SVEUČILIŠTE U OSIJEKU

**FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA OSIJEK**

**Klasifikacija tumora dojke korištenjem modela KNN, SVC i logističke regresije**

*Mislav Ćaleta*

Osijek, siječanj, 2022.

**Sadržaj**

[**1 Uvod** 3](#_Toc97049163)

[**2 Postojeće metode** 4](#_Toc97049164)

[**3 Teoretske osnove korištenih modela** 5](#_Toc97049165)

[**3.1 Logistička regresija** 5](#_Toc97049166)

[**3.2 Stroj potpornih vektora** 6](#_Toc97049167)

[**3.3 K najbližih susjeda** 7](#_Toc97049168)

[**4 Predobradba podataka** 9](#_Toc97049169)

[**4.1 Opis izvornih podataka** 9](#_Toc97049170)

[**4.2 Predobradba** 9](#_Toc97049171)

[**5. Pregled koda korištenog za učenje i testiranje** 11](#_Toc97049172)

[**6. Analiza Rezultata** 14](#_Toc97049173)

[**Izvori** 17](#_Toc97049174)

# **1 Uvod**

U ovom projektnom zadatku pristupa se problemu klasifikacije tumora dojke, na bezopasne i zloćudne tumore, korištenjem tri modela za klasifikaciju. To su modeli : vektor potpornih strojeva, k najbližih susjeda i logistička regresija. Za svaki model provedena je analiza uspješnosti i rezultati modela su uspoređeni. Detaljna analiza i korisničko sučelje za pružanje novih podataka modelima, u svrhu vršenja predviđanja, implementirani su u obliku Streamlit aplikacije. Na početku ovog pisanog rada su navedene neke postojeće metode koje se koriste u ovom području. Zatim su sažeto objašnjene teoretske osnove svakog modela i prikazani su koraci koji su napravljeni prije samog rješavanja problema, što uključuje pregled podataka i predobradbu tih podataka korištenjem programskog jezika Python. Nakon predobradbe podataka, korištenjem istog programskog jezika i biblioteke scikit-learn, spomenuti modeli primjenjeni su na podatke, točnije, naučeni su na dobivenom skupu za učenje i ispitani na skupu za ispitivanje.

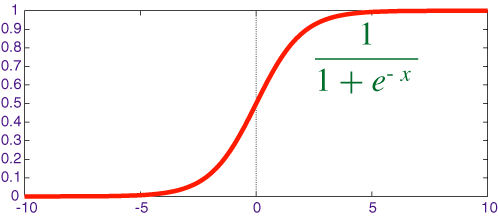
# **2 Postojeće metode**

U rješavanju problema klasifikacije tumora dojke koriste se metode koje se koriste i u ovom radu, ali uz njih se koriste i stablo odlučivanja i različite primjene neuronskih mreža. Ispitivanje se vrši na različitim modelima strojnog učenja kako bi ste postigla što bolja točnost, ali bitnije, što bolji odziv modela. Odziv modela je iznimno bitan za medicinske svrhe kao što je ova. Pacijentu se ne želi reći da tumor nije zloćudan, a da se na kraju dođe do zaključka da ipak je zloćudan.

# **3 Teoretske osnove korištenih modela**

## **3.1 Logistička regresija**

Logistička regresija spada u poopćene linearne modele. Svaki model koji ima nelinearnu funkciju „omotanu“ oko skalarnog produkta vektora težina i vektora značajki naziva se poopćeni linearni model. Nelinearna funkcija koja se koristi, naziva se aktivacijska funkcija modela. Kod logističke regresije ta funkcija je logistička funkcija ili sigmoida.



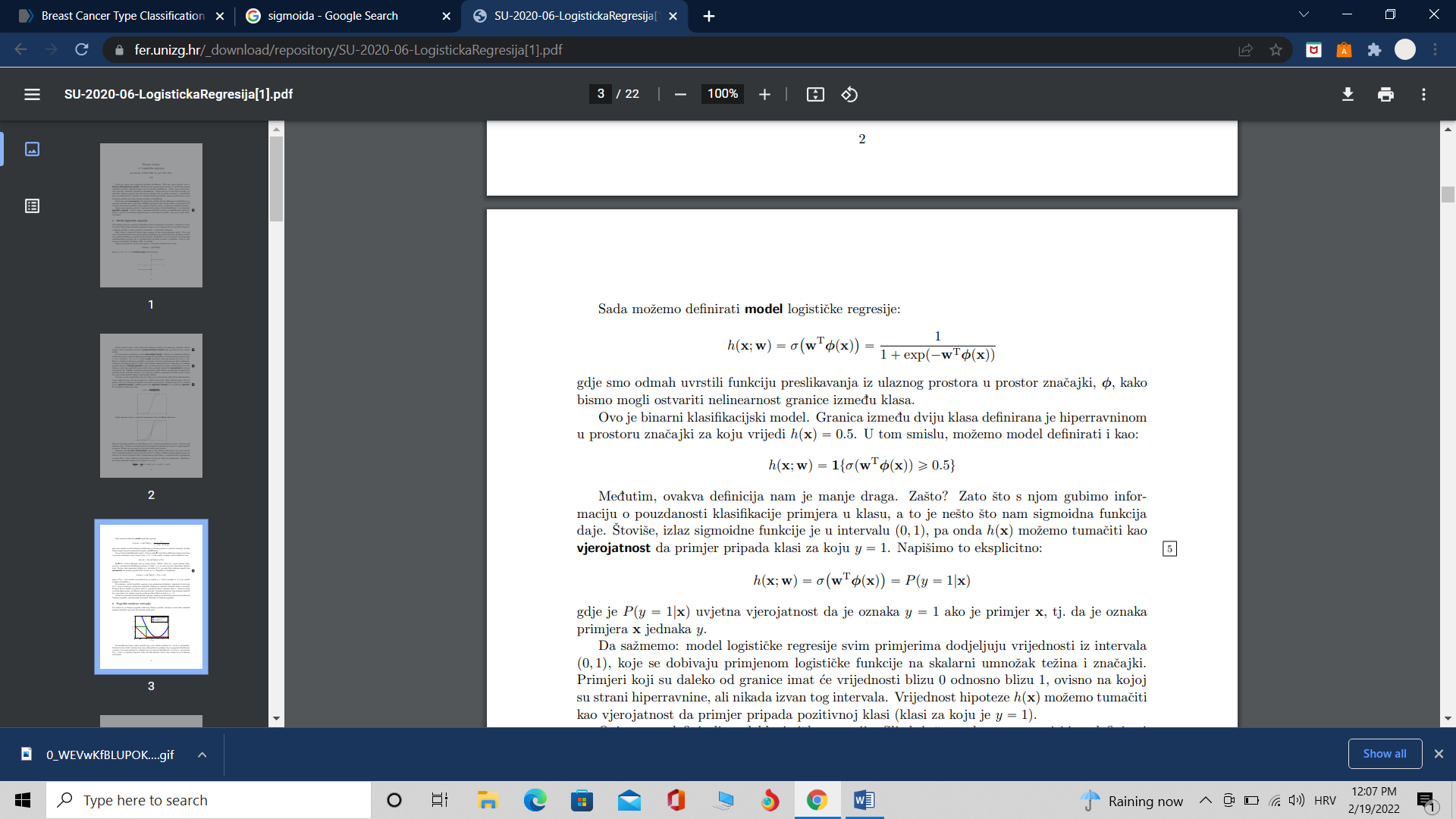
Slika 1: Logistička funkcija

Na slici 1 može se vidjeti graf logističke funkcije. Ako je faktor kojim se množi x veći od 1, sigmoida će postati strmija.

