**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 predvidjanje 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*.

1. **BIBLIOTEKE I PODACI KOJI SU KORIŠĆENI**

Biblioteke koje će biti korišćene su:

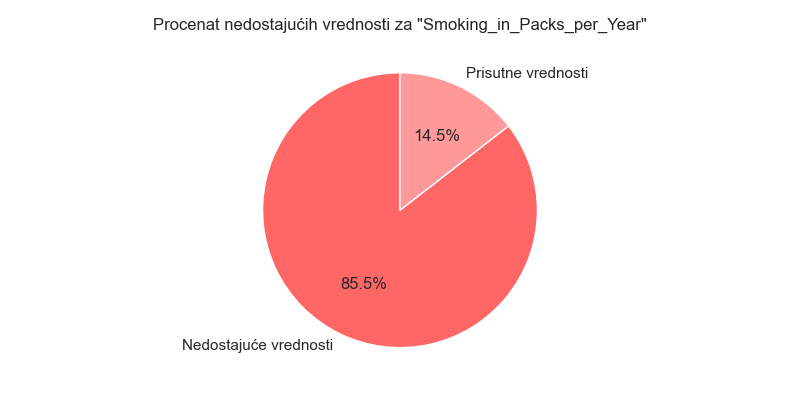
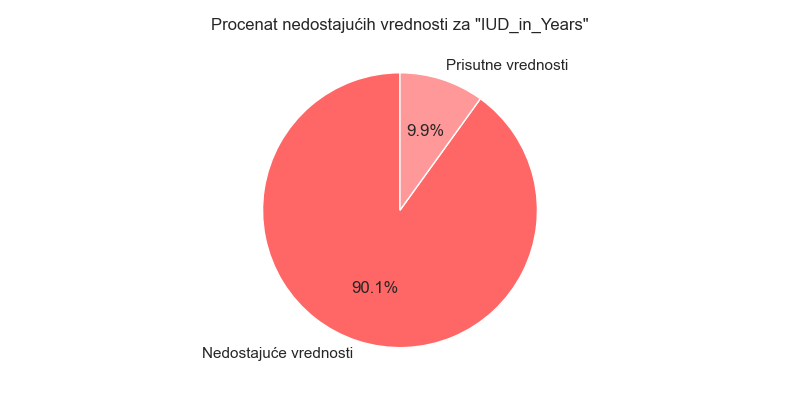
* *Pandas*
* *Sklearn*
* *Numpy*
* *StatsModels*

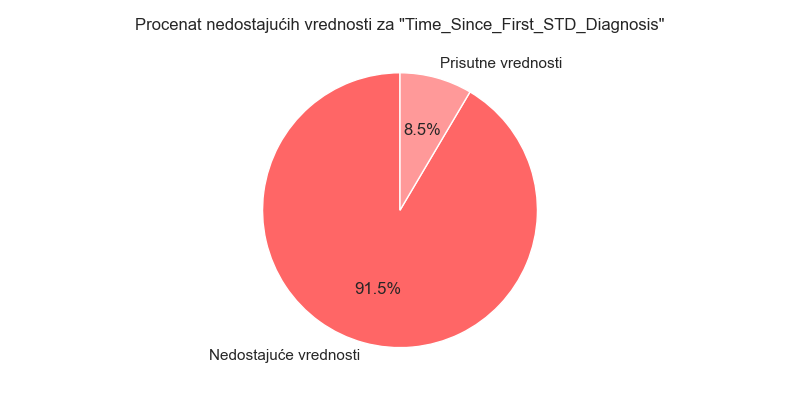
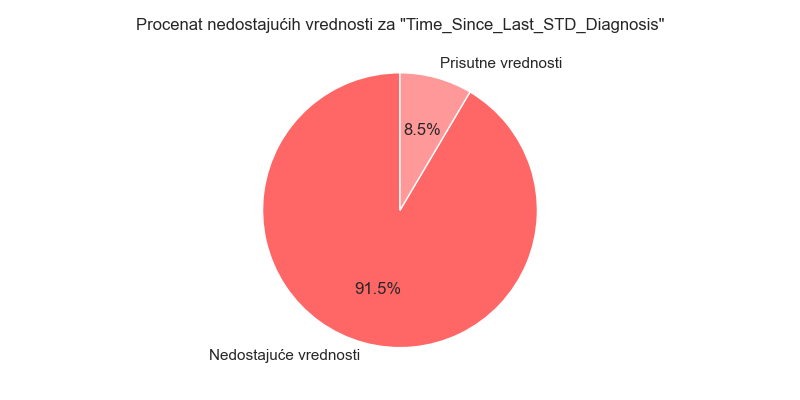
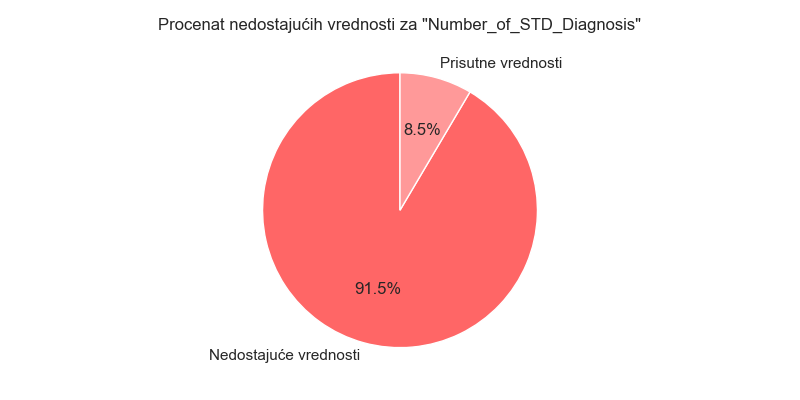
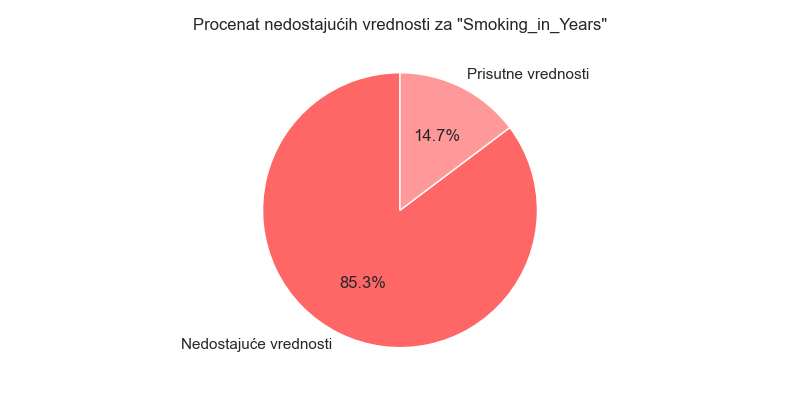
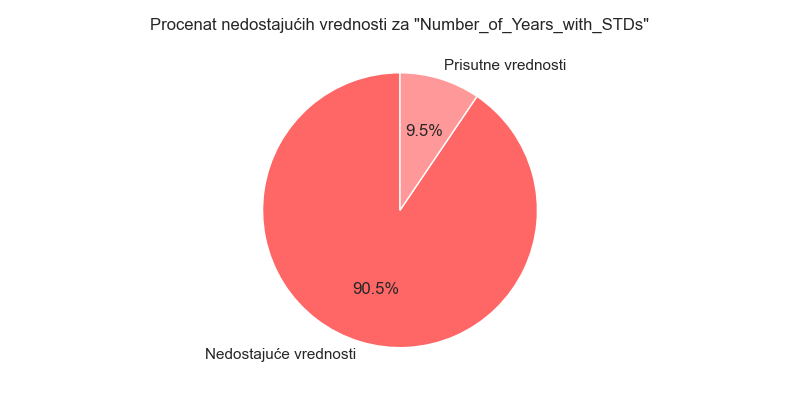
Skup podataka nad kojim će biti primenjeni modeli linearne regresije je *Cervical Cancer Screening* (<https://www.johnsnowlabs.com/marketplace/cervical-cancer-screening/>). Dataset sadrži 36 kolona (atributa) i 858 redova (ispitanih pacijenatkinja). Kolona koja će biti prediktovana je *Is\_Diagnosis\_Cancer* 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 seksualnih 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 žena 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

