|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| TOMASZ CHMIEL | |  | 2024 | PSI | | | | 1 | |
|  | **AUTOR** | **ROK AKAD.** |  | | | **SPEC.** |  | **GRUPA** |
| UCZENIE MASZYNOWE | | | | |  | | | 1 | |
| **TEMAT** | | | **NR SPRAWOZDANIA** | |
| 12.02.2024 | Laboratorium maszynowej analizy danych | | | | | Anna Czaderna-Lekka | | | |
| **DATA WYKONANIA** | **PRZEDMIOT** | | | | | **PROWADZĄCY** | | | |

# Cel

Celem projektu jest zbudowanie modeli klasyfikacyjnych do przewidywania, czy klient zrezygnuje z korzystania z usług firmy, na podstawie dostępnych danych dotyczących cech demograficznych, rodzaju korzystanych usług i szczegółów płatności. Projekt obejmuje następujące etapy:

1. **Analiza eksploracyjna danych (EDA):**
   * Sprawdzenie kompletności danych i identyfikacja brakujących wartości.
   * Obliczenie podstawowych statystyk dla cech ilościowych oraz wizualizacja rozkładu cech kategorycznych.
   * Zbadanie zależności pomiędzy decyzją klienta o rezygnacji z usług a wybranymi cechami, takimi jak rodzaj umowy, rodzaj usługi internetowej czy długość korzystania z usług.
2. **Przetwarzanie danych:**
   * Usunięcie cech, które nie są istotne dla przewidywania (np. identyfikator klienta).
   * Uzupełnienie brakujących danych, jeśli to konieczne.
   * Przekształcenie cech kategorycznych na wartości numeryczne za pomocą kodowania.
   * Normalizacja cech ilościowych w celu zapewnienia spójności skali w modelach wymagających tej operacji.
3. **Budowa modeli:**
   * Stworzenie trzech klasyfikatorów: drzewa decyzyjnego, regresji logistycznej oraz k-najbliższych sąsiadów (KNN).
   * Optymalizacja hiperparametrów każdego z modeli przy użyciu techniki wyszukiwania siatki (GridSearchCV).
4. **Ewaluacja modeli:**
   * Ocena skuteczności modeli za pomocą raportów klasyfikacji, macierzy błędów oraz wskaźników takich jak precyzja, czułość, F1-score i dokładność.
   * Porównanie wyników modeli z domyślnymi parametrami i zoptymalizowanymi ustawieniami, aby ocenić wpływ optymalizacji na ich skuteczność.

# Materiały i metody

**Zbiór danych**

Do realizacji zadania wykorzystano zbiór danych **"Telco-Customer-Churn"**, który zawiera informacje o klientach telekomunikacyjnych oraz ich decyzjach o rezygnacji z usług. Zbiór danych składa się z 7043 próbek, z których każda opisuje jednego klienta. Zawiera różne cechy demograficzne i usługowe, takie jak płeć, wiek, rodzaj umowy, długość korzystania z usług (tenure), miesięczne opłaty oraz dane dotyczące różnych usług (np. telefonia, internet, ochrona urządzeń, wsparcie techniczne).

