Sprawozdanie z projektu

Autorzy: Dorota Perenc, Nikola Marzec, Viktoria Krettek

Celem naszego projektu jest klasyfikacja, czy dany pacjent choruje na cukrzycę.

1. Opis danych

Zbiór danych wykorzystany w analizie pochodzi z pliku “Dataset of Diabetes.csv”, który został pobrany ze strony : Kaggle - diabetes dataset. Zbiór składa się z 1000 obserwacji oraz 12 zmiennych. Dane zawierają informacje medyczne oraz demograficzne dotyczące osób sklasyfikowanych według stanu cukrzycowego na trzy klasy:

- Non-diabetic (N) – osoby zdrowe (brak cukrzycy),
- Diabetic (Y) – osoby chore na cukrzycę,
- Prediabetic (P) – osoby w stanie przedcukrzycowym.

Zmienna objaśniana

- CLASS – zmienna kategoryczna, określająca stan zdrowia pacjenta.

Zmiennie objaśniające

Zbiór zawiera 12 zmiennych numerycznych i jedną zmienną kategoryczną factor. Są to następujące parametry:

- ID - unikalny identyfikator dla każdego rekordu
- No_Pation - inny identyfikator pacjenta
- Gender - płeć osoby badanej (F, M), dane zostały zaklasyfikowane binarnie
- Age - wiek osoby badanej (lata)
- Urea - poziom mocznika we krwi (mg/dL)
- Cr - poziom kreatyniny (mg/dL)
- HbA1c - poziom hemoglobiny glikowanej (%), jedno z kluczowych badań w diagnozie cukrzycy
- Chol - poziom cholesterolu całkowitego we krwi (mg/dL)
- TG - poziom trójglicerydów (mg/dL)
- HDL - poziom lipoprotein o wysokiej gęstości (dobry cholesterol) (mg/dL)
- LDL - poziom lipoprotein o niskiej gęstości (zły cholesterol) (mg/dL)
- VLDL - poziom lipoprotein o bardzo niskiej gęstości (mg/dL)
- BMI - wskaźnik masy ciała, zmienna kluczowa dla analizy otyłości i zespołu metabolicznego

Podsumowanie dla zmiennych numerycznych:

ID	No_Pation	Gender	AGE	Urea	Cr	HbA1c
Min. : 1.0	Min. : 123	Length:1000	Min. :20.00	Min. : 0.500	Min. : 6.00	Min. : 0.900
1st Qu.:125.8	1st Qu.: 24064	Class :character	1st Qu.:51.00	1st Qu.: 3.700	1st Qu.: 48.00	1st Qu.: 6.500
Median :300.5	Median : 34396	Mode :character	Median :55.00	Median : 4.600	Median : 60.00	Median : 8.000
Mean :340.5	Mean : 270551		Mean :53.53	Mean : 5.125	Mean : 68.94	Mean : 8.281
3rd Qu.:550.2	3rd Qu.: 45384		3rd Qu.:59.00	3rd Qu.: 5.700	3rd Qu.: 73.00	3rd Qu.:10.200
Max. :800.0	Max. :75435657		Max. :79.00	Max. :38.900	Max. :800.00	Max. :16.000

Chol	TG	HDL	LDL	VLDL	BMI	CLASS
Min. : 0.000	Min. : 0.30	Min. :0.200	Min. :0.30	Min. : 0.100	Min. :19.00	Length:1000
1st Qu.: 4.000	1st Qu.: 1.50	1st Qu.:0.900	1st Qu.:1.80	1st Qu.: 0.700	1st Qu.:26.00	Class :character
Median : 4.800	Median : 2.00	Median :1.100	Median :2.50	Median : 0.900	Median :30.00	Mode :character
Mean : 4.863	Mean : 2.35	Mean :1.205	Mean :2.61	Mean : 1.855	Mean :29.58	
3rd Qu.: 5.600	3rd Qu.: 2.90	3rd Qu.:1.300	3rd Qu.:3.30	3rd Qu.: 1.500	3rd Qu.:33.00	
Max. :10.300	Max. :13.80	Max. :9.900	Max. :9.90	Max. :35.000	Max. :47.75	

Zmienne wykazują szeroki zakres wartości, zarówno parametry metaboliczne jak i BMI - co sugeruje, że dane pochodzą od osób o bardzo zróżnicowanym stanie zdrowia.

