Tema 2 IA ML Aplicat Alexandru LICURICEANU 332CD

■ credit_risk ■ SalaryPrediction

3.1.1. Analiza tipului de atribute si a plajei de valori a acestora:

Pentru setul de date “credit_risk”:

- Numarul de atribute numerice si categorice care nu au valori lipsa, numarul de valori unice ale atributelor categorice. Per total, din acest set de date nu lipsesc foarte multe valori:

```
loan_rate                       9060
loan_amount                    10000
job_tenure_years                9736
credit_history_length_years    10000
applicant_age                  10000
applicant_income               10000
loan_income_ratio              10000
credit_history_length_months   10000
dtype: int64
```

Numerice

```
residential_status             10000
loan_purpose                   10000
loan_approval_status           10000
loan_rating                    10000
credit_history_default_status  10000
stability_rating               10000
dtype: int64
```

Categorice

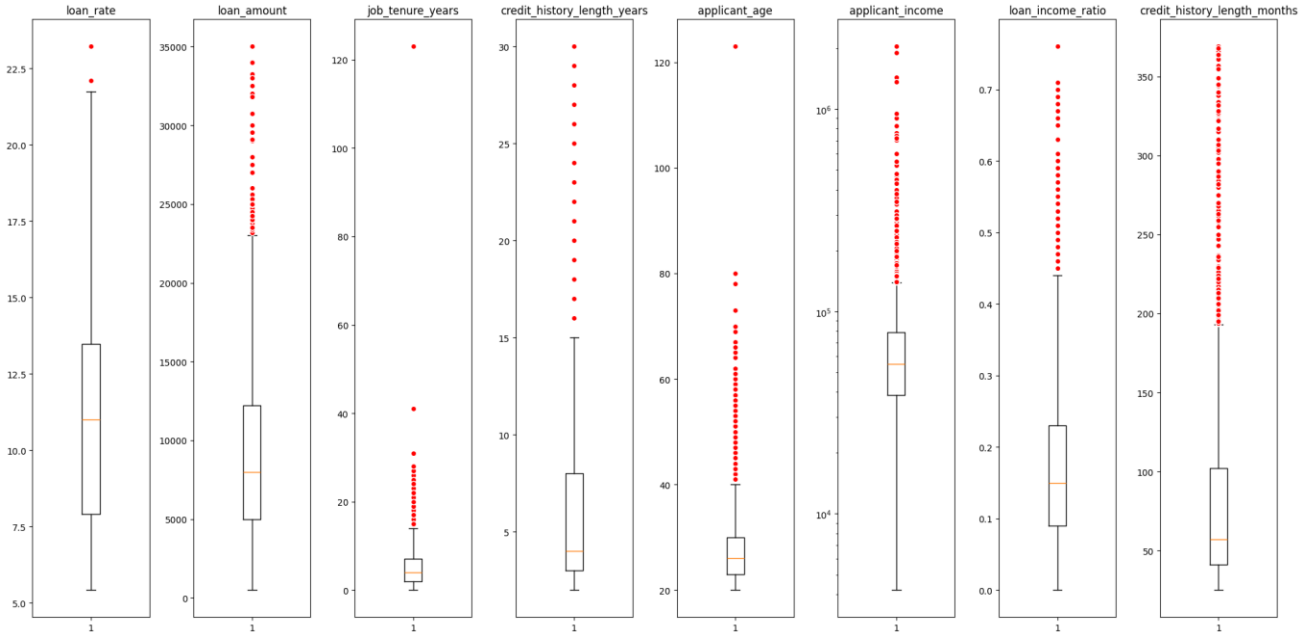
```
residential_status             4
loan_purpose                   6
loan_approval_status           2
loan_rating                    7
credit_history_default_status  2
stability_rating               4
dtype: int64
```

