Predykcja cukrzycy przy użyciu metod Machine Learning

Aleksander Brandt, Daria Bartkowiak

28.04.2025

1 Cel biznesowy i problemowy

Celem projektu jest opracowanie modelu predykcyjnego do wczesnego wykrywania cukrzycy na podstawie danych demograficznych, biomedycznych i stylu życia pacjentów. Model pozwoli na szybszą identyfikację osób z wysokim ryzykiem cukrzycy, umożliwiając wcześniejsze interwencje medyczne, poprawę jakości życia pacjentów oraz optymalizację kosztów opieki zdrowotnej.
Do oceny skuteczności modelu przewidującego cukrzycę wykorzystano trzy kluczowe miary: Recall, aby minimalizować ryzyko przeoczenia chorych pacjentów, F1-score, aby zrównoważyć czułość i precyzję, oraz AUC-ROC, mierzące zdolność modelu do rozróżniania pacjentów z cukrzycą i bez niej. Dzięki temu zapewniono wysoką skuteczność wykrywania cukrzycy przy jednoczesnym ograniczeniu błędnych diagnoz.

2 Opis danych

Dane wykorzystane w projekcie pochodzą ze zbioru **diabetes-clinical-dataset100k-rows**, który zawiera około 100 tysięcy rekordów pacjentów z informacjami klinicznymi dotyczącymi zdrowia i stylu życia. Zmiennymi wejściowymi były m.in.:

- Rok rejestracji danych (**year**),
- Płeć (**gender**),
- Wiek (**age**),
- Lokalizacja geograficzna (**location**),
- Przynależność rasowa (**race**),
- Historia nadciśnienia (**hypertension**),
- Historia choroby serca (**heart_disease**),
- Historia palenia tytoniu (**smoking_history**),
- Wskaźnik masy ciała (**bmi**),
- Poziom hemoglobiny glikowanej (**hbA1c_level**),

- Poziom glukozy we krwi (**blood_glucose_level**),
- Notatki kliniczne (**clinical_notes**).

Zmienną docelową (*target*) była diagnoza cukrzycy (**diabetes**), oznaczona jako 0 (brak cukrzycy) lub 1 (obecność cukrzycy).

3 Eksploracyjna analiza danych

W ramach analizy eksploracyjnej zbadano rozkłady kluczowych zmiennych oraz ich wzajemne korelacje. Tutaj kilka przykładowych wykresów, które udało się stworzyć:

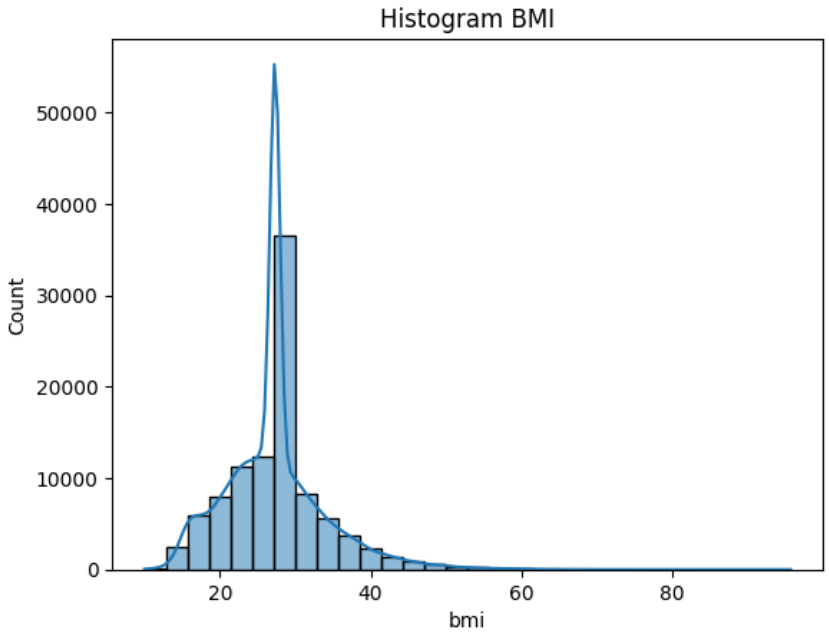


Figure 1: Rozkład wartości BMI w zbiorze danych.

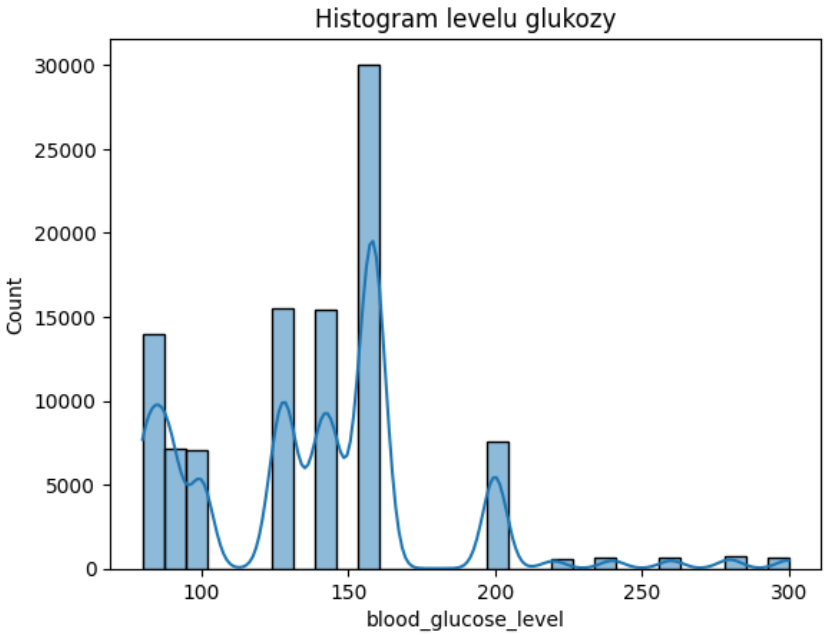


Figure 2: Rozkład poziomu glukozy we krwi.

