Version 1.0
24 Octombrie, 2024

Document de specificații software

[SRS]

Proiect software ce vizează analiza experimentală a unui set de date medicale

Realizat de ↓

- Mihălucă Mădălina-Maria
- Popa Andrei
- Balan Iulia
- Maieczki Petronela-Sînziana
- Cîrja Ioan
- Butnaru Raimond Eduard
- Pintilie Justinian
- Florea Alexandra

Opis

1. Scopul documentului

Acest document este destinat să descrie cu exactitate capabilitățile proiectului software pentru analiza și clasificarea datelor referitoare la Hepatita C, utilizând algoritmi de învățare automată.

Clarifică specificațiile proiectului, obiectivele și constrângerile, ajutând la înțelegerea modului în care produsul va îndeplini cerințele funcționale și nefuncționale, facilitând astfel o implementare precisă și o evaluare corectă a performanței în procesarea datelor (a biomarker-ilor) și clasificarea rezultatelor.

2. Descriere generală

Această aplicație software, realizată în Matlab, este destinată clasificării pacienților cu hepatită C utilizând un set de date disponibil pe Kaggle, care conține biomarkeri și categorii de diagnostic.

Setul de date conține valori de laborator ale donatorilor de sânge și ale pacienților cu Hepatita C, precum și valori demografice, cum ar fi vârsta, etc. Proiectul implică încărcarea și curățarea setului de date, analizarea și preprocesarea variabilelor, inclusiv verificarea și corectarea valorilor lipsă, și calcularea unor statistici descriptive pentru detectarea eventualelor anomalii sau discrepanțe în ceea ce privește datele. Se vor construi diverse partitii ale seturilor de date de antrenament și testare pentru evaluarea performanței modelelor. Algoritmul Random Forest va fi utilizat pentru clasificare, urmând să se realizeze selecția trăsăturilor relevante și evaluarea avantajelor și dezavantajelor acestei abordări, luând în considerare variații în proporțiile de împărțire a datelor pentru antrenare și testare (80%-20%, 70%-30%, 60%-40%, 50%-50%). Analiza noastră experimentelă vizează diferențierea între donatorii de sânge și pacienții cu Hepatita, evidențiind/nu evoluția bolii la Hepatita C, Fibroză și Ciroză.

3. Descrierea generală a produsului

3.1. Situația curentă

Setul de date conține valori de laborator ale donatorilor de sânge și ale pacienților cu Hepatita, precum și valori demografice, cum ar fi vârsta si sexul. Datele au fost obținute din *UCI Machine Learning Repository.* În prezent, datele brute conțin valori lipsă și potențiale anomalii care pot afecta rezultatele analizei. Fără o preprocesare adecvată și un model de clasificare eficient, aceste date nu pot fi utilizate

corespunzător pentru a sprijini clasificarea, diagnosticarea sau prevenirea bolii hepatice.

3.2. Scopul produsului

Algoritmul software are ca scop analiza experimentală și clasificarea datelor medicale (provenite din mediul unui laborator de analize), utilizând tehnici de învățare automată pentru a sprijini diagnosticarea și prevenirea bolii. Acesta va permite preprocesarea datelor brute, identificarea trăsăturilor relevante și aplicarea unui model de clasificare, cum ar fi Random Forest, pentru a oferi predicții precise despre starea pacienților. Prin intermediul acestei soluții, datele vor fi curățate, structurate și analizate într-un mod eficient, oferind informații utile atât cercetătorilor, cât și practicienilor medicali pentru o mai bună înțelegere și gestionare a bolilor hepatice.

3.3. Contextul produsului și motivarea implementării

Bolile hepatice, inclusiv hepatita, reprezintă o problemă de sănătate publică majoră la nivel mondial. Diagnosticarea corectă în timp util a acestor afecțiuni este crucială pentru tratarea eficientă a pacienților și pentru prevenirea complicațiilor grave care pot apărea.

În contextul curent de avansare tehnologică și medicală, potențialul de îmbunătățire a metodelor clasice de diagnosticare a bolilor hepatice este în creștere. Aplicațiile ce sunt bazate pe modele IA(Invatare Automata)/ML(Machine Learning) pot analiza mult mai rapid și chiar mai precis volumele mari de date, identificând tipare care pot scăpa ochiului uman, oferind astfel un sprijin în luarea deciziilor medicale, bazat pe date concrete. Setul de date selectat pentru această aplicație include diferiți biomarkeri care sunt utilizați pentru a determina dacă un pacient poate fi donator sau dacă pacientul respectiv prezintă semne de afecțiuni hepatice.