Nr. Valorilor unice

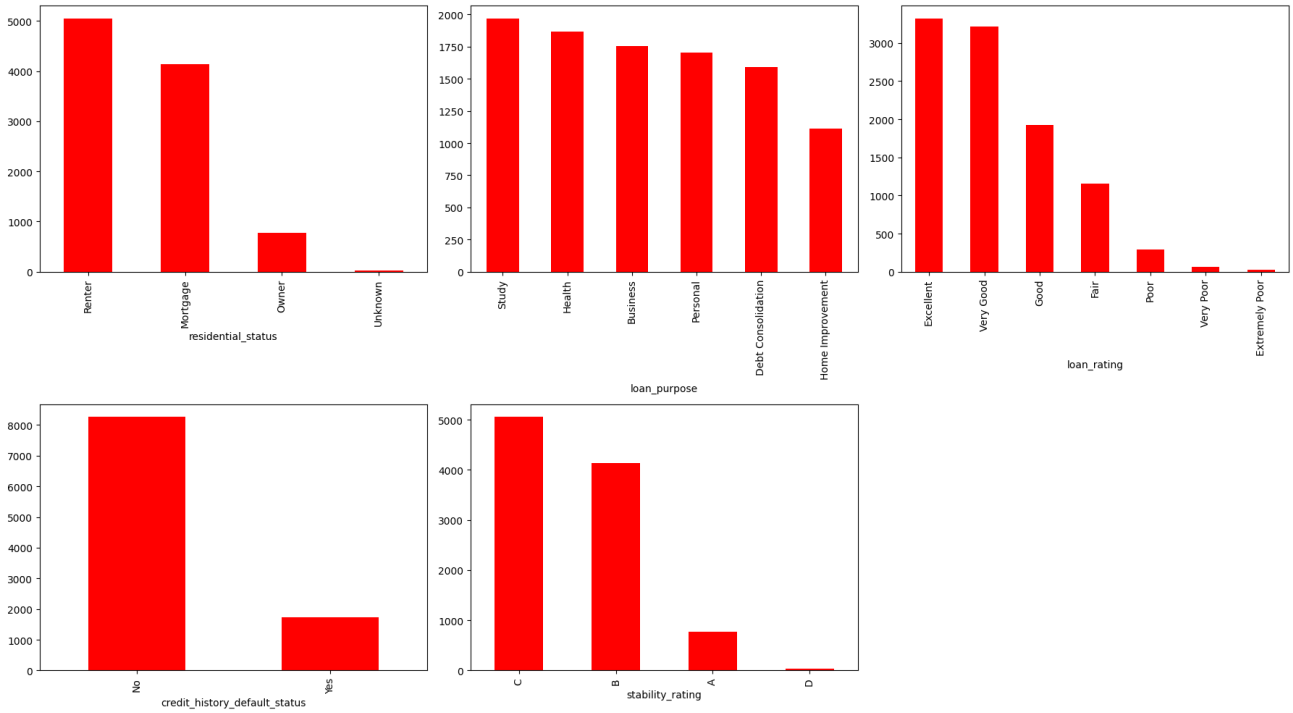
- Valorea medie, deviatia standard a valorilor, valoarea minima, valoarea percentilelor 25%, 50%, 75%, valoarea maxima pentru atributele numerice:

	loan_rate	loan_amount	job_tenure_years	credit_history_length_years	applicant_age	applicant_income	loan_income_ratio	credit_history_length_months
count	9060.000	10000.000	9736.000	10000.000	10000.000	10000.000	10000.000	10000.000
mean	11.007	9568.038	4.786	5.811	27.745	65734.211	0.170	75.761
std	3.266	6350.432	4.353	4.050	6.360	56944.387	0.107	48.677
min	5.420	500.000	0.000	2.000	20.000	4200.000	0.000	25.000
25%	7.900	5000.000	2.000	3.000	23.000	38595.000	0.090	41.000
50%	10.990	8000.000	4.000	4.000	26.000	55000.000	0.150	57.000
75%	13.470	12200.000	7.000	8.000	30.000	78997.000	0.230	102.000
max	23.220	35000.000	123.000	30.000	123.000	2039784.000	0.760	369.000

- Plajele de valori numerice variaza dramatic intre atribute din punct de vedere al valorilor efective, iar aproape toate atributele au foarte multe valori extreme.

