PREDIKCIJA NASTANKA RAKA GRLIĆA MATERICE

1. UVOD

Karcinom grlića materice je maligni tumor koji zahvata grlić materice ili cervikalni kanal. Najčešće se javlja kod žena od treće do pete decenije života i predstavlja drugi najčešći oblik raka kod žena, odmah nakon raka dojke. Svake godine, širom sveta, od ove bolesti oboli približno 500.000 žena. Srbija ima najveću stopu raka grlića materice (24,3 na 100.000 žena) među zemljama bivše Jugoslavije i zauzima drugo mesto u Evropi po učestanosti. U Srbiji se godišnje dijagnostikuje oko 1.500 novih slučajeva, dok oko 500 žena izgubi život zbog ove bolesti.

Cilj ovog projekta je predviđanje raka grlića materice pomoću različitih metoda analize podataka. Postoje mnogi modeli koji mogu biti korisni za rešavanje ovog problema. Ovaj rad će se baviti korišćenjem modela OLS - *Ordinary Least Squares*, WLS - *Weighted Least Squares*, RANSAC - *Random Sample Consensus* kako bi se postiglo što uspešnije predviđanje karcinoma grlića materice. Za ocenjivanje kvaliteta modela biće korišćene metrike *RMSE* i *adjusted R-squared*.

2. BIBLIOTEKE I PODACI KOJI SU KORIŠĆENI

Biblioteke koje će biti korišćene su:

- *Pandas*
- *Sklearn*
- *Numpy*
- *StatsModels*
- *Seaborn*

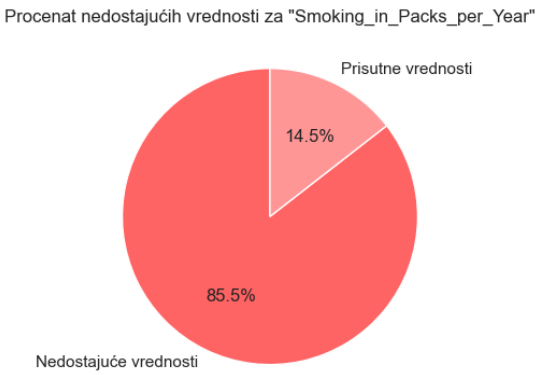
Skup podataka nad kojim će biti primenjeni modeli linearne regresije je *Cervical Cancer Screening* (https://www.johnsnowlabs.com/marketplace/cervical-cancer-screening/). Dataset sadrži 36 kolona (atributa) i 858 redova (ispitanih pacijenatkinja). Kolona koja će biti prediktovana je *Is_Diagnosis_Cancer* koja daje informaciju o tome da li je pacijentkinji dijagnostikovan rak grlića materice ili nije. Najznačajnije nezavisne promenljive su:

- *Age_of_Respondents* - godine pacijentkinje
- *Number_of_Sexual_Partners* - broj partnera
- *First_Sexual_Intercourse* - godine kada je imala prvo seksualno iskustvo
- *Number_of_Pregnancies* - broj trudnoća
- *Is_Smoking* - da li je pacijentkinja pušač ili ne

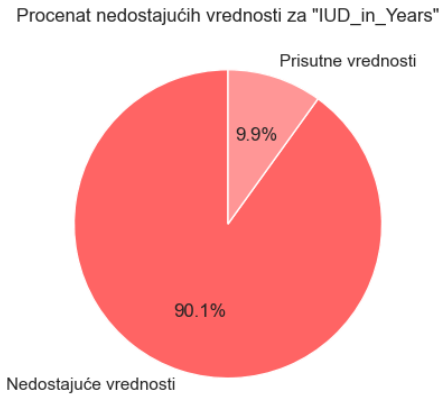
- *Smoking_in_Years* - broj godina koji pacijentkinja puši
- *Smoking_in_Packs_per_Year* - broj kutija cigareta koji konzumira u godini
- *Is_On_Hormonal_Contraceptives* - da li koristi kontraceptivne pilule
- *Hormonal_Contraceptives_in_Years* - broj godina koji konzumira kontraceptivne pilule
- *Is_On_IUD* - da li je koristila spiralu ili ne
- *Is_Diagnosis_HPV* - da li je zaražena HPV-om
- *Is_Diagnosed_with_STDs* - da li je pacijentkinja bolovala od polnih bolesti

3. METODOLOGIJA

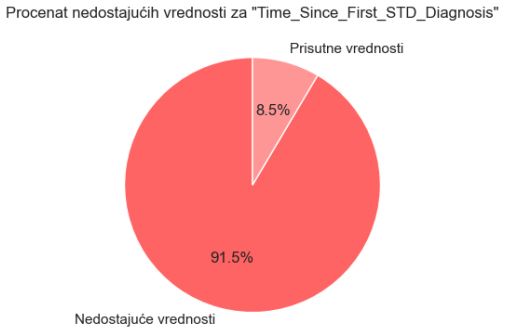
Zbog velikog procenta nedostajućih podataka, u kolonama atributa: *Number_of_STD_Diagnosis, Time_Since_First_STD_Diagnosis, Time_Since_Last_STD_Diagnosi, Number_of_Years_with_STD, Smoking_in_Years, Smoking_in_Packs_per_Year, IUD_in_Years* bilo je razmatrano njihovo uklanjanje.



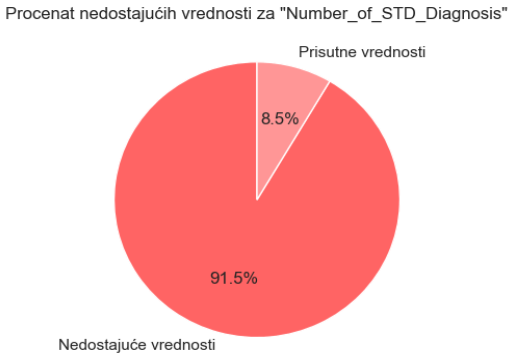
Slika 1 - Nedostajuće vrednosti za kolonu Smoking_in_Packs_per_Year



Slika 2 - Nedostajuće vrednosti za kolonu IUD_in_Years

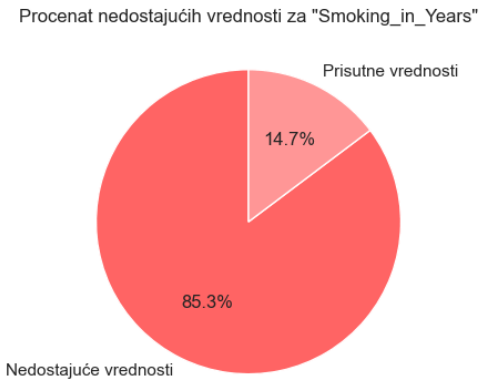


Slika 3 - Nedostajuće vrednosti za kolonu Time_Since_First_STD_Diagnosis

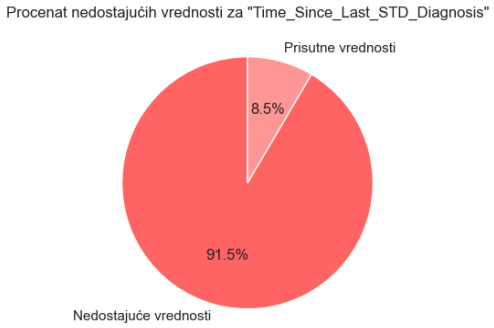


Slika 4 - Nedostajuće vrednosti za kolonu Number_of_STD_Diagnosis

Slika 5 - Nedostajuće vrednosti za kolonu Number_of_Years_with_STD



Slika 6 - Nedostajuće vrednosti za kolonu Smoking_in_Years



Slika 7 - Nedostajuće vrednosti za kolonu Time_Since_Last_STD_Diagnosis

Da bi se utvrdio uticaj ovih nezavisnih promenljivih na zavisnu varijablu određena je matrica korelacije.

Number_of_STD_Diagnosis -0.023611
Time_Since_First_STD_Diagnosis 0.201319
Time_Since_Last_STD_Diagnosis 0.212983
Number_of_Years_with_STDs -0.160779
IUD_in_Years 0.041261
Is_Diagnosis_Cancer 1.000000
Smoking_in_Years 0.271453
Smoking_in_Packs_per_Year 0.391916

Od predloženih kolona za uklanjanje ostavljena je jedino *Smoking_in_Packs_per_Year,* zbog većeg koeficijenta korelacije.

Potom je izvršena obrada nedostajućih vrednosti. Popunjavanje nedostajućih vrednosti je izvršeno primenom KNN imputacije. Kako bi se primenila KNN imputacija podataka bilo je neophodno izvršiti pretvaranje bool tipova u brojne vrednosti tj. true/false u 1/0 i bilo je važno da se izvrši normalizacija ostalih brojnih vrednosti. KNN imputacijom se nedostajucće vrednosti popunjavaju izračunavanjem srednje vrednosti parametara određenog broja najbližih suseda koji su pronađeni. Po default-u se vrednosti popunjavaju primenom Euklidskog rastojanja. KNN imputacija može da bude korisna u ovakvim medicinskim ispitivanjima, jer na osnovu sličnosti u kliničkim slikama pacijentkinja podaci bivaju popunjeni. Ono na šta je bilo neophodno obratiti pažnju je odabir broja suseda, K, na osnovu kojih će se računati srednja vrednost. Manje vrednosti broja suseda mogu biti podložne većem uticaju šuma i mogu dovesti do pretreniranosti modela, dok veće vrednosti K mogu smanjiti osetljivost modela na specifične promene u podacima. S toga su se posmatrale vrednosti metrika za različite vrednosti K kako bi se odabrala najpogodnija vrednost. [3] [4]

Za normalizaciju je upotrebljen MinMaxScaler, jer algoritam KNN imputacije zavisi od udaljenosti između tačaka, a različiti opsezi vrednosti mogu značajno uticati na ove udaljenosti. MinMaxScaler takođe osigurava da svi atributi imaju isti raspon vrednosti, što utiče na bolju procenu nedostajućih vrednosti. Ne samo da se upotrebom MinMaxScaler-a čuva originalna distribucija podataka, već se očuvava i važnost maksimalnih i minimalnih vrednosti atributa za spcifične pacijentkinje. Formula po kojoj se računa normalizovana vrednost podatka je:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Slika 8 - Formula za skaliranje MinMaxScaler-om

Ono što je mana ovakvog pristupa je što bi se unošenjem novih podataka u dataset moralo vršiti skaliranje skroz od početka, jer bi bilo neophodno ponovno izračunavanje minimalnih i maksimalnih vrednosti. Još jedna eventualna mana je osetljivost na outlier-e, jer oni imaju značajan uticaj na minimalne i maksimalne vrednosti. [1] [2]

Outlier je podatak koji značajno odstupa od ostalih podataka. Može biti ili mnogo veći ili mnogo manji od ostalih vrednosti, a njegovo prisustvo može značajno uticati na rezultate algoritama mašinskog učenja. Izolovane vrednosti često nastaju zbog grešaka u merenju.

Postoje dva glavna tipa outliera:

1. Globalni outlieri:
    - Globalni outlieri su izolovani podaci koji su daleko od glavnog skupa podataka.
    - Oni su često lako uočljivi i uklonjivi.

2. Kontekstualni outlieri:
    - Kontekstualni outlieri su podaci koji su neobični u specifičnom kontekstu, ali možda nisu outlieri u nekom drugom kontekstu.
    - Njih je često teže identifikovati, ali ih je moguće uočiti putem dodatnih informacija ili domenskog znanja. Na ovaj način može se utvrditi njihov značaj.
    - U bolnici broj dnevnih prijema pacijenata obično prati određeni obrazac. Međutim, tokom izbijanja epidemije, prirodne katastrofe, velikih nesreća u bolnici može doći do naglog porasta broja prijema pacijenata, što te dane čini kontekstualnim outlier-ima u smislu iskorišćenosti bolnice.

S toga je dobra praksa da se uoče outlier-i i da se obrade na odgovarajući način.

- **Uklanjanje** - Ova metoda podrazumeva identifikaciju i uklanjanje outlier-a iz skupa podataka pre obučavanja modela.
- **Transformacij**a - Ova metoda podrazumeva transformaciju podataka kako bi se smanjio uticaj outlier-a. Uobičajene metode su:

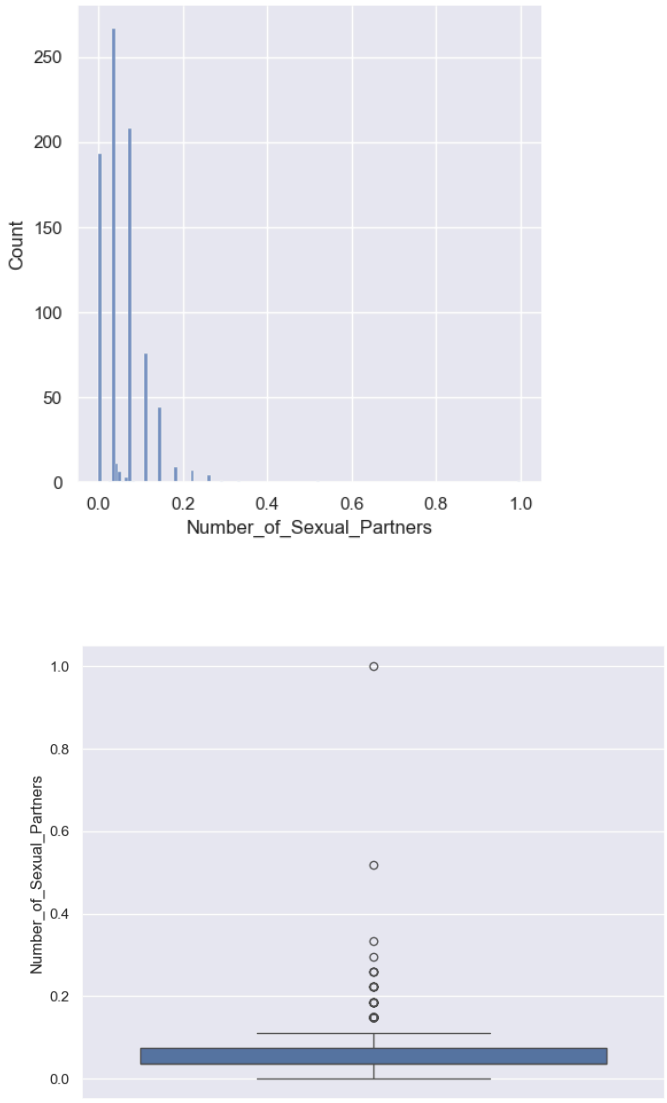
  Skaliranje: Standardizacija ili normalizacija podataka tako da srednja vrednost bude nula i standardna devijacija jedan.

  Winsorizacija: Zamena vrednosti outlier-a najbližom vrednošću koja nije outlier.

