Projekt z przedmiotu Eksploracja Danych

Etap II: Przygotowanie danych + Modelowanie

Temat: Gun Violence Data

Hanna Gibus, Igor Józefowicz, Agnieszka Kulesz

Charakterystyka zbioru danych

Pochodzenie:
https://www.kaggle.com/datasets/jameslko/gun-violence-data

Format:
.csv

Liczba przykładów:
239 677

Ilość zbiorów danych:
1

Cele eksploracji i kryteria sukcesu

Celem eksploracji jest predykcja czy incydent z użyciem broni palnej zakończy się ofiarami śmiertelnymi (`n_killed > 0`). Dodatkowym celem jest określenie, które atrybuty mają największy wpływ na prawdopodobieństwo śmierci w wyniku incydentu. Sukces zostanie osiągnięty, jeżeli model uzyska czułość na poziomie 80% oraz swoistość na poziomie 60%.

Dyskusja kroków dalszego postępowania

Dobór działania eksploracji

Celem eksploracji jest predykcja czy incydent z użyciem broni palnej zakończy się ofiarą śmiertelną (czyli `n_killed > 0`). Problem ten można zaklasyfikować jako klasyfikację binarną, w której klasa pozytywna oznacza incydenty śmiertelne, a klasa negatywna – incydenty bez ofiar śmiertelnych.

Dodatkowym celem eksploracji jest określenie, które atrybuty mają największy wpływ na prawdopodobieństwo wystąpienia incydentu śmiertelnego. Z tego względu szczególną uwagę poświęcono algorytmom, które umożliwiają ocenę ważności cech i ich interpretację.

Dobór algorytmu eksploracji

Podstawowym algorytmem zastosowanym w eksploracji jest drzewo decyzyjne (Decision Tree). Algorytm ten charakteryzuje się prostą strukturą, która umożliwia bezpośrednią interpretację reguł decyzyjnych. Pozwala to na zidentyfikowanie konkretnych progów wartości i cech wpływających na wystąpienie śmiertelnego incydentu. Jest to szczególnie istotne w przypadku eksploracji zorientowanej na wyjaśnialność wyników.
Dodatkowo w celach porównawczych zostały wykorzystane algorytmy zbalansowanego lasu losowego. Las losowy to zbiór drzew decyzyjnych, z których każde trenowane jest na losowym podzbiorze danych oraz losowym podzbiorze cech, a końcowa predykcja podejmowana jest na zasadzie głosowania większościowego. Zbalansowany las losowy to jego odmiana, w której próbki każdej klasy są losowane w równych proporcjach, co czyni ten algorytm bardziej odpornym na niezbalansowane dane.
Choć modele zespołowe są mniej interpretowalne, umożliwiają ocenę istotności cech na podstawie globalnych statystyk (np. znormalizowanego zysku informacyjnego), co pozwala wskazać najważniejsze predyktory.

Dobór metody testowania wyników

Zbiór danych zawiera ponad 240 000 rekordów, co stanowi wystarczającą liczbę do podziału danych na zbiór treningowy i testowy. Jednak pojedynczy, losowy podział może prowadzić do niestabilnych wyników, zależnych od konkretnego doboru próbek. Na przykład, jeśli do zbioru testowego trafią głównie przykłady łatwe do klasyfikacji, model może osiągnąć zawyżoną ocenę jakości.

W celu uzyskania bardziej wiarygodnej oceny predykcji zostanie zastosowana kroswalidacja 10-krotna. W tym podejściu dane dzielone są na 10 równych części, z których każda pełni raz rolę zbioru testowego, a pozostałe – treningowego. Takie podejście minimalizuje wpływ przypadkowego podziału i pozwala uśrednić wyniki, dając rzetelniejszą ocenę skuteczności modelu.

Przygotowanie danych

Dane brakujące i dane do ujednolicenia

Dane zawierają pewną liczbę braków oraz wartości niejednoznacznych, szczególnie w kolumnach tekstowych (`gun_stolen`, `participant_age_group`, `participant_gender`). Brakujące lub nieokreślone wartości zostały ujednolicone jako osobna kategoria „`Unknown`", co pozwala modelowi uwzględnić je bez konieczności odrzucania.

Zamiana na nominalne/numeryczne

Niektóre kolumny zawierają dane w postaci złożonych łańcuchów tekstowych, w których jedna obserwacja może zawierać wiele elementów, np. kolumna `gun_type` posiada wartości w formacie `0::22 LR||1::223 Rem [AR-15]`, gdzie poszczególne typy broni oddzielone są znakiem „`||`", a poprzedzone indeksem uczestnika. W celu przygotowania danych do analizy, z każdego wpisu wyodrębniono unikalne dane, a następnie zakodowano je jako zmienne nominalne (kategoryczne).

Podzbiór danych

W analizowanych danych uwzględniono tylko te kolumny, które mają wysoką potencjalną wartość predykcyjną oraz są kompletne. Wstępna analiza wykazała, że niektóre kolumny cechują się dużym nakładaniem informacji (redundancją) – np. `n_killed` i `n_injured` są skorelowane, jednak `n_killed` stanowi zmienną celu.

Dla uproszczenia i zwiększenia stabilności modelu, ostatecznie w analizie uwzględniono:

- `n_injured` (liczba rannych),
- `n_guns_involved` (liczba użytej broni),
- `state` (stan USA, jako zmienna nominalna),
- nowy atrybut `avg_age` – średni wiek uczestników zdarzenia, wyliczony na podstawie kolumny `participant_age`.
- oraz dodatkowe cechy `male_count` i `female_count` – liczba uczestników zdarzenia w podziale na płeć, wyliczona na podstawie kolumny `participant_gender`.

Kolumny złożone (`participant_type`, `participant_name`, `incident_url`, `address`) zostały pominięte ze względu na brak jednoznacznej wartości predykcyjnej i trudność przekształcenia.

