Predykcja Ryzyka Udaru Mózgu na podstawie danych ze strony kaggle

Projekt nr 2 z przedmiotu Metody Analizy Danych

Julia Łyszkowska, s223487
Tymoteusz Majewski , s223285
Szymon Marciniak, s223526
Natan Misztal, s223309
Wiktor Koprowski, s223372

8 czerwca 2025

Streszczenie

Projekt wykorzystuje metody uczenia maszynowego do przewidywania ryzyka udaru u pacjentów na podstawie cech demograficznych, zdrowotnych i stylu życia. W analizie zastosowano zbiór danych zawierający 5110 obserwacji, uwzględniający takie zmienne jak wiek, płeć, występowanie nadciśnienia, średni poziom glukozy we krwi, wskaźnik masy ciała (BMI) oraz status palenia.

Głównym celem badania było opracowanie modelu klasyfikacyjnego, który na podstawie wprowadzonych danych pacjenta będzie w stanie ocenić prawdopodobieństwo wystąpienia udaru. W projekcie przetestowano różne algorytmy uczenia maszynowego, w tym regresję logistyczną, maszyny wektorów nośnych (SVM), metody ensemble oraz lasy losowe, aby wybrać model o optymalnej skuteczności predykcyjnej.

Wyniki wskazują, że najlepsze rezultaty osiągnęła regresja logistyczna, uzyskując dokładność na poziomie 73,8% i wartość AUC-ROC 83,9%. Mimo że model oparty na lasach losowych (Random Forest) osiągnął wyższą dokładność (93,6%), charakteryzował się bardzo niską czułością (8%), co czyni go niepraktycznym w rzeczywistych zastosowaniach medycznych. Ostateczny klasyfikator może być wykorzystany jako narzędzie wspomagające wczesne wykrywanie zagrożenia udarem, łącząc dobrą skuteczność predykcyjną z interpretowalnością wyników.
Słowa kluczowe: uczenie maszynowe, predykcja udaru, klasyfikacja, czynniki ryzyka, regresja logistyczna, analiza danych medycznych

Spis treści

1 Wprowadzenie

Udar mózgu stanowi jedną z głównych przyczyn śmierci i niepełnosprawności na świecie. Wczesna identyfikacja osób zagrożonych pozwala na wdrożenie działań profilaktycznych, które mogą znacząco zmniejszyć ryzyko wystąpienia udaru. Współczesna medycyna coraz częściej wykorzystuje metody uczenia maszynowego do przewidywania ryzyka chorób na podstawie danych klinicznych i demograficznych.

2 Przedmiot badania

2.1 Cel i zakres badania

Głównym celem badania było opracowanie i porównanie efektywności różnych modeli uczenia maszynowego w przewidywaniu ryzyka udaru mózgu. Szczególny nacisk położono na znalezienie optymalnego rozwiązania łączącego wysoką dokładność predykcji z dobrą interpretowalnością wyników, co ma kluczowe znaczenie dla zastosowań klinicznych. Badanie miało również na celu identyfikację najważniejszych czynników ryzyka wpływających na prawdopodobieństwo wystąpienia udaru.

2.2 Przegląd literatury

W ostatnich latach obserwuje się rosnące zainteresowanie zastosowaniem metod uczenia maszynowego w medycynie, w tym w przewidywaniu ryzyka udaru mózgu. Tradycyjne metody statystyczne, choć użyteczne, często charakteryzują się umiarkowaną dokładnością predykcyjną, co skłania do poszukiwania bardziej zaawansowanych rozwiązań. Algorytmy uczenia maszynowego, dzięki zdolności do analizowania złożonych i obszernych zbiorów danych, mogą identyfikować subtelne wzorce i korelacje, które są trudne do uchwycenia konwencjonalnymi technikami, co prowadzi do bardziej precyzyjnych i spersonalizowanych prognoz ryzyka udaru [4].

Badania wykazały skuteczność różnych algorytmów uczenia maszynowego w przewidywaniu udaru. Na przykład, w pracy [2] autorzy zastosowali sztuczną inteligencję do przewidywania prawdopodobieństwa udaru na podstawie danych z ankiet medycznych, wykorzystując różne podejścia, w tym regresję logistyczną i lasy losowe. Podkreślili oni również znaczenie interpretowalnej sztucznej inteligencji (XAI) dla zrozumienia ważności cech, co jest kluczowe w zastosowaniach klinicznych. Podobnie, analiza w [3] porównała efektywność algorytmów takich jak regresja logistyczna, lasy losowe, maszyny wektorów nośnych (SVM), Extreme Gradient Boosting (XGBoost) i Light Gradient Boosted Machine (LightGBM) w przewidywaniu udaru, wykorzystując dane ze studium Suita. W tej pracy zastosowano również metody wyjaśniania modeli, takie jak SHAP, do identyfikacji kluczowych czynników predykcyjnych, w tym wieku, ciśnienia skurczowego, nadciśnienia i poziomu glukozy we krwi.

Prowadzone są również badania skupiające się na analizie eksploracyjnej danych (EDA) w połączeniu z modelami uczenia maszynowego. Praca autorstwa Fu [5] przedstawia zastosowanie lasów losowych, regresji logistycznej i XGBoost do przewidywania ryzyka udaru, wykorzystując zmienne demograficzne i związane ze stylem życia, takie jak wiek, płeć, BMI i status palenia. Wyniki tych badań wskazują na duży potencjał algorytmów uczenia maszynowego w usprawnianiu wczesnego wykrywania ryzyka udaru i poprawie opieki nad pacjentem.

2.3 Opis zbioru danych

W projekcie wykorzystano publicznie dostępny zbiór danych *Stroke Prediction Dataset*[1], udostępniony na platformie Kaggle przez użytkownika *fedesoriano* .

Dane obejmują informacje o 5110 pacjentach, których celem jest przewidywanie ryzyka wystąpienia udaru mózgu. Choć dokładny okres zbierania danych nie został wskazany, dane mają charakter przekrojowy i dotyczą populacji ogólnej. Nie podano także konkretnego kraju lub regionu, którego dane dotyczą — zbiór ma charakter syntetyczny lub anonimowo zebrany z różnych źródeł.

Zbiór zawiera 11 zmiennych, w tym jedną zmienną objaśnianą `stroke` (0: brak udaru, 1: udar) oraz 10 zmiennych objaśniających o charakterze demograficznym i medycznym. Są to m.in.: płeć, wiek, obecność nadciśnienia i chorób serca, status małżeński, typ zatrudnienia, miejsce zamieszkania, poziom glukozy we krwi, wskaźnik BMI oraz status palenia tytoniu.

Główne cechy zbioru:

- Liczba obserwacji: 5110,
- Liczba zmiennych: 11 (w tym 1 zmienna zależna),
- Typy danych: zmienne ilościowe (np. wiek, glukoza, BMI) i jakościowe (np. płeć, status palenia),

2.4 Zmienne wybrane do analizy

Do analizy predykcyjnej ryzyka udaru mózgu wybrano wszystkie zmienne dostępne w zbiorze danych *Stroke Prediction Dataset* [1], z wyłączeniem identyfikatora pacjenta (`id`), który nie wnosił wartości predykcyjnej. Zmienną zależną, będącą przedmiotem przewidywania, była zmienna binarna `stroke`, przyjmująca wartość 1 w przypadku wystąpienia udaru oraz 0 w przypadku jego braku.

