# Model asanbla

**Tim:**

- **Matija Aleksić**

## Zadatak:

Dostupan je deo policijskih izveštaja o saobraćajnim nesrećama u SAD u periodu 1997 - 2002. Na osnovu dostupnih podataka izvršiti procenu brzine vozila u trenutku sudara (kolona speed). Opis svih atributa je dostupan na pratećoj prezentaciji za ovaj zadatak. Zadatak je uspešno urađen ukoliko se na kompletnom testnom skupu podataka dobije makro **f1_mera** (eng. macro f1 score) veća od 0.30. Zadatak se rešava upotrebom ansambla klasifikatora.

## Analiza podataka:

Opis skupa:

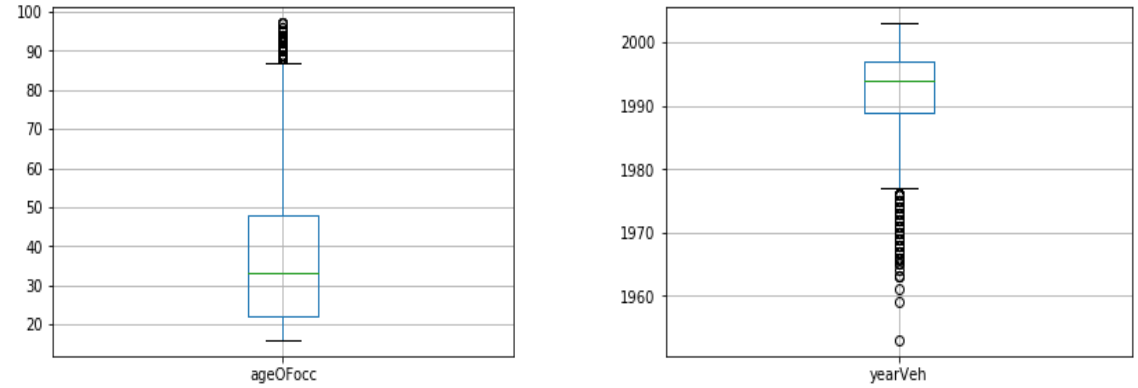
- **speed** – brzina vozila u trenutku sudara (kolona koju je potrebno prediktovati):
  - 1 – 9 km/h
  - 10 – 24
  - 25 – 39
  - 40 – 54
  - 55+
- **weight** – procenjena masa učesnika udesa
- **dead** – da li je učesnik preživeo udes:
  - alive – preživeo
  - dead – nije preživeo
- **airbag** – da li je učesnik imao airbag:
  - none – ne
  - airbag – da
- **seatbelt** – da li je učesnik bio vezan:
  - none – ne
  - belt – da
- **frondal** – da li je u pitanju bio čeoni sudar:
  - 0 – ne
  - 1 – da
- **sex** – pol učesnika:
  - f – ženski
  - m – muški
- **ageOFoc** – starost učesnika
- **yearacc** – godina kada se dogodila nesreća
- **yearVeh** – godina proizvodnje vozila
- **abcat** – da li se aktivirao airbag:
  - unavail – vozilo nije imalo airbag za tog učesnika

    - nodeploy – airbag se nije aktivirao
    - deploy – airbag se aktivirao
- **occRole** – tip učesnika:
    - driver – vozač
    - pass – suvozač
- **deploy** – da li se airbag aktivirao:
    - 0 – airbag nije dostupan za tog učesnika ili se nije aktivirao
    - 1 – airbag se aktivirao
- **injSeverity** – stepen povreda učesnika:
    - 0 – bez povreda
    - 1 – lakše telesne povrede
    - 2 – teže telesne povrede, bez invaliditeta
    - 3 – teže telense povrede, sa invaliditetom
    - 4 – smrt
    - 5 – nepoznato
    - 6 – teške telesne povrede sa smrtnim ishodom (smrt nastupila kasnije)

Od podataka konstruisan je HeatMape, sa kojega je primećeno da postoje neka obeležja koja imaju veliku korelaciju kao sto su: **airbag**, **deploy** i **abcat**. Sva tri obeležja predstavljaju da li se airbag aktivirao prilokom sudara. Na početku sam smatrao da izbacivanjem airbag i deploy pošto podaci postoje u koloni abcat, ali se kasnije pokazalo da izbacivanjem ovih obeležja dovodi do manje tacnog rešenja. Iz svih podataka samo sam uklonio kolonu **dead** zato sto ima ista obelezja u koloni **injSeverity** koja ima podatke o smrtnim slučajevima.

Daljom analizom box plotova podataka uočeni su neki outlieri kod nekih od kolona kao sto us **ageOFocc** i **yearVeh**, ali izbacivanjem istih dolazimo do lošijih rezultata, tako da sam odlučio da ih ostavim u podacima.

<matplotlib.axes._subplots.AxesSubplot at 0x1d18f8399e8> <matplotlib.axes._subplots.AxesSubplot at 0x1d18f8b2828>




Kategorične vrednosti podataka su rešeni sa tehnikom LabelEncodinga, da bi se svi podaci sastojali od numerckih vrednosti.
Podaci su takodje skalirani sa pomoćnom klasom StandardScaler da bi se svi podaci sveli na vrednosti izmedju 1 i 0.

# Odabrani model:

Modeli za klasifikaciju koji su korišceni su sledeci:

- **DecisionTreeClassifier**(max_features = 12, random_state=10)
- **BaggingClassifier**(base_estimator=decision_tree_classifier, max_samples=12, n_estimators=300)
- **RandomForestClassifier**(n_estimators=300, max_features=12)
- **AdaBoostClassifier**(n_estimators=152, learning_rate=1.52, random_state=100)
- **GradientBoostingClassifier**(n_estimators=500, random_state=360, max_features=9)

Kombinacijom ovih modela klasifikacije, napravljen je model asanbla sa modelom za stakovanje:

- **VotingClassifier**(estimators_list, weights=[1, 1.1, 0.8, 0.2, 0.1])

Takodje stavljen je weight atribut koji pomaže pri odabiranju pravog rešenja. Nakon testiranja svakog od modela dosao sam do zakljucka da treba najvise da se slušaju redom: AdaBooster, GradientBooster, RandomForest, BaggigngClassifier, DecisionTree. Takav model je dao najbolji rezultat koji predstavljen **f1_score(average='macro')** merom iznosi 0.32981.