

# Predykcja wypadków samochodowych w Polsce w latach 2018-2023 na podstawie danych pogodowych

Jakub Śliwka 272549  
Informatyka Stosowana  
Wydział Informatyki i Telekomunikacji

10 Czerwiec 2022

## 1 Wstęp

Polska jest jednym z czołowych krajów w Europie pod względem liczby śmiertelnych wypadków samochodowych w przeliczeniu na milion mieszkańców. Pod tym względem tylko Rumunia, Bułgaria, Litwa i Chorwacja mają gorsze statystyki. Wypadki drogowe są jedną z głównych przyczyn zgonów w Polsce, dlatego celem projektu jest analiza wypadków samochodowych oraz stworzenie modelu zdolnego przewidzieć liczbę wypadków na podstawie specjlanych warunków. W kolejnych rozdziałach przybliżę temat wypadków samochodowych, przeanalizuję czynniki zewnętrzne takie pogoda, weekendy czy święta

## 2 Dane użyte w programie

Dane pogodowe używane w projekcie zostały pobrane ze strony <https://danepubliczne.imgw.pl> w formie *zip*. Program automatycznie rozpakowuje pliki oraz łączy je w jeden plik *csv*.

Dane o wypadkach zostały *zescrapowane* ze strony <https://policja.pl/>. Program wyszukuje strony z odpowiednimi datami i dzięki bibliotece *BeautifulSoup* wyszukuje tabelkę z odpowiednimi danymi, po czym łączy i zwraca *DataFrame*.

Dane o świętach w Polsce zostały *zescrapowane* ze strony <https://www.timeanddate.com/holidays/poland/>. Tutaj również została użyta biblioteka w Pythonie *BeautifulSoup* do wyszukania odpowiedniej tabeli zawierającej dane o świętach w Polsce.


Dane o weekendach zostały wyszukane za pomocą biblioteki *Pandas* przy pomocy *DataFrame*ów

## 3 Analiza danych

### 3.1 Wypadki drogowe

#### 3.1.1 Liczba wypadków


Analizę danych warto rozpocząć od samych wypadków drogowych.



visualization/accidents.png

Figure 1: Ilość wypadków w latach 2018-2023

Widząc wykres ?? od razu można zauważyć tendencję spadkową. Ponadto wykres przypomina sinusoidę. Bez żadnych dodatkowych informacji, można dostrzec, że w okresie letnim wypadki zdarzały się dużo częściej, niż w okresie zimowym. Sugeruje to, że w trudniejszych warunkach kierowcy są bardziej ostrożni.



visualization/accidents\_sin.png


Figure 2: Wykres liczby wypadków drogowych z linią regresji oraz funkcją sinusoidalną

Na wykresie ?? 0 oznacza 01.01.2018, natomiast końcowa wartość to 31.12.2023

Na wykresie ?? została dodana linia regresji oraz sinusoida dopasowana do wykresu wypadków. Jak widać wykres ma tendencję spadkową. Funkcja sinusoidalna w przybliżeniu prezentuje się wzorem:

$$y = -0.016x + 84 + 12\sin(0.017x - 231) \quad (1)$$

Warto jednak zbadać dokładniej wykres ?. Poniżej zostały zamieszczone wykresy dzielące okresy na lata oraz miesiące



visualization/accidents\_per\_year.png

Figure 3: Wykres wypadków drogowych podzielony na lata od 2018-2023

visualization/accidents\_per\_month.png

Figure 4: Wykres wypadków drogowych w miesiącach

Wykresy ??, ?? prezentują jak w następnych latach zmieniały się wypadki w Polsce. W 2020 roku na przełomie marca oraz kwietnia można dostrzec nagły spadek wypadków drogowych. Było to spowodowane lockdown'em wprowadzonym wówczas z powodu pandemii. Załamanie "Covidowe" widoczne jest również na wykresie ??. W latach 2022-2023 gdy obostrzenia "Covidowe" były powoli łagodzone, ilość wypadków nie zaczęła ulegać aż takiemu wzrostowi. Na ten fakt mógł mieć wpływ ingerencji rządu oraz zaostrzenia przepisów drogowych. Zmiany taryfikatorów drogowych mogły spowodować, zwiększenie ostrożności kierowców w obawie przed wysokimi mandatami.

Co równie ważne dobrze widoczne różnice o których mowa była wcześniej, czyli różnice w porach roku lato/zima

visualization/accidents\_days\_of\_week.png

Figure 5: Wykres wypadków drogowych w dniach tygodnia

visualization/normal\_vs\_holidays\_vs\_weekends.png

Figure 6: Wypadki drogowe w poszczególnych rodzajach dni

Z wykresu ?? widać, że średnio najwięcej wypadków wypadło w piątek. Może to wynikać z początku weekendu w czasie którym wiele osób planuje wyjazdy. Minimalne załamanie widać w Niedzielę. Może to być spowodowane kilkoma czynnikami, m.in. tym, że niedziela jest uważana za ważny dzień z powodu chrześcijańskiego charakteru Polaków lub "odpoczynkiem" przed nadchodzącym tygodniem pracy. Aż tak dużego znaczenia nie ma natomiast rodzaj dni. Nie ma znacznej różnicy między dnami powszednimi, świętami a weekendami.

