Matematički fakultet

Univerzitet u Beogradu

Seminarski rad:

**Uticaj P-adičnosti na razilke genetskog koda SARS1, SARS2 i MERS koronavirusa**

Mentor: Studenti:

Prof. dr Nenad Mitić Marko Koprivica Milan Bodo

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

[**2.P-adična distanca 3**](#_30j0zll)

[**2.1 P-adično kodiranje nukleotidnih sekvenci 3**](#_1fob9te)

[2.2 P-adična distanca izmedju kodona 3](#_3znysh7)

[**3.Podaci 4**](#_2et92p0)

[**4. Modeli 5**](#_tyjcwt)

[4.1 K- najbližih suseda sa odsecanjem dela sekvence 5](#_3dy6vkm)

[4.1.1 Ocena modela za klasifikaciju po tipu virusa 5](#_1t3h5sf)

[4.1.2 Ocena modela za klasifikaciju po tipu proteina 6](#_4d34og8)

[4.1.3 Ocena modela za klasifikaciju po soju virusa za SARS2 7](#_2s8eyo1)

[4.2 K- najbližih suseda sa celim sekvencama 7](#_17dp8vu)

[4.2.1 Ocena modela za klasifikaciju po tipu virusa 7](#_3rdcrjn)

[4.2.2 Ocena modela za klasifikaciju po tipu proteina 7](#_26in1rg)

[4.2.3 Ocena modela za klasifikaciju po soju virusa za SARS2 7](#_lnxbz9)

[4.3 K- najbližih suseda sa celim sekvencama i jednostavnom metrikom 10](#_39ag2h8odln)

[4.3.1 Ocena modela za klasifikaciju po tipu proteina 10](#_upa2obghezm0)

[4.3.2 Ocena modela za klasifikaciju po soju virusa za SARS2 11](#_fo8oey425amn)

[4.4 Unakrsna provera modela 11](#_ixpwbczdvblo)

[**5. Zaključak 14**](#_44sinio)

[**6.Reference 14**](#_2jxsxqh)

# 1.Uvod

Ovaj seminarski rad bavi se uticajem p-adičnosti na različitosti genetskog koda kod SARS1, MERS i SARS2 virusa, i to na primerima klasifikacije ovih virusa, klasifikacije proteina virusa i klasifikacije sojeva SARS2 virusa po WHO klasifikaciji.

# 2.P-adična distanca

Kako se ∀m∈Z , m≠0 može zapisati kao m=pk•а gde je p prost broj, k ∈{0,1,2,...} i a ceo broj koji nije delji sa p. P-adična distanca broja m u oznaci |m|p  se definiše na sledeći način:

|m|p ≔p-k [1]

i specijalno za m=0, |m|p≔0.

P-adičcna distanca dva cela broja x i y se racuna na sledeći način:

|x-y|p

P-adičcna distanca je ultrametrika, to jest za nju važi:

1. |x-y|p ≥ 0 , |x-y|p＝0 ⇔x＝y

2. |x-y|p＝|y-x|p

3. |x-y|p ≤ max {|x-z|p ,|y-z|p}

Primer:

|63-3|2= |60|2= |22•3•5|2= ½

|63-3|3= |60|3= |22•3•5|3= ⅓

## 2.1 P-adično kodiranje nukleotidnih sekvenci

Nukleotidna sekvenca se deli u kodone(niske od po tri nukleotida), koji se kodiraju u trocifrene brojeve, gde se svakom od nukleotida ({A,C,G i T}) pridružuje po jedan broj iz Z5={0,1,2,3,4}, i to na sledeći način:

C=1 A=2 U=3 G=4

Primetimo da se purinske aminokiseline(A,G) i pirinske aminokisleine(C,U) kodiraju kao cifre iste parnosti, što ćemo korisititi u računanju distanci izmedju kodona.

