

**PROJEKT**

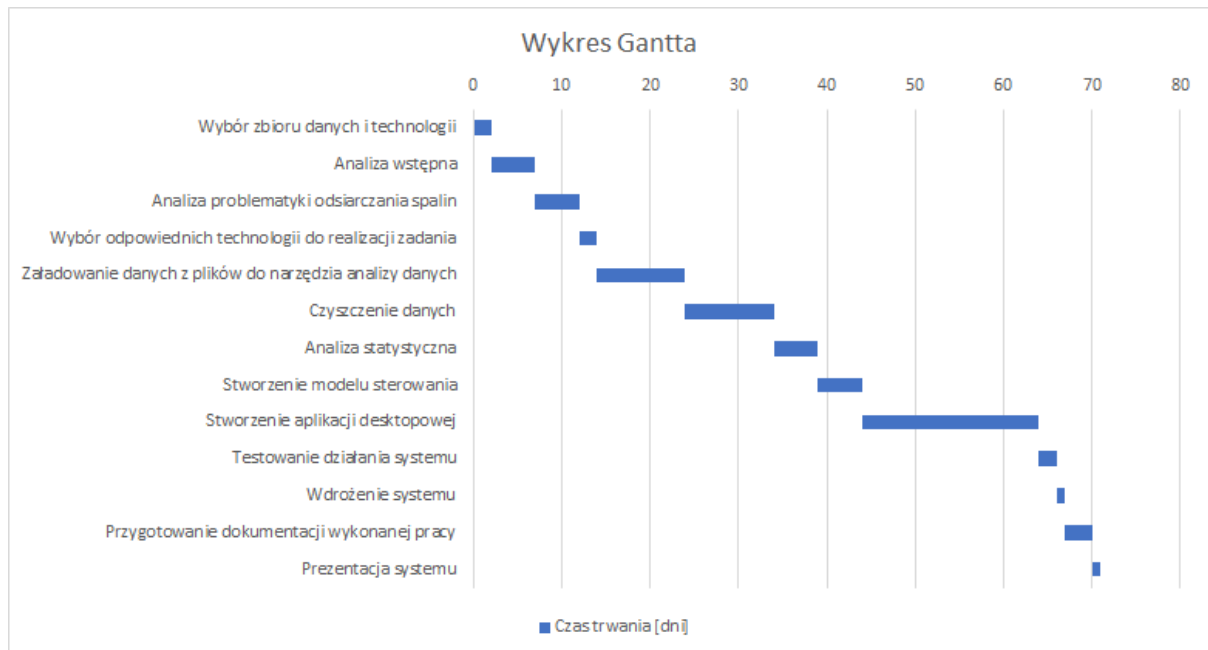
**—**

**SYSTEM STEROWANIA  
INSTALACJĄ ODSIARCZANIA  
SPALIN (IOS)**

<b>Zespół</b>
Marek Dalida
Maciej Gicala
Rafał Gęgotek
Aleksandra Bojęś

<b>1. Harmonogram Projektu</b>	<b>3</b>
<b>2. Podział ról i zadań</b>	<b>3</b>
<b>3. Schemat Komunikacji</b>	<b>4</b>
<b>4. Dobór narzędzi</b>	<b>4</b>
<b>5. Budżet</b>	<b>5</b>
<b>6. Analiza statystyczna</b>	<b>6</b>
<b>7. Porównanie modeli</b>	<b>7</b>
7.1 Wersja dla wytypowanych analitycznie parametrów	7
7.1.1 Decision Tree Regressor z wykorzystaniem PCA	7
7.2 Decision Tree Regressor bez PCA	9
7.3 Decision Tree Regressor z użyciem selekcji danych	9
7.4 Decision Tree Regressor z optymalizacją hiper parametrów	11
7.2 Wersja dla wyselekcjonowanych algorytmicznie parametrów	12
7.2.1 Model regresji Decision Tree	12
7.2.2 Model regresji Decision Tree ze zmienionymi parametrami	12
7.2.3 Model regresji Decision Tree dla 5 najlepszych parametrów	13
7.2.4 Model regresji za pomocą elastycznej siatki dla 5 najistotniejszych parametrów	14
7.2.5 Model regresji wykorzystujący algorytm k-najbliższych sąsiadów dla 5 najistotniejszych parametrów	14
7.2.6 Model regresji wykorzystujący drzewo decyzyjne dla 4 najlepszych parametrów	14
7.2.7 Model regresji wykorzystujący drzewo decyzyjne dla 3 najlepszych parametrów	15
7.2.8 Model k-najbliższych sąsiadów dla 3 najistotniejszych zmiennych	16
7.2.9 Model regresji Decision Tree dla 6 najistotniejszych zmiennych	16
7.2.10 Model regresji Decision Tree dla 7 najistotniejszych zmiennych	16
7.2.11 Model na podstawie drzewa decyzyjnego dla 3 wybranych zmiennych	17
<b>8. Opis modelu</b>	<b>18</b>
<b>9. Wyniki</b>	<b>18</b>
9.1 Aplikacja	18
9.2 Analiza wartości skrajnych	19
<b>10 Wnioski</b>	<b>21</b>

# 1. Harmonogram Projektu



## 2. Podział ról i zadań

Zadania:

a. Implementacja:

- i. Część backendowa:  
Aleksandra Bojęś  
Marek Dalida  
Rafał Gęgotek

- ii. Część frontendowa  
Maciej Gicala

b. Analiza

- i. Aleksandra Bojęś
- ii. Marek Dalida
- iii. Rafał Gęgotek

c. Dokumentacja

- i. Aleksandra Bojęś
- ii. Marek Dalida
- iii. Rafał Gęgotek
- iv. Maciej Gicala

d. Testowanie

- i. Marek Dalida

Role:

- Programista:
  - Aleksandra Bojęś
  - Marek Dalida
  - Rafał Gęgotek
  - Maciej Gicala
- Analityk:
  - Rafał Gęgotek
- Tester:
  - Marek Dalida

