**Univerzitet u Beogradu**

**Fakultet organizacionih nauka**

**A close up of a sign

Description automatically generated**

**Izveštaj**

***Kaplan Meier* analiza**

* **COVID - 19** -

**Predmet:** Bioinformatika

**Profesor:** **Student:**

Dr Veljko Jeremić Brankica Jovanović

4011/19

**Sadržaj:**

[**1.** **Uvod** 3](#_Toc46434045)

[**2.** **Analiza preživljavanja** 4](#_Toc46434046)

[*2.1.* *Kaplan Meier* analiza preživljavanja 5](#_Toc46434047)

[*2.2.* *Kaplan Meier* analiza oporavka 9](#_Toc46434048)

[**3.** **Zaključak** 13](#_Toc46434049)

[**4.** **Reference:** 14](#_Toc46434050)

# **Uvod**

Analiza preživljavanja je deo statistike koji se bavi analizom vremena do pojave nekog događaja. Predmet posmatranja je slučajna promenljiva koja ima pozitivne vrednosti, kao što su na primer u medicini vreme do smrti, dužina boravka u bolnici, vreme do oporavka prilikom primene neke terapije, ili u ekonomiji trajanje štrajka, nezaposlenosti i slično[5](http://www.dmi.uns.ac.rs/site/dmi/download/master/primenjena_matematika/KristinaPopadic.pdf).

Cilj ovog rada je upoznavanje sa osnovnim pojmovima analize preživljavanja i pokazivanje primera primene u analizi Korona virusa COVID-19 na primeru podataka [COVID19\_line\_list\_data1](https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=COVID19_line_list_data.csv).

Programski kod za analizu podataka rađen je u *python* programskom jeziku korišćenjem biblioteka *pandas, numpy, matplotlib i lifelines*.

U prvom delu je prikazana analiza preživljavanja koja je u korelaciji sa starosnom dobi subjekata. Analizi je priključen i Koksov poluparametarski proporcionalni model koji omogućava uključivanje cenzurisanih podataka u procenu.

U drugom delu korišćen je isti skup podataka koji je sređen za potrebe date analize [COVID19\_changed](https://1drv.ms/x/s!Argjw8qp2oFrh8AUwJj8hCfLUYyyVQ?e=0wCdvr), gde je prikazana analiza oporavka pacijenata u vremenskom periodu merenom u danima od početka simptoma, a u skladu sa analizom prikazanom na primeru [Indije](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7343664/" \l "bib11)[2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7343664/" \l "bib11).

A close up of a cake

Description automatically generated

*Slika 1.* Korona virus[p](https://blog.okfn.org/2020/04/16/coronavirus-why-an-open-future-has-never-been-more-important/)

# **Analiza preživljavanja**

Novi koronavirus koji se pojavio 2019. godine nazvan je SARS-CoV-2. Otkriven je u Vuhanu, u Kini u decembru 2019. godine. Radi se o novom soju koronavirusa koji pre nije bio otkriven kod ljudi. Bolest uzrokovana tim virusom naziva se COVID-19. Koronavirusi su virusi koji cirkulišu među životinjama, ali neki od njih mogu preći na ljude. Nakon što pređu sa životinje na čoveka mogu se prenositi među ljudima.[3](https://covid19.rs/%D0%BF%D0%B8%D1%82%D0%B0%D1%9A%D0%B0-%D0%B8-%D0%BE%D0%B4%D0%B3%D0%BE%D0%B2%D0%BE%D1%80%D0%B8-%D0%BE-%D0%B1%D0%BE%D0%BB%D0%B5%D1%81%D1%82%D0%B8-%D1%83%D0%B7%D1%80%D0%BE%D0%BA%D0%BE%D0%B2%D0%B0%D0%BD/)

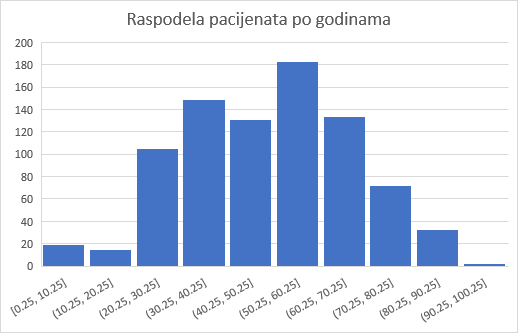
Svetska zdravstvena organizacija je proglasila pandemiju COVID-19. Prijavljeno je više od 15.424.553slučajeva u svetu, što je rezultiralo sa više od 631.237 smrtnih slučajeva širom sveta. Bolest se može manifestovati asimptomatski, može se javiti akutna respiratorna bolest i upala pluća.[4](https://www.worldometers.info/coronavirus/)

Odrasli predstavljaju populaciju sa najvećom stopom zaraze; međutim, novorođenčad, deca i stariji pacijenti takođe mogu biti zaraženi SARS-CoV-2. Najčešći nalaz na rentgenskom snimku grudnog koša kod pacijenata sa upalom pluća bila je upala pluća sa bilateralnim učešćem.[4](https://www.sciencedirect.com/science/article/pii/S1684118220300402)

Analiza preživljavanja Korona virusa COVID-19 prikazana je na primeru podataka [COVID19\_line\_list\_data1](https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=COVID19_line_list_data.csv).

