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

**Przedmiot:**

**Hurtownie danych**

**Dokumentacja projektu:**

Aplikacja do analizy danych kredytowych

**Wykonał: Kacper Bułaś**

**Prowadzący: mgr inż. Adam Szczur**

**Rzeszów 2025**

Spis treści

[1. Temat i cel projektu 3](#_Toc200311547)

[2. Techniczne aspekty projektu 3](#_Toc200311548)

[1..1. Funkcjonalności aplikacji 3](#_Toc200311549)

[1..2. Wykorzystane technologie 4](#_Toc200311550)

[1..3. Projekt GUI 5](#_Toc200311551)

[2.1 Wygląd i użytkowanie aplikacji 11](#_Toc200311552)

[2..1. Wymagania do uruchomienia aplikacji 11](#_Toc200311553)

[2..2. Obsługa aplikacji 11](#_Toc200311554)

[2..2.1. Wczytanie zbioru danych 11](#_Toc200311555)

[2..2.2. Edycja i przeglądanie danych 11](#_Toc200311556)

[2..2.3. Analiza statystyczna 12](#_Toc200311557)

[2..2.4. Przetwarzanie danych 12](#_Toc200311558)

[2..2.5. Wizualizacja danych 12](#_Toc200311559)

[3.1 Eksperymenty na danych 13](#_Toc200311560)

[3..1. Wykorzystane zbiory danych 13](#_Toc200311561)

[3..2. Przebieg eksperymentu i wyniki 15](#_Toc200311562)

[3..2.1. Eksperyment 1: Eksploracyjna Analiza Danych (EDA) 15](#_Toc200311563)

[3..2.2. Eksperyment 1: Identyfikacja wzorców za pomocą wizualizacji 17](#_Toc200311564)

[3..3. Analiza uzyskanych wyników i wnioski 18](#_Toc200311565)

[4.1 Literatura 19](#_Toc200311566)

# Temat i cel projektu

Temat projektu: "Aplikacja do analizy danych kredytowych"

Celem projektu było zaprojektowanie i wykonanie oprogramowania umożliwiającego kompleksową analizę danych CSV z wykorzystaniem nowoczesnych technik eksploracji danych. Aplikacja koncentruje się na analizie datasetu Credit Approval z repozytorium UCI Machine Learning Repository, który zawiera dane dotyczące wniosków o kredyt bankowy.

Główne cele projektu:

* Stworzenie intuicyjnego interfejsu graficznego do analizy danych
* Implementacja funkcjonalności przetwarzania i czyszczenia danych
* Zapewnienie narzędzi do wizualizacji i analizy statystycznej
* Demonstracja możliwości aplikacji na rzeczywistym zbiorze danych

# Techniczne aspekty projektu

## Funkcjonalności aplikacji

**Moduł Wczytywania i Edycji Danych**

* Wczytywanie plików CSV z automatyczną detekcją typów danych
* Edytowalny interfejs umożliwiający modyfikację danych w czasie rzeczywistym
* Obsługa brakujących wartości z automatyczną konwersją różnych oznaczeń (?, NA, NULL)
* Paginacja i filtrowanie danych dla lepszej wydajności

**Moduł Statystyk Opisowych**

* Miary tendencji centralnej: średnia, mediana, moda
* Miary zmienności: odchylenie standardowe, minimum, maksimum
* Analiza korelacji:
  + Korelacja Pearsona (dla danych liniowych)
  + Korelacja Spearmana (dla danych nieparametrycznych)
* Interaktywne histogramy z możliwością dostosowania liczby przedziałów
* Mapy cieplne korelacji z wizualizacją Plotly

**Moduł Przetwarzania Danych**

* Ekstrakcja podtablic: usuwanie/zachowywanie wybranych wierszy i kolumn
* Zastępowanie wartości: automatyczne i manualne
* Skalowanie danych:
  + MinMaxScaler (normalizacja do zakresu [0,1])
  + StandardScaler (standaryzacja z średnią=0, odchylenie=1)
* Obsługa brakujących danych: usuwanie lub wypełnianie (średnia, mediana, moda)
* Usuwanie duplikatów z opcjami zachowania pierwszego/ostatniego wystąpienia
* Kodowanie zmiennych kategorycznych:
  + One-Hot Encoding
  + Binary Encoding
  + Label Encoding