1. **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* bilo je razmatrano njihovo uklanjanje.

* *,

Da bi se utvrdio uticaj ovih nezavisnih promenljivih na zavisnu varijablu odredjena 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*

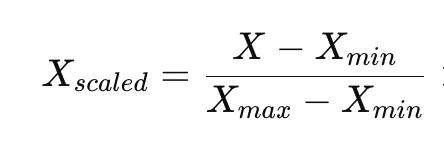
*Smoking\_in\_Years 0.271453*

*Smoking\_in\_Packs\_per\_Year 0.391916*

Od predloženih kolona za uklanjanje ostavljena je jedino *Smoking\_in\_Packs\_per\_Year.*

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 odredjenog broja najbližih suseda koji su pronadjeni. 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 takodje 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:



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 outlayere, jer oni imaju značajan uticaj na minimalne i maksimalne vrednosti. [1] [2]

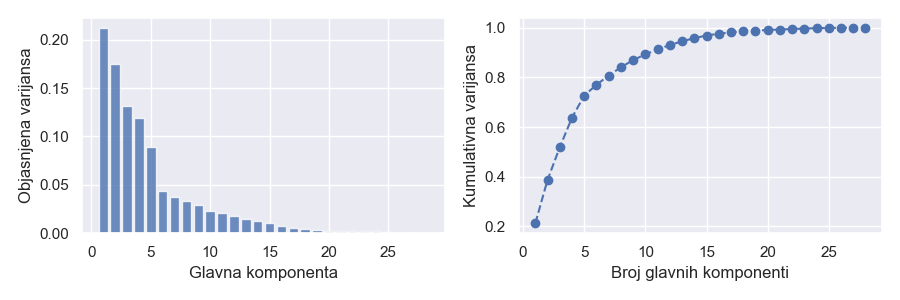
Potom je izvršena podela dataset-a na trening, validacioni i test skup u odnosu 60/20/20. Training skup je bio korišćen za obučavanje modela, dok su validacioni i test skupovi bili korišćeni za procenu performansi treninranog modela.

Nakon toga je kreiran OLS model, koji je predstavljao osnovni model koji je bio unapredjivan raznim metodama. 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 vezane za normalnost grešaka. U ovom slučaju je to opravdano, jer se radi sa velikim skupom podataka.

Zbog velikog broja atributa koji se nalazi 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 odredjivanju 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 izmedju 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]



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 (H₀)**: Homoskedastičnost je prisutna (varijansa reziduala je konstantna).
* **Alternativna hipoteza (Hₐ)**: 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 2.563569532217926×10−6. Na osnovu ovog rezultata je odbačena nulta hipoteza i zaključeno je da postoji heteroskedastičnost u modelu.

**White test** je takodje pokazao da je prisutna heteroskedastičnost, a to je zaključeno jer je rezultat testa 7.302063750636688×10−30 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 podatacima 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 prilagodjavanje 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…

1. 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.

1. 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.

1. **VIZUALIZACIJA PODATAKA**
2. **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 izmedju predvidjenih i posmatranih vrednosti. Manja vrednost uglavnom ukazuje na to da je model bolji. [13] Računa se po formuli:

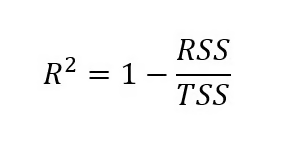
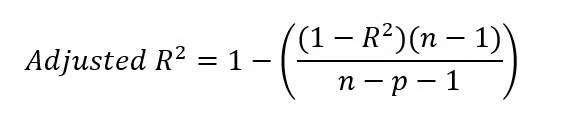
RMSE =

- prediktovana vrednost

- stvarna vrednost

n - broj podataka

* **Adjusted R-Squared** - daje informaciju o tome koliki deo varijanse zavisne promenljive objašnjavaju nezavisne promenljive. On takodje kažnjava uključivanje nebitnih promenljivih u model. Ova metrika se u projektu koristi jer je pogodna za uporedjivanje 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]