Tri karakteristike koje sigmoidu čine dobrim odabirom za aktivacijsku funkciju:

* „gnječi“ izlaz na interval od 0 do 1 (vjerojatnosna interpretacija)
* Oblikom je slična funkciji praga
* Funkcija je derivabilna na cijeloj domeni (postupci optimizacije)

Iz navedenog se dolazi do modela logističke regresije

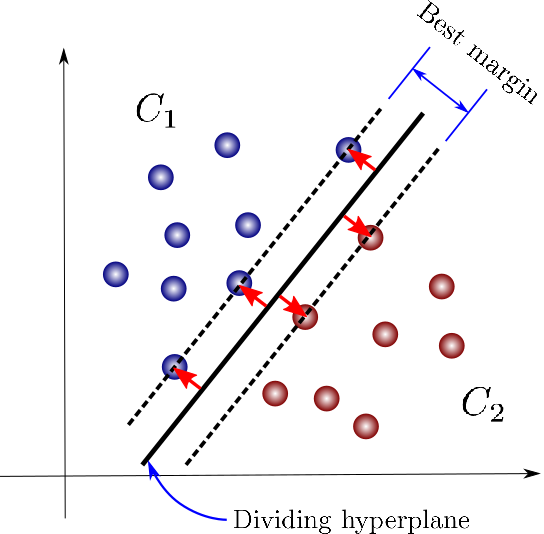


Slika 2: Model logističke regresije

Na slici 2 napisan je model logističke regresije, gdje σ predstavlja aktivacijsku funkciju omotanu oko linearnog modela regresije, odnosno, predstavlja sigmoidu. Vektor w je vektor parametara ili težina, a Φ(x) je funkcija preslikavanja koja kao argument ima primjer x. Klasifikacija se obavlja u prostoru gdje se nalazi granica, to jest, gdje se nalazi linearni model regresije. On predstavlja granicu i ovisno na kojoj strani granice se nalazi primjer spadat će u određenu klasu. Također, udaljenost od same granice, predstavlja sigurnost modela u odabir klase koju je odabrao za odgovarajući primjer. Iz izlaza sigmoide možemo isčitati vjerojatnost da primjer spada u „pozitivnu“ klasu. Sigmoida će preslikati izlaz linearnog modela za neki primjer u tu vjerojatnost. Zato korištenjem logističke regresije možemo doći i do vjerojatnosne interpretacije izlaza modela.

## **3.2 Stroj potpornih vektora**

Stroj potpornih vektora je linearni model i temelji se na maksimalnoj margini. Ako postoje dvije klase primjera koji su linearno odvojivi. Postoji beskonačan broj hipoteza koje možemo izabrati kako bi odvojili te klase. Te hipoteze pripadaju prostoru inačica. Cilj je izabrati neku od tih beskonačno mnogo hipoteza kako bi postigli najbolju moguću generalizaciju. Granicu se želi postaviti u sredinu, između dvije klase, tako da bude maksimalno udaljena od primjera jedne klase i od primjera druge klase. Izbjegava se pristranost prema jednoj klasi. Ta udaljenost od najbližeg primjera sa svake strane do granice naziva se margina.

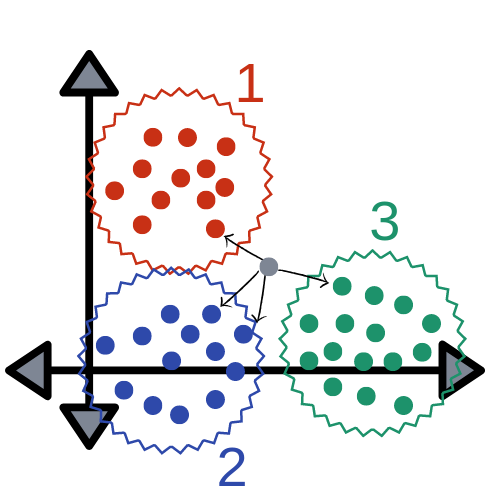


Slika 3: maksimalna margina

Primjeri koji se nalaze na margini nazivaju se potporni vektori. Postoji tvrda margina koja ne propušta primjere u marginu i meka margina koja propušta primjere unutar margine i s druge strane granice, zbog bolje optimizacije i određivanja parametara. To se naziva dualni model.

## **3.3 K najbližih susjeda**

K najbližih susjeda jednostavan je algoritam strojnog učenja. Među skup za učenje postavlja se primjer čija se klasa želi odrediti, zatim se gleda od k susjeda primjera kojoj klasi pripada više najbližih primjera. Kada jednoj klasi pripada više bližih primjera, onda se može smatrati da odgovarajući primjer isto pripada toj klasi.



Slika 4: K najbližih susjeda

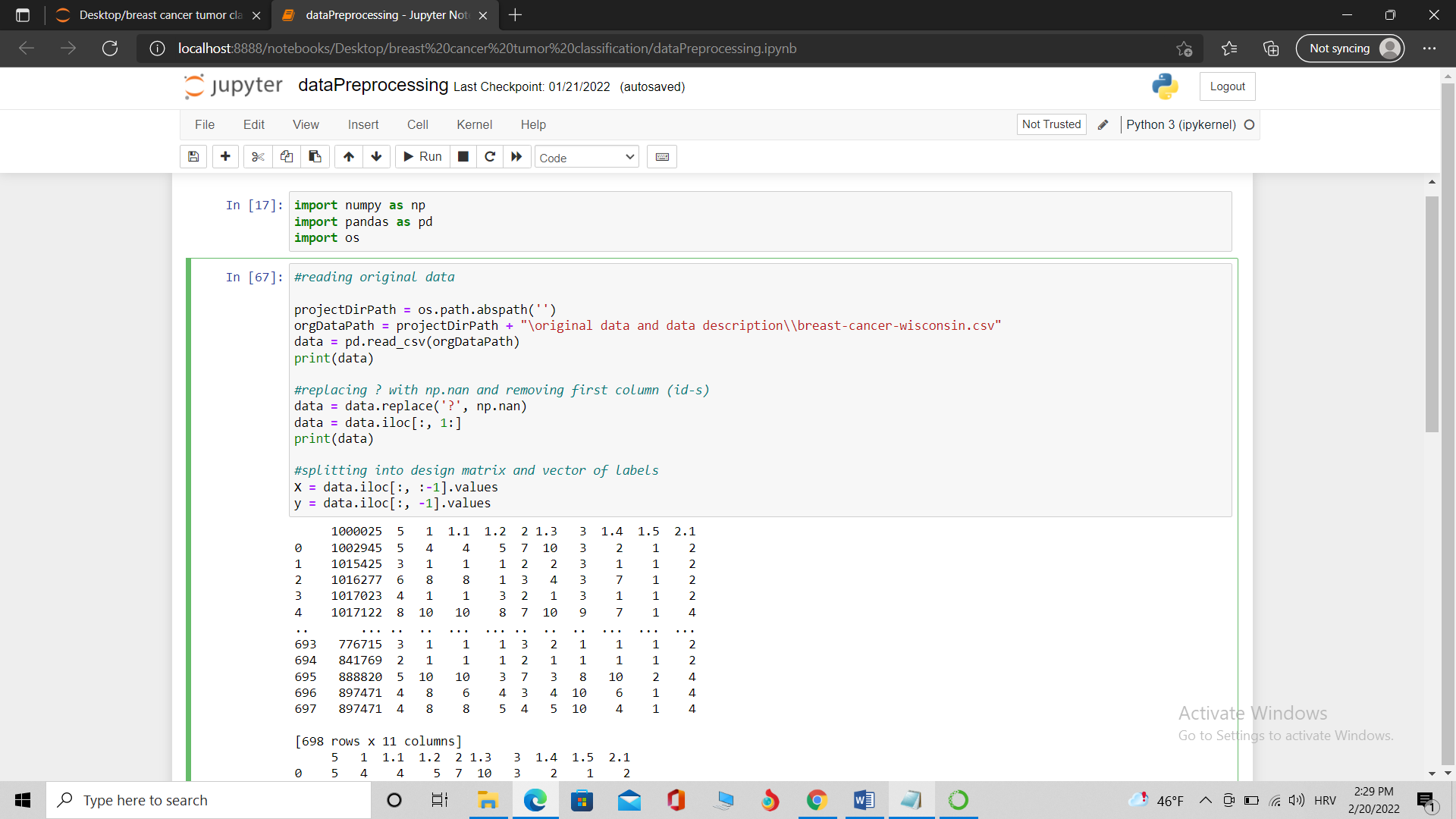
Na slici 4 mogu se vidjeti tri klase i jedan primjer koji treba klasificirati. U ovom slučaju primjer pripada klasi 2, zato što je najviše najbližih primjera u toj klasi. U problemu koji je opisan u ovom radu ne koriste se tri klase već dvije klase, ali metoda je potpuno ista.

