Eksploracja danych

Agnieszka Kulesz, Hania Gibus, Igor Józefowicz

Opis zbioru

Gun violence data

Szczegółowe informacje o incydentach z użyciem broni palnej na terenie Stanów Zjednoczonych w latach 2013–2018

Cel eksploracji

- Predykcja czy incydent z użyciem broni zakończy się ofiarami śmiertelnymi (n_killed > 0)
- Czułość ≥ 80%
- Swoistość ≥ 60%
- Klasyfikacja binarna

Dobór algorytmu eksploracji

Wybrany algorytm - drzewo decyzyjne

- Wysoka interpretowalność
- Jasne reguły decyzyjne
- Identyfikacja progów wartości

Dobór metody testowania wyników

Kroswalidacja 10-krotna

Korzyści:

- Eliminacja wpływu przypadkowego podziału
- Uśrednienie wyników
- Wyższa wiarygodność oceny

Przygotowanie danych

Braki, transformacja i uzupełnienie, podzbiór danych

Braki danych: Brakujące wartości w kolumnach tekstowych (np. `gun_stolen`, `participant_age_group`, `participant_gender`) zostały ujednolicone jako *"Unknown"*

Transformacja danych: Kolumny ze złożonymi łańcuchami tekstowymi (np. *gun_type*) zostały rozbite na unikalne kategorie i zakodowane jako zmienne nominalne.

Podzbiór danych

Wybrane cechy

- **`n_injured`** – liczba rannych
- **`n_guns_involved`** – liczba użytej broni
- **`state`** – stan USA
- **`avg_age`** – średni wiek uczestników
- **`male_count`, `female_count`** – liczba uczestników według płci

Uzupełnienie danych

Cel analizy: Przewidzenie, czy incydent zakończy się ofiarą śmiertelną (*mortality* = *1* jeśli *n_killed* > 0).

Problematyka klas niezbalansowanych

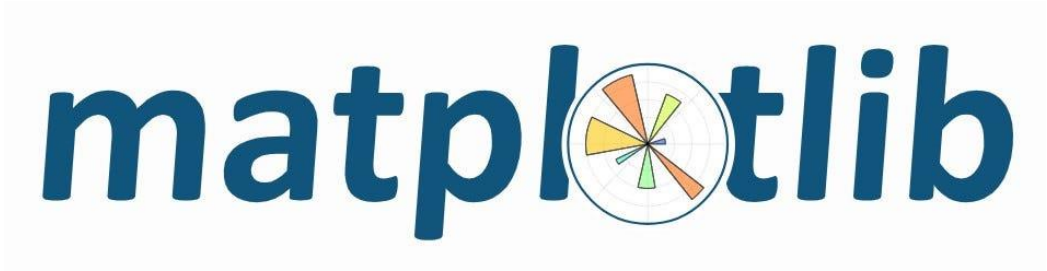
- Incydenty bez ofiar śmiertelnych są liczniejsze, ale odzwierciedlają rzeczywistość

Metody radzenia sobie z nierównowagą klas

- Wagi klas proporcjonalne do ich częstości
- Metoda **SMOTE** – syntetyczne zwiększenie próbek klasy mniejszościowej

