Wrocław, 06.06.2025

Sztuczna inteligencja i inżynieria wiedzy

# Lista 4

ALGORYTMY UCZENIA MASZYNOWEGO

Kacper Tomczyk 272650

# Wstępna analiza zbioru danych

Pierwszym krokiem analizy zbioru danych było jego wczytanie i przygotowanie do dalszych etapów przetwarzania. Zadanie to wykonano za pomocą skryptu w języku **python**, przy wykorzystaniu biblioteki **pandas**.

Dane pochodzą z badań kardiotokograficznych płodów i zawierają 2126 rekordów opisanych 21 cechami diagnostycznymi oraz jedną kolumną klasową (CLASS), która reprezentuje typ wzorca morfologicznego płodu. Już na etapie wstępnej analizy wykryto obecność brakujących wartości w większości kolumn – od kilku do ponad 130 braków w poszczególnych cechach. Najwięcej braków dotyczyło kolumn takich jak „Min”, „DS” czy „UC”, co może wskazywać na problemy z rejestracją lub ekstrakcją niektórych parametrów diagnostycznych:

Tabela 1 – Liczba brakujących wartości dla poszczególnych kolumn

|  |  |
| --- | --- |
| Kolumna | Brakuje |
| LB | 103 |
| AC | 113 |
| FM | 109 |
| UC | 114 |
| DL | 85 |
| DS | 132 |
| DP | 110 |
| ASTV | 111 |
| MSTV | 102 |
| ALTV | 96 |
| MLTV | 109 |
| Width | 103 |
| Min | 126 |
| Max | 88 |
| Nmax | 105 |
| Nzeros | 124 |
| Mode | 102 |
| Mean | 107 |
| Median | 101 |
| Variance | 97 |
| Tendency | 107 |
| CLASS | 0 |

W kolejnej fazie analizy wyznaczono podstawowe statystyki opisowe dla każdej kolumny. Na ich podstawie można zauważyć duże zróżnicowanie zakresów wartości cech – np. wartość cechy „FM” (liczba ruchów płodu na sekundę) sięga maksymalnie 0.481, podczas gdy „Width” (szerokość histogramu FHR) osiąga wartość nawet 180. Również rozstęp między percentylami oraz wysokie odchylenia standardowe w kolumnach takich jak „ALTV” i „Variance” świadczą o znacznej zmienności tych parametrów. Z kolei niektóre cechy jak „DS” (liczba silnych deceleracji) czy „DP” wykazują praktycznie zerową średnią i minimalną wariancję, co może sugerować ich ograniczoną użyteczność w klasyfikacji:

