1. PRINCIPAL COMPONENT ANALYSIS, PCA

Principal Component Analysis (PCA) to technika analizy danych stosowana w statystyce i analizie wielowymiarowej. Jej głównym celem jest redukcja wymiarowości danych, co oznacza zmniejszenie liczby zmiennych lub cech w zestawie danych. PCA działa poprzez transformację oryginalnych zmiennych (cech) w nowy zestaw zmiennych, zwanych głównymi składowymi, które są liniowymi kombinacjami oryginalnych cech.

Proces PCA polega na znalezieniu kierunków (głównych składowych), wzdłuż których wariancja danych jest największa. Pierwsza główna składowa obejmuje kierunek o największej wariancji, a każda kolejna składowa jest ortogonalna do poprzednich i obejmuje maksymalną pozostałą wariancję. Dzięki temu, można zredukować wymiarowość danych poprzez odrzucenie mniej istotnych składowych, zachowując jednocześnie jak najwięcej informacji.

Zastosowania PCA obejmują:

1. **Redukcję wymiarowości:** Pomaga w eliminowaniu zbędnych cech, co może poprawić efektywność analizy danych i przyspieszyć algorytmy uczące maszyny.
2. **Wizualizację danych:** Pozwala na przedstawienie danych w przestrzeni o mniejszej liczbie wymiarów, co ułatwia zrozumienie struktury danych.
3. **Usunięcie współliniarności:** PCA może pomóc w radzeniu sobie z problemem współliniarności, gdy pewne cechy są silnie skorelowane.
4. **Kompresję danych:** Może być używane do kompresji danych, zachowując jednocześnie istotne informacje.

W skrócie, PCA to narzędzie służące do analizy i redukcji wymiarowości danych, co może przyczynić się do lepszego zrozumienia struktury danych i poprawić efektywność analizy danych.

2. KLASTROWANIE (POPRAWA JAKOŚCI, PORÓWNYWANIE WYNIKÓW)

Klastrowanie (ang. clustering) to technika analizy danych, która polega na grupowaniu zbioru danych na podstawie pewnych kryteriów, tak aby obiekty w tych samych grupach (klasterach) były bardziej podobne do siebie nawzajem niż do obiektów spoza tych grup. Klaster to zbiór obiektów, które mają podobne właściwości, a klastrowanie ma na celu znalezienie struktury w danych bez wcześniej określonych etykiet klas.

Podstawową ideą klastrowania jest maksymalizacja podobieństwa wewnątrz klastrów i minimalizacja podobieństwa między klastrami. W skrócie, chcemy, aby obiekty w obrębie jednego klastra były sobie nawzajem bardziej podobne niż do obiektów spoza klastra.

Istnieje wiele różnych podejść do klastrowania, w tym:

1. **Klastrowanie hierarchiczne:** Grupowanie zaczyna się od pojedynczych obiektów, a następnie łączy się je w hierarchiczne struktury klastrów.

-algorytmy aglomeracyjne- definiujemy każdy punkt danych jako osobny klaster. W każdym kroku dwa najbliższe klastry są łączone w jeden klaster.

-algorytmy podziałowe- wszystkie punkty danych są umieszczone w jednym klastrze. W każdym kroku istniejący klaster jest dzielony na dwa

2. **Klastrowanie k-średnich (k-means):** Algorytm ten przyporządkowuje punkty danych do klastrów na podstawie ich odległości od środków klastrów, minimalizując sumę kwadratów odległości między punktami a środkami klastrów.
3. **Metoda K-MEDOIDÓW-** Medoid klastra definiuje się jako obiekt w klastrze, którego średnia odmienność od wszystkich obiektów w klastrze jest minimalna, to znaczy jest to najbardziej centralnie położony punkt w klastrze.

Każdy klaster jest reprezentowany przez jeden ze swoich obiektów

Celem metody jest znalezienie k obiektów reprezentujących k klastrów i minimalizujących przyjętą funkcję kryterialną

-algorytm **k-średnich** nie jest odporny na obserwacje odstające.

-skoro zamiast średniej można użyć odpornej mediany, to zamiast środka ciężkości (centroidu) klastra można użyć jego **medoidu.**

4. **Klastrowanie aglomeracyjne:** Rozpoczyna się od pojedynczych obiektów, a następnie łączy je w klastry, zaczynając od najbardziej podobnych.
5. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** Algorytm ten opiera się na gęstości punktów, identyfikując obszary o większej gęstości jako klastry.
6. **Gaussian Mixture Model (GMM):** Model statystyczny, który zakłada, że dane pochodzą z pewnej liczby rozkładów normalnych (gaussowskich), co umożliwia modelowanie klastrów o różnych kształtach.

Zastosowania klastrowania obejmują segmentację rynku, analizę grup społecznych w mediach społecznościowych, analizę genetyczną, rozpoznawanie wzorców w obrazach, itp.

3. KLASYFIKACJA

Klasyfikacja to rodzaj zadania w uczeniu maszynowym, które polega na przyporządkowywaniu obiektów do określonych klas lub kategorii na podstawie cech, które są wcześniej znane na podstawie zbioru treningowego. Jest to forma uczenia nadzorowanego, co oznacza, że algorytm uczący jest dostarczony z zestawem danych, który zawiera pary wejść (cechy) i odpowiadające im etykiety klas. Celem algorytmu jest nauczenie się relacji między cechami a klasami, aby móc dokonywać poprawnych predykcji dla nowych danych.

Kluczowe elementy związane z klasyfikacją to:

1. **Zbiór treningowy (training set):** Jest to zbiór danych, który zawiera przykłady wejść i odpowiadające im etykiety klas. Algorytm jest trenowany na tym zbiorze w celu zrozumienia relacji między cechami a klasami.
2. **Funkcje lub cechy (features):** Są to atrybuty lub zmienne opisujące obiekty, które są używane do dokonywania predykcji. Na przykład, jeśli klasyfikujemy zwierzęta, cechami mogą być ich waga, wysokość, rodzaj pożywienia, itp.
3. **Etykiety klas (class labels):** Są to oznaczenia przynależności do konkretnych klas. W przypadku zadania binarnej klasyfikacji może to być np. "0" lub "1", natomiast dla wieloklasowej klasyfikacji mogą to być różne etykiety odpowiadające różnym klasom.
4. **Model klasyfikacyjny:** Jest to matematyczna reprezentacja relacji między cechami a etykietami klas. Model jest trenowany na podstawie danych treningowych i następnie może być używany do predykcji klas dla nowych, nieznanych wcześniej danych.