### 3.1.2 Liczba poszkodowanych oraz śmierci

Warto jest zająć się kwestią ilości poszkodowanych oraz śmierci w wypadkach samochodowych. Obie te kwestie są mocno powiązane z ilością wypadków omawianych w poprzedniej sekcji.

visualization/injured\_sin.png

Figure 7: Wykres poszkodowanych w wypadkach drogowych

Wykres ?? wygląda podobnie do wykresu ??. Zauważalny jest również "Covidowe" załamanie. Wzór który można dopasować do wykresu jest podobny to poprzedniego.

$$y = -0.020x + 100 + 22 \sin(0.017x - 235) \quad (2)$$

visualization/deaths.png

Figure 8: Wykres śmierci w wypadkach drogowych

Wykres śmierci ?? również zależny jest od ilości wypadków. Jednak wspólnym czynnikiem który wpływa pośrednio lub bezpośrednio na wypadki, poszkodowanych i śmierci jest pora roku. Co za tym idzie pogoda. Warto zatem przyjrzeć się bliżej pogodzie w latach 2018-2023, być może to właśnie ona miała większy lub mniejszy wpływ na wypadki drogowe.

## 3.2 Pogoda

### 3.2.1 Temperatura powietrza

Na poprzednich wykresach można było zauważyć korelację pomiędzy porami roku, a wypadkami samochodowymi. Dlatego warto najpierw przyjrzeć się temperaturze.

visualization/avg\_temp.png

Figure 9: Temperatura w Polsce w latach 2018-2023

Widząc wykres średniej temperatury dobowej w Polsce można zauważyć korelację między temperaturą dobową, a ilością wypadków w Polsce ???. Przyglądając się obu wykresom można wysnuć wniosek, że pora roku a właściwie temperatura ma wpływ na ilość wypadków.

visualization/temp\_vs\_accidents.png

Figure 10: Temperatura oraz ilość średnia liczba wypadków

visualization/temp\_vs\_accidents\_regress.png

Figure 11: Temperatura, ilość średnia liczba wypadków z linią regresji

Potwierdzenie tego wniosku można znaleźć w wielu źródłach np. <https://www.auto-swiat.pl/porady/kiedy-jest-najwiecej-wypadkow-na-drogach-odpowiedz-cie-zaskoczy/2gvz064>. Podczas trudniejszych warunków pogodowych, droga wymaga od kierowców ciągłego skupienia oraz koncentracji. Natomiast dobra pogoda usypia czujność kierowców. Gdy widoczność jest dobra, nie pada deszcz, a ruch jest jednostajny, koncentracja kierowców maleje z czasem.

### 3.2.2 Opady atmosferyczne

Warto przypatrzeć się również opadom atmosferycznym. Co statystyki mówią o wypadkach podczas deszczu?

visualization/precips.png

Figure 12: Opady atmosferyczne w latach 2018-2023

visualization/normal\_vs\_rainy.png

Figure 13: Wypadki podczas pogody oraz opadów

Wykres ?? prezentuje opady, na którym trudno doszukiwać się jakiegokolwiek korelacji.

Dużo ciekawszy jest natomiast wykres ?? na którym widać minimalną różnicę między średnią ilością wypadków w pogodne dni, a deszczowe. Wykresy ?? oraz ?? potwierdzają wniosek wysnuty w poprzednim rozdziale, że pogoda ma wpływ na ilość wypadków.

Czy rodzaj opadów wpływa na ilość wypadków? Takie pytanie może nasunąć się widząc wykres ???. Warto zatem porównać te sobą deszcz oraz śnieg.

visualization/rain\_vs\_snow.png

Figure 14: Wypadki podczas pogody, deszczu, śniegu

Jak widać opady śniegu mocno odbiegają od opadów deszczu. Średnia ilość wypadków jest prawie taka sama jak w dni pogodne.

Na podstawie danych pogodowych można dojść do wniosku, że ładna pogoda sprzyja wypadkom. Natomiast podczas pory zimowej gdy droga wymaga ciągłego skupienia ilość wypadków jest znacznie mniejsza.

Jak zostało powiedziane, wiele czynników wpływa na wypadki samochodowe. Jednak czy za pomocą danych jakimi są zmienna pogoda, święta oraz weekendy, da się przewidzieć ilość wypadków, rannych oraz śmierci w Polsce? To zagadnienie zostanie omówione w kolejnym rozdziale.