Utworzenie modelu: Drzewo decyzyjne

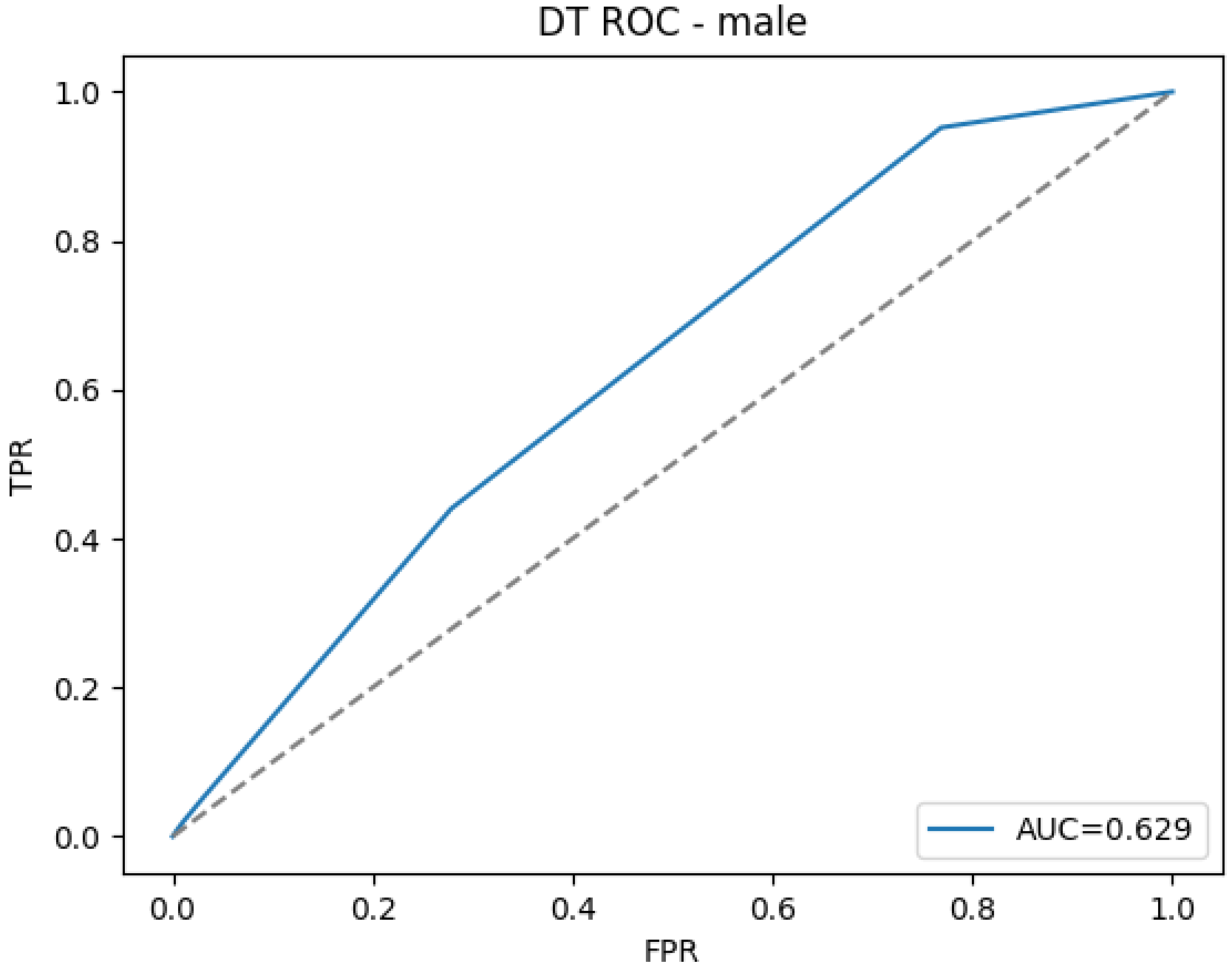
Wykorzystane biblioteki

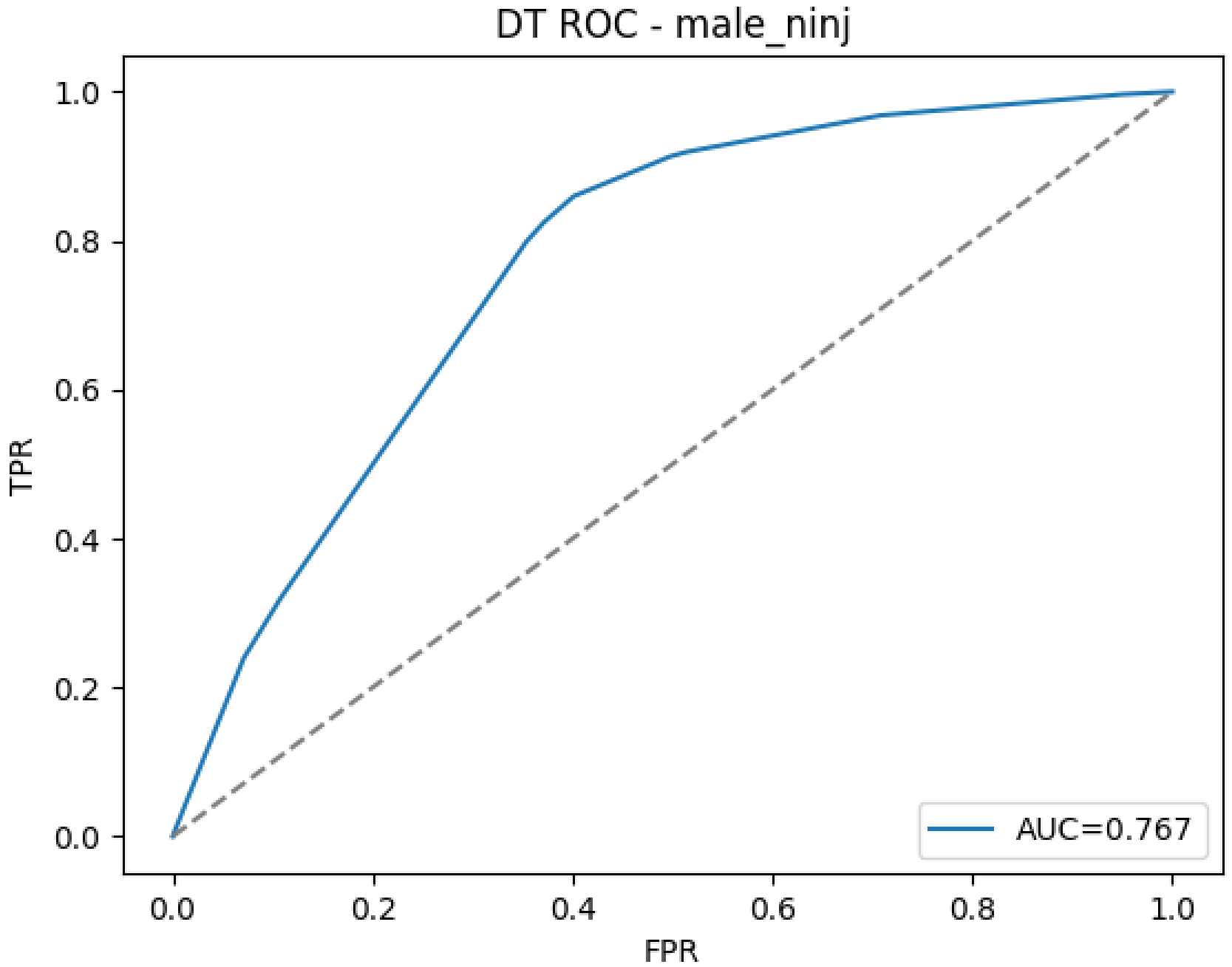


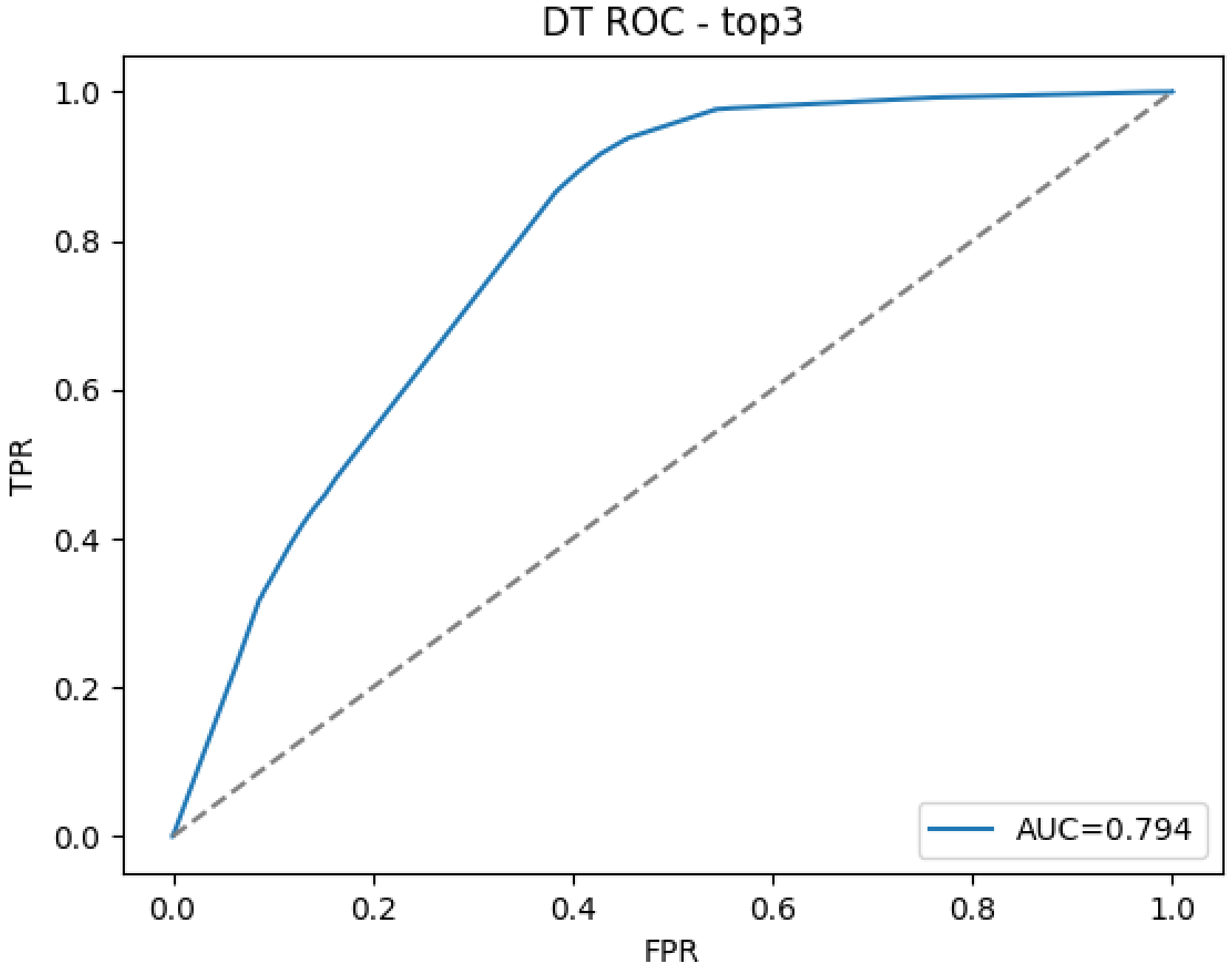
Najważniejsze importy

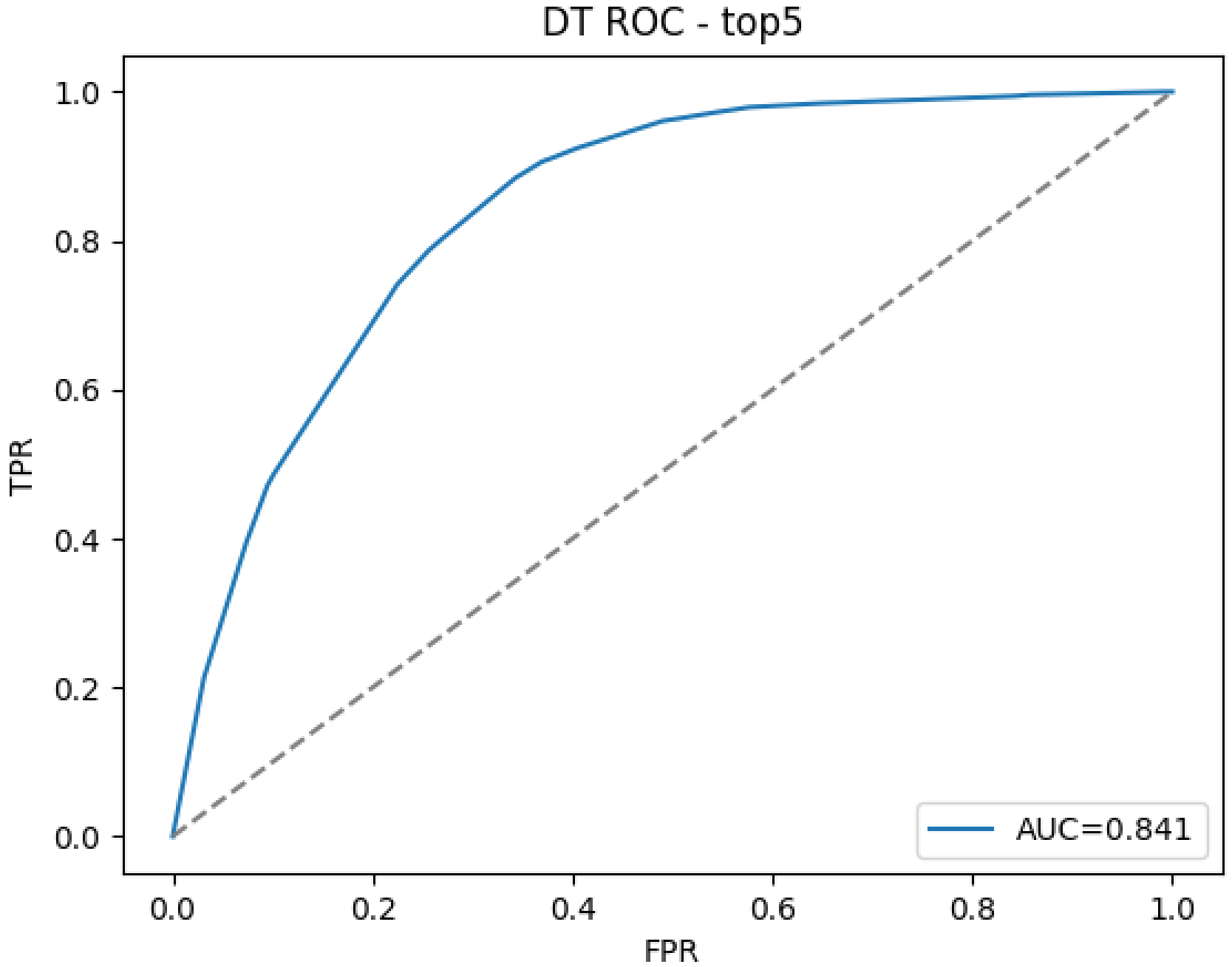
```
from sklearn.model_selection import train_test_split,
GridSearchCV, StratifiedKFold


from sklearn.preprocessing import OneHotEncoder


from sklearn.tree import DecisionTreeClassifier, plot_tree


from sklearn.ensemble import RandomForestClassifier
```