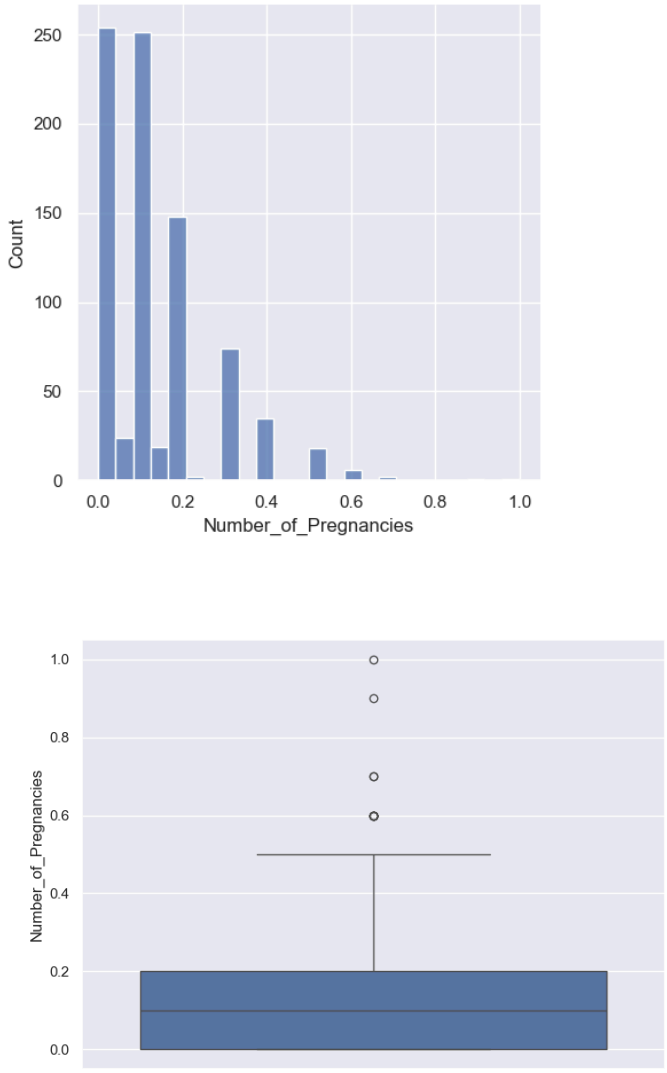
  Logaritamska transformacija: Primena logaritamske transformacije radi smanjenja uticaja ekstremnih vrednosti.

- **Robusna procena:** Ova metoda koristi algoritme koji su manje osetljivi na outlier-e
- Modelovanje outlier-a kao posebne grupe [21]

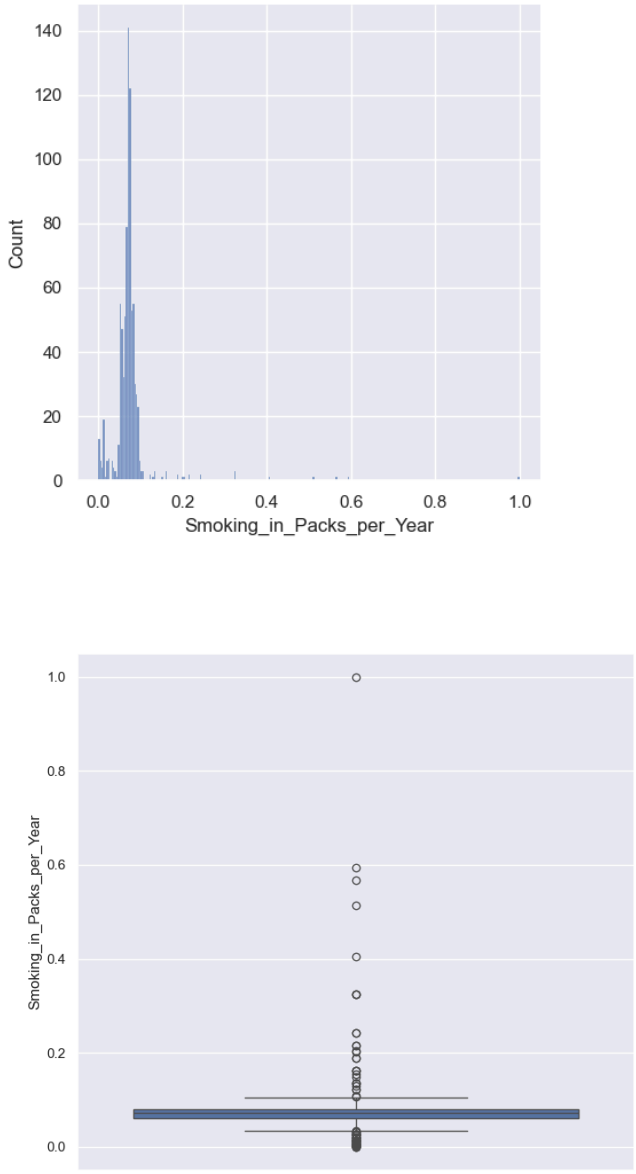
Nakon obrade vrednosti u dataset-u, izvršeno je detektovanje outlier-a i njihovo izbacivanje. Cilj ovoga je bio dobijanje boljih modela i samim tim boljih rezultata metrika. Outlier-i su prvo vizualno uočeni.



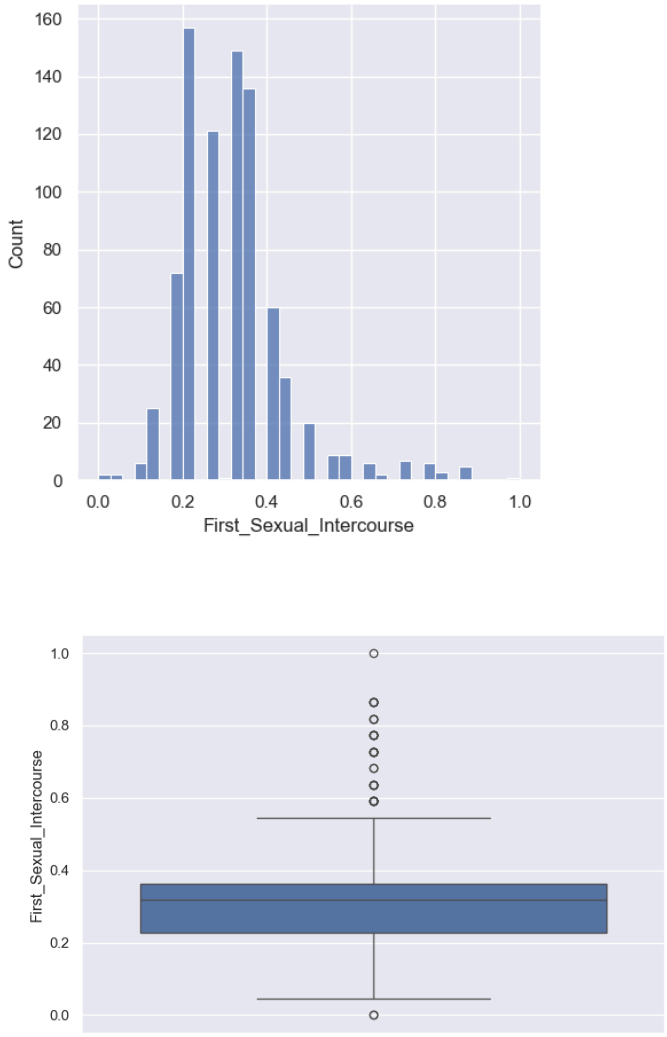
Slika 9 - Prikaz distribucije podataka za kolonu Nuber_of_sexual_Partners



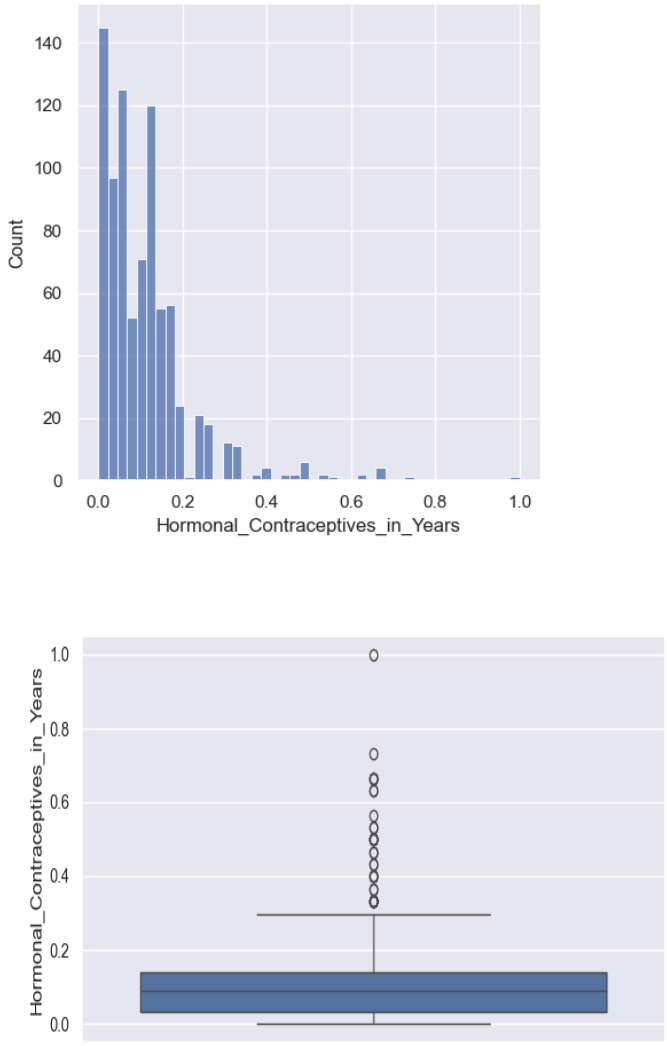
Slika 10 - Prikaz distribucije podataka za kolonu Nuber_of_Pregnancies



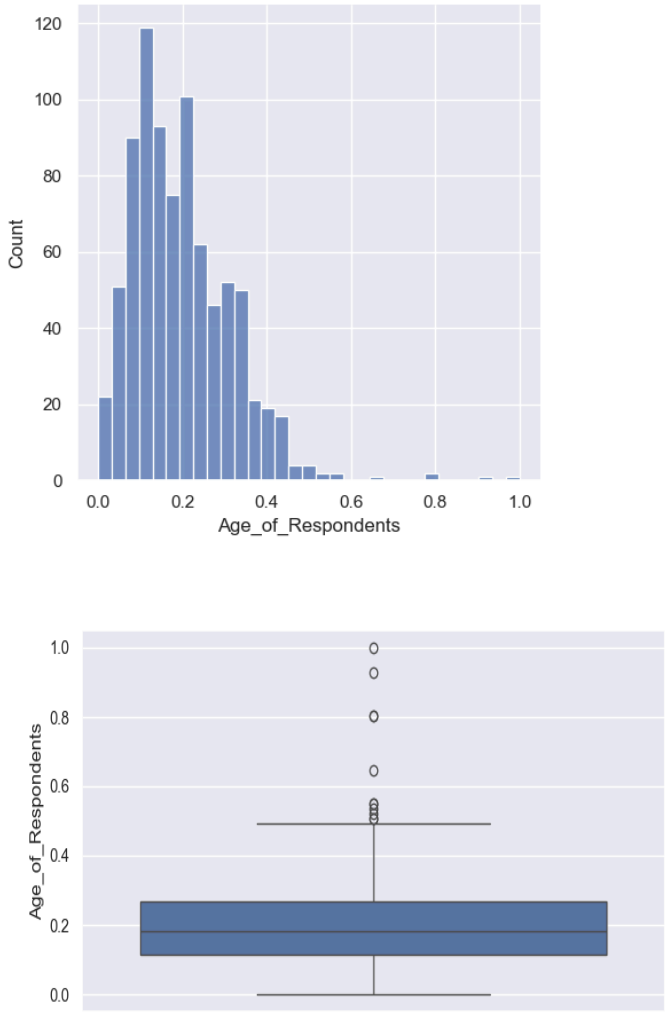
Slika 11 - Prikaz distribucije podataka za kolonu Smoking_in_Packs_per_Year



Slika 12 - Prikaz distribucije podataka za kolonu First_Sexual_Intercourse



Slika 13 - Prikaz distribucije podataka za kolonu Hormonal_Contraceptives_in_Years



Slika 14 - Prikaz distribucije podataka za kolonu Age_of_Respondents

Postoji nekoliko metoda za otkrivanje i uklanjanje outlier-a z-score, IQR i percentil metod. U ovom projektu je odabrana procentil metoda, jer u slučajevima kada podaci nisu normalno raspoređeni, percentili (npr. 95. ili 99. percentili) mogu biti bolji izbor nego metode poput Z-skorova ili IQR-a.

Metoda percentila: Metoda percentila je tehnika koja podrazumeva identifikovanje outlier-a i ograničavanje tih vrednosti na osnovu unapred definisanog procenta. Ovaj metod uključuje izračunavanje granica pomoću percentila i uklanjanje podataka koji premašuju te granice.

Proces korišćenja metode percentila se sastoji od sledećih koraka:

1. **Odredjivanje procentualne granice:** Izabere se procenat koji predstavlja granicu za ekstremne vrednosti. Često korišćene granice su 95. percentil (5% podataka se smatra outlier-ima) ili 99. percentil (1% podataka se smatra outlier-ima). Izbor granice zavisi od specifičnog skupa podataka i željenog nivoa uklanjanja outliera.
2. **Izračunavanja vrednosti granica:** Koristi se izabrani percentil da se izračuna gornja i donja granica. Na primer, ako je izabran 95. percentil, gornja granica bi bila vrednost ispod koje se nalazi 95% podataka, dok bi donja granica bila vrednost iznad koje se nalazi 95% podataka.
3. **Otkrivanje i uklanjanje outlier-a:** Prolazi se kroz skup podataka i identifikuju se sve vrednosti koje premašuju gornju ili donju granicu. Outliere se uklanjaju iz dataset-a.

Primenom metode percentila, ekstremne vrednosti koje se nalaze izvan izabranih granica se efikasno izbacuju. Ovo rezultuje u smanjivanju uticaja outlier-a na analizu podataka i proces modelovanja, omogućeno je stvaranje robusnijih i pouzdanijih rezultata. [18] [19]

Nakon primene percentila u dataset-u je ostalo 748 od početnih 835 podataka, tj. 87 podataka su okarakterisani kao outlier-i.

Potom je izvršena podela dataset-a na trening, validacioni i test skup u odnosu 60/20/20. Training skup je bio korišćen za obučavanje modela, dok su validacioni i test skupovi bili korišćeni za procenu performansi treninranog modela.

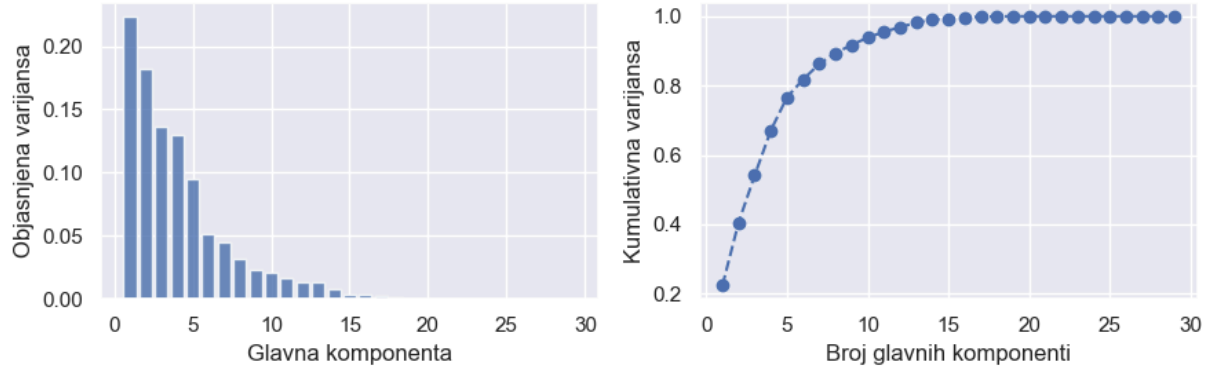
Nakon toga je kreiran OLS model, koji je predstavljao osnovni model. Provereno je da li on zadovoljava L.I.N.E pretpostavke tj. pretpostavke o linearnosti, nezavisnosti grešaka, normalnosti grešaka, savršenoj kolinearnosti i jednakoj varijansi grešaka, pokazalo se da su sve pretpostavke zadovoljene sem one koja je vezana za normalnost grešaka i nepostojanje savršene kolinearnosti u podacima . U ovom slučaju je neispunjavanje pretpostavke o normalnosti greške opravdano, jer se radi sa velikim skupom podataka. Za uklanjanje multikolinearnosti je korišćena PCA (Principal Component Analysis) metoda. PCA eliminiše multikolinearnost između atributa spajanjem visoko koreliranih varijabli u skup nekoreliranih varijabli. Tj. PCA konstruše glavne komponente kao linearne kombinacije ili mešavine početnih varijabli, gde su sve glavne komponente medjusobno nekorelirana.

