Predykcja skali przestępczości

Jakub Gdulski (s27753)

Oleksandr Perekuta (s29072),   
Dawid Kucharski (s27637),

[**1. Cel badania i opis zbioru danych**](#_heading=h.30j0zll) **2**

[Cel badania](#_heading=h.1fob9te) 2

[Opis zbioru danych](#_heading=h.3znysh7) 2

[**2. Metodologia i rozwiązanie**](#_heading=h.2et92p0) **3**

[Podejście do problemu](#_heading=h.tyjcwt) 3

[Użyte algorytmy](#_heading=h.3dy6vkm) 3

[**3. Wstępne przetwarzanie danych**](#_heading=h.4d34og8) **4**

[Imputacja brakujących wartości](#_heading=h.2s8eyo1) 4

[Selekcja ważnych cech](#_heading=h.17dp8vu) 4

[**4. Metoda oceniania jakości modelu**](#_heading=h.3rdcrjn) **4**

[Miary ewaluacyjne](#_heading=h.26in1rg) 4

[Walidacja modeli](#_heading=h.lnxbz9) 4

[**5. Wyniki eksperymentalne**](#_heading=h.35nkun2) **5**

[1. Dla danych oryginalnych](#_heading=h.1ksv4uv)5

[2. Dla danych zredukowanych](#_heading=h.z337ya) 5

[3. Porównanie w postaci graficznej](#_heading=h.3w8zxfj8ur5)10

[**6. Podsumowanie**](#_heading=h.4i7ojhp) **12**

[Wyniki](#_heading=h.2xcytpi) 12

[Wnioski](#_heading=h.na5sztjnsf6p) 12

[Podział pracy](#_heading=h.qaa5xihbonok) 12

### 

# 

# 1. Cel badania i opis zbioru danych

### **Cel badania:**

Celem badania jest opracowanie modelu do predykcji skali przestępczości w nowych regionach, korzystając z danych demograficznych oraz warunków ekonomicznych.

### **Opis zbioru danych:**

Zbiór danych pochodzi z bazy UCI i zawiera informacje o warunkach demograficznych oraz liczbie przestępstw w miastach USA. Główne cechy:

* **Liczba atrybutów:** Ponad 120 cech obejmujących m.in. średnią pensję, warunki mieszkaniowe, średnią wielkość rodziny.
* **Rozkład klas decyzyjnych:** Etykietami są liczby interwencji policji wyrażone w procentach

# 2. Metodologia i rozwiązanie

### **Podejście do problemu:**

1. Klasyfikacja problemu jako zadanie regresji.
2. Analiza wpływu poszczególnych cech na wynik.

### **Użyte algorytmy:**

1. Random Forest
2. Linear Regression
3. Gradient Boosting
4. Lasso Regression
5. Neural Network

**Kroki analizy:**

1. Dane są przetwarzane
2. Modele są trenowane zarówno jak na danych pełnych tak i zredukowanych
3. Modele są oceniane tymi samymi miarami ewaluacyjnymi dla obiektywnej oceny
4. Wyniki regresji porównywano jak dla danych oryginalnych tak i zredukowanych
5. Wyniki oceny modeli przedstawiane w formie graficznej

# 

# 3. Wstępne przetwarzanie danych

### **Imputacja brakujących wartości:**

1. Wartości “?” były zastąpione NaN
2. Użyto średniej jako zastępczej dla zmiennych brakujących numerycznych
3. Zamiana wartości symbolicznych na numeryczne

### **Selekcja ważnych cech:**

Usunięto cechy o niskiej korelacji z predekywaną wartością za pomocą algorytmu Random Forest

# 4. Metoda oceniania jakości modelu

### **Miary ewaluacyjne:**

* Mean Absolute Error (MAE)
* R-squared (R²)
* Mean Squared Error (MSE)

### **Walidacja modeli:**

* Użyto 5-krotnej walidacji krzyżowej do oceny stabilności modelu.

