Escola Tècnica Superior d'Enginyeria
Electrònica i Informàtica La Salle

Treball Final de Màster

Màster Universitari en Enginyeria de Telecomunicació

Predicció del moviment de les cotitzacions d'accions mitjançant *Machine Learning*

Alumne

Martí Finet Carrión

Professor Ponent

Dr. Xavier Vilasís Cardona

ACTA DE L'EXAMEN
DEL TREBALL FI DE CARRERA

Reunit el Tribunal qualificador en el dia de la data, l'alumne

D. Martí Finet Carrión

va exposar el seu Treball de Fi de Carrera, el qual va tractar sobre el tema següent:

Predicció del moviment de les cotitzacions d'accions mitjançant *Machine Learning*

Acabada l'exposició i contestades per part de l'alumne les objeccions formulades pels Srs. membres del tribunal, aquest valorà l'esmentat Treball amb la qualificació de

Barcelona,

VOCAL DEL TRIBUNAL VOCAL DEL TRIBUNAL

PRESIDENT DEL TRIBUNAL

Abstracte

La predicció dels moviments de les cotitzacions de la borsa ha sigut un gran repte per a la d'enginyeria artificial dels darrers anys i fins avui dia. Grans corporacions i bancs d'inversió dediquen molts recursos a aquest camp de recerca, i tot que aconsegueixen batre al mercat, ho fan amb un marge força ajustat. Així, han sigut molts els avanços en algorismes i models de predicció dissenyats per aquest repte, i les dades han obtingut un gran valor, del qual alguns en fan negoci.

Aquest estudi es basa en el disseny d'un model d'intel·ligència artificial que permeti predir el moviment de les cotitzacions de les accions d'un dels principals índex borsaris del mon. Amb aquest objectiu, s'obtenen de manera autònoma les dades històriques de les cotitzacions, així com els indicadors financers de les empreses en qüestió. Aquests, estan dividits en indicadors tècnics i fonamentals, i es fa una breu explicació a mode d'introducció, ja que aquest estudi no es financer i no entra dins l'abast fer un anàlisi dels mateixos.

Per altra banda, s'obtenen noticies d'economia d'un dels principals portal de notícies d'internet i es fan servir per predir els moviments d'un índex borsari. Des de l'obtenció de les notícies, fins a el processament d'aquestes i la codificació per a poder ser usades per un algorisme d'intel·ligència artificial, entren dins l'abast d'aquest estudi. A més, la codificació de text a codis, és fa servir amb un algorisme recent i amb fama entre els enginyers de dades que promet donar grans resultats.

Finalment, l'estudi agafa com a referència altres de molt similars i vol contrastar si els resultats presentats per aquests són realistes. Doncs, tal i com es descriu al llarg d'aquesta memòria, degut a la difícil naturalesa del repte i als problemes observats al llarg del projecte, els resultats han acabat sent poc atractius, però rigorosos.

Agraïments

És amb un immens plaer que reconec l'assistència i els comentaris en la meva tesi a la Dr. Rosa Maria Alsina. Especial gràcies al Dr. Xavier Vilasís Cardona per donar-me l'oportunitat i l'orientació en l'organització de la meva investigació, planificant i supervisant el meu treball. També m'agradaria expressar la meva gratitud per l'etern suport de la meva estimada família i amics, sense el qual no hagués sigut possible arribar fins al final del màster. Us ho dec tot.

Contingut

1 Introducció 13

1.1 Motivació 16

1.2 Abast 16

1.3 Objectius 17

1.3.1 Primer cas 17

1.3.2 Segon cas 17

2 Conceptes bàsics 18

2.1 La Borsa 18

2.1.1 Tipus de ordres 19

2.2 Machine Learning 19

2.2.1 Aprenentatge supervisat. 19

2.2.2 Aprenentatge no supervisat. 22

2.2.3 Aprenentatge semi supervisat. 24

2.2.4 Aprenentatge per reforçament. 25

2.2.5 Casos d'us de ML per a inversors 26

2.2.6 Arbres de decisió 27

2.2.7 Maquines de potenciació del gradient 29

2.2.8 Deep Learning 31

2.2.9 Treballant amb dades de text 34

2.2.10 BERT 36

3 Estat de l'art 39

4 Coneixements financers 43

4.1 Anàlisi Fonamental 43

4.1.1 Informes Financers 44

4.1.2 Indicadors Fonamentals 51

4.2 Anàlisi Tècnic 55

4.2.1 Indicadors Tècnics 57

5 Arquitectura de l'estudi 66

5.1 Python 66

5.2 Arquitectura de sistema 66

5.2.1 Primer cas 67

5.2.2 Segon cas 68

6 Dataset 69
6.1 Primer cas 69
6.1.1 Cotitzacions 69
6.1.2 Indicadors d'anàlisi fonamental 72
6.1.3 Indicadors d'anàlisi tècnic 74
6.2 Segon cas 77
6.2.1 Dades històriques del DJIA 77
6.2.2 *Reddit WorldNews* top 25 Notícies 77
7 Desenvolupament 79
7.1 Primer cas 79
7.1.1 Pandas 79
7.1.2 LABEL 80
7.1.3 Train, Validation i Test 81
7.2 Segon cas 82
7.2.1 Notícies 82
7.2.2 DJIA 83
7.2.3 BERT 83
8 Anàlisi dels resultats 85
8.1 Primer cas 85
8.1.1 XBOOST 85
8.2 Segon cas 87
8.2.1 SVM 87
9 Problemes observats 89
9.1 Component humà 89
9.2 Limitacions tècniques 90
9.3 Interpretació dels indicadors 90
9.4 Manipulació dels resultats 90
9.5 Dataset 91
9.6 Cross-validation 91
10 Conclusions 93
10.1 Línies de futur 95
11 Referències 97

Acrònims

En aquest capítol es llisten els acrònims que apareixen en el document, per ordre alfabètic. Ja no es desplegaran de nou a la resta del document. Si els acrònims provenen de paraules escrites en un idioma diferent del document, s'escriurà en cursiva.

ADX: *Average Directional Index*.

AG: Algoritme Genètic.

ANN: *Artificial Neural Networks*.

API: *Application Programming Interface*.

ARIMA: *Auto Regressive Integrated Moving Average*.

B&H: *Buy And Hold*.

BAI: Benefici Abans d'Impostos.

BSE SENSEX: *Bombay Stock Exchange Sensitive Index*.

CBR: *Case-Based Reasoning*.

CCI: *Commodity Channel Index*.

DL: *Deep Learning.*

DJIA: *Dow Jones Industrial Average*.

DJTA: *Dow Jones Transportation Average*.

EGARCH: *Exponential GARCH*.

EMA: *Exponential Moving Average*.

EMRLF: Mètode de Predicció Lineal de l'Índex Morfològic.

ETF: *Exchange Traded Fund*.

FDP: *Financial Distress Prediction*.

FPE: *Final Prediction Error*.

GARCH: *Generalized AutoRegressive Conditional Heteroskedasticity*.

GBM: *Gradient Boosting Machines*.

GEP: *Programació d'Expressions Genètiques*.

GJR-GARCH: *Glosten-Jagannathan-Runkle GARCH*.

GMM: *Generalized method of moments*.

GP: *Genetic Programming*.

HMM: *Hidden Markov Model*.

HSI: *Hang Seng Index*.

IA: Intel·ligència Artificial.

IGARCH: *Integrated GARCH*.

LDA: *Linear Discriminant Analysis*.

LS-SVM: *Least Squares Support Vector Machine*.

MACD: *Moving Average Convergence / Divergence*.

MEP: Programació d'Expressions Múltiples.

ML: *Machine Learning.*

MLP: *Multilayer Perceptron*.

NASDAQ: *National Association of Securities Dealers Automated Quotation*.

NLP: *Neuro-linguistic programming*.

NSE: *National Stock Exchange of India*.

OBV: *On Balance Volume*.

PCA: *Principal component analysis*.

