Tehnička Dokumentacija Projekta: Klasifikacija Tumora Dojke

Sadržaj

1. Uvod

Ovaj dokument pruža sveobuhvatan tehnički pregled projekta klasifikacije tumora dojke. Cilj projekta je razvoj prediktivnog modela mašinskog učenja za određivanje malignosti (benigni/maligni) tumora, koristeći numeričke karakteristike ćelijskih jezgara dobijenih iz digitalizovanih slika biopsije. Dokumentacija je strukturirana tako da jasno opisuje svaki korak u procesu obuke modela, naglašavajući metodologiju i pristup, a ne samo krajnje performanse, u duhu principa tehničke dokumentacije usmerene na proizvod.

2. Cilj Projekta

Primarni cilj ovog projekta je uspostavljanje robustnog procesa za razvoj i implementaciju modela mašinskog učenja za preciznu klasifikaciju tumora dojke. Sekundarni ciljevi obuhvataju:

- **Standardizacija Procesnih Koraka:** Jasno definisanje faza od prikupljanja podataka do evaluacije modela.
- **Analiza Karakteristika:** Identifikacija i razumevanje uticaja pojedinačnih karakteristika na klasifikaciju.
- **Vizualizacija Podataka:** Kreiranje vizuelnih prikaza za dublje razumevanje strukture podataka i odnosa između atributa i ishoda.

3. Arhitektura Podataka i Karakteristike

Podaci se sastoje od 30 numeričkih karakteristika izvedenih iz digitalizovanih slika biopsije, koje opisuju ćelijska jezgra. Svaki uzorak tumora obuhvata tri grupe merenja za svaku od deset osnovnih karakteristika:

Srednje Vrednosti (Mean Values):

- **Radius (radijus):** Srednja udaljenost od centra do tačaka na obodu.
- **Texture (tekstura):** Standardna devijacija vrednosti sive skale.
- **Perimeter (obim):** Ukupna dužina konture jezgra.
- **Area (površina):** Površina unutar konture jezgra.
- **Smoothness (glatkoća):** Lokalna varijacija dužina radijusa.
- **Compactness (kompaktnost):** Mera kompaktnosti oblika, izračunata kao (obim^2 / površina) - 1.0.
- **Concavity (konkavnost):** Ozbiljnost konkavnih delova konture.
- **Concave Points (konkavne tačke):** Broj konkavnih delova konture.
- **Symmetry (simetrija):** Mera simetrije oblika jezgra.
- **Fractal Dimension (fraktalna dimenzija):** Aproksimacija "obalne linije", koja opisuje složenost konture.

Standardne Greške (Standard Error - SE):

Standardna greška svake od gore navedenih karakteristika, odražavajući varijabilnost merenja.

Najgore (Worst) Vrednosti:

"Najgore" ili najveće (ili srednje za najveća tri) vrednosti svake od gore navedenih karakteristika, često ukazuju na ekstremne promene u ćelijama.

4. Metodologija Klasifikacije: Detaljan Postupak Obuke Modela

Proces obuke modela klasifikacije tumora dojke prati standardni tok rada u mašinskom učenju, sa naglaskom na transparentnost i ponovljivost koraka.

4.1. Faze Projekta:

1. **Razumevanje Podataka i Problem:**
    - Detaljna analiza skupa podataka, uključujući tipove karakteristika i njihovu distribuciju.
    - Definisanje problema klasifikacije (binarna klasifikacija: benigni vs. maligni).
2. **Prikupljanje i Predobrada Podataka:**
    - **Učitavanje Podataka:** Učitavanje skupa podataka iz izvora (u našem slučaju CSV fajl).
    - **Ispitivanje Podataka (EDA - Exploratory Data Analysis):**
        - Provera nedostajućih vrednosti.
        - Analiza distribucije svake karakteristike (histogrami, box plotovi).
        - Identifikacija korelacija između karakteristika.
        - Vizualizacija odnosa između karakteristika i ciljne varijable (tip tumora).
    - **Čišćenje Podataka:** Rešavanje eventualnih anomalija ili grešaka u podacima.
    - **Skaliranje Podataka:** Normalizacija ili standardizacija numeričkih karakteristika kako bi se sprečila dominacija karakteristika sa većim opsegom vrednosti (npr. StandardScaler).
3. **Podela Skupa Podataka:**
    - Podela celokupnog skupa podataka na trening i test setove (npr. 70% za trening, 30% za testiranje). Ovo osigurava da se model evaluira na podacima koje ranije nije video, pružajući realniju procenu njegovih performansi.
4. **Izbor i Obuka Modela:**
    - **Izbor Algoritama:** Odabir jednog ili više algoritama mašinskog učenja pogodnih za binarnu klasifikaciju. Primeri algoritama koji se mogu koristiti uključuju:
        - **Logistička Regresija (Logistic Regression):** Jednostavan, ali efikasan linearni model.
        - **Podržavajuće Vektorske Mašine (Support Vector Machines - SVM):** Moćan algoritam za klasifikaciju, posebno efikasan u prostorima visoke dimenzionalnosti.
        - **Drvo Odluke (Decision Tree):** Intuitivan model koji donosi odluke na osnovu serije pravila.
        - **Nasumična Šuma (Random Forest):** Ansambl metod koji kombinuje više stabala odluke za poboljšanje tačnosti i smanjenje prekomernog prilagođavanja.