### **Podział danych:**

* 75% - zbiór treningowy
* 25% - zbiór testowy

# 

# 5. Wyniki eksperymentalne

## Dla danych oryginalnych

### Porównywanie dokładności modeli pod względem czasu ich trenowania

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **MAE** | **MSE** | **R^2** | **Czas Trenowania (w sekundach)** |
| NeuralNetwork | 0.111767 | 0.028371 | 0.416531 | 38.440783 |
| LinearRegression | 0.096820 | 0.018027 | 0.629265 | 0.020072 |
| RandomForest | 0.089539 | 0.016766 | 0.655202 | 2095.839126 |
| GradientBoosting | 0.089385 | 0.017151 | 0.647285 | 1275.906391 |
| Lasso | 0.089282 | 0.016576 | 0.659106 | 0.339613 |

Wyniki postaci graficznej:

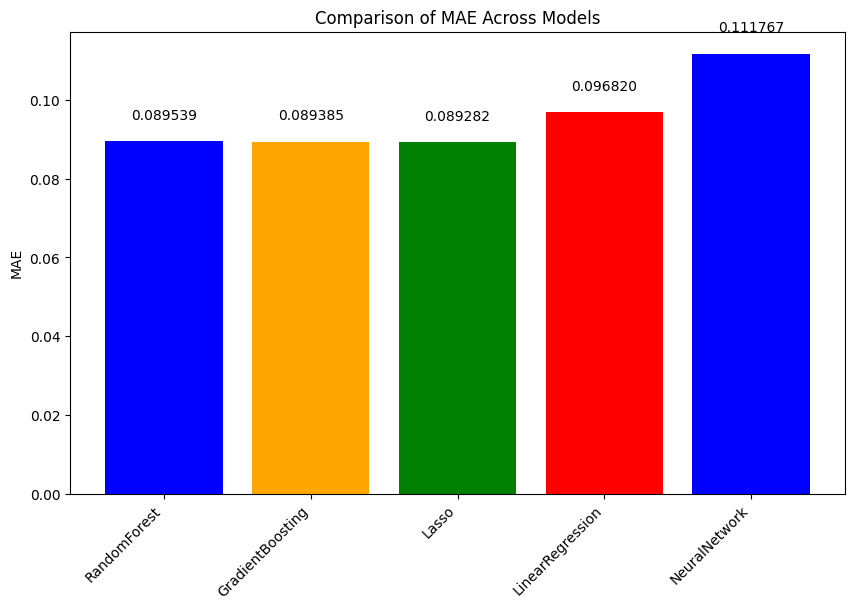
## Dla danych zredukowanych

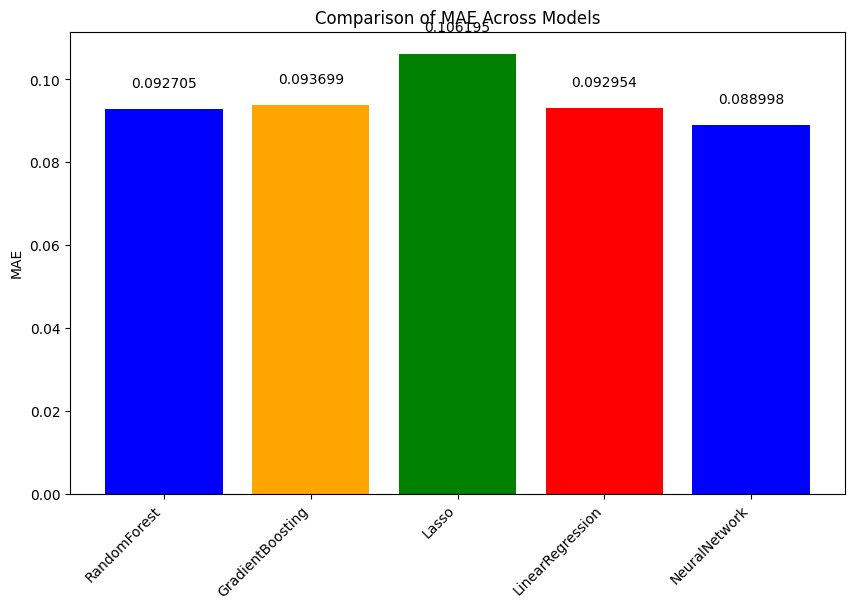
### Porównywanie dokładności modeli pod względem czasu ich trenowania