Uzupełnienie danych

W analizowanym zbiorze danych celem jest przewidzenie czy incydent z użyciem broni palnej zakończy się ofiarą śmiertelną (`mortality = 1`, jeśli `n_killed > 0`). Wstępna eksploracja wykazała, że incydenty bez ofiar śmiertelnych są zdecydowanie liczniejsze, co oznacza, że klasy są wyraźnie niezbalansowane.

Pytanie, które się nasuwa, to czy taki rozkład jest zgodny z rzeczywistością. Analizy oraz dostępne dane potwierdzają, że większość zdarzeń tego typu faktycznie nie kończy się ofiarami śmiertelnymi, dlatego zachowanie oryginalnej proporcji klas jest uzasadnione.

W procesie przygotowania danych uzupełniono brakujące wartości w cechach numerycznych (np. liczba rannych, liczba broni, średni wiek uczestników, liczba mężczyzn i kobiet) oraz w zmiennych kategorycznych (np. stan) stosując odpowiednio medianę, zero lub kategorię „Unknown". Takie podejście pozwala na zachowanie kompletnych danych bez odrzucania rekordów.

Aby poradzić sobie z nierównowagą klas, w modelach zastosowano dwie strategie:

- przekazanie do algorytmu wag klas proporcjonalnych do ich częstości, co pozwala modelowi lepiej uwzględnić rzadziej występującą klasę pozytywną,
- a także wykorzystanie metody SMOTE (Synthetic Minority Over-sampling Technique) w fazie trenowania, która syntetycznie zwiększa liczbę próbek klasy mniejszościowej, poprawiając zdolność klasyfikatora do wykrywania incydentów śmiertelnych.

W przypadku, gdyby zachowanie oryginalnej proporcji klas nie było pożądane, np. z powodu wymagań konkretnego algorytmu lub celu badawczego, można rozważyć alternatywne podejścia, takie jak usuwanie części próbek klasy dominującej lub dalszą augmentację danych.

Aktualne podejście pozwala modelowi uczyć się na reprezentatywnych danych i uwzględniać naturalny rozkład incydentów, co jest istotne dla realnej oceny skuteczności predykcji.

Utworzenie modelu

Model drzewa decyzyjnego został utworzony w Pythonie z wykorzystaniem biblioteki scikit-learn, a dokładniej klasy `DecisionTreeClassifier`. Dla zapewnienia powtarzalności wyników został ustawiony parametr `random_state=42`. W celu uwzględnienia niezbalansowanych klas zastosowano wagi klas (`class_weight='balanced'`), które rekompensują różnice w liczebności próbek każdej klasy podczas trenowania modelu.

W ramach optymalizacji hiperparametrów przeprowadzono eksperymenty z różnymi wartościami parametru `max_depth`, który określa maksymalną głębokość drzewa (liczbę poziomów od korzenia do najdalszego liścia). Przeszukiwanie najlepszego zestawu parametrów odbyło się przy pomocy `GridSearchCV` z 3-krotną walidacją krzyżową na zbiorze treningowym, z oceną opartą na metryce czułości (`recall`).

Dla oceny końcowej modelu zastosowano 10-krotną walidację krzyżową (`StratifiedKFold`), aby uzyskać uśrednione i stabilne miary jakości klasyfikacji, takie jak czułość i swoistość. Wyniki eksperymentów wskazały, że najlepsze parametry to:

- `max_depth=7`
- `class_weight='balanced'`
- `random_state=42`

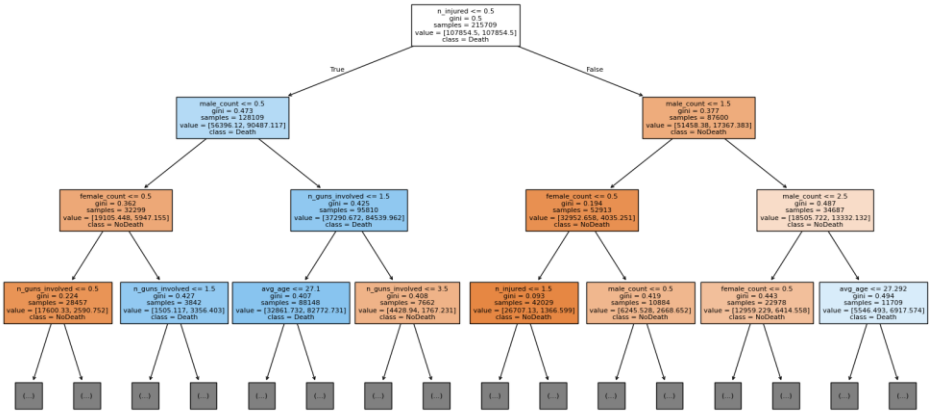
Dla uproszczenia i uniknięcia nadmiernego dopasowania modelu wybrano ostatecznie `max_depth=7` (zgodnie z wynikami wstępnych testów).

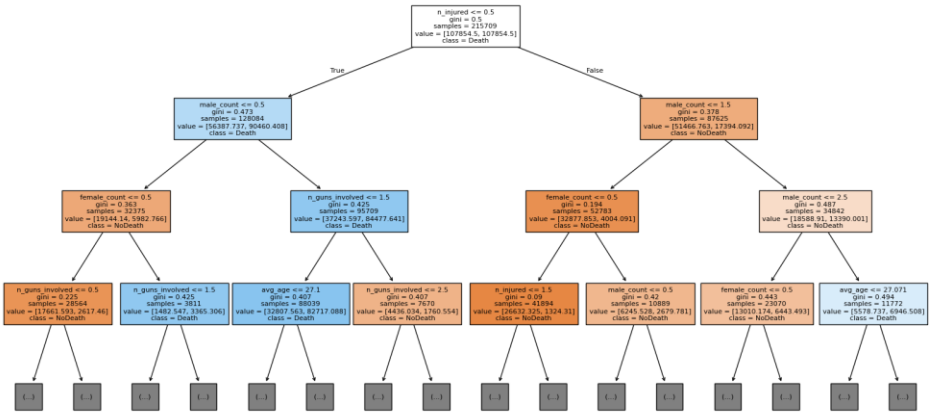
W wyniku 10-krotnej walidacji krzyżowej powstało 10 modeli drzew decyzyjnych, z których każdy został wytrenowany na nieco innym podzbiorze danych, co pozwoliło na ocenę stabilności i uogólnienia modelu.