Tabela 2 – Statystyki dla poszczególnych kolumn

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Kolumna | Liczność | Średnia | Odch. st. | 25% | 50% | 75% | Wart. Maks. | Wart. Min. |
| LB | 2023 | 133.276 | 9.811 | 126.000 | 133.000 | 140.000 | 160.000 | 106.000 |
| AC | 2013 | 0.003 | 0.004 | 0.000 | 0.002 | 0.006 | 0.018 | 0.000 |
| FM | 2017 | 0.009 | 0.045 | 0.000 | 0.000 | 0.003 | 0.481 | 0.000 |
| UC | 2012 | 0.004 | 0.003 | 0.002 | 0.005 | 0.007 | 0.015 | 0.000 |
| DL | 2041 | 0.002 | 0.003 | 0.000 | 0.000 | 0.003 | 0.015 | 0.000 |
| DS | 1994 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 |
| DP | 2016 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.005 | 0.000 |
| ASTV | 2015 | 47.043 | 17.212 | 32.000 | 49.000 | 61.000 | 87.000 | 12.000 |
| MSTV | 2024 | 1.331 | 0.883 | 0.700 | 1.200 | 1.700 | 7.000 | 0.200 |
| ALTV | 2030 | 9.860 | 18.418 | 0.000 | 0.000 | 11.000 | 91.000 | 0.000 |
| MLTV | 2017 | 8.209 | 5.612 | 4.600 | 7.400 | 10.800 | 50.700 | 0.000 |
| Width | 2023 | 70.796 | 39.056 | 37.000 | 68.000 | 101.000 | 180.000 | 3.000 |
| Min | 2000 | 93.412 | 29.554 | 67.000 | 93.000 | 120.000 | 159.000 | 50.000 |
| Max | 2038 | 164.065 | 18.013 | 152.000 | 162.000 | 174.000 | 238.000 | 122.000 |
| Nmax | 2021 | 4.091 | 2.955 | 2.000 | 4.000 | 6.000 | 18.000 | 0.000 |
| Nzeros | 2002 | 0.219 | 0.682 | 0.000 | 0.000 | 0.000 | 10.000 | 0.000 |
| Mode | 2024 | 137.353 | 16.432 | 129.000 | 139.000 | 148.000 | 187.000 | 60.000 |
| Mean | 2019 | 134.617 | 15.608 | 125.000 | 136.000 | 145.000 | 182.000 | 73.000 |
| Median | 2025 | 138.221 | 14.394 | 129.000 | 140.000 | 148.000 | 186.000 | 77.000 |
| Variance | 2029 | 19.130 | 29.320 | 2.000 | 8.000 | 24.000 | 269.000 | 0.000 |
| Tendancy | 2019 | 0.319 | 0.612 | 0.000 | 0.000 | 1.000 | 1.000 | -1.000 |
| CLASS | 2126 | 4.510 | 3.027 | 2.000 | 4.000 | 7.000 | 10.000 | 1.000 |

Następnie, dokonano analizy rozkładu wystąpień wartości dla kolumny CLASS:

Tabela 3 – Rozkład wartości CLASS

|  |  |
| --- | --- |
| Klasa | Wystąpienia |
| 1 | 384 |
| 2 | 579 |
| 3 | 53 |
| 4 | 81 |
| 5 | 72 |
| 6 | 332 |
| 7 | 252 |
| 8 | 107 |
| 9 | 69 |
| 10 | 197 |

Rozkład klas (kolumna CLASS) wskazuje, że zbiór danych jest wyraźnie niezbalansowany. Najliczniejsze klasy to 2 (579 przypadków) oraz 1 (384 przypadki), natomiast klasy takie jak 3 (53), 4 (81) czy 9 (69) są znacznie słabiej reprezentowane. Taki nierównomierny rozkład może wpływać negatywnie na skuteczność modeli klasyfikacyjnych, które mają tendencję do faworyzowania klas dominujących, co należy wziąć pod uwagę przy ewaluacji.

Korelacje pomiędzy cechami zostały przedstawione w postaci macierzy korelacji   
(dla pierwszych pięciu wierszy zbioru danych):