2. Przygotowanie danych

W celu przygotowania danych do analizy i budowy modeli klasyfikacyjnych wykonujemy kroki mające na celu oczyszczenie danych, ujednolicenie ich formatu oraz przystosowanie do wymagań algorytmów uczenia maszynowego. Wykonujemy następujące czynności:

- Kodujemy binarnie zmienną Gender (F-1, M-0)
- Usuwamy zmienne ID oraz No_Pation, które nie mają wpływu na diagnozę pacjenta
- Kodujemy zmienną CLASS jako factor z trzema poziomami
- Sprawdzamy czy wszystkie wiersze są kompletne
- Usuwamy rekordy zawierające wartości NA
- Normalizujemy zmienne wejściowe do przedziału [0,1], aby ujednolicić skalę zmiennych liczbowych
- Wykonujemy dodatkową standaryzację z-score (na etapie dopracowywania modelu przy metodzie k-NN)
- Dzielimy dane losowo (za pomocą set.seed()) na zbiory: testowy - zawierający 195 obserwacji i treningowy - 800 obserwacji
- Sprawdzamy czy odsetek przypadków cukrzyków w zbiorze treningowym i testowym jest podobny.

Podsumowanie dla znormalizowanych zmiennych numerycznych:

```
     Gender            AGE              Urea               Cr              HbA1c             Chol        
 Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.5254   1st Qu.:0.08333   1st Qu.:0.05290   1st Qu.:0.3709   1st Qu.:0.3883  
 Median :0.0000   Median :0.5932   Median :0.10677   Median :0.06801   Median :0.4702   Median :0.4660  
 Mean   :0.4362   Mean   :0.5696   Mean   :0.12040   Mean   :0.07927   Mean   :0.4890   Mean   :0.4724  
 3rd Qu.:1.0000   3rd Qu.:0.6610   3rd Qu.:0.13542   3rd Qu.:0.08438   3rd Qu.:0.6159   3rd Qu.:0.5437  
 Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000  
       TG               HDL               LDL              VLDL              BMI        
 Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000  
 1st Qu.:0.08889   1st Qu.:0.07216   1st Qu.:0.1562   1st Qu.:0.01719   1st Qu.:0.2435  
 Median :0.12593   Median :0.09278   Median :0.2292   Median :0.02292   Median :0.3826  
 Mean   :0.15205   Mean   :0.10364   Mean   :0.2406   Mean   :0.04942   Mean   :0.3674  
 3rd Qu.:0.19259   3rd Qu.:0.11340   3rd Qu.:0.3125   3rd Qu.:0.04011   3rd Qu.:0.4870  
 Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000  
```

Upewnienie się, że odsetki chorych w zbiorach treningowym i testowym są podobne:

```
diabetes_train_labels
Non-diabetic     Diabetic  Prediabetic 
      0.1025       0.8450       0.0525 
```

```
diabetes_test_labels
Non-diabetic     Diabetic  Prediabetic 
  0.10256410   0.84102564   0.05641026 
```

3. Model klasyfikacyjny k-NN

Metoda k-Nearest Neighbors jest algorytmem klasyfikacji, który przypisuje etykietę nowej obserwacji na podstawie jej przynależności do klasy większościowej spośród jej k najbliższych sąsiadów w przestrzeni cech. W naszym projekcie k-NN został przetestowany w różnych konfiguracjach, by sprawdzić, które parametry dają najlepsze wyniki dla klasyfikacji cukrzycy.
Algorytm k-NN jest oparty na odległościach, więc dane zostały odpowiednio przeskalowane. Porównamy normalizację i standaryzację z-score z różnymi parametrami k.

