Ocena zdolności kredytowej klienta na podstawie danych demograficznych i finansowych

Maja Fiszer 223354, Jakub Gilewski 223288,
Dominik Uchman 231032

czerwiec 2025

1

Spis treści

2

1 Streszczenie

W projekcie podjęto próbę oceny zdolności kredytowej klienta na podstawie danych demograficznych i finansowych. Zastosowano trzy klasyczne metody klasyfikacyjne: regresję logistyczną, metodę k-najbliższych sąsiadów (kNN) oraz liniową analizę dyskryminacyjną (LDA). W celu poprawy jakości predykcji zbudowano również model hybrydowy oparty na średniej ważonej wyników uzyskanych z każdej z metod. Projekt obejmował przygotowanie danych, kodowanie zmiennych, podstawowe przekształcenia oraz analizę wyników dla zbioru rzeczywistych obserwacji. Zaprezentowano również przykład działania modeli na danych syntetycznych.

Słowa kluczowe: klasyfikacja, zdolność kredytowa, regresja logistyczna, kNN, LDA, model hybrydowy, dane finansowe, dane demograficzne

2 Wprowadzenie

Ocena zdolności kredytowej stanowi kluczowy element procesu decyzyjnego w bankowości. Na podstawie dostępnych informacji o kliencie bank musi przewidzieć, czy osoba ubiegająca się o kredyt spłaci zobowiązanie w ustalonym terminie. W pracy tej wykorzystujemy 15 atrybutów opisujących klienta: status rachunku, czas trwania kredytu, historię kredytową, cel kredytu, kwotę kredytu, poziom oszczędności, staż zatrudnienia, relację raty do dochodu, status osobisty i płeć, obecność poręczyciela, czas zamieszkania, rodzaj majątku, wiek, liczbę aktualnych kredytów oraz rodzaj pracy. Spośród tych zmiennych wyodrębniono cechy jakościowe i numeryczne.

Głównym celem jest zbudowanie i porównanie czterech modeli klasyfikacyjnych: regresji logistycznej, k-najbliższych sąsiadów (kNN), liniowej analizy dyskryminacyjnej (LDA) oraz modelu hybrydowego (stacking). Do oceny jakości modeli zastosowano podział danych na zbiór treningowy i testowy w proporcji 80/20. Skuteczność klasyfikatorów oceniano za pomocą czterech miar: Accuracy, F1-score, Precision oraz AUC-ROC. W projekcie opisano przygotowanie danych, implementację i przykład działania na obserwacjach syntetycznych.

3 Przedmiot badania

3.1 Cel i zakres badania

Celem pracy jest skonstruowanie modelu klasyfikacyjnego, który na podstawie 15 cech klienta przewidzi, czy otrzyma on decyzję kredytową pozytywną (klasa „1") czy negatywną (klasa „0"). Zbiór danych[1] zawiera $N = 1000$ obserwacji oraz dziewięć zmiennych jakościowych i sześć zmiennych numerycznych. Wykorzystane cechy:

1. **Status rachunku bieżącego (A1)** – kategoria: brak rachunku / ujemne saldo / saldo ≤ 200DM / saldo>200DM.
   Uzasadnienie: obecność rachunku i jego stan odzwierciedlają poziom aktywności finansowej i stabilności klienta.

2. **Czas trwania kredytu (miesiące) (A2)** – liczba miesięcy.
   Uzasadnienie: dłuższy okres kredytowania może obniżać miesięczną ratę, ale zwiększać całkowite ryzyko kredytowe.

3. **Historia kredytowa (A3)** – kategoria: dobra / pożyczki spłacane terminowo / opóźnienia.
   Uzasadnienie: przeszłe doświadczenia ze spłatą są silnym predyktorem przyszłej wypłacalności.

4. **Cel kredytu (A4)** – kategoria: samochód / edukacja / RTVAGD / inne.
   Uzasadnienie: różne cele mogą wiązać się z różnym poziomem ryzyka – kredyty konsumpcyjne są zwykle bardziej ryzykowne niż inwestycyjne.

5. **Kwota kredytu (A5)** – liczba w DM.
   Uzasadnienie: wyższa kwota to większe obciążenie budżetu klienta i większe ryzyko dla instytucji.

6. **Oszczędności/obligacje (A6)** – kategoria: <100DM / 100–500DM / 1000DM.
   Uzasadnienie: obecność oszczędności wskazuje na bufor bezpieczeństwa w sytuacjach kryzysowych.

7. **Staż zatrudnienia (A7)** – kategoria: bezrobotny / <1rok / 1–4lata / 7lat.
   Uzasadnienie: dłuższy staż sugeruje stabilność zatrudnienia i regularność dochodów.

[1] Dua, D., & Graff, C. (1994). *UCI Machine Learning Repository – Statlog (German Credit Data)*. University of California, Irvine. Dostęp online: `https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data` (dostęp: 10.06.2025).

8. **Rata w stosunku do dochodu (A8)** – procentowy udział raty w dochodzie.
   Uzasadnienie: im większy udział raty w dochodzie, tym większe ryzyko niewypłacalności.

9. **Status osobisty i płeć (A9)** – kategoria: mężczyzna_wolny / kobieta_mężatka / itp.
   Uzasadnienie: pewne kombinacje demograficzne mogą korelować ze stabilnością finansową.

10. **Obecność poręczyciela (A10)** – kategoria: brak / poręczyciel / współwnioskodawca.
    Uzasadnienie: poręczyciel lub współkredytobiorca zwiększa szanse na spłatę i zmniejsza ryzyko dla banku.

11. **Czas zamieszkania (lata) (A11)** – liczba lat w bieżącym miejscu.
    Uzasadnienie: dłuższy czas zamieszkania świadczy o stabilizacji życiowej i mniejszym ryzyku migracyjnym.

12. **Rodzaj majątku (A12)** – kategoria: nieruchomość / samochód / brak majątku.
    Uzasadnienie: majątek może być dodatkowym zabezpieczeniem kredytu.

13. **Wiek klienta (A13)** – liczba lat.
    Uzasadnienie: wiek wpływa na stabilność zatrudnienia i długość możliwego okresu spłaty.

14. **Liczba aktualnych kredytów (A14)** – liczba długów w banku.
    Uzasadnienie: większa liczba aktywnych zobowiązań oznacza wyższe ryzyko przeciążenia.

