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

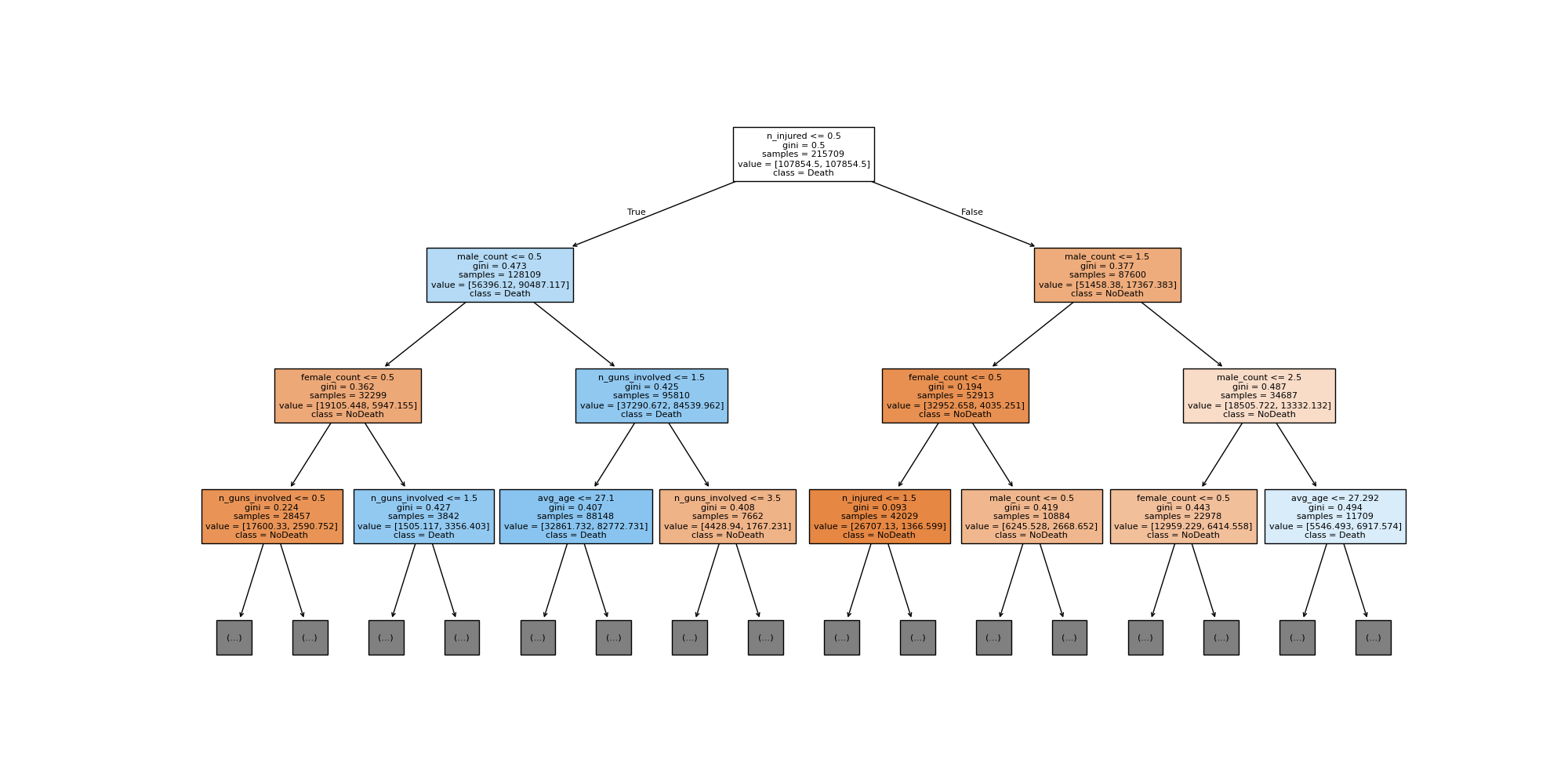
Dla uproszczenia i uniknięcia nadmiernego dopasowania modelu wybrano ostatecznie max\_depth=7 (zgodnie z wynikami wstępnych testów).

