**PRÀCTICA 1**

**ENUNCIAT**

**Descripció de la pràctica a realitzar**

L’objectiu de esta activitat serà la creació de un conjunt de dades a partir de els dades. Heu de tenir compte de les consideracions sobre el lloc web elegit, el codi i el conjunt de dades que s'indica més endavant. Se haurà de presentar una memòria en PDF (màxim 20 pàgines) en la que es resolen els següents apartats:

1. Context. Explicar en quin context específic s'han recollit les dades i argumentar per quin lloc web seleccionat és una font pertinent fiable d'aquesta informació. Indica la direcció del lloc web.

L’objectiu d’aquesta activitat pràctica és treballar un cas pràctic orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d'extracció de dades. S’ha de lliurar un sol fitxer amb l'enllaç GitHub ([https://github.com](https://github.com/)) on hi hagi les solucions incloent els noms dels components de l'equip. Podeu utilitzar la Wiki de GitHub per descriure el vostre equip i els diferents arxius del vostre lliuramentsemestres s’han proposat datasets interessants de lliure disposició que es troben disponibles a GitHub (<https://github.com>) i a Kaggle (<https://www.kaggle.com>):

● <https://github.com/rafoelhonrado/foodPriceScraper>

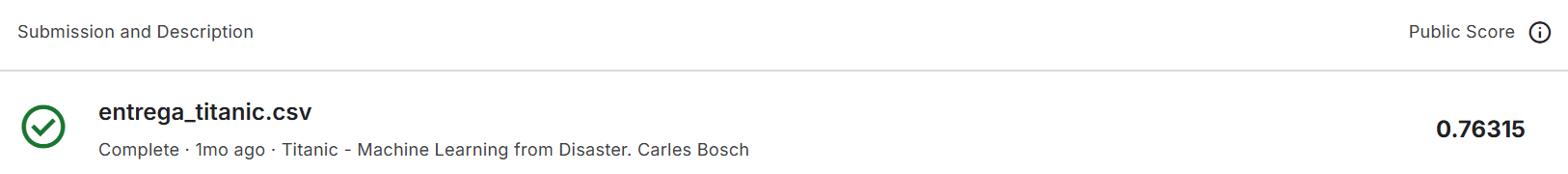
● <https://github.com/tteguayco/Web-scraping>

● Red Wine Quality (<https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>)

● Heart Attack Analysis & Prediction dataset [(https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-predictiondataset](file:///C:\Users\kangu\Desktop\TCVD-PRACTICA2\(https:\www.kaggle.com\datasets\rashikrahmanpritom\heart-attack-analysis-predictiondataset))

● Titanic: Machine Learning from Disaster ([https:/www.kaggle.com/c/titanic](https://www.kaggle.com/c/titanic))

He triat el darrer d’aquests (Titanic) que està disponible a Kaggle (https://www.kaggle.com). que tracta de l'enfonsament del Titanic, ja que l'he trobat força interessant i que és un dels proposats als enunciats de la pràctica de TCVD dels darrers semestres juntament amb els altres datasets anomenats. Aquest exemple, correspon a una competició activa de Kaggle de manera que aquest estiu he aprofitat per a realitzar aquesta pràctica per entrar en aquesta competició. En aquest repte, es demana construir un model predictiu que respongui a la pregunta: "quin tipus de persones tenien més probabilitats de sobreviure?" utilitzant dades de passatgers (és a dir, nom, edat, sexe, classe socioeconòmica, etc.). El resultat de la meva competició a Kaggle ha estat prou bona per ser la primera que faig.



2. Títol. Definir un títol concís y que sigui descriptiu per aquest conjunt de dades.

Titanic - Aprenentatge automàtic del desastre (Titanic - Machine Learning from Disaster)

3. Descripció del conjunt de dades. Desenvolupar una breu descripció del conjunt de dades que s’ha extret. Es necessari que esta descripció mar coherent amb el títol elegit.

