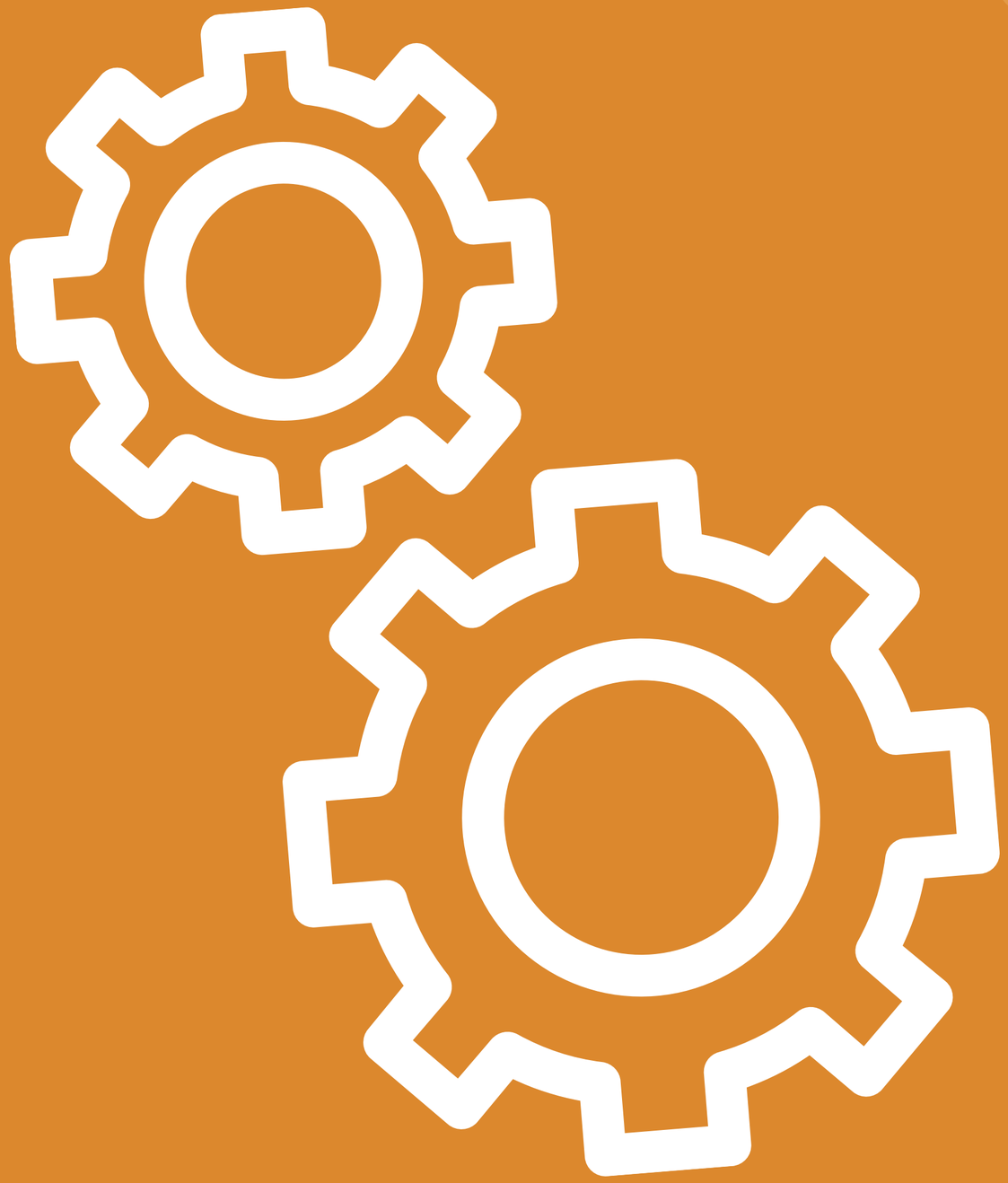
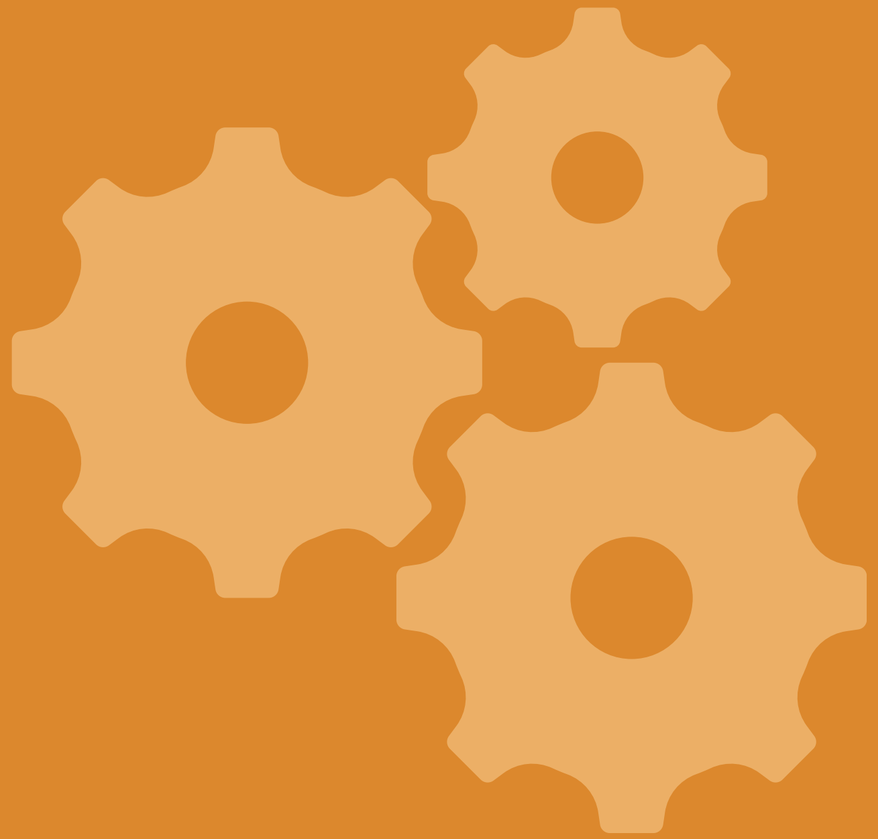


