CardioRiskScore [3,6], што није интуитивно за ниво ризика. Зато се у коду појављује == "Da", који за категорију „Да“ враћа 1, у супротном („Не“) враћа 0, па је опсег вредности [0,3].

Категоризација у ординарном поретку карактеристике CardioRiskScore:

|  |
| --- |
| nivoi\_CardioRiskScore = sort(unique(data\_clean$CardioRiskScore))  category\_cardioRiskScore = c("nema rizika", "nizak rizik", "umeren rizik", "visok rizik")  data\_clean$CardioRiskScore = factor(data\_clean$CardioRiskScore,  levels = nivoi\_CardioRiskScore,  labels = category\_cardioRiskScore,  ordered = TRUE) |

Провера резултата:

|  |
| --- |
| > str(data\_clean$CardioRiskScore)  Ord.factor w/ 4 levels "nema rizika"<..: 3 1 3 2 3 3 2 3 4 1 ... |

Сада за карактеристику CardioRiskScore важи следећа ординална каратегоризација:

|  |  |  |
| --- | --- | --- |
| Категорија | Вредност | Опис |
| нема ризика | 1 | ниједан од фактора није присутан |
| низак ризик | 2 | један од фактора је присутан |
| умерен ризик | 3 | два фактора су присутна |
| јак ризик | 4 | сва три фактора су присутна |

\*фактори су HighBP, HighChol и HeartDiseaseorAttack, а присутност значи да је њихова вредност „Да“.

Расподела категоријске карактеристике CardioRiskScore приказана је на следећем графику:

|  |
| --- |
| ggplot(data\_clean, aes(x = CardioRiskScore, fill = CardioRiskScore)) +  geom\_bar() +  labs(title = "Distribucija kategorija CardioRiskScore", x = "Kategorije", y = "Broj opservacija") +  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  ) +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |

Са графика видимо да највећи део испитаника има низак и умерен ризик, док је мањи део изложен умереном, а само мало део има висок степен ризика. Ова расподела је у складу са очекивањима наспрам доменског знања.

## Инжењеринг карактеристике LifestyleRiskScore

На основу биваријантне анализе карактеристика *Smoker*, *HvyAlcoholConsump* и *PhysActivity* уочено је да свака од ових променљивих има веома слабу, али статистички значајну повезаност са циљном променљивом *Diabetes\_012:*

|  |  |  |
| --- | --- | --- |
| Карактеристике | Крамеров коефицијент | Опис везе |
| Smoker vs Diabetes\_012 | 0.068 | веома слаба |
| HvyAlcoholConsump vs Diabetes\_012 | 0.058 | веома слаба |
| PhysActivity vs Diabetes\_012 | 0.122 | веома слаба |

Иако самостално свака од ових променљивих носи ограничену информацију о ризику, са доменског аспекта оне заједно карактеришу животни стил испитаника. Зато је одлучено да се формира нова карактеристика *LifestyleRiskScore*, која представља степен ризика животног стила.

Карактеристика *LifestyleRiskScore* дефинисана је као збир бинарних вредности сваке од три карактеристике. Тако да се вредности CardioRiskScore крећу у опсегу од 0 до 3, где већа вредност означава већи број присутних фактора животног стила. Изабран је скор приступ уместо бинарне класификације како би се задржала информација о броју присутних фактора,

Формирање је приказано следећим кодом:

|  |
| --- |
| data\_clean$LifestyleRiskScore = as.numeric(data\_clean$Smoker == "Da") +  as.numeric(data\_clean$HvyAlcoholConsump == "Da") +  as.numeric(data\_clean$PhysActivity == "Da") |

Иницијално карактеристике *Smoker*, *HvyAlcoholConsump* и *PhysActivity* су бинарне, то значи да категорија „Не“ има вредност 1, а категорија „Да“ има вредност 2. Сабирањем добија се опсег карактеристике LifestyleRiskScore [3,6], што није интуитивно за ниво ризика. Зато се у коду појављује == "Da", који за категорију „Да“ враћа 1, у супротном („Не“) враћа 0, па је опсег вредности [0,3].

Категоризација у ординарном поретку карактеристике LifestyleRiskScore:

|  |
| --- |
| nivoi\_LifestyleRiskScore = sort(unique(data\_clean$LifestyleRiskScore))  category\_lifestyleRiskScore = c("nema rizika", "nizak rizik", "umeren rizik", "visok rizik")  data\_clean$LifestyleRiskScore = factor(data\_clean$LifestyleRiskScore,  levels = nivoi\_LifestyleRiskScore,  labels = category\_lifestyleRiskScore,  ordered = TRUE) |

Провера резултата:

|  |
| --- |
| > str(data\_clean$LifestyleRiskScore)  Ord.factor w/ 4 levels "nema rizika"<..: 2 3 1 2 2 3 2 3 2 1 ... |

Сада за карактеристику LifestyleRiskScore важи следећа ординална каратегоризација:

|  |  |  |
| --- | --- | --- |
| Категорија | Вредност | Опис |
| нема ризика | 1 | ниједан од фактора није присутан |
| низак ризик | 2 | један од фактора је присутан |
| умерен ризик | 3 | два фактора су присутна |
| јак ризик | 4 | сва три фактора су присутна |

\*фактори су *Smoker*, *HvyAlcoholConsump* и *PhysActivity* , а присутност значи да је њихова вредност „Да“.

Расподела категоријске карактеристике LifestyleRiskScore приказана је на следећем графику:

|  |
| --- |
| ggplot(data\_clean, aes(x = LifestyleRiskScore, fill = LifestyleRiskScore)) +  geom\_bar() +  labs(title = "Distribucija kategorija LifestyleRiskScore", x = "Kategorije", y = "Broj opservacija") +  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  ) +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |

Са графика видимо да већина испитаника спада у категорије „низак ризик“ и „умерен ризик“, док је мањи број у категорији „нема ризика“, а само мали део испитаника припада категорији „висок ризик“. Ова расподела је у складу са очекивањима на основу доменског знања о факторима животног стила у општој популацији.

## Инжењеринг карактеристике HealthScore

На основу биваријантне анализе карактеристика GenHlth, MentHlthCat и PhysHlthCat уочено је да свака од ових променљивих има одређену повезаност са циљном променљивом Diabetes\_012:

|  |  |  |
| --- | --- | --- |
| Карактеристике | Крамеров коефицијент | Опис везе |
| GenHlth vs Diabetes\_012 | 0.219 | слаба |
| MentHlthCat vs Diabetes\_012 | 0, 0540 | веома слаба |
| PhysHlthCat vs Diabetes\_012 | 0,1255 | веома слаба |

Иако самостално свака од ових променљивих носи ограничену информацију о ризику, са доменског аспекта оне заједно описују опште здравствено стање испитаника. Зато је одлучено да се формира нова карактеристика HealthScore која представља степен општег здравља.

Карактеристика HealthScore дефинисана је као збир вредности сваке од три карактеристике GenHlth, MentHlthCat и PhysHlthCat које су ординалне, што значи да веће вредности указују на лошије здравствено стање.

Формирање је приказано следећим кодом:

|  |
| --- |
| data\_clean$HealthScore = (as.numeric(data\_clean$GenHlth) - 1) +  (as.numeric(data\_clean$PhysHlthCat) - 1) +  (as.numeric(data\_clean$MentHlthCat) - 1) |

Како су карактеристике GenHlth, MentHlthCat и PhysHlthCat вишекатегорисјке променљиве, -1 редукује да се категорије рачунају од 0, како би скор био у валиданом опсегу. Да би уочили интеравле категоризације приказали смо дистрибуцију вредности:

|  |
| --- |
| ggplot(data\_clean, aes(x = HealthScore)) +  geom\_histogram(bins = 50, fill = "#7C4B73FF", color="#381A61FF") +  labs(title = "Distribucija HealthScore", x = "HealthScore", y = "Broj opservacija") |
|  |
|  |

Уочавамо да се вредности крећу од 0 до 10 (4+2+2), са кораком 1. Са графика видимо да већина испитаника има ниске и умерене вредности скоровa (0–6), док је мањи број у категоријама са високим и екстремним скоровима (7–10). Ова расподела је у складу са очекивањима на основу доменског знања о факторима здравља у општој популацији: већина људи има релативно задовољавајуће здравствено стање, док само мали број испитаника показује више присутних фактора ризика. Погледајмо квантиле да би нам границе биле јасније.

|  |
| --- |
| > summary(data\_clean$HealthScore)  Min. 1st Qu. Median Mean 3rd Qu. Max.  0.000 3.000 4.000 3.666 4.000 10.000 |

На основу ове расподеле, карактеристика HealthScore је категоризована у четири ординарне групе:

|  |  |  |
| --- | --- | --- |
| Категорија | Опсег | Опис |
| низак ризик | 0 – 2 | мало присутних фактора |
| умерен ризик | 3 - 5 | умерен број присутних фактора |
| висок ризик | 6 – 8 | већи број присутних фактора |
| екстремни ризик | 9 - 10 | Велика већина фактора је присутни |

Код:

|  |
| --- |
| category\_HealthScore = c("nizak rizik", "umeren rizik", "visok rizik", "ekstremni rizik")  interval\_HealthScore = c(-1, 2, 5, 8, 10)  data\_clean$HealthScore <- cut(data\_clean$HealthScore,  breaks = interval\_HealthScore,  labels = category\_HealthScore,  ordered\_result = TRUE) |

Дистрибуција ових категорија је:

|  |
| --- |
| ggplot(data\_clean, aes(x = HealthScore, fill = HealthScore)) +  geom\_bar() +  labs(title = "Distribucija kategorija HealthScore", x = "Kategorije", y = "Broj opservacija") +  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  ) +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |
|  |

## Инжењеринг карактеристике DietScore

На основу биваријантне анализе уочили смо следеће јачине повезаности

|  |  |  |
| --- | --- | --- |
| Карактеристике | Крамеров коефицијент | Опис везе |
| Veggies vs Diabetes\_012 | 0.059 | слаба |
| Fruits vs Diabetes\_012 | 0.042 | слаба |

Иако су појединачне везе између уноса воћа и поврћа и статуса дијабетеса слабе да би деловале као самостални предиктор, χ² тест је показао да ове везе имају статистичку значајност. Са доменског аспекта, карактеристике Fruits и Veggies описују навике у исхрани, које саме по себи не делују изоловано, већ у комбинацији. Зато је одлучено да се формира карактеристика DietScore, која представља степен квалитета исхране испитаника.

Карактеристика DietScore је дефинисана као збир бинарних вредности карактеристика Fruits и Veggies. Вредности DietScore крећу се у опсегу од 0 до 2, где већа вредност означава здравију исхрану. Изабран је скор приступ уместо бинарне класификације како би се задржала информација о броју присутних фактора.

Формирање је приказано следећим кодом:

|  |
| --- |
| data\_clean$DietScore = as.numeric(data\_clean$Fruits == "Da") +  as.numeric(data\_clean$Veggies == "Da") |

Иницијално карактеристике Fruits и Veggies су бинарне, то значи да категорија „Не“ има вредност 1, а категорија „Да“ има вредност 2. Сабирањем добија се опсег карактеристике DietScore [2,4], што није интуитивно за ниво. Зато се у коду појављује == "Da", који за категорију „Да“ враћа 1, у супротном („Не“) враћа 0, па је опсег вредности [0,2].

Категоризација у ординарном поретку карактеристике DietScore:

|  |
| --- |
| nivoi\_DietScore = sort(unique(data\_clean$DietScore))  category\_DietScore = c("nezdrava", "umereno zdrava", "zdrava")  data\_clean$DietScore = factor(data\_clean$DietScore,  levels = nivoi\_DietScore,  labels = category\_DietScore,  ordered = TRUE) |

Провера резултата:

|  |
| --- |
| > str(data\_clean$DietScore)  Ord.factor w/ 3 levels "nezdrava"<"umereno zdrava"<..: 2 1 2 3 3 3 1 2 3 2 ... |

Сада за карактеристику DietScore важи следећа ординална каратегоризација:

|  |  |  |
| --- | --- | --- |
| Категорија | Вредност | Опис |
| нездрава | 1 | ниједан од фактора није присутан |
| умерено здрава | 2 | један од фактора је присутан |
| здрава | 3 | два фактора су присутна |

\*фактори су Fruits и Veggies, а присутност значи да је њихова вредност „Да“.

Расподела категоријске карактеристике DietScore приказана је на следећем графику:

|  |
| --- |
| ggplot(data\_clean, aes(x = DietScore, fill = DietScore)) +  geom\_bar() +  labs(title = "Distribucija kategorija DietScore", x = "Kategorije", y = "Broj opservacija") +  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  ) +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |

Са графика видимо да највећи део испитаника здраво храни. Однос стубића приказује добру дистрибуцију категорија.