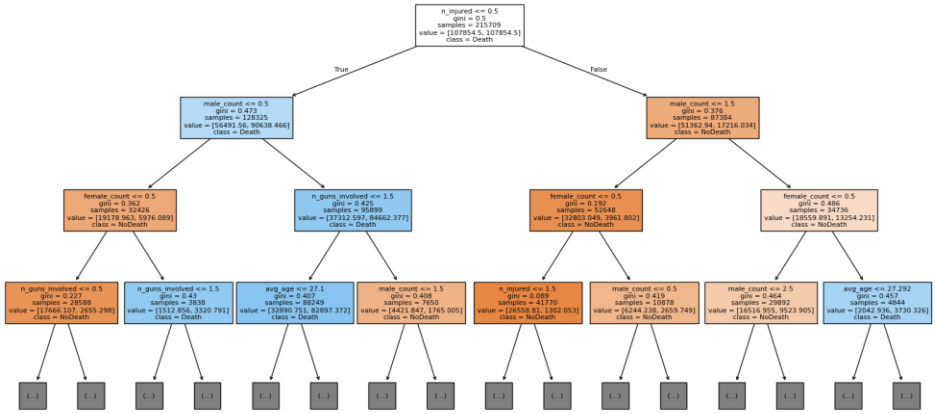
Poniżej zaprezentowano wykresy drzew decyzyjnych dla kolejnych foldów oraz zapisano reguły decyzyjne w formacie tekstowym, co umożliwia interpretację działania modelu.

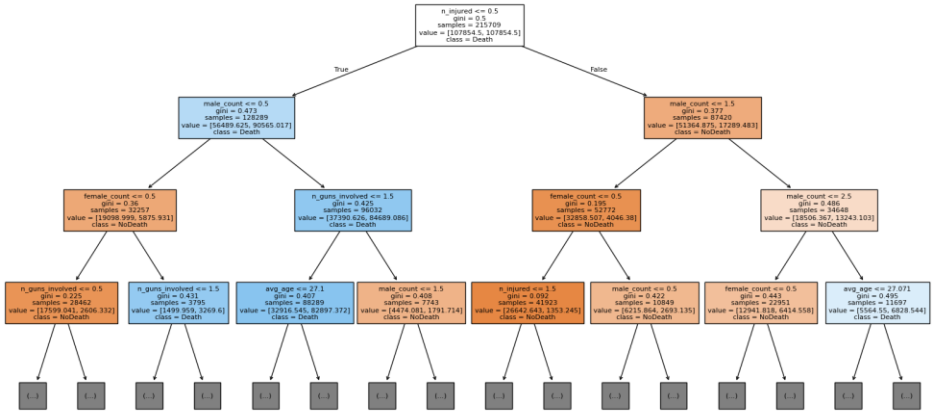
Podsumowując, ostateczny model drzewa decyzyjnego ma następujące parametry: `random_state=42`, `class_weight='balanced'`, `max_depth=7` oraz osiąga średnią czułość około `0.82` i średnią swoistość około `0.63` (wyniki z 10-krotnej kroswalidacji).

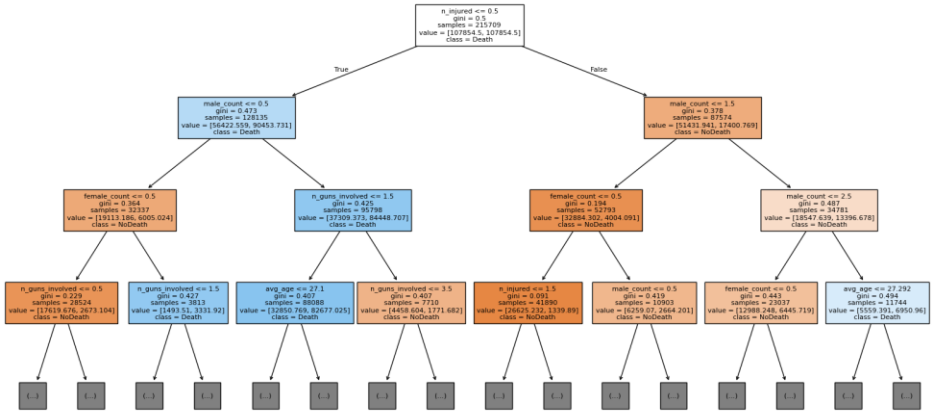
Decision Tree – wykresy z krosswalidacji

