Raport zaliczeniowy
Przewidywanie zużycia paliwa

Aliaksandr Mikulka

MSiD Lab Grupa 3
276751@student.pwr.edu.pl

Raport wygenerowany 19 maja 2024

Opis Problemu

Problemem wybranym do badań jest przewidywanie zużycia paliwa samochodu w zależności od innych czynników. Analiza zbioru danych może wykazać przybliżone zużycie paliwa samochodu na podstawie danych o jego silniku, paliwie które jest wykorzystane, warunkach pogodowych oraz populacji badanego obszaru. Celem projektu jest zbadanie powyższych zależności oraz stworzenie modelu który potrafi przewidzieć zużycie paliwa samochodu.

Zbiór danych i jego przetwarzanie

Zbiór danych

Dla rozwiązania powyższego problemu zostały wykorzystane cztery zbiory danych. Wszystkie te dane były zebrane na terytorium Kanady więc obszarem badanym będzie ten kraj.

Pierwsze dwa zbiory były pobrane ze strony internetowej kaggle.com.

- Pierwszy zbiór [1] zawiera informacje o roku produkcji, marce i modelu, typu samochodu, charakterystykach silnika i transmisji oraz dane o paliwie i wskaźnikach jego zyżycia w różnych przypadkach.
- Drugi zbiór [2] zawiera informacje o średnich temperaturach oraz ilości opadów w każdym miesiącu.
- Trzeci zbiór [3] był pobrany ze strony www150.statcan.gc.ca . Ten zbiór zawiera informacje o ilości samochodów wyprodukowanych w każdym roku.
- Ostatni zbiór [4] był pobrany ze strony wikipedia.org . Zbiór jest zeskrobany z tabeli która zawiera informacje o populacji Kanady.

Przetwarzanie danych do analizy

Każdy zbiór jest przetwarzany w taką postać, w ktorej on będzie praktyczny do wykorzystania i przedstawiania danych.

- Dane o temperaturze (tabela 1) są przetwarzane do postaci, gdzie dla każdego roku są wartości reprezentujące srednią temperaturę, ogólną ilość opadów i śniegu.
- Zbiór o ilości sprzedanych samochodów (tabela 2) zawiera dużo zbędnej informacji. Z tego zbioru są wykorzystane tylko te dane, które reprezentują ilość sprzedanych samochodów w każdym roku.
- Zbiór danych reprezentujący populacje Kanady (tabela 3) jest zeskrobany z tabeli wikipedii. Z tego zbioru potrzebujemy tylko dane o populacji w każdym roku.
- W zbiorze danych o samochodach, ich charakterystykach i wskaźnikach zużycia paliwa (tabela 4) niektóre dane mają takie same wartości, tylko napisane w inny sposób (np. marka napisana wielkimi lub małymi literami, typ samochodu zapisany z innym znakiem rozdielającym). Także kolumna reprezentująca typ paliwa jest przetważana do pełnej nazwy w celu łatwiejszego zrozumienia jakie paliwo jest wykorzystane.

Wszystkie zbiory były przeanalizowane na przedmiot brakujących danych. Tylko w zbiorze reprezentujacy średnie temperatury była znaleziona jedna zgubiona wartość o ilości opadów w 1917 roku. Ta wartość była zastąpiona przez 0 ponieważ takie stare dane nie będa potrzebne dla przewidywania.

Tabela 1. Dane pogodowe

Rok	Średnia temperatura	Opady śniegu	Całkowite opady
1917	3.397697	71993.4	282321.6
1918	4.601561	61197.0	299492.6
1919	4.422720	62713.9	294860.5
1920	4.473663	66841.0	299336.7
1921	5.045527	67793.2	323221.2