15. **Rodzaj pracy (A15)** – kategoria: wykwalifikowany / niewykwalifikowany / urzędnik / pracodawca.
    Uzasadnienie: rodzaj pracy odzwierciedla stabilność oraz poziom dochodów.

3.2 Wstępna analiza danych

3.2.1 Statystyki opisowe zmiennych numerycznych

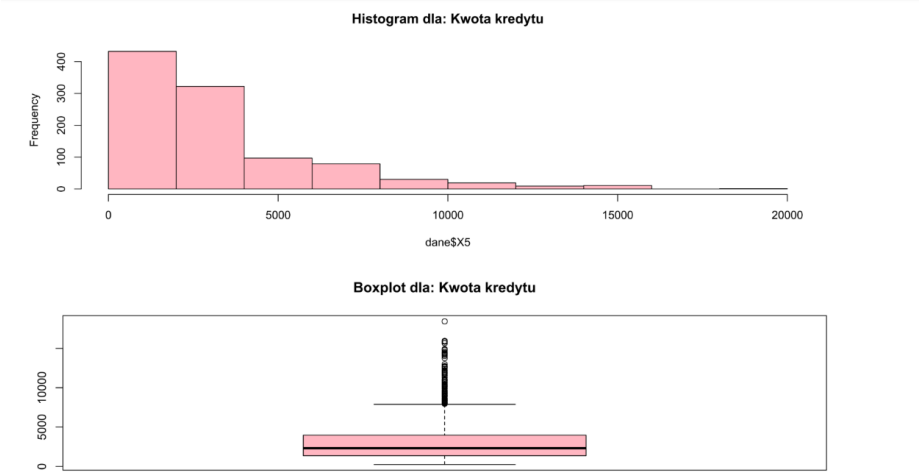
Statystyka	A2	A5	A8
Suma	20903	3 271 258	2973
Średnia	20,903	3271,258	2,973
Mediana	18	2319,5	3
Minimum	4	250	1
Kwartyl 1	12	1365,5	2
Kwartyl 2	18	2319,5	3
Kwartyl 3	24	3972,25	4
Maksimum	72	18424	4
Rozstęp	68	18174	3
Odchylenie standardowe	12,06	2822,74	1,12
Skośność	1,09	1,95	-0,53
Wariancja	145,42	7 967 843,47	1,25
Przedział międzykwartylowy	12	2606,75	2
Współczynnik zmienności	0,58	0,86	0,38

Tabela 1 – Statystyki opisowe dla zmiennych 1–3

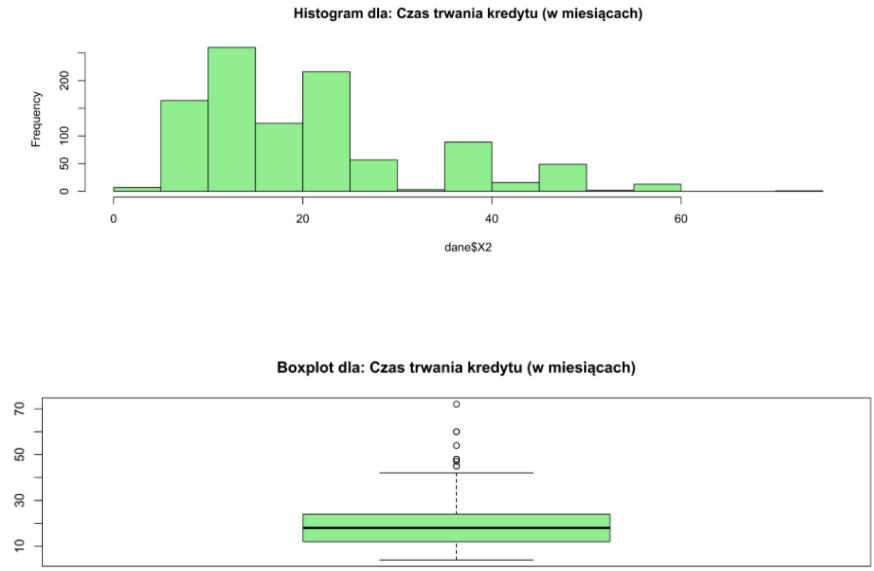
Statystyka	A11	A13	A14
Suma	2845	35546	1407
Średnia	2,845	35,546	1,407
Mediana	3	33	1
Minimum	1	19	1
Kwartyl 1	2	27	1
Kwartyl 2	3	33	1
Kwartyl 3	4	42	2
Maksimum	4	75	4
Rozstęp	3	56	3
Odchylenie standardowe	1,10	11,38	0,58
Skośność	-0,27	1,02	1,27
Wariancja	1,22	129,40	0,33
Przedział międzykwartylowy	2	15	1
Współczynnik zmienności	0,39	0,32	0,41

Tabela 2 – Statystyki opisowe dla zmiennych 4–6

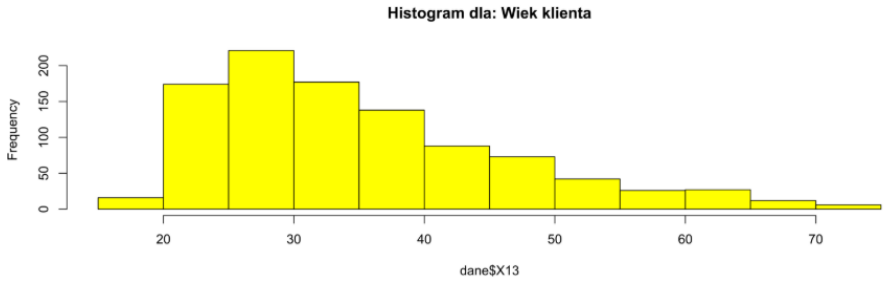
3.2.2 Podstawowe wizualizacje

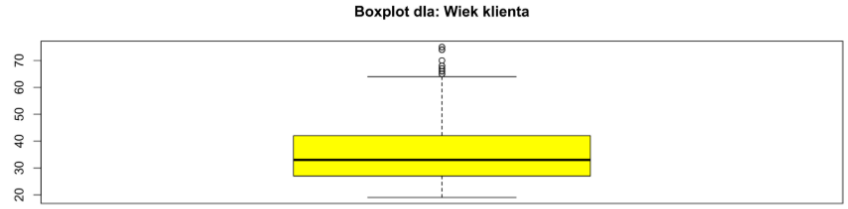


Jak można zauważyć po powyższych wykresach, zdecydowana większość kwot kredytów mieści się w dolnym przedziale wszystkich obserwacji.