Pozostałe 10 zmiennych objaśniających, wykorzystanych w modelach uczenia maszynowego, obejmowało szeroki zakres czynników demograficznych, zdrowotnych i behawioralnych. Zmienne te zostały wybrane w celu kompleksowej oceny ryzyka udaru. Poniżej przedstawiono szczegółowy opis zmiennych wykorzystanych w analizie:

- `gender`: Płeć pacjenta (kobieta, mężczyzna, inne).
- `age`: Wiek pacjenta w latach, zmienna numeryczna.
- `hypertension`: Obecność nadciśnienia (0: brak nadciśnienia, 1: nadciśnienie).
- `heart_disease`: Obecność chorób serca (0: brak chorób serca, 1: choroby serca).
- `ever_married`: Status małżeński (Yes: zamężna/żonaty, No: niezamężna/nieżonaty).
- `work_type`: Typ zatrudnienia (np. Private, Self-employed, Govt_job, Children, Never_worked).
- `Residence_type`: Typ miejsca zamieszkania (Urban: miejskie, Rural: wiejskie).
- `avg_glucose_level`: Średni poziom glukozy we krwi, zmienna numeryczna.
- `bmi`: Wskaźnik masy ciała (Body Mass Index), zmienna numeryczna.

- `smoking_status`: Status palenia tytoniu (np. formerly smoked, never smoked, smokes, Unknown).

Wszystkie te zmienne zostały wykorzystane w procesie trenowania i walidacji modeli predykcyjnych, aby ocenić ich wpływ na prawdopodobieństwo wystąpienia udaru i zidentyfikować kluczowe czynniki ryzyka.

3 Wstępna analiza danych

3.1 Statystyki opisowe

W tej sekcji przedstawiono statystyki opisowe dla zmiennych numerycznych wykorzystanych w analizie. Dla zmiennych binarnych (przyjmujących wartości 0 lub 1), takich jak `hypertension`, `heart_disease` czy `stroke`, przedstawiono średnią (interpretowaną jako proporcję występowania cechy) oraz odchylenie standardowe. Pozostałe statystyki dla zmiennych binarnych zostały pominięte. Dla zmiennych ciągłych przedstawiono pełen zakres statystyk.

Poniższe tabele przedstawiają statystyki opisowe dla kluczowych zmiennych numerycznych:

Tabela 1: Statystyki opisowe dla zmiennych numerycznych (część 1)

Statystyka	**age**	**hypertension**	**heart_disease**	**avg_glucose_level**
Liczba obserwacji	5110	5110	5110	5110
Średnia	43.23	0.10	0.05	106.15
Odchylenie standardowe	22.61	0.30	0.23	45.28
Min	0.08	N/A	N/A	55.12
25% kwartyl	25.00	N/A	N/A	77.25
Mediana	45.00	N/A	N/A	91.89
75% kwartyl	61.00	N/A	N/A	114.09
Max	82.00	N/A	N/A	271.74

Tabela 2: Statystyki opisowe dla zmiennych numerycznych (część 2)

Statystyka	**bmi**	**stroke**
Liczba obserwacji	4909	5110
Średnia	28.89	0.05
Odchylenie standardowe	7.85	0.22
Min	10.30	N/A
25% kwartyl	23.50	N/A
Mediana	28.10	N/A
75% kwartyl	33.10	N/A
Max	97.60	N/A

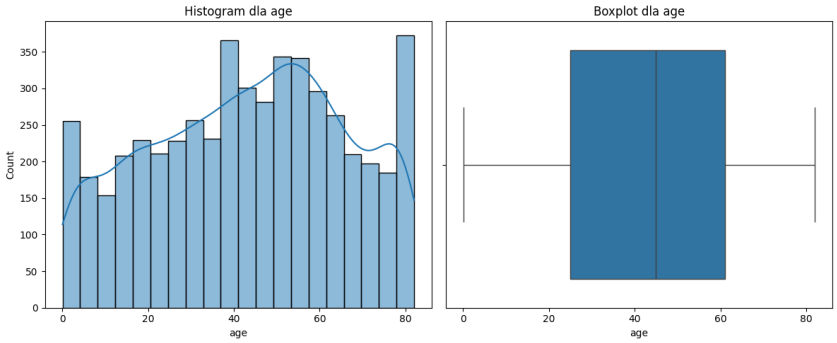
3.2 Wizualizacja Danych

W tej sekcji przedstawiono graficzną reprezentację rozkładów zmiennych zawartych w zbiorze danych. Wizualizacje zostały dobrane odpowiednio do typu zmiennej: histogramy

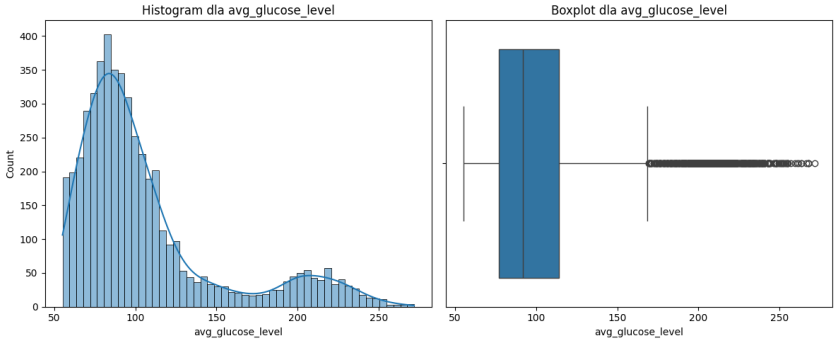
i wykresy pudełkowe (boxploty) dla zmiennych numerycznych ciągłych, wykresy kołowe dla zmiennych binarnych oraz wykresy słupkowe przedstawiające rozkłady dla zmiennych kategorycznych.

3.2.1 Zmienne Numeryczne

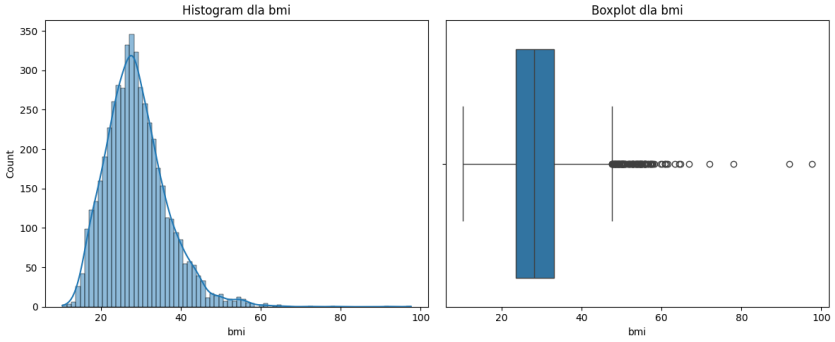
Dla zmiennych numerycznych, takich jak **wiek** (`age`), **średni poziom glukozy we krwi** (`avg_glucose_level`) i **wskaźnik BMI** (`bmi`), zastosowano histogramy, aby pokazać rozkład częstości wartości, oraz boxploty, aby zilustrować rozrzut danych, medianę i obecność wartości odstających.



Rysunek 1: Rozkład wieku



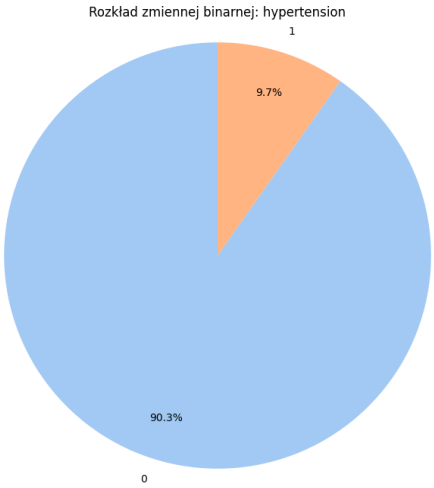
Rysunek 2: Rozkład średniego poziomu glukozy we krwi