WARSZTATY BADAWCZE

Imputacja danych oraz model uczenia maszynowego dla danych dot. żeliwa ADI

Mateusz Król
Mateusz Zacharecki
Patrycja Żak

PLAN PREZENTACJI



1. Cel projektu oraz wybranie zmiennych objaśnianych.
2. Omówienie głównych problemów w danych oraz szczególnych przypadków.
3. Użyte metody imputacji braków danych.
4. Proces modelowania danych.
5. Wnioski oraz możliwe ulepszenia.

CEL PROJEKTU

Celem projektu było uzupełnienie braków danych, co stanowiło najważniejszy kamień milowy postawionego zadania. Dodatkowo, po opracowaniu danych zaimplementowano modele uczenia maszynowego.

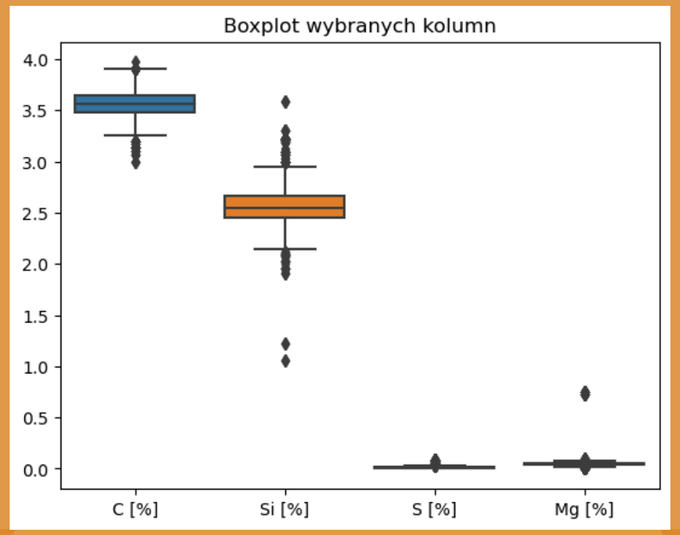
Zmienne objaśniane:

