Inteligență artificială
ML aplicat

Andrei-Gabriel Mitran
andreigabrielmitran@gmail.com

Facultatea de Automatică și Calculatoare, Universitatea Politehnică din București

1 Explorarea Datelor

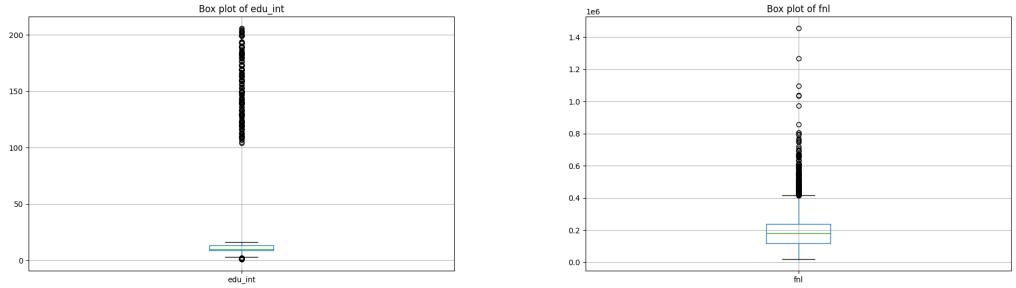
1.1 Analiza tipului de atribute și a plajei de valori a acestora

SalaryPrediction

În analiza datelor, se constată că unele atribute numerice continue prezintă o variație semnificativă, cum ar fi "years", în timp ce altele, cum ar fi "gain" și "loss", au valori aproape constante, de obicei 0. În plus, atributul "hpw" are valori apropiate de media setului de date, în timp ce majoritatea altor atribute prezintă valori extreme (outliers) mai mari decât limita superioară a interquartile range-ului (whisker-ul superior). De asemenea, există outliere sub limita inferioară a interquartile range-ului pentru atributul "hpw".

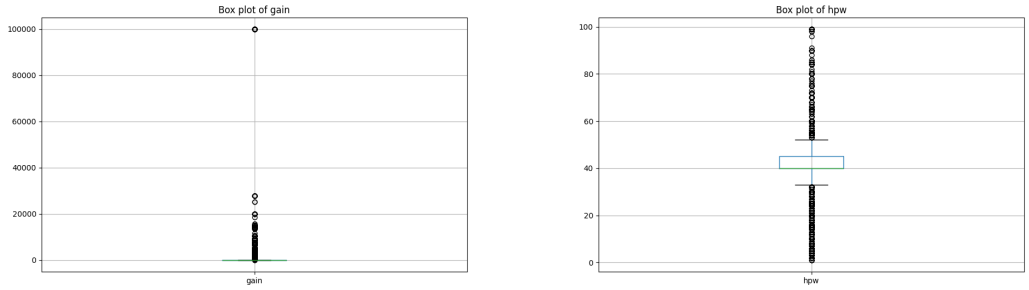
	fnl	hpw	gain	edu_int	years	loss	prod
count	9999.0	9199.0	9999.0	9999.0	9999.0	9999.0	9999.0
mean	190352.902	40.416	979.853	14.262	38.647	84.111	2014.928
std	106070.863	12.517	7003.795	24.771	13.745	394.035	14007.604
min	19214.0	1.0	0.0	1.0	17.0	0.0	-28.0
25%	118282.5	40.0	0.0	9.0	28.0	0.0	42.0
50%	178472.0	40.0	0.0	10.0	37.0	0.0	57.0
75%	237311.0	45.0	0.0	13.0	48.0	0.0	77.0
max	1455435.0	99.0	99999.0	206.0	90.0	3770.0	200125.0

Fig. 1: Numerical attribute analysis



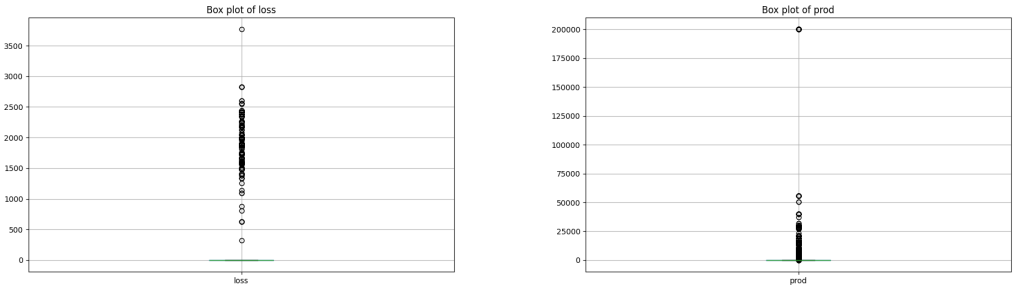
(a) Education (b) Socioeconomic characteristic