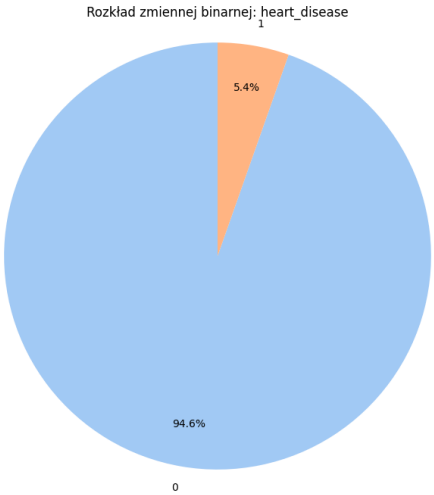
Rysunek 3: Rozkład wskaźnika BMI

3.2.2 Zmienne Binarne

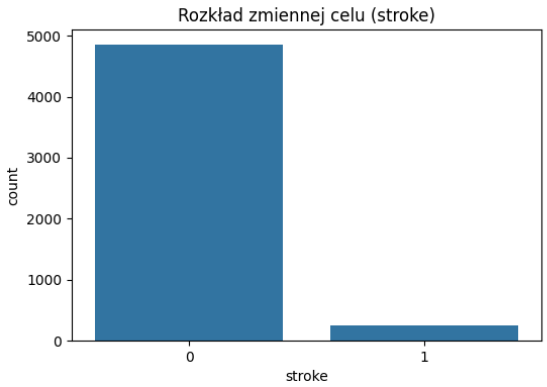
Zmienne binarne, takie jak **nadciśnienie** (`hypertension`), **choroby serca** (`heart_disease`) oraz **wystąpienie udaru** (`stroke`), zostały zwizualizowane za pomocą wykresów kołowych. Wykresy te przedstawiają proporcje (procentowy udział) poszczególnych kategorii.



Rysunek 4: Rozkład występowania nadciśnienia



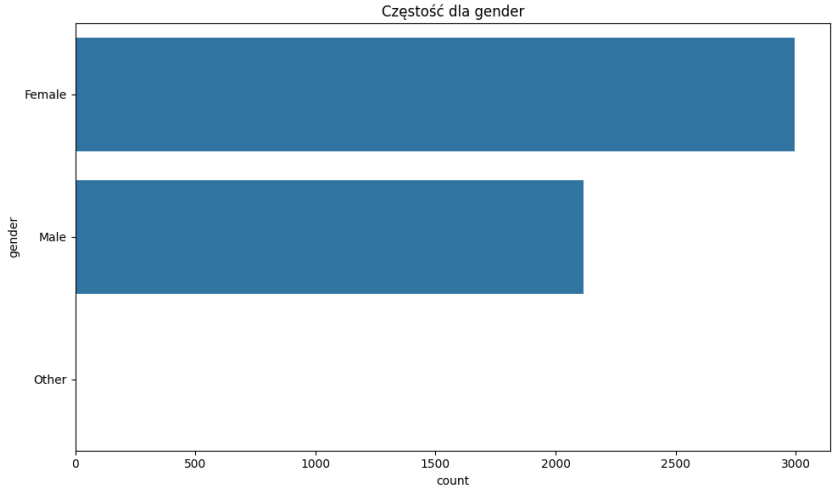
Rysunek 5: Rozkład występowania chorób serca



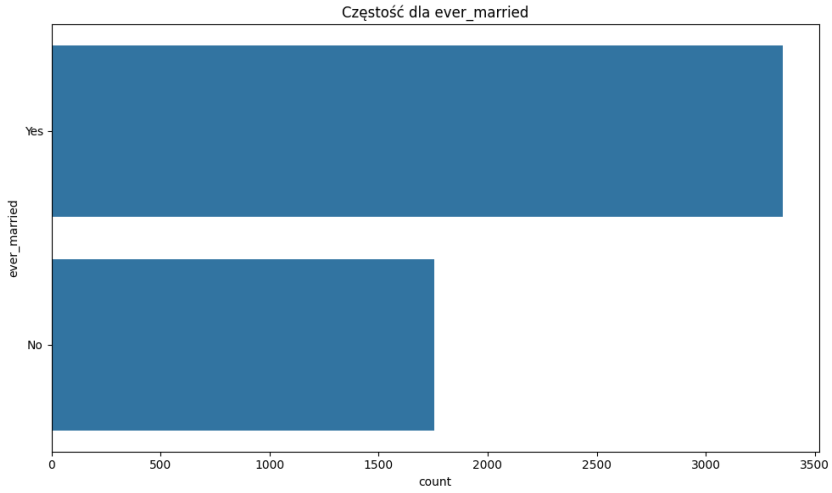
Rysunek 6: Rozkład występowania udaru

3.2.3 Zmienne Kategoryczne

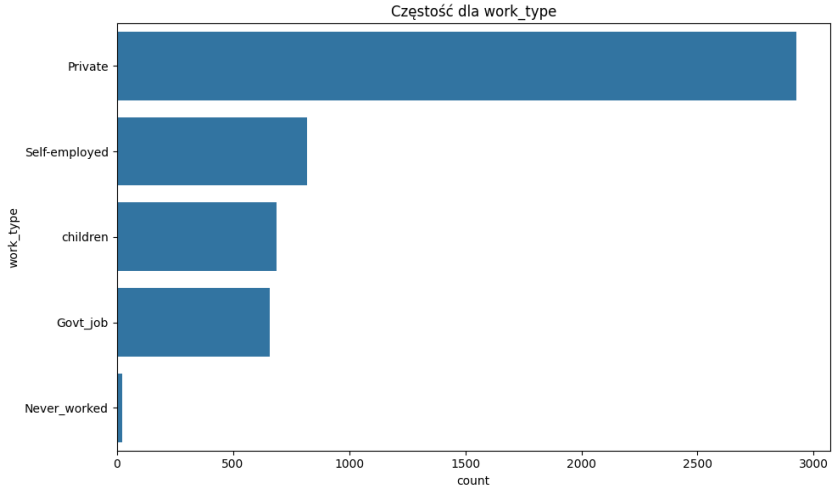
Dla zmiennych kategorycznych, takich jak **płeć** (`gender`), **status małżeński** (`ever_married`), **typ zatrudnienia** (`work_type`), **typ miejsca zamieszkania** (`Residence_type`) i **status palenia tytoniu** (`smoking_status`), zastosowano wykresy słupkowe. Wykresy te ilustrują liczebność każdej kategorii.