Popularne algorytmy klasyfikacyjne obejmują:

- **Maszyny wektorów nośnych (Support Vector Machines, SVM)**
- **Drzewa decyzyjne i ich zbiory (Random Forest)**
- **Regresję logistyczną**
- **Algorytmy k-najbliższych sąsiadów (k-Nearest Neighbors, k-NN)**
- **Sieci neuronowe**

Klasyfikacja jest szeroko stosowana w różnych dziedzinach, takich jak rozpoznawanie obrazów, analiza tekstu, diagnostyka medyczna, prognozowanie, identyfikacja biometryczna, analiza obrazu medycznego i obrazowanie medyczne, rozpoznawanie pisma odręcznego, prognoza odejścia klienta, ocena zdolności kredytowej

4. ODKRYWANIE REGUŁ ASOCJACJI

Odkrywanie reguł asocjacji to technika analizy danych, która ma na celu identyfikowanie istniejących relacji między elementami w zbiorze danych. Celem jest znalezienie interesujących zależności, zwanych regułami asocjacji, które wyrażają, że pewne zdarzenia lub elementy często występują razem. Typowym przykładem zastosowania odkrywania reguł asocjacji jest analiza koszyka zakupowego w sklepach.

Podstawowym pojęciem w odkrywaniu reguł asocjacji są tzw. "reguły asocjacyjne", które mają formę "jeżeli A, to B". Oznacza to, że gdy pewien zestaw produktów lub zdarzeń (oznaczonych jako A) występuje, to z dużym prawdopodobieństwem wystąpi także inny zestaw produktów lub zdarzeń (oznaczonych jako B). Przykładowo, reguła asocjacyjna może brzmieć: "Jeżeli klient kupuje chleb, to z dużym prawdopodobieństwem kupi także masło."

Algorytmy stosowane do odkrywania reguł asocjacji obejmują m.in.:

1. **Apriori:** (eksploracja pozioma) Jest to jedna z najbardziej znanych i używanych technik odkrywania reguł asocjacji. Algorytm Apriori opiera się na zasadzie, że jeśli zbiór przedmiotów występuje często, to jego podzbiory również występują często.
2. **FP-growth (Frequent Pattern growth):** (eksploracja pozioma) To inny popularny algorytm, który tworzy strukturę drzewiastą do efektywnego analizowania często występujących wzorców.
3. **Eclat:** (eksploracja pionowa Jest to algorytm, który przeszukuje dane wertykalnie (kolumnami), zamiast poziomo (rzędami), co może być bardziej efektywne w niektórych przypadkach.

Zastosowania odkrywania reguł asocjacji obejmują:

- Analizę koszyka zakupowego w handlu detalicznym.
- Rekomendacje produktów lub treści w systemach rekomendacyjnych.
- Analizę zachowań klientów w marketingu.
- Analizę baz danych transakcyjnych w poszukiwaniu powiązań między różnymi zdarzeniami.
- Przewidywanie strumieni kliknięć na stronie internetowej: która strona zostanie kliknięta jako następna?
- Błędy oprogramowania do wydobywania: gdzie jest prawdopodobny błąd w programie?
- Znajdowanie wysokiej jakości fraz, jednostek atrybutów w obszernym tekście.
- Znalezienie powtarzających się sekwencji DNA i białek w genomenach.

Odkrywanie reguł asocjacji ma zastosowanie w szerokim zakresie dziedzin, gdzie analiza wzorców i związków między danymi jest istotna do podejmowania decyzji i zrozumienia struktury danych.

5. ODKRYWANIE WZORCÓW SEKWENCYJNYCH

Odkrywanie wzorców sekwencyjnych to obszar analizy danych, który koncentruje się na identyfikowaniu powtarzających się sekwencji zdarzeń lub obiektów w zbiorze danych sekwencyjnych. W przeciwieństwie do odkrywania reguł asocjacji, które analizuje jednoczesne wystąpienie zdarzeń, odkrywanie wzorców sekwencyjnych skupia się na relacjach sekwencyjnych między zdarzeniami w czasie.

Przykłady zastosowań odkrywania wzorców sekwencyjnych obejmują analizę logów transakcyjnych, monitorowanie zachowań użytkowników w aplikacjach internetowych, analizę sekwencji genów w biologii molekularnej, czy śledzenie ruchu pacjentów w opiece zdrowotnej.

Oto kilka kluczowych pojęć związanych z odkrywaniem wzorców sekwencyjnych:

1. **Sekwencje:** Są to uporządkowane listy zdarzeń lub obiektów. Przykładem może być sekwencja produktów dodanych do koszyka zakupowego w sklepie online w ciągu pewnego okresu czasu.
2. **Wzorce sekwencyjne:** Są to powtarzające się, istotne dla analizy, sekwencje zdarzeń. Odkrywanie wzorców sekwencyjnych polega na identyfikowaniu tych wzorców w danych.
3. **Tokeny lub elementy sekwencji:** Są to poszczególne zdarzenia lub obiekty w sekwencji. Na przykład, w przypadku analizy transakcji online, tokenami mogą być produkty dodane do koszyka.
4. **Współczynnik wsparcia (support):** To miara częstości występowania danego wzorca sekwencyjnego w zbiorze danych. Wzorzec jest uważany za istotny, jeśli występuje wystarczająco często.
5. **Długość sekwencji:** Określa ilość zdarzeń w sekwencji. Może to być istotne przy analizie, np. w przypadku odkrywania długotrwałych wzorców zachowań.

Algorytmy stosowane do odkrywania wzorców sekwencyjnych obejmują:

- **PrefixSpan:** Jest to algorytm, który skupia się na przeszukiwaniu przestrzeni wzorców sekwencyjnych przy użyciu struktury drzewiastej.
- **GSP (Generalized Sequential Pattern):** Algorytm ten identyfikuje ogólne wzorce sekwencyjne, które mogą obejmować zdarzenia występujące w różnych sekwencjach.

Odkrywanie wzorców sekwencyjnych ma zastosowanie tam, gdzie ważne jest zrozumienie kolejności i relacji między zdarzeniami w czasie, a także w sytuacjach, gdzie analiza dynamiki sekwencji może dostarczyć cennych informacji.