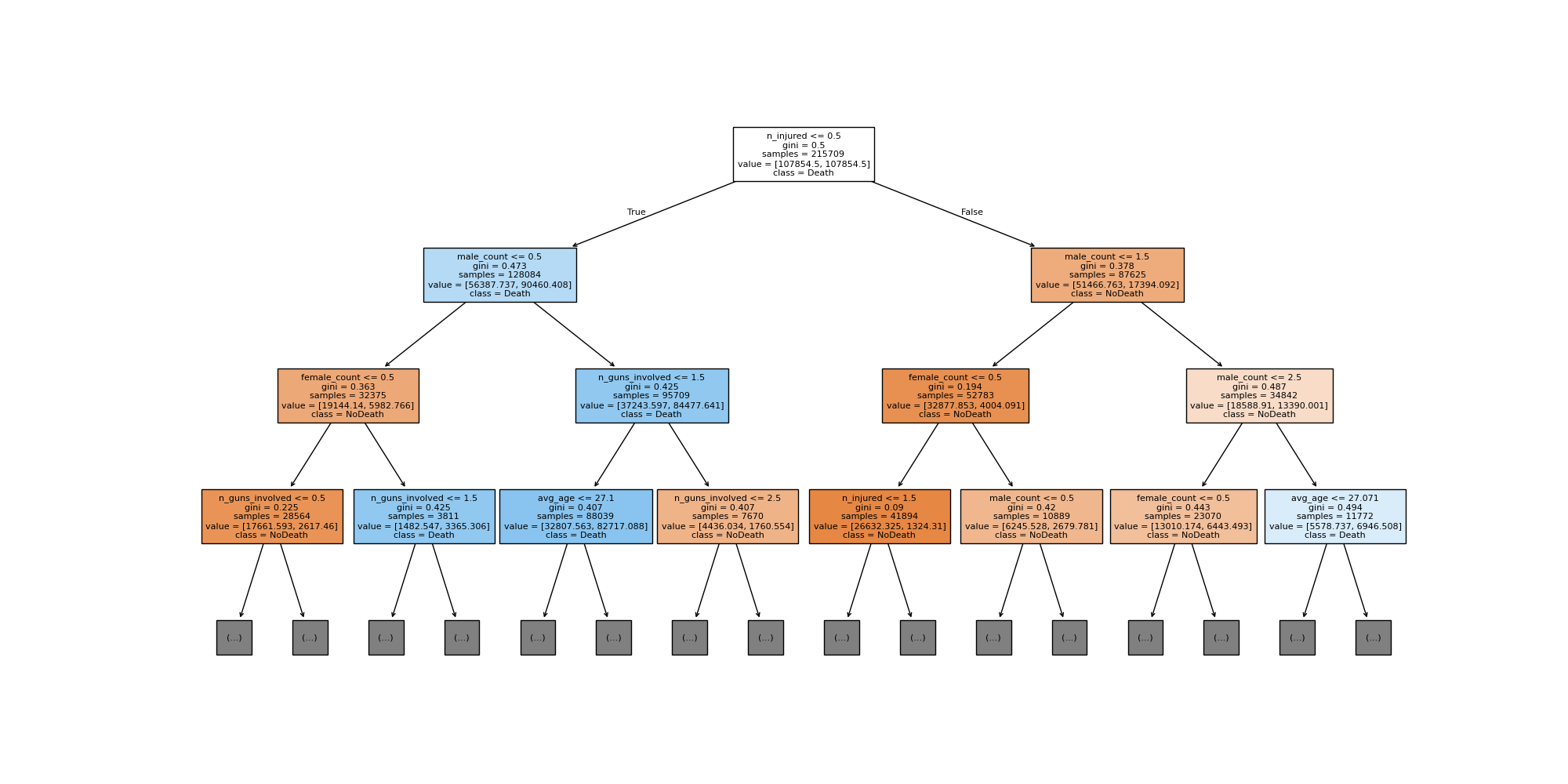
W wyniku 10-krotnej walidacji krzyżowej powstało 10 modeli drzew decyzyjnych, z których każdy został wytrenowany na nieco innym podzbiorze danych, co pozwoliło na ocenę stabilności i uogólnienia modelu.

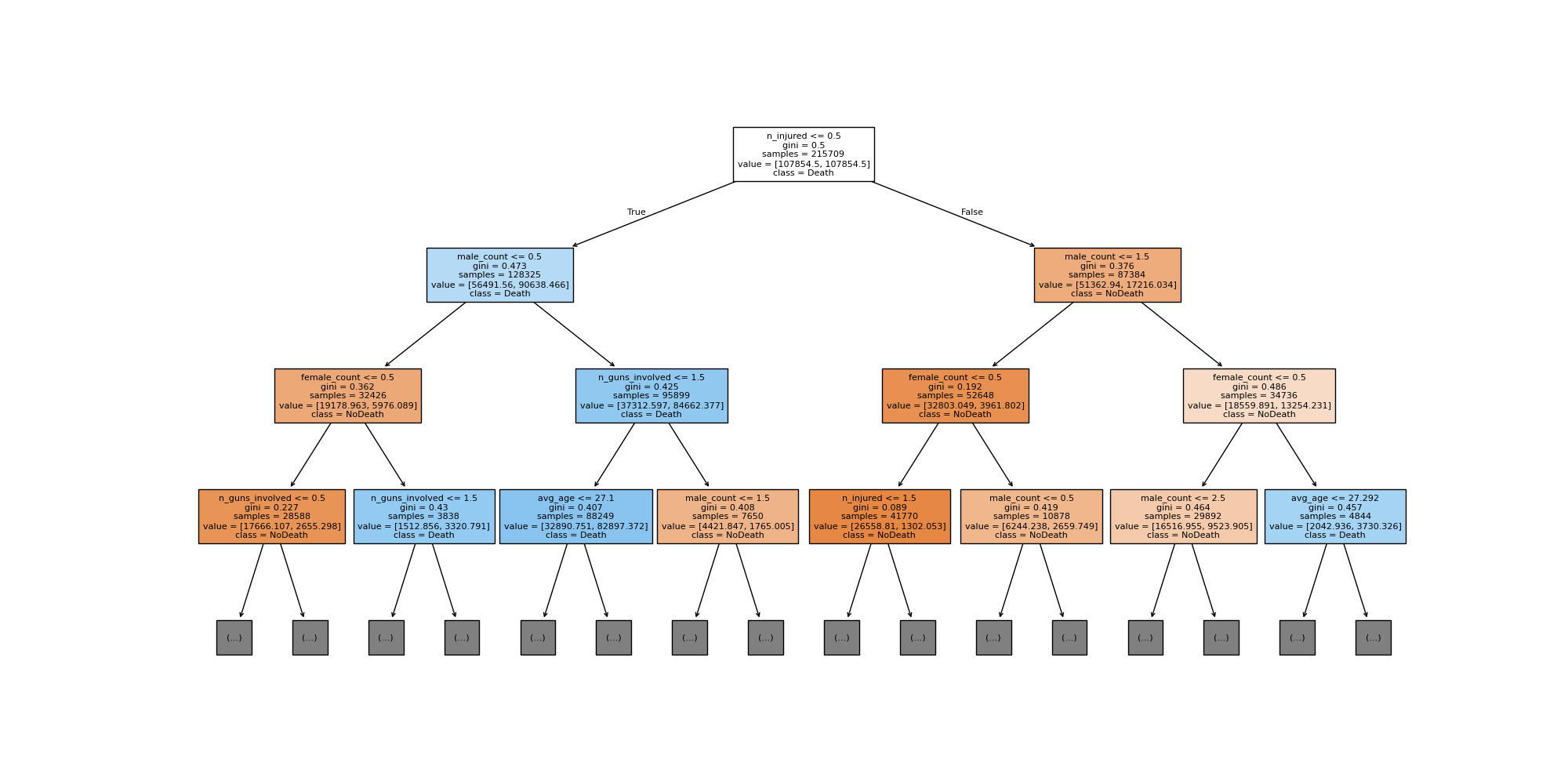
Poniżej zaprezentowano wykresy drzew decyzyjnych dla kolejnych foldów oraz zapisano reguły decyzyjne w formacie tekstowym, co umożliwia interpretację działania modelu.