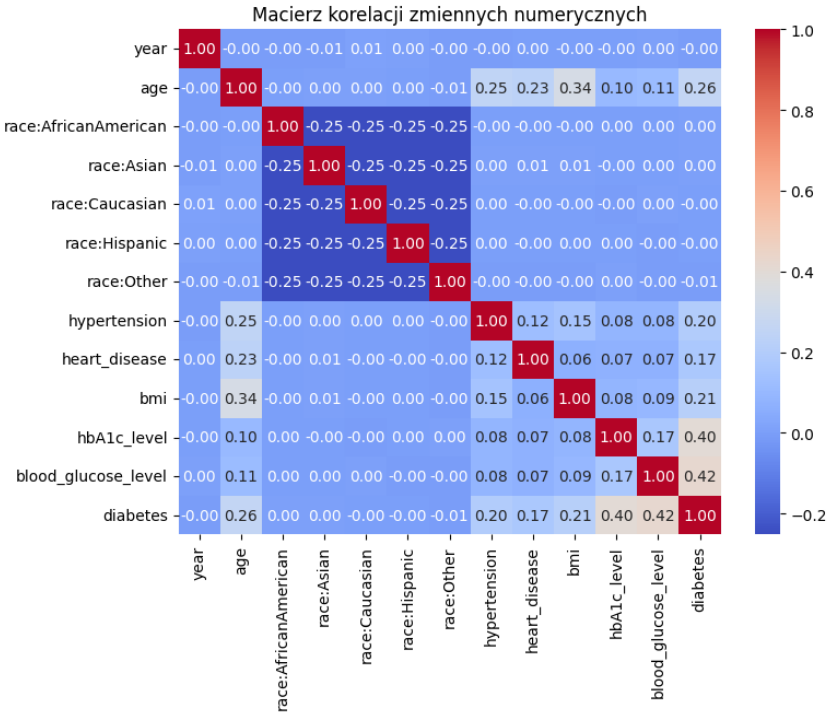


Figure 3: Macierz korelacji między zmiennymi.

Analiza korelacji wykazała, że zmienne takie jak **hbA1c_level** i **blood_glucose_level** mają istotny związek z diagnozą cukrzycy, co potwierdza ich znaczenie w modelowaniu predykcyjnym. Wiek i BMI również mają istotny wpływ – starsze osoby i osoby z wyższą masą ciała są bardziej narażone na cukrzycę. Nadciśnienie i choroby serca wykazują umiarkowaną korelację, ale są powiązane z cukrzycą. Zmienność rasowa oraz rok badań nie mają istotnego wpływu.

4 Wstępne przetwarzanie danych

Podczas przetwarzania danych wykonaliśmy następujące czynności:

- **Oczyszczanie danych**: Usunięto obserwacje, dla których zmienna **gender** miała wartość `Other`, aby zapewnić spójność danych. Brakujące wartości nie były uzupełniane metodami imputacji, a zmienna **clinical-notes** została usunięta jako redundantna i opisowa.

- **Usuwanie wartości odstających**: Dla zmiennej **age** w grupie pacjentów z cukrzycą (`diabetes = 1`) wartości mniejsze niż dolna granica wyznaczona na podstawie rozstępu międzykwartylowego (IQR) zostały zastąpione wartością tej granicy. Dla zmiennej **blood_glucose_level** przeprowadzono analizę wykresu pudełkowego, jednak nie dokonano korekty wartości odstających.

- **Kodowanie zmiennych kategorycznych**: Zmienna **smoking_history** została zakodowana przy użyciu transformacji **Weight of Evidence** (WoE), co umożliwiło uwzględnienie zależności zmiennej kategorycznej od zmiennej docelowej w sposób ilościowy. Następnie oryginalna kolumna **smoking_history** została usunięta.

- **Podział danych**: Dane zostały podzielone na zbiór treningowy (train), walidacyjny (validation) oraz testowy (test) z uwzględnieniem stratyfikacji względem zmiennej docelowej (**diabetes**).

5 Inżynieria cech

Wprowadzono nowe cechy i przekształcenia:

- Utworzono kategorie BMI (niedowaga, prawidłowa masa, nadwaga, otyłość) na podstawie **bmi**. Jednak zmienna ta miała groszą korelację z naszą zmienną docelową, więc zrezygnowaliśmy z tego pomysłu.

- Wzbogacono ramkę danych o szacowany średni poziom glukozy we krwi (**eAG**), obliczany medycznym wzorem eAG = 28.7 × HbA1c - 46.7, który przekłada HbA1c na mg/dL. Dzięki temu uzyskaliśmy miarę długoterminowej glikemii, porównywalną z rzeczywistymi pomiarami glukozy. Wprowadziliśmy też zmienną **glucoseEAGdiff**, pokazującą różnicę między poziomem glukozy a eAG, co pozwala zidentyfikować nietypowe przypadki, jak wysoka glukoza przy niskim HbA1c. Te nowe cechy zwiększają predykcyjność ramki danych w analizie ryzyka cukrzycy.

- Przeprowadzono analizę korelacji Pearsona oraz Spearmana, aby wybrać cechy o wysokim wpływie na predykcję, eliminując mniej istotne (**age** oraz **race...**).

- Dla zmiennej **location** zastosowano podwójną transformację Weight of Evidence (WoE). Najpierw przekształcono kategorie **location** na wartości WoE, odzwierciedlające ich zdolność predykcyjną względem **diabetes**. Następnie te wartości WoE pogrupowano w 5 przedziałów i ponownie zastosowano transformację WoE, tworząc cechę **location-woe-category-woe**. Podział na 5 grup (od 0.0 do 3.0 co 0.6) ułatwia interpretację wpływu lokalizacji, jasno wskazując poziomy ryzyka – od niskiego (niebieskie) do wysokiego (czerwone). Poniższa mapa ukazuje te wartości.

Zastanawialiśmy się nad sprawdzeniem istnienia Diabetes Belt (południowo - zachodnie stany z wyższym odsetkiem cukrzyków), ale mapa WoE nie zgadza się z tym wzorcem, więc zrezygnowaliśmy z tej analizy.

