Opracowanie i implementacja systemu predykcji kosztów leczenia medycznego

Opis przetwarzania wstępnego danych

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830
1334	18	female	31.920	0	no	northeast	2205.98080
1335	18	female	36.850	0	no	southeast	1629.83350
1336	21	female	25.800	0	no	southwest	2007.94500
1337	61	female	29.070	0	yes	northwest	29141.36030

1338 rows × 7 columns

Zakodowanie napisów jako cyfry:

- **sex** → male = 0, female = 1,
- **smoker** → no = 0, yes = 1,

Kolumna **region** początkowo zawierała możliwe 4 wartości: southwest, southeast, northwest, northeast. Zamieniliśmy ją na dwie kolumny **region_north** oraz **region_east** o wartościach 0 lub 1. Dzięki temu możemy lepiej szacować trendy zależne od pochodzenia danej osoby.

Zbiór nie zawierał wartości pustych lub odstajacych.

Ostatnim krokiem przetwarzania wstępnego była normalizacja danych numerycznych za pomocą funkcji MinMaxScaler().

Wybór algorytmów

Bazując na informacjach wyczytanych w publikacjach naukowych oraz implementacjach użytkowników Kaggle postanowiliśmy przetestować 4 modele regresji:

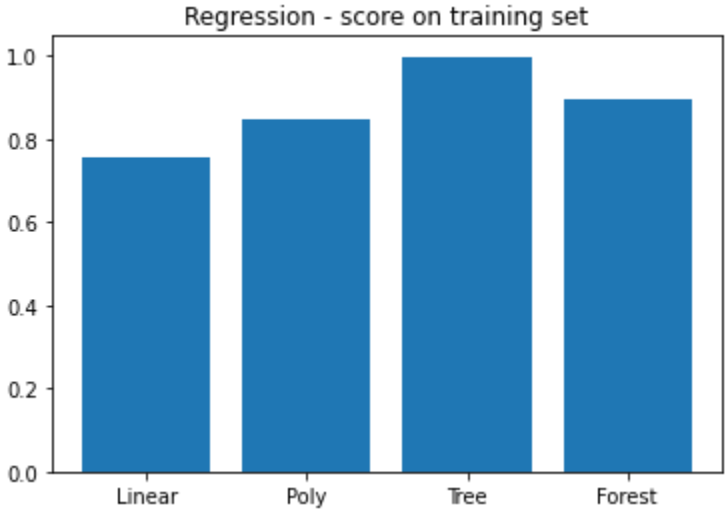
- liniową,
- wielomianową (zdecydowano się na wielomian 2 stopnia),
- regresor drzewa decyzyjnego,
- regresor lasu losowego.

