Predykcja danych o wypadkach na podstawie danych pogodowych

MSiD Lab K01-21c

Mateusz Gazda

12 czerwca, 2023

1. Wstęp

Problemem wybranym do badań jest zależność danych meteorologicznych od danych o wypadkach drogowych w Nowym Jorku. Inspiracją do tych badań jest pytanie: "Czy na podstawie danych pogodowych i przedziału czasowego można predykować liczbę wypadków drogowych, liczbę rannych oraz ofiar śmiertelnych w Nowym Jorku?". Celem projektu jest zbadanie zależności danych o wypadkach:

- Liczba rannych
- Liczba ofiar śmiertelnych
- Liczba rannych pieszych
- Liczba ofiar śmiertelnych wśród pieszych
- Liczba rannych rowerzystów
- Liczba ofiar śmiertelnych wśród rowerzystów
- Liczba rannych kierowców
- Liczba ofiar śmiertelnych wśród kierowców
- Sumaryczna liczba wypadków

od danych meteorologicznych i czasu:

- Godzina
- Dzień tygodnia
- Temperatura
- Punkt rosy
- Wilgotność
- Prędkość wiatru
- Ciśnienie atmosferyczne
- Ilość opadów
- Warunki atmosferyczne

oraz stworzenie modelu predykującego powyższe dane o wypadkach.

2. Zbiór danych

2.1. Pozyskanie danych

Zbiór danych o wypadkach drogowych w Nowym Jorku pozyskano ze strony https://www.kaggle.com/datasets/muzammilrizvi1/motor-vehicle-collisions-crashes [1]. Dane zbierano od 2012-07-01 do 2023-02-25. Dane znajdujące się w tym zbiorze:

- TIMESTAMP - czas odczytu danych pogodowych
- BOROUGH - okręg administracyjny miejsca wypadku
- ZIP CODE - kod pocztowy miejsca wypadku
- LATITUDE - szerokość geograficzna miejsca wypadku
- LONGITUDE - długość geograficzna miejsca wypadku
- LOCATION - koordynaty geograficzne miejsca wypadku
- ON STREET NAME - nazwa ulicy, na której doszło do wypadku
- CROSS STREET NAME - nazwa najbliższej przecznicy w pobliżu miejsca wypadku
- OFF STREET NAME - adres ulicy miejsca wypadku
- NUMBER OF PERSONS INJURED - liczba osób rannych w wypadku
- NUMBER OF PERSONS KILLED - liczba ofiar śmiertelnych
- NUMBER OF PEDESTRIANS INJURED - liczba rannych przechodniów
- NUMBER OF PEDESTRIANS KILLED - liczba ofiar śmiertelnych wśród przechodniów

[1] Motor_Vehicle_Collisions_-_Crashes.csv

- NUMBER OF CYCLIST INJURED - liczba rannych rowerzystów
- NUMBER OF CYCLIST KILLED - liczba ofiar śmiertelnych wśród rowerzystów
- NUMBER OF MOTORIST INJURED - liczba rannych kierowców
- NUMBER OF MOTORIST KILLED - liczba ofiar śmiertelnych wśród kierowców
- CONTRIBUTING FACTOR VEHICLE n - czynnik powodujący wypadek dla danego pojazdu
- COLLISION ID - identyfikator wypadku
- VEHICLE TYPE CODE n - typ danego pojazdu

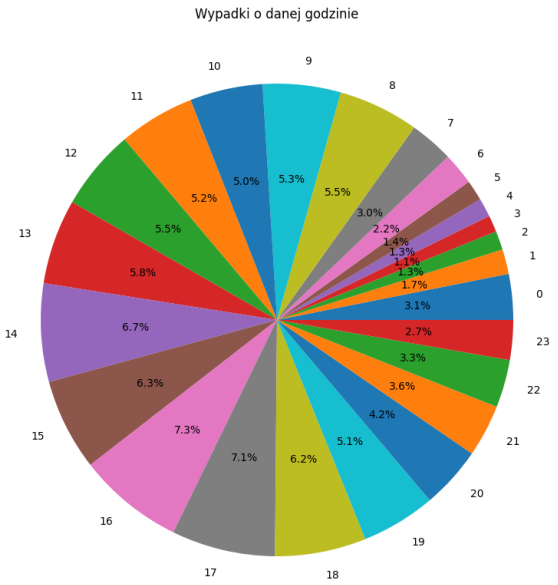
Dane meteorologiczne pozyskano ze strony https://www.wunderground.com/weather/us/ny/new-york-city za pomocą skryptu scrapującego stronę [2]. Dane zawarte w zbiorze:

- Timestamp - czas odczytu danych pogodowych
- Temperature - temperatura w stopniach Celsjusza
- Dew Point - punkt rosy w stopniach Celsjusza
- Humidity - wilgotność w %
- Wind Speed - prędkość wiatru w km/h
- Pressure - ciśnienie atmosferyczne w hPa
- Precip. - opady w mm
- Condition - warunki atmosferyczne

2.2. Wstępne przetwarzanie danych

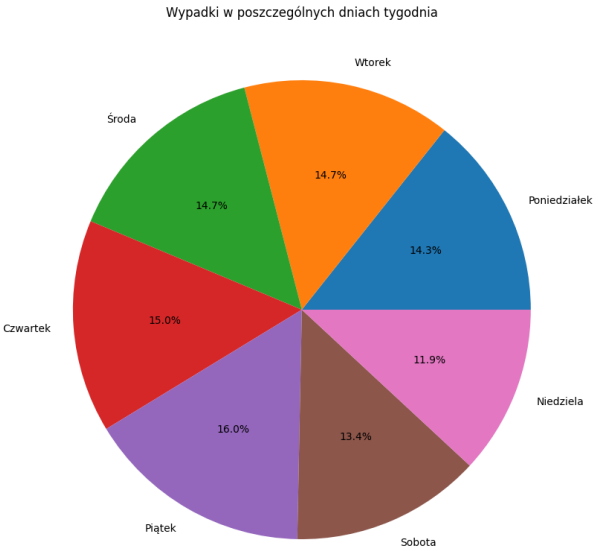
W zbiorze danych o wypadkach drogowych usunięto zbędne kolumny niepotrzebne w analizowanym problemie[3]. Jedynymi brakującymi danymi w podzbiorze były 'NUMBER OF PERSONS INJURED' oraz 'NUMBER OF PERSONS KILLED'. Dane te uzupełniono sumując odpowiednio pozostałe dane o rannych oraz ofiarach śmiertelnych. W celu rozwiązania problemu dane o wypadkach zostały zgrupowane według daty i przedziału godzinowego wynoszącego jedną godzinę. Następnie wartości w kolumnach dla danej grupy zostały zsumowane i dostawiona została dodatkowa kolumna 'NUMBER OF ACCIDENTS' zawierająca liczbę wypadków dla danego przedziału czasowego. W tak powstałym zbiorze danych na podstawie 'Timestamp' została dodana kolumna 'Hour' zawierająca początkową godzinę przedziału czasowego oraz kolumna 'Day of the week' zawierająca numer dnia tygodnia. Wartości z kolumny 'Condition' zostały zamienione na wartości liczbowe i wstawione do kolumny 'Condition code'. Kolumny 'Timestamp' i 'Condition' zostały usunięte.

3. Wstępna analiza danych

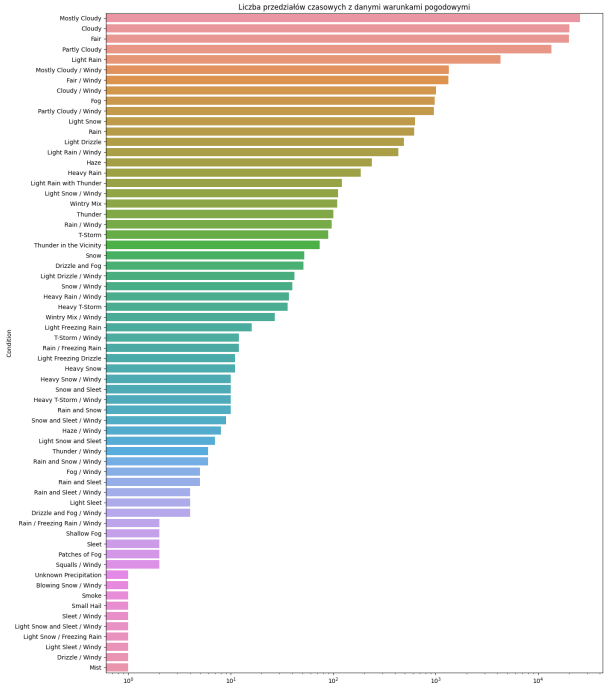


Na powyższym wykresie znajdują się procentowe udziały wypadków o danej godzinie w całym zbiorze. Możemy zauważyć, że najwięcej wypadków przypada na godzinny poranne oraz popołudniowe, kiedy to ruch jest wzmożony w przeciwieństwie do godzin nocnych.

