

Eksploracja danych - etap 2

Krzysztof Nasuta 193328, Filip Dawidowski 193433, Aleks Iwicki 193354

1. Charakterystyka zbioru

- Pochodzenie: [Kaggle](#)
- Liczba przykładów: 5819080
- Format: CSV (3 pliki: `flights.csv` - właściwy zbiór, `airports.csv` - informacje o lotniskach, `airlines.csv` - informacje o liniach lotniczych)
- Ilość zbiorów danych: 1

2. Wprowadzenie

Dataset: **2015 Flight Delays and Cancellations**

Cel: Budowa modelu predykcyjnego klasyfikującego opóźnienia lotów (`ARRIVAL_DELAY` > 15 minut)

Opóźnienia lotów mają znaczący wpływ na funkcjonowanie transportu lotniczego. Niniejszy projekt ma na celu stworzenie modelu uczenia maszynowego przewidującego opóźnienia.

Kluczowe pytania badawcze:

- Które czynniki najsilniej wpływają na opóźnienia?
- Który algorytm osiąga najlepsze wyniki?

3. Założenia wstępne

Podczas przewidywania opóźnień lotów nie będziemy uwzględniać informacji, które nie są dostępne w momencie planowania lotu, takich jak:

- `DEPARTURE_TIME` (nie mylić z `SCHEDULED_DEPARTURE`)
- `DEPARTURE_DELAY`
- `TAXI_OUT`
- `WHEELS_OFF`
- `ELAPSED_TIME`
- `AIR_TIME`
- `WHEELS_ON`
- `TAXI_IN`
- `ARRIVAL_TIME`
- `ARRIVAL_DELAY`

Spowoduje to znaczne obniżenie dokładności modeli, lecz pozwoli na realistyczne przewidywanie, ponieważ przy uwzględnieniu tych cech, modele osiągają niemal 100% dokładności.

4. Przygotowanie Danych

Źródła danych:

- `flights.csv`
- `airlines.csv`
- `airports.csv`

Kroki przetwarzania:

1. Ładowanie danych z pliku flights.csv z ograniczeniem do skonfigurowanej liczby rekordów
2. Definicja zmiennej celu: `DELAYED = 1` jeśli `ARRIVAL_DELAY > 15`
3. Podział na loty opóźnione i nieopóźnione.
4. Balansowanie zbioru danych - równa liczba opóźnionych i nieopóźnionych lotów. Pozostałe loty są usuwane.
5. Podział zbalansowanych danych na zbiór treningowy (80%) i testowy (20%)

Cechy wykorzystane w modelu:

- Kategoryczne: `AIRLINE`, `ORIGIN_AIRPORT`, `DESTINATION_AIRPORT`, `DAY_OF_WEEK`, `MONTH`
- Numeryczne: `YEAR`, `DAY`, `FLIGHT_NUMBER`, `SCHEDULED_DEPARTURE`, `SCHEDULED_TIME`, `DISTANCE`, `SCHEDULED_ARRIVAL`

Usunięte cechy: `DEPARTURE_TIME`, `DEPARTURE_DELAY`, `TAXI_OUT`, `WHEELS_OFF`, `ELAPSED_TIME`, `AIR_TIME`, `WHEELS_ON`, `TAXI_IN`, `ARRIVAL_TIME`, `ARRIVAL_DELAY`

5. Metodologia

Wykorzystane modele:

Wszystkie wykorzystywane modele zostały zaimplementowane w bibliotece `scikit-learn` dostępnej w języku Python. Poniżej przedstawiono klasyfikatory, które zostały użyte w projekcie:

Model	Implementacja
Drzewo Decyzyjne	<code>DecisionTreeClassifier</code>
Las Losowy	<code>RandomForestClassifier</code>
Regresja Logistyczna	<code>LogisticRegression</code>
K-NN	<code>KNeighborsClassifier</code>
Sieć Neuronowa	<code>MLPClassifier</code>