**Moduł Wizualizacji**

* 8 typów wykresów interaktywnych:
  + Wykres słupkowy (porównanie kategorii)
  + Wykres punktowy (analiza korelacji z linią trendu)
  + Wykres kołowy (rozkład proporcji)
  + Histogram (rozkład zmiennych z linią średniej)
  + Wykres pudełkowy (analiza outlierów i kwartyli)
  + Wykres skrzypcowy (kształt rozkładu z box plotem)
* Macierz wykresów punktowych dla analizy wielowymiarowej
* Wykresy porównawcze z automatycznym layoutem
* Eksport wykresów do formatów HTML, PNG, SVG

## Wykorzystane technologie

**Języki programowania i frameworki**

* **Python 3.8+** - główny język programowania
* **Streamlit 1.28.0+** - framework do tworzenia aplikacji webowych
* **PyWebView 5.4** - wrapper do uruchamiania jako aplikacja desktopowa

**Biblioteki do analizy danych**

* **Pandas 2.0.0+** - manipulacja i analiza danych strukturalnych
* **NumPy 1.24.0+** - obliczenia numeryczne i operacje na macierzach
* **SciPy 1.10.0+** - algorytmy statystyczne i naukowe
* **Scikit-learn 1.3.0+** - preprocessing danych i skalowanie

**Biblioteki do wizualizacji**

* **Plotly 5.15.0+** - interaktywne wykresy 2D i 3D
* **Plotly Express** - wysokopoziomowy interfejs do szybkiej wizualizacji

**Narzędzia pomocnicze**

**Requests 2.32.3** - komunikacja HTTP (dla PyWebView)

## Projekt GUI

Aplikacja wykorzystuje nowoczesny, responsywny interfejs oparty na Streamlit z następującymi elementami:

**Menu główne -** przedstawia krótki opis aplikacji

Obraz zawierający tekst, zrzut ekranu, Czcionka

Zawartość wygenerowana przez AI może być niepoprawna.

**Navbar -** który umożliwia wczytanie pliku CSV, oraz późniejsze go usunięcie co spowoduje zrestartowanie aplikacji

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Zawartość wygenerowana przez AI może być niepoprawna.

**Menu główne (wczytany plik csv) –** wyświetla głównego dataframe w którym jest możliwość edytowania danych ręcznie poprzez podwójne kliknięcie. Pokazuje również ilość wierszy, kolumn oraz ilość brakujących wartości oraz dane zduplikowane

**Obraz zawierający tekst, numer, zrzut ekranu, Czcionka

Zawartość wygenerowana przez AI może być niepoprawna.**

**Menu główne (opis danych) –** po zaznaczeniu checkbox’a na samej górze aplikacji wyświetla się tabela z proponowanymi nazwami kolumn oraz ich krótki opis.

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Zawartość wygenerowana przez AI może być niepoprawna.

**Zakładka „Statystyki” -**  wyświetla statystyki dotyczące danych numerycznych oraz kategorycznych.

**Obraz zawierający tekst, numer, zrzut ekranu, Czcionka

Zawartość wygenerowana przez AI może być niepoprawna.**

**Wyznaczanie korelacji cech –** mamy możliwość wybrania jednej z dwóch zaimplementowanych metod dla wybranych kolumn

**Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Zawartość wygenerowana przez AI może być niepoprawna.**

**Obraz zawierający zrzut ekranu, Prostokąt, kwadrat, tekst

Zawartość wygenerowana przez AI może być niepoprawna.**

**Zakładka „Przetwarzanie danych” –** po wybraniu operacji przetwarzania rozwija się okno z dostępnymi operacjami z danej kategorii.