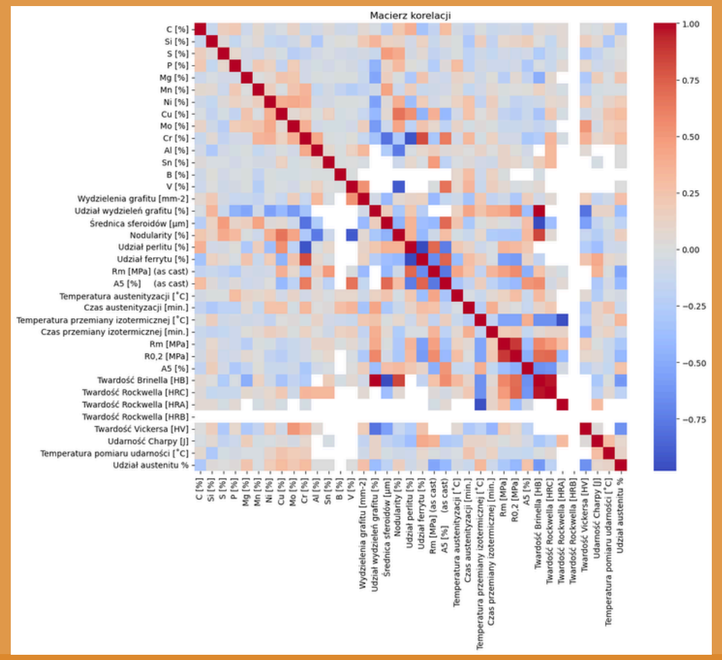
R_m[MPa]

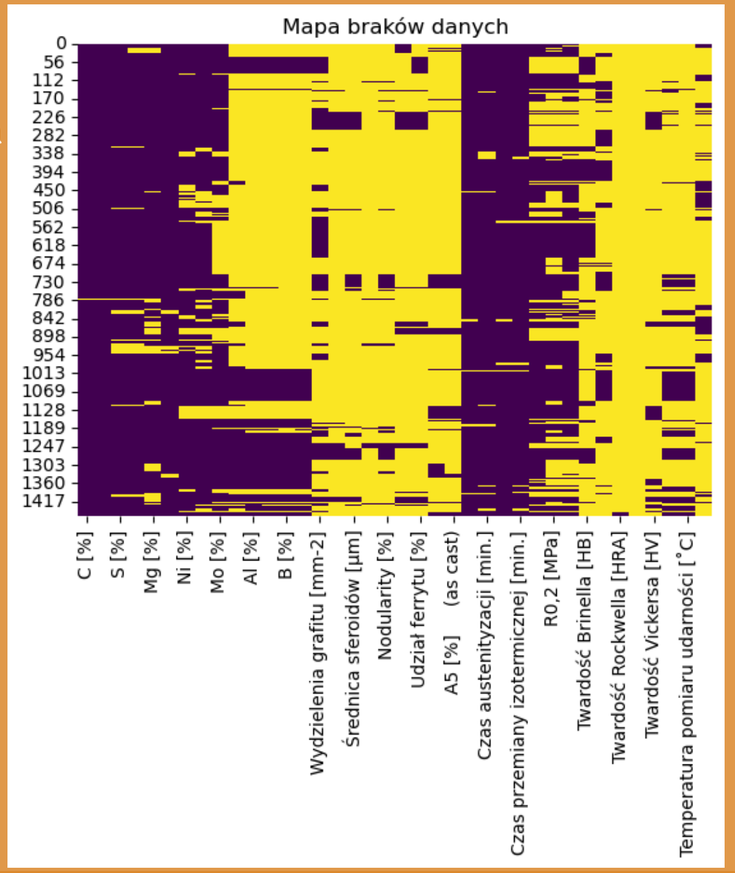
Udarność Charpy [J]

GŁÓWNE PROBLEMY I ROZWIĄZANIA

1. Usunięcie kolumn ze zbędnymi informacjami: Nr źródła, Minimalna grubość ścianki [mm], Martensite volume fraction Xα, Retained austenite volume fraction XγR
2. Sprawdzenie i uzupełnienie udział perlitu + udział ferrytu = 100% oraz sprawdzenie czy nie jest błędne wliczanie węgla do tych 100%.
3. Sprawdzenie wartości fosforu i siarki i wyrzucenie 20 rekordów, gdzie zanieczyszczenie fosforem wyniosło 0.34.
4. Usunięcie 146 rekordów, gdzie poziom węgla i krzemu był poza zakresem dobrze przeprowadzonego procesu.