PNN: *Probabilistic Neural Network*.

QDA: Q*uadratic Discriminant Analysis*.

ROA: *Return On Assets.*

ROE: *Return On Equity*.

ROC: *Rate Of Change*.

RSI: *Relative Strength Index.*

SEC: *Securities and Exchange Commission*.

SMA: *Simple Moving Average*.

SOM: *Self Organizing Map*.

STOCH: *Stochastic Oscillator*.

SVM: *Support Vector Machine*.

SVR: *Support Vector Regression*.

TAEF: *Time-delay Added Evolutionary Forecasting*.

TFM: Treball Final de Màster.

1 Introducció

En primer lloc, el comerç especulatiu en un esforç de vèncer el mercat és tan antic com els mercats mateixos. Avui dia, la gestió d'inversions activa és un gran negoci que mou milions de dòlars a tot el món i es basa la possibilitat de generar un rendiment tan alt com sigui possible amb un nivell de risc acceptable i un temps establert.

Avui en dia, hi ha més de 10.000 fons d'inversió que gestionen aproximadament 3 bilions de dòlars en actius. Un fons d'inversió és una associació d'inversions entre un gestor del fons professional i socis limitats o inversors. Els socis limitats aporten fons, mentre que el soci general gestiona el fons segons l'estratègia del fons per maximitzar els rendiments dels inversors i minimitzar el risc. La firma d'investigació de mercats *Preqin*[1] estima que 1.360 fons d'inversió utilitzen intel·ligència artificial per les seves inversions.

La indústria de la inversió ha evolucionat de manera espectacular durant les últimes dècades i continua fent-ho enmig d'un augment de la competència, els avenços tecnològics i un desafiant entorn econòmic.

El ML està canviant pràcticament tots els aspectes de la vida del esser humà. Avui algorismes de ML realitzen tasques que fins fa poc només podien realitzar experts humans. Pel que fa a les finances, hi ha una clara tendència adaptar aquesta tecnologia disruptiva que transformarà la inversió de tothom durant generacions.

Les tendències que han impulsat el la inversió mitjançant sistemes de ML al nivell actual, inclouen les següents.

- Canvis en la estructura dels mercats financers, com ara la propagació del comerç electrònic i la integració dels mercats entre classes d'actius i geografies.
- El desenvolupament de les estratègies d'inversió emmarcades en termes de factor de risc d'exposició.
- Les revolucions en potència de càlcul, generació de dades i gestió, i mètodes analítics.
- La superació dels pioners en els operadors algorísmics en relació amb els humans.

A més, les crisis financeres del 2001 i del 2008 han afectat la manera en què els inversors afronten la diversificació i gestió de riscos. Donant lloc a inversions passives de baix cost en forma de fons ETFs. Així, el rendiment sumat a la baixa volatilitat després de la crisi del 2008, els inversors conscients dels costos van canviar 2 bilions de dòlars de gestió activa fons d'inversió en ETFs amb gestió passiva.

Per altra banda, la pressió competitiva també es reflecteix en menors honoraris dels fons de cobertura[2] que van baixar de la quota de gestió tradicional del 2% i del 20% de beneficis fins a una mitjana de 1,48% i 17,4%, respectivament, el 2017.

[1] https://www.preqin.com/
[2] https://en.wikipedia.org/wiki/Hedge_fund

El comerç electrònic ha avançat espectacularment en termes de capacitat, volum i cobertura de classes d'actius i geografies, ja que les xarxes van iniciar gestionant els preus amb sistemes informàtics als anys seixanta.

Els mercats de renda variable han liderat aquesta tendència a tot el món. L'any 1997 la SEC[3] va augmentar la competència a les borses afegint xarxes de comunicacions electròniques (ECN). Els ECNs són sistemes automatitzats de negociació alternativa (ATS) que coincideixen amb la compra i la venda comandes a preus especificats, principalment de renda variable i de divises, i registrats com brókers[4]. Fet que va facilitar l'intercanvi d'actius entre individus de diferents zones geogràfiques llocs per intercanviar directament sense intermediaris.

L'Accés al Mercat Directe (DMA[5]) proporciona a al inversor un major control sobre l'execució, permetent-lo fer-ho enviar ordres directament a la borsa, fent servir la infraestructura i mitjançant d'un bróker que és membre d'una borsa. Aquest fet, juntament amb d'altres, com per exemple, la eliminació de certs controls de riscos realitzats pels brókers, van construir la base per al comerç d'alta freqüència (HFT[6]).

HFT es refereix a operacions automatitzades d'instruments financers que s'executen amb un període de temps curt, de microsegons, i on els participants tenen en propietat els actius durant intervals de temps molt curts. L'objectiu és detectar i explotar les ineficiències en la microestructura del mercat i la infraestructura institucional dels llocs de negociació. A més, HFT ha crescut substancialment en el passat deu anys i es calcula que representen aproximadament el 55% del volum de negociació dels mercats de renda variable dels EUA i al voltant del 40% en els mercats de renda variable europeus. Paral·lelament, HFT també ha crescut en mercats de futurs[7] a aproximadament el 80% dels futurs de canvi de divises i els dos terços del tipus d'interès i de la taxa d'interès Volum de futurs a 10 anys de tresoreria (FAS 2016[8]).

Les estratègies HFT tenen com a objectiu obtenir beneficis mitjançant estratègies passives o agressives. Les estratègies passives inclouen el comerç d'arbitratges[9] per beneficiar-se de diferències de preus molt reduïdes per a el mateix actiu o els seus derivats que es negocien en diferents llocs. Les estratègies agressives inclouen ordres d'anticipació o creació d'impulsos.

Per un costat, anticipació d'ordres implica algorismes que presenten petites ordres exploratòries per detectar liquiditat ocultes de grans inversors institucionals i avançar-se a grans ordres per beneficiar-se del moviment de preus posterior.

En canvi, la creació de l'impuls implica l'execució d'un algorisme i cancel·lar una sèrie d'ordres per manipular altres algorismes HFT en comprar (o vendre) més agressivament i es beneficien dels canvis de preu resultants.

---

[3] https://en.wikipedia.org/wiki/U.S._Securities_and_Exchange_Commission
[4] https://en.wikipedia.org/wiki/Broker
[5] https://en.wikipedia.org/wiki/Direct_market_access
[6] https://en.wikipedia.org/wiki/High-frequency_trading
[7] https://en.wikipedia.org/wiki/Futures_exchange
[8] https://www.imf.org/external/index.htm
[9] https://www.investopedia.com/terms/a/arbitrage.asp

Els reguladors han expressat la seva preocupació per la possible relació entre certes estratègies HTF agressives i una major fragilitat i volatilitat del mercat, com la que es va experimentar durant el *Crash Flash* de maig de 2010[10], la volatilitat del mercat del tresor americà d'octubre de 2014 i la caiguda sobtada més de 1.000 punts del DJIA el 24 d'agost de 2015[11]. Al mateix temps, la liquiditat del mercat ha augmentat amb els volums d'activitat comercials a causa de la presència de HFT, que ha reduït els costos globals de les transaccions.

En aquest context, els millors teòrics financers i els científics de dades són contractats per intentar predir el mercat de valors amb l'objectiu d'augmentar el retorn de les inversions. No obstant això, a causa de la naturalesa multidimensional del problema, de la magnitud i de la variació inherent amb el temps, ha estat un repte aclaparadorament difícil per als humans de resoldre, fins i tot amb l'ajut d'eines d'anàlisi de dades convencionals

Tanmateix, amb l'avanç dels recents avenços en les aplicacions de ML, el camp ha anat evolucionant per utilitzar tot tipus de solucions que intenten aprendre el comportament dels mercats per fer prediccions més precises. Alhora, les tres revolucions en la potència de càlcul, les dades i els mètodes de ML han fet que s'hagi fet l'adopció d'estratègies sistemàtiques, basades en dades, no només més convincents i rendibles però una font clau d'avantatge competitiu.

