Eksploracja danych - etap 2

2015 Flight Delays and Cancellations

13.06.2025

Krzysztof Nasuta 193328, Filip Dawidowski 193433, Aleks Iwicki 193354

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.

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

Cel eksploracji

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

Kluczowe pytania badawcze:

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

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`

itp.

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

Porównanie modeli

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.

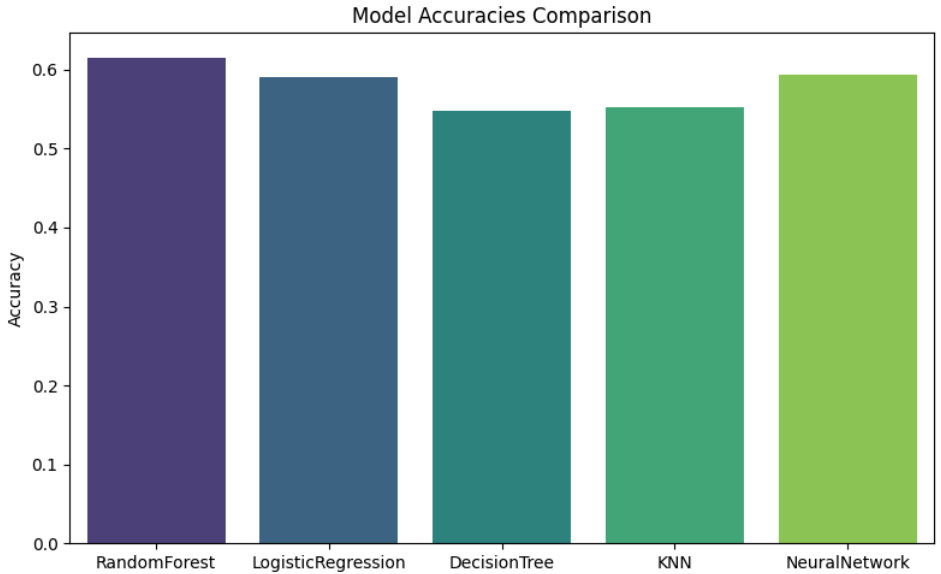
Porównanie dokładności

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.

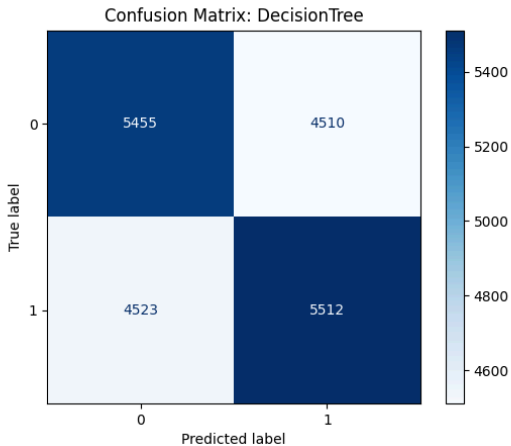
Wykresy przedstawiają wyniki dla pomniejszonego zbioru danych.

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

Porównanie dokładności (ii)

