Benchmark Auto Machine Learning

Robert Gimeno Saborit

Resum— Avui en dia el Machine Learning és un camp que està auge, fàcilment es troben exemples en el dia a dia que utilitzen Machine Learning pel seu funcionament. Alguns exemples serien els següents: reconeixement d'imatges, màrqueting personalitzat, automoció autònoma i detecció de fraus entre molts altres. Aquest projecte vol anar un pas més enllà, i centrar-se en l'Auto ML. Aquest és un concepte innovador i amb molta projecció, ja que pretén millorar la forma que actualment es realitza el Machine Learning. Resumidament, l'Auto ML pretén automatitzar molts dels passos que s'han de realitzar a l'hora de fer Machine Learning, aconseguint així una reducció de temps i millora de resultats obtinguts. Avui en dia existeixen diverses tecnologies que fan ús de l'Auto ML, algunes són de codi obert, és a dir, es permet l'ús i modificació de la tecnologia i d'altres són amb llicència, on s'haurà de pagar una quota per fer ús de la tecnologia. En aquest projecte es fa un estudi de tres tecnologies: Una de codi obert, anomenada **Auto-Weka**, i dues amb llicència propietària, anomenades **DataRobot** i **H2O driveless AI**. Aquest article fa una comparativa entre els productes, extraient així uns resultats que permetran classificar els productes segons la seva actuació.

Paraules clau—Machine Learning, Auto Machine Learning, Auto Weka, Data Robot, H20, Comparativa, Neteja de dades, Feature engineering, Selecció d'algoritmes, Ajust de paràmetres, Exactitud.

Abstract— Nowadays Machine Learning is a field that is booming, you can easily find examples in the everyday that use Machine Learning for its operation. Some examples would be: image recognition, personalized marketing, autonomous automotive and fraud detection among many others. This project wants to go a step further, and focus on the Auto ML. This is an innovative concept with a lot of projection, since it aims to improve the way Machine Learning is currently performed. In short, Auto ML aims to automate many of the steps that must be taken when making Machine Learning, thus achieving a reduction in time and an improvement in the results obtained. Today there are several technologies that make use of the Auto ML, some are are open source, that means, the use and modification of the technology is allowed, and others are licensed, where you will have to pay a fee for making use of the technology. This project focuses on the study of three technologies: one open source, called Auto-Weka, and two with proprietary license, called DataRobot and H2O driveless AI. This article makes a comparison between the products, thus extracting results that will allow to classify the products according to their performance.

Index Terms— Machine Learning, Auto Machine Learning, Auto Weka, Data Robot, H20, Benchmark, Data cleaning, Feature engineering, Algorithm selection, hyperparameter tuning, Acurracy.

1 INTRODUCCIÓ

Actualment vivim en l'època on el concepte de Big Data cobra molta importància. Segons la REF [1] s'utilitza aquest terme per referir-nos a la gran quantitat de dades que es generen avui en dia a Internet. A més, aquestes dades són de gran volum, variabilitat i varietat (aquest terme es coneix com les tres V). Aquest fet dificulta la gestió d'aquestes. Fent que els sistemes tradicionals d'emmagatzemat i explotació de dades quedin obsolets.

Tot i així, han sorgit un seguit de noves tecnologies anomenades Business Intelligence amb l'objectiu de poder administrat tot aquest gran volum de dades i poder-ne treure profit.

Mitjançant la recopilació i l'anàlisi d'aquesta gran quantitat de dades, les empreses poden obtenir una valuosa informació per tal de millorar la presa de decisions.

Dintre de les eines de Business Intelligence existeix una gran ramificació: eines per la captura de dades, per la gestió, pel processament, per l'emmagatzematge, pel consum, etc.

Dins de la branca de consum ens trobem d'intel·ligència artificial, que és l'encarregada de fer que les màquines puguin realitzar funcions cognitives, pròpies dels humans. Com podria ser sentir, comprendre, actuar.

Dintre la intel·ligència artificial, existeix un component anomenat Machine Learning. Aquest component té com a funció donar-li l'habilitat a les màquines de poder aprendre per elles mateixes (REF [2]).

Així i tot, abans d'arribar al punt que les màquines puguin realitzar l'aprenentatge, caldrà abans realitzar tota una feina de programació manual i preparació mitjançant gran quantitat de dades. Aquesta feina prèvia de preparació s'anomena la creació d'un model. Normalment, té un seguit de passos i acostuma a ser bastant laboriós. Més endavant s'expliquen totes aquestes tasques amb més detall.

És per aquesta raó que durant els últims anys ha aparegut l'anomenat Auto Machine Learning (Auto ML), amb l'objectiu d'automatitzar moltes de les tasques necessàries per a la creació d'un model. Reduint així el temps de desenvolupament i millorant el funcionament d'aquest.

