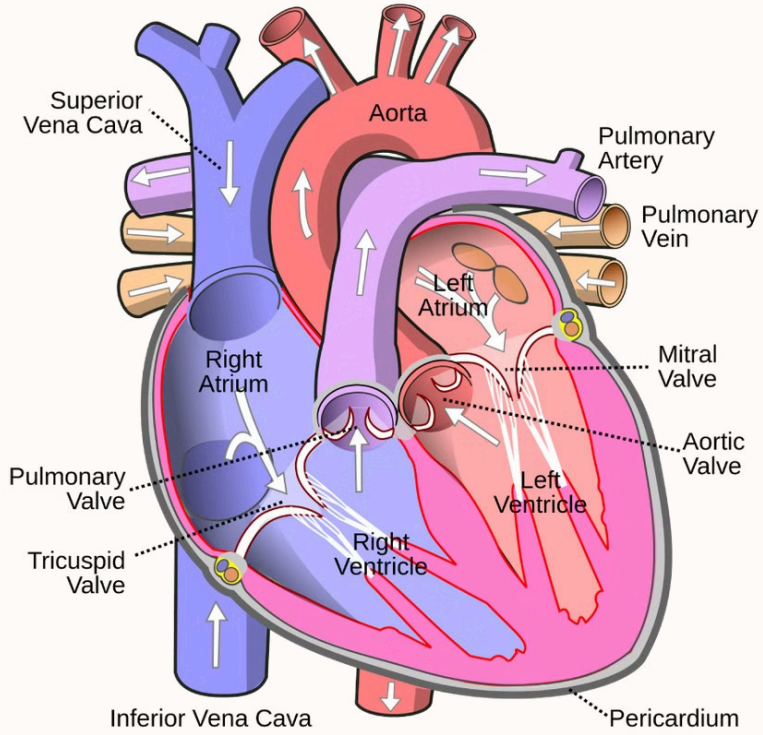
CardioRanger

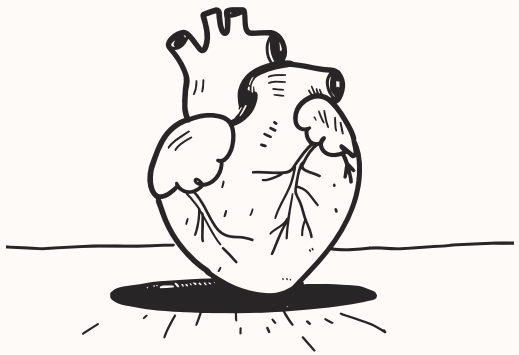
Aplikacja przewidująca potencjalną chorobą wieńcową serca (CHD) oparta na modelu uczenia maszynowego.



Agenda

1. Opis danych oraz cel
2. Baza danych
3. Eksploracyjna analiza danych (EDA)
4. Model i kalibracja
5. Aplikacja

Opis danych i cel



Badanie **Framingham Heart Study** to jedno z najbardziej przełomowych i długofalowych badań epidemiologicznych w historii medycyny. Rozpoczęte w 1948 roku w miasteczku Framingham w stanie Massachusetts, miało na celu zidentyfikowanie czynników ryzyka związanych z chorobami sercowo-naczyniowymi. Badanie objęło pierwotnie 5209 zdrowych mężczyzn i kobiet w wieku 30–62 lat, których stan zdrowia był monitorowany przez kolejne dekady.

Zbiór danych: https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset

Liczba obserwacji: 4240

Liczba kolumn: 16 kolumn (płeć, wiek, wykształcenie, czy pali, czy przyjmuje leki na nadciśnienie, historia udaru, nadciśnienie tętnicze, czy występuję cukrzyca, poziom cholesterolu, skurczowe ciśnienie tętnicze, rozkurczowe ciśnienie tętnicze, wskaźnik masy ciała (BMI), tętno, poziom glukozy, czy uczestnik miał chorobę (target).

Cel : Przewidzenie czy pacjent ma 10-letnie ryzyko przyszłej choroby wieńcowej (CHD).