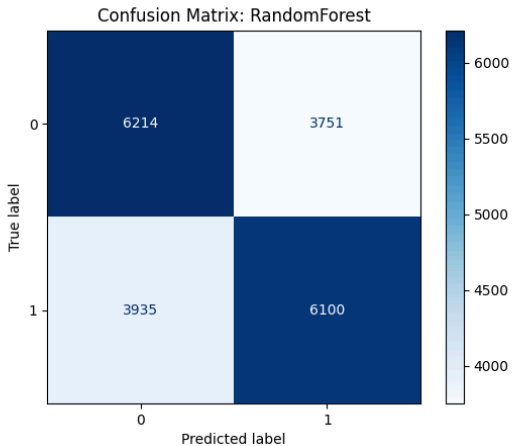


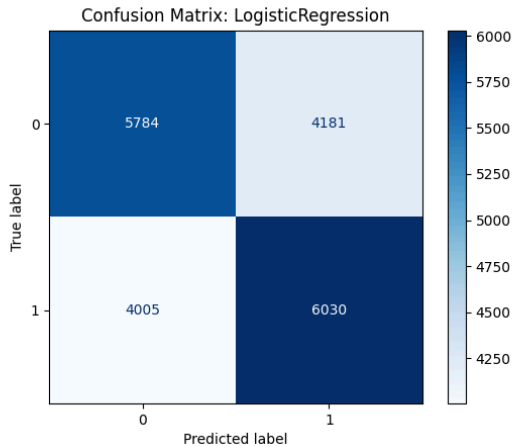
Macierze błędów



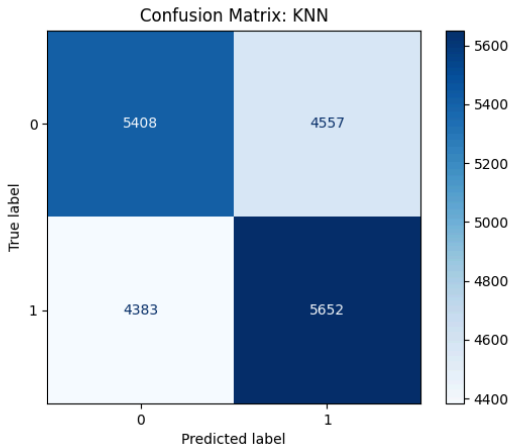
Wykresy przedstawiają wyniki dla pomniejszonego zbioru danych.

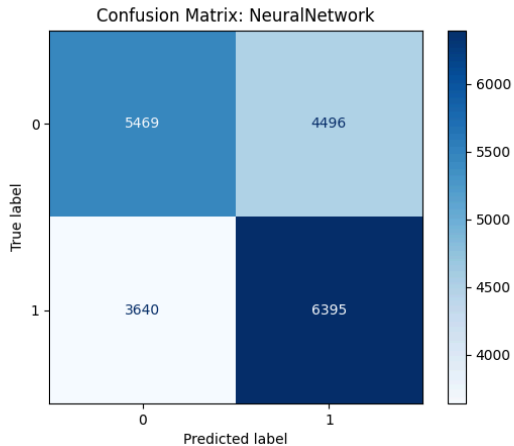
Macierze błędów (ii)



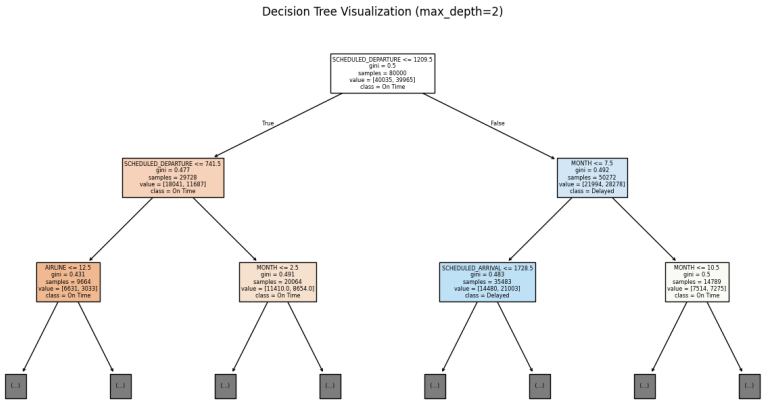


Macierze błędów (iii)