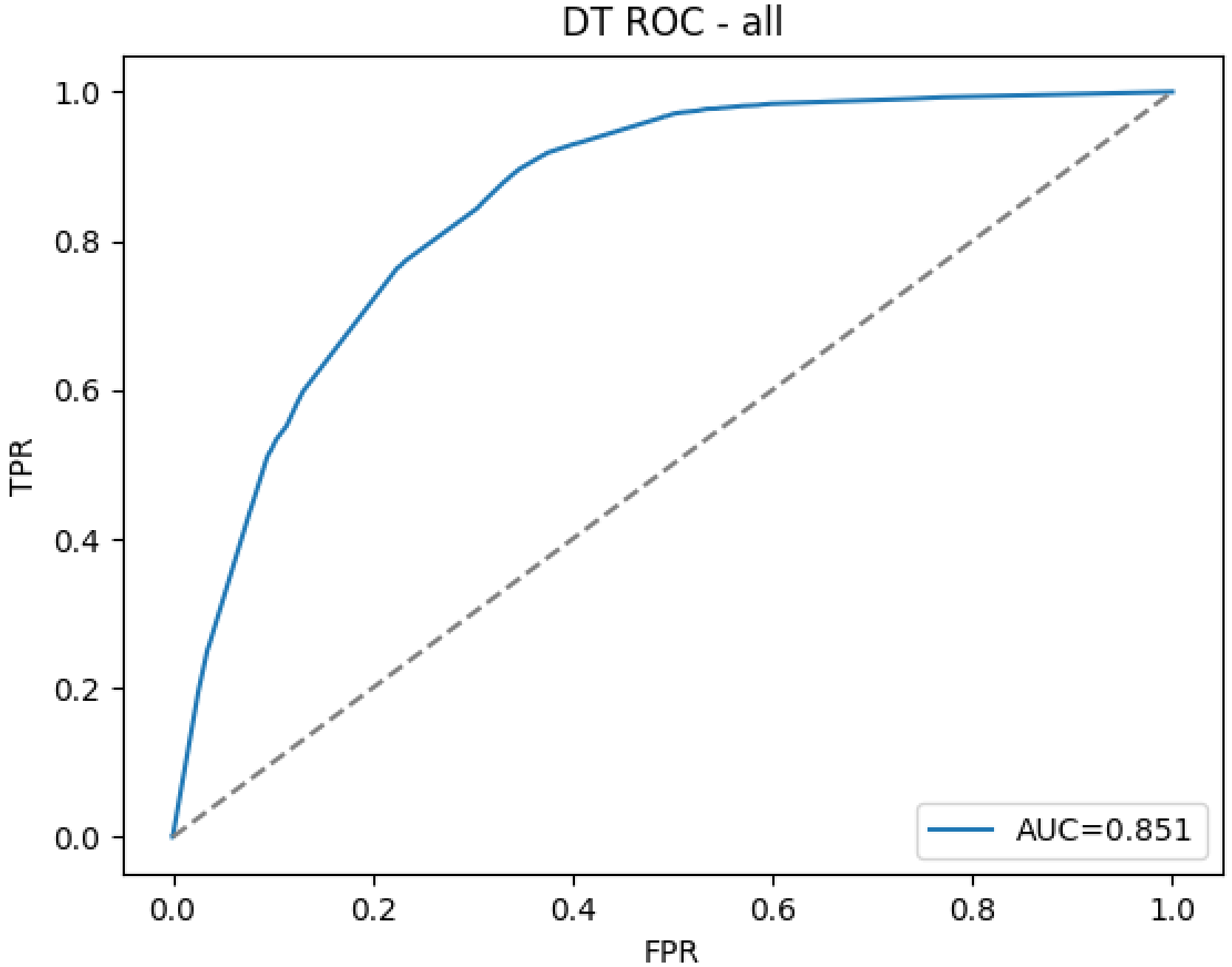
Wyniki i eksperymenty z modelem i zbiorem danych