- **Obuka Modela:** Svaki odabrani model se obučava nezavisno na trening setu podataka. Proces obuke uključuje prilagođavanje internih parametara modela kako bi se minimizovala greška predviđanja.

5. **Evaluacija Modela:**
   - Procena performansi obučenih modela na test setu. Metrike evaluacije su ključne za razumevanje koliko dobro model generalizuje na neviđene podatke.

5. Očekivani Izlazi i Rezultati

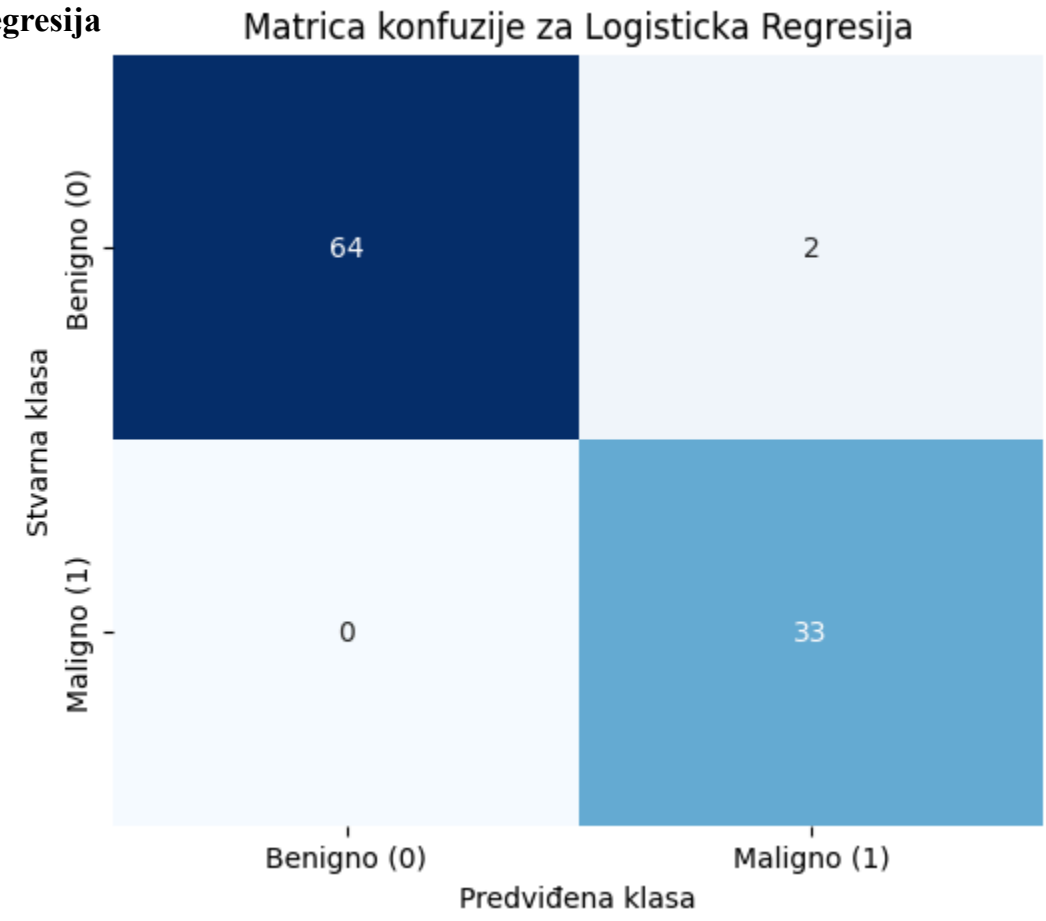
Kao rezultat ovog projekta, očekuju se sledeći artefakti i analize:

5.1. Metrike Evaluacije Modela (za svaki model)

- **Accuracy (tačnost):** Ukupni procenat tačno klasifikovanih uzoraka.
- **Precision (preciznost):** Odnos tačno pozitivnih predviđanja prema ukupnom broju predviđenih pozitivnih (relevantnost pozitivnih predviđanja).
- **Recall (odziv):** Odnos tačno pozitivnih predviđanja prema ukupnom broju stvarnih pozitivnih slučajeva (sposobnost modela da pronađe sve pozitivne slučajeve).
- **F1-Score (F1-mera):** Harmonijska sredina preciznosti i odziva, korisna kada je distribucija klasa neuravnotežena.
- **Matrica Konfuzije:** Tabela koja sumira performanse klasifikacionog modela, prikazujući broj tačnih i netačnih predviđanja za svaku klasu.

5.2. Vizuelna Analiza Važnosti Karakteristika (za svaki model, ako je primenjivo)

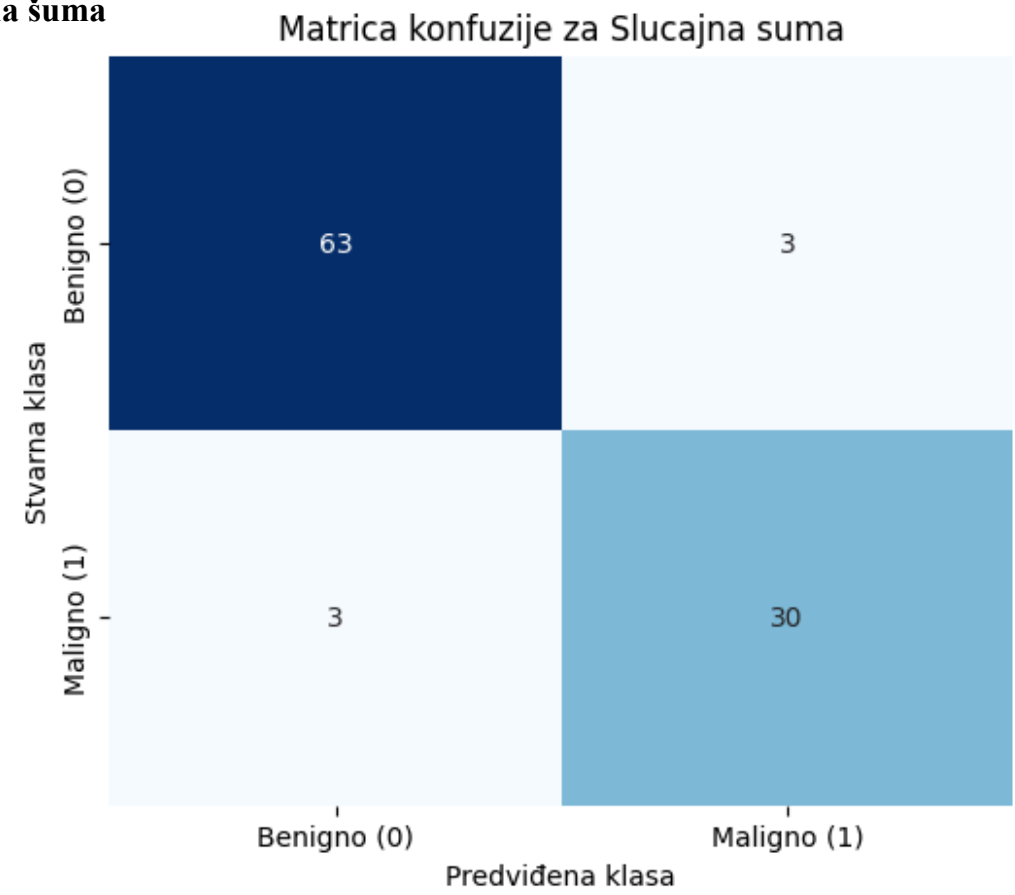
Logistička Regresija



Rezultati:

	preciznost	osetljivost	f1-score	podrška
0	1	0.97	0.98	66
1	0.94	1	0.97	33
tačnost			0.98	99

