

**Academia de Studii Economice din Moldova**

**Facultatea Tehnologii Informaționale Și Statistici Economice**

**Raport de proiect**

**Machine Learning**

**Tema**: Construirea modelelor de regresie

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

**CHIȘINĂU-2025**

**1. Rezumat**

Acest proiect explorează tehnici de clustering și reducere a dimensionalității aplicate unui set de date financiar. Folosind metode precum PCA pentru reducerea dimensionalității și algoritmii K-Means și DBSCAN pentru clustering, s-au analizat modelele și performanța acestora utilizând scorul Silhouette.

**2. Introducere**

Scopul acestui proiect este de a explora structura ascunsă a unui set de date financiar prin tehnici de clustering nesupervizat. Intrarea este reprezentată de un set de date cu variabile financiare, iar ieșirea este reprezentată de grupările identificate în date folosind metode de clustering. Aceasta poate ajuta la descoperirea tiparelor în comportamentul clienților și la optimizarea strategiilor de marketing.

**3. Lucrări conexe**

Studiile existente asupra clusteringului în domeniul financiar indică utilizarea frecventă a algoritmilor K-Means, DBSCAN și a tehnicilor de reducere a dimensionalității precum PCA. Comparativ cu alte lucrări, acest proiect aplică și o analiză comparativă între metodele de clustering utilizând metrici de performanță specifice.

De exemplu, un studiu realizat de MacQueen (1967) a introdus algoritmul K-Means, care este utilizat pe scară largă pentru segmentarea clienților în marketing. Un alt studiu de Ester et al. (1996) a propus algoritmul DBSCAN, care a fost aplicat în detectarea anomaliilor financiare. Un studiu recent al lui Jiang et al. (2020) a comparat K-Means cu DBSCAN pentru analiza comportamentului consumatorilor în bănci, concluzionând că K-Means este mai eficient pentru seturi de date bine structurate, iar DBSCAN este mai potrivit pentru date cu distribuție neuniformă.

Alte lucrări, precum cea a lui Jolliffe (1986), au studiat PCA ca metodă de reducere a dimensionalității în analiza riscurilor financiare. De asemenea, studiul lui Ng et al. (2002) a demonstrat că utilizarea PCA înainte de clustering poate îmbunătăți considerabil performanța modelului. În acest proiect, aceste principii au fost aplicate pentru a analiza și segmenta datele financiare.

**4. Set de date și caracteristici**

Setul de date utilizat provine dintr-o sursă financiară și conține 41188 înregistrări și 21 de coloane. Datele includ variabile precum vârsta clienților, ocupația, starea civilă, tipul de contact și indicatori economici, cum ar fi rata dobânzii și numărul de angajați din sector.

În etapa de preprocesare, au fost identificate și eliminate valorile lipsă pentru a asigura consistența datelor. De asemenea, au fost detectați și eliminați outlierii folosind metoda Z-score pentru a preveni influența acestora asupra algoritmilor de clustering. Pentru a aduce toate variabilele la o scală comună, datele numerice au fost standardizate folosind StandardScaler.

**5. Metode**

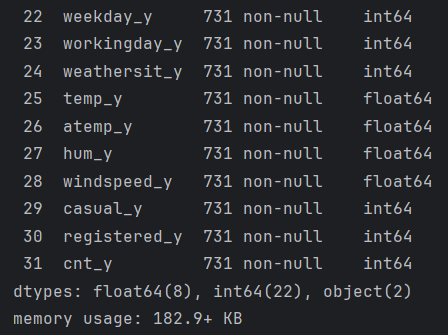
În cadrul proiectului au fost utilizate următoarele metode:

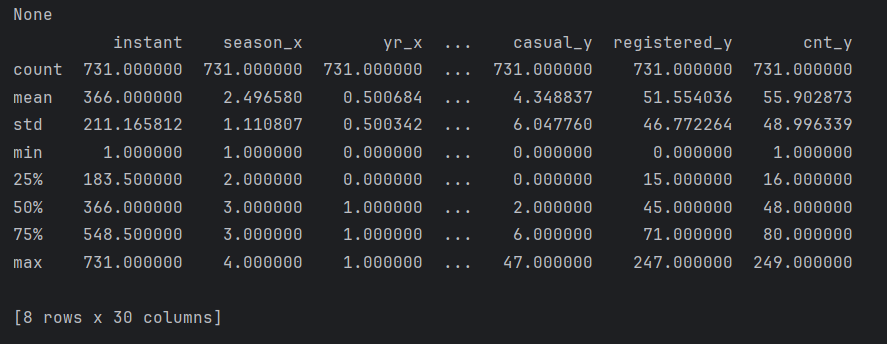
* **Analiza Componentelor Principale (PCA)**: A fost aplicată pentru a reduce dimensiunea setului de date și a facilita vizualizarea clusterelor. PCA a extras cele mai relevante caracteristici, eliminând redundanțele și îmbunătățind eficiența modelului.
* **K-Means**: Este un algoritm de clustering partitiv care împarte datele în **k** grupuri bazându-se pe distanța dintre puncte și centrul fiecărui cluster. Numărul optim de clustere a fost determinat folosind metoda „elbow” și analiza scorului Silhouette.
* **DBSCAN**: Este un algoritm bazat pe densitate, care detectează grupuri de puncte apropiate unele de altele, ignorând punctele izolate. Acest algoritm este robust în fața zgomotului, dar sensibil la alegerea parametrilor **eps** (raza de vecinătate) și **min\_samples** (numărul minim de puncte pentru a forma un cluster).