- Klasyfikacja k-NN z normalizacją oraz parametrem k=29:

	Predicted Non-Diabetic	Predicted Diabetic	Predicted Prediabetic
Actual Non-Diabetic	17	3	0
Actual Diabetic	4	158	2
Actual Prediabetic	3	5	3

Klasa “Diabetic” została rozpoznana z dokładnością 96%: 158 na 164 przypadki zostało sklasyfikowanych poprawnie.
Klasa “Non-Diabetic” została rozpoznana z dokładnością 85%: 17 na 20 przypadków zostało zaklasyfikowanych dobrze.
Klasa “Prediabetic” została rozpoznana z dokładnością 27%: 3 na 11 przypadków zostały sklasyfikowane poprawnie. Najczęściej błędnie przypisywana jako “Diabetic”.
Dokładność algorytmu to około 91,28%.

- Klasyfikacja k-NN z normalizacją oraz parametrem k=27:

	Predicted Non-Diabetic	Predicted Diabetic	Predicted Prediabetic
Actual Non-Diabetic	17	3	0
Actual Diabetic	3	159	2
Actual Prediabetic	3	4	4

Klasa “Diabetic” została rozpoznana z dokładnością 97%: 159 na 164 przypadki zostało sklasyfikowanych poprawnie.
Klasa “Non-Diabetic” została rozpoznana z dokładnością 85%: 17 na 20 przypadków zostało zaklasyfikowanych dobrze.
Klasa “Prediabetic” została rozpoznana z dokładnością 36%: 4 na 11 przypadków zostały sklasyfikowane poprawnie.
Dokładność algorytmu to około 92,31%.

- Klasyfikacja k-NN z normalizacją oraz parametrem k=3:

	Predicted Non-Diabetic	Predicted Diabetic	Predicted Prediabetic
Actual Non-Diabetic	16	1	3
Actual Diabetic	3	158	3
Actual Prediabetic	1	3	7

Klasa “Diabetic” została rozpoznana z dokładnością 96%: 158 na 164 przypadków zostało sklasyfikowanych poprawnie.

Klasa “Non-Diabetic” została rozpoznana z dokładnością 80%: 16 na 20 przypadków zostało zaklasyfikowanych dobrze.
Klasa “Prediabetic” została rozpoznana z dokładnością 64%: 7 na 11 przypadków zostało sklasyfikowanych poprawnie.
Dokładność algorytmu to około 92,82%.

- Klasyfikacja k-NN ze standaryzacją z-score oraz parametrem k=15:

	Predicted Non-Diabetic	Predicted Diabetic	Predicted Prediabetic
Actual Non-Diabetic	15	5	0
Actual Diabetic	3	158	3
Actual Prediabetic	2	5	4

Klasa “Diabetic” została rozpoznana z dokładnością 96%: 158 na 164 przypadków zostało sklasyfikowanych poprawnie.
Klasa “Non-Diabetic” została rozpoznana z dokładnością 75%: 15 na 20 przypadków zostało zaklasyfikowanych dobrze.
Klasa “Prediabetic” została rozpoznana z dokładnością 36%: 4 na 11 przypadków zostały sklasyfikowane poprawnie.
Dokładność algorytmu to około 90,77%.

- Klasyfikacja k-NN ze standaryzacją z-score oraz parametrem k=3:

	Predicted Non-Diabetic	Predicted Diabetic	Predicted Prediabetic
Actual Non-Diabetic	15	2	3
Actual Diabetic	3	159	2
Actual Prediabetic	0	2	9

Klasa “Diabetic” została rozpoznana z dokładnością 97%: 159 na 164 przypadków zostało sklasyfikowanych poprawnie.
Klasa “Non-Diabetic” została rozpoznana z dokładnością 75%: 15 na 20 przypadków zostało zaklasyfikowanych dobrze.
Klasa “Prediabetic” została rozpoznana z dokładnością 82%: 9 na 11 przypadków zostało sklasyfikowanych poprawnie.
Dokładność algorytmu to około 93,85%.