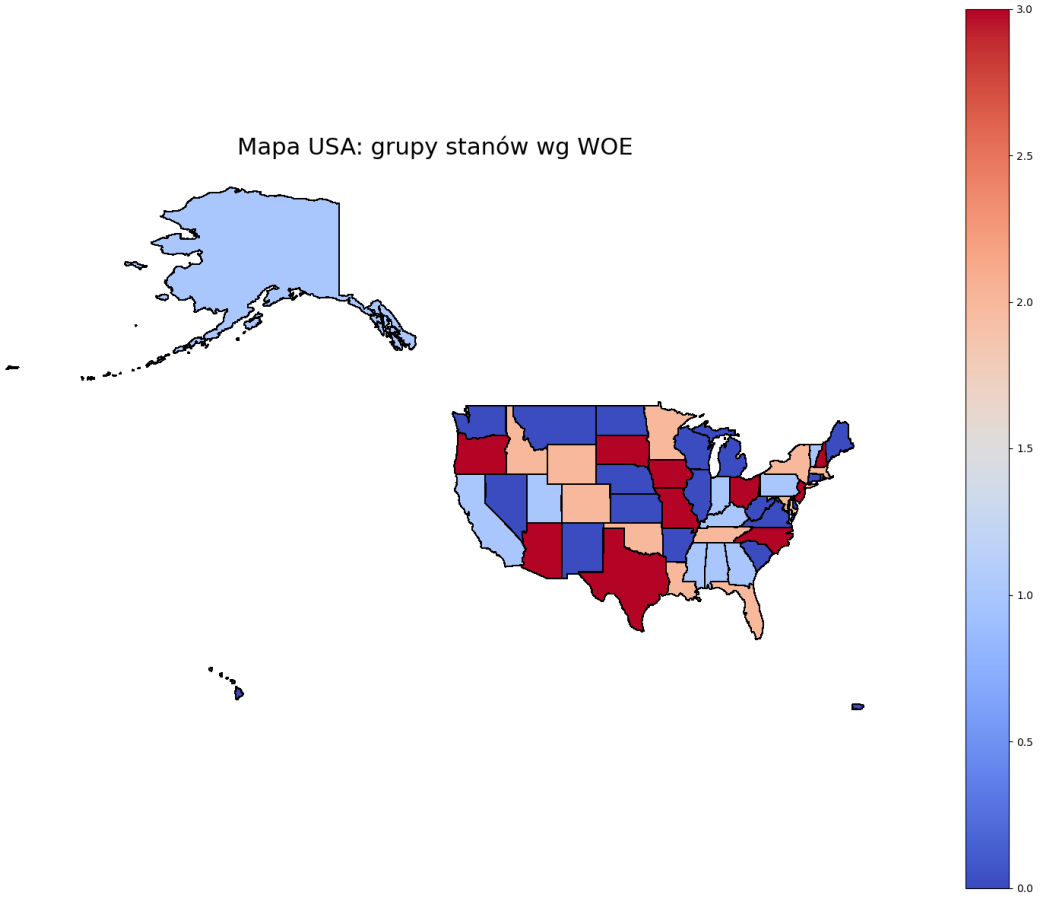


Figure 4: Mapa USA z państwami pokolorowanymi według wartości podwójnego WoE dla cechy **location**. Kolory od niebieskiego (niski WoE) do czerwonego (wysoki WoE) wskazują na różnice w predykcyjnej sile lokalizacji względem diagnozy cukrzycy.

6 Transformacja danych

Aby przygotować dane do trenowania modeli, wykonano dodatkowe transformacje:

- **Standaryzacja zmiennych numerycznych**: Zmienne takie jak **age**, **bmi**, **hbA1c-level** i **blood-glucose-level** zostały przeskalowane do standardowego rozkładu (średnia 0, odchylenie standardowe 1) za pomocą StandardScaler, co zapewniło równomierny wpływ każdej zmiennej na modele.
- **Logarytmowanie zmiennych o skośnym rozkładzie**: Zmienne numeryczne o skośnych rozkładach (**eAG** oraz **blood_glucose_level**) poddano logarytmowaniu. Jednak logarytmowanie pogorszyło korelacje zmiennych ze zmienną 'diabetes' - w tym wypadku zlogarytmowaliśmy tylko zmienną **blood_glucose_level**, która nie ma korelacji liniowej, aby polepszyć działanie niektórych modeli.
- **Balansowanie:**Zbadano rozkład zmiennej docelowej diabetes i stwierdzono nierównowagę klas - pacjenci bez cukrzycy stanowili aż 91,5% wszystkich przypad- ków, co oznacza

stosunek 91,5:8,5 względem osób chorych. Aby temu zaradzić, zas- tosowano kombinację Undersamplingu i Oversamplingu. Najpierw losowo usunięto 12 250 próbek osób bez cukrzycy, a następnie metodą SMOTE wygenerowano 12 250 nowych rekordów pacjentów z cukrzycą. W efekcie uzyskano stosunek chorych do zdrowych na poziomie około 66,5:33,5. Balansowanie przeprowadziliśmy tylko na danych treningowych.