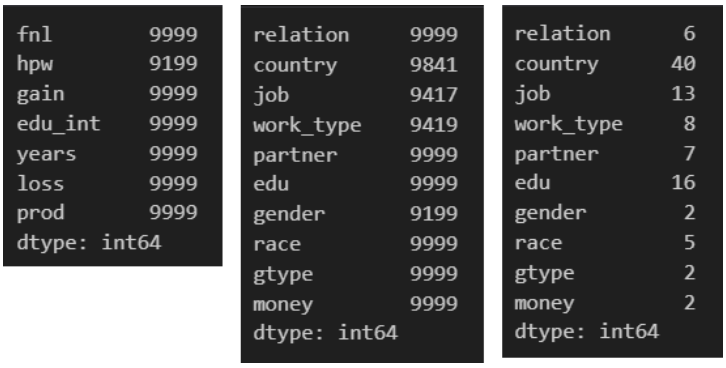


- Distributia atributelor categorice, aproape toate sunt dezechilibrate:



Pentru setul de date “SalaryPrediction”:

- Numarul de atribute numerice si categorice care nu au valori lipsa, numarul de valori unice ale atributelor categorice. Asemanator cu setul de date “credit_risk”, nu lipsesc foarte multe valori:



```
fnl         9999
hpw         9199
gain        9999
edu_int     9999
years       9999
loss        9999
prod        9999
dtype: int64
```

```
relation     9999
country      9841
job          9417
work_type    9419
partner      9999
edu          9999
gender       9199
race         9999
gtype        9999
money        9999
dtype: int64
```

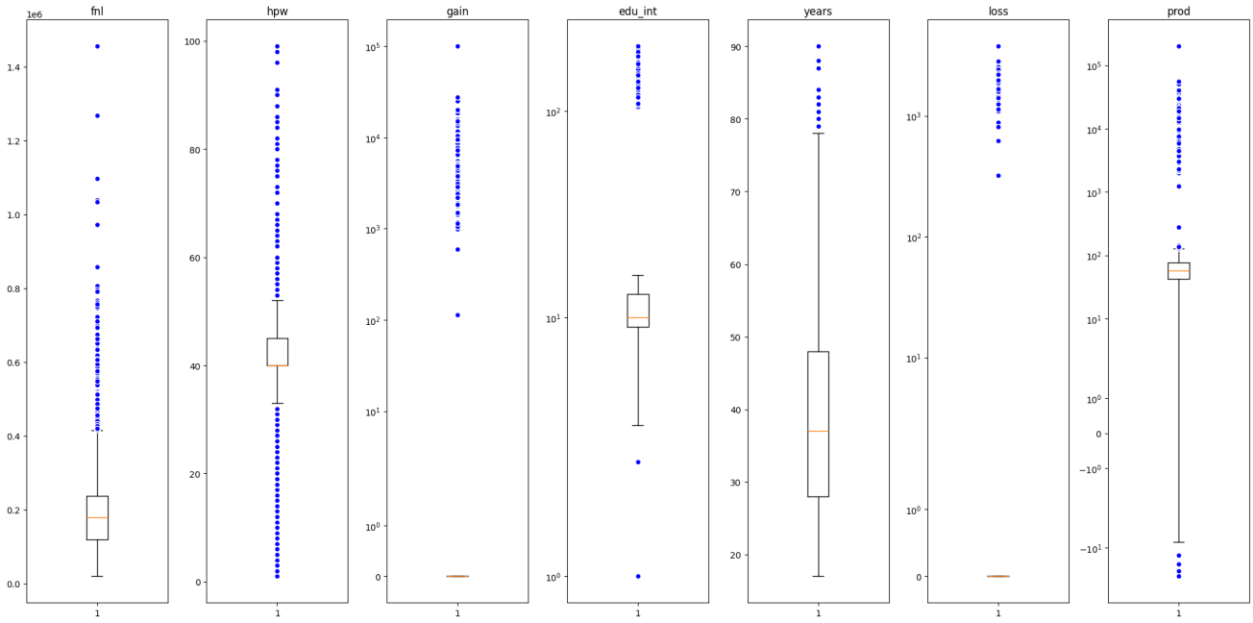
```
relation      6
country      40
job          13
work_type     8
partner       7
edu          16
gender        2
race          5
gtype         2
money         2
dtype: int64
```