# **4 Predobradba podataka**

## **4.1 Opis izvornih podataka**

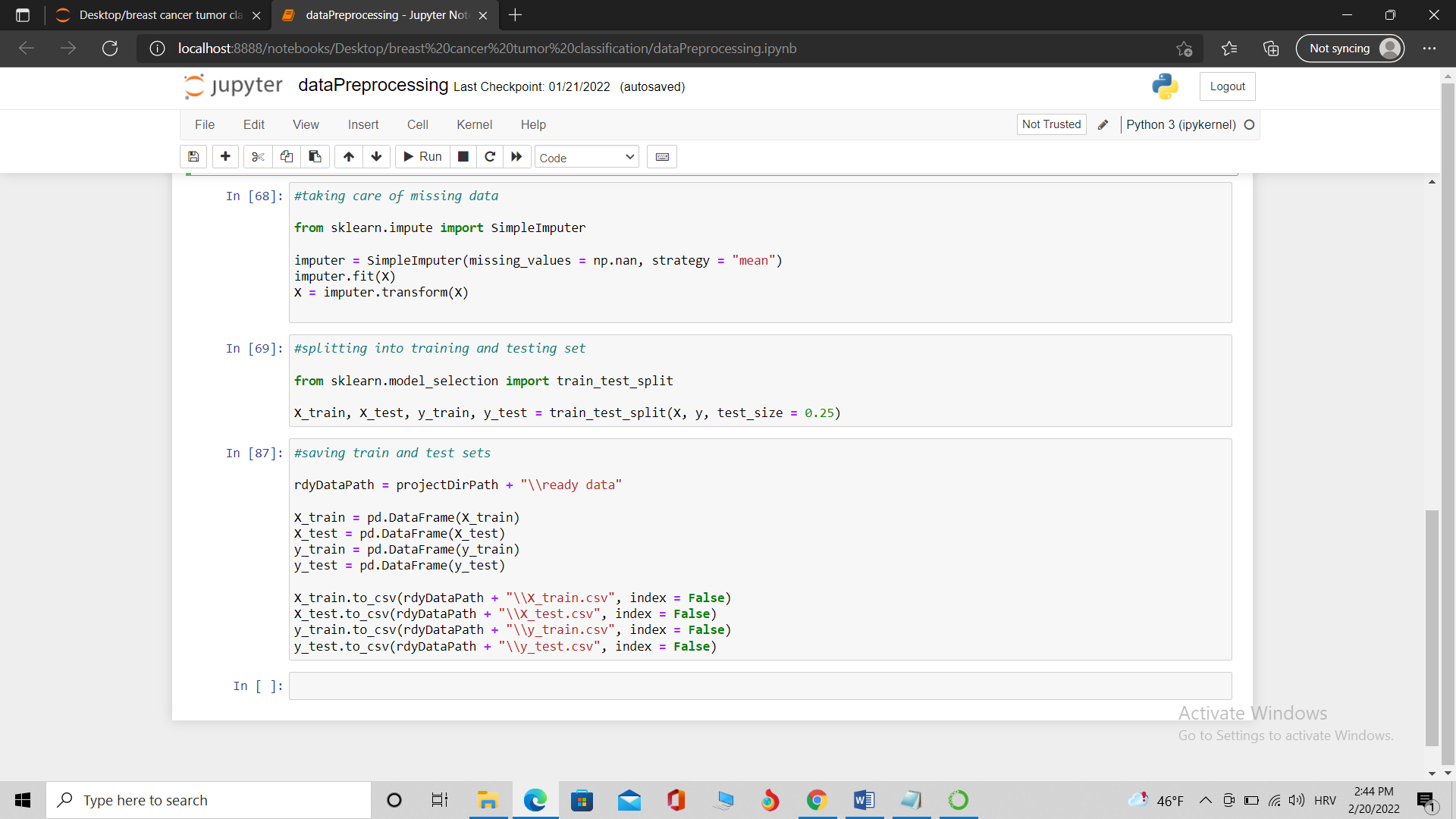
Podaci su preuzeti sa stranice UCI Machine Learning repository [1]. A prikupljeni su u Bolnicama Sveučilišta Wisconsin, na kojima je radio Dr. William H. Wolberg. Postoji 10 atributa od kojih jedan nije koristan za klasifikaciju. Postoje dvije klase, zloćudan i nije zloćudan. Te dvije klase označene su s 2 i 4, gdje klasa 4 predstavlja zloćudan tumor. U skupu podataka postoji 699 primjera mjerenja i također postoje vrijednosti unutar određenih primjera koje nedostaju. Te vrijednosti su tokom predobradbe popunjene.

