- Information Visualization -

Caracteristicile principale ale deceselor datorate problemelor cardiace.

- Project Design –

Cuprins

1. Obiectivele proiectului

Scopul principal al acestui proiect este de a analiza principalele motive pentru care oameni se îmbolnăvesc de probleme cardiace.

Vom observa distributia fiecărei variabile și corelatia acestora pentru stabili cât de important este să menținem anumite măsurători într-un anumit interval ce ne va fi oferit în urma vizualizării diferitelor aspecte ale setului de date. În urma acestui studiu vom lua în considerare și identificarea valorilor extreme (outliers) pentru fiecare variabila si identificarea pacientilor care prezinta aceste valori extreme.

În urma aflării mediei, deviatiei standard si varianței pentru fiecare variabila ne vom face și o părere despre distributia variabilelor care ne poate ajuta la identificarea unor tendinte sau anomalii.

Prin identificarea corelatiilor intre variabilele din setul tau de date vom înțelege mai bine cum variabilele afecteaza una pe alta.

În urma unei analize preliminare a setului de date am identificat anumite caracteristici precum grupul sau starea finala a pacientului. Acestea ne pot îndemna să divizăm setul de date în anumite subdiviziuni pentru a analiza mai bine situația anumitor tipuri de pacienți.

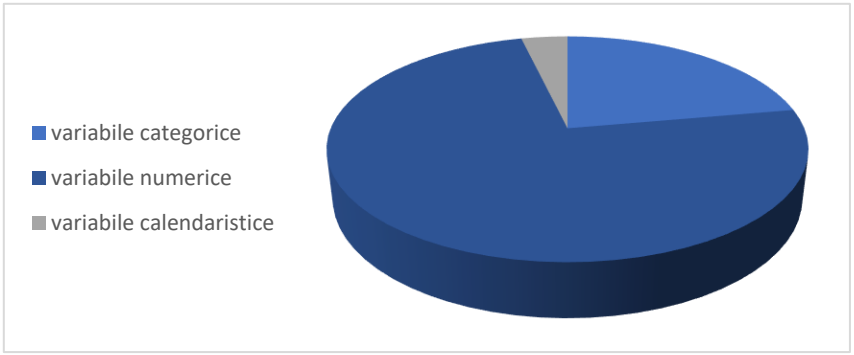
După ce am realizat un top al principalelor deficiențe care ar trebui să ne îngrijore, vom realiza o analiza a unor modele de regresie pentru a identifica variabilele care sunt asociate cu riscul de a dezvolta boli de inima, diabet sau alte deficiențe.

2. ***Prezentarea setului de date***

Setul de date ales de noi a fost extras dintr-o bază de date postgres numită ***MIMIC III Clinical Test Database*** *ce conține o serie de inregistrări medicale variante pornind de la diverse tulburări cardiace până la o listă de tratamente oferite pacienților. Datele noatre au fost extrase prin intermediul unei operații de join asupra tabelelor ce dețin informații cu interes pentru noi. În urma procesării tabelelor, s-a putut extrage un set de date cu 8239 înregistrări care se poate regăsi și studia la urmatoarea adresă.*

50 + features		
10	40	2
categorical columns	numerical columns	date columns
8239 rows		

În urma analizei acestui dataset, am extras următoarele date: ***caracteristici demografice*** *(varsta la momentul internării în spital, sex, etnie, greutate și înalțime),* ***semne vitale*** *(frecvență cardiacă[HR], tensiunea arterială sistolică[SBP], tensiunea arterială diastolică[DBP], tensiunea arterială medie[MBP], frecventa respiratorie[RR], temperatura corpului[BT], saturatia pulsului de oxygen[SPO2], urină), comorbidități (hipertensiune arterială, fibrilație atrială, boala cardiacă ischemică, diabet zaharat[DM], depresie, anemie hipoferică[HA], hiperlipidemie, boală renală cronică[CKD], boală pulmonară obstructivă cronică[BPOC]) și* ***variabile de laborator*** *(hematocrit, globule roșii, hemoglobina corpusculară medie [MCH], concentrația medie a hemoglobinei corpusculare [MCHC], volumul corpuscular mediu [MCV], lățimea distribuției globulelor roșii [RDW], numărul de trombocite, globule albe, neutrofile, bazofile, limfocite, timp de protrombină [PT], raport internațional normalizat [INR], NT-proBNP, creatin kinază, creatinină, azot ureic din sânge [BUN], glucoză, potasiu, sodiu, calciu, clorură, magneziu, gap anionic, bicarbonat, lactate, concentrația ionilor de hidrogen [pH], presiunea parțială a CO2 în sângele arterial și LVEF).*



