Projekt EDA

Maurycy Ebertowski

1. Wstęp

Celem projektu jest przygotowanie danych dla modelu uczenia maszynowego poprzez inżynierię cech oraz eksploracyjną analizę.

2. Źródło danych

Dane zostały pozyskane ze strony https://dane.gov.pl/ , która jest oficjalnym polskim portalem danych publicznych z różnych instytucji takich jak ministerstwa czy urzędy statystyczne. Udostępniane są bezpłatnie i można je wykorzystywać do celów naukowych lub komercyjnych. Dane te są udostępniane w wielu formatach takich jak CSV, XLSL, JSON czy XML a użytkownicy mogą je pobierać bezpośrednio lub integrować z aplikacjami za pomocą API. Dostępny jest interfejs REST API (sposób komunikacji między systemami, który pozwala na wymianę danych za pomocą standardowych żądań http, takich jak GET, POST czy DELETE), który umożliwia programowe pobieranie danych oraz ich automatyczną aktualizację, choć nie wszystkie zestawy są z nim zintegrowane - część można pobrać tylko ręcznie. Przed pobraniem jest możliwość podglądu danych i sprawdzenia jakie informacje zawierają.

3. Wybór danych

Pozyskane dane pochodzą z testów modernizowanego układu zasilania elektrycznego w dźwignicy żurawia zastosowanego jako alternatywa dla silnika diesla. Zawierają informacje o parametrach dźwignicy, kosztach zużycia energii elektrycznej i paliw oraz o warunkach pogodowych.

Dane zawierające informacje takie jak ciężar ładunku, długość wysięgnika, odległość od osi, zużycie paliwa i energii, warunki atmosferyczne czy koszty dzienne można wykorzystać w uczeniu maszynowym na przykład do szacowania dziennego zużycia energii, kosztów eksploatacji czy ryzyka przeciążenia (uczenie nadzorowane) lub do wykrywania anomalii (uczenie nienadzorowane). Dzięki informacji o czasie pomiarów można również analizować ich trendy czy sezonowość.

4. Inżynieria cech i eksploracyjna analiza danych

4.1 Brakujące dane i zbędne kolumny

W pierwszej kolejności sprawdzono ilość brakujących danych w poszczególnych kolumnach i zauważono, że dotyczą one głównie paliwa oraz energii elektrycznej w pomiarach, w których nie korzystano z tego konkretnego rodzaju napędzania. Kolumny z informacjami o zużyciu paliwa lub energii wypełniono wartością 0, ponieważ brak danych oznacza, że korzystano z drugiego sposobu napędzania, natomiast brakujące informacje o kosztach wypełniono medianą z danej kolumny. Pozostałe braki dotyczyły dziennych kosztów operacji dźwignicy i postanowiono usunąć te konkretne pomiary.

Tabela 4.1.1. Nazwy kolumn i ilość brakujących danych

Nazwa kolumny	**Ilość brakujących pomiarów**
Nr. Pomiary	0
Data	0
Ciężar ładunku [T]	0
Długość wysięgnika [m]	0
Odległość od osi [m]	0
Wysokość podnoszenia [m]	0
Maksymalne, chwilowe zużycie ON [l/h]	831
Dzienne zużycie ON [l/8h]	831
Cena hurtowa ON 1000l [PLN]	861
Maksymalne, chwilowe zużycie energii elektrycznej [kW]	117
Dzienne zużycie energii elektrycznej [kW/8h]	117
Cena energii elektrycznej [kWh]	117
Koszt dzienny [PLN]	30
Prędkość wiatru [km/h]	0
Prędkość wiatru [m/s]	0
Temperatura [C]	0
Ciśnienie [hPa]	0

Kolejnym etapem było usunięcie kilku kolumn. Pierwszą z nich była cena energii elektrycznej, gdyż była ona stała dla całego zestawu danych. Drugą kolumną była prędkość wiatru liczona w kilometrach na godzinę, ponieważ mamy kolumnę w jednostce metrów na sekundę i jedna z tych kolumn jest zbędna, bo nie wnosi żadnej nowej informacji. Trzecią kolumną był numer pomiaru, gdyż takie dane również nie wnoszą żadnej przydatnej informacji.