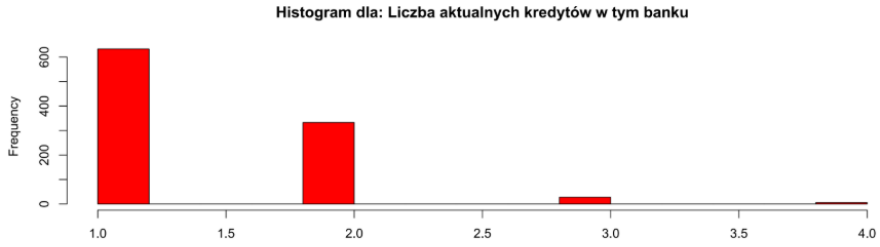


Większość klientów preferuje czas trwania kredytu między 5 a 30 miesięcy.

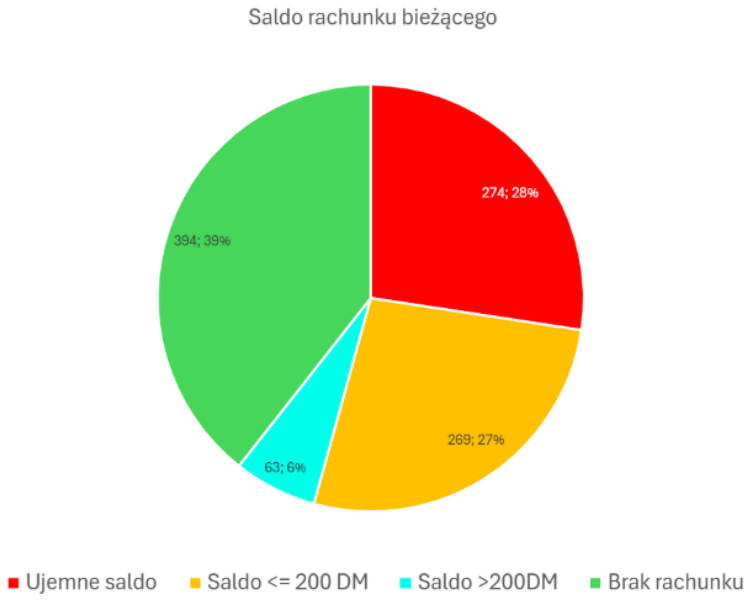




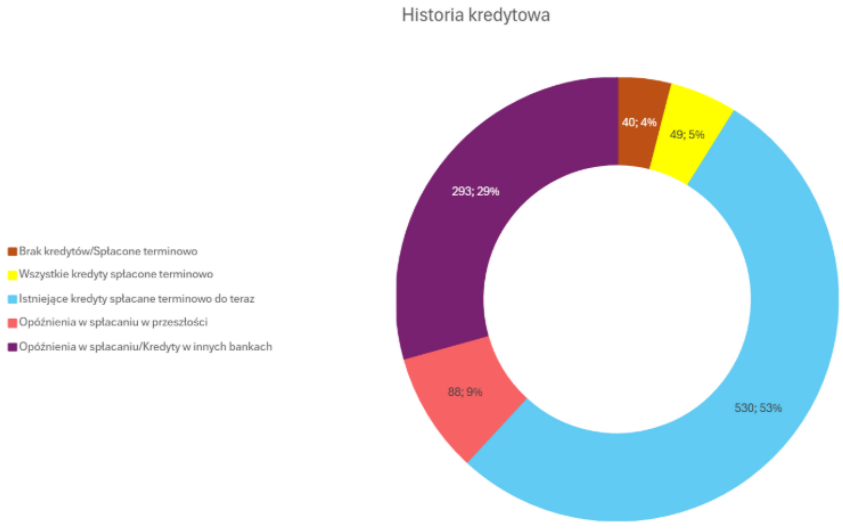
Najwięcej klientów jest w przedziale 20-40 lat.

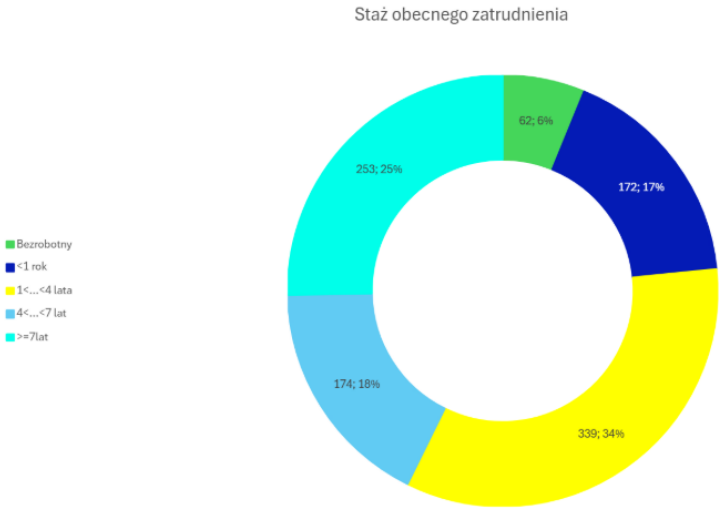


Zdecydowana większość klientów posiada aktualnie 1 lub 2 inne kredyty w tym banku. Pojedyncze osoby posiadają więcej (3 lub 4).



Większość klientów ma negatywny status rachunku lub go nie posiada, co wskazuje na podwyższone ryzyko kredytowe.





3.2.3 Braki danych

W naszym zbiorze danych posiadamy komplet, nie występują żadne braki.

3.2.4 Obserwacje odstające

Obserwacje odstające metodą przedziału międzykwartylowego						
Cecha charakterystyczna	A2	A5	A8	A11	A13	A14
Przedział międzykwartylowy	12	2606,75	2	2	15	1
Minimum	4,00	250,00	1,00	1,00	19,00	1,00
Dolny próg	-6	-2544,63	-1	-1	4,5	-0,5
Górny próg	42	7882,375	7	7	64,5	3,5
Maksimum	72,00	18 424,00	4,00	4,00	75,00	4,00
Czy są obserwacje odstające?	TAK	TAK	NIE	NIE	TAK	TAK

Tabela 3 – Łącznie zaobserwowano 171 odstających obserwacji, co stanowi 2,44% wszystkich 7000 obserwacji cech ilościowych.