Odkrywanie wzorców sekwencyjnych a reguł asocjacji różnice:

1. **Kontekst czasowy:**
   - **Odkrywanie wzorców sekwencyjnych:** Skupia się na relacjach sekwencyjnych między zdarzeniami w czasie. Analizuje kolejność występowania zdarzeń i szuka powtarzających się sekwencji.
   - **Odkrywanie reguł asocjacji:** Skupia się na jednoczesnym występowaniu zdarzeń, bez brania pod uwagę ich kolejności w czasie. Reguły asocjacyjne opisują, które zdarzenia są często obecne razem.
2. **Typ danych:**
   - **Odkrywanie wzorców sekwencyjnych:** Najczęściej stosowane w danych sekwencyjnych, takich jak logi czasowe, transakcje w sklepach, czy dane czasowe w ogólności.
   - **Odkrywanie reguł asocjacji:** Stosowane w różnych rodzajach danych, niekoniecznie posiadających strukturę sekwencyjną. Przykłady to analiza koszyka zakupowego, bazy danych transakcyjnych itp.
3. **Reprezentacja wyników:**
   - **Odkrywanie wzorców sekwencyjnych:** Wynikiem są sekwencje zdarzeń lub wzorców sekwencyjnych, które pokazują, w jakiej kolejności występują pewne zdarzenia.
   - **Odkrywanie reguł asocjacji:** Wynikiem są reguły opisujące częstość wspólnego występowania zdarzeń, bez określania ich kolejności.

4. **Przykłady zastosowań:**
   - **Odkrywanie wzorców sekwencyjnych:** Stosowane tam, gdzie ważna jest analiza sekwencji zdarzeń w czasie, np. w analizie zachowań użytkowników na stronie internetowej, ruchu pacjentów w opiece zdrowotnej, czy analizie sekwencji genów.
   - **Odkrywanie reguł asocjacji:** Powszechne w analizie transakcji w handlu detalicznym, analizie danych koszyka zakupowego, rekomendacjach produktów itp.

Podsumowując, główna różnica polega na tym, czy analiza skupia się na kolejności zdarzeń w czasie (w przypadku odkrywania wzorców sekwencyjnych) czy jednoczesnym wspólnym występowaniu zdarzeń (w przypadku odkrywania reguł asocjacji). Obydwa podejścia są użyteczne w różnych kontekstach i dla różnych typów danych.

6. ANALIZA I RODZAJE WYKRESÓW

Tabela 1. Rodzaje wykresów dla rozkładu jednej zmiennej.

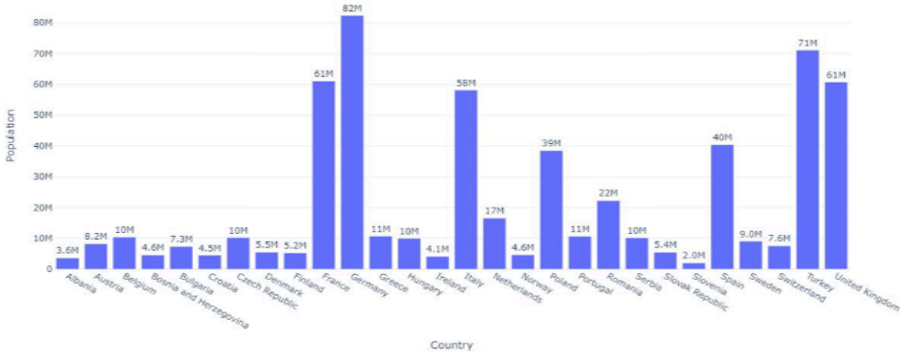
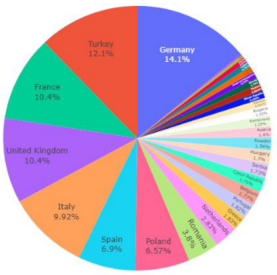
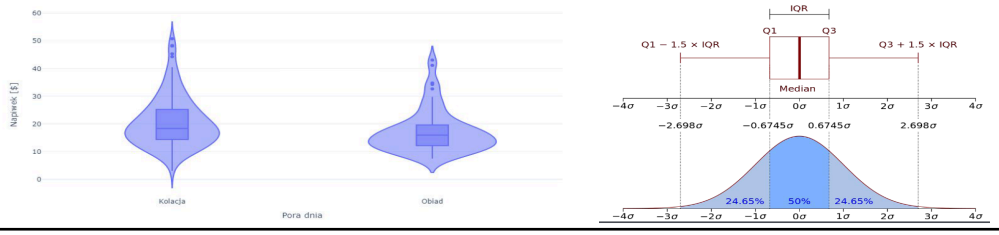
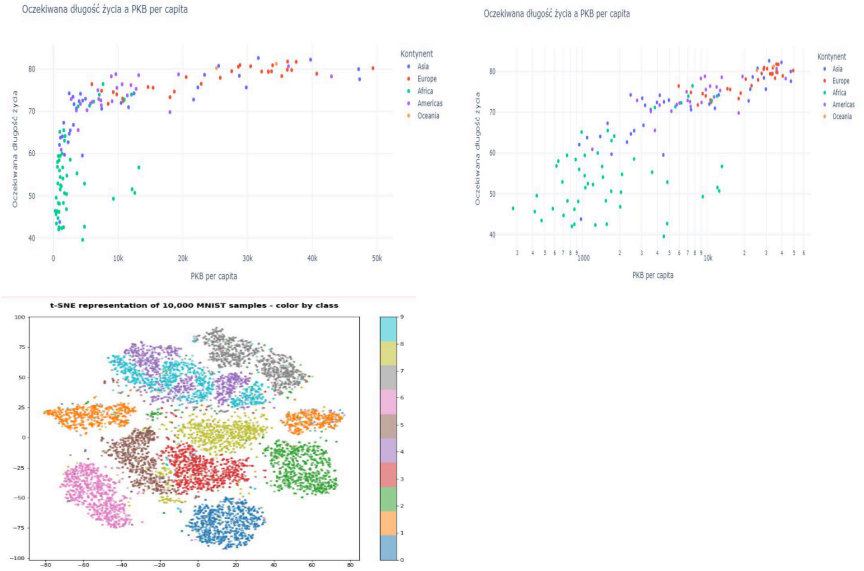
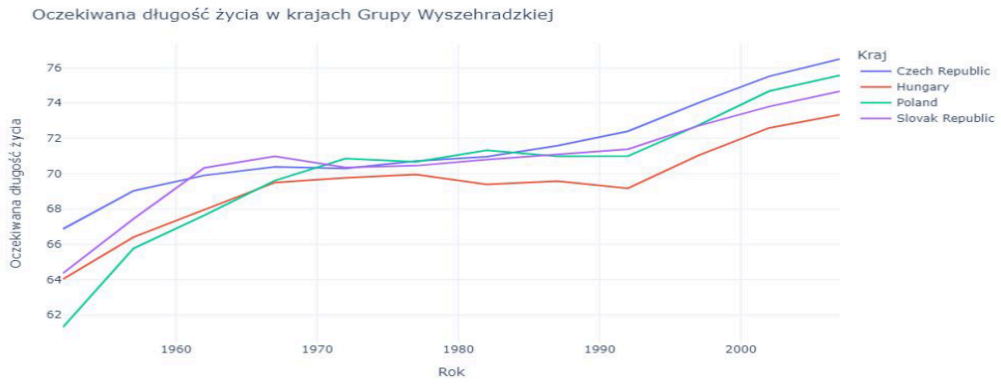
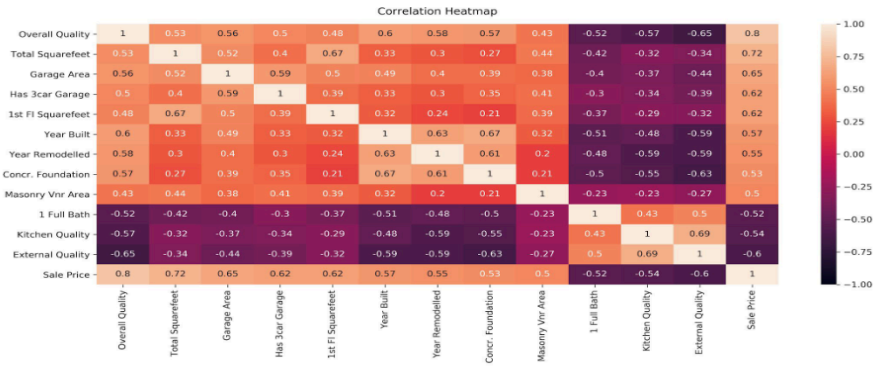
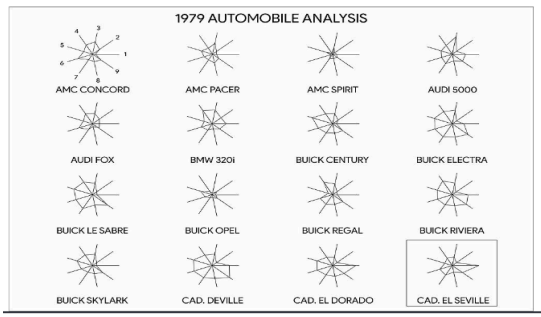
Słupkowy	
Kołowy	
Pudełkowy	
Skrzypcowy	

