**Uniwersytet Rzeszowski**

**Wydział Nauk Ścisłych i Technicznych**

**Instytut Informatyki**

**Dawid Olko**

**Nr albumu studenta do125148**

***Aplikacja GUI wykorzystująca elementy eksploracji danych***

**Praca projektowa z Hurtowni danych**

Prowadzący: mgr. Szczur

Rzeszów r. a. 2024/2025

**Dlaczego baza „Adult” jest dobra do tego zadania?**

1. **Różnorodność typów danych:**
   * **Dane numeryczne:** Atrybuty takie jak age, education-num, capital-gain, capital-loss oraz hours-per-week umożliwiają obliczenie miar statystycznych (min, max, mediana, odchylenie standardowe) oraz wykonanie operacji skalowania i standaryzacji. Dzięki temu możesz łatwo zademonstrować przetwarzanie danych ciągłych.
   * **Dane kategoryczne:** Kolumny takie jak workclass, education, marital-status, occupation, relationship, race, sex, native-country dostarczają możliwości przeprowadzenia transformacji, usuwania niepotrzebnych kolumn lub zamiany wartości, a także generowania wykresów słupkowych czy analiz grupowych.
2. **Problem brakujących danych:**
   * W zbiorze występują brakujące lub nieprawidłowe wartości, reprezentowane najczęściej jako „?”. To stanowi idealny przypadek do zaimplementowania strategii radzenia sobie z brakami – zarówno poprzez usuwanie rekordów, jak i uzupełnianie danych (np. średnią, medianą lub stałą wartością).
3. **Zadania ekstrakcji i modyfikacji danych:**
   * Możliwość wyboru określonych zakresów wierszy i kolumn (np. wyciągnięcie podtablicy dla określonej grupy wiekowej czy konkretnego zawodu) umożliwia przetestowanie funkcji ekstrakcji danych.
   * Dane kategoryczne i numeryczne umożliwiają ćwiczenie operacji takich jak usuwanie kolumn, zmiana wartości (np. korygowanie „?” na bardziej użyteczne dane lub przypisywanie etykiet).
4. **Wykonanie transformacji i skalowania:**
   * Zestaw zmiennych numerycznych, często z bardzo zróżnicowanymi zakresami (np. capital-gain z przewagą zer oraz sporadyczne wysokie wartości), wymaga zastosowania metod normalizacji i standaryzacji, co jest kluczowym elementem zadania.
5. **Elementy wizualizacyjne:**
   * Różnorodność danych (zarówno ciągłych, jak i kategorycznych) umożliwia tworzenie różnych typów wykresów, takich jak histogramy rozkładu, wykresy słupkowe dla zmiennych kategorycznych oraz wykresy rozproszenia dla zależności między zmiennymi.
6. **Zadania uczenia maszynowego:**
   * Baza ta jest klasycznym przykładem problemu klasyfikacji, gdzie celem jest przewidzenie, czy dana osoba zarabia więcej niż 50 tys. dolarów rocznie. To pozwala na zastosowanie różnych algorytmów klasyfikacyjnych (np. drzewa decyzyjne, k-NN, SVM) oraz algorytmów grupowania (np. k-means) do eksploracji danych i odkrywania ukrytych struktur.
   * Możliwość zastosowania reguł asocjacyjnych (np. algorytm Apriori) pozwala dodatkowo na analizę powiązań między różnymi atrybutami.

**Problematyka bazy „Adult” i powiązanie z wymaganiami projektu**

1. **Odczyt danych z pliku CSV:**
   * Baza „Adult” jest dostępna w formacie CSV, co sprawia, że operacje wczytywania danych za pomocą bibliotek takich jak pandas są proste i intuicyjne.
2. **Obliczanie miar statystycznych:**
   * Dzięki obecności danych numerycznych (np. age, capital-gain, hours-per-week) jest możliwość przeprowadzenia pełnej analizy statystycznej, pokazując liczbowe podsumowanie oraz rozkłady danych.
3. **Korelacja cech:**
   * Zestaw zmiennych numerycznych pozwala obliczyć korelacje między cechami, co może być wizualizowane przy pomocy macierzy korelacji czy wykresów, umożliwiając głębszą analizę zależności w danych.
4. **Ekstrakcja podtablic:**
   * Możesz wyodrębnić podzbiory danych na podstawie określonych kryteriów, np. analizować tylko osoby z określonego przedziału wiekowego czy z konkretnego typu zatrudnienia. Umożliwia to bardziej szczegółowe badanie wybranych fragmentów zbioru.
5. **Usuwanie i zamiana kolumn:**
   * Niektóre kolumny, jak fnlwgt, mogą być mniej przydatne w analizie i warto je usunąć lub zastąpić przetworzonymi danymi. Również kolumny zawierające wartości „?” wymagają interwencji – zastąpienia lub usunięcia.
6. **Skalowanie i standaryzacja:**
   * Dane, takie jak capital-gain i capital-loss, często wykazują duże rozbieżności, co czyni operacje skalowania i standaryzacji niezbędnymi przed zastosowaniem metod uczenia maszynowego.
7. **Obsługa brakujących danych:**
   * Obecność braków danych (np. „?” w kolumnie workclass lub occupation) zmusza do wdrożenia strategii czyszczenia danych, co jest kluczowym elementem projektu.
8. **Dodawanie nowych kolumn:**
   * Na podstawie istniejących atrybutów możesz tworzyć nowe kolumny (np. grupy wiekowe, przekształcone etykiety edukacyjne), co stanowi ćwiczenie z zakresu feature engineeringu i dodatkowo wzbogaca analizę.
9. **Wizualizacja danych:**
   * Możliwość generowania wykresów (histogramy, wykresy słupkowe, wykresy rozproszenia) pozwala na szybką i czytelną prezentację wyników analizy, co jest szczególnie przydatne w interaktywnym GUI.
10. **Klasyfikacja i grupowanie:**
    * Problematyka klasyfikacji (czy osoba zarabia >50K) oraz możliwość zastosowania algorytmów klastrowania sprawiają, że baza „Adult” jest idealna do demonstracji metod uczenia maszynowego i eksploracji danych.

**Podsumowanie**

Baza „Adult” to kompletny zestaw danych, który:

* Dostarcza zarówno danych ciągłych, jak i kategorycznych,
* Wprowadza wyzwania związane z brakującymi danymi i różnymi skalami wartości,
* Umożliwia implementację szerokiego zakresu operacji przetwarzania danych, od podstawowych obliczeń statystycznych po zaawansowane techniki transformacji i klasyfikacji.

Dzięki tym cechom, zbiór ten idealnie wpisuje się w wymagania projektu, pozwalając na wszechstronne ćwiczenie technik ETL oraz eksploracji danych przy jednoczesnym zastosowaniu interfejsu graficznego. Realizacja projektu na bazie tego zbioru nie tylko pozwoli Ci na praktyczne zastosowanie zdobytej wiedzy, ale również umożliwi dogłębną analizę problematyki przetwarzania danych w realnych scenariuszach.