Numerice Categorice Nr. Valorilor unice

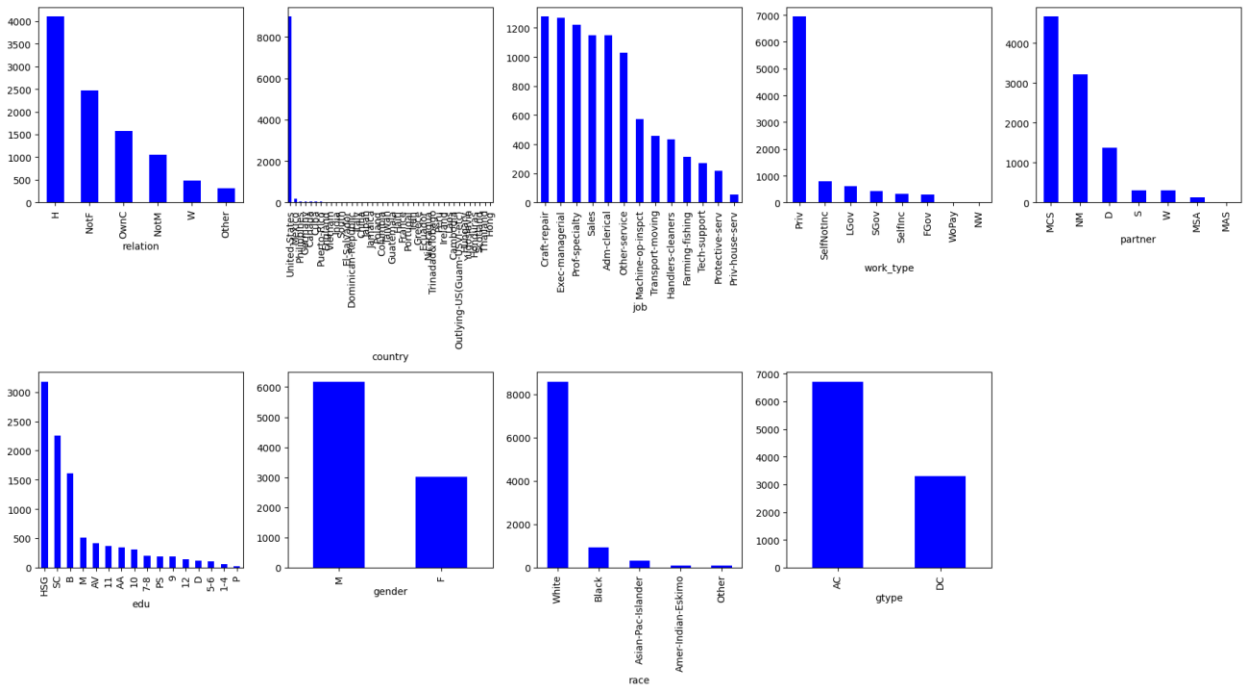
- Valorea medie, deviatia standard a valorilor, valoarea minima, valoarea percentilelor 25%, 50%, 75%, valoarea maxima pentru atributele numerice:

	fnl	hpw	gain	edu_int	years	loss	prod
count	9999.000	9199.000	9999.000	9999.000	9999.000	9999.000	9999.000
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	19214.000	1.000	0.000	1.000	17.000	0.000	-28.000
25%	118282.500	40.000	0.000	9.000	28.000	0.000	42.000
50%	178472.000	40.000	0.000	10.000	37.000	0.000	57.000
75%	237311.000	45.000	0.000	13.000	48.000	0.000	77.000
max	1455435.000	99.000	99999.000	206.000	90.000	3770.000	200125.000

- Din nou, avem plaje valori foarte diferite intre atributele numerice, multe valori extreme si de asemenea foarte multe valori de 0 pentru atributele “gain” si “loss”:

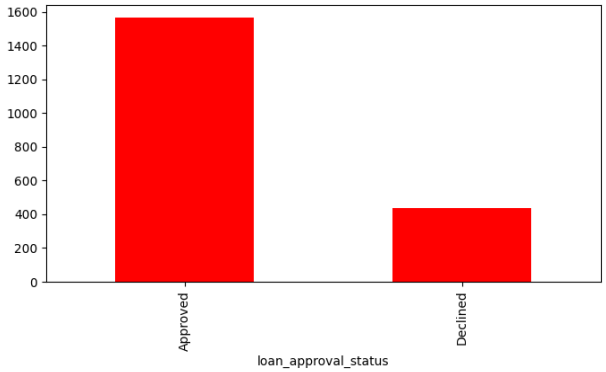


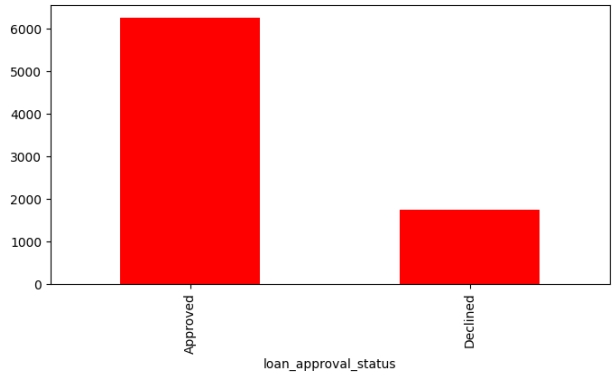
- Distributia atributelor categorice, toate sunt dezechilibrate:



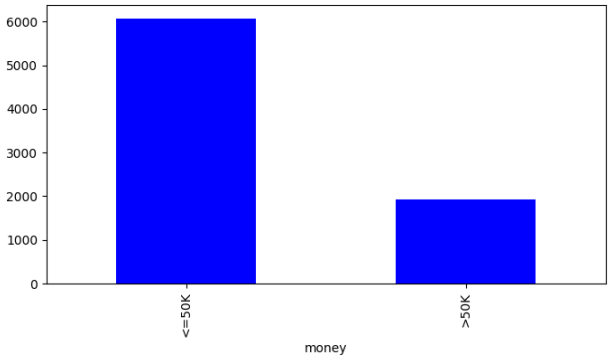
3.1.2. Analiza echilibrului de clase. Stanga train, dreapta test.

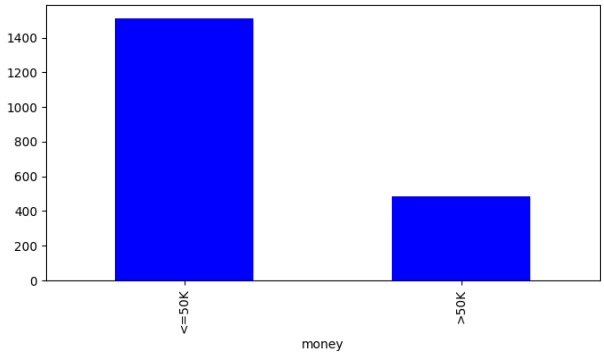
Pentru setul de date “credit_risk”:





Pentru setul de date “SalaryPrediction”:

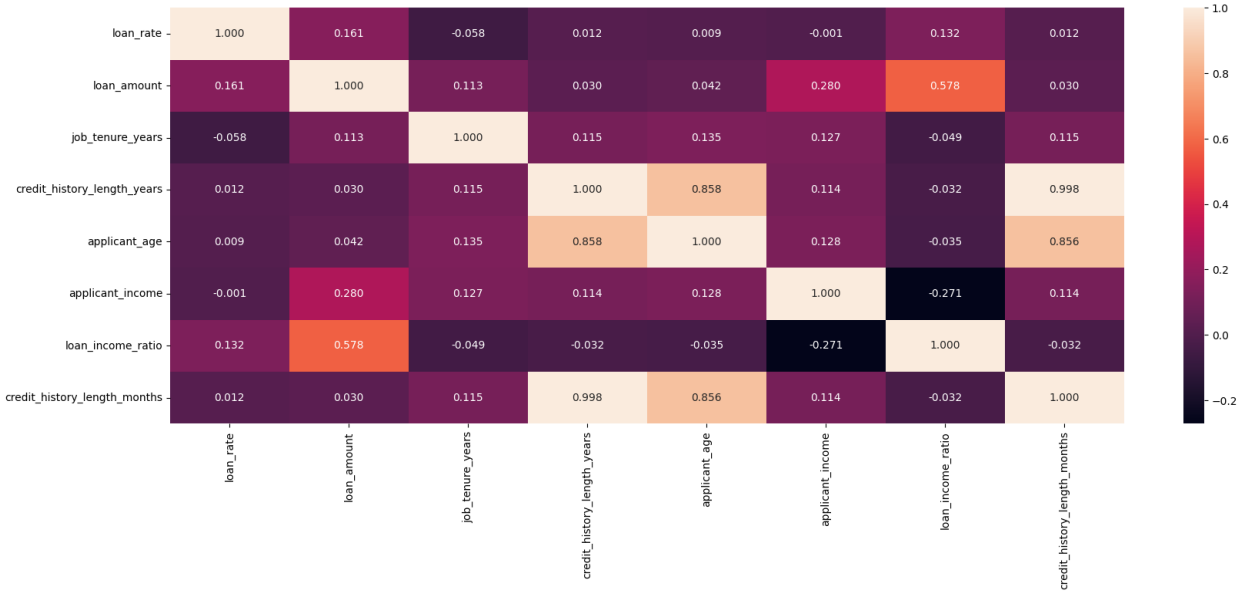




Atat pentru “credit_risk”, cat si pentru “SalaryPrediction”, numarul exemplelor clasei “Approved” respectiv “<=50K” este de aproximativ 3 ori mai mare decat numarul exemplelor pentru “Declined” si “>50K”, facand clasificarea mai dificila si crescand probabilitatea de clasificari eronate. Deoarece seturile de train si test sunt oarecum proportionale, in cel mai probabil caz clasificatorul va reusi sa obtina acuratete ridicata, insa modelul mai mult ca sigur va face overfitting.

3.1.3. Analiza corelatiei intre atribute.

Pentru setul de date “credit_risk”:



Numerice

Categorice

Pentru variabilele numerice, sunt cateva care au valori de corelatie intre 0.5/0.7 si 1, ceea ce ar indica o corelatie puternica intre ele, insa majoritatea sunt sub valoarea de 0.5.

Pentru cele categorice, valorile de 0 rezultate din testul Chi-Squared indica faptul ca ipoteza ca atributele sunt independente poate fi respinsa.

Pentru setul de date “SalaryPrediction”:

Numerice

Categorice

Spre deosebire de “credit_risk”, aici majoritatea variabilelor numerice au valori de corelatie care se apropie de 0, ceea ce indica independenta intre variabile.

Pentru cele categorice, testul Chi-Squared din nou indica faptul ca ipoteza nula poate fi respinsa, exceptand atributele “work_type” si “country”.

In final, am ales sa nu renunt la atribute numerice sau la cele categorice deoarece nu am reusit sa obtin nici timp mai bun de executie, nici metrici mai bune, chiar din contra, daca renunt la unele atribute, modelul ofera mai multe clasificari gresite.

3.3.1. Arbore de decizie.

Hiperparametrii:

Versiune	Dataset	Max_depth	Min_samples_leaf	Criterion	Class_weight
sklearn	Credit_risk	100	20	entropy	-
	SalaryPrediction	40	20	entropy	>50K: 0.6 <=50K: 0.4
Laborator	Credit_risk	100	40	entropy	-
	SalaryPrediction	100	60	entropy	-

Pentru arborii de decizie, nu am folosit hiperparametrul de “class_weight” la “credit_risk” deoarece nu oferea rezultate mai bune. La versiunea de la laborator nu am implementat aceasta functionalitate.