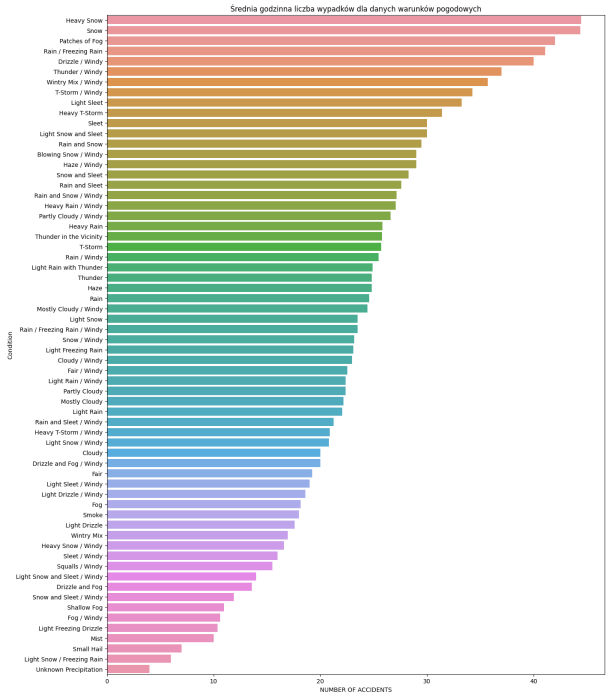
[2] weather_data_scraper.ipynb
[3] raport_jupyter.ipynb



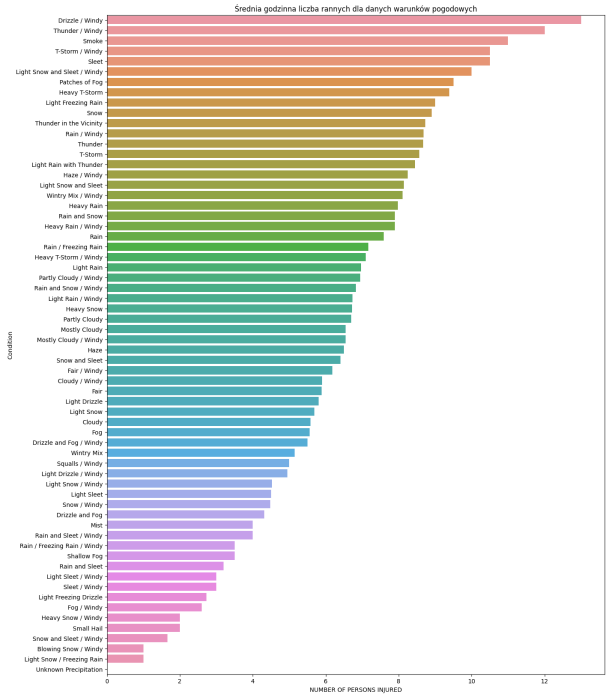
Powyższy wykres przedstawia rozłożenie liczby wypadków w ciągu całego tygodnia. Możemy zauważyć, że jest ono w przybliżeniu równomierne z lekkim nasileniem w piątek oraz spadkiem w niedziele.