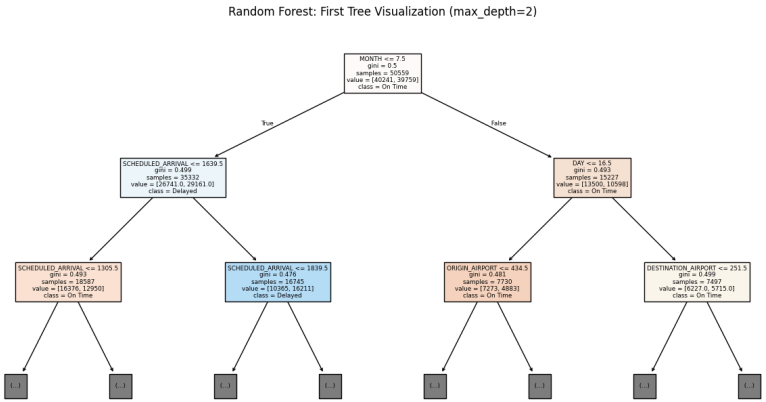


Wizualizacja modeli - Drzewo Decyzyjne



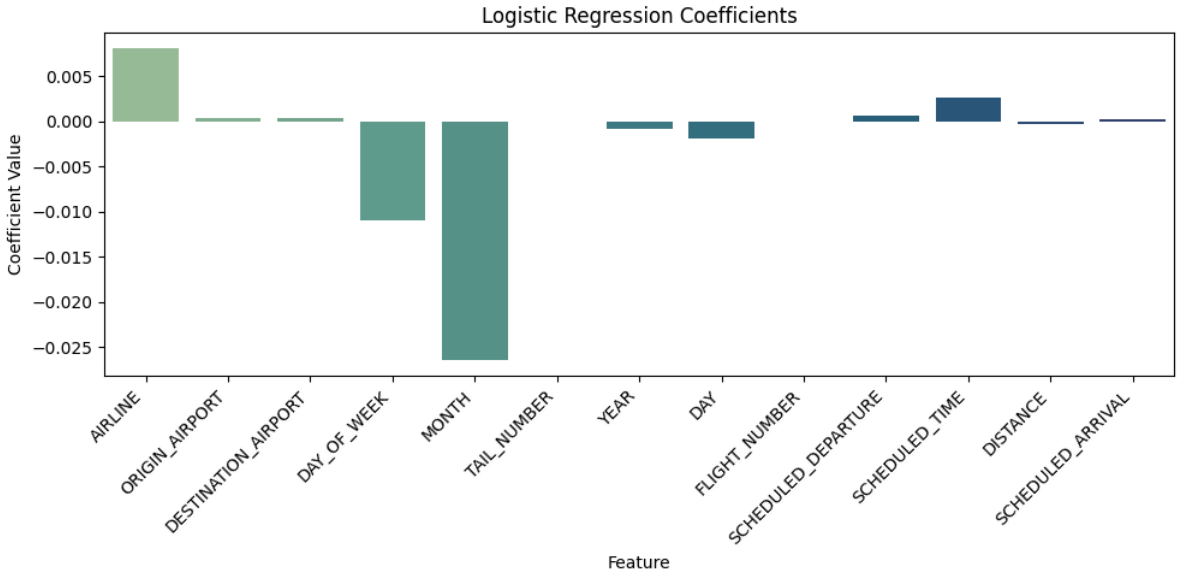
Graf przedstawia strukturę wytrenowanego drzewa decyzyjnego.

Wizualizacja modeli - Random Forest



Graf przedstawia strukturę pierwszego z drzew w wytrenowanym lesie losowym.

Wizualizacja modeli - Regresja Logistyczna



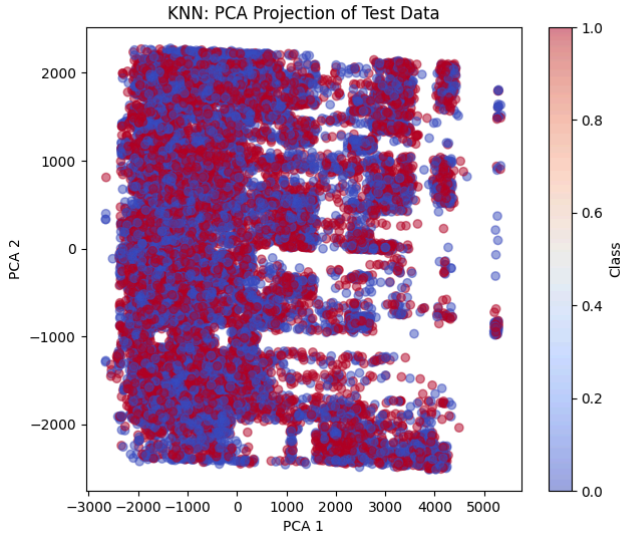
Wizualizacja modeli - Regresja Logistyczna (ii)

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:

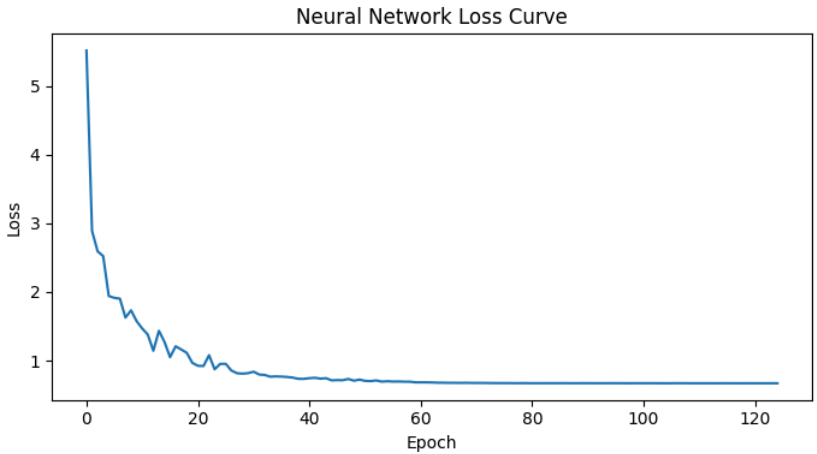
- `MONTH`
- `DAY_OF_WEEK`
- `SCHEDULED_TIME`
- `AIRLINE`

Wizualizacja modeli - 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.

Wizualizacja modeli - 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.

Eksperymenty ze zbiorem danych

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.

