**Instytut Informatyki  
Wydział Nauk Ścisłych i Technicznych  
Uniwersytet Rzeszowski**

**Przedmiot:**

**Hurtownie danych**

**Dokumentacja projektu:**

***DataFusion***

**Wykonali: Dawid Olko i Piotr Smoła**

**Prowadzący: mgr inż. Adam Szczur**

**Rzeszów 2025**

Spis treści

[1. Temat i cel projektu 3](#_Toc196247571)

[2. Techniczne aspekty projektu 4](#_Toc196247572)

[2.1. Funkcjonalności aplikacji 4](#_Toc196247573)

[2.2. Wykorzystane technologie 5](#_Toc196247574)

[2.3. Projekt GUI 5](#_Toc196247575)

[3. Wygląd i użytkowanie aplikacji 5](#_Toc196247576)

[3.1. Wymagania do uruchomienia aplikacji 5](#_Toc196247577)

[3.2. Obsługa aplikacji 5](#_Toc196247578)

[3.2.1. Wczytanie zbioru danych 5](#_Toc196247579)

[3.2.2. Funkcjonalność 2 5](#_Toc196247580)

[3.2.3. Funkcjonalność 3 5](#_Toc196247581)

[4. Eksperymenty na danych 6](#_Toc196247582)

[4.1. Wykorzystane zbiory danych 6](#_Toc196247583)

[4.2. Przebieg eksperymentów i wyniki 6](#_Toc196247584)

[4.3. Analiza uzyskanych wyników i wnioski 7](#_Toc196247585)

[5. Literatura 7](#_Toc196247586)

# Temat i cel projektu

**Temat projektu:**

Projekt dotyczy stworzenia aplikacji opartej na graficznym interfejsie użytkownika (GUI), która ma na celu wspomaganie procesów analizy i obróbki danych. Program wykorzystuje dane z dwóch publicznych zbiorów: **UCI Adult Dataset** (dane o zarobkach) oraz **Chronic Kidney Disease Dataset** (dane o chorobach nerek). Aplikacja umożliwia użytkownikowi wykonanie różnych operacji na danych, takich jak obliczanie miar statystycznych, przekształcanie danych (np. kodowanie zmiennych kategorycznych, standaryzacja), a także umożliwia usuwanie, zamienianie i filtrowanie danych.

**Cel projektu:**

Celem projektu jest zaprojektowanie oraz wykonanie aplikacji do analizy danych, która pozwala na:

* **Ładowanie danych** z wybranych zestawów danych (UCI Adult Dataset oraz Chronic Kidney Disease Dataset).
* **Obliczanie miar statystycznych** (średnia, mediana, odchylenie standardowe, korelacja itp.) zarówno dla danych numerycznych, jak i kategorycznych.
* **Podstawowe operacje na danych**:
  + Zastępowanie wartości w kolumnach danych (np. zmiana wartości w kolumnie z <=50K na >50K).
  + Usuwanie lub zachowanie wybranych wierszy oraz kolumn.
  + Wykonywanie operacji standaryzacji (np. za pomocą StandardScaler, MinMaxScaler).
* **Reprezentacja danych** w postaci wykresów, takich jak wykresy słupkowe, punktowe, kołowe, oraz mapa cieplna.
* **Obsługa danych brakujących**: umożliwienie ich wypełniania średnią, medianą lub usuwania.
* **Wykonywanie analiz predykcyjnych** i klastrowania (Logistic Regression, K-Means).

# Techniczne aspekty projektu

## Funkcjonalności aplikacji

Poniżej znajdują się wszystkie wymagania (wariant dla projektu 2‑osobowego) wraz z krótkim opisem sposobu, w jaki zostanie zrealizowany w aplikacji **DataFusion**.

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Funkcjonalność** | **Status** | **Uszczegółowienie implementacji** |
| 1 | **Odczyt danych z CSV** | ✅ | load\_dataset() obsługuje *Adult* i *Chronic Kidney Disease*. |
| 2 | **Miary statystyczne (numeryczne i kategoryczne)** | ✅ | Numeryczne: min, max, mean, median, std, variance, mode, skewness, kurtosis. Kategoryczne: słownik liczebności każdej kategorii + mode. Obliczenia realizuje compute\_statistics(). |
| 3 | **Korelacje atrybutów (dwie metody)** | ✅ | compute\_correlation() generuje macierze **Pearson** i **Spearman**, prezentowane w zakładce *Data & Stats*. |
| 4 | **Ekstrakcja podtablic (dwa warianty)** | ✅ | extract\_subtable() pozwala usunąć wskazane wiersze/kolumny albo zachować tylko podane (flaga keep). |
| 5 | **Zastępowanie wartości** | ✅ | replace\_values() zamienia wskazaną wartość w kolumnie; replace\_all\_values() podmienia wszystkie wartości w kolumnie. |
| 6 | **Skalowanie i standaryzacja** | ✅ | scale\_columns() korzysta z **StandardScaler** lub **MinMaxScaler**. |
| 7 | **Obsługa braków danych** | ✅ | handle\_missing\_values() umożliwia usunięcie wierszy z brakami (*remove*) albo wypełnienie ich średnią, medianą lub modą. |
| 8 | **Usuwanie duplikatów** | ✅ | Przycisk w zakładce *Data Cleaning & Transformation* wywołuje remove\_duplicates(). |
| 9 | **Kodowanie symboliczne (trzy metody)** | ✅ | • **One‑Hot** (one\_hot\_encoding()),  • **Binary** (binary\_encoding() – kody numeryczne),  • **Target** (target\_encoding() – średnia zmiennej celu dla kategorii). |
| 10 | **Wizualizacje danych (≥ 4 rodzaje)** | ✅ | Dostępne: Histogram, Boxplot, Bar Chart, Line Plot, Pie Chart. Wykresy generuje generate\_plot(). |
| 12 | **Intuicyjne GUI** | ✅ | PySimpleGUI, pięć zakładek: Creators & Info, Data & Stats, Replacement & Subtable, Scaling & Visualisation, Data Cleaning & Transformation. |
| 13 | **Dokumentacja projektu** | ✅\* | Niniejszy opis stanowi część rozdziału 2.1 dokumentacji tworzonej wg wzoru *dokumentacjaHD\_2.docx*. |