## **4.2 Predobradba**



Slika 5: Predobradba 1. dio

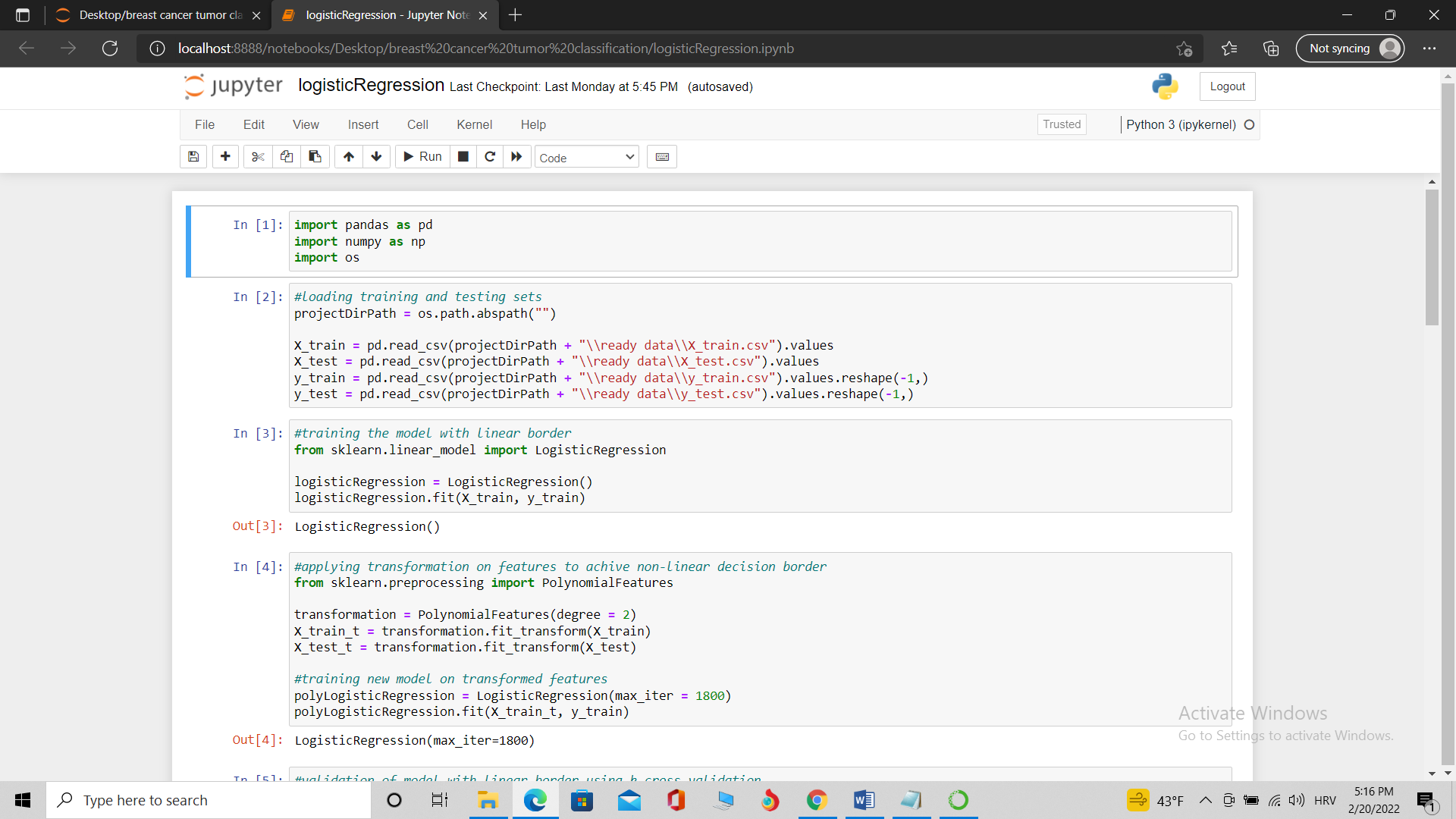
Na slici 5 može se vidjeti prvi dio programskog koda koji je napisan u svrhu obrade poadataka. Korištene su python biblioteke: numpy, pandas i os. Pročitana je csv datoteka korištenjem pandas biblioteke, zatim je zamjenjena oznaka za vrijednosti koje nedostaju ? s np.nan kako bi se kasnije lakše mogle popuniti te vrijednosti. Na kraju se skup podataka rastavlja na vektor oznaka y i matricu dizajna X.



Slika 6: Predobradba 2. dio

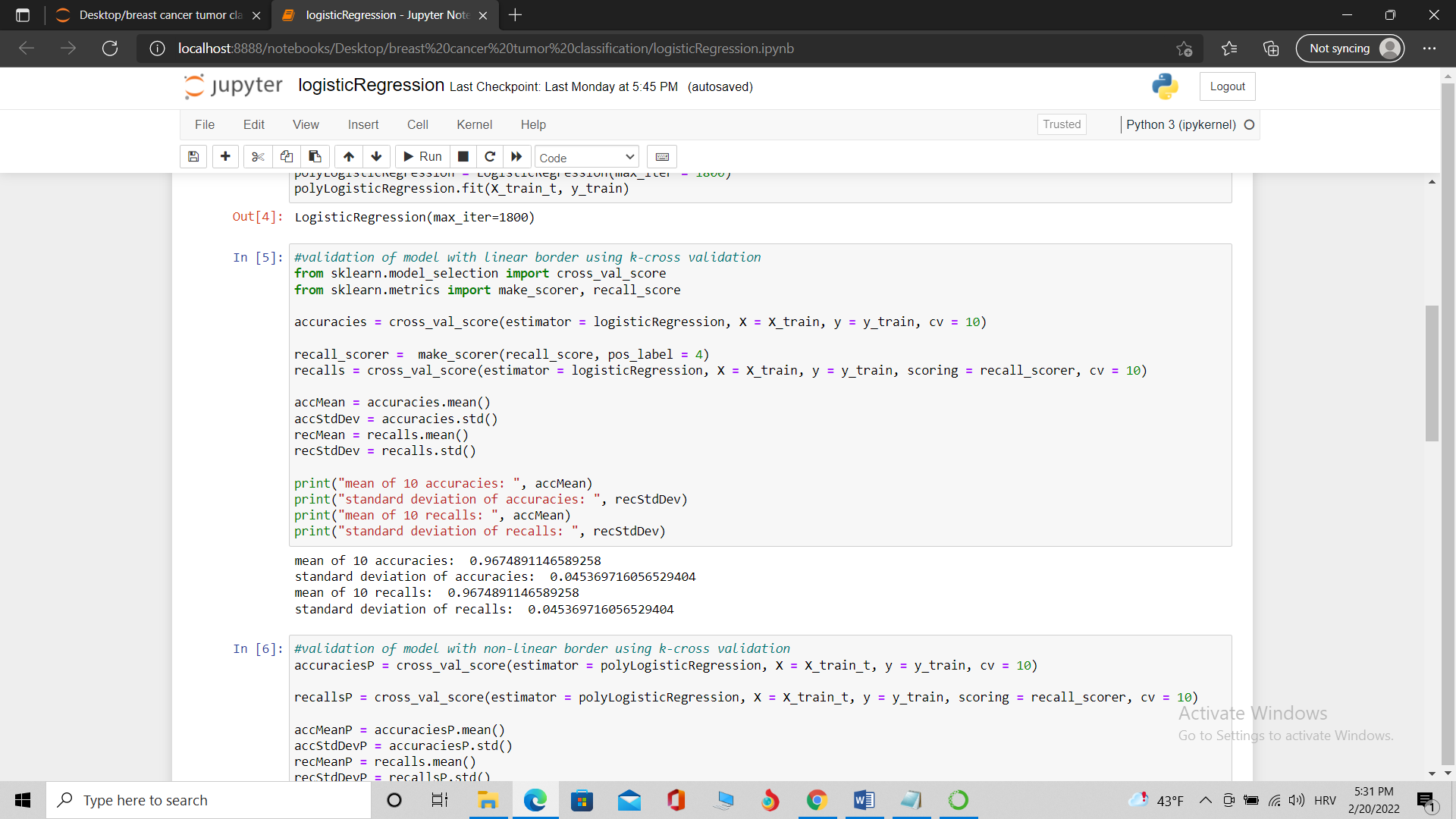
Na slici 6 prikazan je drugi dio predobradbe. Pomoću biblioteke scikit-learn, napravljen je objekt klase SimpleImputer koji je korišten za popunjavanje vrijednosti koje nedostaju. Zatim su matrica dizajna i vektor oznaka rastavljeni na skup za učenje i skup za testiranje. Na kraju su ti skupovi spremljeni u određene csv datoteke kako bi se kasnije mogli iskoristiti.