Baza danych

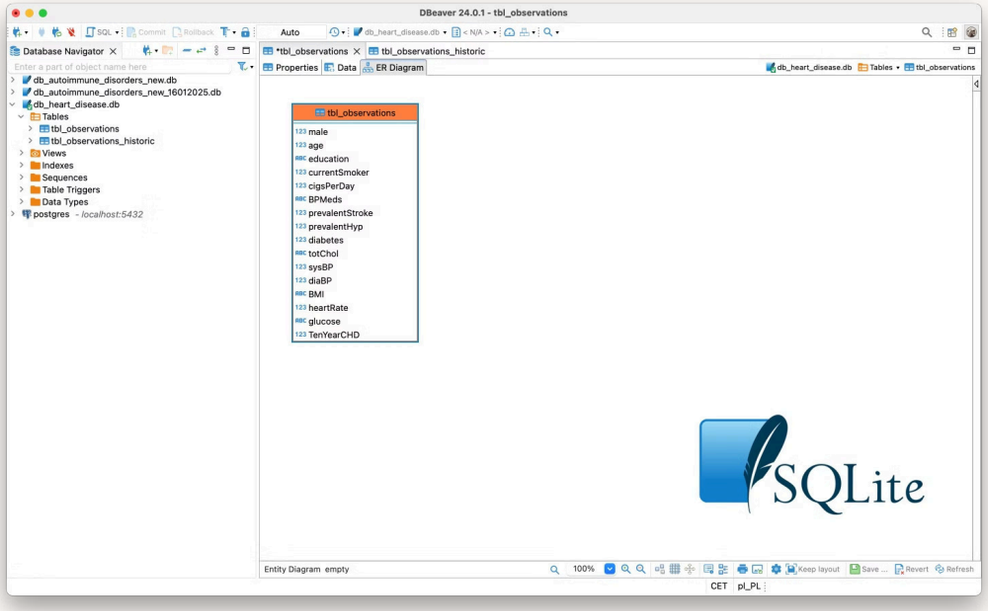
• **Technologia:** Baza danych została zbudowana w **SQLite**, lekkim i wydajnym systemie bazodanowym, który nie wymaga osobnego serwera i doskonale sprawdza się w aplikacjach lokalnych oraz analitycznych.

• **Struktura:** Baza zawiera **dwie tabele**:

tbl_observations – główna tabela zawierająca **4240 rekordów**, wykorzystywana do budowy modelu predykcyjnego.

tbl_observations_historic – tabela przeznaczona do **archiwizacji danych historycznych**, nieużywana w modelu.

• **Zawartość:** Obie tabele posiadają **16 kolumn**, przechowujących kluczowe zmienne medyczne i demograficzne wykorzystywane w analizie ryzyka chorób serca.



Eksploracyjna analiza danych (EDA)

Przygotowanie danych:

- Zmieniono nazwy niektórych zmiennych i ich typy.
- Usunięto **15,21%** przypadków z brakami danych, pozostawiając **3 658** rekordów.
- Brak duplikatów i błędnych wartości.
- Zmienną kategoryczną (edukacja) zakodowano metodą **OneHotEncoder**.
- Cechy numeryczne poddano standaryzacji (**StandardScaler**).

Wnioski z EDA:

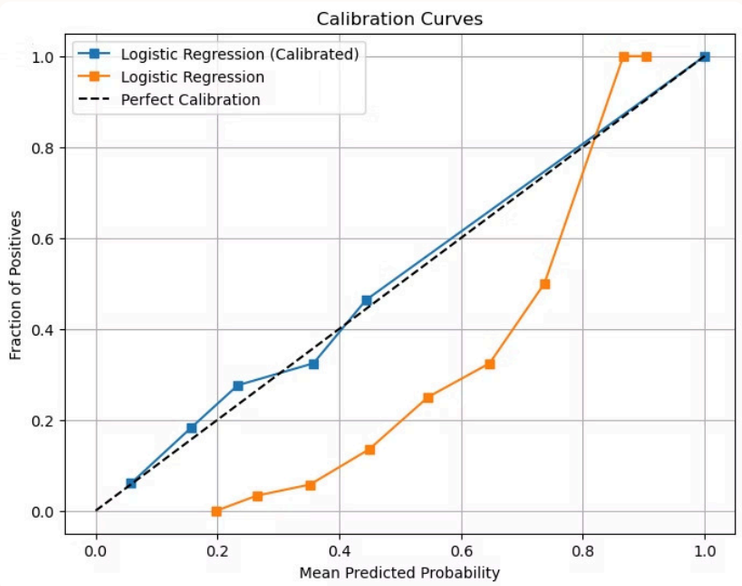
- Usunięto zmienne **currentSmoker** (palenie tytoniu) i **heartRate** (tętno) ze względu na brak istotnego wpływu na zmienną zależną.
- Największy wpływ na ryzyko choroby serca mają: **age** (wiek), **sysBP** (ciśnienie skurczowe), **prevalentHyp** (nadciśnienie tętnicze) oraz **CigsPerDay** (ilość papierosów spalanych dziennie).