- *E-mail de contacte: robertgimeno7@gmail.com*
- *Menció realitzada: Tecnologies de la Informació.*
- *Treball tutoritzat per: Victor García Font*
- *Curs 2018/19*

En aquest article segueix la següent estructura: en la primera secció es defineix el projecte i s'explica planificacio/metodologia seguida. En la segona secció s'exposa el background necessari per entendre els conceptes bàsics. En la tercera secció es troba l'estat de l'art, on s'exposen estudis semblants i que aporta aquest nou estudi respecte a els altres. En la quarta secció comença la comparativa fent ús de la documentació aportada per cada un dels productes. En la cinquena secció es troba la part pràctica de la comparativa. En la sisena secció hi ha les conclusions. Finalment, en la setena secció es troben els agraïments del treball.

1.1 Objectius

El projecte té un seguit de 5 objectius bàsics:

- Adquirir coneixements sobre l'Auto Machine Learning. Entendre què és, com funciona i perquè és útil.
- Prova de 3 tecnologies que utilitzin Auto Machine learning. Tant el rendiment, la funcionalitat, característiques rellevants, etc.
- Puntuar cada producte depenent de la seva actuació seguint unes mètriques objectives.
- Comparació de les tecnologies d'Auto Machine Learning en les diverses propietats.
- Extreure conclusions sobre si realment és notable la diferència entre tecnologies de codi obert i tecnologies amb llicència comercial.

1.2 Metodologia

Per tal de complir els objectius descrits anteriorment, el projecte segueix una metodologia concreta. El mètode utilitzat ha estat el de comparació normativa (REF [4]). Aquest es basa a definir unes propietats més destacades i veure cada un dels productes quina puntuació obté a cada un de les propietats.
En aquesta comparativa s'han inclòs tres productes diferents que fan ús d'Auto ML: Auto Weka, DataRobot i H2O Driveless AI.
A l'hora de fer la comparació s'ha utilitzat dues fonts. En primer lloc, s'ha fet ús de la documentació oficial aportada per cada una de les companyies dels productes. En segon lloc, s'ha dut a terme una Prova de Concepte (PoC). Aquesta PoC s'ha basat a escollir diferents datasets de característiques variades sobre diversos àmbits. S'ha escollit un total de 5 datasets extrets de REF [3].
Un cop recollides totes les dades, s'han analitzat i extret uns resultats per tal d'arribar a unes certes conclusions.

2.3 Planificació

L'inici del projecte va ser el 9 de febrer de 2019 i la data de finalització ha estat el 30 de juny de 2019, un total de quasi 5 mesos.
Per tal de poder dur a terme el projecte dins dels terminis establerts ha fet falta una prèvia planificació exhaustiva. L'eina utilitzada per fer la planificació ha estat Microsoft Project 2013 (REF [5]).
S'han planificat un total de 4 fases amb un seguit de tasques específiques per tal de poder controlar el desenvolupament del projecte:

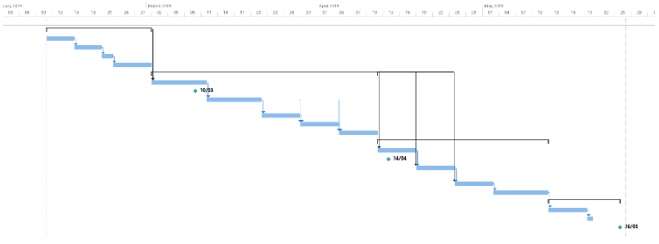
Fase 1 – Inici:
- Planificació
- Context
- Descripció
- Objectius
- Metodologia

Fase 2 – Documentació:
- Machine Learning
- Auto ML
- Auto-Weka
- DataRobot
- H20 Driveless IA

Fase 3 – Implementació:
- PoC Auto-Weka
- PoC DataRobot
- PoC H20 Driveless IA
- Anàlisis Resultats

Fase 4 – Tancament:
- Conclusions
- Agraïments

A més, s'han tingut en compte les entregues parcials que s'han de realitzar periòdicament. També s'ha planificat acabar el projecte amb antelació a la data d'entrega, per si es produeix qualsevol imprevist o s'han de fer millora hi hagi marge de temps.



Il·lustració 1. Diagrama de Gantt

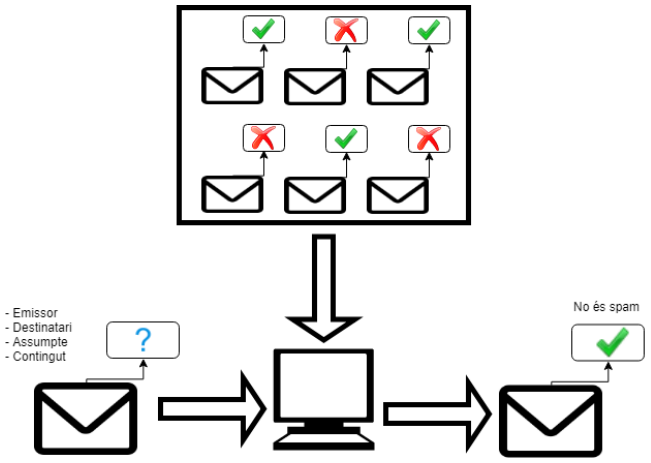
Per informació més detallada sobre la planificació es pot consultar al document adjunt anomenat *planificació.mpp*.