\* Ostateczna kompozycja dokumentu zostanie przygotowana po zakończeniu testów.

## Wykorzystane technologie

|  |  |
| --- | --- |
| Technologia | Rola w projekcie |
| **Python 3.x** | Główny język implementacji aplikacji. |
| **PySimpleGUI** | Budowa interfejsu graficznego (zakładki, przyciski, tabele, pola tekstowe). |
| **Pandas** | Struktury danych *DataFrame*, wczytywanie/eksport CSV, filtrowanie i transformacje tabelaryczne. |
| **NumPy** | Szybkie obliczenia tablicowe i operacje numeryczne wykorzystywane przy statystykach i skalowaniu. |
| **SciPy (stats)** | Zaawansowane miary statystyczne: skośność (skew) i kurtoza (kurtosis). |
| **Matplotlib** + FigureCanvasTkAgg | Tworzenie wszystkich wykresów oraz ich osadzanie w oknie GUI. |
| **Seaborn** | Rozszerzone wizualizacje (boxploty, heatmapa korelacji, palety kolorów). |
| **scikit‑learn** | Preprocessing (**StandardScaler**, **MinMaxScaler**, **LabelEncoder**), podział danych (train\_test\_split), modele ML (**LogisticRegression**, **KMeans**). |

## Projekt GUI

Kilka rysunków / screenów z krótkim opisem

# Wygląd i użytkowanie aplikacji

## Wymagania do uruchomienia aplikacji

m.in. wymagania sprzętowe i programowe

## Obsługa aplikacji

à la manual ...

### Wczytanie zbioru danych

...

### Funkcjonalność 2

...

### Funkcjonalność 3

...

# Eksperymenty na danych

## Wykorzystane zbiory danych

**a) Adult Census Income – opis przygotował Dawid Olko**

* **Pochodzenie** – UCI Machine Learning Repository, oryginalnie zestaw powstał na bazie spisu powszechnego USA z 1994 r.
* **Tematyka / problem** – predykcja rocznego dochodu gospodarstwa domowego (<=50K lub >50K) na podstawie danych demograficznych i zawodowych.
* **Rozmiar** – 48 842 rekordy, 14 atrybutów wejściowych + 1 cel (dochód).
* **Struktura cech**  
   – age (int), workclass (8 kategorii), fnlwgt (waga finalna), education + education‑num, marital‑status, occupation (15 kategorii), relationship, race, sex, capital‑gain, capital‑loss, hours‑per‑week, native‑country (41 krajów), income (cel).
* **Problematyka jakości danych** – wartość ? oznacza brak (ok. 7 % przypadków w niektórych kolumnach), umiarkowana nierównowaga klas (~76 % w kategorii <= 50 K).
* **Zastosowanie w projekcie** – obliczanie statystyk, wizualizacja rozkładów, klasyfikacja przy użyciu Logistic Regression oraz segmentacja K‑Means.

**b) Chronic Kidney Disease – opis przygotował Piotr Smoła**

* **Pochodzenie** – UCI Machine Learning Repository; dane kliniczne zebrane w szpitalach indyjskich (bliższe źródło: Tavakkoli et al.).
* **Tematyka / problem** – diagnostyka i wczesne wykrywanie przewlekłej choroby nerek (klasyfikacja: ckd vs notckd).
* **Rozmiar** – 400 rekordów, 24 atrybuty + etykieta classification.
* **Struktura cech (wybrane)**  
   – parametry medyczne: age, bp (blood pressure), sg (specific gravity), al (albumin), su (sugar), bgr, bu, sc, sod, pot, hemo, pcv, wc, rc.  
   – atrybuty symboliczne: rbc, pc, pcc, ba, htn, dm, cad, appet, pe, ane.
* **Problematyka jakości danych** – braki dochodzące do 25 % w wybranych kolumnach, umiarkowana nierównowaga klas (250 × ckd, 150 × notckd).
* **Zastosowanie w projekcie** – testowanie procedur czyszczenia danych (usuwanie braków, imputacja), skalowanie zmiennych i eksperymenty z wizualizacją (heatmapy, boxploty).

## Przebieg eksperymentów i wyniki

opis eksperymentu/ów (np. budowy modelu klasyfikatora i klasyfikacji danych)

każdy z autorów opisuje swój własny eksperyment na wybranych przez siebie danych

wyniki np. w formie tabelki

## Analiza uzyskanych wyników i wnioski

opis uzyskanych wyników, analiza, wnioski, porównania …

każdy z autorów osobno

# Literatura