RSS - suma kvadrata reziduala

TSS - ukupna suma kvadrata

n - broj podataka

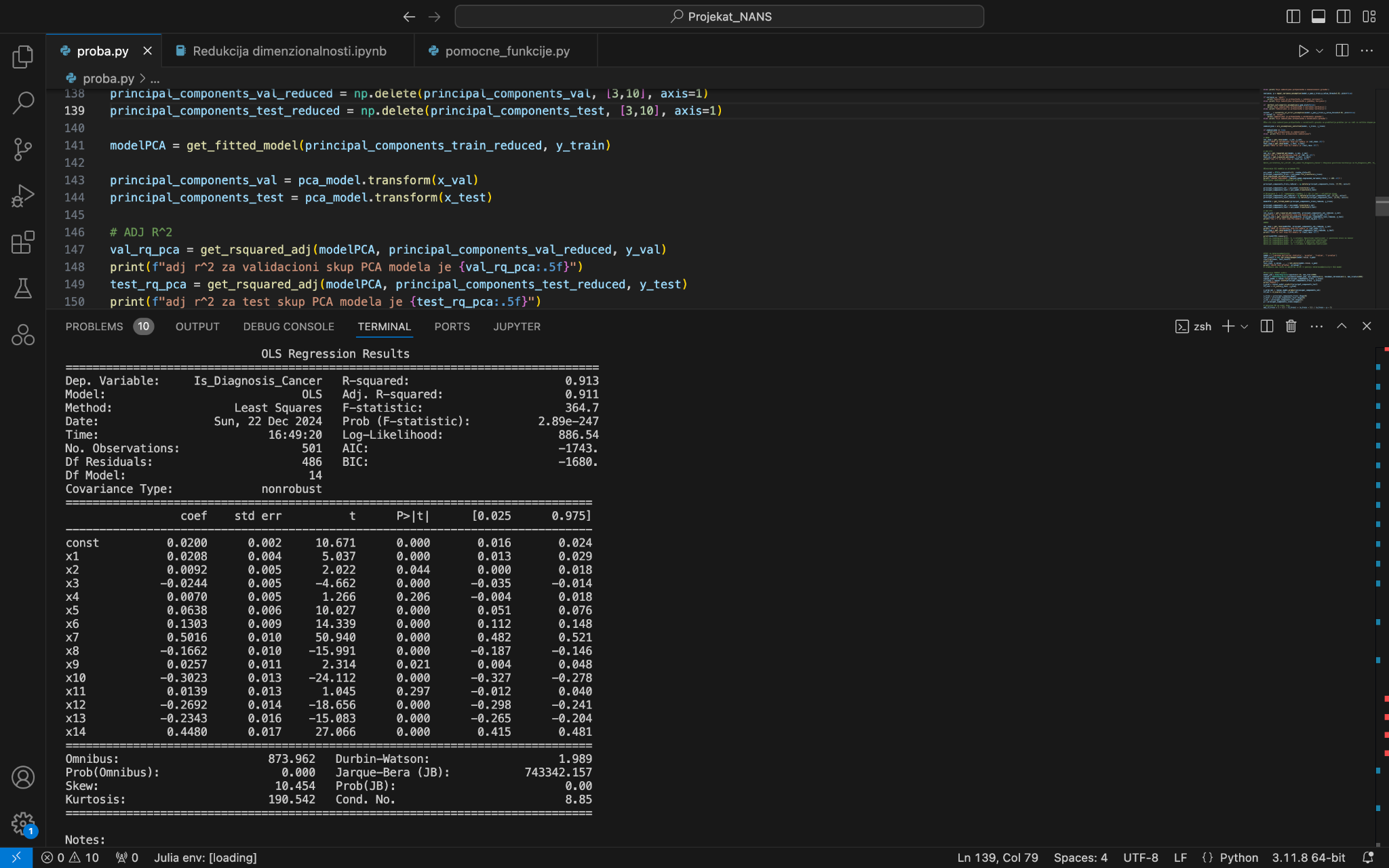
p - broj nezavisnih promenljivih

1. **REZULTATI**

Za OLS model rezultati metrika su:

* Validacioni skup
* RMSE je 0.05635
* Adjusted R-Squared je 0.86753
* Test skup
* RMSE je 0.09977
* Adjusted R-Squared je 0.31636

Nakon što primene PCA i analize karakteristika modela, uočeno je da su p-vrednosti nekih komponenti veće od 0.05 što je ukazivalo na to da model potencijalno može da bude bolji bez njih.



One su potom izbačene i dobijeni su bolji rezultati metrika.

Za PCA model metrike su:

* Validacioni skup
* RMSE je 0.05776
* Adjusted R-Squared je 0.87617
* Test skup
* RMSE je 0.09848
* Adjusted R-Squared je 0.40745

Za RANSAC model metrike su:

* Validacioni skup
* RMSE je 0.05844
* Adjusted R-Squared je 0.87158
* Test skup
* RMSE je 0.09936
* Adjusted R-Squared je 0.38882

Za WLS model metrike su:

* Validacioni skup
* RMSE je 0.05776
* Adjusted R-Squared je 0.87454
* Test skup
* RMSE je 0.09848
* Adjusted R-Squared je 0.39965

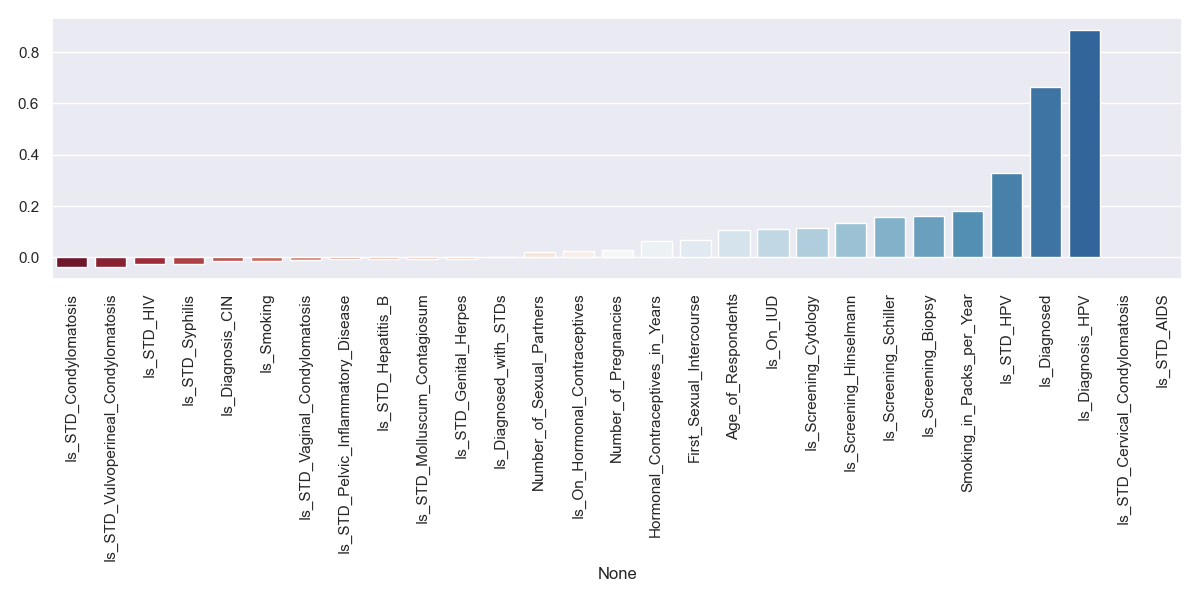
Najbolji rezultati metrika su dobijeni korišćenjem PCA modela, razlozi za to su sposobnog PCA algoritma da smanji uticaj šumova i mogućnost smanjivanja negativnih efekata multikolinearnosti medju atributima.

