# Projekt - Środowisko uruchomieniowe AutoML

# Prognozowanie Cukrzycy

S24980 Jakub Zawierucha

S21254 Mikołaj Kowalczyk

S25147 Kamil Krzemieniewski

# Wstęp

## Cel projektu

Celem projektu jest stworzenie modelu uczenia maszynowego, który na podstawie dostępnych danych medycznych i demograficznych będzie w stanie przewidywać ryzyko wystąpienia cukrzycy u pacjentów. Projekt ma na celu wykorzystanie zestawu danych, który zawiera informacje takie jak wiek, płeć, wskaźnik masy ciała (BMI), ciśnienie krwi, historia palenia, obecność chorób serca oraz poziomy glukozy we krwi i HbA1c, w celu opracowania dokładnego i efektywnego narzędzia diagnostycznego.

## Opis wybranej bazy danych

Zestaw **danych do przewidywania cukrzycy**to zbiór danych medycznych i demograficznych od pacjentów, wraz z ich statusem cukrzycowym (pozytywnym lub negatywnym). Dane obejmują takie cechy, jak **wiek**, **płeć**, **wskaźnik** **masy** **ciała** (BMI), **nadciśnienie**, **choroby serca**, **historia palenia**, **poziom HbA1c i poziom glukozy we krwi**. Ten zestaw danych może być używany do budowania modeli uczenia maszynowego w celu przewidywania cukrzycy u pacjentów na podstawie ich historii medycznej i informacji demograficznych. Może to być przydatne dla pracowników służby zdrowia w identyfikacji pacjentów, którzy mogą być narażeni na ryzyko rozwoju cukrzycy i w opracowywaniu spersonalizowanych planów leczenia. Ponadto zestaw danych może być używany przez badaczy do eksploracji związków między różnymi czynnikami medycznymi i demograficznymi a prawdopodobieństwem rozwoju cukrzycy

## Model uczenia maszynowego

## 

**Uzasadnienie wyboru modelu XGBClassifier**

W projekcie do przewidywania cukrzycy zastosowano **XGBClassifier**, algorytm oparty na metodzie gradient boosting, który łączy zalety wydajności i wysokiej dokładności w klasyfikacji danych. Wybór tego modelu został uzasadniony następującymi powodami:

**1. Wysoka dokładność i odporność na overfitting (przeuczenie)**

XGBClassifier (XGBoost) jest jednym z najczęściej używanych algorytmów w konkursach uczenia maszynowego, takich jak Kaggle, ze względu na swoje zdolności do optymalizacji wyników. Dzięki wbudowanym mechanizmom, takim jak regularyzacja L1 i L2, XGBoost skutecznie zapobiega problemowi nadmiernego dopasowania do danych.

**2. Obsługa danych o zróżnicowanych cechach**

Model doskonale radzi sobie z danymi o różnej skali i typach. W przypadku analizy medycznej, gdzie cechy takie jak wiek, płeć, poziomy HbA1c czy glukozy różnią się między sobą pod względem zakresu wartości, XGBClassifier automatycznie znajduje optymalne sposoby przetwarzania i ważenia tych cech.

**3. Wbudowana analiza cech**

XGBClassifier oferuje możliwość oceny ważności cech poprzez ich wpływ na wynik modelu. Zastosowanie tej funkcji umożliwiło zidentyfikowanie, że poziom HbA1c ma największy wpływ na wyniki predykcji. Takie informacje są kluczowe w projektach medycznych, ponieważ pozwalają lepiej zrozumieć, które wskaźniki najbardziej przyczyniają się do diagnozy cukrzycy.

**4. Skalowalność i wydajność**

XGBoost został zaprojektowany z myślą o przetwarzaniu dużych zbiorów danych. Jego implementacja efektywnie wykorzystuje wielowątkowość i optymalizacje pamięci, co sprawia, że jest odpowiedni zarówno dla małych, jak i dużych zbiorów danych.

**5. Adaptacja do nierównomiernych danych**

W przypadku problemów klasyfikacyjnych, takich jak przewidywanie cukrzycy, często występuje nierównowaga w danych (np. liczba osób z cukrzycą może być mniejsza niż osób zdrowych). XGBClassifier oferuje możliwość dostosowania wag klas, co poprawia jakość predykcji w takich sytuacjach.

## Opis aplikacji

**1. Wykorzystane technologie**

* **FLASK** – technologia użyta do stworzenia API, które umożliwia komunikację między modelem a interfejsem użytkownika.
* **XGBClassifier** – model uczenia maszynowego zastosowany do predykcji ryzyka cukrzycy, charakteryzujący się wysoką dokładnością i możliwością interpretacji wyników.

**2. Opis funkcjonalności**

Aplikacja oferuje prosty interfejs w postaci strony internetowej, na której użytkownik wprowadza swoje dane:

* **Płeć** (*gender*),
* **Wiek** (*age*),
* **Nadciśnienie** (*hypertension*),
* **Choroby serca** (*heart\_disease*),
* **Historia palenia** (*smoking\_history*),
* **Wskaźnik masy ciała (BMI)** (*bmi*),
* **Poziom HbA1c** (*HbA1c\_level*),
* **Poziom glukozy we krwi** (*blood\_glucose\_level*).

Po wprowadzeniu tych danych API generuje wynik informujący, czy użytkownik może być zagrożony cukrzycą.

Najistotniejszym wskaźnikiem wpływającym na wynik jest **HbA1c\_level**, który odzwierciedla długoterminowy poziom glukozy we krwi. Dzięki temu użytkownicy mogą lepiej zrozumieć swoje ryzyko oraz podjąć działania profilaktyczne.

### Rozkład wieku:

Obraz zawierający tekst, zrzut ekranu, diagram, Wykres

Opis wygenerowany automatycznie

### Rozkład płci:

Obraz zawierający tekst, zrzut ekranu, Prostokąt, diagram

Opis wygenerowany automatycznie

### Rozkład BMI:

Obraz zawierający tekst, diagram, zrzut ekranu, Wykres

Opis wygenerowany automatycznie

### Rozkład nadciśnienia:

Obraz zawierający tekst, zrzut ekranu, Prostokąt, diagram

Opis wygenerowany automatycznie

### Rozkład występowania chorób serca:

Obraz zawierający tekst, zrzut ekranu, Prostokąt, diagram

Opis wygenerowany automatycznie

### Rozkład historii palenia:

Obraz zawierający tekst, zrzut ekranu, diagram, Wykres

Opis wygenerowany automatycznie

### Macierz konfuzji:

Obraz zawierający tekst, zrzut ekranu, diagram, design

Opis wygenerowany automatycznie

### Stopień znaczenia poszczególnych zmiennych:

Obraz zawierający tekst, zrzut ekranu, diagram, wyświetlacz

Opis wygenerowany automatycznie

Załączniki:

Link do repozytorium GIT

<https://github.com/PuszkaColi/SUML_Project>