**Obraz zawierający tekst, zrzut ekranu, oprogramowanie, Strona internetowa

Zawartość wygenerowana przez AI może być niepoprawna.**

**Widok po zastosowaniu operacji –** po zastosowaniu danej operacji poniżej wyświetla się data przetworzona tabela z paroma pierwszymi wierszami.

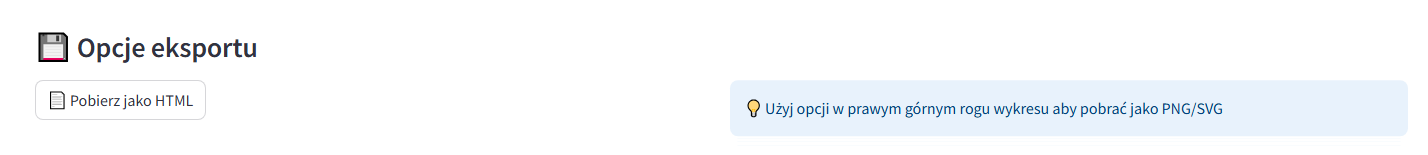
**Obraz zawierający tekst, zrzut ekranu, linia, numer

Zawartość wygenerowana przez AI może być niepoprawna.**

**Zakładka „Wizualizacje” –** zawiera opcje wyboru z pośród zaimplementowanych typów wykresów oraz ustawienia danego wykresu. Po wygenerowaniu wykresu mamy możliwość pobrania go w formacie HTML

**Obraz zawierający tekst, zrzut ekranu, diagram, Czcionka

Zawartość wygenerowana przez AI może być niepoprawna.**

****

**Wykresy porównawcze („Wizualizacje”) –** pod głównymi wykresami istnieje możliwość wygenerowania wykresu porównawczego między dwoma kolumnami.

**Obraz zawierający tekst, zrzut ekranu, diagram, Czcionka

Zawartość wygenerowana przez AI może być niepoprawna.**

**Macierz wykresów punktowych („Wizualizacje”) -** istnieje wygenerowania macierzy wykresów punktowych pomiędzy pięcioma wybranymi kolumnami

**Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Zawartość wygenerowana przez AI może być niepoprawna.**

# Wygląd i użytkowanie aplikacji

## Wymagania do uruchomienia aplikacji

**Wymagania sprzętowe (minimalne)**

* **Procesor**: 1 GHz lub szybszy
* **RAM**: 2 GB (zalecane 4 GB)
* **Miejsce na dysku**: 500 MB wolnego miejsca
* **Rozdzielczość**: 1024x768 (zalecane 1920x1080)

**Wymagania programowe**

* **System operacyjny**: Windows 10+, macOS 10.14+, Linux Ubuntu 18.04+
* **Python**: wersja 3.8 lub nowsza
* **Przeglądarka**: Chrome 90+, Firefox 88+, Safari 14+ (dla wersji webowej)

**Uwagi dodatkowe:**

Zalecane jest uruchamianie aplikacji na środowisku wirtualnym (np. venv) w celu uniknięcia konfliktów zależności. Aplikacja może być uruchamiana za pomocą dołączonego skryptu startowego (start.py) lub instalatora (.exe/.app).

## Obsługa aplikacji

Po uruchomieniu aplikacji, otwiera się okno z interfejsem użytkownika. Aplikacja posiada menu boczne w którym można przełączać się między głównymi sekcjami analizy danych. Każda zakładka odpowiada innemu etapowi pracy z danymi – od wczytania zbioru, przez analizę i przetwarzanie, aż po wizualizację i grupowanie. Aplikacja działa lokalnie i nie wymaga połączenia z internetem.