### 3. Schemat Komunikacji

Komunikacja uwzględnia korzystanie z aplikacji:

- Microsoft Teams
- repozytorium github
- tablice Trello

### 4. Dobór narzędzi

a. Języki programowania

Język	Zastosowanie
Python	analiza danych i stworzenie aplikacji desktopowej

b. Biblioteki i frameworki

Nazwa	Typ	Zastosowanie	Opis
scikit learn	biblioteka	analiza danych i stworzenie modelu	Biblioteka służąca do uczenia maszynowego
pandas	biblioteka	analiza danych i stworzenie modelu	Biblioteka ułatwiająca analizę i manipulację na danych
tkinter	biblioteka	aplikacja desktopowa	Biblioteka do tworzenia interfejsu graficznego dla języka python

c. Środowiska programistyczne

Projekt będzie tworzony z wykorzystaniem następujących środowisk programistycznych:

- Jupyter notebook

## 5. Budżet

Czynność	Czas trwania [dni]	Koszt roboczogodziny	Wycena [zł]
Wybór zbioru danych i technologii	2	100	1600
Analiza wstępna	5	100	4000
Analiza problematyki odsiarczania spalin	5	100	4000
Wybór odpowiednich technologii do realizacji zadania	2	100	1600
Łaďadowanie danych z plików do narzędzia analizy danych	10	100	8000
Czyszczenie danych	10	100	8000
Analiza statystyczna	5	100	4000
Stworzenie modelu	5	100	4000
Stworzenie aplikacji desktopowej	20	100	16000
Testowanie działania systemu	2	100	1600
Wdrożenie systemu	1	100	800
Przygotowanie dokumentacji wykonanej pracy	3	100	2400
Prezentacja systemu	1	100	800
<b>Podsumowanie</b>		<b>całkowity koszt:</b>	<b>56800</b>

## 6. Analiza statystyczna

Po załadowaniu danych do programu zostały one wyświetlone w programie w celu weryfikacji ich poprawności.

```
In [68]: splitDataToArray(dataKominSO2)

Out[68]: [['2019-12-01 00:00:00', '180'],
          ['2019-12-01 00:01:00', '165'],
          ['2019-12-01 00:02:00', '165'],
          ['2019-12-01 00:03:00', '165'],
          ['2019-12-01 00:04:00', '174'],
          ['2019-12-01 00:05:00', '172'],
          ['2019-12-01 00:06:00', '175'],
          ['2019-12-01 00:07:00', '168'],
          ['2019-12-01 00:08:00', '173'],
          ['2019-12-01 00:09:00', '175'],
          ['2019-12-01 00:10:00', '181'],
          ['2019-12-01 00:11:00', '188']]
```

Po analizie zauważyliśmy, że występuje około 17 minutowe przesunięcie zanim zanieczyszczenia pojawią się w na kominieSO2, dlatego czas w danych dla tej zmiennej został skorygowany o to przesunięcie.

```
In [77]: for i in range(len(dataKominSO2)):
          dateCol = dataKominSO2[i][0]
          epochColumn = mapToEpoch(dateCol)
          date_time = datetime.datetime.fromtimestamp(epochColumn)
          epochColumn += 60*17
          date_time_2 = datetime.datetime.fromtimestamp( epochColumn )
          print("DATE MOVED FROM " + str(date_time) + " TO " + str(date_time_2))
          dataKominSO2[i][0] = str(date_time_2)

DATE MOVED FROM 2019-12-01 00:00:00 TO 2019-12-01 00:17:00
```

Następnie zostało sprawdzone czy w zbiorze występują wartości brakujące

```
In [80]: #Sprawdzanie czy są wartości null
          df_merge_col.isnull().sum()

Out[80]: date                0
          data14HTA23CT001    0
          dataKominSO2        0
          data73HTA10CQ002    0
          data14HTA23CT002    0
          data14HTA21CF902    0
          data14HTQ21CF001    0
          14HTJ11CF912        0
          14HTJ10CL001        0
          dtype: int64
```

Kolejnym krokiem było sprawdzenie czy w zbiorze nie występują puste łańcuchy znaków oraz rzutowanie odczytów danych na typ zmiennoprzecinkowy.

Poniżej zostały przedstawione statystyki zbioru danych:

	date	data14HTA23CT001	dataKominSO2	data73HTA10CQ002	data14HTA23CT002	data14HTA21CF902	data14HTQ21CF001	14HTJ11CF912	14HTJ10CL001
count	4.462300e+04	44623.000000	44623.000000	44623.000000	44623.000000	44623.000000	44623.000000	44623.000000	44623.000000
mean	1.576494e+09	84.549089	193.242588	1323.601282	84.877776	273.374269	306.774421	1045.467046	41.576494
std	7.729017e+05	37.132035	247.572639	775.864865	37.264173	124.420114	196.624675	603.903381	2.729017
min	1.575156e+09	0.000000	0.000000	-199.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.575825e+09	99.000000	117.000000	1224.000000	99.000000	330.000000	221.000000	913.000000	41.575825
50%	1.576494e+09	100.000000	180.000000	1572.000000	101.000000	330.000000	294.000000	1038.000000	50.000000
75%	1.577164e+09	101.000000	210.000000	1871.000000	102.000000	330.000000	477.000000	1588.000000	60.000000
max	1.577833e+09	133.000000	2015.000000	2546.000000	132.000000	330.000000	1177.000000	2132.000000	71.577833

## 7. Porównanie modeli

### 7.1 Wersja dla wytypowanych analitycznie parametrów

Na podstawie ustaleń z prowadzącym wytypowano 7 podstawowych parametrów dla których miała zostać przeprowadzone rozwiązywanie problemu regresji.

