Uczenie maszynowe na podstawie zestawu danych “Cleveland Heart Disease”

Wykonanie: Jakub Matysek, Małgorzata Pach

Wstęp

Celem tego projektu jest zaprojektowanie, implementacja oraz weryfikacja modelu uczenia maszynowego zdolnego do wspierania diagnostyki kardiologicznej. W obszarze medycy prewencyjnej szczególne znaczenie ma możliwość nieinwazyjnego i wczesnego przewidywania choroby serca, bez konieczności kosztownych i inwazyjnych procesów ewentualnego dalszego leczenia, ma kluczowe znaczenie. Skuteczne narzędzia predykcyjne mogą przyczynić się do poprawy wyników leczenia, zwiększenia zaufania pacjentów do rekomendacji lekarskich oraz zmniejszenia śmiertelności poprzez wcześniejsze wykrycie nieprawidłowości.

Do realizacji celu powyższego zadania wykorzystano zbiór danych klinicznych „Cleveland Heart Disease”. Dane te obejmują zestaw kluczowych parametrów medycznych, takich jak wiek pacjenta, poziom cholesterolu, ciśnienie tętnicze, wyniki elektrokardiogramu (EKG), parametry testów wysiłkowych oraz inne wskaźniki diagnostyczne.

Analizowany model opiera się na klasyfikacji binarnej. Jego rolą jest ustalenie, czy pacjent spełnia kryterium obecności choroby serca, co prowadzi do dwóch możliwych wyników:

- Grupa 0 – wynik negatywny – brak ryzyka lub brak stwierdzonej obecności choroby serca.
- Grupa 1 – wynik pozytywny – stwierdzone ryzyko lub obecność choroby serca, bez stopnia zaawansowania.

Przetwarzanie wstępne

Jakość danych wejściowych jest bardzo ważna dla ostatecznej wydajności modelu, o czym mówi zasada „Garbage In, Garbage Out”. Zbiór danych składa się z 14 atrybutów medycznych opisujących stan zdrowia pacjenta. Surowe dane wymagały czyszczenia i transformacji przed przystąpieniem do budowy modelu. Przetwarzanie polegało na trzech – niżej opisanych – elementach: czyszczenie danych
i identyfikacja braków, imputacja brakujących wartości oraz transformacja zmiennej celowej.

Czyszczenie danych i identyfikacja braków

W surowej wersji zbioru danych występowały niepoprawne lub nietypowe wartości, m.in. oznaczenia „?”, które podczas wczytywania pliku CSV nie mogły zostać zinterpretowane jako liczby.
W implementacji zdefiniowano te wartości jako wartości brakujące (NaN), co umożliwiło ich dalszą analizę.

Wstępna weryfikacja danych zbioru ujawniła niewielką liczbę brakujących wartości w dwóch zmiennych:

- Ca – liczba głównych naczyń zabarwionych we fluoroskopii – 4 braki
- Thal – wynik scyntygrafii talu – 2 braki

Imputacja brakujących wartość

Gdy braki stanowią nieliczną część całości dokładnie tak jak w tym przypadku (6 na ponad 300 obserwacji), najlepszym rozwiązaniem jest ich wypełnienie czyli imputacja, rezygnując z usuwania rekordów, co mogłoby prowadzić do utraty cennych informacji klinicznych. Brakujące wartości zostały uzupełnione modą , czyli najczęściej występującą wartością w danej kolumnie. Jest to metoda preferowana dla zmiennych o charakterze kategorycznym lub dyskretnym (jak Ca czy Thal). Dane kategoryczne to te które mają sens liczbowy np. wiek czy wynik ciśnienia, natomiast dyskretne to te które przyjmują wartości opisowe jak płeć czy typ bólu. W przeciwieństwie do imputacji średnią, moda nie generuje wartości ułamkowych, dzięki czemu pozwala zachować), naturalny charakter i interpretowalność danych.