## Инжењеринг карактеристике SocioEconomicStatus

На основу биваријантне анализе уочили смо следеће јачине повезаности са циљном променљивом Diabetes\_012:

|  |  |  |
| --- | --- | --- |
| Карактеристике | Крамеров коефицијент | Опис везе |
| AnyHealthcare vs Diabetes\_012 | 0.016 | занемарљива |
| NoDocbcCost vs Diabetes\_012 | 0.0395 | веома слаба |
| IncomeCat vs Diabetes\_012 | 0.120 | слаба |
| EducationCat vs Diabetes\_012 | 0.079 | веома слаба |
| EducationCat vs IncomeCat | 0.21 | умерена |

Иако појединачне везе ових карактеристика са дијабетесом нису снажне да би деловале као самостални предиктори, χ² тест је показао статистичку значајност. Са доменског аспекта IncomeCat указује на економску моћ испитаника, EducationCat на образовни ресурс и AnyHealthcare/NoDocbcCost приступ здравственом систему. Тако да заједно описују труктурне услове у којима се здравље одржава или нарушава. Због тога је оправдано третирати IncomeCat и EducationCat као једану промељиву SocioEconomicStatus. Променљиве AnyHealthcare и NoDocbcCost показују занемарљив утицај на дијабетес и због тога нису укључене у формирање SocioEconomicStatus карактеристике

EducationCat има категорије: ниско, основно, средње, високо. IncomeCat има категорије: ниска, ниско-средња, средња, висока. Што је показано следећим кодом:

|  |
| --- |
| unique(data\_clean$EducationCat)  unique(data\_clean$IncomeCat) |

Карактеристика SocioEconomicStatus дефинисана је као збир ове две карактеристике, тако да се вредности крећу у опсегу од 0 до 6.

Формирање је приказано следећим кодом:

|  |
| --- |
| data\_clean$SocioEconomicStatus = (as.numeric(data\_clean$EducationCat) - 1) +  (as.numeric(data\_clean$IncomeCat) - 1) |

Како су карактеристике EducationCat и IncomeCat вишекатегорисјке променљиве, -1 редукује да се категорије рачунају од 0, како би скор био у валиданом опсегу. Да би уочили интеравле категоризације приказали смо дистрибуцију вредности:

|  |
| --- |
| ggplot(data\_clean, aes(x = SocioEconomicStatus)) +  geom\_histogram(bins = 50, fill = "#7C4B73FF", color="#381A61FF") +  labs(title = "Distribucija SocioEconomicStatus", x = "SocioEconomicStatus", y = "Broj opservacija") |
|  |

Уочавамо да се вредности крећу од 1 до 6, са кораком 1. Иако скуп није такав може да постоји и вредност 0, па ћемо и њу узети у обзир пликом граница. Са графика видимо да већина испитаника има средње вредности економског статуса, али гледајући појединачно доминира највећи економски статус. Да би границе категорија низак, средњи, висок биле јасније анализирамо квантиле:

|  |
| --- |
| > summary(data\_clean$SocioEconomicStatus)  Min. 1st Qu. Median Mean 3rd Qu. Max.  0.000 3.000 4.000 4.337 6.000 6.00 |

Оно што је прво уочљиво, а на графику није, да постоји и вредност 0. Гледајући остале вредности уочавамо благо асиметричну расподелу. Већина испитаника се налази између 3 и 6, што значи да доминирају средње и високе вредности социоекономског статуса. Закључујемо следећу категоризацију:

|  |  |  |
| --- | --- | --- |
| Категорија | Опсег | Опис |
| низак | 0 – 3 | до 1st Qu. |
| средњи | 4 – 5 | између 1st и 3rd Qu. |
| висок | 6 | 3rd Qu. и горе |

Категоризација у ординарном поретку карактеристике DietScore:

|  |
| --- |
| category\_SocioEconomicStatus = c("nizak", "srednji", "visok")  interval\_SocioEconomicStatus = c(-1, 3, 5, 6)  data\_clean$SocioEconomicStatus <- cut(data\_clean$SocioEconomicStatus,  breaks = interval\_SocioEconomicStatus,  labels = category\_SocioEconomicStatus,  ordered\_result = TRUE) |

Провера резултата:

|  |
| --- |
| > str(data\_clean$SocioEconomicStatus)  Ord.factor w/ 3 levels "nizak"<"srednji"<..: 1 1 2 1 1 3 2 1 1 1 ... |

Сада за карактеристику DietScore важи следећа ординална каратегоризација:

|  |  |  |
| --- | --- | --- |
| Категорија | Вредност | Опис |
| низак | 1 | ниједан од фактора (Income и Education) није присутан у средњем/високом нивоу |
| средњи | 2 | један од фактора је присутан у средњем/високом нивоу |
| висок | 3 | оба фактора су присутна у средњем/високом нивоу |

Расподела категоријске карактеристике SocioEconomicStatus приказана је на следећем графику:

|  |
| --- |
| ggplot(data\_clean, aes(x = SocioEconomicStatus, fill = SocioEconomicStatus)) +  geom\_bar() +  labs(title = "Distribucija kategorija SocioEconomicStatus", x = "Kategorije", y = "Broj opservacija") +  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  ) +  scale\_fill\_paletteer\_d("MetBrewer::Archambault") + theme(legend.position="none") |
|  |

Расподела категорија SocioEconomicStatus показује да је највећи број испитаника у средњој категорији, док ниске и високе категорије имају приближно исти број испитаника. Ово указује на добро избалансиран однос , што омогућава да се различити нивои социоекономског статуса квалитетно анализирају у даљем раду.

# Биваријантна анализа нових карактеристика

## Однос са циљаном карактеристиком Diabetes\_012

Након поглавља инжењерства карактеристика анализирали смо однос тих карактеристика са циљаном променљивом како би добили ове увиде о скупу података.

### PhysHlthCat vs Diabetes\_012

У оквиру ове биваријантне анализе испитиван је однос између категоријске променљиве PhysHlthCat, која описује ниво физичких здравствених потешкоћа испитаника, и циљне променљиве Diabetes\_012, која разликује испитанике без дијабетеса, са предијабетесом и са дијабетесом. PhysHlthCat садржи категорије „нема проблема“, „благи проблеми“, „умерени проблеми“, „тешки проблеми“. Дистрибуцију степена физичких потешкоћа унутар сваке класе дијабетеса приказали смо стубичастим дијаграмом и топлотним дијаграмом.

|  |  |
| --- | --- |
| ggplot(data\_clean, aes(x = PhysHlthCat, fill = Diabetes\_012 )) +  geom\_bar(position = "dodge") +  labs(  title = "Distribucija kategorija PhysHlthCat po kategorijama Diabetes\_012",  y = "Broj opservacija"  ) + scale\_fill\_paletteer\_d("MetBrewer::Archambault") +  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  ) +  theme(legend.position="bottom")  data\_clean %>%  count(Diabetes\_012, PhysHlthCat) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = PhysHlthCat, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija PhysHlthCat po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "PhysHlthCat",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) | |
|  |  |

Стубичасти дијаграм показује јасну разлику у расподели категорија физичког здравља у односу на статус дијабетеса. Код испитаника без дијабетеса доминира категорија „нема проблема“, са знатно већим бројем посматрања у односу на све остале категорије. Како се ниво физичких проблема повећава (од благих ка тешким), број испитаника без дијабетеса опада. Супротно томе, код испитаника са дијабетесом уочава се веће учешће категорија благи, умерени и тешки проблеми у односу на групу без дијабетеса. Иако је и код ове групе најзаступљенија категорија „нема проблема“, релативни удео тежих физичких потешкоћа је израженији него код недијабетичара. Категорија предијабетеса бројчано је мање заступљена у целом узорку, али показује сличан образац.

Дијаграм топлотне мапе омогућава јаснији увид у интензитет односа између посматраних категорија. Најтамније поље јавља се код комбинације *„нема дијабетеса“* и *„нема физичких проблема“*, што указује на високу концентрацију испитаника у овој групи. Иако не постоји савршено линеаран градијент кроз све категорије, јасан је образац повећаног присуства физичких потешкоћа код особа са дијабетесом.

На основу претходних тумачења дајемо претпоставку да између физичких потешкоћа и статуса дијабетеса постоји јасна повезаност. Како графичка анализа има само описни карактер потврдићемо претпоставку статистичким моделима.

|  |
| --- |
| chi\_sq\_test(data\_clean$PhysHlthCat, data\_clean$Diabetes\_012)  cramer\_v(data\_clean$PhysHlthCat, data\_clean$Diabetes\_012) |
|  |

χ² тест независности показује статистички значајну повезаност између променљивих Diabetes\_012 и PhysHlthCat χ² = 7983,5, a p-value < 0,00000000000000022 што је доста испод границе статистичке значајности од 0,05. Јачина ове повезаности износи 0,1255 (Крамеров коефицијент) што је слаба јачина везе. Овај самостални утицај нећемо разматрати, али због постојања статистички значајне везе (χ² тест) узећемо у разматрање мултиваријантне анализе.

### MentHlthCat vs Diabetes\_012

У оквиру ове биваријантне анализе испитиван је однос између категоријске променљиве MentHlthCat, која описује ниво металних здравствених потешкоћа испитаника, и циљне променљиве Diabetes\_012, која разликује испитанике без дијабетеса, са предијабетесом и са дијабетесом. MentHlthCat садржи категорије „нема проблема“, „благи проблеми“, „умерени проблеми“, „тешки проблеми“. Дистрибуцију степена менталних потешкоћа унутар сваке класе дијабетеса приказали смо стубичастим дијаграмом и топлотним дијаграмом.

|  |  |
| --- | --- |
| ggplot(data\_clean, aes(x = MentHlthCat, fill = Diabetes\_012 )) +  geom\_bar(position = "dodge") +  labs(  title = "Distribucija kategorija MentHlthCat po kategorijama Diabetes\_012",  y = "Broj opservacija"  ) + scale\_fill\_paletteer\_d("MetBrewer::Archambault") +  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  ) +  theme(legend.position="bottom")  data\_clean %>%  count(Diabetes\_012, MentHlthCat) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = MentHlthCat, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija MentHlthCat po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "MentHlthCat",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) | |
|  |  |

Стубичасти дијаграм указује на јасне разлике у расподели категорија менталног здравља у односу на статус дијабетеса. Код испитаника без дијабетеса доминира категорија „нема проблема“ (148 972), док се број испитаника постепено смањује са порастом интензитета менталних потешкоћа. Ипак, и у овој групи је присутан значајан број испитаника са благим, умереним и тешким проблемима. Код испитаника са дијабетесом уочава се другачија структура расподеле. Иако је и у овој групи најбројнија категорија „нема проблема“ (23 361), релативно је веће учешће категорија благи, умерени и тешки проблеми у поређењу са испитаницима без дијабетеса. Посебно је приметно да број испитаника са тешким менталним потешкоћама остаје висок и у поређењу са категоријом умерених проблема, што указује на веће оптерећење менталним здрављем код дијабетичара

Дијаграм топлотне мапе визуелно потврђује обрасце уочене на стубичастом дијаграму. Највећи интензитет боје присутан је код комбинације „нема дијабетеса“ и „нема проблема“, што одражава највећи број опсервација у овој групи.

На основу претходних тумачења дајемо претпоставку да између менталних потешкоћа и статуса дијабетеса постоји јасна повезаност. Како графичка анализа има само описни карактер потврдићемо претпоставку статистичким моделима.

|  |
| --- |
| chi\_sq\_test(data\_clean$MentHlthCat, data\_clean$Diabetes\_012)  cramer\_v(data\_clean$MentHlthCat, data\_clean$Diabetes\_012) |
|  |

χ² тест независности показује статистички значајну повезаност између променљивих Diabetes\_012 и MentHlthCat χ² = 1476,6, a p-value < 0,00000000000000022 што је доста испод границе статистичке значајности од 0,05. Јачина ове повезаности износи 0, 0540 (Крамеров коефицијент) што је веома слаба јачина везе. Овај самостални утицај нећемо разматрати, али због постојања статистички значајне везе (χ² тест) узећемо у разматрање мултиваријантне анализе.