WLS model je takodje dao unapredjene rezultate u odnosu na početni OLS model razlog za to je karakteristika WLS modela da smanjuje heteroskedastičnost medju podacima.

RANSAC model je dao bolje rezultate nego početni OLS model jer nije osetljiv na prisustvo outlier-a.

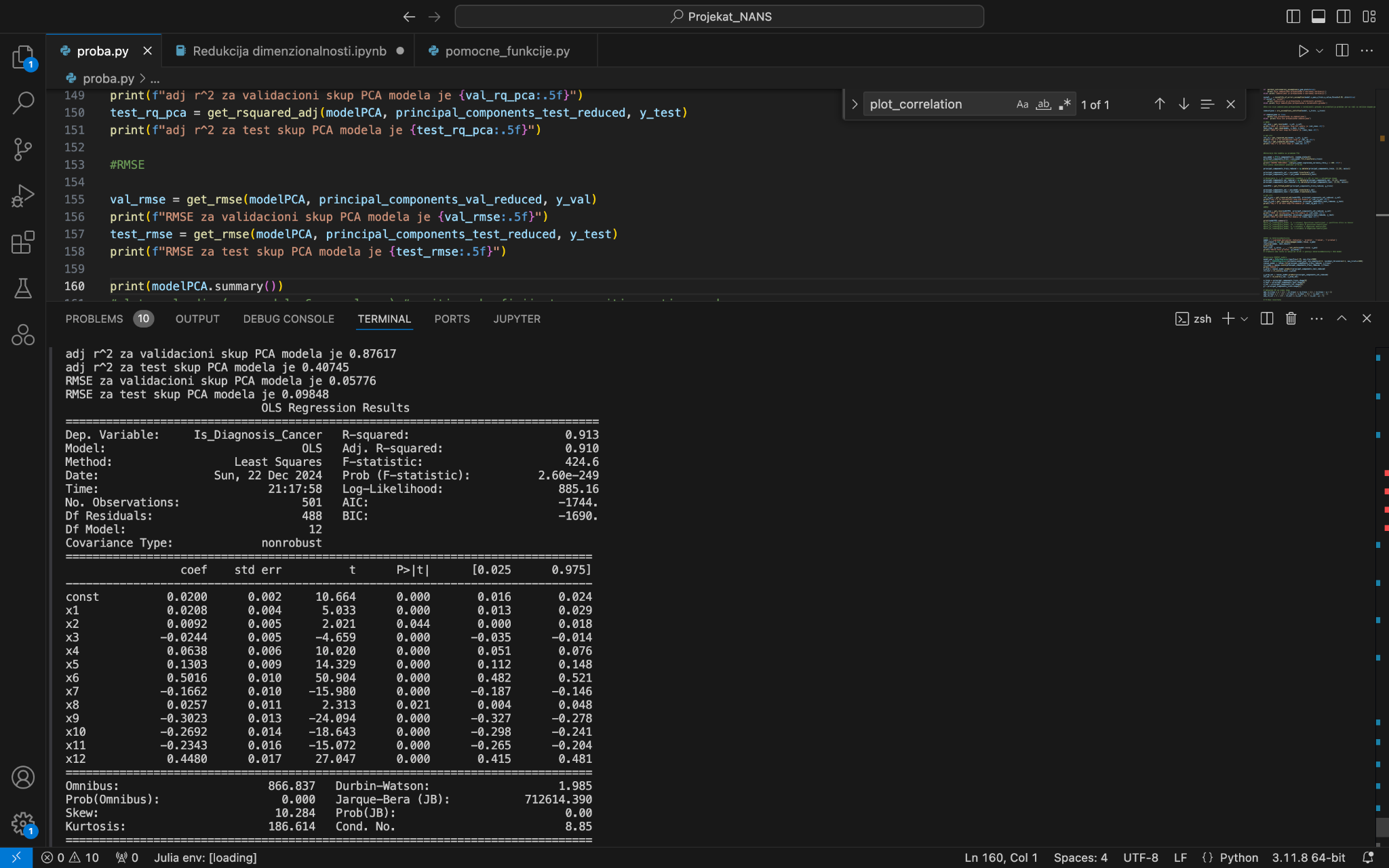
U nastavku će biti komentarisani uticaji atributa na ciljanu kolonu *Is\_Diagnosis\_Cancer*.

