**Statystyki oraz predykcje dotyczące Środowiska**

Autorzy

*Paweł Tryfon [s19288]*

*Filip Dzięcioł [s18651]*

*Oskar Kielczyk [s20008]*

Spis treści

[Wprowadzenie 2](#_Toc93089465)

[Cel projektu 2](#_Toc93089466)

[Wybrana technologia 2](#_Toc93089467)

[Back-End: Flask (Flask RESTful) 2](#_Toc93089468)

[Front-End: TypeScript (React.js) 2](#_Toc93089469)

[Uczenie maszynowe: Python + Scikit-Learn 2](#_Toc93089470)

[Dane 3](#_Toc93089471)

[Zbiory danych 3](#_Toc93089472)

[Sposób przygotowania danych 3](#_Toc93089473)

[Dane przetworzone 3](#_Toc93089474)

[Metoda 5](#_Toc93089475)

[Parametry modeli ML 5](#_Toc93089476)

[Opis funkcjonalności 5](#_Toc93089477)

[Statystyki oraz predykcje dotyczące emisji CO2 w Polsce 5](#_Toc93089478)

[Statystyki oraz predykcje dotyczące emisji CO2 w dowolnym kraju świata 5](#_Toc93089479)

[Statystyki oraz predykcje dotyczące różnych współczynników na świecie 5](#_Toc93089480)

[Losowe ciekawostki na temat środowiska 5](#_Toc93089481)

[Podział ról 6](#_Toc93089482)

[Załączniki 6](#_Toc93089483)

# Wprowadzenie

## Cel projektu

Z każdym rokiem na świecie zwiększa się liczba emitowanych ton dwutlenku węgla (CO2) – samochody spalinowe, zwiększający się konsumpcjonizm oraz rosnąca populacja powodują negatywne dla skutki natury, tj. efekt cieplarniany, czy pogarszające się powietrze oraz jakość wód.

W celu zwiększenia uświadamiania użytkowników Internetu postanowiono stworzyć aplikację pozwalającą na pokazywanie statystyk oraz predykcji związanych ze środowiskiem, jak i działaniami ze strony ludzkości.

## Wybrana technologia

Ze względu na dziedzinę biznesową oraz technologiczną postanowiono wykorzystać następujące technologie:

### *Back-End: Flask (Flask RESTful)*

Aplikacji po stronie serwera została stworzona w języku Python. Do serwowania naszych danych został użyty framework Flask, wraz z rozszerzeniem Flask RESTful. Takie rozwiązanie zostało zastosowane z uwagi na zastosowanie aplikacji typu SPA po stronie przeglądarki.   
  
Serwer wystawia 6 końcówek:

* dane surowe dla Polski `GET /api/poland/raw`
* predykcja emisji dla Polski `GET /api/poland/predict/:year`
* dane surowe dla kontynentów `GET /api/continents/:continentId/raw`
* predykcja emisji dla kontynentu `GET /api/continents/:continentId/predict/:year`
* dane surowe dla innych krajów `GET /api/countries/:countryId/raw`
* predykcja emisji dla kraju `GET /api/countries/:countryId/predict/:year`

### *Front-End: TypeScript (React.js)*

Aplikacja po stronie przeglądarki została w stworzona w języku TypeScript. Do stworzenia komponentów został zastosowany framework React.js wraz z HTML i CSS. Takie rozwiązanie zostało zastosowane ze względu na duże możliwości oferowane przez React.js oraz dużą ilość bibliotek ułatwiających tworzenie widoków po stronie przeglądarki.

Aplikacja przeglądarkowa posiada 3 widoki:

* widok dla Polski `/poland`
* widok dla kontynentów `/continents`
* widok dla innych krajów `/countries`

### *Uczenie maszynowe: Python + Scikit-Learn*

Modele zostały przygotowane z wykorzystaniem języka Python. Wykorzystywane algorytmy pochodzą z biblioteki scikit-learn. Dodatkowo użyto biblioteki tj. „Pandas” do odczytania danych w formacie csv, czy „Pickle” do eksportu gotowego modelu. Taki stos technologiczny umożliwił wytrenowanie modelu w łatwy i szybki sposób.

Wypracowane modele:

- poland\_model.sv – Model utworzony za pomocą regresji liniowej, która świetnie dopasowała się do charakteru danych dot. Emisji CO2 w Polsce

- continents\_model.sv – Model utworzony za pomocą losowego lasu decyzyjnego, który okazał się być najskuteczniejszym algorytmem.

- world\_model.sv – Podobnie jak w przypadku kontynentów, model utworzony za pomocą losowego lasu decyzyjnego, który również okazał się być najskuteczniejszym algorytmem.

# Dane

## Zbiory danych

Na potrzeby projektu użyto otwartych zbiorów danych ze strony kaggle.com:

- World CO2 Emissions Analysis

- Country to Continent

Pierwszy ze zbiorów zawiera setki współczynników dotyczących danych z większości krajów świata zarówno zagregowane na różnych poziomach (np. kraje o wysokim dochodzie, Unia Europejska, czy Kraje Afryki Centralnej) zbierane w latach 1970-2020.