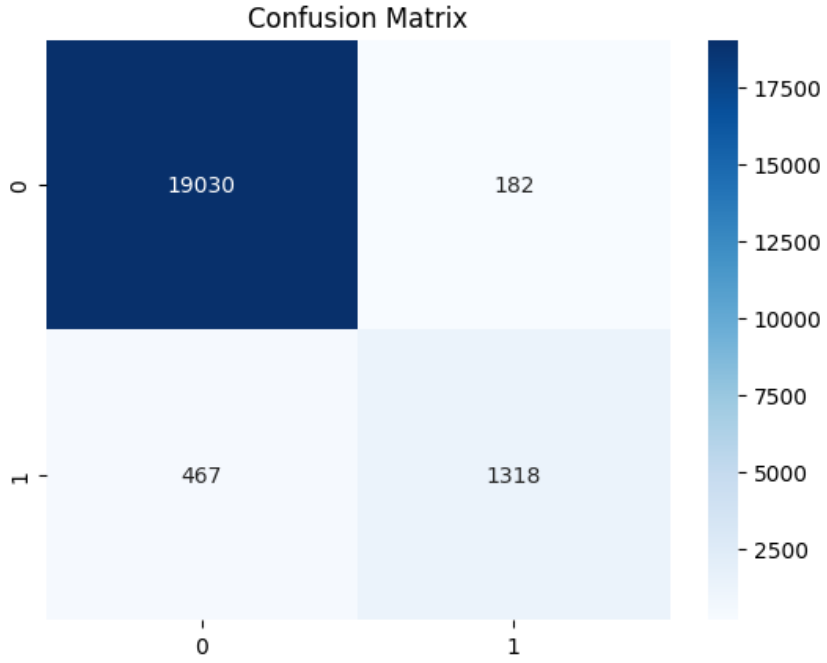
7 Budowa modeli

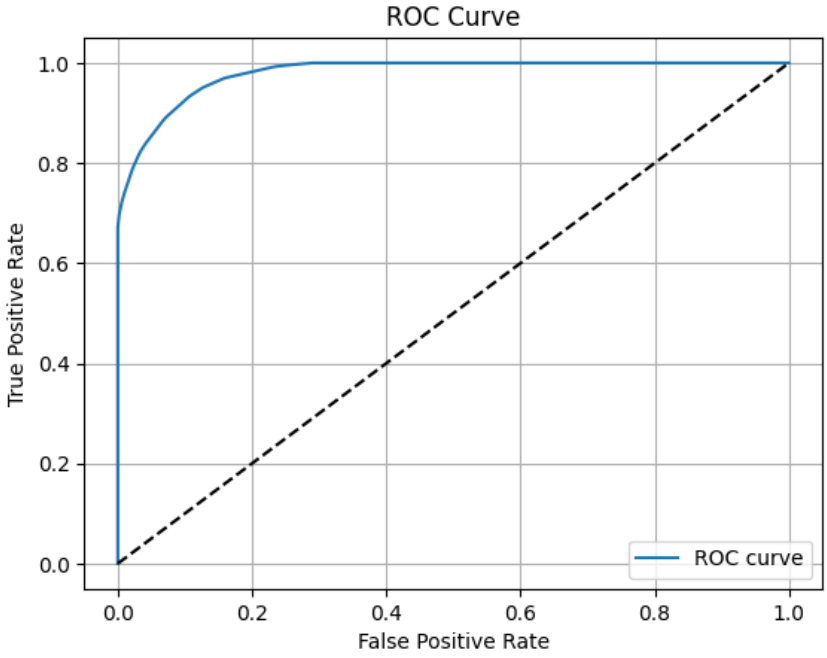
Pracę nad modelami rozpoczęliśmy od analizy modeli bazowych: **regresja logistyczna** oraz **drzewo decyzyjne**. Podczas optymalizacji modeli próbowlaiśmy różnych proporcji podczas balansowania danych, re-weighting oraz różnych wartości threshold. Zarówno model logistyczny jak i drzewa decyzyjnego najlepiej sprawdzał się przy częściowym zbalansowaniu oraz bez zmiany wag poszczególnych klas. Re-weighting zmniejszał nam wartość parametru Recall, na którym najbardziej nam zależy. Zweryfikowaliśmy także, czy model jest przeuczony analizując ROC AUC dla danych treningowych - nie zaobserwowaliśmy nic niepokojącego.
Po przeprowadzeniu wstępnej oceny klasyfikatorów, przeszliśmy do testowania bardziej złożonych modeli uczenia maszynowego. Wśród nich znalazły się drzewa decyzyjne, metody boostowane (takie jak AdaBoost, Gradient Boosting oraz XGBoost), modele liniowe, probabilistyczne, a także klasyfikator SVM. Dla każdego modelu przeanalizowaliśmy wy- niki w kontekście najważniejszych metryk: recall (jako główny priorytet), F1-score oraz ROC AUC. Jednak żaden z nich nie zapewniał jednocześnie satysfakcjonujących wyników we wszystkich ob- szarach. W związku z tym zdecydowaliśmy się na podejście oparte na **stacking ensemble**, które pozwala połączyć predykcje kilku modeli bazowych w jeden, bardziej złożony mo- del nadrzędny. Model stackingowy został zbudowany na podstawie najlepiej rokujących klasyfikatorów wyłonionych w poprzednim etapie analizy, czyli **DecisionTreeClassifier, AdaBoostClassifier, GradientBoostingClassifier, XGBClassifier**.

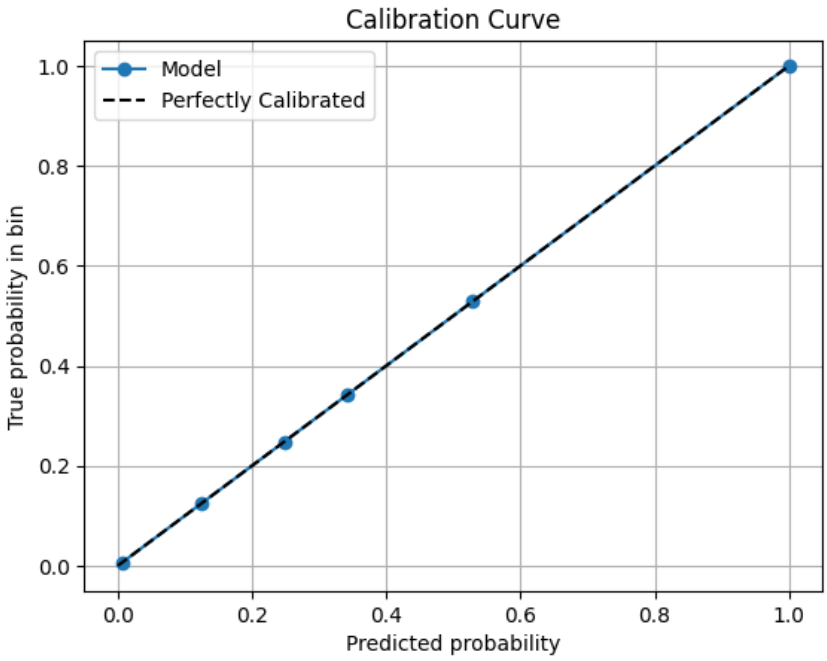
8 Tuning, wyniki i ocena modeli

W celu poprawy skuteczności modelu stackingowego, szczególnie pod kątem metryki recall, przeprowadziliśmy tuning hiperparametrów. Użyliśmy RandomizedSearchCV z walidacją krzyżową, skupiając się na dostrojeniu modeli bazowych oraz meta-modelu. Dodatkowo aby nasz model był interpretowalny poddaliśmy go kalibracji do początkowego stosunku klas z użyciem CalibratedClassifierCV. Otrzymaliśmy następujące wyniki:

1. Recall: 0.7383753501400561
2. F1 Score: 0.8024353120243531
3. AUC-ROC: 0.9795193946827118


Confusion Matrix
19030
182
467
1318
0
1
0
1
17500
15000
12500
10000
7500
5000
2500


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve



Sprawdziliśmy także wariancje naszego modelu:

1. Train Recall: 0.7503
2. Test Recall: 0.7384
3. Różnica Recall (train - test): 0.0119

Błąd modelu na klasie pozytywnej (1) nie różni się znacząco między treningiem a testem. Różnica 0.0119 sugeruje, że model: nie przeucza się (low variance) i nie ma też wyraźnego niedouczenia (low bias). Ddatkowo krzywa kalibracji wskazuje nam, że udało nam się poprawnie skalibrować model. Ogólne wyniki były bardzo zadowalające.

9 Interpretacja

Analiza SHAP wykazała, że cechy **eAG**, **blood_glucose_level** oraz **age** miały największy wpływ na predykcje modelu, co potwierdzono w analizie ważności cech.