2 Background

2.1 Machine Learning

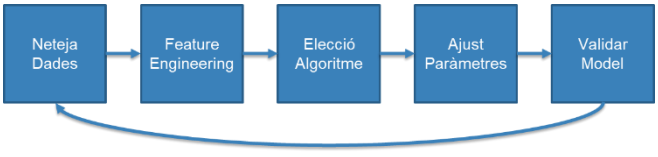
El Machine Learning és l'encarregat de fer que les màquines aprenguin a realitzar una tasca sense que hagin sigut explícitament programades per realitzar-la. Per poder aconseguir-ho la màquina ha de rebre un entrenament previ, anomenat creació del model. Tot aquest entrenament previ no és una tasca trivial, cal dedicar una gran quantitat de temps i coneixements per poder aconseguir un model adequat.
Un exemple fàcil per entendre millor el concepte seria el següent, extret de REF [2]: El nostre objectiu és filtrar si un correu és spam o no. Machine Learning el que farà és

preparar la màquina prèviament mitjançant una gran quantitat de correus i analitzant les seves característiques. D'aquesta manera la maquina tindrà un model. Finalment, segons tota aquesta experiència, quan la màquina rebi un correu, sabrà si és spam o no.



Il·lustració 2. Diagrama detecció correu spam

Segons REF [6] la creació d'un model es divideix en diverses etapes:



Il·lustració 3. Seqüencia per la creació d'un model

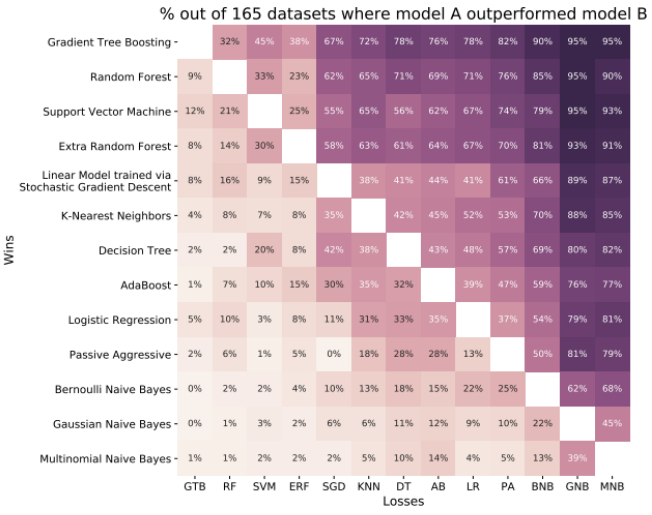
En primer lloc s'han de **Netejar les Dades** amb les quals entrenarem la nostra màquina. Normalment les dades sense processar solen estar de forma no estructurada; amb dades corruptes, duplicades, inexactes, incompletes. Si utilitzéssim aquestes dades sense netejar-les, ens produirien una quantitat d'errors immensa. Aleshores, el primer pas per poder tractar-les serà arreglar totes les dades incoherents de forma que obtinguem un dataset consistent i organitzat.

Seguidament ve l'etapa anomenada **Feature Engineering**. Aquest és el procés de modificar les dades o crear-ne de noves per fer que els algoritmes d'aprenentatge automàtic funcionin millor. És un pas crucial en el procés d'aprenentatge de màquines, perquè l'atribut correcte pot facilitar la dificultat de modelatge, i per tant, millorar la qualitat del model.

La tercera etapa és **l'Elecció de l'Algoritme**. Dins del Machine Learning existeixen una gran quantitat d'algoritmes que permeten construir un model. Segons REF [7] aquest algoritmes se solen ordenar en tres grups:

- **Supervisat**: El dataset d'entrenament conté les dades d'entrada i les respostes correctes, en funció d'aquestes l'algorisme generalitza per respondre correctament. Per exemple ho serien els algoritmes de regressió (calcular un valor numèric) i classificació (etiquetar a una classe).
- **No supervisat**: En aquest cas les respostes correctes no són proporcionades, sinó que l'algorisme tracta d'identificar similituds entre les dades de manera que tinguin alguna cosa en comú per poder classificar-les. Per exemple ho serien els algoritmes de clustering.
- **Reforç**: Quan l'algoritme obté un resultat incorrecte, és notificat, però no se li diu com corregir-lo. Ha d'explorar i provar diferents possibilitats fins que trobi la resposta correcta.

Cada un dels algoritmes funciona diferent i serveixen per casos particulars. Com bé sabem poden existir infinitat de tipus de dataset. Depenent de les característiques d'aquestes dades i l'objectiu que es vol aconseguir, un algoritme podrà funcionar millor que un altre.



*Il·lustració 4. Imatge extreta de: **arXiv:1708.05070** [q-bio.QM] "Comparativa rendiment dels algoritmes"*

Com es pot veure en la il·lustració 4, d'un conjunt de 165 datasets, es van construir per cada un d'ells un model amb els principals algoritmes. En la imatge es mostra una comparativa sobre quin model obtenia millor resultat que l'altre.