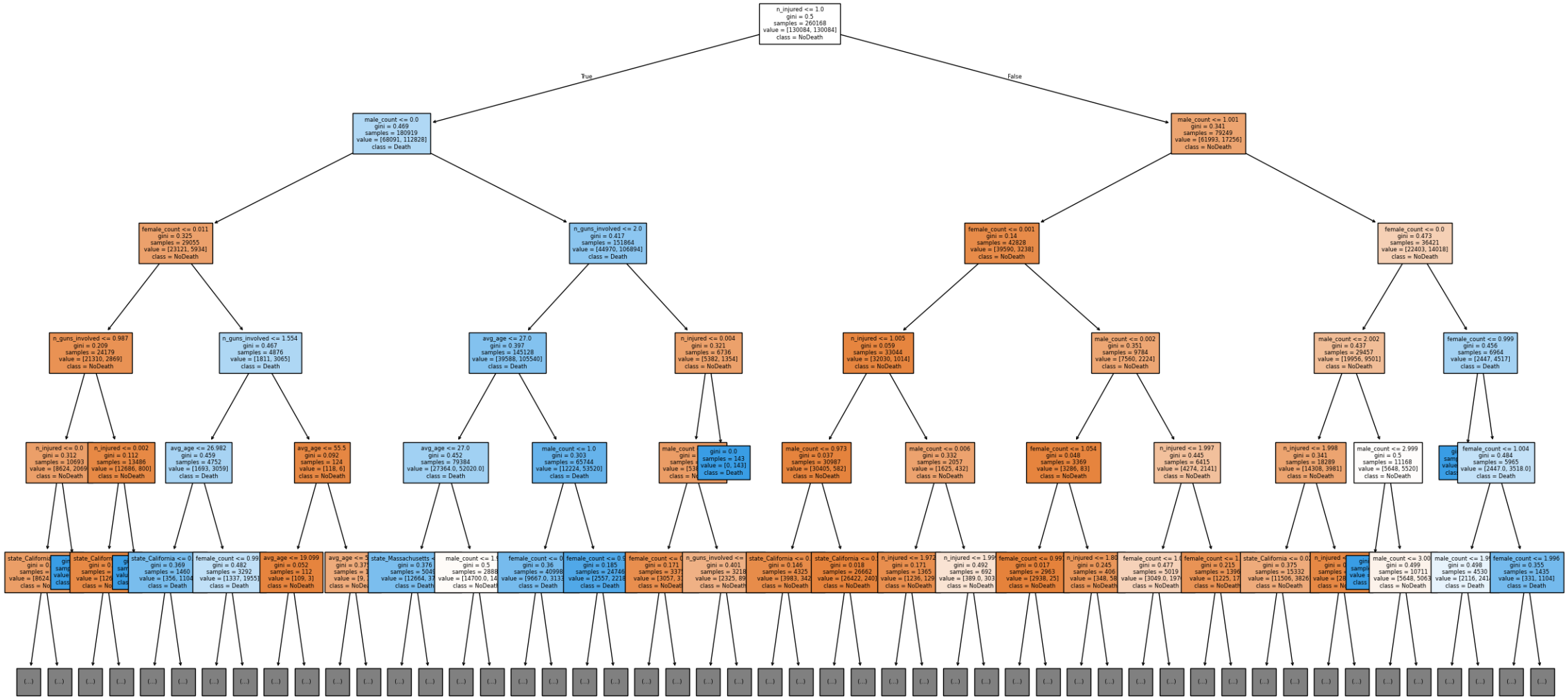

DT ROC - male
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
AUC=0.629

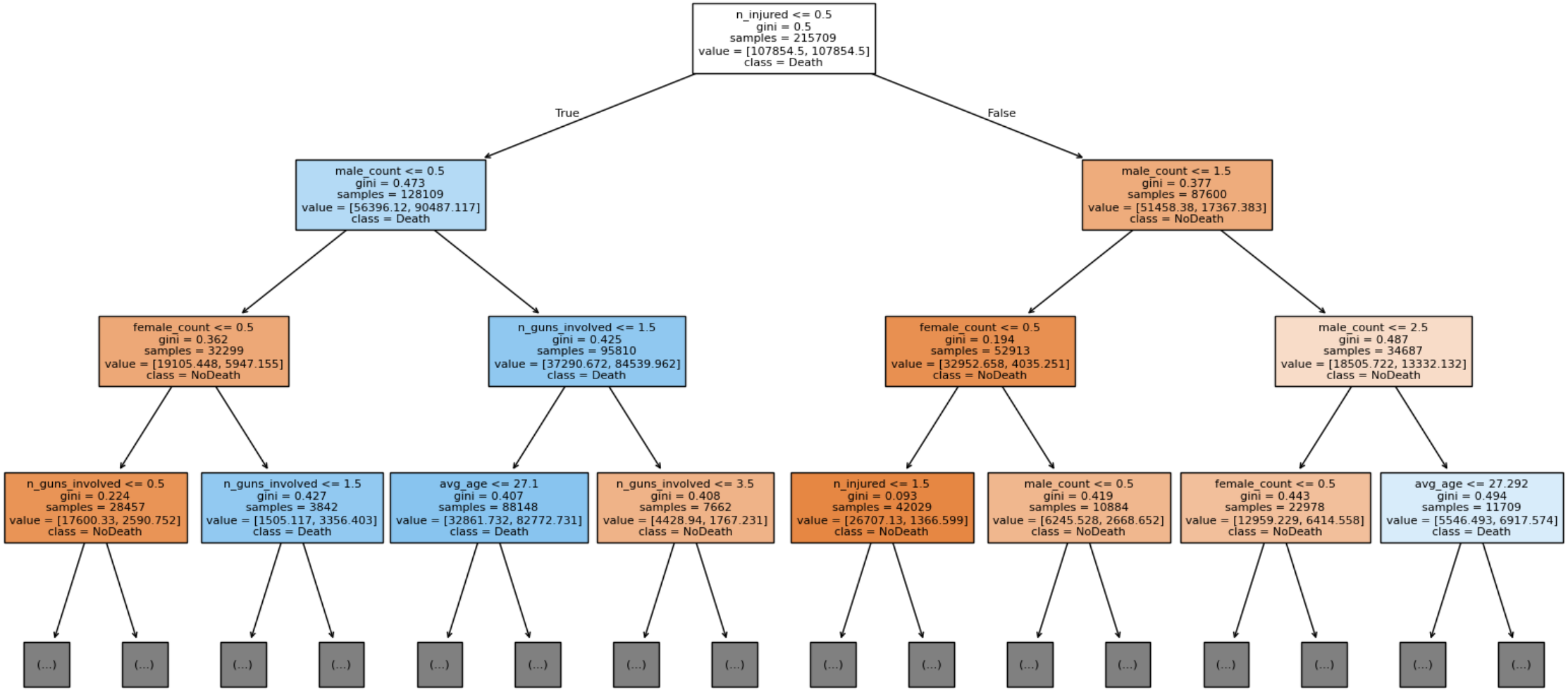

DT ROC - male_ninj
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
AUC=0.767