Zbog velikog broja atributa i multikolinearnosti u prikupljenim podacima, izvršeno je smanjenje dimenzionalnosti problema - uklanjanje neinformativnih atributa i zadržavanje relevantnih atributa. Za ovaj problem je bio korišćen PCA algoritam. PCA prilikom smanjenja dimenzionalnosti kreira novi set atributa koji je manji od početnog, a koji čuva većinu informacija koje su sadržane u podacima. Glavne komponente su linearne kombinacije originalnih promenljivih iz skupa podataka i raspoređene su u opadajućem redosledu po važnosti. PCA pronalazi komponente koje sadrže najviše informacija i formira PC komponente tako da najveći značaj imaju informativni atributi, a mali značaj imaju neinformativni atributi.

Gde su informativni atributi oni koji imaju najveću varijabilnost, a to su široko rasuti podaci. [5] Ono što je značajno u ovom algoritmu je određivanja broja komponenti. Postoji nekoliko pristupa u određivanju ovog broja, a to su:

1. ‘metod lakta’ - gde se na osnovu grafika odabere broj komponenti koji će biti zadržan
2. postizanje ciljane varijanse podataka koji će biti obuhvaćeni (uglvanom je ukupna varijansa između 95% i 99% podataka).

U ovom projektu je odabran drugi pristup tj. ukupna varijansa podataka je 95.5% iz razloga što je uočeno poboljšanje metrika kojim se ocenjuje model kada je varijansa 95.5% . Kada bi se primenio ‘metod lakta’ i kada bi broj zadržanih komponenti bio 5-6 metrike bi značajno opale, a ovaj projekat nije imao to za cilj. [6]



Slika 15 - Odnos varijanse i broja komponenti kod PCA metode

U nastavku je izvršeno ispitivanje heteroskedastičnost. Heteroskedastičnost podrazumeva nejednako rasipanje reziduala, tj. odnosi se na promene u širini rasipanja reziduala. Ona predstavlja izazov za rešavanje, jer OLS modeli podrazumevaju homoskedastičnost tj. varijansa je konstantna. Za ispitivanje postojanja heteroskedastičnosti se koriste Brojš-Pagan-Godfri (Breusch-Pagan-Godfrey) i Vajtov (White) test. [9]

Breusch-Pagan test se temelji na sledećim hipotezama:

- **Nulta hipoteza (H_0)**: Homoskedastičnost je prisutna (varijansa reziduala je konstantna).
- **Alternativna hipoteza (H_a)**: Homoskedastičnost nije prisutna (tj. postoji heteroskedastičnost).

Primenom ovog testa dobijena je p-vrednost koja je mnogo manja od praga 0.05 i ona je iznosila 3.6697674237759055×10−17. Na osnovu ovog rezultata je odbačena nulta hipoteza i zaključeno je da postoji heteroskedastičnost u modelu.

White test je takođe pokazao da je prisutna heteroskedastičnost, a to je zaključeno jer je rezultat testa 1.4973732331606613×10−30 mnogo manji od praga 0.05. [10]

Za rešavanje problema heteroskedastičnosti je preporučeno korišćenje modela poput **WLS**. WLS je sličan OLS modelu, ali daje veću važnost (ili "težinu") određenim podacima u odnosu na druge. WLS dodeljuje težine na osnovu varijanse greške, omogućavajući preciznije modelovanje podataka sa heteroskedastičnošću (neujednačenom varijansom). Podaci sa manjom varijabilnošću ili većom pouzdanošću dobijaju veće težine. Prilikom prilagođavanja regresione linije, WLS pridaje veću važnost tačkama podataka sa većim težinama, što znači da one imaju jači uticaj na konačni rezultat. Ovo omogućava pronalaženje tačnijeg regresionog modela, uproks postojanja heteroskedastičnosti. [11]

Naredni model koji je realizovan u projektu je **RANSAC** model. RANSAC je robusni algoritam koji se koristi u mašinskom učenju za procenu parametara modela u prisustvu outlier-a. Posebno je koristan kada postoji velika količina šumova, a cilj je pronaći model koji dobro odgovara podacima koji su unutar modela (inliers). Pošto je RANSAC iterativni algoritma funkcioniše tako što se nasumično uzimaju podskupovi podataka i vrši se prilagođavanje modela tom podskupu. Zatim se model koristi za klasifikaciju preostalih podataka na inlier i na outlier. Algoritam nastavlja sa iteracijama, birajući nove nasumične podskupove podataka, sve dok se ne pronađe zadovoljavajući model. [12]

Parametri RANSAC modela su:

1. Estimator

Model koji će biti korišćen za fitovanje podataka. Može biti bilo koji regresioni model iz scikit-learn biblioteke poput LinearRegressor-a, HuberRegressor-a, SVR… U projektu je bio korišćen HauberRegressor. On je odabran jer koristi funkciju gubitka (Huber loss) koja kombinuje prednosti kvadratne greške (za manje greške) i apsolutne greške (za velike greške). Ovo smanjuje uticaj outlier-a na model. [17]

2. Residual-Threshold

Smanjenje ovog praga utiče na smanjenje broja validnih podataka. Ovaj prag omogućava da model ne bude osetljiv na outlier-e te oni neće uticati na konačnu procenu parametara.

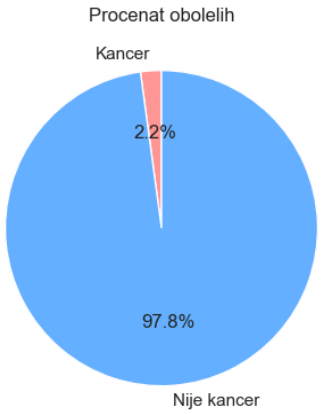
3. Max trials

Ovaj parametar govori o broju iteracija. Kada model ne konvergira, povećanjem ovog parametra se može doći do boljeg modela, posebno kada je prisutan veći broj outlier-a.

4. VIZUALIZACIJA PODATAKA

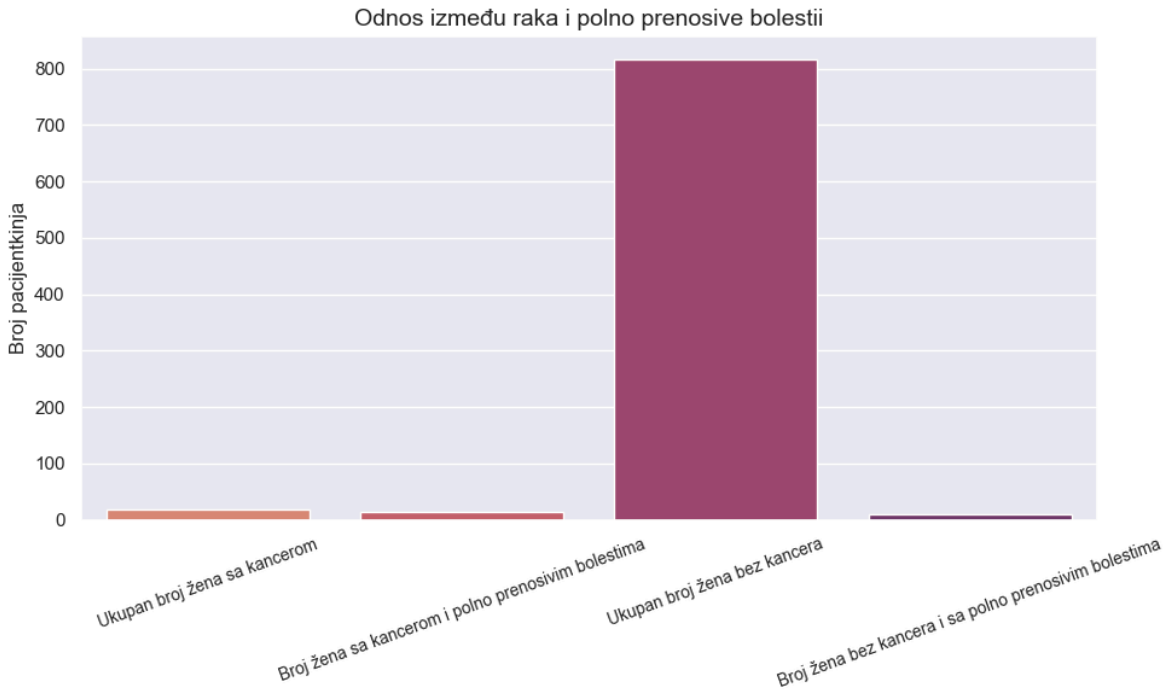
U nastavku će biti vizuelno interpretirani odnosi između zavisne varijable i određenih atributa kako bi se uočile određene karakteristike u podacima koje će posle biti upoređene sa rezultatima koji su dobijeni upotrebom modela.

Može se zaključiti da većina ispitanih pacijentkinja nije imala rak.



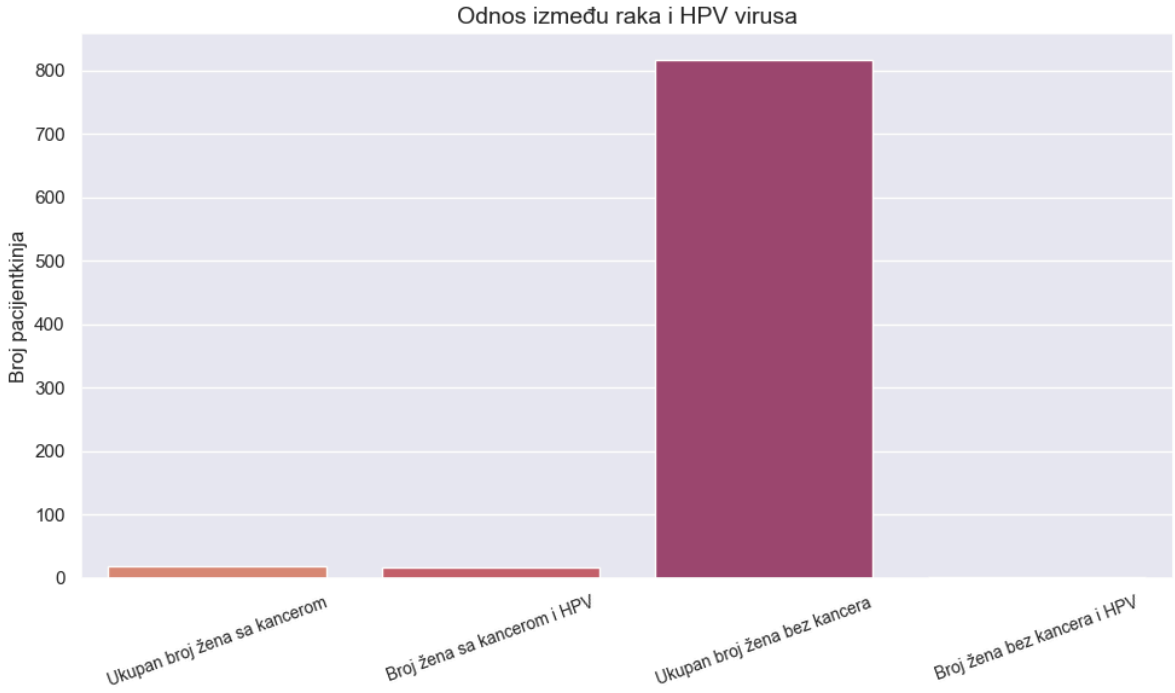
Slika 16 - Procenat obolelih pacijentkinja

Skoro svim pacijentkinjama koje su dobile kancer, bila je dijagnostikovana bar jedna polno prenosiva bolest.



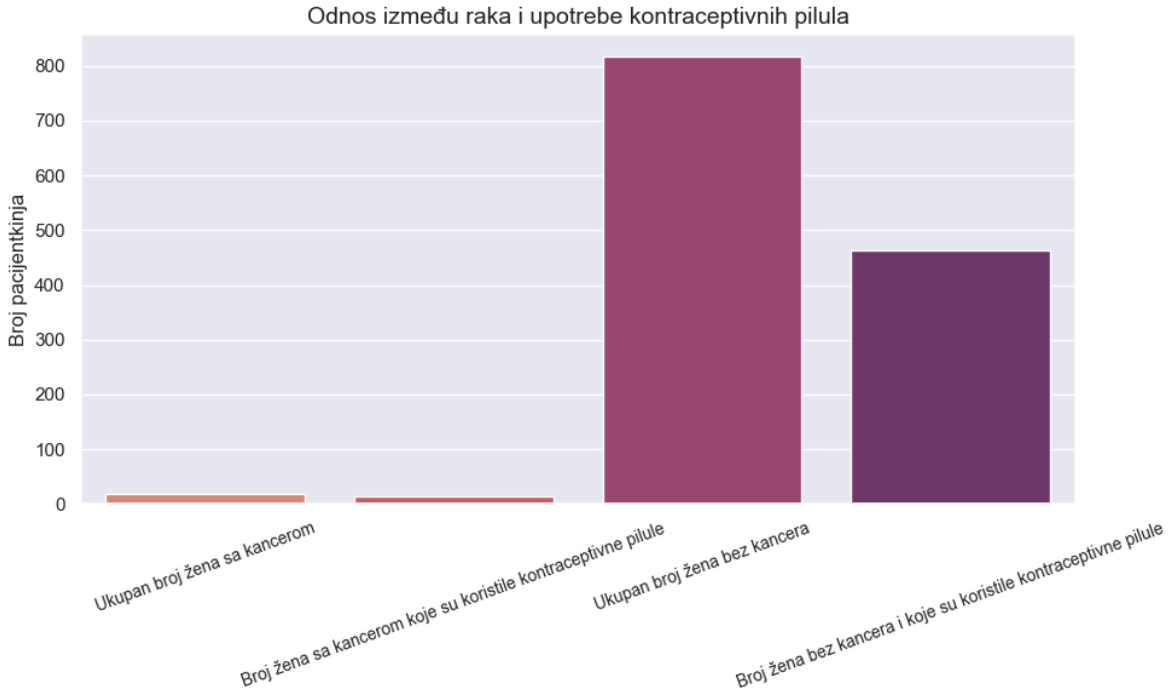
Slika 17 - odnos između raka i polno prenosivih bolesti

Po ovim rezultatima se može zaključiti da su žene koje su bile zaražene HPV virusom gotovo uvek oboljevale od raka grlića materice.



