Wczesna detekcja cukrzycy typu 2

Paulina Bogacz, Julita Janik, Anna Wygoda

June 20, 2025

1 Opis zbioru danych

Wykorzystany zbiór danych pochodzi z wieloletniego projektu badawczego realizowanego przez *National Institute of Diabetes and Digestive and Kidney Diseases*. Zbiór danych został przekazany do szerszego użytku badawczego przez Vincenta Sigillito.

Zbiór został pozyskany z platformy *Kaggle* i jest publicznie dostępny pod adresem: `https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data`

Badania prowadzone były wśród kobiet z plemienia Pima, zamieszkujących stan Arizona, w wieku 21 lat i starszych.

Struktura danych:

- Zbiór danych zawiera 768 obserwacji oraz 9 zmiennych.
- Wszystkie pomiary są typu numerycznego (num).
- Wśród zmiennych znajduje się 8 predyktorów numerycznych, będących podstawowymi pomiarami klinicznymi i demograficznymi.
- Zmienna docelowa *diabetes* (klasa) określa wynik testu OGTT jako pozytywny (1) lub negatywny (0).
- W oryginalnym zbiorze danych w niektórych pomiarach występują wartości równe 0, które z punktu widzenia fizjologii nie są możliwe.
- W analizie wartości zerowe zostały potraktowane jako dane brakujące, a następnie usunięte z zestawu danych w celu zachowania poprawności i wiarygodności modelowania.

Zmienna	Jednostka	Opis kliniczny
`Pregnancies`		Liczba przebytych ciąż
`Glucose`	mg/dL (OGTT)	Glikemia 2 godziny po doustnym obciążeniu glukozą
`BloodPressure`	mm Hg	Spoczynkowe ciśnienie tętnicze rozkurczowe
`SkinThickness`	mm	Grubość fałdu skórno-tłuszczowego nad mięśniem trójgłowym
`Insulin`	µU/mL	Stężenie insuliny w surowicy 2 godziny po OGTT
`BMI`	kg/m^2	Wskaźnik masy ciała (BMI)
`DiabetesPedigreeFunction`		Współczynnik genetycznej predyspozycji do cukrzycy typu 2
`Age`	lata	Wiek pacjentki
`Outcome`		Wynik testu OGTT (0/1)

2 Sformułowanie celu analizy danych

Celem niniejszej analizy jest opracowanie i ocena skuteczności modeli umożliwiających binarną klasyfikację medyczną pacjentek. Model ma za zadanie przeprowadzić klasyfikację pacjentek na podstawie zestawu pomiarów medycznych, przypisując je do klas odpowiadających obecności lub braku cukrzycy typu 2. W praktyce model może zostać wykorzystany do odróżniania pacjentek wymagających dalszej diagnostyki od tych o niskim ryzyku rozwoju choroby.

Uzasadnienie wyboru zastosowanych narzędzi analizy danych

Zmienna opisująca wynik testu, przyjmuje wartości binarne (0 lub 1), co kwalifikuje problem jako zadanie binarnej klasyfikacji.

1. Pierwszym zastosowanym narzędziem jest metoda k-NN (k-Nearest Neighbors). Algorytm ten cechuje się prostotą implementacji oraz skutecznością, a także charakteryzuje się krótkim czasem uczenia modelu.

2. Kolejną zastosowaną metodą jest naiwny klasyfikator Bayesa. Algorytm ten przy niewielkich zbiorach danych treningowych oraz nie wymaga skomplikowanego strojenia hiperparametrów.

3. Ostatnią zastosowaną metodą są drzewa decyzyjne zbudowane za pomocą algorytmu C5.0. Metoda ta jest efektywna również przy mniejszych zbiorach danych oraz cechuje się wysoką wydajnością w zadaniach klasyfikacyjnych.

3 Specyfikacja zastosowanych modeli analitycznych

Dane zostały podzielone na zbiór treningowy oraz testowy w proporcji 80:20.