## 2.2 P-adična distanca izmedju kodona

Za računanje distance izmedju kodona korišćena je kombinacija 5-adične i 2-adične metrike [[1]](#_2jxsxqh) i to na sledeći način. Ako se kodoni razlikuju na prvoj nukleotidi, distanca se uvećava za 50=1, ako se kodoni razlikuju na drugoj nukleotidi, distanca se uvećava za 5-1=1/5, dok ako se kodoni razlikuju na trećoj nukleotidi i apsolutna razlika kodova na toj poziciji je 2 (što znači da su obe nukleotide na trećoj poziciji ili purinske ili pirinske) distnaca je uvećava za 2-1\*5-2=1/2\*1/25 ili ako ta razlika nije 2 distanca se uvećava za 20\*5-2=1\*1/25. Distanca izmedju sekvenci predstavlja zbir distanci izmedju kodona.

2.3 Implentacija računanja P-adične distance izmedju kodona

Računanje distanci izmedju svih sekvenci bez odsecanja u Pajton programskom jeziku nije uspela da se izvrši,te je to pokušano C++ programskom jeziku gde je uspelo. Kod tog programa se nalazi u *create\_distances\_traingle.cpp,*  u njemu se pravi dijagonalna matrica rastojanja izmedju svih sekvenci, koja se posle koristi za sve vrste klasifikacija.

# 3.Podaci

Istraživanje je izvođeno na bazi podataka od 13202 instanci. Najvažniji atributi su ime virusa, ime proteina, i nukleotidna sekvenca.

| Virus | Broj instanci |
| --- | --- |
| SARS1 | 18 |
| MERS | 1859 |
| SARS2 | 11325 |

Tabela 3.1 raspodela instanci po tipu virusa

Još za SARS2 razlikovali su se i sojevi virusa:

| Soj SARS2 virusa | Broj instanci |
| --- | --- |
| Alpha, Delta, Epsilon, Gamma, Iota, Omricon | 1500 |
| Eta | 628 |
| Beta | 586 |
| Zeta | 467 |
| Lambda | 410 |
| Kappa | 157 |
| Mu | 65 |
| Theta | 20 |

Tabela 3.2 raspodela instanci po soju SARS2 virusa

Raspodela proteina virusa:

| Protein | Broj instanci |
| --- | --- |
| ORF1ab polyprotein | 5755 |
| ORF1a polyprotein | 4153 |
| surface glycoprotein | 1666 |
| nucleocapsid phosphoprotein | 640 |
| ORF3a protein | 221 |
| membrane glycoprotein | 136 |
| ORF4b protein | 128 |
| ORF3 protein | 104 |
| ORF5 protein | 104 |
| ORF8 protein | 76 |
| ORF4a protein | 67 |
| ORF8b protein | 59 |
| envelope protein | 40 |
| ORF7a protein | 25 |
| ORF6 protein | 10 |
| ORF7b protein | 9 |
| ORF1b polyprotein | 7 |
| ORF10 protein | 2 |

Tabela 3.3 raspodela instanci po tipu proteina

Sve sekvence proteina su jedinstvene, nukleotidno kompletne i bez višeznačnih karaktera.

# 4. Modeli

Podaci u svim modelima su podeljenji u trening i test skup u u odnosu 2 : 1, stratifikovanim uzorkom.

## 4.1 K- najbližih suseda sa odsecanjem dela sekvence

Primenjen je algoritma K-najbližih suseda, gde je broj suseda 3 i korišćena je metrika objašnjena u delu [2.2](#_3znysh7) za računanje razdaljine izmedju instanci. Ovaj model je primenjen za klasifikaciju po tipu virusa, po tipu proteina i po soju virusa za SARS2.

Sve nukleotidne sekvence su odsečene na dužinu od 78 kodona.

### 4.1.1 Ocena modela za klasifikaciju po tipu virusa

Ovaj model na test skupu ima tačnost od 0.99 i matricu konfuzije kao u tabeli 4.1.1.1.

