Anàlisis de la relació entre les emisions de Co2 i l'augment de la temperatura, mitjançant tècniques i tecnologies de Big Data. Informe Inicial

Miquel Freixes Faya

1 OBJECTIUS

L'OBJECTIU principal d'aquest treball és analitzar com ha augmentat la temperatura a la superfície terrestre al llarg dels anys. No obstant això, també es vol comparar aquestes temperatures amb altres dos paràmetres, per veure quina implicació tenen en l'augment. També es pot donar el cas que no es trobi cap relació entre aquests tres volums de dades. Concretament els dos volums de dades que s'analitzaran són:

- La quantitat de Co2 que ha produït cada país al llarg d'un any sencer. [1]
- La petjada ecològica mitjana de cada país al llarg de cada any. [2] Això és una forma de representar l'àrea de vegetació i/o aigua que es necessita per satisfer les necessitats d'una persona durant un any. [3] El conjunt de dades escollit agafarà una mitjana de totes les persones de cada país.

Aquests objectius van enfocats a aprendre a processar diferents Datasets, aconseguint un que estigui destinat a entrenar un algorisme de Machine Learning. Aquest algorisme ha d'acabar sent capaç de predir les tendències en la temperatura a cada país. Per tant, un altre objectiu serà entendre com es pot entrenar un algorisme i extreure els resultats que pugui treure. Com a objectiu addicional es vol fer aquest anàlisis amb diversos algorismes per poder fer una comparativa de l'efectivitat i ràtio d'encert de cada un d'ells.

2 PLANIFICACIÓ

Tal com es mostra en el diagrama de Gantt de la Figura 1, aquest TFG té diverses fases bastant ben diferenciades. Aquestes són:

- Una primera fase de recerca: Aquí cal investigar si s'han fet treballs semblants al del TFG, ja sigui sobre el canvi climàtic, sobre processament de datasets. També cal fer una recerca sobre quins algorismes de Machine Learning caldrà entrenar i com entrenar-los, a causa del gran nombre d'algorismes existents. Cal mirar l'estat de l'art, a més d'aprofitar-lo per aprendre les millors tècniques a aplicar i com aplicar-les als datasets escollits.
- La segona va destinada al processament dels Datasets. Tot i que els datasets escollits tenen les dades netejades, cada un té columnes diferents i fins i tot algun dataset té dades de períodes temporals que altres no tenen. La idea bàsica és seleccionar les columnes i files més rellevants per la investigació que es vol fer i solucionar algun problema en la compatibilitat de les dades. Un exemple de problema a solucionar ve donat pel Dataset de les temperatures. Aquest Dataset et diu les temperatures mitjanes al llarg dels anys de moltes ciutats arreu del món, però no dels països. En canvi els altres dos datasets tracten de les emissions en cada país i no de les ciutats. En acabar es tindrà un únic dataset amb les dades més rellevants en un període de temps determinat, en principi es vol agafar les dades entre l'any 1990 i el 2013.
- La tercera agafa el dataset resultant de la fase anterior i l'utilitza per entrenar un algorisme de Machine Learning. Aquest algorisme s'escollirà al llarg de la primera i segona fase segons els tipus de dades que s'acabin analitzant. No es destinarà el 100% del dataset pel fet que es reservarà una part per comprovar si l'algorisme funciona correctament. Concretament es té pensat destinar un 80% a l'entrenament i un 20% per les comprovacions posteriors.
- La quarta fase acabarà sent la representació dels resultats que s'hauran extret de l'algorisme, ja sigui a partir de gràfics estadístics, diagrames o gràfics lineals. També es farà una reflexió sobre el bon o mal funcionament de l'algorisme i es trauran conclusions.
- Hi ha una última fase que només es farà si hi ha su-

• E-mail de contacte: m.freixes.faya@gmail.com
• Menció realitzada: Tecnologies de la Informació
• Treball tutoritzat per: Jordi Casas Roma (departament)
• Curs 2018/19

ficient temps. Aquesta consistirà a repetir la tercera i la quarta amb diferents algorismes i després comparar els resultats que extreu cada un. Posteriorment es farà una anàlisi de les característiques de cada un com l'eficiència, la dificultat d'entrenar-lo o el percentatge d'error de cada un.

Totes aquestes fases s'aniran documentant per tal d'anar fent l'informe final i com es pot veure en el diagrama es té pensat deixar bastant de temps per poder acabar l'informe i/o tenir marge per si apareixen situacions inesperades o complicacions.