- **Metoda k-NN** została przeprowadzona przy zastosowaniu dwóch różnych metod normalizacji: normalizacji min-max oraz standaryzacji Z-score.
  W ramach tej metody przeprowadzono dobór optymalnego hiperparametru k, porównując modele dla wartości $k \in \{5, 7, 11, 18, 24, 26\}$.
- **Naiwny klasyfikator Bayesa** był testowany na danych nieznormalizowanych ze względu na specyfikę modelu. Dla poprawy jakości modelu zastosowano wygładzanie Laplace'a.
- **Algorytm C5.0** (drzewa decyzyjne) analizowano na danych nieznormalizowanych, zgodnie ze specyfiką tej metody. Próby ulepszenia modelu polegały na modyfikacji liczby iteracji w adaptacyjnym boostingu oraz zastosowaniu macierzy wag błędów o różnym koszcie dla poszczególnych typów błędów.

4 Opis weryfikacji jakości zbudowanych modeli

Ze względu na fakt, że każdy z zastosowanych modeli dotyczy problemu klasyfikacji, weryfikacja ich jakości została przeprowadzona na podstawie macierzy krzyżowej, która przedstawia liczbę poprawnych oraz niepoprawnych klasyfikacji na zbiorze testowym.

Aby rzetelnie ocenić skuteczność naszego modelu klasyfikacyjnego, wyróżniamy cztery podstawowe typy predykcji, które odnosimy do specyfiki diagnozowania cukrzycy:

- **True Positive (TP):** Pacjentka faktycznie chora na cukrzycę została prawidłowo zaklasyfikowana jako chora.
- **True Negative (TN):** Pacjentka zdrowa została poprawnie rozpoznana jako zdrowa.
- **False Positive (FP):** Zdrowa pacjentka została błędnie zakwalifikowana jako chora, co może prowadzić do niepotrzebnych badań..
- **False Negative (FN):** Pacjentka chora została błędnie zaklasyfikowana jako zdrowa, co jest najpoważniejszym błędem, ponieważ może skutkować brakiem odpowiedniego leczenia.

Na podstawie tych typów klasyfikacji wyliczamy kluczowe miary oceny modelu:

Precyzja: Wskazuje, jaka część pacjentek, które model oznaczył jako chore, faktycznie cierpi na cukrzycę. Wysoka precyzja pomaga zmniejszyć liczbę fałszywych alarmów i ograniczyć zbędne badania.

$$\text{Precyzja} = \frac{TP}{TP + FP}$$

Czułość: Kluczowy wskaźnik w diagnostyce medycznej, mierzący zdolność modelu do wykrycia wszystkich prawdziwych przypadków cukrzycy. Im wyższa czułość, tym rzadziej model pomija pacjentki faktycznie chore.

$$\text{Czułość} = \frac{TP}{TP + FN}$$

F1-score: Jest to średnia harmoniczna precyzji i czułości, służąca do oceny modeli, które muszą wypracować kompromis między dokładnym wykrywaniem cukrzycy a ograniczeniem liczby fałszywych diagnoz.

$$\text{F1-score} = 2 \cdot \frac{\text{Precyzja} \cdot \text{Czułość}}{\text{Precyzja} + \text{Czułość}}$$

4.1 Metoda k-NN

Poniżej przedstawione zostały wyniki zebrane w macierzach krzyżowych dla najlepszych wariantów każdej metody.
W metodzie k-NN, przy **normalizacji min-max**, optymalnym parametrem jest $k = 5$.

actual		**predicted**		**Row Total**
		no	**yes**	
no	Count	40	7	47
yes	Count	12	20	32
Column Total		52	27	79

- Wśród 32 pacjentek, u których faktycznie występuje cukrzyca, model prawidłowo zaklasyfikował 20 z nich. Natomiast w 12 przypadkach błędnie uznał pacjentki za zdrowe. Te błędy mogą prowadzić do przeoczenia choroby i braku wdrożenia odpowiedniego leczenia.

- Spośród 47 pacjentek, które rzeczywiście były zdrowe, model poprawnie rozpoznał brak cukrzycy w 40 przypadkach, natomiast 7 osób zostało błędnie zaklasyfikowanych jako chore. Ten rodzaj błędu jest mniej niebezpieczny z punktu widzenia zdrowia pacjentki, ale może być kosztowny ze względu na konieczność zlecenia dodatkowych badań.