Zmienną objaśnianą był poziom zanieczyszczenia z komina oznaczony jako *dataKominSO2*. Natomiast siedem zmiennych niezależnych dotyczyło parametrów:

- data14HTA23CT001
- dataKominSO2
- data73HTA10CQ002
- data14HTA23CT002
- data14HTA21CF902
- data14HTQ21CF001
- 14HTJ11CF912
- 14HTJ10CL001

#### 7.1.1 Decision Tree Regressor z wykorzystaniem PCA

Pierwszą czynnością jaka została wykonana po określeniu parametrów jakie miały zostać użyte w uczeniu maszynowym była redukcja wymiarowości z wykorzystaniem modelu PCA.

Analiza głównych składowych w skrócie **PCA** (Principal component analysis) jest jedną z najpopularniejszych metod ekstrakcji cech. Jest to algorytm, który redukuje wymiarowość danych poprzez rzutowanie ich na przestrzeń o mniejszej liczbie wymiarów.

Model został sprawdzony na 7 parametrach:

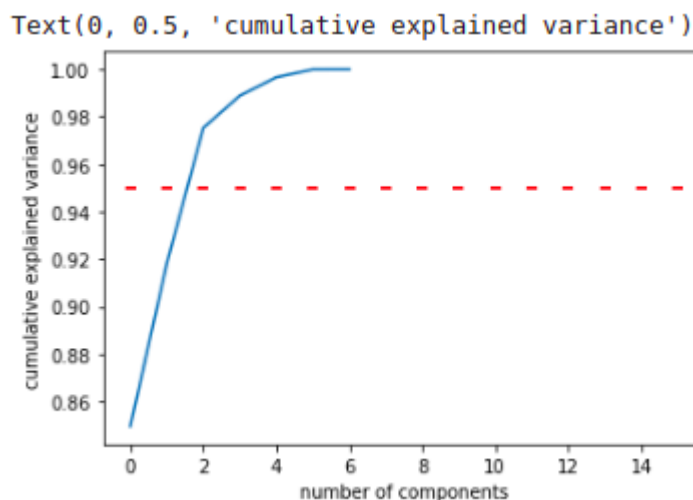
	data14HTA23CT001	data73HTA10C0002	data14HTA23CT002	data14HTA21CF902	\
0	0.759398	0.707832	0.765152	1.0	
1	0.759398	0.706375	0.765152	1.0	
2	0.759398	0.714026	0.765152	1.0	
3	0.759398	0.715847	0.765152	1.0	
4	0.759398	0.719490	0.765152	1.0	
...	...	...	...	...	
44618	0.744361	0.852823	0.757576	1.0	
44619	0.744361	0.854281	0.757576	1.0	
44620	0.744361	0.851730	0.757576	1.0	
44621	0.744361	0.859745	0.757576	1.0	
44622	0.744361	0.859745	0.757576	1.0	

	data14HT021CF001	14HTJ11CF912	14HTJ10CL001
0	0.316907	0.497655	0.615385
1	0.263381	0.497655	0.615385
2	0.255735	0.497655	0.615385
3	0.254036	0.497655	0.615385
4	0.251487	0.500000	0.615385
...	...	...	...
44618	0.271878	0.526735	0.756410
44619	0.265930	0.526735	0.756410
44620	0.271028	0.526735	0.756410
44621	0.267630	0.526735	0.756410
44622	0.265081	0.526735	0.756410

[44623 rows x 7 columns]

W ramach wyniku algorytmu możemy sprawdzić, jaka ilość wariancji jest wyjaśniona za pomocą k pierwszych składowych.



Tak jak widać na wykresie powyżej 95% wariancji jest wyjaśniane przez dwie najbardziej zależne składowe. W kolejnym etapie te dwie zmienne zostaną użyte do policzenia predykcji wykorzystaniem modelu regresji typu Decision Tree.

Jako sprawdzenie efektywności modelu został wykonana predykcja dla konkretnych danych testowych oraz sprawdzenie skuteczności modelu:

- miara skuteczności R2 = **0.55**

- predykcja:

dane wejściowe: [101.0 , 1744, 101.0, 330.0, 373, 1061, 48.0 ]

współczynniki dla parametrów wejściowych: [0.031, 1, 0.031, 0.1667, 0.1916, 0.5975, 0]

predykcja: 282.0



## 7.2 Decision Tree Regressor bez PCA

Zastosowanie modelu regresji typu Decision Tree dla wszystkich wcześniej wytypowanych parametrów dało w rezultacie wskazania:

- miara skuteczności  $R^2 = 0.72$

- predykcja:

dane wejściowe: [101.0 , 1744, 101.0, 330.0, 373, 1061, 48.0 ]

współczynniki dla parametrów wejściowych: [0.031, 1, 0.031, 0.1667, 0.1916, 0.5975, 0]

predykcja: 64.94221864304409

## 7.3 Decision Tree Regressor z użyciem selekcji danych

Na wstępie w tym etapie dane zostały przefiltrowane z wartości mniejszych niż 220 i dla parametru 'dataKominS02'.

```
[ ] 1 select_data = dataAll.loc[dataAll['dataKominS02'] < 220.0]
    2 print (select_data)
```

	date	data14HTA23CT001	dataKominS02	data73HTA10CQ002	\
0	1.575156e+09	101.0	180.0	1744.0	
1	1.575156e+09	101.0	165.0	1740.0	
2	1.575156e+09	101.0	165.0	1761.0	
3	1.575156e+09	101.0	165.0	1766.0	
4	1.575156e+09	101.0	174.0	1776.0	
...	...	...	...	...	
44597	1.577832e+09	101.0	197.0	2033.0	
44619	1.577833e+09	99.0	215.0	2146.0	
44620	1.577833e+09	99.0	215.0	2139.0	
44621	1.577833e+09	99.0	215.0	2161.0	
44622	1.577833e+09	99.0	197.0	2161.0	