**Techniki analizy**

* **Regresja logistyczna**: Regresja logistyczna została zastosowana do przewidywania, czy dany klient zdecyduje się na rezygnację z usług ("Churn"), na podstawie cech takich jak płeć, rodzaj umowy, długość korzystania z usług, całkowite opłaty i inne zmienne demograficzne. Regresja logistyczna jest odpowiednia, ponieważ jest to zadanie klasyfikacji binarnej (rezygnacja vs. brak rezygnacji).
* **GridSearchCV**: Aby znaleźć optymalne ustawienia dla modelu regresji logistycznej, zastosowano przeszukiwanie siatki (GridSearchCV), testując różne wartości parametrów, takich jak C (współczynnik regularyzacji) i solver (algorytm optymalizacji). Optymalizacja parametrów ma na celu poprawienie wydajności modelu i uzyskanie jak najlepszego dopasowania do danych.
* **Ewaluacja modelu**: Model został oceniony na podstawie:
  + **Raportu klasyfikacji**: Zawiera miary takie jak dokładność, precyzja, recall, F1-score dla obu klas (Churn = 0, 1).
  + **Macierzy błędów**: Ukazuje, ile razy model prawidłowo lub błędnie sklasyfikował przypadki.
  + **Krzywa ROC** i **AUC**: Służy do oceny jakości modelu pod względem jego zdolności do rozróżniania klas (czy klient zrezygnuje, czy nie).
* **Procedury przygotowania danych**:
  + **Wczytanie danych**: Zbiory danych zostały wczytane z plików CSV za pomocą funkcji pd.read\_csv().
  + **Czyszczenie danych**: W przypadku brakujących wartości (np. w kolumnie "TotalCharges" lub "tenure"), dane zostały odpowiednio przetworzone — np. brakujące wartości w kolumnie "TotalCharges" zostały usunięte.
  + **Kodowanie danych kategorialnych**: Zmienna kategoryczna "Churn" została zakodowana na wartości numeryczne (0 = "No", 1 = "Yes"), a inne zmienne kategoryczne (np. płeć, rodzaj umowy) zostały przekonwertowane na format numeryczny przy pomocy **one-hot encoding** (np. zmienne "InternetService", "DeviceProtection" były kodowane na 0 lub 1).
  + **Skalowanie danych**: Zastosowano **StandardScaler** do standaryzacji cech, co poprawia konwergencję algorytmu regresji logistycznej, zwłaszcza gdy różne cechy mają różne zakresy wartości.
  + **Podział na zbiory treningowy i testowy**: Zbiór danych został podzielony na zbiór treningowy (75%) i testowy (25%), co umożliwia ocenę modelu na danych, które nie były użyte do jego treningu.

# Wyniki i dyskusja

**1. Jaki procent klientów odchodzi, a ilu kontynuuje korzystanie z aktywnych usług?**

Wyniki pokazują, że:

* **26.58%** klientów odeszło (Churn = 1).
* **73.42%** klientów kontynuuje korzystanie z usług (Churn = 0).

Oznacza to, że stosunkowo duży odsetek klientów zdecydował się na rezygnację z usług, ale nadal większość klientów pozostaje z firmą.

**2. Czy istnieją wzorce wśród odchodzących klientów w zależności od płci?**

Wyniki pokazują, że:

* **Dla kobiet**: 26.96% odchodzących, a 73.04% kontynuuje korzystanie z usług.
* **Dla mężczyzn**: 26.20% odchodzących, a 73.80% kontynuuje korzystanie z usług.

Zatem nie ma znaczącej różnicy między płciami, jeśli chodzi o procent klientów, którzy odchodzą. Mężczyźni mają minimalnie mniejszy odsetek odchodzących klientów niż kobiety.

**3. Czy są jakieś wzorce/preferencje wśród odchodzących klientów w zależności od rodzaju oferowanych usług?**

Wyniki pokazują różnice w wskaźnikach churn w zależności od rodzaju usług:

* **PhoneService**:
  + Klienci, którzy **nie korzystają z telefonu** (PhoneService = No), mają **25% churn rate**.
  + Klienci, którzy **korzystają z telefonu** (PhoneService = Yes), mają **26.75% churn rate**.
* **InternetService**:
  + **DSL**: 18.99% odchodzących, 81.00% kontynuuje.
  + **Fiber optic**: 41.89% odchodzących, 58.11% kontynuuje.
  + **Brak internetu**: 7.43% odchodzących, 92.57% kontynuuje.

**Wniosek**: Klienci z **Fiber optic** mają najwyższy wskaźnik churn, podczas gdy ci bez internetu (No InternetService) mają najmniejszy wskaźnik churn.

* **OnlineSecurity**:
  + **Brak online security**: 41.78% odchodzących.
  + **Brak internetu**: 7.43% odchodzących.

**Wniosek**: Klienci, którzy **nie korzystają z OnlineSecurity**, mają wyższy wskaźnik churn.

* **OnlineBackup**:
  + **Brak backupu**: 39.94% odchodzących.
  + **Backup w ofercie**: 21.57% odchodzących.

**Wniosek**: Klienci, którzy **korzystają z OnlineBackup**, mają znacznie niższy wskaźnik churn.

* **DeviceProtection**:
  + **Brak ochrony urządzenia**: 39.14% odchodzących.
  + **Ochrona urządzenia w ofercie**: 22.54% odchodzących.