Caracteristicile demografice și semnele vitale extrase au fost înregistrate în primele 24 de ore de la fiecare internare și au fost măsurate variabilele de laborator pe toată durata ședерii la unitățiile de la terapie intensivă. Comorbiditățile au fost identificate folosind codurile ICD-9. Pentru variabile de laborator au fost realizate măsurători multiple.

	data type
group	*categorical data*
ID	*numerical data*
dod	*time series*
outcome	*categorical data*
age	*numerical data*
gendera	*categorical data*
BMI	*numerical data*
dofa	*time series*
hypertensive	*categorical data*
atrialfibrillation	*categorical data*
CHD with no MI	*categorical data*
diabetes	*categorical data*
deficiencyanemias	*categorical data*
depression	*categorical data*
Hyperlipemia	*categorical data*
Renal failure	*categorical data*
COPD	*categorical data*
heart rate	*numerical data*
Systolic blood pressure	*numerical data*
Diastolic blood pressure	*numerical data*
Respiratory rate	*numerical data*
temperature	*numerical data*
SP O2	*numerical data*
Urine output	*numerical data*
hematocrit	*numerical data*
RBC	*numerical data*
MCH	*numerical data*
MCHC	*numerical data*
MCV	*numerical data*
RDW	*numerical data*
Leucocyte	*numerical data*
Platelets	*numerical data*
Neutrophils	*numerical data*
Basophils	*numerical data*
Lymphocyte	*numerical data*
PT	*numerical data*
INR	*numerical data*
NT-proBNP	*numerical data*
Creatine kinase	*numerical data*
Creatinine	*numerical data*
Urea nitrogen	*numerical data*
glucose	*numerical data*
Blood potassium	*numerical data*
Blood sodium	*numerical data*
Blood calcium	*numerical data*
Chloride	*numerical data*
Anion gap	*numerical data*
Magnesium ion	*numerical data*
PH	*numerical data*
Bicarbonate	*numerical data*
Lactic acid	*numerical data*
PCO2	*numerical data*
EF	*numerical data*

3. ***Alegerea graficelor potrivite***

În urma stabilirii tipurilor de date întâlnite în setul de date ales trebuie să stabilim principalele grafice pe care le vom realiza pe parcursul studiului unei anumite variabile.

Atunci când vine vorba de reprezentarea unei variabile într-un mod vizual, alegerea tipului de grafic potrivit este esențială pentru a transmite informațiile în mod eficient și ușor de înțeles. Alegerea unui grafic nepotrivit poate duce la o interpretare incorectă a datelor și poate afecta deciziile care sunt luate în urma analizei.

Alegerea celui mai bun tip de grafic se realizează prin stabilirea tipurilor de date ale variabilelor și analizarea mai multor tipuri de grafice.

În setul de date ales de noi există 3 tipuri principale de variabile: categorice, numerice și interval de timp. În funcție de tipul de date stabilit avem următoarele posibilități de grafice ce se pot regăsi în tabelul alăturat.

	variabile categorice	variabile numerice	variabile de tip date calendaristice
line chart		x	x
pie chart	x		
bar chart	x		
frequency chart	x		
strip chart	x	x	x
box and whisker chart		x	x
quartile chart	x	x	
histogram		x	x
scatter plot		x	
percentage chart		x	
radar chart		x	
Gantt chart			x
heatmap calendar chart			x

4. ***Tehnologiile utilizate***

Tehnologiile folosite de echipa noastră pentru proiect vor fi ***Pyspark****(python) si* ***PowerBI****.*

