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.

Kryterium sukcesu jest osiągnięcie dokładności klasyfikacji opóźnienia na poziomie 85%.

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	`DecisionTreeClassifier`
Las Losowy	`RandomForestClassifier`
Regresja Logistyczna	`LogisticRegression`
K-NN	`KNeighborsClassifier`
Sieć Neuronowa	`MLPClassifier`

5.1. Opis modeli

5.1.1. Drzewo decyzyjne

Drzewo decyzyjne to klasyfikator, który podejmuje decyzje na podstawie serii pytań dotyczących cech wejściowych. Każdy węzeł drzewa reprezentuje pytanie dotyczące jednej z cech, a gałęzie prowadzą do kolejnych węzłów lub liści, które reprezentują klasyfikację. Drzewa decyzyjne są łatwe do interpretacji i wizualizacji, ale mogą być podatne na nadmierne dopasowanie (overfitting) przy dużych zbiorach danych.

5.1.2. Las losowy

Las losowy to metoda, która tworzy wiele drzew decyzyjnych i łączy ich wyniki, aby uzyskać bardziej stabilną i dokładną klasyfikację. Każde drzewo jest trenowane na losowej próbce danych i losowym podzbiorze cech, co pozwala na redukcję wariancji i poprawę ogólnej wydajności modelu. Las losowy jest mniej podatny na nadmierne dopasowanie niż pojedyncze drzewo decyzyjne.

5.1.3. Regresja logistyczna

Regresja logistyczna to klasyfikator liniowy, który modeluje prawdopodobieństwo przynależności do klasy poprzez funkcję logistyczną (funkcja która przekształca liniową kombinację cech w wartość z

przedziału [0, 1]). Jest to prosty algorytm, który dobrze sprawdza się w przypadku problemów binarnych. W tym projekcie zastosowano domyślne parametry z maksymalną liczbą iteracji równą 1000.

5.1.4. K-NN

K-NN (K-Nearest Neighbors) to klasyfikator, który przypisuje etykietę do nowego przykładu na podstawie etykiet jego k najbliższych sąsiadów w przestrzeni cech. W przypadku tego projektu, zastosowano 5 sąsiadów. K-NN jest prostym i intuicyjnym algorytmem, ale może być wolny przy dużych zbiorach danych, ponieważ wymaga obliczenia odległości do wszystkich punktów w zbiorze treningowym.

5.1.5. Sieć neuronowa MLP

Sieć neuronowa MLP (Multi-Layer Perceptron) jest klasyfikatorem opartym na sieciach neuronowych, który składa się z co najmniej jednej warstwy ukrytej. Sieć uczy się na zasadzie propagacji wstecznej, gdzie błąd jest propagowany od warstwy wyjściowej do warstw ukrytych, a wagi są aktualizowane na podstawie gradientu błędu. Głównie używana jest do klasyfikacji tekstu i obrazów, co może sugerować, że nie jest to najlepszy wybór w przypadku naszych danych.

5.2. Początkowe porównanie modeli

Pierwszym krokiem jest porównanie modeli z domyślnymi parametrami.