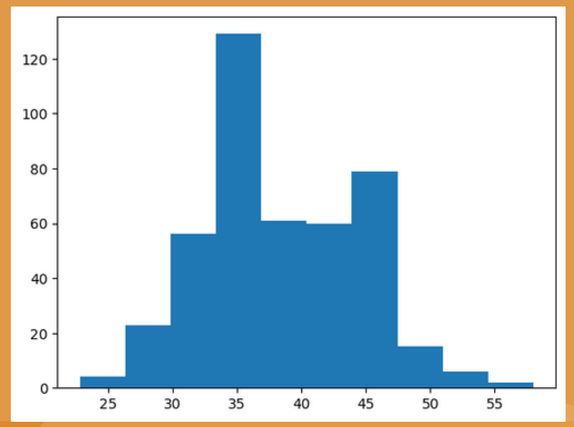


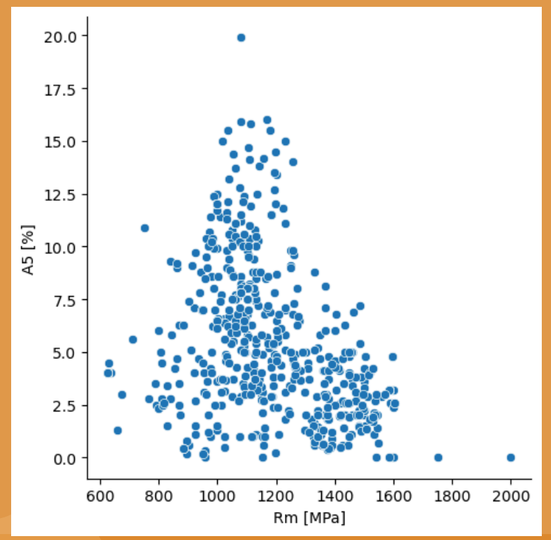


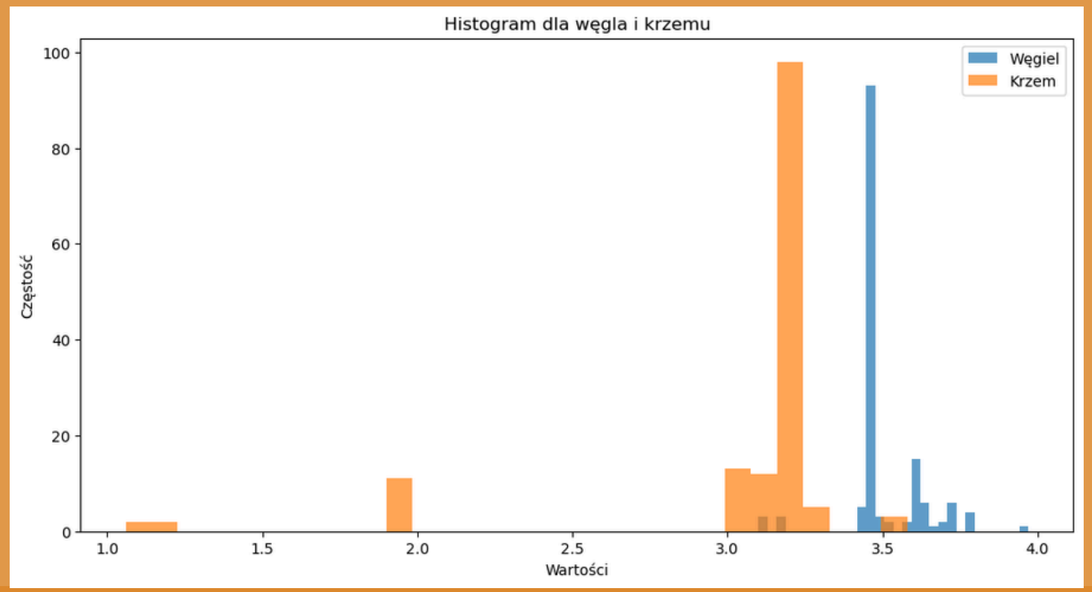
GŁÓWNE PROBLEMY I ROZWIĄZANIA

5. Uzupełniono brakujące wartości kolumn nikiel, miedz, chrom, aluminium, cyna wartością 0.0001.
6. Uzupełniono wartości kolumny 'Średnica sferoidów [µm]' na bazie wartości z kolumny 'Wielkość sferoidów' stosując zasadę 4: 18, 5: 9, 6: 4.5, 6 i 7: 3.
7. Uzupełniono brakujące wartości kolumny 'Temperatura pomiaru udarności [°C]' wartością 22.
8. Usunięto niepoprawne procesy, gdzie Mg < 0.025, Temperatura austenityzacji < 900, Czas austenityzacji <= 1.
9. Uzupełniono brakujące wartości kolumn Siarka, Fosfor, Mangan i Czas przemiany izotermicznej [min.] medianą obs. nieodstających, ponieważ w tych kolumnach jest mało braków i nie ma wyraźnej korelacji ze zmiennymi.
10. Imputacja twardości Brinella korzystając z przybliżonych wzorów oraz kolumn HV, HRC, HRA.