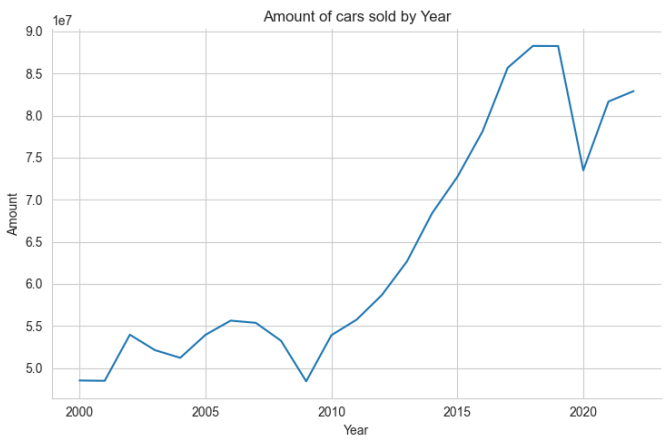
Tabela 2. Dane o sprzedaży samochodów

Rok	Ilość sprzedanych samochodów
1946	313373
1947	646494
1948	660518
1949	875066
1950	1315365

Analiza danych i wyszukiwanie zależności

Następnym etapem jest wyszukiwanie zależności pomiędzy danymi które mogą pomóc w przewidywaniu zużycia. Sprzedaży samochodów będą mieć zakres od 2000 do 2022 roku.

Analiza danych



Rysunek 1. Ilość sprzedanych samochodów według roku

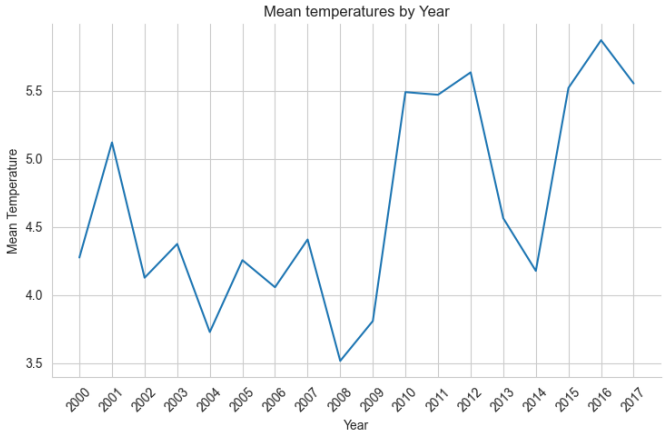
Dane pogodowe będą mieć zakres od 2000 do 2017 roku.

Tabela 3. Coroczna Populacja

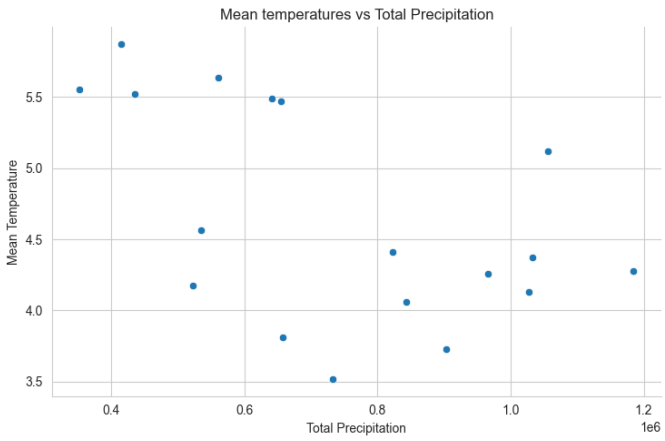
Rok	Populacja
1900	5500000
1901	5600000
1902	5760000
1903	5930000
1904	6100000

Tabela 4. Dane o samochodach i zużyciu paliwa

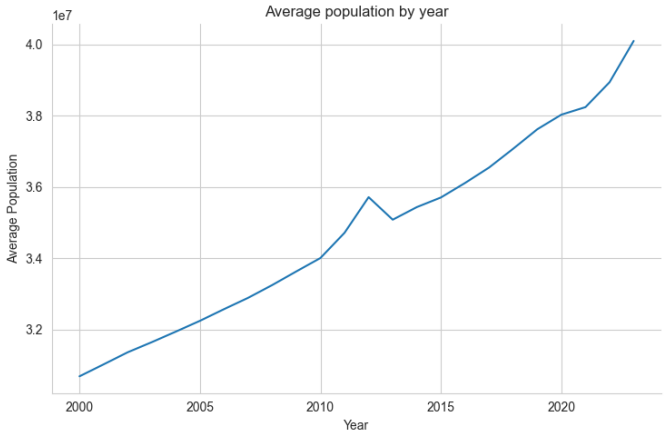
Rok produkcji	Marka	Model	Typ Samochodu
2000	ACURA	1.6EL	COMPACT
2000	ACURA	1.6EL	COMPACT
2000	ACURA	3.2TL	MID-SIZE
Pojemność silnika	**Cylindry**	**Transmisja**	**Paliwo**
1.6	4	A4	Regular gasoline
1.6	4	M5	Regular gasoline
3.2	6	AS5	Premium gasoline
Zużycie paliwa	**HWY**	**COMB**	**Emisje (mpg)**
9,2	6.7	8.1	186
8.5	6.5	7.6	175
12.2	7.4	10.0	230