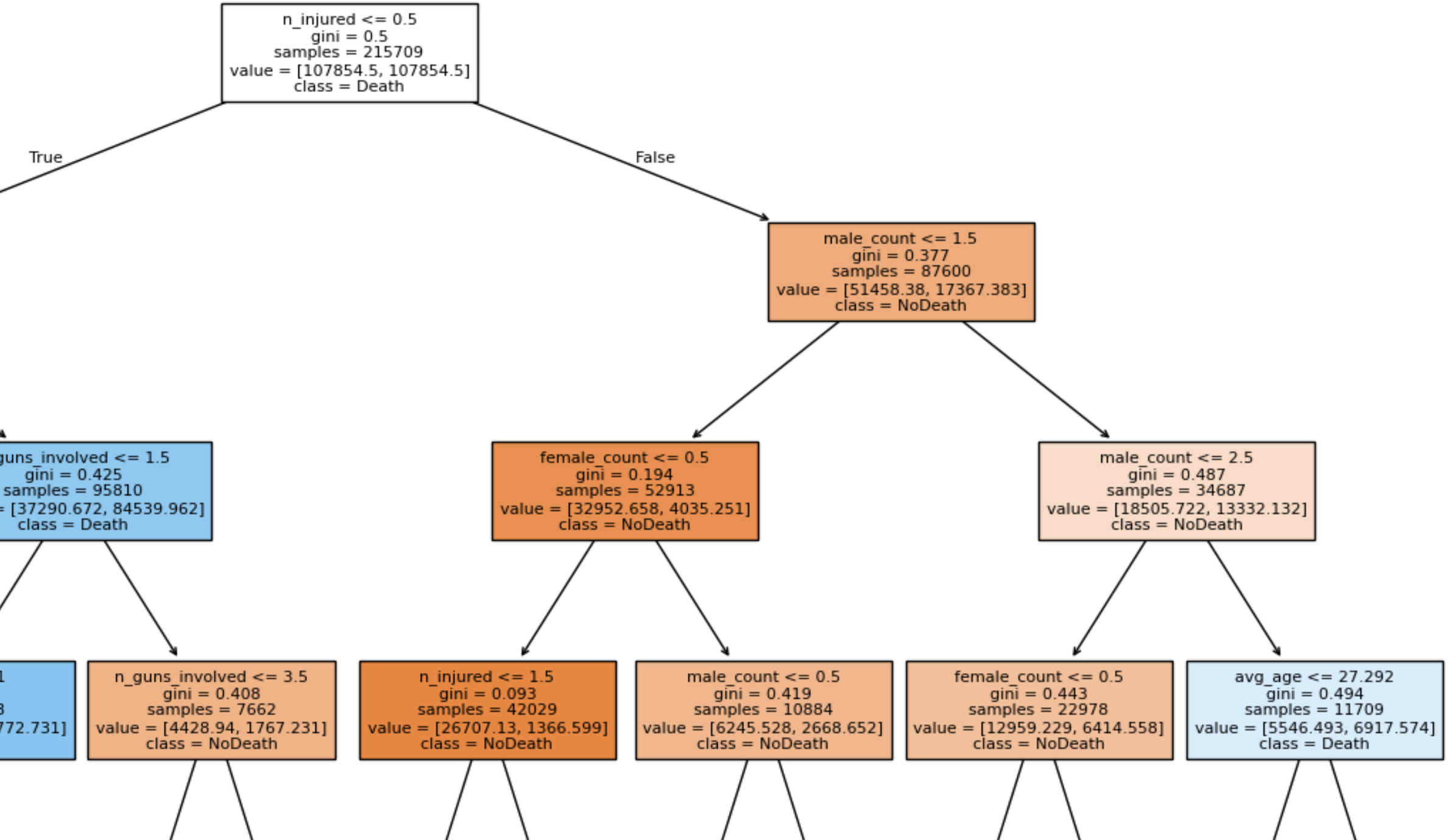

DT ROC - top3
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
AUC=0.794

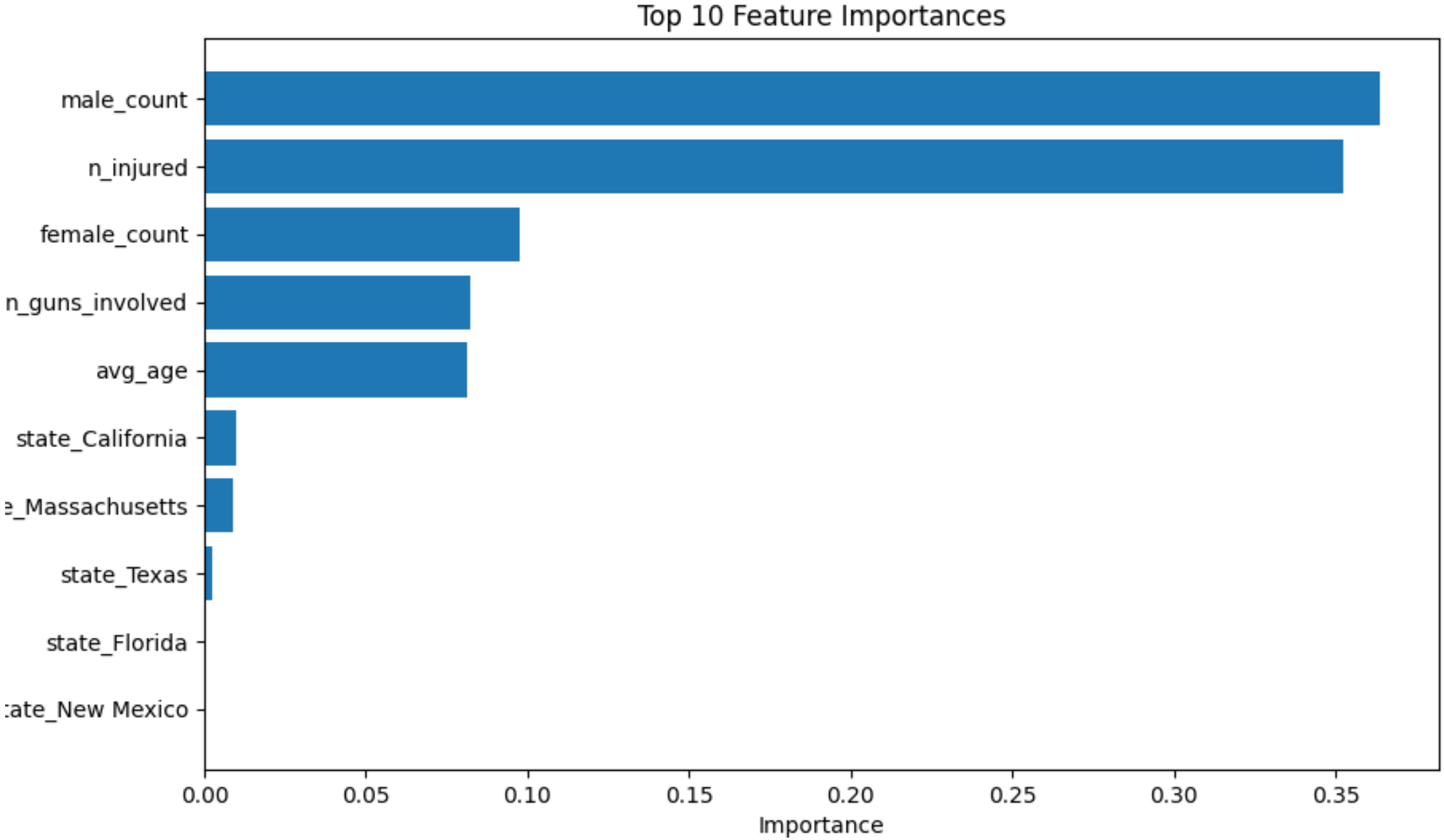

DT ROC - top5
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
AUC=0.841


DT ROC - all
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
AUC=0.851

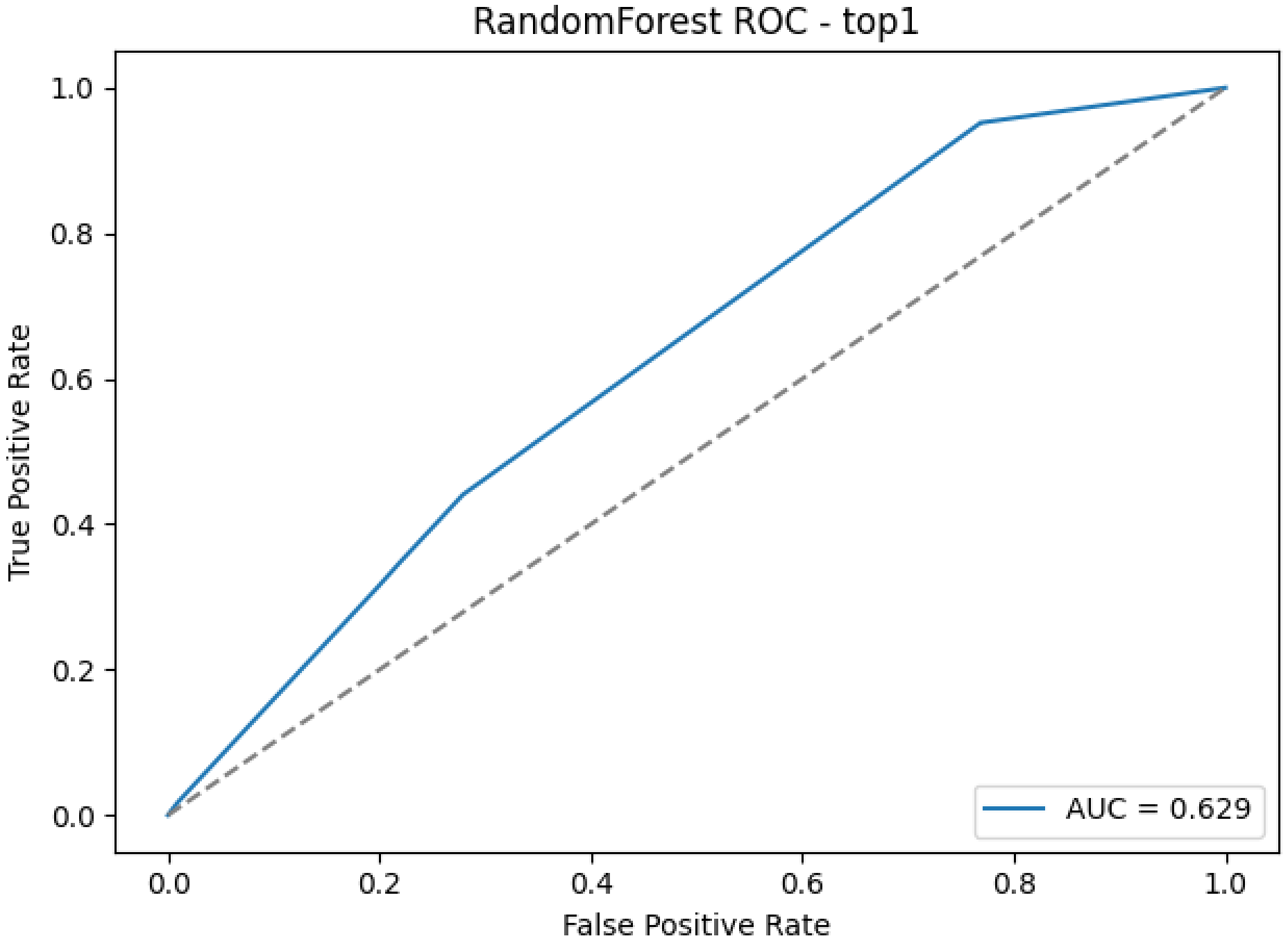

n_injured <= 1.0
gini = 0.5
samples = 260168
value = [130084, 130084]
class = NoDeath
True
False
male_count <= 0.0
gini = 0.469
samples = 180919
value = [68091, 112828]
class = Death
male_count <= 1.001
gini = 0.341
samples = 79249
value = [61993, 17256]
class = NoDeath
female_count <= 0.011
gini = 0.325
samples = 29055
value = [23121, 5934]
class = NoDeath
n_guns_involved <= 2.0
gini = 0.417
samples = 151864
value = [44970, 106894]
class = Death
female_count <= 0.001
gini = 0.14
samples = 42828
value = [39590, 3238]
class = NoDeath
female_count <= 0.0
gini = 0.473
samples = 36421
value = [22403, 14018]
class = NoDeath
n_guns_involved <= 0.987
gini = 0.209
samples = 24179
value = [21310, 2869]
class = NoDeath
n_guns_involved <= 1.554
gini = 0.467
samples = 4876
value = [1811, 3065]
class = Death
avg_age <= 27.0
gini = 0.397
samples = 145128
value = [39588, 105540]
class = Death
n_injured <= 0.004
gini = 0.321
samples = 6736
value = [5382, 1354]
class = NoDeath
n_injured <= 1.005
gini = 0.059
samples = 33044
value = [32030, 1014]
class = NoDeath
male_count <= 0.002
gini = 0.351
samples = 9784
value = [7560, 2224]
class = NoDeath
male_count <= 2.002
gini = 0.437
samples = 29457
value = [19956, 9501]
class = NoDeath
female_count <= 0.999
gini = 0.456
samples = 6964
value = [2447, 4517]
class = Death
n_injured <= 0.0
gini = 0.312
samples = 10693
value = [8624, 2069]
class = NoDeath
n_injured <= 0.002
gini = 0.112
samples = 13486
value = [12686, 800]
class = NoDeath
avg_age <= 26.982
gini = 0.459
samples = 4752
value = [1693, 3059]
class = Death
avg_age <= 55.5
gini = 0.092
samples = 124
value = [118, 6]
class = NoDeath
avg_age <= 27.0
gini = 0.452
samples = 79384
value = [27364.0, 52020.0]
class = Death
male_count <= 1.0
gini = 0.303
samples = 65744
value = [12224, 53520]
class = Death
gini = 0.0
samples = 143
value = [0, 143]
class = Death
male_count <= 0.973
gini = 0.037
samples = 30987
value = [30405, 582]
class = NoDeath
male_count <= 0.006
gini = 0.332
samples = 2057
value = [1625, 432]
class = NoDeath
female_count <= 1.054
gini = 0.048
samples = 3369
value = [3286, 83]
class = NoDeath
n_injured <= 1.997
gini = 0.445
samples = 6415
value = [4274, 2141]
class = NoDeath
n_injured <= 1.998
gini = 0.341
samples = 18289
value = [14308, 3981]
class = NoDeath
male_count <= 2.999
gini = 0.5
samples = 11168
value = [5648, 5520]
class = NoDeath
female_count <= 1.004
gini = 0.484
samples = 5965
value = [2447.0, 3518.0]
class = Death
state_California <= 0.0
gini = 0.375
samples = 15332
value = [11506, 3826]
class = NoDeath
female_count <= 0.017
gini = 0.146
samples = 4325
value = [3983, 342]
class = NoDeath
state_California <= 0.0
gini = 0.018
samples = 26662
value = [26422, 240]
class = NoDeath
n_injured <= 1.972
gini = 0.171
samples = 1365
value = [1236, 129]
class = NoDeath
n_injured <= 1.99
gini = 0.492
samples = 692
value = [389.0, 303]
class = NoDeath
female_count <= 0.99
gini = 0.017
samples = 2963
value = [2938, 25]
class = NoDeath
n_injured <= 1.80
gini = 0.245
samples = 406
value = [348, 58]
class = NoDeath
female_count <= 1.0
gini = 0.477
samples = 5019
value = [3049.0, 197
class = NoDeath
female_count <= 1.
gini = 0.215
samples = 1396
value = [1225, 17
class = NoDeath
male_count <= 3.00
gini = 0.499
samples = 10711
value = [5648, 5063
class = NoDeath
male_count <= 1.9
gini = 0.498
samples = 4530
value = [2116, 241
class = Death
female_count <= 1.996
gini = 0.355
samples = 1435
value = [331, 1104]
class = Death