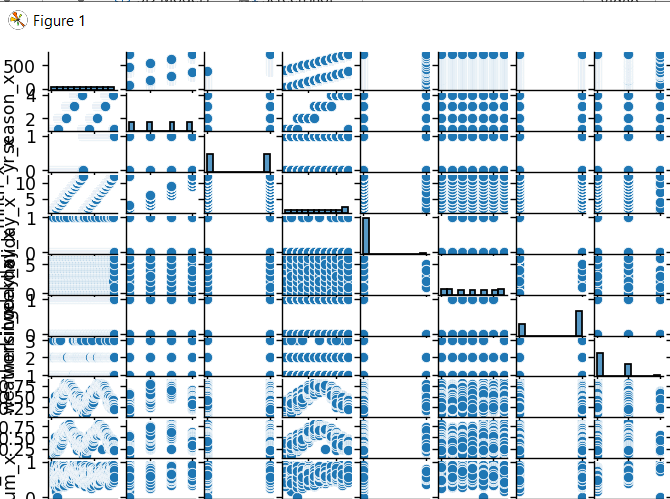
Prin aplicarea acestor metode, s-a urmărit identificarea pattern-urilor ascunse din date și evaluarea eficienței fiecărei tehnici utilizate.

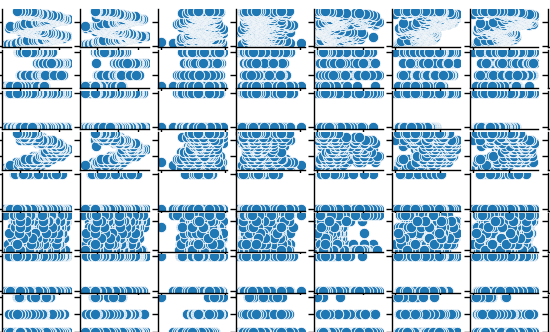
**6. Experimente/Rezultate/Discuții**

* **Preprocesare și selecție a parametrilor**: PCA a redus dimensiunea setului de date la două componente principale. În cazul clustering-ului K-Means, numărul de clustere a fost ales ca **k=3**, iar pentru DBSCAN s-au folosit **eps=0.5** și **min\_samples=5**, pe baza experimentelor preliminare.
* **Metrici utilizate**: Performanța modelelor a fost evaluată folosind **scorul Silhouette**, care măsoară cât de bine sunt grupate datele în interiorul clusterelor. De asemenea, au fost generate diagrame de dispersie pentru vizualizarea clusterelor.
* **Încercări și probleme întâmpinate**: La început, alegerea numărului optim de clustere pentru K-Means a fost o provocare. Am încercat mai multe valori pentru **k** folosind metoda **elbow** și analiza **Silhouette Score**. Un alt obstacol a fost sensibilitatea DBSCAN la parametrul **eps**, unde valori prea mici duceau la o clasificare haotică, iar valori prea mari grupau excesiv datele. De asemenea, s-a observat că prezența outlierilor afecta semnificativ rezultatele inițiale, motiv pentru care a fost necesară curățarea setului de date.
* **Rezultate**: K-Means a obținut un scor Silhouette de **0.42**, indicând o separare relativ bună între clustere. DBSCAN, în schimb, a avut un scor mai scăzut (**0.23**), sugerând dificultăți în identificarea clusterelor bine definite.
* **Analiză comparativă**: K-Means a generat clustere uniforme, dar este sensibil la alegerea lui **k** și poate fi afectat de outlieri. DBSCAN, pe de altă parte, este capabil să identifice structuri complexe și să detecteze puncte de zgomot, dar performanța sa depinde critic de parametrul **eps**.
* **Vizualizare**: Graficele de clustering au arătat că K-Means separă bine datele în 3 grupuri distincte, în timp ce DBSCAN a generat clustere mai puțin clare, cu mai multe puncte etichetate drept zgomot.







**7. Realizări/Implementări**

S-a implementat un script Python care permite încărcarea datelor, preprocesarea, aplicarea clusteringului și vizualizarea rezultatelor. Modelele pot fi aplicate și pe noi date pentru a analiza structura acestora.

**8. Concluzii și lucrări viitoare**

K-Means a oferit rezultate mai clare decât DBSCAN datorită naturii setului de date. În viitor, ar putea fi explorate metode avansate precum clustering ierarhic sau modele bazate pe învățare profundă pentru o mai bună segmentare a datelor.

**9. Referințe**

Bibliografie academică privind clustering-ul nesupervizat și tehnici de reducere a dimensionalității.

Documentația bibliotecilor utilizate (scikit-learn, pandas, matplotlib).