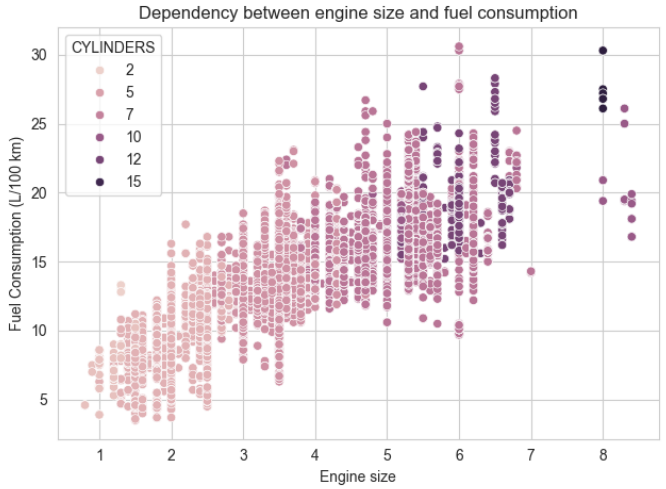
Rysunek 2. Średnie temperatury według roku



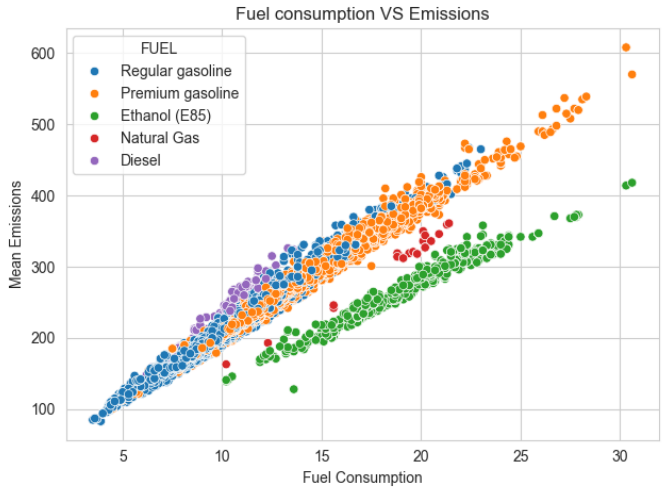
Rysunek 3. Zależność pomiędzy średnią temperaturą i całkowitymi opadami



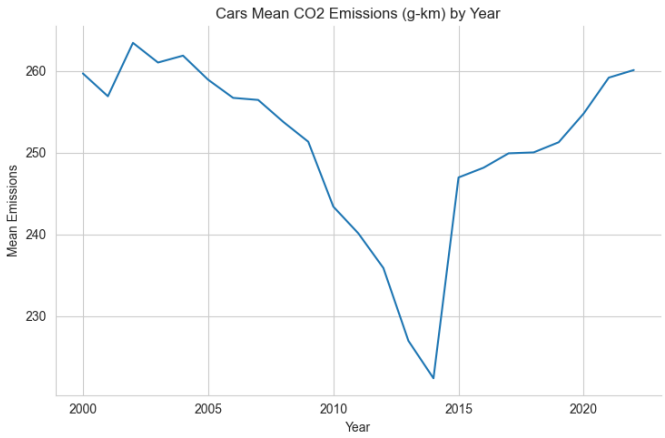
Rysunek 4. Średnia populacja według roku



