Projektowanie Zaawansowane - projekt zaliczeniowy

Michał Kotowski
Bartosz Górny
Dawid Bieńkowski
Bartek Turkosz

Czerwiec 2020

1 Założenia projektowe

1.1 Problem badawczy

Oszacowanie ilości wypadków drogowych w Polsce w latach 2019 - 20XX na podstawie danych z lat 2000-2018 przy wykorzystaniu matematycznych modeli rozwiązywania zadań.

1.2 Teoria opracowania

Po przeanalizowaniu założonego celu oraz pobranych danych doszliśmy do wniosku, iż realizowane zadanie ma charakter czysto regresyjny. Wybraliśmy odpowiednio dwie metody: regresja liniowa metodą najmniejszych kwadratów błędów i regresja wielomianowa wykorzystując aproksymację wielomianową średnio kwadratową.

- Regresja liniowa: rozwiązanie zadania przedstawioną metodą polega na wyznaczeniu linii trendu w postaci $y = ax+b$, gdzie a i b to współczynniki, których poszukujemy realizując minimalizacje podanej sumy:

$$S(a,b) = \sum_{i=1}^{n}[y_i - y(x_i)]^2 = \sum_{i=1}^{n}[y_i - (ax_i + b)]^2$$

Należy zaznaczyć, że różnice miedzy dokładnymi y_i oraz wartościami obliczonymi z równania prostej są podniesione do kwadratu, aby uniknąć możliwości, że będą się nawzajem znosiły na skutek różnicy znaków

- Regresja wielomianowa: jest to sposób obliczenia zależności między zmienną zależną a jedną lub więcej zmiennymi niezależnymi występującymi w wyższych potęgach. W przypadku jednej zmiennej niezależnej równianie regresji przyjmuję postać:

$$y^2 = a + b_1 * x + b_2 * x^2$$

Wykorzystując aproksymację średnio kwadratową, wykorzystując wzór:

$$E = \sum_{i=1}^{n} \left[y_i - (a_m x_i^m + a_{m-1} x_i^{m-1} + \ldots + a_1 x_i + a_0) \right]^2$$

należy wyznaczyć a_i funkcji E takich że pochodne cząstkowe względem a_i są równe 0. Poniżej przedstawiamy rozpisany proces wyznaczania a_i dla wielomianu kwadratowego.

$$E = \sum_{i=1}^{n} \left[y_i - (a_2 x_i^2 + a_1 x_i + a_0) \right]^2$$

$$\frac{\partial E}{\partial a_0} = -2 \sum_{i=1}^{n} (y_i - a_2 x_i^2 - a_1 x_i - a_0) = 0$$

$$\frac{\partial E}{\partial a_1} = -2 \sum_{i=1}^{n} (y_i - a_2 x_i^2 - a_1 x_i - a_0) x_i = 0$$

$$\frac{\partial E}{\partial a_2} = -2 \sum_{i=1}^{n} (y_i - a_2 x_i^2 - a_1 x_i - a_0) x_i^2 = 0$$

$$\left(\sum_{i=1}^{n} 1 \right) a_0 + \left(\sum_{i=1}^{n} x_i \right) a_1 + \left(\sum_{i=1}^{n} x_i^2 \right) a_2 = \sum_{i=1}^{n} y_i$$

$$\left(\sum_{i=1}^{n} x_i \right) a_0 + \left(\sum_{i=1}^{n} x_i^2 \right) a_1 + \left(\sum_{i=1}^{n} x_i^3 \right) a_2 = \sum_{i=1}^{n} x_i y_i$$

$$\left(\sum_{i=1}^{n} x_i^2 \right) a_0 + \left(\sum_{i=1}^{n} x_i^3 \right) a_1 + \left(\sum_{i=1}^{n} x_i^4 \right) a_2 = \sum_{i=1}^{n} x_i^2 y_i$$

1.3 Przedstawienie danych

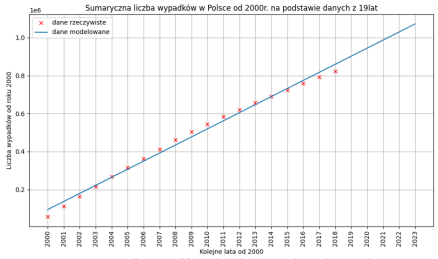
Dane zostały pobrane w formacie .csv ze strony Głównego urzędu statystycznego: *https://stat.gov.pl/* zatem przedstawiają one realne wartości dla analizowanego problemu. Poniżej zestawienie tabelaryczne.