Fig. 2: Boxplots of Education and Socioeconomic characteristic



(a) Capital Gain (b) Hours Per Week

Fig. 3: Boxplots of Capital Gain and Hours Per Week



(a) Capital Loss (b) Productivity

Fig. 4: Boxplots of Capital Loss and Productivity

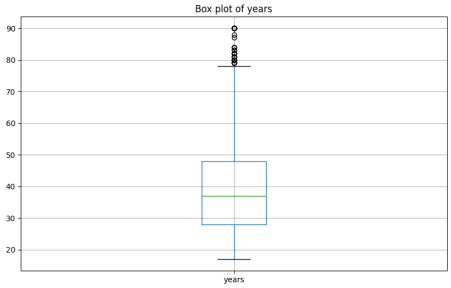
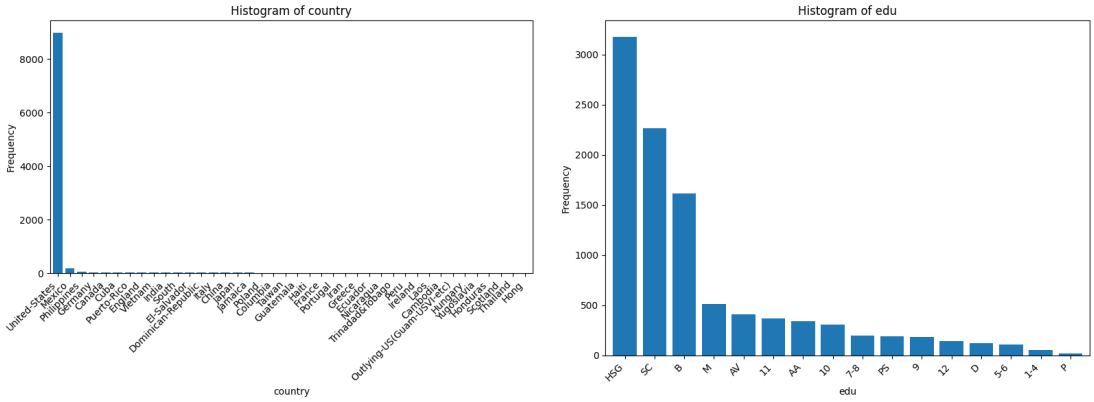


Fig. 5: Years Old

În prelucrarea datelor, am considerat "?" drept valoare lipsă. Cu excepția atributului "gender", toate celelalte atribute nu ar fi avut valori lipsă. Atributele "country", "race" și "work type" prezintă o valoare care apare mult mai des decât celelalte, în timp ce "edu", "partner" și "relation" au valori care formează o curbă descrescătoare. "Job" este caracterizat prin variație, în timp ce "gender" și "gtype" sunt atribute binare.

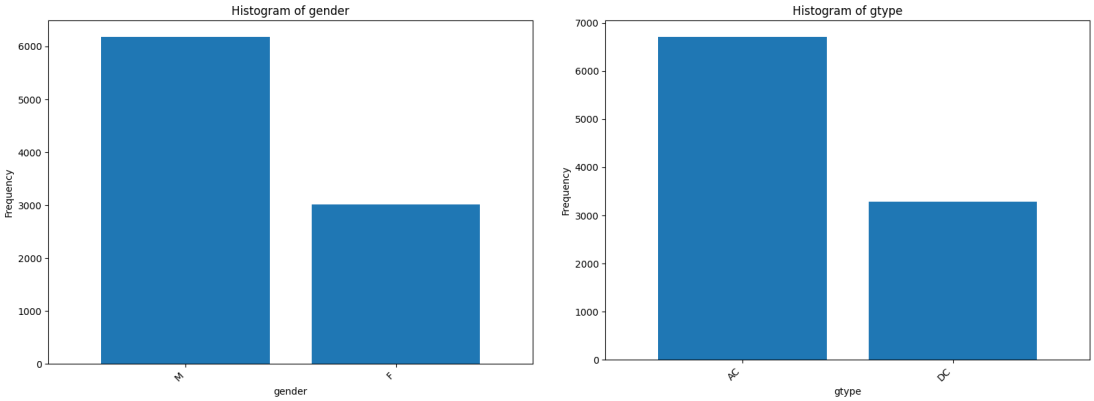
Attribute	Examples without Missing Values	Unique Values
relation	9999	6
country	9841	40
job	9417	13
work_type	9419	8
partner	9999	7
edu	9999	16
gender	9199	2
race	9999	5
gtype	9999	2