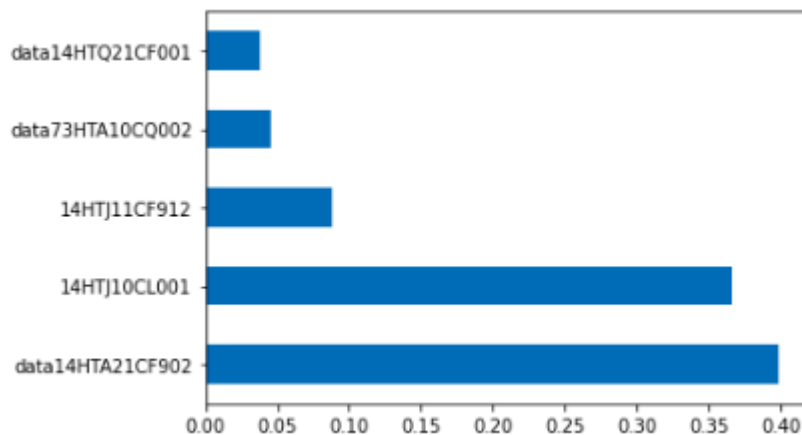
	data14HTA23CT002	data14HTA21CF902	data14HTQ21CF001	14HTJ11CF912	\
0	101.0	330.0	373.0	1061.0	
1	101.0	330.0	310.0	1061.0	
2	101.0	330.0	301.0	1061.0	
3	101.0	330.0	299.0	1061.0	
4	101.0	330.0	296.0	1066.0	
...	...	...	...	...	
44597	102.0	330.0	605.0	1889.0	
44619	100.0	330.0	313.0	1123.0	
44620	100.0	330.0	319.0	1123.0	
44621	100.0	330.0	315.0	1123.0	
44622	100.0	330.0	312.0	1123.0	

	14HTJ10CL001
0	48.0
1	48.0
2	48.0
3	48.0
4	48.0
...	...
44597	60.0
44619	59.0
44620	59.0
44621	59.0
44622	59.0

[36521 rows x 9 columns]

W następnym kroku przy użyciu Extra Tree Regressora została zbadana zależność, w jakim stopniu które parametry wpływają na końcową predykcję zmiennej objaśnianej. W efekcie ich wpływ został zobrazowany na poniższym wykresie (dla 5 najbardziej wpływowych cech).



Z powyższego wykresu można zauważyć, że największy wpływ na zmienną zależną mają zmienne data14HTF15CF001 i 14HTJ10CL001.

W kolejnych etapach regresja była obliczana dla różnej liczby najbardziej istotnych cech określonych przy pomocy algorytmu Extra Tree Regressor. Wyniki prezentują się następująco

- Dla 4 najbardziej istotnych parametrów:**  
 ['data14HTA21CF902', 'data14HTQ21CF001', '14HTJ11CF912', '14HTJ10CL001']  
 - data scaled mean: [0.81060212 0.40059644 0.45738349 0.58773421]  
 - data scaled std: [0.39182435 0.25780679 0.26509404 0.29408413]  
 - Mean squared error = 1484.46  
 - Median absolute error = 10.0  
 - Explain variance score = 0.74  
 - R2 score = 0.74
- Dla 5 najbardziej istotnych parametrów:**  
 ['data73HTA10CQ002', 'data14HTA21CF902', 'data14HTQ21CF001', '14HTJ11CF912', '14HTJ10CL001']  
 - data scaled mean: [0.53614701 0.81060212 0.40059644 0.45738349 0.58773421]  
 - data scaled std: [0.28562096 0.39182435 0.25780679 0.26509404 0.29408413]  
 - Mean squared error = 1189.26  
 - Median absolute error = 8.67  
 - Explain variance score = 0.79  
 - R2 score = 0.79
- Dla 6 najbardziej istotnych parametrów:**  
 ['data14HTA23CT001', 'data73HTA10CQ002', 'data14HTA21CF902', 'data14HTQ21CF001', '14HTJ11CF912', '14HTJ10CL001']  
 - data scaled mean: [0.642470 0.536147 0.810602 0.4005964 0.4573834 0.587734]  
 - data scaled std: [0.3102210 0.2856209 0.3918243 0.2578067 0.2650940 0.294084]  
 - Mean squared error = 848.94  
 - Median absolute error = 8.0  
 - Explain variance score = 0.85  
 - R2 score = 0.85

- **Dla 7 najbardziej istotnych parametrów:**

```
['data14HTA23CT001', 'data73HTA10CQ002', 'data14HTA23CT002',
'data14HTA21CF902', 'data14HTQ21CF001', '14HTJ11CF912', '14HTJ10CL001']
- data scaled mean: [0.64247 0.53614 0.650245 0.81060 0.40059 0.45738 0.58773]
- data scaled std: [0.31022 0.28562 0.313959 0.391824 0.257806 0.265094 0.29408]
- Mean squared error = 839.85
- Median absolute error = 8.0
- Explain variance score = 0.86
- R2 score = 0.86
```

Dla 7 najlepszych cech została również przeprowadzona predykcja na danych testowych:

```
- dane wejściowe: [101.0, 1744, 101.0, 330.0, 373, 1061, 48.0]
- współczynniki dla parametrów wejściowych:
  [0.0312 1. 0.03125 0.16627358 0.19162736 0.59728774 0.]
- predykcja: 1.7902551233793391
```

## 7.4 Decision Tree Regressor z optymalizacją hiper parametrów

Model danych użytych w poprzednim punkcie dla 7 najlepszych parametrów został wzbogacony o optymalizację hiper parametrów przy pomocy algorytmu GridSearch. Algorytm weryfikował wszystkie kombinacje dla wskazanych poniżej wartości odnoszących się parametrów składowych modelu regresji Decision Tree.

```
parameters={"splitter":["best","random"],
            "max_depth" : [1,3,5,7,9,11,12],
            "min_samples_leaf": [1,2,3,4,5,6,7,8,9,10],
            "min_weight_fraction_leaf": [0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9],
            "max_features": ["auto", "log2", "sqrt", None],
            "max_leaf_nodes": [None, 10, 20, 30, 40, 50, 60, 70, 80, 90] }
```

