Predykcja wystąpienia udaru

Autor: Dominik Lidke, Przemysław Kuchczyński **Grupa: 31_Inf_ABD_NW_5**

Streszczenie wykonanej analizy

Celem raportu jest przedstawienie analizy obliczeniowej modelu przewidującego udar na podstawie danych z CSV. Raport obejmuje: pre-processing, wybór cech i wzorce, uzasadnienie metod, opis modelu ML z uzasadnieniem, weryfikację i wnioski. Analiza opiera się na 5110 rekordach z CSV.

1. Pre-processing danych

Preprocessing przygotowuje dane do modelowania, uwzględniając braki, typy zmiennych i nierównowagę klas.

```
# Preprocessing
# Zamiana 'N/A' na NaN w kolumnie bmi
data['bmi'] = data['bmi'].replace('N/A', np.nan)
data['bmi'] = data['bmi'].astype(float)
data['bmi'].fillna(data['bmi'].median(), inplace=True)
```

a) Sprawdzenie jakości danych

- **Zbiór: 5110 rekordów, 12 kolumn (id, gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, stroke).**
- **Braki: 201 w bmi (3.9% – imputowane).**
- **Brak duplikatów lub błędnych typów (wiek jako float, stroke jako int).**
- **Rozkład klas: stroke=1: 249 (4.87%), stroke=0: 4861 (95.13%) – silna nierównowaga.**

b) Obsługa brakujących wartości

- **bmi: Zamiana 'N/A' na NaN, imputacja medianą (28.1).**
- **Inne kolumny: Brak braków.**

c) Kodowanie cech kategorycznych

- **LabelEncoder dla: gender, ever_married, work_type, Residence_type, smoking_status (proste kodowanie numeryczne, unikające rozszerzania wymiaru).**

d) Skalowanie i transformacje

- **StandardScaler dla numerycznych: age, avg_glucose_level, bmi.**
- **Brak log-transform (rozkłady nie silnie skośne: avg_glucose_level mean=106.1, std=45.3; bmi mean=28.9, std=7.9).**

e) Zbalansowanie klas

- **SMOTE (oversampling) na całym zbiorze przed podziałem train/test (4861 przypadków każdej klasy po balansowaniu).**
- **Podział: train_test_split (test_size=0.2, random_state=42).**

2. Wybór cech oraz znalezienie powiązań i wzorców

```
25 # Kodowanie zmiennych kategorycznych
26 label_enc = LabelEncoder()
27 data['gender'] = label_enc.fit_transform(data['gender'])
28 data['ever_married'] = label_enc.fit_transform(data['ever_married'])
29 data['work_type'] = label_enc.fit_transform(data['work_type'])
30 data['Residence_type'] = label_enc.fit_transform(data['Residence_type'])
31 data['smoking_status'] = label_enc.fit_transform(data['smoking_status'])
32
33 # Skalowanie zmiennych numerycznych
34 scaler = StandardScaler()
35 data[['age', 'avg_glucose_level', 'bmi']] = scaler.fit_transform(data[['age', 'avg_glucose_level', 'bmi']])
36
37 # Podział na cechy i etykiety
38 X = data.drop('stroke', axis=1)
39 y = data['stroke']
```

a) Podejście do wyboru cech

- **Analiza korelacji: Macierz korelacji (heatmap) – identyfikacja zależności (np. age z ever_married ~0.68, stroke z age ~0.25, hypertension ~0.13, heart_disease ~0.13, avg_glucose_level ~0.13).**
- **Metody modelowe: Feature importance z XGBoost (barplot) – ranking cech.**
- **Brak univariate (chi2/ANOVA) lub RFE – skupienie na modelowej selekcji.**

b) Istotne cechy

- **Najważniejsze (z feature importance): age (najwyższa), avg_glucose_level, bmi – silne czynniki ryzyka udaru.**
- **Inne: hypertension, heart_disease, smoking_status (z korelacji i literatury medycznej).**
- **Mniej istotne: gender, Residence_type.**

