4. System rekomendacji:

Stwórz system rekomendacji oparty na współpracy lub na treści, wykorzystując techniki uczenia maszynowego.

Projekt dotyczący stworzenia **systemu rekomendacji opartego na współpracy lub na treści** z wykorzystaniem technik uczenia maszynowego jest bardzo praktycznym i popularnym zagadnieniem w dziedzinie analizy danych, e-commerce, platform streamingowych oraz mediów społecznościowych. Systemy rekomendacyjne są stosowane do przewidywania preferencji użytkowników na podstawie ich wcześniejszych działań lub na podstawie cech przedmiotów, które ich interesują. Projekt zakłada wybór podejścia **– filtrację opartą na współpracy** lub **filtrację opartą na treści –** a następnie implementację modelu, który będzie w stanie sugerować użytkownikom produkty, filmy, książki, artykuły itp.

Celem projektu jest:

- Zbudowanie systemu rekomendacyjnego opartego na współpracy (Collaborative Filtering) lub na analizie treści (Content-Based Filtering) przy użyciu technik uczenia maszynowego.
- Przetestowanie systemu na rzeczywistych danych, takich jak zestawy filmów (np. **MovieLens**), produktów (np. **Amazon**), książek lub muzyki.
- Analiza efektywności modelu rekomendacyjnego oraz porównanie różnych podejść i algorytmów rekomendacji.

Opis problemu

Systemy rekomendacyjne mają na celu przewidywanie, jakie produkty (filmy, książki, artykuły, itd.) mogą zainteresować użytkownika na podstawie jego zachowań, historii interakcji lub podobieństwa do innych użytkowników. Istnieją dwa główne podejścia do rekomendacji:

- **Filtracja oparta na współpracy (Collaborative Filtering) –** rekomendacje są generowane na podstawie zachowań i preferencji innych użytkowników, którzy mają podobne gusta. Działa to na zasadzie „użytkownicy tacy jak Ty lubili także…”.
- **Filtracja oparta na treści (Content-Based Filtering) –** rekomendacje są oparte na cechach przedmiotów, które użytkownik wcześniej polubił. System analizuje atrybuty produktów (np. gatunki filmów, tematy książek, cechy techniczne produktów) i sugeruje podobne przedmioty.

Etapy realizacji projektu

Wybór podejścia rekomendacyjnego

i. Filtracja oparta na współpracy

- **Collaborative Filtering** opiera się na wykorzystaniu danych o interakcjach użytkowników z produktami, gdzie ważne są oceny (ratingi), zakupy, oglądane filmy, odtwarzane utwory, kliknięcia w produkty itp.
- **Podejścia do filtracji opartej na współpracy**:

- **Collaborative Filtering z wykorzystaniem sąsiedztwa (User-based i Item-based)**:
  - **User-based** analizuje podobieństwo użytkowników do siebie i sugeruje produkty, które preferują podobni użytkownicy.
  - **Item-based** skupia się na analizie podobieństwa między przedmiotami, bazując na preferencjach użytkowników. Jeśli użytkownik ocenił film X, który jest podobny do filmu Y, to system zasugeruje film Y.
- **Matrix Factorization (Funkcje ukryte) –** podejście to polega na zastosowaniu metod takich jak **SVD (Singular Value Decomposition)** lub **ALS (Alternating Least Squares)**, które dekomponują macierz użytkowników i przedmiotów, aby znaleźć ukryte wzorce, co pozwala przewidywać nowe interakcje.

ii. Filtracja oparta na treści

- W **Content-Based Filtering** system analizuje cechy przedmiotów, które użytkownik wcześniej oceniał lub oglądał. Rekomendacje są oparte na podobieństwie cech przedmiotów do tych, które użytkownik preferuje.
- **Podejścia do filtracji opartej na treści**:
  - **Modele TF-IDF –** popularne w analizie tekstu, stosowane do oceny ważności słów w opisie przedmiotu (np. książki, filmu).
  - **Modele wektorów osadzania (word embeddings) –** np. **Word2Vec**, które mogą uchwycić bardziej subtelne relacje między treściami, takimi jak podobieństwo kontekstowe między różnymi produktami.
  - **Klasyfikacja cech –** techniki takie jak regresja logistyczna, **drzewa decyzyjne**, **SVM (Support Vector Machine)** czy **Random Forest** mogą być używane do klasyfikowania i sugerowania przedmiotów na podstawie cech.