11. Uzupełnienie danych w oparciu o zmienne z wysoką korelacją.

Twardość Rockwella







DODATKOWE METODY IMPUTACJI

Na początku sprawdzono trzy metody na kolumnach, w których pozostały nadal braki danych:

1. Algorytm MICE (Multivariate imputation by chained equations)
2. MLPRegressor (Multi-layer Perceptron regressor)
3. RandomForest

Następnie przetestowano również inne podejście, które składało się z:

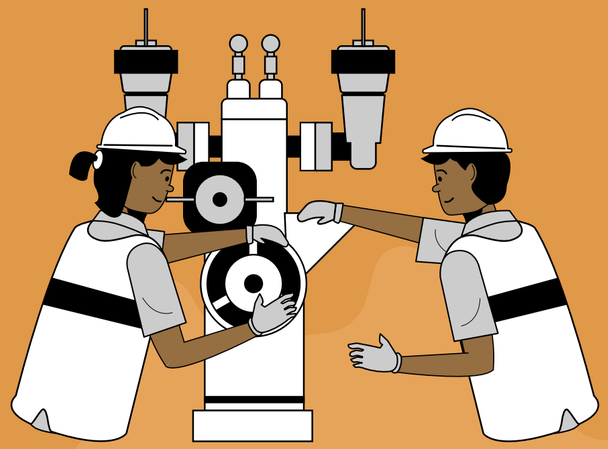
1. RandomForest dla złożonych procesów fizycznych: Rm [MPa], A5 [%], Twardość Brinella [HB]
2. CatBoostRegressor dla danych opisujących skład chemiczny (pierwiastki) żeliwa.
3. KNNImputer dla zmiennych jakościowych i pozostałych zmiennych

PRZYGOTOWANIE MODELU

Przetestowano trzy modele:

1. RandomForestRegressor
2. XGBRegressor
3. LGBMRegressor

oraz użyto GridSearchCV i BayesSearch do optymalizacji hiperparametrów.