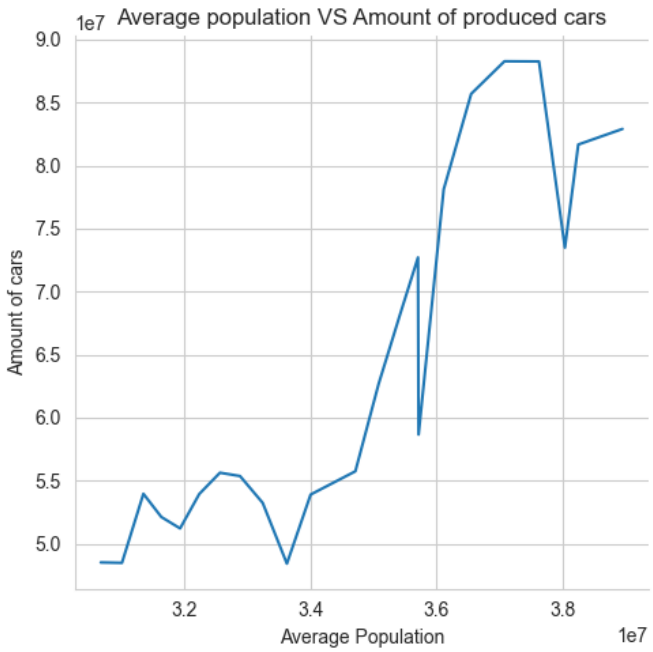
Rysunek 5. Zależność pomiędzy pojemnością silnika i zużyciem paliwa



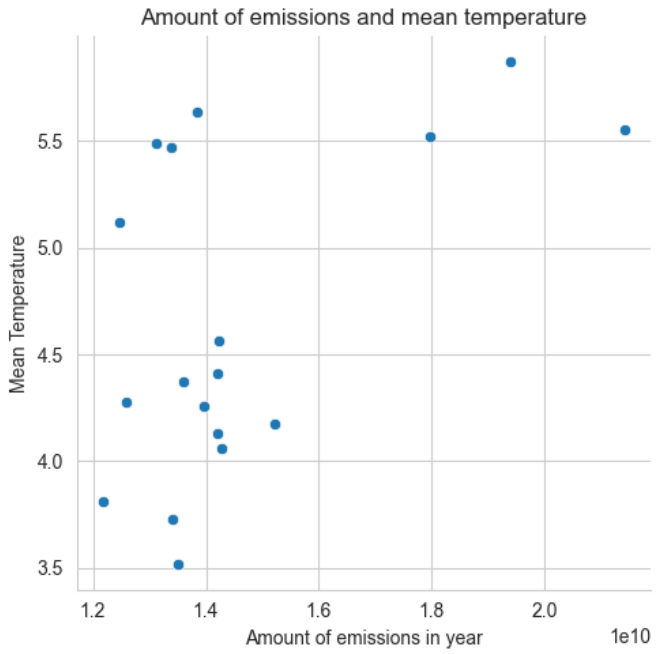
Rysunek 6. Zależność pomiędzy zużyciem paliwa i emisją



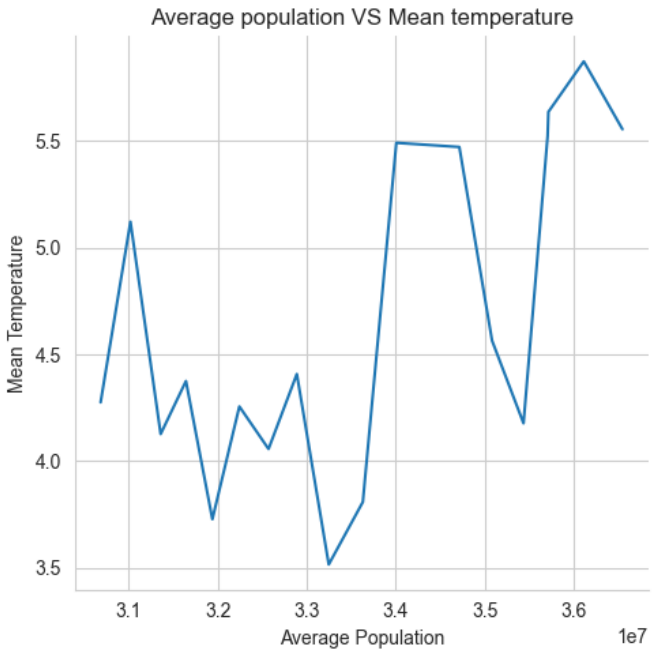
Rysunek 7. Średnia emisja samochodów według roku