- Rozważyliśmy łącznie 5 konfiguracji. Najwyższą dokładność algorytm osiągnął przy standaryzacji z-score i parametrze k=3. Klasa “Diabetic” była sklasyfikowana bardzo dobrze przez wszystkie metody, zawsze osiągała dokładność powyżej 95%. Klasa “Non-Diabetic” osiągnęła maksymalnie 85% dokładności z normalizacją i k=29. Klasa “Prediabetic” sprawiła najwięcej trudności - wiele przypadków zostało błędnie przypisanych do klasy “Diabetic”, ale przy standaryzacji z-score i k =3 osiąga dokładność równą 82%.

4. Klasyfikacja metodą Naive Bayes

Klasyfikator Naive Bayes to probabilistyczna metoda uczenia maszynowego, oparta na twierdzeniu Bayesa z założeniem niezależności atrybutów. Dla każdego obiektu wyliczane jest prawdopodobieństwo przynależności do każdej z klas na podstawie wartości jego cech. Obiekt zostaje przypisany do klasy z najwyższym prawdopodobieństwem.
W analizie zastosowaliśmy klasyczny klasyfikator Naive Bayes dostępny w pakiecie e1071, a następnie wygładzenie Laplace'a w celu poprawy modelu.

- Klasyfikacja Naive Bayes bez wygładzenia:

	Predicted Non-Diabetic	Predicted Diabetic	Predicted Prediabetic
Actual Non-Diabetic	18	2	0
Actual Diabetic	8	155	1
Actual Prediabetic	0	2	9

Klasa "Diabetic" została rozpoznana z dokładnością 95%: 155 na 164 przypadków zostało sklasyfikowanych poprawnie.
Klasa "Non-Diabetic" została rozpoznana z dokładnością 90%: 18 na 20 przypadków zostało zaklasyfikowanych dobrze.
Klasa "Prediabetic" została rozpoznana z dokładnością 82%: 9 na 11 przypadków zostało sklasyfikowanych poprawnie.
Dokładność algorytmu to około 93,33%.

- Klasyfikacja Naive Bayes z wygładzeniem Laplace'a i parametrem laplace=1:

	Predicted Non-Diabetic	Predicted Diabetic	Predicted Prediabetic
Actual Non-Diabetic	18	2	0
Actual Diabetic	8	155	1
Actual Prediabetic	0	2	9

Klasa "Diabetic" została rozpoznana z dokładnością 95%: 155 na 164 przypadki zostało sklasyfikowanych poprawnie.
Klasa "Non-Diabetic" została rozpoznana z dokładnością 90%: 18 na 20 przypadków zostało zaklasyfikowanych dobrze.
Klasa "Prediabetic" została rozpoznana z dokładnością 82%: 9 na 11 przypadków zostało sklasyfikowanych poprawnie.
Dokładność algorytmu to około 93,33%.

- Klasyfikacja Naive Bayes z wygładzeniem Laplace'a i parametrem laplace=0.5:

	Predicted Non-Diabetic	Predicted Diabetic	Predicted Prediabetic
Actual Non-Diabetic	18	2	0
Actual Diabetic	8	155	1
Actual Prediabetic	0	2	9

Klasa "Diabetic" została rozpoznana z dokładnością 95%: 155 na 164 przypadków zostało sklasyfikowanych poprawnie.
Klasa "Non-Diabetic" została rozpoznana z dokładnością 90%: 18 na 20 przypadków zostało zaklasyfikowanych dobrze.
Klasa "Prediabetic" została rozpoznana z dokładnością 82%: 9 na 11 przypadków zostało sklasyfikowanych poprawnie.
Dokładność algorytmu to około 93,33%.