Fig. 6: Categorical Attributes



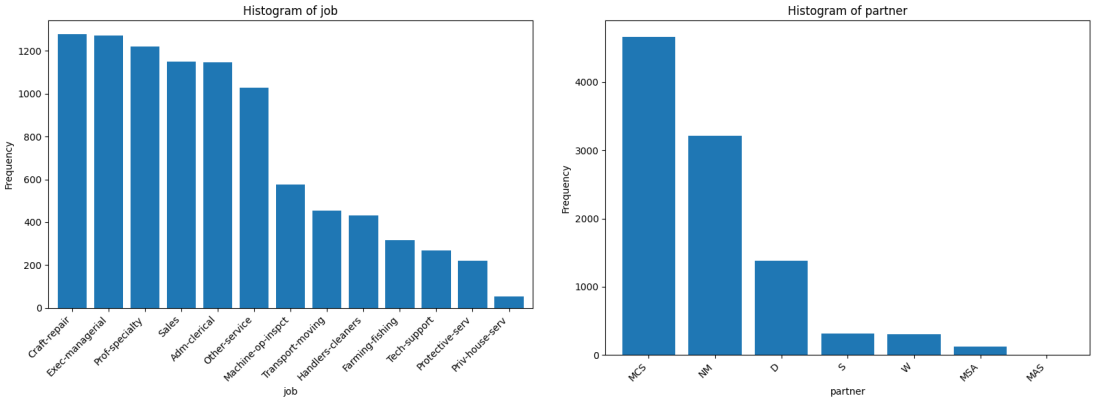
(a) Country (b) Education

Fig. 7: Histograms of Country and Education



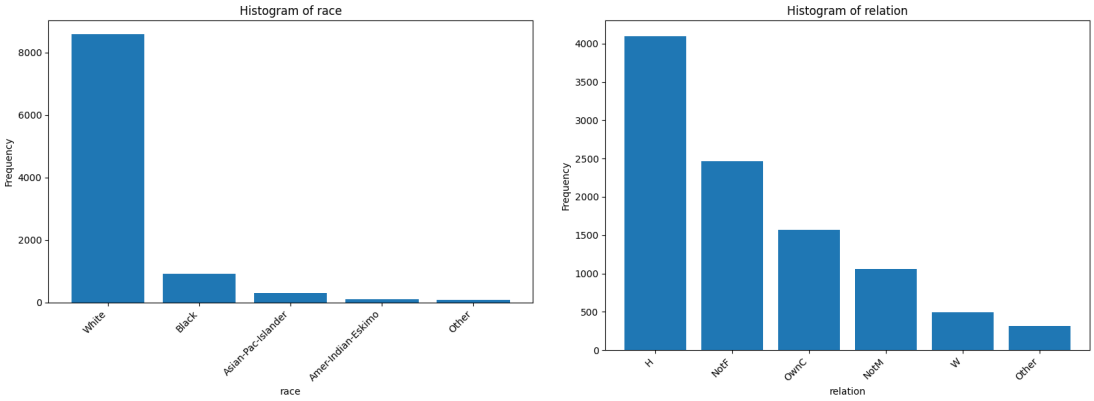
(a) Gender (b) Group Type

Fig. 8: Histograms of Gender and Group Type



(a) Job (b) Partner

Fig. 9: Histograms of Job and Partner



(a) Race (b) Relation

Fig. 10: Histograms of Race and Relation

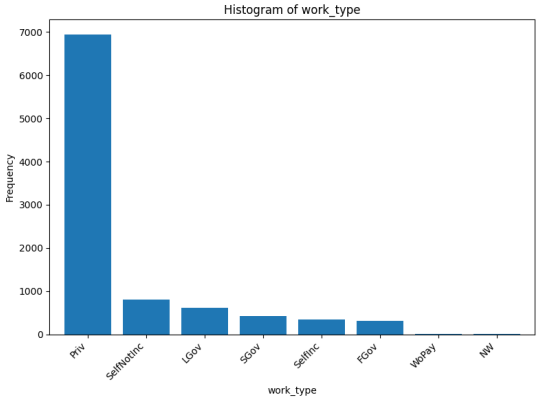


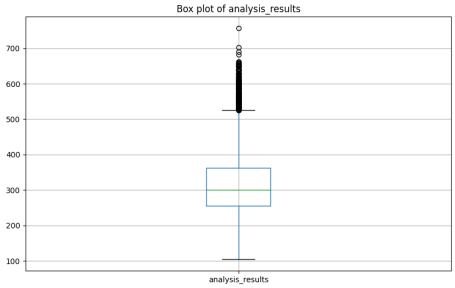
Fig. 11: Work Type

AVC

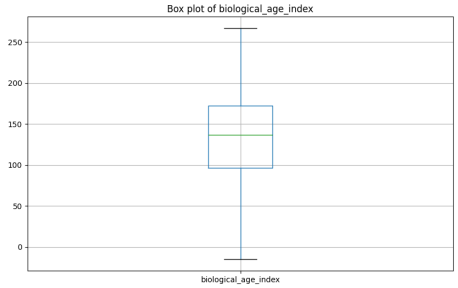
În setul de date, se observă variații semnificative, iar valori extreme (outliers) sunt întâlnite doar în partea superioară a distribuției.

	mean_blood_sugar_level	body_mass_indicator	years_old	analysis_results	biological_age_index
count	5110.0	4909.0	5110.0	4599.0	5110.0
mean	106.148	28.893	46.569	323.523	134.784
std	45.284	7.854	26.594	101.577	50.399
min	55.12	10.3	0.08	104.83	-15.109
25%	77.245	23.5	26.0	254.646	96.711
50%	91.885	28.1	47.0	301.032	136.375
75%	114.09	33.1	63.75	362.823	172.507
max	271.74	97.6	134.0	756.808	266.986