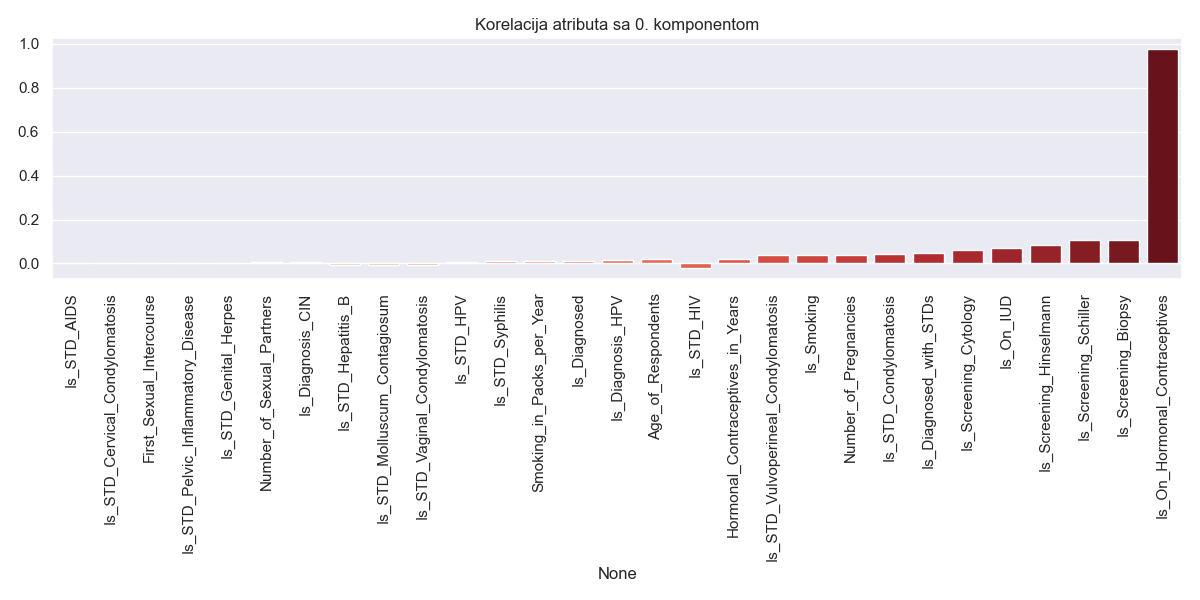
Na grafiku je prikazano koliko svaki atribut pojedinačno utiče na kolonu *Is\_Diagnosis\_Cancer* tj. posmatra se korelacija sa kolonom koja govori da li je oboljenje kancer ili ne.



Najveća pozitivna korelacija je sa atributoma *Is\_Diagnosis\_HPV* i *Is\_Diagnosed.* Dok je najveća negativna korelacija sa kolonom *Is\_STD\_Condylomatosis* vrlo mala oko 0.04 te njen uticaj 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 glvane komponente zavisiti dosta od ovih atributa.

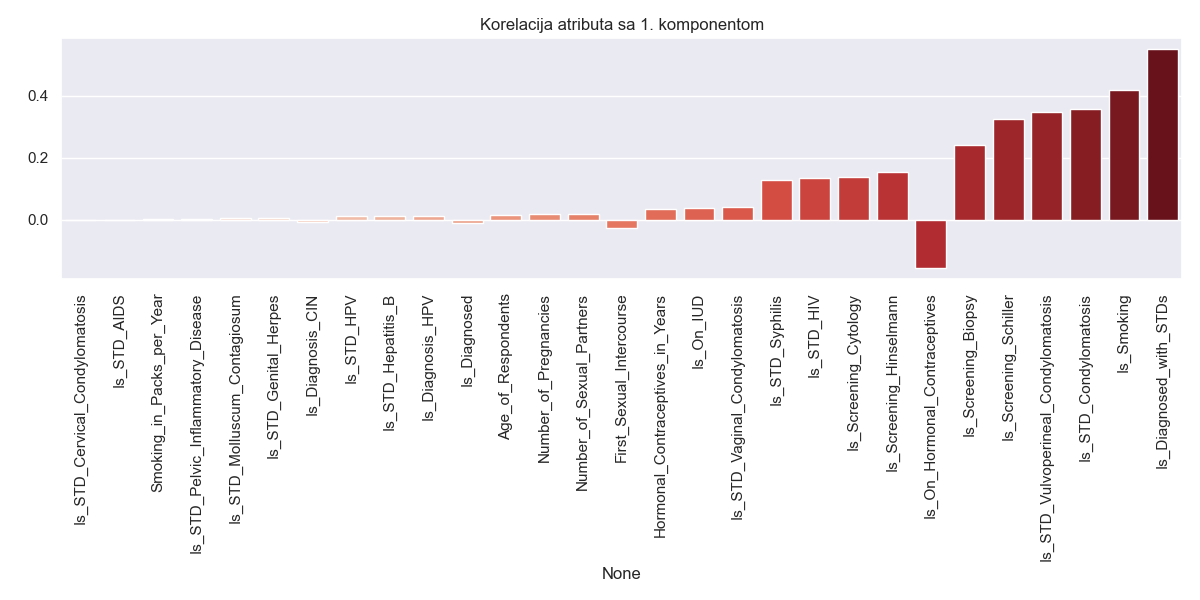
Potom je bilo neophodno odrediti glavne komponente i posmatrati korelacije glavnih komponenti i originalnih atributa. Bilo je neophodno posmatranje i znaka koeficijenata glavnih komponenti da bi se utvrdilo da li je dejstvo inverzno ili ne, ovi podaci su pročitani iz summary-ja.



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. 

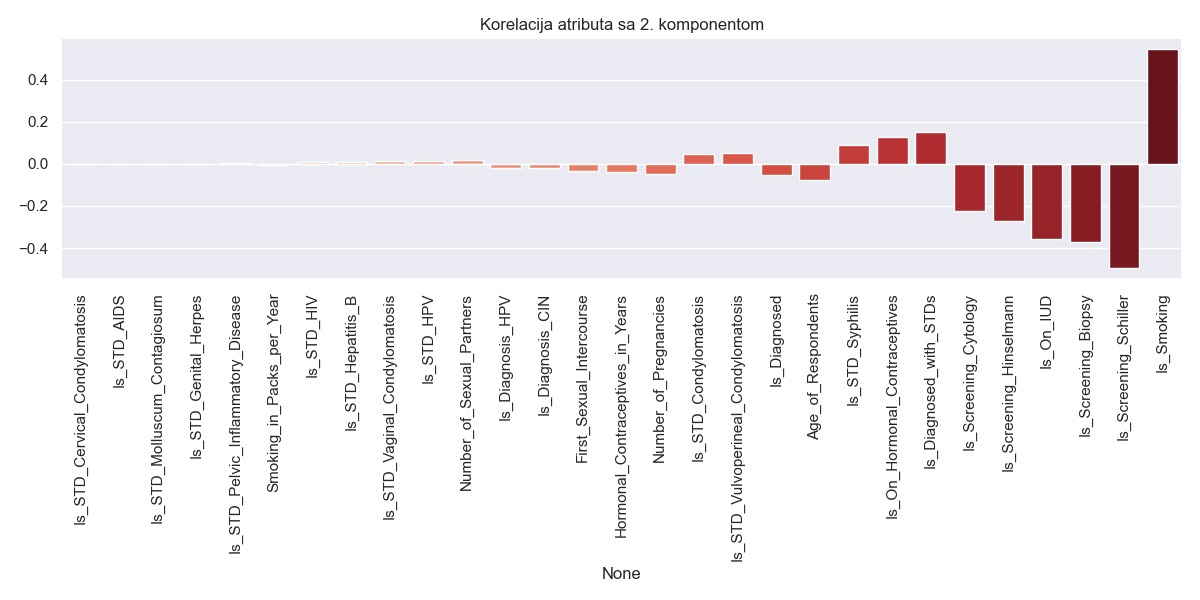
*Slika PC1*

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…Na osnovu blago 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 blago utiču na dobijanje karcinoma grlića materice.