| **Model** | **MAE** | **MSE** | **R^2** | **Czas Trenowania (w sekundach)** |
| --- | --- | --- | --- | --- |
| NeuralNetwork | 0.088425 | 0.018668 | 0.616068 | 141.065065 |
| LinearRegression | 0.093886 | 0.018066 | 0.628457 | 0.007766 |
| RandomForest | 0.092255 | 0.018298 | 0.623683 | 4578.577290 |
| GradientBoosting | 0.091558 | 0.018684 | 0.615753 | 8733.697478 |
| Lasso | 0.106195 | 0.020387 | 0.580715 | 0.262445 |

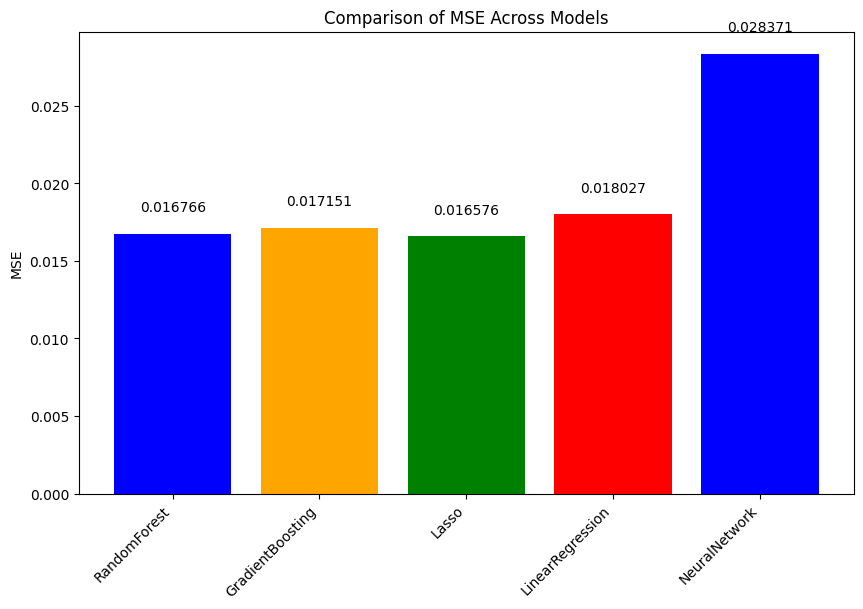
## Porównanie w postaci graficznej

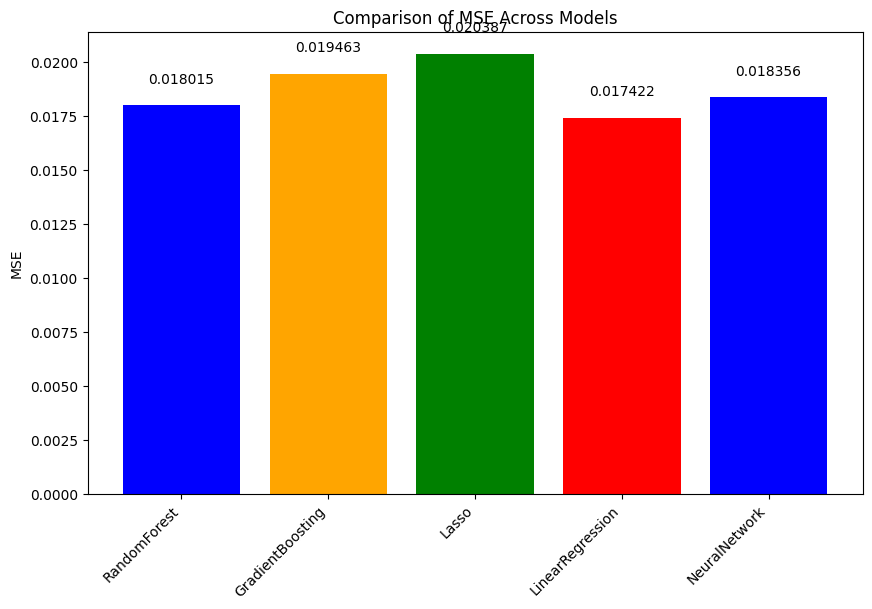
#### MAE (dane oryginalne / zredukowane)



