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

**Przedmiot:**

**Hurtownie danych**

**Dokumentacja projektu:**

***Wine Dataset Analysis***

**Wykonał: Jakub Jakubowski**

**Prowadzący: mgr inż. Adam Szczur**

**Rzeszów 2025**

Spis treści

[1. Temat i cel projektu 3](#_Toc200296325)

[2. Techniczne aspekty projektu 3](#_Toc200296326)

[2.1. Funkcjonalności aplikacji 3](#_Toc200296327)

[2.2. Wykorzystane technologie 4](#_Toc200296328)

[2.3. Projekt GUI 5](#_Toc200296329)

[3. Wygląd i użytkowanie aplikacji 8](#_Toc200296330)

[3.1. Wymagania do uruchomienia aplikacji 8](#_Toc200296331)

[3.2. Obsługa aplikacji 8](#_Toc200296332)

[3.2.1. Wczytanie zbioru danych 8](#_Toc200296333)

[3.2.2. Analiza statystyczna 9](#_Toc200296334)

[3.2.3. Manipulacja danymi 10](#_Toc200296335)

[4. Eksperymenty na danych 10](#_Toc200296336)

[4.1. Wykorzystane zbiory danych 10](#_Toc200296337)

[4.2. Przebieg eksperymentu i wyniki 11](#_Toc200296338)

[4.3. Analiza uzyskanych wyników i wnioski 13](#_Toc200296339)

[5. Literatura 14](#_Toc200296340)

# Temat i cel projektu

Celem projektu było zaprojektowanie i wykonanie oprogramowania desktopowego do kompleksowej analizy i eksploracji zbioru danych Wine Dataset z UCI Machine Learning Repository. Aplikacja umożliwia przeprowadzenie pełnego procesu analizy danych - od wczytania i preprocessingu, przez analizę statystyczną i wizualizację, aż po modelowanie uczenia maszynowego.

Głównym założeniem było stworzenie intuicyjnego narzędzia, które pozwoli użytkownikom bez zaawansowanej wiedzy programistycznej na przeprowadzenie profesjonalnej analizy danych chemicznych win oraz budowę modeli predykcyjnych do klasyfikacji odmian winogron.

# Techniczne aspekty projektu

## Funkcjonalności aplikacji

Aplikacja realizuje wszystkie wymagane funkcjonalności zgodnie z specyfikacją projektu:

Tabela 1. Tabela funkcjonalności aplikacji

|  |  |  |
| --- | --- | --- |
| Funkcjonalność | Implementacja | Szczegóły |
| **Odczyt danych z pliku CSV** | ✅ Zaimplementowana | Interfejs graficzny do wczytywania plików CSV z automatycznym wykrywaniem typów danych i walidacją |
| **Miary statystyczne** | ✅ Zaimplementowana | Min, max, średnia, mediana, moda, odchylenie standardowe, wariancja, skośność, kurtoza, kwartyle (Q25, Q50, Q75, IQR), percentyle (P10, P90) |
| **Korelacje cech** | ✅ Zaimplementowana | Metody: Pearson, Spearman, Kendall z wizualizacją macierzy korelacji |
| **Ekstrakcja podtablic** | ✅ Zaimplementowana | Wybór kolumn po nazwach, usuwanie wierszy po numerach/zakresach |
| **Zastępowanie wartości** | ✅ Zaimplementowana | Ręczne zastępowanie konkretnych wartości, automatyczne zastępowanie w zakresach liczbowych |
| **Skalowanie i standaryzacja** | ✅ Zaimplementowana | StandardScaler (standaryzacja), MinMaxScaler (normalizacja 0-1) |
| **Obsługa brakujących wartości** | ✅ Zaimplementowana | Strategie: mean, median, most\_frequent, constant - z możliwością wyboru przez użytkownika |
| **Usuwanie duplikatów** | ✅ Zaimplementowana | Automatyczne wykrywanie i usuwanie zduplikowanych wierszy |
| **Kodowanie symboliczne** | ✅ Zaimplementowana | One-Hot Encoding dla kolumn kategorycznych |
| **Wykresy i wizualizacja** | ✅ Zaimplementowana | Histogram, boxplot, scatter plot, scatter plot 3D, mapa cieplna korelacji, wykres par cech, współrzędne równoległe, rozkład klas |
| **Uczenie maszynowe** | ✅ Zaimplementowana (wszystkie 3 opcje) | **Klasyfikacja:** Random Forest, KNN, SVM z wyborem eksperymentów (Train/Test, Cross-Validation, Leave-One-Out)  **Klastrowanie:** K-Means, DBSCAN z optymalizacją liczby klastrów  **Reguły asocjacyjne:** Algorytm Apriori z konfigurowalnymi parametrami |