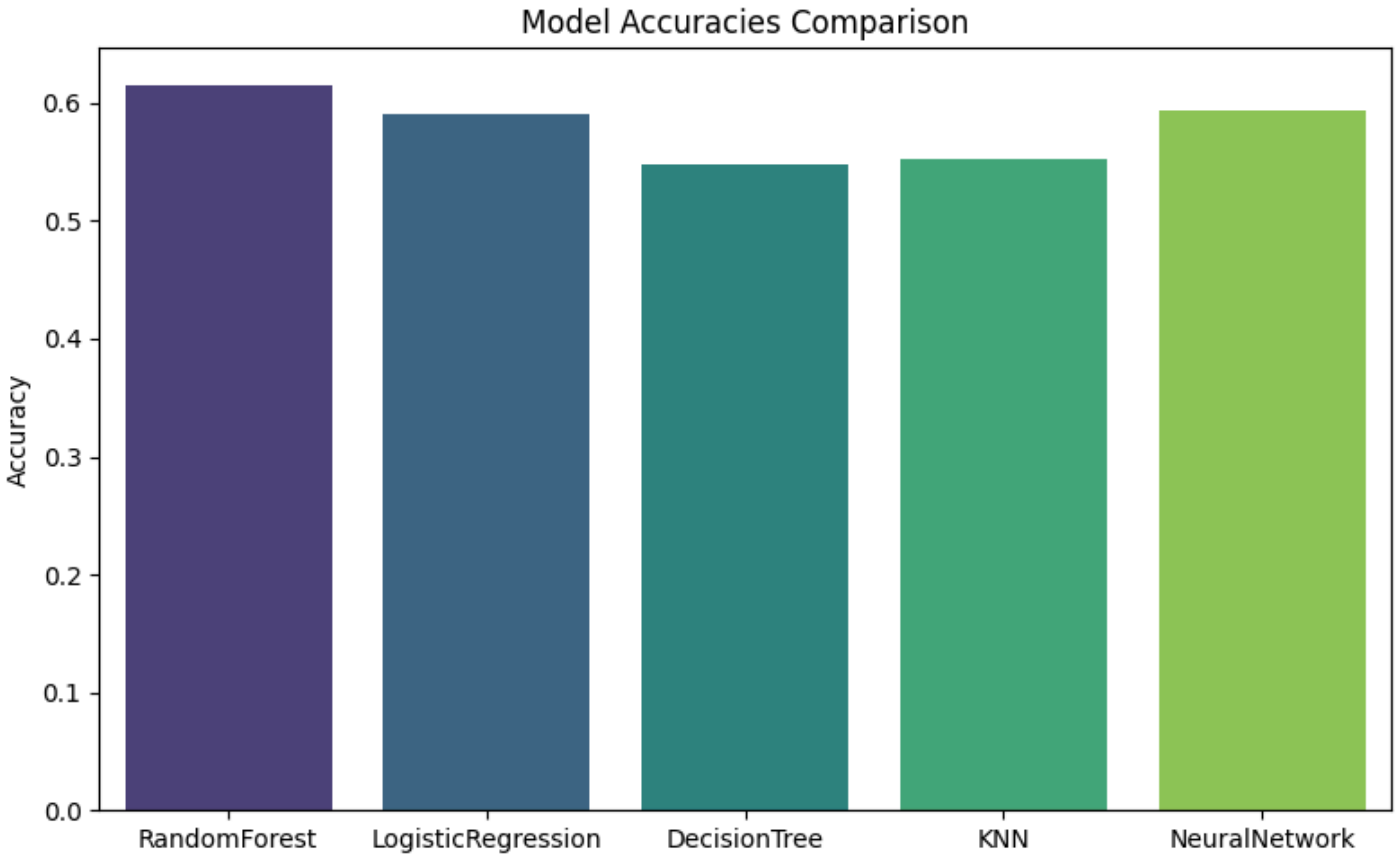
- Dla `RandomForest` utworzono 100 drzew, maksymalna głębokość nie jest ograniczona, a minimalna liczba próbek do podziału to 2.
- Dla `LogisticRegression` zastosowano domyślne parametry, z maksymalną liczbą iteracji równą 1000.
- Dla `DecisionTree` zastosowano domyślne parametry, z maksymalną głębokością 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.

5.2.1. Dokładność modeli

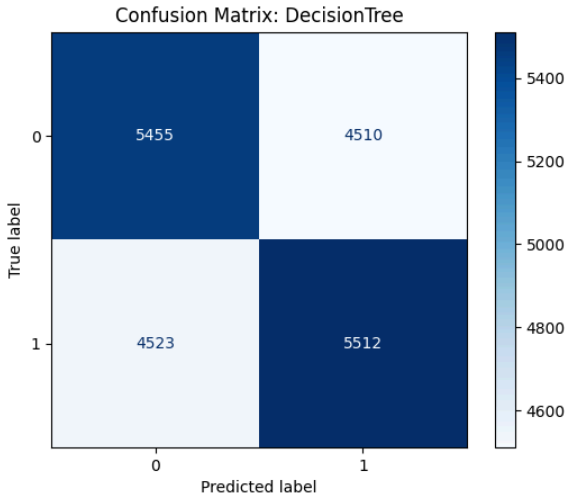
Rozmiar zbioru danych w tym teście nie został ograniczony, aby umożliwić uzyskanie najlepszych wyników. Zbiór ten jest zbalansowany, zawiera po 1 023 498 lotów opóźnionych i nieopóźnionych.

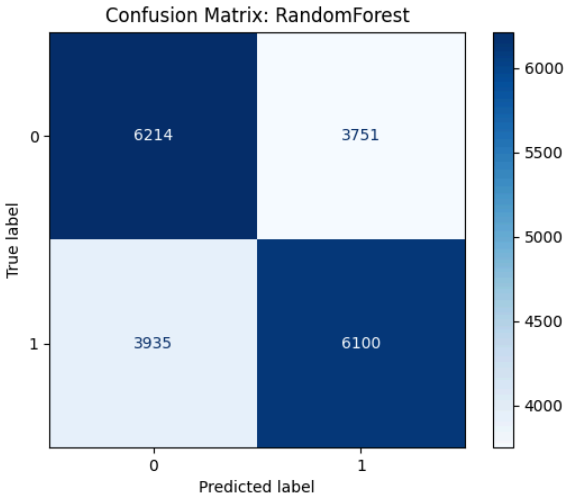
Wykresy przedstawiają wyniki dla pomniejszonego zbioru danych.