**Wniosek**: Klienci, którzy **korzystają z ochrony urządzenia**, mają niższy wskaźnik churn.

* **TechSupport**:
  + **Brak wsparcia technicznego**: 41.65% odchodzących.
  + **Wsparcie techniczne w ofercie**: 15.20% odchodzących.

**Wniosek**: Klienci, którzy **korzystają z wsparcia technicznego**, mają znacznie niższy wskaźnik churn.

* **StreamingTV**:
  + **Brak TV na żądanie**: 33.54% odchodzących.
  + **Streaming TV w ofercie**: 30.11% odchodzących.
* **StreamingMovies**:
  + **Brak Streaming Movies**: 33.73% odchodzących.
  + **Streaming Movies w ofercie**: 29.95% odchodzących.

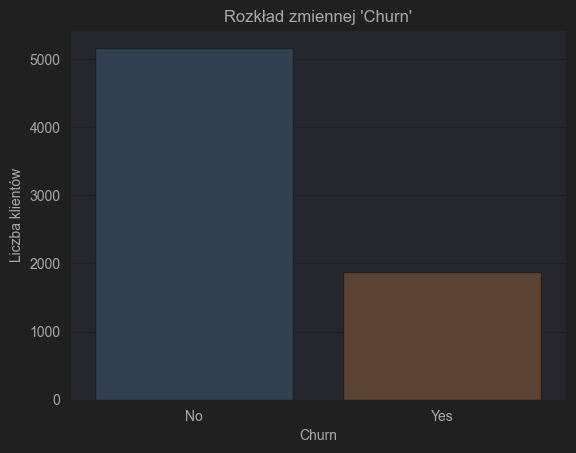
**Wniosek**: Wskaźnik churn dla klientów korzystających z usług typu **Streaming TV** i **Streaming Movies** jest nieco niższy niż dla tych, którzy nie korzystają z tych usług.

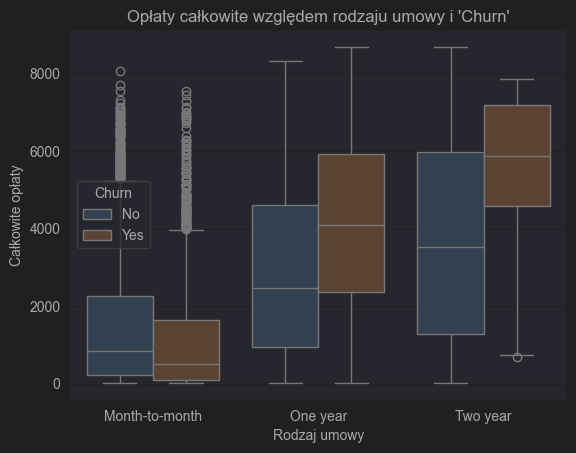
**4. Jakie typy usług są najbardziej dochodowe?**

