# Raport Final

## Introducere

Acest proiect a avut ca scop crearea unui model de regresie pentru a prezice variabila „hours-per-week” folosind datele din setul de date Census. Procesul a inclus preprocesarea datelor, explorarea caracteristicilor, selectarea variabilelor relevante, implementarea mai multor modele de regresie, ajustarea hiperparametrilor și evaluarea performanțelor.   
  
Modelul final selectat este modelul de regresie cu arbore de decizie ajustat (Tuned Decision Tree Regression), care a oferit cele mai bune rezultate între modelele testate.

## Preprocesarea Datelor

1. Setul de date:  
- Variabila țintă: `hours-per-week`.  
- S-au utilizat date preprocesate, incluzând eliminarea valorilor lipsă, codificarea variabilelor categorice și standardizarea caracteristicilor numerice.  
  
2. Divizarea datelor:  
- Seturi: Train (70%), Validation (15%), Test (15%).  
  
3. Selecția variabilelor:  
- Au fost selectate primele 10 caracteristici corelate cu `hours-per-week` și normalizate.  
  
4. Tratamentul outlierilor:  
- Au fost eliminate 396 cazuri cu reziduuri mai mari de ±20 ore pentru a reduce zgomotul și a crește performanțele modelului.

## Modele Implementate

1. Linear Regression (OLS): Model de referință.  
2. SGDRegressor: Implementat conform cerințelor.  
3. Decision Tree Regression: Tunat și selectat ca model final.  
4. Random Forest Regression: Testat opțional, dar cu performanțe inferioare modelului cu arbore de decizie.

## Performanțele Modelelor

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metrică** | **Linear Regression** | **Tuned Decision Tree** | **Random Forest** | **SGDRegressor** |
| **Train MAE** | 7.99 | 7.14 | 6.19 | 7.96 |
| **Train R²** | 0.07 | 0.20 | 0.45 | 0.067 |
| **Validation MAE** | 7.92 | 7.28 | 8.20 | 7.88 |
| **Validation R²** | 0.06 | 0.205 | 0.06 | 0.059 |
| **Test MAE** | 8.05 | 7.31 | 8.38 | 8.01 |
| **Test R²** | 0.065 | 0.177 | 0.046 | 0.066 |

## Analiza Modelului Final: Tuned Decision Tree Regression

1. Parametrii selectați (prin GridSearchCV):  
- `max\_depth = 5`  
- `min\_samples\_split = 2`  
- `min\_samples\_leaf = 10`  
- `max\_features = None`  
  
2. Performanțe:  
- Train MAE: 7.14 | Validation MAE: 7.28 | Test MAE: 7.31  
- Train R²: 0.20 | Validation R²: 0.205 | Test R²: 0.177  
  
3. Importanța caracteristicilor:  
- Caracteristicile cele mai importante au fost:  
 - `education-num`   
 - `capital-gain`  
 - `age`   
 - `is\_highly\_educated`

## Vizualizări

1. Actual vs. Predicted:  
- Se observă o aliniere rezonabilă între valorile reale și cele prezise pentru modelul final.  
2. Plotul reziduurilor:  
- Reziduurile sunt distribuite relativ uniform în jurul liniei de zero, indicând o bună generalizare a modelului.  
3. Feature Importance:  
- `education-num` și `capital-gain` au contribuit cel mai mult la predicții.

## Concluzii

1. Modelul Final:  
- Tuned Decision Tree Regression este cel mai performant model bazat pe:  
 - MAE și RMSE cele mai scăzute pe seturile de validare și test.  
 - R² mai mare comparativ cu celelalte modele, sugerând o explicabilitate mai bună a variației din date.  
  
2. Avantaje:  
- Generalizare mai bună față de Linear Regression și SGDRegressor.  
- Performanță robustă, chiar și după eliminarea outlierilor.  
  
3. Limitări și Îmbunătățiri viitoare:  
**Model Complexity:** Explorarea unor modele mai complexe, precum Gradient Boosting sau XGBoost, poate aduce îmbunătățiri suplimentare.  
**Feature Engineering:** Adăugarea interacțiunilor între variabile sau termeni polinomiali poate captura mai bine relațiile non-liniare.  
**Outlier Handling:** Analiza și tratarea suplimentară a outlierilor extremi.