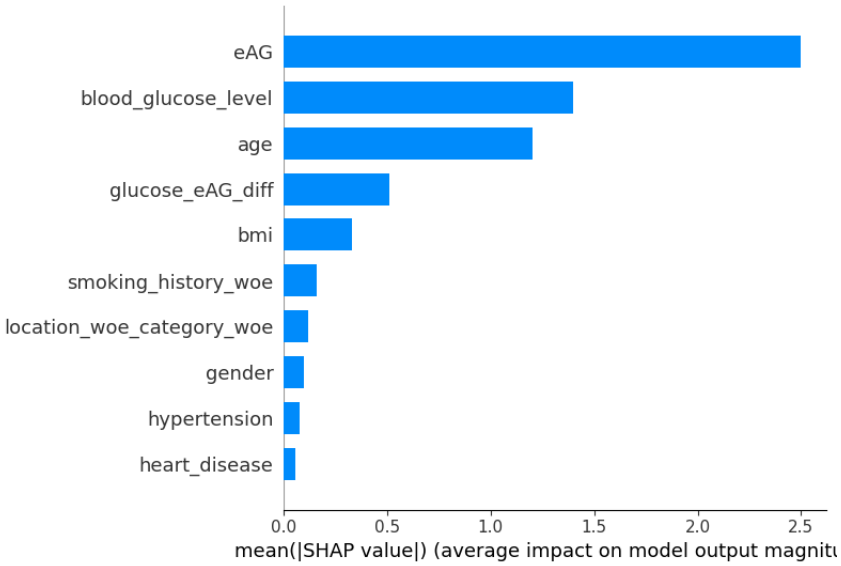


Figure 5: Podsumowanie ważności cech według analizy SHAP.

W przypadku fałszywego negatywu (cukrzyk sklasyfikowany jako zdrowy), cechy takie jak **location_woe_category_woe** oraz niskie wartości **blood_glucose_level** i **eAG** obniżyły predykcję ryzyka, mimo że **heart_disease** i **bmi** wskazywały na pewne zagrożenie, co sugeruje niedoszacowanie wpływu niektórych zmiennych.

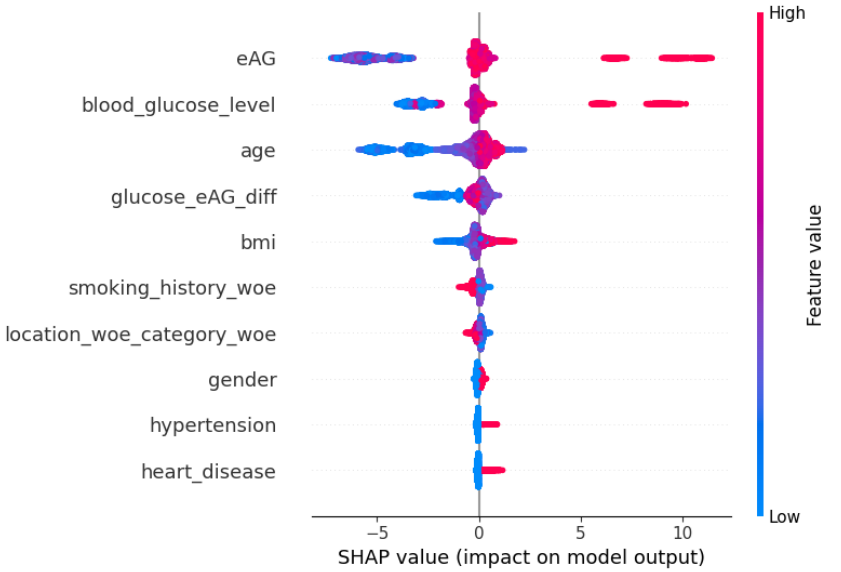


Figure 6: Wykres SHAP dla ogólnej interpretacji modelu.

9.1 Interpretacja decyzyjności modelu

W celu zrozumienia, jak model podejmuje decyzje klasyfikacyjne, przeanalizowano przypadki predykcji za pomocą wykresów SHAP force plot, które pokazują wpływ poszczególnych cech na decyzję modelu w konkretnych przypadkach. Oto przykłady naszych wyników:

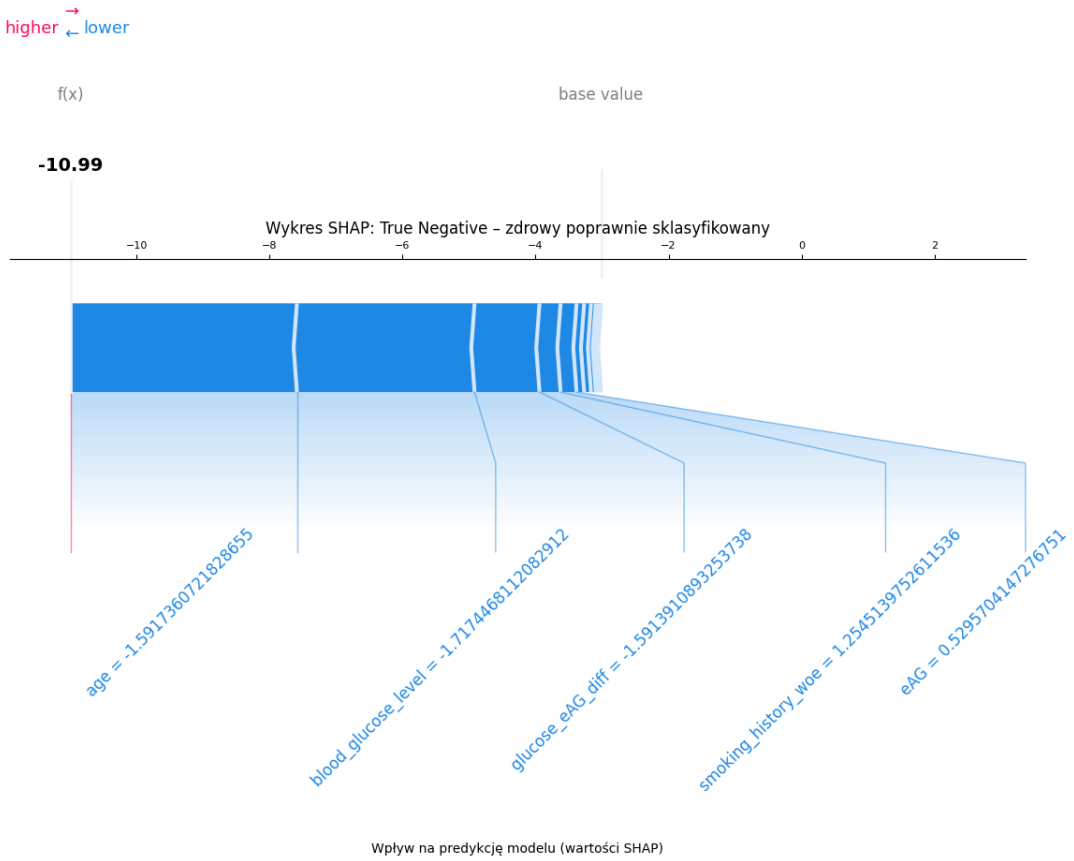


Figure 7: Wykres SHAP dla przypadku True Negative.