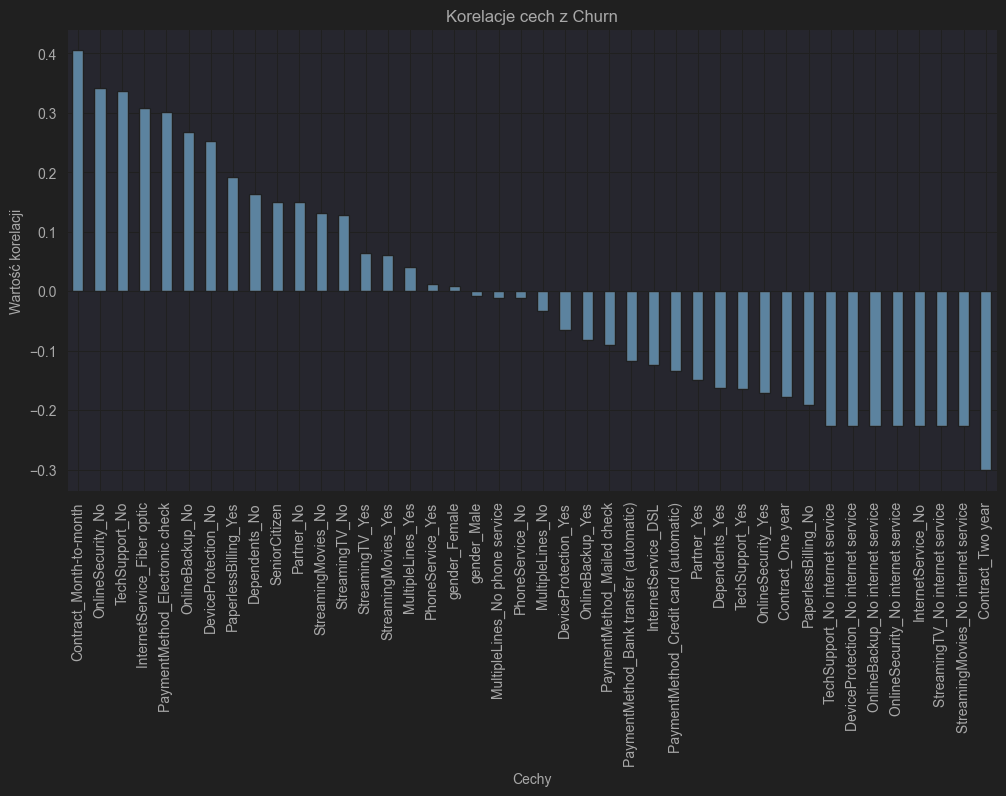
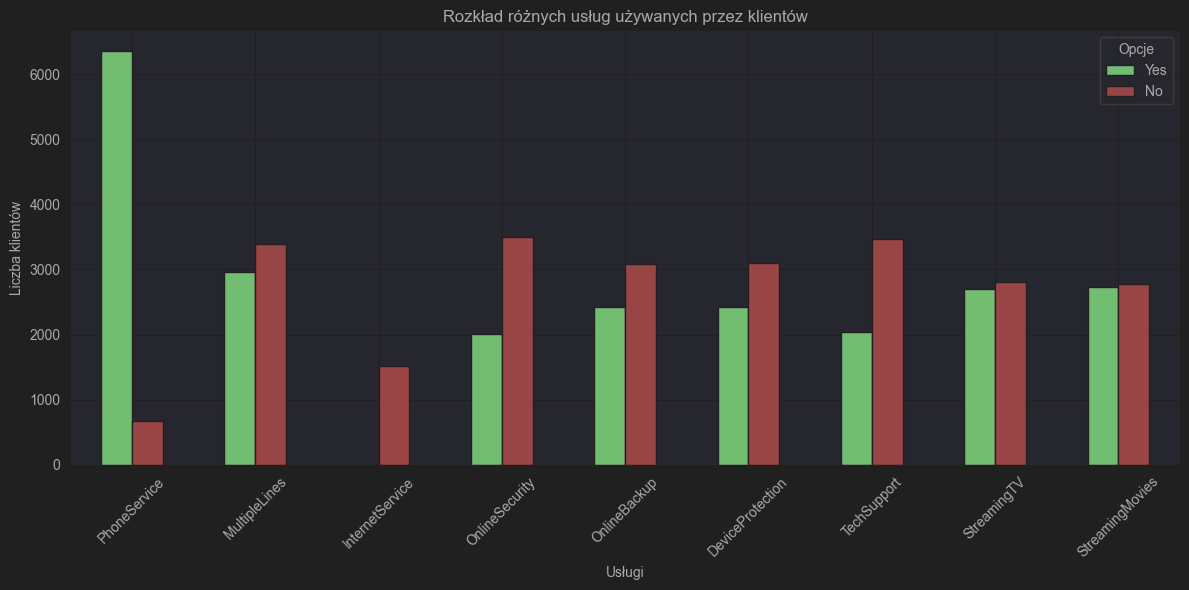
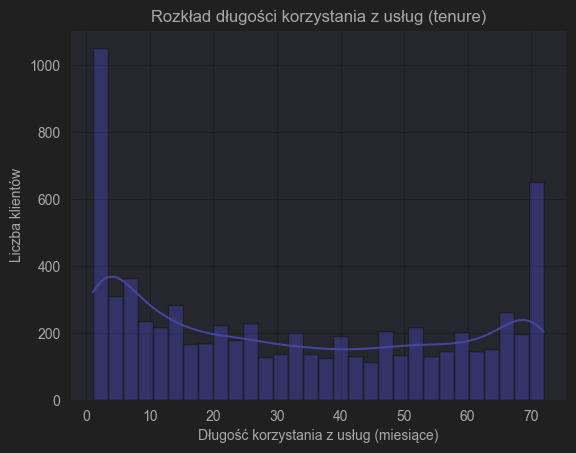
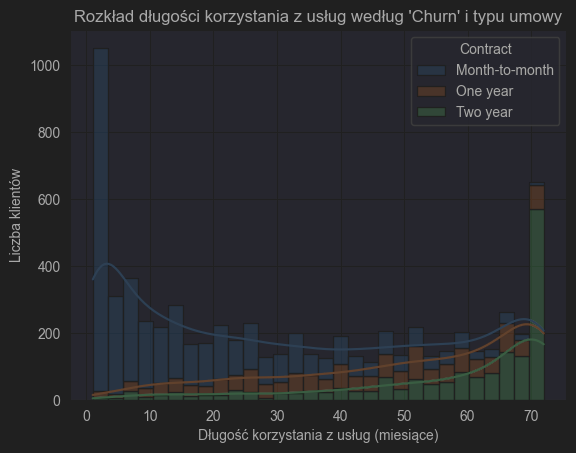
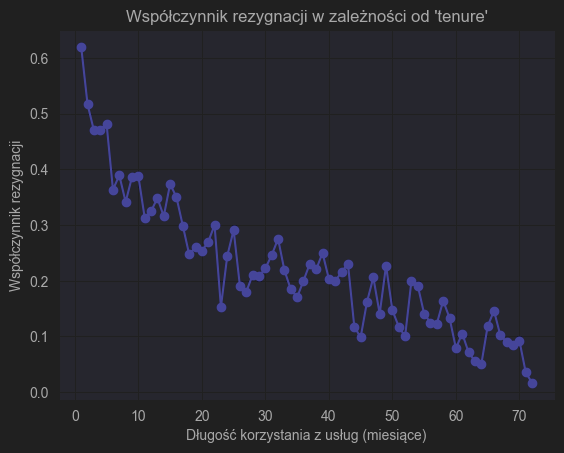
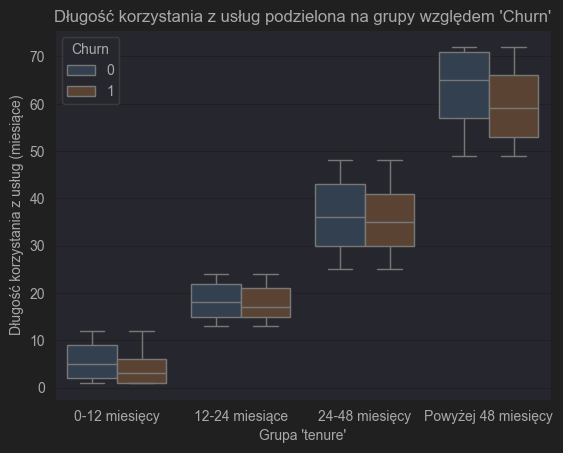
Na podstawie średnich opłat miesięcznych możemy określić, które usługi generują najwyższe dochody:

* **PhoneService**:
  + Klienci, którzy **korzystają z telefonu** (PhoneService = Yes), płacą średnio **67.24** jednostek waluty miesięcznie.
  + Klienci, którzy **nie korzystają z telefonu** (PhoneService = No), płacą średnio **41.99** jednostek waluty miesięcznie.
* **InternetService**:
  + **Fiber optic**: Średnia miesięczna opłata wynosi **91.50** jednostek waluty.
  + **DSL**: Średnia miesięczna opłata wynosi **58.09** jednostek waluty.
  + **Brak internetu**: Średnia miesięczna opłata wynosi **21.08** jednostek waluty.

**Wniosek**: **Fiber optic** jest najbardziej dochodową usługą, generującą najwyższe miesięczne opłaty.

** *A screenshot of a graph

Description automatically generated* 

***Klasyfikatory przed optymalizacją:***

**1. Drzewo Decyzyjne (wersja domyślna):**

* **Precision (0)**: 0.80, **Recall (0)**: 0.80, **F1-Score (0)**: 0.80
* **Precision (1)**: 0.46, **Recall (1)**: 0.47, **F1-Score (1)**: 0.46
* **Accuracy**: 0.71
* **Macro avg**: 0.63
* **Weighted avg**: 0.71

Wnioski:

* Model dobrze klasyfikuje klasę 0 (klienci, którzy nie odeszli), ale ma **znacznie gorsze wyniki** dla klasy 1 (klienci, którzy odeszli).
* **F1-score dla klasy 1** wynosi tylko 0.46, co sugeruje, że model nie radzi sobie dobrze z przewidywaniem rezygnacji klientów.
* **Accuracy** 71% oznacza, że model poprawnie klasyfikuje około 71% wszystkich przypadków, ale widać, że problemy występują przy klasyfikacji klasy 1.