Tabela 2. Rodzaje wykresów dla reprezentacji relacji między zmiennymi.

Punktowy	
Liniowy	
Mapa cieplna	
Radarowy	 **Zmienne** 1. Cena 2. Zasięg (mpg) 3. Rekord napraw z 1978 r. (1 = najgorszy, 5 = najlepszy) 4. Rekord napraw z 1977 r. (1 = najgorszy, 5 = najlepszy) 5. Miejsce nad głową 6. Przestrzeń na tylnym siedzeniu 7. Przestrzeń bagażnika 8. Waga 9. Długość

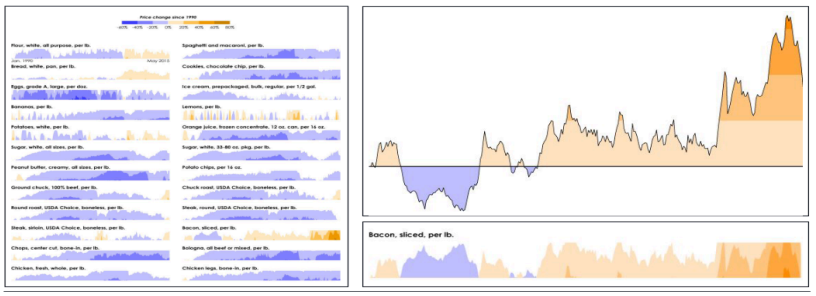
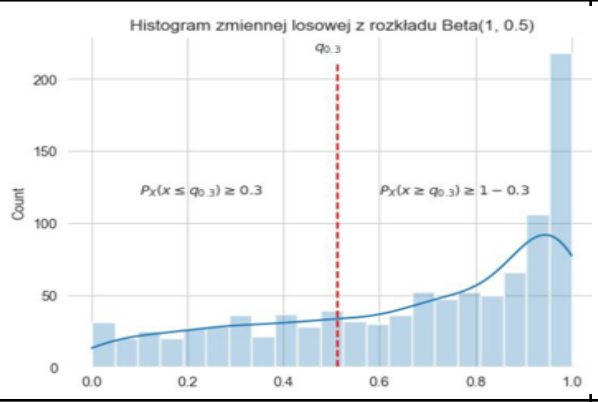
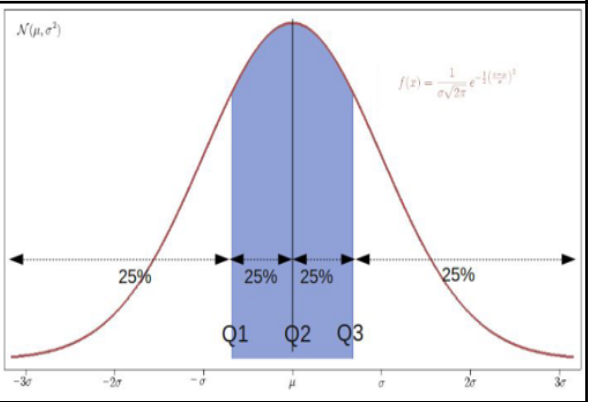
Horizon Graph	

Tabela 3. Sposoby na analizowanie wykresów.