W przypadku **standaryzacji Z-score**, optymalnym parametrem jest $k = 7$.

actual		**predicted**		**Row Total**
		no	**yes**	
no	Count	44	3	47
yes	Count	18	14	32
Column Total		62	17	79

- Spośród 32 pacjentek faktycznie chorujących na cukrzycę, 14 przypadków zostało poprawnie sklasyfikowanych jako chore, natomiast w 18 przypadkach model błędnie uznał je za zdrowe. Ten rodzaj błędu prowadzi do niewykrycia choroby i braku leczenia, co stanowi zagrożenie dla zdrowia pacjentki.

- Wśród 47 pacjentek, które w rzeczywistości były zdrowe, 44 zostały prawidłowo zaklasyfikowane, natomiast w 3 przypadkach model błędnie przewidział cukrzycę. Ten rodzaj błędu może skutkować dodatkowymi badaniami diagnostycznymi oraz potencjalnym kosztami.

4.2 Naiwny klasyfikator Bayesowski

W przypadku zastosowania naiwnego klasyfikatora Bayesowskiego, rodzaj normalizacji danych (lub jej brak) nie wpływa na poprawę jakości modelu. Ponadto, zastosowanie wygładzenia Laplace'a nie przyczyniło się do zwiększenia efektywności uzyskanych wyników.

actual		**predicted**		**Row Total**
		no	**yes**	
no	Count	40	11	51
yes	Count	7	21	28
Column Total		47	32	79

- Spośród 28 pacjentek faktycznie chorujących na cukrzycę, model prawidłowo rozpoznał 21 przypadków, natomiast w 7 przypadkach błędnie zaklasyfikował je jako zdrowe.

- Spośród 51 pacjentek zdrowych, model poprawnie sklasyfikował 40 jako zdrowe, jednak w 11 przypadkach błędnie przewidział cukrzycę.

4.3 Drzewa decyzyjne

Zastosowano drzewa decyzyjne zbudowane za pomocą algorytmu C5.0. Modyfikacja liczby iteracji w adaptacyjnym boostingu nie poprawia jakości modelu.

actual		**predicted**		**Row Total**
		no	**yes**	
no	Count	35	12	47
yes	Count	5	27	32
Column Total		40	39	79

- Analiza macierzy pomyłek wskazuje, że spośród 32 pacjentek faktycznie chorujących na cukrzycę, 27 zostało prawidłowo sklasyfikowanych, natomiast 5 błędnie uznano za zdrowe.

- Spośród 47 pacjentek zdrowych, 35 zostało poprawnie zaklasyfikowanych, a 12 błędnie sklasyfikowano jako chore.

Możliwą metodą poprawy modelu jest nadanie różnym typom błędnych klasyfikacji odpowiednich wag. Błąd polegający na zaklasyfikowaniu pacjentki chorej jako zdrowej (False Negative) może być obarczony kilkukrotnie większą wagą niż błąd typu False Positive, w którym zdrowe pacjentki zostają skierowane na dodatkowe badania w celu weryfikacji (choć błędnej) diagnozy cukrzycy. W przedstawionej analizie przyjęto, że waga błędu False Negative jest pięciokrotnie większa, co odzwierciedla większe ryzyko zdrowotne związane z przeoczeniem choroby.

Przy założeniu, że błędna klasyfikacja pacjentki chorej jako negatywnej jest obarczona wagą pięciokrotnie większą niż inne błędy, liczba tego typu przypadków uległa redukcji — z 5 do 2 przypadków. Jednakże kosztem tego była znaczna utrata ogólnej dokładności modelu.

actual		**predicted**		**Row Total**
		no	**yes**	
no	Count	26	21	47
yes	Count	2	30	32
Column Total		28	51	79

- Wśród 47 pacjentek zdrowych, tylko 26 zostało prawidłowo rozpoznanych jako zdrowe. Aż 21 pacjentek zostało błędnie sklasyfikowanych jako chore.
- Spośród 32 pacjentek chorujących na cukrzycę, 30 zostało poprawnie sklasyfikowanych jako chore.