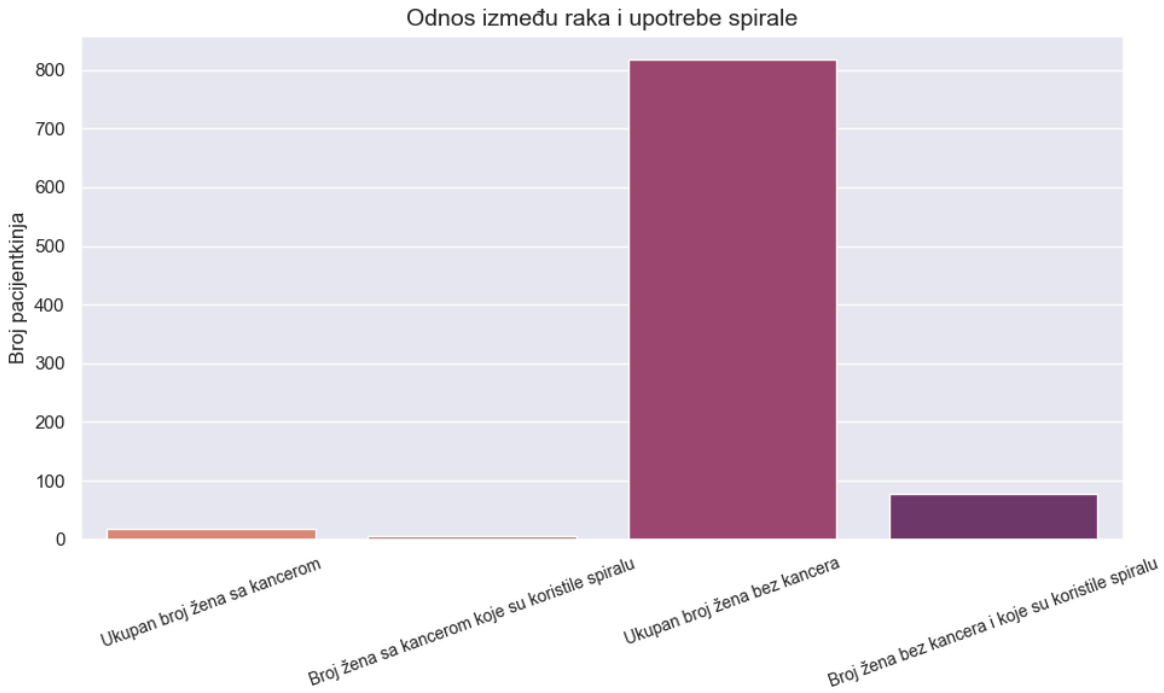
Slika 18 - odnos između raka i HPV virusa

Po ovim rezultatima se može zaključiti da su pacijentkinje kojima je dijagnostikovan karcinom većinski koristile kontraceptivne pilule. Ali, se takođe može primetiti da veliki broj žena kojima nije dijagnostikovan rak takođe koriste kontraceptivne pilule. Moguće je da postoje specifične vrste kontraceptivnih pilula koje podstiču razvoj raka grlića materice, neke vrste koje su jače ili imaju neki specifičan sastojak, kao i da vreme upotrebe odredjenih kontraceptivnih pilula utiče na razvoj raka grlića.



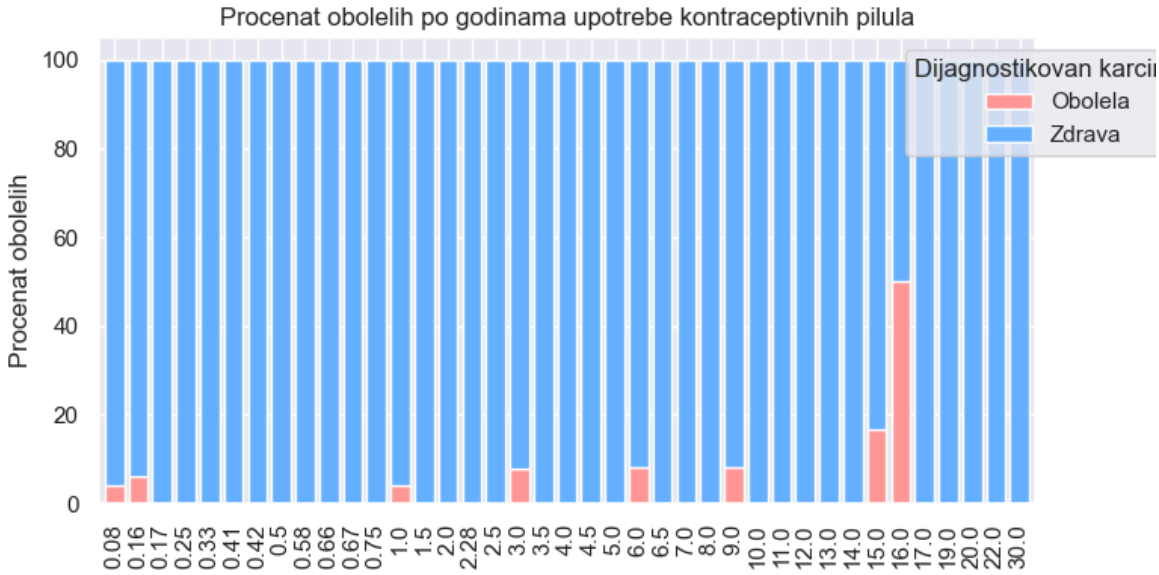
Slika 19 - odnos između pojave raka i upotrebe kontraceptivnih pilula

Po ovim rezultatima se može zaključiti da spirala nema značajan uticaj na pojavu raka grlića materice.



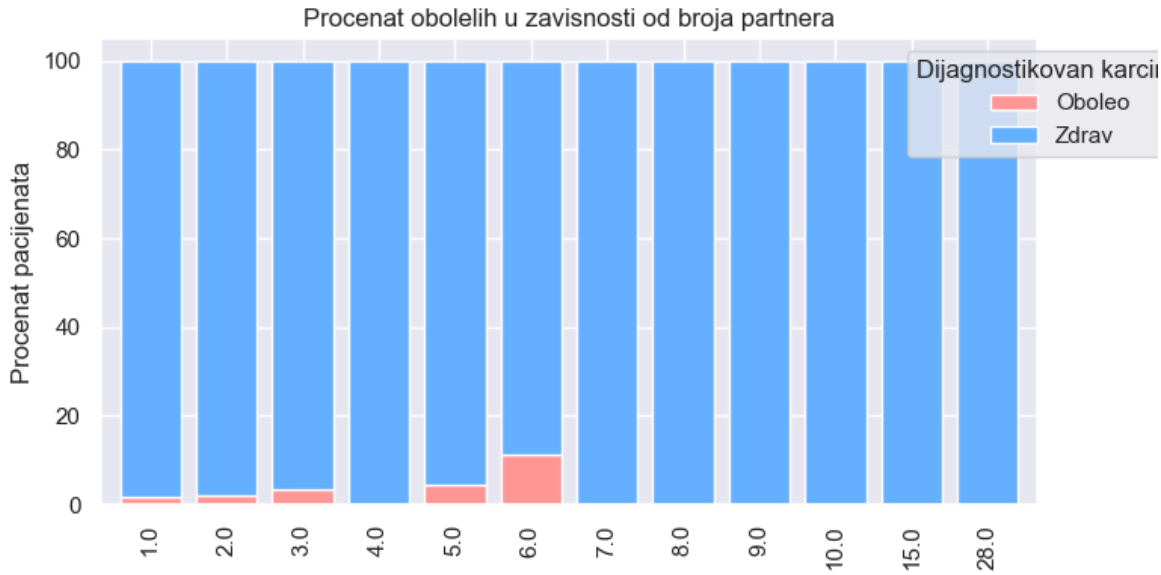
Slika 20 - odnos između pojave raka i upotrebe spirale

Na osnovu rezultata se može zaljučiti da žene koje duži vremenski period konzumiraju kontraceptivne pilule imaju veći rizik za dobijanje raka grlića materice .



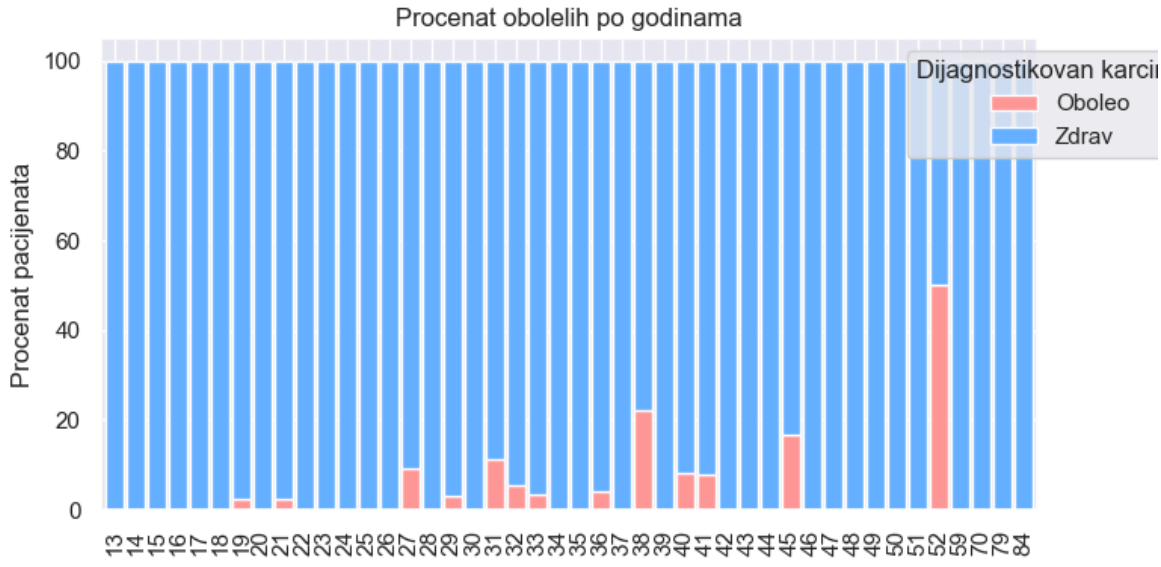
Slika 21 - procenat obolelih pacijentkinja u zavisnosti od godina korišćenja kontraceptivnih pilula

Na osnovu rezultata se može zaljučiti da se povećanjem broja partnera blago povećava rizik za dobijanje raka grlića materice.



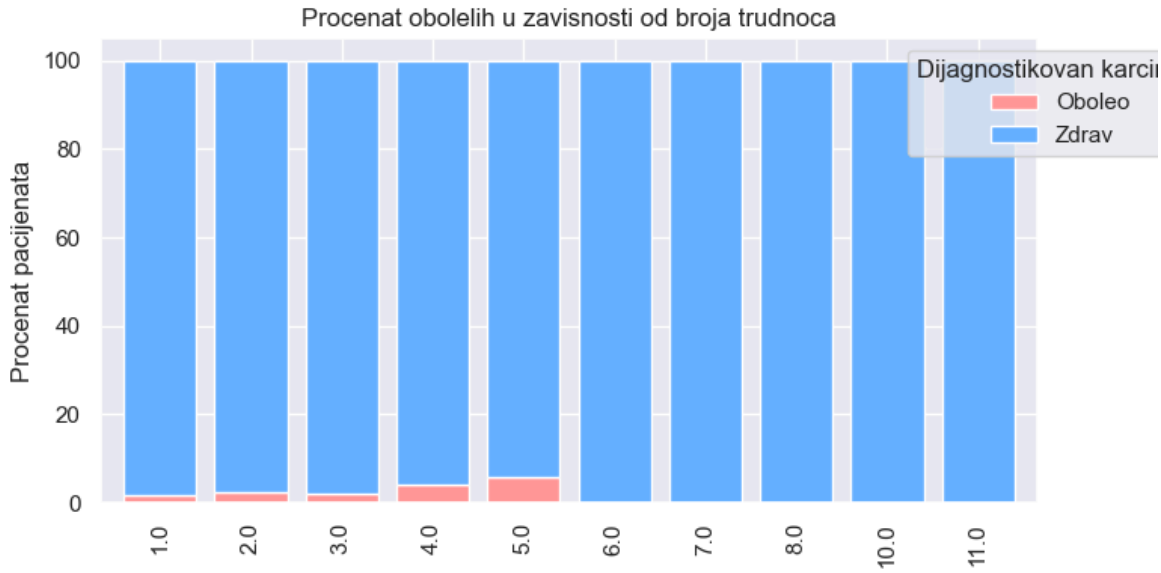
Slika 22 - procenat obolelih pacijentkinja u zavisnosti od broja partnera

Na osnovu rezultata se može zaljučiti da se kod žena u trećoj, četvrtoj i početkom pete decenije života najčešće javlja rak grlića materice.



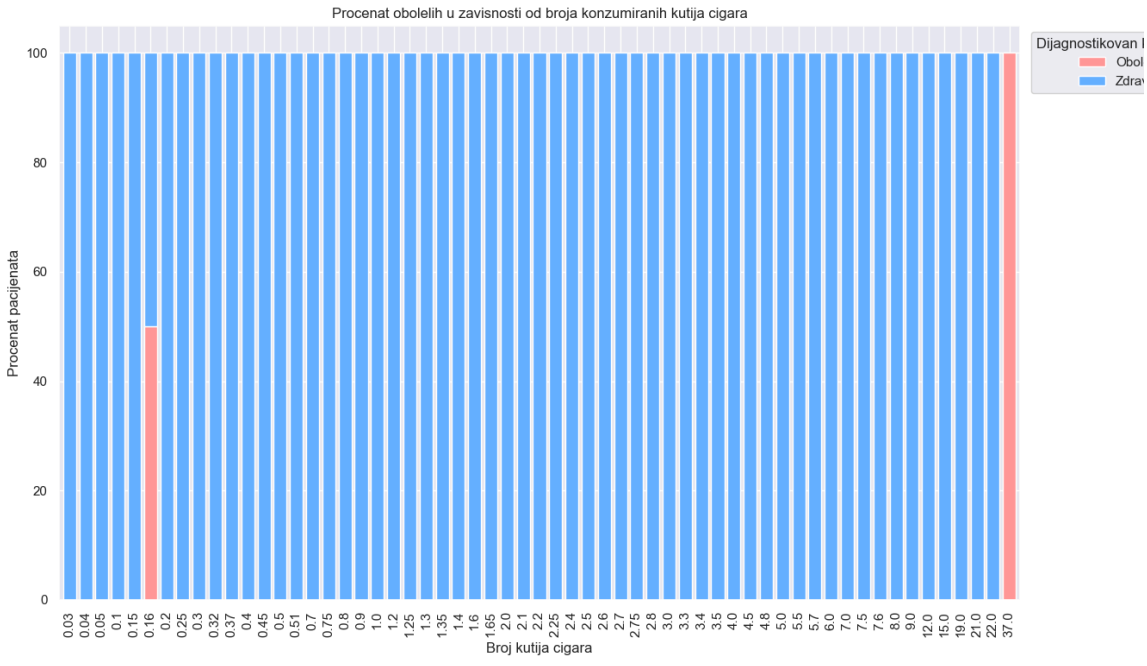
Slika 23 - procenat obolelih pacijentkinja u zavisnosti od godina

Na osnovu rezultata se može zaljučiti da se povećanjem broja trudnoća blago povećava rizik za dobijanje raka grlića materice.



Slika 24 - procenat pacijentkinja u zavisnosti od broja trudnoća

Na osnovu rezultata se može zaljučiti da su pacijentkinje koje su konzumirale najviše kutija cigareta godišnje rizična grupa za dobijanje raka grlića materice.



Slika 25 - procenat pacijentkinja u zavisnosti od kutija cigara koje konzumiraju po godini

5. METRIKE

Metrike korišćene u projektu su:

- **RMSE (Root Mean Squared Error)** - daje informaciju o tome koliko model greši u proseku (u stvarnim jedinicama) tj. kolika je razlika između predviđenih i posmatranih vrednosti. Manja vrednost uglavnom ukazuje na to da je model bolji. [13] Računa se po formuli:

$$\text{RMSE} = \frac{1}{n}\sqrt{\Sigma(\hat{y}i - yi)^2}$$

  $\hat{y}i$ - prediktovana vrednost
  yi - stvarna vrednost
  n - broj podataka

- **Adjusted R-Squared** - daje informaciju o tome koliki deo varijanse zavisne promenljive objašnjavaju nezavisne promenljive. On takođe kažnjava uključivanje nebitnih promenljivih u model. Ova metrika se u projektu koristi, jer je pogodna za upoređivanje modela sa različitim brojem varijabli. Adjusted R-Squared je bolji za ovakva upoređenja of R-Squared, jer ne favorizuje modele sa većim brojem prediktora, kao što to radi standardni R-Squared. Kako bi se izračunao adjusted R-Squared prvo mora da se izračuna R-Squared. [14]

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$Adjusted\ R^2 = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - p - 1}\right)$$

Slika 26 - Formule za R-Squared and adjusted R-Squared

RSS - suma kvadrata reziduala
TSS - ukupna(totalna) suma kvadrata
n - broj podataka
p - broj nezavisnih promenljivih

6. REZULTATI

Za OLS model rezultati metrika su:

- Validacioni skup
  - ➔ RMSE je 0.03609
  - ➔ Adjusted R-Squared je 0.91750
- Test skup
  - ➔ RMSE je 0.05788
  - ➔ Adjusted R-Squared je 0.83973

Napomena

Ovaj OLS model nije pouzdan. S toga je uradjena primena PCA metode i kreiranje drugih modela.