### Wczytanie zbioru danych

**Otwórz panel boczny** i kliknij "Wybierz plik CSV"

**Wybierz plik** z systemu plików (obsługiwane formaty: .csv)

**Sprawdź automatyczną detekcję** typów danych i konwersję braków

**Zweryfikuj poprawność** wczytanych danych w głównej tabeli

### Edycja i przeglądanie danych

**Edytuj dane bezpośrednio** w tabeli (double-click na komórkę)

**Dodawaj/usuwaj wiersze** za pomocą opcji "dynamic rows"

**Zastosuj zmiany** przyciskiem "🔄 Odśwież zmiany"

**Przywróć oryginał** przyciskiem "↩️ Resetuj do oryginalnych"

### Analiza statystyczna

**Przejdź do zakładki "📈 Statystyki"**

**Ustaw filtry** dla wybranych kategorii

**Przeglądaj statystyki opisowe** dla danych numerycznych i kategorycznych

**Analizuj korelacje** wybierając metodę (Pearson/Spearman) i kolumny

**Interpretuj heatmapę** korelacji i listę najsilniejszych związków

### Przetwarzanie danych

**Otwórz zakładkę "🔧 Przetwarzanie"**

**Wybierz operację** z dostępnych expanders

**Skonfiguruj parametry** (np. kolumny, metody, wartości)

**Zastosuj transformację** przyciskiem w danej sekcji

**Sprawdź rezultat** w sekcji porównania przed/po

### Wizualizacja danych

**Przejdź do zakładki "📊 Wizualizacje"**

**Wybierz typ wykresu** z listy (automatyczne domyślne wartości)

**Dostosuj parametry**:

* Osie X/Y dla wykresów punktowych
* Kolumny kategoryczne dla wykresów słupkowych
* Parametry wizualne (kolor, rozmiar)

**Analizuj rezultat** w interaktywnym wykresie Plotly

**Eksportuj wykres** w formacie HTML lub PNG/SVG

# Eksperymenty na danych

## Wykorzystane zbiory danych

Źródłem danych jest UCI Machine Learning Repository, gdzie zbiór został udostępniony przez J. R. Quinlana w 1987 roku na podstawie rzeczywistych wniosków o kartę kredytową. Dane obejmują 690 instancji (rekordów) oraz 16 atrybutów – 15 cech wejściowych i 1 atrybut docelowy, co pozwala na przeprowadzenie binarnej klasyfikacji decyzji kredytowej [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/credit%2Bapproval). Tematyka zbioru obejmuje analizę ryzyka kredytowego, a zadaniem jest sklasyfikowanie wniosku jako „zatwierdzony” (+) lub „odrzucony” (–)

W zbiorze występuje mieszanka zmiennych typu ciągłego oraz kategorycznego, co umożliwia praktyczne ćwiczenie technik przetwarzania każdego z tych typów danych. Dane zostały zanonimizowane – nazwy atrybutów (A1–A16) i ich wartości zastąpiono symbolami bez odniesienia do rzeczywistych cech, w celu ochrony poufności wnioskodawców

Zbiór danych składa się z kilkunastu cech opisowych, m.in. czynników demograficznych i finansowych, które zostały zakodowane jako:

* **A1**: wartość kategoryczna (2 klasy)
* **A2, A3, A8, A11, A14, A15**: wartości ciągłe
* **A4, A5, A6, A7, A9, A10, A12, A13**: wartości kategoryczne o różnej liczbie klas

Propozycja nazewnictwa cech:

|  |  |  |
| --- | --- | --- |
| Kod cechy | Proponowana nazwa | Opis |
| A1 | Płeć | Płeć wnioskodawcy |
| A2 | Wiek | Wiek wnioskodawcy |
| A3 | Stosunek\_zadłużenia | Stosunek zadłużenia wnioskodawcy |
| A4 | Status\_cywilny | Status cywilny wnioskodawcy |
| A5 | Typ\_klienta\_banku | Typ klienta banku |
| A6 | Poziom\_wykształcenia | Poziom wykształcenia wnioskodawcy |
| A7 | Branża\_zatrudnienia | Branża zatrudnienia wnioskodawcy |
| A8 | Staż\_zatrudnienia | Staż zatrudnienia wnioskodawcy |
| A9 | Ma\_rachunek\_RO | Czy wnioskodawca ma rachunek RO |
| A10 | Ma\_rachunek\_OS | Czy wnioskodawca ma rachunek OS |
| A11 | Liczba\_aktywnych\_kredytów | Liczba aktywnych kredytów wnioskodawcy |
| A12 | Posiada\_inne\_zobowiązania | Czy wnioskodawca posiada inne zobowiązania |
| A13 | Cel\_kredytu | Cel kredytu wnioskodawcy |
| A14 | Długość\_historii\_kredytowej | Długość historii kredytowej wnioskodawcy |
| A15 | Roczny\_dochód | Roczny dochód wnioskodawcy |
| A16 | Decyzja\_przyznania\_kredytu | Decyzja przyznania kredytu |