Com a resultat, els la inversió mitjançant aquest sistemes no només troben una aplicació més àmplia als fons de cobertura, indústria que va ser pionera en aquestes estratègies, si no que la majoria de fons d'inversió en fan us. En particular, la anàlisis predictiu, mitjançant ML i l'automatització algorítmica tenen un paper cada vegada més destacat en tots passos del procés d'inversió entre classes d'actius, des de la generació d'idees i la investigació fins a formulació d'estratègies i construcció de carteres, execució comercial i gestió de riscos.

Les estimacions de la mida de la indústria varien perquè no hi ha una definició clara i a més, molts fan servir un enfocament humà-màquina, amb sistemes híbrids. Morgan Stanley va estimar el 2017 que les estratègies d'inversió mitjançant ML van créixer al 15% anual durant els darrers sis anys i controlen prop d'1,5 bilions de. Altres informes suggereixen la indústria quantitativa[12] de fons de cobertura va superar els 1 bilió de dòlars d'euros, gairebé duplicant la seva grandària des del 2010 en detriment de fons de cobertura tradicionals. En canvi, el capital total de la indústria dels fons de cobertura va assolir els 3.218 milions de dòlars d'acord amb l'últim informe global d'investigació sobre fons de cobertura.

També, la firma d'investigació de mercat *Preqin* estima que gairebé 1.500 fons de cobertura tenen la majoria dels seus oficis amb ajuda de models informàtics. Els fons de cobertura quantitatius són ara responsable del 27% de totes les operacions borsàries nord-americanes dels inversors, un 14% més que el 2013.

En els últims anys, no obstant això, els fons s'han mogut cap a un ML veritable, on sistemes intel·ligents poden analitzar grans quantitats de dades a gran velocitat i millorar-se a través d'aquest tipus anàlisis.

[10] https://en.wikipedia.org/wiki/2010_Flash_Crash
[11] https://en.wikipedia.org/wiki/2015%E2%80%9316_stock_market_selloff
[12] https://en.wikipedia.org/wiki/Quantitative_fund

1.1 Motivació

La gestió del patrimoni és una de les coses que més i millor és capaç de transformar la vida de les persones en tots els sentits, i a més d'una forma permanent.

L'educació financera és un dels assumptes més importants que existeixen i no obstant això és força ignorada per la major part de la població. Hauria de ser d'ensenyament obligatori a les escoles perquè gestionar diners és una cosa que absolutament tothom, sense excepció, ha de fer al llarg de la seva vida. No només és impossible viure sense gestionar diners sinó que la forma en què es gestioni és una de les coses que més determinen la vida d'una persona, per bé o per malament, en tots els sentits.

Els diners no són un fi en si mateix, sinó un mitjà per poder viure la vida que cada un desitja viure. I no és un mitjà qualsevol, sinó un dels que més poder tenen per transformar i definir la vida d'una persona en un sentit o un altre.

El futur econòmic, i no econòmic, d'una persona ho determina més la bona o mala gestió dels diners que faci aquesta persona, que els ingressos que arribi a tenir al llarg de la seva vida. Es podria dir que la situació actual de qualsevol persona és el resultat de les decisions financeres preses en el passat.

És evident que si es guanyen molts diners i a més es gestiona bé els resultats són espectaculars, però entre aquestes dues situacions: guanyar molts diners i gestionar malament; o guanyar pocs diners i gestionar bé, a llarg termini és preferible la segona.

Qualsevol persona, per pocs diners que guanyi, pot arribar a tenir un bon patrimoni si fa una bona gestió dels seus diners, i aquest patrimoni és el que li permetrà transformar la seva vida en el que desitgi.

Gestionar bé els diners dóna a la persona llibertat i independència veritable i més, de forma indirecta, millora la situació de la resta de la Humanitat.

1.2 Abast

En aquest context de cerca de la llibertat financera, la primera pregunta que sol fer-se una persona quan comença a pensar en gestionar els seus diners és on invertir. En base als objectius vitals de la persona es defineixen estratègies d'inversió. En aquest estudi no es parla d'aquestes estratègies ni de la gestió de carteres o patrimoni, si no que es base en intentar dissenyar un sistema capaç de predir els moviments de les cotitzacions de les accions de qualsevol índex borsari.

La predicció de les cotitzacions del mercat de valors és l'art que té com a objectiu d'intentar determinar el valor o comportament futur d'una empresa o un altre instrument financer cotitzat en un intercanvi financer. El benefici d'una predicció amb alt percentatge d'encert es maximitzar els guanys dels inversors.

Cal destacar que aquest estudi no és financer, i per tant es centra en la cerca i desenvolupament del sistema d'intel·ligència artificial capaç de predir el moviment de les cotitzacions d'empreses.

1.3 Objectius

A continuació es defineixen els objectius que persegueix aquest estudi. El principal és el de desenvolupar un sistema basat en intel·ligència artificial capaç de predir el moviment de les cotitzacions d'accions, agafant com a referència els estudis similars trobats i contrastar els resultats presentats en aquests.

S'ha estudiat dos cassos per aquest estudi, donada la dificultat a la que s'enfronta.

1.3.1 Primer cas

Estudiar i dissenyar un sistema capaç de predir el moviment del preu de les accions de les empreses d'un dels principals índex borsaris del mon, com per exemple el DJIA, basat en el històric de preu d'aquestes, i també en indicadors extrets d'anàlisi fonamental i tècnic.

1.3.2 Segon cas

Estudiar i dissenyar un sistema capaç de predir el moviment del preu d'un dels principals índex borsaris del mon, com per exemple el DJIA, basat en notícies d'un conegut diari digital.

2 Conceptes bàsics

En aquest capítol es fa una explicació sobre conceptes necessaris per seguir el desenvolupament del projecte. Està dividit en dues parts, la primera sobre els mercats de valors, i l'altre sobre ML.

Evidentment, es fa èmfasis en la segona, en la que es veu breument que és i les seves diferents branques, així com una petita introducció de cada una. En capítols posterior d'aquest estudi, és dedica una part a l'explicació de coneixements sobre finances, també necessaris. Tot i que aquest estudi no pretén incloure teoria financera, com es mostra, s'obtenen indicadors que seran dades que no només farà servir el model de ML, si no que a més la seva comprensió esdevé molt rellevant en el resultat del mateix.

2.1 La Borsa

La negociació d'actius financers es produeix en les borses[13], és una organització privada que brinda les facilitats necessàries per a que els seus membres, atesos als mandats dels seus clients, introdueixin ordres i realitzin negociacions de compra i venda de valors, per exemple com accions de companyies o societats anònimes, bons públics i privats, certificats, títols de participació i una àmplia varietat d'instruments d'inversió.

És a dir, és un mercat central on es troben compradors i venedors, i els compradors competeixen entre ells per obtenir l'oferta més alta mentre els venedors competeixen pel més baix oferta.

Aquestes, poden basar-se en sistemes bilaterals de negociació o en sistemes de comandes, on es poden enviar comandes de compra i venta, seguint certes regles. Moltes utilitzen intermediaris que proporcionen liquiditat i poden actuar com a distribuïdors que comercien per compte propi o com a com a brokers en nom d'altres.

Il·lustració 1.- El toro de *Wall Street* és una escultura de bronze creada per *Arturo Di Modica* l'any 1987, símbol de la força i el poder. Avui dia es fa servir el qualificatiu toro per a un mercat on els preus pugen o s'espera que pugin. En canvi, en el cas contrari es diu que es un mercat ós.

[13] https://en.wikipedia.org/wiki/Stock_exchange

2.1.1 Tipus de ordres

Els inversors poden col·locar diversos tipus de comandes de compra o venda. Algunes comandes garanteixen immediatament execució, mentre que altres poden indicar un llindar de preu o altres condicions per a que es s'executin. Les comandes solen ser vàlides per al mateix dia de negociació, llevat que s'indiqui el contrari.