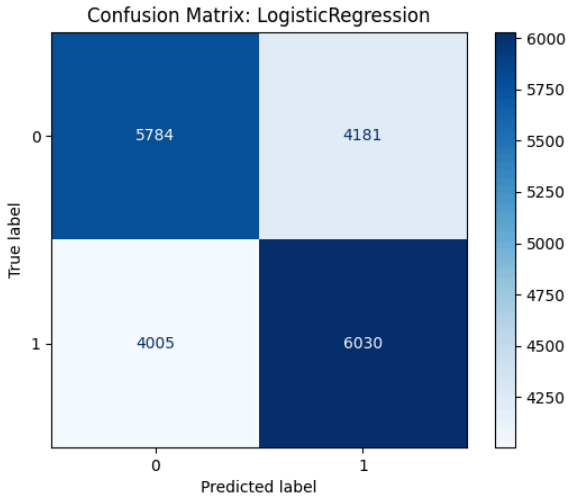
Model	Decision Tree	Random Forest	Logistic Regression	K-NN	Neural Network
Dokładność	`59.09%`	`65.43%`	`58.80%`	`57.96%`	`58.96%`

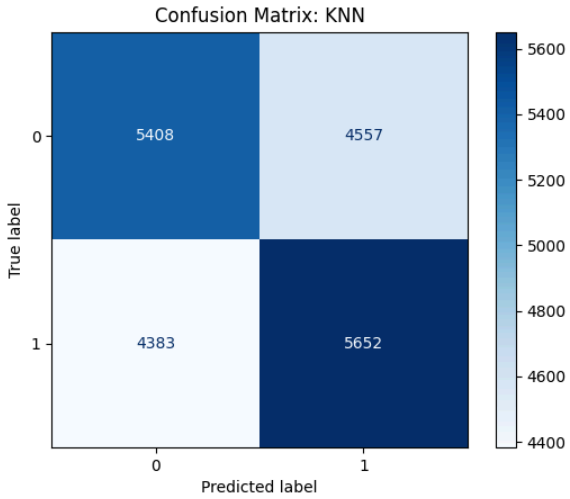


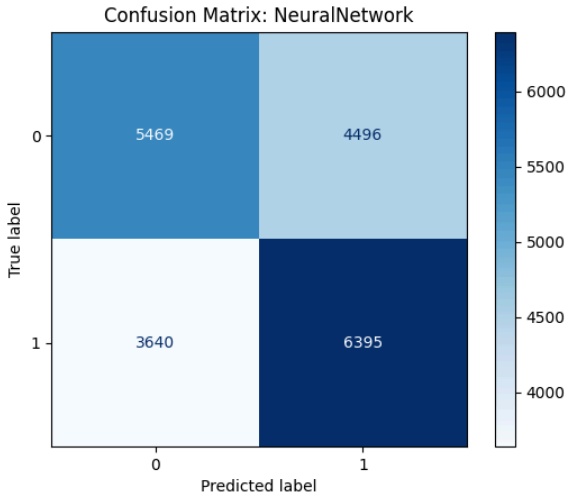
5.2.2. Macierze błędów





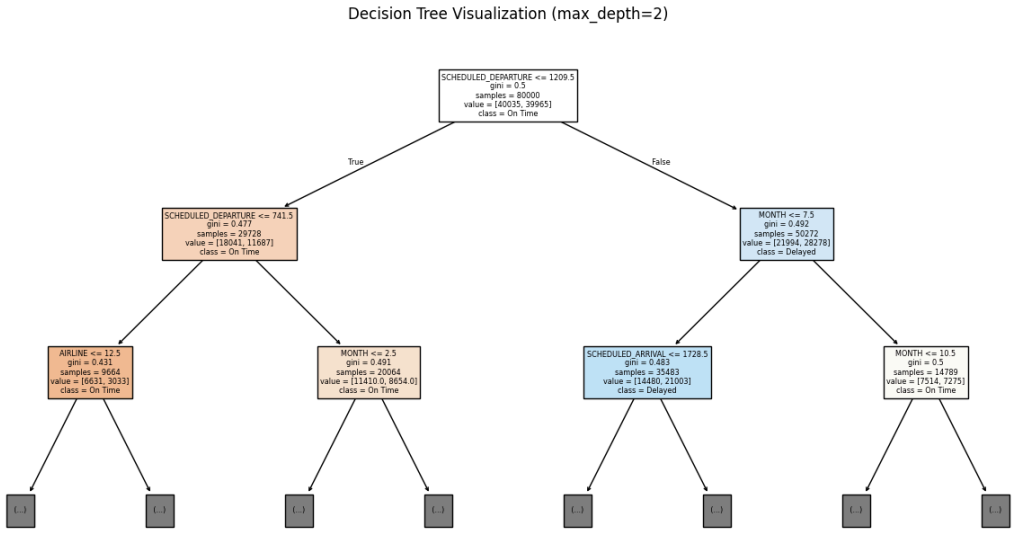

Confusion Matrix: LogisticRegression
5784
4181
4005
6030
True label
Predicted label
0
1
6000
5750
5500
5250
5000
4750
4500
4250