5.1. Początkowe porównanie modeli

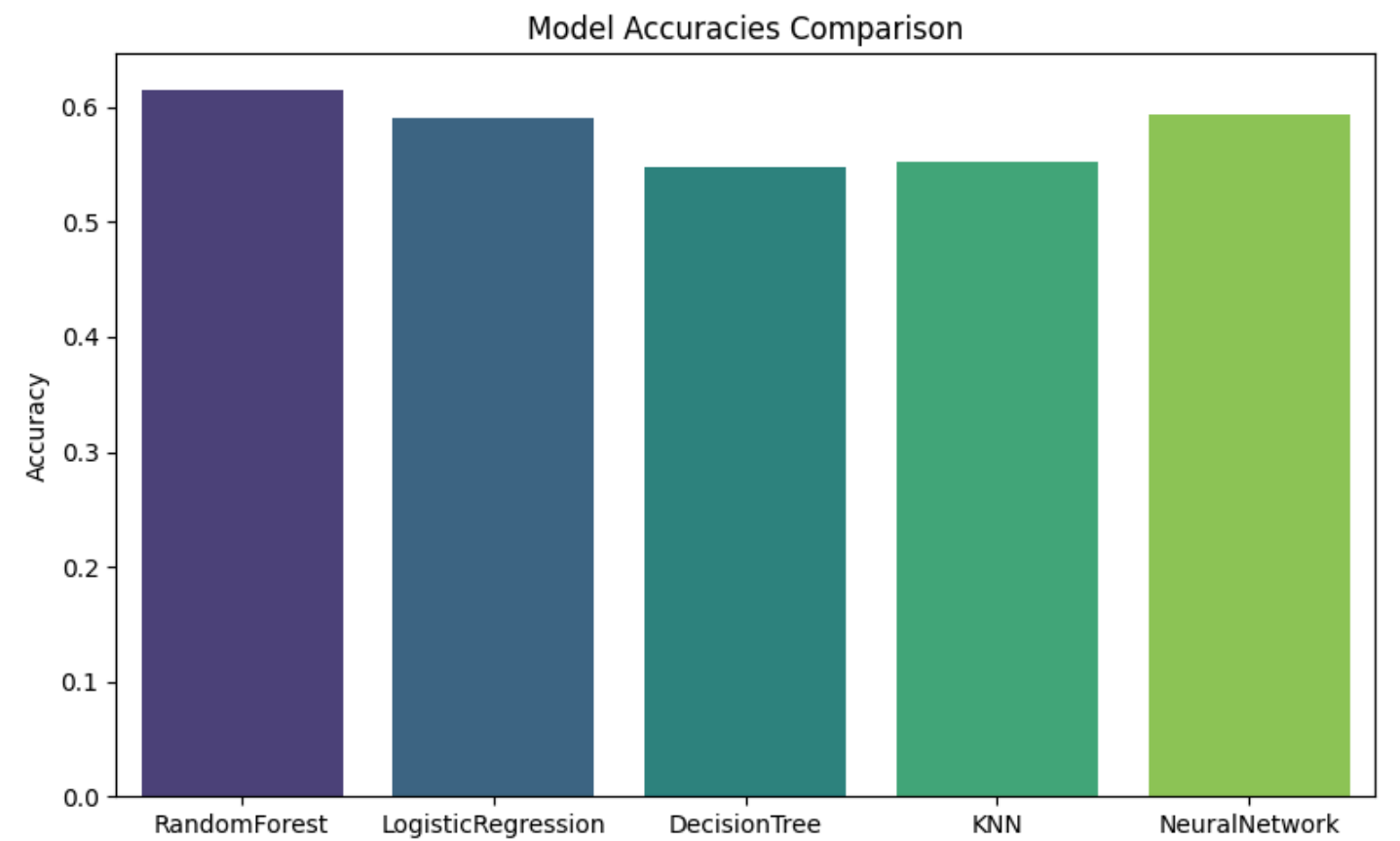
Pierwszym krokiem jest porównanie modeli z domyślnymi parametrami.

- Dla `RandomForest` utworzono 100 drzew, maksymalna głębokość nie jest ograniczona, a minimalna liczba próbek do podziału to 2.
- Dla `LogisticRegression` zastosowano domyślne parametry, z maksymalną liczbą iteracji równą 1000.
- Dla `DecisionTree` zastosowano domyślne parametry, z maksymalną głębokością nieograniczoną.
- Dla `KNN` zastosowano 5 sąsiadów i wagę równą „uniform” - każdy sąsiad ma równy wpływ na klasyfikację.
- Dla `NeuralNetwork` zastosowano dwie warstwy ukryte o rozmiarach 100 i 50, maksymalną liczbę iteracji równą 500.