**2. K-Najbliżsi Sąsiedzi (wersja domyślna):**

* **Precision (0)**: 0.81, **Recall (0)**: 0.83, **F1-Score (0)**: 0.82
* **Precision (1)**: 0.50, **Recall (1)**: 0.46, **F1-Score (1)**: 0.48
* **Accuracy**: 0.73
* **Macro avg**: 0.65
* **Weighted avg**: 0.73

Wnioski:

* **F1-score** dla klasy 1 wynosi 0.48, co oznacza, że model nie jest w stanie efektywnie klasyfikować klientów, którzy odchodzą.
* Podobnie jak w przypadku drzewa decyzyjnego, **precision** i **recall** dla klasy 1 są słabe.
* Model ma jednak **nieco lepszą dokładność** (73%) w porównaniu do drzewa decyzyjnego.

**3. Regresja Logistyczna (ustawienia domyślne):**

* **Precision (0)**: 0.81, **Recall (0)**: 0.87, **F1-Score (0)**: 0.84
* **Precision (1)**: 0.56, **Recall (1)**: 0.44, **F1-Score (1)**: 0.50
* **Accuracy**: 0.76
* **Macro avg**: 0.69
* **Weighted avg**: 0.75

Wnioski:

* Regresja logistyczna radzi sobie **najlepiej** w klasyfikacji klasy 0 (klienci pozostający). **Recall** dla tej klasy wynosi 0.87, co oznacza, że model poprawnie identyfikuje większość klientów, którzy nie odchodzą.
* Model ma **lepszy wynik F1-score** dla klasy 1 (0.50), ale wciąż nie radzi sobie zbyt dobrze z przewidywaniem rezygnacji klientów.
* **Accuracy** 76% oznacza, że model ogólnie radzi sobie dobrze, ale nadal jest trudność w przewidywaniu churnu (rezygnacji).

**Który niezoptymalizowany klasyfikator działał najlepiej?**

Na podstawie wyników:

* **Regresja logistyczna** działa najlepiej z trzech niezoptymalizowanych klasyfikatorów, osiągając **najwyższą dokładność** (76%) oraz **najlepszy wynik F1** dla klasy 0 (0.84). Chociaż wyniki dla klasy 1 są nadal niewystarczające (F1 = 0.50), **regresja logistyczna** daje ogólnie lepsze wyniki niż **drzewo decyzyjne** (F1 = 0.46) i **k-najbliżsi sąsiedzi** (F1 = 0.48).

**Macierz błędów dla każdego z klasyfikatorów przed optymalizacją:A screenshot of a graph

Description automatically generated** **A screenshot of a graph

Description automatically generated** A screenshot of a computer

Description automatically generated

***Klasyfikatory po optymalizacji:***

**1. Drzewo Decyzyjne (po optymalizacji):**

* **Precision (0)**: 0.85, **Recall (0)**: 0.85, **F1-Score (0)**: 0.85
* **Precision (1)**: 0.58, **Recall (1)**: 0.58, **F1-Score (1)**: 0.58
* **Accuracy**: 0.78
* **Macro avg**: 0.72
* **Weighted avg**: 0.78

**Wnioski**:

* **Dla klasy 0** (klienci, którzy nie odeszli), model osiągnął bardzo dobre wyniki z **precision** i **recall** na poziomie 0.85, co oznacza, że model skutecznie klasyfikuje klientów, którzy pozostają.
* **Dla klasy 1** (klienci, którzy odeszli), wyniki są poprawione w porównaniu do wersji przed optymalizacją: **precision** i **recall** wynoszą 0.58, a **F1-score** wynosi 0.58, co jest lepszym wynikiem niż w wersji domyślnej (0.46).
* **Accuracy** wzrosło do **78%**, co wskazuje na znaczną poprawę w ogólnym dopasowaniu modelu.