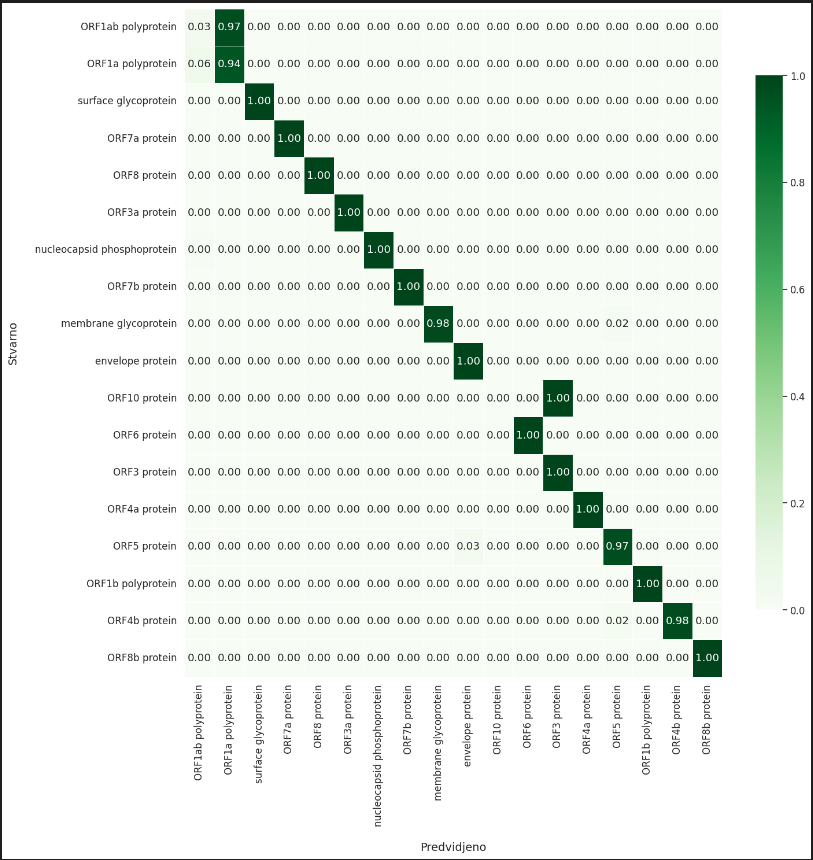
|  | Predvidjen MERS | Predvidjen SARS1 | Predvidjen SARS2 |
| --- | --- | --- | --- |
| Stavrno MERS | 610 | 0 | 1 |
| Stvarno SARS1 | 0 | 4 | 2 |
| Stvarno SARS2 | 1 | 0 | 3739 |

Tabela 4.1.1.1 Matrica konfuzije za klasifikaciju po tipu virusa

Primećujemo da ovaj model ima problem u razlikovanju malobrojnog SARS1 i SARS2 virusa. Tu postoji mogućnost za popravljanje ovog modela.

### 4.1.2 Ocena modela za klasifikaciju po tipu proteina

Ovaj model na test skupu ima tačnost od 0.56 i matricu konfuzije kao u tabeli(toplotnoj mapi) 4.1.2.1

Tabela 4.1.2.1 Matrica konfuzije za klasifikaciju po tipu proteina

Primetno je da se najbrojniji *ORF1ab polyprotein* skoro u potpunosti klasifikuje kao drugi po brojnosti *ORF1a polyprotein*, što najviše kvari tačnost ovog modela.

### 4.1.3 Ocena modela za klasifikaciju po soju virusa za SARS2

Ovaj model na test skupu ima tačnost od 0.2 te matricu konfuzije nema potrebe ni prikazivati.

## 4.2 K- najbližih suseda sa celim sekvencama

Primenjen je algoritma K-najbližih suseda, gde je broj suseda 3 i korišćena je metrika objašnjena u delu 2.2 za računanje razdaljine izmedju instanci. Ovaj model je primenjen za klasifikaciju po tipu virusa, po tipu proteina i po soju virusa za SARS2, korišćene su cele sekvence.

### 4.2.1 Ocena modela za klasifikaciju po tipu virusa

Rezultati ovog modela su slični kao i rezultati modela iz [4.1.1](#_1t3h5sf).