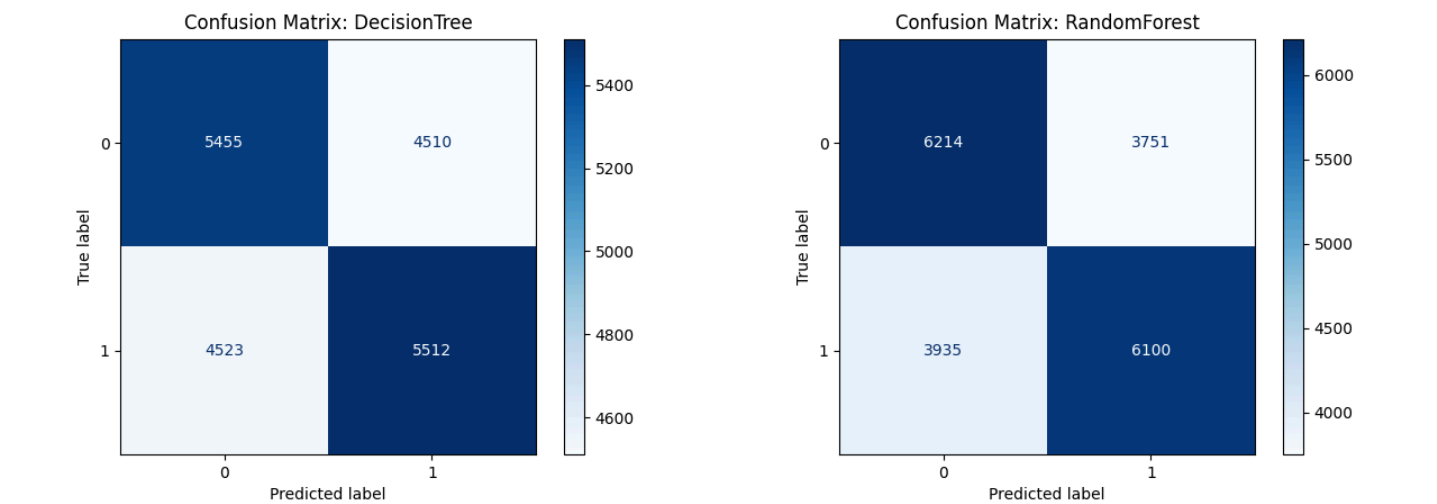
5.1.1. Dokładność modeli

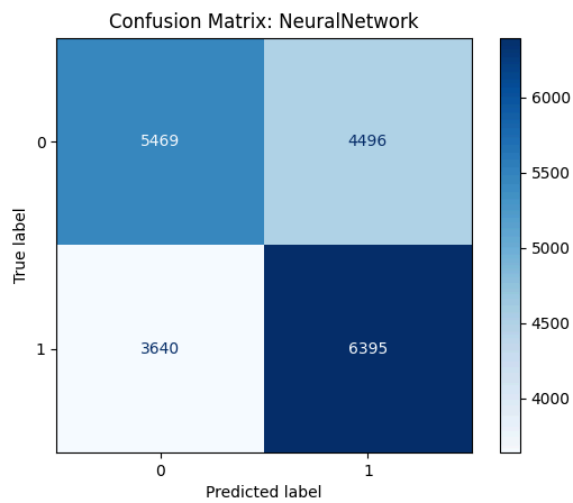
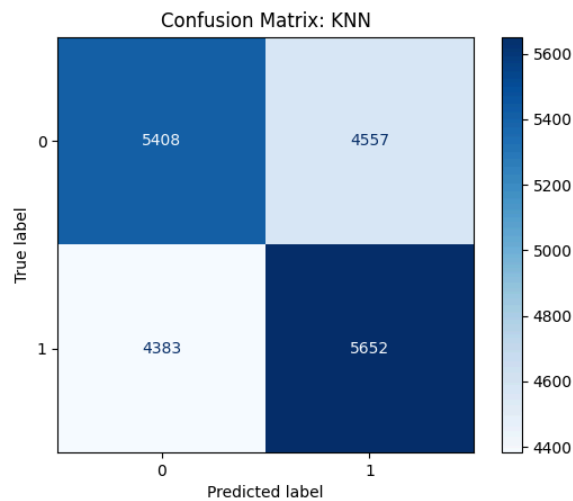
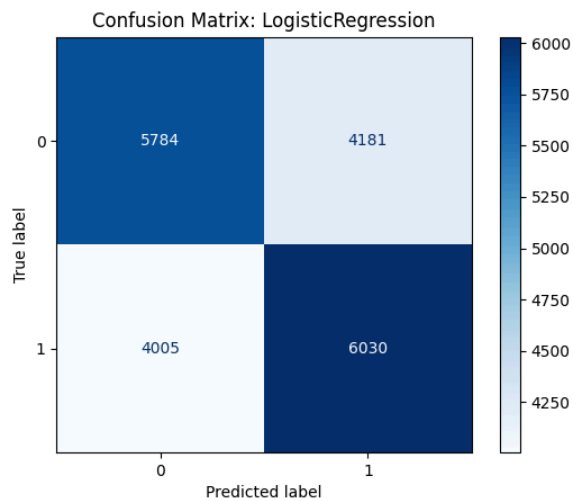
Rozmiar zbioru danych został ograniczony do 100 000 rekordów, aby przyspieszyć proces uczenia modeli. Zbiór ten jest zbalansowany, zawiera po 50 000 lotów opóźnionych i nieopóźnionych.

Model	Decision Tree	Random Forest	Logistic Regression	K-NN	Neural Network
Dokładność	54.84%	61.57%	59.07%	55.30%	59.32%