Transformacja zmiennej celowej – binaryzacja

Ważnym krokiem w przygotowaniu danych było przejście od problemu klasyfikacji wieloklasowej (pięć stopni zaawansowania choroby) do klasyfikacji binarnej (stwierdzono ryzyko/brak ryzyka). Podjęto taką decyzję z uwagi na dwa główne powody: praktyczny (użyteczność kliniczna) oraz techniczny (balans danych).

Oryginalna zmienna celu zawierała wartości od 0 do 4, wskazujące stopień zaawansowania zwężenia naczyń, gdzie 0 to brak choroby, a 1-4 to rosnący stopień zaawansowania choroby. Ze względu na ograniczoną liczbę przypadków w klasach 2, 3 i 4, problem uproszczono do znalezienia samej obecności choroby, zmieniając go na zadanie klasyfikacji binarnej.

Uzasadnienie Praktyczne

Oryginalna zmienna docelowa zawierała wartości od 0 do 4. Z punktu widzenia wstępnej diagnostyki, kluczowe jest rozstrzygnięcie, czy pacjent powinien zostać skierowany na dalszą, szczegółową diagnostykę kardiologiczną (grupa 1), czy nie (grupa 0). Uproszczenie problemu do dwóch klas zwiększa użyteczność modelu jako narzędzia wspomagania decyzji.

Uzasadnienie Techniczne

Oryginalny zbiór danych charakteryzował się silnym niezbalansowaniem klas, ponieważ wyższe stopnie zaawansowania choroby (klasy 2, 3 i 4) miały bardzo małą liczebność. Trenowanie modelu na tak niezbalansowanych danych skutkowałoby niską wiarygodnością, zwłaszcza w detekcji rzadkich, ale poważnych przypadków. Łącząc klasy 1, 2, 3 i 4 w jedną, ogólną klasę "1" (Obecność Choroby), stworzono bardziej zbalansowany problem binarny. To zapewniło wystarczającą ilość danych dla obu grup – zdrowych (0) i chorych (1).

Analiza Balansu Klas

Po transformacji dokonano analizy danych w zmiennej docelowej:

- Pacjenci Zdrowi (0): 164 przypadki
- Pacjenci Chorzy (1): 139 przypadków

Uzyskany zbiór jest stosunkowo zrównoważony, co oznacza, że nie było konieczności stosowania zaawansowanych technik balansowania danych (np. Oversampling, Undersampling), które mogłyby sztucznie wprowadzić wszelkie nieprawidłowości bądź złą analizę przedstawionych danych.

Analiza eksploracyjna danych

Analiza eksploracyjna danych (EDA) miała na celu zrozumienie struktury zbioru, identyfikację zależności między cechami, a zmienną docelową oraz ocenę potencjału predykcyjnego. Analiza wizualna, pozwoliła zidentyfikować czynniki, które są najsilniej powiązane z diagnozą choroby serca. Analizie poddano zarówno zmienne numeryczne, jak i kategoryczne.

Analiza zmiennych numerycznych

Macierz korelacji oraz wizualizacje rozkładów pozwoliły wyodrębnić cechy o szczególnie wysokiej mocy diagnostycznej:

- *Tętno maksymalne (thalach):* Jest jedną z najważniejszych predyktorów. Wskazuje wyraźną negatywną korelację z chorobą. Pacjenci ze zdiagnozowaną chorobą serca osiągają niższe maksymalne tętno podczas wysiłku (mediana ok. 140 bpm) w porównaniu do osób zdrowych (mediana ok. 160 bpm).
- *Obniżenie odcinka ST (oldpeak):* Wyższe wartości tego parametru (reprezentującego stopień obniżenia odcinka ST na EKG w trakcie wysiłku) są silnie powiązane z obecnością choroby. Dla osób zdrowych mediana oldpeak wynosi niemal 0, co czyni tę cechę wysoce dyskryminacyjną.
- *Wiek (age):* Pacjenci z chorobą byli statystycznie starsi (mediana ok. 58 lat) niż pacjenci zdrowi (ok. 52 lata). Wiek stanowi ważny predyktor, choć nie jest on dominujący.
- *Cholesterol (chol) i Ciśnienie (trestbps):* W analizowanym zbiorze cechy te wykazywały niewielką wartość dyskryminacyjną. Rozkłady dla obu klas w znacznej mierze pokrywały się.