En aquesta competició, he tingut accés a aquests dos conjunts de dades, que són similars i que inclouen informació sobre els passatgers, com ara el nom, l'edat, el sexe, la classe socioeconòmica, etc. Un conjunt de dades es titula train.csv i l'altre es titula test.csv

* conjunt d'entrenament (train.csv)
* conjunt de prova (test.csv)

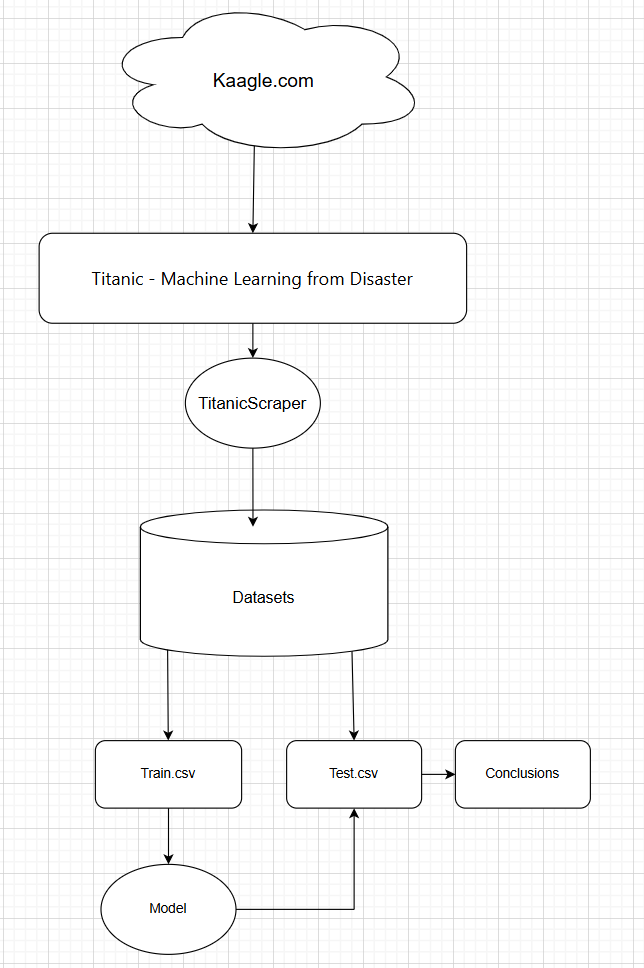
**El conjunt d'entrenament (train)** l’he utilitzat per a crear un model d'aprenentatge automàtic. El model es basa en "característiques" com el gènere i la classe dels passatgers.

Train.csv dóna els detalls d'un subconjunt dels passatgers a bord (891 per ser exactes) i, sobretot, ens diu si van sobreviure o no, també conegut com la "veritat terrestre".

**El conjunt de proves (test)** l’he utilitzat per a per veure el rendiment del model creat amb dades no vistes. Per al conjunt de proves, no proporcionem la veritat terrestre per a cada passatger. El conjunt de proves, test.csv, conté informació similar, però no revela la "veritat terrestre" per a cada passatger.

El que s’he fet és predir uns resultats de la supervivència dels passatgers del conjunt de passatgers del conjunt de proves (test) i per a cada passatger d’aquest conjunt utilitzant el model entrenat amb el conjunt d’entrenament, predir si van sobreviure o no a l'enfonsament del Titanic.

4. Representació gràfica. Dibuixar un esquema o diagrama que reflecteix visualment el dataset i el projecte escollit.



5. Contingut. Explicar els camps que s'inclouen en el dataset i el període de temps al que pertanyen els dades. Les dades estan referides al viatge inaugural del Titanic (Abril, 1914)

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable (eng)** | **Variable**  **(cat)** | **Definició** | **clau** |
| PassengerId | Identificació | Identificació del passatger |  |
| survival | supervivència | Supervivència | 0 = No, 1 = Sí |
| pclass | pclass | Classe de bitllet | 1 = 1r, 2 = 2n, 3 = 3r |
| sex | sexe | Sexe |  |
| Age | Edat | Edat en anys |  |
| sibsp | germans | N, de germans/cònjuges a bord del Titanic |  |
| parch | parch | N, de pares/fills a bord del Titanic |  |
| ticket | bitllet | N. de bitllet |  |
| fare | tarifa | Tarifa de passatgers |  |
| cabin | cabina | Número de cabina |  |
| embarked | embarcat | Port d'embarcament | C = Cherbourg, Q = Queenstown, S = Southampton |