Pomimo zaobserwowania kilku przypadków wartości odstających postanowiono nie korygować tych wyników. Uznano, że nietypowe wyniki nie są spowodowane błędami pomiarowymi, lecz faktycznie odzwierciedlają rzeczywisty stan. Dlatego też odstępstwa, widoczne na wykresach, są naturalną cechą danych, a ich usunięcie mogłoby prowadzić do utraty istotnych informacji o nietypowych,

ale rzeczywistych przypadkach klientów, co mogłoby pogorszyć skuteczność klasyfikacji.

4 Opis metod

4.1 Regresja logistyczna

Regresja logistyczna modeluje prawdopodobieństwo przynależności obserwacji do klasy „pozytywnej" ($Y = 1$) za pomocą tzw. szans (odds). Najpierw definiujemy szansę jako stosunek prawdopodobieństwa sukcesu do prawdopodobieństwa porażki:

$$\text{Odds} \;=\; \frac{p}{1-p} \;=\; \exp\bigl(\alpha + \beta^{\top} x\bigr),$$

gdzie $p = P(Y = 1 \mid x)$, zaś α to wyraz wolny, a $\beta = (\beta_1, \ldots, \beta_d)^{\top}$ wektor współczynników regresji. Odwrotne przekształcenie na p ma postać:

$$p \;=\; \frac{\text{Odds}}{1+\text{Odds}} \;=\; \frac{\exp(\alpha + \beta^{\top} x)}{1 + \exp(\alpha + \beta^{\top} x)}.$$

Funkcja logitowa (logarytm szans) wyraża się jako:

$$(p) \;=\; \ln \frac{p}{1-p} \;=\; \alpha + \beta^{\top} x.$$

Parametry (α, β) estymuje się przez minimalizację funkcji straty (log-loss):

$$L(\alpha, \beta) \;=\; -\sum_{j=1}^{n} \Bigl[y^{(j)} \ln p^{(j)} \;+\; \bigl(1 - y^{(j)}\bigr) \ln\bigl(1 - p^{(j)}\bigr) \Bigr],$$

gdzie $p^{(j)} = P\bigl(Y = 1 \mid x^{(j)}\bigr)$ oraz $y^{(j)} \in \{0, 1\}$ to etykieta j–tej obserwacji.

Bibliografia: Kleinbaum, D. G., Klein, M. (2010).

4.2 k-najbliższych sąsiadów (kNN)

Klasyfikator kNN przydziela etykietę na podstawie większości etykiet wśród k najbliższych punktów w przestrzeni cech. Zanim jednak obliczymy odległości, konieczna jest normalizacja cech, aby uniezależnić je od skali pomiaru. W pracy zastosowano dwie typowe metody:

- **Normalizacja min–maks**:

$$x_{\text{norm}} \;=\; \frac{x - \min(x)}{\max(x) - \min(x)},$$

dzięki czemu każda cecha przyjmuje wartości w przedziale $[0, 1]$.

- **Standaryzacja (Z–score)**:

$$x_{\text{std}} \;=\; \frac{x - \overline{x}}{\operatorname{sd}(x)},$$

gdzie $\overline{x}$ to średnia cechy, a $\operatorname{sd}(x)$ jej odchylenie standardowe.

Do wyznaczania dystansu między dwiema obserwacjami $x^{(i)}$ i $x^{(j)}$ zastosowano odległość euklidesową:

$$d\big(x^{(i)},\, x^{(j)}\big) \;=\; \sqrt{\sum_{m=1}^{d} \big(x_m^{(i)} - x_m^{(j)}\big)^2}.$$

Dla danej próby x score kNN definiuje się (w kontekście klasyfikacji binarnej) jako odsetek sąsiadów o etykiecie 1:

$$S_{\text{kNN}}(x) \;=\; \frac{1}{k} \sum_{i=1}^{k} \mathbf{1}\big(y^{(i)} = 1\big).$$

Bibliografia: Baesens, B., Van Gestel, T., Stepanova, M., Suykens, J., & Vanthienen, J. (2003).

4.3 Liniowa analiza dyskryminacyjna (LDA)

Metoda LDA zakłada rozkład cech w klasach jako wielowymiarowy rozkład normalny, co w praktyce często nie jest spełnione, dlatego bywa uznawana za metodę naiwną. Założenia LDA:

- Dla każdej klasy $k \in \{0, 1\}$ wektor cech x ma rozkład wielowymiarowy normalny:

$$x \mid \big(Y = k\big) \;\sim\; \mathcal{N}\big(\mu_k,\, \Sigma\big),$$

gdzie μ_k to wektor średnich cech w klasie k, a Σ to wspólna macierz kowariancji (identyczna dla obu klas).

- Apriori prawdopodobieństwo klasy k oznaczamy $\pi_k = P(Y = k)$.

Funkcja dyskryminacyjna dla klasy k wyznaczona jest wzorem:

$$\delta_k(x) \;=\; x^T\Sigma^{-1}\mu_k \;-\; \tfrac{1}{2}\,\mu_k^T\Sigma^{-1}\mu_k \;+\; \ln \pi_k.$$

Reguła klasyfikacyjna (Bayesa) w dwóch klasach:

$$\text{przypisz } Y=1, \text{ jeśli } \delta_1(x) > \delta_0(x); \quad \text{w przeciwnym razie } Y=0.$$

Dla dowolnych dwóch klas i, j różnicę funkcji dyskryminacyjnych zapisujemy:

$$\delta_{ij}(x) \;=\; \delta_i(x) \;-\; \delta_j(x).$$

Zn znak $\delta_{ij}(x)$ decyduje o przynależności do klasy i (>0) lub j (<0). Ponadto, z tych wartości można zbudować probabilistyczny score:

$$P_{\text{LDA}}\big(Y=k \mid x\big) \;=\; \frac{\exp\big(\delta_k(x)\big)}{\exp\big(\delta_0(x)\big) \;+\; \exp\big(\delta_1(x)\big)}.$$

Bibliografia: Izenman, A. J. (2008).