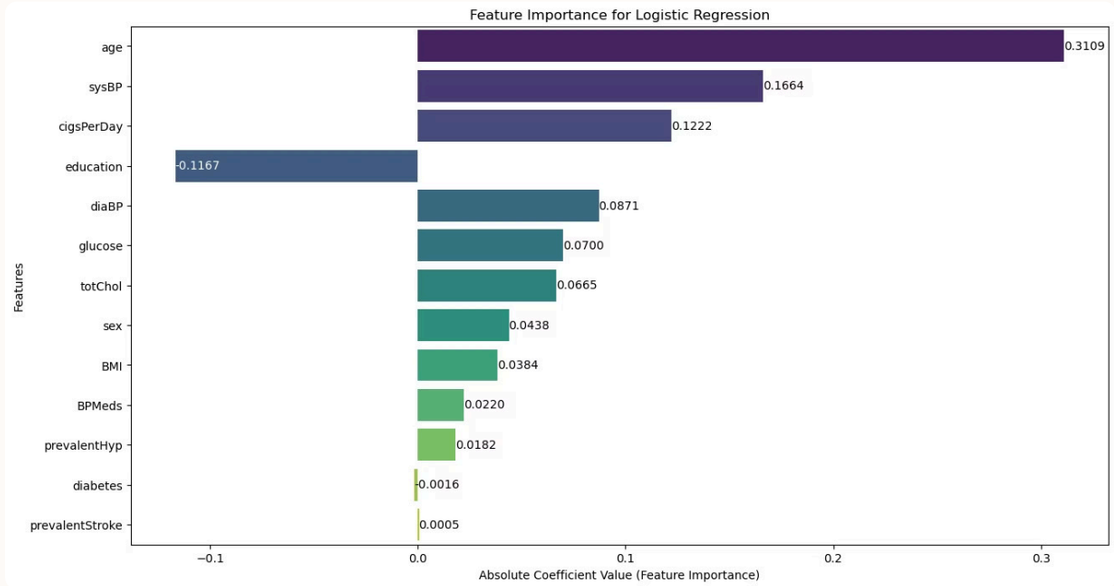
Model

W procesie tworzenia aplikacji przetestowanych zostało łącznie 4 algorytmy uczenia maszynowego na zbiorze danych repróbkowanych metodą SMOTE: **Regresja Logistyczna**, **KNN**, **Lasy Losowe** oraz **CatBoost**. Kluczową metryką ewaluacyjną, na podstawie której porównywana była skuteczność modeli, był wskaźnik **AUC**.

Najlepsze wyniki osiągnął model **Regresji Logistycznej**, uzyskując **AUC** na poziomie **75%**. Optymalne parametry obejmowały parametr kary L2, współczynnik regularyzacji C równy 0.001 oraz algorytm optymalizacji funkcji kosztu ***liblinear*.**

Aby zwiększyć stabilność i wiarygodność predykcji, finalny model został dodatkowo **skalibrowany** przy użyciu **regresji izotonicznej**. Istotność zmiennych w modelu została szczegółowo przeanalizowana, co pozwoliło na lepsze zrozumienie czynników wpływających na przewidywanie ryzyka choroby wieńcowej serca.