Za PCA model metrike su:

- Validacioni skup
  - ➔ RMSE je 0.04881 (ukazuje da je greška prilikom predikcije raka oko 0.048, gde je 1 vrednost zavisne promenljive kada je rak prisutan, a 0 kada nije prisutan)
  - ➔ Adjusted R-Squared je 0.86875 (ukazuje da model objašnjava oko 87% varijabilnosti zavisne promenljive u validacionim podacima)
- Test skup
  - ➔ RMSE je 0.04450
  - ➔ Adjusted R-Squared je 0.91761

Za RANSAC model metrike su:

- Validacioni skup
  - ➔ RMSE je 0.04212
  - ➔ Adjusted R-Squared je 0.90228
- Test skup
  - ➔ RMSE je 0.03861
  - ➔ Adjusted R-Squared je 0.938

Za WLS model metrike su:

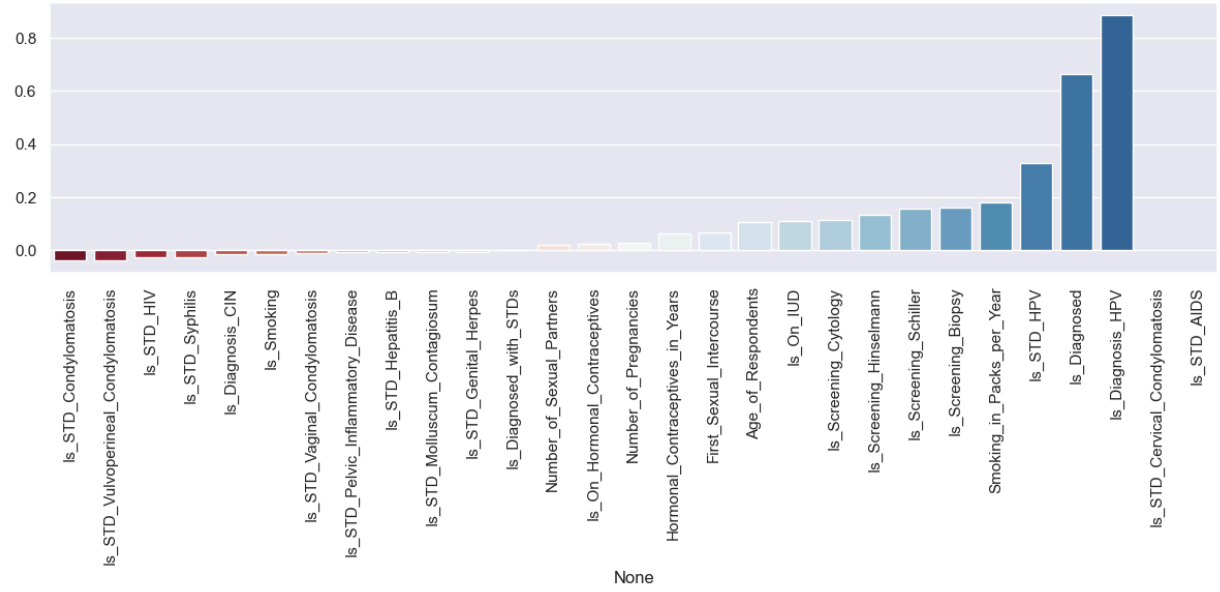
- Validacioni skup
  - ➔ RMSE je 0.04881
  - ➔ Adjusted R-Squared je 0.86875
- Test skup
  - ➔ RMSE je 0.0445
  - ➔ Adjusted R-Squared je 0.91761

Najbolji rezultati metrika su dobijeni korišćenjem RANSAC model, ovo je model koji nije osetljiv na prisustvo preostalih outlier-a - nakon uklanjanja.

PCA i WLS modela su dali takodje veoma dobre rezultate metrika, razlozi za to su sposobnot PCA algoritma da smanji uticaj šumova i mogućnost smanjivanja negativnih efekata multikolinearnosti među atributima. S druge strane, WLS model je smanjio uticaj prisutne heteroskedastičnost među podacima.

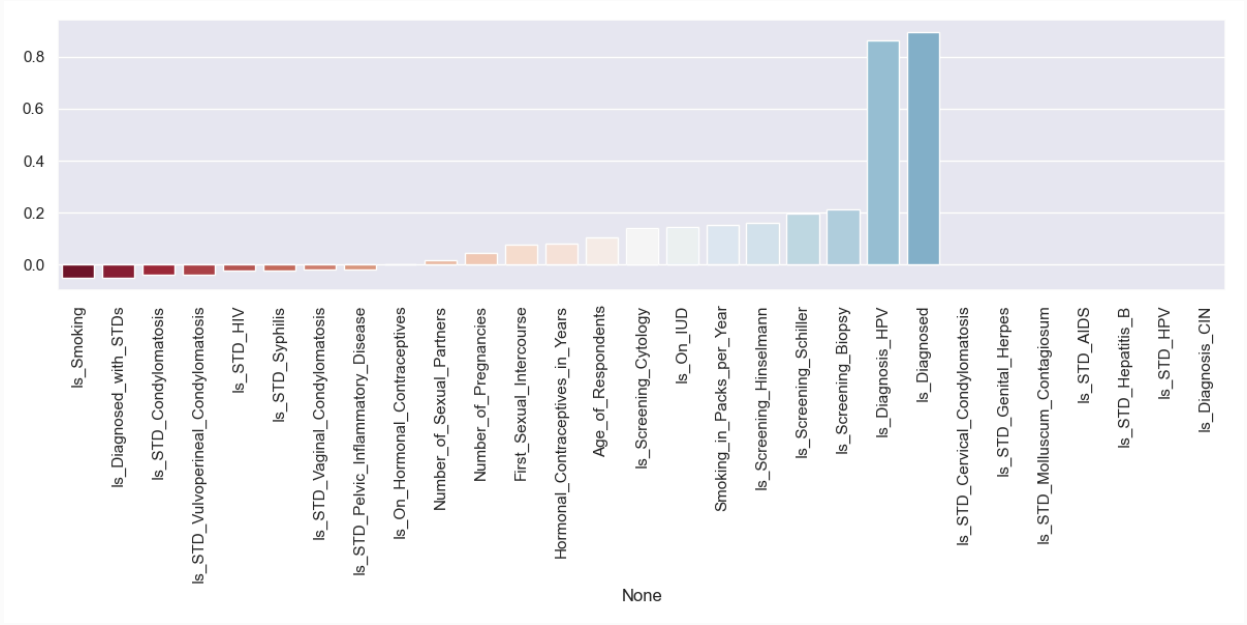
U nastavku će biti komentarisani uticaji atributa na ciljanu kolonu *Is_Diagnosis_Cancer*.

Na grafiku je prikazano koliko svaki atribut pojedinačno utiče na kolonu *Is_Diagnosis_Cancer* tj. posmatra se korelacija sa kolonom koja govori da li je oboljenje kancer ili ne. Ovde se posmatra dataset koji sadrži outlier-e.



Slika 27 - uticaj atributa na kolonu Is_Diagnosis_Cancer za dataset sa outlierima

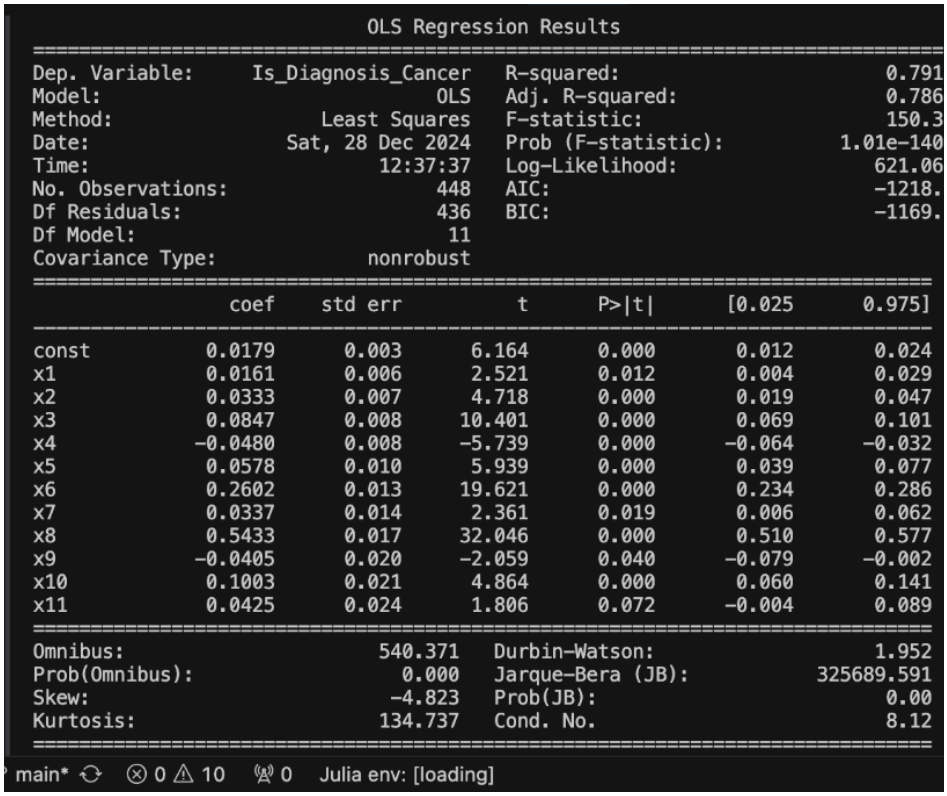
Na grafiku je prikazano koliko svaki atribut pojedinačno utiče na kolonu *Is_Diagnosis_Cancer* tj. posmatra se korelacija sa kolonom koja govori da li je oboljenje kancer ili ne. Ovde se posmatra dataset koji ne sadrži outlier-e.



Slika 28 - uticaj atributa na kolonu Is_Diagnosis_Cancer za dataset bez outlierima

Najveća pozitivna korelacija je sa atributoma *Is_Diagnosis_HPV* i *Is_Diagnosed.* Dok su najveće negativne korelacije sa kolonama *Is_STD_Condylomatosis*, *Is_Diagnosed_with_STDs i Is_Smoking* vrlo mala oko 0.04 te njihovi uticaji verovatno neće biti od velikog značaja. Kolone *Is_Diagnosis_HPV* i *Is_Diagnosed* pojedinačno utiču najviše na pojavu raka grlića materice s toga se očekuje da će glavne komponente zavisiti dosta od ovih atributa.

Potom je bilo neophodno odrediti glavne komponente i posmatrati korelacije glavnih komponenti i originalnih atributa. Bilo je neophodno posmatranje i znaka koeficijenata glavnih komponenti da bi se utvrdilo da li je dejstvo inverzno ili ne, ovi podaci su pročitani iz summary-ja.



OLS Regression Results

Dep. Variable:	Is_Diagnosis_Cancer	R-squared:	0.791
Model:	OLS	Adj. R-squared:	0.786
Method:	Least Squares	F-statistic:	150.3
Date:	Sat, 28 Dec 2024	Prob (F-statistic):	1.01e-140
Time:	12:37:37	Log-Likelihood:	621.06
No. Observations:	448	AIC:	-1218.
Df Residuals:	436	BIC:	-1169.
Df Model:	11		
Covariance Type:	nonrobust		

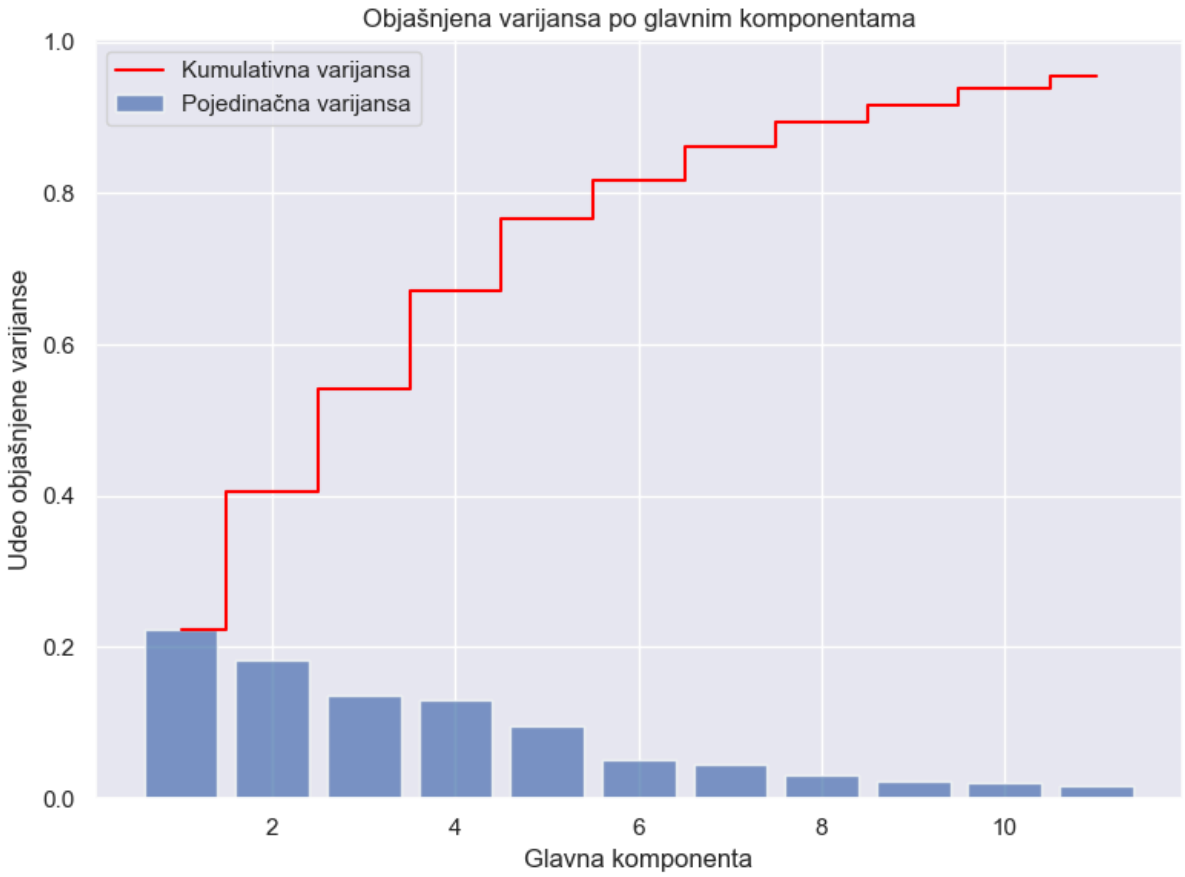
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0179	0.003	6.164	0.000	0.012	0.024
x1	0.0161	0.006	2.521	0.012	0.004	0.029
x2	0.0333	0.007	4.718	0.000	0.019	0.047
x3	0.0847	0.008	10.401	0.000	0.069	0.101
x4	-0.0480	0.008	-5.739	0.000	-0.064	-0.032
x5	0.0578	0.010	5.939	0.000	0.039	0.077
x6	0.2602	0.013	19.621	0.000	0.234	0.286
x7	0.0337	0.014	2.361	0.019	0.006	0.062
x8	0.5433	0.017	32.046	0.000	0.510	0.577
x9	-0.0405	0.020	-2.059	0.040	-0.079	-0.002
x10	0.1003	0.021	4.864	0.000	0.060	0.141
x11	0.0425	0.024	1.806	0.072	-0.004	0.089

Omnibus:	540.371	Durbin-Watson:	1.952
Prob(Omnibus):	0.000	Jarque-Bera (JB):	325689.591
Skew:	-4.823	Prob(JB):	0.00
Kurtosis:	134.737	Cond. No.	8.12

main* ⊗ 0 ⚠ 10 0 Julia env: [loading]