4.4 Model hybrydowy (średnia ważona score'ów)

W modelu hybrydowym łączymy probabilistyczne wyniki (score'y) z trzech wyżej opisanych metod: regresji logistycznej, kNN oraz LDA. Definiujemy:

$$S_{\text{hyb}}(x) \;=\; w_{\log}\,p_{\log}(x) + w_{\text{kNN}}\,p_{\text{kNN}}(x) + w_{\text{LDA}}\,p_{\text{LDA}}(x), \quad w_{\log}+w_{\text{kNN}}+w_{\text{LDA}}=1,\; w_i \ge 0,$$

gdzie:

- $p_{\log}(x) = P_{\text{LR}}\big(Y=1 \mid x\big)$ – probabilistyczny wynik z regresji logistycznej,
- $p_{\text{kNN}}(x) = S_{\text{kNN}}(x)$ – score kNN (proporcja sąsiadów o etykiecie 1),
- $p_{\text{LDA}}(x) = P_{\text{LDA}}\big(Y=1 \mid x\big)$ – wynik probabilistyczny z LDA.

Klasyfikację wykonujemy, porównując $S_{\text{hyb}}(x)$ z progiem 0.5:

$$\text{jeśli } S_{\text{hyb}}(x) \ge 0.5 \implies Y=1, \quad \text{w przeciwnym razie } Y=0.$$

Wagi $(w_{\log}, w_{\mathrm{kNN}}, w_{\mathrm{LDA}})$ dobierane są metodą metaregresji (stacking) na zbiorze walidacyjnym. Tworzymy wtedy macierz:

$$X_{\text{meta}} = \begin{bmatrix} p_{\log}(x^{(1)}) & p_{\mathrm{kNN}}(x^{(1)}) & p_{\mathrm{LDA}}(x^{(1)}) \\ p_{\log}(x^{(2)}) & p_{\mathrm{kNN}}(x^{(2)}) & p_{\mathrm{LDA}}(x^{(2)}) \\ \vdots & \vdots & \vdots \\ p_{\log}(x^{(n_{\text{val}})}) & p_{\mathrm{kNN}}(x^{(n_{\text{val}})}) & p_{\mathrm{LDA}}(x^{(n_{\text{val}})}) \end{bmatrix}, \quad y_{\text{meta}} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n_{\text{val}})} \end{bmatrix}.$$

Następnie na danych $(X_{\text{meta}}, y_{\text{meta}})$ trenujemy model regresji logistycznej, którego współczynniki końcowe $(\widehat{w}_{\log}, \widehat{w}_{\mathrm{kNN}}, \widehat{w}_{\mathrm{LDA}})$ służą jako optymalne wagi dla hybrydowego algorytmu.

Bibliografia: Wolpert, D. H. (1992). „Stacked Generalization," *Neural Networks*, 5(2), 241–259.

5 Metody oceny modeli i miary skuteczności

W celu rzetelnej oceny jakości klasyfikatorów zastosowano kilka popularnych miar skuteczności predykcji. Ponieważ problem klasyfikacji zdolności kredytowej może wiązać się z nierównomiernym rozkładem klas oraz różnymi kosztami błędów, konieczne było zastosowanie więcej niż jednej miary.

- **Accuracy** – odsetek poprawnie sklasyfikowanych przypadków spośród wszystkich obserwacji. Miara ta jest intuicyjna, lecz w przypadku niezbalansowanych zbiorów danych może być myląca.
- **Precision (precyzja)** – stosunek liczby poprawnych klasyfikacji pozytywnych do liczby wszystkich przypadków sklasyfikowanych jako pozytywne. Miara ta pokazuje, jak wiele z przypadków zakwalifikowanych do klasy „1" rzeczywiście powinno się tam znaleźć.
- **Recall (czułość)** – stosunek liczby poprawnie sklasyfikowanych przypadków pozytywnych do liczby wszystkich rzeczywistych przypadków pozytywnych. Informuje o zdolności modelu do wykrywania wszystkich przypadków klasy „1".
- **F1-score** – średnia harmoniczna precyzji i czułości. Jest szczególnie użyteczna przy niezbalansowanych danych, ponieważ uwzględnia kompromis między wykrywaniem pozytywnych przypadków a błędną klasyfikacją.

- **AUC-ROC** – obszar pod krzywą ROC (Receiver Operating Characteristic), który wskazuje zdolność modelu do odróżniania klas. Wartość 0,5 oznacza losowy model, a 1 – idealny klasyfikator.

Dla każdego modelu obliczono wartości powyższych miar, a także zaprezentowano je w formie tabel oraz wykresów porównawczych. Ocena została przeprowadzona przy użyciu **podziału danych na zbiór treningowy i testowy w proporcji 80/20**.