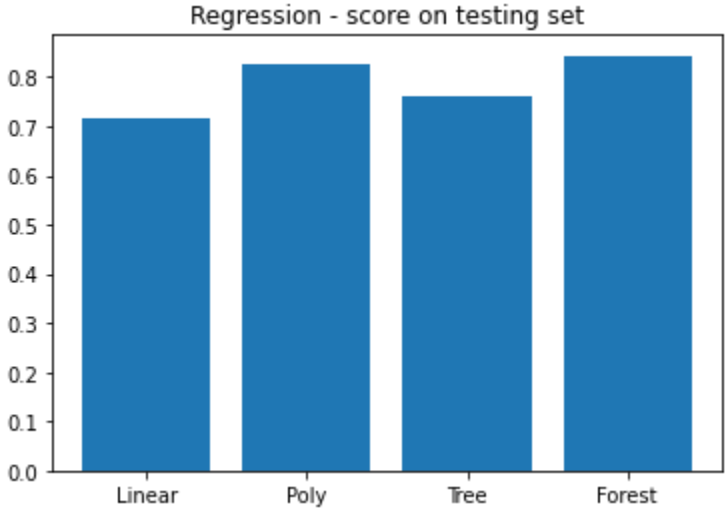
Miarami oceny ich jakości

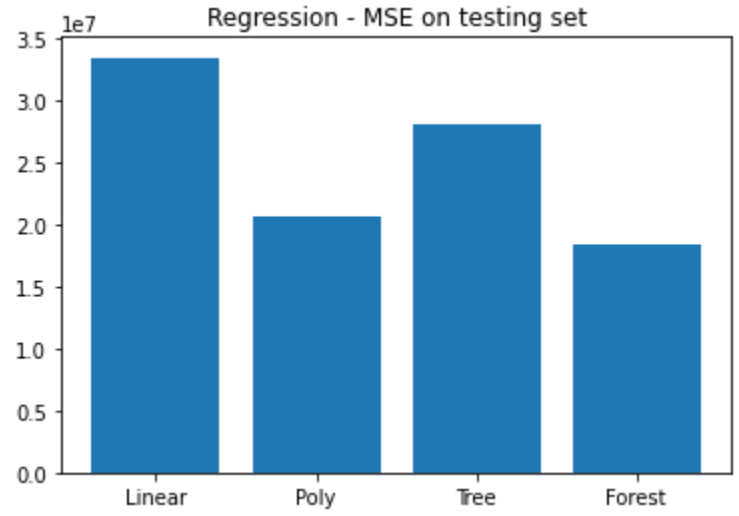
Oceniając najlepszy algorytm do celów naszego projektu posłużyliśmy się 3 metrykami:

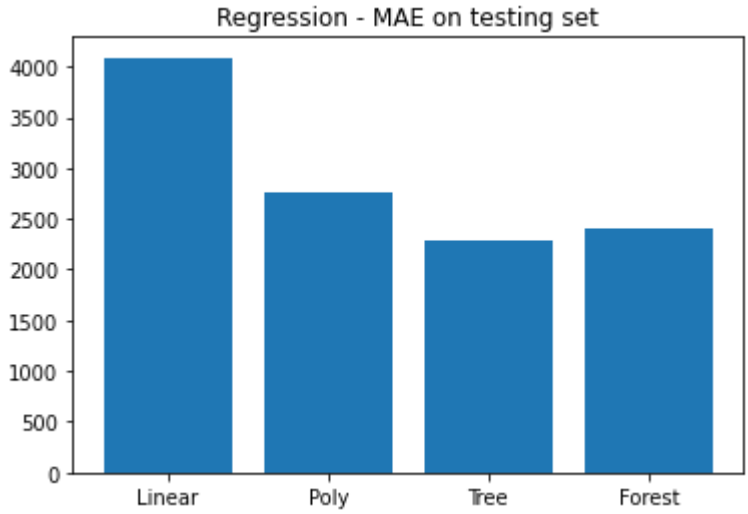
- **współczynnik determinacji** – liczony zarówno na zbiorze treningowym oraz testowym,
- **maksymalny błąd bezwzględny** – liczony tylko na zbiorze testowym,
- **błąd średniokwadratowy** – liczony tylko na zbiorze testowym

Porównanie algorytmów









	Score – train	Score – test	MSE	MAE
Linear	0.756	0.716	33447093	4088
Polynomial	0.85	0.825	20694876	2759
Tree	0.999	0.761	28137564	2276
Forest	0.895	0.843	18443504	2396

Uzasadnienie wyboru algorytmu zaimplementowanego w aplikacji

- Najgorszym wyborem okazała się regresja liniowa, dlatego została odrzucona z konkursu na najlepszy model.
- Mimo świetnego wyniku na zbiorze treningowym, regresor drzewa decyzyjnego okazał się przetrenowany, osiągając bardzo słabe wyniki na zbiorze testowym,
- Z pozostałych dwóch, lepsze wyniki uzyskał regresor lasu losowego, dlatego postanowiliśmy finalnie, że to ten model zostanie wykorzystany w aplikacji.