L'ordre de mercat garanteix l'execució immediata de l'ordre, al preu que prevalgui en aquell moment. Per contra, una ordre limitada només s'executa si el preu de mercat és més alt que el límit establert en el cas de una ordre de venta. Contràriament, si és de compra només s'executa si el preu és més baix que l'establert en l'ordre.

Per altra costat, una ordre de parada s'executa quan el preu del mercat puja per sobre o cau per sota a un preu especificat per a una comanda de compra o venta. Es pot utilitzar un ordre de parada de compra per limitar les pèrdues de vendes curtes. Tot i que, aquestes comandes poden estar limitades.

2.2 Machine Learning

Segons la definició de Wikipedia[14],

" L'Aprenentatge automàtic és un camp de la intel·ligència artificial que està dedicat al disseny, l'anàlisi i el desenvolupament d'algorismes i tècniques que permeten que les màquines evolucionin. És a dir, tracta de crear programes capaços de generalitzar comportaments a partir del reconeixement de patrons o classificació. L'aprenentatge automàtic està relacionat amb el camp de l'estadística, però també coincideix amb els mètodes de construcció de models, o l'aprenentatge estadístic. "

També hi ha punts de contacte amb la informàtica teòrica. Això és degut a la complexitat computacional dels problemes. Alguns camps on s'ha aplicat aquest tipus d'aprenentatge són les aplicacions dedicades al processament del llenguatge natural, als algorismes de cerca, la diagnosi mèdica, la bioinformàtica, la detecció de fraus i la classificació. Qualsevol sistema que es consideri intel·ligent ha de tenir l'habilitat d'aprendre, és a dir, de millorar automàticament amb l'experiència.

Els programes utilitzats són sistemes d'aprenentatge capaços d'adquirir coneixements d'alt nivell i estratègies per la resolució de problemes mitjançant exemples, de forma anàloga a com ho faria la ment humana.

Els algorismes d'aprenentatge automàtic es classifiquen d'acord amb allò que s'espera que el programa aprengui i també segons el grau d'interacció amb l'usuari.

2.2.1 Aprenentatge supervisat.

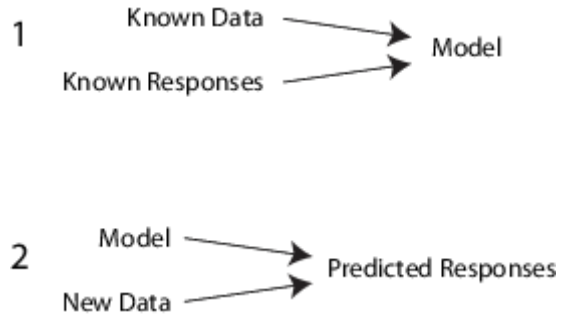
L'aprenentatge supervisat[15] és el tipus de ML que s'utilitza més sovint. El terme supervisat implica la presència d'una variable de resultat que guia l'aprenentatge procés, és a dir, ensenya a l'algorisme la solució correcta a la tasca que s'està aprenent.

[14] https://ca.wikipedia.org/wiki/Aprenentatge_autom%C3%A0tic
[15] https://en.wikipedia.org/wiki/Supervised_learning

A més, té com a objectiu generalitzar una relació funcional entre dades d'entrada i sortida que s'obtenen de mostres individuals i apliquen a dades noves.

La variable de sortida és també, depenent del camp, anomenada de forma intercanviable l'etiqueta, objectiu, resultat, variable endògena o de banda esquerra. Fins i tot, algunes tasques estan representades per diversos resultats, també anomenats problemes de multi etiqueta. També es coneixen les dades d'entrada d'un problema d'aprenentatge supervisat com a característiques, variables exògenes i de la dreta.

La solució a un problema d'aprenentatge supervisat és una funció que representa el que el model ha après sobre la relació entre l'entrada i la sortida de la mostra i s'aproxima a la relació real. Així, es pot utilitzar per trobar associacions estadístiques o relacions de causalitat entre variables d'interès més enllà de les de mostra, o es pot utilitzar per predir resultats per a noves dades d'entrada.



Il·lustració 2.- En l'aprenentatge supervisat, amb un conjunts de dades donat aprèn quin hauria de ser el resultat adient.

Els dos objectius s'enfronten a un important compromís: els models més complexos tenen més parts mòbils i són capaços de representar relacions més matisades, però també poden ser més difícils d'inspeccionar. Però, també és probable que sobre aprenguin, i aprenguin sorolls aleatoris particulars pertanyents a la mostra, a diferència d'una funció més general que representi una relació d'entrada relació, més aplicable a altres mostres.

D'altra banda, els models massa simples perdran senyals i lliuraran resultats esbiaixats. Aquest compromís es coneix com el compromís de variació del biaix[16], però això també s'aplica a les altres formes de ML, on models excessivament complexes poden funcionar malament més enllà de les dades d'entrenament.

Aquest tipus de aprenentatge es categoritza en models de classificació i de regressió.

2.2.1.1 Classificació.

Els problemes de classificació tenen variables de resultat categòriques. La majoria dels predictors donen una puntuació per indicar si una observació pertany a una classe determinada. En el segon pas, aquests puntuacions es tradueixen en prediccions reals.

En el cas binari, on etiquetarem les classes positives i negatives, la puntuació normalment varia entre zero o està normalitzada en conseqüència.

[16] https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff

Un cop les puntuacions es converteixen en prediccions 0 o 1, hi pot haver quatre resultats, ja que cadascuna de les dues classes existents pot ser predit correctament o incorrectament. Amb més de dues classes, pot haver-hi més casos si es diferencia entre els diversos possibles errors.

		Actual (Truth)	
		Positive	**Negative**
Prediction	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Taula 1.- Tots els indicadors d'error es calculen a partir del desglossament de les prediccions a través dels quatre camps de la matriu de confusió[17] de 2 x 2 que associa les classes reals i predites

Accuracy	$= \frac{\text{\# Correct Predictions}}{\text{\# Cases}}$	$= \frac{TP + TN}{TP + FP + TN + FN}$
True Positive Rate (Sensitivity, Recall)	$= \frac{\text{\# Correct Positive Predictions}}{\text{\# Positive Cases}}$	$= \frac{TP}{TP + FN}$
False Negative Rate (Miss Rate)	= 1 - True Positive Rate	
True Negative Rate (Specificity)	$= \frac{\text{\# Correct Negative Predictions}}{\text{\# Negative Cases}}$	$= \frac{TN}{TN + FP}$
False Positive Rate (Fall-Out)	= 1 - True Negative Rate	

Taula 2.- Diferents mètriques que es poden obtenir en problemes de classificació.

[17] https://en.wikipedia.org/wiki/Confusion_matrix

Finalment, un classificador no necessàriament ha de proporcionar un resultat amb probabilitats, sinó que pot produir puntuacions relatives entre si per distingir els casos positius de negatius. Per tant, el llindar és una variable de decisió que pot i ha de ser optimitzada tenint en compte tenir en compte els costos i els beneficis de prediccions correctes i incorrectes. Un llindar inferior implica prediccions més positives, amb una taxa de falsificació positiva potencialment ascendent, i per a un llindar més alt, és probable que sigui el contrari.

2.2.1.2 Regressió.

Els problemes de regressió pretenen predir una variable contínua. Són un conjunt de processos estadístics per estimar les relacions entre variables.

Inclou moltes tècniques per a la modelització i l'anàlisi de diverses variables, quan el focus se centra en la relació entre una variable dependent i una, o més variables independents o predictors. Més específicament, l'anàlisi de regressió ajuda a entendre com el valor típic de la variable dependent o variable de criteri, canvia quan qualsevol de les variables independents és variada, mentre que les altres variables independents es mantenen fixes. La regressió s'utilitza per predir una mesura basant-nos en el coneixement d'una altra. Hi ha diversos tipus de regressions.