6 Rezultaty

6.1 Model: Regresja logistyczna

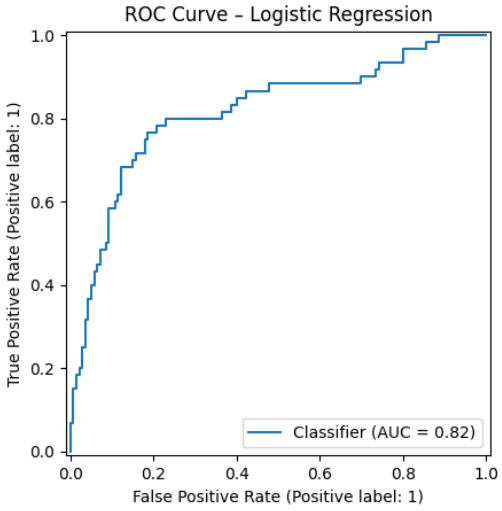


Tabela 4 – Krzywa ROC dla modelu regresji logistycznej

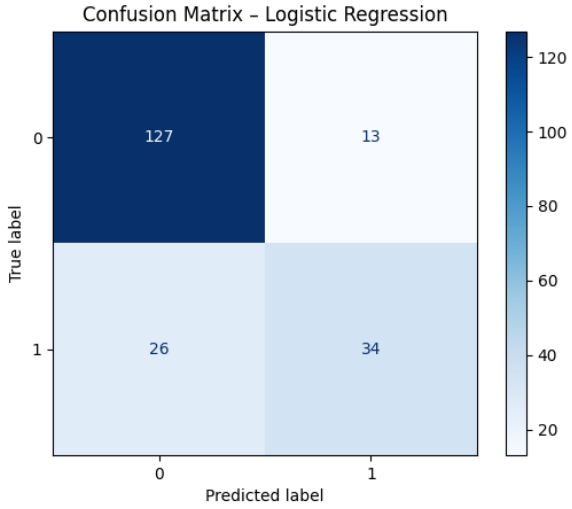


Tabela 5 – Macierz pomyłek dla modelu regresji logistycznej

6.2 Model: k-najbliższych sąsiadów

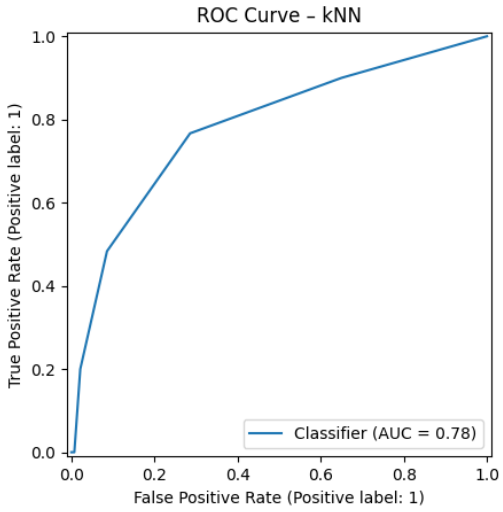


Tabela 6 – Krzywa ROC dla modelu kNN

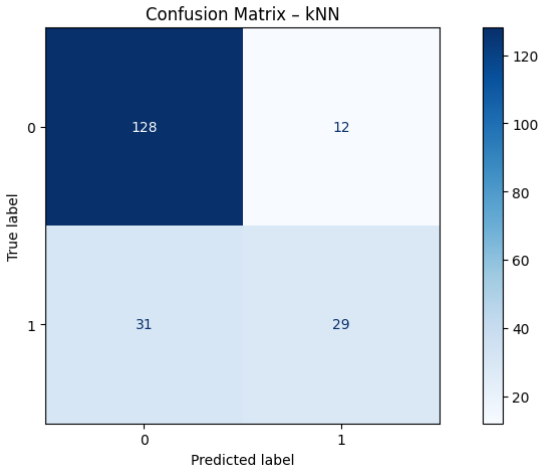


Tabela 7 – Macierz pomyłek dla modelu kNN

6.3 Model: Liniowa analiza dyskryminacyjna

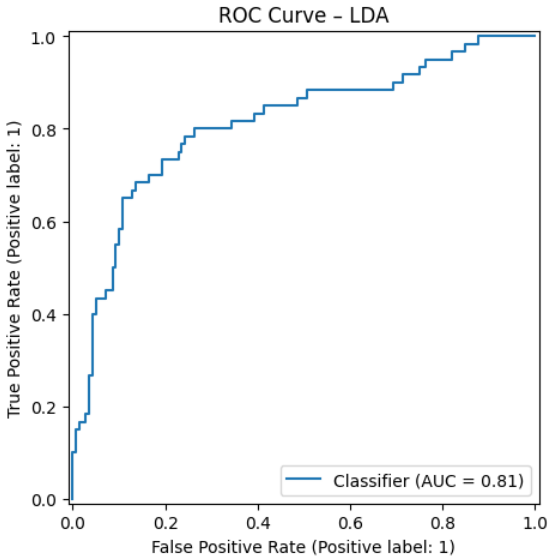


Tabela 8 – Krzywa ROC dla modelu LDA

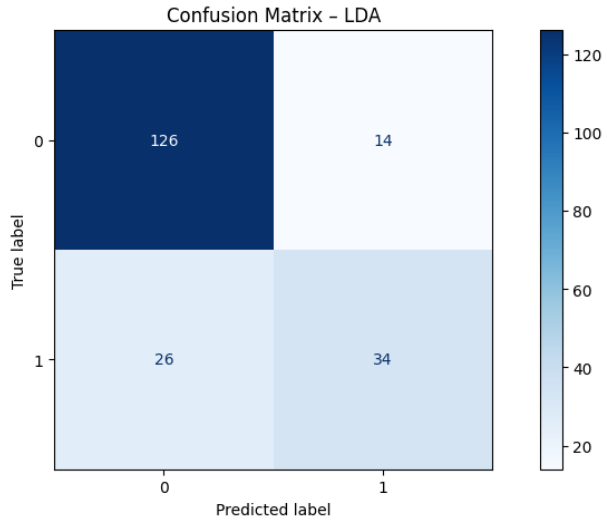


Tabela 9 – Macierz pomyłek dla modelu LDA

6.4 Model hybrydowy

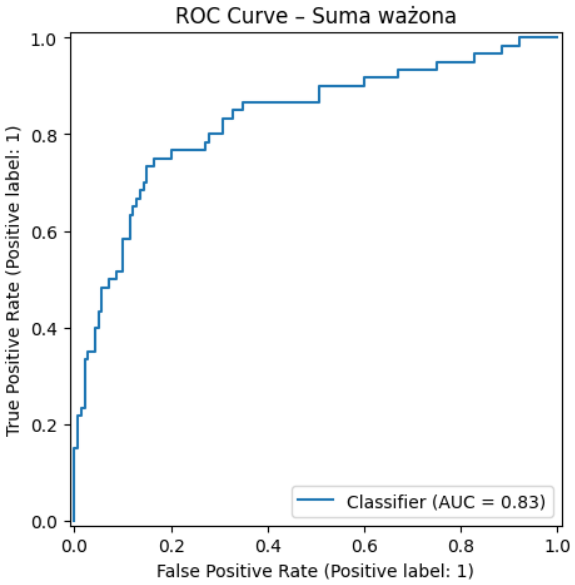


Tabela 10 – Krzywa ROC dla modelu hybrydowego

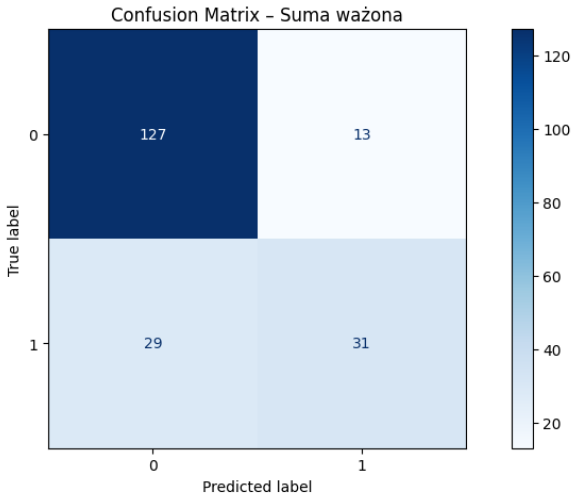
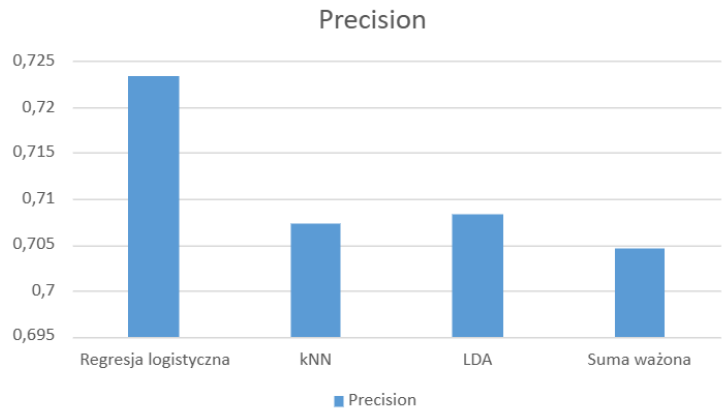


Tabela 11 – Macierz pomyłek dla modelu hybrydowego

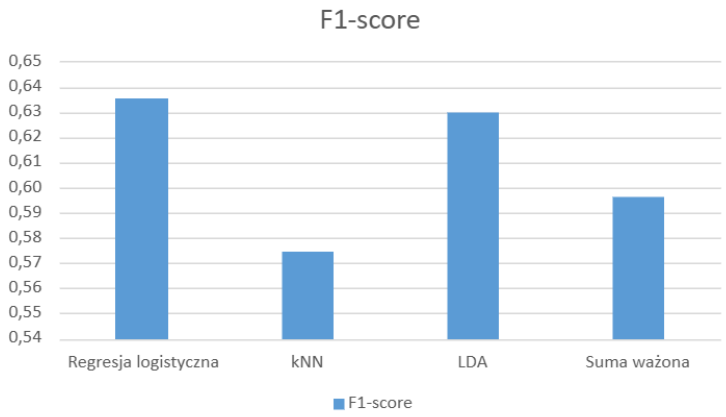
6.5 Porównanie miar skuteczności

	Regresja logistyczna	kNN	LDA	Suma ważona
Accuracy	0.8050	0.7850	0.8000	0.7900
Precision	0.7234	0.7073	0.7083	0.7045
Recall	0.5667	0.4833	0.5667	0.5167
F1-score	0.6355	0.5743	0.6296	0.5962
AUC-RO	0.8170	0.7845	0.8113	0.8299

Tabela 12 – Porównanie wartości miar skuteczności dla poszczególnych modeli



Rysunek 1 – Precyzja modeli klasyfikacyjnych



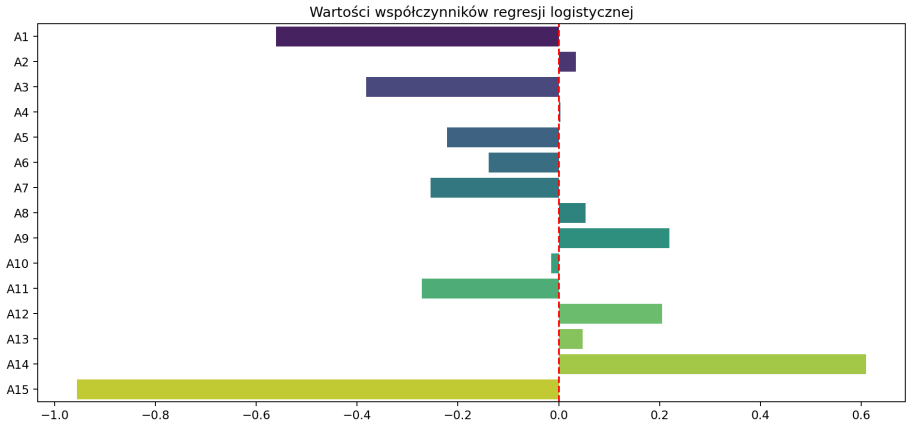
Rysunek 2 – F1-score modeli klasyfikacyjnych

6.6 Wnioski