6. Propietari. Presentar al propietari del conjunt de dades. Es necessari incloure cites d'anàlisi anteriors o, en el seu defecte, justificar aquesta recerca amb anàlisi similar. Indicar quins passos s'han seguit per actuar d'acord amb els principis ètics i legals en el context del projecte elegit.

Kaggle ([www.kaggle.com](http://www.kaggle.com)) 🡪 Les dades de Kaggle-Titanic estan sota la llicència Apache 2.0

Aquesta és una llicència permissiva les condicions principals de la qual requereixen la preservació dels avisos de copyright i de llicència. Els col·laboradors proporcionen una concessió expressa de drets de patent. Les obres amb llicència, les modificacions i les obres més grans es poden distribuir sota diferents termes i sense codi font.

En aquest cas concret, donat que les dades formen part d’una competició en la pròpia web. La llicència és de Kaggle i està subjecta a les regles de la competició (Subject to Competition Rules)

Citació: Will Cukierski. (2012). Titanic - Machine Learning from Disaster. Kaggle. <https://kaggle.com/competitions/titanic>

7. Inspiració. Explicar per què pot ser interessant aquest conjunt de dades i quines preguntes es pretén respondre amb elles . És necessari comparar amb els anàlisis anteriors o similars presentats a l'apartat 6.

Aquest conjunt de dades és interessant perquè permet aprofundir quines variables vam influir més en el grau de supervivència del desastre del Titanic i possibilita fer una anàlisi exploratòria de com aquests van influir. Recordem una mica que va passar per posar-nos en context i veure la pregunta que es pretén resoldre amb aquesta anàlisi de dades.

El 15 d'abril de 1912, durant el seu viatge inaugural, el RMS Titanic, àmpliament considerat "insubmergible", es va enfonsar després de xocar amb un iceberg. Malauradament, no hi havia prou bots salvavides per a tothom a bord, la qual cosa va provocar la mort de 1502 de 2224 passatgers i tripulants. Tot i que hi havia algun element de sort implicat en la supervivència, sembla que alguns grups de persones tenien més probabilitats de sobreviure que altres.

Utilitzant el model entrenat amb el conjunt d’entrenament, he predit si van sobreviure o no a l'enfonsament del Titanic els passatgers del conjunt test. La feina realitzada m’ha permès predir aquests resultats i m’ha animat força participar en la competició plantejada per Kaggle.

Per respondre a la pregunta de “**quants passatgers del conjunt test van sobreviure?”,** he fet servir els patrons que he trobat a les dades de train.csv, he fet una predicció de si els altres 418 passatgers a bord (que es troben a test.csv ) van sobreviure o no.

8. Llicència. Seleccioneu una llicència adequada per al conjunt de dades resultant i justificar el motiu de la seva elecció. Exemples de llicències que poden considerar-se:

● Alliberat Sota CC0: Públic Domini llicència.

● Alliberat Sota CC BY-NC-SA 4.0 llicència.

● Alliberat Sota CC BY-SA 4.0 llicència.

● Base de dades alliberat sota llicència de base de dades oberta, continguts individuals sota llicència de contingut de base de dades.

● Altra (especificar quina?).

La llicència és de Kaggle i està subjecta a les regles de la competició (Subject to Competition Rules)

9. Codi. Codi implementat per a l'obtenció del conjunt de dades, preferiblement en Python o, alternativament, en R.

● El codi haurà ubicar-se en la carpeta /font del dipòsit.