- Regressions lineal (simple o múltiple)
- Regressions no lineal

2.2.2 Aprenentatge no supervisat.

L'aprenentatge sense supervisió[18] afegeix valor mitjançant el descobriment d'estructures de les dades sense una variable de resultat per guiar el procés de cerca. Aquesta tasca contrasta amb el context d'aprenentatge supervisat vist prèviament.

Els algorismes d'aprenentatge no supervisats poden ser útils quan es disposa d'un conjunt de dades que només conté funcions i cap mesura del resultat, o quan es vol extreure informació independentment de el resultat. En lloc de predir resultats futurs, l'objectiu és estudiar un document informatiu representació de les dades útils per resoldre una altra tasca, inclosa l'exploració d'un conjunt de dades.

Sovint, els algoritmes sense supervisió tenen com a objectiu aprendre una nova representació de les dades d'entrada útil per a algunes altres tasques. Això inclou elaborar etiquetes que identifiquin punts comuns entre les observacions, o una descripció resumida que capta informació rellevant que requereix punts de dades o característiques. Els algorismes d'aprenentatge no supervisats també difereixen dels algorismes d'aprenentatge supervisats en els supòsits que fan sobre la naturalesa de l'estructura que volen descobrir. La reducció de la dimensionalitat i l'agrupació són les tasques principals d'aprenentatge sense supervisió.

2.2.2.1 algorismes de clusterització

Els algorismes de clusterització identifiquen i agrupen observacions o característiques similars en lloc d'identificar noves funcions. Els algorismes difereixen en com defineixen el similitud d'observacions i les seves suposicions sobre els grups resultants.

[18] https://en.wikipedia.org/wiki/Unsupervised_learning

És a dir, utilitzen una mesura de similitud per identificar observacions o atributs de dades que contenen informació similar. Resumeixen un conjunt de dades assignant un gran nombre de *data points* i a un nombre menor de clústers, perquè els membres del clúster estiguin més de relacionats entre si amb que amb membres d'altres clústers. A més, els algorismes de clúster poden variar principalment pel que fa al tipus de clústers que produeixen, fet que implica supòsits diferents sobre el procés de generació de dades, que s'indiquen a continuació:

- Agrupament de K-mitjans[19]: els *data points* pertanyen a un dels *k* grups de mateix mida que prenen una forma el·líptica.
- Models de mescles gaussianes[20]: s'han generat *data points* per qualsevol de les diverses distribucions normals multivariants.
- Clústers basats en la densitat[21]: els clústers són de forma arbitrària i només es defineixen per l'existència d'un nombre mínim de *data points* propers.
- Grups jeràrquics[22]: els *data points* pertanyen a diversos superconjunts, o *sets*, de grups formats per la fusió successiva de grups més petits.

2.2.2.2 La reducció de la dimensionalitat

La reducció de la dimensió transforma les dades existents en un nou, més petit conjunt, minimitzant la pèrdua d'informació. Existeix una àmplia gamma d'algoritmes que difereixen només en com mesuren la pèrdua d'informació, tant si s'apliquen transformacions lineals o no lineals, o les limitacions que imposen al nou conjunt de dades.

Així, produeix noves dades que capturen la informació més important continguts a les dades d'origen. En comptes d'agrupar les dades existents en clústers, aquests algorismes transformen les dades existents en un conjunt de dades nou que utilitza significativament menys característiques o observacions per representar la informació original. Els algorismes varien segons a la naturalesa del nou conjunt de dades que generaran, com es mostra a la llista següent:

- PCA[23]: troba la transformació lineal que captura la major part de la variància del conjunt de dades existent.
- Aprenentatge múltiple[24]: identifica una transformació no lineal que produeix una dimensió inferior representació de les dades.
- Codificadors automàtics[25]: utilitza xarxes neuronals per comprimir dades no lineals amb pèrdua mínima d'informació .

Més endavant s'explica amb més detall algorismes lineals, no lineals i xarxes neuronals d'aprenentatge no supervisat, incloses les aplicacions importants de processament del llenguatge natural.

[19] https://en.wikipedia.org/wiki/K-means_clustering
[20] https://en.wikipedia.org/wiki/Mixture_model
[21] https://en.wikipedia.org/wiki/DBSCAN
[22] https://en.wikipedia.org/wiki/Hierarchical_clustering
[23] https://en.wikipedia.org/wiki/Principal_component_analysis
[24] https://en.wikipedia.org/wiki/Nonlinear_dimensionality_reduction
[25] https://en.wikipedia.org/wiki/Autoencoder

2.2.2.3 Aplicacions

Hi ha diversos usos útils de l'aprenentatge sense supervisió que es pot aplicar a la inversió en accions, per exemple els següents:

- Agrupació d'accions, o altres instruments financers, amb característiques de risc i rendiment similars.
- Trobar una petita quantitat de factors de risc que tenen un gran impacte en el rendiment d'un nombre molt més gran de valors.
- Identificar patrons de negociació i de preus que difereixen sistemàticament i poden suposar riscos més alts.
- Identificació de temes latents en un cos de documents que inclouen els aspectes més importants d'aquests documents.

2.2.3 Aprenentatge semi supervisat.

Recapitulant, s'ha vist una visió de les diferències clau entre el ML supervisat i no supervisat. Però, hi una tercer mètode que aprofita la barreja de tots dos, un enfocament anomenat aprenentatge semi-supervisat[26].

La diferència més gran entre l'aprenentatge automàtic supervisat i no supervisat és que els algorismes d'aprenentatge automàtic supervisat es formen sobre conjunts de dades que inclouen etiquetes afegides per un humà que guia l'algorisme per entendre quines característiques són importants per al problema. D'altra banda, els algorismes d'aprenentatge automàtic no supervisats es formen sobre dades no etiquetades i han de determinar la importància de les característiques pel seu compte basant-se en patrons inherents a les dades.

Així, els algorismes d'aprenentatge semi-supervisats es formen en una combinació de dades etiquetades i no etiquetades. Això és útil per alguns motius. En primer lloc, el procés d'etiquetatge de quantitats massives de dades per a l'aprenentatge supervisat requereix sovint un temps desmesurat i car. A més, massa etiquetatge pot imposar biaixos humans al model. Això vol dir que incloure moltes dades no etiquetades durant el procés d'entrenament tendeix a millorar la precisió del model final tot reduint el temps i els costos que es dediquen a construir-lo.

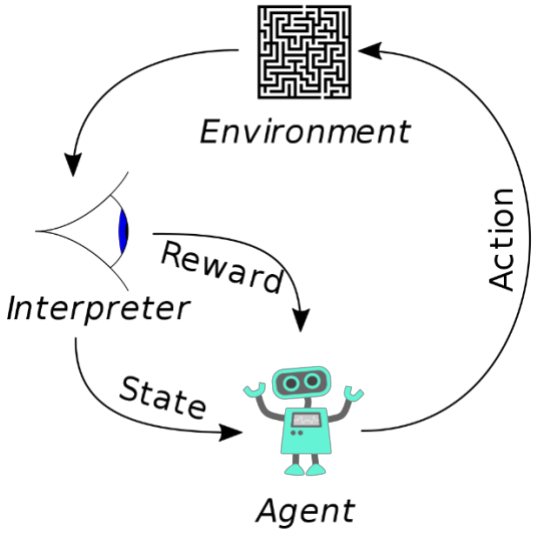
A mode d'il·lustració, utilitzant la classificació com a exemple, es compara com funcionen aquests tres enfocaments a la pràctica:

- Classificació supervisada: l'algorisme aprèn a assignar etiquetes basades en les etiquetes que un humà va introduir durant el procés de formació.
- Agrupació no supervisada: l'algorisme analitza les similituds inherents entre les dades per agrupar-les en grups.
- Classificació semi-supervisada: les dades etiquetades s'utilitzen per identificar que hi ha en els grups específics a les dades i classificar-les. L'algorisme s'entrena després amb dades no etiquetades per definir els límits d'aquests tipus de dades i fins i tot pot identificar nous característiques de les dades que no estaven especificades a les etiquetes introduïdes per humans.