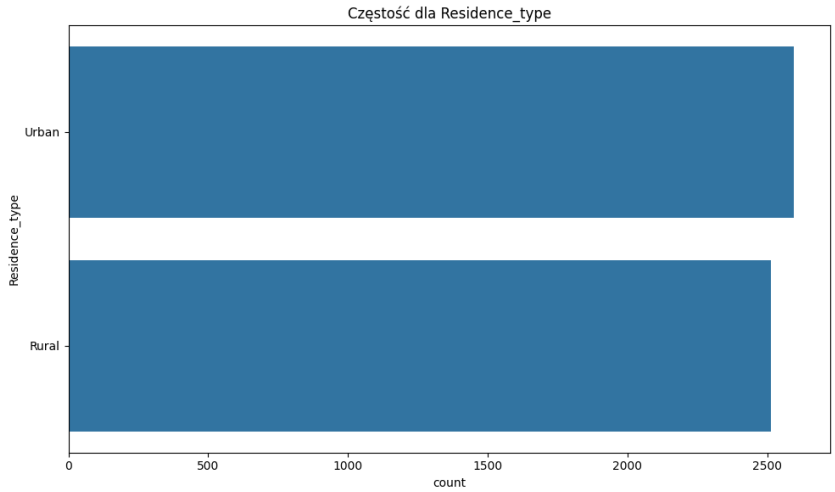
Rysunek 7: Rozkład płci



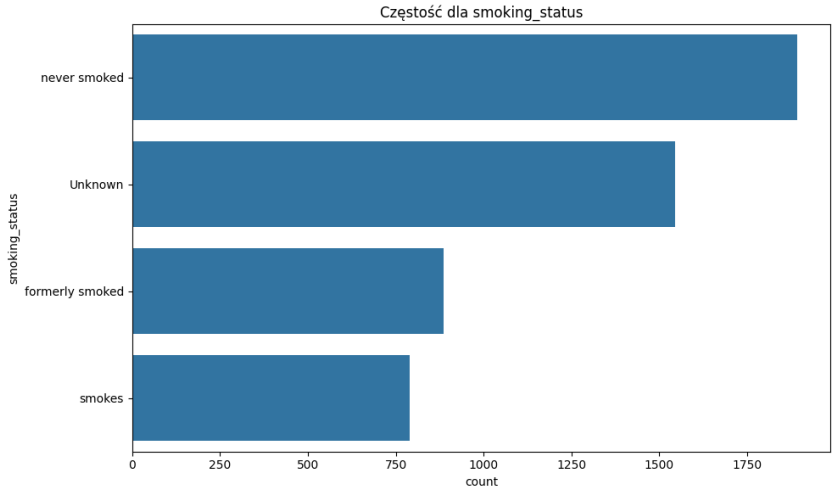
Rysunek 8: Rozkład statusu małżeńskiego



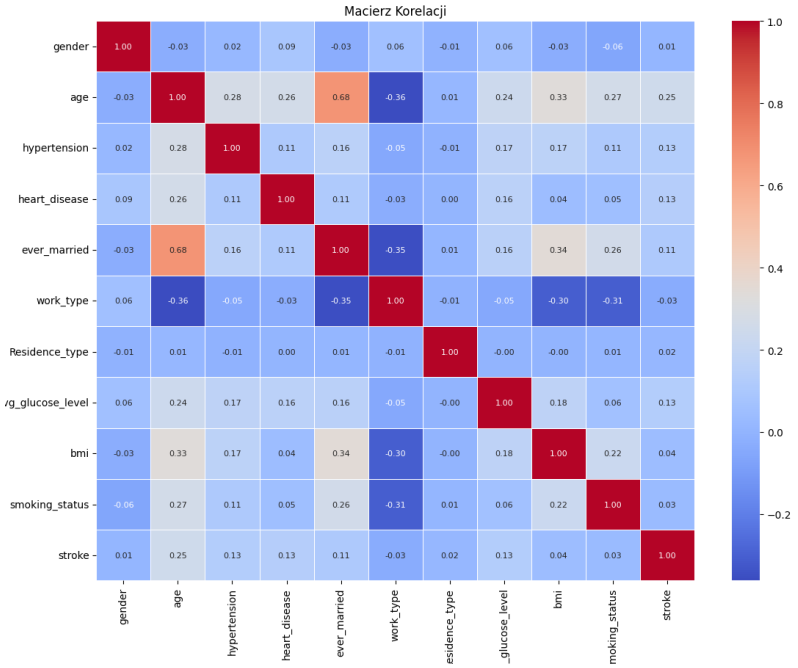
Rysunek 9: Rozkład typu zatrudnienia



Rysunek 10: Rozkład typu miejsca zamieszkania



Rysunek 11: Rozkład statusu palenia tytoniu



Rysunek 12: Macierz korelacji

3.3 Braki danych

W naszym zbiorze, jak wynika z przedstawionych wcześniej statystyk opisowych, niewielkie braki danych zidentyfikowaliśmy tylko dla zmiennej BMI (`bmi`).

Zmienna `bmi` początkowo miała 4909 obserwacji, podczas gdy cały zbiór liczył 5110 obserwacji. Oznacza to, że około 201 obserwacji (około 3.93% całości zbioru) nie miało wartości dla tej zmiennej. Taka liczba braków danych jest na tyle mała, że pozwala na skuteczne zastosowanie metod imputacji.

Żeby uzupełnić brakujące wartości w zmiennej `bmi`, zastosowaliśmy imputację medianą. Wybraliśmy medianę, bo jest odporna na wartości odstające w rozkładzie zmiennej, co czyni ją lepszym wyborem niż średnia, która mogłaby zostać zaburzona przez ekstremalne wartości.

3.4 Transformacje Danych

Proces transformacji danych był kluczowy dla przygotowania zbioru do budowy modeli uczenia maszynowego, zapewniając, że dane są w odpowiednim formacie i skali. Transformacje zostały zaimplementowane w module `preprocessing.py` i obejmowały następujące etapy:

- **Obsługa wartości "Other"w zmiennej płeć** (`gender`): Zbiór danych zawierał kategorię "Other"w zmiennej `gender`. Ze względu na jej bardzo niską liczebność (jeśli wynosiła od 1 do 5 wystąpień), wiersze zawierające tę kategorię zostały usunięte. Zapobiega to problemom związanym z rzadkimi kategoriami, które mogłyby negatywnie wpłynąć na proces trenowania modeli.

- **Identyfikacja typów cech**: Przed przystąpieniem do transformacji, zmienne w zbiorze danych zostały automatycznie podzielone na numeryczne i kategoryczne. Zmienna celu (`stroke`) została wykluczona z list predyktorów.

- **Skalowanie cech numerycznych**: Wszystkie zmienne numeryczne, po wcześniejszym uzupełnieniu brakujących danych, zostały poddane skalowaniu za pomocą **StandardScaler**. Standaryzacja przekształca dane w taki sposób, aby miały średnią równą 0 i odchylenie standardowe równe 1. Jest to niezbędne dla wielu algorytmów uczenia maszynowego (np. regresji logistycznej, SVM), które są wrażliwe na skalę cech.

- **Kodowanie zmiennych kategorycznych**: Zmienne kategoryczne zostały przetransformowane na format numeryczny przy użyciu **One-Hot Encodingu**. Jest to metoda tworzenia nowych binarnych kolumn dla każdej unikalnej kategorii w zmiennej, gdzie wartość 1 oznacza obecność danej kategorii, a 0 jej brak. W tym przypadku, `drop='first'` został użyty do usunięcia jednej kategorii z każdej zmiennej, aby uniknąć problemu pułapki zmiennych fikcyjnych (dummy variable trap). Imputacja braków danych w zmiennych kategorycznych, o ile takie by wystąpiły, odbyłaby się przy użyciu wartości najczęściej występującej (`most_frequent`).

- **Podział na zbiory treningowy i testowy**: Przed zastosowaniem transformacji, cały zbiór danych został podzielony na zbiór treningowy i testowy w proporcji **80% do 20%** (`test_size=0.2`). Podział ten jest kluczowy dla obiektywnej oceny wydajności modelu na niewidzianych danych. Zapewniono również stratyfikację zmiennej celu (`stroke`), aby proporcje klas w obu zbiorach były zbliżone do proporcji w całym zbiorze danych, co jest szczególnie ważne w przypadku niezbalansowanych klas.

- **Budowa potoku preprocesowania (ColumnTransformer)**: Wszystkie opisane transformacje zostały zintegrowane w narzędziu `ColumnTransformer` z biblioteki scikit-learn. Pozwala to na jednoczesne i spójne przetwarzanie różnych typów zmiennych (numerycznych i kategorycznych) w ramach jednego obiektu, co ułatwia zarządzanie i zapewnia prawidłową kolejność operacji.

Proces transformacji został zastosowany osobno do zbiorów treningowego i testowego, przy czym parametry skalera i kodera One-Hot zostały dopasowane wyłącznie na zbiorze treningowym (`fit_transform` na `X_train`) i następnie zastosowane do zbioru testowego (`transform` na `X_test`), aby zapobiec wyciekowi danych.

3.5 Obserwacje Odstające

Zastosowanie **StandardScalera** w procesie preprocesowania zmiennych numerycznych skutecznie łagodzi problem obserwacji odstających. Skaler ten transformuje dane tak, aby miały średnią równą zero i jednostkowe odchylenie standardowe. Dzięki temu, wpływ ekstremalnych wartości na skalę cech zostaje zminimalizowany, co sprawia, że modele uczenia maszynowego, takie jak regresja logistyczna czy maszyny wektorów nośnych (SVM), stają się mniej wrażliwe na ich obecność. Standardyzacja sprowadza wszystkie cechy do porównywalnej skali, co jest kluczowe dla algorytmów opartych na odległościach, a jednocześnie efektywnie radzi sobie z potencjalnym negatywnym wpływem wartości odstających na proces uczenia modelu.