3.4. Beneficii

Printre beneficii se enumeră:

- **Diagnostice îmbunătățite** - Utilizarea tehnicilor de ML cum ar fi, spre exemplu, Random Forest, aplicația poate clasifica datele medicale pentru a identifica pacienții predispuși la boli hepatice, aducând astfel o diagnosticare mai rapidă, mai precisă și mult mai eficientă.

- **Costuri și timp de analiză reduse** - Analizele tradiționale sunt costisitoare și consumatoare de timp. Utilizând algoritmi automați pentru prelucrarea datelor, aplicația poate reduce semnificativ timpul necesar pentru a obține un diagnostic, reducând astfel costurile și facilitând accesul la diagnosticare pentru mai mulți pacienți.
- **Prevenirea și intervenția timpurie** - Prin identificarea pacienților cu risc încă din fazele incipiente ale bolii, se poate interveni mai devreme, prevenind agravarea afecțiunii și reducând necesitatea unor tratamente mai agresive și costisitoare în stadiile avansate.
- **Validarea și analiza datelor** - Aplicația va efectua o serie de pași pentru pre-procesarea datelor, cum ar fi eliminarea valorilor lipsă și corectarea datelor incorecte. În plus, analiza statistică (medii, dispersii, valori minime și maxime) și indicatorii specifici vor oferi o imagine de ansamblu asupra caracteristicilor setului de date, permițând identificarea tiparelor importante și a corelațiilor relevante pentru diagnostic

