


Informatyka Stosowana		
Laboratorium 4	Problemy przy pracy na prawdziwych danych – część 2	 POLITECHNIKA BYDGOSKA Wydział Telekomunikacji, Informatyki i Elektrotechniki
Przedmiot	Eksploracyjna analiza danych	
Prowadzący	mgr inż. Gracjan Kątek	

1. Wprowadzenie

Praca z prawdziwymi danymi jest jednym z kluczowych wyzwań w analizie danych. Dane rzadko występują w idealnym stanie, dlatego ich wstępne przygotowanie i przekształcenie jest kluczowe przed przystąpieniem do bardziej zaawansowanych analiz. To laboratorium skupia się na podstawowych technikach czyszczenia i manipulowania danymi, które są niezbędne w codziennej pracy.

A. Manipulacja nazwami kolumn

Czasami zbiory danych mają niespójne lub trudne do odczytania nazwy kolumn, np. z wielkimi literami, spacjami lub innymi znakami specjalnymi.

Ujednolicanie nazw kolumn:

- Ułatwia czytelność kodu.
- Redukuje błędy wynikające z literówek lub niespójności nazw.

B. Konwersja typu danych

W zbiorach danych często spotykamy zmienne przechowywane w niewłaściwym typie, np.:

- Liczby przechowywane jako ciągi znaków (object).
- Daty zapisane jako teksty.

Takie dane wymagają konwersji, aby umożliwić ich dalsze wykorzystanie w analizie lub modelowaniu. Konwersja typów pozwala:

- Wykonywać operacje matematyczne na liczbach.
- Porównywać wartości w danych.
- Optymalizować wydajność obliczeń.

C. Mapowanie wartości w kolumnach

W wielu przypadkach kategorie w danych są zapisane w formacie tekstowym, co może być problematyczne dla algorytmów uczenia maszynowego, które lepiej radzą sobie z wartościami liczbowymi. Mapowanie wartości polega na przypisaniu każdej kategorii odpowiedniej liczby.

Zalety mapowania:

- Upraszcza dalsze analizy.
- Przygotowuje dane do modelowania statystycznego lub uczenia maszynowego.

D. Znaczenie technik w praktyce

Te podstawowe operacje są pierwszym krokiem w każdej analizie danych. Bez właściwego przygotowania i czyszczenia danych, dalsze analizy, wizualizacje czy modelowanie mogą być błędne lub mało efektywne.

Opanowanie takich technik pozwala na:

- Spójne i logiczne przygotowanie danych.
- Automatyzację procesów czyszczenia i standaryzacji.
- Usprawnienie pracy z dużymi zbiorami danych.

2. Zadania do samodzielnego wykonania

Zadanie 1: Zastępowanie nazw kolumn

Wczytaj dane z dołączonego pliku, a następnie zastąp wszystkie duże litery w nazwach kolumn małymi. Wyświetl pierwsze pięć wierszy za pomocą metody `pd.DataFrame.head()` i funkcji `print()`.

Zadanie 2: Konwersja typu danych

Wczytaj dane z dołączonego pliku, a następnie spójrz na rozkład zmiennej `engine`. Zauważ, że zmienna `engine` jest typu `object`. Usuń ostatnie 3 znaki z każdego elementu tej zmiennej i przekonwertuj na typ `int`. Wydrukuj pierwsze pięć wierszy za pomocą metody `pd.DataFrame.head()` i funkcji `print()` z kolumn: `name` oraz `engine`.

Zadanie 3: Mapowanie wartości

Wczytaj dane z dołączonego pliku, a następnie sprawdź rozkład zmiennej `transmission` i wykonaj następujące mapowanie tej zmiennej:

- `Manual` -> `0`
- `Automatic` -> `1`

Wydrukuj pierwsze pięć wierszy zmiennej `transmission` za pomocą metody `pd.DataFrame.head()` i funkcji `print()`.

Zadanie 4: Mapowanie wartości z wykorzystaniem Label Encoding i One Hot Encoding z biblioteki scikit-learn

Wczytaj dane z dołączonego pliku, a następnie sprawdź rozkład zmiennej `transmission` i wykonaj następujące mapowanie tej zmiennej z wykorzystaniem Label Encoding i One Hot

Encoding z biblioteki scikit-learn. Wydrukuj pierwsze pięć wierszy zmiennej `transmission` za pomocą metody `pd.DataFrame.head()` i funkcji `print()` dla każdego z mapowań oraz opisz, jakie są w nich różnice.