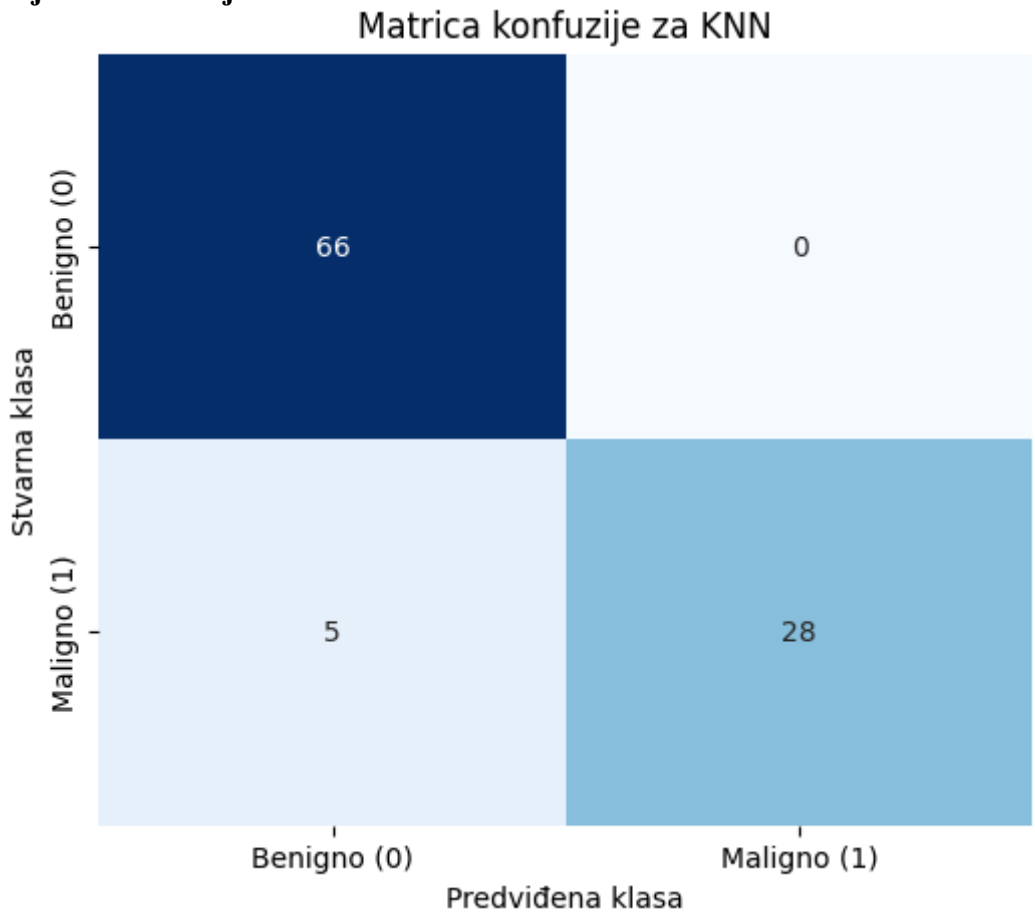
Slučajna šuma



Rezultati:

	preciznost	osetljivost	f1-score	podrška
0	0.95	0.95	0.95	66
1	0.91	0.91	0.91	33
tačnost			0.94	99