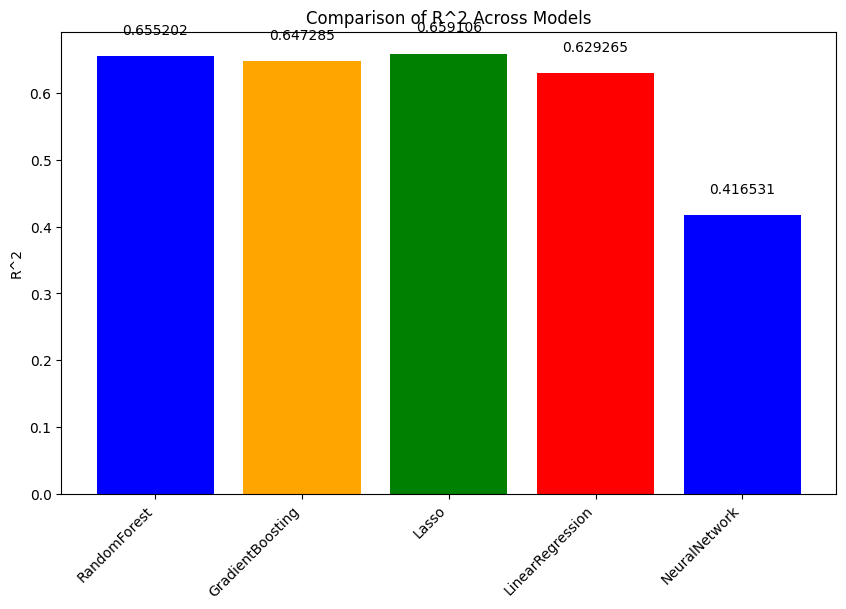


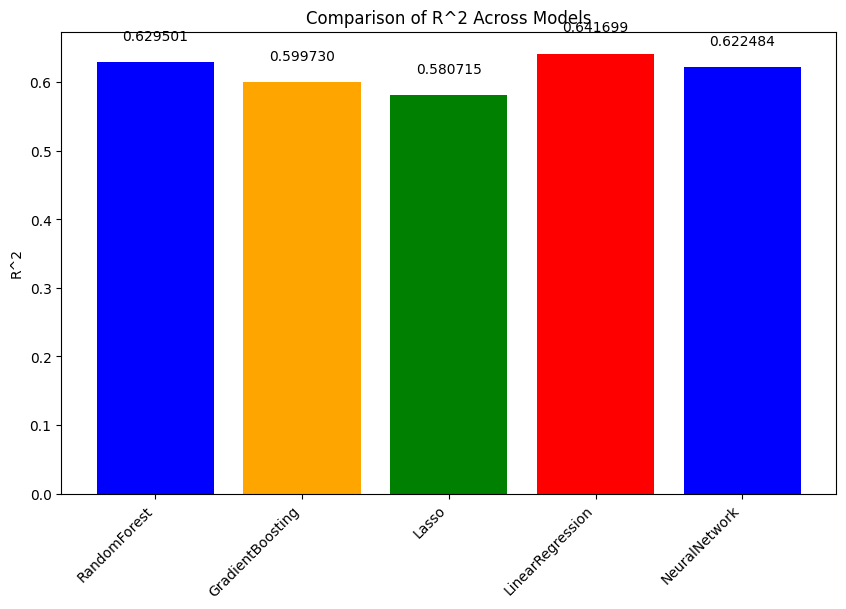
#### MSE (dane oryginalne / zredukowane)