WYNIKI

R_m[MPa]

Wyniki dla Random Forest (Grid) :
MSE: 4524.06
MAE: 40.04
R^2 Score: 0.89

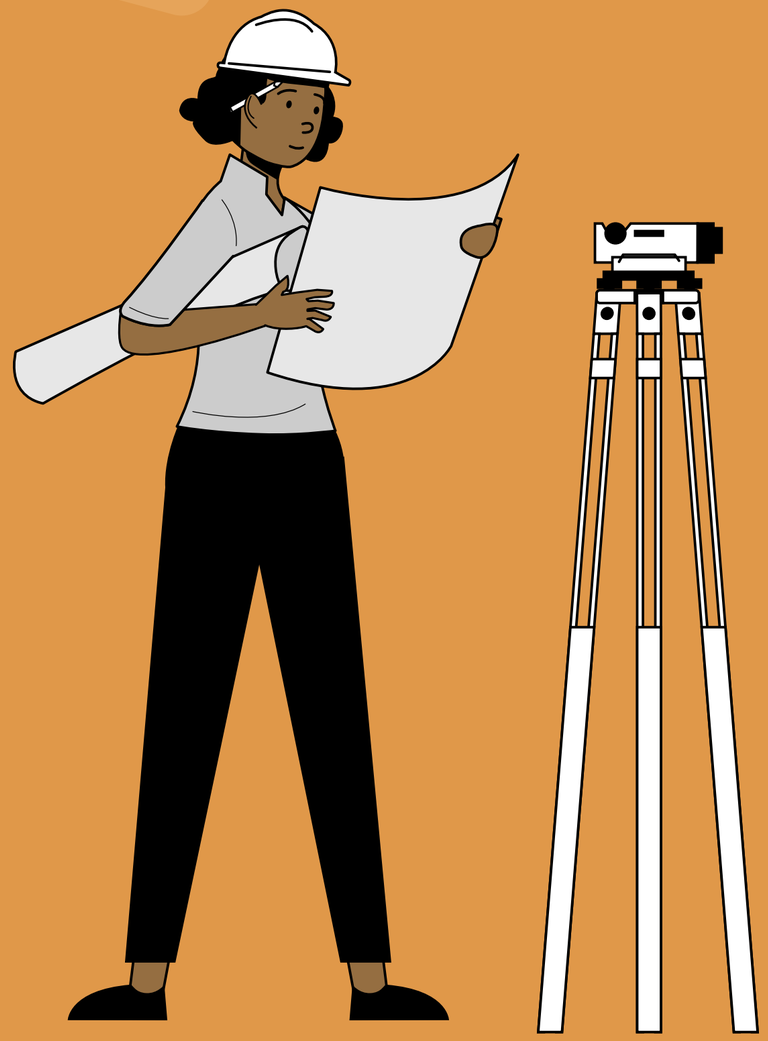
Wyniki dla Random Forest (Bayes):
MSE: 4573.27
MAE: 40.15
R^2 Score: 0.89

Wyniki dla XGB (Grid):
MSE: 3441.63
MAE: 36.59
R^2 Score: 0.91

Wyniki dla XGB (Bayes):
MSE: 3557.91
MAE: 35.62
R^2 Score: 0.91

Wyniki dla LGB (Grid):
MSE: 4011.20
MAE: 42.45
R^2 Score: 0.90

Wyniki dla LGB (Bayes):
MSE: 3667.79
MAE: 39.65
R^2 Score: 0.91



Udarność Charpy [J]