### EducationCat vs Diabetes\_012

У оквиру ове биваријантне анализе испитиван је однос између категоријске променљиве EducationCat, која представља степен образовања испитаника, и циљне променљиве Diabetes\_012, која класификује испитанике на оне без дијабетеса, са предијабетесом и са дијабетесом. EducationCat садржи категорије „ниско“, „основно“, „средње“, „високо“. Дистрибуцију степена образовања унутар сваке класе дијабетеса приказали смо стубичастим дијаграмом и топлотним дијаграмом.

|  |  |
| --- | --- |
| ggplot(data\_clean, aes(x = EducationCat, fill = Diabetes\_012 )) +  geom\_bar(position = "dodge") +  labs(  title = "Distribucija kategorija EducationCat po kategorijama Diabetes\_012",  y = "Broj opservacija"  ) + scale\_fill\_paletteer\_d("MetBrewer::Archambault") +  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  ) +  theme(legend.position="bottom")  data\_clean %>%  count(Diabetes\_012, EducationCat) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = EducationCat, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija EducationCat po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "EducationCat",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) | |
|  |  |

Дијаграми показују јасну разлику у расподели нивоа образовања у односу на статус дијабетеса. У свим групама доминирају испитаници са средњим и високим образовањем, што је и очекивано с обзиром на структуру узорка. Највећи број испитаника без дијабетеса припада категорији високог образовања, док је нешто мањи, али и даље значајан број присутан у категорији средњег образовања. Код испитаника са дијабетесом приметан је релативно већи удео особа са средњим и основним образовањем у односу на групу без дијабетеса. Иако апсолутно највећи број дијабетичара и даље има високо образовање, у поређењу са недијабетичарима приметна је промена у структури дистрибуције, са већом заступљеношћу нижих нивоа образовања. Категорија ниског образовања је слабо заступљена у узорку и у свим групама има занемарљив број посматрања, те се на основу ове категорије не могу доносити поуздани закључци.

На основу претходних тумачења дајемо претпоставку да између нивоа образовања и статуса дијабетеса постоји јасна повезаност али недовољно јака да се користи као самостални предиктор.. Како графичка анализа има само описни карактер потврдићемо претпоставку статистичким моделима.

|  |
| --- |
| chi\_sq\_test(data\_clean$EducationCat, data\_clean$Diabetes\_012)  cramer\_v(data\_clean$EducationCat, data\_clean$Diabetes\_012) |
|  |

χ² тест независности показује статистички значајну повезаност између променљивих Diabetes\_012 и *EducationCat* χ² = 3161, a p-value < 0,00000000000000022 што је доста испод границе статистичке значајности од 0,05. Јачина ове повезаности износи 0, 079 (Крамеров коефицијент) што је веома слаба јачина везе. Овај самостални утицај нећемо разматрати, али због постојања статистички значајне везе (χ² тест) узећемо у разматрање мултиваријантне анализе.

Важно је напоменути да је током примене χ² теста добијено упозорење да апроксимација χ² расподелом може бити нетачна. Ово упозорење је последица неравномерне расподеле опсервација по категоријама, односно присуства поља контингентне табеле са малим очекиваним фреквенцијама. Међутим, с обзиром на велику величину узорка, резултат теста и даље пружа валидан индикатор постојања зависности.

### IncomeCat vs Diabetes\_012

У оквиру ове биваријантне анализе ћемо испитивати да ли је ново настала карактеристика IncomeCat добро повезана са Diabetes\_012 пошто је варијабла Income дала солидну повезаност са дијабетесом, одлучили смо да је оптимизујемо и сада тестирамо категоричку променљиву IncomeCat:

|  |
| --- |
| ggplot(data\_clean, aes(x = Diabetes\_012, fill =IncomeCat )) +  geom\_bar(position = "dodge")+  scale\_fill\_manual(values = colors) +  labs(title = "Distribucija IncomeCat u odnosu na PhysActivity",  y = "Broj opservacija",  fill = "Nivo primanja") |
|  |

На основу стубичастог графа,евидентно је да највише опсервација припада испитаницима који немају дијабетес,али је евидентно и то да се пропорционално односи између нивоа примања разликују у односу на то који тип дијабетеса испитаник има.Наравно да бисмо даље одредили однос морамо се послужити и другим анализама.

|  |
| --- |
| data\_clean %>%  count(Diabetes\_012, IncomeCat) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = IncomeCat, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija IncomeCat po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "IncomeCat",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) |
|  |

Топлотна мапа нам је дала бољи увид у односе,иако је евидентно да су бројеви опсервација по типу дијабетеса различити, евиденто је и то да се односи разликују, тј пропорције опсервација по нивоу примања су другачије по свим типовима дијабетеса што нам даје претпоставку да су солдино повезани,па ћемо употребити тестове за бољу анализу.

|  |
| --- |
| chi\_sq\_test(data\_clean$IncomeCat, data\_clean$Diabetes\_012)   Pearson's Chi-squared test  data: tabela X-squared = 7369.1, df = 6, p-value < 2.2e-16  cramer\_v(data\_clean$IncomeCat, data\_clean$Diabetes\_012) 0.1206558 |

Резулати теста нам говоре да повезаност постоји али није уопште јака, тј слаба је.

χ² тест =7369 ,p = 2.2е-16, на основу Хи теста имамо назнаке о повезаности двеју варијабли, са друге стране Крамер нам даје ниску вредност од 0.12 па се опет ова променљива може употребити само за комбинацију са неком додатном варијаблом за значајнију предикцију.

### AgeCat vs Diabetes\_012

У овој анализи ћемо одређивати повезаност новонастале варијабле AgeCat, настале из варијабле Age, како би се постигла боља распоређеност уз мање нивоа старосних доби, са варијаблом Diabetes\_012. AgeCat је катогоричка варијабла па се служимо стубичним графом и топлотном мапон за визуеалну анализу повезаности:

|  |
| --- |
| ggplot(data\_clean, aes(x = Diabetes\_012, fill =AgeCat)) +  geom\_bar(position = "dodge")+  scale\_fill\_manual(values = colors) +  labs(title = "Distribucija AgeCat u odnosu na PhysActivity",  y = "Broj opservacija",  fill = "Starosna dob") |
|  |

На основу Стубичног графа видимо да се број опсервација знатно разликује између класа дијабетеса,мада оно што је такође уочљиво јесте то да су односи међу старосним групама различиити у односу на тип дијабетеса. Зато ћемо урадити још неке анализе да бисмо утврдили да ли повезаност постоји.

|  |
| --- |
| data\_clean %>%  count(Diabetes\_012, AgeCat) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = AgeCat, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija starosne dobi u odnosu na tipove Diabetes\_012",  x = "Diabetes\_012",  y = "AgeCat",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) |
|  |

На основу топлотне мапе рекло би се да повезаност постоји,јер је евидентна промена односа старосних категорија у односу на типове дијабетеса, рецимо оних који имају дијабетес најмање има међу младима, док највише међу старијима, сениорима и зрелим старосним групама. Користимо статисчке тестове да проверимо претпоставку о повезаности.

|  |
| --- |
| data: tabela X-squared = 8705.2, df = 6, p-value < 2.2e-16  > cramer\_v(data\_clean$AgeCat, data\_clean$Diabetes\_012) [1] 0.131139 |

На основу тестова Хи нам говори да статистичка повезаност постоји и да је релативно значајна према вредности χ² тест = 8705 ,p = 2.2е-16. Са друге стране Крамеров коефицијент V= 0.13 нам ипак каже да чак иако повезаност постоји није довљоно јака да буде предиктивна, односно да је потребно користити ову карактеристику у комбинацији са другима како би добили јак предиктиван однос.

### CardioRiskScore vs Diabetes\_012

У оквиру ове биваријантне анализе испитиван је однос између категоријске променљиве CardioRiskScore, која описује степен кардиолошког ризика испитаника, и циљне променљиве Diabetes\_012, која разликује испитанике без дијабетеса, са предијабетесом и са дијабетесом. CardioRiskScore садржи категорије „нема ризика“, „низак ризик“, „умерен ризик“, „јак ризик“. Дистрибуцију степена кардиолошког ризика унутар сваке класе дијабетеса приказали смо стубичастим дијаграмом и топлотним дијаграмом.

|  |  |
| --- | --- |
| ggplot(data\_clean, aes(x = CardioRiskScore, fill = Diabetes\_012 )) +  geom\_bar(position = "dodge") +  labs(  title = "Distribucija kategorija CardioRiskScore po kategorijama Diabetes\_012",  y = "Broj opservacija"  ) + scale\_fill\_paletteer\_d("MetBrewer::Archambault") +  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  ) +  theme(legend.position="bottom")  data\_clean %>%  count(Diabetes\_012, CardioRiskScore) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = CardioRiskScore, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija CardioRiskScore po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "CardioRiskScore",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) | |
|  |  |

Стубичасти дијаграм приказује расподелу категорија CardioRiskScore у односу на статус дијабетеса. Код испитаника без дијабетеса највећи број посматрања је у категорији „нема ризика“, док број испитаника опада како се степен кардиолошког ризика повећава, са најмањим бројем у категорији „висок ризик“. Супротно томе, код испитаника са дијабетесом приметно је учешће категорија „умерен ризик“ и „висок ризик“ у односу на групу без дијабетеса, што указује на присуство више кардиолошких фактора ризика. Испитанци са предијабетесом имају расподелу између две екстремне групе, са већим учешћем категорије „умерен ризик“ у односу на „нема ризика“.

Топлотна мапа пружа јаснији увид у интензитет односа између категорија. Најтамнија поља јављају се код комбинација „нема дијабетеса“ и „нема ризика“, што указује на високу концентрацију испитаника у овој групи, и код „умерен ризик“ и „дијабетес“, што означава да кардиолошки фактори ризика имају већи удео код особа са дијабетесом. Иако није присутан савршено линеаран градијент, види се јасан образац повећаног кардиолошког ризика са стањем дијабетеса.

На основу графичких анализа можемо претпоставити да постоји повезаност између CardioRiskScore и статуса дијабетеса. Како графичка анализа има описни карактер, потврда ове повезаности биће извршена коришћењем статистичких модела.

|  |
| --- |
| chi\_sq\_test(data\_clean$CardioRiskScore, data\_clean$Diabetes\_012)  cramer\_v(data\_clean$CardioRiskScore, data\_clean$Diabetes\_012) |
|  |

χ² тест независности показује статистички значајну повезаност између променљивих Diabetes\_012 и CardioRiskScore χ² = 26242, a p-value < 0,00000000000000022 што је доста испод границе статистичке значајности од 0,05. Јачина ове повезаности износи 0,2276 (Крамеров коефицијент) што је слаба јачина везе.

### LifestyleRiskScore vs Diabetes\_012

У оквиру ове биваријантне анализе испитиван је однос између категоријске променљиве LifestyleRiskScore, која описује степен ризика животног стила испитаника, и циљне променљиве Diabetes\_012, која разликује испитанике без дијабетеса, са предијабетесом и са дијабетесом. CardioRiskScore садржи категорије „нема ризика“, „низак ризик“, „умерен ризик“, „јак ризик“. Дистрибуцију степена ризика животног стила унутар сваке класе дијабетеса приказали смо стубичастим дијаграмом и топлотним дијаграмом.

|  |  |
| --- | --- |
| ggplot(data\_clean, aes(x = LifestyleRiskScore, fill = Diabetes\_012 )) +  geom\_bar(position = "dodge") +  labs(  title = "Distribucija kategorija LifestyleRiskScore po kategorijama Diabetes\_012",  y = "Broj opservacija"  ) + scale\_fill\_paletteer\_d("MetBrewer::Archambault") +  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  ) +  theme(legend.position="bottom")  data\_clean %>%  count(Diabetes\_012, LifestyleRiskScore) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = LifestyleRiskScore, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija LifestyleRiskScore po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "LifestyleRiskScore",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) | |
|  |  |

Код испитаника без дијабетеса највећи број посматрања је у категорији „низак ризик“, док су категорије „нема ризика“ и „умерен ризик“ мање заступљене, а категорија „висок ризик“ има најмањи број испитаника. Код испитаника са дијабетесом приметно је учешће категорија „умерен ризик“ и „висок ризик“ у односу на групу без дијабетеса, што указује на присуство више фактора ризичног животног стила. Испитанци са предијабетесом имају расподелу између категорија „низак ризик“ и „умерен ризик“, са већим учешћем „умерен ризик“ него „низак ризик“.

Топлотна мапа пружа јаснији увид у интензитет односа између категорија. Најтамнија поља јављају се код комбинације „низак ризик“ и „нема дијабетеса“, што указује на високу концентрацију испитаника у овој групи, као и код „умерен ризик“ и „дијабетес“, што означава да фактори ризичног животног стила имају већи удео код особа са дијабетесом. Иако није присутан савршено линеаран градијент, види се јасан образац повећаног ризика животног стила са стањем дијабетеса.