Fig. 12: Numerical attribute analysis

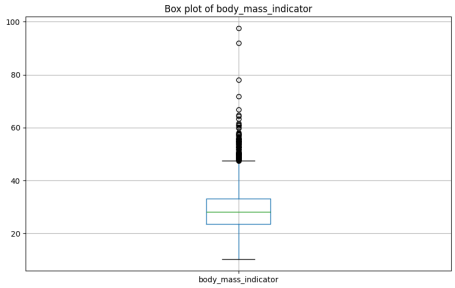


(a) Analysis Results

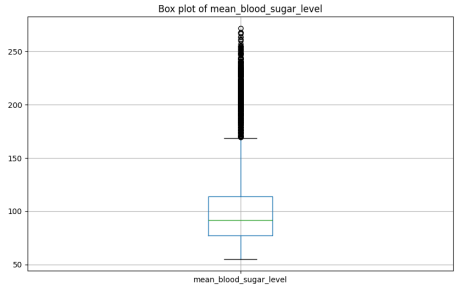


(b) Biological Age Index

Fig. 13: Boxplots of Analysis Results and Biological Age Index



(a) Body Mass Indicator



(b) Mean Blood Sugar Level

Fig. 14: Boxplots of Body Mass Indicator and Mean Blood Sugar Level

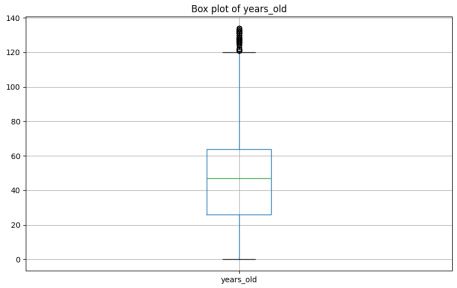
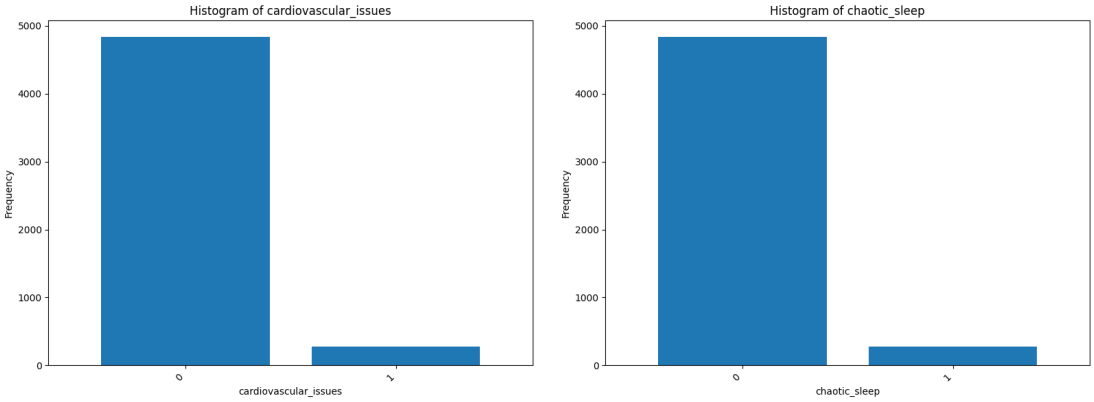


Fig. 15: Years Old

În ceea ce privește valorile lipsă, acestea sunt prezente doar în atributul "married". Se observă prezența unui număr semnificativ de atribute binare, cu o valoare mult mai rară comparativ cu cealaltă. De remarcat este distribuția aproape egală a atributului "living area", în timp ce categoria de job-uri are o valoare comună, 3 care sunt aproape egale, dar mai rar întâlnite, și una extrem de rar întâlnită.

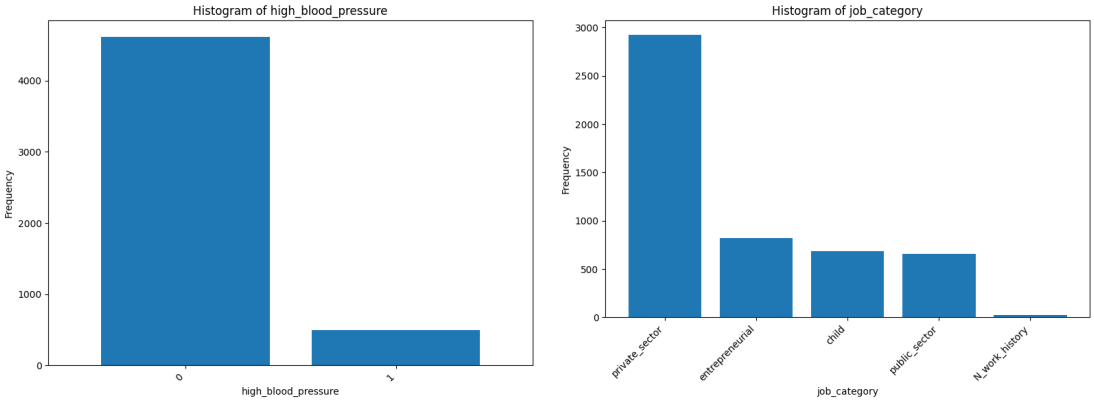
Attribute	Examples without Missing Values	Unique Values
cardiovascular_issues	5110	2
job_category	5110	5
sex	5110	2
tobacco_usage	5110	4
high_blood_pressure	5110	2
married	4599	2
living_area	5110	2
chaotic_sleep	5110	2