- Dla klasyfikacji Naive Bayes bez wygładzenia oraz klasyfikacji z parametrami wygładzenia Laplace'a równymi 0,5 oraz 1 zostały uzyskane identyczne macierze tabele krzyżowe. Modyfikacja parametru nie wpłynęła więc w żaden sposób na proces klasyfikacji - model za każdym razem podjął identyczne decyzje co do przypisania rekordów do klas. Powodem tego może być: duża liczebność danych treningowych w każdej klasie, brak zerowych wartości prawdopodobieństw a posteriori dla żadnej z klas lub stabilny rozkład cech w zbiorze.
W rezultacie wszystkie modele osiągnęły dokładność ogólną na poziomie 93,33%.

5. Klasyfikacja metodą drzew decyzyjnych

C5.0 to algorytm klasyfikacyjny oparty na drzewach decyzyjnych. Tworzy on drzewo decyzyjne, gdzie każdy węzeł dzieli dane na podstawie wartości jednego z predyktorów, a liście odpowiadają przewidzianej klasie. Metoda działa iteracyjnie, wybierając atrybuty maksymalizujące informację - tzn. zmniejszającą niepewność - entropię. W analizie użyliśmy drzewa decyzyjnego bez boostingu oraz z boostingiem (10 iteracji).

- Drzewo decyzyjne bez boostingu:
Drzewo podstawowe zostało zbudowane na 800 przypadkach i ma 11 węzłów. Do klasyfikacji wykorzystano między innymi takie atrybuty jak: HbA1c, BMI, TG, HDL. Najważniejszymi regułami klasyfikacji są:
    - Jeśli BMI > 25 i HbA1c >6.3 -> "Diabetic"
    - Jeśli BMI <= 25 i HbA1c <= 5.6 i Chol <= 4.9 -> "Non-Diabetic"

Dokładność drzewa na zbiorze treningowym wynosi 99,8%- 2 błędy na 800 przypadków.

	Predicted Non-Diabetic	Predicted Diabetic	Predicted Prediabetic
Actual Non-Diabetic	19	1	0
Actual Diabetic	1	163	0
Actual Prediabetic	0	0	11

Klasa “Diabetic” została rozpoznana z dokładnością 99%: 163 na 164 przypadki zostały sklasyfikowane poprawnie.
Klasa “Non-Diabetic” została rozpoznana z dokładnością 95%: 19 na 20 przypadków zostało zaklasyfikowanych dobrze.
Klasa “Prediabetic” została rozpoznana z dokładnością 100%: 11 na 11 przypadków zostało sklasyfikowanych poprawnie.
Dokładność algorytmu to około 98,97%.

- Drzewo decyzyjne z boostingiem:
  Boosting polega na tworzeniu serii drzew, z których każde jest lepsze od poprzedniego. Model zbudował 10 drzew o średnim rozmiarze 7.8. Drzewa wykorzystywały więcej atrybutów, w tym: LDL, VLDL, HDL itd.

  Na zbiorze treningowym uzyskano dokładność 100%.

	Predicted Non-Diabetic	Predicted Diabetic	Predicted Prediabetic
Actual Non-Diabetic	19	1	0
Actual Diabetic	1	163	0
Actual Prediabetic	0	1	10

  Klasa “Diabetic” została rozpoznana z dokładnością 99%: 163 na 164 przypadki zostały sklasyfikowane poprawnie.
  Klasa “Non-Diabetic” została rozpoznana z dokładnością 95%: 19 na 20 przypadków zostało zaklasyfikowanych dobrze.
  Klasa “Prediabetic” została rozpoznana z dokładnością 91%: 10 na 11 przypadków zostało sklasyfikowanych poprawnie.
  Dokładność algorytmu to około 98,46%.