[26] https://en.wikipedia.org/wiki/Semi-supervised_learning

2.2.4 Aprenentatge per reforçament.

L'aprenentatge per reforçament[27] és el quart tipus de ML. El seu objectiu és escollir l'acció que produeix la recompensa més alta, donat un conjunt de dades d'entrada que descriuen un context o entorn. És un procés dinàmic i interactiu, ja que el corrent de recompenses positives i negatives afecta l'aprenentatge dels algoritmes, i les accions realitzades ara poden influir tant en l'entorn com en recompenses futures.



Il·lustració 3.- Típic escenari d'aprenentatge per reforçament, on un agent pren accions en un entorn, que s'interpreta en una recompensa i una representació de l'estat, que es reenvia a l'agent.

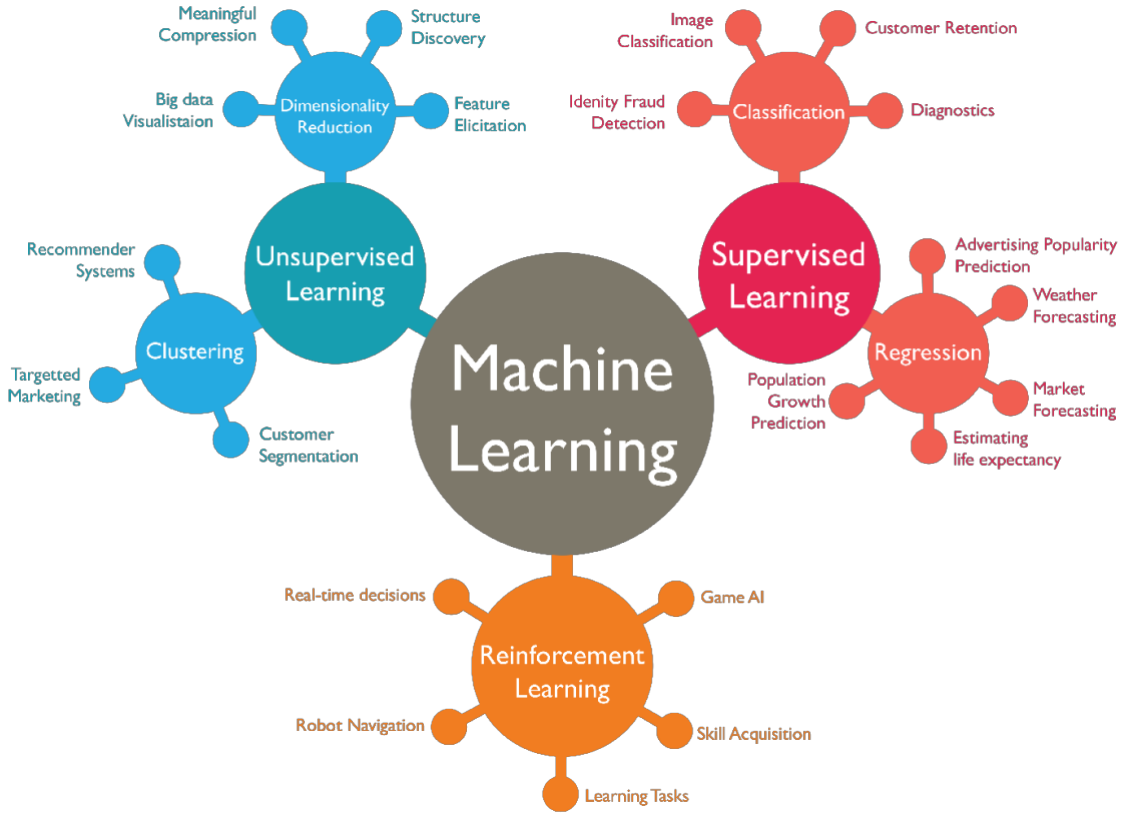
El compromís entre l'explotació el curs d'una acció ha après a produir una certa recompensa i l'exploració de noves accions que puguin augmentar la recompensa en el futur dóna lloc a un enfocament de prova i error. L'aprenentatge de reforçament optimitza l'agent aprenentatge utilitzant la teoria de sistemes dinàmics i, en particular, el control òptim processos de decisió de Markov[28] amb informació incompleta.

Aquest tipus d'aprenentatge difereix de l'aprenentatge supervisat, on les dades d'entrenament disponibles ten establertes tant el context com la decisió correcta per a l'algorisme. Està dissenyat per a escenaris interactius on els resultats només es posi a disposició al llarg del temps i l'aprenentatge es produeix de forma contínua a mesura que l'agent adquireixi noves experiències. No obstant, alguns dels progressos més notables en IA implica un reforç que utilitza *deep learning* per aproximar les relacions entre accions, entorns i recompenses futures. També difereix de la l'aprenentatge no supervisat, ja que la el resultat de les conseqüències esta disponible, encara que amb retard.

Finalment, l'aprenentatge de reforçament és especialment adequat per a la predicció d'accions perquè el concepte d'un agent maximitzador de retorn en un entorn dinàmic incert, té molt en comú amb un inversor o una estratègia d'inversió que interactua amb els mercats financers.

[27] https://en.wikipedia.org/wiki/Reinforcement_learning
[28] https://en.wikipedia.org/wiki/Markov_decision_process

Aquest enfocament s'ha aplicat amb èxit als agents de joc, sobretot al joc de Go[29], però també a videojocs complexos. També s'utilitza en robòtica, per exemple, la conducció autònoma de cotxes o per personalitzar serveis com ara ofertes de llocs web basats en la interacció dels usuaris.



Il·lustració 4.- Esquema il·lustratiu dels diferents algorismes de ML.

2.2.5 Casos d'us de ML per a inversors

Els inversors poder fer servir algorismes de ML per a totes les fases del procés d'inversió i a continuació es fa un petit recull dels diversos casos d'us. Entre les aplicacions clau es pot incloure les següents.

- Mineria de dades[30] per identificar patrons i extreure característiques.
- Aprenentatge supervisat per generar factors de risc o *alphas*[31] i crear idees comercials.
- Agregació de senyals individuals en una estratègia.
- Assignació d'actius segons perfils de risc après per un algorisme.
- La prova i avaluació de les estratègies, fins i tot mitjançant l'ús de dades sintètiques[32].
- El refinament interactiu i automatitzat d'una estratègia mitjançant l'aprenentatge de reforçament.

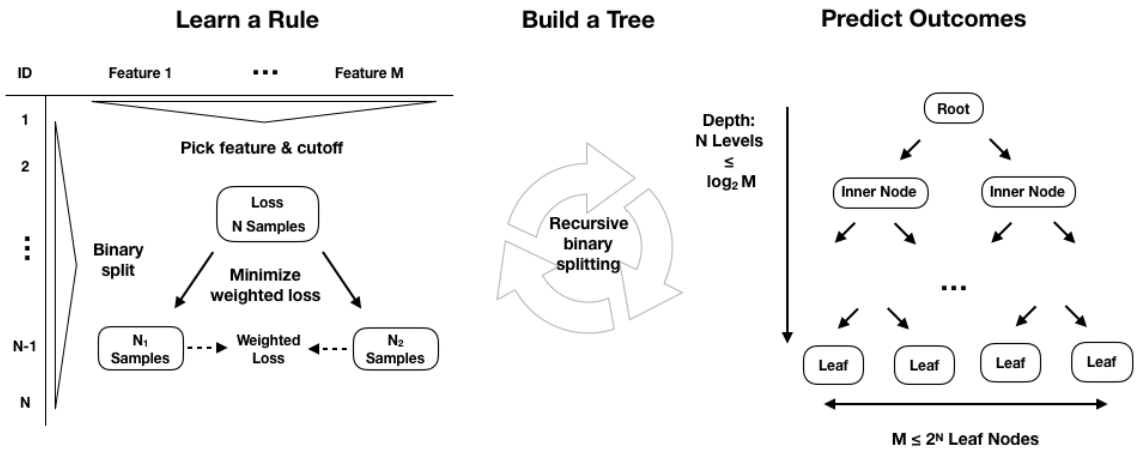
[29] https://deepmind.com/research/alphago/
[30] https://en.wikipedia.org/wiki/Data_mining
[31] https://en.wikipedia.org/wiki/Alpha_(finance)
[32] https://en.wikipedia.org/wiki/Synthetic_data

2.2.6 Arbres de decisió

Els arbres de decisió són un algorisme de ML que prediu el valor d'una variable de resultat basat en regles de decisió apreses a partir de dades d'entrenament. L'algorisme es pot aplicar a problemes de regressió i classificació canviant la funció objectiva que regeix com l'arbre aprèn les regles de decisió.