Na podstawie porównania wyników czterech modeli klasyfikacyjnych można stwierdzić, że **regresja logistyczna** ogólnie osiągnęła najlepsze rezultaty. Model ten charakteryzuje się najwyższą skutecznością oraz najlepszym bilansem między precyzją a czułością, co potwierdza również najwyższy F1-score. LDA wypadła porównywalnie, natomiast model kNN wyraźnie odstaje pod względem wszystkich miar. Model hybrydowy uzyskał co prawda najwyższą wartość AUC-ROC, jednak inne wskaźniki sugerują, że może być mniej stabilny w klasyfikacji.

W kontekście oceny **zdolności kredytowej** - najlepszym modelem jest również regresja logistyczna. Osiąga ona najwyższe wartości Precision (0.7234) oraz F1-score (0.6355), co oznacza, że najtrafniej przewiduje klientów, którzy rzeczywiście spłacą kredyt. Jest to kluczowe z punktu widzenia ryzyka finansowego, ponieważ ogranicza liczbę błędnych decyzji kredytowych (False Positives). Mimo że model hybrydowy osiągnął najwyższe AUC-ROC (0.8299), jego niższe Precision i Recall wskazują, że nie sprawdza się tak dobrze przy konkretnym progu klasyfikacyjnym.

W ramach analizy regresji logistycznej sprawdzono także, które cechy mają największe znaczenie w przewidywaniu przynależności do jednej z dwóch klas. W tym celu oceniono istotność statystyczną każdej cechy na podstawie wartości *p-value*. Poniżej przedstawiono wykres wartości współczynników regresji logistycznej dla wszystkich cech użytych w modelu. Dodatkowo, na podstawie wartości p-value, wyznaczono trzy najbardziej istotne cechy mające największy wpływ na klasyfikację.