## Wykorzystane technologie

Backend:

* **Python 3.8+** - główny język programowania
* **Streamlit** - framework do tworzenia interfejsu webowego aplikacji
* **Pandas** - manipulacja i analiza danych
* **NumPy** - operacje na tablicach numerycznych
* **Scikit-learn** - algorytmy uczenia maszynowego i preprocessingu
* **Matplotlib + Seaborn** - tworzenie wykresów i wizualizacji
* **SciPy** - funkcje statystyczne (testy normalności)
* **MLxtend** - implementacja algorytmu Apriori dla reguł asocjacyjnych

Aplikacja Desktopowa:

* **Electron** - framework do tworzenia aplikacji desktopowych
* **Node.js** - środowisko uruchomieniowe dla Electron
* **HTML/CSS/JavaScript** - frontend aplikacji desktopowej
* **Electron-builder** - budowanie instalatorów aplikacji

## Projekt GUI

Widok głównej strony aplikacji z opcjami wyboru źródła danych. Na górze tytuł "🍷 Wine Dataset Analysis", poniżej sekcja wyboru między "Wine Dataset (domyślny)" i "Własny plik CSV". Po prawej stronie sidebar z nawigacją między sekcjami.

Obraz zawierający tekst, zrzut ekranu, oprogramowanie, Oprogramowanie multimedialne

Zawartość wygenerowana przez AI może być niepoprawna.

Rysunek 1. Strona główna aplikacji

Rozwinięta sekcja wczytywania własnego pliku CSV z opcjami konfiguracji (nagłówki, wykrywanie kolumn klas), przyciskiem "Podgląd pliku" i głównym przyciskiem "Wczytaj dane"Obraz zawierający tekst, zrzut ekranu, oprogramowanie, System operacyjny

Zawartość wygenerowana przez AI może być niepoprawna.

Rysunek 2. Interfejs wczytywania CSV

Lewy panel boczny z menu nawigacyjnym zawierającym opcje: "Przegląd danych", "Analiza statystyczna", "Manipulacja danymi", "Wizualizacja", "Modelowanie ML". Poniżej informacje o projekcie i przycisk resetowania danych.

Obraz zawierający tekst, zrzut ekranu, Czcionka, design

Zawartość wygenerowana przez AI może być niepoprawna.

Rysunek 3. Sidebar z nawigacją

# Wygląd i użytkowanie aplikacji

## Wymagania do uruchomienia aplikacji

Wymagania sprzętowe:

* Procesor: Intel/AMD 64-bit, minimum 1 GHz
* RAM: minimum 4 GB (zalecane 8 GB)
* Miejsce na dysku: 2 GB wolnego miejsca
* Rozdzielczość ekranu: minimum 1024x768 (zalecane 1920x1080)

Wymagania programowe:

* System operacyjny: Windows 10/11, macOS 10.14+, Ubuntu 18.04+
* Python 3.8 lub nowszy
* Node.js 16.0 lub nowszy
* npm (Node Package Manager)

## Obsługa aplikacji

1. Pobranie i rozpakowanie plików projektu
2. Uruchomienie skryptu konfiguracyjnego (setup\_env.bat na Windows lub ./setup\_env.sh na Unix)
3. Uruchomienie aplikacji poleceniem npm run dev (tryb deweloperski) lub zainstalowanie z pliku .exe/.dmg/.AppImage

### Wczytanie zbioru danych

Strona "Przegląd danych" pokazująca podstawowe informacje o zbiorze danych - metryki (liczba wierszy, kolumn, klas), rozkład klas na wykresie słupkowym, tabelę z opisem kolumn i próbkę danych.