Els arbres de decisió aprenen i apliquen seqüencialment un conjunt de regles que divideix els punts de dades en subconjunts i després fan una predicció per a cada subconjunt. Les prediccions es basen en el resultat dels valors del subconjunt de mostres d'entrenament que resulten de l'aplicació d'una determinada seqüència de regles. Així, els models de classificació prediuen una probabilitat estimada a partir de les freqüències relativa de classes o del valor de la classe majoritària directament, mentre que els models de regressió calculen la predicció de la mitjana del resultat dels valors dels punts de dades disponibles.

Cadascuna d'aquestes normes es basa en una característica particular i utilitza un llindar per dividir les mostres en dos grups amb valors per sota o per sobre del llindar respecte d'aquesta funció. Un arbre binari representa, de forma natural, la lògica del model on l'arrel és el punt d'inici per a totes les mostres, els nodes representen l'aplicació de les regles de decisió i les dades es mouen al llarg de les vores, ja que es divideix en subconjunts més petits fins a arribar a un node de fulla on el model fa una predicció. El camí de l'arrel a les fulles crea una transparència sobre com funcionen les funcions i les seus valors porten a decisions específiques del model.



Il·lustració 5.- La figura mostra com el model aprèn una regla. Durant l'entrenament, l'algorisme escaneja les funcions i, per a cada funció, busca un tall que divideixi les dades per reduir al mínim la pèrdua que resulta de les prediccions fetes amb els subconjunts que resultin de la divisió, ponderada pel nombre de mostres de cada subconjunt.

Per construir un arbre sencer durant l'entrenament, l'algorisme d'aprenentatge repeteix aquest procés dividint l'espai de les funcions, és a dir, el conjunt de valors possibles per a les variables d'entrada p, X_1, X_2, ..., X_p, en regions autònomes i exhaustives, representades per cadascuna un node de fulla. Malauradament, l'algorisme no serà capaç d'avaluar tots els partició possible de l'espai de característiques donat el nombre explosiu de possibles combinacions de seqüències de característiques i llindars.

L'aprenentatge basat en l'arbre té una visió de dalt a baix, un enfocament cobdiciós, conegut com divisió binària recursiva per superar aquesta limitació computacional. Aquest procés és recursiu, ja que utilitza subconjunts de dades resultants de les particions anteriors.

És superior perquè comença al node arrel de l'arbre, on totes les observacions encara pertanyen a un una sola regió i després crea successivament dues noves branques de l'arbre afegint una divisió a l'espai predictiu.

És cobdiciós perquè l'algorisme selecciona la millor regla en la forma d'una combinació de llindar de característiques basant-se en l'impacte immediat sobre la funció objectiva, en lloc de mirar cap endavant i avaluar la pèrdua de diversos passos.

El nombre de mostres d'entrenament continua reduint-se a mesura que les separacions recursives afegeixen nous nodes l'arbre. Si les regles dividien les mostres de manera uniforme, resultant en un arbre perfectament equilibrat amb un igual nombre de fills per a cada node, llavors hi hauria 2^n nodes a nivell *n*, cadascun que conté una fracció corresponent del nombre total d'observacions.

A la pràctica, això és poc probable, de manera que el nombre de mostres al llarg d'algunes branques pot disminuir ràpidament i els arbres tendeixen a créixer a diferents nivells de profunditat al llarg de diferents camins.

Per arribar a una predicció d'una nova observació, el model utilitza les regles que va inferir durant l'entrenament per decidir quin node de fulla ha d'assignar-se al punt de dades, i després utilitza la mitjana, per a la regressió, o el mode, per a la classificació, de les observacions d 'entrenament a la regió corresponent de l'espai de funcions.

Un nombre menor de mostres d'entrenament en un una regió determinada de l 'espai característic, és a dir, en un node de fulla donat, redueix la confiança a la zona predicció i pot reflectir l'excedència.

La separació recursiva continuaria fins que cada node de fulla només contingui una única mostra i l'error d'entrenament s'hagi reduït a zero.

Els arbres de decisió no només es poden visualitzar per inspeccionar el camí de decisió d'una característica determinada, però també proporcionen una mesura resumida de l'aportació de cada funció al model adequat les dades d'entrenament.

La importància de la funció captura la quantitat de partides produïdes per la funció optimitzen la mètrica del model utilitzada per avaluar la qualitat de la divisió.

La importància d'una característica es calcula com la reducció total normalitzada d'una mètrica i té en compte el nombre de mostres afectades per una divisió.

Els arbres de decisió tenen una forta tendència a sobreentrenar-se, sobretot quan un conjunt de dades té una gran quantitat nombre de característiques relatives al nombre de mostres. Aquest fet, produeix un augment de l'error de predicció perquè el model no només aprèn el senyal continguts a les dades d'entrenament, però també el soroll. Hi ha diverses maneres d'afrontar el risc d'excedència, com es mostra a continuació.

- La reducció de la dimensionalitat millora la relació de característiques-mostres representant les funcions existents amb menys característiques, més informatives i menys sorolloses.
- *Ensemble models*[33], com ara els *random forest*[34], combinen diversos arbres mentre que aleatoritzen la construcció d'arbres.
- Els arbres de decisió ofereixen diversos hiperparàmetres de regularització[35] per limitar el creixement d'un arbre i la complexitat associada. Mentre cada ruptura augmenta el nombre de nodes, també redueix el nombre de mostres disponibles per node per donar suport a una predicció. Per a cada nivell addicional, és el doble del nombre de mostres necessitat per poblar els nous nodes amb la mateixa densitat de mostra.
- *Tree-pruning*[36] és una eina per reduir la complexitat d'un arbre mitjançant la eliminació de nodes o parts senceres d'un arbre que aportin poc valor, però que augmentin la variància del model. La poda de complexitat de costos, per exemple, comença amb un arbre gran i redueix de forma recursiva la seva mida substituint essencialment els nodes amb fulles executant la construcció d'arbre al revés. Els diferents passos produeixen una seqüència d'arbres que es poden comparar mitjançant la validació creuada per seleccionar la mida ideal.

2.2.7 Maquines de potenciació del gradient

Existeixen uns algorismes de ML, anomenats GBM, per a conjunts d'arbres de decisió que sovint produeixen els millors resultats. La diferència clau és aquesta potenciació que modifica les dades que s'utilitzen per formar cada arbre en funció de l'acumulació errors produïts pel model abans d'afegir el nou arbre. A diferència dels *random forests,* que entrena molts arbres independentment l'un de l'altre utilitzant diferents versions del conjunt de dades d'entrenament, augmenta el rendiment seqüencialment utilitzant versions de ponderació de les dades.

L'estat de l'art d'aquests tipus d'algorismes també adopta les estratègies aleatorització de *random forests.*

Aquests algorismes, són uns dels de ML amb més èxit durant els últims anys. Per exemple, recentment dominen competències de ML per a dades estructurades.

El potenciador és un algorisme d'aprenentatge conjunt que combina la base d'entrenadors, generalment arbres de decisió, en un conjunt.