W rezultacie najlepsze ustawienie zawierało kombinacje:

```
'max_depth': 7,
'max_features': 'auto',
'max_leaf_nodes': None,
'min_samples_leaf': 1,
'min_weight_fraction_leaf': 0.1,
'splitter': 'best'
```

Dla takich ustawień model regresji uzyskał następujące wyniki:

```
- data scaled mean: [0.64247 0.53614 0.65024 0.8106 0.40059 0.457383 0.58773]
- data scaled std: [0.310221 0.28562 0.31395 0.391824 0.2578 0.265094 0.294084]
- Mean squared error = 1229.44
- Median absolute error = 14.84
- Explain variance score = 0.79
- R2 score = 0.79
```

## 7.2 Wersja dla wyselekcjonowanych algorytmicznie parametrów

Poniższe modele zostały przeprowadzone dla zbioru zmiennych, które zostały wyselekcjonowane na podstawie korelacji między parametrami.

Przy pomocy metody `corrcoef` z biblioteki `numpy` zostało wybrane 8 najlepszych parametrów, a są nimi:

```
['d14HTJ10CL001', 'd14HTA21CT002', 'd14HTA21CT001', 'd14QEA31CP001',  
'd14HTA23CT002', 'd14HTA23CT001', 'd14HTF15CF001', 'd14HTA21CP001']  
oraz zestaw 10 najlepszych z uwzględnieniem powyższych oraz dwóch jako:  
['d14HTJ11CF912', 'd73HTA10CQ002']
```

### 7.2.1 Model regresji Decision Tree

```
[0.58773421 0.75875893 0.76159124 0.73242245 0.65024537 0.6424704  
 0.72342445 0.3317823 ]  
[0.29408413 0.36648354 0.36784927 0.36378972 0.31395996 0.31022105  
 0.34948078 0.05141652]  
Mean squared error = 765.61  
Median absolute error = 8.0  
Explain variance score = 0.87  
R2 score = 0.87
```

Otrzymane wyniki można uznać za obiecujące. Wynik miary dopasowania R-kwadrat wyniósł aż 0,87, natomiast błąd średniokwadratowy wyniósł 765,61.

### 7.2.2 Model regresji Decision Tree ze zmienionymi parametrami

W tym modelu zdecydowaliśmy się zmodyfikować część parametrów dla drzewa decyzyjnego. Maksymalna głębokość drzewa została wyznaczona na 5 węzłów.

```
[0.55599615 0.7171747 0.71982142 0.69464844 0.61446056 0.6071404  
 0.68318584 0.33454694]  
[0.31565752 0.39598368 0.39744101 0.39179658 0.33915658 0.33513539  
 0.37727076 0.05410825]  
Mean squared error = 1372.83  
Median absolute error = 18.5  
Explain variance score = 0.75  
R2 score = 0.75
```

W tym przypadku można zaobserwować, że zmiana parametrów modelu nie polepszyła wyników. Wynik R-kwadrat pogorszył się i wyniósł 0,75, natomiast błąd średniokwadratowy również uległ pogorszeniu i wyniósł 1372,83.

### 7.2.3 Model regresji Decision Tree dla 5 najlepszych parametrów

```
[0.58773421 0.76159124 0.72342445 0.3317823 0.65024537]
[0.29408413 0.36784927 0.34948078 0.05141652 0.31395996]
Mean squared error = 821.74
Median absolute error = 8.0
Explain variance score = 0.86
R2 score = 0.86
```

Po powyższych wynikach można zaobserwować, że model z wykorzystaniem 5 najbardziej istotnych parametrów uzyskał dobre wyniki. Współczynnik dopasowania R-kwadrat wyniósł 0,86, natomiast błąd średniokwadratowy wyniósł 821,74. Można uznać, że utworzony model jest nieznacznie gorszy w porównaniu do modelu pełnego.

- Testowanie wyników metryki R-kwadrat dla różnych modeli z wykorzystaniem cross-walidacji

Poniżej zostały przedstawione wyniki r-kwadrat z wykorzystaniem cross-walidacji dla 5 różnych typów regresorów: regresji liniowej, regresji typu LASSO, regresji za pomocą siatki elastycznej, regresji z wykorzystaniem drzewa decyzyjnego, regresji za pomocą algorytmu k najbliższych sąsiadów i regresji Gradient Boosting. Wyświetlony został średni wynik R-kwadrat dla danej metody regresji, a w nawiasie znajduje się odchylenie standardowe.

Najlepsze 5 parametrów to: d14HTJ10CL001, d14HTA21CT001, d14HTF15CF001,

```
ScaledLR: 0.727496 (0.008463)
ScaledLASSO: 0.721400 (0.006558)

ScaledEN: 0.691094 (0.005567)
ScaledKNN: 0.877953 (0.006651)

ScaledCART: 0.854350 (0.007950)

ScaledGBM: 0.828925 (0.007327)
d14HTA21CP001, d14HTA23CT002
```

Z powyższych wyników można zauważyć, że najlepsze wyniki były otrzymywane dla regresji z wykorzystaniem algorytmu k-najbliższych sąsiadów średni wynik R-kwadrat 0,877. Najgorszy wynik miary dopasowania R-kwadrat był zwracany dla regresji metodą LASSO i wynosił średnio 0,69, można zauważyć że dla tej metody najmniejszą wartość miało także odchylenie standardowe.

### 7.2.4 Model regresji za pomocą elastycznej siatki dla 5 najistotniejszych parametrów

```
[17.92958426 24.67401647 23.63469973 -1.16665969 21.0242168 ]
75.02204096315248
Mean squared error = 2944.52
Median absolute error = 47.88
Explain variance score = 0.49
R2 score = 0.49
```

Otrzymane wyniki dla regresji z wykorzystaniem metody LASSO okazały się niezbyt zadowalające. Wynik miary dopasowania R-kwadrat wyniósł zaledwie 0,49, natomiast błąd średniokwadratowy wyniósł aż 2944,52.