# **5. Pregled koda korištenog za učenje i testiranje**

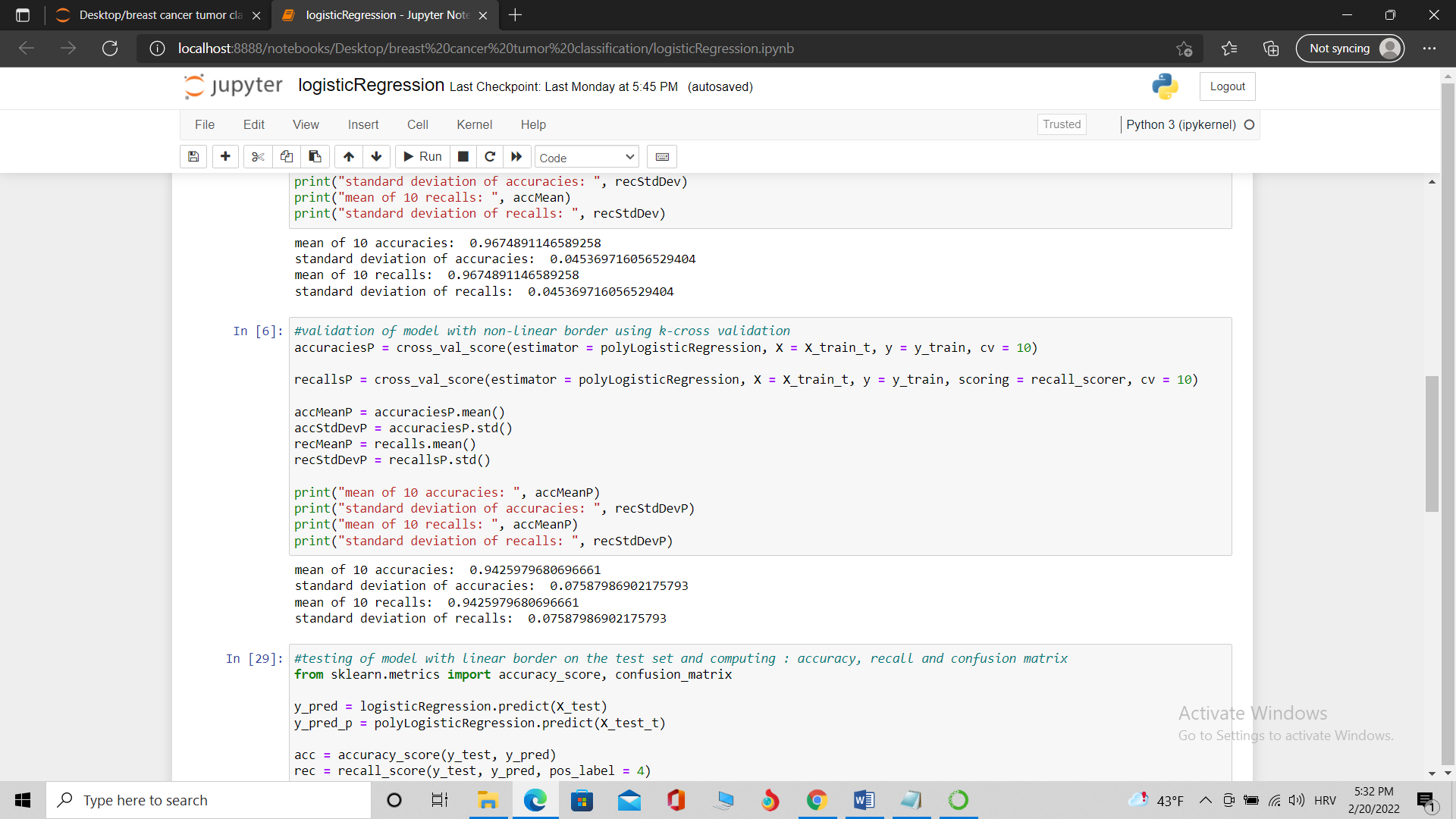


Slika 7: Logistička regresija 1. dio

Na slici 7. prikazan je dio učenja modela logističke regresije prije i nakon polinomskog proširenja (preslikavanja u prostor značajki). Skupovi za učenje i za testiranje koji su spremljeni u predobradbi su učitani i iskorišteni. Primjenjeno je kvadratno polinomsko proširenje za postizanje nelinearne granice u prostoru primjera.

