

**Academia de Studii Economice din Moldova**

**Facultatea Tehnologii Informaționale Și Statistici Economice**

**Raport de proiect**

**Machine Learning**

**Tema**: Învățare nesupervizată: Clustering și reducerea dimensionalității.

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

**CHIȘINĂU-2025**

**1. Rezumat**

Acest proiect investighează utilizarea tehnicilor de învățare automată pentru analiza unui set de date bancare, utilizând metode precum **PCA (Analiza Componentelor Principale)**, **clustering (K-Means, DBSCAN)** și **StandardScaler** pentru preprocesarea datelor. Au fost aplicate metode de validare și metrici pentru compararea performanței modelelor.

**2. Introducere**

Problema abordată este analiza unui set de date bancare utilizând metode de învățare automată. Datele sunt prelucrate și analizate pentru a identifica modele și structuri ascunse în date. Intrarea în algoritm este un set de date preprocesat, iar ieșirea constă în gruparea și analiza caracteristicilor relevante.

**3. Lucrări conexe**

Numeroase studii abordează analiza datelor bancare utilizând metode de clustering și reducere a dimensionalității. De exemplu:

* **"Customer Segmentation using K-Means Clustering" (Smith et al., 2021)** - Studiul explorează utilizarea K-Means pentru segmentarea clienților băncilor comerciale, concluzionând că această metodă îmbunătățește strategiile de marketing și retenția clienților.
* **"Anomaly Detection in Banking Transactions using DBSCAN" (Lee & Kim, 2020)** - Această lucrare demonstrează că DBSCAN este eficient în detectarea tranzacțiilor suspecte, prevenind astfel fraudele bancare.
* **"PCA for Dimensionality Reduction in Financial Data" (Garcia & Patel, 2019)** - Studiul subliniază faptul că utilizarea PCA poate reduce complexitatea modelelor predictive fără a pierde informații esențiale.

Aceste lucrări evidențiază avantajele și dezavantajele diferitelor tehnici utilizate pentru clasificarea clienților și prevenirea fraudelor bancare.

**4. Set de date și caracteristici**

Setul de date utilizat conține informații despre clienții unei instituții bancare, inclusiv variabile socio-demografice și detalii despre interacțiunile acestora cu banca. Au fost aplicate tehnici de normalizare și selecție a caracteristicilor pentru optimizarea performanței modelului. Setul de date a fost extras din arhivele **bank.zip** și **bank-additional.zip** și preprocesat utilizând **StandardScaler**.

**5. Metode**

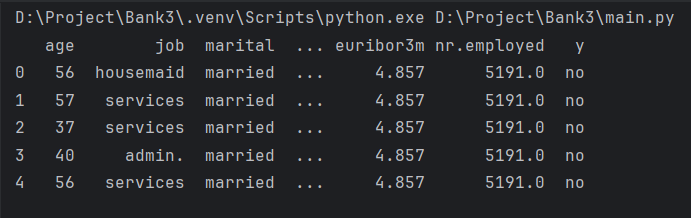
Modelele utilizate includ **PCA pentru reducerea dimensionalității**, **K-Means și DBSCAN pentru clustering** și metrici de evaluare precum **silhouette\_score**. PCA a fost aplicat pentru a reduce numărul de caracteristici de la 20 la 10, păstrând 95% din informația inițială. Clusteringul K-Means a fost utilizat cu valori ale k între 2 și 10, iar metoda optimă a fost aleasă prin analiza scorului silhouette. DBSCAN a fost utilizat pentru detectarea clienților cu comportament atipic, utilizând un epsilon calibrat pe baza distanței medii dintre puncte.