● Se han de indicar les llibreries i versions utilitzades. P. ex., en Python poden obtenir-se mitjançant l'ordre: pip3 congelar > requisits.txt

● A la memòria en PDF, se han de comentar els aspectes més rellevants sobre com el codi realitza el procés de recollida de dades, què dificultats presenta el lloc web escollit, y cóm s’ha resolt

10. Conjunt de dades. Publicar el conjunt de dades obtingut en format CSV a Zenodo, incloent una breu descripció del mateix. Obtenir i adjuntar l'enllaç del DOI del dataset (https://doi.org/…). El conjunt de dades també haurà incloure en la carpeta /dataset del dipòsit. Si existeix alguna circumstància que impedeixi publicar obertament el dataset real a Zenodo, haureu de:

a. Comentar esta circumstància y justificar el motiu.

b. Generar un conjunt de dades simulat i publicar-lo a Zenodo, obtenint l'enllaç del DOI.

c. Comunicar al professor el dataset real de forma privada (p. ex., en el repositori privat o en una carpeta de Google Drive privada).

11. Vídeo . **Realitzar un vídeo breu explicatiu de la pràctica (màxim 10 minuts)** , que haurà de contar amb la participació dels dos integrants del grup . En el vídeo haurà de realitzar una presentació del projecte, destacant els punts més rellevants, tant de las respostes a els apartats com a codi utilitzats per extraure els dades. Indica l'enllaç del vídeo (https://drive.google.com/…), que haurà d'ubicar-se al Google Drive de la UOC.

**Format i data de lliurament**

En referència a la entrega de la pràctica, se demana:

a. La memòria de la pràctica, que haurà de ser un únic document PDF , la extensió de la qual no ha de superar les 20 pàgines. En la primera pàgina ha de contenir:

● Los noms de els integrants del grup.

● El enllaç al lloc web elegit.

● El enllaç al dipòsit con el codi de la pràctica.

● El enllaç al conjunt de dades publicat en Zenodo.

● El enllaç al vídeo de presentació de la pràctica.

A continuació, la memòria ha de contenir las respostes a els 11 apartats.

A la taula següent es mostra la meva aportació als diferents apartats, ja que **d’acord amb la tutora i donades les meves circumstàncies personals, se m’ha permès realitzar sol les dues pràctiques d’aquesta assignatura de TCVD.**

|  |  |
| --- | --- |
| **Aportacions** | **Firma** |
| Investigació prèvia | Carles M. Bosch Herrera |
| Redacció de las respostes | Carles M. Bosch Herrera |
| Desenvolupament del codi | Carles M. Bosch Herrera |
| Participació en el vídeo | Carles M. Bosch Herrera |

Aquest document se lliurarà a l’Espai de Entrega PR1 del aula Canvas

b. Un repositori Git amb el codi de la pràctica a la rama “ main” . El dipòsit se crearà en GitHub ( https://github.com/ ). Haurà de ser un dipòsit privat , per lo que haurà d'accedir al professor en el moment de la entrega. El repositori haurà de contenir:

b.1. Un document README.md : Estarà situat a la carpeta arrel i haurà de contenir:

* Els noms de els integrants del grup
* Un apartat on es descriuen els arxius que componen el dipòsit.
* Un apartat on es descriu com utilitzar el codi generat. Heu d'incloure informació sobre els possibles paràmetres que admetin l'script i un o diversos exemples replicables del seu ús.
* El DOI de Zenodo del conjunt de dades generat.

b.2. Un arxiu requisits.txt con las llibreries necessàries per executar el codi.

b.3. Carpeta /font : Haurà de contenir el codi Python o R implementat per a l'obtenció de les dades.

b.4. Carpeta /dataset : Haurà de contenir el conjunt de dades resultant en format CSV.

c. Un vídeo explicatiu , **la durada no ha de superar els 10 minuts** . L' enllaç del mateix s'ha d'indicar a l'apartat 11 del document PDF.

El document PDF (memòria) s’ha de pujar a l'Espai d'Entrega PR1 de l'aula Canvas **abans de les 23:59h CET del dia 16 de abril**. No se acceptaran entregues fora de termini. No podrà modificar cap element de la pràctica passada la data de lliurament (repositori, arxius de Google Drive, etc.).