Tema 2 IA

*3.1. Explorarea Datelor*

*3.1.1. Analiza tipului de atribute și a plajei de valori a acestora*

Atributele continue sunt specificate în temă, pentru a fi apoi analizate.

Plajele de date ale atributelor continue (pentru diabetes, respective credit\_risk) arată astfel:

A graph of a number of objects

Description automatically generated with medium confidence

În cazul psychological-rating, BodyMassIndex și CognitionScore, nu avem atât de multe outliere, distribuția fiind concentrate în jurul unei valori mici(0).

În cazul Age valorile nu variază la fel de tare (nu se poate observa clar aici scara la care variază, deoarece lucrăm cu valori mari în rest) și outlierele nu există (nu avea sens să existe outliere la vârstă, nu am luat vampiri la sondaje)

În cazul Body\_Stats se văd foarte clar outlierele, fiind multe și mergând mult peste box și whiskers. Distribuția este largă și se poate observa cu outlierele sunt ”concentrate la capete”. Există, deci, multe persoane cu valori peste medie.

În cazul Metabolical\_Rate nu vedem outliere, probabil asta implică lipsa/numărul mic de oameni din statistici care să aibă probleme metabolice severe sau că până și în cazul acelor persoane nu există o variație foarte mare față de medie.

A graph with numbers and a line

Description automatically generated

Aici observăm cel mai clar outlierele la applicant\_income, iar de data asta majoritatea lor sunt concentrate în apropierea boxului. Cu cât mai sus ajungem, cu atât mai puțini oameni care se încadrează în acea categorie(nu oricine e bogat; totodată observăm mulți oameni upper middle class).

În cazul tuturor celorlalte categorii se aplică ce am menționat la graficul pentru diabet în primul paragraf.

În implementare am comentat și o afișare pentru fiecare tabel, for better viewing și pentru a se înțelege mai bine categoriile și outlierele.

**Plajele de date ale atributelor discrete (categorice):**

A graph of different colored bars

Description automatically generated

În cazul multor atribute (PreCVA, RoutineChecks, alcoholAbuse, cholesterol\_ver, vegetables, HighBP, Unprocessed\_fructose, myocardial\_infarction, Cardio, ImprovedAveragePulmonaryCapacity, Diabetes) întâlnim doar valori între 0 și 1, cu o tendință spre 0, ceea ce indică fie lipsa acestor condiții ori faptul că nu sunt documentate.

Excepție face cholesterol\_ver, care are un număr mai echilibrat de instanțe între valorile 0 și 1, indicând o distribuție mai uniformă.

HealthScore are valori între 1 și 5, cu o distribuție concentrată spre 3, indicând un grad de sănătate cât se poate de mediu (spre scăzut, avem și mulți de 2).

SalaryBracket are valori între 1 și 8 și putem vedea o creștere spre 8, implicând că mulți participanți au salarii mari.

Pentru credit\_risk nu mi se face histograma, deci în cod le am pe toate făcute individual. Voi afișa una for reference și voi spune interpretările la toate după.

A graph of a number of blue rectangular objects

Description automatically generated

Aici observăm un număr mare de oameni care plătesc chirie, respectiv cu ipotecă. Multe cheltuieli.

În cazul loan\_purpose repartiția este echilibrată(motivele sunt variate, deci nu există unul mai des întâlnit decât altele pentru împrumut).

La loan\_approval\_status majoritatea au aprobare(apx. 70-80%).

La loan\_rating sunt multe ratinguri bune(mai ales very good și excellent) și foarte puține negative.

La credit\_history, majoritatea nu au avut card de credit anterior.

Stabilty\_rating este oarecum echilibrat, majoritatea având C, urmat de B. Foarte puțini au D.

*3.2. Analiza echilibrului de clase*

A graph of a number of blue squares

Description automatically generated with medium confidence

A graph with blue squares

Description automatically generated with medium confidence

Cele două tabele seamănă, având predominant diabetul de tip 0, urmat de 2 apoi foarte puțini având 1. În cazul \_test, unde este și un eșantion mai mic, se găsesc mai multe cazuri de diabet de tip 2.

*3.3 Corelație dintre atribute*

Analiza de corelație arată astfel:

(pentru diabet)

A screenshot of a computer

Description automatically generated

Pe diagonala principală avem 1 iar în rest scorul de corelație.

A close-up of a grid

Description automatically generated

(pentru credit\_risk)

A screen shot of a computer

Description automatically generated

A screenshot of a grid

Description automatically generated

Aceste valori vor fi testate cu Chi Squared pentru a se decide dacă sunt sau nu corelate.

În urma aplicării algoritmului, tabelul care ne determină atributele redundante arată astfel:

A screenshot of a computer screen

Description automatically generated

Punctele în culori calde pot fi considerate puternic corelate, cele albastru deschis pot fi considerate moderat corelate, iar cele albastru închis nu sunt corelate.

Tabelul pare că trebuie modificat. Există prea puține valori corelate.

Atributele care sunt corelate pot fi apoi eliminate.