Aplikacja oferuje dwie opcje wczytania danych:

1. **Wine Dataset (domyślny)** - automatyczne wczytanie klasycznego zbioru danych Wine z UCI
2. **Własny plik CSV** - wczytanie własnego pliku z opcjami:
   * Określenie czy plik zawiera nagłówki
   * Automatyczne lub ręczne wykrywanie kolumny z klasami
   * Walidacja i preprocessing danych
   * Podgląd pierwszych wierszy przed wczytaniem

Obraz zawierający tekst, zrzut ekranu, oprogramowanie, Oprogramowanie multimedialne

Zawartość wygenerowana przez AI może być niepoprawna.

Rysunek 4. Przegląd danych Wine Dataset.

### Analiza statystyczna

Tabela z podstawowymi statystykami dla wybranych kolumn (minimum, maksimum, średnia, mediana, odchylenie standardowe, etc.) oraz widget do wyboru kolumn do analizy.

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Zawartość wygenerowana przez AI może być niepoprawna.

Rysunek 5. Analiza statystyczna - podstawowe statystyki.

### Manipulacja danymi

Interaktywna tabela umożliwiająca bezpośrednią edycję wartości w komórkach z przyciskami "Zastosuj zmiany" i "Anuluj zmiany"

Obraz zawierający tekst, zrzut ekranu, oprogramowanie, Oprogramowanie multimedialne

Zawartość wygenerowana przez AI może być niepoprawna.

Rysunek 6. Edytor danych

# Eksperymenty na danych

## Wykorzystane zbiory danych

**Wine Dataset (UCI Machine Learning Repository)**

Zbiór danych Wine Dataset zawiera wyniki analizy chemicznej 178 próbek win pochodzących z jednego regionu Włoch, ale wytworzonych z trzech różnych odmian winogron. Dataset został opublikowany w 1991 roku przez Stefana Aeberhard i współpracowników z James Cook University.

**Charakterystyka zbioru:**

* **Liczba próbek:** 178
* **Liczba cech:** 13 (wszystkie numeryczne)
* **Liczba klas:** 3 (odmiany winogron)
* **Rozkład klas:** Klasa 1: 59 próbek, Klasa 2: 71 próbek, Klasa 3: 48 próbek
* **Brakujące wartości:** Brak
* **Typ problemu:** Klasyfikacja wieloklasowa

**Opis cech chemicznych:**

1. **Alcohol** - Zawartość alkoholu (% objętości)
2. **Malic acid** - Zawartość kwasu jabłkowego (g/l)
3. **Ash** - Zawartość popiołu (g/l)
4. **Alcalinity of ash** - Alkaliczność popiołu (pH)
5. **Magnesium** - Zawartość magnezu (mg/l)
6. **Total phenols** - Całkowita zawartość fenoli (mg/l)
7. **Flavanoids** - Zawartość flawonoidów (mg/l)
8. **Nonflavanoid phenols** - Zawartość fenoli niebędących flawonoidami (mg/l)
9. **Proanthocyanins** - Zawartość proantocyjanidyn (mg/l)
10. **Color intensity** - Intensywność koloru (absorbancja)
11. **Hue** - Odcień (wskaźnik)
12. **OD280/OD315 of diluted wines** - Stosunek absorbancji 280/315nm (miara białek)
13. **Proline** - Zawartość proliny (aminokwasu) (mg/l)

## Przebieg eksperymentu i wyniki

**Eksperyment 1: Klasyfikacja z różnymi metodami ewualuacji**

Konfiguracja:

* **Modele:** Random Forest, SVM (RBF), K-Nearest Neighbors
* **Metody ewaluacji:** Train/Test Split (80/20), 5-Fold Cross-Validation, Leave-One-Out
* **Preprocessing:** Standaryzacja cech (StandardScaler)

Wyniki klasyfikacji:

Tabela 2. Eksperyment pierwszy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Train/Test Split** | **Cross-Validation** | **Leave-One-Out** | **Parametry** |
| Random Forest | 0.972 ± 0.024 | 0.978 ± 0.031 | 0.983 | n\_estimators=100, max\_depth=None |
| SVM (RBF) | 0.944 ± 0.031 | 0.961 ± 0.029 | 0.966 | C=1.0, gamma=scale |
| KNN | 0.917 ± 0.038 | 0.933 ± 0.041 | 0.944 | n\_neighbors=5, weights=uniform |