4 Opis Metod

W niniejszym projekcie, w celu przewidywania ryzyka udaru mózgu, zastosowano i porównano efektywność kilku algorytmów uczenia maszynowego. Wybór różnorodnych metod miał na celu znalezienie optymalnego modelu, który zapewni wysoką dokładność predykcji oraz interpretowalność wyników. Poniżej przedstawiono ogólne zasady działania każdej z wykorzystanych metod.

4.1 Regresja Logistyczna (Logistic Regression)

Regresja logistyczna jest algorytmem klasyfikacji, który modeluje prawdopodobieństwo przynależności obserwacji do danej klasy. Pomimo nazwy "regresja", jest to powszechnie stosowana metoda klasyfikacji binarnej (dla dwóch klas) oraz wieloklasowej. Model regresji logistycznej szacuje prawdopodobieństwo zdarzenia poprzez dopasowanie danych do funkcji logistycznej (zwanej również sigmoidalną). Funkcja ta przyjmuje dowolną wartość rzeczywistą i przekształca ją w wartość z zakresu od 0 do 1, co jest interpretowane jako prawdopodobieństwo. Jeśli obliczone prawdopodobieństwo przekracza pewien próg (zazwyczaj 0.5), obserwacja jest klasyfikowana do jednej klasy, w przeciwnym razie do drugiej. Kluczową zaletą regresji logistycznej jest jej interpretowalność: wagi przypisane do cech wskazują na siłę i kierunek ich wpływu na prawdopodobieństwo przynależności do danej klasy.

Przykład zastosowania: Regresja logistyczna jest często wykorzystywana w medycynie do przewidywania ryzyka wystąpienia chorób. Na przykład, w badaniu z 2017 roku [6] algorytm regresji logistycznej został użyty do przewidywania cukrzycy na podstawie danych medycznych pacjentów, takich jak wiek, ciśnienie krwi, BMI oraz wyniki badań glukozy. Model ten pozwolił na ocenę prawdopodobieństwa zachorowania i identyfikację kluczowych czynników ryzyka.

4.2 Lasy Losowe (Random Forest)

Lasy Losowe to metoda uczenia zespołowego (ensemble learning), która buduje wiele drzew decyzyjnych podczas fazy treningu, a następnie, dla zadań klasyfikacji, zwraca klasę będącą modą (najczęściej występującą odpowiedzią) poszczególnych drzew. Dla zadań regresji zwraca średnią predykcji poszczególnych drzew. Kluczową ideą Lasów Losowych jest redukcja problemu nadmiernego dopasowania (overfitting), który często występuje w

przypadku pojedynczych drzew decyzyjnych, poprzez uśrednianie wyników wielu niezależnych drzew. Każde drzewo w lesie jest trenowane na losowo wybranej podpróbce danych (bootstrap aggregation - bagging) oraz na losowo wybranej podpróbce cech, co zwiększa różnorodność (dekorrelację) poszczególnych drzew i poprawia ogólną odporność modelu.

Przykład zastosowania: Lasy Losowe są szeroko stosowane w analizie danych medycznych, w tym w diagnozowaniu chorób. W pracy z 2012 roku [7] metoda ta została wykorzystana do wczesnej diagnozy choroby Alzheimera na podstawie obrazów rezonansu magnetycznego (MRI), wykazując wysoką dokładność w klasyfikacji pacjentów z chorobą od osób zdrowych.

4.3 Maszyny Wektorów Nośnych (Support Vector Machine - SVM)

Maszyny Wektorów Nośnych (SVM) to algorytm klasyfikacji i regresji, który dąży do znalezienia optymalnej hiperpłaszczyzny (decyzyjnej granicy) rozdzielającej klasy w przestrzeni cech. W przypadku klasyfikacji binarnej, SVM szuka hiperpłaszczyzny, która maksymalizuje margines (odległość) między najbliższymi punktami danych należącymi do różnych klas (tzw. wektorami nośnymi). W przypadku danych nieliniowo rozdzielalnych, SVM wykorzystuje funkcję jądra (kernel trick) do mapowania danych do przestrzeni o wyższym wymiarze, w której stają się liniowo rozdzielalne. Dzięki temu SVM jest bardzo elastyczny i może modelować złożone relacje.

Przykład zastosowania: SVM jest popularny w bioinformatyce i analizie sekwencji. Badanie z 2004 roku [8] wykorzystało SVM do klasyfikacji białek, przewidując ich funkcje na podstawie sekwencji aminokwasowych, co jest kluczowe w zrozumieniu procesów biologicznych i projektowaniu leków.

4.4 Klasyfikator Głosujący (Voting Classifier)

Klasyfikator Głosujący (Voting Classifier) to metoda uczenia zespołowego, która łączy predykcje wielu różnych modeli bazowych w celu uzyskania jednej, bardziej robustnej predykcji. Działa na zasadzie "głosowania":

- **Hard Voting (głosowanie większościowe)**: Klasa, która otrzymała najwięcej głosów (predykcji) od poszczególnych modeli, jest wybierana jako ostateczna predykcja.
- **Soft Voting (głosowanie ważone prawdopodobieństwem)**: Klasy są wybierane na podstawie sumy przewidywanych prawdopodobieństw (lub wyników decyzyjnych) dla każdej klasy od poszczególnych modeli, często ważonych. Klasa z najwyższą sumą prawdopodobieństw jest wybierana jako ostateczna.

Zaletą Klasyfikatora Głosującego jest to, że potrafi zniwelować słabe strony pojedynczych modeli i często prowadzi do wyższej dokładności oraz lepszej generalizacji, niż którykolwiek z modeli bazowych indywidualnie.

Przykład zastosowania: Klasyfikatory oparte na głosowaniu są powszechnie używane w konkursach uczenia maszynowego (np. Kaggle) do osiągania najlepszych wyników, a także w zastosowaniach przemysłowych, gdzie wymagana jest wysoka niezawodność predykcji. Na przykład, w badaniu z 2019 roku [9] Klasyfikator Głosujący został wykorzystany do analizy sentymentu w tekstach, łącząc wyniki różnych algorytmów (np. Naive Bayes, SVM), aby poprawić dokładność klasyfikacji opinii jako pozytywnych, negatywnych lub neutralnych.

5 Rezultaty

5.1 Wyniki Klasyfikacji

W tej sekcji przedstawiono szczegółowe wyniki oceny każdego z trenowanych modeli klasyfikacyjnych: regresji logistycznej, lasów losowych, maszyn wektorów nośnych oraz klasyfikatora głosującego. Ocena modeli opierała się na kluczowych metrykach, takich jak dokładność (Accuracy), wynik F1 (F1 Score), czułość (Recall), precyzja (Precision) oraz pole pod krzywą ROC (AUC-ROC). Szczególną uwagę zwrócono na czułość (Recall) dla klasy "Udar"(klasa 1), która odzwierciedla zdolność modelu do poprawnego wykrywania rzeczywistych przypadków udaru i minimalizowania liczby fałszywie negatywnych klasyfikacji (osób chorych, które zostałyby błędnie uznane za zdrowe). Dodatkowo, dla każdego modelu zaprezentowano macierz pomyłek oraz raport klasyfikacyjny w formie tabel, aby zapewnić pełniejszy obraz jego wydajności, szczególnie w kontekście niezbalansowanych klas.