На основу графичких анализа можемо претпоставити да постоји повезаност између LifestyleRiskScore и статуса дијабетеса. Како графичка анализа има описни карактер, потврда ове повезаности биће извршена коришћењем статистичких модела.

|  |
| --- |
| chi\_sq\_test(data\_clean$LifestyleRiskScore, data\_clean$Diabetes\_012)  cramer\_v(data\_clean$LifestyleRiskScore, data\_clean$Diabetes\_012) |
|  |

χ² тест независности показује статистички значајну повезаност између променљивих Diabetes\_012 и LifestyleRiskScore χ² = 1546,1, a p-value < 0,00000000000000022 што је доста испод границе статистичке значајности од 0,05. Јачина ове повезаности износи 0,0552 (Крамеров коефицијент) што је веома слаба јачина везе.

### HealthScore vs Diabetes\_012

У оквиру ове биваријантне анализе испитиван је однос између категоријске променљиве HealthScore, која описује степен општег здравља испитаника, и циљне променљиве Diabetes\_012, која разликује испитанике без дијабетеса, са предијабетесом и са дијабетесом. HealthScore садржи категорије „низак ризик“, „умерен ризик“, „висок ризик“, „екстремни ризик“. Дистрибуцију степена општег здравља унутар сваке класе дијабетеса приказали смо стубичастим дијаграмом и топлотним дијаграмом.

|  |  |
| --- | --- |
| ggplot(data\_clean, aes(x = HealthScore, fill = Diabetes\_012 )) +  geom\_bar(position = "dodge") +  labs(  title = "Distribucija kategorija HealthScore po kategorijama Diabetes\_012",  y = "Broj opservacija"  ) + scale\_fill\_paletteer\_d("MetBrewer::Archambault") +  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  ) +  theme(legend.position="bottom")  data\_clean %>%  count(Diabetes\_012, HealthScore) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = HealthScore, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija HealthScore po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "HealthScore",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) | |
|  |  |

Стубичасти дијаграм приказује јасну разлику у расподели категорија HealthScore у односу на статус дијабетеса. Код испитаника без дијабетеса доминира категорија „умерен ризик“, са убедљиво највећим бројем посматрања, док су категорије „низак ризик“ и „висок ризик“ знатно мање заступљене. Категорија „екстремни ризик“ је код ове групе присутна у занемарљивом броју случајева. Уочава се да са порастом нивоа здравственог ризика опада број испитаника без дијабетеса. Супротно томе, код испитаника са дијабетесом уочава се релативно веће учешће категорија „висок ризик“ и „екстремни ризик“ у поређењу са групом без дијабетеса. Иако је и у овој групи најзаступљенија категорија „умерен ризик“, приметан је већи број испитаника који припадају здравствено ризичнијим категоријама, што указује на лошији укупни здравствени статус код особа са дијабетесом. Категорија предијабетеса је бројчано мање заступљена у целом узорку, али показује интермедијаран образац, односно вредности које се налазе између групе без дијабетеса и групе са дијабетесом.

Дијаграм топлотне мапе омогућава јаснији увид у интензитет односа између категорија HealthScore и статуса дијабетеса. Најтамније поље односи се на комбинацију „нема дијабетеса“ и „умерен ризик“, што указује на високу концентрацију испитаника у овој групи. Са преласком ка категоријама „висок“ и „екстремни ризик“, интензитет боје је израженији код дијабетичара него код недијабетичара, што додатно потврђује уочени образац.

Иако расподела није строго линеарна кроз све категорије, приметан је јасан тренд повећаног здравственог ризика код особа са дијабетесом у односу на особе без дијабетеса. На основу графичке анализе може се претпоставити да између HealthScore категорија и статуса дијабетеса постоји значајна повезаност. С обзиром на то да је ова анализа искључиво описног карактера, наведена претпоставка ће бити додатно испитана и потврђена применом одговарајућих статистичких модела.

|  |
| --- |
| chi\_sq\_test(data\_clean$HealthScore, data\_clean$Diabetes\_012)  cramer\_v(data\_clean$HealthScore, data\_clean$Diabetes\_012) |
|  |

χ² тест независности показује статистички значајну повезаност између променљивих Diabetes\_012 и HealthScore χ² = 4846,8, a p-value < 0,00000000000000022 што је доста испод границе статистичке значајности од 0,05. Јачина ове повезаности износи 0,0978 (Крамеров коефицијент) што је веома слаба јачина везе.

### SocioEconomicStatus vs Diabetes\_012

У оквиру ове биваријантне анализе испитиван је однос између категоријске променљиве SocioEconomicStatus, која описује социјално економски статус, и циљне променљиве Diabetes\_012, која разликује испитанике без дијабетеса, са предијабетесом и са дијабетесом. SocioEconomicStatus садржи категорије „низак“, „средњи“, „висок“. Дистрибуцију социјално економскоог статуса унутар сваке класе дијабетеса приказали смо стубичастим дијаграмом и топлотним дијаграмом.

|  |  |
| --- | --- |
| ggplot(data\_clean, aes(x = SocioEconomicStatus , fill = Diabetes\_012 )) +  geom\_bar(position = "dodge") +  labs(  title = "Distribucija kategorija SocioEconomicStatus po kategorijama Diabetes\_012",  y = "Broj opservacija"  ) + scale\_fill\_paletteer\_d("MetBrewer::Archambault") +  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  ) +  theme(legend.position="bottom")  data\_clean %>%  count(Diabetes\_012, SocioEconomicStatus) %>%  group\_by(Diabetes\_012) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Diabetes\_012, y = SocioEconomicStatus, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija SocioEconomicStatus po kategorijama Diabetes\_012",  x = "Diabetes\_012",  y = "SocioEconomicStatus",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) | |
|  |  |

Стубичасти дијаграм приказује јасне разлике у расподели категорија социоекономског статуса у односу на статус дијабетеса. Код испитаника без дијабетеса доминира категорија „средњи“ социоекономски статус, са убедљиво највећим бројем посматрања, док су категорије „низак“ и „висок“ заступљене у нешто мањем, али и даље значајном обиму. Ова расподела указује да је већина испитаника без дијабетеса концентрисана у средњем социоекономском слоју.

Код испитаника са дијабетесом уочава се другачији образац расподеле. Иако је и у овој групи најзаступљенија категорија „средњи“ социоекономски статус, приметно је релативно веће учешће испитаника са „ниским“ социоекономским статусом у поређењу са групом без дијабетеса, док је категорија „висок“ социоекономски статус најмање заступљена. Овакав образац може указивати на повезаност нижег социоекономског статуса са већом учесталошћу дијабетеса. Категорија предијабетеса је бројчано мање заступљена у целом узорку, али показује интермедијаран образац расподеле. Број испитаника у овој групи налази се између вредности забележених код испитаника без дијабетеса и оних са дијабетесом, при чему је и овде најизраженија категорија „средњи“ социоекономски статус.

Дијаграм топлотне мапе омогућава јаснији увид у интензитет односа између социоекономског статуса и статуса дијабетеса. Најтамнија поља односе се на комбинације „нема дијабетеса“ и „средњи“ социоекономски статус, као и „нема дијабетеса“ и „висок“ социоекономски статус, што указује на високу концентрацију испитаника у овим категоријама. Са преласком ка статусу дијабетеса, интензитет боје је израженији у категорији „низак“ социоекономски статус, што додатно потврђује уочени образац са стубичастог дијаграма.

Иако расподела није строго линеарна кроз све категорије социоекономског статуса, приметан је јасан тренд већег учешћа нижег социоекономског статуса код особа са дијабетесом у односу на особе без дијабетеса. На основу графичке анализе може се претпоставити да између социоекономског статуса и статуса дијабетеса постоји значајна повезаност. С обзиром на то да је анализа описног карактера, ова претпоставка ће бити додатно испитана применом одговарајућих статистичких метода.

|  |
| --- |
| chi\_sq\_test(data\_clean$SocioEconomicStatus, data\_clean$Diabetes\_012)  cramer\_v(data\_clean$SocioEconomicStatus, data\_clean$Diabetes\_012) |
|  |

χ² тест независности показује статистички значајну повезаност између променљивих Diabetes\_012 и SocioEconomicStatus χ² = 6876,8, a p-value < 0,00000000000000022 што је доста испод границе статистичке значајности од 0,05. Јачина ове повезаности износи 0,1165 (Крамеров коефицијент) што је слаба јачина везе.

## Однос карактеристикама које нису циљане

### IncomeCat VS EducationCat

У овој анализи ћемо утврђивати у каквој су релацији новонастале варијабле за ниво образовања и примања испитаника. EducationCat je настао из Education јер је било потребно смањити број нивоа унутар варијабле и притом побољшати баланс међу класама, исто је то случај и код IncomeCat који је настаио из Income, Како су обе категоријске променљиве, тако ћемо корисити Стубични граф и топлотну мапу за визуелизацију:

|  |
| --- |
| ggplot(data\_clean, aes(x = EducationCat, fill =IncomeCat )) +  geom\_bar(position = "dodge")+ scale\_fill\_manual(values = colors) +  labs(title = "Distribucija nivoa primanja u odnosu na nivo obrazovanje", y = "Broj opservacija",   fill = "Primanja") |
|  |

На основу графа видимо да број испитаника који имају ниско и основно образовање,је готово занемарујући, што и не чуди јер је истраживање рађено у развијеној земљи (САД), са друге стране оно што је уочљиво јесте да постоји разлика у односима нивоа примања између средњег и високог нивоа образовања,али за даље утврђивање односа потребне су нам још неке анализе.

|  |
| --- |
| data\_clean %>%  count(EducationCat, IncomeCat) %>%  group\_by(EducationCat) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = EducationCat, y = IncomeCat, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija visine primanja u odnosu na tipove nivo obrazovanja",  x = "EducationCat",  y = "IncomeCat",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) |
|  |

На основу топлотне мапе видимо да постоје разлике у нивоима примања међу свим нивоима образовања, иако је скуп и даље тесно асиметричан. Како бисмо добили потврду о претпоставци повезаности кориситмо статистичке тестове:

|  |
| --- |
| chi\_sq\_test(data\_clean$EducationCat,data\_clean$IncomeCat)  data: tabela X-squared = 35868, df = 9, p-value < 2.2e-16  > cramer\_v(data\_clean$EducationCat,data\_clean$IncomeCat)  0.2173442 |

Као што се види у резултатима,очигледно је да статистичка повезаност и значајно одступање од независтности варијабли постоји, то нам говори мала вредност параметра “p”,али и висока X² вредност, што нам даје јасан увид да имамо јаку повезаност. Наравно Крамеров коефицијент нам говори да иако повезаност постоји, она јесте утицајна међу датим варијабла,али не јако, већ у неком умереном интезитету па би је опет требало користити са још неким предикторима.

### MentHlthCat VS Stroke

У оквиру ове биваријантне анализе испитиван је однос између категоријске променљиве MentHlthCat, која описује ниво менталних потешкоћа укључујући и стрес испитаника, и бинарне категоријске променљиве Stroke, која разликује испитанике који су доживели мождани удар од оних који нису. MentHlthCat садржи категорије „нема проблема“, „благи проблеми“, „умерени проблеми“ и „тешки проблеми“. Дистрибуцију степена менталних потешкоћа унутар сваке класе можданог удара приказали смо стубичастим дијаграмом и топлотним дијаграмом.

|  |  |
| --- | --- |
| ggplot(data\_clean, aes(x = MentHlthCat, fill = Stroke )) +  geom\_bar(position = "dodge") +  labs(  title = "Distribucija kategorija MentHlthCat po kategorijama Stroke",  y = "Broj opservacija"  ) + scale\_fill\_paletteer\_d("MetBrewer::Archambault") +  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  ) +  theme(legend.position="bottom")  data\_clean %>%  count(Stroke, MentHlthCat) %>%  group\_by(Stroke) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = Stroke, y = MentHlthCat, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija MentHlthCat po kategorijama Stroke",  x = "Stroke",  y = "MentHlthCat",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) | |
|  |  |

Стубичасти дијаграм показује јасну диспропорцију у расподели категорија менталног здравља у односу на статус можданог удара. Код испитаника који нису имали мождани удар (Stroke = "Ne") апсолутно доминира категорија „нема проблема“ (169063), са знатно већим бројем посматрања у односу на све остале категорије. Како се ниво менталних проблема повећава, број испитаника без можданог удара нагло опада.

Код испитаника који су доживели мождани удар (Stroke = "Da"), такође је најзаступљенија категорија оних без менталних проблема (6220), али је приметан другачији образац у репу дистрибуције. Удео испитаника са „тешким проблемима“ (1525) је неочекивано висок у односу на „умерене проблеме“ (1024) и „благе проблеме“ (1505) унутар ове групе, што сугерише да особе са историјом можданог удара релативно чешће пријављују озбиљније менталне потешкоће него општа популација.