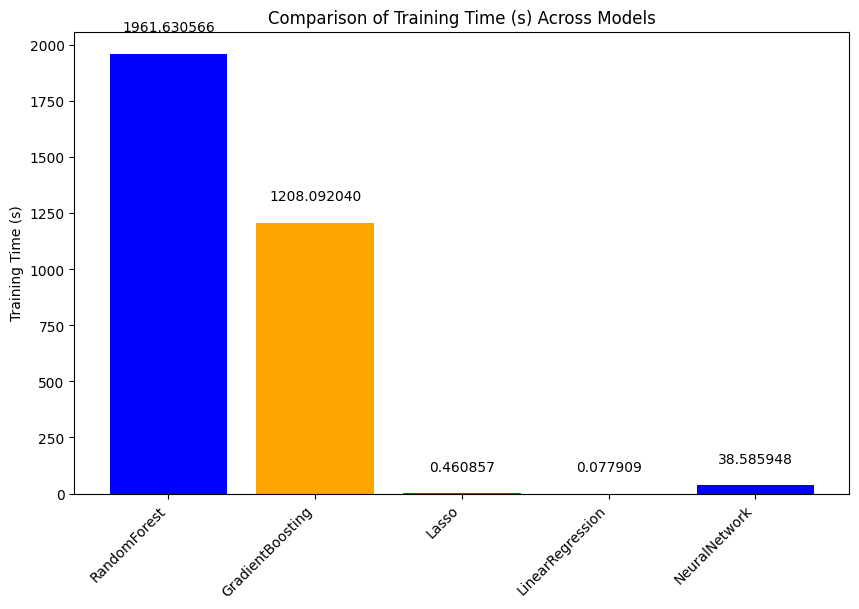


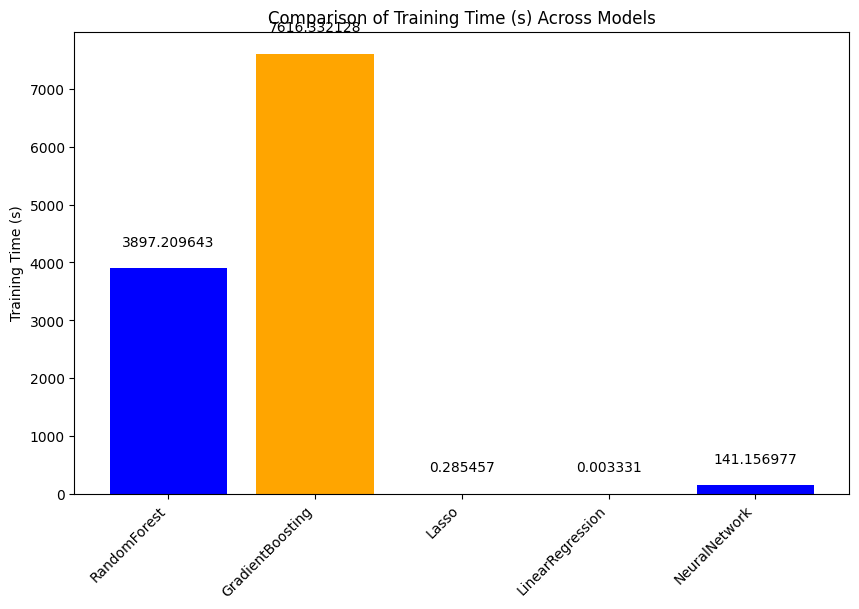
#### R^2 (dane oryginalne / zredukowane)