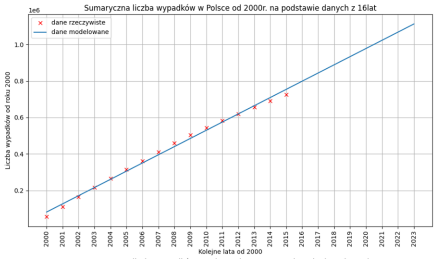
Tabela 1: Dane z GUS na temat ilości wypadków w Polsce w latach 2000-2018.

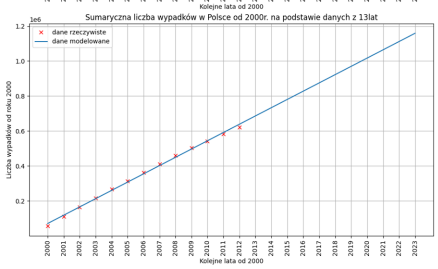
Lp.	*Rok*	*Dane*	*Sumarycznie*
1	2000	57 331	57 331
2	2001	53 799	111 130
3	2002	53 559	164 689
4	2003	51 078	215 767
5	2004	51 069	266 836
6	2005	48 100	314 936
7	2006	46 876	361 812
8	2007	49 536	411 348
9	2008	49 054	460 402
10	2009	44 196	504 598
11	2010	38 832	543 430
12	2011	40 131	583 561
13	2012	37 062	620 623
14	2013	35 847	656 470
15	2014	34 970	691 440
16	2015	32 967	724 407
17	2016	33 664	758 071
18	2017	32 760	790 831
19	2018	31 674	822 505