Дијаграм топлотне мапе омогућава јаснији увид у интензитет ових фреквенција. Најтамније поље јавља се код комбинације нема можданог удара и нема менталних проблема, што указује на високу концентрацију испитаника у овој групи. Међутим, боја указује и на то да, иако су апсолутне бројке код „Stroke = Da“ мале, постоји конзистентно присуство менталних потешкоћа.

На основу претходних тумачења дајемо претпоставку да између менталних потешкоћа и можданог удара постоји одређена повезаност. Како графичка анализа има само описни карактер, потврдићемо претпоставку статистичким моделима.

|  |
| --- |
| chi\_sq\_test(data\_clean$MentHlthCat, data\_clean$Stroke)  cramer\_v(data\_clean$MentHlthCat, data\_clean$Stroke) |
|  |

χ² тест независности показује статистички значајну повезаност између променљивих Stroke и MentHlthCat. Вредност χ² теста износи 1175,9, а p-value < 0,00000000000000022 што је доста испод границе статистичке значајности од 0,05. Јачина ове повезаности износи 0,0682 (Крамеров коефицијент) што је веома слаба јачина везе.

### PhysHlthCat vs DiffWalk

У оквиру ове биваријантне анализе испитиван је однос између категоријске променљиве PhysHlthCat, која описује ниво физичких здравствених потешкоћа испитаника, и променљиве DiffWalk, која указује на то да ли испитаник има озбиљне потешкоће при ходању или пењању уз степенице. PhysHlthCat садржи категорије „нема проблема“, „благи проблеми“, „умерени проблеми“ и „тешки проблеми“. Дистрибуцију степена физичких потешкоћа унутар категорија потешкоћа са кретањем приказали смо стубичастим дијаграмом и топлотним дијаграмом.

|  |  |
| --- | --- |
| ggplot(data\_clean, aes(x = PhysHlthCat, fill = DiffWalk )) +  geom\_bar(position = "dodge") +  labs(  title = "Distribucija kategorija PhysHlthCat po kategorijama DiffWalk",  y = "Broj opservacija"  ) + scale\_fill\_paletteer\_d("MetBrewer::Archambault") +  geom\_text(  stat = "count",  aes(label = ..count..),  position = position\_dodge(width = 0.8),  vjust = -0.3  ) +  theme(legend.position="bottom")  data\_clean %>%  count(DiffWalk, PhysHlthCat) %>%  group\_by(DiffWalk) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = DiffWalk, y = PhysHlthCat, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija kategorija PhysHlthCat po kategorijama DiffWalk",  x = "DiffWalk",  y = "PhysHlthCat",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) | |
|  |  |

Стубичасти дијаграм показује изузетно снажну диференцијацију у физичком статусу испитаника у зависности од тога да ли имају потешкоће са кретањем. Код испитаника који немају потешкоће са ходом (DiffWalk = "Ne"), доминира категорија „нема проблема“ (148334), док број испитаника експоненцијално опада ка тежим категоријама. Насупрот томе, код групе са потешкоћама у кретању (DiffWalk = "Da"), примећујемо потпуно другачији тренд: категорија „тешки проблеми“ је најбројнија (16573), што је више у односу на испитанике са тешким физичким проблемима који немају потешкоће са ходом (9515).

Дијаграм топлотне мапе додатно наглашава ову корелацију. Док је највећа концентрација података (најтамније поље) очекивано код здраве популације, десна колона графика јасно приказује да се интензитет физичких проблема значајно појачава код особа са дијагнозом DiffWalk. Код здравих људи фреквенција опада са тежином проблема, док код особа са потешкоћама у ходу фреквенција расте са тежином физичких проблема.

На основу претходних тумачења дајемо претпоставку да између физичких потешкоћа и потешкоћа са кретањем постоји веома изражена јака повезаност. Како графичка анализа има само описни карактер, потврдићемо претпоставку статистичким моделима.

|  |
| --- |
| chi\_sq\_test(data\_clean$PhysHlthCat, data\_clean$DiffWalk)  cramer\_v(data\_clean$PhysHlthCat, data\_clean$DiffWalk) |
|  |

χ² тест независности показује статистички значајну повезаност између променљивих DiffWalk и PhysHlthCat. Вредност χ² теста износи изузетно високих 56980, а p-value < 0,00000000000000022, што је знатно испод границе статистичке значајности од 0,05. Јачина ове повезаности износи 0,4745 (Крамеров коефицијент), што указује на јаку повезаност између нивоа физичких потешкоћа и потешкоћа у кретању.

Овакав резултат је у складу са доменским знањем, јер се повећање физичких здравствених проблема природно одражава на способност кретања. Због изражене међусобне повезаности, ове променљиве носе сличну информацију и њихово истовремено укључивање у модел може довести до редундантности. Међутим PhysHlthCat је искључена из даље анализе као засебан предиктор јер је већ укључена у степен здравља (HealthScore).

С друге стране, променљива DiffWalk представља последицу здравља који није директно обухваћен HealthScore-ом и показује јачи самостални утицај на циљну променљиву. С тога ће у даљој анализи бити размотрена могућа синергија променљивих DiffWalk и HealthScore, односно њихов заједнички допринос предикцији статуса дијабетеса.

### AgeCat vs HealthScore

У овој анализи ћемо утврђивати повезаност између две новонастале варијабле, односно AgeCat која је настала из Age као потреба за сједињавањем старосних група са 13 на 4, као и HealthScore коју смо добили рачунањем физичке,психичке и генералнe оценe стања испитаника.У основи истражујемо повезаност старосне групе и општег здравственог стања испитаника. Обе од ових варијабли су категоричке,тако да користимо стубичасти дијаграм и топлотну мапу за визуализацију односа:

|  |
| --- |
| ggplot(data\_clean, aes(x = AgeCat, fill =HealthScore )) +    geom\_bar(position = "dodge")+    scale\_fill\_manual(values = colors) +    labs(title = "Distribucija zdravstvenog stanja a u odnosu na starosnu grupu",    y = "Broj opservacija",    fill = "Zdravstveno stanje") |
|  |

На основу стубичастог графа јасно је да је у свим старосним групама однос броја опсервација готово исти, осим код зреле група која мало предњачи са бројем испитаника са ниским здравственим ризицима. Како бисмо даље утврдили однос користимо додатне анализе.

|  |
| --- |
| data\_clean %>%  count(AgeCat, HealthScore) %>%  group\_by(AgeCat) %>%  mutate(percent = n / sum(n) \* 100) %>%  ggplot(aes(x = AgeCat, y = HealthScore, fill = percent)) +  geom\_tile() +  geom\_text(aes(label = paste0(round(percent, 1), "%")), color = "white") +  labs(  title = "Distribucija zdravstvenog stanja a u odnosu na starosnu grupu",  x = "AgeCat",  y = "HealthScore",  fill = "Procenat (%)"  ) +  scale\_fill\_gradient(low = colorS[1], high = colorS[2]) |
|  |

Топлотна мапа нам указује на то да постоје извесне промене у односу здравственог стања међу старосним групама. Па тако евидентноје да је однос између екстремног ризика и ниског ризика код сениора 1:27,док је код младих 1:46, што ће рећи да одређена повезаност постоји,сада колика тачна остаје да се види статичким тестовима.

|  |
| --- |
| chi\_sq\_test(data\_clean$AgeCat,data\_clean$HealthScore) data: tabela X-squared = 1590.4, df = 9, p-value < 2.2e-16  cramer\_v(data\_clean$AgeCat,data\_clean$HealthScore)  0.0457672 |

На основу тестова добили смо слабе резултате. Наиме X²= 1590 што говори о неком слабијем одступању од независтности варијабли,а p има вредност близу 0,па говори да постоји статистичка значајност. Крамер нам даје много слаб резултат,тј 0.046 па би се рекло да нема готовог никаквог утицаја међу варијаблама,па се ова повезаност занемарује,осим у употреби са још неким предиктором.

## Табела закључака

|  |  |
| --- | --- |
| PhysHlthCat vs Diabetes\_012 | Повезаност ових варијабли није нарочито велика, што показује Хи тест χ² = 7983,5,као и p=2.2e-16 који говоре да постоји статистичка повезаност,али нема нарочитог утицаја V=0,12 па је ово најбоље проверавати даље са другим варијаблама. |
| MentHlthCat vs Diabetes\_012 | Повезаност ових варијабли је слаба, иако Хи тест даје резултат од χ² = 1476,као и p=2.2e-16, што би рекло да постоји статистичка повезаност, Крамер коеф. V=0.054 је низак, па остаје само да се можда употреби у комбинацији са другим варијаблама. |
| EducationCat vs Diabetes\_012 | Повезаност ових варијабли је слаба, иако Хи тест даје резултат од χ² = 3161,као и p=2.2e-16, што би рекло да постоји статистичка повезаност, Крамер коеф. V= 0, 079 је низак, па остаје само да се можда употреби у комбинацији са другим варијаблама. |
| IncomeCat vs Diabetes\_012 | Повезаност ових варијабли постоји, али нема јаке везе за предиктиван однос,што су тестови и показали,Хи је дао χ² = 7369,као и p=2.2e-16 што показује повезаност, али Крамер V=0.12 даје нам објашњење да се мора употребљавати само уз још неку варијаблу. |
| AgeCat vs Diabetes\_012 | Повезаност ових варијабли постоји, али нема јаке везе за предиктиван однос,што су тестови и показали,Хи је дао χ² = 8705,као и p=2.2e-16 што показује повезаност, али Крамер V=0.13 даје нам објашњење да се мора употребљавати само уз још неку варијаблу. |
| CardioRiskScore vs Diabetes\_012 | Повезаност ових варијабли је умерено јака,тј. нема предиктиван однос,али има добру статистичку повезаност што је Хи тест показао χ² = 26242,као и p=2.2e-16 што показује повезаност,али Крамер говори да јачина вози није више од умерене (Крамер V=0.23) |
| LifestyleRiskScore vs Diabetes\_012 | Повезаност ових варијабли је слаба, иако Хи тест даје резултат од χ² = 1546,као и p=2.2e-16, што би рекло да постоји статистичка повезаност, Крамер коеф. V= 0, 055 је низак, па остаје само да се можда употреби у комбинацији са другим варијаблама. |
| HealthScore vs Diabetes\_012 | Повезаност ових варијабли је слаба, иако Хи тест даје резултат од χ² =4846,као и p=2.2e-16, што би рекло да постоји статистичка повезаност, Крамер коеф. V= 0, 098 је низак, па остаје само да се можда употреби у комбинацији са другим варијаблама. |
| SocioEconomicStatus vs Diabetes\_012 | Повезаност ових варијабли није нарочито велика, што показује Хи тест χ² = 6877,као и p=2.2e-16 који говоре да постоји статистичка повезаност,али нема нарочитог утицаја V=0,11 па је ово најбоље проверавати даље са другим варијаблама. |
| IncomeCat VS EducationCat | Повезаност ових варијабли је умерено јака,тј. нема предиктиван однос,али има добру статистичку повезаност што је Хи тест показао χ² = 35868,као и p=2.2e-16 што показује повезаност,али Крамер говори да јачина вози није више од умерене (Крамер V=0.22) |
| MentHlthCat VS Stroke | Повезаност ових варијабли је слаба, иако Хи тест даје резултат од χ² = 1176,као и p=2.2e-16, што би рекло да постоји статистичка повезаност, Крамер коеф. V= 0, 068 је низак, па остаје само да се можда употреби у комбинацији са другим варијаблама. |
| PhysHlthCat vs DiffWalk | Повезаност између двеју варијабли постоји, и може бити предиткиван, што нам Хи и Крамер доказују, Хи показао χ² = 56980,као и p=2.2e-16 што је високо и показује да постоји узајамна зависност, а Крамер показује да је веза јака V=0.47 |
| AgeCat vs HealthScore | Повезаност ових варијабли је слаба, иако Хи тест даје резултат од χ² =1590,као и p=2.2e-16, што би рекло да постоји статистичка повезаност, Крамер коеф. V= 0, 046 је низак, па остаје само да се можда употреби у комбинацији са другим варијаблама. |

# Селекција предиктора (Feature Engeenering)

Селекција скупа предиктора представља кључан корак у процесу припреме података, где се за циљ има смањење димензионалности скупа података уз задржавање релевантних информација за предикцију циљане променљиве (Diabetes\_012). Дакле, важи:

{X1​,X2​,…,Xp​} à {Xi1​​,Xi2​​,…,Xik​​} , k < p

На основу следећег кода установили смо да data\_clean скуп садржи 32 карактеристике, од којих је једна циљана:

|  |
| --- |
| > dim(data\_clean)  [1] 253096 32 |

Списак тих карактеристика је:

|  |
| --- |
| names(data\_clean) |
| [1] "Diabetes\_012" "HighBP" "HighChol"  [4] "CholCheck" "BMI" "Smoker"  [7] "Stroke" "HeartDiseaseorAttack" "PhysActivity"  [10] "Fruits" "Veggies" "HvyAlcoholConsump"  [13] "AnyHealthcare" "NoDocbcCost" "GenHlth"  [16] "MentHlth" "PhysHlth" "DiffWalk"  [19] "Sex" "Age" "Education"  [22] "Income" "PhysHlthCat" "MentHlthCat"  [25] "EducationCat" "CardioRiskScore" "LifestyleRiskScore"  [28] "HealthScore" "DietScore" "IncomeCat"  [31] "SocioEconomicStatus" "AgeCat" |

То значи да је почетни скуп предиктора X1 = { HighBP , HighChol , CholCheck, BMI, Smoker, Stroke, HeartDiseaseorAttack , PhysActivity, Fruits, Veggies , HvyAlcoholConsump , AnyHealthcare, NoDocbcCost , GenHlth, MentHlth , PhysHlth , DiffWalk , Sex, Age , Education, Income, PhysHlthCat, MentHlthCat , EducationCat , CardioRiskScore, LifestyleRiskScore, HealthScore, DietScore, IncomeCat, SocioEconomicStatus , AgeCat }, где је p = 31.