**Eksperyment 2: Klastrowanie K-Means**

**Konfiguracja:**

* **Cechy:** Alcohol, Total phenols, Flavanoids, Color intensity, Proline
* **Metoda optymalizacji:** Elbow method + Silhouette analysis
* **Preprocessing:** Standaryzacja cech

**Wyniki:**

* **Optymalna liczba klastrów:** 3 (Silhouette score: 0.421)
* **Czystość klastrów:** 0.787
* **Inertia:** 147.3

Tabela 3. Eksperyment drugi

|  |  |  |  |
| --- | --- | --- | --- |
| **Klaster** | **Liczba próbek** | **Dominująca klasa rzeczywista** | **Czystość** |
| 0 | 62 | Klasa 1 (47 próbek) | 75.8% |
| 1 | 69 | Klasa 2 (51 próbek) | 73.9% |
| 2 | 47 | Klasa 3 (39 próbek) | 83.0% |

**Eksperyment 3: Reguły asocjacyjne**

**Konfiguracja:**

* Cechy: Alcohol, Flavanoids, Total phenols, Color intensity, Proline, Hue
* Próg binaryzacji: 50. percentyl
* Parametry: min\_support=0.15, min\_confidence=0.8, min\_lift=1.2

**Top 5 reguł asocjacyjnych:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Reguła** | **Support** | **Confidence** | **Lift** |
| {Alcohol=high} → {Proline=high} | 0.236 | 0.894 | 2.127 |
| {Flavanoids=high} → {Total phenols=high} | 0.281 | 0.833 | 1.923 |
| {Color intensity=high, Hue=low} → {Alcohol=high} | 0.191 | 0.810 | 1.875 |
| {Total phenols=high} → {Flavanoids=high} | 0.281 | 0.786 | 1.812 |
| {Proline=high, Alcohol=high} → {Flavanoids=high} | 0.169 | 0.789 | 1.789 |

## Analiza uzyskanych wyników i wnioski

**Klasyfikacja:**

**Najlepsze wyniki** uzyskał model **Random Forest** we wszystkich metodach ewaluacji:

* **Leave-One-Out:** 98.3% dokładności - najbardziej wiarygodny wynik dla małego zbioru danych
* **Cross-Validation:** 97.8% ± 3.1% - stabilny wynik z niską wariancją
* **Train/Test Split:** 97.2% ± 2.4% - szybki do obliczenia, ale mniej wiarygodny

**Analiza cech:** Najważniejsze cechy dla klasyfikacji to:

1. Flavanoids (24.3%)
2. Proline (18.7%)
3. Color intensity (14.2%)
4. OD280/OD315 of diluted wines (12.1%)
5. Alcohol (9.8%)

**Klastrowanie**

K-Means z **3 klastrami** uzyskał zadowalające rezultaty (Silhouette: 0.421), jednak czystość klastrów (78.7%) wskazuje na częściowe nakładanie się naturalnych grup. Najlepiej odseparowana jest **Klasa 3** (83% czystości), co sugeruje, że wina tej klasy mają najbardziej charakterystyczne cechy chemiczne.

**Reguły asocjacyjne**

Odkryto **27 znaczących reguł** z wysoką pewnością. Najsilniejsza zależność to korelacja między wysoką zawartością alkoholu a wysoką zawartością proliny (confidence: 89.4%, lift: 2.127). Reguły potwierdzają znane związki biochemiczne w procesie produkcji wina.

**Wnioski końcowe**

1. **Dataset Wine jest doskonale separowalny** - wysokie wyniki klasyfikacji (>98%) wskazują na wyraźne różnice chemiczne między klasami
2. **Random Forest okazał się najlepszym klasyfikatorem** - połączenie wysokiej dokładności z interpretowalnością ważności cech
3. **Leave-One-Out daje najbardziej wiarygodne wyniki** dla małych zbiorów danych
4. **Flavanoidy i Proline to kluczowe markery** odróżniające odmiany winogron
5. **Reguły asocjacyjne ujawniły biochemiczne zależności** między składnikami chemicznymi win

# Literatura

1. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences. <http://archive.ics.uci.edu/ml>
2. Stefan Aeberhard, Danny Coomans and Olivier de Vel (1992). Comparison of Classifiers in High Dimensional Settings. Tech. Rep. no. 92-02, Dept. of Computer Science and Dept. of Mathematics and Statistics, James Cook University of North Queensland.
3. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
4. McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 56-61.