Es pot dir que hi ha algoritmes que en general solen funcionar millor que d'altres però tot i així en casos específics no és així, per tant, no es pot obviar cap algoritme. Escollir l'algoritme adequat doncs, no serà senzill.

El següent pas és l'**Ajust de Paràmetres**. Tots els algoritmes tenen certs paràmetres que poden ser modificats, afectant així al rendiment del model. Els paràmetres que venen per defecte no sempre fan que l'algoritme obtingui el seu màxim rendiment. Per tant, ajustar els paràmetres farà que puguem obtenir una millor actuació del model.

Un cop realitats tots els passos anteriors caldrà **Validar el Model** i veure realment si funciona realment. Normalment, es divideix el dataset en dues seccions, una més gran per fer l'entrament i una de més petita per realitzar la validació. Per tal de realitzar la comprovació, majoritariament s'utilitza el concepte d'exactitud. Existeixen

d'altres, com la ràtio de falsos positius, falsos negatius, pèrdua logarítmica, etc. En aquest estudi s'utilitzarà l'exactitud. Aquesta indica quants casos el nostre model obten el resultat esperat. És un valor que va de 0% a 100%. El valor indicarà el percentatge de casos on el resultat obtingut és el correcte.
En cas que l'exactitud no sigui suficient o es vulgui incrementar-la caldrà retornar a un dels passos anteriors i tornar a repetir el procés explicat.

2.2 Auto Machine Learning

La funció doncs de l'auto ML és automatitzar totes les tasques descrites anteriorment. Aconseguint així una gran reducció del temps en la realització del procés de Machine Learning. A més, assegura una qualitat elevada del model construit, ja que darrere de les tecnologies que utilitzen Auto ML solen treballar data scientists amb gran coneixement en l'àmbit de Machine Learning.

3 Estat de l'art

En el mercat actual existeixen gran varietat d'eines que utilitzen Auto ML. En l'apartat d'open source trobem com a més destacades auto-sklearn, Auto Weka, TPLOT i BigML. En l'apartat de eines comercials trobem DataRobot, H20 Driveless AI, Google AutoML entre d'altres.
A continuació es farà un estudi previ de les tres tecnologies seleccionades (Auto Weka, Data Robot, H2O driveless IA) per fer una posterior comparativa practica entre elles.
Actualment, no hi ha cap article acadèmic on es faci una comparativa entre aquestes 3 eines.
Tot i així, hi ha diverses comparatives entre diferents eines d'Auto ML. És un exemple el document REF [8] on es comparen TPOT, auto-ml, auto-sklearn i H2O. Aquesta comparativa és realment de gran qualitat.
En una pàgina web REF [9], es fa una altra comparativa entre DataRobot i H2O però de forma menys científica i menys rigorositat.
A diferència dels altres estudis, en aquest s'ha volgut estudiar altres àmbits i característiques dels productes, no tan sols el seu resultat final.

4 Comparativa

Les propietats escollides per fer la comparativa han estat les següents. A la taula també esta explicat el criteri que s'ha seguit per tal de fer la puntuació:

Propietat	Descripció / Puntuació
Exactitud	Percentatge d'encert que té el model construït. És a dir, qualitat obtinguda en executar-se les proves. La puntuació és la mitjana de l'exactitud obtinguda en cada una de les proves. S'utilitza la mitjana perquè és un valor representatiu de tot el conjunt de proves.
Transparència	Grau de secretisme que té el producte en mostrar com funciona o com obté els resultats. Nul – 0%: El producte no mostra cap informació segons com funciona o com obté els resultats. Normal – 50%: El producte mostra certa informació segons com funciona o com obté els resultats. Total – 100%: El producte té transparència total.
Usabilitat	Facilitat que té el producte a l'hora de ser utilitzat. Nul – 0%: El producte no té cap UI. Normal – 50%: El producte té un UI senzilla. Total – 100: El producte té una UI completa.
Anàlisi	Facilitats que aporta el producte a l'hora d'obtenir anàlisis amb els resultats Nul – 0%: El producte no ofereix cap tipus d'anàlisi. Normal – 50% El producte ofereix algun tipus d'anàlisi senzill. Total – 100% El producte ofereix gran quantitat d'eines i dades per poder analitzar els resultats.
Grau d'automatització	Passos del Machine Learning que el producte és capaç d'automatitzar. Normal – 50%: Automatitza alguns dels passos de Machine Learning. Total – 100%: Automatitza tots els passos de Machine Learning.
Llibreria d'algoritmes	Ventall d'algoritmes que fa ús cada un de les tecnologies Normal – 50% : El producte utilitza algoritmes d'una sola llibreria d'algoritmes. Total – 100%: El producte utilitza algoritmes de diverses llibreries d'algoritmes.
Desplegament	Possibilitat que ofereix el producte perquè pugui ser utilitzat en local o en el cloud. Normal – 50%: En local. Total – 100%: En local i en el cloud.
Explotació del model	Un cop obtingut el model, les opcions que dóna el producte per fer ús d'aquest. Normal – 50%: El producte genera una única opció per fer ús del model. Total – 100%: El producte genera diverses opcions per fer ús del model .
Preu	Preu del producte.