Селекција карактеристика је извршена на основу претходне биваријантне анализе и доменског знања. Биваријантна анализа је идентификовала карактеристике које показују статистички значајну везу са циљном променљивом *Diabetes\_012*, док је доменско знање омогућило елиминисање карактеристика које би могле повећати шум или редудансу у моделу. Алтернативно, могли би се применити формални приступи као што су best subset selection, forward, backward или stepwise selection на почетном скупу од 31 карактеристике. Међутим, због великог броја посматрања и карактеристика, ови приступи би били изузетно компјутерски захтевни и у овом случају нису неопходни. Применом биваријантне анализе и доменског знања обезбеђена је стабилност и интерпретабилност модела, а истовремено задовољен критеријум релевантности предиктора.

Критеријуми за селекцију које смо применили су:

* карактеристике са веома слабом везом (Крамеров коефицијен < 0.05) елиминисане су из даље анализе, како би се смањио шум и избегло укључивање нерелевантних информација у модел
* све карактеристике садржане у некој композитној карактеристици нису задржане као самосталне карактеристике, како би се избегла редундантност и преклапање информација
* карактеристике чијом категоризацијом су добијене нове карактеристике нису задржане

Биваријантна анализа је показала да карактеристике *AnyHealthcare*, *NoDocbcCost*, *Sex* и DietScore имају занемарљиву јачину везе са циљном променљивом (Крамеров коефицијент < 0.05) и стога су искључене. Карактеристике садржане у композитним карактеристикама су HighBP, HighChol, GenHlth, HeartDiseaseorAttack, MentHlthCat, PhysHlthCat, Fruits, Veggies, Smoker, HvyAlcoholConsump, PhysActivity, IncomeCat, EducationCat па тако нису задржане у скупу. Карактеристике које су категоризиване MentHlth, PhysHlth, Income, Education, Age нису задржане.

LifestyleRiskScore, CholCheck су карактеристике са веома слабом повезаношћу (0,05 < Крамеров коефицијент < 0,1). Њихово укључивање у модел доводи до увећања шума и потенцијалне редудансе информација, што може смањити стабилност и интерпретабилност модела, стога смо изабрали задржимо само једну. Са аспекта домеснког знања LefestyleScore је погоднији за задржавање.

У складу са претходним закључцима следи да се скуп X1 слика у скуп X2 = {BMI, Stroke, DiffWalk, CardioRiskScore, LifestyleRiskScore, HealthScore, SocioEconomicStatus , AgeCat } где је k=8 задовољавајући услов k < p.

Следећа табела јасније приказује преглед закључака за сваку карактеристику иницијалног скупа X1.

|  |  |  |  |
| --- | --- | --- | --- |
| Карактеристика | Статистички тест (Крамеров коефицијент / χ² тест) | Одлука | Образложење |
| HighBP | 0,27 | Одбачен | Садржан у CardioRiskScore |
| HighChol | 0,21 | Одбачен | Садржан у CardioRiskScore |
| CholCheck | 0.068 | Одбачен | Крамеров коефицијент < 0.05 |
| BMI | Pr(>F) < 2e-16 | Задржан |  |
| GenHlth | 0.219 | Одбачен | Садржан у HealthScore |
| HeartDiseaseorAttack | 0,18 | Одбачен | Садржан у CardioRiskScore |
| Stroke | 0.107 | Задржан |  |
| MentHlth | F = 717.1 Pr(>F) < 2e-16 | Одбачен | Категоризован у MentHlthCat |
| PhysHlth | F = 4079 Pr(>F) < 2e-16 | Одбачен | Категоризован у PhysHlthCat |
| DiffWalk | 0.2244245 | Задржан |  |
| Fruits | 0.042 | Одбачен | Садржан у DietScore |
| Veggies | 0.059 | Одбачен | Садржан у DietScore |
| Smoker | 0.068 | Одбачен | Садржан у LifestyleRiskScore |
| HvyAlcoholConsump | 0.058 | Одбачен | Садржан у LifestyleRiskScore |
| PhysActivity | 0.122 | Одбачен | Садржан у LifestyleRiskScore |
| AnyHealthcare | 0.016 | Одбачен | Крамеров коефицијент < 0.05 |
| NoDocbcCost | 0.0395 | Одбачен | Крамеров коефицијент < 0.05 |
| Income | 0.124 | Одбачен |  |
| Sex | 0.031 | Одбачен | Крамеров коефицијент < 0.05 |
| Age | F=4560 Pr(>F) < 2e-16 | Одбачен | Категоризован у AgeCat |
| Education | 0.095 | Одбачен | Категоризован у EducationCat |
| MentHlthCat | 0.054 | Одбачен | Садржан у HealthScore |
| PhysHlthCat | 0,12 | Одбачен | Садржан у HealthScore |
| IncomeCat | 0.12 | Одбачен | Садржан у SocioEconomicStatus |
| EducationCat | 0, 079 | Одбачен | Садржан у SocioEconomicStatus |
| AgeCat | 0.13 | Задржан |  |
| CardioRiskScore | 0.23 | Задржан |  |
| LifestyleRiskScore | 0, 055 | Задржан |  |
| HealthScore | 0, 098 ≈ 0,1 | Задржан |  |
| DietScore | 0.04477093 | Одбачен | Крамеров коефицијент < 0.05 |
| SocioEconomicStatus | 0,11 | Задржан |  |

Код којим је спроведено формирање скупа X2 :

|  |
| --- |
| selected\_features <- c("BMI", "Stroke", "DiffWalk", "CardioRiskScore",  "LifestyleRiskScore", "HealthScore",  "SocioEconomicStatus", "AgeCat", "Diabetes\_012")  data\_selected <- data\_clean[, selected\_features] |

Прво смо кроз вектор selected\_features узели све називе колона које желимо да укључимо, а затим из скупа података data\_clean узели све редове и изабране колоне формирајући скуп података data\_selected. Димензије скупа су 253096 опсервација са 9 колона:

|  |
| --- |
| > dim(data\_selected)  [1] 253096 9 |

# Моделовање

## Дефинисање тренинг и тест скупа

У процесу изградње модела за анализу и предикцију, један од кључних корака представља адекватна подела скупа података на тренинг и тест скуп. Циљ ове поделе јесте обезбеђивање поузданог учења модела, као и објективне процене његових перформанси над подацима који нису коришћени током фазе обучавања.

Тренинг скуп користи се за откривање образаца и односа између променљивих, односно за изградњу модела, док тест скуп служи као независна основа за проверу способности модела да генерализује резултате на нове, до тада невидљиве податке тј. да предвиђа.

У ууниваријантној анализи уочена је неуравнотежена расподела класа циљане променљиве Diabetes\_012. Овде ћемо поновити резултате како би сагледали структуту података и очували односе приликом поделе на тренинг и тест скуп. Kод:

|  |
| --- |
| raspodela\_Diabetes\_012 = data\_selected %>%  count(Diabetes\_012) %>%  mutate(Udeo = round(n / sum(n) \* 100, 2))  print(raspodela\_Diabetes\_012) |

.резултат

|  |  |  |
| --- | --- | --- |
| Diabetes\_012 | n | Udeo |
| nema dijabetes | 213207 | 84,24 |
| predijabetes | 4619 | 1,82 |
| dijabetes | 35270 | 13,94 |

Потврдили смо присуство неуравнотежене расподеле категорија, из тог разлога примењена је стратификована подела на тренинг и тест скуп, којом се обезбеђује очување релативне заступљености сваке категорије у оба скупа.

Дефинисали смо семе за генератор случаних бројева чиме обезбеђујемо да при сваком покретању кода буду изабрани исти подаци. Индексом (train\_index) одредили смо низ позиција или редова у оригиналном скупу података који ће бити укључени у тренинг скуп. Функција createDataPartition из R пакета caret креира такав индекс на основу циљне променљиве Diabetes\_012 , водећи рачуна о стратификованој подели тј. очувању процентних удела. На основу добијеног индекса поделили смо скуп за тренинг, а преостали подаци иду у тестни скуп. Опис је реализован следећим код:

|  |
| --- |
| set.seed(83762021)  train\_index <- createDataPartition(  data\_selected$Diabetes\_012,  p = 0.8,  list = FALSE  )  data\_train = data\_selected[train\_index, ]  data\_test = data\_selected[-train\_index, ] |

Провера одрживости расподела:

|  |
| --- |
| Табеле расподела |
| raspodela\_trening = data\_train %>%  count(Diabetes\_012) %>%  mutate(Udeo = round(n / sum(n) \* 100, 2))%>%  mutate(Skup = "Trening skup")  raspodela\_test = data\_test %>%  count(Diabetes\_012) %>%  mutate(Udeo = round(n / sum(n) \* 100, 2))%>%  mutate(Skup = "Test skup")  raspodela\_Diabetes\_012 = data\_selected %>%  count(Diabetes\_012) %>%  mutate(Udeo = round(n / sum(n) \* 100, 2))%>%  mutate(Skup = "Ceo skup")  raspodela\_tabele = bind\_rows(raspodela\_trening, raspodela\_test, raspodela\_Diabetes\_012) |

|  |
| --- |
| Графички приказ |
| ggplot(raspodela\_tabele, aes(x = Diabetes\_012, y = Udeo, fill = Skup)) +  geom\_col(position = "dodge") +  labs(  title = "Uporedna raspodela kategorija po skupovima",  y = "Procenat (%)",  x = "Diabetes\_012"  ) + scale\_fill\_paletteer\_d("MetBrewer::Archambault") |
|  |

На основу графика јасно се види да смо поделом задржали расподелу почетног скупа.

## Балансирање података циљане карактеристике Diabetes\_012

У нашем скупу података, циљна променљива Diabetes\_012 представља класе које означавају да ли испитаник има дибетес, предијабетес или га нема уопште.. Униваријантном анализом закључено је да број опсервација у овим категоријама није равнимерно распоређен. Оваква расподела може довести до пристрасности модела тј. модел ће имати тенденцију да боље предвиђа категорију која је више заступљена, док ће ретке категорије бити недовољно предвиђене. Да би модел учио подједнако о свим категоријама циљане променљиве, користе се технике балансирања података, као што су oversampling (понављање или генерисање додатних примера мање заступљених класа) или undersampling (смањење броја примера у доминантној класи).

Користили смо технику која комбинује oversampling мање заступљених класа и undersampling доминантне класе. Oversampling обезбеђује да се за ретке класе креирају нови синтетички примери на основу постојећих, чиме се повећава број примера и омогућава моделу да учи на свим категоријама подједнако. С друге стране, undersampling подразумева смањење броја примера доминантне класе, што може довести до губитка неких информација. Међутим, овај корак је неопходан како би се спречила пристрасност модела ка доминантној класи и обезбедила боља предвидљивост за ретке категорије.