### 7.2.5 Model regresji wykorzystujący algorytm k-najbliższych sąsiadów dla 5 najistotniejszych parametrów

```
Mean squared error = 519.19
Median absolute error = 6.6
Explain variance score = 0.91
R2 score = 0.91
```

Otrzymane wyniki dla algorytmu k-najbliższych sąsiadów z wykorzystaniem 5 najistotniejszych okazały się najlepszymi osiągniętymi dotychczas. Wynik miary dopasowania modelu R-kwadrat wyniósł 0,91, natomiast błąd średniokwadratowy osiągnął wartość 519,19.

### 7.2.6 Model regresji wykorzystujący drzewo decyzyjne dla 4 najlepszych parametrów

```
[0.58773421 0.76159124 0.72342445 0.3317823 ]
[0.29408413 0.36784927 0.34948078 0.05141652]
Mean squared error = 900.16
Median absolute error = 9.0
Explain variance score = 0.84
R2 score = 0.84
```

Otrzymane wyniki dla modelu z wykorzystaniem drzewa decyzyjnego dla 4 najistotniejszych zmiennych to R-kwadrat równy 0,84 i błąd średniokwadratowy równy 900,16. Widać, że wyniki znacząco pogorszyły się w porównaniu do modelu z wykorzystaniem drzewa decyzyjnego dla 5 parametrów.

### 7.2.7 Model regresji wykorzystujący drzewo decyzyjne dla 3 najlepszych parametrów

```
[0.58773421 0.76159124 0.72342445]
[0.29408413 0.36784927 0.34948078]
Mean squared error = 864.36
Median absolute error = 8.75
Explain variance score = 0.85
R2 score = 0.85
```

Otrzymane wyniki dla modelu z wykorzystaniem drzewa decyzyjnego dla 3 najistotniejszych zmiennych to R-kwadrat równy 0,85 i błąd średniokwadratowy równy 864,36. Widać po otrzymanych wynikach, że wraz ze zmniejszeniem ilości zmiennych wynik się polepszył w kontekście modelu z 4 parametrami, ale pogorszył w porównaniu do modelu z 5 parametrami.

- Testowanie wyników metryki R-kwadrat dla różnych modeli z wykorzystaniem cross-walidacji dla 3 najistotniejszych zmiennych

Poniżej zostały przedstawione wyniki r-kwadrat z wykorzystaniem cross-walidacji dla 3 różnych typów regresorów: regresji liniowej, regresji typu LASSO, regresji za pomocą siatki elastycznej, regresji z wykorzystaniem drzewa decyzyjnego, regresji za pomocą algorytmu k najbliższych sąsiadów i regresji Gradient Boosting. Wyświetlony został średni wynik R-kwadrat dla danej metody regresji, a w nawiasie znajduje się odchylenie standardowe.

Najistotniejsze parametry to: d14HTJ10CL001, d14HTA21CT001, d14HTF15CF001.

```
ScaledLR: 0.703161 (0.019028)
ScaledLASSO: 0.702500 (0.018804)
ScaledEN: 0.676492 (0.017230)

ScaledKNN: 0.853115 (0.010110)

ScaledCART: 0.837738 (0.015129)

ScaledGBM: 0.775346 (0.017289)
```

Z powyższych wyników można zauważyć, że najlepsze wyniki pod względem miary R-kwadrat daje algorytm k-najbliższych sąsiadów, ma on również najmniejsze odchylenie standardowe. Najgorsze wyniki dają regresje z wykorzystaniem metody LASSO oraz siatki elastycznej.

### 7.2.8 Model k-najbliższych sąsiadów dla 3 najistotniejszych zmiennych

```
Mean squared error = 814.69
Median absolute error = 8.8
Explain variance score = 0.86
R2 score = 0.86
```

Dla 3 najistotniejszych zmiennych model oparty o algorytm k najbliższych sąsiadów osiągnął wynik R-kwadrat = 0,86 i błąd średniokwadratowy 814,69. Wyniki można uznać za dobre jednak są one zauważalnie niższe niż w najlepszym modelu.

### 7.2.9 Model regresji Decision Tree dla 6 najistotniejszych zmiennych

Używane zmienne w modelu to: d14HTJ10CL001, d14HTA21CT001, d14HTF15CF001, d14HTA21CP001, d14HTA23CT002, d14HTA21CT002.

```
[0.55599615 0.71982142 0.68318584 0.33454694 0.61446056 0.7171747 ]
[0.31565752 0.39744101 0.37727076 0.05410825 0.33915658 0.39598368]
Mean squared error = 708.99
Median absolute error = 6.0
Explain variance score = 0.87
R2 score = 0.87
```

Otrzymany wynik R-kwadrat wyniósł 0,87 i można go uznać za dobry jednak okazał się on gorszy od najlepszego modelu z wykorzystaniem k-najbliższych sąsiadów. Błąd średniokwadratowy wyniósł 708,99.

### 7.2.10 Model regresji Decision Tree dla 7 najistotniejszych zmiennych

Używane zmienne w modelu to: d14HTJ10CL001, d14HTA21CT001, d14HTF15CF001, d14HTA21CP001, d14HTA23CT002, d14HTA21CT002, d14HTA23CT001.

```
[0.58773421 0.76159124 0.72342445 0.3317823 0.65024537 0.75875893
0.6424704 ]
[0.29408413 0.36784927 0.34948078 0.05141652 0.31395996 0.36648354
0.31022105]
Mean squared error = 754.0
Median absolute error = 8.0
Explain variance score = 0.87
R2 score = 0.87
```

Otrzymany wynik R-kwadrat wyniósł 0,87 i można go uznać za dobry jednak okazał się on gorszy od najlepszego modelu z wykorzystaniem k-najbliższych sąsiadów. Błąd średniokwadratowy wyniósł 754 czyli trochę gorzej niż w przypadku 6 najlepszych zmiennych.