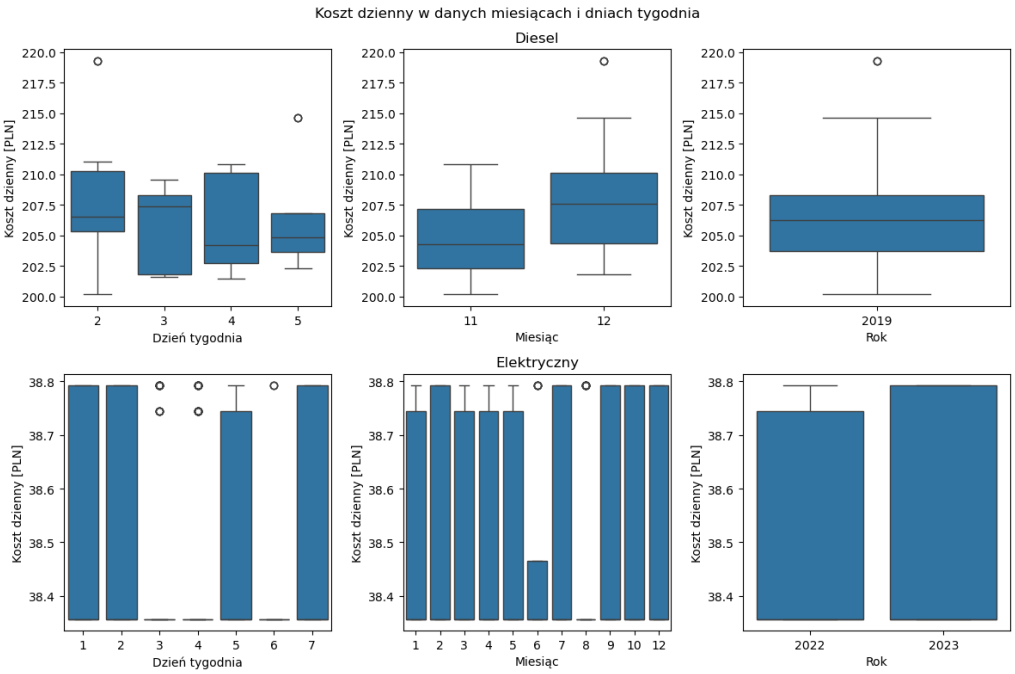
4.2 Wizualizacja kosztów operowania dźwignicy

Wizualizacja kosztów w czasie pozwala zobaczyć czy istnieje jakaś powtarzalność. Tym co jednak najbardziej rzuca się w oczy jest koszt dzienny w przypadku korzystania z napędu elektrycznego, który różni się o minimalne wartości w poszczególnych pomiarach. W zależności od specyfikacji problemu i celu dla modelu uczenia maszynowego, w konkretnych projektach można całkowicie usunąć z zestawu informacje o energii elektrycznej i przyjąć, że jest ona stała niezależnie od okoliczności.

Rysunek 4.2.1. Wykres kosztów dziennych z podziałem na rodzaj napędzania.

4.3 Rozbicie daty na osobne kolumny

Na powyższych wykresach można dostrzec pewną okresowość, więc postanowiono dodać kolumny z informacją o roku, miesiącu oraz dniu tygodnia danego pomiaru, korzystając z zapisanych dat i sprawdzić czy istnieją jakieś zależności z kosztami.

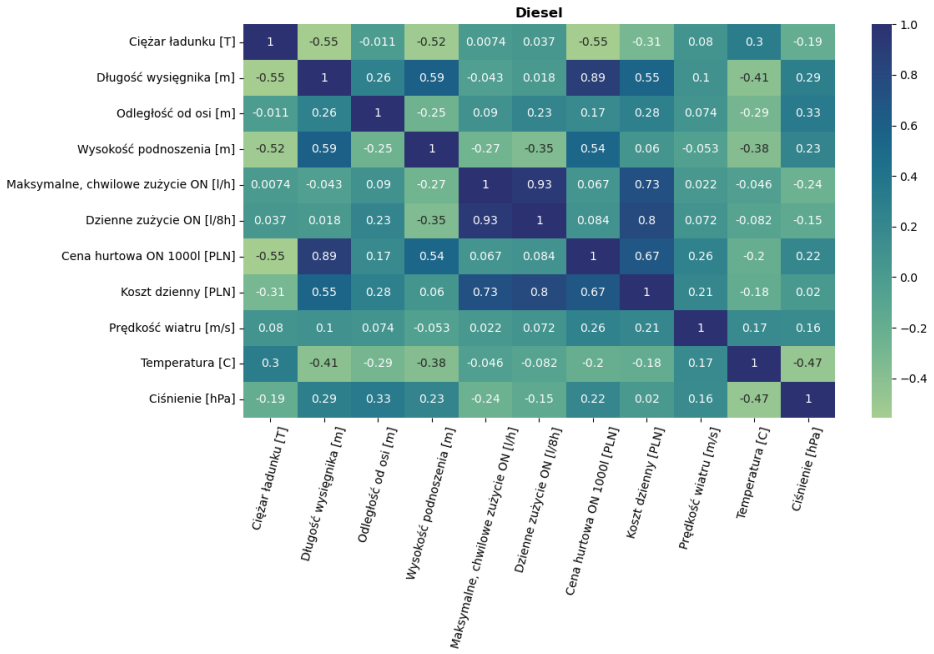


Rysunek 4.3.1. Wykresy pudełkowe przedstawiające rozkład kosztów w poszczególnych ramach czasowych.