Технике балансирања смо применили на тренинг скуп, како би тест скуп остао репрезентативан и омогућио реалну процену перформанси модела. Тренутна расподела вредности у тренинг скупу је:

|  |
| --- |
| > print(raspodela\_trening)  Diabetes\_012 n Udeo Skup  1 nema dijabetes 170566 84.24 Trening skup  2 predijabetes 3696 1.83 Trening skup  3 dijabetes 28216 13.94 Trening skup |

Како је класа педијабетеса веома ретка 1,83% до 15000 примера је довољно да модел научи о овој класи, без претераног генерисања синтетичких података. Класа нема дијабетес је веома бројна 170566, како би је балансирали у односу на остале класе треба да има дупло више узорака у односу на најмању па је можемо је свети на 30000. Реализација балансирања:

|  |
| --- |
| predijabetes = subset(data\_train, Diabetes\_012 == "predijabetes")  dijabetes = subset(data\_train, Diabetes\_012 == "dijabetes")  nemadijabetes = subset(data\_train, Diabetes\_012 == "nema dijabetes")  set.seed(83762021)  predijabetes\_oversampling = predijabetes[sample(nrow(predijabetes), 15000, replace = TRUE), ]  nemadijabetes\_undersampling = nemadijabetes[sample(nrow(nemadijabetes), 30000), ]  data\_train\_balanced = rbind(predijabetes\_oversampling, dijabetes, nemadijabetes\_undersampling) |

Прво смо издвојили скупове података по класама. Сетовали смо семе за генерисање случајног понављања, обезбеђујући да увек буде исто приликом поновног покретања кода. На скуп са класом предијабетес примењен је oversampling чиме је број приимера повећан на 15000, а за скуп са класом нема дијабетес undersampling чиме је изабрано раддом 30000 примера те класе. За класу дијабетес није рађена техника сходно добром броју примерака. На крају сва три скупа спојена су у балансирани скуп data\_train\_balanced.

Расподела балансираног скупа је:

|  |  |  |
| --- | --- | --- |
| raspodela\_data\_train\_balanced = data\_train\_balanced %>%  count(Diabetes\_012) %>%  mutate(Udeo = round(n / sum(n) \* 100, 2)) | | |
| Diabetes\_012 | n | Udeo |
| nema dijabetes | 30000 | 40,97 |
| predijabetes | 15000 | 20,49 |
| dijabetes | 28216 | 38,54 |

## Регуларизација и РСА

Регуларизација и редукција димензија применом анализе главних компоненти (PCA) нису примењене у овом истраживању из више разлога. Пре свега, скуп предиктора је већ редукован на осам карактеристика применом биваријантне анализе и доменског знања, чиме је елиминисан вишак слабо информативних и редундантних променљивих. С обзиром на велики број посматрања и релативно мали број пажљиво одабраних предиктора, ризик од пренаглашавања (overfitting) је значајно умањен, те примена регуларизације није била неопходна.

Такође, примена PCA није разматрана јер би довела до губитка интерпретабилности модела, што је од посебног значаја у контексту анализе фактора ризика за дијабетес. PCA трансформише оригиналне карактеристике у линеарне комбинације које је тешко директно повезати са конкретним здравственим и социоекономским факторима. С обзиром да је један од циљева рада био разумевање утицаја појединачних предиктора на појаву дијабетеса, задржавање оригиналних карактеристика представља методолошки оправдан избор.

## Тренирање модела

У овом поглављу описујемо процес креирања предиктивних модела за циљану променљиву Diabetes\_012, користећи претходно одабрани скуп предиктора: {BMI, Stroke, DiffWalk, CardioRiskScore, LifestyleRiskScore, HealthScore, SocioEconomicStatus, AgeCat}.

Процес тренирања модела обухвата три кључна корака:

* Избор модела где дефинишемо који алгоритми ће се користити за предикцију и образлажемо њихов избор.
* Поновно узорковање где примењујемо технике као што су stratified cross-validation или други методи за обезбеђивање стабилних и поузданих резултата модела.
* Финално тренирање где обучавамо коначне моделе на целокупном балансираном тренинг скупу који ће се користити за предвиђање на тест скупу.

### Избор модела

Наша предиктивна варијабла јесте Diabetes\_012, она је по природи категоријска, јер нам говори да ли испитаник има или нема дијабетес или је у питању предијабетес. Са тим у вези морамо употребити класификационе моделе. Моделе које ћемо употребити у овом случају су:

* Логистичка регресија- Moдел који предвиђа припадност некој класи, поред тога што предвиђа којој класи припада нека опсервација, он рачуна и вероватноћу припадности те опсервације предвиђеној класи. У нашем случају добро је користити логистичку регресију зато што нам може дати увид у то како неки предиктор попут SocioEconomicStatus утиче на повећање односно смањивање вероватноће да нека опсервација припада управо класи за коју модел тврди да припада. Оно што је такође велики плус код овог модела јесте то што су наши подаци приклоњени овом моделу, тј. користили смо се статистичким методама Крамеров коефицијент и X² тест на које се сама Логистичка регресија и ослања за претпоставке о некој варијабли. Овај модел нам даје статистичку чврстину и коефицијенте,и покушава да направи праволонијску границу између класа.
* Classification and Regression Trees (CART)- CART модел у основи представља стабло одлуке, оно функционише по принципу прага, тј ако је нека вредност изнад неке референтне вредности она се сматра припадницом неке класе, ако не онда није. Са друге стране CART нам може одлично послужити како би визуализовали кретање опсервације кроз испитивања њених вредности.Овај модел за разлику од логистичке регресије покушава да што боље обликује границу међу класама тако да буде јасно дефинисана. Проблем са CART-ом је такав да се он често превише прилагођава подацима, што доводи до поремећаја прагова унутар стабла, и може довести до погрешних претпоставки припадности класи. Зато ћемо користити још један алгоритам написан испод.
* Random forest- Овај модел у основи представља скуп CART стабала, са тим што овде не може доћи до нарочитог overfitting-a,тј прилагођавања стабала подацима. Оно за шта је random forest нарочито добар јесте да би открио скривене везе међу предикторима, тј може нам помоћи да одредимо уз коју варијаблу AgeCat може добити на значајности,као на пример AgeCat+LifestyleRiskScore би био критичан фактор.Са друге стране Random forest усредњује одлуке стотине стабала чиме смањује варијансу у подацима, што је свакако добро по наш скуп који носи различите варијансе по свим варијаблама. Овај модел нам је кључан због тачности предикције,тј максимално ће бити оптимизована.

### Унакрсна валидација (k-fold)

Да бисмо реално проценили колико добро ће модел предвиђати нове податке, користимо методе поновног узорковања. Тест грешка представља просечну грешку модела приликом предвиђања одговора на новим подацима која нису коришћена приликом обуке. Овим добијамо показатељ колико је модел поуздан када се примени на податке који нису део тренинг скупа.

Применили смо k-fold унакрсну валидацију (k-fold cross-validation). Ова метода подразумева да се тренинг скуп података дели на подскупова, при чему се модел више пута тренира на подскупова, а преостали подскуп користи за валидацију. Процена тачности модела добија се просеком резултата из свих итерација.

За овај скуп података изабрано је k = 10 [[референца](#КВалидација)], јер пружа добар компромис између bias-a и варијансе процене. Поред тога, применом стратификоване поделе осигурава се да у сваком скупу пропорције класа циљане променљиве остају приближно једнаке као у целом скупу података.

#### Избор једначине логистичке регресије

За логистичку регресију, k-fold унакрсну валидацију смо користили не само да проценимо грешку, већ и да изаберемо најбољу комбинацију предиктора и коефицијената, тј. да добијемо оптималну једначину логистичке регресије. Модел који даје најмању просечну тест грешку сматрали смо финалним и користили га за предвиђање на тест скупу. Овај приступ осигурава да модел није прекомерно прилагођен тренинг подацима и има добру предиктивну моћ на новим подацима.  
Сви изабрани предиктори укључени су у моделе без додатног одбацивања, с обзиром на то да је њихова повезаност са циљаном променљивом потврђена у оквиру биваријантне анализе и поглавља селекције.

Тестирали смо следеће моделе:

* Модел 1 са свим укљученим параметрима без синергија и полиномијалних чланова
* Модел 2 са укљученом синергијом БМИ и HealthScore (доменска хипотеза)
* Модел 3 са БМИ као полимијалним чланом, испитујући више степена истог

Са доменског аспекта однос БМИ и дијабетеса није линеаран, након одређеног прага гојазности ризик расте брже.

* Модел 4 са БМИ као полимијалним чланом, испитујући више степена истог и синергијским утицајем са HealthScore

Тумачења смо реализовали кроз код:

|  |
| --- |
| set.seed(83762021)  stepen = 1:4 |

|  |
| --- |
| Унакрсна валидација модел 1 |
| model\_1 = glm(  Diabetes\_012 ~ BMI + Stroke + DiffWalk + CardioRiskScore +  LifestyleRiskScore + HealthScore + SocioEconomicStatus + AgeCat,  data = data\_train\_balanced,  family = binomial  )  cv\_error\_model1 = cv.glm(data\_train\_balanced, model\_1, K = 10)$delta[1] |
| Унакрсна валидација модел 2 |
| model\_2 = glm(  Diabetes\_012 ~ BMI \* HealthScore + Stroke + DiffWalk + CardioRiskScore +  LifestyleRiskScore + SocioEconomicStatus + AgeCat,  data = data\_train\_balanced,  family = binomial  )  cv\_error\_model2 = cv.glm(data\_train\_balanced, model\_2, K = 10)$delta[1] |
| Унакрсна валидација модел 3 |
| cv\_error\_model3 = numeric(length(stepen))  for (s in stepen) {  model <- glm(  Diabetes\_012 ~ poly(BMI, s) + Stroke + DiffWalk + CardioRiskScore +  LifestyleRiskScore + HealthScore + SocioEconomicStatus + AgeCat,  data = data\_train\_balanced,  family = binomial  )  cv = cv.glm(data\_train\_balanced, model, K = 10)  cv\_error\_model3[s] = cv$delta[1]  } |
| Унакрсна валидација модел 4 |
| cv\_error\_model4 = numeric(length(stepen))  for (s in stepen) {  model <- glm(  Diabetes\_012 ~ poly(BMI, s)\*HealthScore + Stroke + DiffWalk + CardioRiskScore +  LifestyleRiskScore + SocioEconomicStatus + AgeCat,  data = data\_train\_balanced,  family = binomial  )  cv = cv.glm(data\_train\_balanced, model, K = 10)  cv\_error\_model4[s] = cv$delta[1]  } |

Резултате смо приказали у табели cv\_rezultati\_logisticka:

|  |
| --- |
| cv\_rezultati\_logisticka = data.frame(  Model = c(  "Model 1: linearni",  "Model 2: BMI × HealthScore",  paste0("Model 3: poly(BMI,", stepen, ")"),  paste0("Model 4: poly(BMI,", stepen, ") × HealthScore")  ),  CV\_Error = c(  cv\_error\_model1,  cv\_error\_model2,  cv\_error\_model3,  cv\_error\_model4  )  ) |
|  |

Из добијених резултата видимо да сваки од модела показује сличну CV\_Error вредност (~0.18), што указује да сви модели имају добру предиктивну снагу за наш скуп података. Међутим, најнижу грешку показује Модел 3 са БМИ четвртог степена (CV\_Error = 0.1807798). Ово значи да модел боље хвата нелинеарни однос између BMI и ризика за дијабетес, што је у складу са доменским знањем да се ризик након одређеног прага гојазности брже повећава.

С обзиром на то, за финалну једначину логистиче регресије гласи Diabetes\_012 ~ poly(BMI, 4) + Stroke + DiffWalk + CardioRiskScore + LifestyleRiskScore + HealthScore + СocioEconomicStatus + AgeCat.

#### Процена грешке модела CART и Random Forest

|  |
| --- |
| Дефинисање промељивих и подела на подскупове |
| k = 10  podskupovi <- createFolds(data\_train\_balanced$Diabetes\_012, k = k, list = TRUE, returnTrain = FALSE)  cv\_error\_cart <- numeric(k)  cv\_error\_rf <- numeric(k) |

У обе валидације смо пролазили к пута, узимали i-ти скуп за тест, а остале за тренинг. Тренирали модел функцијом rpart за модел cart и randomForest за модел Random Forest. Пустили смо предикцију над тестним скупом и памтили у вектор грешака средњу вредност. Реализација кроз код:

|  |
| --- |
| CART model |
| for(i in 1:k){  test\_idx <- podskupovi[[i]]  train\_idx <- setdiff(1:nrow(data\_train\_balanced), test\_idx)    model\_cart <- rpart(Diabetes\_012 ~ BMI + Stroke + DiffWalk + CardioRiskScore +  LifestyleRiskScore + HealthScore + SocioEconomicStatus + AgeCat,  data = data\_train\_balanced[train\_idx, ], method = "class")    preds <- predict(model\_cart, data\_train\_balanced[test\_idx, ], type = "class")  cv\_error\_cart[i] <- mean(preds != data\_train\_balanced$Diabetes\_012[test\_idx])  } |
| Random Forest model |
| for(i in 1:k){  test\_idx <- podskupovi[[i]]  train\_idx <- setdiff(1:nrow(data\_train\_balanced), test\_idx)    model\_rf <- randomForest(Diabetes\_012 ~ BMI + Stroke + DiffWalk + CardioRiskScore +  LifestyleRiskScore + HealthScore + SocioEconomicStatus + AgeCat,  data = data\_train\_balanced[train\_idx, ],  ntree = 500)    preds <- predict(model\_rf, data\_train\_balanced[test\_idx, ])  cv\_error\_rf[i] <- mean(preds != data\_train\_balanced$Diabetes\_012[test\_idx])  } |

Резултати унакрне валидације модела CART дати су у облику вектора cv\_error\_cart који садржи вредности грешке за сваку од десет подскупова подела:

|  |
| --- |
| > cv\_error\_cart  [1] 0.4361426 0.4414697 0.4348539 0.4379951 0.4319858 0.4372354 0.4367659 0.4326687  [9] 0.4312252 0.4347173 |

Ови подаци нам показују да CART модел има просечну валидациону грешку од око 0.436, што значи да модел погреши у предвиђању категорије дијабетеса у приближно 43% случајева на подацима који нису коришћени за тренинг. У поређењу са логистичком регресијом (где смо добили CV грешку око 0.18), CART модел показује већу грешку, што указује да овај модел није тако прецизан за наш скуп података.