Tabela 4 – Macierz korelacji kolumn

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | LB | AC | FM | UC | DL | DS | DP | ASTV | MSTV | ALTV | MLTV | Width | Min | Max | Nmax | Nzeros | Mode | Mean | Median | Variance | Tendency | CLASS |
| LB |  | -0.07 | -0.03 | -0.13 | -0.16 | -0.06 | -0.11 | 0.3 | -0.26 | 0.28 | -0.04 | -0.13 | 0.36 | 0.28 | -0.11 | -0.02 | 0.71 | 0.72 | 0.79 | -0.13 | 0.29 | 0.14 |
| AC | -0.07 |  | 0.04 | 0.1 | -0.11 | -0.04 | -0.13 | -0.28 | 0.19 | -0.37 | -0.15 | 0.29 | -0.15 | 0.39 | 0.19 | -0.01 | 0.24 | 0.27 | 0.27 | 0.12 | 0.02 | -0.29 |
| FM | -0.03 | 0.04 |  | -0.07 | 0.05 | -0.01 | 0.26 | -0.09 | 0.13 | -0.07 | 0.02 | 0.15 | -0.15 | 0.1 | 0.17 | -0.03 | -0.06 | -0.1 | -0.06 | 0.19 | 0.01 | 0.09 |
| UC | -0.13 | 0.1 | -0.07 |  | 0.28 | 0 | 0.08 | -0.24 | 0.29 | -0.31 | -0.05 | 0.14 | -0.11 | 0.12 | 0.08 | 0.06 | -0.1 | -0.18 | -0.14 | 0.24 | -0.07 | -0.1 |
| DL | -0.16 | -0.11 | 0.05 | 0.28 |  | 0.1 | 0.22 | -0.13 | 0.57 | -0.27 | -0.23 | 0.52 | -0.55 | 0.22 | 0.39 | 0.26 | -0.34 | -0.52 | -0.39 | 0.57 | 0.02 | 0.41 |
| DS | -0.06 | -0.04 | -0.01 | 0 | 0.1 |  | 0.02 | 0.02 | 0.03 | -0.03 | -0.03 | 0.04 | -0.07 | -0.02 | -0.01 | 0.06 | -0.21 | -0.15 | -0.15 | 0.13 | -0.04 | 0.06 |
| DP | -0.11 | -0.13 | 0.26 | 0.08 | 0.22 | 0.02 |  | 0.05 | 0.27 | -0.14 | -0.23 | 0.26 | -0.28 | 0.13 | 0.23 | 0.06 | -0.45 | -0.49 | -0.45 | 0.51 | -0.21 | 0.27 |
| ASTV | 0.3 | -0.28 | -0.09 | -0.24 | -0.13 | 0.02 | 0.05 |  | -0.43 | 0.46 | -0.32 | -0.26 | 0.28 | -0.1 | -0.17 | -0.17 | 0.07 | 0.08 | 0.12 | -0.14 | -0.02 | 0.29 |
| MSTV | -0.26 | 0.19 | 0.13 | 0.29 | 0.57 | 0.03 | 0.27 | -0.43 |  | -0.47 | 0.07 | 0.66 | -0.62 | 0.41 | 0.5 | 0.27 | -0.32 | -0.45 | -0.33 | 0.55 | -0.05 | 0.07 |
| ALTV | 0.28 | -0.37 | -0.07 | -0.31 | -0.27 | -0.03 | -0.14 | 0.46 | -0.47 |  | -0.18 | -0.44 | 0.42 | -0.28 | -0.28 | -0.12 | 0.16 | 0.22 | 0.18 | -0.28 | 0.04 | 0.31 |
| MLTV | -0.04 | -0.15 | 0.02 | -0.05 | -0.23 | -0.03 | -0.23 | -0.32 | 0.07 | -0.18 |  | 0.12 | -0.16 | 0.01 | 0.06 | 0.11 | 0.07 | 0.12 | 0.05 | -0.16 | 0.14 | -0.22 |
| Width | -0.13 | 0.29 | 0.15 | 0.14 | 0.52 | 0.04 | 0.26 | -0.26 | 0.66 | -0.44 | 0.12 |  | -0.9 | 0.7 | 0.75 | 0.3 | -0.16 | -0.28 | -0.16 | 0.61 | 0.13 | 0.16 |
| Min | 0.36 | -0.15 | -0.15 | -0.11 | -0.55 | -0.07 | -0.28 | 0.28 | -0.62 | 0.42 | -0.16 | -0.9 |  | -0.3 | -0.66 | -0.3 | 0.35 | 0.49 | 0.4 | -0.54 | -0.25 | -0.16 |
| Max | 0.28 | 0.39 | 0.1 | 0.12 | 0.22 | -0.02 | 0.13 | -0.1 | 0.41 | -0.28 | 0.01 | 0.7 | -0.3 |  | 0.53 | 0.17 | 0.22 | 0.18 | 0.29 | 0.44 | -0.15 | 0.07 |
| Nmax | -0.11 | 0.19 | 0.17 | 0.08 | 0.39 | -0.01 | 0.23 | -0.17 | 0.5 | -0.28 | 0.06 | 0.75 | -0.66 | 0.53 |  | 0.28 | -0.1 | -0.23 | -0.12 | 0.44 | 0.11 | 0.13 |
| Nzeros | -0.02 | -0.01 | -0.03 | 0.06 | 0.26 | 0.06 | 0.06 | -0.17 | 0.27 | -0.12 | 0.11 | 0.3 | -0.3 | 0.17 | 0.28 |  | -0.08 | -0.09 | -0.06 | 0.18 | 0.1 | 0.08 |
| Mode | 0.71 | 0.24 | -0.06 | -0.1 | -0.34 | -0.21 | -0.45 | 0.07 | -0.32 | 0.16 | 0.07 | -0.16 | 0.35 | 0.22 | -0.1 | -0.08 |  | 0.89 | 0.93 | -0.32 | 0.42 | -0.1 |
| Mean | 0.72 | 0.27 | -0.1 | -0.18 | -0.52 | -0.15 | -0.49 | 0.08 | -0.45 | 0.22 | 0.12 | -0.28 | 0.49 | 0.18 | -0.23 | -0.09 | 0.89 |  | 0.95 | -0.41 | 0.31 | -0.2 |
| Median | 0.79 | 0.27 | -0.06 | -0.14 | -0.39 | -0.15 | -0.45 | 0.12 | -0.33 | 0.18 | 0.05 | -0.16 | 0.4 | 0.29 | -0.12 | -0.06 | 0.93 | 0.95 |  | -0.29 | 0.38 | -0.11 |
| Variance | -0.13 | 0.12 | 0.19 | 0.24 | 0.57 | 0.13 | 0.51 | -0.14 | 0.55 | -0.28 | -0.16 | 0.61 | -0.54 | 0.44 | 0.44 | 0.18 | -0.32 | -0.41 | -0.29 |  | -0.08 | 0.29 |
| Tendency | 0.29 | 0.02 | 0.01 | -0.07 | 0.02 | -0.04 | -0.21 | -0.02 | -0.05 | 0.04 | 0.14 | 0.13 | -0.25 | -0.15 | 0.11 | 0.1 | 0.42 | 0.31 | 0.38 | -0.08 |  | 0.08 |
| CLASS | 0.14 | -0.29 | 0.09 | -0.1 | 0.41 | 0.06 | 0.27 | 0.29 | 0.07 | 0.31 | -0.22 | 0.16 | -0.16 | 0.07 | 0.13 | 0.08 | -0.1 | -0.2 | -0.11 | 0.2 | 0.08 |  |