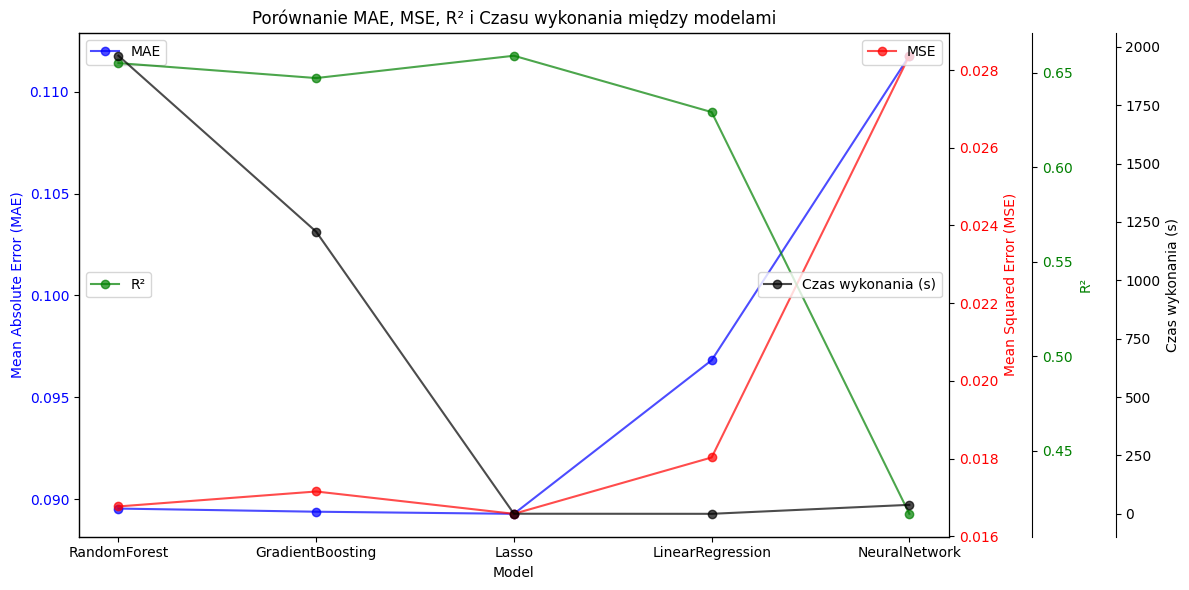


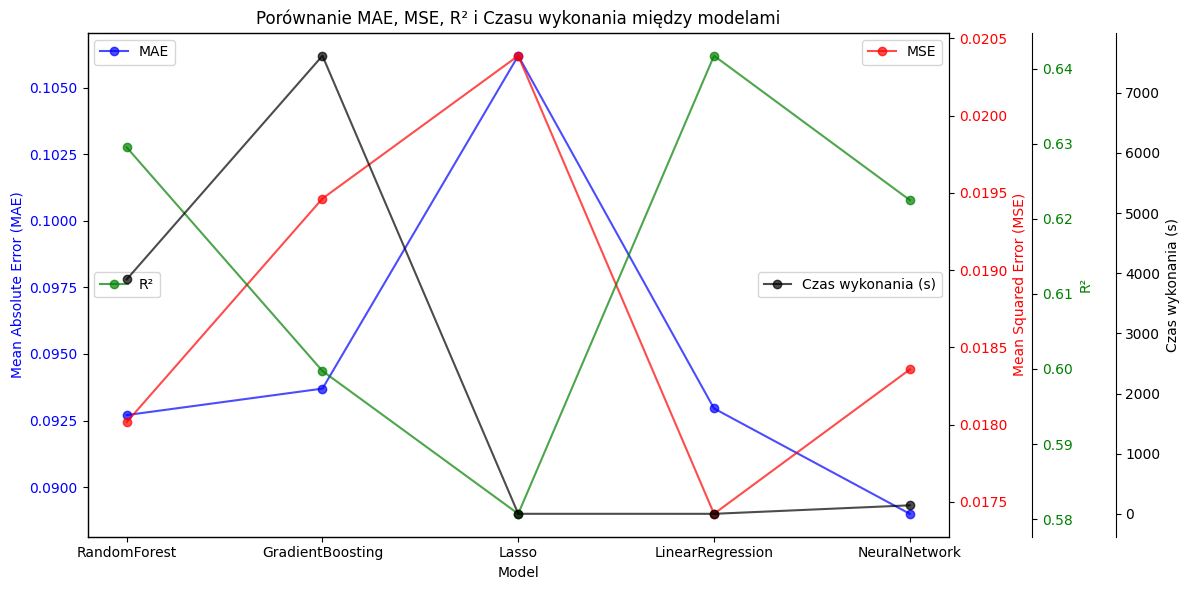
#### Training Time (dane oryginalne / zredukowane)