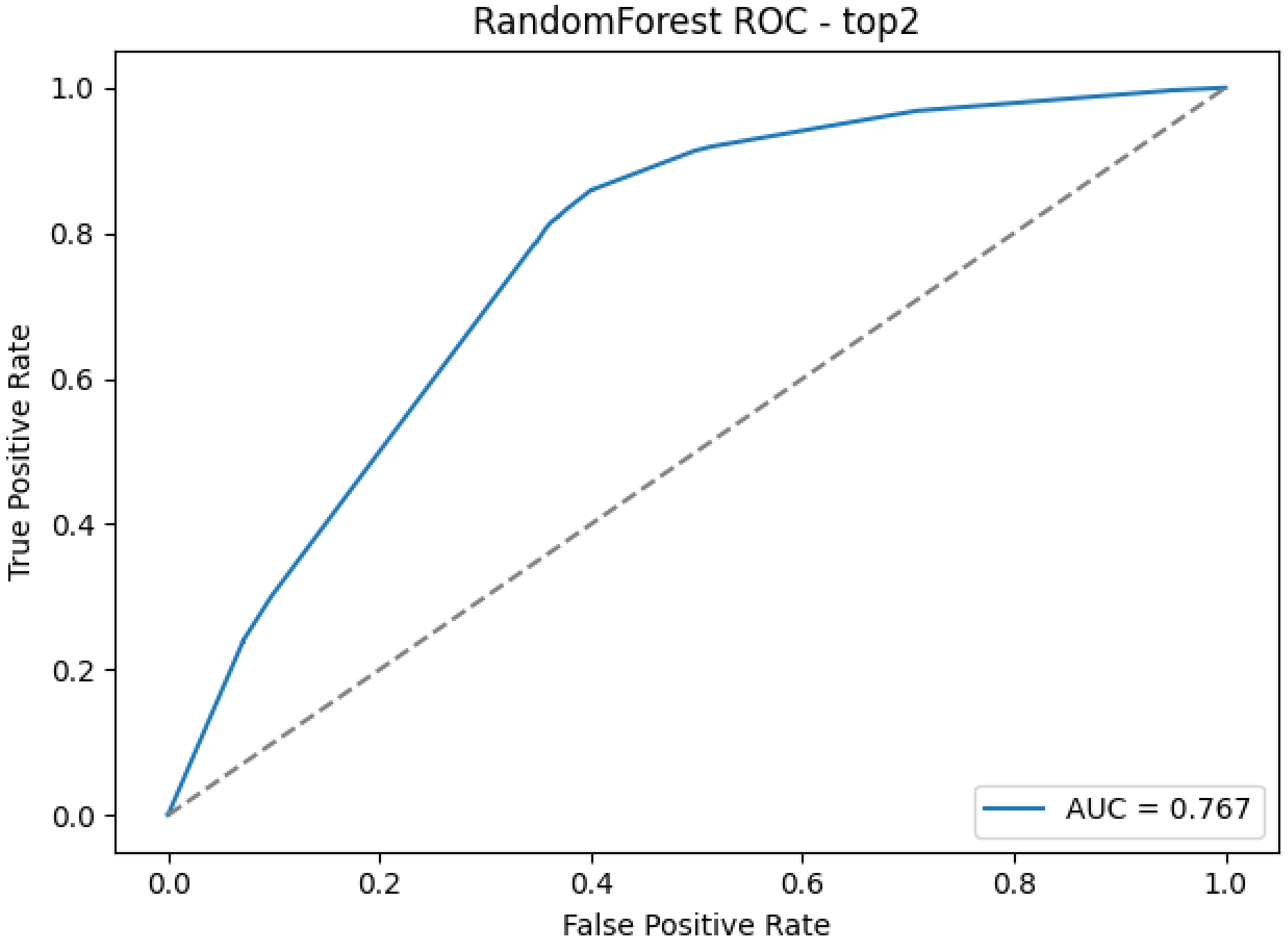

n_injured <= 0.5
gini = 0.5
samples = 215709
value = [107854.5, 107854.5]
class = Death
True
False
male_count <= 0.5
gini = 0.473
samples = 128109
value = [56396.12, 90487.117]
class = Death
male_count <= 1.5
gini = 0.377
samples = 87600
value = [51458.38, 17367.383]
class = NoDeath
female_count <= 0.5
gini = 0.362
samples = 32299
value = [19105.448, 5947.155]
class = NoDeath
n_guns_involved <= 1.5
gini = 0.425
samples = 95810
value = [37290.672, 84539.962]
class = Death
female_count <= 0.5
gini = 0.194
samples = 52913
value = [32952.658, 4035.251]
class = NoDeath
male_count <= 2.5
gini = 0.487
samples = 34687
value = [18505.722, 13332.132]
class = NoDeath
n_guns_involved <= 0.5
gini = 0.224
samples = 28457
value = [17600.33, 2590.752]
class = NoDeath
n_guns_involved <= 1.5
gini = 0.427
samples = 3842
value = [1505.117, 3356.403]
class = Death
avg_age <= 27.1
gini = 0.407
samples = 88148
value = [32861.732, 82772.731]
class = Death
n_guns_involved <= 3.5
gini = 0.408
samples = 7662
value = [4428.94, 1767.231]
class = NoDeath
n_injured <= 1.5
gini = 0.093
samples = 42029
value = [26707.13, 1366.599]
class = NoDeath
male_count <= 0.5
gini = 0.419
samples = 10884
value = [6245.528, 2668.652]
class = NoDeath
female_count <= 0.5
gini = 0.443
samples = 22978
value = [12959.229, 6414.558]
class = NoDeath
avg_age <= 27.292
gini = 0.494
samples = 11709
value = [5546.493, 6917.574]
class = Death
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