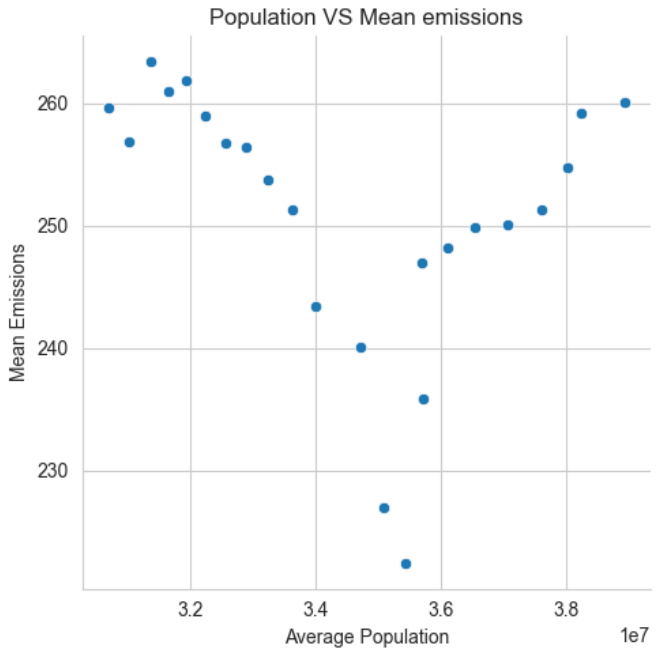
Rysunek 9. Zależność liczby ludności od ilości wyprodukowanych samochodów



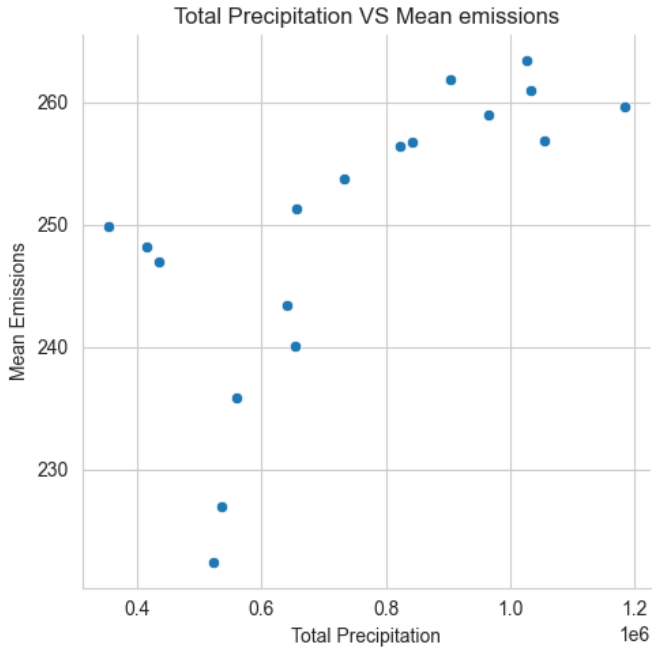
Rysunek 8. Zależność między ilością emisji a średnią temperaturą



Rysunek 10. Zależność populacji od średniej temperatury



Rysunek 11. Zależność pomiędzy liczbą ludności a wielkością emisji



Rysunek 12. Zależność pomiędzy opadami i ilością emisji

Wyszukiwanie zależności

Na podstawie statystyk trudno wywnioskować zależności pomiędzy średnią temperaturą, populacją a danymi pogodowymi, jednak to nie oznacza, że żadnej zależności nie ma. Na niektórych wykresach można zobaczyć podobeństwo zależności.

Jednak możemy zobaczyć bardzo dobrą zależność pomiędzy pojemnością silnika, ilością cylindrów, wykorzystanym typem paliwa a wskaźnikiem zużycia paliwa. Także, na wykresie 6 można zobaczyć interesującą zależność od typu wykorzystanego paliwa z którego można wywnioskować, że diesel robi najwięcej emisji przy takim samym zużyciu, benzyna premialna ma największy wskaźnik zużycia/emisji (z tego powodu że benzyna premialna jest wykorzystana w samochodach premium klasy, np. Porsche, Ferrari itd., które mają duże silniki i wysoką moc) oraz to, że etanol jest najbardziej ekologicznym paliwem, ponieważ przy dowolnym zużyciu paliwa zawsze produkuje znacznie mniej emisji niż inne paliwa.