3.3.2. Multi-Layer Perceptron.

Hiperparametrii:

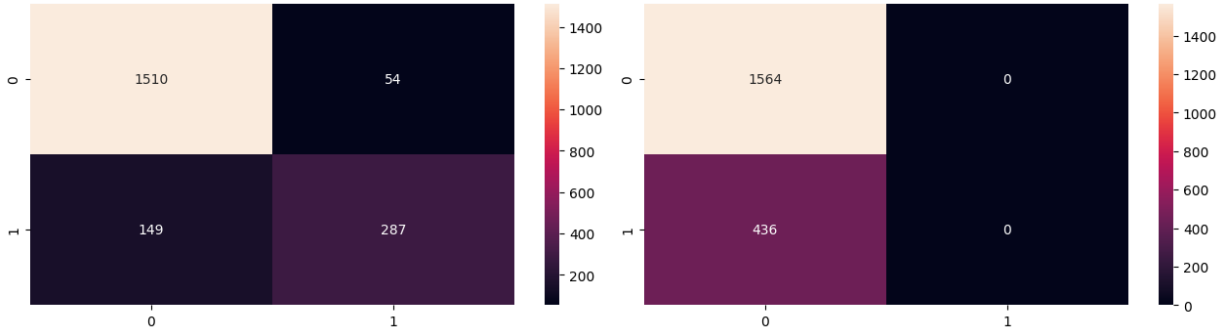
Versiune	Dataset	Hidden_layer_sizes	activation	solver	alpha
sklearn	Credit_risk	(100, 150)	ReLU	adam	0.0008
	SalaryPrediction	(784, 128)	ReLU	adam	0.0004
Laborator	Credit_risk	(10, 100)	ReLU	SGD	0
	SalaryPrediction	(10, 100)	ReLU	SGD	0

Versiune	Dataset	Learning_rate	Max_iter	Batch_size	Early_stopping
sklearn	Credit_risk	0.04	5000	100	True
	SalaryPrediction	0.005	20	128	True
Laborator	Credit_risk	0.005	-	128	False
	SalaryPrediction	0.005	-	128	False

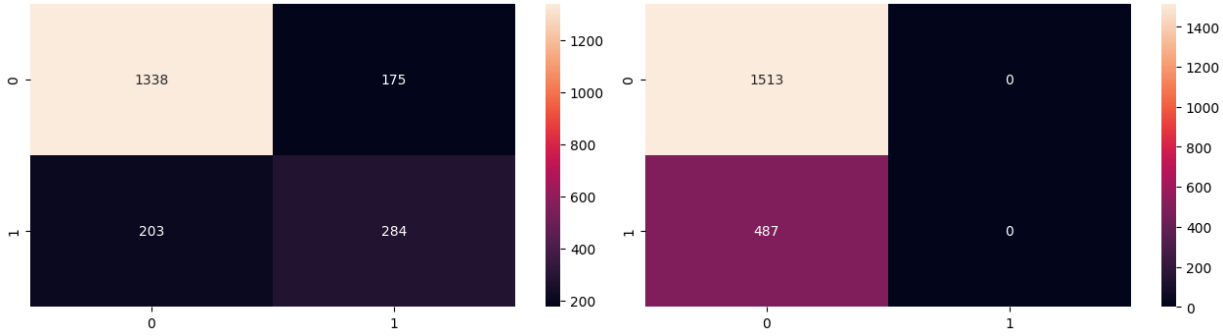
Stanga sklearn, dreapta laborator.

3.3.3. Evaluarea algoritmilor - arbore de decizie:

Pentru dataset-ul “credit_risk”:

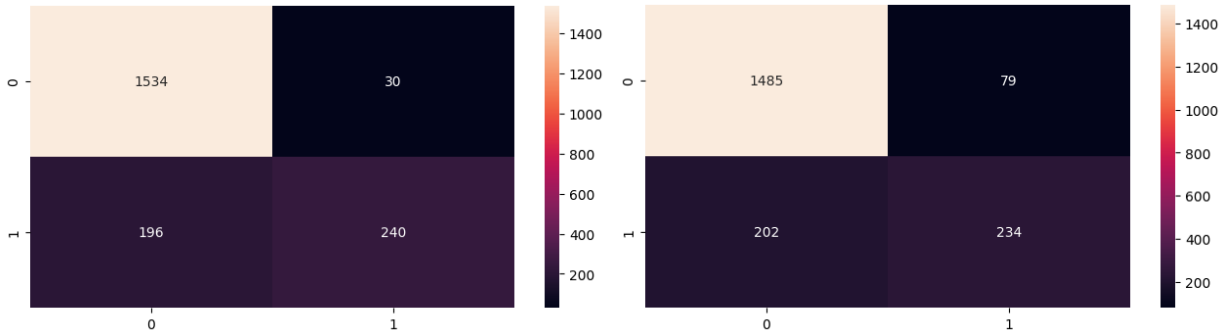


Pentru dataset-ul “SalaryPrediction”:

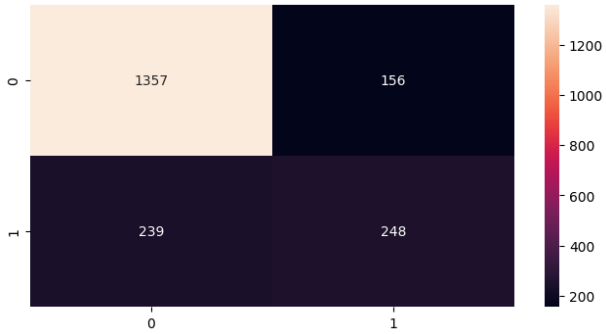


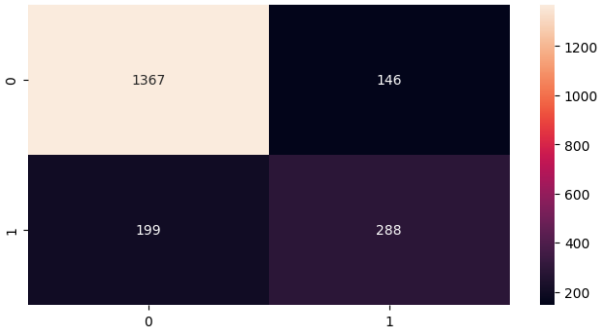
3.3.4. Evaluare algoritmilor – Multi-Layer Perceptron:

Pentru dataset-ul “credit_risk”:



Pentru dataset-ul “SalaryPrediction”:





Evaluarea metricilor - “credit_risk”:

Versiune	Algoritm	Clasa	Precision	Recall	F1-score	Accuracy
sklearn	Arbore de decizie	Approved	0.91	0.97	0.94	0.90
		Declined	0.84	0.66	0.74	
	MLP	Approved	0.89	0.98	0.93	0.89
		Declined	0.89	0.55	0.68	
Laborator	Arbore de decizie	Approved	0.78	1.00	0.88	0.78
		Declined	0.00	0.00	0.00	
	MLP	Approved	0.88	0.95	0.91	0.86
		Declined	0.75	0.54	0.62	

Evaluarea metricilor - “SalaryPrediction”:

Versiune	Algoritm	Clasa	Precision	Recall	F1-score	Accuracy
sklearn	Arbore de decizie	<=50K	0.87	0.88	0.88	0.81
		>50K	0.62	0.58	0.60	
	MLP	<=50K	0.87	0.90	0.89	0.83
		>50K	0.66	0.59	0.63	
Laborator	Arbore de decizie	<=50K	0.76	1.00	0.86	0.76
		>50K	0.00	0.00	0.00	
	MLP	<=50K	0.85	0.90	0.87	0.80
		>50K	0.61	0.51	0.56	

Parerea mea este ca algoritmii din biblioteca sklearn produc rezultate mai bune deoarece cu siguranta sunt mai bine implementate decat versiunile din laboratoare ale algoritmilor. De asemenea, versiunile din sklearn folosesc mai multi hiperparametrii care ar putea oferi rezultate mai bune, iar diferenta intre metrici dintre arbore si MLP este foarte mica pe acelasi

set de date. Chiar si versiunea de la laborator de MLP se apropie de cea din sklearn, iar seturile de date cu care am lucrat, mai mult ca sigur au influentat clasificatorii, fiind dezechilibrate.