SPRAWOZDANIE
Zajęcia: Uczenie Maszynowe
Prowadzący: prof. dr hab. Vasyl Martsenyuk

Laboratorium Nr 2 Data 26.02.2025 Temat: Praktyczne Zastosowanie Drzew Decyzyjnych i Metod Ensemble w Analizie Danych Wariant 10	Krzysztof Świerczek Informatyka II stopień, stacjonarne, 1semestr, gr. A

1. Polecenie: Praktyczne zastosowanie regresji liniowej w analizie danych.

2. Github:
https://github.com/Krzycho165/STUDIA

In [1]:
```
# 1. KS Opracować przepływ pracy uczenia maszynowego zagadnienia klasyfikacji (p
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier, export_text
from sklearn.metrics import classification_report, accuracy_score

data = pd.read_csv('Smoker_Epigenetic_df.csv')
data.dropna()

data['Smoking Status'] = data['Smoking Status'].map({'current': 1, 'former': 0,
data['Gender'] = data['Gender'].map({'f': 0, 'm': 1})

X = data.drop(columns=['GSM', 'Smoking Status'])
y = data['Smoking Status']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_

clf = DecisionTreeClassifier(random_state=42)
clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)
print("Accuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_pred))
print("\nDecision Tree:\n")
print(export_text(clf, feature_names=list(X.columns)))
```

```
Accuracy: 0.656934306569343

Classification Report:
              precision    recall  f1-score   support

          -1       0.39      0.36      0.37        39
           1       0.75      0.78      0.76        98

    accuracy                           0.66       137
   macro avg       0.57      0.57      0.57       137
weighted avg       0.65      0.66      0.65       137


Decision Tree:

|--- cg02839557 <= 0.04
|   |--- Age <= 32.00
|   |   |--- class: -1
|   |--- Age >  32.00
|   |   |--- cg00213748 <= 0.92
|   |   |   |--- cg03052502 <= 0.99
|   |   |   |   |--- class: 1
|   |   |   |--- cg03052502 >  0.99
|   |   |   |   |--- class: -1
|   |   |--- cg00213748 >  0.92
|   |   |   |--- class: -1
|--- cg02839557 >  0.04
|   |--- cg01707559 <= 0.06
|   |   |--- cg03695421 <= 0.58
|   |   |   |--- class: 1
|   |   |--- cg03695421 >  0.58
|   |   |   |--- class: -1
|   |--- cg01707559 >  0.06
|   |   |--- cg00050873 <= 0.57
|   |   |   |--- cg01707559 <= 0.22
|   |   |   |   |--- cg03052502 <= 0.45
|   |   |   |   |   |--- Age <= 62.50
|   |   |   |   |   |   |--- cg03443143 <= 0.38
|   |   |   |   |   |   |   |--- cg02004872 <= 0.26
|   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |   |--- cg02004872 >  0.26
|   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |--- cg03443143 >  0.38
|   |   |   |   |   |   |   |--- Age <= 48.50
|   |   |   |   |   |   |   |   |--- cg02842889 <= 0.39
|   |   |   |   |   |   |   |   |   |--- cg01707559 <= 0.22
|   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |--- cg01707559 >  0.22
|   |   |   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |   |   |--- cg02842889 >  0.39
|   |   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |   |--- Age >  48.50
|   |   |   |   |   |   |   |   |--- cg02004872 <= 0.10
|   |   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |   |   |--- cg02004872 >  0.10
|   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |--- Age >  62.50
|   |   |   |   |   |   |--- cg02494853 <= 0.08
|   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |--- cg02494853 >  0.08
```

```
|   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |--- cg03052502 >  0.45
|   |   |   |   |   |--- cg03155755 <= 0.45
|   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |--- cg03155755 >  0.45
|   |   |   |   |   |   |--- cg03683899 <= 0.33
|   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |--- cg03683899 >  0.33
|   |   |   |   |   |   |   |--- class: -1
|   |   |   |--- cg01707559 >  0.22
|   |   |   |   |--- cg00455876 <= 0.35
|   |   |   |   |   |--- cg03695421 <= 0.27
|   |   |   |   |   |   |--- cg03155755 <= 0.26
|   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |--- cg03155755 >  0.26
|   |   |   |   |   |   |   |--- cg02004872 <= 0.12
|   |   |   |   |   |   |   |   |--- cg02004872 <= 0.10
|   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |--- cg02004872 >  0.10
|   |   |   |   |   |   |   |   |   |--- cg02494853 <= 0.03
|   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |--- cg02494853 >  0.03
|   |   |   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |   |--- cg02004872 >  0.12
|   |   |   |   |   |   |   |   |--- cg00213748 <= 0.10
|   |   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |   |   |--- cg00213748 >  0.10
|   |   |   |   |   |   |   |   |   |--- cg00214611 <= 0.30
|   |   |   |   |   |   |   |   |   |   |--- cg00050873 <= 0.46
|   |   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |   |--- cg00050873 >  0.46
|   |   |   |   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |   |   |   |--- cg00214611 >  0.30
|   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |--- cg03695421 >  0.27
|   |   |   |   |   |   |--- cg02842889 <= 0.39
|   |   |   |   |   |   |   |--- cg02233190 <= 0.46