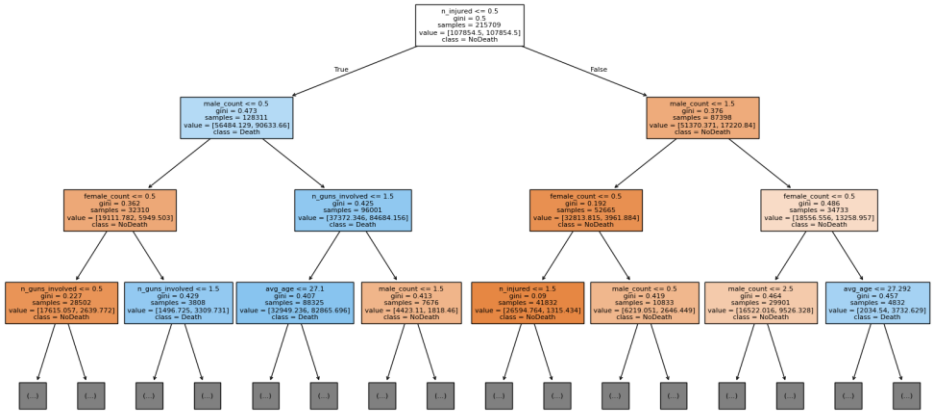


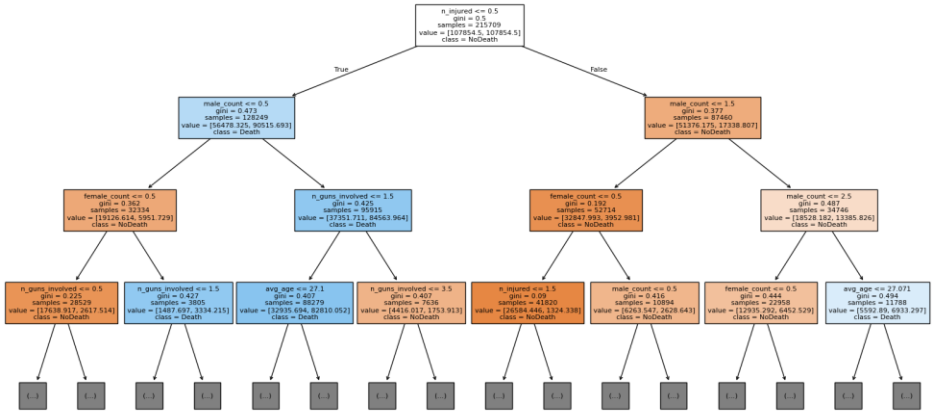


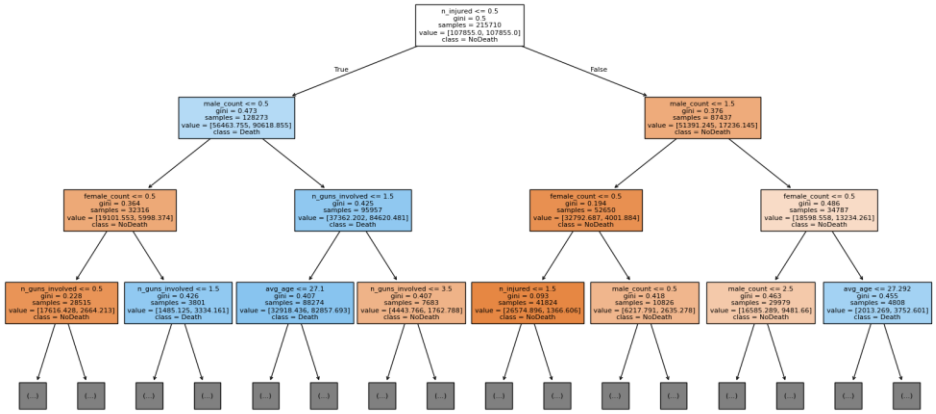

n_injured <= 0.5
gini = 0.5
samples = 215709
value = [107854.5, 107854.5]
class = Death
True
False
male_count <= 0.5
gini = 0.473
samples = 128325
value = [56491.56, 90638.466]
class = Death
male_count <= 1.5
gini = 0.376
samples = 87384
value = [51362.94, 17216.034]
class = NoDeath
female_count <= 0.5
gini = 0.362
samples = 32426
value = [19378.963, 5976.089]
class = NoDeath
n_guns_involved <= 1.5
gini = 0.425
samples = 95899
value = [37312.597, 84662.377]
class = Death
female_count <= 0.5
gini = 0.392
samples = 52648
value = [32803.049, 3961.802]
class = NoDeath
female_count <= 0.5
gini = 0.486
samples = 34736
value = [18559.891, 13254.231]
class = NoDeath
n_guns_involved <= 0.5
gini = 0.222
samples = 28588
value = [17666.107, 2655.298]
class = NoDeath
n_guns_involved <= 1.5
gini = 0.43
samples = 3838
value = [1512.856, 3320.791]
class = Death
avg_age <= 27.1
gini = 0.407
samples = 88249
value = [32890.751, 82897.372]
class = Death
male_count <= 1.5
gini = 0.408
samples = 7650
value = [4421.847, 1765.005]
class = NoDeath
n_injured <= 1.5
gini = 0.089
samples = 41770
value = [26558.81, 1302.053]
class = NoDeath
male_count <= 0.5
gini = 0.419
samples = 10878
value = [6244.238, 2659.749]
class = NoDeath
male_count <= 2.5
gini = 0.464
samples = 29892
value = [16516.955, 9523.905]
class = NoDeath
avg_age <= 27.292
gini = 0.457
samples = 4844
value = [2042.936, 3730.326]
class = Death
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

