# Ensemble


## Tim [sigma]:

- Marko Vukotić SW-71/2018
- Miloš Popović SW-24/2018


## 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.

## Analiza podataka:

U podacima postoje razne nekonzistentnosti:

- **Weight -** se kreće od 0 do 57000 i sadrži nullove
- **abcat** - Sadrži broj pored ostalih kategoričkih obeležja
- sva ostala polja imaju haotično null vrednosti

Postoje redudantna polja *abcat* označava da li je airbag otvoren, nije otvoren ili ne postoji. Dok *airbag* i *deployed* u kombinaciju teoretski daju identične informacije. Još jedan problem je što značenja kolona ne podudaraju.

Takođe se može zaključiti na osnovu kolone injSeverity. Kao u prošlom slučaju ni ovde se ne poklapaju logične vrednosti.

## Pretprocesiranje podataka:

Pošto nema previše podataka bez vrednosti, izbacili smo sva polje gde nema vrednosti, (malo manje od 10% skupa)

Uradjen je one hot encoding samo nad poljem **abcat** (i nad **injSeverity** ali nije korišćeno u rezultatu)

## Rezultati

Inicijalno su testirani modeli (trenirano nad svim obelezjima iz train skupa bez one hot encodinga) i dobijenisu rezultati (korišćena cross validacija) (macro-f1)

- AdaBoostClassifier - 0.27
- BaggingClassifier - 0.30
- ExtraTreesClassifier - 0.28
- GradientBoostingClassifier - 0.29
- RandomForestClassifier - 0.29

Odlučili smo se da AdaBoostClassifier dodatno finetunujemo

**Prvi pokušaj:**

Model je treniran nad sledećim kolonama (empirijski određeno)
['weight', 'dead', 'frontal', 'yearacc', 'yearVeh', 'abcat0', 'abcat1', 'injSeverity']
'abcat0' i 'abcat1' je one hot encoding od kolone 'abcat'

Parametri:

- n_estimators=100
- learning_rate=0.5
- base_estimator = DecisionTreeClassifier(max_depth=17)
- random_state=999 (Zbog evaluacije)

Rezultat cross validacijom: 0.4125478574481879
Rezultat na platformi: 0.433739407236186

**Drugi pokušaj:**

Null polja nisu izbačena već zamenjana sa najčešćim vrednostima iz te kolone

Parametri AdaBoostClassifier :

- n_estimators=400
- learning_rate=0.125
- base_estimator = DecisionTreeClassifier(max_depth=17)
- random_state=999 (Zbog evaluacije)

Rezultat cross validacijom: 0.40689206116042975
Rezultat na platformi: 0.438147007157506

Kao mera uspešnosti uzima se macro-f1