Zauważalne są silne dodatnie korelacje między niektórymi cechami histogramu FHR, np. „Mode”, „Mean” i „Median” są silnie powiązane z cechą „LB” (baseline FHR), co jest logiczne, ponieważ wszystkie te cechy odnoszą się do częstości rytmu serca. Interesujące są również silne korelacje „DL” z „MSTV” i „Tendency”, co może sugerować istotne relacje pomiędzy zwolnieniami a zmiennością krótkoterminową. Z drugiej strony, wiele cech (np. „DS”, „DP”) wykazuje bardzo słabą korelację z innymi zmiennymi oraz z klasą, co może oznaczać niską wartość predykcyjną.

Tabela 5 – Wariancja kolumn wejściowych

|  |  |
| --- | --- |
| Kolumna | Wariancja |
| LB | 9.63E+01 |
| AC | 1.47E-05 |
| FM | 2.03E-03 |
| UC | 8.72E-06 |
| DL | 8.73E-06 |
| DS | 3.00E-09 |
| DP | 3.61E-07 |
| ASTV | 2.96E+02 |
| MSTV | 7.81E-01 |
| ALTV | 3.39E+02 |
| MLTV | 3.15E+01 |
| Width | 1.53E+03 |
| Min | 8.73E+02 |
| Max | 3.24E+02 |
| Nmax | 8.73E+00 |
| Nzeros | 4.65E-01 |
| Mode | 2.70E+02 |
| Mean | 2.44E+02 |
| Median | 2.07E+02 |
| Variance | 8.60E+02 |
| Tendency | 3.75E-01 |

Analiza wariancji ujawniła, że niektóre cechy wykazują bardzo niską zmienność (np. „DS” czy „DP”), co może sugerować, że dostarczają one niewiele informacji i mogłyby zostać usunięte lub potraktowane jako stałe podczas dalszego przetwarzania. Z drugiej strony, cechy takie jak „Width”, „Min” i „Variance” wykazują bardzo wysoką wariancję, co może wskazywać na ich istotną rolę w różnicowaniu przypadków.