*Slika PC2*

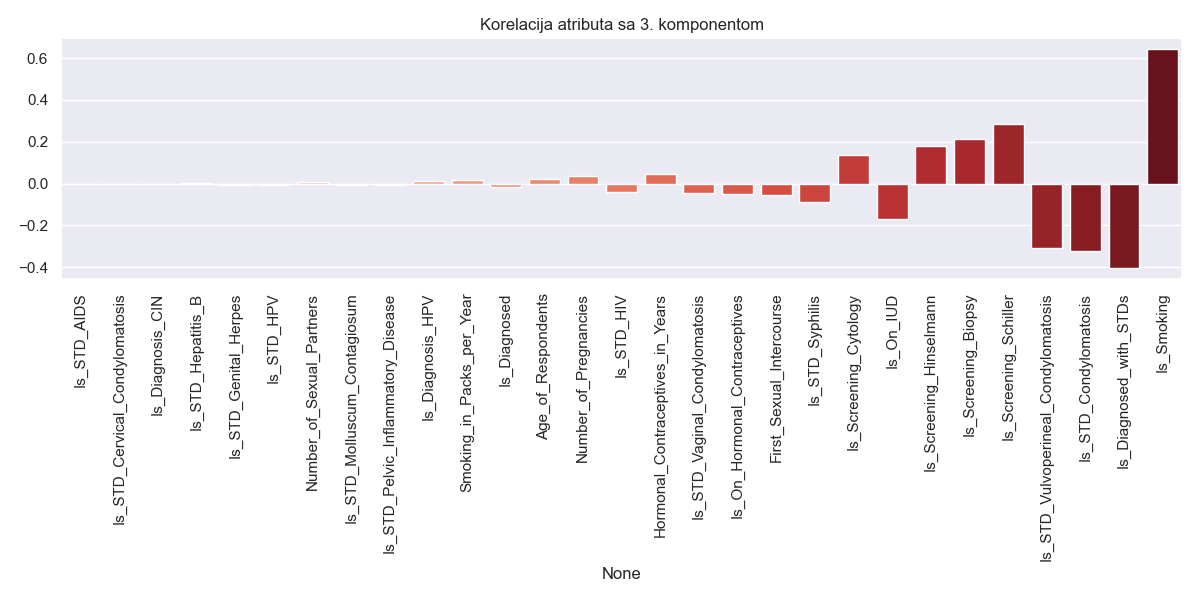
Sa grafika se može zaključiti da ova komponenta ima najveću korelaciju sa time da li je pacijentkinja pušač i da li su korišćeni medicinski mehanizmi poput Schiller-ove metode i biopsije za predikciju raka grlića materice i da li je korišćena spirala…Na osnovu negativnog koeficijenta uz PC3 se može zaključiti da se smanjuje verovatnoća da pacijentkinja ima rak kada PC3 raste. Tj. da upotreba spirale povećava verovatnoću za dobijanje raka. Po tumačenju rezultata ove komponente se zaključuje da to da li je pacijentkinja pušač ne podstiče dobijanje raka grlića materice.



