Heart Disease Evaluation App

# Opis projektu

Heart Disease Evaluation App to aplikacja webowa, która umożliwia użytkownikom ocenę ryzyka wystąpienia choroby wieńcowej serca (CHD) lub zawału serca (MI) na podstawie różnych wskaźników zdrowotnych. Aplikacja pozwala na wprowadzenie danych dotyczących zdrowia, takich jak BMI, nawyki zdrowotne, historia medyczna oraz obecny stan zdrowia, a następnie przewiduje ryzyko chorób serca za pomocą wytrenowanego modelu uczenia maszynowego.

# Aplikacja Heart Disease Evaluation App może być przydatna dla:

**Pacjentów z grupy ryzyka**

Osoby z czynnikami ryzyka, takimi jak nadwaga, palenie tytoniu, wysokie ciśnienie krwi czy historia chorób serca w rodzinie, mogą monitorować swoje zdrowie i ocenić ryzyko wystąpienia CHD lub MI.

**Profesjonalistów medycznych**

Lekarze i dietetycy mogą korzystać z aplikacji jako narzędzia wspomagającego diagnozę oraz dostosowywanie planów leczenia i porad zdrowotnych dla swoich pacjentów.

**Osób dbających o zdrowie**

Osoby, które pragną prowadzić zdrowy tryb życia i prewencyjnie kontrolować stan swojego serca, mogą wykorzystać aplikację do regularnego monitorowania wskaźników zdrowotnych i nawyków.

# Dane

Dane użyte do stworzenia modelu zostały pobrane z artykułu dostępnego na stronie:

<https://www.kaggle.com/datasets/vaisakhnair/heart-disease-data>

Dane, które posłużą nam do oszacowania, czy dana osoba ma zwiększone szanse na atak serca:

**Sex (Płeć)**

Płeć osoby sprawdzanej (mężczyzna/kobieta). Badania pokazują, że płeć ma znaczenie w kontekście ryzyka chorób serca. Mężczyźni i kobiety mogą mieć różne profile ryzyka.

**Age Category (Kategoria Wieku)**

Wiek osoby podzielony na kategorie wiekowe: 18-24 lata, 25-29 lat, 30-34 lata, 35-39 lat, 40-44 lata, 45-49 lat, 50-54 lata, 55-59 lat, 60-64 lata, 65-69 lat, 70-74 lata, 75-79 lat, 80 lat i więcej. Wiek jest jednym z kluczowych czynników ryzyka chorób serca. Starszy wiek zazwyczaj zwiększa ryzyko.

**Race (Pochodzenie Etniczne)**

Pochodzenie etniczne osoby, podzielone na grupy: White (Biali), Black (Czarni), Asian (Azjaci), American Indian / Alaskan Native (Indianie Amerykańscy / Rdzenni mieszkańcy Alaski), Other (Inne), Hispanic (Latynosi). Różne grupy etniczne mogą mieć różne predyspozycje genetyczne i socjoekonomiczne wpływające na zdrowie serca.

**BMI (Body Mass Index)**

Współczynnik masy ciała, obliczany przez podzielenie masy ciała (w kilogramach) przez kwadrat wysokości (w metrach). Wysoki wskaźnik BMI jest powiązany z wyższym ryzykiem chorób serca ze względu na związek z otyłością.

**General Health (Ogólny Stan Zdrowia)**

Samoocena zdrowia osoby, podzielona na kategorie: Excellent (Doskonały), Very good (Bardzo dobry), Good (Dobry), Fair (Średni), Poor (Zły). Samoocena zdrowia jest wskaźnikiem ogólnego stanu zdrowia osoby i może odzwierciedlać inne ukryte problemy zdrowotne.

**Smoking (Palenie)**

Czy osoba paliła w swoim życiu więcej niż 100 papierosów. Kategoria Yes/No. Palenie tytoniu jest jednym z głównych czynników ryzyka chorób serca.

**Alcohol Drinking (Spożywanie Alkoholu)**

Czy osoba regularnie spożywa alkohol. Kategoria Yes/No. Regularne spożywanie alkoholu, szczególnie w dużych ilościach, może zwiększać ryzyko chorób serca.

**Stroke (Udar)**

Czy osoba miała kiedykolwiek udar. Kategoria Yes/No. Historia udaru może wskazywać na problemy z układem krążenia, które zwiększają ryzyko chorób serca.

**Diabetic (Cukrzyca)**

Czy osoba ma cukrzycę. Kategoria Yes/No. Cukrzyca znacząco zwiększa ryzyko chorób serca.

**Asthma (Astma)**

Czy osoba ma astmę. Kategoria Yes/No. Chociaż astma jest chorobą układu oddechowego, może mieć wpływ na ogólny stan zdrowia i serca.