### 4.2.2 Ocena modela za klasifikaciju po tipu proteina

Primena modela na cele sekvence je ovde napravila značajna poboljšanja u prepoznavanju tipa proteina. Tačnost ovog modela je 0.99 i matrica konfuzije je prikazana u tabeli 4.2.2.1.

### 4.2.3 Ocena modela za klasifikaciju po soju virusa za SARS2

Primena modela na cele sekvence je u ovom slučaju klasifikacije napravila još značajnija poboljšanja u prepoznavanju soja SARS2 virusa u odnosu na model [4.1.3](#_2s8eyo1). Tačnost ovog modela je 0.96 i matrica konfuzije je prikazana u tabeli 4.2.3.1.

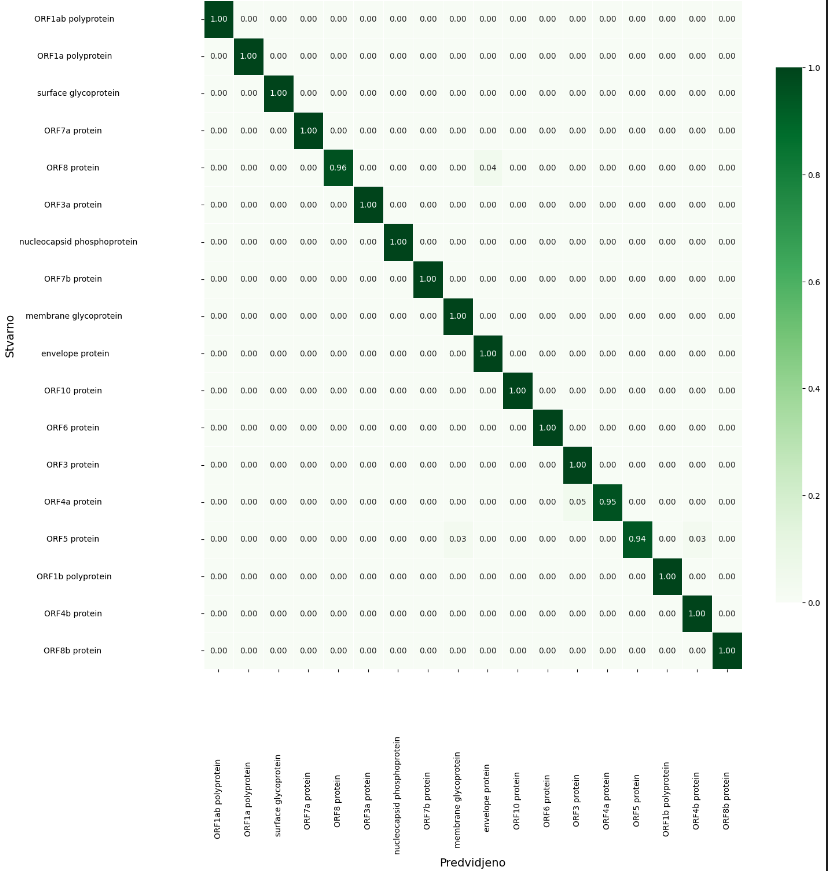


Tabela 4.2.2.1 Matrica konfuzije za klasifikaciju po tipu proteina

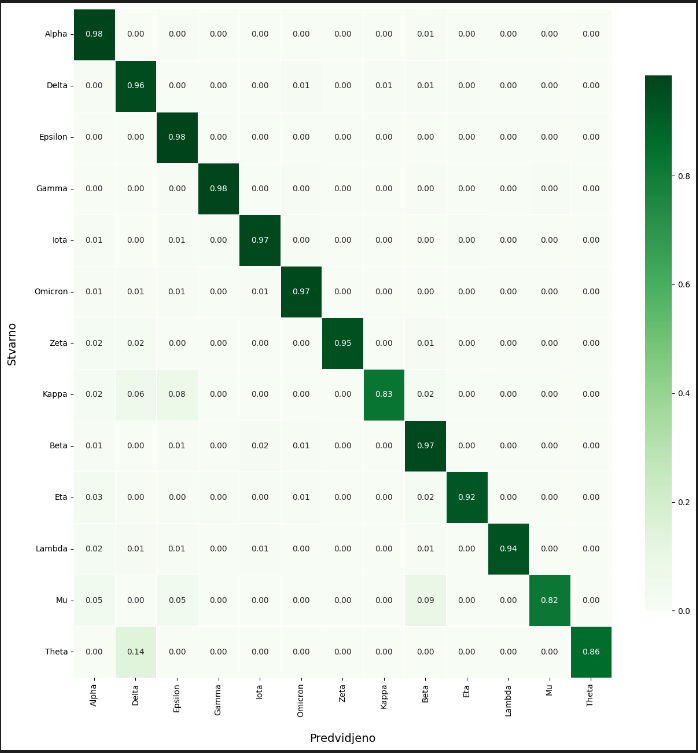


Tabela 4.2.3.1 Matrica konfuzije za klasifikaciju po soju SARS2 virusa