Skup podataka sadrži informacije o potvrđenim pojedinačnim slučajevima Covid-19 u svetu . Podaci o nešto više od 1000 subjekata su prikupljani u periodu od 20. januara 2020. do 28. februara 2020 iz više zemalja.

Na *slici 1*. prikazana je raspodela pacijenata po godinama u podacima. Primećuje se da je najviše pacijenata u srednjoj do starijoj starosnoj dobi.



*Slika 1. raspodela pacijenata po godinama*

### *2.1. Kaplan Meier* analiza preživljavanja

Procena *Kaplana-Meiera* jedna je od najboljih opcija koja se koristi za merenje procenata živih subjekata tokom određenog vremena nakon tretmana ili bolesti. Vreme koje počinje od definisane tačke do nastanka određenog događaja, na primer smrt se naziva vremenom preživljavanja, a analiza grupnih podataka kao analizom preživljavanja.[7](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/)

Vreme preživljavanja procenjuje se od dana pojave simptoma do smrti. Pacijenti su smatrani cenzurisanim ako su se oporavili ili ostali na hospitalizaciji nakon 28. februara 2020. godine.

Na *slici 2*. prikazana je verovatnoća preživljavanja COVID-19 u zavisnosti od godina pacijenata. Primećuje se da se verovatnoća preživljavanja kod pacijenata mlađe starosne dobi kreće oko 1 odnosno 100%. Verovatnoća preživljavanja počinje da opada oko 60-te godine sa vrednošću oko 0.9 i oko 85-te godine je na 0.5 odnosno 50%, odnosno veća je verovatnoća da će se desiti smrt ukoliko je osoba starija od 60 godina.

**A screenshot of a map

Description automatically generated**

*Slika 2. Verovatnoća preživljavanja COVID-19 u zavisnosti od godina pacijenata*

Na *slici 3*. prikazana je verovatnoća preživljavanja u zavisnosti od godina i pola. Primećuje se da su žene u blagoj prednosti. Njihova verovatnoća preživljavanja kreće se oko 1 sve do približno 75-e godine. Verovatnoća preživljavanja kod muškog pola počinje da opada oko 50-e godine.

A close up of a map

Description automatically generated

*Slika 3. Verovatnoća preživljavanja COVID-19 u zavisnosti od godina pacijenata i pola*

A close up of a map

Description automatically generated

*Slika 4 -*

* *Pacijent 47*. new confirmed imported COVID-19 pneumonia patient in Dazhou, Sichuan: male, 36, return to Dazhou from Wuhan on 01/17/2020, symptom onset on 01/18/2020 and hospitalized.
* *Pacijent 49*. Death from COVID-19 pneumonia in Wuhan: male, 61, symptom onset on 12/20/2019, hospitalized on 12/27/2019, death on 01/09/2020.

*Slika 4.* pokazuje dva primera primene Kaplan Meier analize za slučajeve u zavisnosti od godina, kada je drugi slučaj imao verovatnoću da će preživeti oko 70%, a smrt se, na žalost, desila. To potvrđuje da model nije idealan, kao ni podaci, a pored toga se srećemo i sa nepoznatim virusom koji ne pokazuje konstantnost. Potrebno je više podataka i više vremena kako bismo mogli da analiziramo ovaj virus.

A screenshot of a social media post

Description automatically generated

A picture containing monitor, screen, holding, television

Description automatically generated

*Slika 5. Sumarna statistika*

Sumarna statistika prikazana na *slici 5* ukazuje na značaj atributa u predviđanju. Vidimo da su pol i mogućnost da ste iz Vuhana statistički značajne tj. p vrednost je < 0.05.

### *Kaplan Meier* analiza oporavka

Analiza je sprovedena na skupu podataka [COVID19\_changed](https://1drv.ms/x/s!Argjw8qp2oFrh8AUwJj8hCfLUYyyVQ?e=0wCdvr), koji je za potrebe ove analize preprocesiran odnosno sređen iz izvornog skupa podataka.

Originalni podaci su u okviru *recovered* kolone imali datum oporavka, zatim 0 (ukoliko se pacijent nije oporavio) i mali broj 1 (ako se pacijent jeste oporavio bez poznavanja datuma).

U koloni *death* je postojao datum smrti i 1 (ako se desila smrt bez poznatog datum), kao i 0 (ako pacijent nije umro).

Cilj pripreme podataka bio je da se naprave dve kolone - kolona *recovered* (0 ili1) i druga kolona sa datumom događaja *recoveredDate*.

Preprocesiranje skupa podataka u *Excel*-u obuhvatilo je:

* dupliranje kolone *recovered*, preimenovanje u *recoveredDate*
* obrisane su vrste kod kojih je recovered=1 zato sto ne znamo datum oporavka
* *recovered* - postaviti recovered = 1 gde postoji datum
* *recovered* - obrisati vrste u kojima je *recovered*= 0 i nije nastupila smrt (death=0– pacijent se jos uvek leci)
* *recoveredDate* - u slucaju smrti (death=1, recovered=0) stavljamo najveci moguci datum 28.2.2020. koji smatramo da se ne koristi za analizu vec samo za konzistentnost tipa podataka u toj koloni
* Ovim preprocesiranjem postižemo da Kaplan Meier kriva ne ide do 100% oporavka, već uključujemo i broj potvrdjenih umrlih.