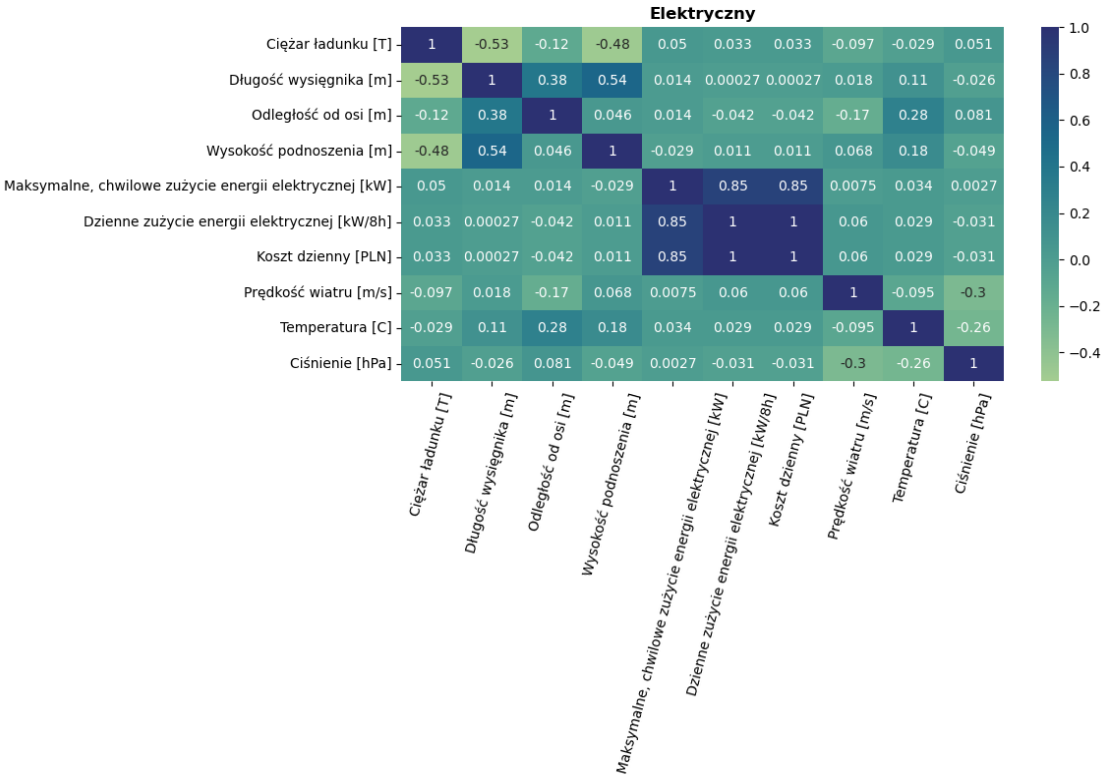
Po krótkiej analizie wykresów pudełkowych zdecydowano się na usunięcie kolumny przechowującej informację o roku, w którym dokonano pomiaru.

4.4 Korelacje między zmiennymi

Dzięki mapom korelacji Pearsona pomiędzy poszczególnymi zmiennymi, można szybko zobaczyć związki między cechami. Silna dodatnia lub ujemna korelacja może wskazywać na ich redundancję i po dokładniejszej analizie mogłoby się okazać, że jedna z nich jest niepotrzebna. Jeżeli zmienna nie ma wysokiej korelacji z żadną inną kolumną niekoniecznie oznacza to, że nic nie wnosi do zestawu danych, gdyż może mieć związek nieliniowy. Na mapach wyświetlone są tylko zmienne ciągłe (bez kategorycznych).



Rysunek 4.4.1. Mapa korelacji Pearsona dla zmiennych ciągłych i zasilania spalinowego.



Rysunek 1.4.2. Mapa korelacji Pearsona dla zmiennych ciągłych i zasilania elektrycznego.

5. Wybór zmiennych dla modelu uczenia maszynowego

Zmienną, która najbardziej nadaje się na bycie zmienną objaśnianą w modelu uczenia maszynowego niewątpliwie jest koszt dzienny. Kolejne kolumny, które są dobrymi kandydatami na bycie zmienną TARGET to dzienne zużycie paliwa lub energii elektrycznej. Z kolei do zbioru cech objaśniających (FEATURE) bardzo dobre będą dane o specyfikacji pracy dźwignicy (wysokość, ciężar, długość wysięgnika, odległość od osi) oraz dane o warunkach atmosferycznych (temperatura, prędkość wiatru, ciśnienie).

6. Wnioski

W ramach eksploracyjnej analizy danych udało się zidentyfikować istotnie oraz nieistotne dane wpływające na koszt dzienny operowania dźwignicy lub na zużycie paliwa i energii elektrycznej. Zauważono również okresowość kosztów. Wykorzystanie map korelacji pozwoliło określić jak silnie konkretne zmienne są ze sobą powiązane relacją liniową.

W ramach inżynierii cech usunięte zostały niepotrzebne kolumny oraz utworzone zostały nowe zmienne kategoryczne związane z datą danego pomiaru. Dodatkowo uzupełniono brakujące dane, aby cały zestaw był bardziej kompletny.

Dane są teraz lepiej przygotowane do dalszych etapów takich jak dobór modelu uczenia maszynowego i jego trening. W zależności od przyjętego celu, możliwe jest przetestowanie różnych algorytmów uczenia nadzorowanego lub nienadzorowanego.

7. Źródła

https://dane.gov.pl/pl/dataset/3089,dane-z-testow-dzwignicy, dostęp 02.05.2025