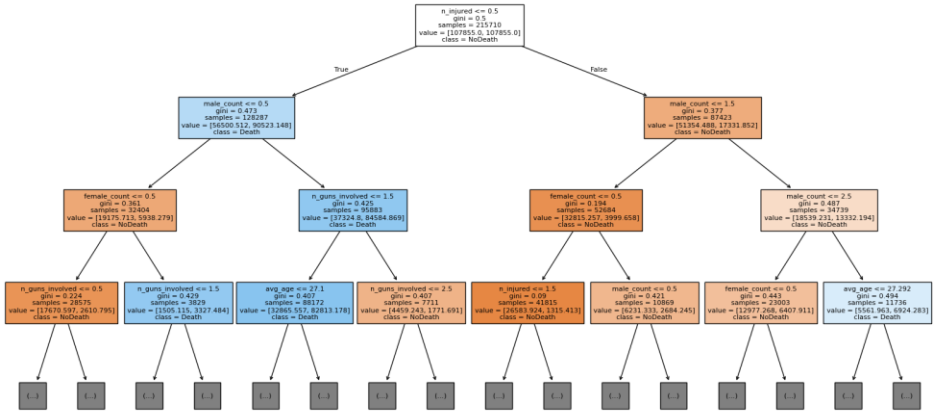

n_injured <= 0.5
gini = 0.5
samples = 215709
value = [107854.5, 107854.5]
class = Death
True
False
male_count <= 0.5
gini = 0.473
samples = 128289
value = [56489.625, 90565.017]
class = Death
male_count <= 1.5
gini = 0.377
samples = 87420
value = [51364.875, 17289.483]
class = NoDeath
female_count <= 0.5
gini = 0.38
samples = 32257
value = [19098.999, 5875.931]
class = NoDeath
n_guns_involved <= 1.5
gini = 0.425
samples = 96032
value = [37390.626, 84689.086]
class = Death
female_count <= 0.5
gini = 0.195
samples = 52772
value = [32858.507, 4046.38]
class = NoDeath
male_count <= 2.5
gini = 0.486
samples = 34648
value = [18506.367, 13243.103]
class = NoDeath
n_guns_involved <= 0.5
gini = 0.225
samples = 28462
value = [17599.041, 2606.332]
class = NoDeath
n_guns_involved <= 1.5
gini = 0.431
samples = 3795
value = [1499.959, 3269.6]
class = Death
avg_age <= 27.1
gini = 0.407
samples = 88289
value = [32916.545, 82897.372]
class = Death
male_count <= 1.5
gini = 0.408
samples = 7743
value = [4474.081, 1791.714]
class = NoDeath
n_injured <= 1.5
gini = 0.092
samples = 41923
value = [26642.643, 1353.245]
class = NoDeath
male_count <= 0.5
gini = 0.422
samples = 10849
value = [6215.864, 2693.135]
class = NoDeath
female_count <= 0.5
gini = 0.443
samples = 22951
value = [12941.818, 6414.558]
class = NoDeath
avg_age <= 27.071
gini = 0.495
samples = 11697
value = [5564.55, 6828.544]
class = Death
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

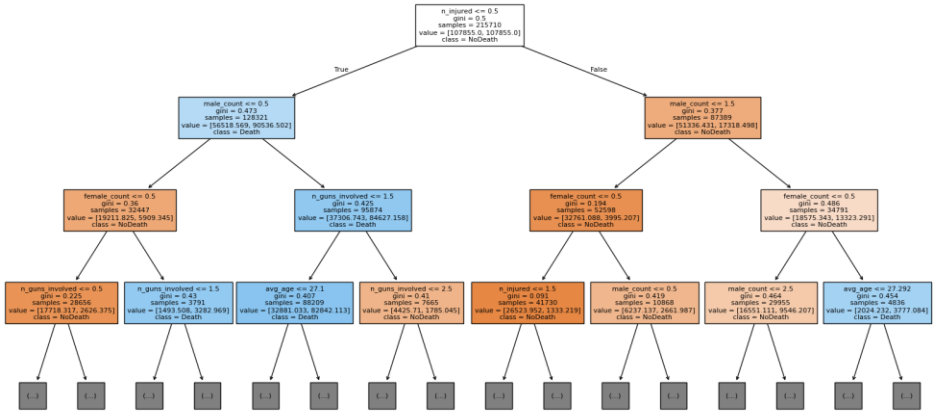

n_injured <= 0.5
gini = 0.5
samples = 215709
value = [107854.5, 107854.5]
class = Death
True
False
male_count <= 0.5
gini = 0.473
samples = 128135
value = [56422.559, 90453.731]
class = Death
male_count <= 1.5
gini = 0.378
samples = 87574
value = [51431.941, 17400.769]
class = NoDeath
female_count <= 0.5
gini = 0.364
samples = 32337
value = [19313.186, 6005.024]
class = NoDeath
n_guns_involved <= 1.5
gini = 0.425
samples = 95798
value = [37309.373, 84448.707]
class = Death
female_count <= 0.5
gini = 0.194
samples = 52793
value = [32884.302, 4004.091]
class = NoDeath
male_count <= 2.5
gini = 0.487
samples = 34781
value = [18547.639, 13396.678]
class = NoDeath
n_guns_involved <= 0.5
gini = 0.229
samples = 26524
value = [17619.676, 2673.104]
class = NoDeath
n_guns_involved <= 1.5
gini = 0.427
samples = 5813
value = [1493.51, 3331.92]
class = Death
avg_age <= 27.1
gini = 0.407
samples = 88088
value = [32850.769, 82677.025]
class = Death
n_guns_involved <= 3.5
gini = 0.407
samples = 7710
value = [4458.604, 1771.682]
class = NoDeath
n_injured <= 1.5
gini = 0.091
samples = 41890
value = [26625.232, 1339.89]
class = NoDeath
male_count <= 0.5
gini = 0.419
samples = 10903
value = [6259.07, 2664.201]
class = NoDeath
female_count <= 0.5
gini = 0.443
samples = 23037
value = [12988.248, 6445.719]
class = NoDeath
avg_age <= 27.292
gini = 0.494
samples = 11744
value = [5559.391, 6950.96]
class = Death
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)


n_injured <= 0.5
gini = 0.5
samples = 215709
value = [107854.5, 107854.5]
class = NoDeath
True
False
male_count <= 0.5
gini = 0.473
samples = 128311
value = [56484.129, 90633.66]
class = Death
male_count <= 1.5
gini = 0.376
samples = 87398
value = [51370.371, 17220.84]
class = NoDeath
female_count <= 0.5
gini = 0.362
samples = 32310
value = [19111.782, 5949.503]
class = NoDeath
n_guns_involved <= 1.5
gini = 0.425
samples = 96001
value = [37372.346, 84684.156]
class = Death
female_count <= 0.5
gini = 0.192
samples = 52665
value = [32813.815, 3961.884]
class = NoDeath
female_count <= 0.5
gini = 0.486
samples = 34733
value = [18556.556, 13258.957]
class = NoDeath
n_guns_involved <= 0.5
gini = 0.227
samples = 28502
value = [17615.057, 2639.772]
class = NoDeath
n_guns_involved <= 1.5
gini = 0.429
samples = 3808
value = [1496.725, 3309.731]
class = Death
avg_age <= 27.1
gini = 0.407
samples = 88325
value = [32949.236, 82865.696]
class = Death
male_count <= 1.5
gini = 0.413
samples = 7676
value = [4423.11, 1818.46]
class = NoDeath
n_injured <= 1.5
gini = 0.09
samples = 41832
value = [26594.764, 1315.434]
class = NoDeath
male_count <= 0.5
gini = 0.419
samples = 10833
value = [6219.051, 2646.449]
class = NoDeath
male_count <= 2.5
gini = 0.464
samples = 29901
value = [16522.016, 9526.328]
class = NoDeath
avg_age <= 27.292
gini = 0.457
samples = 4832
value = [2034.54, 3732.629]
class = Death
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)