Vreme oporavka (u danima) procenjuje se od dana pojave simptoma do oporavka. Pacijenti su smatrani cenzurisanim ako su umrli ili ostali na hospitalizaciji nakon 28. februara 2020. godine.

A screenshot of a social media post

Description automatically generated

*Slika 6. Verovatnoća oporavka od COVID-19 u danima*

Na *slici 6*. odnosno na inverznoj *slici 7.* (koja logičnije prikazuje verovatnoću oporavka) vidimo da u prvih 10 dana nema oporavka kod pacijenata zaraženih korona virusom. Nakon 10 dana, verovatnoća

raste i dostiže maksimum koji je potreban za oporavak od 35-40 dana. Broj pacijenata koji nije postigao oporavak čini oko 33%.

A screenshot of a social media post

Description automatically generated

*Slika 6. Verovatnoća oporavka od COVID-19 u danima*

A screenshot of a cell phone

Description automatically generated

Na *slici 8*. prikazana je verovatnoća oporavka u zavisnosti od pola po danima. Primećuje se da nema statistički značajne razlike u broju dana do oporavka, mada je verovatnoća oporavka veća kod muškaraca za oko 20% nego kod žena.

A close up of a device

Description automatically generated

* *Pacijent 48.* - new confirmed COVID-19 patient in US: male, in his 60s, husband of 1/24-No.1, Chicago, admitted to hospital 1/28/2020, symptom onset 1/28/2020
* *Pacijent 49.* - new confirmed COVID-19 patient in Hong Kong: male, 25, lives in Kowloon City, now in hospital, had not travelled outside Hong Kong, symptom onset 1/28/2020 (cough, runny nose, diarrhea), went to hospital 1/29/2020

*Slika 9* pokazuje dva primera primene Kaplan Meier analize za slučajeve u zavisnosti od godina, gde se može primetiti da mlađi pacijent ima veću verovatnoću oporavka. Vreme oporavka je u oba slučaja isto odnosno kreće se u rasponu do 30 dana.

A screenshot of a cell phone

Description automatically generated

A picture containing monitor, screen, holding, television

Description automatically generated

Sumarna statistika prikazana na *slici 5* ukazuje na značaj atributa u predviđanju. Vidimo da je mogućnost da ste iz Vuhana statistički značajna tj. p vrednost je < 0.05.

# **Zaključak**

Metoda Kaplana-Mejera je korisna metoda koja može igrati značajnu ulogu u generisanju informacija zasnovanih na dokazima o vremenu preživljavanja.

Na osnovu malog uzorka podataka, koristeći ovu metodu pokazano je da starosna dob pacijenta ima značajnu korelaciju sa verovatnoćom oporavka – za starost manju od 40 godina, ovaj virus, je sa veoma malom verovatnoćom izazivao smrtni ishod, dok sa povećanjem starosne dobi ka granici od 90 je verovatnoća smrti rasla.

Sprovedena analiza oporavka na osnovu broja dana od početka simptoma, sugerišse da je kod 50% pacijenata potvrdjen oporavak nakon 3 nedelje a oko 70% pacijenata se oporavilo u roku od 40 dana ili više dana. 30% pacijenata na ovom uzorku je preminulo.

U zadatku su opisane osnovne Kaplan Meier krive koje su prikazale statistički značajne atribute i odgovorile na pitanja vezana za koronavirus. Možemo zaključiti da je potrebno dalje praćenje atributa i prikupljanje sređenih podataka koji su nesumnjivo važni u statističkim analizama.

# **Reference:**

1. [https://www.kaggle.com/](https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=COVID19_line_list_data.csv)
2. [https://www.ncbi.nlm.nih.gov/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7343664/#bib11)
3. [https://covid19.rs/](https://covid19.rs/%D0%BF%D0%B8%D1%82%D0%B0%D1%9A%D0%B0-%D0%B8-%D0%BE%D0%B4%D0%B3%D0%BE%D0%B2%D0%BE%D1%80%D0%B8-%D0%BE-%D0%B1%D0%BE%D0%BB%D0%B5%D1%81%D1%82%D0%B8-%D1%83%D0%B7%D1%80%D0%BE%D0%BA%D0%BE%D0%B2%D0%B0%D0%BD/)
4. <https://www.worldometers.info/coronavirus/>
5. <http://www.dmi.uns.ac.rs/site/dmi/download/master/primenjena_matematika/KristinaPopadic.pdf>
6. <https://matematika.pmf.uns.ac.rs/wp-content/uploads/zavrsni-radovi/primenjena_matematika/NikolinaLjiljak.pdf>
7. [https://www.ncbi.nlm.nih.gov/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/)
8. [https://blog.okfn.org/](https://blog.okfn.org/2020/04/16/coronavirus-why-an-open-future-has-never-been-more-important/)