Analiza zmiennych kategorycznych

Analiza wykresów słupkowych i rozkładów dla zmiennych kategorycznych wykazała, że niektóre zmienne kategoryczne są silnie powiązane z występowaniem choroby

- *Rodzaj bólu w klatce piersiowej (cp):* Rodzaj bólu jest tutaj kluczowy. Choć typowy (1) i nietypowy dławicowy ból (2 oraz 3) mogą występować u osób zdrowych, klasa 4 (bezobjawowy/asymptomatic) jest najsilniejszym wskaźnikiem choroby. Zdecydowana większość przypadków w tej kategorii to osoby chore.
- *Płeć (sex):* W analizowanym zbiorze danych mężczyźni (1) charakteryzowali się znacznie wyższym ryzykiem choroby niż kobiety (0). Jest to zgodne z ogólnymi trendami demograficznymi dla chorób kardiologicznych.
- *Thal (Wynik Scyntygrafii):* Wynik badania izotopowego jest wysoce diagnostyczny. Wyniki "Wada Odwracalna" (Reversable defect) oraz "Wada Stała" (Fixed defect) są silnie powiązane z pozytywną diagnozą, podczas gdy wynik "Normal" dominował w grupie osób zdrowych.

Wnioski z EDA i Wybór Cech

Na podstawie przeprowadzonej analizy można stwierdzić, że największą moc predykcyjną posiadają cechy związane z testami wysiłkowymi i objawami nietypowymi: thalach, oldpeak, cp oraz wynik thal. Atrybuty te stanowią kluczowe elementy dalszego procesu budowy modelu uczenia maszynowego.

Uzasadnienie Wyboru Metod i Narzędzi

W procesie analitycznym zastosowano zestaw standardowych, lecz kluczowych decyzji technicznych, których celem było zapewnienie rzetelności analizy, reprodukowalności wyników oraz maksymalizacji wiarygodności uzyskanego modelu predykcyjnego. Wybór metod został podyktowany zarówno charakterystyką danych medycznych, jak i dobrymi praktykami stosowanymi w uczeniu maszynowym.

Przygotowanie cech

Kodowanie Zmiennych Kategorycznych (One-Hot Encoding)

Zmienne kategoryczne o charakterze nominalnym, takie jak **cp** (rodzaj bólu w klatce piersiowej), **restecg** (wynik spoczynkowego EKG) oraz **thal** (wynik scyntygrafii talu), zostały przekształcone z wykorzystaniem metody **One-Hot Encoding** (kodowania zero-jedynkowego).

Uzasadnienie wyboru metody:

Wartości tych zmiennych nie posiadają charakteru porządkowego ani hierarchicznego – ich numeryczne oznaczenia służą wyłącznie identyfikacji kategorii. Bezpośrednie wykorzystanie takich wartości w modelach uczenia maszynowego, w szczególności w modelach liniowych (np. regresji logistycznej), mogłoby prowadzić do błędnej interpretacji relacji liczbowych pomiędzy kategoriami i w konsekwencji zniekształcenia wag cech.

Implementacja:

Zastosowano funkcję `pd.get_dummies` z parametrem `drop_first=True`, co pozwoliło uniknąć problemu pułapki zmiennych fikcyjnych. Usunięcie jednej kategorii referencyjnej zapobiega idealnej współliniowości cech, co jest niezbędne dla stabilnej estymacji parametrów i poprawnego działania modeli regresyjnych.

Podział Danych i Reprodukowalność

Stratyfikowany Podział na Zbiory Treningowy i Testowy

Zbiór danych został podzielony na część treningową (75%) oraz testową (25%). Podział ten przeprowadzono w sposób stratyfikowany, z wykorzystaniem parametru `stratify=y`.