Nul – 0%: Cal llicència de pagament.
Total – 100%: És gratis.

Taula 1. Propietats estudiades

Prèviament a la implementació s'ha realitzat una tasca de documentació on s'han vist els detalls més destacats de cada un dels productes.

4.1 Auto Weka

Segons REF [10] Weka és una plataforma de Machine Learning de codi obert àmpliament utilitzada. Aquesta té un mòdul anomenat Auto Weka que incorpora Auto ML amb l'objectiu d'ajudar als usuaris a construir els seus models.
Aquest mòdul està totalment integrat amb la plataforma Weka, fent que la seva instal·lació sigui senzilla si disposem de la plataforma Weka ja instal·lada.
Auto-Weka no fa Neteja de Dades ni Feature Engineering. Pel que fa a la selecció d'algoritmes, Auto Weka fa una cerca automàtica dels algorismes d'aprenentatge propis de Weka i els seus respectius paràmetres per maximitzar el rendiment del model.
El desplegament és en local, és a dir, es fa en la màquina de l'usuari. Permet l'execució en paral·lel per a un millor rendiment. A més, l'usuari pot configurar la quantitat de recursos que vol dedicar a l'execució del programa. Per tal de construir un model adequat, es recomanen vàries hores d'execució.
Un cop acabada l'execució, la sortida d'Auto Weka conté el codi Java per tal d'executar el model creat. També pots desar el model dins la plataforma (REF [11]).
El codi font d'Auto Weka està publicat a GitHub (REF [12]).

4.2 DataRobot

Segons REF [13] Data Robot és una plataforma d'Auto ML amb llicència comercial. Té una interfície molt gràfica i és realment senzilla d'utilitzar. L'usuari només ha d'introduir el dataset, seleccionar la variable que vol predir i el programari farà la resta.
DataRobot utilitza les últimes biblioteques de Machine Learning de codi obert, incloent-hi scikit-learn, H2O, TensorFlow, Vowpal Wabbit, Spark ML i XGBoost.
DataRobot, a diferència de Auto Weka, automatitza tots els passos del Machine Learning.
El desplegament pot ser tant en local com en el cloud. Si es fa en local caldrà disposar de màquines d'altes prestacions (Sistema Desktop/Server Intel Xeon E5-2699 v3, 36 Cores amb un total de 256 GB de RAM i 4 TB de memòria SSD com a mínim). Si el desplegament es fa en el cloud, està hostejat per AWS.
Un cop acabada l'execució, DataRobot publica un REST API endopoint on l'usuari podrà connectar-se per fer ús del model. També permet obtenir un executable del model.

4.3 H2O Driveless AI

Segons REF [14] aquest producte és més semblant a DataRobot que Auto Weka. També té llicència comercial. La interfície és gràfica i amigable per l'usuari. Un cop introduït el dataset permet a l'usuari visualitzar les dades de forma clara, seguidament l'usuari ha de seleccionar la variable a predir i esperar que s'executi l'experiment.
H2O utilitza vàries biblioteques d'algoritmes com XGBoost, GLM, TensorFlow, RuleFit i FTRL.
El desplegament pot ser tant en local com en el cloud. El producte està dissenyat per tal que s'executi en targetes gràfiques (tot i això, amb CPUs també pot funcionar). Per tant, qualsevol màquina podrà executar el producte.
Un cop acabada l'execució, es pot generar tant en codi Python com en Java el model escollit.

5 Implementació

Per poder fer ús de les tecnologies s'han seguit diversos procediments.
En el cas d'Auto Weka, s'ha descarregat el producte directament des de la pàgina oficial de la companyia.
En el cas de Data Robot s'ha demanat a la companyia fer ús d'un període de prova de 15 dies que l'empresa ofereix. En aquest demo l'empresa facilita una llicència de prova vàlida per un cert temps on et pots connectar a un dels seus servidors i fer ús del producte.
Per últim, en el cas de H2O driveless IA, s'ha volgut fer el mateix que amb DataRobot. Des de la pàgina web oficial ells ofereixen provar una demo del seu producte durant 21 dies. El problema ha estat que després de completar tots els formularis. Finalment, la companyia no ha lliurat cap llicència de prova. Per tant, no s'hn pogut realitzar les proves pràctiques per aquest producte.
El datasets utilitzats han estat els següents:

- El primer relacionat amb maquinària defectuosa. On es recullen característiques de diferents plaques d'acer. Es predirà si la placa d'acer serà defectuosa o no.
- El segon relacionat amb frau d'assegurances. On es té informació sobre les reclamacions que es fan a les assegurances. Es predirà si aquestes reclamacions són fraudulentes o no.
- El tercer relacionat amb el manteniment de maquinària. On es recullen les dades de les màquines. Es predirà si la maquinària es trencarà o no.
- El quart relacionat amb el blanqueig de capital. On es recullen transaccions monetàries. Es predirà si les transaccions estan relacionades amb el blanqueig de diners o no.
- El cinquè relacionat sobre el reingrés hospitalari. On es recullen dades sobre els clients que han passat per l'hospital. Es predirà si un client recaurà i tornarà l'hospital o no.