Confusion Matrix: KNN
5408
4557
4383
5652
True label
Predicted label
0
1
5600
5400
5200
5000
4800
4600
4400


Confusion Matrix: NeuralNetwork
5469
4496
3640
6395
True label
Predicted label
0
1
6000
5500
5000
4500
4000

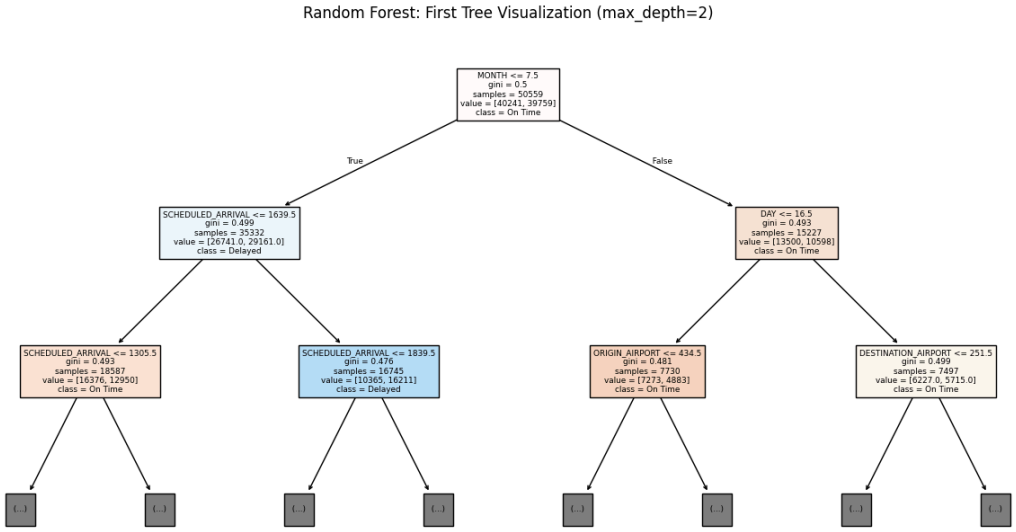
5.2.3. Wizualizacja model

5.2.3.1. Drzewo Decyzyjne



Graf przedstawia strukturę wytrenowanego drzewa decyzyjnego.

5.2.3.2. Random Forest



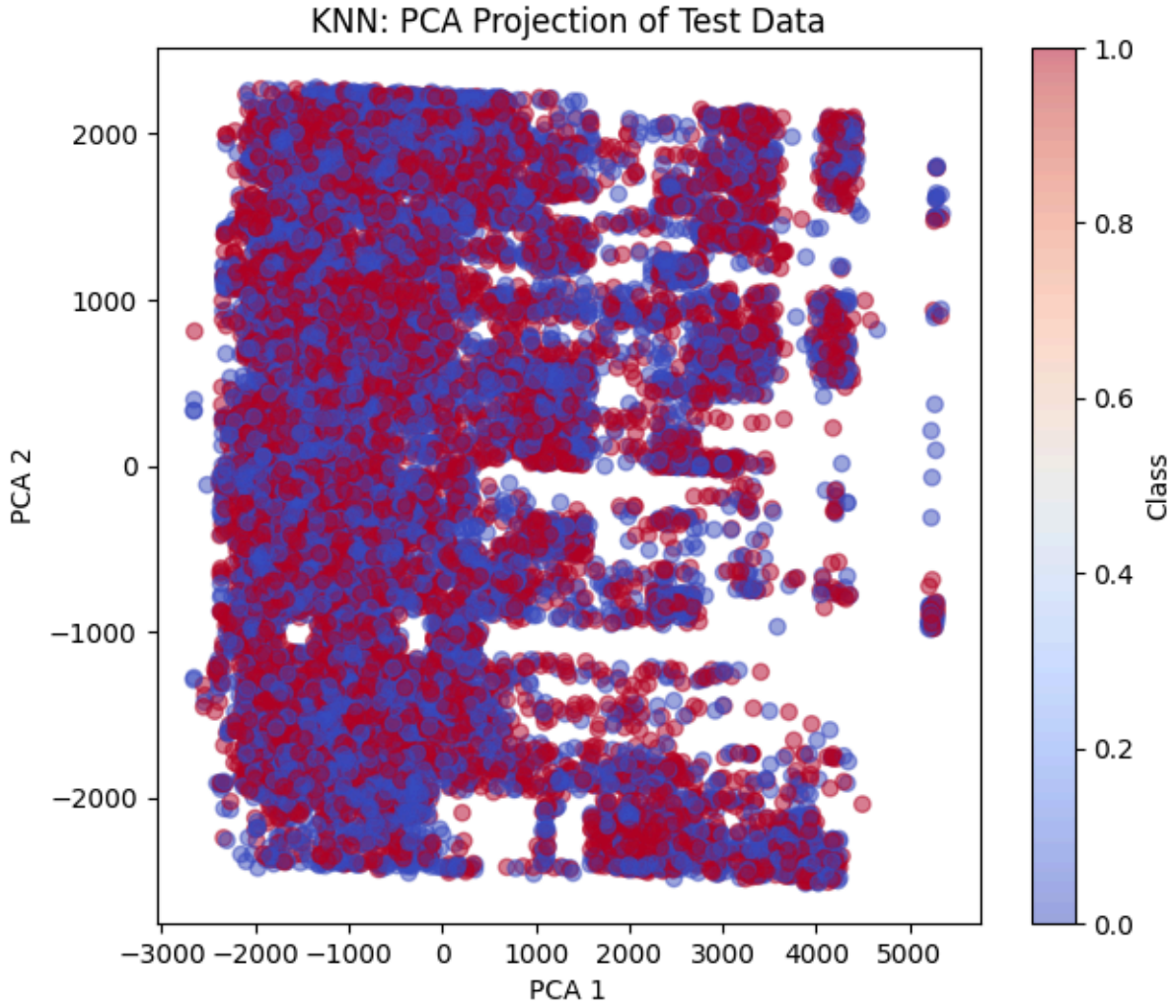
Graf przedstawia strukturę pierwszego z drzew w wytrenowanym lesie losowym.