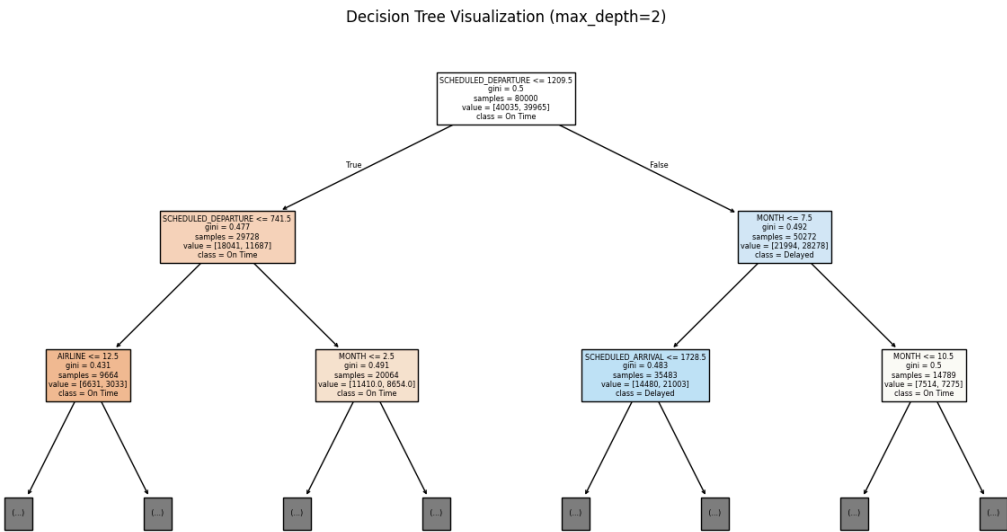
5.1.2. Macierze błędów





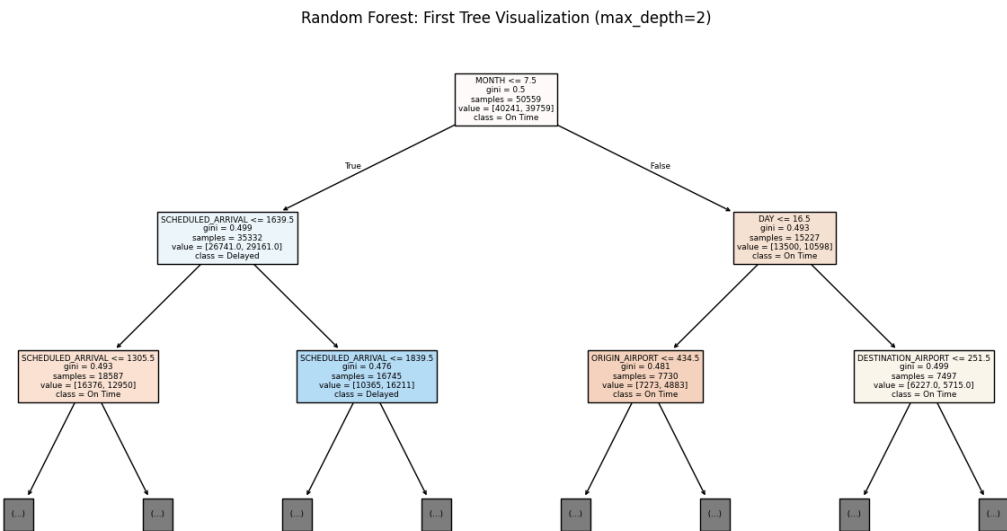
5.1.3. Wizualizacja model

5.1.3.1. Drzewo Decyzyjne



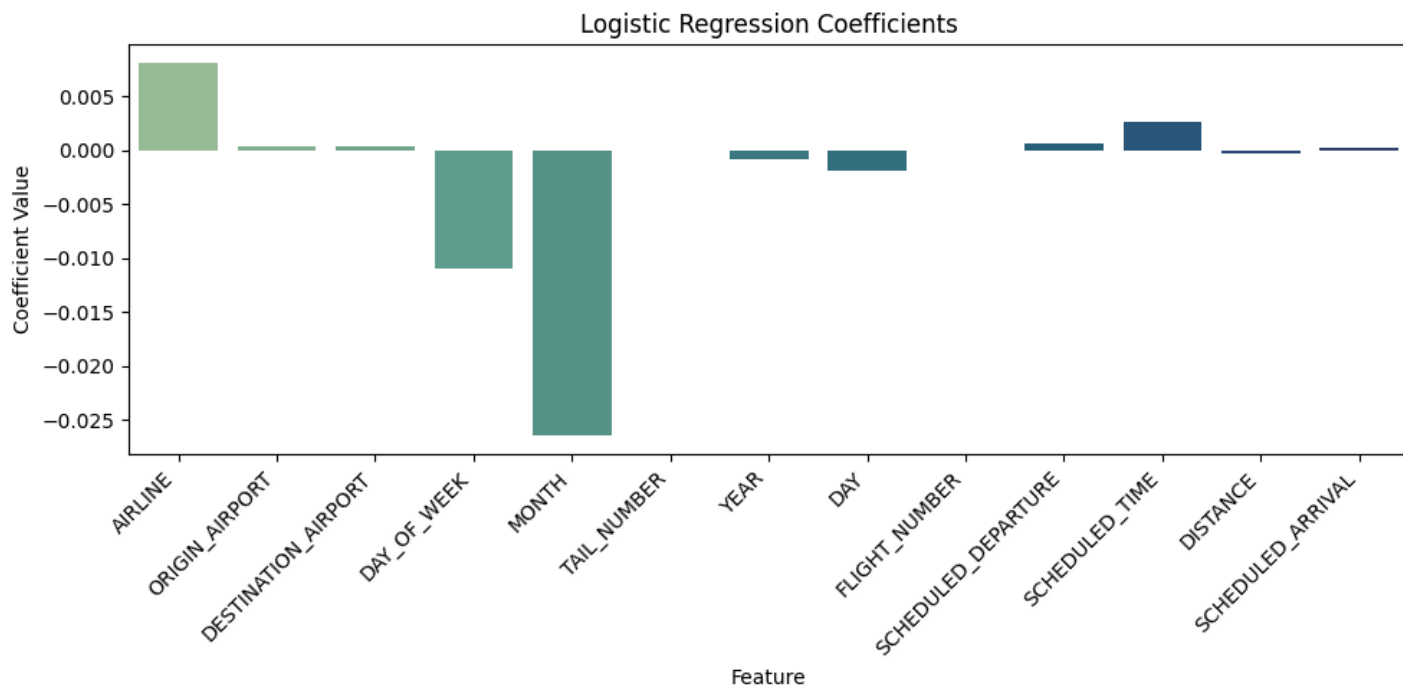
Graf przedstawia strukturę wytrenowanego drzewa decyzyjnego.