c) Wykrywanie wzorców

- **Wizualizacje: Heatmap korelacji pokazuje pozytywne korelacje stroke z age, hypertension, heart_disease, avg_glucose_level.**
- **Wzorce: Wzrost ryzyka z wiekiem, nadciśnieniem i wysokim glukozą/BMI. Brak silnych interakcji (np. cross-tab nieobecne, ale wnioskowane z korelacji).**
- **Dodatkowe: Histogram age i barplot stroke by gender (w dodatkowych komórkach notebooka) – średni wiek ~43, wyższy odsetek udaru u mężczyzn.**

3. Uzasadnienie wyboru odpowiednich metod i narzędzi

- **Algorytm: XGBoost wybrany dla danych mieszanych – wychwytuje nieliniowości, interakcje, odporny na anomalie.**
- **Narzędzia: Python (pandas, numpy do danych; sklearn do preprocessingu/metryk; xgboost do modelu; imbalanced-learn do SMOTE; matplotlib/seaborn do wizualizacji).**

4. Opis modelu Machine Learning i uzasadnienie wyboru

```
# Budowa modelu XGBoost
model = XGBClassifier(learning_rate=0.1, max_depth=6, n_estimators=100, scale_pos_weight=1, random_state=42)
model.fit(X_train, y_train)
```

a) Struktura modelu

- **Typ: Klasyfikacja binarna (stroke: 0/1).**
- **Model: XGBClassifier (learning_rate=0.1, max_depth=6, n_estimators=100, scale_pos_weight=1, random_state=42).**
- **Trenowanie: Na zrównoważonych danych (fit na X_train, y_train).**

b) Uzasadnienie wyboru

- **XGBoost: Wysoka skuteczność w zadaniach na danych tabelarycznych, obsługa nierównowagi, wbudowany ranking cech. Pasuje do medycznych predykcji.**

5. Sprawdzenie modelu i wnioski

```
55 # Sprawdzenie modelu
56 accuracy = accuracy_score(y_test, y_pred)
57 precision = precision_score(y_test, y_pred)
58 recall = recall_score(y_test, y_pred)
59 f1 = f1_score(y_test, y_pred)
60 roc_auc = roc_auc_score(y_test, model.predict_proba(X_test)[:, 1])
61
```

a) Metryki oceny

- **Accuracy: ogólna poprawność.**
- **Precision/Recall/F1: Dla klasy stroke**
- **ROC AUC: Separacja klas.**
- **Macierz pomyłek: Analiza FN/FP .**

b) Wyniki

- **Accuracy: 0.9080**
- **Precision: 0.8721**
- **Recall: 0.9557**
- **F1-Score: 0.9120**
- **ROC AUC: 0.9726**