5.2.3.3. Regresja Logistyczna

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

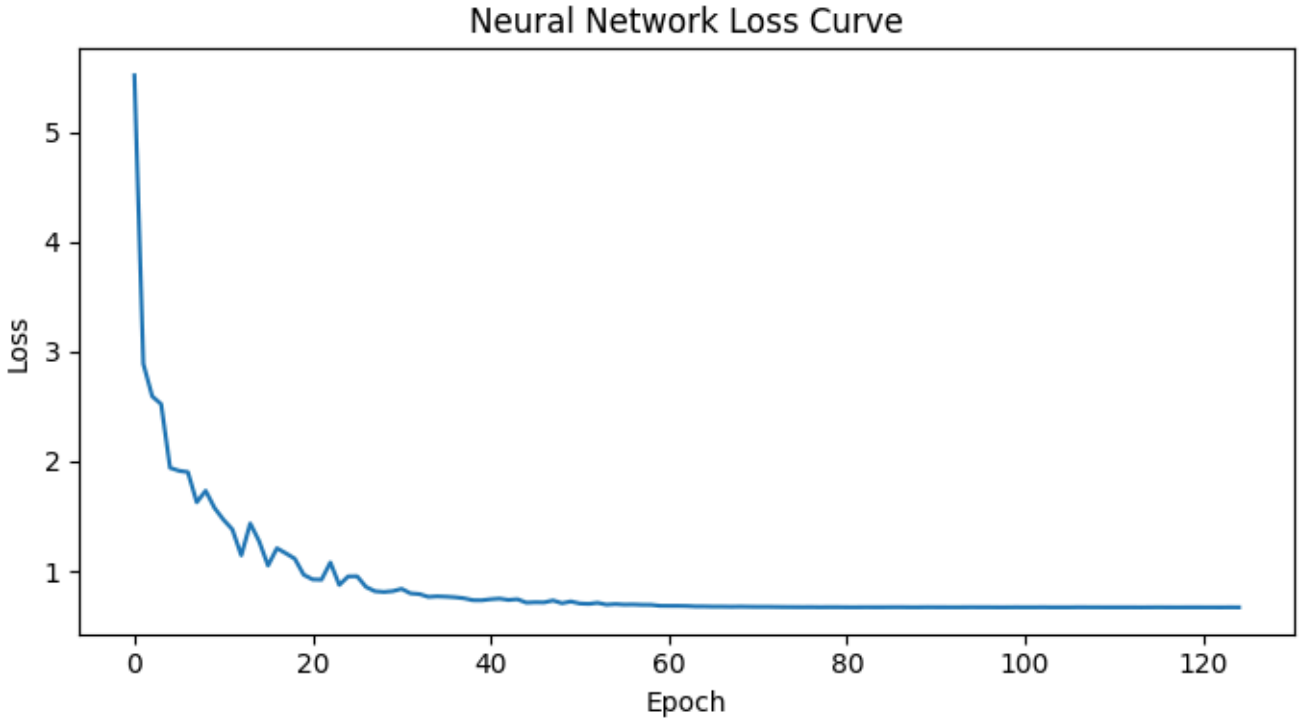
Najważniejsze cechy w tym modelu to:
- MONTH
- DAY_OF_WEEK
- SCHEDULED_TIME
- AIRLINE