KWANTYLE	KWARTYLE
	
Kwantylem rzędu *p*, gdzie 0 <= p <= 1, w rozkładzie empirycznym P_X zmiennej losowej X nazywamy każdą liczbę x_p, dla której spełnione są nierówności $P_X((-\infty, q_p]) \geq p$ oraz $P_X([q_p, \infty)) \geq 1 - p$	Kwantyle dzielące obserwacje na (mniej więcej) równe ćwiartki (kwarty): • Q1 - kwantyl rzędu ¼ (0.25) = dolny kwartyl • Q2 - kwantyl rzędu ½ (0.5) = mediana • Q3 - kwantyl rzędu ¾ (0.75) = górny kwartyl

7. METODY ODKRYWANIA OBSERWACJI ODSTAJĄCYCH (DO SPRAWDZENIA)

Odkrywanie obserwacji odstających to kluczowy krok w analizie danych, służący identyfikacji nietypowych lub nietypowo zachowujących się punktów w zbiorze danych. Istnieje kilka metod wykrywania obserwacji odstających:

1. Metody statystyczne: Metody te opierają się na wartościach odstających od ogólnego rozkładu danych, takie jak kryterium Tukeya, odległość Cooka, z-score, odchylenie bezwzględne

2. Metody gęstościowe: Bazują na pomiarze odległości między punktami danych. Przykłady to odległość euklidesowa, indeks odległości od najbliższych sąsiadów (KNN) lub lokalne odchylenia gęstości danych próbek względem sąsiadów np. Local outlier factor, Las izolacji

3. Metody predykcyjne: Wykorzystują algorytmy klasyfikacyjne, które identyfikują punkty danych, które nie pasują dobrze do określonych klas lub klastrów, np. Klasyfikacje jednoklasowe np. za pomocą maszyny wektorów nośnych

4. Metody oparte na klastrowaniu: Polegają na znalezieniu obserwacji, które nie zostały przydzielone do żadnego klastra i zostają uznane za obserwacje odstające.

5. Techniki uczenia maszynowego: Wykorzystują modele uczenia maszynowego, takie jak autoenkodery czy algorytmy detekcji anomalii, które uczą się reprezentacji danych i identyfikują punkty odstające.

8. WSPÓŁCZYNNIKI KORELACJI (PEARSONA, SPEARMANA)

Współczynnik korelacji Pearsona	**Współczynnik korelacji Spearmana**
Oznaczenie: ***rr***	Oznaczenie: ***ρρ***
Mierzy liniową zależność między dwiema zmiennymi ciągłymi	Mierzy ogólną zależność monotoniczną między dwiema zmiennymi, niekoniecznie liniową. Jest oparty na porządku wzajemnym rang obserwacji
Przyjmuje wartości w przedziale od -1 do 1, gdzie: ●r=1 i r=1 oznacza doskonałą dodatnią korelację liniową, ●r=−1 i r=−1 oznacza doskonałą ujemną korelację liniową, ●r=0 i r=0 oznacza brak liniowej zależności między zmiennymi	Przyjmuje wartości w przedziale od -1 do 1, podobnie jak współczynnik Pearsona
Zastosowanie: Współczynnik korelacji Pearsona jest często używany w analizie statystycznej, zwłaszcza w kontekście analizy regresji i związków liniowych. Jest wrażliwy na wartości odstające.	*Zastosowanie:* Współczynnik korelacji Spearmana jest stosowany w przypadkach, gdy zależy nam na wykryciu ogólnej zależności między zmiennymi, a niekoniecznie liniowej. Jest bardziej odporny na wartości odstające i sprawdza się w przypadku danych niemiarowych.

Oba współczynniki mają swoje zastosowanie w analizie danych w zależności od charakterystyki danych i celu badania. Wybór między nimi zależy od tego, czy zakładamy liniową czy ogólną zależność między zmiennymi, oraz od odporności na wartości odstające.

9. RODZAJE ZMIENNYCH (KATEGORYCZNE, BINARNE, CIĄGŁE)

Zmienna kategoryczna:

W analizie danych, zmienna kategoryczna to zmienna, która przyjmuje ograniczony zestaw kategorii lub grup. Te kategorie nie mają porządku, co oznacza, że nie można jednoznacznie ustalić hierarchii między nimi. Analiza zmiennych kategorycznych często obejmuje używanie wskaźników takich jak częstość występowania poszczególnych

kategorii, testy statystyczne (np. test chi-kwadrat) oraz metody wizualizacyjne, takie jak wykresy słupkowe czy diagramy kołowe.

Zmienna binarna:

Zmienna binarna w analizie danych przyjmuje tylko dwie wartości, co może być użyteczne w przypadku klasyfikacji. Analiza zmiennych binarnych obejmuje często obliczanie proporcji jednej z wartości względem drugiej, a także stosowanie miar jakości modelu klasyfikacyjnego, takich jak czułość, swoistość czy krzywa ROC.

Zmienna ciągła:

W analizie danych, zmienna ciągła to zmienna, która może przyjąć nieskończoną liczbę wartości w określonym przedziale. Analiza zmiennych ciągłych obejmuje często obliczanie miar centralnych (średniej, mediany), miar rozproszenia (wariancji, odchylenia standardowego) oraz stosowanie histogramów czy wykresów gęstości dla wizualizacji rozkładu danych.

10. BŁĘDY PODCZAS TWORZENIA WYKRESÓW

Oś Y na wykresie nie zaczyna się w zerze
Odległość pomiędzy kolejnymi punktami powinny być zachowane
wykresy powinny być prawdomówne. Współczynnik kłamstwa nie powinien być większy ani mniejszy niż 1
Wykres zawiera tylko te elementy, które są niezbędne do jego zrozumienia
Ludzki mózg lepiej odbiera zmiany danych jako zmiany jasności niż np. zmiany odcienia (najczęstszym zaburzeniem widzenia jest nierozróżnianie czerwieni i zieleni)

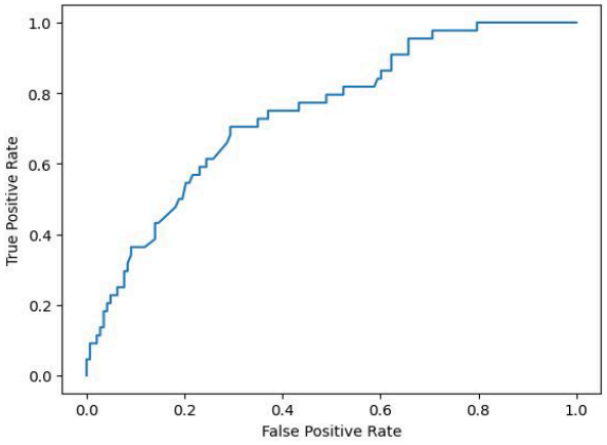
11. DANE USTRUKTURYZOWANE I NIEUSTRUKTURYZOWANE