Резултати унакрне валидације модела Random Forest дати су у облику вектора cv\_error\_rf који садржи вредности грешке за сваку од десет подскупова подела:

|  |
| --- |
| > cv\_error\_rf  [1] 0.4032236 0.4174293 0.4045343 0.4049440 0.4038514 0.4039066 0.4057635 0.4030320  [9] 0.4092337 0.4037148 |

Ови подаци показују да Random Forest модел има просечну валидациону грешку око 0.404, што значи да модел погреши у предвиђању категорије дијабетеса у приближно 40% случајева на независним подацима. У поређењу са CART моделом (CV грешка ~ 0.436) и логистичком регресијом (CV грешка ~ 0.180), Random Forest показује бољу прецизност од CART-а, али је и даље мање прецизан од оптимизованог логистичког модела који користи полиномијалне чланове за BMI. Већа тачност Random Forest-а у односу на CART је очекивана јер Random Forest користи већи број стабала и агрегира предвиђања, што смањује варијансу и чини модел робуснијим. Међутим, у односу на логистичку регресију са поли члановима и доменским хипотезама, Random Forest није дао нижу грешку, што указује да је за овај скуп података најпогоднија комбинација експлицитног укључивања полиномијалних и интеракционих чланова у логистичку регресију.

### Финално тренирање

Сада вршимо тренирање сва три модела на основу data\_train скупа, такво да нам касније на основу data\_test скупа да резултуће класификационе метрике и опише колико су одабрани модели истренирани скупом података заправо добри.

Крећемо од лoгистичке регресије чија једначина гласи:

|  |
| --- |
| lr\_model=Diabetes\_012 ~ poly(BMI, 4) + Stroke + DiffWalk + CardioRiskScore + LifestyleRiskScore + HealthScore + SocioEconomicStatus + AgeCat. |

Креирамо модел са glm funkcijom

|  |
| --- |
| lr\_model = glm( Diabetes\_012 ~ BMI + Stroke + DiffWalk + CardioRiskScore + LifestyleRiskScore + HealthScore + SocioEconomicStatus + AgeCat, data = data\_train\_balanced, family = binomial ) |

Модел смо истренирали са data\_train\_balanced скупом података. Када смо креирали модел, позивамо функцију summary која нам даје увид личну карту модела:

|  |
| --- |
| summary(lr\_model)  Call: glm(formula = Diabetes\_012 ~ BMI + Stroke + DiffWalk + CardioRiskScore +   LifestyleRiskScore + HealthScore + SocioEconomicStatus +   AgeCat, family = binomial, data = data\_train\_balanced)  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) -2.887819 0.058053 -49.745 < 2e-16 \*\*\* BMI 0.084249 0.001569 53.708 < 2e-16 \*\*\* StrokeDa 0.115890 0.042006 2.759 0.0058 \*\*  DiffWalkDa 0.100499 0.025217 3.985 6.74e-05 \*\*\* CardioRiskScore 0.574752 0.010643 54.002 < 2e-16 \*\*\* LifestyleRiskScore.L -0.360793 0.042904 -8.409 < 2e-16 \*\*\* LifestyleRiskScore.Q -0.140054 0.033165 -4.223 2.41e-05 \*\*\* LifestyleRiskScore.C -0.146844 0.019251 -7.628 2.39e-14 \*\*\* HealthScore 0.137799 0.004549 30.294 < 2e-16 \*\*\* SocioEconomicStatus -0.144541 0.006419 -22.518 < 2e-16 \*\*\* AgeCat.L 1.056643 0.025307 41.752 < 2e-16 \*\*\* AgeCat.Q -0.283610 0.020533 -13.813 < 2e-16 \*\*\* AgeCat.C -0.019246 0.016491 -1.167 0.2432  --- Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  (Dispersion parameter for binomial family taken to be 1)   Null deviance: 99100 on 73215 degrees of freedom Residual deviance: 79142 on 73203 degrees of freedom AIC: 79168  Number of Fisher Scoring iterations: 4 |

На основу прегледа основних података о моделу,видимо да су нам све варијабле у PR(>|z|) колони такве да имају вредност \*\*\*,тј мање су од <0.001, што би рекло да су поприлично значајне по предикцију,осим AgeCat.C, тј кубни тренд старости, који није нарочито значајна са вредношћу од 0.2432. Са друге стране ако се погледа колона Estimate видимо како дате варијабле утичу на вероватноћу појаве дијабетеса,оне варијабле које имају позитивну вредност у Estimate колони,утичу тако да се дијабетес појави, док оне са негативном утичу да је вероватноћа појаве дијабетеса мања.

Сада је на реду CART модел:

Он ће такође бити трениран скупом података data\_train\_balanced, креирамо га уз помоћ rpart функције која је део истоимене библиотеке предвиђене управо за креирање CART модела:

|  |
| --- |
| cart\_model <- rpart(Diabetes\_012 ~ BMI + Stroke + DiffWalk + CardioRiskScore +  LifestyleRiskScore + HealthScore + SocioEconomicStatus + AgeCat,  data = data\_train\_balanced,  method = "class") |

Сада када смо креирали модел, употребићемо функцију rpart.plot из истоименог пкаета да добијемо основне податке о самом моделу:

|  |
| --- |
| rpart.plot(cart\_model) |
|  |

CART модел је идентификовао CardioRiskScore као најдоминантнији предиктор. Први праг раздвајања постављен је на вредност 0.5, где је модел са великом сигурношћу (70**%**) идентификовао здраве појединце. Даља гранања преко BMI и HealthScore параметара додатно су прецизирала групу са дијабетесом, док је класа предијабетеса остала најтежа за прецизну класификацију.

Сада ћемо истренирати RF модел:

Њега као и друге моделе тренирамо скупом data\_train\_balanced, креирамо га уз помоћ randomForest функције,која припада истоименој библиотеци, и која је предвиђена за интеракцију са RF моделима:

|  |
| --- |
| library(randomForest)  rf\_model<-randomForest(Diabetes\_012 ~ BMI + Stroke + DiffWalk + CardioRiskScore +  LifestyleRiskScore + HealthScore + SocioEconomicStatus + AgeCat,  data = data\_train\_balanced,  ntree = 500,  importance = TRUE) |

У овом случају, до података о утицајности варијабли унутар модела долазимо мало другачијом командом од претходне две,тј конфузионом матрицом:

|  |
| --- |
| print(rf\_model$confusion)  nema dijabetes predijabetes dijabetes class.error nema dijabetes 21614 161 8225 0.2795333 predijabetes 4818 2589 7593 0.8274000 dijabetes 6732 177 21307 0.2448611 |

Анализа матрице конфузије Random Forest модела открива значајне разлике у прецизности класификације међу групама. Модел је показао високу успешност у идентификацији особа са дијабетесом (грешка од свега 24.5**%**) и здравих појединаца (грешка од 28**%**). Међутим, највећи изазов представља категорија предијабетеса са стопом грешке од 82.7**%**. Већина особа из ове категорије класификована је или као здрава или као оболела, што указује на чињеницу да су биометријске и животне карактеристике (попут БМИ и кардио ризика) код предијабетичара веома сличне граничним вредностима осталих класа. Ово потврђује да је за прецизнију диференцијацију предијабетеса потребно укључити додатне, специфичније клиничке параметре.

## Метрике модела

Сада је потребно ова 3 модела упоредити по њиховим перформансама, за то ћемо употребити data\_test скуп података и употребити стандардне класификационе метрике пошто се ради о класификационим моделима попут Прецизности (Precision), Опозива (Recall),F1 скора.

Прво је потребно креирати предикције за сваки од модела:

|  |
| --- |
| lm\_preds <- predict(lr\_model, newdata = data\_test)  cart\_preds <- predict(cart\_model, newdata = data\_test, type = "class")  rf\_preds <- predict(rf\_model, newdata = data\_test) |

Након креирања датих предикција потребно је креирати конфузиону матрицу на основу њих и правих података:

|  |
| --- |
| nivoi <- levels(as.factor(data\_test$Diabetes\_012))      stvarni\_f <- factor(data\_test$Diabetes\_012, levels = nivoi)  cart\_f <- factor(cart\_preds, levels = nivoi)  rf\_f <- factor(rf\_preds, levels = nivoi)      glm\_preds\_zaokruzeno <- round(lm\_preds)  glm\_preds\_zaokruzeno[glm\_preds\_zaokruzeno < 0] <- 0  glm\_preds\_zaokruzeno[glm\_preds\_zaokruzeno > 2] <- 2  glm\_f <- factor(nivoi[glm\_preds\_zaokruzeno + 1], levels = nivoi)      library(caret)    izvuci\_sve\_metrike <- function(pred, stvarni, ime\_modela) {    cm <- confusionMatrix(pred, stvarni, mode = "everything")      izvestaj <- as.data.frame(cm$byClass[, c("Precision", "Recall", "F1")])  izvestaj$Model <- ime\_modela  izvestaj$Klasa <- nivoi    return(izvestaj)  }      finalna\_tabela <- rbind(  izvuci\_sve\_metrike(glm\_f, stvarni\_f, "Logisticka Regresija"),  izvuci\_sve\_metrike(cart\_f, stvarni\_f, "Stablo (CART)"),  izvuci\_sve\_metrike(rf\_f, stvarni\_f, "Random Forest")  )  finalna\_tabela[is.na(finalna\_tabela)] <- 0  print(finalna\_tabela) |

Како би се омогућила објективна компарација перформанси модела који генеришу различите типове излаза (линеарни log-odds, вероватноће и директне класе), примењен је поступак унификације нивоа предвиђања. За логистичку регресију (ГЛМ) извршено је мапирање континуираних вредности на најближе целобројне категорије (0, 1, 2), док су код модела стабала одлучивања (ЦАРТ) и Random Forest-а излазне лабеле усклађене са референтним вредностима из тест скупа. Финална евалуација спроведена је коришћењем функције confusionMatrix из пакета caret, чиме је осигурано да се метрике Precision, Recall и F1-score рачунају на идентичан начин за сваку од три посматране категорије: одсуство дијабетеса, предијабетес и дијабетес.

|  |
| --- |
|  |

Метрике су показале да је Random Forest најефикаснији модел за класификацију дијабетеса, пружајући најбољи баланс између прецизности и осетљивости (F1 = 0.43). Визуелизација стабла одлучивања потврдила је да су кардиоваскуларни ризик и БМИ кључни биолошки маркери који детерминишу исход. Међутим, немогућност сва три модела да прецизно идентификују предијабетес указује на чињеницу да ова прелазна фаза болести захтева много специфичније клиничке параметре који нису обухваћени овим моделом, што представља значајан простор за даља истраживања.

# Референце

[Скуп податка](https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?resource=download)

[Подаци о америчким нивоима примања код становништва за 2015. годину](https://www.pewresearch.org/social-trends/2015/12/09/the-american-middle-class-is-losing-ground/)

[Подаци истраживања Интернационалне Дијабетес Федерације](https://idf.org/europe/our-network/our-members/serbia/?utm_source=chatgpt.com)

[Истраживање распрострањености дијабетеса код деце у Србији](https://pubmed.ncbi.nlm.nih.gov/29774417/)

[Информације о значењу карактеристика у скупу података](https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators)

[Извор боја за графике](https://r-graph-gallery.com/color-palette-finder)

[Биваријантна анализа](https://medium.com/gen-ai-adventures/bivariate-analysis-exploring-relationships-between-variables-6d4ab1cc84dc)

[Балансирање података](https://www.rdocumentation.org/packages/DMwR/versions/0.4.1/topics/SMOTE)

[Избор вредности к, за к унакрну валидацију](https://cran.r-project.org/web/packages/cvms/vignettes/picking_the_number_of_folds_for_cross-validation.html)