Uzasadnienie:

Stratyfikacja zapewnia zachowanie identycznych proporcji klas zmiennej docelowej w obu zbiorach, co jest kluczowe dla wiarygodnej oceny skuteczności modelu. Nawet przy relatywnie zrównoważonym rozkładzie klas (około 54% przypadków zdrowych i 46% chorych), losowy podział bez stratyfikacji mógłby prowadzić do zafałszowania wyników ewaluacji, szczególnie w kontekście detekcji przypadków pozytywnych. Zachowanie struktury klas zwiększa stabilność i interpretowalność uzyskanych rezultatów.

Środowisko Programistyczne

Całość analizy została przeprowadzona w języku Python, z wykorzystaniem powszechnie stosowanych i uznanych bibliotek: Pandas (manipulacja i przygotowanie danych), Matplotlib/Seaborn (wizualizacja danych) oraz Scikit-learn (implementacja algorytmów uczenia maszynowego). Taki dobór narzędzi zapewnia wysoką reprodukowalność analizy, dostęp do zoptymalizowanych implementacji algorytmów oraz zgodność z aktualnymi standardami w obszarze analizy danych i uczenia maszynowego.

Wybór modeli klasyfikacyjnych

W ramach procesu modelowania przetestowano pięć algorytmów uczenia maszynowego należących do różnych rodzin. Celem było porównanie modeli o różnym stopniu złożoności oraz sprawdzenie, które podejście najlepiej radzi sobie z binarnym problemem klasyfikacji w danych klinicznych, przy zachowaniu kompromisu pomiędzy skutecznością a interpretowalnością.

Logistic Regression

Regresja logistyczna została wykorzystana jako model bazowy. Jest to algorytm często stosowany w analizach medycznych, umożliwiający zarówno klasyfikację binarną, jak i wyznaczenie prawdopodobieństwa wystąpienia choroby. Jego zaletą jest liniowa postać oraz łatwa interpretacja współczynników, które wskazują kierunek i siłę wpływu poszczególnych cech.

Random Forest

Random forest zastosowano jako metodę zespołową opartą na wielu drzewach decyzyjnych. Algorytm ten cechuje się wysoką odpornością na nadmierne dopasowanie do danych treningowych oraz zdolnością do modelowania nieliniowych i złożonych zależności pomiędzy cechami. Jest to szczególnie istotne w przypadku danych klinicznych, gdzie interakcje pomiędzy parametrami (np. wiekiem, tętnem i wynikami testów wysiłkowych) często nie mają charakteru liniowego.

K-Nearest Neighbors (KNN)

Algorytm KNN został uwzględniony jako model geometryczny, oparty na miarach odległości w przestrzeni cech. Klasyfikuje on obserwacje na podstawie podobieństwa do innych przypadków, co pozwala traktować go jako użyteczny punkt odniesienia dla metod nieliniowych. Jego zastosowanie umożliwia ocenę, czy lokalna struktura danych klinicznych niesie istotną informację predykcyjną.

Gaussian Naive Bayes

Model Gaussian Naive Bayes został przetestowany jako algorytm probabilistyczny oparty na założeniu niezależności cech. Pomimo że założenie to nie jest w pełni spełnione w danych medycznych, model ten często osiąga dobre wyniki przy niewielkiej liczbie obserwacji i niskiej złożoności obliczeniowej.

Decision Tree

Drzewo decyzyjne zapewnia bardzo wysoką interpretowalność, umożliwiając wygenerowanie przejrzystych reguł decyzyjnych typu *if–then*, co jest istotne z punktu widzenia zastosowań klinicznych. Algorytm ten charakteryzuje się również podatnością na nadmierne dopasowanie do danych treningowych. Jego uwzględnienie pozwoliło na bezpośrednie porównanie pojedynczego drzewa
z metodą zespołową Random Forest, która wykorzystuje jego strukturę, jednocześnie ograniczając problem overfittingu.

Ocena modeli

Ocenę wszystkich wytrenowanych modeli przeprowadzono na niezależnym zbiorze testowym, obejmującym dane, które nie były wykorzystywane w procesie trenowania. Takie podejście pozwoliło na obiektywną ocenę zdolności generalizacji modeli oraz ich rzeczywistej skuteczności w warunkach zbliżonych do praktycznego zastosowania klinicznego.