**2. K-Najbliżsi Sąsiedzi (po optymalizacji):**

* **Precision (0)**: 0.82, **Recall (0)**: 0.89, **F1-Score (0)**: 0.86
* **Precision (1)**: 0.60, **Recall (1)**: 0.46, **F1-Score (1)**: 0.52
* **Accuracy**: 0.78
* **Macro avg**: 0.71
* **Weighted avg**: 0.77

**Wnioski**:

* **Dla klasy 0** (klienci, którzy nie odeszli), model uzyskał wysokie wyniki w zakresie **precision** (0.82), **recall** (0.89) i **F1-score** (0.86), co świadczy o skuteczności w klasyfikowaniu klientów, którzy pozostają.
* **Dla klasy 1** (klienci, którzy odeszli), wyniki są poprawione, ale nadal pozostają słabe: **precision** (0.60) i **recall** (0.46), a **F1-score** wynosi tylko 0.52.
* **Accuracy** wynosi **78%**, co wskazuje na ogólnie dobrą jakość modelu, choć nadal jest pole do poprawy dla klasy 1.

**3. Regresja Logistyczna (po optymalizacji):**

* **Precision (0)**: 0.87, **Recall (0)**: 0.84, **F1-Score (0)**: 0.85
* **Precision (1)**: 0.59, **Recall (1)**: 0.63, **F1-Score (1)**: 0.61
* **Accuracy**: 0.79
* **Macro avg**: 0.73
* **Weighted avg**: 0.79

**Wnioski**:

* **Dla klasy 0** (klienci, którzy nie odeszli), model uzyskał bardzo wysokie wyniki: **precision** (0.87), **recall** (0.84), **F1-score** (0.85), co oznacza, że jest bardzo skuteczny w klasyfikacji klientów, którzy pozostają.
* **Dla klasy 1** (klienci, którzy odeszli), wyniki są również poprawione w porównaniu do wersji przed optymalizacją: **precision** wynosi 0.59, **recall** 0.63, a **F1-score** 0.61, co jest znaczną poprawą w porównaniu do wersji domyślnej (F1-score = 0.50).
* **Accuracy** wzrosło do **79%**, co świadczy o lepszym dopasowaniu modelu do danych.

**Macierz błędów dla każdego z klasyfikatorów po optymalizacji:**

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated A screenshot of a graph

Description automatically generated

# Podsumowanie

Na podstawie wyników po optymalizacji, **regresja logistyczna** okazała się najlepszym klasyfikatorem. Jej dokładność predykcji wynosi **79%**, co jest najwyższym wynikiem spośród zoptymalizowanych klasyfikatorów.

**Który model zoptymalizowany wykazał największą poprawę klasyfikacji względem klasyfikatora niezoptymalizowanego?**

Aby odpowiedzieć na to pytanie, porównamy poprawę **F1-score** i **accuracy** dla każdego klasyfikatora:

**1. Drzewo Decyzyjne (przed i po optymalizacji):**

* **Accuracy**: Zwiększyło się z **71%** do **78%**.
* **F1-score dla klasy 1**: Zwiększyło się z **0.46** do **0.58**.

**2. K-Najbliżsi Sąsiedzi (przed i po optymalizacji):**

* **Accuracy**: Zwiększyło się z **73%** do **78%**.
* **F1-score dla klasy 1**: Zwiększyło się z **0.48** do **0.52**.

**3. Regresja Logistyczna (przed i po optymalizacji):**

* **Accuracy**: Zwiększyło się z **76%** do **79%**.
* **F1-score dla klasy 1**: Zwiększyło się z **0.50** do **0.61**.

**Największa poprawa:**

* **Regresja logistyczna** wykazała **największą poprawę** w klasyfikacji **klasy 1** (odchodzących klientów), gdzie **F1-score** wzrosło o **0.11** (z 0.50 do 0.61). Ponadto, poprawa w **accuracy** wyniosła **3%** (z 76% do 79%).
* **Drzewo decyzyjne** poprawiło się o **7%** w **accuracy** (71% do 78%), a poprawa F1 dla klasy 1 wyniosła **0.12** (0.46 do 0.58).
* **K-Najbliżsi Sąsiedzi** poprawili **accuracy** o **5%** (73% do 78%), ale poprawa F1 dla klasy 1 była najmniejsza (tylko **0.04**).

Przeprowadzona optymalizacja wskazuje na istotność dostosowywania hiperparametrów w celu uzyskania lepszych wyników klasyfikacji, również w zadaniach przewidywania churnu (rezygnacji klientów).