Podsumowując, analiza wykazała, że zbiór danych zawiera zarówno cechy potencjalnie użyteczne (zróżnicowane, skorelowane z klasą), jak i cechy, które mogą nie wnosić wartości dodanej (nisko wariancyjne lub o dużym stopniu braków).

# Przygotowanie oraz wstępne przetwarzanie danych

Dane wejściowe wczytywane są z pliku **csv** za pomocą funkcji **read\_csv()**, zaimplementowanej w bibliotece pandas. Wczytane dane przedstawione są za pomocą obiektu typu **DataFrame**.

Po wczytaniu danych, głównym etapem przygotowania jest ich oczyszczenie oraz przeskalowanie, co odbywa się w funkcji preprocess\_data(). Pierwszym krokiem w tej funkcji jest zamiana pustych pól (również tych zawierających tylko spacje) na wartość **NaN**, aby możliwe było ich dalsze przetwarzanie:



Następnie, brakujące wartości w zbiorze danych są uzupełniane średnią wartością z danej kolumny, co jest prostą, ale skuteczną metodą imputacji:



Po oczyszczeniu danych, zbiór cech (X) zostaje oddzielony od etykiet klasowych (y), co przygotowuje dane do procesu uczenia:



W dalszej części funkcji zastosowano możliwość przeskalowania danych z wykorzystaniem dwóch strategii: normalizacji (Normalizer) lub standaryzacji (StandardScaler), wybieranych na podstawie argumentu --scaler przekazywanego z linii komend:

A computer screen with text

AI-generated content may be incorrect.

Po wybraniu odpowiedniego skalera dane są przetwarzane:

A black screen with white text

AI-generated content may be incorrect.

Ostatecznie dane są dzielone na zbiór treningowy i testowy przy pomocy funkcji train\_test\_split() z **sklearn**, a proporcja podziału ustalana jest przez użytkownika za pomocą parametru CLI --ratio:

A screenshot of a computer program

AI-generated content may be incorrect.

# Klasyfikacja

Do rozwiązania problemu klasyfikacji wzorców morfologicznych płodu zastosowano dwa klasyczne algorytmy uczenia nadzorowanego: naiwny klasyfikator Bayesa (Naive Bayes) oraz drzewo decyzyjne (Decision Tree). W obu przypadkach użytkownik ma możliwość wyboru konkretnych wariantów modelu oraz konfiguracji hiperparametrów za pomocą argumentów CLI programu.

1. **Naive Bayes**

Algorytm Naive Bayes został zaimplementowany w trzech odmianach dostępnych w bibliotece scikit-learn:

* **GaussianNB** – wariant zakładający rozkład normalny zmiennych wejściowych. Parametrem dostosowywanym jest **var\_smoothing**, kontrolujący stabilność obliczeń w przypadku małych wariancji.
* **MultinomialNB** – odpowiedni do danych dyskretnych, oparty o rozkład wielomianowy. Użytkownik może ustawić wartość parametru **alpha**, czyli współczynnika Laplace'owego wygładzania
* **BernoulliNB** – wariant dla danych binarnych; oprócz **alpha**, posiada także parametr **binarize**, który określa próg binarnego przekształcenia danych:

1. **Decision Tree**

W przypadku klasyfikatora drzewa decyzyjnego wykorzystano klasę **DecisionTreeClassifier** z scikit-learn. Z poziomu programu możliwe jest dostosowanie następujących hiperparametrów ( w nawiasach podano domyślne wartości):

* **max\_depth** (None) – maksymalna głębokość drzewa, ograniczająca jego złożoność,
* **min\_samples\_split** (2) – minimalna liczba próbek potrzebna do rozdzielenia węzła,
* **min\_samples\_leaf** (1) – minimalna liczba próbek w liściu drzewa,
* **max\_leaf\_nodes** (None) – maksymalna liczba liści,
* **criterion** (gini) – funkcja oceny podziału (gini lub entropy),
* **splitter** (best) – strategia wyboru podziału (best lub random),
* **max\_features** (None) – liczba cech brana pod uwagę przy każdym podziale.