n_injured <= 0.5
gini = 0.5
samples = 215709
value = [107854.5, 107854.5]
class = NoDeath
True
False
male_count <= 0.5
gini = 0.473
samples = 128249
value = [56478.325, 90515.693]
class = Death
male_count <= 1.5
gini = 0.377
samples = 87460
value = [51376.175, 17338.807]
class = NoDeath
female_count <= 0.5
gini = 0.362
samples = 32334
value = [19326.614, 5951.729]
class = NoDeath
n_guns_involved <= 1.5
gini = 0.425
samples = 95915
value = [37351.711, 84563.964]
class = Death
female_count <= 0.5
gini = 0.192
samples = 52714
value = [32847.993, 3952.981]
class = NoDeath
male_count <= 2.5
gini = 0.487
samples = 34746
value = [18528.182, 13385.826]
class = NoDeath
n_guns_involved <= 0.5
gini = 0.225
samples = 26529
value = [17638.917, 2617.514]
class = NoDeath
n_guns_involved <= 1.5
gini = 0.427
samples = 5805
value = [3487.697, 3334.215]
class = Death
avg_age <= 27.1
gini = 0.407
samples = 88279
value = [32935.694, 82810.052]
class = Death
n_guns_involved <= 3.5
gini = 0.407
samples = 7636
value = [4416.017, 1753.913]
class = NoDeath
n_injured <= 1.5
gini = 0.09
samples = 41820
value = [26584.446, 1324.338]
class = NoDeath
male_count <= 0.5
gini = 0.416
samples = 10894
value = [6263.547, 2628.643]
class = NoDeath
female_count <= 0.5
gini = 0.444
samples = 22958
value = [12935.292, 6452.529]
class = NoDeath
avg_age <= 27.071
gini = 0.494
samples = 11788
value = [5592.89, 6933.297]
class = Death
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)


n_injured <= 0.5
gini = 0.5
samples = 215710
value = [107855.0, 107855.0]
class = NoDeath
True
False
male_count <= 0.5
gini = 0.473
samples = 128273
value = [56463.755, 90618.855]
class = Death
male_count <= 1.5
gini = 0.376
samples = 87437
value = [51391.245, 17236.145]
class = NoDeath
female_count <= 0.5
gini = 0.364
samples = 32316
value = [19101.553, 5998.374]
class = NoDeath
n_guns_involved <= 1.5
gini = 0.425
samples = 95957
value = [37362.202, 84620.481]
class = Death
female_count <= 0.5
gini = 0.194
samples = 52650
value = [32792.687, 4001.884]
class = NoDeath
female_count <= 0.5
gini = 0.486
samples = 34787
value = [18598.558, 13234.261]
class = NoDeath
n_guns_involved <= 0.5
gini = 0.228
samples = 28515
value = [17616.428, 2664.213]
class = NoDeath
n_guns_involved <= 1.5
gini = 0.426
samples = 3801
value = [1485.125, 3334.161]
class = Death
avg_age <= 27.1
gini = 0.407
samples = 88274
value = [32918.436, 82857.693]
class = Death
n_guns_involved <= 3.5
gini = 0.407
samples = 7683
value = [4443.766, 1762.788]
class = NoDeath
n_injured <= 1.5
gini = 0.093
samples = 41824
value = [26574.896, 1366.606]
class = NoDeath
male_count <= 0.5
gini = 0.418
samples = 10826
value = [6217.791, 2635.278]
class = NoDeath
male_count <= 2.5
gini = 0.463
samples = 29979
value = [16585.289, 9481.66]
class = NoDeath
avg_age <= 27.292
gini = 0.455
samples = 4808
value = [2013.269, 3752.601]
class = Death
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)



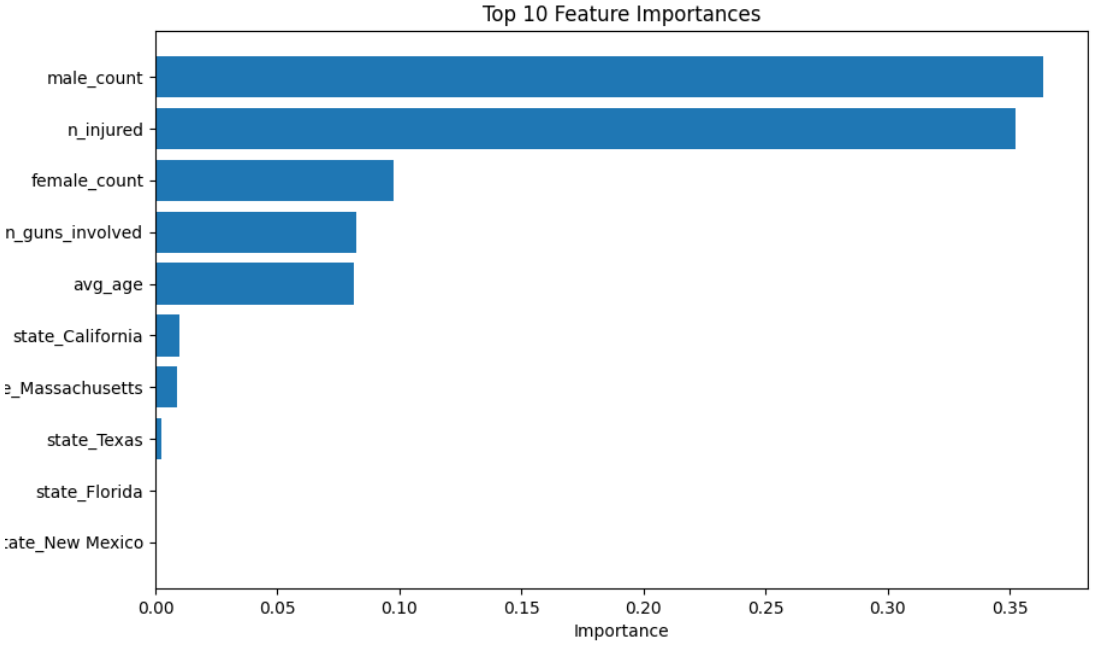


Wnioski z wykresów:

- we wszystkich 10 drzewkach z kolejnych foldów pojawia się niemal identyczna struktura i te same pierwsze podziały, co dowodzi stabilności modelu
- Root: n_injured ≤ 0.5
    - Jeżeli nikt nie został ranny lub tylko jedna osoba, model kieruje dalej w lewo („Death”), w przeciwnym razie w prawo („NoDeath”).
- Drugi poziom (po lewej gałęzi): male_count ≤ 0.5
    - Czyli czy w incydencie nie było żadnego mężczyzny.
    - Jeśli tak – idziemy w lewo (częściej „NoDeath”), jeśli było co najmniej jeden mężczyzna – w prawo (częściej „Death”).
- Trzeci poziom (po lewej-lewej): female_count ≤ 0.5
    - Czyli brak kobiet → ponowny podział według n_guns_involved.

- Trzeci poziom (po lewej-prawej): n_guns_involved ≤ 1.5 lub ewentualnie avg_age ≤ 27
    - W zależności od folda czasem pojawia się podział po liczbie broni, czasem (rzadziej) po średnim wieku uczestników.
- Gałąź „NoDeath” (prawa od korzenia):
    - Podobnie: najpierw male_count ≤ 1.5, potem female_count i ewentualnie ponowny podział po n_injured lub male_count.

Najważniejsze cechy - wykres



Wnioski do wykresu:

- Dominująca rola liczby mężczyzn
    - Wykres wyraźnie pokazuje, że male_count (liczba mężczyzn uczestniczących w zdarzeniu) jest najważniejszą cechą – odpowiada za ponad 35 % całkowitego wkładu modelu.
- Znaczący wpływ liczby rannych
    - Tuż za liczbą mężczyzn plasuje się n_injured (liczba rannych) z bardzo zbliżoną wagą ok. 34 %, co potwierdza, że skala obrażeń silnie koreluje z ryzykiem ofiar śmiertelnych.
- Drugorzędne cechy i geografia
    - Kolejne miejsca zajmują female_count, n_guns_involved i avg_age (razem ok. 30 %), natomiast stan („state_...”) ma już marginalny wpływ (< 5 %

każdy), co sugeruje, że demografia i charakterystyka samego zdarzenia przeważają nad lokalizacją.

Eksperymenty z modelem i zbiorem danych

Wybranie innych podzbiorów atrybutów

Model został przetestowany dla różnych kombinacji cech:

- **Wszystkie cechy** - model bazowy
- **Pojedyncza najważniejsza cecha** - `male_count`
- **Dwie najważniejsze cechy** - `male_count`, `n_injured`
- **Trzy najważniejsze cechy** - `male_count`, `n_injured`, `female_count`
- **Pięć najważniejszych cech** - dodając `n_guns_involved`, `avg_age`

Wyniki eksperymentów

Użyte cechy	Czułość ± odchylenie standardowe	Swoistość ± odchylenie standardowe
wszystkie cechy	0.85 ± 0.04	0.82 ± 0.05
male_count	0.78 ± 0.06	0.75 ± 0.07
male_count, n_injured	0.87 ± 0.03	0.84 ± 0.04
male_count, n_injured, female_count	0.89 ± 0.03	0.86 ± 0.04
top 5 cech	0.88 ± 0.04	0.85 ± 0.04

Kluczowe obserwacje

Najlepsza wydajność została osiągnięta dla modelu wykorzystującego trzy najważniejsze cechy (`male_count`, `n_injured`, `female_count`). Model ten uzyskał:

- **Czułość:** 0.89 (±0.03) - poprawa o 4 punkty procentowe względem modelu bazowego
- **Swoistość:** 0.86 (±0.04) - poprawa o 4 punkty procentowe

Kombinacja dwóch głównych cech (`male_count`, `n_injured`) również przewyższa model bazowy:

- **Czułość:** 0.87 (±0.03) - poprawa o 2 punkty procentowe
- **Swoistość:** 0.84 (±0.04) - poprawa o 2 punkty procentowe