3 METODOLOGIA

A l'hora de fer projectes hi ha un gran nombre de metodologies per escollir, però en l'àmbit de la informàtica les més utilitzades són les Agile. Això és degut a l'increment del nombre de projectes àgils acabats en èxit, respecte dels projectes on s'han aplicat les metodologies tradicionals. [2] Dins de la metodologia àgil hi ha diverses metodologies com Scrum, Kanban, XP... En aquest treball se seguirà la Kanban, a causa del fet que la metodologia Scrum no es podria aprofitar al màxim amb només una persona en el projecte. La principal raó per no escollir Scrum ha estat que molts papers els hauria de fer la mateixa persona, com el de Scrum Master, Product Owner... Algunes de les principals característiques no tindrien sentit. [3]

Kanban es basa en una taula de progrés on es fiquen les tasques per fer, les que s'estan fent i les fetes. Això permetrà un seguiment constant de com avança el projecte. També permet treballar segons prioritats, no per temps, i això s'ajusta molt bé al tipus de projecte que es desenvoluparà. Degut al fet que el projecte està format per tasques que es poden paral·lelitzar, i d'altres que no es poden fer sense una prèvia, apareixeran prioritats al llarg del desenvolupament. [4]

La fase del projecte destinada a l'anàlisi dels datasets es farà utilitzant el software d'Apache Spark a partir de Databricks. S'ha escollit Databricks perquè ofereix un petit clúster amb 6Gb de Ram gratuïtament per estudiants, uns recursos més que suficients per processar el volum de dades desitjat. A part d'això, tot el progrés de cada etapa s'anirà guardant en un Git per tal de no perdre res i poder mantenir un històric dels canvis. En el Git es guardarà els informes, els codis i algorismes necessaris i els datasets amb les dades. Per acabar, els codis es faran en el llenguatge de Python. Aquest llenguatge és molt utilitzat a l'hora de treballar amb algorismes de Machine Learning. A més, s'utilitzaran algorismes de la pàgina web de Scikit Learn, a causa de la gran varietat d'algorismes que ofereix i a la facilitat d'implementar-los. [8]

REFERÈNCIES

[1] https://www.kaggle.com/unitednations/international-greenhouse-gas-emissions

[2] https://www.kaggle.com/footprintnetwork/national-footprint-accounts-2018

[3] "Ecological Footprint", *WWF*. [En línia]. Disponible a: https://wwf.panda.org/knowledge_hub/teacher_resources/webfieldtrips/ecological_balance/eco_footprint/. [Accedit Octubre 4, 2018].

[4] "Sample Research", *The Standish Group*. [En línia]. Disponible a: https://www.standishgroup.com/sample_research [Accedit Octubre 4, 2018].

[5] "Scrum Reference Card now available in English and Spanish", *Agile Methodology*. [En línia]. Disponible a: http://agilemethodology.org/. [Accedit Octubre 4, 2018].

[6] "The Complete Guide to Understanding Agile Testing", *QA Symphony*. [En línia].Disponible a: https://www.qasymphony.com/blog/agile-methodology-guide-agile-testing/. [Accedit Octubre 4, 2018].

[7] A.Géron, *Hands-On Machine Learning with Scikit-Learn & TensorFlow*. Sebastopol, CA: O'Reilly Media, 2017.

[8] "Supervised Learning", *Scikit-Learn*. Disponible a: http://scikit-learn.org/stable/supervised_learning.html#supervised-learning. [Accedit Octubre 5, 2018].

4 ANNEX

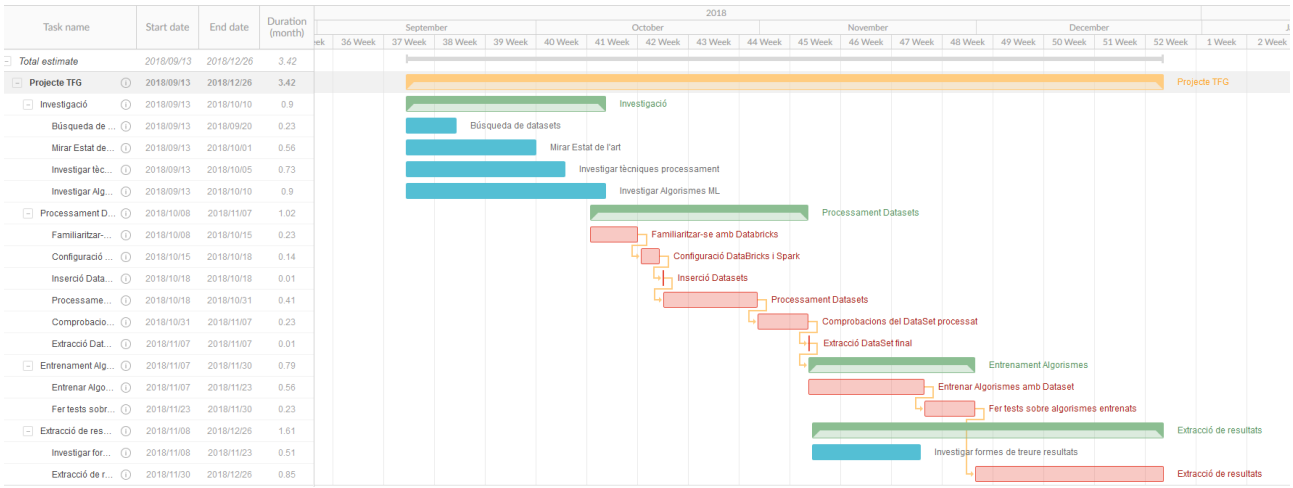

Fig. 1: Diagrama de Gantt