Wykres przedstawia przypadek True Negative, gdzie model prawidłowo zaklasyfikował pacjenta jako zdrowego ($f(x) = -10.99$). Cechy takie jak **age**, **blood_glucose_level**, **glucose_eAG_diff**, **smoking_history_woe** oraz **eAG** znacząco obniżają predykcję ryzyka (niebieskie strzałki skierowane w lewo), co potwierdza prawidłowe rozpoznanie braku cukrzycy.

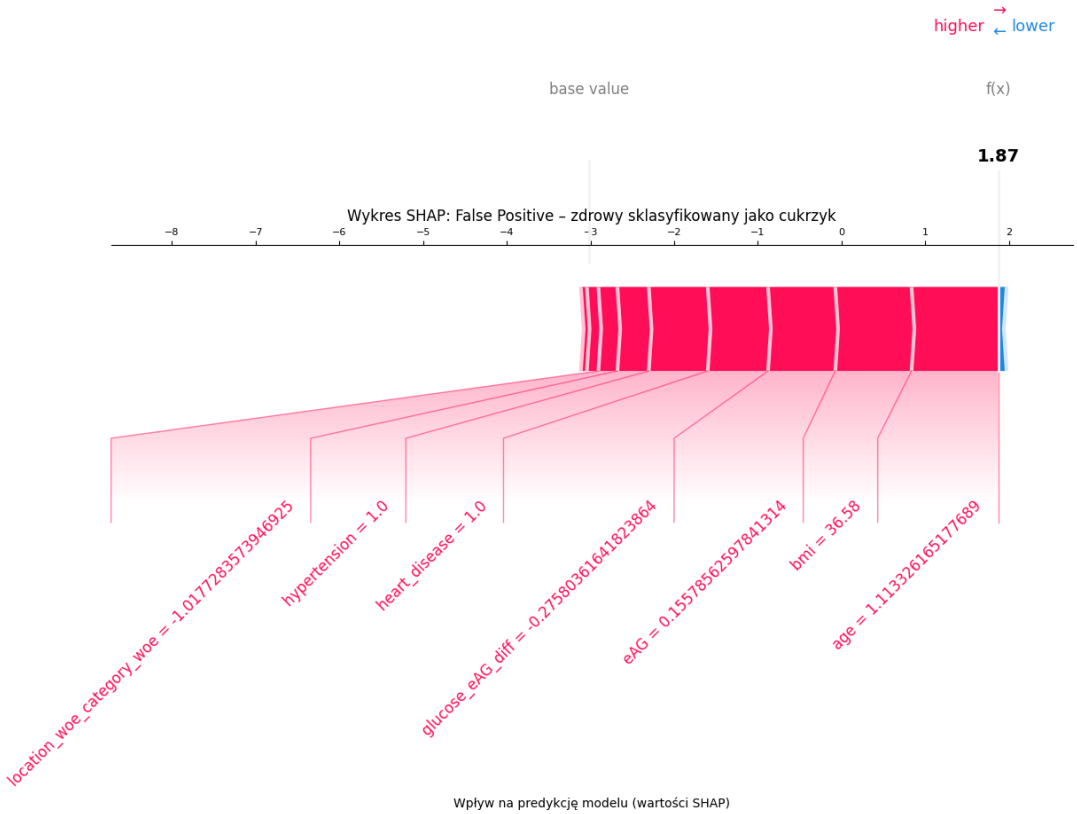


Figure 8: Wykres SHAP dla przypadku False Positive.

Wykres ilustruje przypadek False Positive, gdzie zdrowy pacjent został błędnie sklasyfikowany jako cukrzyk ($f(x) = 1.87$). Kluczowe cechy zwiększające ryzyko to **hypertension** (1.0), **heart_disease** (1.0), wysokie **bmi** (36.58) oraz **age** (czerwone strzałki), które przeważyły nad przeciwnym wpływem **location_woe_category_woe** i **glucose_eAG_diff** (niebieskie strzałki). Sugeruje to, że model może przeceniać wpływ chorób współistniejących w takich przypadkach.

Figure 9: Wykres SHAP dla przypadku True Positive (Próbka 1).

Pierwszy wykres przedstawia przypadek True Positive, gdzie model prawidłowo zaklasyfikował pacjenta jako chorego na cukrzycę ($f(x) = 10.47$). Dominującymi czynnikami są wysoki poziom **blood_glucose_level** (2.7908209594949324) i **eAG** (1.8378117375500914), które znacząco zwiększają predykcję ryzyka (czerwone strzałki skierowane w prawo). Pozostałe cechy mają mniejszy wpływ, co potwierdza kluczową rolę pomiarów glikemii w diagnozie.