Zbieranie i przetwarzanie danych

Należy wybrać odpowiedni zbiór danych, na którym system będzie testowany. Przykładowe źródła danych:

- **MovieLens –** popularny zbiór danych o filmach i ocenach użytkowników,
- **Goodreads –** dane o książkach i preferencjach czytelników,
- **Amazon Product Data –** dane o produktach i ich ocenach,
- **Last.fm –** dane o preferencjach muzycznych użytkowników.

Dane te mogą zawierać:

- **Oceny użytkowników** (ratingi),
- **Interakcje** (kliknięcia, zakupy, obejrzenia),
- **Atrybuty przedmiotów** (gatunki, cechy, opisy).

Przetwarzanie danych - Dane należy odpowiednio przetworzyć, co może obejmować:

- **Czyszczenie danych**: usunięcie brakujących lub nieprawidłowych danych,
- **Normalizację ocen**: jeśli różni użytkownicy mają różne skale ocen, warto znormalizować te dane,
- **Feature engineering**: dla filtracji opartej na treści należy wygenerować reprezentacje cech przedmiotów, np. za pomocą TF-IDF lub wektorów osadzania.

Implementacja algorytmu rekomendacji

W zależności od wybranego podejścia, należy zaimplementować odpowiedni algorytm:

i. Collaborative Filtering (Filtracja oparta na współpracy)

- **User-based Collaborative Filtering**: Algorytm ten opiera się na podobieństwie między użytkownikami. Można wykorzystać **miarę podobieństwa kosinusowego** do porównania ocen użytkowników.
- **Item-based Collaborative Filtering**: Algorytm oparty na podobieństwie przedmiotów. Można użyć podobnych technik jak w przypadku podobieństwa użytkowników, ale w odniesieniu do przedmiotów.
- **Matrix Factorization**: Wykorzystanie **SVD** lub **ALS** do rozbicia macierzy ocen na mniejsze komponenty, co pozwala na przewidywanie brakujących ocen.

ii. Content-Based Filtering (Filtracja oparta na treści)

- **Wektorowe reprezentacje**: Na przykład, dla rekomendacji książek można wygenerować wektory reprezentujące opisy książek, a następnie sugerować książki o podobnych wektorach do tych, które użytkownik wcześniej ocenił wysoko.
- **Modele klasyfikacyjne**: Użycie modeli klasyfikacyjnych (np. **Random Forest**, **SVM**, **k-NN**) do przewidywania, czy użytkownik polubi dany przedmiot na podstawie jego cech.

Ocena jakości systemu rekomendacji

Należy zastosować odpowiednie miary do oceny jakości systemu rekomendacyjnego:

- **Precision** i **Recall**: Do oceny, na ile rekomendacje były trafne i kompletne.
- **Mean Absolute Error (MAE)** i **Root Mean Square Error (RMSE)**: W przypadku przewidywania ocen.
- **ROC i AUC**: Dla klasyfikacyjnych podejść, aby mierzyć jakość klasyfikacji binarnej.

Testy należy przeprowadzić na zestawie testowym, który nie był używany do trenowania modelu, aby sprawdzić zdolność modelu do generalizacji.

Eksperymenty i tuning - Warto przeprowadzić eksperymenty, zmieniając parametry modelu oraz sprawdzając wpływ różnych strategii na jakość rekomendacji. Na przykład:

- Dla Collaborative Filtering można porównać różne miary podobieństwa (np. kosinusowa, Pearsona).
- Dla Content-Based Filtering można eksperymentować z różnymi technikami przetwarzania tekstu (np. TF-IDF, Word2Vec).

Rozszerzenia projektu - System rekomendacji hybrydowej

1. Dodanie rekomendacji hybrydowej

Zaproponowane rozszerzenie to implementacja **systemu rekomendacji hybrydowej**, który łączy zalety filtracji opartej na współpracy i filtracji opartej na treści. W zależności od dostępnych danych i kontekstu użytkownika, system może:

- Początkowo korzystać z filtracji opartej na treści, a następnie przechodzić do filtracji opartej na współpracy, gdy dostępnych będzie więcej danych o użytkowniku.
- Połączyć wyniki z obu metod, aby uzyskać bardziej precyzyjne rekomendacje.

Przykład podejścia hybrydowego:

- Można wykorzystać model macierzy czynników (Matrix Factorization) w filtracji opartej na współpracy oraz wektory osadzania (word embeddings) w filtracji opartej na treści. Następnie wyniki obu metod można zintegrować poprzez algorytm wagowy lub metodę boostingową.