Optymalizacja hiperparametrów (ii)

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

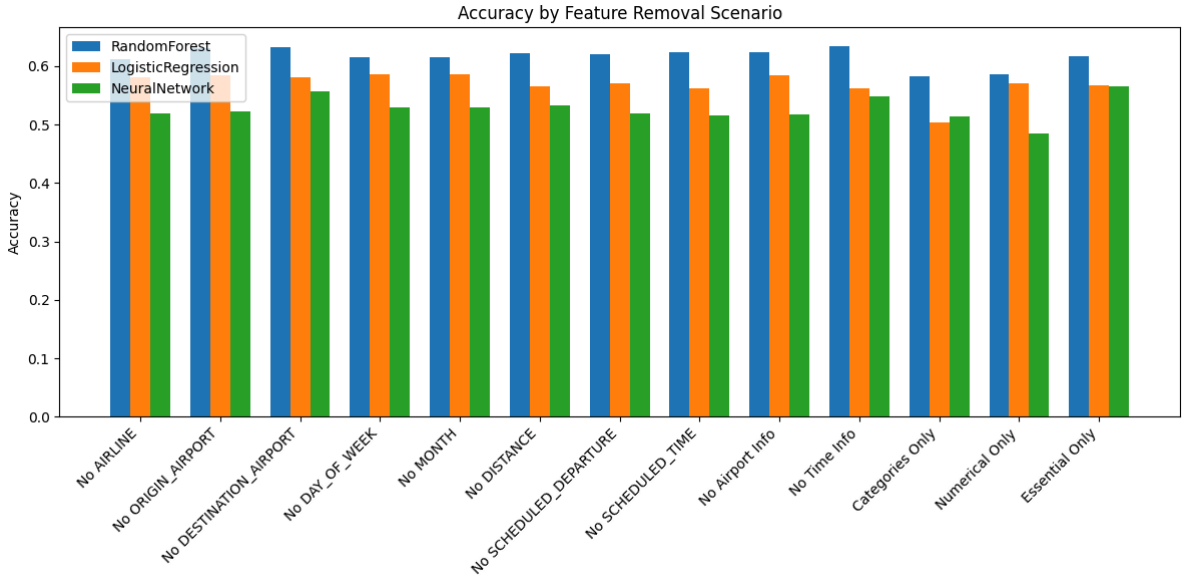
Neural Network

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

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:

Usunięcie cech (ii)



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%

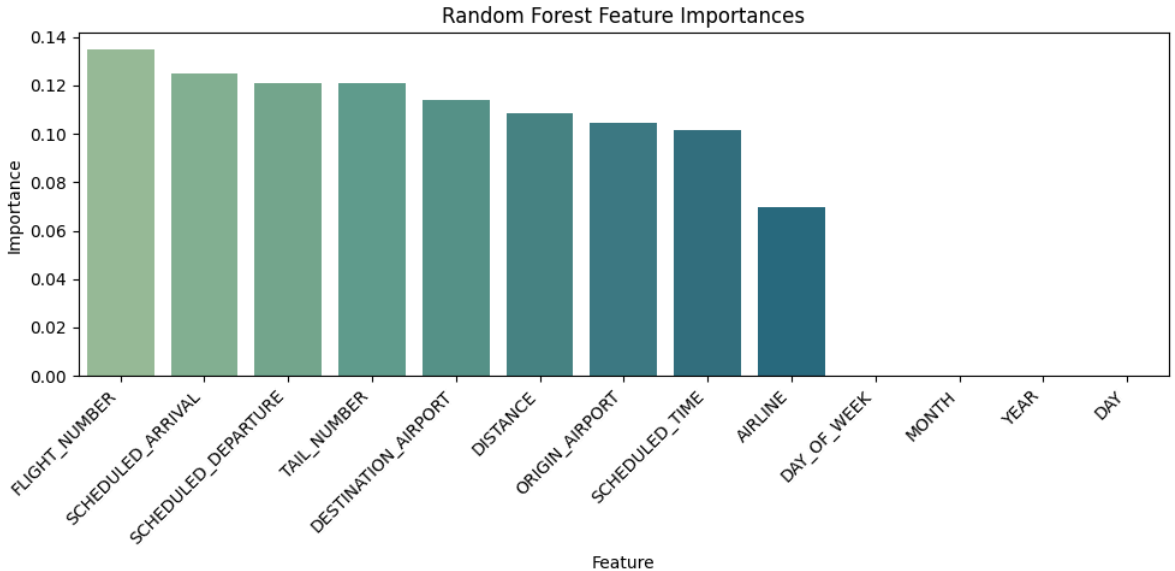
Neural Network (dokładność nominalna: 62.50%)

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%

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.

Ważność cech (ii)



Ważność cech (iii)

Najważniejszymi cechami są:

- `FLIGHT_NUMBER`
- `SCHEDULED_ARRIVAL`
- `SCHEDULED_DEPARTURE`
- `TAIL_NUMBER`

Atrybuty MONTH, YEAR i DAY zostały uznane za nieistotne.

Podsumowanie

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.

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.

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.

Dziękujemy za uwagę!