### 7.2.11 Model na podstawie drzewa decyzyjnego dla 3 wybranych zmiennych

Wybrane zmienne to najlepsze 2 parametry z pierwszego modelu, na podstawie tych które podała prowadząca oraz dodatkowa najbardziej istotna z poprzedniego modelu.

```
[0.76605847 0.55599615 0.68318584]
[0.42333544 0.31565752 0.37727076]
Mean squared error = 1240.88
Median absolute error = 11.22
Explain variance score = 0.78
R2 score = 0.78
```

Wynik R-kwadrat wyniósł 0,78 co jest wynikiem gorszym w porównaniu do większości modeli opartych o drzewo decyzyjne. Błąd średniokwadratowy wyniósł 1240 co również patrząc z perspektywy pozostałych modeli jest raczej wynikiem przeciętnym.

### 7.2.12 Model końcowy z dodatkową zmienną

Na końcu do modelu końcowego (z punktu 7.2.5) dodano ilość dozowanego wapnia w celce na prośbę prowadzącej.

```
Mean squared error = 599.68
Median absolute error = 11.0
Explain variance score = 0.9
R2 score = 0.9
```

Wartość wskaźnika R-kwadra wyniosła tyle samo, błąd średniokwadratowy okazała się większy. Mode został odrzucony.

## 8. Opis modelu

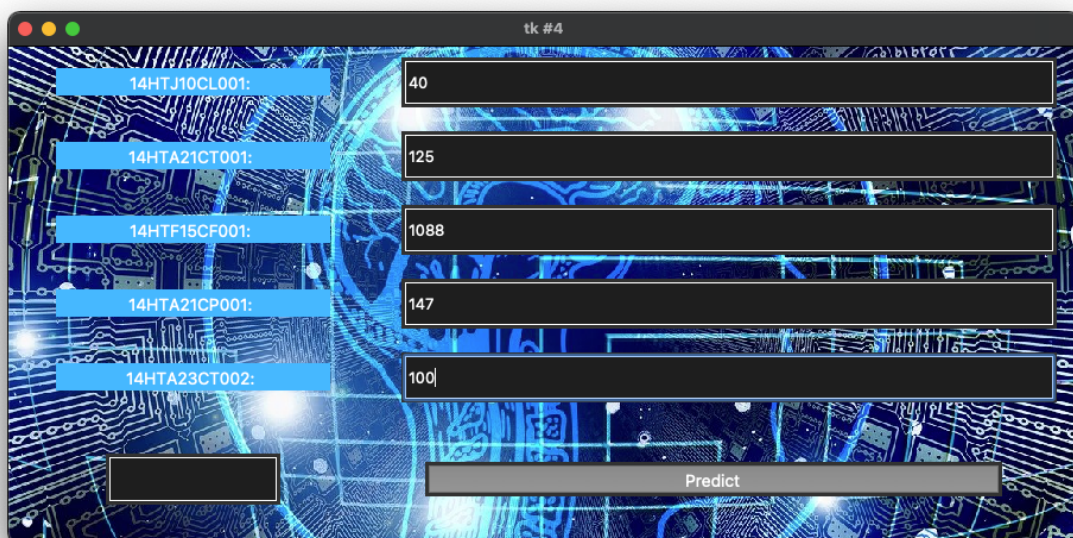
Najlepszym modelem z badanych okazał się ten oparty na algorytmie k najbliższych sąsiadów. Osiągnął on najlepsze wyniki zarówno biorąc pod uwagę współczynnik R-kwadrat, jak i błąd średniokwadratowy (wartości odpowiednio: 0,91 i 519,19). Najwyższą skuteczność uzyskiwał dla wybranych pięciu zmiennych niezależnych:

- 14HTJ10CL001
- 14HTA21CT001
- 14HTF15CF001
- 14HTA21CP001
- 14HTA23CT002

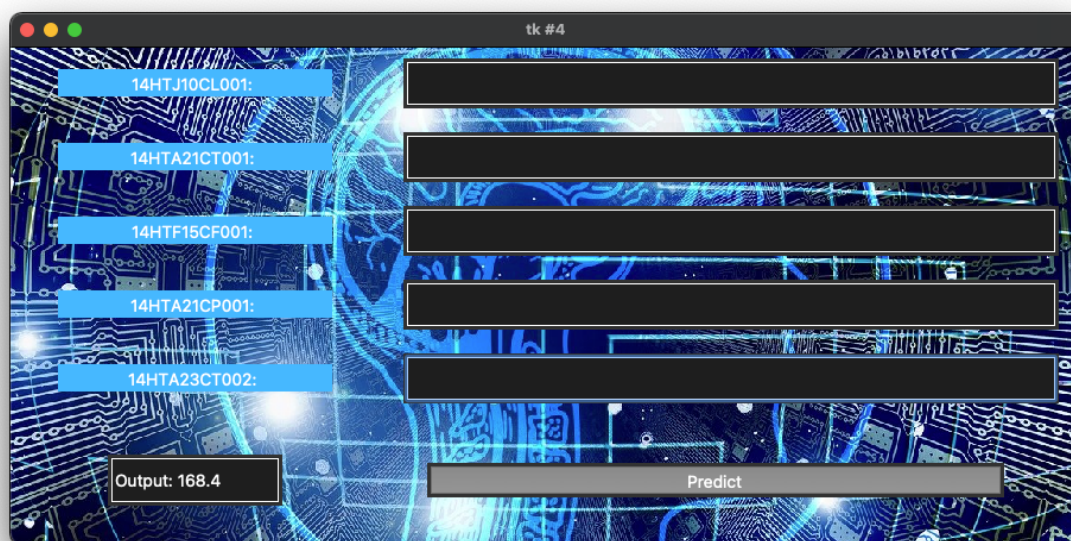
## 9. Wyniki

### a. Aplikacja

W celu ułatwienia użytkownikowi korzystania z modelu została stworzona aplikacja desktopowa. Umożliwia ona wprowadzenie konkretnych wartości dla zmiennych modelu, wykonuje obliczenia i wyświetla wartość zmiennej zależnej wyznaczoną przez model.