DANE USTRUKTURYZOWANE	DANE NIEUSTRUKTURYZOWANE
● dane uporządkowane w sposób umożliwiający niezawodną identyfikację poszczególnych stwierdzeń faktów oraz ich wszystkich składników ● można je przedstawić w formie tabelarycznej, w której mogą być przechowywane w relacyjnej bazie danych ● liczby, daty i tekst ● wg Gartner stanowią ok 20% danych w przedsiębiorstwach ● wymagają niewiele przestrzeni dyskowej ● istnieją ustandaryzowane sposoby zarządzania i zabezpieczania takich danych	● nie mają wstępnie zdefiniowanego modelu danych lub nie są zorganizowane we wstępnie zdefiniowany sposób ● nie da się w łatwy (nieprzetworzony) sposób przedstawić ich w formie tabelarycznej ● obrazy, audio, video, emaile, sformatowane pliki tekstowe, arkusze kalkulacyjne ● wg Gartner stanowią aż 80% danych w przedsiębiorstwach
PRZYKŁADY	
A. ECOMMERCE	
●katalog produktów ●ceny ●dane klienta	●zachowanie klienta i wzorce wydawania pieniędzy ●zadowolenie klienta z usługi
B. SŁUŻBA ZDROWIA	
●formularze dla pacjentów ●dane o ubezpieczeniu	●zdjęcia RTG i tomografia komputerowa ●notatki z wizyt

●dane do płatności	●zalecenia lekarskie
C. BANKOWOŚĆ	
●operacje finansowe ●dane klienta	●logi z rozmów telefonicznych ●nagrania audio i video z komunikacji między klientami a bankiem

12. KRZYWA ROC

KRZYWA ROC



- Krzywa ROC powstaje poprzez zmianę progu, powyżej którego następuje pozytywna klasyfikacja.
- Krzywa ROC pokazuje kompromis pomiędzy swoistością a czułością klasyfikatora dla różnych progów funkcji decyzyjnej.
- Nie każdy klasyfikator posiada zdefiniowaną funkcję decyzyjną np. drzewo decyzyjne nie ma.
- Im większe pole pod krzywą - tym lepszy klasyfikator.

PYTANIA Z UBIEGŁEGO KOLOKWIUM (DO SPRAWDZENIA)

1. Dane zostały poklastrowane za pomocą algorytmu K-Means z k=3, jednak wyniki klastrowania Cię nie satysfakcjonują. Co można zrobić, żeby poprawić jakość klastrowania (zaproponuj kilka rozwiązań)? Jak sprawdzić, czy wyniki klastrowania poprawiły się po dokonaniu zmian?

   Jeśli wyniki klastrowania algorytmem K-Means z k=3 nie są satysfakcjonujące, istnieje kilka sposobów na ich poprawę:

   a) Zmiana liczby klastrów (k): Spróbuj zmienić liczbę klastrów i ponownie przeprowadzić klastrowanie, np. zwiększyć lub zmniejszyć liczbę klastrów, aby zobaczyć, czy nowe podziały lepiej odzwierciedlają strukturę danych.

   b) Inne algorytmy klastrowania: Wypróbuj inne algorytmy klastrowania, takie jak aglomeracyjna analiza skupień (hierarchiczna), DBSCAN czy algorytmy oparte na gęstości.

   c) Normalizacja danych: Upewnij się, że dane są odpowiednio znormalizowane, szczególnie jeśli mają różne skale lub jednostki. Czasami normalizacja może poprawić wyniki klastrowania.

d) Wybór odpowiednich cech: Może się okazać, że niektóre cechy w danych są mniej istotne lub wpływają negatywnie na klastrowanie. Przeprowadź analizę i wybierz najbardziej istotne cechy do klastrowania.

e) Inicjalizacja centroidów: Zmiana początkowego ustawienia centroidów może wpłynąć na wyniki klastrowania. Możesz spróbować różnych metod inicjalizacji, np. k-means++.

Aby sprawdzić poprawę wyników klastrowania po dokonaniu zmian, możesz zastosować:

a) Metryki ewaluacji klastrowania: Skorzystaj z metryk takich jak indeks Daviesa-Bouldina, indeks Silhouette czy wewnętrzna suma kwadratów (WCSS), aby porównać jakość klastrów przed i po zmianach.

b) Wizualizacja klastrów: Wykorzystaj wykresy, np. wykres punktowy (scatter plot), aby zobaczyć, czy nowe klastry lepiej odzwierciedlają naturalną strukturę danych niż poprzednie.

c) Analiza stabilności klastrów: Możesz również przeprowadzić analizę stabilności klastrów, gdzie sprawdzisz, czy klastry są stabilne w różnych próbach losowych lub dla różnych podziałów danych.

Kombinacja tych metod pozwoli ci ocenić poprawę jakości klastrowania po dokonaniu zmian i wybrać najlepsze podejście dla swoich danych.

2. Pewien sklep internetowy zebrał dane dotyczące swoich klientów. Dane zawierają informacje takie jak płeć, wiek, roczny przychód oraz syntetycznie wyznaczoną liczbę punktów (od 0 do 100) odzwierciedlającą poziom wydatków każdego klienta w tym sklepie. Twoim zadaniem jest dokonanie predykcji wydatków(wyrażonych jako punkty) dla nowych klientów sklepu na podstawie płci, wieku i rocznego przychodu. Zdefiniuj ten problem w języku metod ekstrakcji danych, zaproponuj algorytm(y), z których warto skorzystać jako pierwszych i uzasadnij swój wybór. W jaki sposób wybrał(a)byś algorytm, który najlepiej rozwiązuje postawione zadanie?

W celu rozwiązania tego problemu można zastosować kilka algorytmów, ale pierwszymi wybieranymi mogą być:

a) Regresja liniowa: Jest to prosty, ale skuteczny sposób przewidywania wartości numerycznych na podstawie liniowego modelu związku między zmiennymi niezależnymi a zmienną zależną.

b) Drzewa decyzyjne lub Las Losowy: Drzewa decyzyjne mogą dobrze radzić sobie z przewidywaniem wartości numerycznych. Las losowy, będący zbiorem drzew decyzyjnych, może zwiększyć dokładność predykcji.