## 4 Predykcja wypadków w Polsce

Moim celem jest przewidzenie dokładnej ilości wypadków przy odpowiednich zadanych warunkach. Dlatego wybór padł na *Regressor*'y

### 4.1 Support Vector Regression

Support Vector Regression (SVR) jest rodzajem Support Vector Machine (SVM).

Ponieważ SVR jest algorytmem opartym na odległości, skalowanie jest ważnym etapem przetwarzania wstępnego, który może poprawić dokładność i stabilność modelu. Dane które przyjmuje SVR powinny być wstępnie przetworzone, przeskalowane, ponieważ model bazuje na odległości. Do skalowania użyty zostały Standard-Scaler. Moduł ten skaluje dane tak, aby ich średnia wynosiła 0, a odchylenie standardowe wynosiło 1.

SVR oryginalnie został stworzony jako single-output. Aby "przerobić" SVR na multioutput można było użyć narzędzi dostępnych w Sklearn: *MultiOutputRegressor* lub *RegressionChain*.

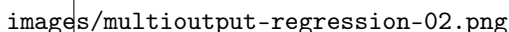


Figure 15: [www.geeksforgeeks.org](http://www.geeksforgeeks.org)

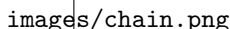


Figure 16: [www.geeksforgeeks.org](http://www.geeksforgeeks.org)

W projekcie nie została użyta żadna z powyższych metod. W przypadku MultiOutputRegressorów ciężko było o metrykę oceniającą każde wyjście z osobna. Natomiast w przypadku ChainRegression score  $R^2$  wychodził poniżej 0.1. Można powiedzieć, że w projekcie została zastosowana metoda ???. Tworząc 3 osobne modele dla ilości wypadków, rannych czy śmierci. Idea jest ta sama jednak takie rozwiązanie pozwoliło mi oceniać każde z wyjść z osobna.

Ogólny wynik modelu jest dosyć słaby. Score  $R^2$  dla liczby wypadków wskazywał 0.34, dla poszkodowanych 0.31, natomiast dla liczby śmierci 0.06. Po wielu testach jakim był poddawany SVR można z całą pewnością dojść do wniosku, że liczba śmierci na drogach jest praktycznie do nie przewidzenia. Liczba wypadków w niektórych przypadkach przewidywana jest niemal dokładnie, jednak model w większości zaniżał dane.

Parametry dla których SVR działa najlepiej to: kernel="rbf", C=4, gamma=0.1

### 4.2 Random Forest Regression

Random Forest to metoda uczenia się zespołowego, która łączy predykcje z wielu drzew decyzyjnych w celu uzyskania dokładniejszych i stabilnych wyników. Random Forest używa się zarówno do Regresji jak i Klasyfikacji.

Random Forest składa się z wielu Drzew Decyzyjnych. Każde drzewo ma tak zwane bootstrap samples. Bootstrap samples to podzbiór oryginalnego zbioru. Podzbiór ten to losowo wybierane dane. Ważne jest żeby podzbiór ten finalnie zawierał również powtórzenia. Oczekuje się 63,2% unikalnych danych.

Dla każdego subsetu wybiera się również podzbiór niepowtarzalnych cech. Każde drzewo jest trenowane. Wynik w przypadku regresji jest średnią, natomiast w przypadku klasyfikacji wybierana jest odpowiedź występująca najczęściej. Cały proces nazywany jest baggingiem

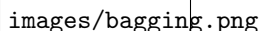


Figure 17: [en.wikipedia.org](http://en.wikipedia.org)

Random Forest w przeciwieństwie do SVR nie potrzebuje Feature Scaling. Model ten oferuje wsparcie w MultiOutput. Jednak potencjał ten nie został wykorzystany.

Cechy na których Random Forest działał najlepiej to: średnia temperatura, opady deszczu, opady śniegu, czy weekend, czy święto, miesiąc, dzień tygodnia.

Początkowe testy (bez miesiąca i dnia tygodnia) z MultiOutput'em wypadały praktycznie nierzadnie. Przewidywania praktycznie nie działały. Rozwiązaniem okazało się rozdzielenie na 3 osobne modele przewidujące wypadki, poszkodowanych, śmierci oraz dodanie miesiąca i dnia tygodnia w którym doszło do wypadku. Score  $R^2$  dla ilość wypadków oscylował w okolicach 0.44. Wynik podobnie jak w przypadku SVR dla poszkodowanych jest nieco niższy 0.38. Natomiast dla śmierci wynik oscylował w okolicach 0.1.

Parametry dla których Rando Forest Regressor dział najlepiej to: max depth=5, n estimators=100

## 5 Zakończenie

Z przeprowadzonych testów można wywnioskować, że jest możliwa predykcja ilości wypadków samochodowych w Polsce. Żeby uzyskać lepsze rezultaty należy jeszcze bardziej uszczegółwić dane lub zebrać ich większą ilość. Modele radzą sobie również z predykcją osób poszkodowanych.