Model, dążąc do minimalizacji groźniejszych błędów, generuje więcej fałszywych alarmów, w których zdrowe pacjentki są sklasyfikowane jako chorujące na cukrzyce. Może to prowadzić do niepotrzebnych badań diagnostycznych oraz kosztów z tym związanych.

5 Podsumowanie

Celem przeprowadzonej analizy było stworzenie modeli predykcyjnych umożliwiających wczesne wykrywanie cukrzycy typu 2 na podstawie zebranych pomiarów. Postawiony problem – binarna klasyfikacja obecności cukrzycy – został rozwiązany przy użyciu trzech modeli: k-NN, naiwnego klasyfikatora Bayesa oraz drzew decyzyjnych C5.0.

5.1 Metryki. Interpretacja wyników

Przedstawiamy wyniki uzyskane w poszczególnych modelach:

Metoda	**Dokładność**	**Czułość**	**Precyzja**	**F1-score**
k-NN (Normalizacja min–max)	0.7595	0.6250	0.7407	0.6780
k-NN (Standaryzacja Z-score)	0.7342	0.4375	**0.8235**	0.5714
Naiwny klasyfikator Bayesowski	0.7722	0.7500	0.6563	0.7000
Drzewa decyzyjne C5.0	**0.7848**	0.8438	0.6923	**0.7606**
Drzewa decyzyjne C5.0 (waga FN $\times 5$)	0.7089	**0.9375**	0.5882	0.7229

- Spośród zastosowanych metod najwyższą dokładność osiągnięto przy użyciu drzew decyzyjnych 78.48% Naiwny klasyfikator Bayesowski wykazał zbliżony wynik.
- Najwyższą czułość osiągnął model drzewa decyzyjnego C5.0 z zastosowaniem wagi 5 dla błędów False Negative osiągając wartość aż 93.75%. Wysoka czułość jest kluczowa w kontekście diagnostycznym, ponieważ minimalizuje ryzyko przeoczenia zachorowania na cukrzycę, co mogłoby prowadzić do braku leczenia.
- Najwyższą precyzję osiągnął k-NN ze standaryzacją Z-score 82.35%, co oznacza, że w tym przypadku liczba fałszywych alarmów (False Positive) była najmniejsza. Wysoka precyzja jest istotna, aby ograniczyć dodatkowe badania.

Metryką stanowiącą kompromis między czułością a precyzją jest *F1- score*. Model drzewa decyzyjnego C5.0, który nie uwzględniał wag, osiągnął najwyższy F1-score na poziomie 76.06%, co czyni ten model najbardziej wyważonym w zastosowaniach ogólnych.

5.2 Wnioski

W problemie wczesnej detekcji cukrzycy typu 2 najlepszym wyborem spośród analizowanych modeli jest drzewo decyzyjne C5.0 w podstawowej wersji, bez modyfikacji wag błędów klasyfikacji. Model ten uzyskał najwyższy F1-score, co świadczy o najlepszej równowadze między skutecznością wykrywania choroby a ograniczeniem liczby niepotrzebnych skierowań na dalszą diagnostykę. Wysoka czułość na poziomie 84,38% oznacza, że model skutecznie identyfikuje pacjentki rzeczywiście chore, natomiast precyzja wynosząca 69,23% pozwala ograniczyć liczbę fałszywych alarmów, co ma istotne znaczenie z perspektywy kosztów i komfortu pacjentek.

Choć model C5.0 z pięciokrotnie zwiększoną wagą błędów fałszywie negatywnych osiąga jeszcze wyższą czułość na poziomie 93,75 %, towarzyszy temu zauważalny spadek precyzji do 58,82%, co wiąże się z większą liczbą niepotrzebnych skierowań zdrowych osób na kosztowne badania. Taki kompromis może być nieopłacalny i nieefektywny w praktyce. Z tego względu klasyczna wersja modelu C5.0 stanowi najbardziej wyważone i praktyczne rozwiązanie w kontekście wczesnego wykrywania cukrzycy typu 2.