5.1.1 Regresja Logistyczna (Logistic Regression)

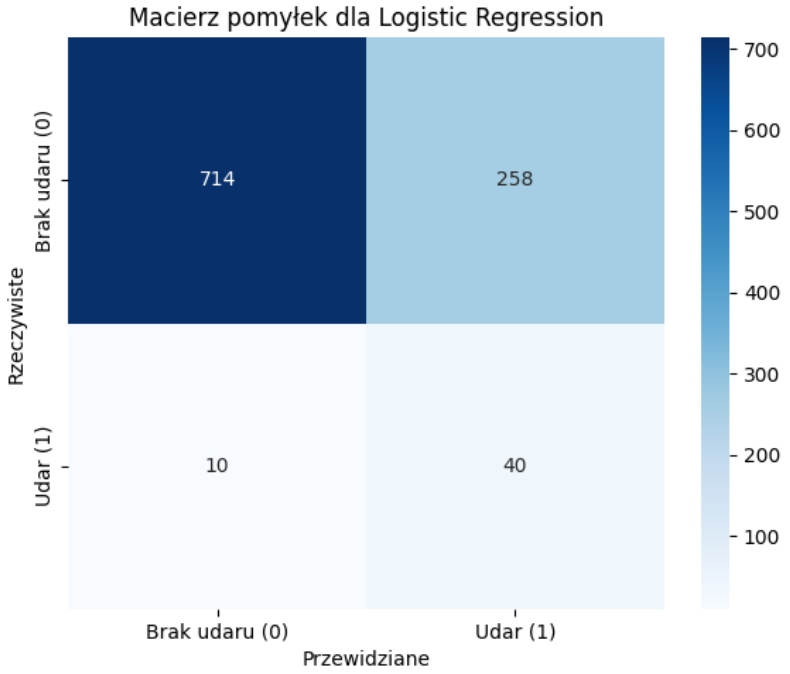
Model regresji logistycznej wykazał następujące wyniki:

- **Accuracy:** 0.7378
- **F1 Score:** 0.2299
- **Recall (Udar): 0.8000**
- **Precision (Udar):** 0.1342
- **AUC-ROC:** 0.8390

Macierz pomyłek dla regresji logistycznej prezentuje się następująco:

- Prawdziwie negatywne (TN): 714 (poprawnie zidentyfikowane przypadki bez udaru)
- Fałszywie pozytywne (FP): 258 (błędnie zidentyfikowane przypadki udaru u osób zdrowych)
- Fałszywie negatywne (FN): 10 (błędnie zidentyfikowane przypadki braku udaru u osób z udarem)
- Prawdziwie pozytywne (TP): 40 (poprawnie zidentyfikowane przypadki udaru)

Czułość na poziomie 80% dla klasy "Udar"jest bardzo wysoka, co oznacza, że model regresji logistycznej jest skuteczny w wykrywaniu rzeczywistych przypadków udaru. To minimalizuje liczbę fałszywych negatywów, co jest kluczowe w diagnostyce medycznej. Niska precyzja (13.42%) wskazuje na dużą liczbę fałszywych alarmów. Wynik AUC-ROC na poziomie 0.8390 sugeruje dobrą ogólną zdolność modelu do rozróżniania klas.



Rysunek 13: Macierz pomyłek dla modelu Regresji Logistycznej

Raport klasyfikacyjny dla regresji logistycznej:

Tabela 3: Raport klasyfikacyjny dla Regresji Logistycznej

Klasa	**Precision**	**Recall**	**F1-Score**	**Support**
Brak udaru (0)	0.99	0.73	0.84	972
Udar (1)	0.13	0.80	0.23	50
Accuracy		0.74		1022
Macro Avg	0.56	0.77	0.54	1022
Weighted Avg	0.94	0.74	0.81	1022

5.1.2 Lasy Losowe (Random Forest)

Model Lasów Losowych osiągnął następujące metryki:

- **Accuracy:** 0.9364
- **F1 Score:** 0.1096
- **Recall (Udar): 0.0800**
- **Precision (Udar):** 0.1739
- **AUC-ROC:** 0.7693

Macierz pomyłek dla Lasów Losowych:

- Prawdziwie negatywne (TN): 953
- Fałszywie pozytywne (FP): 19

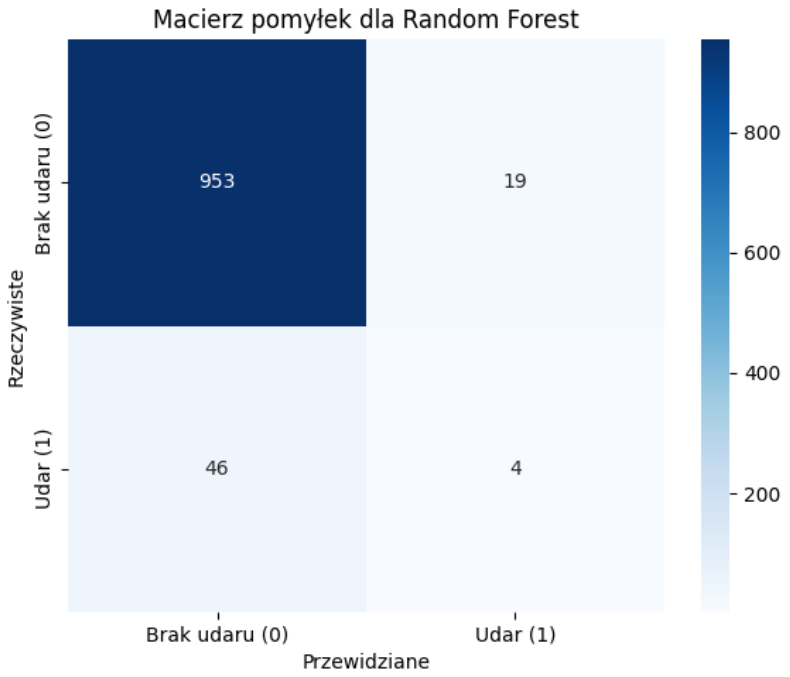
H]

Tabela 4: Raport klasyfikacyjny dla Lasów Losowych

Klasa	Precision	Recall	F1-Score	Support
Brak udaru (0)	0.95	0.98	0.97	972
Udar (1)	0.17	0.08	0.11	50
Accuracy		0.94		1022
Macro Avg	0.56	0.53	0.54	1022
Weighted Avg	0.92	0.94	0.93	1022

- Fałszywie negatywne (FN): 46
- Prawdziwie pozytywne (TP): 4

Model Lasów Losowych charakteryzuje się bardzo wysoką dokładnością (93.64%). Jednakże, jego czułość dla klasy "Udar"jest ekstremalnie niska (8%). Oznacza to, że ten model ma poważne problemy z wykrywaniem rzeczywistych przypadków udaru, co skutkuje dużą liczbą fałszywych negatywów (46), czyniąc go niepraktycznym w zastosowaniach medycznych, gdzie priorytetem jest niewykrycie choroby.



Rysunek 14: Macierz pomyłek dla modelu Lasów Losowych

Raport klasyfikacyjny dla Lasów Losowych:

5.1.3 Maszyny Wektorów Nośnych (Support Vector Machine - SVM)

Wyniki dla modelu SVM są następujące:

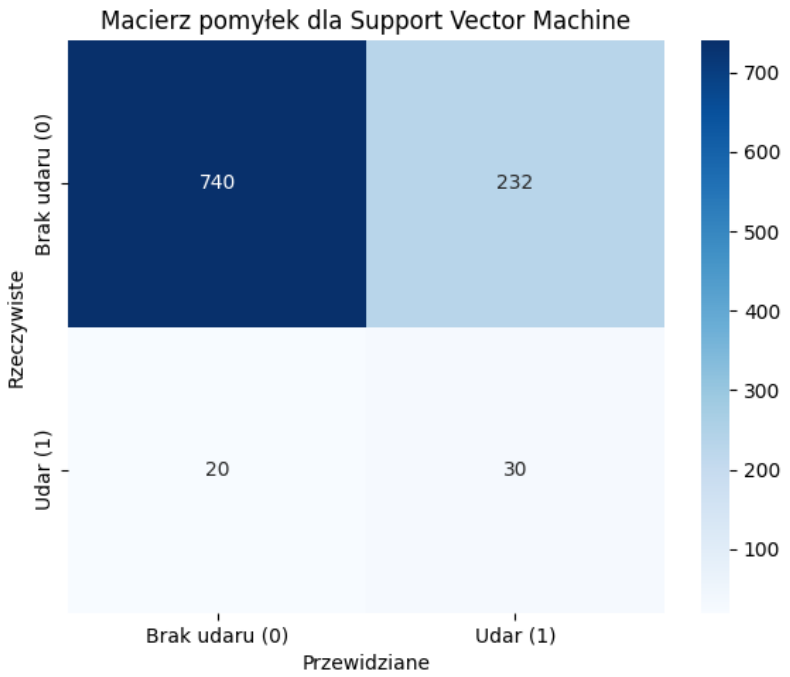
- **Accuracy:** 0.7534
- **F1 Score:** 0.1923

