Sprawozdanie: Analiza danych zakupowych i klasyfikacja klientów

# 1. Opis danych

Dane pochodziły z pliku CSV zawierającego informacje o klientach odwiedzających stronę internetową. Zmienna docelowa to 'kupil', przyjmująca wartości 0 (nie kupił) i 1 (kupił). Wśród pozostałych zmiennych znalazły się dane liczbowe (wiek, przychód, czas na stronie, liczba wizyt, godzina wejścia, dzień tygodnia) oraz kategoryczne (region, typ klienta).

# 2. Eksploracja danych

Dane zostały przeanalizowane eksploracyjnie. Sporządzono histogramy dla zmiennych liczbowych, które ujawniły charakterystyczne rozkłady wieku, przychodu czy aktywności klientów. Szczególnie ciekawa była analiza liczby zakupów w zależności od wieku, która pokazała większą aktywność zakupową w określonych przedziałach wiekowych.

# 3. Przetwarzanie danych

W celu przygotowania danych do modelowania zakodowano zmienne kategoryczne metodą one-hot encoding. Dodatkowo, zastosowano technikę undersamplingu, aby zbalansować liczebność klas w zmiennej docelowej. Usunięto brakujące dane (nie występowały), a zmienne liczbowo-kategoryczne ujednolicono.

# 4. Podział danych

Dane podzielono na zbiór treningowy i testowy w proporcji 80:20. Dzięki temu możliwe było skuteczne uczenie modeli oraz ich późniejsza ocena na niezależnym zbiorze danych.

# 5. Budowa i ocena modeli

Zbudowano i przetestowano trzy modele klasyfikacyjne: regresję logistyczną, Random Forest oraz XGBoost. Wyniki ich skuteczności przedstawiały się następująco:

• Regresja logistyczna: accuracy ~75%, f1-score ~0.75, AUC ~0.79  
• Random Forest: accuracy ~98%, f1-score ~0.98, AUC ~0.999  
• XGBoost: accuracy ~83%, f1-score ~0.83, AUC ~0.91

Model Random Forest osiągnął najwyższe wyniki, jednak wstępna analiza sugerowała możliwość przeuczenia (overfittingu).

# 6. Tuning hiperparametrów

Przeprowadzono tuning hiperparametrów dla modelu Random Forest przy użyciu GridSearchCV. Dobierano m.in. liczbę drzew, maksymalną głębokość drzew oraz sposób podziału danych. Działania te miały na celu poprawę skuteczności modelu i zminimalizowanie ryzyka nadmiernego dopasowania do danych treningowych.

# 7. Overfitting – sprawdzenie i interpretacja

Model Random Forest uzyskał 100% trafności na zbiorze treningowym, co wskazywało na ryzyko overfittingu. Po sprawdzeniu skuteczności na zbiorze testowym (accuracy: 98%, AUC: 0.998), zauważono nieznaczne obniżenie skuteczności, co potwierdzało obecność przeuczenia, choć model nadal działał bardzo dobrze. Zmiana strategii na undersampling zmniejszyła trafność do 79%, ale pozwoliła uzyskać bardziej realistyczne, zrównoważone wyniki.

# 8. Wnioski

Na podstawie analizy danych i oceny modeli można stwierdzić, że:  
- Dane wymagają balansu klas dla uczciwej oceny modeli.  
- Random Forest był najskuteczniejszy, ale najbardziej podatny na overfitting.  
- Regresja logistyczna zapewnia stabilne wyniki przy niższej złożoności.  
- XGBoost łączy zalety obu podejść, dając wysoką skuteczność przy umiarkowanej podatności na przeuczenie.  
- Przyszłe analizy warto rozszerzyć o dodatkowe metody walidacji krzyżowej i interpretowalność modeli.