# 

## 4.3 K- najbližih suseda sa celim sekvencama i jednostavnom metrikom

Primenjen je algoritma K-najbližih suseda, gde je broj suseda 3 i korišćena jednostvana metrika (ako su kodoni razlličiti distanca se povećava za jedan). Ovaj model je primenjen za klasifikaciju po tipu virusa, po tipu proteina i po soju virusa za SARS2, korišćene su cele sekvence.

### 4.3.1 Ocena modela za klasifikaciju po tipu proteina

Primena ovog modela dala je tačnost od 0.98 i matricu konfuzije kao u tabeli 4.3.1.1

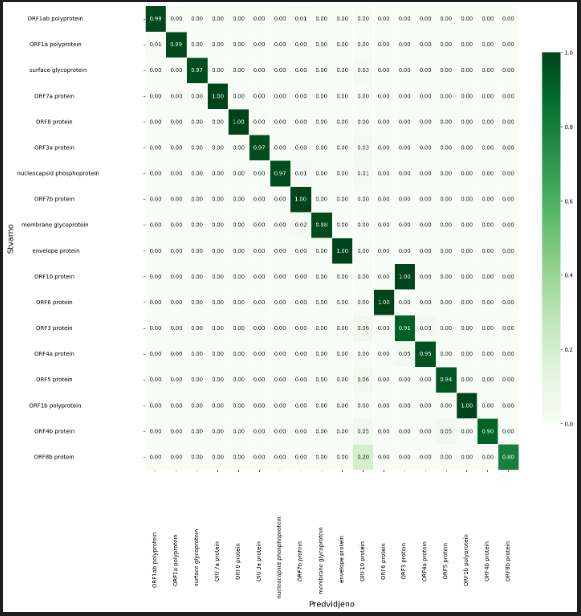


Tabela 4.3.1.1 Matrica konfuzije za klasifikaciju po tipu proteina

Možemo primetiti da i ovaj model radi dobro, najveća razlika je u tome što ORF10 protein u potpunosti prepoznaje kao ORF3 protein.

### 4.3.2 Ocena modela za klasifikaciju po soju virusa za SARS2

Primena ovog modela dala je tačnost od 0.81 i matricu konfuzije kao u tabeli 4.3.2.1