## Sposób przygotowania danych

Ze względu na wymogi uczenia maszynowego, aplikacji, jak i niską jakość danych wykorzystano metody pre-processingu danych przy wykorzystaniu *Pythona* oraz bibliotek tj.:

- *Pandas* – przygotowanie danych w formie tabel (*DataFrame*)

- *Scikit-Learn* – zmiana wartości kategorycznych na numeryczne (*LabelEncoder*)

- *GoogleTrans* – interfejs tłumaczący wiadomości ze świata pobrane z pierwszego zbioru danych

- inne biblioteki, tj. *os*, czy *re*

W związku z niską jakością danych pod kątem kontynentów należało pobrać drugi zestaw danych, który pozwoli na wyodrębnienie kontynentów.

## Dane przetworzone

Ostatecznie powstało wiele wyczyszczonych zbiorów danych:

* *df\_poland\_co2* – dane emisji CO2 dla Polski w latach 1970-2012.
* *df\_continents\_co2* – dane emisji CO2 dla każdego kraju w latach 1970-2012.
* *df\_continents\_co2\_grouped* – dane emicji CO2 dla każdego kraju w latach 1970-2012 wzbogacone o indexy oraz nazwy kontynentów.
* *df\_continents\_co2\_grouped\_by\_continent* – dane emisji CO2 zsumowane per kontynent dla lat 1970-2012.
* *df\_world\_indicators* – dane wszystkich współczynników dostępnych w zbiorze danych (50) wyczyszczone dla modelu uczenia maszynowego (predykcja danego wskaźnika na bazie innych).
* *df\_news* – przetłumaczone z języka angielskiego na polski ciekawostki dotyczące środowiska, które były dostępne w zbiorze danych.

W celu zachowania logiki aktualizacji danych kategorycznych na numeryczne powstały także słowniki index (dana numeryczna) – wartość (dana kategoryczna).

# Metoda

## Parametry modeli ML

Na potrzeby projektu zostały wytrenowane 3 modele ML:

* Model emisji CO2 w Polsce:

- parametry wejściowe: rok

- rezultat: emisja CO2 dla podanego parametru

- ze względu na prostą strukturę danych wykorzystano regresje liniową

* Model emisji CO2 z podziałem na regiony, w tym wypadku kontynenty:

- parametry wejściowe: rok, identyfikator regionu

- rezultat: emisja CO2 dla podanych parametrów

- ze względu na najwyższy współczynnik precyzji wybrano losowy las decyzyjny do którego przykazywany

- dokładność klasyfikacji wyniosła około 0.99 przy wielkości zbioru testowego ustawionego na 50%

* Model emisji CO2 na całym świecie:

- parametry wejściowe: rok, identyfikator kraju

- rezultat: emisja CO2 dla podanych parametrów oraz na podstawie dostępnych wskaźników ekonomicznych

- w tym modelu zdecydowano się na uczenie za pomocą sieci neuronowej.

## Opis funkcjonalności

### Statystyki oraz predykcje dotyczące emisji CO2 w Polsce

W związku z posiadaniem danych statystycznych dotyczących ilości emisji ton CO2 w Polsce w latach 1970-2012 postanowiono użyć prostej Regresji Liniowej w celu przewidzenia nieznanych przyszłych lat, a także możliwość wskazania znanych wartości z przeszłych lat.

Polska, jako kraj zleceniodawcy, została przygotowana jako osobne menu.

### Statystyki oraz predykcje dotyczące emisji CO2 w dowolnym kraju świata

Podobnie jak w przypadku Polski – możliwość sprawdzenia emisji dwutlenku węgla w dowolnym kraju świata w latach 1970-2012 wraz z przewidzeniem przyszłych lat.

### Statystyki oraz predykcje dotyczące różnych współczynników na świecie

Zebrane zostały statystyki z lat 1970-2020 dotyczące różnych wskaźników dla całego świata, co pozwoli na predykcję dowolnych innych czynników.

### Losowe ciekawostki na temat środowiska

W celu uświadamiania użytkowników na temat różnorakich procesów oraz faktów dotyczących środowiska pobrano, oczyszczono oraz przetłumaczono ciekawostki, które będą pojawiały się, w sposób losowo wybrany, na stronie.

# Podział ról

**Cały zespół**

- Dyskusje dotyczące zakresu projektu oraz wymagań funkcjonalnych

- Konsultacje z każdym uczestnikiem projektu (zapewnianie transparentności oraz kultury współpracy kros-funkcjonalnej)

- Przeglądy oraz akceptacja kodu pozostałych uczestników projektu

**Filip Dzięcioł**

- Zarządzanie projektem, zadaniami oraz organizacja prac

- Inżynieria Danych – znajdywanie zbiorów danych, pobieranie oraz przygotowanie ich na potrzeby projektu (dla modeli ML oraz aplikacji)

**Oskar Kielczyk**

- Dopasowanie wyczyszczonych danych na potrzeby algorytmów uczenia maszynowego

- Tworzenie modeli uczenia maszynowego, eksport oraz wystawianie modeli na potrzeby inferencji

**Paweł Tryfon**

- Tworzenie aplikacji od strony Front-End’u oraz Back-End’u

- Tworzenie oprawy wizualnej oraz interaktywnych wykresów

# Załączniki

1. Link do zespołu w Trello

<https://trello.com/b/o3iwqDO6/sml-g1-project>

1. Link do repozytorium Git

<https://github.com/s18651/sml-project/>

1. Zbiory danych

<https://www.kaggle.com/manchunhui/world-co2-emissions-analysis>

<https://www.kaggle.com/statchaitya/country-to-continent>