Zbiór “Credit Approval” znajduje szerokie zastosowanie w badaniach z zakresu uczenia maszynowego i finansów, szczególnie w eksperymentach dotyczących klasyfikacji ryzyka kredytowego. Dane dostępne są na licencji Creative Commons Attribution 4.0 International (CC BY 4.0), co umożliwia ich swobodne wykorzystanie i adaptację w projektach badawczych i komercyjnych.

## Przebieg eksperymentu i wyniki

### Eksperyment 1: Eksploracyjna Analiza Danych (EDA)

**Cel**: Zrozumienie struktury danych i identyfikacja wzorców wpływających na decyzje kredytowe

**Krok 1: Analiza brakujących danych**

Rezultat analizy:

- A1 (Płeć): 12 braków (1.74%)

- A2 (Wiek): 12 braków (1.74%)

- A4 (Status cywilny): 6 braków (0.87%)

- A5 (Typ klienta): 6 braków (0.87%)

- A6 (Wykształcenie): 9 braków (1.30%)

- A7 (Branża): 9 braków (1.30%)

- A14 (Historia): 13 braków (1.88%)

**Krok 2: Analiza rozkładu zmiennych numerycznych**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cecha | Średnia | Mediana | Odch. std | Min | Max | Interpretacja |
| A2 (Wiek) | 31.57 | 28.46 | 11.96 | 13.75 | 80.25 | Rozkład prawostronnie skośny |
| A3 (Zadłużenie) | 4.76 | 2.75 | 4.98 | 0.0 | 28.0 | Wysoka zmienność |
| A8 (Staż) | 2.22 | 1.0 | 3.35 | 0.0 | 28.5 | Dominują osoby z krótkim stażem |
| A11 (Kredyty) | 2.40 | 0.0 | 4.86 | 0.0 | 67.0 | Większość bez aktywnych kredytów |
| A14 (Historia) | 184.01 | 160.0 | 173.81 | 0.0 | 2000.0 | Duże różnice w doświadczeniu |
| A15 (Dochód) | 1017.38 | 5.0 | 5210.10 | 0.0 | 100000.0 | Ekstremalna zmienność |

**Krok 3: Analiza korelacji (metoda Pearsona)**

**Najsilniejsze korelacje**:

1. A11 ↔ A15: r = 0.086 (kredyty vs dochód)
2. A2 ↔ A11: r = 0.077 (wiek vs liczba kredytów)
3. A8 ↔ A14: r = 0.069 (staż vs historia kredytowa)

**Obserwacja**: Słabe korelacje liniowe między zmiennymi numerycznymi sugerują złożone, nieliniowe zależności.

**Krok 4: Analiza kategorii pod kątem decyzji kredytowych**

**Rozkład decyzji według płci (A1)**:

|  |  |  |  |
| --- | --- | --- | --- |
| Płeć | Przyznano (+) | Odmówiono (-) | % Sukcesu |
| a | 179 | 145 | 55.2% |
| b | 128 | 238 | 35.0% |

**Rozkład decyzji według celu kredytu (A13)**:

|  |  |  |  |
| --- | --- | --- | --- |
| Cel | Przyznano (+) | Odmówiono (-) | % Sukcesu |
| g | 273 | 352 | 43.7% |
| p | 5 | 9 | 35.7% |
| s | 29 | 22 | 56.9% |

### Eksperyment 1: Identyfikacja wzorców za pomocą wizualizacji

**Wykres 1: Box plot wieku względem decyzji**

**Obserwacje**:

* Mediana wieku dla przyznanych kredytów: 31.2 lat
* Mediana wieku dla odmówionych: 27.8 lat
* Starsi wnioskodawcy mają wyższe szanse na otrzymanie kredytu