4. Specificații funcționale

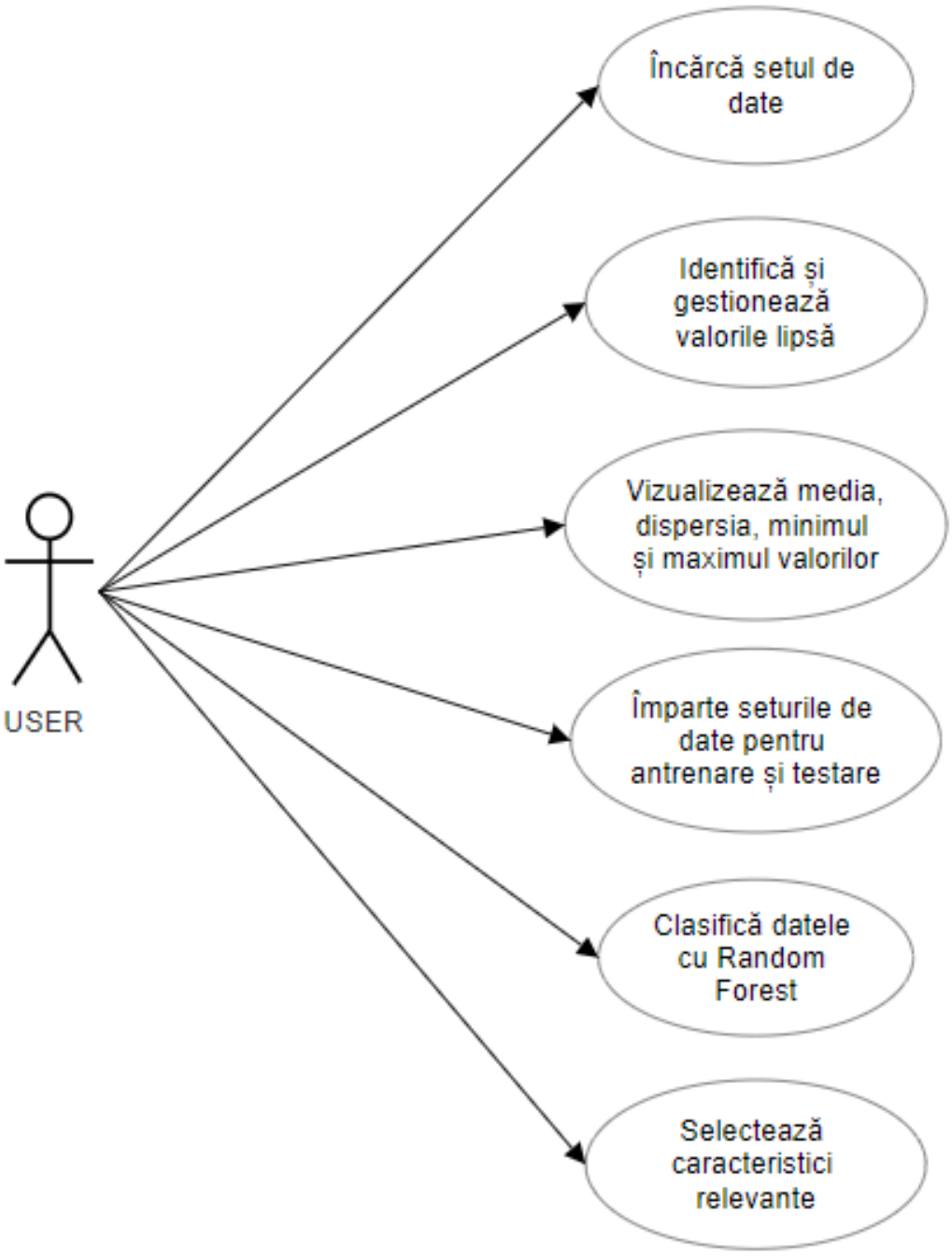
4.1. Actori

În cazul acestui proiect singurul actor este utilizatorul(User-ul). Mai jos sunt descrise toate acțiunile pe care le poate executa.

Acțiuni User:

- Încarcă setul de date
- Identifică și gestionează valorile lipsă
- Vizualizează media, dispersia, minimul și maximul valorilor
- Împarte seturile de date pentru antrenare și testare
- Clasifică datele cu Random Forest
- Selectează caracteristici relevante

4.2. Diagrama cazurilor de utilizare



4.3. Descrierea cazurilor de utilizare

4.3.1. Încărcarea setului de date;

USE CASE:	**Încărcare set de date**
Descriere:	Setul de date este încărcat pentru analiza și procesarea ulterioară.
Prioritate:	Esențial
Declanșator:	Cercetătorul sau utilizatorul selectează fișierul setului de date.
Precondiție:	Utilizatorul are acces la setul de date.
Calea de bază:	Utilizatorul încarcă fișierul CSV cu datele despre medicale.
Calea alternativă:	Dacă fișierul este corupt sau formatul nu este recunoscut, se returnează un mesaj de eroare.
Postcondiții:	Datele sunt încărcate și pregătite pentru preprocesare.
Calea pentru excepții:	Dacă fișierul nu este disponibil sau nu poate fi încărcat, sistemul returnează un mesaj de eroare și revine la starea inițială.

4.3.2. Identificarea și gestionarea valorilor lipsă;

USE CASE:	**Identificare și gestionarea a valorilor lipsă**
Descriere:	Se identifică și se gestionează valorile lipsă sau nedefinite din setul de date despre hepatită.
Prioritate:	Esențial
Declanșator:	Utilizatorul începe analiza setului de date și verifică dacă există date lipsă.
Precondiție:	Datele sunt încărcate și accesibile pentru procesare.
Calea de bază:	Valorile lipsă sunt gestionate fie prin eliminarea eșantioanelor incomplete, fie prin completarea lor folosind metode statistice.
Calea alternativă:	Dacă nu există valori lipsă, se trece direct la analiza statistică.
Postcondiții:	Setul de date este complet și pregătit pentru analiza statistică și preprocesare.
Calea pentru excepții:	Dacă prea multe date lipsesc, se sugerează utilizatorului eliminarea unor coloane sau reîncărcarea datelor.

4.3.3. Vizualizarea mediei, dispersiei, a minimului și a maximului;

USE CASE:	**Analiza statistică a variabilelor de intrare**
Descriere:	Calcularea unor statistici de bază (media, dispersia, valorile minime și maxime) pentru variabilele de intrare.
Prioritate:	Esențial
Declanșator:	Utilizatorul dorește să verifice distribuția valorilor biomarkerilor și să identifice eventuale eșantioane izolate sau discrepanțe.
Precondiție:	Datele sunt complete și pregătite pentru analiză.
Calea de bază:	1. Se identifică și se raportează eventualele valori anormale sau discrepanțe în setul de date. 2. Se realizează o vizualizare simplă, cum ar fi histogramele, pentru a observa distribuția valorilor.
Calea alternativă:	Dacă datele sunt deja normalizate sau corectate, se poate trece direct la pasul următor de preprocesare.
Postcondiții:	Caracteristicile principale ale setului de date sunt identificate, și eventualele eșantioane izolate sunt raportate.
Calea pentru excepții:	Dacă setul de date conține erori majore sau valori neașteptate, utilizatorul este informat să revizuiască preprocesarea.