5.1.3.2. Random Forest



Graf przedstawia strukturę pierwszego z drzew w wytrenowanym lesie losowym.

5.1.3.3. Regresja Logistyczna

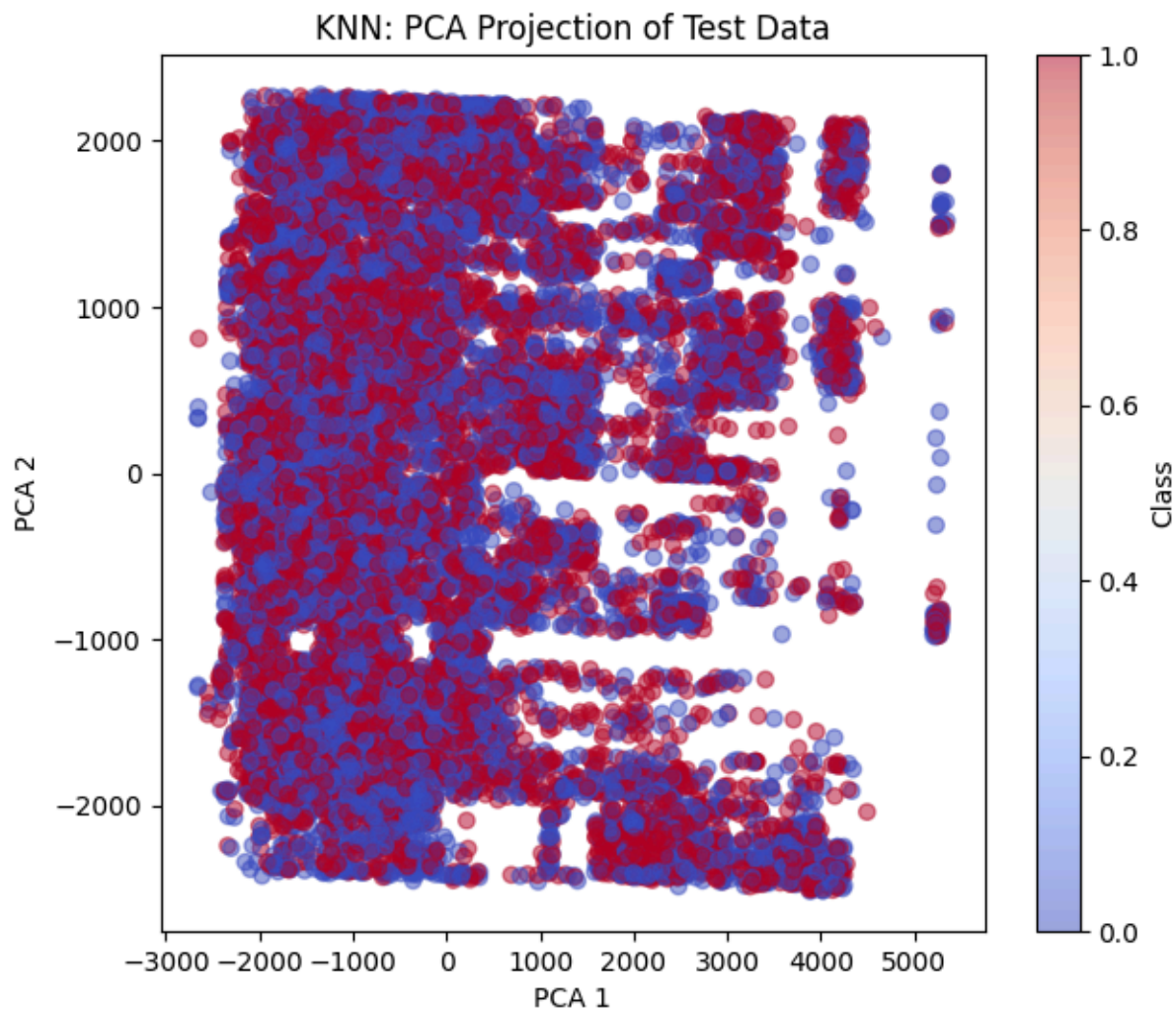


Wykres przedstawia współczynniki regresji logistycznej dla poszczególnych cech. Większa wartość współczynnika oznacza większy wpływ danej cechy na prawdopodobieństwo opóźnienia lotu.