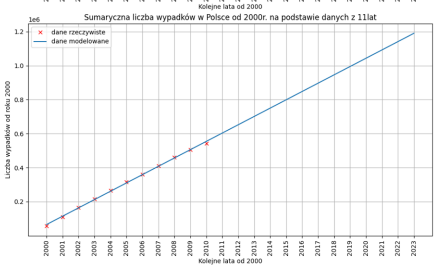
2 Przedstawienie wyników

2.1 Regresja liniowa - zrealizowana na podstawie kodu źródłowego na przeprowadzonych zajęciach

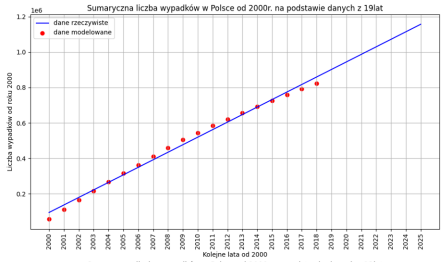


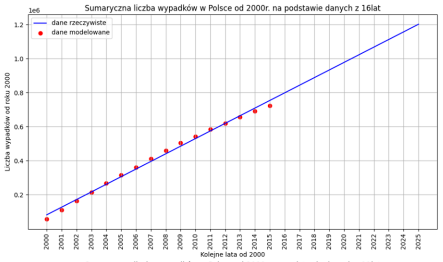


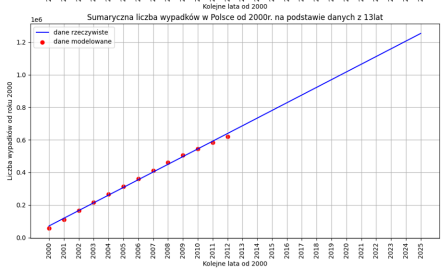


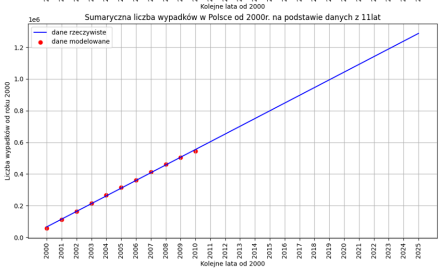


2.2 Regresja liniowa - zrealizowana przy wykorzystaniu biblioteki sklearn

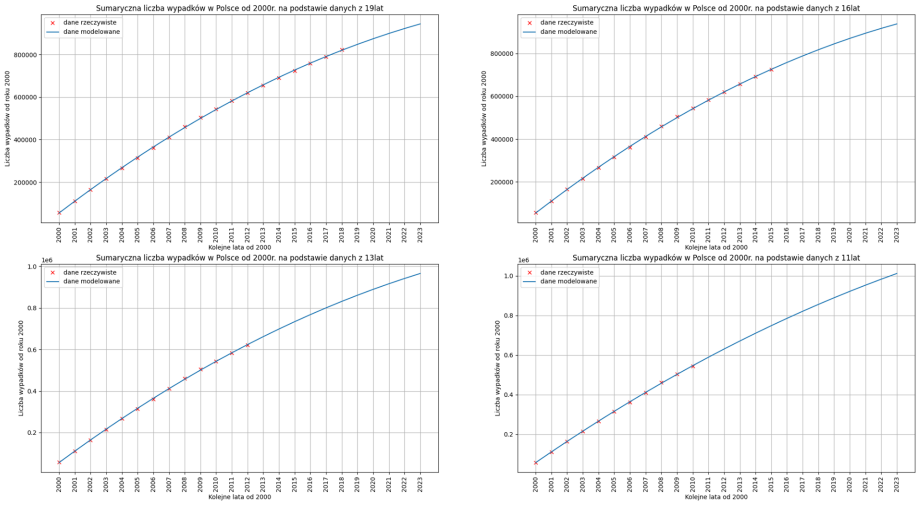




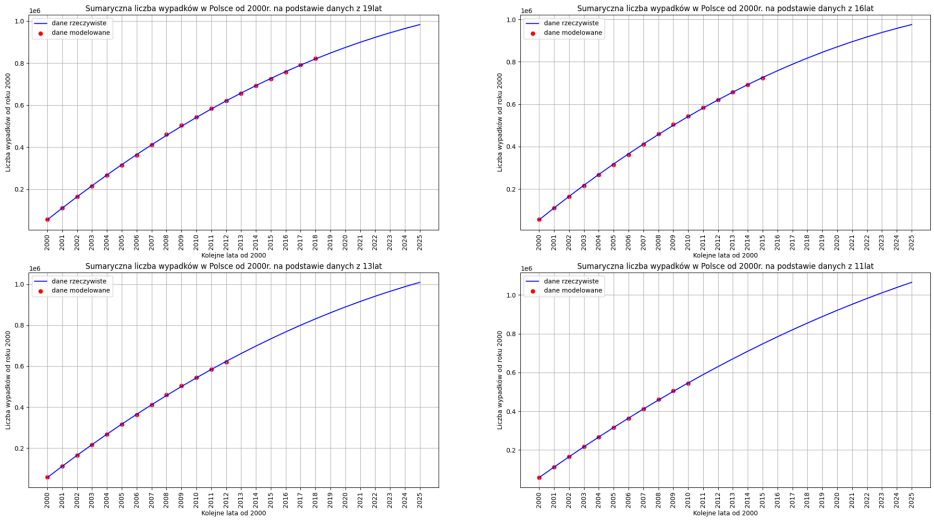




2.3 Regresja wielomianowa - zrealizowana na podstawie kodu źródłowego na przeprowadzonych zajęciach



2.4 Regresja wielomianowa - (kwadratowa) zrealizowana przy wykorzystaniu biblioteki sklearn



3 Wnioski

Ze względu na charakter badanej wartości oraz niewielką ilość zmiennych, po przeprowadzonej analizie można wywnioskować iż modelem przedstawiającym wyniki najbardziej zbliżone to rzeczywistych jest model regresji wielomianu kwadratowego, zarówno ten realizowany za pomocą biblioteki sklearn jak również algorytm pisany ręcznie (obie funkcje nieznacznie się różnią) . W dalszych latach można zaobserwować spłaszczenie wykresu, wartości nie rosną już tak gwałtownie. To oznacza ze model regresji liniowej, im dalej w przyszłość tym z większą niedokładnością szacowane są dane, natomiast wielomian kwadratowy pozwala nam na zniwelowanie poziomy błędu. Ponadto należy stwierdzić, że wprowadzenie dodatkowych zmiennych niezależnych takich jak np. opady roczne, ilość pojazdów ogółem czy średnia temperatura pozwoliła by na doprecyzowanie modelu.

4 Podział zadań

Proces decyzyjny dotyczący realizowanego problemu projektowego realizowaliśmy poprzez tzw. burzę mózgów przez komunikator, następnie podzieliliśmy się pracą na dwa mniejsze zespoły.

- Michał Kotowski: analiza problemu i zebranie danych, regresja wielomianowa kod, przygotowanie dokumentacji
- Bartosz Górny: analiza problemu i zebranie danych, regresja wielomianowa kod, przygotowanie dokumentacji
- Dawid Bieńkowski: analiza problemu i zebranie danych, regresja liniowa kod
- Bartek Turkosz: analiza problemu i zebranie danych, regresja liniowa kod

5 Źródła:

$http://www.math.uni.wroc.pl/\ dpilarcz/dydaktyka/bio12/W_4.pdf$
$http://www.if.pw.edu.pl/\ agatka/numeryczne/wyklad_05.pdf$
$https://bdl.stat.gov.pl/BDL/metadane/cechy/2423$
$https://www.latex-tutorial.com/tutorials/pgfplotstable/$
$https://scikit-learn.org/stable/$