4.3.4. Împărțirea seturilor de date pentru antrenare și testare;

USE CASE:	**Împărțire seturi de date pentru antrenare și testare**
Descriere:	Setul de date despre hepatită este împărțit în seturi de antrenare și testare pentru modelarea ulterioară.
Prioritate:	Esențial
Declanșator:	Utilizatorul dorește să pregătească datele pentru antrenarea și testarea unui model de clasificare.
Precondiție:	Datele sunt curățate și analizate.
Calea de bază:	1. Utilizatorul împarte setul de date folosind funcția în proporții de 80%-20%, 70%-30%, 60%-40% și 50%-50%. 2. Se verifică distribuția valorilor țintă în ambele seturi pentru a se asigura că nu există dezechilibre majore.
Calea alternativă:	În cazul dezechilibrelor de distribuție, se poate utiliza stratificarea pentru a menține distribuția uniformă a clasei.
Postcondiții:	Seturile de date pentru antrenare și testare sunt pregătite și echilibrate pentru următoarea etapă de clasificare.
Calea pentru excepții:	Dacă distribuția este dezechilibrată, utilizatorul este avertizat și se recomandă reîmpărțirea setului de date.

4.3.5. Clasifică datele cu Random Forest;

USE CASE:	**Clasificare de date cu Random Forest**
Descriere:	Datele despre hepatită sunt clasificate folosind algoritmul Random Forest.
Prioritate:	Esențial
Declanșator:	Utilizatorul dorește să clasifice pacienții în funcție de biomarkerii lor și să determine dacă sunt donatori sau au boli hepatice.
Precondiție:	Seturile de date de antrenare și testare sunt pregătite.
Calea de bază:	1. Modelul este antrenat pe setul de date de antrenare. 2. Performanța modelului este evaluată pe setul de testare folosind metrici precum acuratețea, precizia și matricea de confuzie.
Calea alternativă:	Dacă performanța este slabă, utilizatorul poate ajusta parametrii modelului.
Postcondiții:	Modelul Random Forest este antrenat și evaluat, iar rezultatele sunt raportate.
Calea pentru excepții:	Dacă modelul nu converge sau are o performanță slabă, se recomandă reanalizarea setului de date.

4.3.6. Selectează caracteristici relevante;

USE CASE:	**Selecția caracteristicilor relevante**
Descriere:	Se identifică trăsăturile biomarkerilor relevanți pentru clasificarea eficientă a pacienților.
Prioritate:	Esențial
Declanșator:	Utilizatorul dorește să optimizeze modelul prin selecția caracteristicilor relevante.
Precondiție:	Modelul de clasificare a fost construit și evaluat.
Calea de bază:	Caracteristicile irelevante sau redundante sunt eliminate din setul de date.
Calea alternativă:	Dacă selecția de trăsături nu îmbunătățește modelul, se recomandă utilizarea unor tehnici de reducere a dimensionalității.
Postcondiții:	Setul de date conține doar trăsăturile relevante pentru o clasificare eficientă.
Calea pentru excepții:	Dacă nu există diferențe semnificative în performanță, modelul poate fi recalibrat cu toate caracteristicile.

5. Specificații non-funcționale

5.1. Specificațiile interfeței cu utilizatorul

- Utilizatorul trebuie să fie familiarizat cu Matlab, fiind capabil să pornească un script și să înțeleagă rezultatele generate.
- Interacțiunea cu programul trebuie să fie minimă, cu instrucțiuni clare și ușor de urmărit.

5.2. Specificațiile de performanță

- Programul trebuie să ruleze eficient pe orice sistem care suportă Matlab, fără a necesita resurse hardware suplimentare.
- Timpul de execuție trebuie să fie optim pentru a gestiona seturi de date mari, iar calculele complexe să fie finalizate într-un interval rezonabil.

5.3.Disponibilitatea și fiabilitatea

- Programul trebuie să producă întotdeauna rezultatul corect și să fie capabil să semnaleze erorile umane, cum ar fi date lipsă sau formate incorecte.
- În cazul apariției unor probleme, programul trebuie să returneze mesaje de eroare clare pentru a ajuta utilizatorul să corecteze situația.