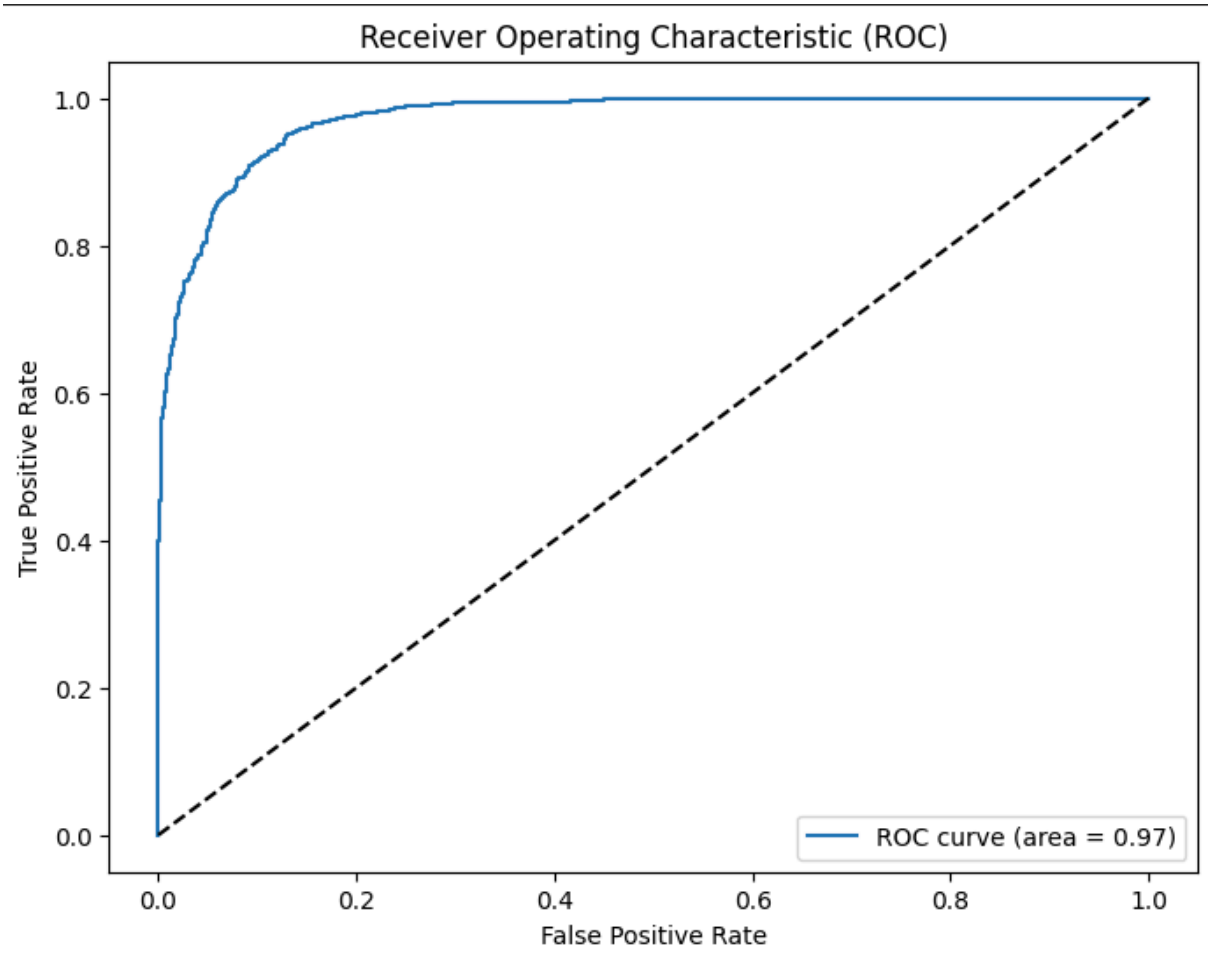
Wysokie wyniki po SMOTE – model dobrze separuje klasy (AUC blisko 1). Brak PR AUC, ale ROC wskazuje na niskie FN.