W projekcie zachowano wyraźny podział pomiędzy fazą trenowania modelu a fazą klasyfikacji danych testowych. Proces uczenia realizowany jest po wywołaniu komendy **train** i skutkuje zapisaniem wytrenowanego modelu do pliku **.pkl** przy użyciu mechanizmu serializacji (pickle).

Model ten może być później niezależnie załadowany i użyty do klasyfikacji nowych danych za pomocą komendy **classify**. Umożliwia to oddzielenie potencjalnie czasochłonnego procesu uczenia od procesu predykcji. Istnieje możliwość zapisania wyników klasyfikacji do pliku .csv, który zawiera wszystkie wiersze oraz kolumny źródłowego zestawu danych, oraz dodatkową kolumnę, zawierającą przewidzianą klasyfikację danego wiersza.

# Porównanie oraz analiza wyników

Poniżej przedstawiono macierz eksperymentów wraz z wynikami klasyfikacji. Każdy wiersz odpowiada jednemu wytrenowanemu modelowi, który jest scharakteryzowany przez swój typ, swoje hiperparametry oraz ustawienia dla przetwarzania danych wejściowych:

Tabela 6 – Wyniki eksperymentów

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Name | Model | Submodel | Scaler | Ratio | Hyperparams | Accuracy | Precision | Recall | F1 Score |
| NB1 | Naive-Bayes | Gaussian | (None) | 0.2 | --smoothing 1e-8 | 0.5399 | 0.5895 | 0.5399 | 0.5519 |
| NB1S | Naive-Bayes | Gaussian | Standardization | 0.2 | --smoothing 1e-9 | 0.6315 | 0.7227 | 0.6315 | 0.6428 |
| NB1N40 | Naive-Bayes | Gaussian | Normalization | 0.4 | --smoothing 1e-9 | 0.6005 | 0.6174 | 0.6005 | 0.6044 |
| NB1N10 | Naive-Bayes | Gaussian | Normalization | 0.1 | --smoothing 1e-9 | 0.5915 | 0.6053 | 0.5915 | 0.592 |
| NB1N | Naive-Bayes | Gaussian | Normalization | 0.2 | --smoothing 1e-9 | 0.6174 | 0.6354 | 0.6174 | 0.6206 |
| NB2N | Naive-Bayes | Multinomial | Normalization | 0.2 | --alpha 1.0 | 0.2676 | 0.0721 | 0.2676 | 0.1136 |
| NB3N | Naive-Bayes | Bernoulli | Normalization | 0.2 | --alpha 0.5 --binarize 0.1 | 0.4272 | 0.3028 | 0.4272 | 0.3389 |
| DT1 | Decision Tree |  | (None) | 0.2 | --max\_depth 5 --min\_samples\_split 4 --criterion gini --splitter best | 0.7394 | 0.7343 | 0.7394 | 0.7275 |
| DT1S | Decision Tree |  | Standardization | 0.2 | --max\_depth 5 --min\_samples\_split 4 --criterion gini --splitter best | 0.7394 | 0.7336 | 0.7394 | 0.7273 |
| DT1N | Decision Tree |  | Normalization | 0.2 | --max\_depth 5 --min\_samples\_split 4 --criterion gini --splitter best | 0.7113 | 0.6849 | 0.7113 | 0.6937 |
| DT2N | Decision Tree |  | Normalization | 0.2 | --max\_depth 10 --min\_samples\_leaf 3 --criterion entropy --splitter best | 0.6995 | 0.7021 | 0.6995 | 0.6953 |
| DT3N | Decision Tree |  | Normalization | 0.2 | --max\_leaf\_nodes 20 --min\_samples\_split 2 --criterion gini  --splitter random | 0.6526 | 0.6084 | 0.6526 | 0.6137 |