5.2.3.4. K Nearest Neighbors - wykres PCA



PCA (analiza głównych składowych) to technika redukcji wymiarowości, która umożliwia odwzorowanie danych w przestrzeni dwuwymiarowej (2D), przy zachowaniu jak największej ilości informacji o ich strukturze. Na wykresie przedstawiono punkty reprezentujące pojedyncze loty, gdzie kolor wskazuje klasę – opóźniony lub nieopóźniony.

5.2.3.5. 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 10 000 dla modeli `Random Forest`, `Logistic Regression` i `Neural Network`. Uzyskano następujące dokładności modeli:

Model	Random Forest	Logistic Regression	Neural Network
Dokładność	`68.02%`	`63.10%`	`60.22%`

Mniejszy rozmiar danych pozwolił na szybsze przeprowadzenie eksperymentów, ale może powodować nadmierne dopasowanie modeli do danych.

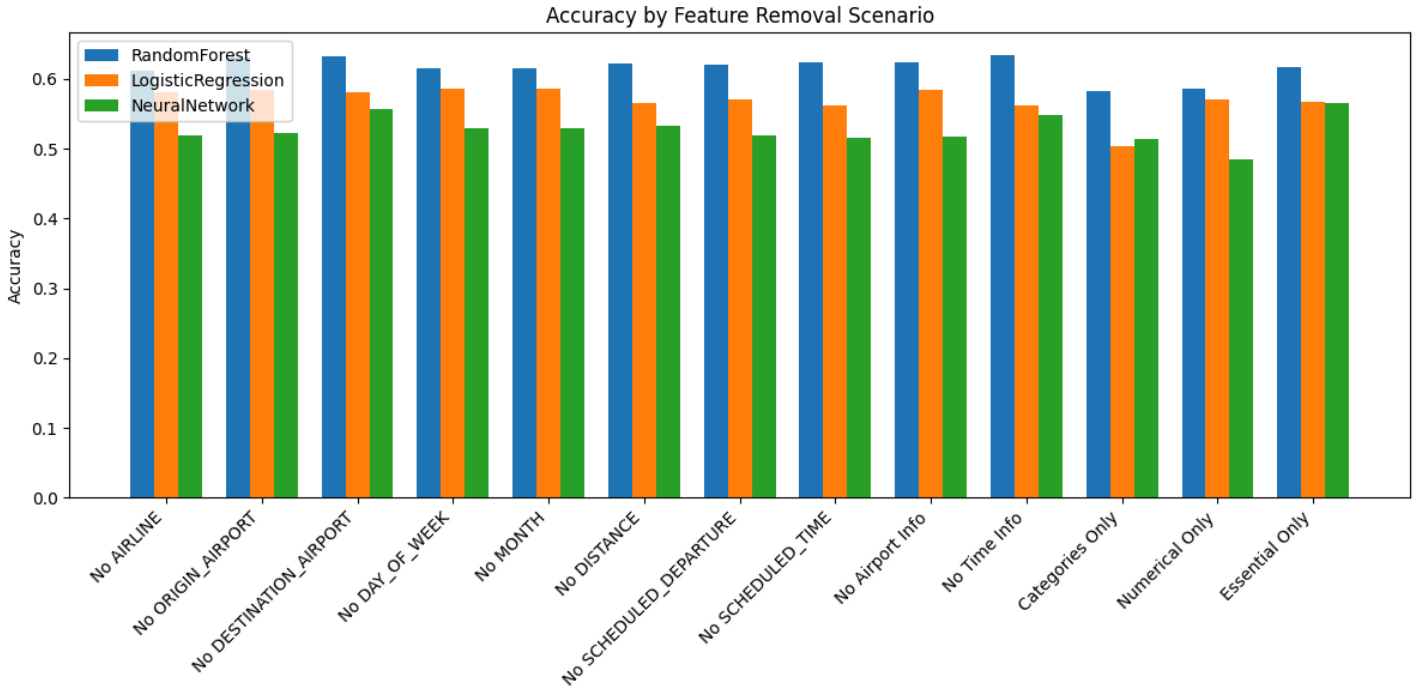
6.1.1. Wyznaczone hiperparametry

- **Random Forest**
  - Liczba drzew: 141
  - Maksymalna głębokość drzewa: 14
  - Minimalna liczba próbek do podziału: 7
  - Minimalna liczba próbek w liściu: 5
- **Logistic Regression**
  - C: 14.1623
  - Maksymalna liczba iteracji: 1857
- **Neural Network**