Wyniki dla Random Forest (Grid) :
MSE: 188.70
MAE: 6.60
R^2 Score: 0.70

Wyniki dla Random Forest (Bayes):
MSE: 194.69
MAE: 6.70
R^2 Score: 0.69

Wyniki dla XGB (Grid):
MSE: 146.67
MAE: 6.18
R^2 Score: 0.77

Wyniki dla XGB (Bayes):
MSE: 137.69
MAE: 6.21
R^2 Score: 0.78

Wyniki dla LGB (Grid):
MSE: 174.85
MAE: 6.95
R^2 Score: 0.72

Wyniki dla LGB (Bayes):
MSE: 156.99
MAE: 6.72
R^2 Score: 0.75

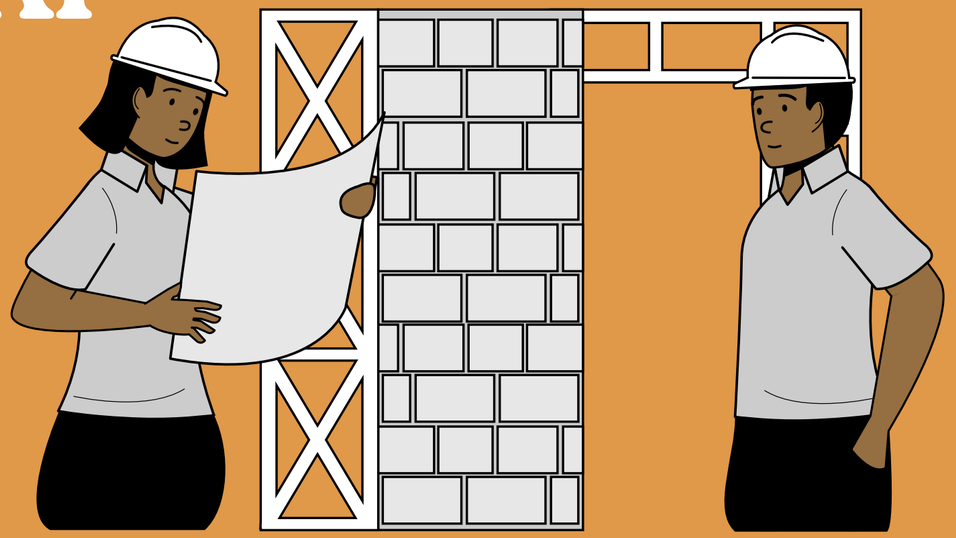
WNIOSKI

DLA R_m[MPa]:

- XGB (Grid) - najlepszy model wg MSE i R2
- XGB (Bayes) - najlepszy model wg MAE

DLA Udarność Charpy [J]:

- XGB (Grid) - najlepszy model wg MSE i R2
- XGB (Bayes) - najlepszy model wg MAE



MOŻLIWE ULEPSZENIA

1. Większa indywidualność przy wyborze metod imputacji.
2. Wybór metod intuicyjnych typu DALIA.

WYNIKI NA DANYCH TESTOWYCH

DLA R_m[MPa]:

- colsample_bytree: 1.0,
- learning_rate: 0.05,
- max_depth: 6,
- n_estimators: 300,
- subsample: 0.6
- random_state: 123

HIPERPARAMETRY

DLA Udarność Charpy [J]:

- colsample_bytree: 0.6,
- learning_rate: 0.026164,
- max_depth: 20,
- n_estimators: 486,
- subsample: 0.6,
- random_state: 24

DLA R_m[MPa]:

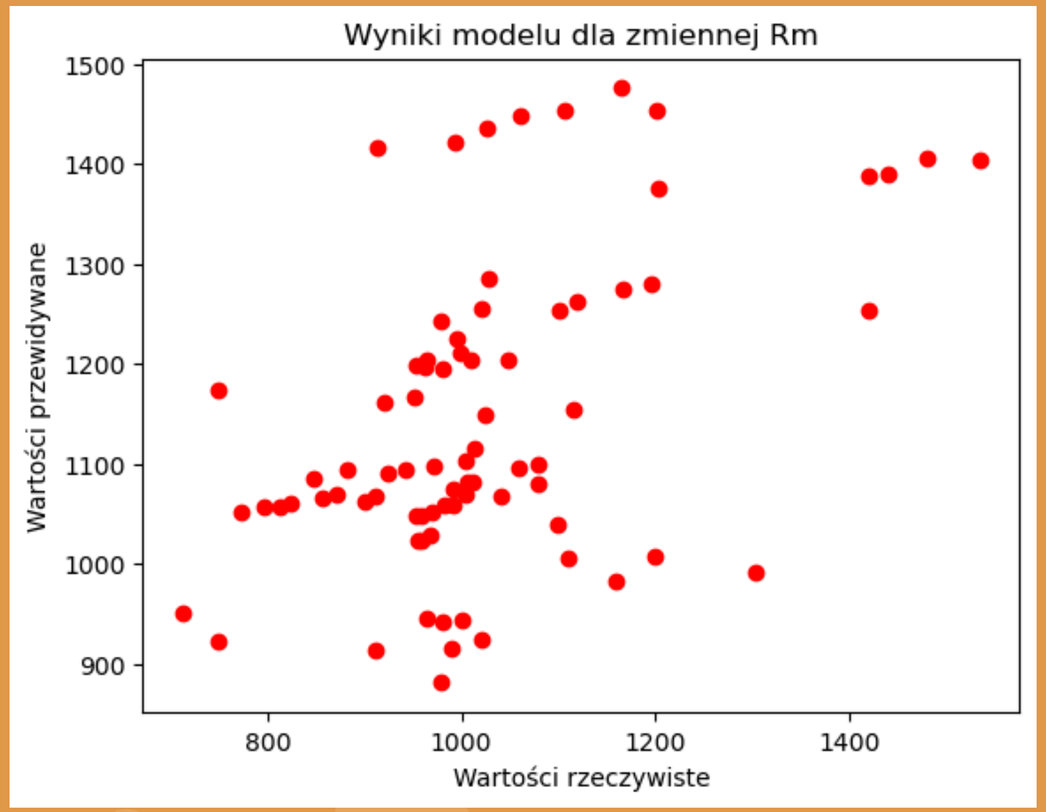
- MSE: 26323.43
- MAE: 132.70
- R^2 Score: -0.63