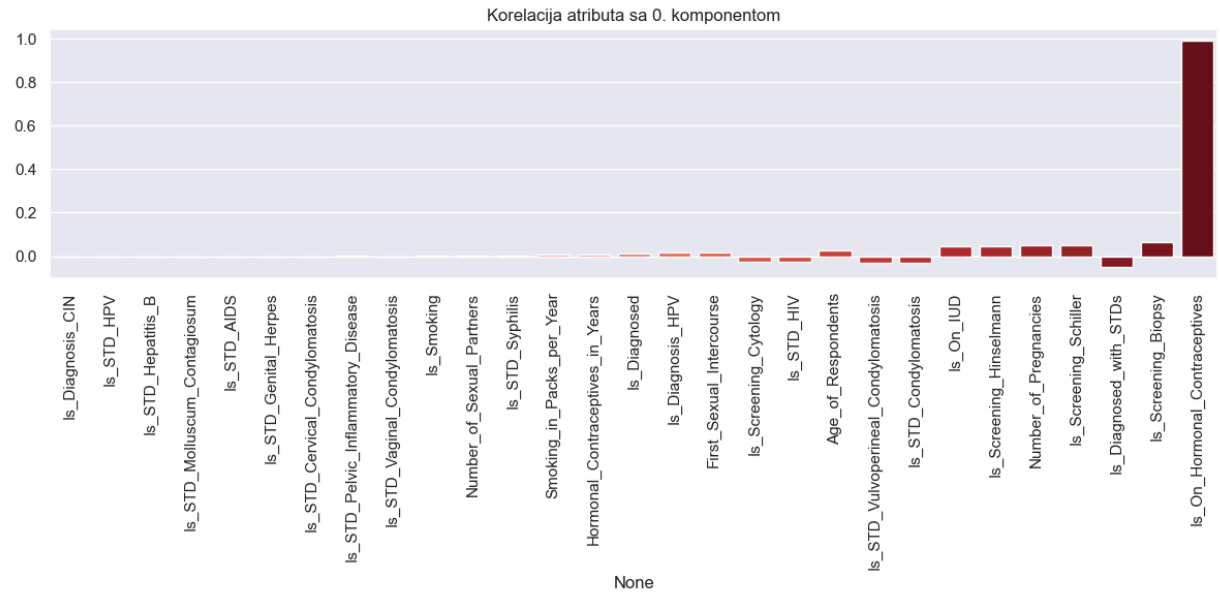
Slika 29 - summary PCA modela

U nastavku je prikazan grafik koji daje informacije o tome koliko varijanse svaka komponenta pojedinačno objašnjava. S toga komponente koje objašnjavaju više varijanse su značajnije.



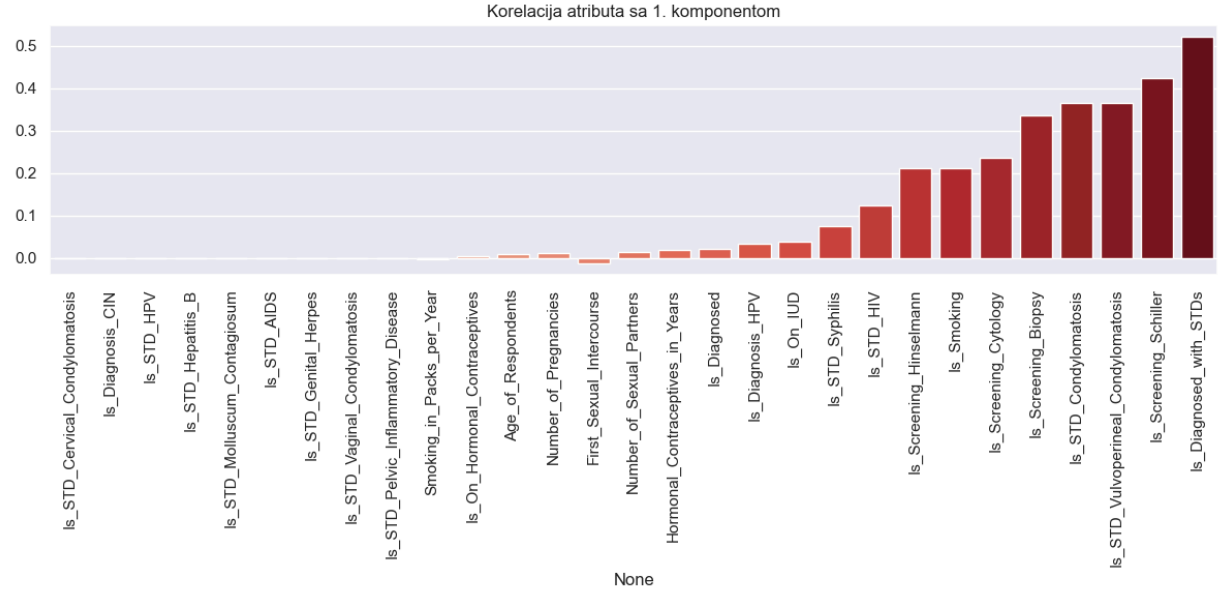
Slika 30 - Količina varijanse koju svaka komponenta pojedinčno objašnjava

Sa grafika se može zaključiti da ova komponenta ima najveću korelaciju sa time da li je pacijentkinja koristila kontraceptivne pilule iz čega sledi da PC1 raste kada se konzumiraju kontraceptivne pilule. Na osnovu pozitivnog koeficijenta uz PC1 se može zaključiti da verovatnoća za dobijanje kancera raste kako raste PC1. Iz čega se može zaključiti da konzumiranje kontraceptivnih pilule podstiče dobijanje raka grlića materice.



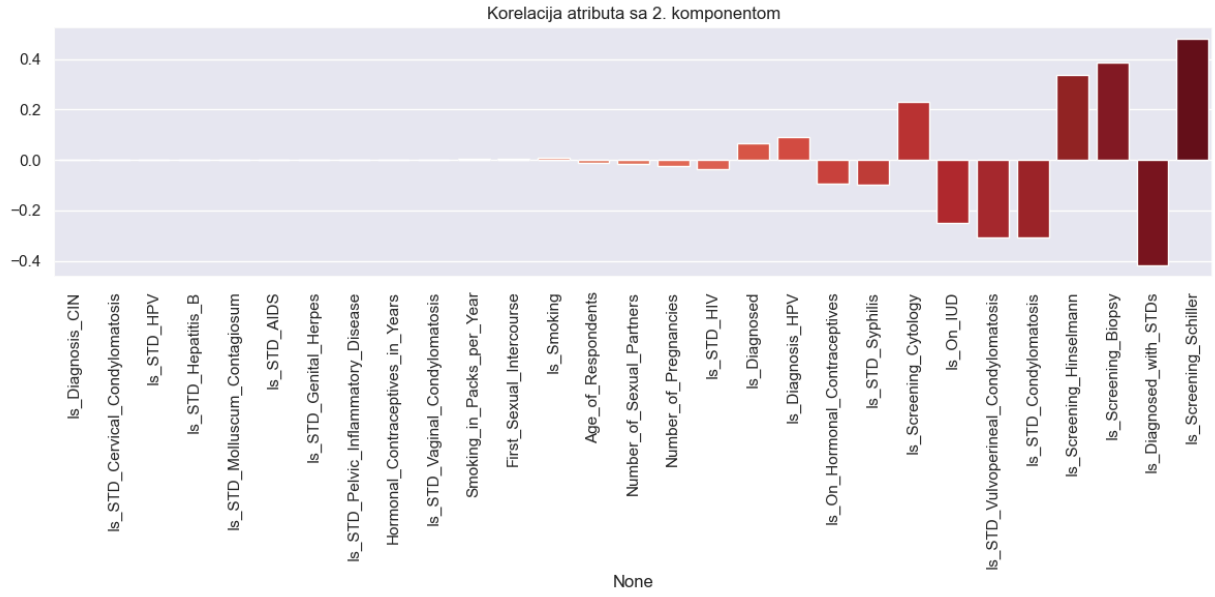
Slika 31 - PC1

Sa grafika se može zaključiti da ova komponenta ima najveću korelaciju sa time da li je pacijentkinji dijagnostikovana neka polno prenosiva bolest (posebno kondilomi), da li je pacijentkinja pušač, da li je upotrebljavana Schillerova metoda za dijagnostikovanje karcinoma… Na osnovu pozitivnog koeficijenta uz PC2 se može zaključiti da verovatnoća za dobijanje kancera raste kako raste PC2. Iz čega sledi da polno prenosive bolesti i pušenje utiču na dobijanje karcinoma grlića materice.



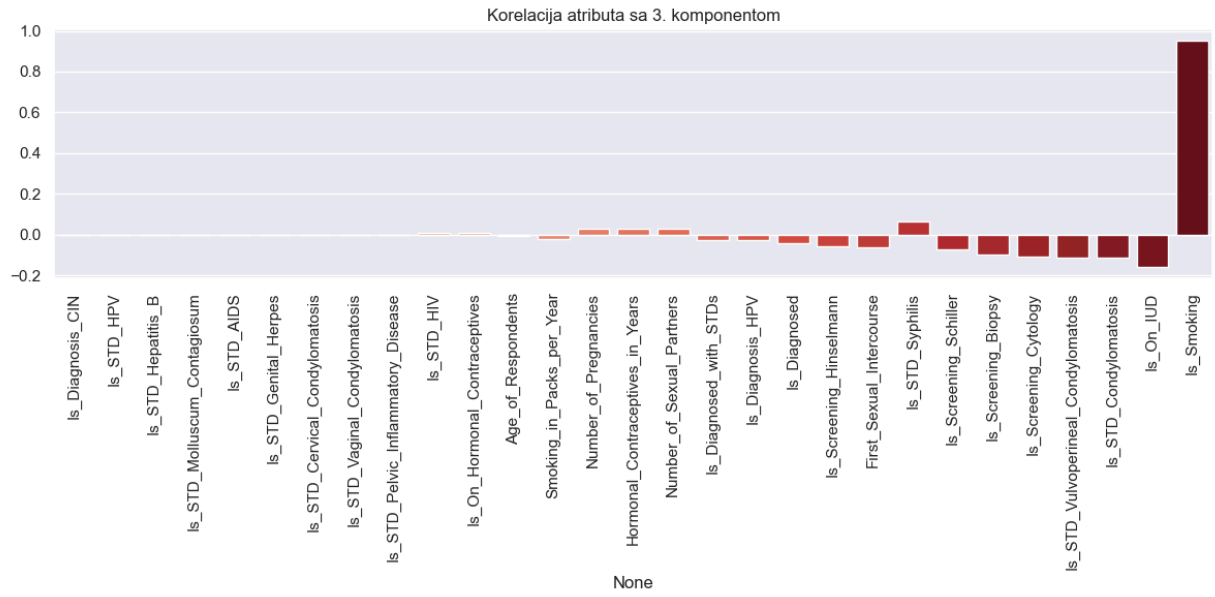
Slika 32 - PC2

Sa grafika se može zaključiti da ova komponenta ima najveću korelaciju sa time da li su korišćeni medicinski mehanizmi poput Schiller-ove metode, biopsije, Hinselman-ove metode za predikciju raka grlića materice, da li je koristila spiralu i da li je pacijentkinja prethodno bolovala od polno prenosivih bolesti (posebno značajno je bolovanje od kondiloma)…Na osnovu pozitivnog koeficijenta uz PC3 se može zaključiti da se povećava verovatnoća da pacijentkinja ima rak kada PC3 raste. Može zaključiti da kada se koristi metodologija poput biopsije, Schiller-a i Hinselmann-a povećava se verovatnoća da kancer bude dijagnostikovan. Dok upotreba spirale i prethodno bolovanje od kondiloma smanjuju verovatnoću za dijagnostikovanje raka.



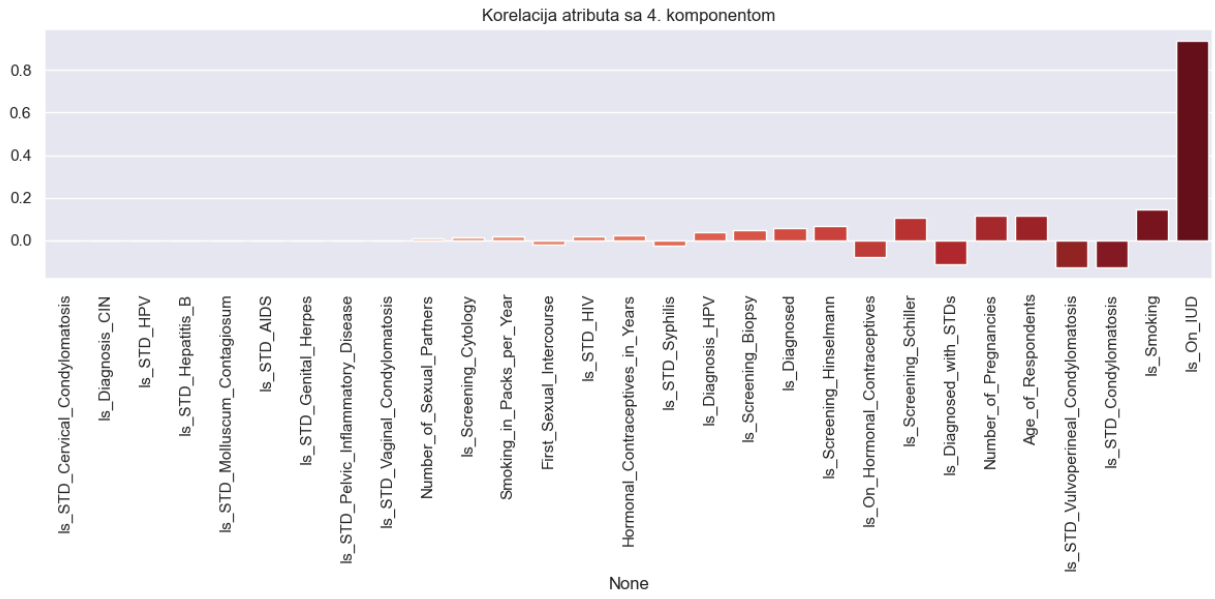
Slika 33 - PC3

Sa grafika se može zaključiti da ova komponenta ima najveću korelaciju sa time da li je pacijentkinja pušač i da li je pacijentkinja koristila spiralu…Na osnovu negativnog koeficijenta uz PC4 se može zaključiti da se smanjuje verovatnoća da pacijentkinja ima rak kada PC4 raste. Tj. da pušenje smanjuje verovatnoću za razvoj raka, dok upotreba spirale povećava.