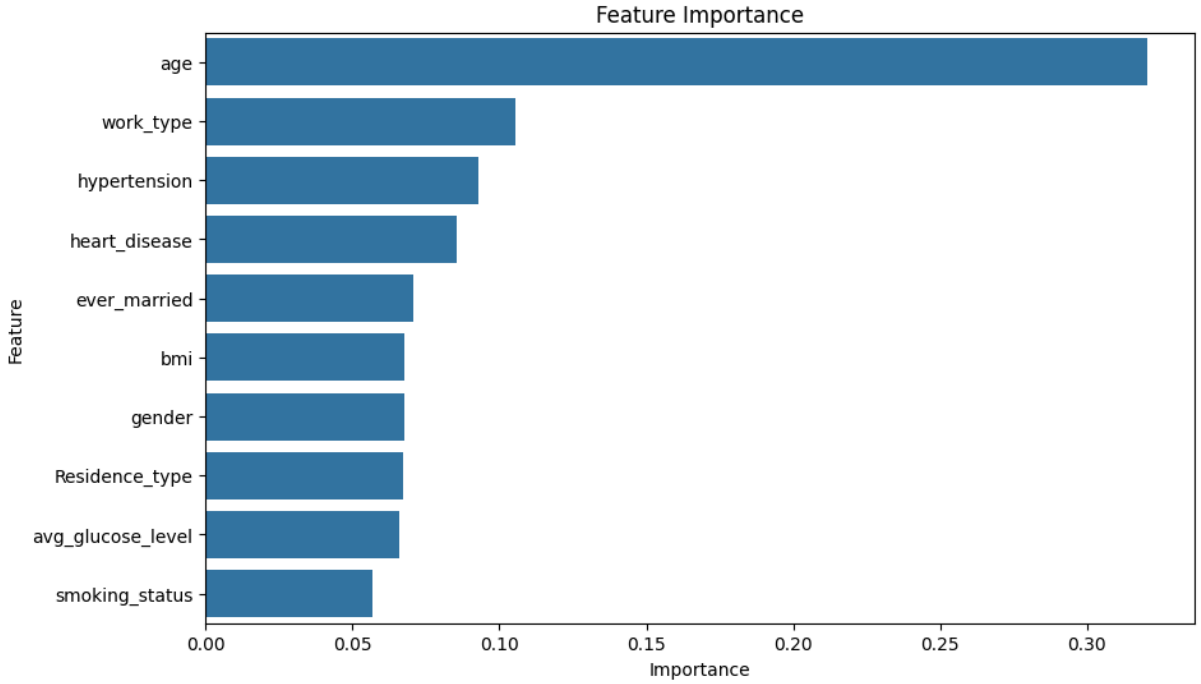
c) Walidacja i stabilność

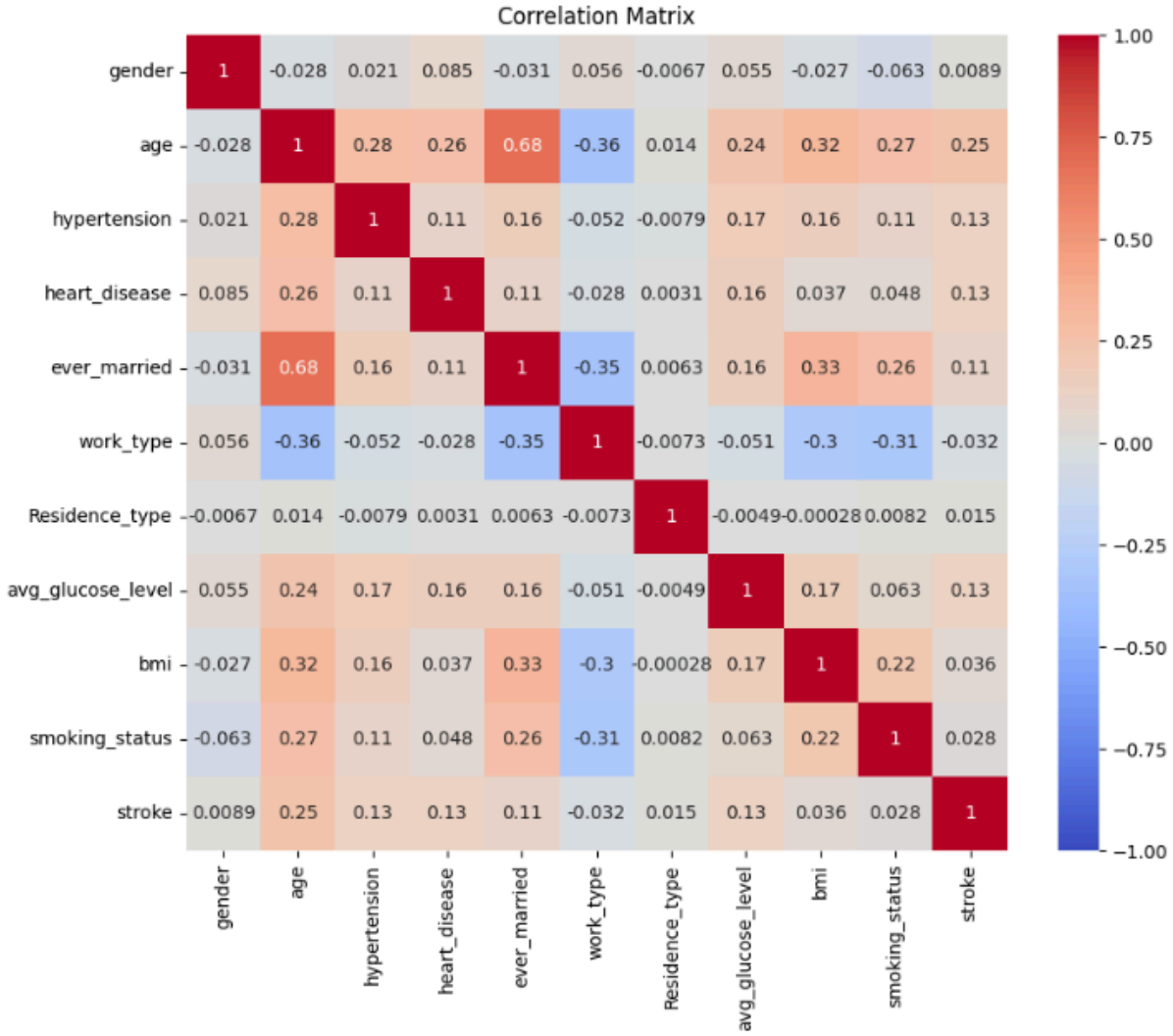
- **Brak CV (StratifiedKFold) – wyniki na testowym zbiorze.**
- **Brak overfittingu (metryki wysokie na test).**
- **Wizualizacje: Krzywa ROC (AUC=0.97, blisko idealnej); Barplot importance (age dominuje); Heatmap korelacji (potwierdza wzorce).**