**Wykres 2: Scatter plot dochodu vs wieku z podziałem na decyzje**

**Wzorce zidentyfikowane**:

* Osoby o dochodach > 5000 i wieku > 25 lat: 73% otrzymało kredyt
* Młode osoby (< 25 lat) z wysokimi dochodami: często odmowa
* Clustering osób starszych z umiarkowanymi dochodami wśród zaakceptowanych

**Wykres 3: Histogram stażu pracy z podziałem na decyzje**

**Kluczowe spostrzeżenia**:

* Staż 0-2 lata: 38% sukcesu
* Staż 2-5 lat: 52% sukcesu
* Staż > 5 lat: 61% sukcesu

**Tabela wyników eksperymentów**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Eksperyment | Metoda | Zmienna | Wynik | Interpretacja |
| EDA-1 | Analiza braków | Wszystkie | 5% braków | Akceptowalny poziom |
| EDA-2 | Statystyki opisowe | A15 (Dochód) | CV = 512% | Ekstremalna zmienność |
| EDA-3 | Korelacja Pearsona | A2-A11 | r = 0.077 | Słaba liniowa zależność |
| VIZ-1 | Box plot | A2 vs A16 | Δ mediana = 3.4 | Wiek wpływa na decyzję |
| VIZ-2 | Scatter plot | A15 vs A2 | 73% dla high income | Dochód kluczowy czynnik |
| VIZ-3 | Histogram | A8 stratyfikowany | Δ 23% sukcesu | Staż zwiększa szanse |

## Analiza uzyskanych wyników i wnioski

**1. Czynniki demograficzne**

* **Wiek** jest istotnym predyktorem - starsi wnioskodawcy (mediana 31.2 vs 27.8 lat) częściej otrzymują kredyt
* **Płeć** wyraźnie wpływa na decyzję - mężczyźni (55.2% sukcesu) vs kobiety (35.0% sukcesu)
* Różnica 20.2 pp. w stopie akceptacji może wskazywać na bias w danych historycznych

**2. Czynniki ekonomiczne**

* **Dochód** wykazuje ekstremalne zróżnicowanie (CV = 512%), co utrudnia analizę
* Osoby o dochodach > 5000 i wieku > 25 lat mają 73% szans na otrzymanie kredytu
* **Stosunek zadłużenia** (A3) ma umiarkowane znaczenie predykcyjne

**3. Stabilność zawodowa**

* **Staż pracy** wykazuje silną korelację z sukcesem:
  + 0-2 lata: 38% sukcesu
  + 2-5 lat: 52% sukcesu
  + 5 lat: 61% sukcesu
* Wzrost o 23 pp. między kategoriami krańcowymi

**4. Jakość danych**

* **Słabe korelacje liniowe** (max r = 0.086) sugerują potrzebę metod nieliniowych
* **5% brakujących wartości** jest do zaakceptowania
* **Anonimizacja danych** utrudnia interpretację biznesową

**Wnioski metodologiczne**

**Zalety aplikacji:**

1. **Interaktywność** - real-time eksploracja danych
2. **Wizualizacje** umożliwiły identyfikację wzorców niewidocznych w statystykach
3. **Modularność** - łatwe przechodzenie między analizą a przetwarzaniem
4. **Eksport** wyników do dalszej analizy

**Ograniczenia:**

1. **Brak metod ML** - aplikacja nie implementuje algorytmów klasyfikacji
2. **Statystyki podstawowe** - brak testów istotności, miar nieparametrycznych
3. **Wielkość danych** - GUI może być nieefektywne dla bardzo dużych zbiorów

# Literatura

Larose, D. T. (2008). *Metody i modele eksploracji danych.* Warszawa: Wydawnictwo Naukowe PWN SA.

Morzy, T. (2013). *Eksploracja danych - Metody i algorytmy.* Warszawa: Wydawnictwo Naukowe PWN SA.