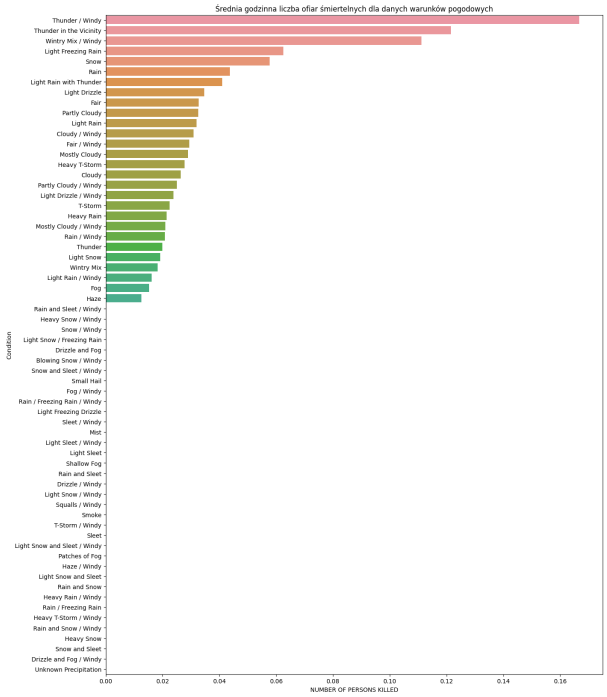
Na podstawie powyższego wykresu możemy zauważyć, że warunki meteorologiczne w zbiorze danych nie występują w równomiernych liczbach. Dominują odczyty z godzinami pochmurnymi, warunkami umiarkowanymi oraz lekkimi opadami.



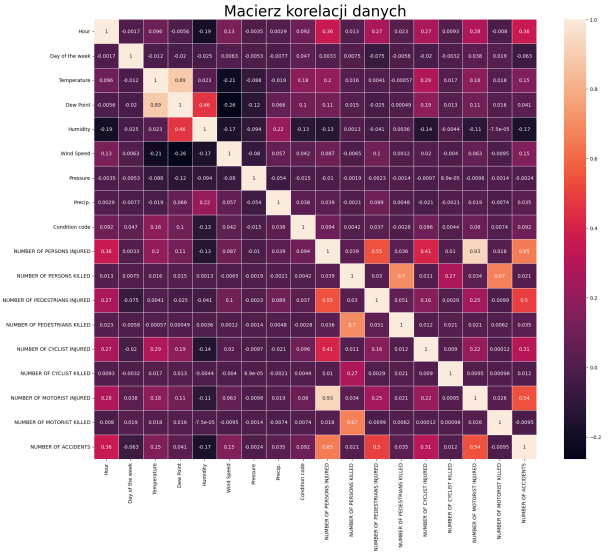
Średnio na godzinę najwięcej wypadków przypada dla odczytów meteorologicznych zawierających opady śniegu i marznącego deszczu oraz innych nieprzyjazdnych warunków związanych z opadami. Odczyty z godzinami pochmurnymi i warunkami umiarkowanymi, których jest najwięcej w zbiorze danych, zawierają średnie wartości. Najmniejsze średnie godzinne liczby wypadków przypadają na lekkie opady.



Średnie godzinne liczby rannych rozkładają się podobnie do średnich godzinnych liczb wypadków. Największa średnia przypada dla mżawki z wiatrem. Tutaj również dominują warunki związane z opadami i burzami.



W przeciwieństwie do średnich godzinnych liczb wypadków oraz rannych znaczące wartości średniej godzinnej liczby ofiar śmiertelnych przypadają tylko dla części warunków atmosferycznych, są one również o rząd mniejsze od dwóch poprzednich parametrów. Na powyższym wykresie znacząco wyróżniaja się burza z piorunami w połącznieniu z wiatrem. Warunkami z najwięszymi wartościami średniej są również marznący deszcz oraz inne warunki zimowe.



Z macierzy korelacji danych możemy wyczytać, że największe zależności danych występują pomiędzy wielkościami, które chcemy estymować, oraz pomiędzy wielkościami na podstawie, których chcemy predykować. Najwięszke zależności między parametrami należącymi do różnych zbiorów to godzina, temperatura oraz liczba poszczególnych rannych, liczba wypadków.

4. Modele

W celu przygotowania modelów dane zostały podzielone na treningowe i testowe w stosunku 3:1.

Skorzystano z modelu regresji liniowej ‘LinearRegression‘ oraz z ‘PolynomialFeatures‘ z biblioteki ‘scikit-learn‘ w celu stworzenia uogólnionego modelu regresji. Model regresji liniowej zakłada liniową zależność między zmiennymi niezależnymi, a zmienną zależną. Model ‘LinearRegression‘ dopasowuje prostą regresji liniowej do danych treningowych za pomocą metody najmniejszych kwadratów. Dzięki wykorzystaniu ‘PolynomialFeatures‘ zamiast prostej, możemy dopasowywać wielomiany wyższych stopni.