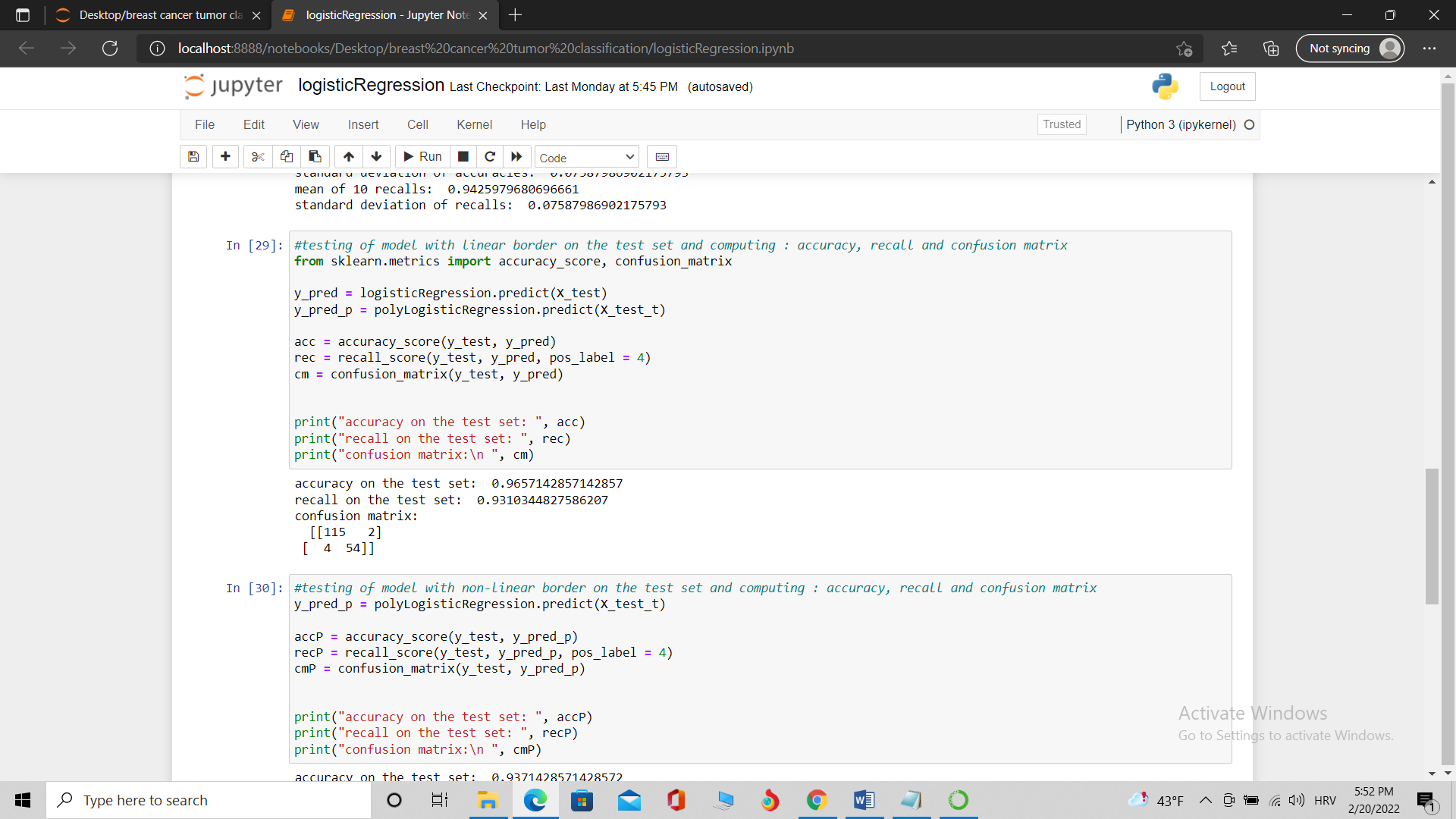


Slika 8: Logistička regresija 2. dio

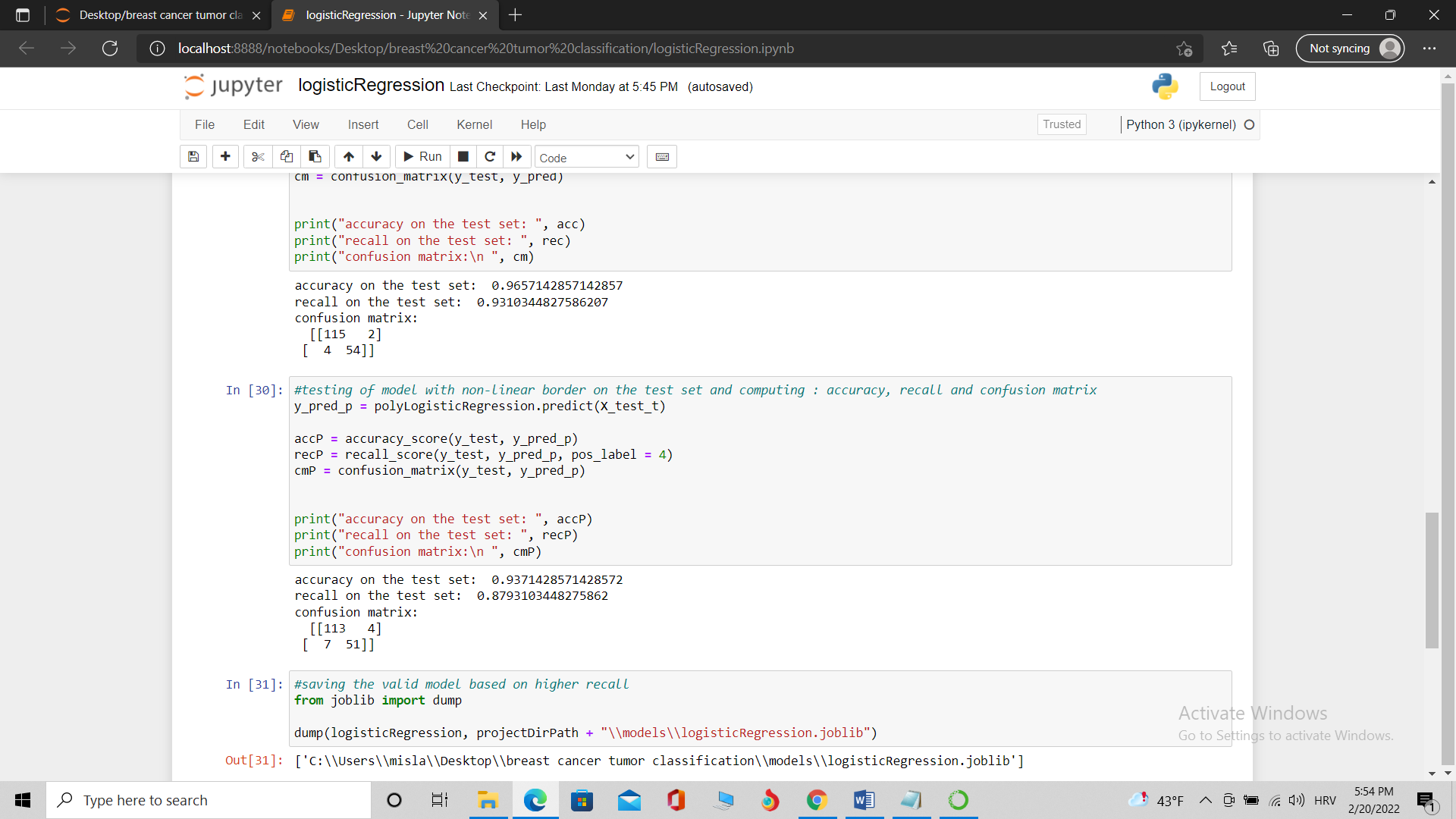


Slika 9: Logistička regresija 3. dio

Na slici 8 i 9 pokazana je primjena k unakrsne validacije, gdje se skup za učenje razdvaja na određeni broj preklopa i svaku iteraciju se jedan preklop ostavlja kao skup za testiranje dok se ostali ostavljaju kao skup za učenje. Na taj način se dobije dodatna sigurnost da su mjere modela točne. Mjere koje su korištene su srednja vrijednost točnosti i odziva i standardne devijacije tih mjera, budući da je napravljeno deset iteracija tokom validacije.

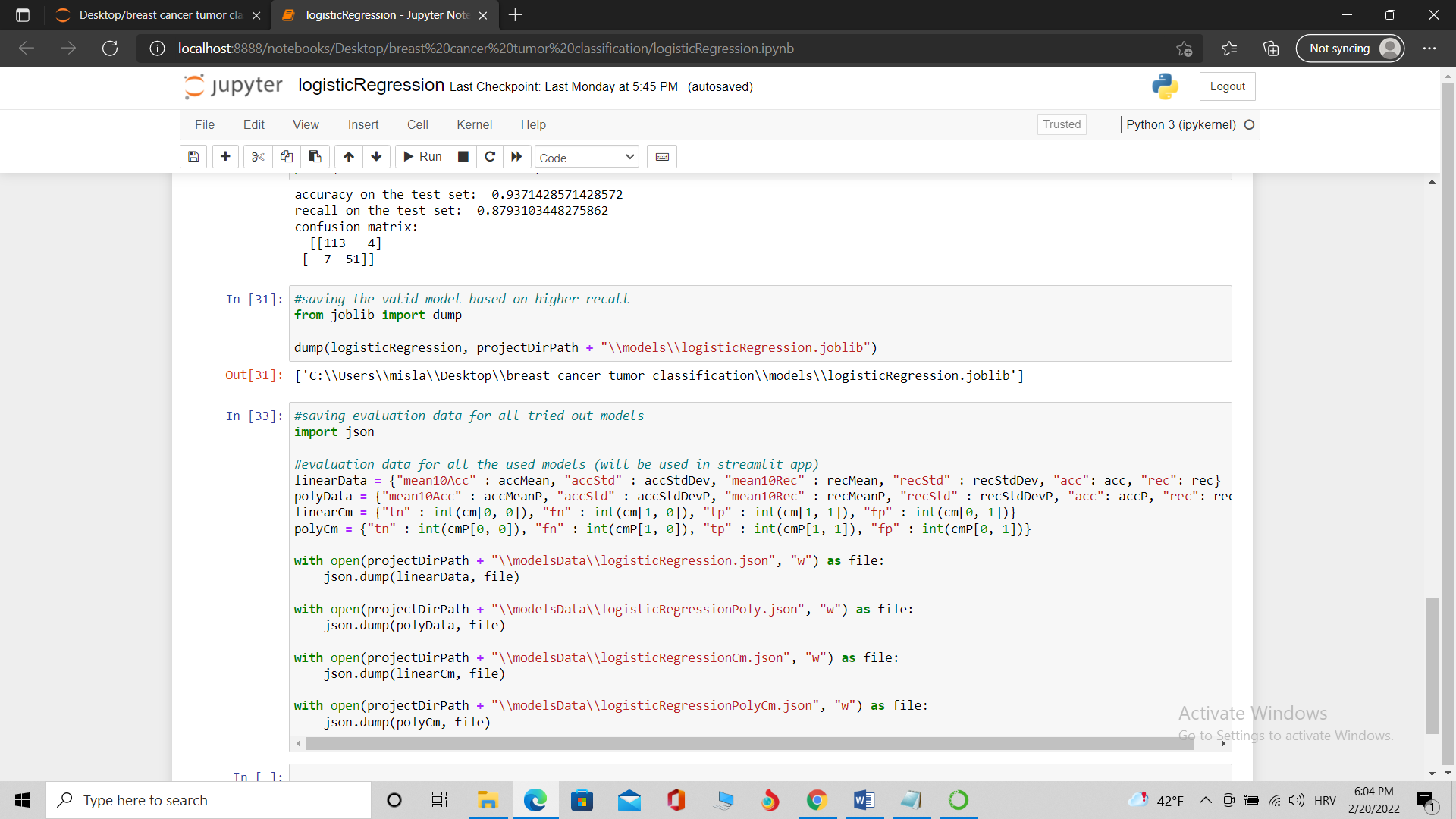


Slika 10: Logistička regresija 4. dio



Slika 11: Logistička regresija 5. dio

Na slici 10 i 11 prikazano je testiranje oba ranije spomenuta modela na skupu za testiranje i dobivene mjere zajedno s matricom zabune.