Fig. 16: Categorical Attributes



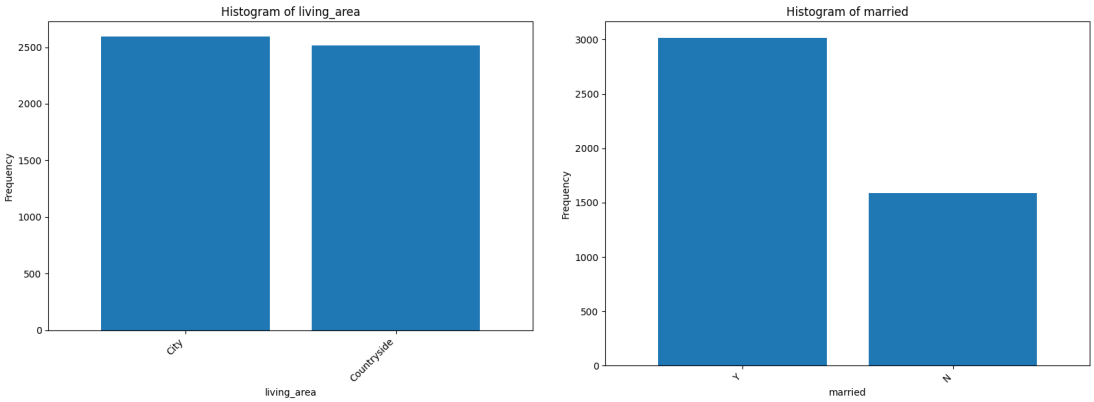
(a) Cardiovascular Issues (b) Chaotic Sleep

Fig. 17: Histograms of Cardiovascular Issues and Chaotic Sleep



(a) High Blood Pressure (b) Job Category

Fig. 18: Histograms of High Blood Pressure and Job Category



(a) Living Area (b) Married

Fig. 19: Histograms of Living Area and Married

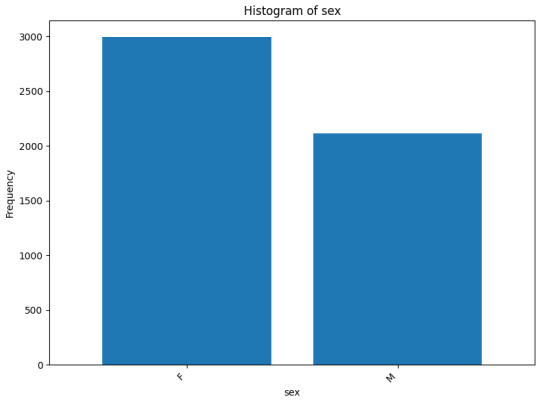


Fig. 20: Sex

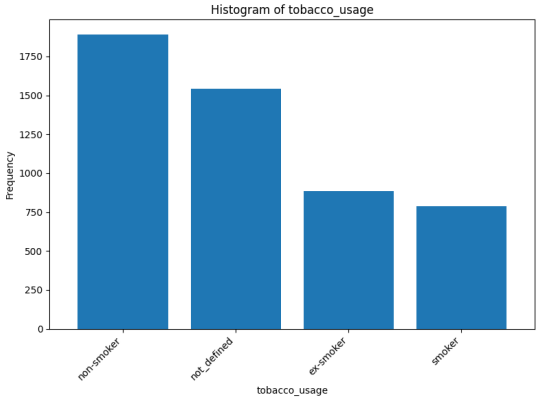
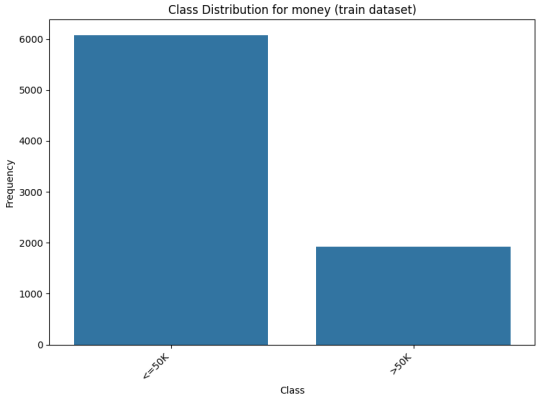


Fig. 21: Tobacco Usage

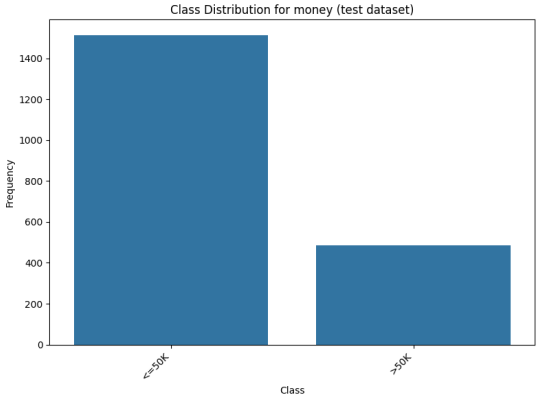
1.2 Analiza echilibrului de clase

Salary Prediction

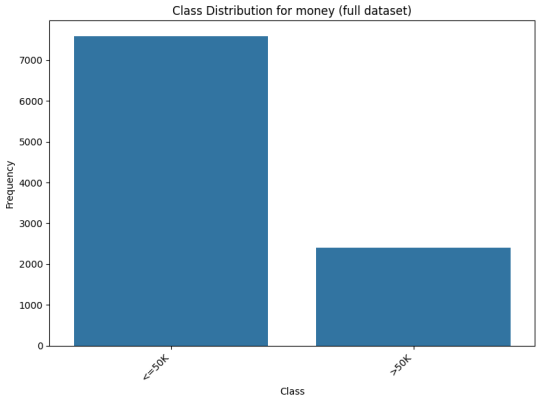
În distribuția claselor, observăm că clasa ”>50K” cuprinde aproape o treime din numărul de instanțe din clasa ”≤50K”. Acest raport între clase este menținut atât pe setul de date de antrenare (train), cât și pe cel de testare (test).