Wybór najlepszego algorytmu można dokonać przez:

- Ewaluację modeli: Przetestowanie różnych algorytmów na zbiorze walidacyjnym i porównanie ich wyników za pomocą metryk takich jak błąd średniokwadratowy (RMSE), średni błąd bezwzględny (MAE) czy współczynnik determinacji (R^2).

- Kros-walidację: Zastosowanie kroswalidacji, która pomoże w ocenie wydajności modelu dla różnych podziałów danych i uniknięciu nadmiernego dopasowania.

- Tuning hiperparametrów: Dostrojenie parametrów modelu dla każdego algorytmu może znacznie poprawić jego wydajność.

- Złożoność modelu: Zwrócenie uwagi na złożoność modelu - zbyt prosty model może nie być w stanie uchwycić złożonych wzorców, podczas gdy zbyt skomplikowany może dopasować się do szumu.

Na podstawie wyników testów wybrałbym algorytm, który osiąga najlepsze wyniki pod względem metryk ewaluacyjnych na zbiorze walidacyjnym. Ważne jest również zrozumienie, jakie założenia stoją za wybranymi algorytmami i ich odpowiedniość do struktury danych oraz problemu przewidywania wydatków na podstawie cech klientów.

3. Opisz na czym polega proces odkrywania reguł asocjacyjnych. Wytłumacz pojęcia “zbiór częsty” oraz “silna reguła asocjacyjna”. Zaprezentuj te pojęcia na przykładzie małej bazy transakcji

Proces odkrywania reguł asocjacyjnych jest techniką analizy danych, która pozwala odkrywać relacje pomiędzy zmiennymi w zbiorze danych, zwłaszcza w przypadku danych transakcyjnych. Najbardziej znany przykład to analiza koszyka zakupowego, gdzie staramy się zrozumieć, które produkty są często kupowane razem.

a) Zbiór częsty: Zbiór częsty to zestaw elementów (takich jak produkty) występujących razem w co najmniej z góry określonej liczbie transakcji. Na przykład, jeśli zestaw produktów A i B pojawia się w co najmniej 50% wszystkich transakcji, to jest to zbiór częsty.

b) Silna reguła asocjacyjna: Jest to zależność między zbiorami elementów, która wyraża się w postaci "jeśli X, to Y", gdzie X i Y są zbiorami produktów lub zmiennych, a reguła jest uznawana za silną, jeśli ma wysoką miarę wsparcia (support) i ufności (confidence).

    - Wsparcie (support): Określa częstość występowania zbioru elementów wśród wszystkich transakcji. Im wyższe wsparcie, tym częstszy jest zbiór.
    - Ufność (confidence): Oznacza prawdopodobieństwo, że reguła jest poprawna. Im wyższa ufność, tym większe prawdopodobieństwo, że jeśli klient kupił X, to kupi również Y.

Na przykład, zakładając prostą bazę transakcji:

Numer transakcji	Produkty
Transakcja 1	chleb, mleko, jajka

Transakcja 2	chleb, masło
Transakcja 3	chleb, mleko, masło
Transakcja 4	mleko, jajka
Transakcja 5	masło, jajka

Załóżmy, że chcemy znaleźć silne reguły asocjacyjne. Przyjmijmy minimalne wsparcie na poziomie 40% i minimalną ufność na poziomie 80%.

Zbiory częste:

- {chleb, mleko} (wsparcie: 40%, występuje w transakcjach 1, 3)
- {chleb, masło} (wsparcie: 40%, występuje w transakcjach 2, 3)
- {mleko, jajka} (wsparcie: 40%, występuje w transakcjach 1, 4)
- {masło, jajka} (wsparcie: 40%, występuje w transakcjach 3, 5)

Silne reguły asocjacyjne:

- {chleb} -> {mleko} (wsparcie: 40%, ufność: 66.7%, występuje w transakcjach 1, 3)
- {mleko} -> {chleb} (wsparcie: 40%, ufność: 66.7%, występuje w transakcjach 1, 3)

4. Załóżmy, że chcesz zbadać czynniki, które potencjalnie wpływają na gotowanie ryżu. Czego użył(a)byś jako zmiennej odpowiedzi w tym eksperymencie? Jak zmierzył(a)byś tę odpowiedź? Wymień wszystkie potencjalne źródła zmienności, które mogą mieć wpływ na odpowiedź

Jako zmienną odpowiedzi w badaniu dotyczącym gotowania ryżu można użyć stopnia ugotowania ryżu, czyli stopnia jego miękkości lub twardości. Można zmierzyć tę odpowiedź na różne sposoby:

1. Subiektywna ocena: Poprzez ocenę osób degustujących ryż, którzy oceniają stopień miękkości lub twardości na podstawie ich doświadczenia i odczuć.

2. Pomiar fizyczny: Za pomocą przyrządu, który mierzy stopień ugotowania ryżu, na przykład penetrometrem, który określa stopień jego miękkości poprzez wbicie igły w ziarna.

Potencjalne źródła zmienności, które mogą mieć wpływ na stopień ugotowania ryżu, to:

a) Czas gotowania: Czas gotowania ryżu może znacząco wpłynąć na jego stopień ugotowania. Dłuższe gotowanie może doprowadzić do bardziej miękkiego ryżu.

b) Ilość wody: Proporcja wody do ryżu może mieć wpływ na stopień miękkości. Więcej wody może prowadzić do bardziej miękkiego ryżu, podczas gdy mniejsza ilość wody może sprawić, że ryż będzie twardszy.

c) Rodzaj ryżu: Różne rodzaje ryżu (np. długoziarnisty, krótkoziarnisty) mogą wymagać różnego czasu gotowania i mają różną tendencję do miękkości.

d) Temperatura gotowania: Temperatura podczas gotowania ryżu może wpłynąć na jego stopień ugotowania. Gotowanie w niższej temperaturze może wydłużyć czas gotowania.

e) Metoda gotowania: Gotowanie ryżu na gazie, na parze, w kuchence mikrofalowej czy w specjalnych urządzeniach (np. ryżowarach) może wpłynąć na stopień miękkości.

f) Przechowywanie ryżu: Ryż przechowywany w różnych warunkach (wilgotność, temperatura) może zmienić jego gotowanie i stopień miękkości.

Analiza tych czynników może pomóc zrozumieć, jak różne warunki gotowania wpływają na stopień ugotowania ryżu i jak optymalizować proces gotowania, aby uzyskać pożądany efekt końcowy.