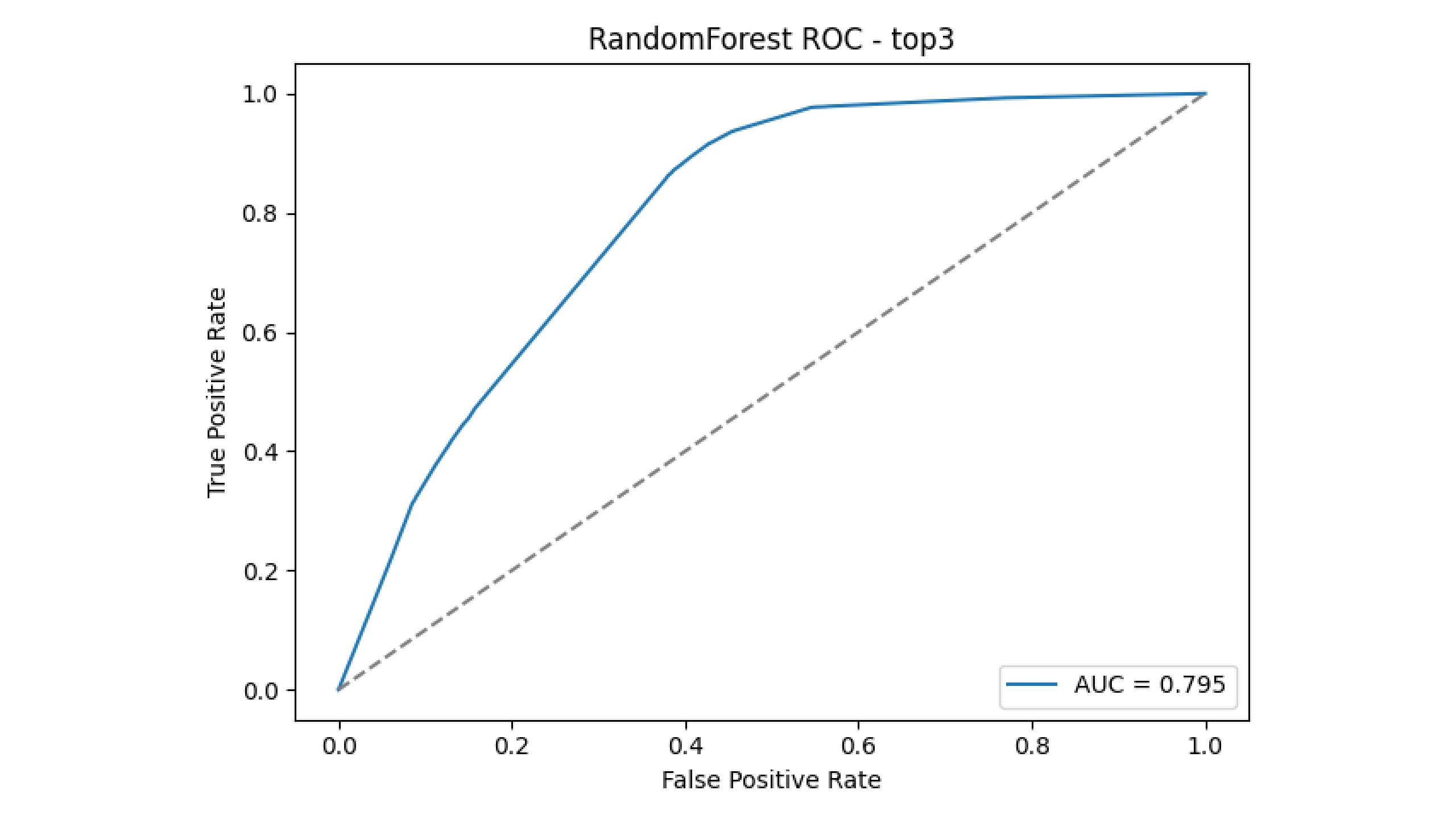

n_injured <= 0.5
gini = 0.5
samples = 215709
value = [107854.5, 107854.5]
class = Death
True
False
male_count <= 1.5
gini = 0.377
samples = 87600
value = [51458.38, 17367.383]
class = NoDeath
guns_involved <= 1.5
gini = 0.425
samples = 95810
= [37290.672, 84539.962]
class = Death
female_count <= 0.5
gini = 0.194
samples = 52913
value = [32952.658, 4035.251]
class = NoDeath
male_count <= 2.5
gini = 0.487
samples = 34687
value = [18505.722, 13332.132]
class = NoDeath
1
3
772.731]
n_guns_involved <= 3.5
gini = 0.408
samples = 7662
value = [4428.94, 1767.231]
class = NoDeath
n_injured <= 1.5
gini = 0.093
samples = 42029
value = [26707.13, 1366.599]
class = NoDeath
male_count <= 0.5
gini = 0.419
samples = 10884
value = [6245.528, 2668.652]
class = NoDeath
female_count <= 0.5
gini = 0.443
samples = 22978
value = [12959.229, 6414.558]
class = NoDeath
avg_age <= 27.292
gini = 0.494
samples = 11709
value = [5546.493, 6917.574]
class = Death

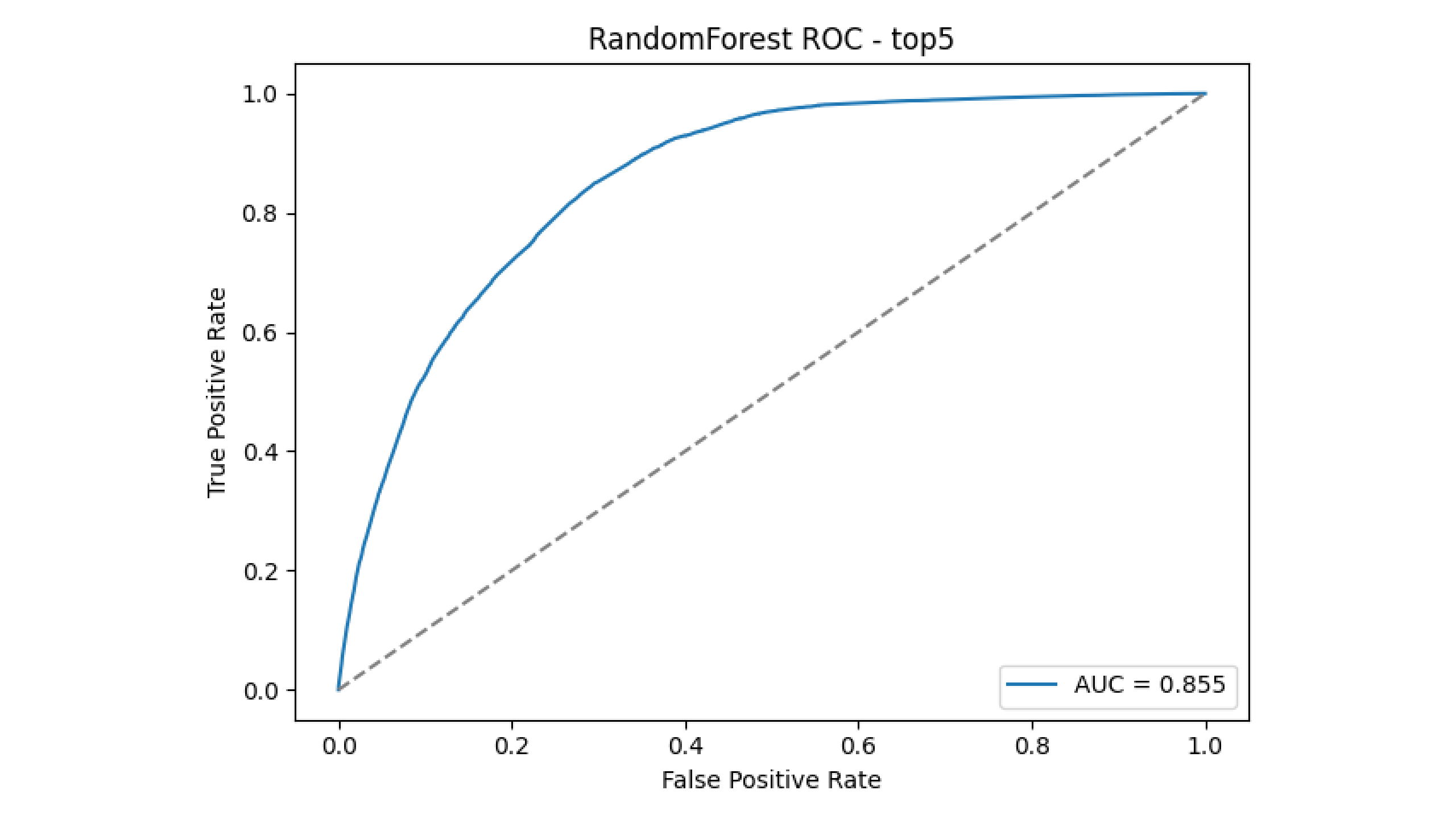

Top 10 Feature Importances
male_count
n_injured
female_count
n_guns_involved
avg_age
state_California
e_Massachusetts
state_Texas
state_Florida
:ate_New Mexico
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Importance