Rysunek 3 – Porównanie wartości współczynników regresji logistycznej

Cecha	Współczynnik regresji	p-value
Status rachunku bieżącego (A1)	-0.561	0.00000
Historia kredytowa (A3)	-0.382	0.00001
Czas trwania kredytu (miesiące) (A2)	0.034	0.00004

Tabela 13 – Top 3 najbardziej istotne cechy w regresji logistycznej

Na podstawie analizy regresji logistycznej można stwierdzić, że najbardziej istotnym czynnikiem wpływającym na decyzję kredytową jest **status rachunku**

bieżącego. Negatyczny współczynnik tej cechy oznacza, że niekorzystny stan konta znacząco obniża szanse na uzyskanie kredytu. Równie silny, negatywny wpływ ma historia kredytowa – osoby z gorszą przeszłością finansową mają zauważalnie mniejsze szanse na pozytywną decyzję. Czas trwania kredytu z kolei wykazuje niewielki, ale dodatni wpływ, co sugeruje, że dłuższy okres spłaty może nieznacznie zwiększać prawdopodobieństwo przyznania kredytu, choć ten efekt jest słabszy niż w przypadku pozostałych dwóch cech.

7 Przykład użycia modeli na danych syntetycznych

W celu przetestowania działania wytrenowanych modeli klasyfikacyjnych wygenerowano 10 przykładowych instancji danych wejściowych, reprezentujących potencjalnych klientów ubiegających się o kredyt. Dla każdej z tych instancji wykonano predykcję klasy.

nr	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15
1	A13	24	A32	A43	5152	A61	A74	4	A93	A101	2	A123	25	1	A173
2	A13	24	A32	A43	1377	A62	A75	4	A92	A101	2	A124	47	1	A173
3	A11	12	A32	A40	1228	A61	A73	4	A92	A101	2	A121	24	1	A172
4	A13	12	A34	A40	1480	A63	A71	2	A93	A101	4	A124	66	2	A171
5	A14	24	A32	A42	3062	A63	A75	4	A93	A101	3	A124	32	1	A173
6	A14	24	A32	A40	1474	A62	A72	4	A94	A101	3	A121	33	1	A173
7	A11	18	A32	A42	4153	A61	A73	2	A93	A102	3	A123	42	1	A173
8	A11	36	A33	A49	2145	A61	A74	2	A93	A101	1	A123	24	2	A173
9	A12	6	A32	A45	454	A61	A72	3	A94	A101	1	A122	22	1	A172
10	A14	12	A32	A43	804	A61	A75	4	A93	A101	4	A123	38	1	A173

Tabela 14 – Wartości przykładowych instancji

- **Regresja logistyczna** przyjęła następujące klasy: 0, 1, 0, 1, 1, 0, 0, 0, 1, 1
- **kNN** przyjął następujące klasy: 0, 1, 0, 0, 1, 0, 0, 0, 0, 0
- **LDA** przyjęła następujące klasy: 0, 1, 0, 1, 1, 0, 0, 0, 0, 1
- **Model hybrydowy** przyjął następujące klasy: 0, 1, 0, 1, 1, 0, 0, 0, 0, 1

8 Podsumowanie

W przeprowadzonym projekcie oceniono skuteczność trzech klasycznych metod klasyfikacyjnych (regresja logistyczna, kNN oraz LDA) oraz modelu hybrydowego

w przewidywaniu zdolności kredytowej klienta na podstawie danych demograficznych i finansowych. Dane obejmowały 1000 obserwacji i 15 cech, zarówno jakościowych, jak i numerycznych.

Analiza statystyczna wykazała różnorodność wartości oraz obecność obserwacji odstających, które jednak zostały pozostawione ze względu na ich realny charakter. Dane były kompletne, bez braków.

Wyniki klasyfikacji wskazały, że spośród badanych metod najlepszą skuteczność w ocenie zdolności kredytowej osiągnęła regresja logistyczna, szczególnie pod względem miary precyzji (Precision). Oznacza to, że model regresji logistycznej najdokładniej identyfikuje klientów faktycznie zdolnych do spłaty kredytu, minimalizując liczbę fałszywych pozytywów, co jest bardzo istotne z punktu widzenia instytucji finansowej.

Model hybrydowy, oparty na średniej ważonej wyników trzech metod, również prezentował dobre wyniki, choć nie przewyższył jednoznacznie regresji logistycznej w najważniejszych metrykach.

Podsumowując, regresja logistyczna jest najbardziej efektywnym i stabilnym modelem do oceny zdolności kredytowej na badanym zbiorze danych, oferując wysoką precyzję i praktyczną użyteczność w decyzjach kredytowych.

9 Załączniki

Poniżej znajduje się lista plików załączonych do projektu:

- **Analiza Danych.xlsx** – plik programu Excel zawierający:
  - surowe dane wejściowe,
  - statystyki opisowe,
  - wykrywanie wartości odstających,
  - wykresy porównujące skuteczność klasyfikatorów.
- **BiH.R** – skrypt w języku R, który generuje:
  - histogramy oraz wykresy typu boxplot,
- **METDAN/** – folder zawierający program w języku Python:
  - trenowanie modeli klasyfikacyjnych (logistyczna, kNN, LDA, hybryda),
  - obliczanie miar skuteczności modeli,
  - generowanie wykresów ROC oraz macierzy pomyłek.

- **regresja_logistyczna/** – folder zawierający skrypt Pythona:
  - obliczanie współczynników regresji oraz wartości istotności (p-value) dla każdej z cech,
  - wizualizację wpływu cech na wynik klasyfikacji.

10 Bibliografia

1. Kleinbaum, D. G., & Klein, M. (2010). *Logistic Regression: A Self-Learning Text* (3rd ed.). Springer.
2. Baesens, B., Van Gestel, T., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). "Benchmarking state-of-the-art classification algorithms for credit scoring." *Journal of the Operational Research Society*, 54(6), 627–635.
3. Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer.
4. Wolpert, D. H. (1992). "Stacked Generalization." *Neural Networks*, 5(2), 241–259.
5. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, 12, 2825–2830.
6. Dua, D., & Graff, C. (1994). *UCI Machine Learning Repository – Statlog (German Credit Data)*. University of California, Irvine. Dostęp online: `https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data` (dostęp: 10.06.2025).