(a) Train Class Distribution



(b) Test Class Distribution

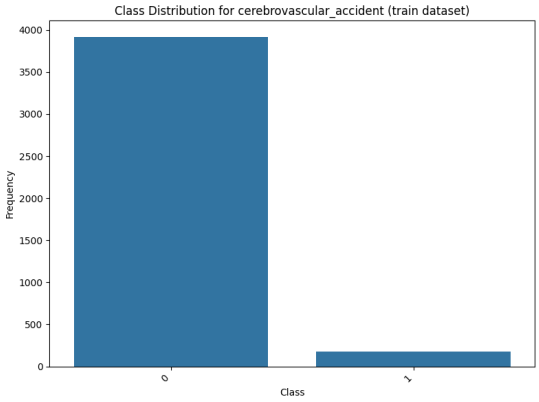


(c) Full Class Distribution

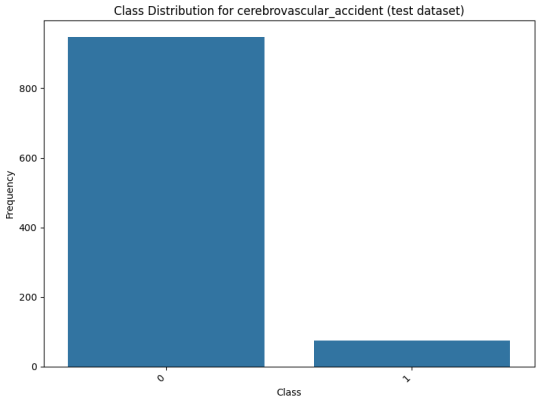
Fig. 22: Class Distribution for Salary Prediction

AVC

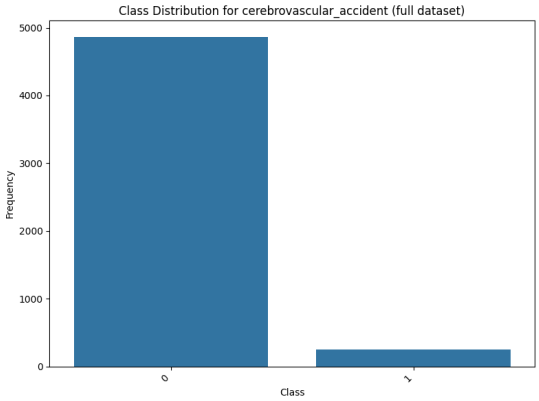
În setul de date, clasa care definește oamenii care nu au suferit un AVC este mult mai prezentă decât cei care au suferit un AVC. Acest lucru indică un dezechilibru în setul de date. Același dezechilibru este observat atât în setul de date de antrenare (train), cât și în cel de testare (test).



(a) Train Class Distribution



(b) Test Class Distribution



(c) Full Class Distribution

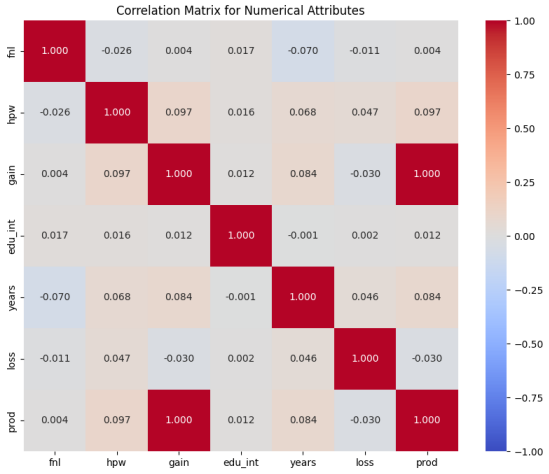
Fig. 23: Class Distribution for Salary Prediction

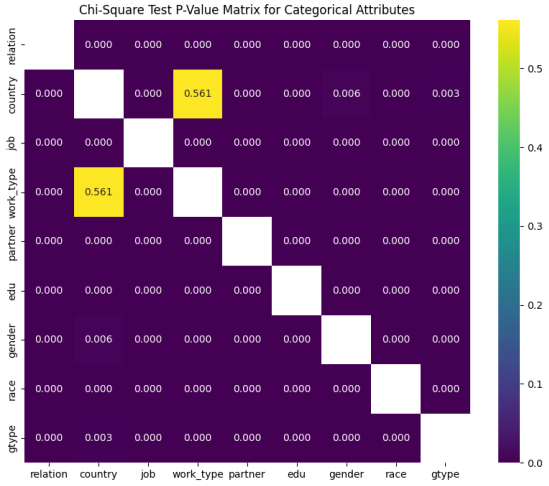
1.3 Analiza corelației între atribute

SalaryPrediction

În setul de date, observăm o puternică corelație între atributele "gain" și "prod". Există posibilitatea să se considere eliminarea unuia dintre acestea pentru a evita multicolinearitatea, însă în algoritmii discutați această eliminare nu a avut loc.