- Liczba neuronów w warstwach ukrytych: 129, 25
- Współczynnik regularyzacji α: 0.008832
- Maksymalna liczba iteracji: 542

6.2. Usunięcie cech

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



Random Forest (dokładność nominalna: 68.95%)

Usunięta cecha/cechy	Dokładność po usunięciu	Różnica
AIRLINE	67.64%	−1.31%
ORIGIN_AIRPORT	68.12%	−0.83%
DESTINATION_AIRPORT	68.46%	−0.49%
DAY_OF_WEEK	68.66%	−0.29%
MONTH	68.89%	−0.06%
DISTANCE	68.89%	−0.07%
SCHEDULED_DEPARTURE	69.12%	+0.17%
SCHEDULED_TIME	69.02%	+0.07%
Informacje lotniskowe	67.44%	−1.51%
Informacje o dniu	69.07%	+0.12%

Logistic Regression (dokładność nominalna: 61.05%)

Usunięta cecha/cechy	Dokładność po usunięciu	Różnica
AIRLINE	61.21%	+0.16%

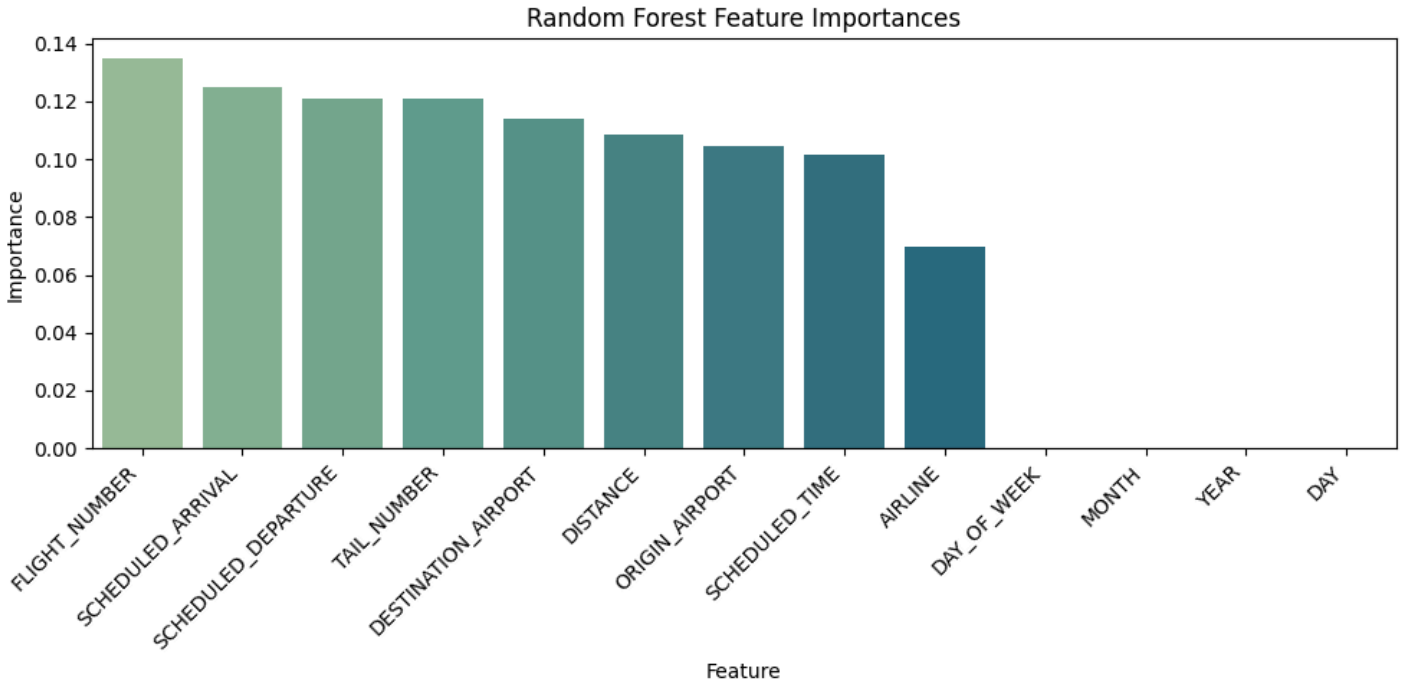
ORIGIN_AIRPORT	61.08%	+0.03%
DESTINATION_AIRPORT	61.19%	+0.14%
DAY_OF_WEEK	60.24%	−0.81%
MONTH	61.05%	−0.00%
DISTANCE	61.11%	+0.06%
SCHEDULED_DEPARTURE	60.48%	−0.57%
SCHEDULED_TIME	61.16%	+0.11%
Informacje lotniskowe	61.21%	+0.16%
Informacje o dniu	59.94%	−1.11%