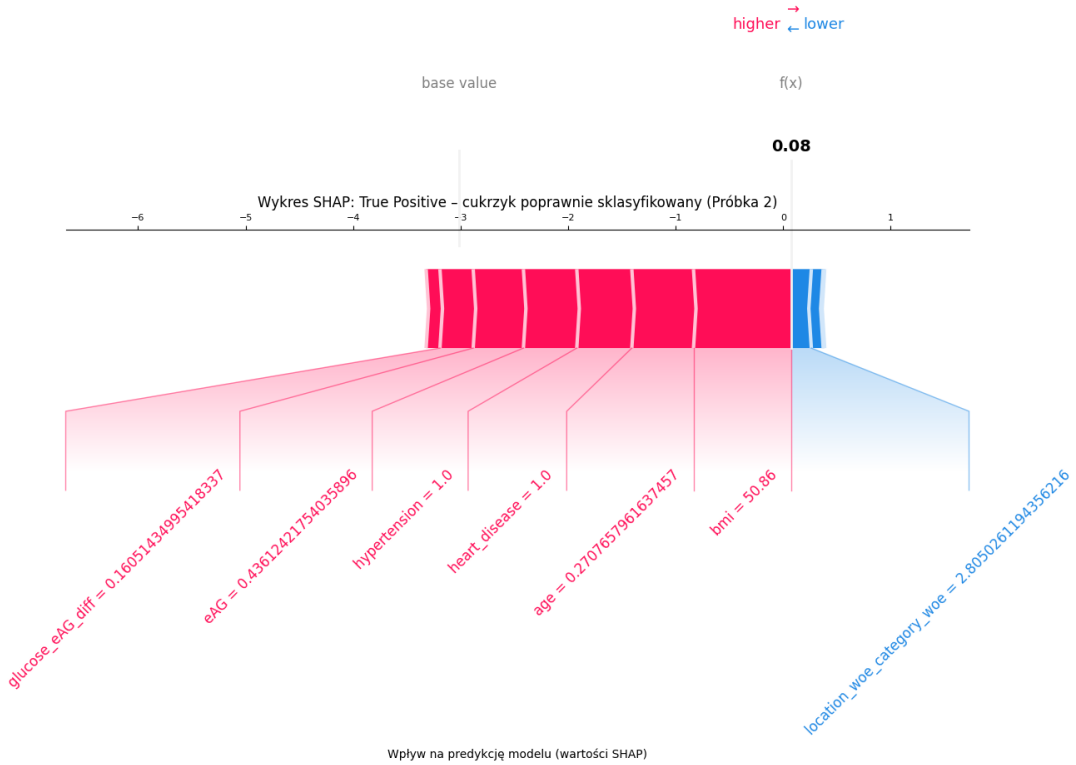


Figure 10: Wykres SHAP dla przypadku True Positive (Próbka 2).

Drugi wykres True Positive ($f(x) = 7.03$) potwierdza skuteczność modelu. Kluczowe cechy to **blood_glucose_level** (1.4078352046889) i **eAG** (1.183697950388824), które zwiększają predykcję (czerwone strzałki). Cechy takie jak **location_woe_category_woe** (-0.8939200597941) nieznacznie obniżają ryzyko (niebieskie strzałki), ale nie zmieniają decyzji modelu.

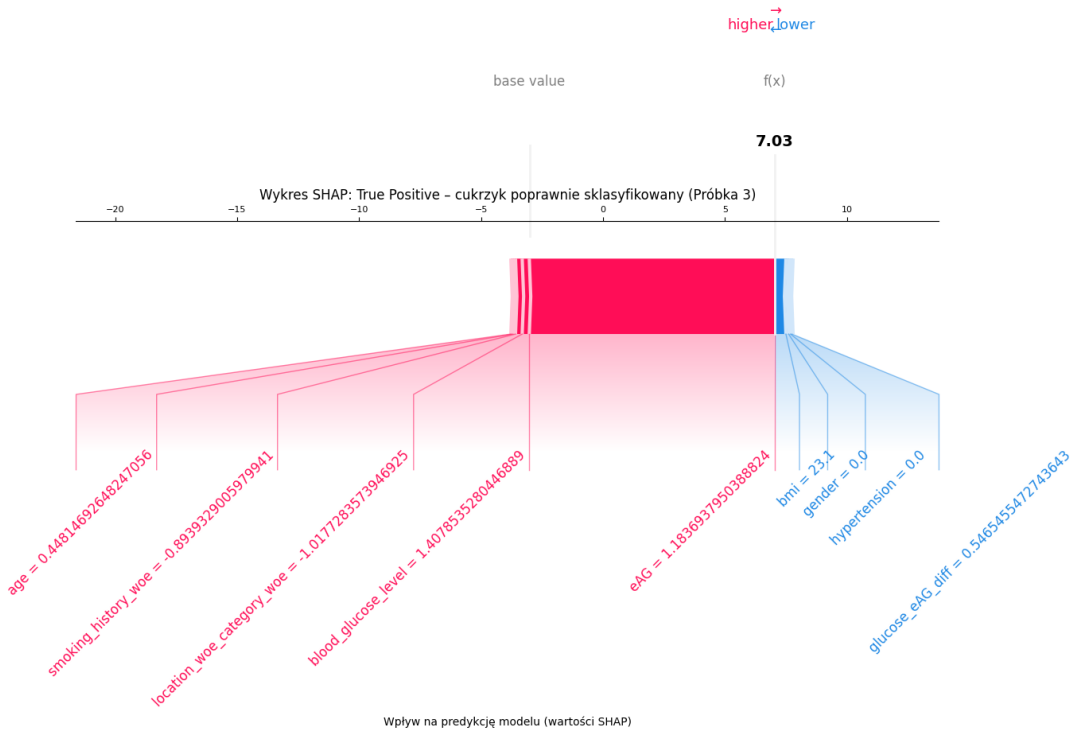


Figure 11: Wykres SHAP dla przypadku True Positive (Próbka 3).

Trzeci wykres True Positive ($f(x) = 9.52$) pokazuje prawidłową klasyfikację. Największy wpływ mają **blood_glucose_level** (2.3027011729644) i **eAG** (2.491490405556129), które zwiększają predykcję (czerwone strzałki). **glucose_eAG_diff** (-0.97066627324765) działa przeciwnie (niebieskie strzałki), ale jego wpływ jest mniejszy.