Najważniejsze cechy w tym modelu to:

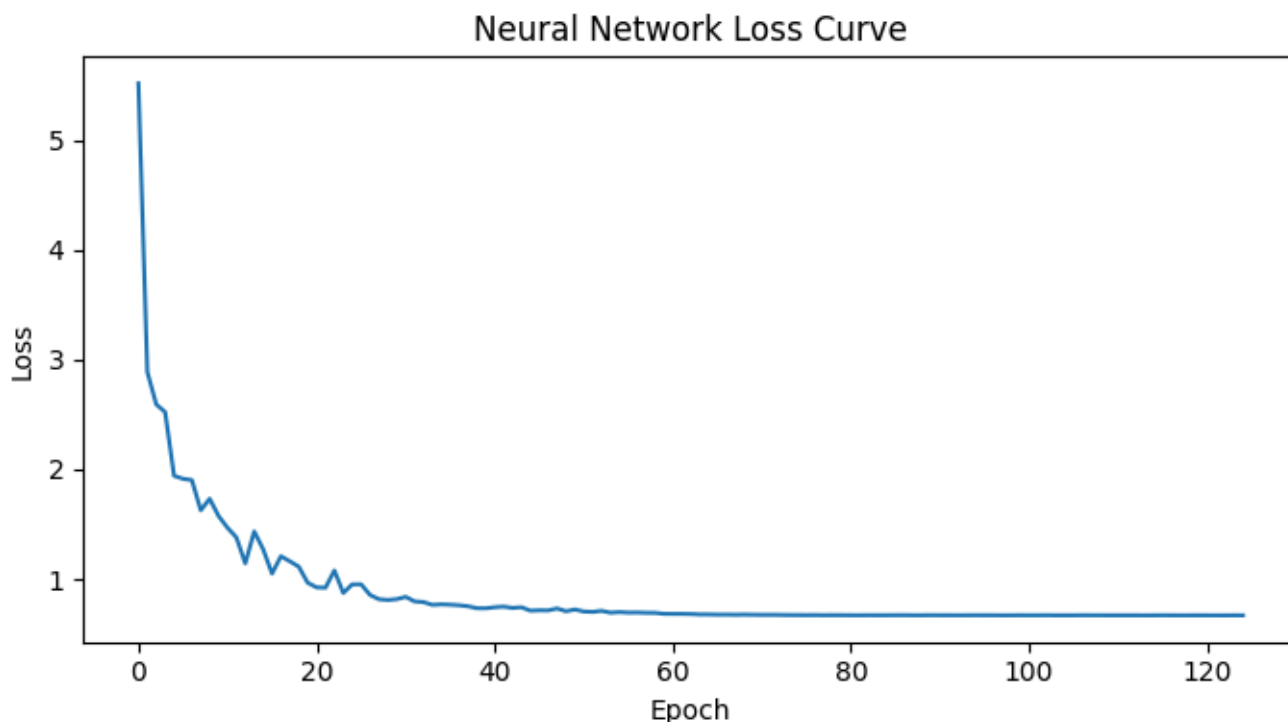
- MONTH
- DAY_OF_WEEK
- SCHEDULED_TIME
- AIRLINE

5.1.3.4. K Nearest Neighbors - wykres PCA



PCA (analiza głównych składowych) jest techniką redukcji wymiarowości, która pozwala na wizualizację danych w przestrzeni 2D. Wykres przedstawia punkty reprezentujące loty, gdzie kolor wskazuje na klasę (opóźniony lub nieopóźniony).

5.1.3.5. Sieć Neuronowa



Wykres przedstawia krzywą strat dla sieci neuronowej podczas treningu. Widać, że strata maleje wraz z kolejnymi epokami, co sugeruje, że model uczy się poprawnie.

6. Eksperymenty ze zbiorem danych

6.1. Optymalizacja hiperparametrów

Dokonano optymalizacji hiperparametrów przy użyciu algorytmu genetycznego oraz ustalono testową wielkość zbioru 1000 dla modeli Random Forest, Logistic Regression i Neural Network. Uzyskano następujące dokładności modeli:

Model	Random Forest	Logistic Regression	Neural Network
Dokładność	68.45%	72.61%	72.02%

Mniejszy rozmiar danych pozwolił na szybsze przeprowadzenie eksperymentów, jednak nie możemy bezpośrednio porównywać wyników z poprzednimi modelami. Mniejszy zbiór danych może skutkować nadmiernym dopasowaniem modeli.