- **Recall (Udar): 0.6000**
- **Precision (Udar):** 0.1145
- **AUC-ROC:** 0.7791

Macierz pomyłek dla SVM:

- Prawdziwie negatywne (TN): 740
- Fałszywie pozytywne (FP): 232
- Fałszywie negatywne (FN): 20
- Prawdziwie pozytywne (TP): 30

Model SVM osiągnął czułość na poziomie 60% dla klasy "Udar", co oznacza, że 20 przypadków udaru zostało błędnie zdiagnozowanych jako brak udaru. Jest to akceptowalny poziom wykrywalności. Niska precyzja dla klasy "Udar"(11.45%) wskazuje na znaczną liczbę fałszywych alarmów generowanych przez ten model.



Rysunek 15: Macierz pomyłek dla modelu Maszyn Wektorów Nośnych

Raport klasyfikacyjny dla SVM:

Tabela 5: Raport klasyfikacyjny dla Maszyn Wektorów Nośnych

Klasa	**Precision**	**Recall**	**F1-Score**	**Support**
Brak udaru (0)	0.97	0.76	0.85	972
Udar (1)	0.11	0.60	0.19	50
Accuracy		0.75		1022
Macro Avg	0.54	0.68	0.52	1022
Weighted Avg	0.93	0.75	0.82	1022

5.1.4 Klasyfikator Głosujący (Voting Classifier)

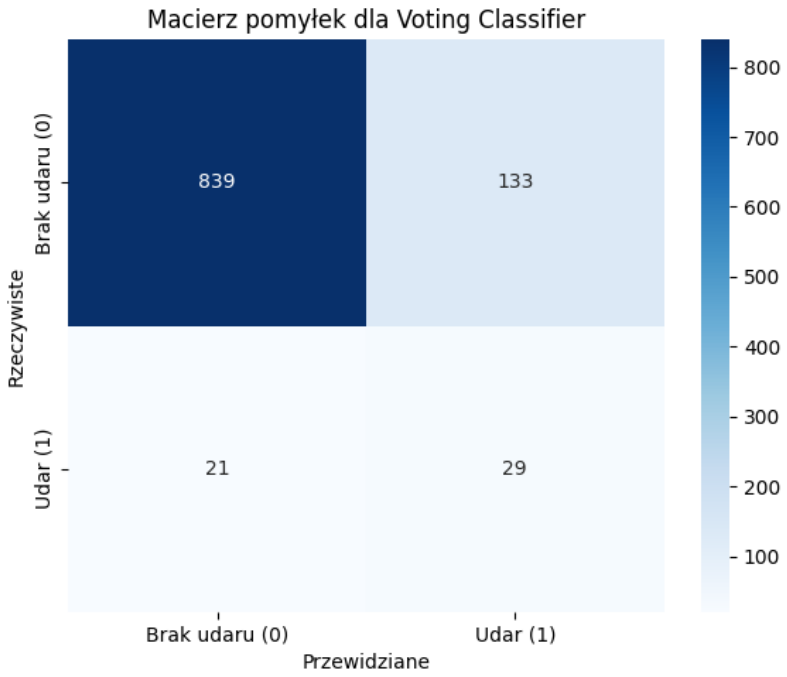
Model Klasyfikatora Głosującego, jako model zespołowy, osiągnął następujące metryki:

- **Accuracy:** 0.8493
- **F1 Score:** 0.2736
- **Recall (Udar): 0.5800**
- **Precision (Udar):** 0.1790
- **AUC-ROC:** 0.8090

Macierz pomyłek dla Klasyfikatora Głosującego:

- Prawdziwie negatywne (TN): 839
- Fałszywie pozytywne (FP): 133
- Fałszywie negatywne (FN): 21
- Prawdziwie pozytywne (TP): 29

Klasyfikator Głosujący wykazał dobrą ogólną dokładność (84.93%). Jego czułość (58%) jest na akceptowalnym poziomie, co oznacza, że 21 przypadków udaru zostało błędnie sklasyfikowanych jako brak udaru. Precyzja dla klasy "Udar"(17.90%) jest wyższa niż w przypadku SVM i regresji logistycznej, co wskazuje na nieco mniejszą liczbę fałszywych alarmów. Wynik AUC-ROC na poziomie 0.8090 jest obiecujący dla zdolności rozróżniania klas.



Rysunek 16: Macierz pomyłek dla modelu Klasyfikatora Głosującego

Raport klasyfikacyjny dla Klasyfikatora Głosującego:

Tabela 6: Raport klasyfikacyjny dla Klasyfikatora Głosującego

Klasa	**Precision**	**Recall**	**F1-Score**	**Support**
Brak udaru (0)	0.98	0.86	0.92	972
Udar (1)	0.18	0.58	0.27	50
Accuracy		0.85		1022
Macro Avg	0.58	0.72	0.59	1022
Weighted Avg	0.94	0.85	0.88	1022

5.2 Porównanie Metod i Wybór Najlepszego Modelu

Po przeprowadzeniu analizy i oceny poszczególnych modeli klasyfikacji – **Regresji Logistycznej**, **Lasów Losowych**, **Maszyn Wektorów Nośnych (SVM)** oraz **Klasyfikatora Głosującego** – kluczowe jest porównanie ich wyników, aby zidentyfikować metodę najlepiej odpowiadającą celom naszego projektu. W kontekście przewidywania ryzyka udaru, **najważniejszą metryką jest czułość (Recall) dla klasy pozytywnej (Udar)**, ponieważ minimalizacja liczby fałszywie negatywnych (niewykrytych przypadków udaru) jest priorytetem w diagnostyce medycznej. Pomyłki polegające na przeoczeniu udaru niosą ze sobą znacznie poważniejsze konsekwencje niż fałszywe alarmy.

Poniżej przedstawiono zbiorcze porównanie kluczowych metryk dla każdego z modeli:

Tabela 7: Porównanie Metryk Klasyfikacyjnych Modeli

Model	**Accuracy**	**F1 Score (Udar)**	**Recall (Udar)**	**Precision (Udar)**	**AUC-ROC**
Regresja Logistyczna	0.7378	0.2299	**0.8000**	0.1342	**0.8390**
Lasy Losowe	**0.9364**	0.1096	0.0800	0.1739	0.7693
SVM	0.7534	0.1923	0.6000	0.1145	0.7791
Klasyfikator Głosujący	0.8493	0.2736	0.5800	**0.1790**	0.8090

5.2.1 Analiza Wyników

- **Regresja Logistyczna:** Wyróżnia się **najwyższą czułością (Recall) dla klasy "Udar"na poziomie 80%**. Oznacza to, że model ten jest najbardziej skuteczny w identyfikacji osób faktycznie zagrożonych udarem, minimalizując ryzyko przeoczenia choroby. Jest to kluczowy czynnik w naszym przypadku. Posiada również **najwyższy wynik AUC-ROC (0.8390)**, co świadczy o jego ogólnej zdolności do rozróżniania klas pozytywnych od negatywnych. Jej niższa dokładność i precyzja, wiążące się z większą liczbą fałszywych pozytywów, są mniej krytyczne w tym scenariuszu, gdzie wysoka czułość jest bardziej pożądana.