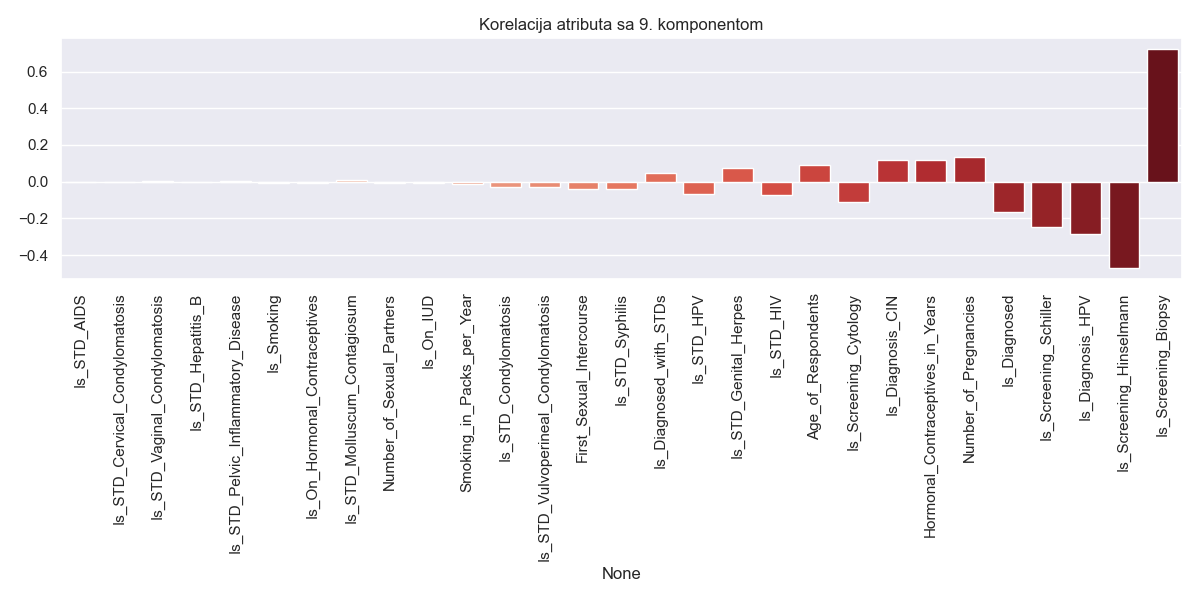
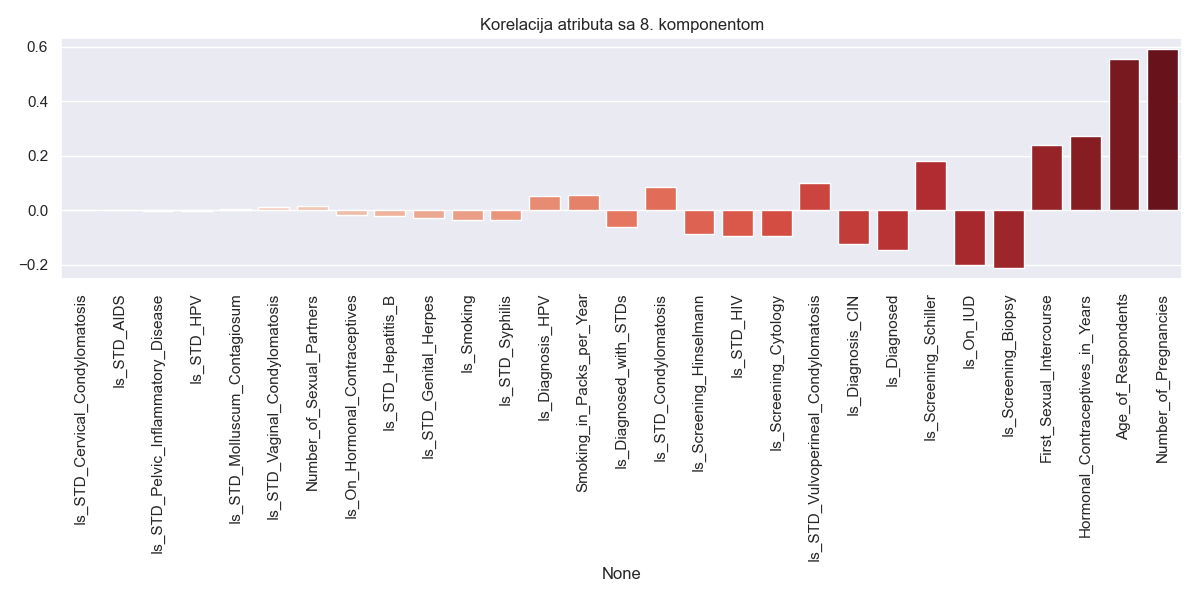
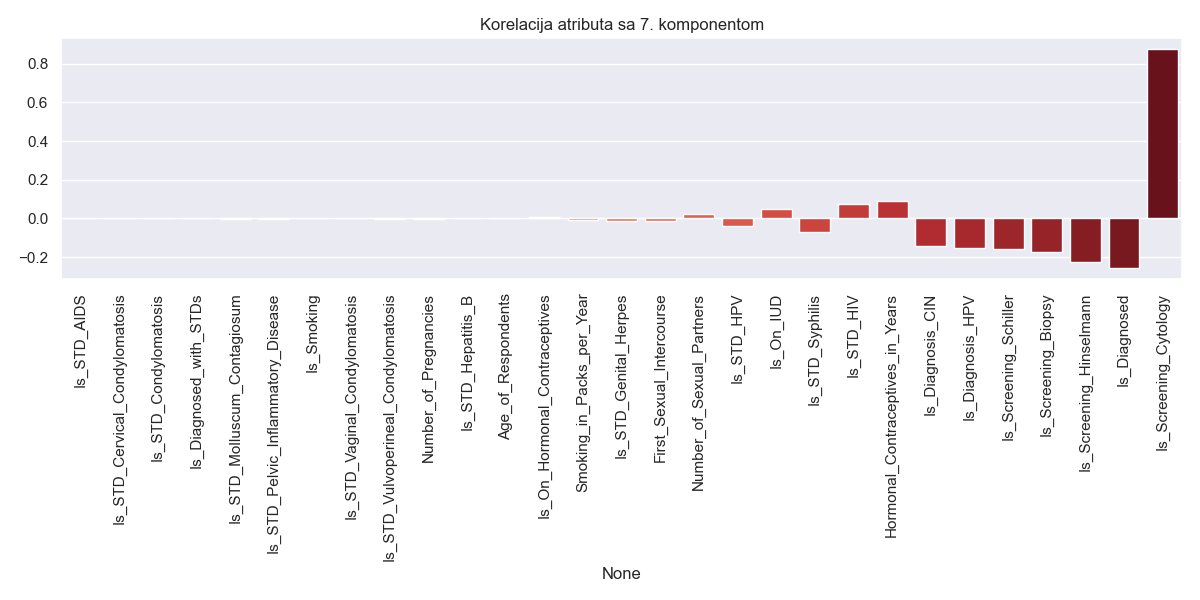
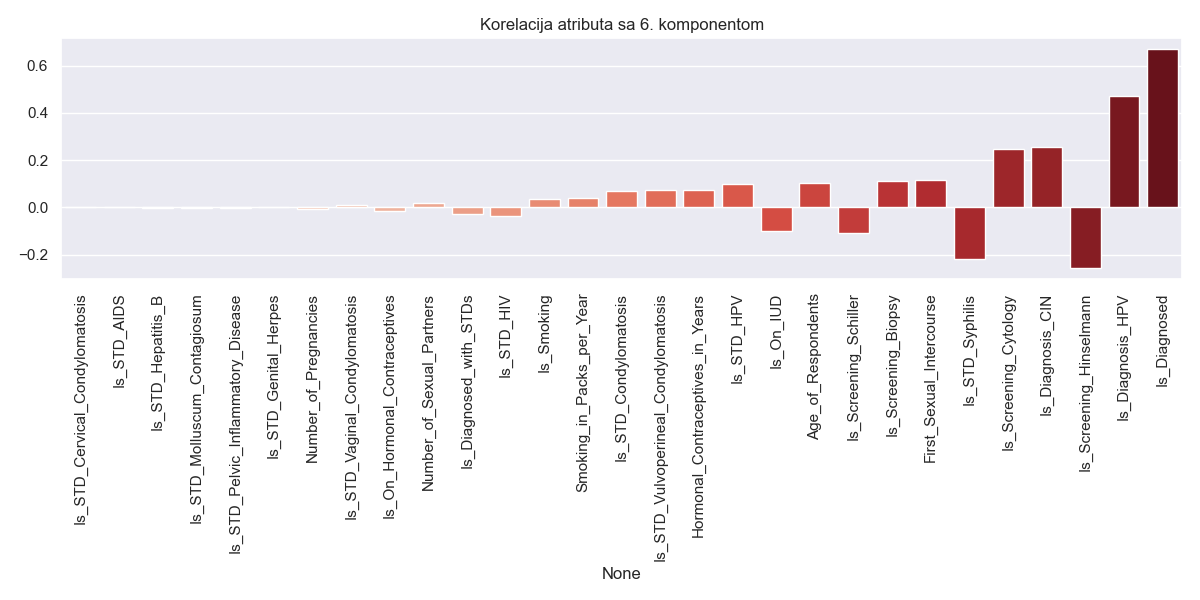
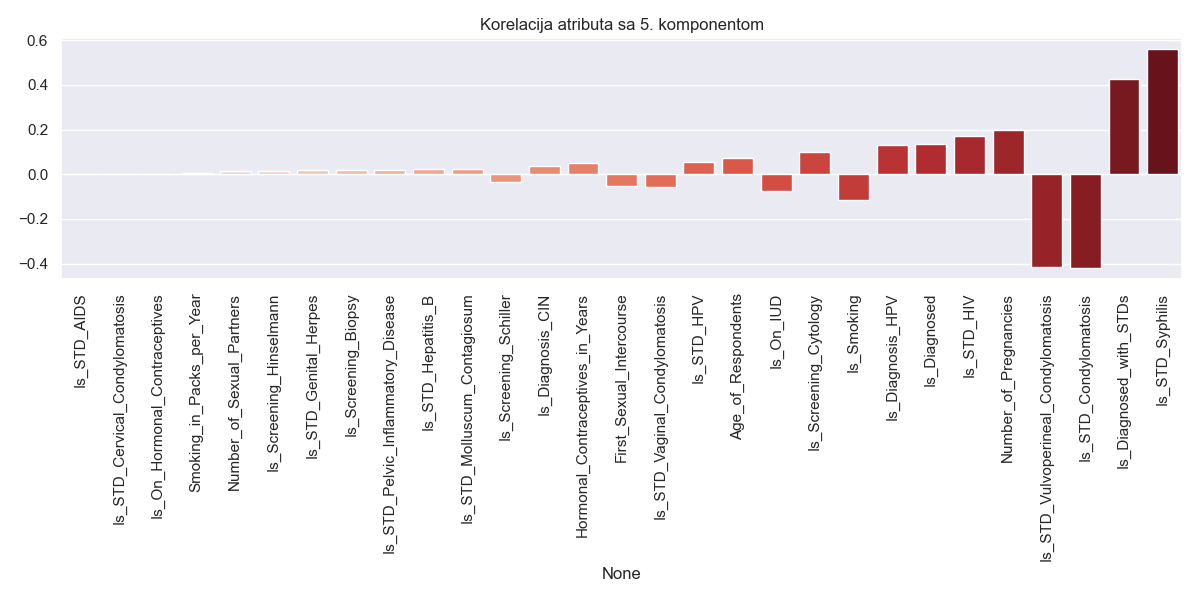
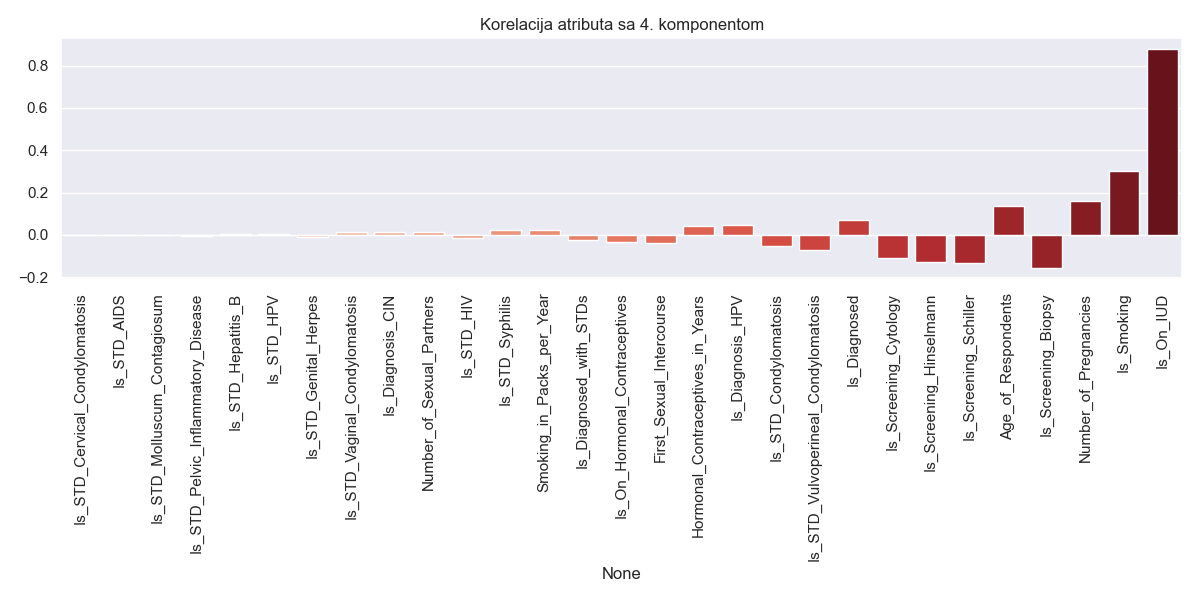
*Slika PC3*

Sa grafika se može zaključiti da ova komponenta ima najveću korelaciju sa time da li je pacijentkinja pušač i da li su pacijentkinji dijagnostikovane polne bolesti…Na osnovu pozitivnog koeficijenta uz PC4 se može zaključiti da se povećava verovatnoća da pacijentkinja ima rak kada PC4 raste. Tj. da pušenje podstiče razvoj raka.

Pošto je u komponenti najvećeg prioriteta primećeno da pacijentkinje koje puše imaju veću verovatnoću da razviju rak, kao i u komponentama sa velikim koeficijentima može se zaključiti da pušenje podstiče pojavu raka grlića materice.



*Slika PC4*



1. **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 uporedjivanje korišćenih metodologija i dobijenih rezultata.

U radu nije izvršeno splitovanje na training, test i validacioni skup podataka, već je izvršena *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, PCA, 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 izvršena i dodatno smanjenje dimenzionalnosti problema primenom PCA. S toga se može zaključiti da je robusnost modela smanjena zbog odbacivanja velikog dela podataka.

1. **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/