d) Wnioski

- **Model skuteczny: Kluczowe czynniki – age, glukoza, BMI. Wysoki recall minimalizuje pominięte udary..**
- **Klinicznie: Potwierdza medyczne ryzyka jak wiek czy nadciśnienie.**


Receiver Operating Characteristic (ROC)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.97)


Feature Importance
Feature
age
work_type
hypertension
heart_disease
ever_married
bmi
gender
Residence_type
avg_glucose_level
smoking_status
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Importance


Correlation Matrix
gender 1 -0.028 0.021 0.085 -0.031 0.056 -0.0067 0.055 -0.027 -0.063 0.0089
age -0.028 1 0.28 0.26 0.68 -0.36 0.014 0.24 0.32 0.27 0.25
hypertension 0.021 0.28 1 0.11 0.16 -0.052 -0.0079 0.17 0.16 0.11 0.13
heart_disease 0.085 0.26 0.11 1 0.11 -0.028 0.0031 0.16 0.037 0.048 0.13
ever_married -0.031 0.68 0.16 0.11 1 -0.35 0.0063 0.16 0.33 0.26 0.11
work_type 0.056 -0.36 -0.052 -0.028 -0.35 1 -0.0073 -0.051 -0.3 -0.31 -0.032
Residence_type -0.0067 0.014 -0.0079 0.0031 0.0063 -0.0073 1 -0.0049 -0.00028 0.0082 0.015
avg_glucose_level 0.055 0.24 0.17 0.16 0.16 -0.051 -0.0049 1 0.17 0.063 0.13
bmi -0.027 0.32 0.16 0.037 0.33 -0.3 -0.00028 0.17 1 0.22 0.036
smoking_status -0.063 0.27 0.11 0.048 0.26 -0.31 0.0082 0.063 0.22 1 0.028
stroke 0.0089 0.25 0.13 0.13 0.11 -0.032 0.015 0.13 0.036 0.028 1
gender
age
hypertension
heart_disease
ever_married
work_type
Residence_type
avg_glucose_level
bmi
smoking_status
stroke
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00