|   |   |   |   |   |   |   |   |--- Age <= 31.00
|   |   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |   |   |--- Age >  31.00
|   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |--- cg02233190 >  0.46
|   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |--- cg02842889 >  0.39
|   |   |   |   |   |   |   |--- cg03695421 <= 0.28
|   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |   |--- cg03695421 >  0.28
|   |   |   |   |   |   |   |   |--- cg03052502 <= 0.52
|   |   |   |   |   |   |   |   |   |--- cg03052502 <= 0.45
|   |   |   |   |   |   |   |   |   |   |--- cg02842889 <= 0.42
|   |   |   |   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |   |   |   |--- cg02842889 >  0.42
|   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 4
|   |   |   |   |   |   |   |   |--- cg03052502 >  0.45
|   |   |   |   |   |   |   |   |   |--- cg01707559 <= 0.28
|   |   |   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |   |   |   |--- cg01707559 >  0.28
|   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 2
|   |   |   |   |   |   |   |   |--- cg03052502 >  0.52
|   |   |   |   |   |   |   |   |   |--- cg03244189 <= 0.35
```

```
|   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |--- cg03244189 >  0.35
|   |   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |--- cg00455876 >  0.35
|   |   |   |   |   |--- cg00455876 <= 0.37
|   |   |   |   |   |   |--- cg02842889 <= 0.40
|   |   |   |   |   |   |   |--- cg03695421 <= 0.21
|   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |   |--- cg03695421 >  0.21
|   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |--- cg02842889 >  0.40
|   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |--- cg00455876 >  0.37
|   |   |   |   |   |   |--- cg02494853 <= 0.09
|   |   |   |   |   |   |   |--- cg00212031 <= 0.58
|   |   |   |   |   |   |   |   |--- cg03706273 <= 0.13
|   |   |   |   |   |   |   |   |   |--- cg02494853 <= 0.06
|   |   |   |   |   |   |   |   |   |   |--- cg02842889 <= 0.33
|   |   |   |   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |   |   |   |   |--- cg02842889 >  0.33
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 8
|   |   |   |   |   |   |   |   |   |--- cg02494853 >  0.06
|   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |--- cg03706273 >  0.13
|   |   |   |   |   |   |   |   |   |--- cg02842889 <= 0.49
|   |   |   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |   |   |--- cg02842889 >  0.49
|   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |--- cg00212031 >  0.58
|   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |--- cg02494853 >  0.09
|   |   |   |   |   |   |   |--- class: -1
|   |   |   |--- cg00050873 >  0.57
|   |   |   |   |--- cg02233190 <= 0.02
|   |   |   |   |   |--- class: -1
|   |   |   |   |--- cg02233190 >  0.02
|   |   |   |   |   |--- cg02494853 <= 0.11
|   |   |   |   |   |   |--- Age <= 48.50
|   |   |   |   |   |   |   |--- cg03695421 <= 0.64
|   |   |   |   |   |   |   |   |--- cg02494853 <= 0.05
|   |   |   |   |   |   |   |   |   |--- cg03052502 <= 0.98
|   |   |   |   |   |   |   |   |   |   |--- cg02494853 <= 0.03
|   |   |   |   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |   |   |   |   |--- cg02494853 >  0.03
|   |   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |--- cg03052502 >  0.98
|   |   |   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |   |   |--- cg02494853 >  0.05
|   |   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |   |--- cg03695421 >  0.64
|   |   |   |   |   |   |   |   |--- cg03244189 <= 0.11
|   |   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |   |   |--- cg03244189 >  0.11
|   |   |   |   |   |   |   |   |   |--- cg02842889 <= 0.05
|   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |--- cg02842889 >  0.05
|   |   |   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |--- Age >  48.50
|   |   |   |   |   |   |   |--- cg03695421 <= 0.20
|   |   |   |   |   |   |   |   |--- class: -1
```

```
|   |   |   |   |   |   |--- cg03695421 >  0.20
|   |   |   |   |   |   |   |--- cg02839557 <= 0.04
|   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |   |--- cg02839557 >  0.04
|   |   |   |   |   |   |   |   |--- cg00212031 <= 0.02
|   |   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |   |   |--- cg00212031 >  0.02
|   |   |   |   |   |   |   |   |   |--- Age <= 67.50
|   |   |   |   |   |   |   |   |   |   |--- cg03052502 <= 0.29
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 2
|   |   |   |   |   |   |   |   |   |   |--- cg03052502 >  0.29
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 5
|   |   |   |   |   |   |   |   |   |--- Age >  67.50
|   |   |   |   |   |   |   |   |   |   |--- cg02004872 <= 0.02
|   |   |   |   |   |   |   |   |   |   |   |--- class: -1
|   |   |   |   |   |   |   |   |   |   |--- cg02004872 >  0.02
|   |   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |--- cg02494853 >  0.11
|   |   |   |   |   |--- class: -1
```