Najpierw dobrano odpowiedni stopień wielomianu, dla którego otrzymano najlepsze wyniki. Za początkowy wskaźnik oceny modelu posłużył współczynnik determinacji R^2, który pokazuje jak dobrze model jest dopasowany do danych, czym bliżej wartości 1.0 tym lepiej.

Dla modeli predykujących liczbę poszczególnych ofiar śmiertelnych nie otrzymano satysfakcjonujących wyników dla żadnego stopnia wielomianu. W dalszej części pominięto te dane, uznając je za niemożliwe do predykowania na podstawie posiadanych danych. Dla reszty danych najlepsze współczynniki R^2 otrzymano dla wielomianu drugiego

stopnia, dalsze obliczenia kontynuowano dla tego właśnie stopnia.

Dodatkowymi parametrami oceny modelu były:

- MSE (mean squared error) - błąd średniokwadratowy, MSE jest wartością oczekiwaną kwadratu błędu, czyli różnicy między estymatorem a wartością estymowaną.
- MAE (mean absolute error) - średni błąd bezwzględny, średnia bezwzględna różnica między przewidywaniem modelu a wartością docelową.
- MAPE (mean absolute percentage error) - średni bezwzględny błąd procentowy, średnia wielkość błędów prognoz wyrażona w procentach

W celu próby poprawy wyników parametr zawierający godzinę zamieniono na parametr zawierający porę dnia (6 wartości). Jak można zauważyć wyniki w niewielkim stopniu się polepszyły.

W celu dalszego polepszenia wyników zredukowano liczbę kategorii pogodowych. Zauważono, że wiele kategorii występuje w dwóch wersjach, różniących się dopiskiem 'Windy'. Dodano kolumnę 'Is windy', w której znajduje się wartość 1 jeśli w kategorii danego wpisu występuje 'Windy' oraz 0 w przeciwnym wypadku. Następnie usunięto dopiski 'Windy' z kategorii pogodowych i przerobiono je spowrotem na wartości numeryczne. Zredukowanie liczby kategorii nieznacznie pogorszyło jakość modelów.

Następnie sprawdzono jak skalowanie danych wpłynie na jakość modeli. Do tego celu użyto StandardScaler oraz MinMaxScaler z biblioteki scikit-learn. Skalowanie danych nie wpłynęło na jakość modelów.

Następnie dla tak zmodyfikowanych danych sprawdzono jak brak poszczególnego parametru wpłynie na jakość modelów.

Nawet dla najlepszego modelu (model prognozujący godzinną liczbę wypadków bez korzystania z temperatury) średni błąd bezwzględny prognozowanej wielkości stanowi około 42% średniej wartośći tej wielkości.

Pomimo zabiegów zwiększających precyzję modelów, po wynikach wskaźników oceniających modele, należy stwierdzić, że na podstawie posiadanych danych pogodowych nie jesteśmy wstanie trafnie przewidywać danych o wypadkach.

5. Wnioski

W wyniku badań, wielkości:

- Liczba ofiar śmiertelnych
- Liczba ofiar śmiertelnych wśród pieszych
- Liczba ofiar śmiertelnych wśród rowerzystów
- Liczba ofiar śmiertelnych wśród kierowców

uznano za niemożliwe do predykowania na podstawie posiadanych danych meteorologicznych. Wielkości te zależą od wielu czynników oraz ich udział w zbiorze danych jest zbyt mały.

Dla pozostałych wielkości otrzymano lepsze wyniki, jednakże były one dalekie od satysfakcjonujących. Parametrem mającym największy pozytywny wpływ na jakość modelu okazała się pora dnia. Najlepszy model uzyskano dla predykcji liczby wypadków po usunięciu parametru związanego z temperaturą. Niemniej jednak jego współczynnik R^2 wyniósł jedynie 0.363918, a średni błąd absolutny stanowił około 42% wartości średniej estymowanej wielkości.

W świetle badań należy przyjąć, że warunki meteorologiczne mają wpływ na liczbę rannych oraz wypadków, nie są one jednak wystarczające do trafnego estymowania tych wielkości. Większy wpływ na te wielkości ma pora dnia.