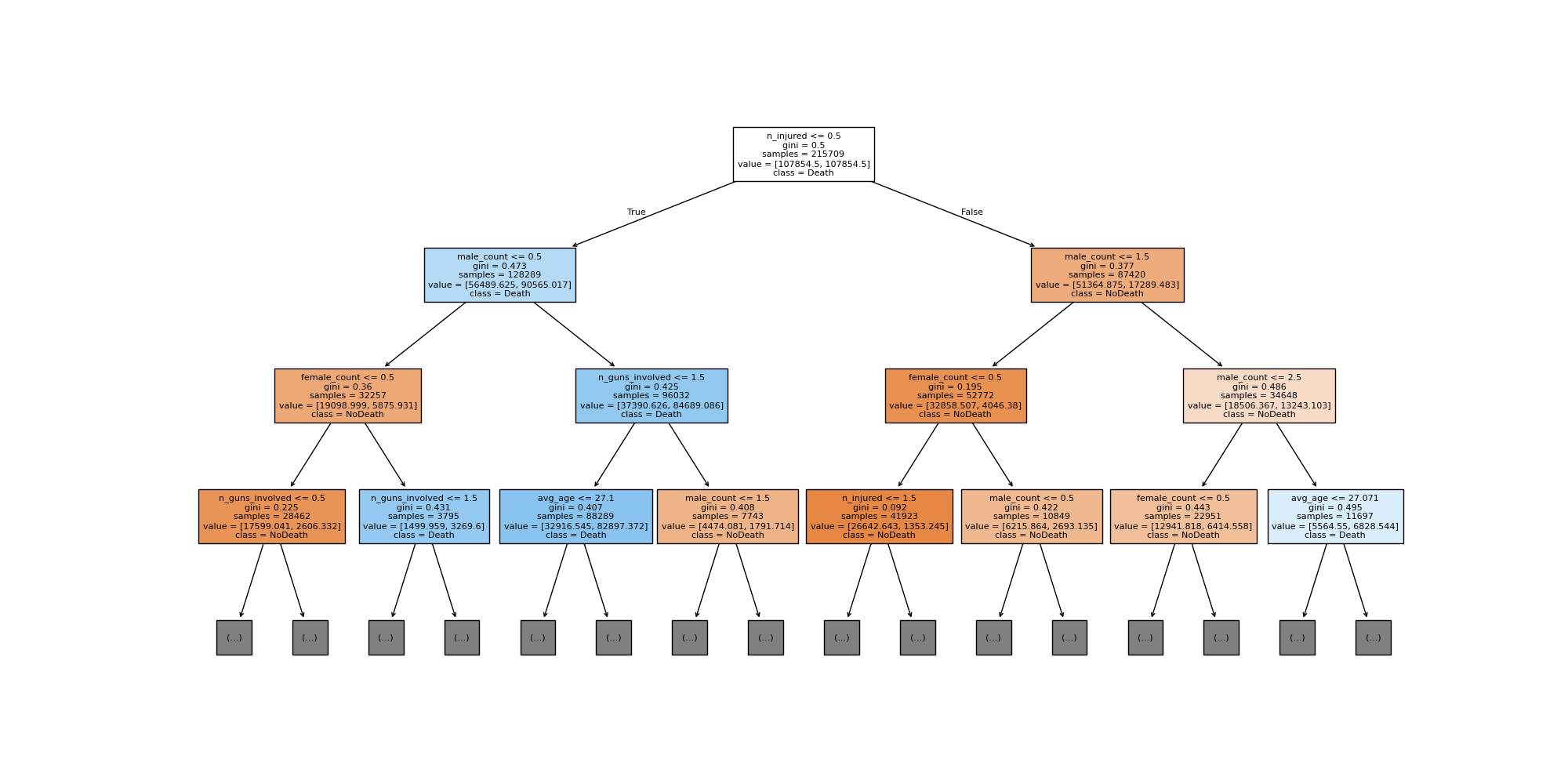
Podsumowując, ostateczny model drzewa decyzyjnego ma następujące parametry:  
random\_state=42, class\_weight='balanced', max\_depth=7  
oraz osiąga średnią czułość około 0.82 i średnią swoistość około 0.63 (wyniki z 10-krotnej kroswalidacji).