In [2]:
```python
# 2. KS wykonać klasyfikację ensemble (używając modeli Random Forrest, Boosting,
from sklearn.impute import SimpleImputer
from sklearn.ensemble import RandomForestClassifier, BaggingClassifier, AdaBoost
from sklearn.metrics import classification_report, accuracy_score
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_

imputer = SimpleImputer(strategy='mean')
X_train = imputer.fit_transform(X_train)
X_test = imputer.transform(X_test)

models = {
    "Random Forest": RandomForestClassifier(n_estimators=100, random_state=42),
    "Bagging": BaggingClassifier(n_estimators=100, random_state=42),
    "Boosting (AdaBoost)": AdaBoostClassifier(n_estimators=100, random_state=42)
    "Gradient Boosting": GradientBoostingClassifier(n_estimators=100, random_sta
}

for name, model in models.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    print(f"=== {name} ===")
    print(f"Accuracy: {accuracy_score(y_test, y_pred):.2f}")
    print(classification_report(y_test, y_pred))
```

```
C:\Users\krzys\AppData\Local\Programs\Python\Python310\lib\site-packages\sklearn
\impute\_base.py:635: UserWarning: Skipping features without any observed values:
['Gender']. At least one non-missing value is needed for imputation with strategy
='mean'.
  warnings.warn(
C:\Users\krzys\AppData\Local\Programs\Python\Python310\lib\site-packages\sklearn
\impute\_base.py:635: UserWarning: Skipping features without any observed values:
['Gender']. At least one non-missing value is needed for imputation with strategy
='mean'.
  warnings.warn(
```

```
=== Random Forest ===
Accuracy: 0.72
              precision    recall  f1-score   support

          -1       0.41      0.17      0.24        53
           1       0.76      0.91      0.83       152

    accuracy                           0.72       205
   macro avg       0.58      0.54      0.53       205
weighted avg       0.67      0.72      0.68       205

=== Bagging ===
Accuracy: 0.76
              precision    recall  f1-score   support

          -1       0.55      0.30      0.39        53
           1       0.79      0.91      0.85       152

    accuracy                           0.76       205
   macro avg       0.67      0.61      0.62       205
weighted avg       0.73      0.76      0.73       205

=== Boosting (AdaBoost) ===
Accuracy: 0.75
              precision    recall  f1-score   support

          -1       0.52      0.32      0.40        53
           1       0.79      0.89      0.84       152

    accuracy                           0.75       205
   macro avg       0.65      0.61      0.62       205
weighted avg       0.72      0.75      0.72       205

=== Gradient Boosting ===
Accuracy: 0.75
              precision    recall  f1-score   support

          -1       0.53      0.40      0.45        53
           1       0.81      0.88      0.84       152

    accuracy                           0.75       205
   macro avg       0.67      0.64      0.65       205
weighted avg       0.73      0.75      0.74       205
```