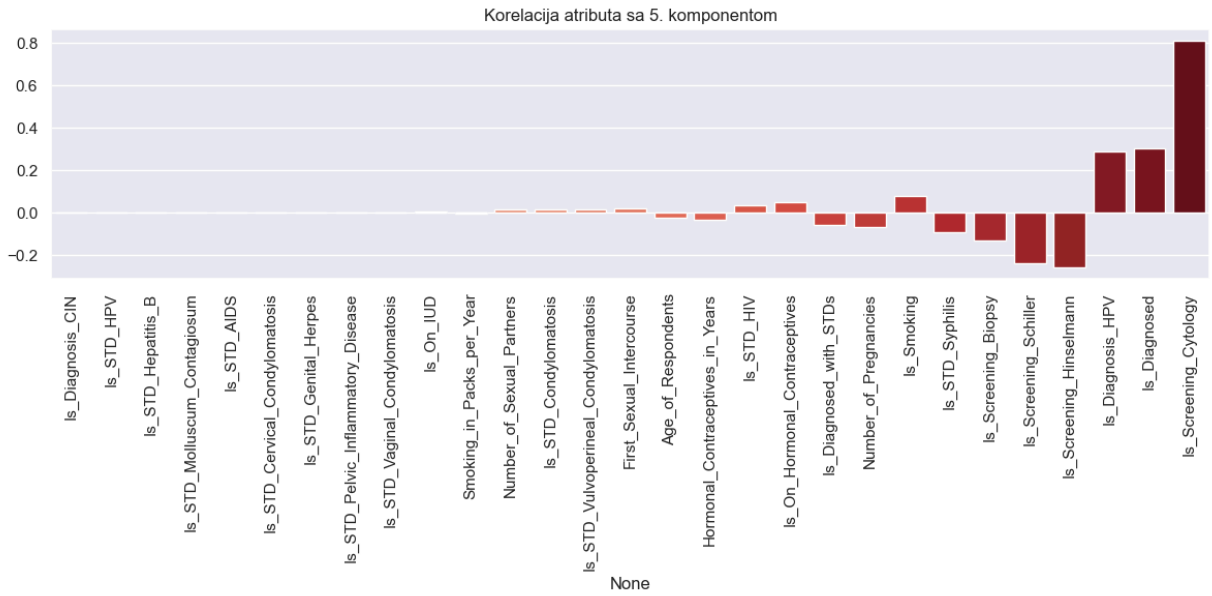
Slika 34 - PC4

Sa grafika se može zaključiti da komponenta PC5 ima najveću korelaciju sa time da li je pacijentkinja koristila spiralu, da li je pacijentkinja pušač i da li je bolovala od kondiloma…Na osnovu pozitivnog koeficijenta uz PC5 se može zaključiti da se povećava verovatnoća da pacijentkinja ima rak kada PC5 raste. Tj. da upotreba spirale i pušenje pospešuju razvoj raka grlića materice, dok bolovanje od kondiloma smanjuje verovatnoću za razvoj grlića materice.



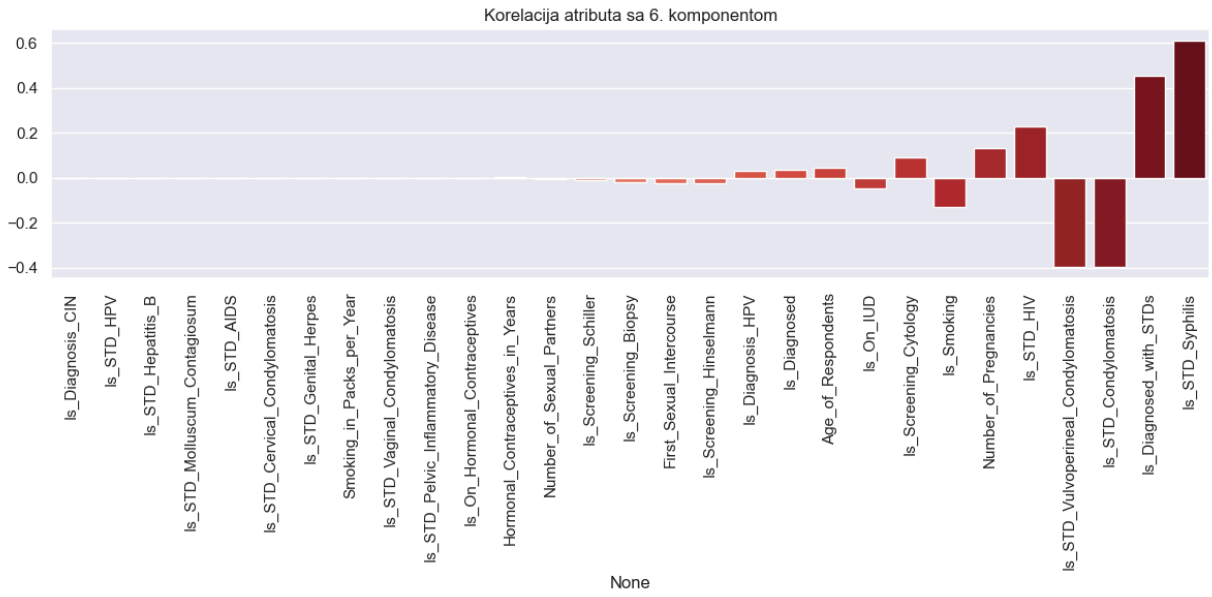
Slika 35 - PC5

Sa grafika se može zaključiti da komponenta PC6 ima najveću korelaciju sa time da li je pacijentkinja bolovala od nekog virusa koji izaziva polno prenosive bolesti (posebno od HPV virusa), da li je metoda za dijagnostikovanje bila citologija…Na osnovu pozitivnog koeficijenta uz PC6 se može zaključiti da se povećava verovatnoća da se kod pacijentkinje javi rak kada PC6 raste. Tj. da činjenica da je pacijentkinja bolovala od polno prenosivih bolesti (posebno od HPV virusa) povećava verovatnoću da dobije rak, dok negativna korelacija sa bolovanjem od sifilisa smanjuje verovatnoću da pacijentkinji bude otkriven rak grlića. Kada je korišćena citologija za dijagnostikovanje postoji povećana šansa da pacijentkinji bude otkriven rak.



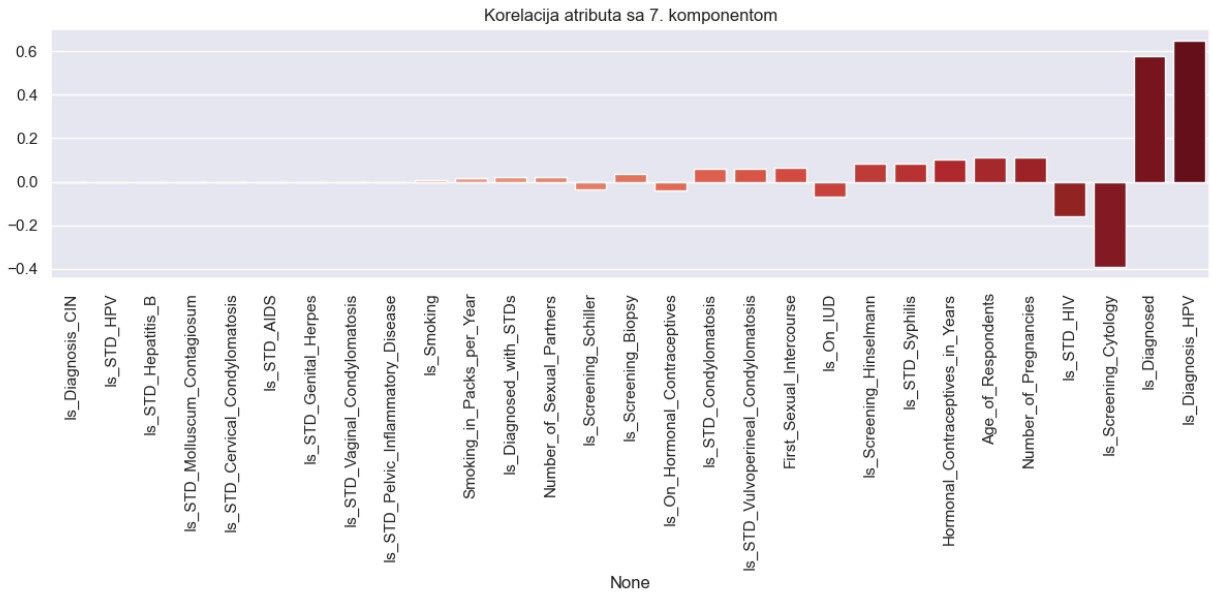
Slika 36 - PC6

Sa grafika se može zaključiti da komponenta PC7 ima najveću korelaciju sa time da li je pacijentkinji dijagnostikovana neka polno prenosiva bolest (posebno su značajane sifilis i kondilomi) …Na osnovu pozitivnog koeficijenta uz PC7 se može zaključiti da se povećava verovatnoća da pacijentkinja ima rak kada PC7 raste. Tj. da činjenica da je pacijentkinja bolovala od sifilisa povećava verovatnoću da dobije rak, dok negativna korelacija sa kondilomima ukazuje na to da se smanjuje verovatnoća da kancer bude dijagnostikovan kada je pacijentkinja bolovala od njih.



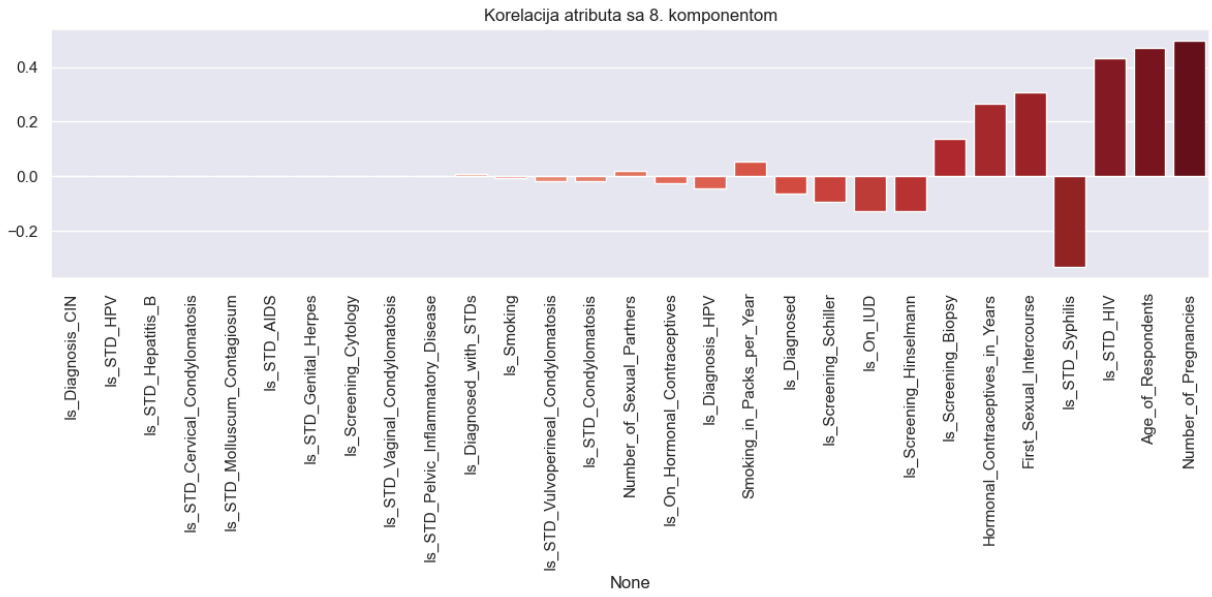
Slika 37 - PC7

Sa grafika se može zaključiti da komponenta PC8 ima najveću korelaciju sa time da li je pacijentkinja uopšte bolovala od virusa koji izaziva polno prenosive bolesti (posebno da li je bila zaražena HPV virusom) i da li je korišćena citologija za predikciju raka grlića materice … Na osnovu pozitivnog koeficijenta uz PC8 se može zaključiti da se povećava verovatnoća da pacijentkinja ima rak kada PC8 raste. Tj. kada je je bolovala HPV virusa ili bilo kog drugog virusa koji izaziva polno prenosive bolesti povećava se verovatnoća da se pojavi rak grlića materice, dok činjenica da se upotrebljava citologija za dijagnostikovanje raka smanjuje verovatnoću da bolest bude rak.



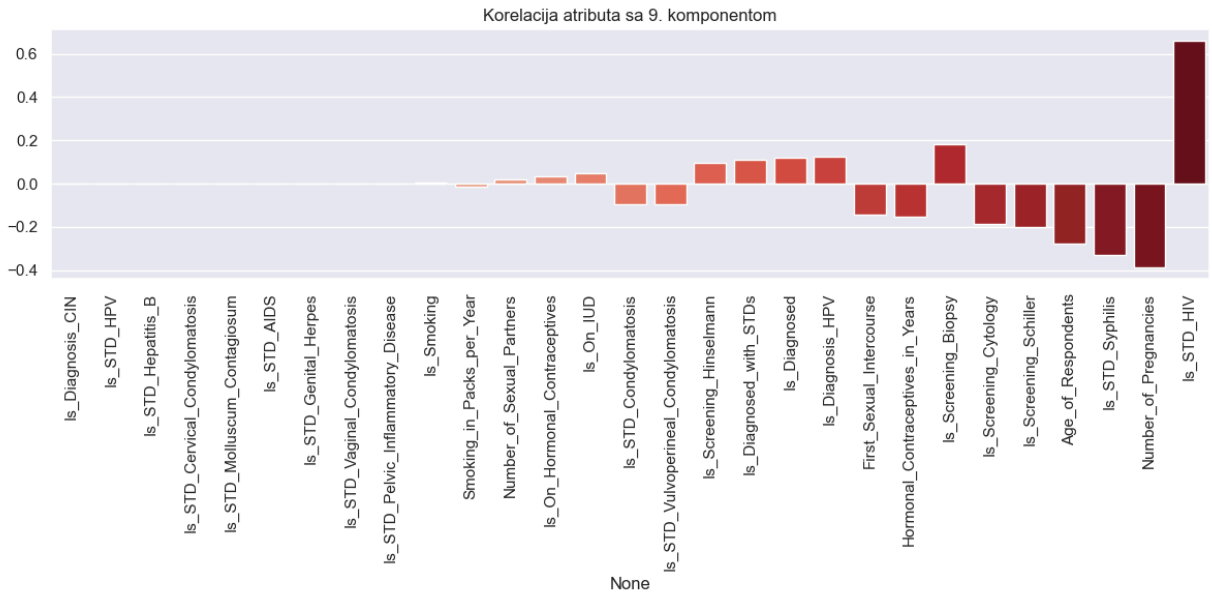
Slika 38 - PC8

Sa grafika se može zaključiti da komponenta PC9 ima najveću korelaciju sa atributima koji daju informaciju o broju trudnoća pacijentkinje, godinama pacijentkinje, kolliko godina je pacijentkinja konzumirala kontraceptivne pilule, da li bolovala od HIV-a ili sifilisa… Na osnovu negativnog koeficijenta uz PC9 se može zaključiti da se smanjuje verovatnoća da pacijentkinja ima rak opada kada PC9 raste. Tj. kako se povćava broj trudnoća pacijentkinje, godine starosti i godine tokom kojih konzumira kontraceptivne pilule, tako se smanjuje verovatnoća da joj se dijagnostikuje rak grlića materice. Pacijentkinje kojima je prethodno dijagnostikovan sifilis imaju manju verovatnoću da obole od raka.



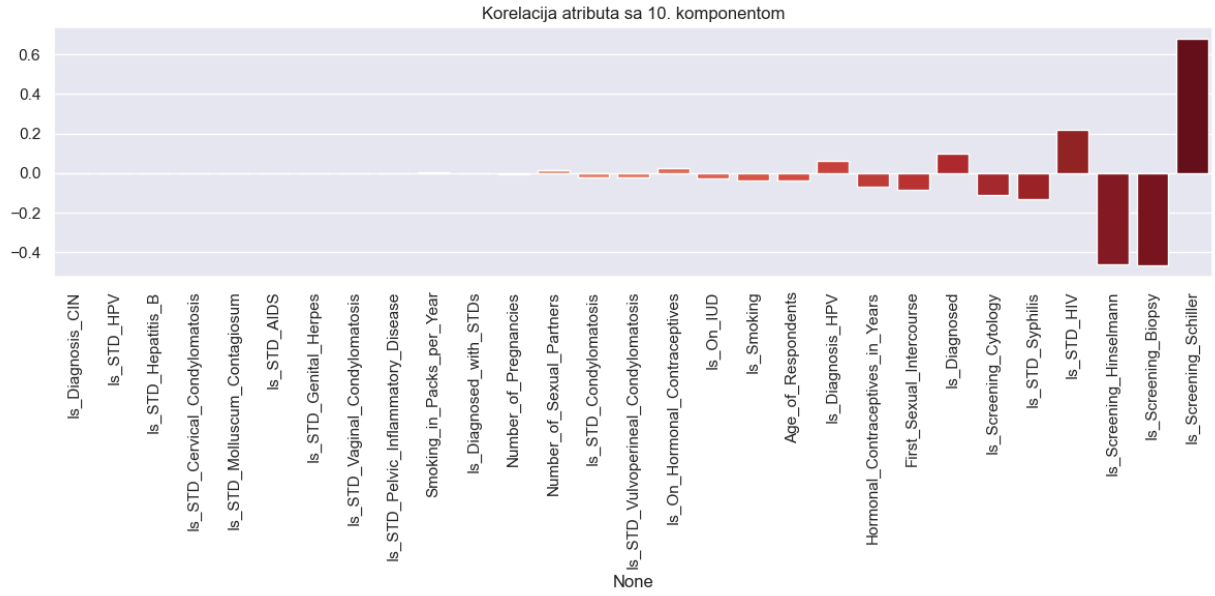
Slika 39 - PC9

Sa grafika se može zaključiti da komponenta PC10 ima najveću korelaciju sa time da li je pacijentkinja bolovala od HIV-a ili od sifilisa, brojem trudnoća pacijentkinje i godinama života pacijentkinje… Na osnovu pozitivnog koeficijenta uz PC10 se može zaključiti da se povećava verovatnoća da pacijentkinja ima rak kada PC10 raste. Tj. pacijentkinja koja je bolovala od HIV-a ima veću verovatnoću za razvijanje raka grlića materice, dok povećanje broja trudnoća, godina pacijentkinje i bolovanje od sifilisa rezultuju smanjenom verovatnoćom za razvoj raka grlića materice.