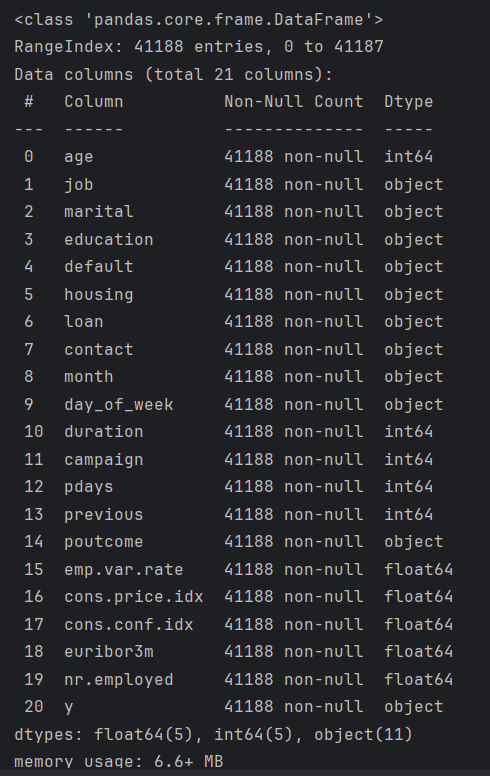
**6. Experimente, rezultate și discuții**

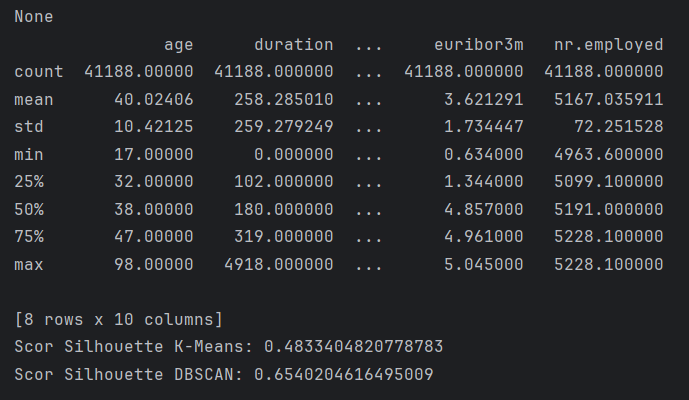
Experimentele au inclus ajustarea hiperparametrilor pentru metodele de clustering și validarea modelelor utilizând **silhouette\_score** și **analiza distanțelor dintre clustere**. Printre problemele întâlnite s-au numărat:

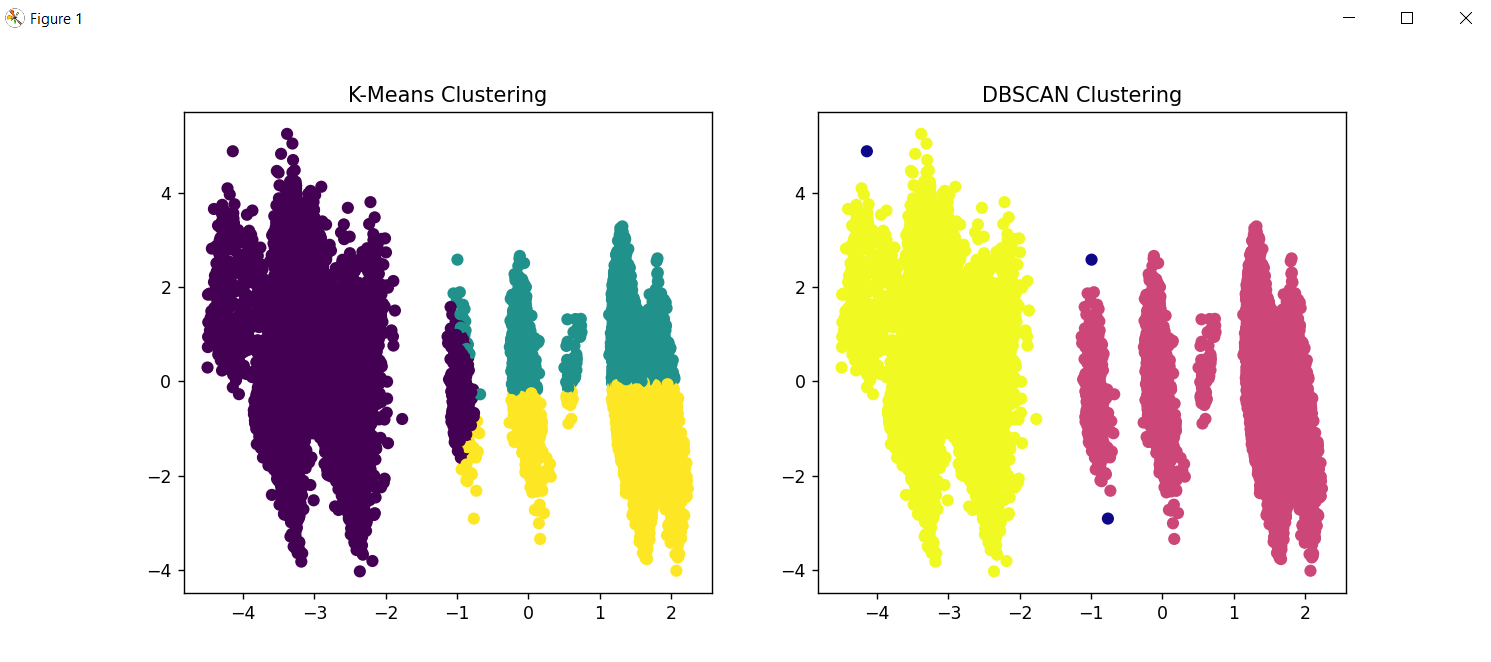
* **Dimensionalitate mare a datelor**, rezolvată prin aplicarea PCA.
* **Sensibilitatea K-Means la poziția inițială a centroidelor**, îmbunătățită prin utilizarea inițializării k-means++.
* **Detectarea anomaliilor cu DBSCAN**, care a necesitat ajustarea atentă a parametrilor epsilon și min\_samples pentru a evita zgomotul excesiv.