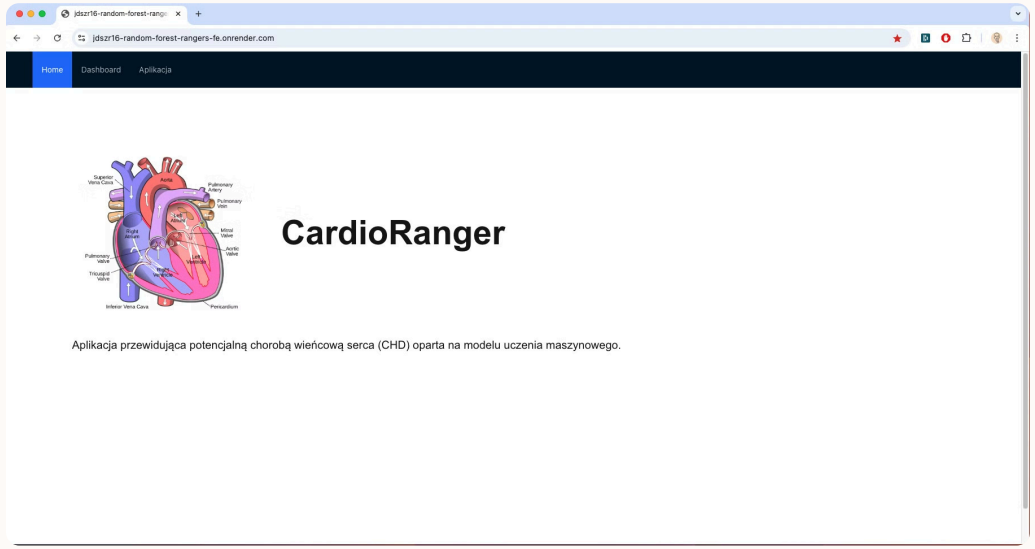


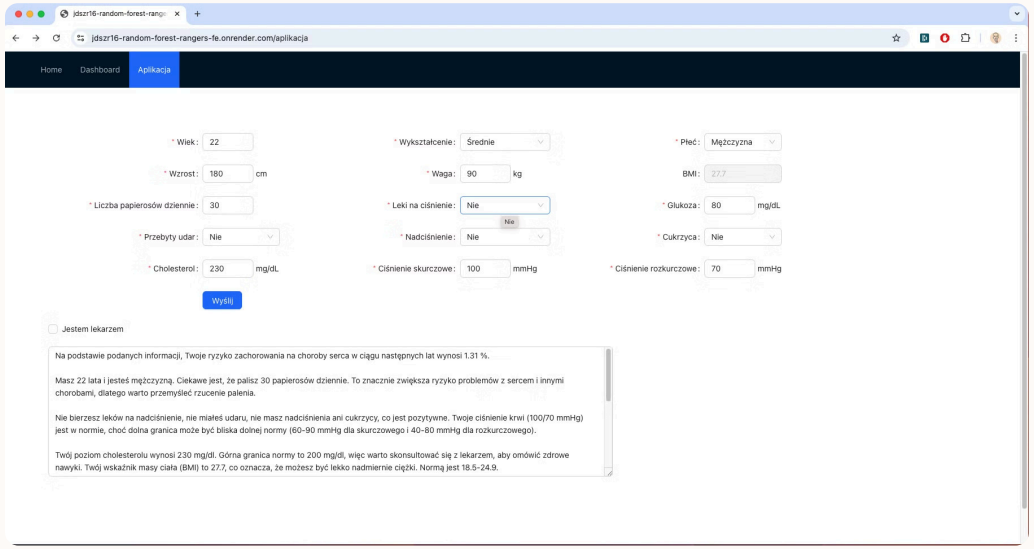


Aplikacja

Dashboard umożliwia interaktywne analizowanie czynników wpływających na zdrowie serca. **CardioRanger** to dashboard stworzony w **Streamlit**, który wizualizuje kluczowe wskaźniki zdrowotne i pozwala na eksplorację danych zapisanych w bazie **SQLite**.

Aplikacja





CardioRanger - Twój przewodnik po zdrowiu serca!