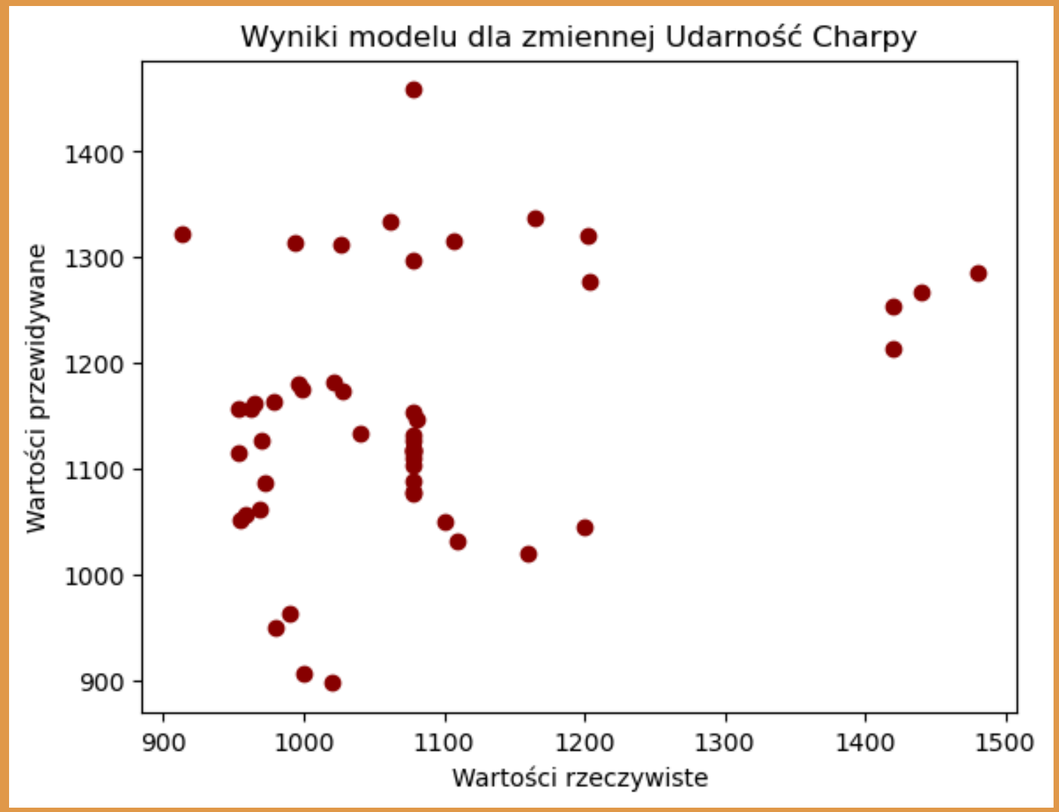
WYNIKI

DLA Udarność Charpy [J]:

- MSE: 20781.52
- MAE: 115.35
- R^2 Score: -0.29

WYNIKI NA DANYCH TESTOWYCH





DZIĘKUJEMY ZA UWAGĘ