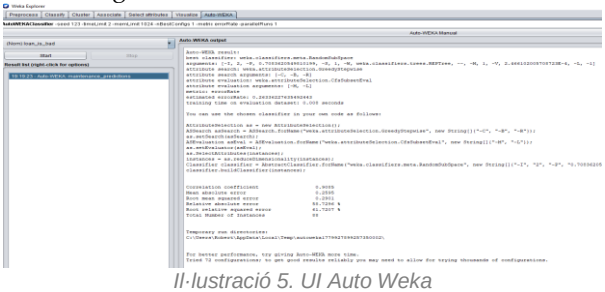
5.1 PoC Auto Weka

Auto Weka té una interfície molt simple però funcional. Per fer ús del producte només cal fer un de dues pestanyes, la primera anomenada pre-process. En aquest apartat s'introdueix el dataset. Un cop introduït el dataset, dona informació sobre aquest: Atributs, visualització de les dades, rang de valors, etc.

Seguidament, cal anar al mòdul d'Auto Weka. Allà simplement seleccionarem la variable que volem predir i les mètriques de l'experiment (duració de l'experiment, recursos dedicats, paral·lelisme entre d'altres).

Finalment es llança l'experiment i s'espera 15 minuts que finalitzi. S'ha escollit aquest temps perquè amb 15 minuts eren suficients per poder construir un model adequat segons les dimensions dels datasets. S'ha comprovat que fent l'experiment durant 1h, el resultat obtingut era el mateix.

Un cop finalitzar l'experiment es mostren els resultats obtinguts. En primer lloc es mostra l'algoritme utilitzat amb els seus paràmetres corresponents. A continuació, indica el codi en Java necessari per crear i posteriorment fer ús del model. Per últim, hi ha l'apartat on es mostren certs indicadors que recullen informació sobre el millor model obtingut.



Il·lustració 5. UI Auto Weka

Els resultats obtinguts per cada un dels datasets han estat els següents, per a cada entrada hi ha el dataset amb el qual s'ha fet la prova, la exactitud màxima obtinguda i l'algoritme amb el qual s'ha obtingut aquesta exactitud:

Dataset	Exactitud	Algoritme escollit
Maquinària defectuosa	92,14%	weka.classifiers.trees.RandomForest
Frau d'assegurances de cotxe	84,49%	weka.classifiers.trees.RandomForest
Manteniment maquinària	90,85%%	weka.classifiers.meta.RandomSubSpace
Blanqueig de capital	95,78%	weka.classifiers.trees.RandomForest
Reingrés hospitalari	78,25%	weka.classifiers.functions.Logistic

Taula 2. Resultats obtinguts Auto Weka

5.2 PoC DataRobot

DataRobot té una interfície més detallada i cridanera. Aquesta interfície és molt user-friendly fent que intuïtivament l'usuari pugui fer ús de l'eina. Igual que Auto Weka, per poder fer ús del producte s'ha de fer unes simples passes.

Cal introduir el dataset. Després de llegir les dades, permet obtenir informació clara i de manera molt gràfica dels diferents atributs del dataset. A continuació, se selecciona la variable que es vol predir i comença l'experiment.

Un cop finalitat l'experiment es mostren els resultats. A diferència de Auto Weka, els resultats de l'experiment són molt més detallats. DataRobot llista tots els models construïts, ordenats segons el seu rendiment.

Per cada model es pot accedir als seus detalls, allà hi ha gran quantitat d'informació analítica útil, mostrada de forma molt visual. Mostra indicadors com ROC curves, life charts, matrius de confusió, etc.

A diferència d'Auto Weka, DataRobot et mostra l'algoritme utilitzat per construir el model però no diu l'ajust de paràmetres realitzat.

Il·lustració 6. UI DataRobot

Els resultats obtinguts per cada un dels datasets han estat els següents, per a cada entrada hi ha el dataset amb el qual s'ha fet la prova, la exactitud màxima obtinguda i l'algoritme amb el qual s'ha obtingut aquesta exactitud:

Dataset	Exactitud	Algoritme escollit
Maquinària defectuosa	92,17%	eXtreme Gradient Boosted Trees Classifier with Early Stopping
Frau d'assegurances de cotxe	90,48%	eXtreme Gradient Boosted Trees Classifier with Early Stopping
Manteniment maquinària	98,90%	Elastic-Net Classifier (L2 / Binomial Deviance)
Blanqueig de capital	94,77%	RandomForest Classifier (Gini)
Reingrés hospitalari	81,78%	AVG Blender