**Kidney Disease (Choroby Nerek)**

Czy osoba ma jakiekolwiek problemy z nerkami. Kategoria Yes/No. Choroby nerek są często powiązane z chorobami serca ze względu na wspólne czynniki ryzyka.

**Skin Cancer (Rak Skóry)**

Czy osoba miała raka skóry. Kategoria Yes/No. Historia raka skóry może wskazywać na ogólny stan zdrowia i podatność na inne choroby.

**Physical Health (Zdrowie Fizyczne)**

Stan zdrowia fizycznego osoby, mierzony liczbą dni w miesiącu, w których osoba nie czuła się dobrze (0-30 dni). Liczba dni, w których osoba czuje się fizycznie źle, może być wskaźnikiem innych problemów zdrowotnych.

**Mental Health (Zdrowie Psychiczne)**

Stan zdrowia psychicznego osoby, mierzony liczbą dni w miesiącu, w których osoba nie czuła się dobrze (0-30 dni). Liczba dni złego samopoczucia psychicznego może wpływać na zdrowie serca, zwłaszcza w kontekście stresu i depresji.

**Difficulty Walking (Trudności z Chodzeniem)**

Czy osoba ma trudności z chodzeniem. Kategoria Yes/No. Trudności w poruszaniu się mogą wskazywać na problemy z układem krążenia lub inne poważne problemy zdrowotne.

**Physical Activity (Aktywność Fizyczna)**

Czy osoba podejmuje jakąś aktywność fizyczną poza pracą. Kategoria Yes/No. Regularna aktywność fizyczna jest ważnym czynnikiem zapobiegającym chorobom serca.

**Sleep Time (Czas Snu)**

Ilość godzin snu w ciągu doby (0-24 godziny). Odpowiednia ilość snu jest kluczowa dla zdrowia serca.

Model przewidujący ryzyko ataku serca wykorzystuje powyższe zmienne, aby oszacować prawdopodobieństwo wystąpienia tego zdarzenia u danej osoby. Każda zmienna wnosi istotne informacje na temat zdrowia fizycznego, psychicznego oraz stylu życia, co pozwala na dokładniejszą i bardziej spersonalizowaną ocenę ryzyka.

# Technologie

**Kedro: Zarządzanie przepływem danych i modelami uczenia maszynowego**

* Kedro to framework open-source, który ułatwia zarządzanie przepływem danych w projektach uczenia maszynowego. Pomaga w organizacji kodu, śledzeniu zależności i utrzymaniu standardów jakości oprogramowania. Dzięki Kedro możliwe jest łatwe definiowanie i wizualizowanie całego przepływu danych, co zwiększa przejrzystość i skalowalność projektów.

**Streamlit: Tworzenie interfejsu użytkownika aplikacji webowej**

* Streamlit to framework do tworzenia interaktywnych aplikacji webowych dla projektów uczenia maszynowego i analizy danych. Umożliwia szybkie budowanie interfejsów użytkownika bez potrzeby zaawansowanej wiedzy z zakresu web developmentu. Dzięki Streamlit można w prosty sposób prezentować wyniki modeli, umożliwiając użytkownikom interakcję z danymi i modelami w czasie rzeczywistym.

# Funkcjonalności

**Interfejs użytkownika: Intuicyjny interfejs do wprowadzania danych**

* Aplikacja oferuje prosty i przejrzysty interfejs użytkownika, który umożliwia łatwe i szybkie wprowadzanie danych niezbędnych do predykcji ryzyka chorób serca. Dzięki wykorzystaniu Streamlit, użytkownicy mogą intuicyjnie wprowadzać informacje takie jak płeć, wiek, pochodzenie etniczne, BMI oraz inne istotne dane zdrowotne. Interfejs jest zaprojektowany w sposób umożliwiający bezproblemową nawigację i minimalizujący ryzyko popełnienia błędów przy wprowadzaniu danych.

**Przetwarzanie danych: Kodowanie etykiet, usuwanie wartości odstających**

* Wprowadzone dane są automatycznie przetwarzane w celu przygotowania ich do analizy. Proces ten obejmuje kodowanie etykiet (np. zamiana wartości kategorii na numeryczne) oraz usuwanie wartości odstających, które mogą negatywnie wpływać na wyniki predykcji. Automatyczne przetwarzanie danych zapewnia, że wszystkie dane są w odpowiednim formacie i są gotowe do analizy przez model uczenia maszynowego. Kodowanie etykiet i normalizacja danych pozwala na lepszą interpretację zmiennych przez model.