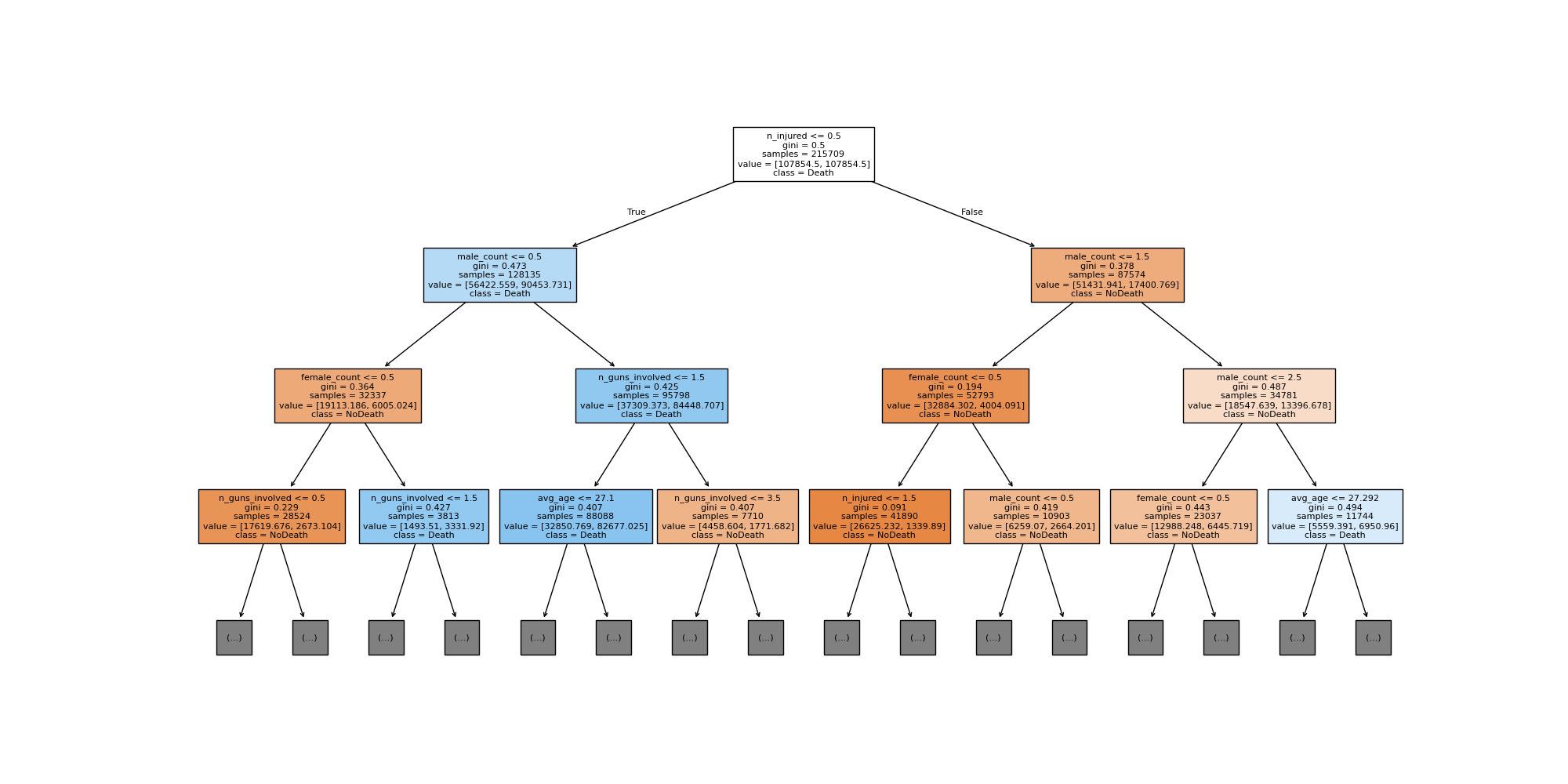
## Decision Tree – wykresy z krosswalidacji