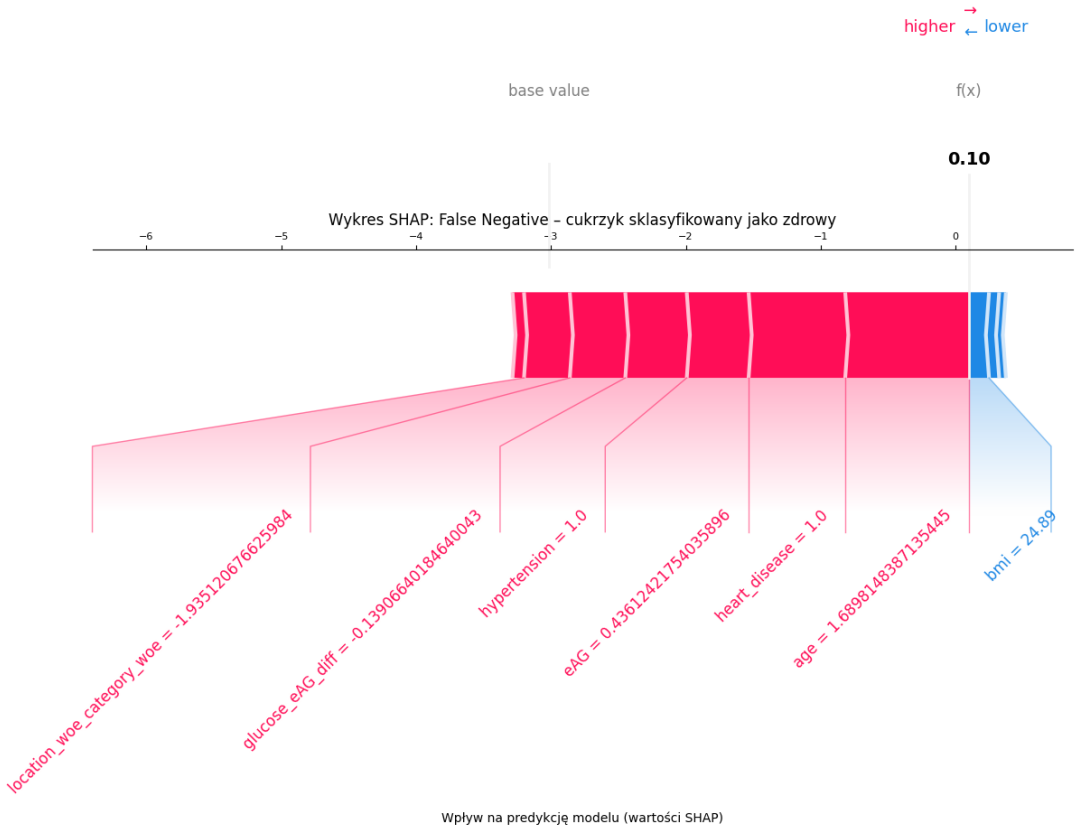


Figure 12: Wykres SHAP dla przypadku False Negative (Próbka 1).

Pierwszy wykres pokazuje przypadek False Negative, gdzie pacjent z cukrzycą został błędnie uznany za zdrowego ($f(x) = 0.10$). Cechy takie jak **location_woe_category_woe** (-1.93512067662984), **glucose_eAG_diff** (-0.13906640184640043) i **eAG** (0.4361242175403589) znacząco obniżają predykcję ryzyka (niebieskie strzałki), mimo że **hypertension** (1.0), **heart_disease** (1.0), **bmi** (24.89) i **age** (1.02307448385761) wskazują na pewne zagrożenie (czerwone strzałki). Wskazuje to na potencjalne niedoszacowanie wpływu niektórych zmiennych przez model.

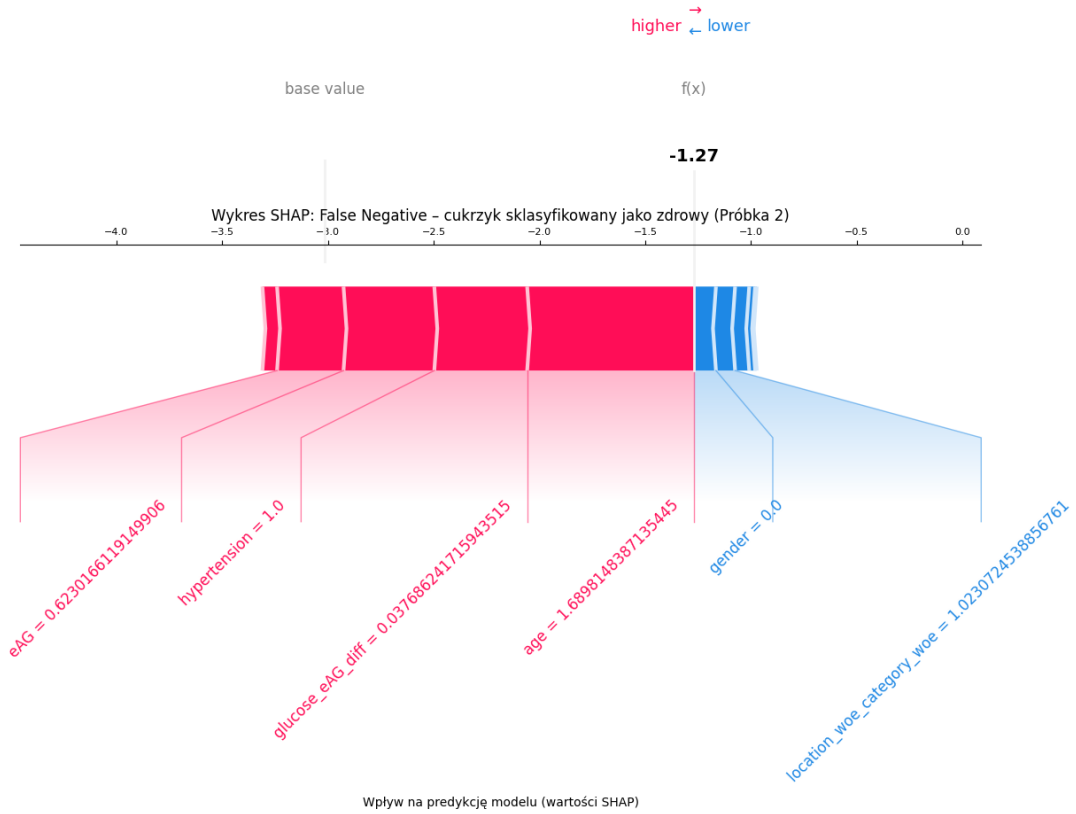


Figure 13: Wykres SHAP dla przypadku False Negative (Próbka 2).

Drugi przypadek False Negative ($f(x) = -1.27$) również pokazuje błędną klasyfikację cukrzyka jako zdrowego. Cechy takie jak **location_woe_category_woe** (-1.23072458385761) i **gender_woe** (-1.024572438586761) obniżają ryzyko (niebieskie strzałki), mimo że **blood_glucose_level** (1.25453197526153) i **bmi** (27.32) wskazują na pewne ryzyko (czerwone strzałki). Model może niedoszacowywać wpływu czynników ryzyka w takich przypadkach.

10 Wyniki walidacji:

Wyniki istotnych metryk:

1. **Recall:** 0.728
2. **F1 Score:** 0.801
3. **AUC ROC:** 0.976

Test modelu na zbiorze testowym wykazał, że model dobrze radzi. Recall różni się o mniej niż 2 punkty procentowe – wydajność na niewidzianym zbiorze testowym odpowiada walidacji.