## Naive-Bayes

W przypadku klasyfikatora Naive Bayes, najlepsze wyniki uzyskano dla wariantu Gaussian z zastosowaniem standaryzacji cech (**NB1S**). Osiągnął on accuracy = 0.6315 oraz F1-score = 0.6428, co stanowi znaczną poprawę względem wariantu bez przetwarzania (**NB1**: 0.5399 / 0.5519). Użycie normalizacji również przyniosło lepsze rezultaty niż brak przetwarzania – wariant **NB1N** osiągnął F1-score = 0.6206, przy accuracy = 0.6174. Analiza wariantów z różnymi proporcjami zbioru testowego (**NB1N10**, **NB1N40**) wykazała, że zarówno zbyt mała (0.1), jak i zbyt duża (0.4) proporcja testowa może nieco pogorszyć wyniki względem ustawienia domyślnego 0.2.

Warianty Multinomial (**NB2N**) oraz Bernoulli (**NB3N**) okazały się znacznie mniej skuteczne – szczególnie **NB2N**, dla którego accuracy spadło do zaledwie 0.2676, a F1-score do 0.1136. Sugeruje to, że założenia rozkładu wielomianowego i binarnego nie są odpowiednie dla danych kardiotokograficznych, które są zmiennymi ciągłymi, a nie kategoriami lub binariami.

## Drzewo decyzyjne

Drzewa decyzyjne wypadły znacznie lepiej niż Naive Bayes we wszystkich konfiguracjach. Najwyższą skuteczność uzyskano w wariantach **DT1** i **DT1S**, które różnią się jedynie skalowaniem (odpowiednio: brak skalowania i standaryzacja). Oba osiągnęły accuracy = 0.7394 i F1-score ≈ 0.727, co wskazuje na stabilność działania modelu niezależnie od metody skalowania. Normalizacja danych (**DT1N**) wpłynęła nieco negatywnie na wyniki (F1-score = 0.6937), co może wynikać z tego, że drzewo decyzyjne nie wymaga standaryzacji ani normalizacji, ponieważ nie opiera się na odległościach, lecz na wartościach progowych.

Pozostałe warianty drzewa (**DT2N** i **DT3N**) różniły się głównie parametrami struktury drzewa. **DT2N**, wykorzystujące większą głębokość drzewa i entropię jako kryterium podziału, osiągnęło F1-score = 0.6953. **DT3N**, wykorzystujące losowy wybór podziałów i ograniczenie liczby liści, uzyskało nieco niższe wartości (F1 = 0.6137), co sugeruje pogorszenie uogólniania modelu na danych testowych.

## Wnioski

Wyniki wskazują, że drzewa decyzyjne są znacznie bardziej efektywne w tym zadaniu niż wszystkie warianty Naive-Bayes. Wśród nich, najlepiej sprawdza się klasyczny wariant drzewa z niską głębokością i kryterium Gini (DT1, DT1S). W przypadku Naive-Bayes, model Gaussian z odpowiednim skalowaniem (zwłaszcza standaryzacja) daje najlepsze wyniki, ale nie dorównuje drzewom pod względem skuteczności. Warto zauważyć, że nie wszystkie techniki przetwarzania danych poprawiają rezultaty – szczególnie w przypadku modeli drzewiastych normalizacja może działać wręcz niekorzystnie.

Warto wspomnieć, że uzyskane wyniki, nawet dla najlepszego wariantu drzewa decyzyjnego (F1-score ≈ 0.727), nie są zadowalająco wysokie, biorąc pod uwagę 10-klasową naturę problemu. Potencjalną ścieżką poprawy może być zastosowanie bardziej zaawansowanych metod klasyfikacji, takich jak Support Vector Machines (SVM), Random Forest czy XGBoost, które lepiej radzą sobie z bardziej złożonymi strukturami danych i mogą zapewnić lepsze uogólnienie oraz większą odporność na nadmierne dopasowanie.

# Kod

Kod z rozwiązaniem dostępny jest na repozytorium Github: <https://github.com/therockey/SIIW-Lab04>