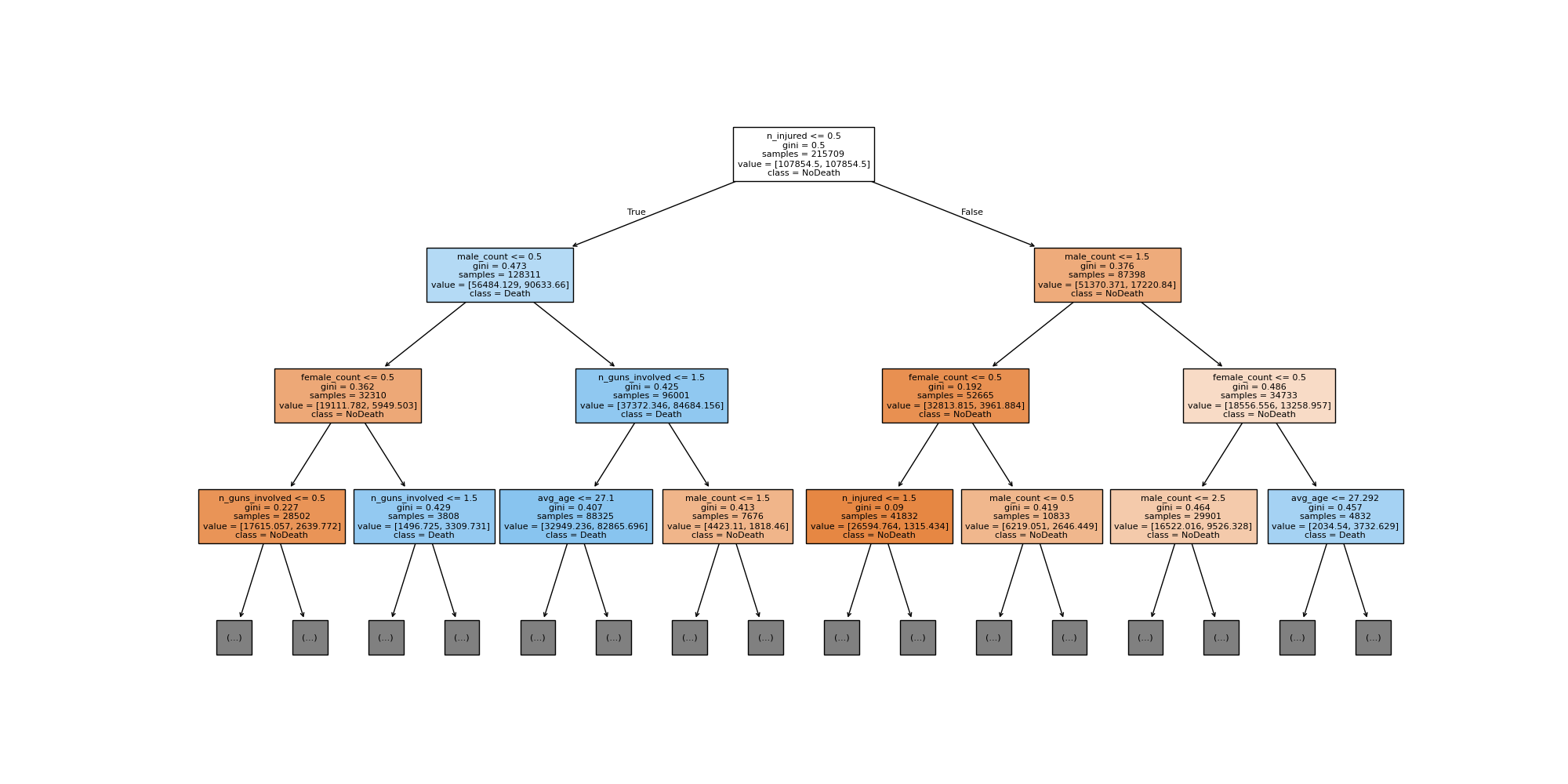


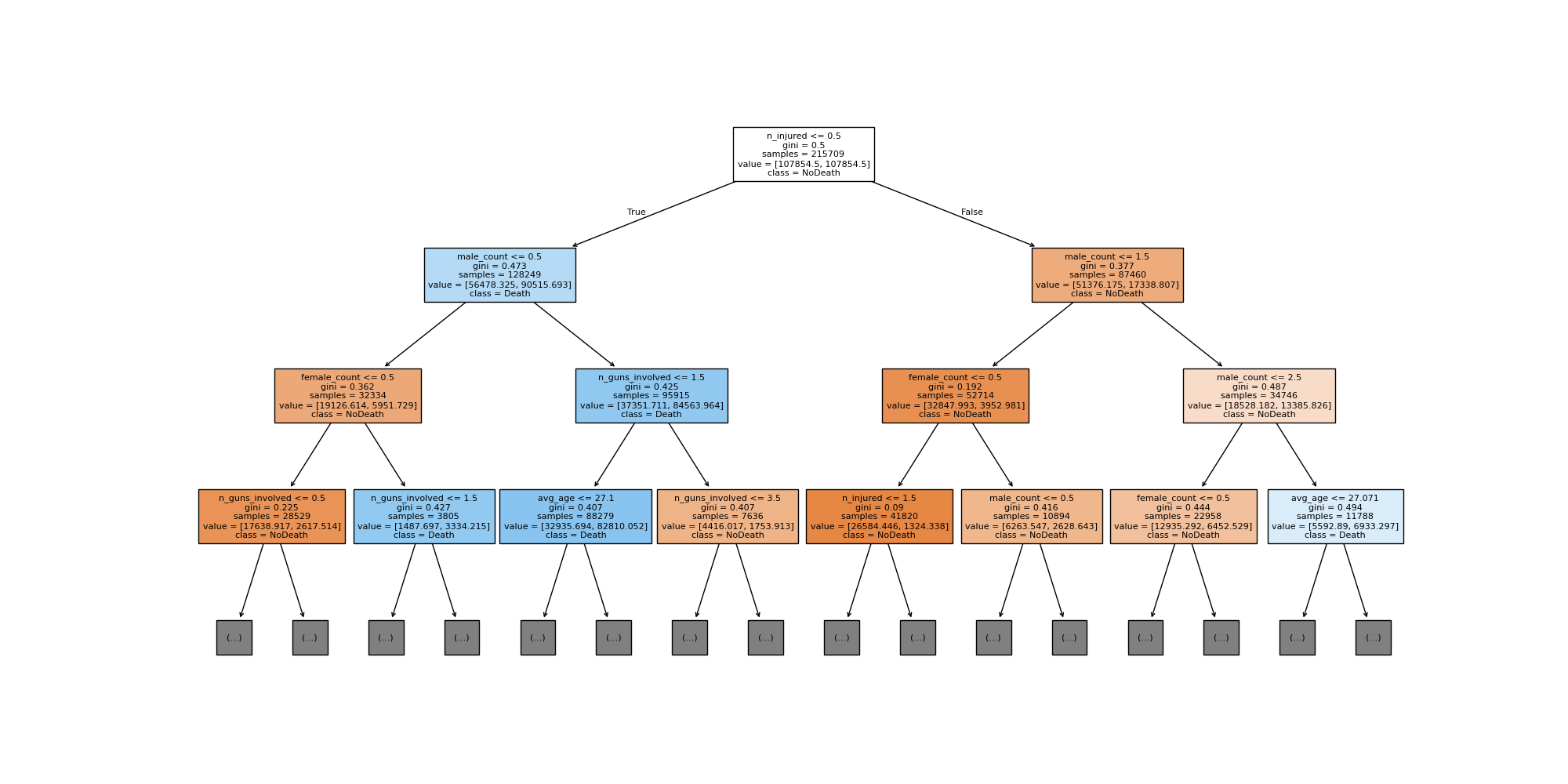


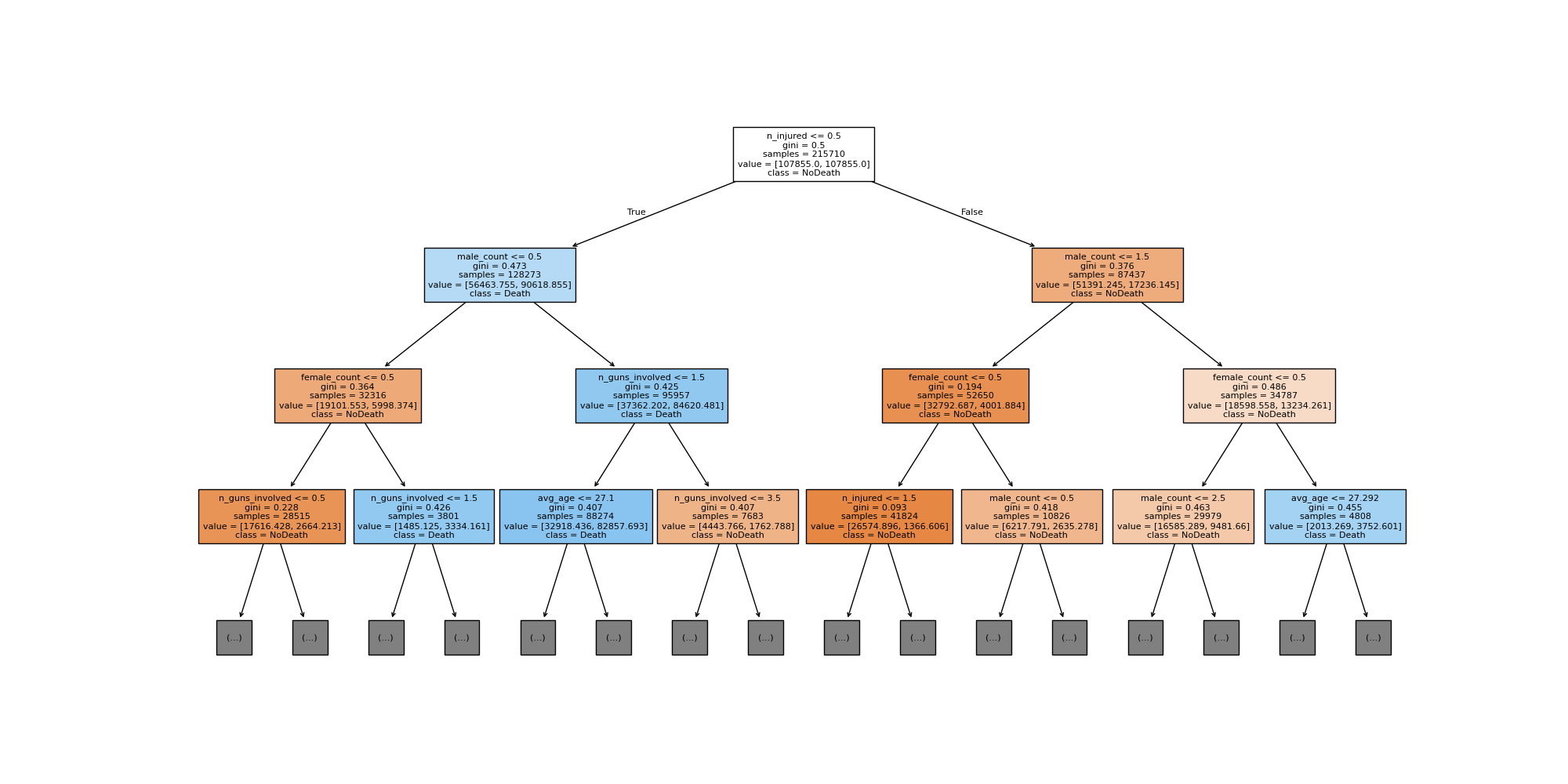


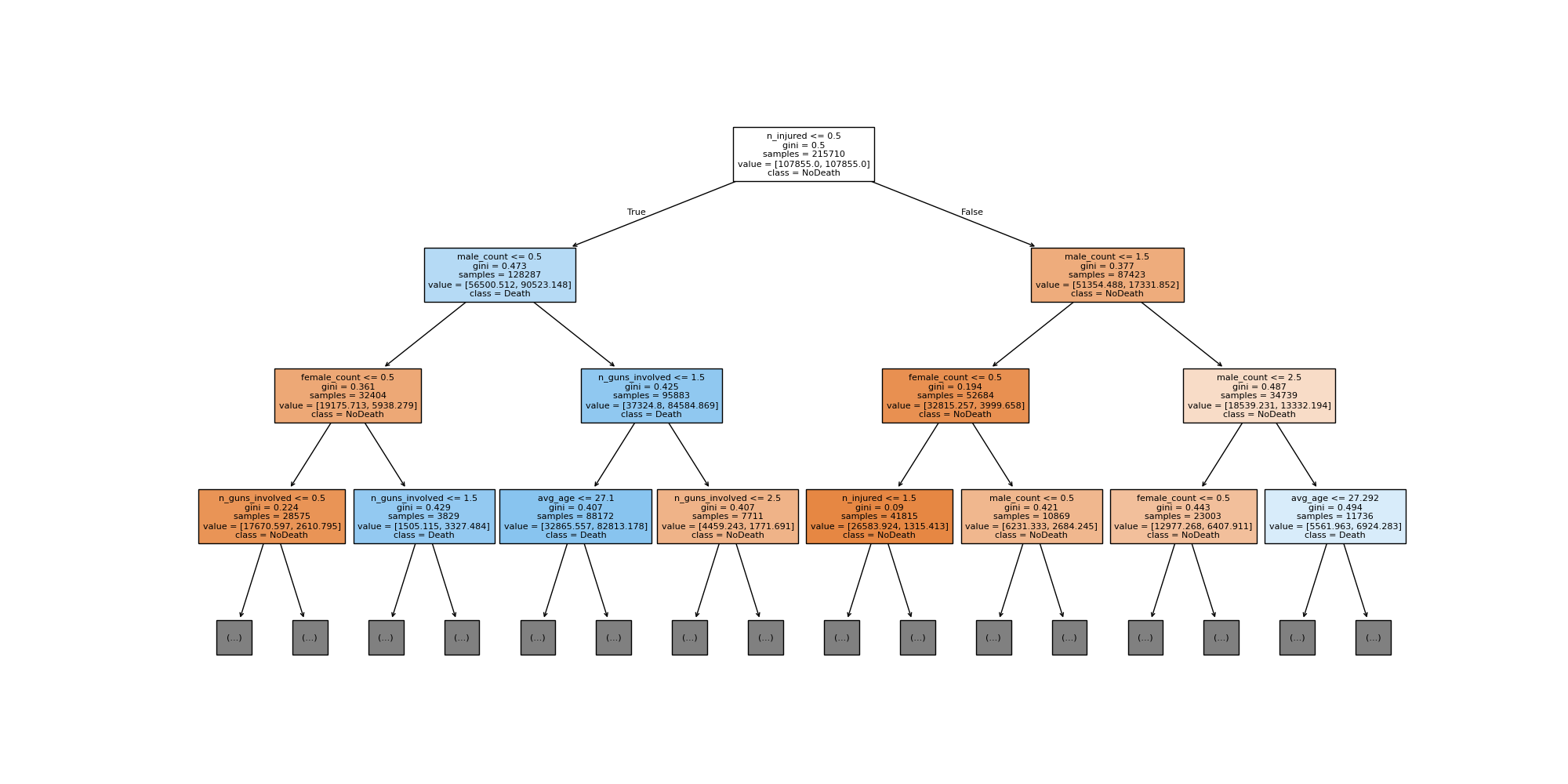