În ceea ce privește atributele non-numerice, majoritatea sunt corelate, cu excepția atributelor "country" și "work type".

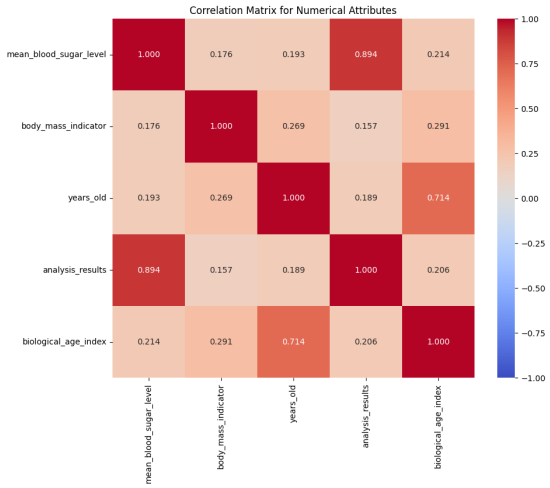




AVC

În analiza setului de date, se remarcă o corelație relativ mare între ”biological age index” și ”years old”, precum și între ”analysis results” și ”mean blood sugar level”.

În ceea ce privește atributele categorice, toate par să fie destul de corelate cu excepția ”living area”, care nu prezintă o corelație semnificativă cu niciun alt atribut.



2 Preprocesarea Datelor

2.1 Date lipsă pentru un atribut într-un eșantion

Pentru a trata datele lipsă din seturile de date, am aplicat o metodă de imputare. Am identificat mai întâi atributele care conțin valori lipsă și am decis să folosesc o abordare multivariată pentru a estima aceste valori lipsă. Această abordare permite să luăm în considerare informațiile din toate atributele disponibile pentru a estima valorile lipsă într-un mod mai precis.

2.2 Valori extreme pentru un atribut într-un eșantion

Pentru a detecta și trata valorile extreme din setul de date, am calculat intervalul interquartil (IQR) pentru fiecare atribut numeric. Valorile care se află în afara intervalului $(Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR)$ au fost considerate outlier-i și au fost înlocuite cu `NaN`, astfel încât să poată fi tratate ulterior.

2.3 Atribute redundante (puternic corelate)

Pentru a identifica atributele care sunt puternic corelate între ele, am efectuat o analiză de corelație asupra setului de date. Am observat că unele perechi de atribute au o corelație semnificativă. Cu toate acestea, nu am eliminat niciun atribut în acest stadiu al procesului nostru.

2.4 Plaje valorice de mărimi diferite pentru atributele numerice

Am observat că atributele numerice din setul de date au plaje valorice semnificativ diferite. Pentru a evita problemele care pot apărea din cauza acestor diferențe de scară, am aplicat o operație de standardizare asupra acestor atribute. Astfel, am asigurat că toate atributele numerice au aceeași scală și sunt comparabile între ele în procesul de analiză ulterior.

Acestea sunt pașii principali pe care i-am urmat în preprocesarea datelor pentru a asigura că setul nostru de date este pregătit corect pentru analiza ulterioară și antrenarea modelelor de machine learning.

3 Utilizarea algoritmilor de Învățare Automată

SMOTE (Synthetic Minority Over-sampling Technique) este utilizat pentru a echilibra setul de date înainte de antrenarea modelelor. Această tehnică creează noi exemple sintetice pentru clasa minoritară, asigurând o distribuție echilibrată a claselor în setul de date de antrenament.

"my train" implementează modele de învățare automată personalizate, bazate pe codul de la laborator, care utilizează metode specifice de preprocesare și optimizare. În contrast, "train" utilizează biblioteca scikit-learn pentru antrenarea modelelor, specificând hiperparametri relevanți specifici fiecărui tip de model.

Tipul de encodare folosit pentru fiecare atribut categoric:

- **În `my_train`:** Encodare one-hot.
- **În `train`:** Encodare one-hot.

Regresie Logistică

Setările algoritmului de optimizare de tip gradient descent folosit:

- **În `my_train`:**
  - Tip de optimizator: Stochastic Gradient Descent (SGD).
  - Learning rate: Valoare fixă specificată manual, cu ajustare ulterioară în timpul antrenării.
- **În `train`:**
  - Algoritmul de optimizare: Algoritmul implicit al lui `LogisticRegression`, care este o variantă a gradient descent.
  - Numărul maxim de iterații: 1000.

Metoda de regularizare folosită:

- **În `my_train`:** Nu s-a specificat utilizarea unei metode de regularizare (neschimbat din implementarea din laborator).
- **În `train`:** Nu s-a specificat utilizarea unei metode de regularizare (L2 default).