**Predykcja: Wykorzystanie wytrenowanego modelu Random Forest Classifier do przewidywania ryzyka chorób serca**

* Aplikacja korzysta z wytrenowanego modelu **Random Forest Classifier**, który został stworzony na podstawie danych dotyczących zdrowia. Model ten wykorzystuje zaawansowane techniki uczenia maszynowego, aby dokładnie przewidywać ryzyko wystąpienia chorób serca u danej osoby. Po przetworzeniu danych wejściowych model generuje predykcje, które informują użytkownika o prawdopodobieństwie wystąpienia choroby.

**Wizualizacja: Czytelne wyświetlanie wyników predykcji i danych wejściowych**

* Wyniki predykcji oraz wprowadzone dane są prezentowane w sposób czytelny i zrozumiały dla użytkownika. Użytkownik widzi procentowe szanse oszacowane przez wytrenowany model i może manipulować pojednyczymi cechami sprawdzając ich wpływ na wynik.

# Uzasadnienie funkcjonalności aplikacji

Heart Disease Evaluation App ma na celu wspieranie użytkowników w ocenie ryzyka wystąpienia choroby wieńcowej serca (CHD) lub zawału serca (MI). Aplikacja umożliwia szybkie i łatwe wprowadzenie danych zdrowotnych oraz nawyków życiowych, a następnie przewiduje ryzyko wystąpienia tych schorzeń za pomocą zaawansowanego modelu uczenia maszynowego.

Praktyczne wykorzystanie aplikacji jest wszechstronne. Może ona służyć jako narzędzie wspomagające pracę lekarzy, pozwalając na wczesne wykrywanie potencjalnych problemów zdrowotnych u pacjentów. Użytkownicy indywidualni mogą z kolei samodzielnie monitorować swoje zdrowie i podejmować proaktywne kroki w celu zmniejszenia ryzyka wystąpienia chorób serca, takie jak zmiana diety czy zwiększenie aktywności fizycznej.

Aplikacja może także być używana w programach zdrowotnych i kampaniach profilaktycznych, gdzie edukacja i świadomość ryzyka chorób serca są kluczowe. Dzięki łatwej dostępności i intuicyjnemu interfejsowi, Heart Disease Evaluation App jest narzędziem, które może znacząco przyczynić się do poprawy zdrowia publicznego i indywidualnego.

Zastosowany model uczenia maszynowego:

**Random Forest Classifier** to algorytm uczenia maszynowego, który może być szczególnie skuteczny w przewidywaniu ryzyka chorób serca. Działa poprzez tworzenie wielu drzew decyzyjnych (stąd "forest" w nazwie), które są trenowane na różnych losowych próbkach danych wejściowych. Każde drzewo decyzyjne dokonuje swojej predykcji, a finalna decyzja klasyfikacyjna jest podejmowana na podstawie większości głosów ze wszystkich drzew.

W kontekście oceny ryzyka chorób serca, Random Forest Classifier mógłby działać w następujący sposób:

Model jest trenowany na zbiorze danych historycznych pacjentów, zawierających informacje takie jak BMI, nawyki zdrowotne, historia medyczna oraz wyniki zdrowotne (czy pacjent doświadczył CHD lub MI).

Algorytm tworzy wiele drzew decyzyjnych, z których każde analizuje różne losowe podzbiory danych. Dzięki temu każde drzewo może skupić się na różnych aspektach danych, co zwiększa różnorodność i odporność modelu na nadmierne dopasowanie.

Gdy użytkownik wprowadzi swoje dane zdrowotne do aplikacji, każdy z tych drzew decyzyjnych dokonuje predykcji, czy użytkownik jest w grupie ryzyka chorób serca. Finalna predykcja to agregacja (np. większościowy głos) wyników z wszystkich drzew.

Wynik przedstawiony użytkownikowi to ocena ryzyka oparta na agregacji predykcji ze wszystkich drzew decyzyjnych. Im więcej drzew zgadza się co do wyniku, tym bardziej pewna jest predykcja.

Random Forest jest popularnym wyborem w zastosowaniach medycznych ze względu na jego wysoką dokładność, odporność na nadmierne dopasowanie i zdolność do radzenia sobie z dużą ilością różnorodnych danych.

# Instalacja i uruchomienie

* Klonowanie repozytorium:
  + git clone https://github.com/dsuszek/pjatk-suml-final-project/tree/final-project
* Tworzenie i aktywacja środowiska w Anaconda:
  + conda create --name kedro python=3.10 -y
  + conda activate kedro
* Instalacja zależności:
  + pip install -r requirements.txt
* Wytrenowanie modelu przy pomocy Kedro:
  + Kedro run
* Uruchomienie aplikacji:
  + streamlit run homepage.py

Te kroki umożliwią skonfigurowanie i uruchomienie aplikacji, która pozwoli użytkownikom na ocenę ryzyka chorób serca na podstawie ich danych zdrowotnych.

# Diagram

# Obraz zawierający tekst, diagram, Prostokąt, zrzut ekranu Opis wygenerowany automatycznie