In [ ]:

3.Wnioski płynące z ćwiczenia

Ćwiczenie polegało na opracowaniu przepływu pracy w uczeniu maszynowym dla problemu klasyfikacji, wykorzystując pojedyncze drzewo decyzyjne. W trakcie realizacji tego zadania zdobyłem umiejętności związane z przygotowaniem danych, budową modelu klasyfikacyjnego oraz jego oceną.
Na początku przeprowadzono wczytanie i wstępne przetwarzanie danych przy użyciu biblioteki `pandas`. Proces obejmował usunięcie brakujących wartości oraz zakodowanie zmiennych kategorycznych, co było kluczowe dla prawidłowego działania modelu klasyfikacyjnego. Zastosowanie funkcji `map()` pozwoliło na przekształcenie wartości zmiennych „Smoking Status" oraz „Gender" na format liczbowy, umożliwiając ich wykorzystanie w modelu uczenia maszynowego.
Kolejnym krokiem było podzielenie zbioru danych na część treningową i testową przy użyciu `train_test_split()`. Dzięki temu możliwe było sprawdzenie skuteczności modelu na nieznanych wcześniej danych, co jest kluczowym elementem oceny jego ogólności i zdolności do przewidywania nowych obserwacji.
Do klasyfikacji wykorzystano `DecisionTreeClassifier` z biblioteki `sklearn.tree`. Model ten został wytrenowany na zbiorze treningowym, a następnie zastosowany do przewidywania klas w zbiorze testowym. Ocena skuteczności modelu została przeprowadzona przy użyciu metryk takich jak **dokładność (accuracy)** oraz **raport klasyfikacji (classification report)**, który zawierał szczegółowe informacje na temat precyzji, czułości i wartości F1-score dla każdej klasy.
Dodatkowo, miałem możliwość zapoznania się ze strukturą wytrenowanego drzewa decyzyjnego, korzystając z funkcji `export_text()`, która pozwala na wizualizację reguł podejmowania decyzji w modelu. Analiza struktury drzewa umożliwiła lepsze zrozumienie sposobu działania modelu oraz identyfikację kluczowych cech wpływających na klasyfikację.
Ćwiczenie pozwoliło na praktyczne zapoznanie się z procesem budowy modeli klasyfikacyjnych w uczeniu maszynowym. Nauczyłem się, jak poprawnie przygotować dane, wytrenować model drzewa decyzyjnego i ocenić jego skuteczność. Dodatkowo, możliwość analizy reguł decyzyjnych pozwoliła na lepsze zrozumienie działania klasyfikatorów opartych na drzewach. Wiedza zdobyta podczas tego zadania jest istotna w kontekście budowy i interpretacji modeli klasyfikacyjnych, które znajdują szerokie zastosowanie w analizie danych i systemach predykcyjnych.