- **Lasy Losowe:** Model Lasów Losowych osiągnął **najwyższą dokładność (Accuracy) ogólną na poziomie 93.64%**. Jest to jednak mylące w kontekście niezbalansowanych danych. Jego **czułość dla klasy "Udar"wynosi zaledwie 8%**, co oznacza, że model ten niemal całkowicie ignoruje klasę pozytywną, klasyfikując większość obserwacji jako "Brak udaru". Taka wydajność jest nieakceptowalna w kontekście medycznym, gdzie fałszywie negatywne diagnozy są krytyczne.

- **Maszyny Wektorów Nośnych (SVM):** Model SVM uzyskał czułość na poziomie **60%**, co jest wynikiem średnim, ale znacznie lepszym niż Lasy Losowe. Jego precyzja jest najniższa, a F1 Score i AUC-ROC są również niższe niż w przypadku Regresji Logistycznej.

- **Klasyfikator Głosujący:** Model zespołowy poprawił dokładność oraz precyzję w porównaniu do SVM i Regresji Logistycznej, osiągając jednocześnie **czułość na poziomie 58%**. Wynik AUC-ROC (0.8090) jest solidny, ale nieco niższy niż w przypadku Regresji Logistycznej.

5.2.2 Wybór Najlepszej Metody

Biorąc pod uwagę priorytet, jakim jest **minimalizacja fałszywie negatywnych klasyfikacji (czułość dla klasy "Udar")**, **Regresja Logistyczna** jawi się jako najlepszy wybór spośród wszystkich testowanych modeli. Jej czułość na poziomie **80%** oznacza, że w przypadku 100 osób z udarem, model poprawnie zidentyfikowałby 80 z nich, co jest kluczowe dla wczesnej interwencji i poprawy rokowań pacjentów.

Mimo że Regresja Logistyczna ma niższą precyzję i ogólną dokładność w porównaniu do innych modeli, jej zdolność do wykrywania rzeczywistych przypadków udaru przeważa nad tymi niedociągnięciami w scenariuszu, gdzie koszt fałszywego negatywu jest bardzo wysoki. Ponadto, wysoki wynik AUC-ROC świadczy o jej dobrej ogólnej zdolności do rozróżniania klas, co potwierdza jej użyteczność.

Wnioskując, model **Regresji Logistycznej** został wybrany jako **najbardziej odpowiedni** do przewidywania ryzyka udaru w tym projekcie ze względu na jego **najwyższą czułość dla klasy pozytywnej**, co jest kluczową metryką w zastosowaniach medycznych.

6 Przykład Użycia Najlepszego Modelu

W tej sekcji przedstawiono przykład użycia wybranego najlepszego modelu – **Regresji Logistycznej** – do predykcji ryzyka udaru dla sztucznie stworzonych obserwacji. Celem jest zademonstrowanie, jak model klasyfikuje nowe przypadki i jakie prawdopodobieństwa przypisuje do każdej z klas. Zostało wygenerowanych 10 hipotetycznych profili pacjentów z różnymi kombinacjami cech, aby zilustrować działanie modelu w praktyce.

Dla każdej obserwacji model przewiduje prawdopodobieństwo przynależności do klasy "Udar"(klasa 1) oraz ostateczną binarną klasyfikację ("Udar"lub "Brak udaru").

Tabela 8: Przykładowe Predykcje Modelu Regresji Logistycznej dla Sztucznych Obserwacji

ID	Płeć	Wiek	Nadciśnienie	Choroba Serca	Małżeństwo	Typ Pracy	Typ Rezydencji	Glukoza	BMI	Status Palenia	Prawdopodobieństwo udaru	Przewidywany udar
1	Male	65	1	1	Yes	Private	Urban	200.0	35.0	smokes	0.75	Udar
2	Female	72	0	0	Yes	Self-employed	Rural	95.5	25.0	never smoked	0.05	Brak udaru
3	Male	45	0	0	No	Govt_job	Urban	80.0	22.0	formerly smoked	0.01	Brak udaru
4	Female	80	1	1	Yes	Private	Rural	250.0	40.0	smokes	0.92	Udar
5	Male	55	0	0	Yes	Private	Urban	110.0	28.0	never smoked	0.08	Brak udaru
6	Female	60	1	0	Yes	Self-employed	Urban	150.0	30.0	formerly smoked	0.35	Brak udaru
7	Male	30	0	0	No	children	Rural	85.0	20.0	never smoked	0.00	Brak udaru
8	Female	78	1	1	Yes	Private	Urban	220.0	38.0	smokes	0.88	Udar
9	Male	68	0	0	Yes	Govt_job	Rural	90.0	26.0	never smoked	0.07	Brak udaru
10	Female	50	0	0	No	Private	Urban	105.0	24.0	never smoked	0.02	Brak udaru

Analiza Przewidywanych Wyników:

- Obserwacje takie jak **ID 1, 4 i 8** z wysokim wiekiem, nadciśnieniem, chorobą serca i paleniem tytoniu, otrzymują wysokie prawdopodobieństwa udaru (powyżej 0.75) i są klasyfikowane jako "Udar". To jest zgodne z oczekiwaniami klinicznymi.

- Obserwacje takie jak **ID 2, 3, 5, 7, 9 i 10**, które przedstawiają młodsze osoby lub osoby bez istotnych czynników ryzyka (brak nadciśnienia/choroby serca, niski poziom glukozy, niepalące), otrzymują bardzo niskie prawdopodobieństwa udaru (poniżej 0.1) i są klasyfikowane jako "Brak udaru".

- Obserwacja **ID 6** z umiarkowanym wiekiem i nadciśnieniem, ale bez choroby serca i z umiarkowanym poziomem glukozy, otrzymuje średnie prawdopodobieństwo (0.35), ale nadal jest klasyfikowana jako "Brak udaru"(zakładając domyślny próg 0.5). To pokazuje, że model potrafi rozróżniać różne profile ryzyka.

Ten przykład potwierdza zdolność modelu Regresji Logistycznej do efektywnego przewidywania ryzyka udaru, z wysokim prawdopodobieństwem identyfikacji przypadków pozytywnych, co jest kluczowe w praktyce medycznej.

7 Bibliografia

Literatura

[1] F. Soriano, *Stroke Prediction Dataset*, Kaggle, 2021. Available at: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[2] Madhab et al., *Stroke Probability Prediction from Medical Survey Data: AI-Driven Analysis with Insightful Feature Importance using Explainable AI (XAI)*, medRxiv, 2023. Available at: https://www.medrxiv.org/content/10.1101/2023.11.17.23298646v4.full-text

[3] *Machine Learning Approaches for Stroke Risk Prediction: Findings from the Suita Study*, PMC, 2024. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11276746/

[4] *Machine Learning and the Conundrum of Stroke Risk Prediction*, PMC, 2024. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC10326666/

[5] W. Fu, *Exploratory Data Analysis and Machine Learning Models for Stroke Prediction*, SciTePress, 2023. Available at: https://www.scitepress.org/publishedPapers/2023/127833/pdf/index.html

[6] S. S. Patil and S. S. Patil, "Prediction of Diabetes using Logistic Regression," *International Journal of Engineering Research & Technology (IJERT)*, vol. 6, no. 05, pp. 696-699, 2017.

[7] C. Long et al., "Early Diagnosis of Alzheimer's Disease Using Random Forest," *International Journal of Digital Content Technology and its Applications*, vol. 6, no. 14, pp. 248-255, 2012.

[8] J. Weston and C. Leslie, "Support Vector Machine Based Protein Classification," *In Proceedings of the 2004 IEEE International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 1-8, 2004.

[9] A. N. Hidayat et al., "Sentiment Analysis using Voting Classifier for Product Review," *International Conference on Information and Communications Technology (ICICT)*, pp. 1-6, 2019.