6.1.1. Wyznaczone hiperparametry

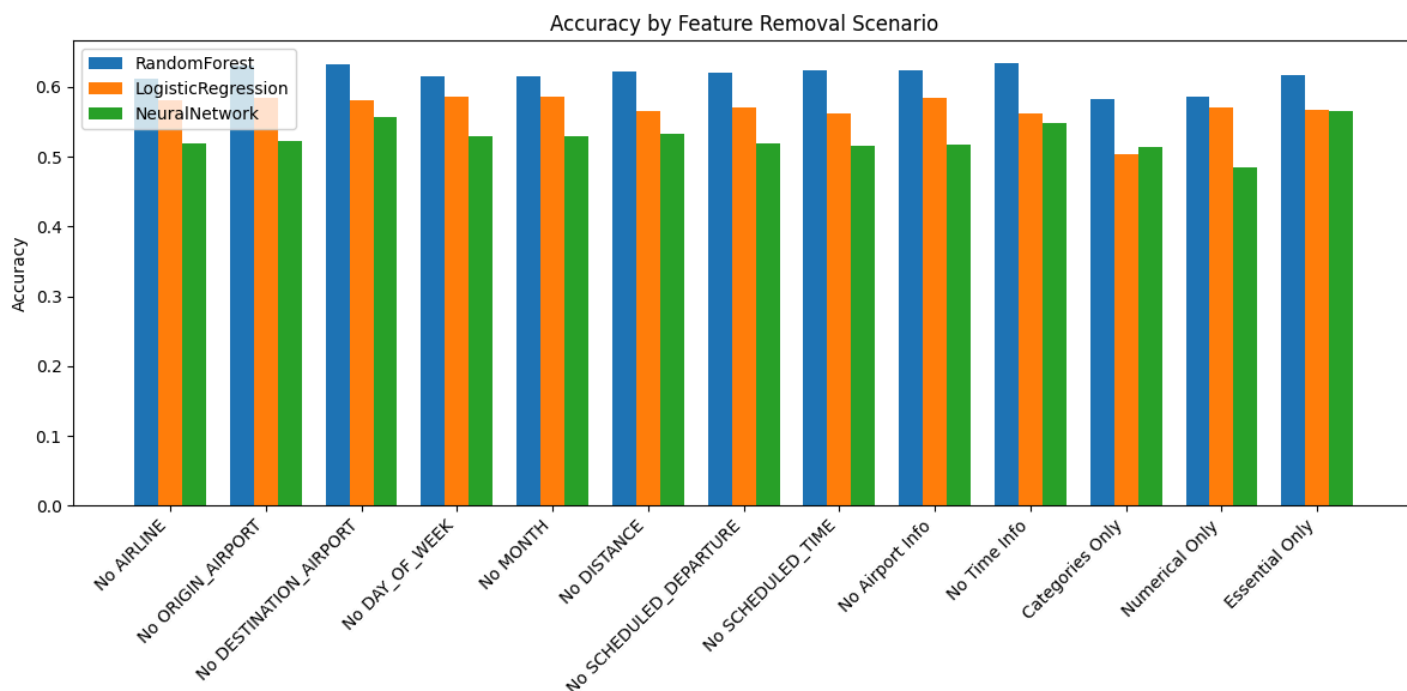
- **Random Forest**
 - Liczba drzew: 28
 - Maksymalna głębokość drzewa: 4
 - Minimalna liczba próbek do podziału: 8
 - Minimalna liczba próbek w liściu: 3
- **Logistic Regression**
 - C: 17.825
 - Maksymalna liczba iteracji: 1000

• Neural Network

- Liczba neuronów w warstwach ukrytych: 100, 83
- Współczynnik regularyzacji α : 0.00125
- Maksymalna liczba iteracji: 785

6.2. Usunięcie cech

Dokonano analizy wpływu usunięcia poszczególnych cech na dokładność modeli **Random Forest**, **Logistic Regression** oraz **Neural Network** działających wielkośći zbioru 10000. Dokładności poszczególnych modeli po usunięciu cech przedstawiono na poniższym wykresie oraz tabeli:



Random Forest (dokładność nominalna: 62.24%)

Usunięta cecha/cechy	Dokładność po usunięciu	Różnica
AIRLINE	61.18%	−1.06%
ORIGIN_AIRPORT	63.00%	+0.76%
DESTINATION_AIRPORT	63.29%	+1.05%
DAY_OF_WEEK	61.48%	−0.76%
MONTH	61.48%	−0.76%
DISTANCE	62.24%	−0.00%
SCHEDULED_DEPARTURE	62.08%	−0.16%
SCHEDULED_TIME	62.39%	+0.15%
Informacje lotniskowe	62.39%	+0.15%
Informacje o dniu	63.44%	+1.20%

Logistic Regression (dokładność nominalna: 58.61%)