MLP

Arhitectura

- **În `my_train`:**
  - Numărul de straturi: 3.
  - Dimensiunea fiecărui strat:
    - Primul strat ascuns: 64 neuroni.
    - Stratul de ieșire: 2 neuroni.
  - Tipul funcțiilor de activare folosite:
    - Pentru straturile ascunse: ReLU (Rectified Linear Unit).
    - Pentru stratul de ieșire: Sigmoid.
- **În `train`:**
  - Numărul de straturi: 2.
  - Dimensiunea fiecărui strat:
    - Primul strat ascuns: 64 neuroni.
    - Al doilea strat ascuns: 32 neuroni.
    - Stratul de ieșire: 2 neuroni.
  - Tipul funcțiilor de activare folosite:
    - Pentru straturile ascunse: ReLU (Rectified Linear Unit).
    - Pentru stratul de ieșire: Sigmoid.

Configurarea Optimizatorului

- **În `my_train`:**
  - Tip de optimizator: Stochastic Gradient Descent (SGD).
  - Learning rate: 0.005.
  - Număr de epoci de antrenare: 20.
  - Dimensiunea batch-urilor de antrenare: 32.
- **În `train`:**
  - Algoritmul de optimizare: Adam.
  - Learning rate: 0.001.
  - Numărul maxim de iterații: 1000.
  - Dimensiunea batch-urilor de antrenare: 32.

Metode de Regularizare Folosite

- **În** `my_train`**:**
  - Utilizare early stopping (încercare manuală).
- **În** `train`**:**
  - Metode de regularizare folosite: Nu s-a menționat utilizarea unei metode de regularizare explicite (alpha=0.0001 implică regularizare L2).

Matricea de confuzie

Setul de date: SalaryPrediction

train-LR

	Predicted Negative	Predicted Positive
Actual Negative	4936	1142
Actual Positive	751	5327

train-MLP

	Predicted Negative	Predicted Positive
Actual Negative	5666	412
Actual Positive	354	5724

my_train-LR

	Predicted Negative	Predicted Positive
Actual Negative	1728	4350
Actual Positive	1945	4133

my_train-MLP

	Predicted Negative	Predicted Positive
Actual Negative	3258	2820
Actual Positive	124	5954

Setul de date: AVC

train-LR

	Predicted Negative	Predicted Positive
Actual Negative	3099	815
Actual Positive	615	3299

train-MLP

	Predicted Negative	Predicted Positive
Actual Negative	3768	146
Actual Positive	10	3904

my_train-LR

	Predicted Negative	Predicted Positive
Actual Negative	2325	1589
Actual Positive	1382	2532

my_train-MLP

	Predicted Negative	Predicted Positive
Actual Negative	1761	2153
Actual Positive	60	3854

Setul de date: SalaryPrediction

train-LR (Test)

	Predicted Negative	Predicted Positive
Actual Negative	1225	288
Actual Positive	128	359

train-MLP (Test)

	Predicted Negative	Predicted Positive
Actual Negative	1297	216
Actual Positive	190	297

my_train-LR (Test)

	Predicted Negative	Predicted Positive
Actual Negative	1089	1089
Actual Positive	323	323

my_train-MLP (Test)

	Predicted Negative	Predicted Positive
Actual Negative	824	689
Actual Positive	15	472

Setul de date: AVC

train-LR (Test)

	Predicted Negative	Predicted Positive
Actual Negative	775	172
Actual Positive	30	45

train-MLP (Test)

	Predicted Negative	Predicted Positive
Actual Negative	876	71
Actual Positive	59	16

my_train-LR (Test)

	Predicted Negative	Predicted Positive
Actual Negative	369	369
Actual Positive	48	48

my_train-MLP (Test)

	Predicted Negative	Predicted Positive
Actual Negative	437	510
Actual Positive	1	74

Tabel comparativ al algoritmilor

Setul de date: SalaryPrediction

Algoritm	**Acuratețe generală**	**Precizie (Clasa 0)**	**Recall (Clasa 0)**	**F1 (Clasa 0)**
train-LR	0.844	0.87	0.81	0.84
train-MLP	0.937	0.94	0.93	0.94
my_train-LR	0.3735	0.2287	0.6632	0.3401
my_train-MLP	0.648	0.4065	0.9692	0.5728

Table 1: Tabel comparativ pentru setul de date SalaryPrediction

Setul de date: AVC

Algoritm	**Acuratețe generală**	**Precizie (Clasa 0)**	**Recall (Clasa 0)**	**F1 (Clasa 0)**
train-LR	0.802	0.96	0.82	0.88
train-MLP	0.873	0.94	0.93	0.93
my_train-LR	0.6125	0.1151	0.64	0.1951
my_train-MLP	0.5	0.1267	0.9867	0.2246

Table 2: Tabel comparativ pentru setul de date AVC

Observăm că algoritmul Multilayer Perceptron (MLP), indiferent dacă este implementat de la zero sau utilizând biblioteca scikit-learn, demonstrează o performanță remarcabilă pe ambele seturi de date. Acest lucru este evident în special în analiza metricilor de evaluare, precum F1-score, care evidențiază capacitatea MLP de a echilibra precizia și recall-ul, oferind astfel o imagine comprehensivă a performanței sale. Este notabil faptul că MLP reușește să obțină rezultate solide și consistente, sugerând adaptabilitatea și robustețea sa în fața diversității de seturi de date și de contexte de utilizare.