Modele przewidywające

Wstępne przetwarzanie danych

Dlatego aby stworzyć model, który przewidzi zużycie paliwa najpierw potrzebujemy dodać wszystkie dane do jednego zbioru danych. Głównym zbiorem danych w danym przypadku to zbiór samochodów wraz z ich zużyciem paliwa. Dane samochodu łączone są z danymi o średniej temperaturze, populacji i sprzedanych samochodach na podstawie roku produkcji samochodu, tzn. zestaw danych podstawowych dla każdego unikalnego roku produkcji będzie miał tą samą wartość, która odpowiada jego wartości i rokowi w oryginalnym zbiorze.

Dla trenowania został wybrany okres od 2000 do 2017 roku, ponieważ w tym scenariuszu mamy wszystkie potrzebne dane.

Następnie, dla trenowania były wyrzucone niepotrzebne dane, które nie mają wpływu na zużycie paliwa: *Marka, Model, Typ Samochodu, Transmisja*. Także musimy usunąć inne wskaźniki zużycia paliwa i emisji: *HWY (L/100 km), COMB (L/100 km), COMB (mpg), Emisje*, ponieważ one są wyliczane na podstawie zużycia paliwa.

Ostatnim krokiem jest zmiana wartości typów paliwa. Ponieważ modele mogą trenować się tylko na podsawie danych numerycznych, typy paliwa są zmieniane na odpowiednie liczby:

- 1 - Benzyna zwykłą
- 2 - Benzyna premialna
- 3 - Diesel
- 4 - Ethanol
- 5 - Naturalny gaz

Tworzenie modeli przewidywających

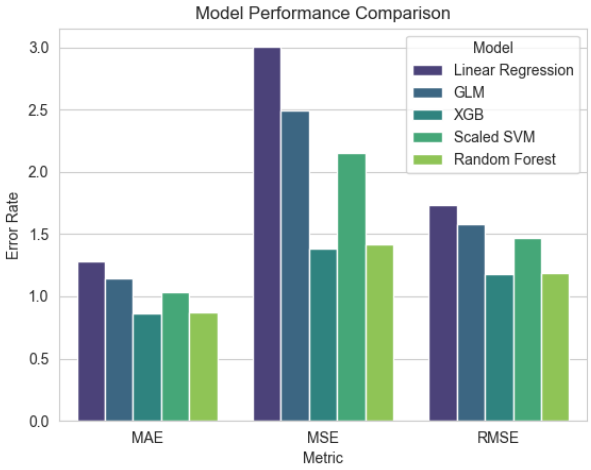
Dla tworzenia zbiorów trenujących i testowych była wykorzystana metoda *train_test_split* z biblioteki scikit-learn. Następnie za pomocą tych zbiorów były trenowane pięć różnych modeli: **Regresja liniowa**, **Uogólniony model liniowy (GLM)**, **XGBoost**, **Maszyna wektorów nośnych (Scaled SVM)**, wykorzystująca dodatkową bibliotekę doprowodzająca dane do jednej skali w celu uzyskania dobrych wyników uczenia się, i **Metoda lasu losowego (Random Forest)**.

Po trenowaniu i przewidywaniu możliwych wartości spalania paliwa, były otrzymane rezultaty pokazane w tabeli 5:

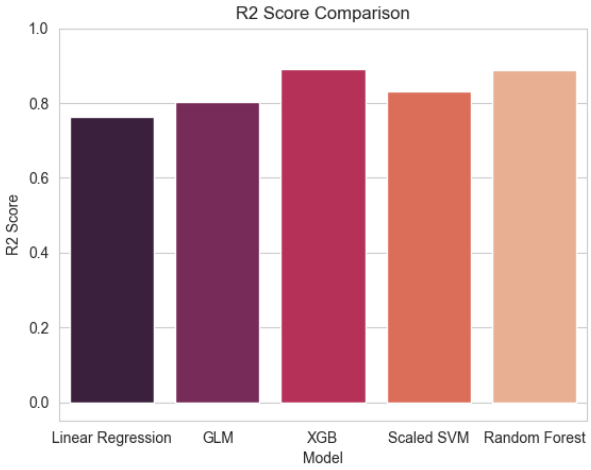
Tabela 5. Porównanie wydajności modeli

Model	MAE	MSE	RMSE	R^2
Regresja liniowa	1.303577	3.156626	1.776690	0.754250
GLM	1.374618	3.376737	1.837590	0.737114
XGBoost	**0.887038**	**1.519095**	**1.232516**	**0.881735**
Scaled SVM	1.070279	2.328741	1.526021	0.818702
Random forest	0.894831	1.556078	1.247428	0.878856

MAE - Średni błąd bezwzględny
MSE - Średni błąd kwadratowy
RMSE - Pierwiastek średniokwadratowy błędu



Rysunek 13. Porównanie wydajności modeli

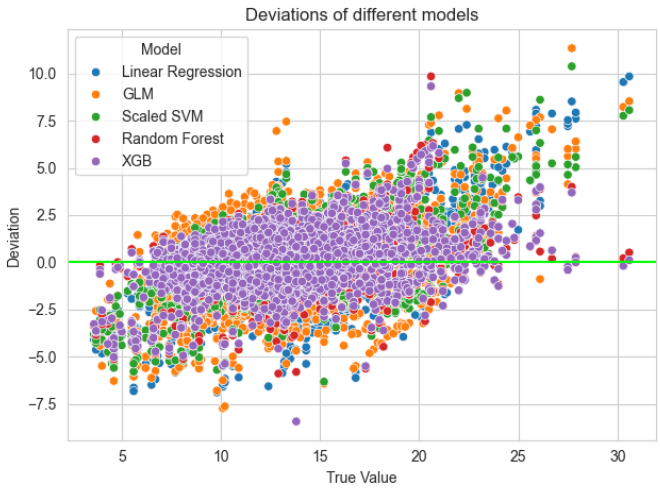


Rysunek 14. Porównanie wyników R^2

Współczynnik R^2 reprezentuje procentową zmienność zmiennej zależnej, która jest wyjaśniona przez niezależne zmienne w modelu. Wartość R^2 jest zazwyczaj pomiędzy 0 a 1, gdzie 1 oznacza idealne dopasowanie, a wartość bliższa 0 oznacza, że model słabo wyjaśnia dane. Wysoki współczynnik R^2 wskazuje na to, że model dobrze pasuje do obserwowanych danych.

Na podstawie wyników z tabeli 5 i wykresów 13 oraz 14, można wywnioskować, że w naszym przypadku najlepsze przewidywanie danych osiąga model **XGBoost**.

Wykres 15 pokazuje odchylenie przewidzianych wartości od wartości prawdziwych.



Rysunek 15. Odchylenia różnych modeli

Wnioski

Analiza i modelowanie zużycia paliwa w samochodach na podstawie zgromadzonych danych z różnych źródeł dostarczyło istotnych wniosków. Przeprowadzone badania wykazały, że model **XGBoost** był najbardziej efektywny w przewidywaniu zużycia paliwa, co zostało potwierdzone najwyższym wynikiem współczynnika R^2 i błędów MAE, MSE i RMSE. Wysoka wartość współczynnika R^2 wskazuje na dobrą zdolność modelu do wyjaśniania zmienności danych oraz możliwość stosowania tego modelu w przyszłości dla oszacowania przybliżonego zużycia paliwa bez potrzeby na dodatkowe badania. Dodatkowo, integracja danych z różnych źródeł, takich jak warunki pogodowe, demografia i sprzedaż samochodów, pozwoliła na stworzenie bardziej złożonego i dokładnego modelu.

Źródła

[1] *Fuel Consumption Dataset*. adr.: https://www.kaggle.com/datasets/ahmettyilmazz/fuel-consumption/code.

[2] *Canada Tempearture Data*. adr.: https://www.kaggle.com/datasets/sarahquesnelle/canada-data/data.

[3] *New motor vehicle sales*. adr.: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2010000101.

[4] *Demographics of Canada*. adr.: https://en.wikipedia.org/wiki/Demographics_of_Canada.