5. Wyobraź sobie, że masz do czynienia z danymi tekstowymi. Aby reprezentować słowa, wykorzystujesz word embeddings, w wyniku czego każde słowo kodowane jest jako 1000-wymiarowy wektor a słowa bliskoznaczne znajdują się blisko siebie w tej 1000-wymiarowej przestrzeni. Chcesz zredukować wymiarowość tych wielowymiarowych danych, tak aby podobne słowa znajdowały się blisko siebie również po redukcji. W takim przypadku, który z poniższych algorytmów najprawdopodobniej wybierzesz? Odpowiedź uzasadnij.

   a. Analiza składowych głównych (PCA): Ta metoda zmniejsza liczbę zmiennych w zbiorze danych, zachowując jednocześnie jak najwięcej informacji (poprzez maksymalizację wariancji).
   b. Skalowanie wielowymiarowe (MDS): Ta metoda ma na celu zachowanie odległości pomiędzy parami punktów w oryginalnych danych w rozmaitości o niższej wymiarowości.
   c. Isomap: Ta metoda wykorzystuje kombinację MDS i teorii grafów, aby zachować odległości geodezyjne między punktami w oryginalnych danych.
   d. Local Linear Embedding (LLE): Ta metoda ma na celu zachowanie lokalnych relacji między punktami w oryginalnych danych poprzez skonstruowanie ważonego grafu najbliższych sąsiadów każdego punktu.
   e. t-SNE: Metoda wykorzystuje model probabilistyczny do zachowania lokalnych relacji między punktami w oryginalnych danych.

W przypadku reprezentacji słów za pomocą word embeddings i chęci redukcji wymiarowości, aby zachować podobieństwa między słowami nawet po redukcji, najprawdopodobniej wybrałbym t-SNE (e. t-SNE - t-distributed Stochastic Neighbor Embedding).

t-SNE jest szczególnie skuteczne w zachowaniu lokalnych relacji między punktami w wysokowymiarowej przestrzeni, co jest kluczowe w przypadku reprezentacji słów za pomocą word embeddings. Chociaż niekoniecznie zachowuje odległości globalne, doskonale radzi sobie z zachowaniem podobieństw między punktami w oryginalnych danych w przestrzeni o niższej wymiarowości.

Choć PCA (a. Analiza składowych głównych) również redukuje wymiarowość danych, zachowuje ona głównie globalne zależności, co może prowadzić do utraty informacji o lokalnych relacjach między słowami. Skalowanie wielowymiarowe (MDS), Isomap i LLE również starają się zachować odległości lub relacje między punktami, ale t-SNE jest często bardziej skuteczne w zachowywaniu lokalnych struktur danych, co jest istotne w przypadku reprezentacji słów.

W przypadku analizy danych tekstowych i reprezentacji słów za pomocą word embeddings, t-SNE jest często wybieranym algorytmem do redukcji wymiarowości ze względu na jego zdolność do zachowania podobieństw między punktami, co jest kluczowe dla znaczenia semantycznego słów.

6. Wytłumacz na czym polega walidacja krzyżowa.

Walidacja krzyżowa to technika oceny wydajności modelu, która pozwala na oszacowanie jego zdolności do generalizacji na nieznanym zbiorze danych. Jest to ważny krok w procesie uczenia maszynowego, mający na celu uniknięcie nadmiernego dopasowania modelu do konkretnego zbioru danych (overfitting) lub zbyt optymistycznego oszacowania jego skuteczności.

Proces walidacji krzyżowej wygląda następująco:

1. Podział danych: Zbiór danych dzieli się na kilka równych części (foldów). Standardowo stosuje się walidację krzyżową k-fold, gdzie zbiór danych dzieli się na k równych części.

2. Iteracyjne trenowanie i testowanie: Model jest trenowany k razy, każdorazowo na k-1 foldach danych, a następnie testowany na pozostałym jednym foldzie.

3. Ocena wydajności: Dla każdej iteracji model jest testowany na foldzie, na którym nie był uczony, a następnie oceniana jest jego skuteczność, na przykład za pomocą średniej błędów lub innych metryk oceny modelu.

4. Uśrednienie wyników: Ostateczna skuteczność modelu jest obliczana poprzez uśrednienie wyników uzyskanych z poszczególnych iteracji walidacji krzyżowej.

7. Wytłumacz czym jest imputacja danych. W jaki sposób można wykorzystać algorytm k najbliższych sąsiadów do imputacji danych?

Imputacja danych to proces uzupełniania brakujących lub uszkodzonych danych, wykorzystując informacje dostępne w pozostałych częściach zbioru danych. Ma na celu zastąpienie brakujących wartości wiarygodnymi lub przewidywanymi danymi, co umożliwia dalszą analizę lub stosowanie modeli uczenia maszynowego.

Algorytm k najbliższych sąsiadów (k-NN) może być wykorzystany do imputacji danych poprzez:

1. Wybór sąsiadów: Dla każdej brakującej wartości wybierane są k najbliższe obserwacje (sąsiedzi) z pełnego zbioru danych, które posiadają wartości dla pozostałych zmiennych.

2. Uśrednianie lub wagowanie: Uzyskane wartości od sąsiadów są uśredniane lub ważone na podstawie odległości od brakującej wartości. Może to obejmować średnią wartość lub inną funkcję agregującą.

3. Metoda ważonej imputacji: W przypadku wagowanej imputacji, sąsiadów można ważyć na podstawie odległości od brakującej wartości, przy czym bliżsi sąsiedzi mają większy wpływ na imputowaną wartość.

4. Uwzględnienie różnych metryk odległości: Można stosować różne metryki odległości, takie jak odległość euklidesowa, manhattan czy kosinusowa, aby określić podobieństwo między obserwacjami.

5. Dobór liczby sąsiadów (k): Wybór liczby sąsiadów ma wpływ na dokładność imputacji. Większa liczba sąsiadów może prowadzić do bardziej stabilnej imputacji, ale może być również bardziej podatna na szumy.

6. Algorytm k-NN jest stosunkowo prosty i elastyczny, dzięki czemu można go zastosować do różnych typów danych i problemów imputacji. Jednak jego skuteczność może być uzależniona od struktury danych, liczby brakujących wartości oraz doboru odpowiednich parametrów, takich jak liczba sąsiadów czy metryka odległości.