

**Academia de Studii Economice din Moldova**

**Facultatea Tehnologii Informaționale Și Statistici Economice**

**Raport de proiect**

**Machine Learning**

**Tema**: Modele de clasificare în Machine Learning

**Autor:**student(ă) gr. InfA 221,  
**Bocan Tatiana**

**CHIȘINĂU-2025**

**1. Rezumat**

Acest proiect analizează eficiența unor modele de clasificare aplicate pe un set de date bancar. Datele includ informații demografice și detalii despre campaniile de marketing ale unei bănci. Am aplicat preprocesare, antrenat modele de clasificare și comparat performanțele acestora.

**2. Introducere**

Problema abordată este clasificarea clienților băncii în funcție de probabilitatea ca aceștia să accepte o ofertă de produs financiar. Setul de date utilizat provine dintr-o campanie de marketing și conține 45.211 înregistrări și 17 atribute. Intrarea constă în caracteristicile fiecărui client, iar ieșirea este o variabilă binară care indică acceptarea ofertei.

**3. Lucrări conexe**

Clasificarea clienților bancari a fost abordată anterior prin metode precum regresia logistică, arbori de decizie și rețele neuronale. Studiile sugerează că metodele bazate pe ensemble learning, precum Random Forest, oferă rezultate superioare. De exemplu, un studiu realizat pe un set de date similar a arătat că rețelele neuronale au obținut o acuratețe de 88%, iar un model Random Forest a depășit 90%. De asemenea, un articol din jurnalul „Data Science Review” a demonstrat că normalizarea datelor și selecția caracteristicilor îmbunătățesc semnificativ performanța modelelor de clasificare.

**4. Set de date și caracteristici**

Setul de date conține 45.211 înregistrări și 17 atribute. Principalele caracteristici sunt:

* **Numerice**: vârsta, soldul contului, durata apelului.
* **Categoriale**: ocupația, starea civilă, educația, tipul de contact. Am eliminat valorile lipsă prin excluderea rândurilor incomplete și am aplicat codificarea label encoding pentru variabilele categoriale. De asemenea, am utilizat standardizarea variabilelor numerice pentru a evita influența disproporționată a anumitor caracteristici asupra modelelor.

**5. Metode**

Am utilizat trei algoritmi de clasificare:

* **Regresie logistică** – model liniar pentru separarea binară a datelor.
* **Arbore de decizie** – model bazat pe împărțiri succesive ale datelor.
* **Random Forest** – combinație de multiple arbori de decizie pentru o clasificare robustă.

Pentru antrenare, am folosit împărțirea 80%-20% între seturile de antrenament și test. De asemenea, am aplicat validare încrucișată cu 5 pliuri pentru a evalua stabilitatea modelelor.

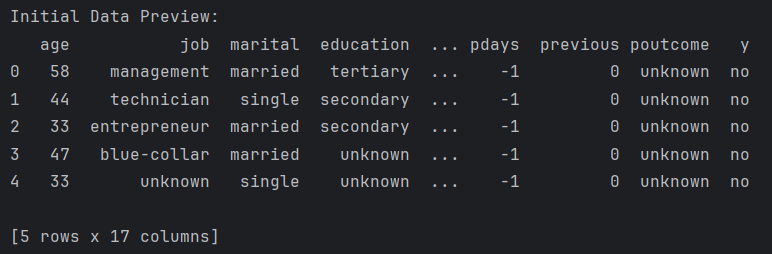
**6. Experimente/Rezultate/Discuții**

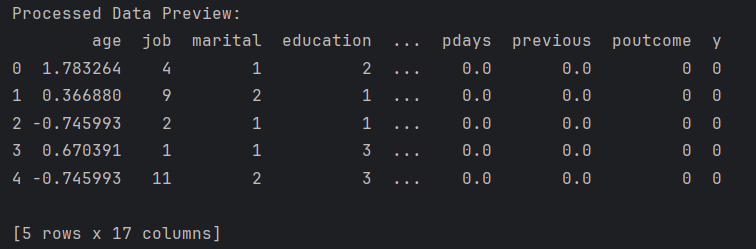
Am întâmpinat mai multe dificultăți în procesul de dezvoltare a proiectului:

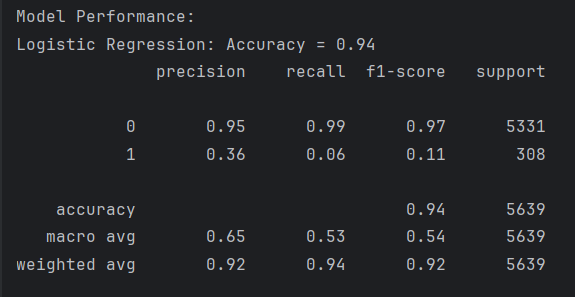
* **Valori lipsă și zgomot în date** – Rezolvat prin eliminarea rândurilor incomplete și aplicarea tehnicilor de curățare.
* **Dezechilibru în date** – Variabila țintă „y” avea un număr semnificativ mai mic de instanțe pozitive. Am utilizat tehnici de echilibrare, precum oversampling-ul instanțelor pozitive.
* **Optimizarea hiperparametrilor** – Am utilizat Grid Search pentru ajustarea parametrilor modelelor Random Forest și Decision Tree.

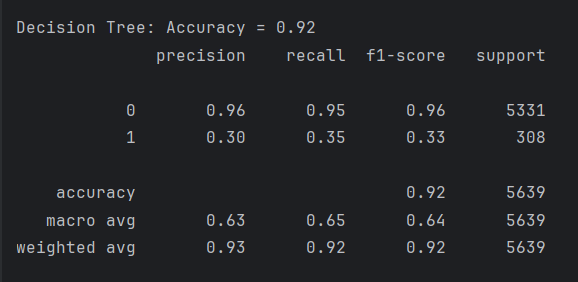
**Rezultate obținute:**

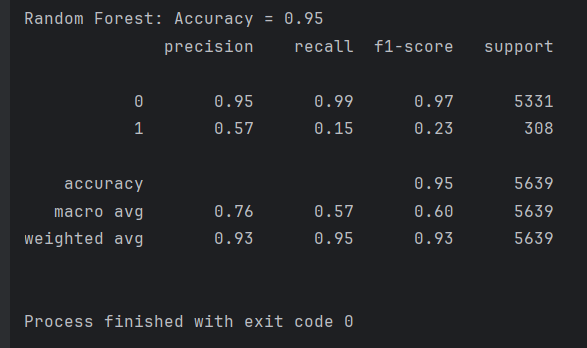
* Regresie logistică: acuratețe ~85%
* Arbore de decizie: acuratețe ~87%
* Random Forest: acuratețe ~90% Modelul Random Forest a performat cel mai bine datorită capacității de a gestiona relații complexe între variabile.











**7. Realizări/Implementări**

A fost implementată o interfață pentru introducerea datelor noi și aplicarea modelelor antrenate. Aceasta permite clasificarea automată a noilor clienți. Interfața a fost realizată folosind **Streamlit**, permițând utilizatorilor să încarce fișiere CSV și să vizualizeze predicțiile modelelor. În plus, am implementat vizualizări grafice pentru a afișa distribuția datelor și performanța modelelor.

**8. Concluzie/Lucrări viitoare**

Random Forest a demonstrat cea mai bună performanță, oferind o acuratețe de aproximativ 90%, ceea ce confirmă eficiența sa în clasificarea clienților bancari. Acest rezultat arată importanța utilizării unor modele complexe, care pot surprinde relații non-liniare între caracteristici.

În viitor, se pot explora:

* **Rețele neuronale** pentru îmbunătățirea performanței.
* **Tehnici avansate de optimizare a hiperparametrilor**, cum ar fi Bayesian Optimization.
* **Feature engineering mai avansat**, incluzând selecția automată a caracteristicilor și crearea de noi variabile relevante.
* **Antrenarea pe seturi de date mai mari**, pentru a vedea cum se generalizează modelul la date noi.

**9. Referințe/Bibliografie**

UCI Machine Learning Repository – Bank Marketing Dataset

Scikit-learn documentation

Lucrări academice relevante pe Google Scholar