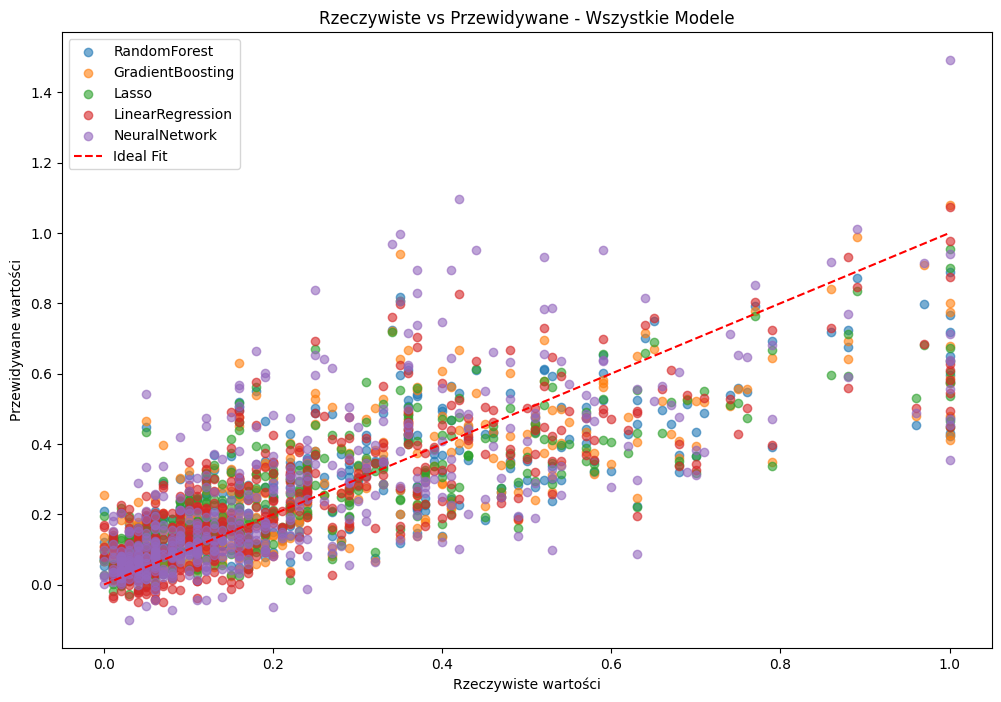


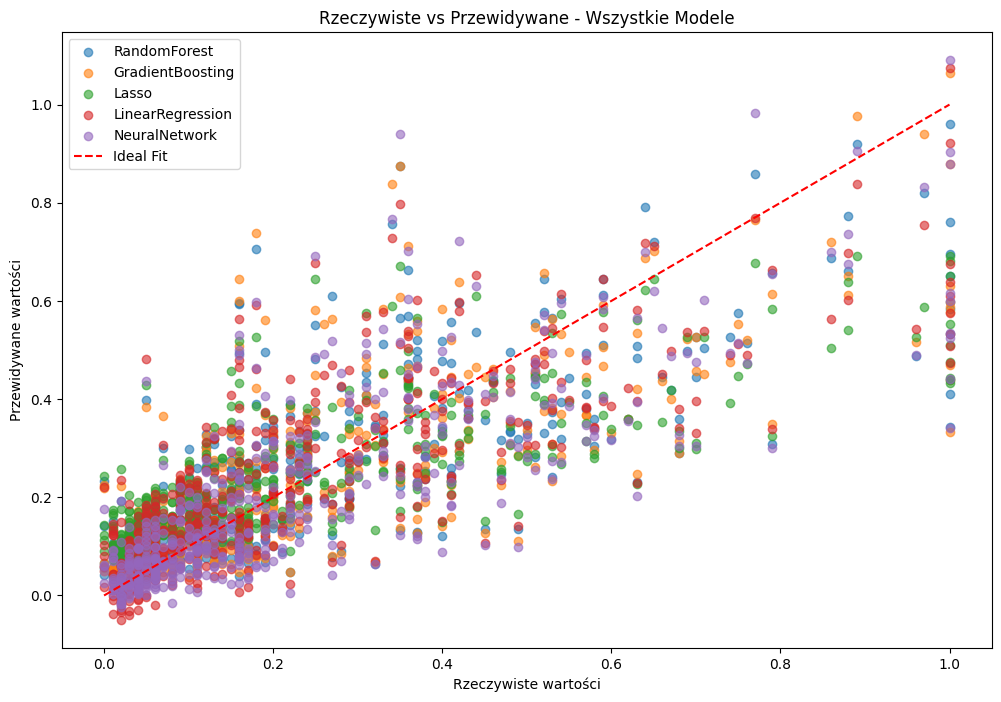
#### Porównanie wszystkich modeli względem dokładności (dane oryginalne / zredukowane)





* Przewidywane vs Rzeczywiste wartości (Wszystkie modele)





# 6. Podsumowanie

## **Wyniki:**

1.Dla danych oryginalnych:

* Lasso Regression osiągnęło najlepsze wyniki i najkrótszym czasem trenowania
* Sieci neuronowe charakteryzowały się najdłuższym czasem trenowania i stosunkowo niższą skutecznością.
* Gradient Boosting oraz Random Forest miały porównywalne wyniki, ale różniły się znacznie czasem trenowania.

2. Dla danych zredukowanych:

* Sieci neuronowe poprawiły wyniki, ale trenowanie było dłuższe
* Lasso Regression miało najkrótszy czas trenowania, ale gorsze wartości R²

## **Wnioski:**

* Redukcja danych wpłynęła na zwiększenie czasu trenowania modeli i dokładności, szczególnie w przypadku algorytmów wymagających większej mocy obliczeniowej.
* Lasso Regression okazało się najbardziej efektywne w przypadku danych pełnych, dzięki równowadze między czasem trenowania a dokładnością.
* Sieci neuronowe były mniej efektywne, wymagają dalszej optymalizacji, aby były konkurencyjne względem klasycznych metod regresji.

Wnioski te podkreślają istotność wyboru odpowiednich algorytmów w zależności od wymagań projektu.

## **Podział pracy:**

**Oleksandr Perekuta**

budowanie modeli (Neural Network), wizualizacja wyników, napisanie raportu, sprawdzenie poprawności wyników

**Jakub Gdulski**

wybranie ważnych cech i usunięcie nieważnych, budowanie modeli (Lasso, Gradient Boosting), tworzenie wizualizacji, analiza wyników

**Dawid Kucharski**

załadowanie danych, budowanie modeli (Random Forest, Linear Regression), napisanie raportu