Skuteczności Modeli

W celu porównania jakości predykcyjnej pięciu zaimplementowanych algorytmów, posłużono się metryką dokładności wyznaczoną na zbiorze testowym. Zestawienie skuteczności prezentuje się następująco:

- Logistic Regression: 82,89%
- Random Forest: 78,95%
- Gaussian Naive Bayes: ok. 75%
- Decision Tree: ok. 72%
- KNN: ok. 66%

Uzyskane rezultaty wskazują, że w analizowanym problemie najlepiej sprawdził się model liniowy, pomimo dostępności bardziej złożonych metod nieliniowych.

Szczegółowa Analiza Logistic Regression

Logistic Regression osiągnęła najwyższą skuteczność spośród wszystkich testowanych algorytmów. Wyniki te sugerują, że po zastosowaniu odpowiedniego przygotowania danych (imputacja braków oraz kodowanie zmiennych kategorycznych), relacje pomiędzy cechami diagnostycznymi a zmienną celową mają charakter w dużej mierze liniowy.

Kluczowe metryki dla Regresji Logistycznej na zbiorze testowym przedstawiają się następująco:

- **Precyzja (Precision): 0,82** Model rzadko generuje fałszywe alarmy – w przypadku przewidzenia choroby, w 82% przypadków decyzja ta jest poprawna.
- **Czułość (Recall): 0,80** Model poprawnie identyfikuje 80% wszystkich pacjentów z rzeczywistą chorobą serca. Jest to parametr o kluczowym znaczeniu klinicznym, odnoszący się bezpośrednio do zdolności wykrywania patologii.

Interpretacja Wyników Pozostałych Modeli

Niższa skuteczność Random Forest w porównaniu do Logistic Regression sugeruje, że analizowany zbiór danych może być zbyt niewielki, aby w pełni wykorzystać potencjał bardziej złożonego modelu zespołowego bez ryzyka nadmiernego dopasowania (overfittingu). Ponadto, jeśli granica decyzyjna między klasami ma charakter liniowy, prostszy model parametryczny naturalnie osiąga lepsze wyniki. Słaby rezultat algorytmu KNN wskazuje natomiast, że w wielowymiarowej przestrzeni cech pacjenci nie tworzą wyraźnych, jednorodnych klastrów, co ogranicza skuteczność metod opartych na lokalnym sąsiedztwie.

Wnioski

Celem projektu było stworzenie modelu uczenia maszynowego wspierającego wstępną diagnostykę chorób serca w sposób nieinwazyjny. Wykorzystano zbiór „Cleveland Heart Disease", obejmujący 14 parametrów medycznych. Dane zostały wstępnie oczyszczone, uzupełnione o brakujące wartości oraz przekształcone do formy binarnej – wskazującej obecność lub brak choroby.

Analiza wykazała, że największą moc predykcyjną mają: maksymalne tętno (thalach), obniżenie odcinka ST (oldpeak), rodzaj bólu w klatce piersiowej (cp) oraz wynik scyntygrafii talu (thal). Te cechy zostały uwzględnione w budowie modeli.

Przetestowano pięć algorytmów: Logistic Regression, Random Forest, Decision Tree, KNN oraz Gaussian Naive Bayes. Podział danych na zbiory treningowy i testowy był stratyfikowany, co zapewniło zachowanie proporcji klas i obiektywną ocenę skuteczności.

Najlepszy wynik uzyskała regresja logistyczna (dokładność 82,89%, precyzja 0,82, czułość 0,80), co wskazuje na liniowy charakter zależności między cechami a ryzykiem choroby. Bardziej złożone modele nieliniowe osiągnęły nieco niższą skuteczność, prawdopodobnie ze względu na ograniczoną liczbę obserwacji.

Podsumowując, projekt potwierdził, że odpowiednie przygotowanie danych i selekcja kluczowych cech pozwalają na skuteczne wspomaganie wstępnej diagnostyki kardiologicznej, przy czym regresja logistyczna okazała się najbardziej efektywnym i interpretowalnym narzędziem.