Pyspark *este o librărie Python pentru procesarea datelor în cadrul platformei Apache Spark. Apache Spark este un framework de procesare a datelor în memorie, conceput pentru a procesa cantități mari de date în mod eficient și rapid, utilizând un model distribuit de calcul.*

Pyspark este util pentru proiectul nostru, deoarece oferă o mulțime de funcționalități puternice pentru procesarea și analiza datelor. Printre aceste funcționalități, vom aminti de:

1. ***Preprocesarea datelor****: Pyspark poate fi utilizat pentru a încărca, curăța și transforma datele într-un format adecvat pentru analiza ulterioară. Datorită capacității sale de a procesa cantități mari de date, pyspark este potrivit pentru prelucrarea datelor în timp real sau a datelor istorice.*
2. ***Analiza datelor****: Pyspark oferă o serie de funcții pentru analiza datelor, cum ar fi agregarea, filtrarea și sortarea datelor. De asemenea, permite utilizarea de algoritmi de machine learning pentru modelarea datelor și dezvoltarea de modele predictibile.*
3. ***Vizualizarea datelor****: Pyspark poate fi integrat cu alte librării de vizualizare a datelor, cum ar fi Matplotlib sau Seaborn, pentru a crea grafice și diagrame care să ajute la înțelegerea datelor.*

Power BI este un instrument puternic pentru proiectele de data science, care permite utilizatorilor să vizualizeze, să analizeze și să partajeze datele într-un mod interactiv și ușor de înțeles. Power BI poate fi utilizat în cadrul proiectului nostru pentru a integra și analiza date pentru a crea vizualizări personalizate.

5. ***Etapele principale***

<table>
<tr><td>step 0</td><td colspan="4">extragerea datelor din tabelele bazei de date</td></tr>
<tr><td colspan="5"></td></tr>
<tr><td colspan="3">PySpark & python</td><td colspan="2">Power BI</td></tr>
<tr><td>step 1</td><td>analizarea datelor lipsă</td><td rowspan="3"><u>PROCESAREA DATELOR</u></td><td colspan="2" rowspan="3"></td></tr>
<tr><td>step 2</td><td>stabilirea modalităților de completare a setului de date</td></tr>
<tr><td>step 3</td><td>procesarea datelor extrase</td></tr>
<tr><td>step 4</td><td colspan="4">stabilirea strategiei de analiza a datelor și tipurile de grafice folosite</td></tr>
<tr><td>step 5</td><td>calcularea statisticilor descriptive (mediana,media, deviatia standard)</td><td rowspan="4"><u>ANALIZA EXPLORATORIE A DATELOR</u></td><td>step 6</td><td>calcularea statisticilor descriptive (mediana,media, deviatia standard)</td></tr>
<tr><td>step 7</td><td>distributia fiecarei variabile (histrograms, density plot, bar plot, box plot)</td><td>step 8</td><td>distributia fiecarei variabile (histrograms, density plot, bar plot, box plot)</td></tr>
<tr><td>step 9</td><td>identificarea anomaniilor si valorilor extreme</td><td>step 10</td><td>identificarea anomaniilor si valorilor extreme</td></tr>
<tr><td>step 11</td><td>relatiile dintre variabile setului de date (line plots scatter plots, correlation matrix)</td><td>step 12</td><td>relatiile dintre variabile setului de date (line plots scatter plots, correlation matrix)</td></tr>
<tr><td>step 13</td><td>crearea unui model de pentru a identifica variabilele care sunt asociate cu riscul de a dezvolta boli de inima, diabet sau alte deficiențe.</td><td><u>REZULTATE FINALE</u></td><td>step 14</td><td>crearea unui dashboard interactiv în care să prezentăm informațiile descoperite în urma analizei exploratorie a datelor.</td></tr>
</table>

6. ***Concluzii***

Rezultatul final al acestui proiect este de a scoate in evidenta caracteristicile importante care sunt semnificative pentru decesul indiviziilor care au participat la testele realizate, de a realiza o serie de modele de regresie care să ne ofere mai multe detalii cu privire caracteristicile corelate în prezicerea boliilor cardiovasculare și nu numai.