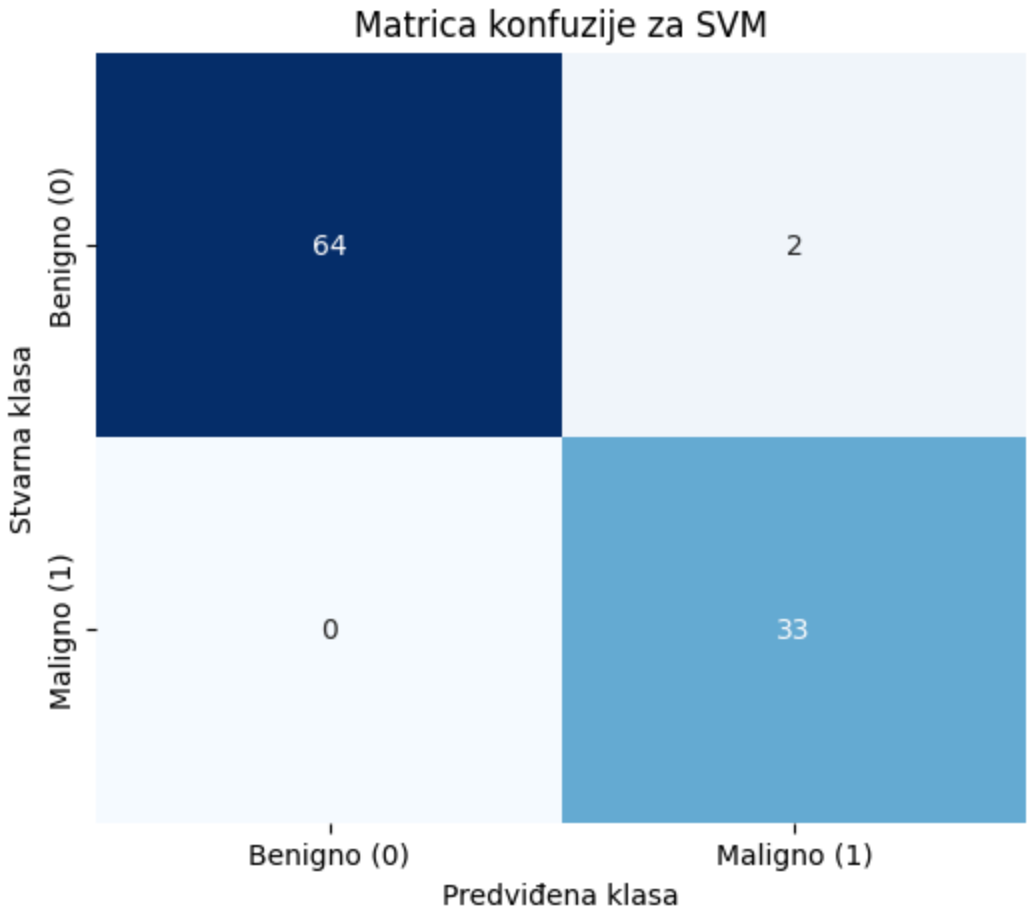
Model K-Najbližih komšija



Rezultati:

	preciznost	osetljivost	f1-score	podrška
0	0.93	1	0.96	66
1	1	0.85	0.92	33
tačnost			0.95	99

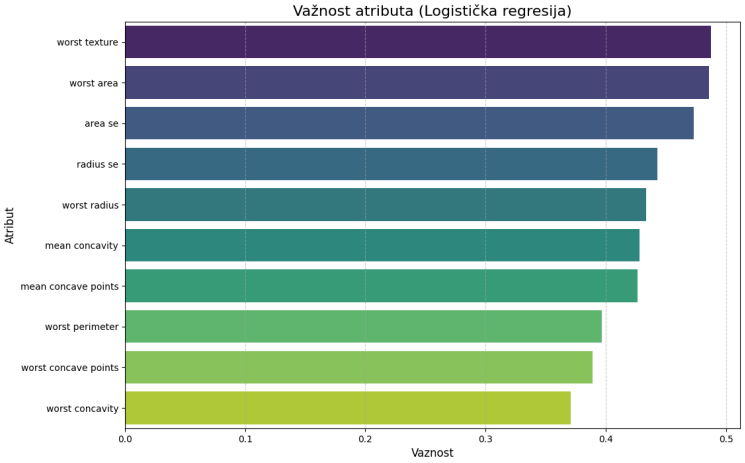
Model potpornih vektora(SVM):

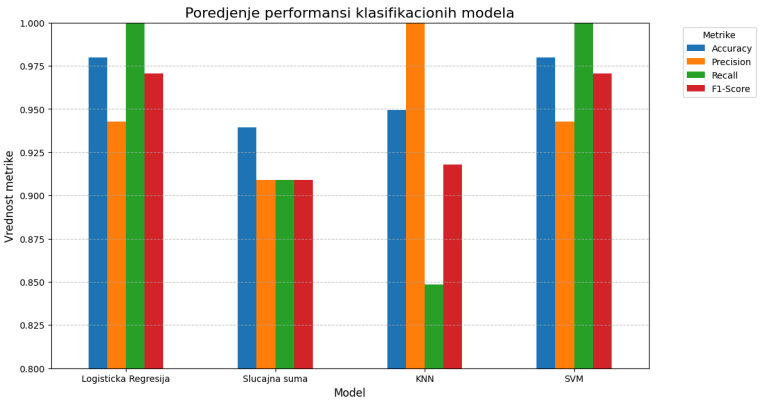


Rezultat

	preciznost	osetljivost	f1-score	podrška
0	1	0.97	0.98	66
1	0.94	1	0.97	33
tačnost			0.98	99

5.3. Interpretacija Rezultata i Preporuke (za svaki model):





Na osnovu rezultata evaluacije modela, uključujući metrike preciznosti, osetljivosti, F1-mere i tačnosti za Logističku Regresiju, Slučajnu Šumu, Model K-Najbližih Komšija i Model Potpornih Vektora (SVM), u nastavku će biti predstavljen uvod u analizu ovih performansi.