Taula 3. Resultats obtinguts DataRobot

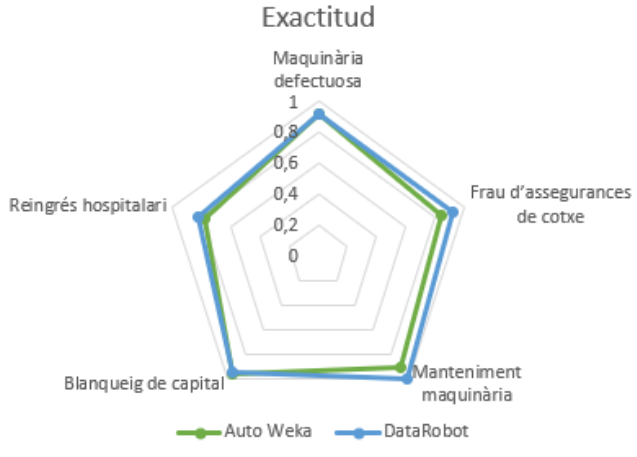
5.3 PoC H2O Driveless AI

Tot i que no s'ha pogut provar el producte H2O driveless AI, la mateixa empresa té publicats diferents vídeos mostrant el funcionament del seu producte REF [15].
S'ha pogut comprovar que té una interfície amb una funcionalitat i estrucurada de manera semblant a DataRobot. Per tant, els procediments a l'hora d'executar els experiments són els mateixos.
De forma similar a DataRobot, H2O Driveless AI mostra gran quantitat d'indicadors analítics per evaluar i entendre el model construït.
En contrapartida, a diferència de DataRobot, H2O Driveless AI és més transparents pel que fa a l'explicació de l'obtenció del model. En aquest cas, H2O et mostra en tot moment els procediments i valors escollits amb els quals s'han aconseguit els resultats.

Il·lustració 7. UI H20 Driveless AI

6 ANÀLISIS DEL RESULTAT

En primer lloc es mostra un gràfic on es pot veure l'exactitud obtinguda per a cada producte en les diferents proves realitzades. Aquesta es la propietat de la comparativa on s'ha dedicat més esforç i té més importància, es per aquesta raó que la anàlisis d'aquesta propietat serà més exhaustiu. Recordar que aquesta propietat d'exactitud, H2O driveless AI queda exclòs.



Il·lustració 8. Comparació exactitud

Com es pot veure en la il·lustació 8, ambdós tecnologies obtenen uns resultats similars. Tot i així, en la majoria dels casos, DataRobot obté una petita millora de resultat.
A continuació hi ha una taula on es recullen les puntuacions obtingudes de cada producte per a cada una de les propietats.

Propietat	Auto Weka	DataRobot	H2O driveless AI
Exactitud	Puntuació: 88,30%	Puntuació: 91,62%	N/A
Transparència	Puntuació: 100%	Puntuació: 50%	Puntuació: 100%
Usabilitat	Puntuació: 50%	Puntuació: 100%	Puntuació: 100%
Anàlisi	Puntuació: 50%	Puntuació: 100%	Puntuació: 100%
Grau d'automatització	Puntuació: 50%	Puntuació: 100%	Puntuació: 100%
Llibreria d'algoritmes	Puntuació: 50%	Puntuació: 100%	Puntuació: 100%
Desplegament	Puntuació: 50%	Puntuació: 100%	Puntuació: 100%
Explotació del model	Puntuació: 50%	Puntuació: 100%	Puntuació: 50%
Preu	Puntuació: 100%	Puntuació: 0%	Puntuació: 0%

Taula 4. Puntuacions obtingudes

A partir dels resultats obtinguts es poden obtenir un seguit de conclusions.
En general podem dir que els dos productes de pagament que hem estudiat són productes més complets i treballats. Que obtenen millors resultats ja que explorem més quantitat d'algoritmes i per tant construeixen models amb més exactitud. A més, l'experiència d'usuari també és notablement superior.
Tot i així, encara que el producte open source sigui més senzill, el producte és 100% funcional. És veritat que obté resultats pitjors pel que fa l'exactitud dels models construïts. Aquesta és una de les mètriques més importants de l'experiment, però aquesta diferència no és relativament gran, fent que en casos d'ús on no requereixin una exactitud extremadament elevada, Auto Weka pot ser útil.
Auto Weka és una molt bona opció per començar a treballar en l'àmbit del Auto ML i poder fer les primeres passes. Aquest producte ens permetrà obtenir models de gran qualitat sense necessitat de tenir grans coneixements en Machine Learning. Actualment, Auto Weka, és l'únic producte gratis que utilitza Auto ML i té una interfície gràfica.
Un altre punt important és que Auto Weka, a diferència de DataRobot i H2O driveless AI, no fa Neteja de dades ni Feature Engineering. Això significa que tots aquests pas-

sos s'han de fer de manera manual. Per tant, com més gran sigui el dataset, menor serà l'estalvi de temps.
En definitiva, es pot concloure que cada un de les tecnologies s'adequa a les necessitats de l'usuari. Si l'usuari necessita una solució en l'àmbit empresarial, és a dir, amb datasets de gran tamany, dades semi-estrucutrades o necessitat d'indicadors que ens ajudin a entendre els detalls dels models construïts, caldrà utilitzar productes de pagament.
En contrapartida, si necessitem fer ús d'una ena d'Auto ML en un àmbit més personal o acadèmic, un producte open source serà suficient.