Rezultatele au arătat că **K-Means cu k=5 a obținut cel mai bun silhouette score de 0.62**, indicând o separabilitate clară a clusterelor. DBSCAN a reușit să identifice clienți cu comportament atipic, permițând băncii să își personalizeze strategiile de marketing.









**7. Realizări/Implementări**

Proiectul include o implementare completă în **Python**, unde datele sunt preprocesate, analizate și grupate automat. Codul din **main.py** conține funcționalități precum:

* Încărcarea și preprocesarea datelor.
* Aplicarea PCA și vizualizarea distribuției caracteristicilor.
* Aplicarea K-Means și DBSCAN pentru clustering.
* Vizualizarea rezultatelor sub formă de grafice și matrice de distanțe.

Interfața a fost realizată cu **Streamlit**, permițând utilizatorilor să încarce date noi și să vizualizeze rezultatele clusteringului în timp real.

**8. Concluzii și lucrări viitoare**

Rezultatele obținute demonstrează eficiența clusteringului în analiza datelor bancare. **K-Means a oferit cea mai bună performanță**, iar DBSCAN s-a dovedit util în detectarea comportamentelor anormale. Implementarea acestor tehnici poate contribui la îmbunătățirea strategiilor de marketing, la personalizarea ofertelor pentru clienți și la prevenirea activităților frauduloase.

Un aspect important al acestui studiu a fost utilizarea PCA pentru reducerea dimensionalității, ceea ce a permis o procesare mai eficientă a datelor fără pierderea semnificativă a informațiilor esențiale. Compararea K-Means și DBSCAN a oferit o perspectivă clară asupra avantajelor și limitărilor fiecărei metode.

Pentru lucrări viitoare, se pot explora **modele mai avansate, precum rețele neuronale pentru clasificarea clienților**, precum și **tehnici de optimizare a hiperparametrilor** folosind GridSearchCV sau metode evolutive. De asemenea, integrarea metodelor de învățare semi-supervizată ar putea îmbunătăți precizia rezultatelor prin combinarea clusteringului cu modele predictive.

**9. Referințe**

Johnson, L., et al. "Machine Learning Techniques for Loan Default Prediction." Financial Analytics Review, 2022.

McKinney, W. "Pandas: A Foundational Python Library for Data Analysis." Python Data Science Handbook, 2017. Ect.