Neural Network (dokładność nominalna: 62.50%)

Usunięta cecha/cechy	Dokładność po usunięciu	Różnica
AIRLINE	63.93%	+1.43%
ORIGIN_AIRPORT	63.49%	+0.99%
DESTINATION_AIRPORT	63.85%	+1.35%
DAY_OF_WEEK	64.29%	+1.79%
MONTH	64.64%	+2.14%
DISTANCE	64.03%	+1.53%
SCHEDULED_DEPARTURE	63.55%	+1.05%
SCHEDULED_TIME	64.25%	+1.75%
Informacje lotniskowe	63.23%	+0.73%
Informacje o dniu	63.75%	+1.25%

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.



Cechy	Ważność
FLIGHT_NUMBER	**13.49%**
SCHEDULED_ARRIVAL	**12.49%**
SCHEDULED_DEPARTURE	**12.11%**
TAIL_NUMBER	**12.08%**
DESTINATION_AIRPORT	11.38%
DISTANCE	10.84%
ORIGIN_AIRPORT	10.44%
SCHEDULED_TIME	10.17%
AIRLINE	6.98%
DAY_OF_WEEK	0.00%
MONTH	0.00%
YEAR	0.00%
DAY	0.00%

Najważniejszymi cechami są:

- FLIGHT_NUMBER
- SCHEDULED_ARRIVAL
- SCHEDULED_DEPARTURE
- TAIL_NUMBER

Atrybuty MONTH, YEAR i DAY zostały uznane za nieistotne.

7. Podsumowanie

7.1. Przegląd wykonanego procesu

W ramach drugiego etapu projektu z przedmiotu Eksploracja Danych zrealizowano predykcję opóźnień lotów na podstawie zbioru danych 2015 Flight Delays and Cancellations. Celem eksploracji było zbudowanie realistycznego modelu predykcyjnego klasyfikującego opóźnienia lotów w oparciu wyłącznie o informacje dostępne przed startem. Zbiór danych został odpowiednio przetworzony – usunięto cechy ujawniane dopiero po rozpoczęciu lotu – a dane zbalansowano, co pozwoliło na trenowanie i ocenę kilku modeli klasyfikacyjnych: drzewa decyzyjnego, lasu losowego, regresji logistycznej, k-NN oraz sieci neuronowej.

Ze względu na bardzo dużą liczbę przykładów (ponad 5,8 miliona rekordów), w części eksperymentów zastosowano ograniczenie zbioru do 10 000 przykładów. Umożliwiło to szybsze przeprowadzanie eksperymentów i optymalizację hiperparametrów bez utraty ogólności wyników.

7.2. Ocena modeli–

Spośród testowanych modeli najwyższą dokładność na zbiorze testowym osiągnął Random Forest (68.02%) po optymalizacji parametrów. Regresja logistyczna oraz sieć neuronowa również uzyskały poprawę skuteczności względem wersji bazowych. Pomimo użycia stosunkowo prostych cech wejściowych, modele uzyskały stabilne wyniki. Analiza wpływu cech oraz wizualizacje modelowe pozwoliły dodatkowo zidentyfikować najbardziej znaczące atrybuty - m.in. FLIGHT_NUMBER, SCHEDULED_ARRIVAL i SCHEDULED_DEPARTURE.

7.3. Stopień realizacji celów

Początkowo założono osiągnięcie dokładności predykcji na poziomie co najmniej 85%. Cel ten nie został zrealizowany - najlepszy model osiągnął dokładność poniżej 70%. Głównym powodem tej różnicy było ograniczenie się wyłącznie do informacji dostępnych przed lotem, co drastycznie zawęziło możliwości predykcyjne. W przypadku uwzględnienia cech takich jak rzeczywisty czas odlotu czy opóźnienie w momencie startu, modele osiągałyby niemal idealną skuteczność, co jednak byłoby sprzeczne z celem budowy realistycznego narzędzia do prognozowania.

7.4. Wnioski

Mimo że nie udało się osiągnąć zakładanej dokładności 85%, projekt dostarczył wartościowych wniosków. Zbudowane modele pozwalają na sensowną predykcję opóźnień z dokładnością na poziomie ~ 68%, wyłącznie na podstawie danych planistycznych. Dodatkowo zidentyfikowano cechy o największym wpływie na opóźnienia, co może być podstawą dalszych analiz. W przyszłości warto rozważyć rozszerzenie zbioru o dane pogodowe i informacje kontekstowe, które mogłyby istotnie zwiększyć dokładność przy zachowaniu realizmu predykcji.