Logistička regresija

Veoma snažan model koji se ističe kao **najbolji "detektiv"**. Postigla je savršen rezultat po najvažnijem kriterijumu, čineći je jednim od dva najbolja modela.

- **(+) Glavna snaga:** Apsolutno savršena osetljivost (**recall**). Model je uspeo da **pronađe svaki postojeći maligni tumor** (0 propuštenih slučajeva), što ga čini savršenim "detektivom".
- **(-) Glavna slabost:** Jedina mana su **dve "lažne uzbune"**, gde su benigni uzorci proglašeni malignim. Ovo je prihvatljiv kompromis s obzirom na ispunjenje glavnog cilja.
- **Presuda: Jedan od dva najbolja modela i jasan izbor za implementaciju.** U potpunosti zadovoljava zahtev da se rizik od propuštene dijagnoze svede na nulu.

Slučajna šuma

Izuzetno pouzdan i robustan model, čija je glavna odlika **opreznost**.

- **(+) Glavna snaga:** Model je pokazao relativan balans, ne favorizujući drastično jednu vrstu greške u odnosu na drugu (3 lažne uzbune i 3 propuštena slučaja).
- **(-) Glavna slabost: Tri propuštena maligna slučaja** su ključna mana koja ga diskvalifikuje, jer su druga dva modela pokazala da je moguće imati nula takvih grešaka.
- **Presuda: Nije preporučljiv za ovaj zadatak.** Iako generalno dobar, postoje očigledno superiornije alternative.

K-Najbližih komšija

Objektivno najslabiji model u ovoj grupi.

- **(+) Glavna snaga:** Neočekivana snaga u ovom testu je bila **nula "lažnih uzbuna"**, što znači da je bio veoma pouzdan kada je predviđao benigne slučajeve.
- **(-) Glavna slabost: Pet propuštenih malignih slučajeva** je najgori rezultat od svih modela i predstavlja nedopustivo visok rizik.
- **Presuda: Najlošiji izbor.** Visok broj najkritičnijih grešaka ga čini nepouzdanim i neprikladnim za implementaciju

Model potpornih vektora(SVM)

SVM je, identično kao i Logistička Regresija, pružio besprekorne performanse po pitanju detekcije, svrstavajući se tako na sam vrh.

- **(+) Glavna snaga:** Poput Logističke Regresije, njegova najveća snaga je **savršena detekcija svih malignih slučajeva** (0 propuštenih slučajeva).
- **(-) Glavna slabost:** Takođe identično, jedina slabost su **dve "lažne uzbune"**, što je u skladu sa definisanim prioritetima projekta.
- **Presuda: Deli prvo mesto sa Logističkom Regresijom i predstavlja optimalan izbor.** Njegova sposobnost da ne propusti nijedan maligan slučaj ga čini idealnim rešenjem.

Na osnovu detaljne analize i definisanog prioriteta da je **propuštanje maligne dijagnoze najkritičnija greška**, zaključuje se sledeće:

Najbolji modeli za ovaj zadatak su Logistička Regresija i Model potpornih vektora.

Iako ovi modeli generišu oderđen broj lažnih uzbuna(po dva slučaja), oni su superioirniji jer su bili **savršeni** u primarnom cilju - pronalaženje malignih slučajeva. **Sa nula propuštenih slučajeva**, oni u potpunosti eliminišu rizik od najopasnije moguće greške i time predstavljaju optimalno rešenje za ovaj problem.

6. Zaključak

Ovaj projekat pruža solidnu osnovu za klasifikaciju tumora dojke, sa fokusom na jasan i ponovljiv metodološki pristup. Dalji razvoj može uključivati istraživanje naprednijih algoritama, primenu tehnika za smanjenje dimenzionalnosti, korišćenje većih i raznovrsnijih skupova podataka, te potencijalnu integraciju sa kliničkim sistemima za podršku dijagnostici. Kontinuirana evaluacija i iterativno poboljšanje su ključni za optimizaciju modela i maksimizovanje njegove korisnosti.