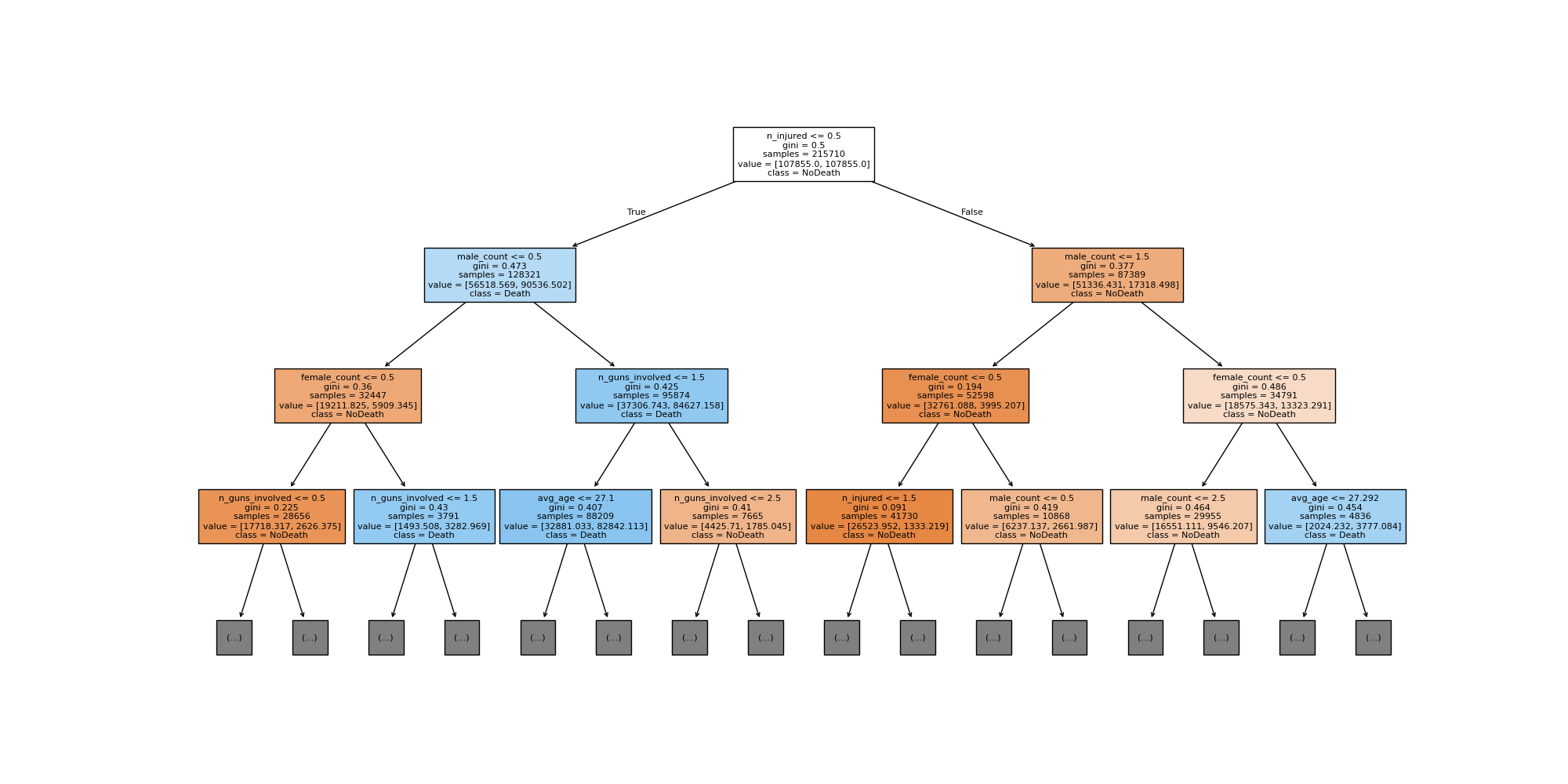








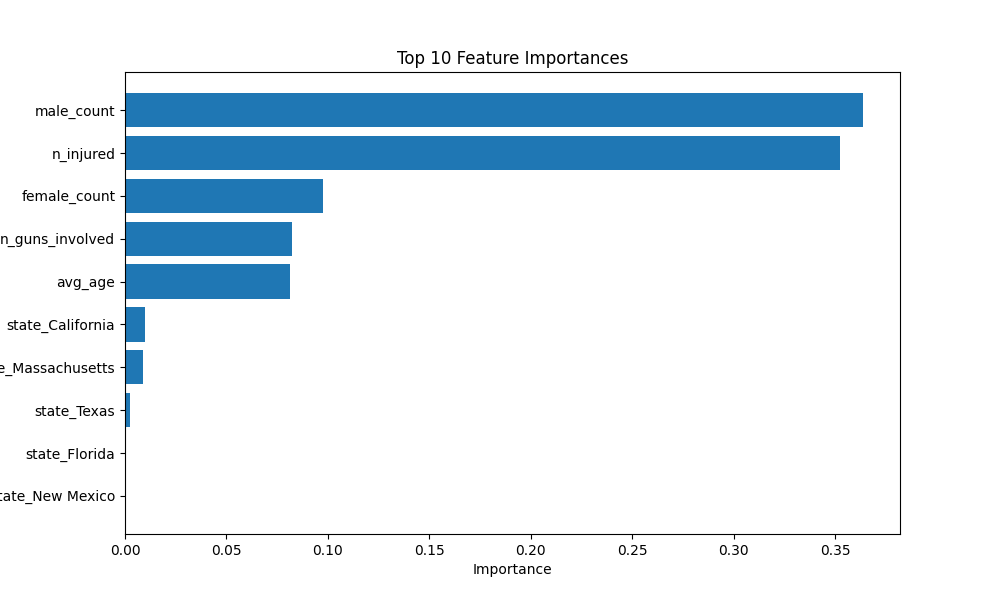




Wnioski z wykresów:

* 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

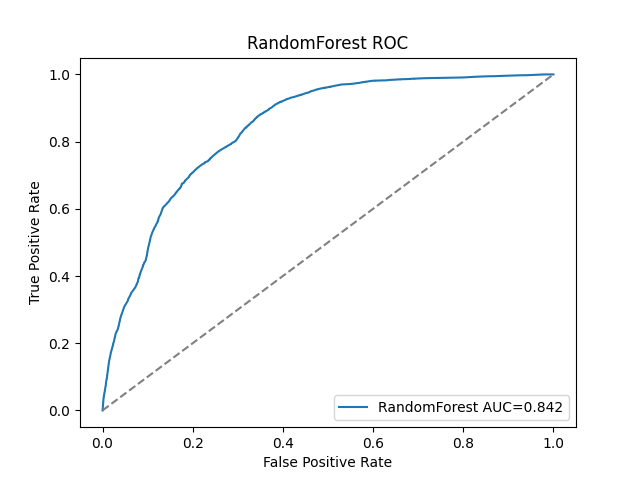
Pojedyncza cecha male\_count wykazała umiarkowaną wydajność predykcyjną:

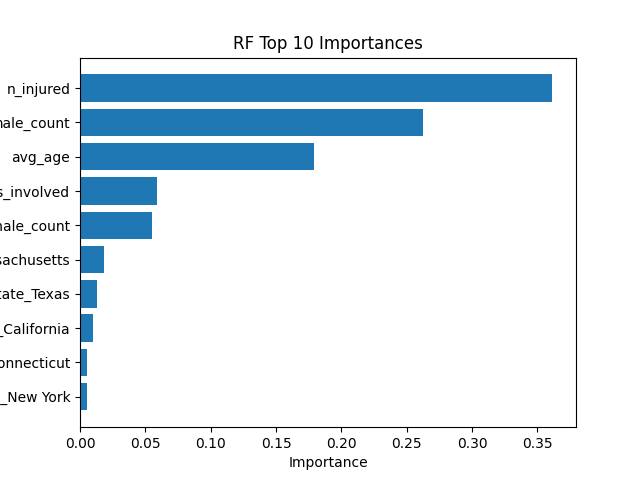
* **Czułość:** 0.78 (±0.06)
* **Swoistość:** 0.75 (±0.07)

### 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.