Slika 40 - PC10

Sa grafika se može zaključiti da komponenta PC11 ima najveću korelaciju sa time da li su metode za lečenje bili Schillerova metoda, biopsija ili Hinselmanova metoda i da li je pacijentkinja bolovala od HIV… Na osnovu pozitivnog koeficijenta uz PC10 se može zaključiti da se povećava verovatnoća da pacijentkinja ima rak kada PC10 raste. Tj. pacijentkinja koja je bolovala od HIV-a ima veću verovatnoću za razvijanje raka grlića materice, a povećava se verovatnoća da bude dijagnostikovan rak grlića kada se koristi Schiller-ova metoda, dok se upotrebom biopsije i Hinselmann-ove metode smanjuje verovatnoća za otkrivanje raka.



Slika 41 - PC11

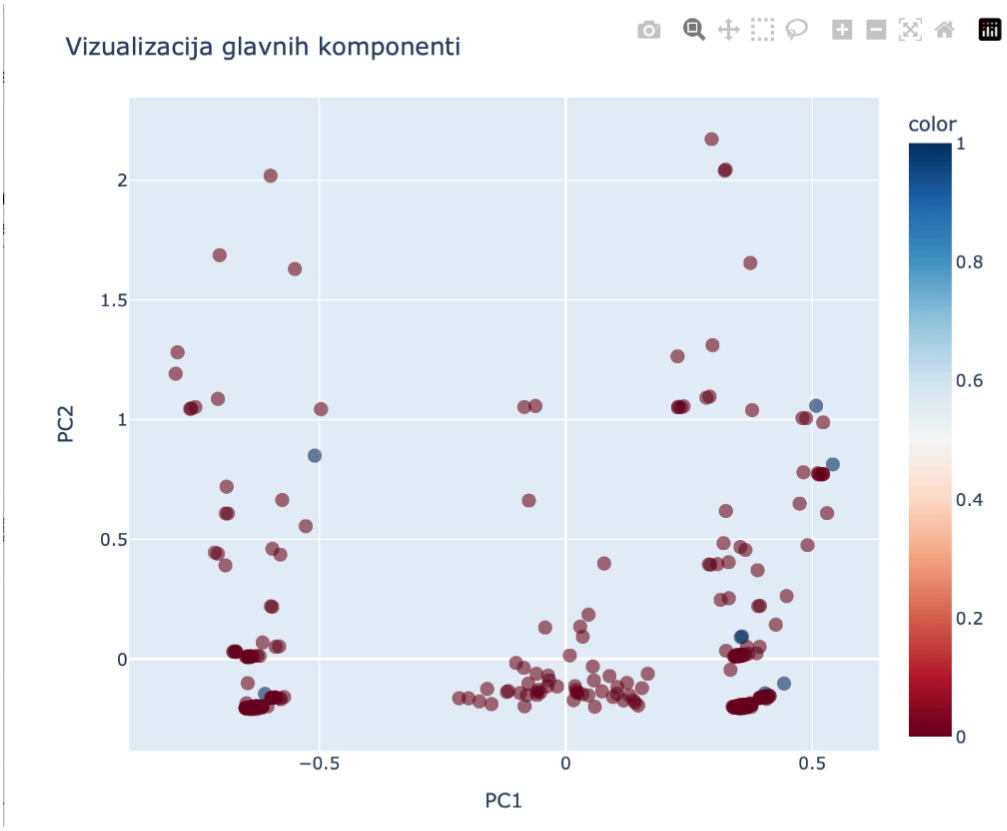
Zaključak na osnovu rezultata

- Pošto je zadržan veliki broj komponenti u PCA algoritmu, jer je cilj bio da bude objašnjen veliki deo varijanse (preko 95%), došlo je do nekih kontradiktornosti u rezultatima npr. na osnovu jedne komponente se zaključi da pušenje podstiče pojavu raka grlića materice, dok se na osnovu druge zaključi da ne podstiče razvoj raka. Ovo se desilo jer veliki broj komponenti sadrži informacije o specifičnim, manje bitnim obrascima. Kako bi se ove nelogičnosti prevazišle nacrtan je grafik koji daje informacije o tome koji procenat varijanse svaka komponenta objašnjava. S toga su komponente sa većim procentom podataka koji objašnjavaju uzete za relevantnije.
- Analizom podataka se može zaključiti da na razvoj karcinoma grlića materice najvišu utiču kontraceptivne pilule, pušenje, upotreba spirale, polno prenosive bolesti

(najznačajnija je bolest izazvana HPV virusom). Takođe je uočeno da je upotreba metoda poput Schillera i biopsije povezana sa dijagnostikovanjem raka grlića materice.

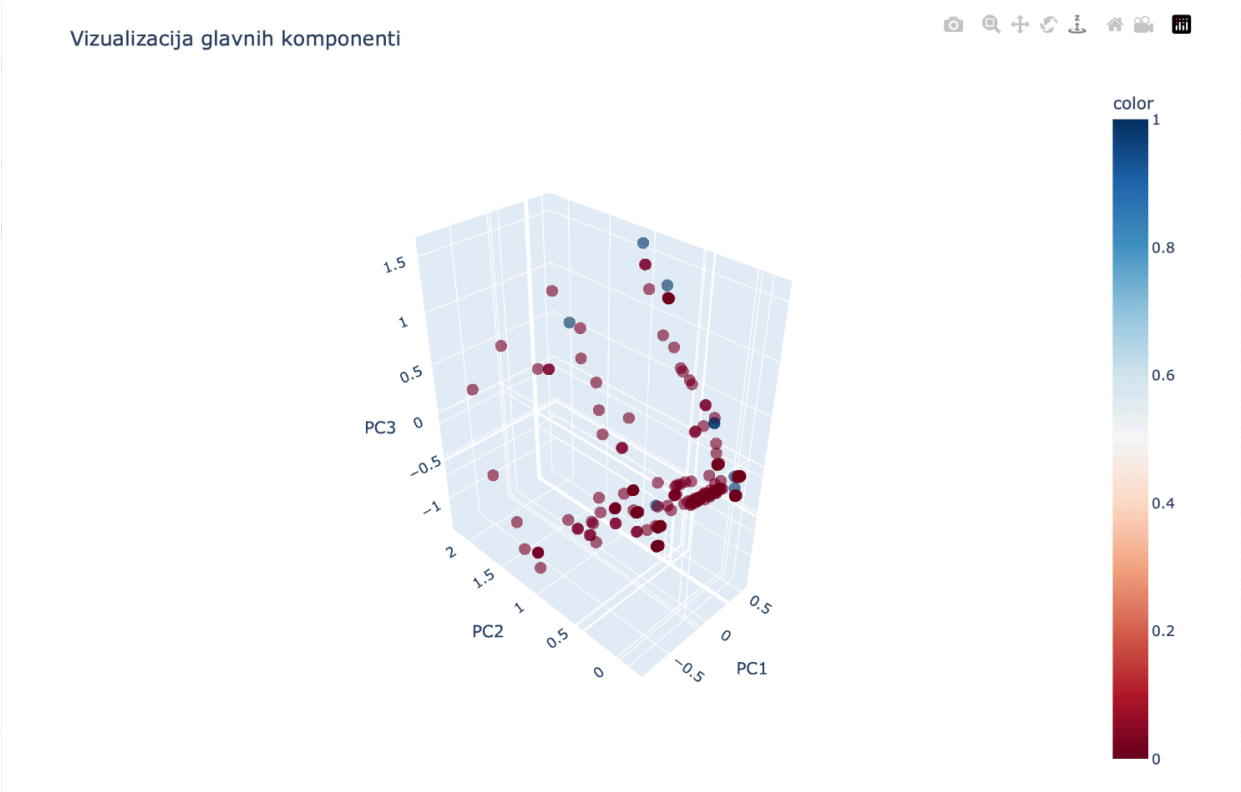
- Ovakvi rezultati su donekle i bili očekivani nakon što su podaci vizualizovani. Upotrebom vizualizacije se moglo očekivati da će na pojavu raka grlića materice dominantno uticati polno prenosive bolesti (pogotovo bolest koju izaziva HPV virus), konzumiranje cigareta, kotraceptivne pilule…

Na graficima u nastavku je prikazana struktura podataka. Crvene tačke označavaju žene kojima nije dijagnostikovan rak, dok plave označavaju pacijentkinje kojima je otkriven karcinom. Na 2D grafiku se vidi da žene kojima je dijagnostikovan rak za prvu i drugu komponentu imaju visoke vrednosti.



Slika 42 - Vizualizacija glavnih komponenti u 2D

Na 3D grafiku se vidi da žene kojima je dijagnostikovan rak za prvu, drugu i treću komponentu imaju visoke vrednosti.



Slika 43 - Vizualizacija glavnih komponenti u 3D

7. KOMPARACIJA

Ovaj rad (https://pmc.ncbi.nlm.nih.gov/articles/PMC9185380/#sec3-sensors-22-04132) je koristio isti dataset kao što je korišćen u ovom projektu i u nastavku će biti izvršeno upoređivanje korišćenih metodologija i dobijenih rezultata.

U radu nije izvršeno splitovanje na training, test i validacioni skup podataka, već je izvršen postupak *cross validation.* Mana ovakvog pristupa validiranja je bila ta što nije dobijena informacija o tome koliko će dobijeni model biti dobar kada se budu pravile nove predikcije za nove podatke.

Dok su u projektu korišćeni OLS, WLS, RANSAC modeli u sličnom radu su korišćeni modeli Gradient Boosting i XGBoost.

Za Gradient Boosting model metrike su:

- ➔ RMSE je 0
- ➔ R-Squared je 1

Za XGBoost model metrike su:

- ➔ RMSE je 0.14805144
- ➔ R-Squared je 0.68628035

Gradient Boosting je popularni algoritam za "boostovanje" u mašinskom učenju koji se koristi za zadatke klasifikacije i regresije. Boostovanje je jedan od metoda učenja zasnovanih na ansamblima. Model se trenira sekvencijalno, pri čemu svaki novi model pokušava da ispravi greške prethodnog modela. Generalno je overfit-ovanje modela jedan od glavnih problema u mašinskom učenju, ali je veoma česta pojava kod Gradient-Boosting-a, s toga mislim da je u ovom radu došlo do overfit-ovanja jer su dobijeni savršeni rezultati. [15][16]

Za **XGBoost** algoritam su dobijeni lošiji rezultati metrika nego što su dobijeni ovde na projektu. U radu je pretprocesiranje podataka izvršeno na drugačiji način nego što je ovde u projektu, nedostajuće vrednosti nisu popunjavane već izbacivane kao i outlier-i, nakon toga je izvršeno i dodatno smanjenje dimenzionalnosti problema primenom PCA. S toga se može zaključiti da je robusnost modela smanjena zbog odbacivanja velikog dela podataka.

U nastavku je izvršena analiza korelacije atributa sa zavisnom promenljivom pomoću matrice korelacije u radu, dok su u projektu PC komponente analizirane pojodinačno kako bi se uočili latentni faktori. S toga kad bi se uporedilo koliko svaki atribut pojedinačno utiče na kolonu *Is_Diagnosis_Cancer* nad dataset-om koji sadrži outlier-e iz projekta sa ciljanom kolonom matrice korelacije iz rada dobili bi se slični rezultati. Npr. koeficijent korelacije između zavisne promenljive i polno prenosive bolesti izazvane HPV-om je oko 0.9, koefcijent koji predstavlja vezu između zavisne promenljive i atributa vezanog za informaciju da li pacijent ima bilo koju bolest je oko 0.7. Ali, kada bi se posmatrale zavisnosti kolone *Is_Diagnosis_Cancer* i ostalih atributa nad dataset-om koji ne sadrži outlier-e dobili bi se drugačiji koeficijenti korelacije, dok bi dominantni atributi ostali isti (*Is_Diagnosed* i *Is_Diagnosis_HPV*). Mogući razlog za različite koeficijente je upotreba drugačijih metoda za detekciju outlier-a (u radu nije navedeno koja konkretno metoda je upotrebljena) i drugačiji kriterijum za klasifikovanje podatka kao outlier.

Dodatnom analizom pojedinačnih PC komponenti u projektu je uočeno da su konkretno polno prenosive bolesti, pušenje i kontracepcija najopasniji za razvoj raka.

8. LITERATURA

[1] https://medium.com/@chandradip93/minmaxscaler-7ee697b9e89
[2] https://medium.com/@iamkamleshrangi/how-min-max-scaler-works-9fbebb9347da
[3] https://medium.com/@kyawsawhtoon/a-guide-to-knn-imputation-95e2dc496e
[4]https://www.geeksforgeeks.org/how-knn-imputer-works-in-machine-learning/#step-3-find-the-nearest-neighbors
[5] https://www.geeksforgeeks.org/principal-component-analysis-pca/
[6] https://www.baeldung.com/cs/pca

[7] https://www.statology.org/breusch-pagan-test/
[8] https://matematika.pmf.uns.ac.rs/wp-content/uploads/zavrsni-radovi/primenjena_matematika/NevenaNesic.pdf
[9] https://www.geeksforgeeks.org/how-to-perform-a-breusch-pagan-test-in-python/
[10] https://spureconomics.com/white-test-for-heteroscedasticity/
[11] https://www.geeksforgeeks.org/weighted-least-squares-regression-in-python/
[12] https://medium.com/@chandu.bathula16/machine-learning-concept-69-random-sample-consensus-ransac-e1ae76e4102a
[13] https://www.statology.org/mse-vs-rmse/
[14] https://www.datacamp.com/tutorial/adjusted-r-squared
[15] https://medium.com/@meir412_37692/d-a-r-t-your-new-weapon-against-overfitting-in-boosting-models-9ea4e6aa435b
[16] https://www.geeksforgeeks.org/ml-gradient-boosting/
[17] https://scikit-learn.org/dev/modules/generated/sklearn.linear_model.HuberRegressor.html
[18] https://tahera-firdose.medium.com/treating-outliers-using-iqr-and-percentile-approach-part-2-9d8c4ec55af7#:~:text=The%20percentile%20method%20is%20a,with%20the%20corresponding%20threshold%20values.
[19] https://ncss-tech.github.io/soil-range-in-characteristics/why-percentiles.html#:~:text=Percentiles%20require%20no%20distributional%20assumptions,max%20of%20the%20observed%20data.
[20] https://www.geeksforgeeks.org/applying-pca-to-logistic-regression-to-remove-multicollinearity/
[21] https://www.geeksforgeeks.org/machine-learning-outlier/