Usunięta cecha/cechy	Dokładność po usunięciu	Różnica
AIRLINE	58.01%	−0.60%

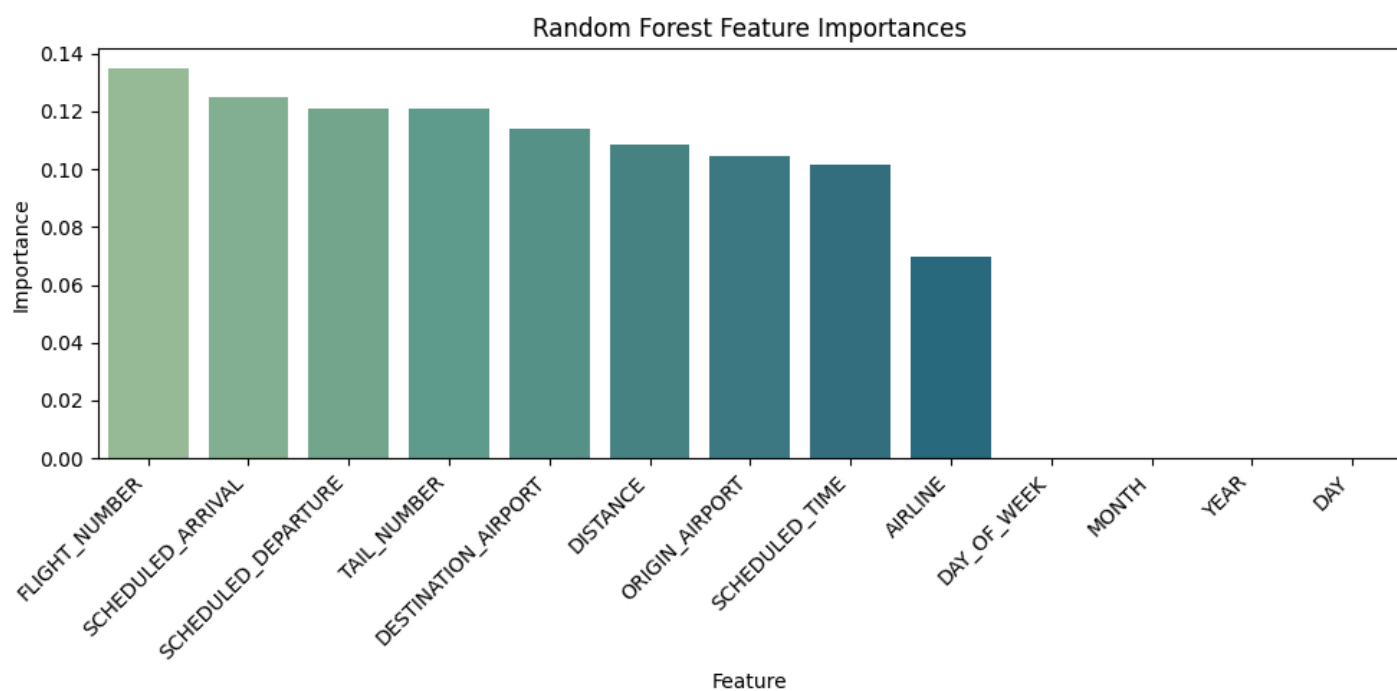
ORIGIN_AIRPORT	58.46%	−0.15%
DESTINATION_AIRPORT	58.01%	−0.60%
DAY_OF_WEEK	58.61%	−0.00%
MONTH	58.61%	−0.00%
DISTANCE	56.50%	−2.11%
SCHEDULED_DEPARTURE	57.10%	−1.51%
SCHEDULED_TIME	56.19%	−2.42%
Informacje lotniskowe	58.46%	−0.15%
Informacje o dniu	56.19%	−2.42%

Neural Network (dokładność nominalna: 54.22%)

Usunięta cecha/cechy	Dokładność po usunięciu	Różnica
AIRLINE	51.96%	−2.26%
ORIGIN_AIRPORT	52.27%	−1.95%
DESTINATION_AIRPORT	55.74%	+1.52%
DAY_OF_WEEK	53.02%	−1.20%
MONTH	53.02%	−1.20%
DISTANCE	53.32%	−0.90%
SCHEDULED_DEPARTURE	51.96%	−2.26%
SCHEDULED_TIME	51.51%	−2.71%
Informacje lotniskowe	51.66%	−2.56%
Informacje o dniu	54.83%	−0.61%

6.3. Ważność cech

Ważność cech została obliczona dla modelu **Random Forest** i przedstawiona na poniższym wykresie. Wartości te wskazują, jak duży wpływ ma dana cecha na decyzje podejmowane przez model. Wyższa wartość oznacza większy wpływ na klasyfikację.



Najważniejszymi cechami są:

- FLIGHT_NUMBER
- SCHEDULED_ARRIVAL
- SCHEDULED_DEPARTURE
- TAIL_NUMBER

Atrybuty MONTH, YEAR i DAY zostały uznane za nieistotne.

7. Podsumowanie