

Домаћи задатак из предмета Основи машинског учења

Титаник

25. новембар 2023.

1 Опис проблема

15. априла 1912., највећи путнички брод икада, сударио се са сантом леда током свог првог путовања. Када је **Титаник** потонуо погинуло је 1502 од 2224 људи (укључујући и путнике и посаду). Ова трагедија инспирисала је увођење нових безбедносних мера за путничке бродове. Један од разлога за тако велики број преминулих је тај што на броду није било довољно чамаца за спасавање за тако велики број путника и посаде. Иако је за преживљавање овог потопа пука срећа имала одређен фактор, неке групе путника су имале боље шансе за преживљавање од осталих.

Циљ овог домаћег задатка је да се направи модел машинског учења који на основу података путника (пол, старост, социјално-економски статус, итд.) предвиђа да ли је он преживео потоп.

Све потребне датотеке су доступне на **Титаник**. У оквиру овог задатка добијате два скупа података који садрже податке о путницима као што су име, старост, пол, класа итд. Један скуп података је **train.csv**, а други **test.csv**.

Датотека **train.csv** садржи податке о 891 путнику, као и информацију о томе да ли је путник преживео или не, док датотека **Test.csv** садржи сличне податке као и **train.csv** за осталих 418 путника, али не садржи информацију о томе да ли је путник преживео или не. Овај податак треба да буде предвиђање вашег модела који ћете обучити на основу података из **train.csv**.

2 Опис скупа података

Колоне достављеног скупа за обучавање представљају различите атрибуте, чије је значење дато у наставку:

- **PassengerID:** Нумерички податак који представља јединствени идентификациони број путника.
- **Survived:** Податак о томе да ли је путник преживео потоп. Могуће вредности су: 0 - није преживео, 1 - јесте преживео.
- **Pclass:** Нумерички податак о класи путника. 1 - прва класа, 2 - друга класа, 3 - трећа класа.
- **Name:** Текстуални податак са именом и презименом путника.
- **Sex:** Текстуални податак са полом путника. Могуће вредности су: male - мушко и female - женско.
- **Age:** Нумерички податак о старости путника.
- **SibSp:** Нумерички податак о укупном броју рођака путника на броду.
- **Parch:** Нумерички податак о броју деце/родитеља путника на броду.
- **Ticket:** Текстуални податак са бројем карте путника.
- **Fare:** Нумерички податак о цени карте путника.
- **Cabin:** Текстуални податак са називом кабине путника.
- **Embarked:** Текстуални податак о месту укрцавања путника. C - Cherbourg, Q - Queenstown, S - Southampton.

Није нужно корисити све атрибуте из скупа података за обучавање модела, такође водити рачуна о томе да ли постоје сви подаци за сваког путника.

3 Предаја домаћег задатка

Након што обучите модел на основу `train.csv` података, потребно је генерисати предвиђања за путнике из `test.csv` и сачувати их у `submission.csv`. Датотека `submission.csv` треба да садржи две колоне:

- **PassengerID**: нумерички податак који представља ID путника из `test.csv`.
- **Survived**: Предикција вашег модела да ли је путник преживео потоп. Могуће вредности су: 0 - није преживео, 1 - јесте преживео.

Пример изгледа ове датотеке је у оквиру `gender_submission.csv`.

Потребно је предати датотеку `submission.csv` као и комплетан код којим сте модел обучавали и генерисали предвиђања. Датотеке сачувати у оквиру директоријума са називом `Ime_Prezime_Broj_Indeksa`, директоријум запаковати у `zip` датотеку с истим називом и предати на следећу адресу: `ml.fink2023@gmail.com`, са насловом “ОМУ-Титаник”.

Напомена: Датотеке предати искључиво у задатом формату.

Рок за предају је 25. децембар 2023. године у 23:59.