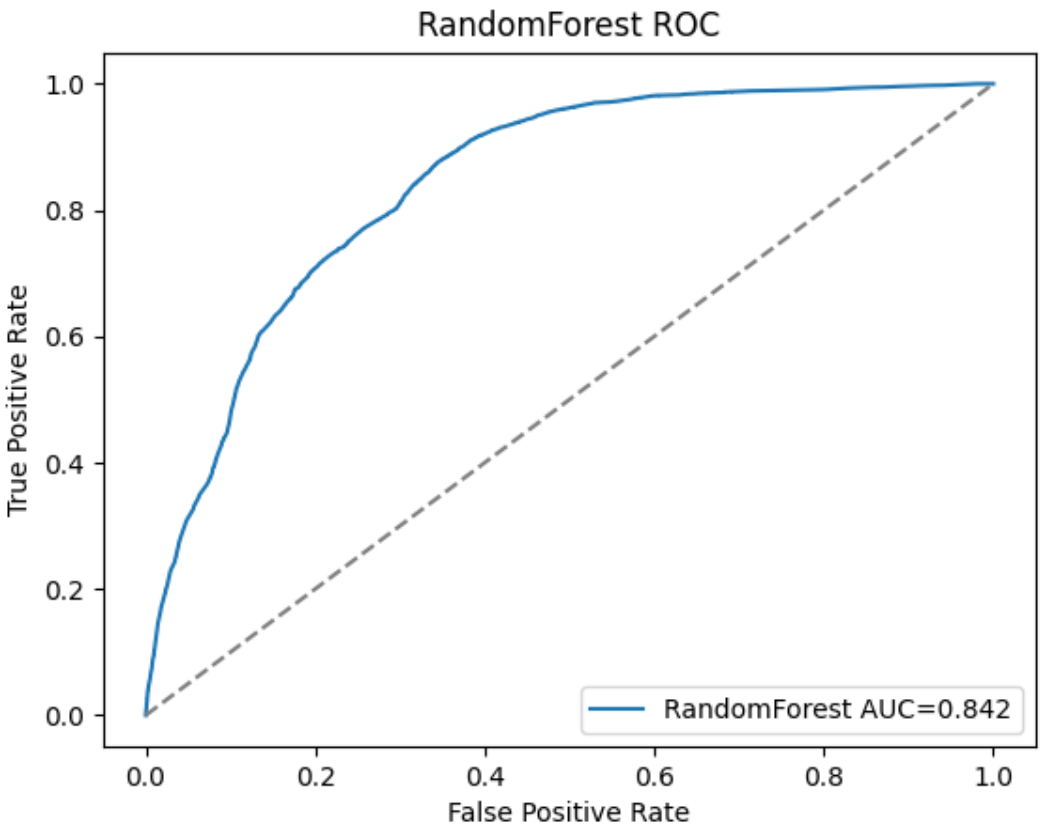
Pojedyncza cecha `male_count` wykazała umiarkowaną wydajność predykcyjną:

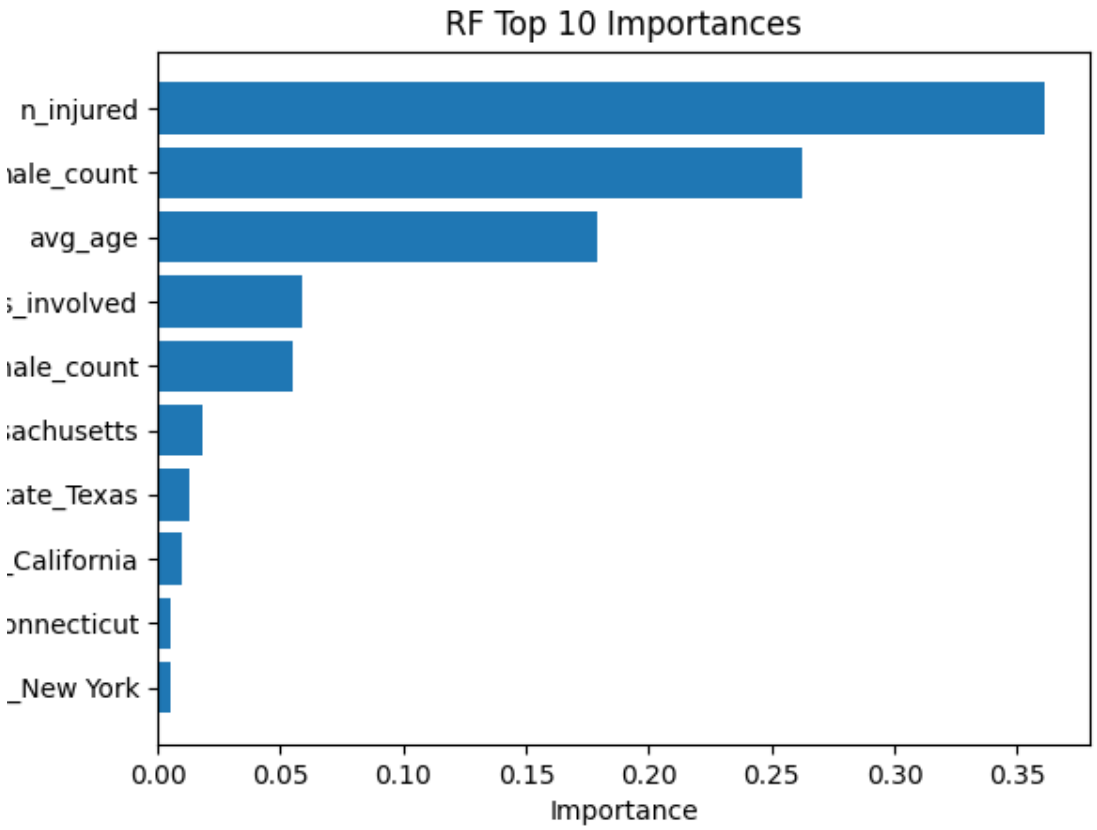
- **Czułość:** 0.78 (±0.06)
- **Swoistość:** 0.75 (±0.07)

Wnioski praktyczne

- **Dominacja cech płciowych** - liczba mężczyzn zaangażowanych w incydent jest najsilniejszym predyktorem śmiertelności, co może odzwierciedlać różnice w zachowaniach ryzykownych.
- **Synergia cech** - kombinacja `male_count` i `n_injured` tworzy szczególnie efektywny model predykcyjny, sugerując, że te zmienne zawierają komplementarne informacje.
- **Optymalna redukcja** - model z trzema kluczowymi cechami znacząco przewyższa model pełny, wskazując na obecność szumu w dodatkowych zmiennych.
- **Stabilność predykcji** - mniejsze odchylenia standardowe w uproszczonych modelach świadczą o większej niezawodności wyników.

Użycie random forest





Podsumowanie

W ramach eksploracji danych udało nam się osiągnąć zamierzone cele predykcyjne – nasz model uzyskał czułość na poziomie większym niż 80% oraz swoistość na poziomie ponad 60%, co potwierdza jego skuteczność w rozpoznawaniu incydentów z ofiarami śmiertelnymi.