Label	Value
14HTJ10CL001:	40
14HTA21CT001:	125
14HTF15CF001:	1088
14HTA21CP001:	147
14HTA23CT002:	100



## b. Analiza wartości skrajnych

Dalsza ocena zbudowanego modelu polegała na wykorzystaniu stworzonej aplikacji do predykcji wartości zmiennej zależnej dla skrajnych wartości zmiennych niezależnych. Poniżej znajdują się statystyki opisowe wspomnianych zmiennych:

	d14HTJ10CL001	d14HTA21CT001	d14HTF15CF001	d14HTA21CP001	d14HTA23CT002
<b>count</b>	29486.000000	29486.000000	29486.000000	29486.000000	29486.000000
<b>mean</b>	43.367700	97.895713	760.385844	-83.351116	77.422031
<b>std</b>	24.621704	54.052894	419.909480	58.654337	42.734454
<b>min</b>	0.000000	0.000000	0.000000	-446.000000	0.000000
<b>25%</b>	42.000000	122.000000	932.000000	-130.000000	98.000000
<b>50%</b>	54.000000	127.000000	967.000000	-94.000000	101.000000
<b>75%</b>	60.000000	129.000000	988.000000	-38.000000	102.000000
<b>max</b>	78.000000	136.000000	1113.000000	638.000000	126.000000

i wartości zmiennej zależnej przewidywane dla wszystkich kombinacji skrajnych wartości zmiennych niezależnych:

	<b>d14HTJ10CL001</b>	<b>d14HTA21CT001</b>	<b>d14HTF15CF001</b>	<b>d14HTA21CP001</b>	<b>d14HTA23CT002</b>	<b>dKominSO2</b>
<b>1</b>	0.0	0.0	0.0	-446.0	0.0	194.8
<b>2</b>	0.0	0.0	0.0	-446.0	126.0	194.8
<b>3</b>	0.0	0.0	0.0	638.0	0.0	33.6
<b>4</b>	0.0	0.0	0.0	638.0	126.0	33.6
<b>5</b>	0.0	0.0	1113.0	-446.0	0.0	165.8
<b>6</b>	0.0	0.0	1113.0	-446.0	126.0	164.0
<b>7</b>	0.0	0.0	1113.0	638.0	0.0	33.6
<b>8</b>	0.0	0.0	1113.0	638.0	126.0	33.6
<b>9</b>	0.0	136.0	0.0	-446.0	0.0	194.8
<b>10</b>	0.0	136.0	0.0	-446.0	126.0	194.8
<b>11</b>	0.0	136.0	0.0	638.0	0.0	45.2
<b>12</b>	0.0	136.0	0.0	638.0	126.0	45.2
<b>13</b>	0.0	136.0	1113.0	-446.0	0.0	165.8
<b>14</b>	0.0	136.0	1113.0	-446.0	126.0	164.0
<b>15</b>	0.0	136.0	1113.0	638.0	0.0	39.8
<b>16</b>	0.0	136.0	1113.0	638.0	126.0	39.8
<b>17</b>	78.0	0.0	0.0	-446.0	0.0	194.8
<b>18</b>	78.0	0.0	0.0	-446.0	126.0	194.8
<b>19</b>	78.0	0.0	0.0	638.0	0.0	33.6
<b>20</b>	78.0	0.0	0.0	638.0	126.0	33.6
<b>21</b>	78.0	0.0	1113.0	-446.0	0.0	164.0
<b>22</b>	78.0	0.0	1113.0	-446.0	126.0	164.0
<b>23</b>	78.0	0.0	1113.0	638.0	0.0	33.6
<b>24</b>	78.0	0.0	1113.0	638.0	126.0	33.6
<b>25</b>	78.0	136.0	0.0	-446.0	0.0	194.8
<b>26</b>	78.0	136.0	0.0	-446.0	126.0	194.8
<b>27</b>	78.0	136.0	0.0	638.0	0.0	45.2
<b>28</b>	78.0	136.0	0.0	638.0	126.0	45.2
<b>29</b>	78.0	136.0	1113.0	-446.0	0.0	164.0
<b>30</b>	78.0	136.0	1113.0	-446.0	126.0	164.0
<b>31</b>	78.0	136.0	1113.0	638.0	0.0	39.8
<b>32</b>	78.0	136.0	1113.0	638.0	126.0	39.8

## 10. Wnioski

- W ramach projektu poddano analizie dane pochodzące z systemu sterowania instalacją odsiarczania spalin. Zbudowano różne modele regresji i porównano ich wyniki. Dla najlepszego z nich zbudowano aplikację desktopową ułatwiającą użytkownikom korzystanie z modelu i próby predykcji wartości stężenia siarki przy kominie.
- Najlepszy model udało się uzyskać dla modelu z wykorzystaniem algorytmu k-najbliższych sąsiadów oraz trenowanie dla 5 najbardziej istotnych zmiennych. Współczynnik dopasowania R-kwadrat wyniósł aż 0,91, natomiast błąd średniokwadratowy wyniósł ponad 500 i były to wartości najlepsze ze wszystkich modeli.
- Bardzo dobrze z predykcją stężenie na kominie radziły sobie modele oparte o model drzewa decyzyjnego, niestety żaden z tych modeli nie uzyskał tak dobrych wyników jak najlepszy model z wykorzystaniem algorytmu k-najlepszych sąsiadów. Wyniki współczynnika dopasowania R-kwadrat wynosiły mniej niż 0,90.
- Modelami, które najslabiej radziły sobie z predykcją stężenia na kominie były regresja metodą LASSO i metodą siatki elastycznej
- Projekt został zrealizowany zgodnie z zaplanowanym wcześniej harmonogramem. Jedyną zmianą dokonaną w trakcie trwania projektu, względem początkowych ustaleń, było zweryfikowanie wykorzystywanych technologii.