Przyczyną tych wyników może być zależność danych o wypadkach od wielu innych czynników. Na jakość modelów wpłym mogła mieć również mała dokładność danych atmosferycznych, w szczególności zbyt duże przedziały czasowe pomiędzy pomiarami oraz lokalny i gwałtowny charakter niektórych warunków meteorologicznych. W przypadku predykcji liczby ofiar śmiertelnych, mały udział tej wielkości w ogólnej liczbie wpisów oraz zbyt skomplikowany charakter tego parametru mogły mieć kluczowy wpływ na wyniki.

Predicted data	1	2	3	4
NUMBER OF PERSONS INJURED	0.174865	0.180920	-0.135320	-874.261164
NUMBER OF PERSONS KILLED	0.000369	-0.001052	-2.823215	-173.065275
NUMBER OF PEDESTRIANS INJURED	0.096076	0.125276	-0.060100	-660.566032
NUMBER OF PEDESTRIANS KILLED	0.001195	-0.010803	-3.194683	-70.581429
NUMBER OF CYCLIST INJURED	0.156451	0.159790	-1.516552	-55.502386
NUMBER OF CYCLIST KILLED	0.000055	-0.001180	-1.465744	-161.829083
NUMBER OF MOTORIST INJURED	0.114844	0.114614	-0.051060	-631.072011
NUMBER OF MOTORIST KILLED	0.000551	-0.000429	-0.068240	-86.416565
NUMBER OF ACCIDENTS	0.183575	0.330818	-0.716101	-49.843611

Tabela 1: Wartości R^2 dla poszczególnych stopni wielomianu

Predicted data	R^2	MSE	MAE	MAPE
NUMBER OF PERSONS INJURED	0.180920	18.613267	3.266432	1.299319e+15
NUMBER OF PEDESTRIANS INJURED	0.125276	2.047888	0.997222	1.812917e+15
NUMBER OF CYCLIST INJURED	0.159790	0.619836	0.568059	1.254049e+15
NUMBER OF MOTORIST INJURED	0.114614	13.302406	2.756289	1.849806e+15
NUMBER OF ACCIDENTS	0.330818	138.443287	9.307358	8.878276e-01

Tabela 2: Ocena modelów uogólnionej regresji wielomianu 2 stopnia

Predicted data	R^2	MSE	MAE	MAPE
NUMBER OF PERSONS INJURED	0.196020	18.270124	3.222515	1.214965e+15
NUMBER OF PEDESTRIANS INJURED	0.133840	2.027836	0.984308	1.767050e+15
NUMBER OF CYCLIST INJURED	0.164005	0.616726	0.564249	1.246057e+15
NUMBER OF MOTORIST INJURED	0.124229	13.157934	2.734353	1.788558e+15
NUMBER OF ACCIDENTS	0.361155	132.167051	8.912290	7.841644e-01

Tabela 3: Ocena modelów po wprowadzeniu pory dnia

Predicted data	R^2	MSE	MAE	MAPE
NUMBER OF PERSONS INJURED	0.189968	18.407652	3.223606	1.215275e+15
NUMBER OF PEDESTRIANS INJURED	0.128599	2.040107	0.984621	1.767171e+15
NUMBER OF CYCLIST INJURED	0.163809	0.616871	0.564221	1.245626e+15
NUMBER OF MOTORIST INJURED	0.120199	13.218494	2.734905	1.791044e+15
NUMBER OF ACCIDENTS	0.355893	133.255614	8.914296	7.842595e-01

Tabela 4: Ocena modelów po redukcji liczby kategorii warunków pogodowych

Predicted data	Without parameter	R^2	MSE	MAE	MAPE
NUMBER OF ACCIDENTS	Temperature	0.363918	131.595363	8.906754	7.831794e-01
NUMBER OF CYCLIST INJURED	Humidity	0.177221	0.606977	0.563631	1.243361e+15
NUMBER OF MOTORIST INJURED	Precip.	0.133989	13.011300	2.733827	1.782298e+15
NUMBER OF PEDESTRIANS INJURED	Temperature	0.144362	2.003202	0.984191	1.769910e+15
NUMBER OF PERSONS INJURED	Temperature	0.207475	18.009821	3.219689	1.213232e+15

Tabela 5: Wyniki dla najlepszych modelów bez danego parametru