- Przypisanie kar różnym rodzajom błędów - stworzenie macierzy kosztów
  Nadajemy odpowiednie kary za dane błędy:
  - Błędnego zaklasyfikowania “Diabetic” jako “Non-Diabetic”- koszt=3
  - Błędnego zaklasyfikowania “Prediabetic” jako “Diabetic”- koszt=2
  - Błędnego zaklasyfikowania “Prediabetic” jako “Non-Diabetic”- koszt=1

  Model zoptymalizowany pod kątem opisanej macierzy dał wynik:

	Predicted Non-Diabetic	Predicted Diabetic	Predicted Prediabetic
Actual Non-Diabetic	19	1	0
Actual Diabetic	0	164	0
Actual Prediabetic	0	0	11

Klasa “Diabetic” została rozpoznana z dokładnością 100%: 164 na 164 przypadki zostały sklasyfikowane poprawnie.
Klasa “Non-Diabetic” została rozpoznana z dokładnością 95%: 19 na 20 przypadków zostało zaklasyfikowanych dobrze.
Klasa “Prediabetic” została rozpoznana z dokładnością 100%: 11 na 11 przypadków zostało sklasyfikowanych poprawnie.
Dokładność algorytmu to około 99,49%.

- Porównaliśmy trzy podejścia oparte na algorytmie C5.0. Najlepszy wynik uzyskała klasyfikacja kosztowa zoptymalizowana względem najistotniejszych błędów klasyfikacji. Dzięki zdefiniowaniu macierzy kosztów algorytm uzyskał najlepszy wynik a wszystkie klasy zostały rozpoznane z wysoką skutecznością.

6. Podsumowanie wyników z rozważanych modeli klasyfikacji

W przeprowadzonej analizie porównaliśmy kilka metod klasyfikacyjnych w celu przewidywania stanu zdrowia pacjenta w kontekście cukrzycy: k-NN, Naive Bayes oraz drzewa decyzyjne C5.0.

Metoda k-NN okazała się wrażliwa na dobór parametru k oraz sposób przetwarzania danych. Najlepsze wyniki uzyskano po standaryzacji z-score zmiennych przy parametrze k = 3 - osiągnięto ogólną dokładność ok. 93,85%, a szczególnie dobrze została rozpoznana klasa “Prediabetic” (82%).

Naive Bayes wykazał zaskakująco stabilne wyniki - niezależnie od zastosowanego wygładzania Laplace’a, wyniki klasyfikacji pozostały identyczne. Ogólna dokładność klasyfikacji wyniosła 93,33%, z bardzo dobrą skutecznością rozpoznawania klasy “Diabetic” (95%) i “Prediabetic” (82%). Metoda ta ma jednak tendencję do zbyt silnego upraszczania zależności pomiędzy cechami, co może wpływać na błędne przypisania.

Najwyższą skutecznością wykazały się drzewa decyzyjne C5.0. Klasyczne drzewo bez boostingu osiągnęło dokładność 98,97%, a drzewo z boostingiem – 98,46%, co świadczy o ich wysokim potencjale predykcyjnym. Boosting poprawił klasyfikację trudniejszych przypadków kosztem większej złożoności modelu.
Najlepsze rezultaty przyniosła klasyfikacja kosztowa, która uzyskała dokładność 99,49%. Dzięki zastosowaniu macierzy kosztów błędów (przypisując wyższy koszt najbardziej niepożądanym pomyłkom), model ten optymalizuje klasyfikację pod kątem bezpieczeństwa pacjenta. Wszystkie przypadki z klasy “Prediabetic” i “Diabetic” zostały sklasyfikowane poprawnie (100%), a klasa “Non-Diabetic” z 95% skutecznością.
To czyni ten model najbardziej efektywnym pod względem precyzji i bezpieczeństwa diagnostycznego - minimalizuje on ryzyko błędnej klasyfikacji przypadków wymagających interwencji medycznej.