7 Conclusions

Durant aquest treball s'ha presentat un concepte d'avanguarda conegut com Auto Machine Learning. S'ha fet un pas previ d'explicació de diversos conceptes per poder entendre de que tracta. Seguidamet s'han estudiat tres productes que fan ús d'aquest innovador terme per fer una posterior comparativa seguint unes mètriques.
Respecte a els objectius proposats durant el projecte, en general s'han pogut complir tots amb excepció de part d'un d'ells. El de realitzar una prova amb cada una de les tecnologies, on finalment un dels productes no ha pogut ser provat.
Com a idees futures per continuar aquest TFG es podrien afegir més tecnologies conegudes que fan ús de l'Auto ML, de forma que la comparativa abastaria gran part del mercat. Per altra banda, encara que el nombre de datasets utilitzats ha estat suficients per poder veure la tendència general, es podrien afegir uns quants per fer la comparativa més robusta.

8 Agraïments

En primer lloc agraïr a la companyia DataRobot la qual m'han facilitat una llicència de prova per fer ús del seu producte. Agraïr també a les altres companyies (Auto Weka i H2O) per tota la documentació publicada per poder entendre els seus productes.
Agrair també al meu tutor per totes les recomanacions que m'ha donat al llarg del treball per tal de millorar-lo i els consells que m'ha donat quan han sorgit imprevistos.

Bibliografia

[1] C. M. Jiménez, «Big data. Un nuevo paradigma de análisis de datos,» 2014. [Online]. Available: https://www.powerdata.es/big-data. [Last access: 15 Feb 2019].

[2] A. Ng, «What is Machine Learning?,» Coursera, 2019. [En línea]. Available: https://www.coursera.org/learn/machine-learning/supplement/aAgxl/what-is-machine-learning. [Last access: 27 Feb 2019].

[3] uiah, «Comparative Study,» Febrer 2019. [Online]. Available: http://www.uiah.fi/projekti/metodi/172.htm.

[4] Kaggle, «Datasets,» Març 2019. [Online]. Available: https://www.kaggle.com/datasets.

[5] Microsoft, «Microsoft Project,» Febrer 2019. [Online]. Available: https://products.office.com/es-es/project/project-and-portfolio-management-software.

[6] W. L. C. Z. M. A. V. i. J. H. M. Randal S. Olson, «Data-driven advice for applying machine learning to bioinformatics problems,» 2018. [Online]. Available: https://arxiv.org/pdf/1708.05070.pdf. [Last access: 29 May 2019].

[7] S. Marsland, Machine Learning, An Algorithmic Prespective, Cambridge, UK: Taylor & Francis Group, 2009.

[8] A. B. i. A. Allen, «Benchmarking Automatic Machine Learning Frameworks,» 7 Ago 2018. [Online]. Available: https://arxiv.org/pdf/1808.06492.pdf. [Last access: 6 Jun 2019].

[9] A. D. Marco, «Automated Machine Learning with H2O Driverless AI,» 3 Mai 2018. [Online]. Available: https://albertodema.wordpress.com/2018/05/03/automated-machine-learning-with-h2o-driverless-ai/. [Last access: 6 Jun 2019].

[10] C. T. H. H. H. F. H. i. K. L.-B. Lars Kotthof, «Auto-WEKA 2.0: Automatic model selection,» 2016. [Online]. Available: https://www.cs.ubc.ca/labs/beta/Projects/autoweka/papers/16-599.pdf. [Last access: 03 Apr 2019].

[11] C. T. F. H. Lars Kotthoff, «User Guide for Auto-WEKA version 2.6,» 2017. [Online]. Available: https://www.cs.ubc.ca/labs/beta/Projects/autoweka/manual.pdf. [Last access: 03 04 2019].

[12] L. Kotthoff, «autoweka,» 2017. [Online]. Available: https://github.com/automl/autoweka. [Last access: 03 Apr 2019].

[13] DataRobot, «DataRobot Product Datasheet,» 2017. [Online]. Available: https://3gp10c1vpy442j63me73gy3s-wpengine.netdna-ssl.com/wp-content/uploads/2017/07/DataRobot_Product_Datasheet_0617-1-1.pdf. [Last access: 07 04 2019].

[14] H2O.ai, «Driverless AI User Guide (EN),» 2019. [Online]. Available: http://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/index.html. [Last access: 12 Apr 2019].

[15] H2O.ai, «H2O Driverless AI Demo,» 1 Apr 2019. [Online]. Available: https://www.youtube.com/watch?v=wcyMBRRLmqs. [Last access: 28 Apr 2019].