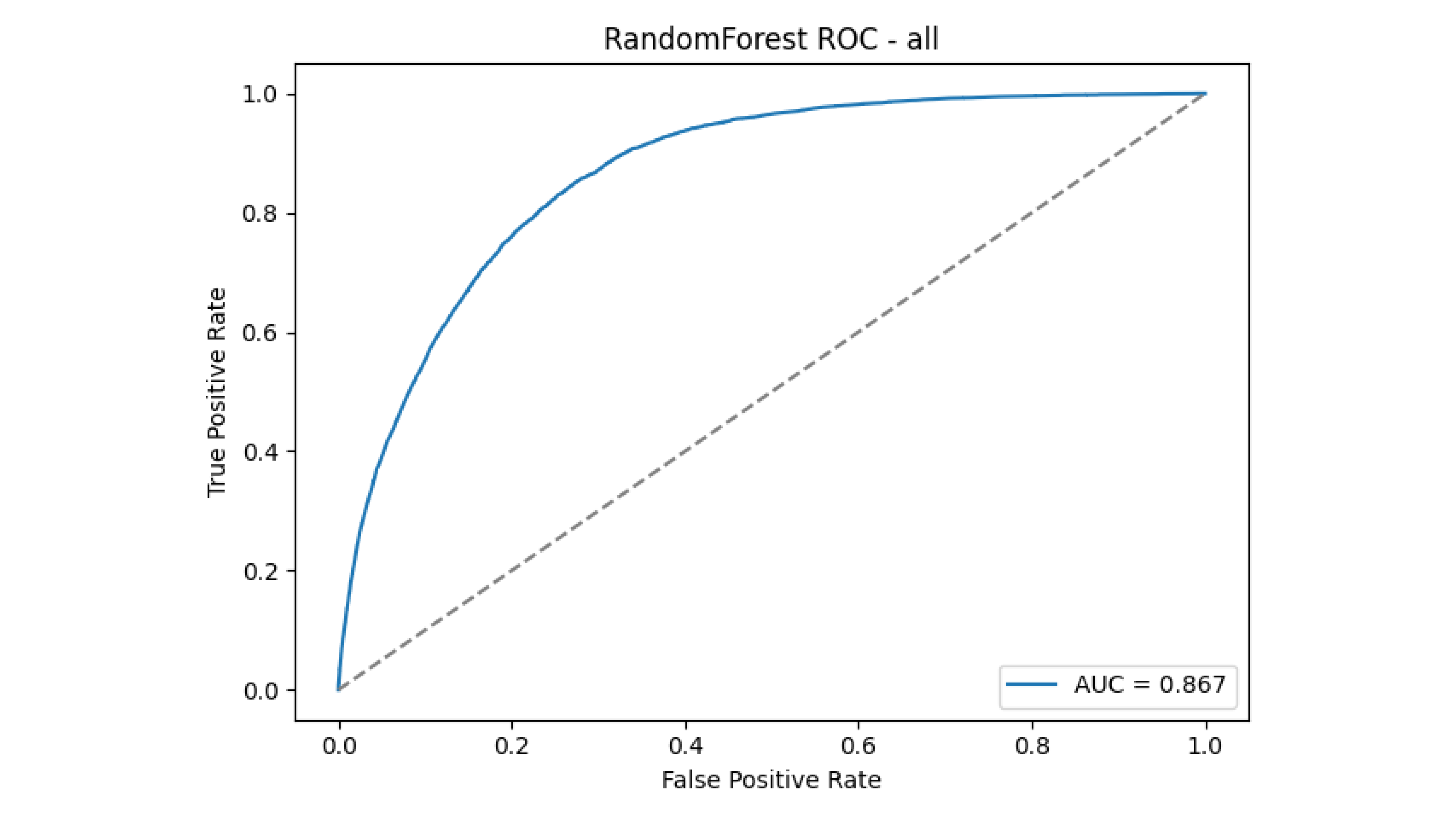

Las losowy

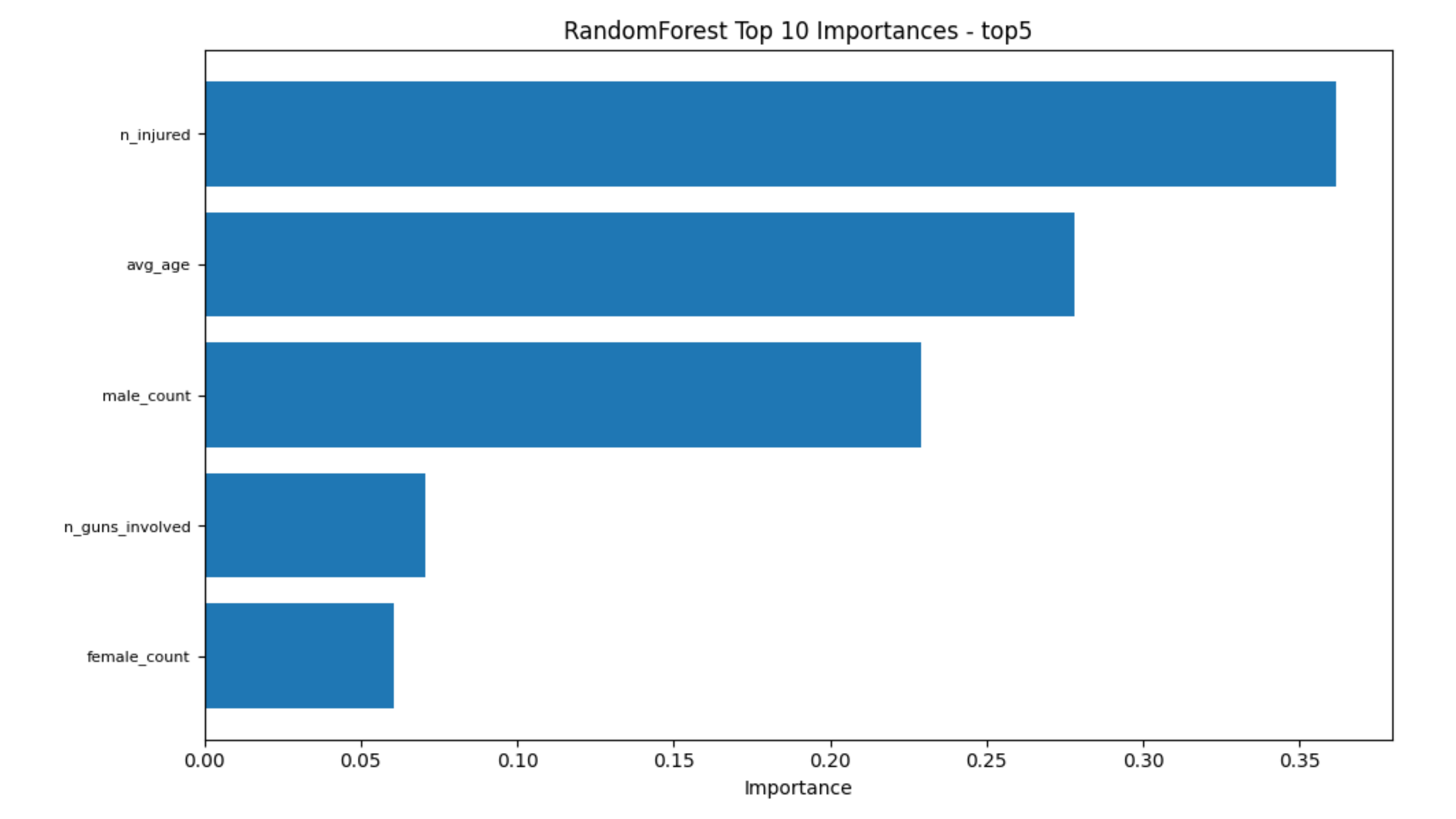

RandomForest ROC - top1
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.629


RandomForest ROC - top2
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.767


RandomForest ROC - top3
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.795


RandomForest ROC - top5
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.855


RandomForest ROC - all
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.867


RandomForest Top 10 Importances - top5
n_injured
avg_age
male_count
n_guns_involved
female_count
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Importance

Las losowy - podsumowanie

Zestaw cech	AUC	Recall	Specificity
Wszystkie	0.867	0.830	0.747
Top 1	0.629	0.440	0.721
Top 2	0.767	0.814	0.638
Top 3	0.795	0.863	0.620
Top 5	0.855	0.834	0.718

Obie metody - podsumowanie

Zestaw cech	Decision Tree (AUC / Recall / Spec)	Random Forest (AUC / Recall / Spec)
Wszystkie	0.851 / 0.772 / 0.768	0.867 / 0.823 / 0.753
Top 1	0.629 / 0.440 / 0.721	0.629 / 0.440 / 0.721
Top 2	0.767 / 0.825 / 0.629	0.767 / 0.814 / 0.639
Top 3	0.794 / 0.865 / 0.617	0.795 / 0.871 / 0.613
Top 5	0.841 / 0.801 / 0.732	0.855 / 0.835 / 0.718

Podsumowanie projektu ✨🎉

- Udało nam się osiągnąć cele eksploracji danych:
  - Czułość (recall) ≥ 80%
  - Swoistość (specificity) ≥ 60%
- Wykonaliśmy 10 eksperymentów z różnymi cechami
  - 5 z drzewami decyzyjnymi
  - 5 z lasami losowymi

Dziękujemy za uwagę

Agnieszka Kulesz, Hania Gibus, Igor Józefowicz