Домаћи задатак из предмета Основи машинског учења

Титаник

25. децембар 2021.

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 безбедоносних 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 задтака је креирање модела машинског учења који ће на основу података путника (пол, старост, социјално-економсски статус итд.) предвидети да ли је он преживео потоп.

У оквиру овог задатка добијате два скупа података који садрже податке о путницима као што су име, старост, пол, класа итд. Један скуп података се зове `train.csv`, а други `test.csv`.

`Train.csv` садржи податке о 891 путнику, као и инфромацију о томе да ли је путник преживео или не. `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 предикција вашег модела који ћете обучити на основу података из `train.csv`.

2 Опис скупа података

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

- **PassengerID**: нумерички податак који представља јединствени Id путника.
- **Survived**: податак о томе да ли је путник преживео потоп. Могуће вредности су: 0 - није преживео и 1 - јесте преживео.
- **Pclass**: нумерички податак о класи путника. 1 - прва класа, 2 - друга класа и 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, укупан број рођака путника на броду.
- **Parch**: нумерички податак, број деце / родитеља путника на броду.
- **Ticket**: текстуални податак, број карте путика.
- **Fare**: нумерички податак, цена карте путника.
- **Cabin**: текстуалн податак, назив кабине путника.
- **Embarked**: тексутални податак са местом укрцавања путника. C - Cherbourg, Q - Queenstown, S - Southampton.

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

3 Предаја домаћег задатка

Након што обучите модел на основу `train.csv` података, потребно је да генеришете предикције за путнике из `test.csv` и сачувате их у `submission.csv`. `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 дат је у оквиру `example_submission.csv`.

Потребно је предати `Submission.csv` као и комплетан код којим се модел обучава и генеришу предикције. Датотеке сачувати у оквиру директоријума са називом `Ime_Prezime_Broj_Indeksa`, директоријум зиповати у зип датотеку са истим називом и то предати на слањем на nradovanovic@kg.ac.rs са насловом ОМУ - Титаник.

Напомена: Датотеке предати искључиво у задатом формату.