Slika 12: Logistička regresija 6. dio

U zadnjem dijelu na slici 12. prikazano je spremanje modela i mjera modela za korištenje unutar streamlit aplikacije. Za stroj potpornih vektora i za k najbližih susjeda, koraci su slični uz neke specifičnosti kao postavljanje broja susjeda, ali programski kod svodi se na iste korake. Modeli koji su korišteni u izradi aplikacije birani su po najboljem odzivu. Detaljna analiza modela može se pogledati u Streamlit aplikaciji na linku

<https://share.streamlit.io/mislavcaleta/breast-cancer-tumor-classification/main/streamlitApp.py>

# **6. Analiza Rezultata**

Logistička regresija – validacija

|  |  |  |
| --- | --- | --- |
|  | Linearna granica | Nelinearna granica |
| Srednja vrijednost 10 točnosti tokom k-unakrsne validacije | 0.9675 | 0.9426 |
| Srednja vrijednost 10 odziva tokom k-unakrsne validacije | 0.9509 | 0.9509 |
| Standardna devijacija točnosti | 0.021 | 0.0297 |
| Standardna devijacija odziva | 0.0454 | 0.0759 |

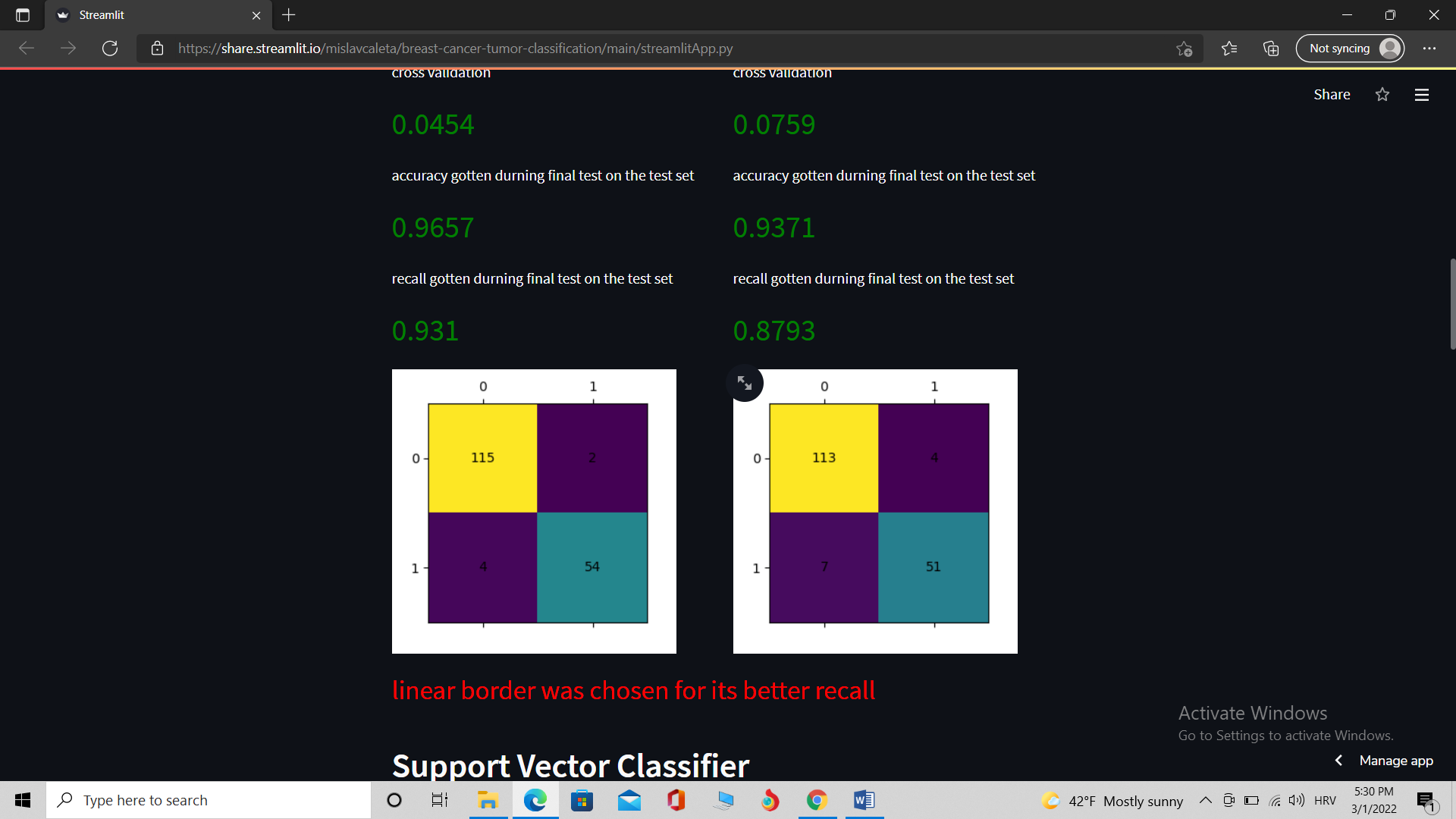
Odzivi su isti, ali točnost modela s linearnom granicom je veća, također je standardna devijacija odziva manja, što je bolje. Zbog navedenih razloga izabran je model s linearnom granicom

Logistička regresija – testiranje odabranog modela na testnim podacima

Točnost – 0.9657

Odziv – 0.931

Matrica zabune:

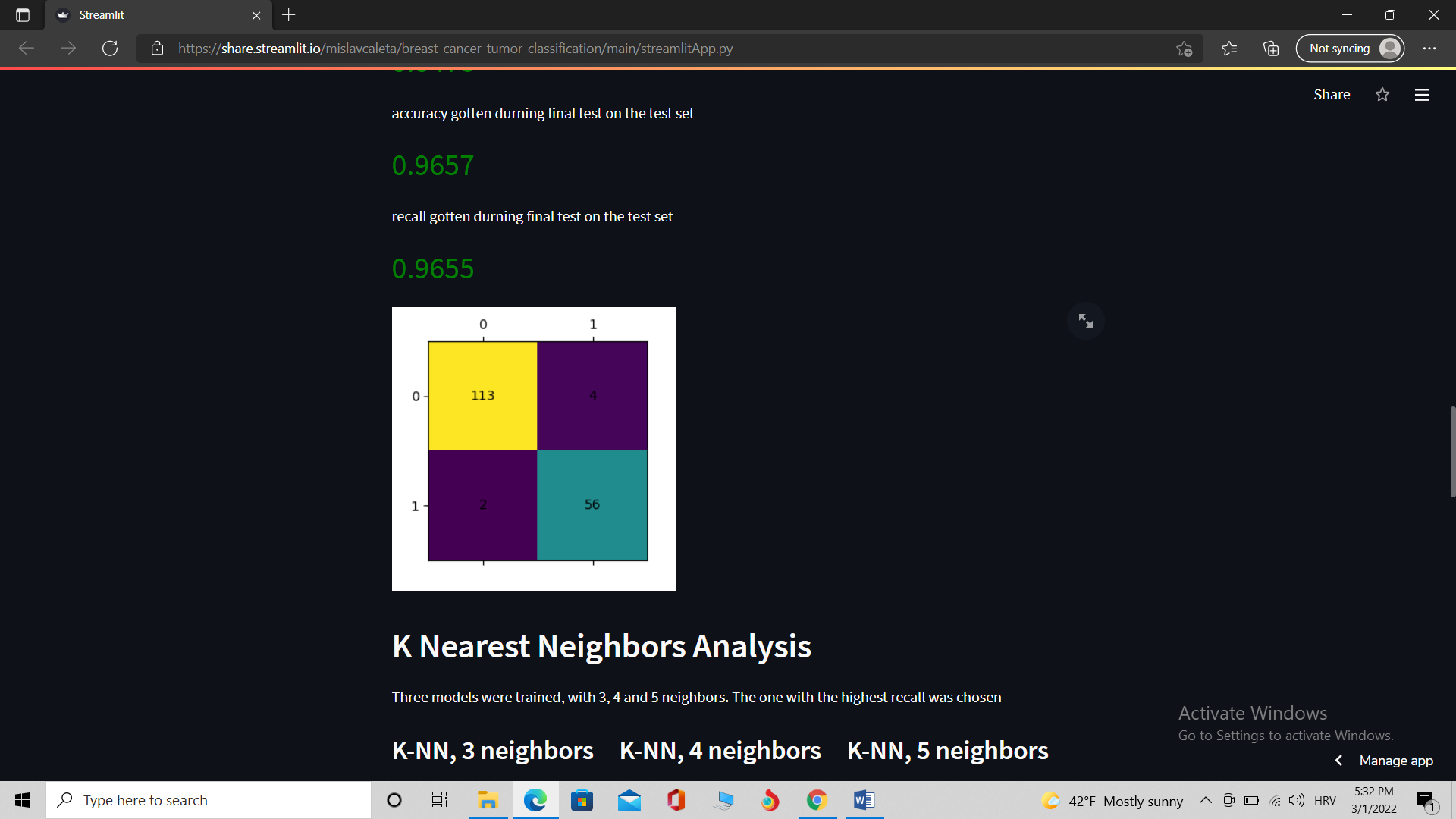


Stroj potpornih vektora – testiranje modela na testnim podacima (odmah je izabran model bez validacije)

Točnost – 0.9657

Odziv – 0.9655

Matrica zabune:



K najbližih susjeda – validacija

|  |  |  |  |
| --- | --- | --- | --- |
|  | 3 susjeda | 4 susjeda | 5 susjeda |
| Srednja vrijednost 10 točnosti tokom k-unakrsne validacije | 0.9751 | 0.9675 | 0.9732 |
| Srednja vrijednost 10 odziva tokom k-unakrsne validacije | 0.9725 | 0.945 | 0.9725 |
| Standardna devijacija točnosti | 0.0228 | 0.0258 | 0.0176 |
| Standardna devijacija odziva | 0.0511 | 0.0653 | 0.0371 |

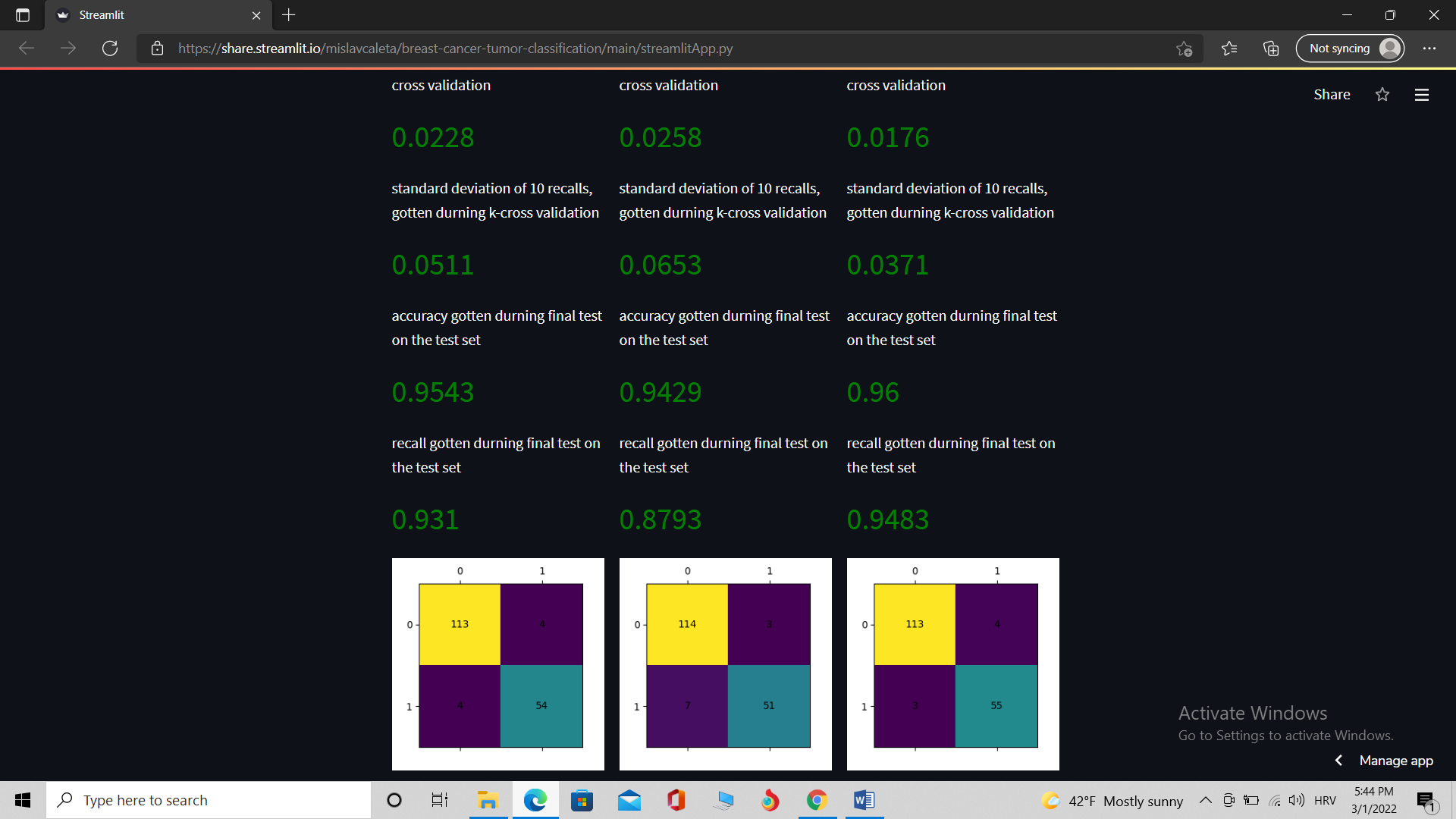
Odziv je jednak kod modela s 3 susjeda i s 5 susjeda, ali standardna devijacija manja je kod modela s 5 susjeda, tako da je odabran model s 5 susjeda.

K najbližih susjeda – testiranje odabranog modela na testnim podacima

Točnost – 0.96

Odziv – 0.9483

Matrica zabune:



# **Izvori**

[1]

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29