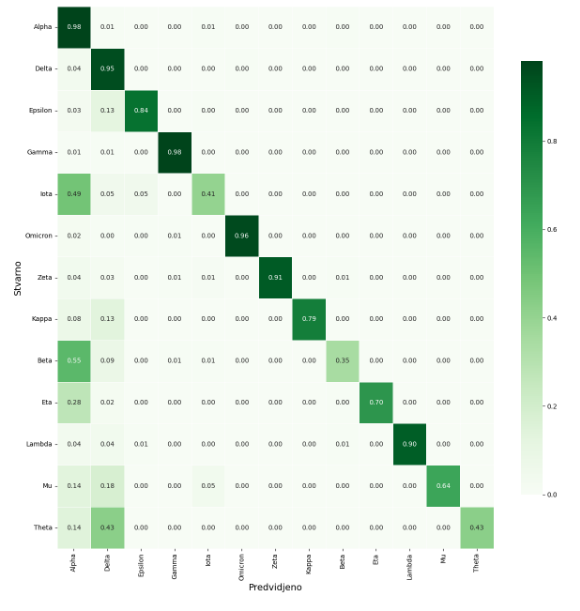


Tabela 4.3.2.1 Matrica konfuzije za klasifikaciju po soju virusa za SARS2

Možemo primetiti da ovaj model značajno lošije prepoznaje sojeve SARS2 virusa u odnosu na model [4.2.3](#_lnxbz9).

## 4.4 Unakrsna provera modela

Izvršena je unakrsna provera hiperparametra k modela za sve 3 klasifikacije. Provereni su modeli za k = 2,...,9. Originalni skup podataka se delio na 4 podskupa (skoro) jednake veličine i svaki od podskupova je korišćen kao validacioni skup. Dobijene su sledeće prosešne preciznosti:

* Za klasifikacija po tipu virusa:  
  + K=2 - 0.9975
  + **K=3 - 0.99775**
  + **K=4 - 0.99775**
  + K=5 - 0.997
  + K=6 - 0.99675
  + K=7 - 0.9965
  + K=8 - 0.9965
  + K=9 - 0.99625

Opaska: Svi klasifikatori na svakom od validacionih skupova su imali preciznost preko 99%.

* Za klasifikacija po tipu proteina:  
  + K=2 - 0.99525
  + **K=3 - 0.99875**
  + K=4 - 0.9985
  + K=5 - 0.9985
  + K=6 - 0.99825
  + K=7 - 0.99825
  + K=8 - 0.99825
  + K=9 - 0.99825

Opaska: Svi klasifikatori na svakom od validacionih skupova su imali preciznost preko 99%.

* Za klasifikaciju sars2 korona virusa po Svetskoj zdravstvenoj organizaciji:  
  + K=2 - 0.94825
  + K=3 - 0.951
  + **K=4 - 0.95125**
  + K=5 - 0.9505
  + K=6 - 0.9495
  + K=7 - 0.94875
  + K=8 - 0.9455
  + K=9 - 0.9445

Opaska: Svi klasifikatori na svakom od validacionih skupova su imali preciznost preko 89%.

Preciznost je približno ista za svako k iz skupa {2, 3, …, 9}, odakle sledi da odabir broja najbližih suseda (iz zadatog skupa) ne utiče bitno na preciznost klasifikacije.

# 

# 5. Zaključak

Uporedjivanjem modela [4.2.2](#_26in1rg) i modela [4.3.1](#_upa2obghezm0) može se primetiti da model 4.2.2 ima malo veću tačnost (0.99 naspram 0.98), ali i još značajnije, bolju matricu konfuzije. Model 4.3.1 ima problema sa prepoznavanjem proteina ORF10 i ORF8b. Kako su oba modela primenjena na celim sekvencama virusa i jedino se razlikuju u primenjenoj metrici (model 4.2.2 koristi p-adičnu metriku), možemo zaključiti da p-adična metrika ima uticaj na razlikovanje proteina SARS1, SARS2 i MERS koronavirusa.

Uporedjivanjem modela iz [4.2.3](#_lnxbz9) i [4.3.2](#_fo8oey425amn) može se primetiti da model 4.2.3 ima nešto bolje karakteristike od modela 4.3.2 (tačnost od 0.96 naspram 0.81 i značajno bolju matricu konfuzije. Kako je jedina razlika izmedju ovih modela primenjena metrika, možemo zaključiti da postoji uticaj P-adičnosti na razilke genetskog koda SARS1, SARS2 i MERS koronavirusa.

# 6.Reference

Dragovich,, Branko, and Nataša Ž Mišić. “P-Adic Hierarchical Properties of the Genetic Code.” *BioSystems,*, vol. 185, no. 104017, 2019, [1]