[33] https://en.wikipedia.org/wiki/Ensemble_learning
[34] https://es.wikipedia.org/wiki/Random_forest
[35] https://codeburst.io/what-is-regularization-in-machine-learning-aed5a1c36590
[36] https://en.wikipedia.org/wiki/Decision_tree_pruning

La potenciació es va desenvolupar inicialment per a problemes de classificació, però també es pot utilitzar per a la regressió i s'ha anomenat un dels les idees d 'aprenentatge més potents introduïdes en els últims anys. És un mètode general o de modificació que es pot aplicar a molts models d'aprenentatges estadístic.

La motivació per al desenvolupament dels potenciadors va ser trobar un mètode per combinar les sortides de molts models febles, un predictor es diu feble quan es realitza una mica millor que endevinar a l'atzar, en una predicció conjunta més potent, és a dir, potenciada. En general, l'augment aprèn una hipòtesi additiva, H_M, d'una forma similar a la regressió lineal. No obstant això, ara cadascun dels elements *m = 1, ..., M* de la suma és un aprenent de base feble, anomenada h_t que necessita formació.

$$H_M(x) = \sum_{m=1}^{M} h_t(x)$$

Els potenciadors procedeixen seqüencialment mitjançant la formació de la base d'entrenadors sobre les dades que es modifiquen repetidament per reflectir els resultats d'aprenentatge acumulats. L'objectiu és assegurar que el següent aprenentatge bàsic compensi les deficiències actuals del conjunt, ja que el conjunt fa prediccions utilitzant una mitjana ponderada de les prediccions dels models febles.

El primer algoritme de potenciació va venir amb una prova matemàtica que millora el rendiment dels aprenents febles i la van desenvolupar Robert Schapire i Yoav Freund l'any 1990. El 1997 es va crear una solució pràctica per als problemes de classificació de l'algorisme de potenciació adaptativa, AdaBoost[37], que va guanyar el premi Göedel el 2003.

Cinc anys més tard, aquest algorisme es va estendre a funcions d'objectius arbitraris quan Leo Breiman, que va inventar *random forests,* va connectar l'enfocament de descens del gradient, i Jerome Friedman va sorgir amb els potenciadors del gradient el 1999. Són nombroses les implementacions d'aquest optimitzades que acceleren l'entrenament, milloren l'eficiència dels recursos i permeten l'algorisme escalar a conjunts de dades molt grans, com ara XGBoost[38], LightGBM i CatBoost, que han aparegut en els darrers anys i s'han establert amb força com a solució per a dades estructurades.

Els GBM han demostrat un rendiment en molts reptes de classificació i regressió. Probablement siguin els algorisme d'aprenentatge conjunt més populars com a predictor independent en un conjunt divers de competicions de ML, així com en produccions del món real, per exemple, per predir tarifes de clics per als anuncis en línia.

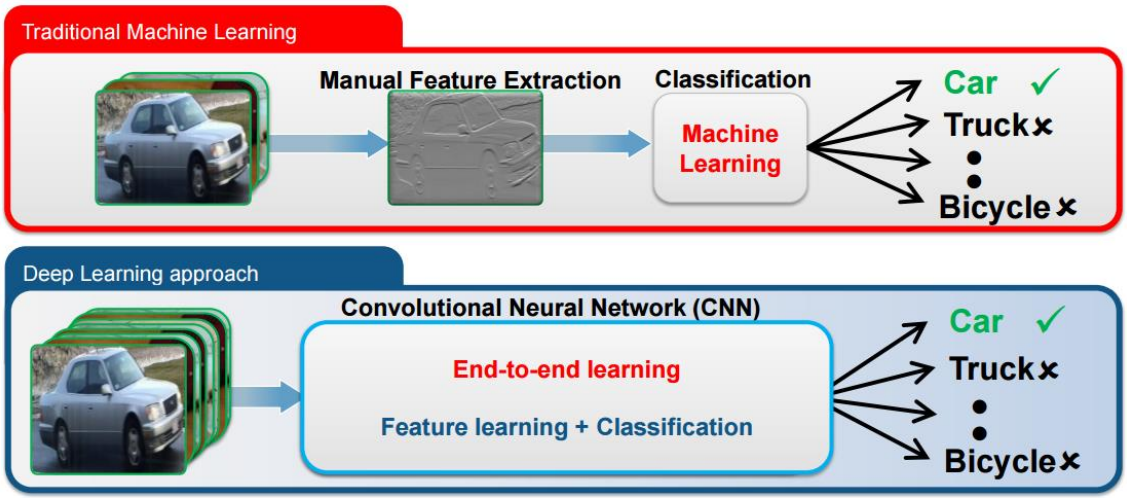
L'èxit de GBM es basa en la seva capacitat d'aprendre de forma complexa i funcional les relacions de manera incremental. La flexibilitat d'aquest algorisme requereix la gestió del risc de sobreentrenament mitjançant l'ajustament d'hiperparàmetres que limiten la tendència del model inherent a aprendre el soroll, en detriment del senyal de les dades d'entrenament.

[37] https://en.wikipedia.org/wiki/AdaBoost
[38] https://en.wikipedia.org/wiki/XGBoost

2.2.8 *Deep Learning*

El DL és una tècnica d'aprenentatge automàtic de extrem a extrem. Això, fa que sigui molt atractiu per a moltes aplicacions. Una de les més famoses, ja que ha aconseguit treure millors resultats que els tradicionals algoritmes de ML, és a la classificació d'imatges.

El DL té una avantatge clara i és que intenta aprendre a trobar característiques en les dades en brut, és a dir, sense processar per un expert, fent automàticament l'extracció de característiques i classificant-les directament a partir de les imatges, alhora.



Il·lustració 6.- Exemple amb un esquema comparatiu de les tècniques tradicionals d'ML i l'enfoc a DL, en una tasca de classificació d'imatges.

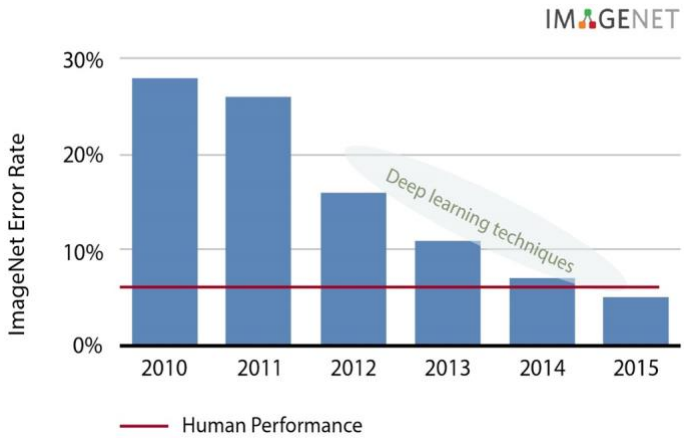
Per tant, un dels avantatges és que no cal ser un expert en entendre les dades, fet que estalvia molt de temps i ho fa accessible a més gent, i a més té la capacitat d'abordar representacions més complexes en imatges.

L'altre avantatge principal és l'elevat nivell de detall que aconsegueix, que no es podria fer amb enfocaments tradicionals d'ML i, per tant, és un dels motius pels quals el DL és tan popular.

Per exemple, un dels reptes de classificació d'imatges més famosos és l'ILSVRC[39] on hi ha una competència entre investigadors per classificar correctament el conjunt de dades IMAGENET, que el 2010 consistia en més de 15 milions d'imatges d'alta resolució etiquetades pertanyents a aproximadament 22.000 categories.

Així, abans del 2012, els millors resultats eren al voltant del 25% utilitzant tècniques tradicionals d'ML. Però l'any 2012, la coneguda xarxa neuronal AlexNet [1] va aconseguir el millor resultat en aquesta competició i, des de llavors, s'han aconseguit els millors resultats mitjançant tècniques de DL, superant els resultats de l'home, o experts.

[39] http://www.image-net.org/challenges/LSVRC/



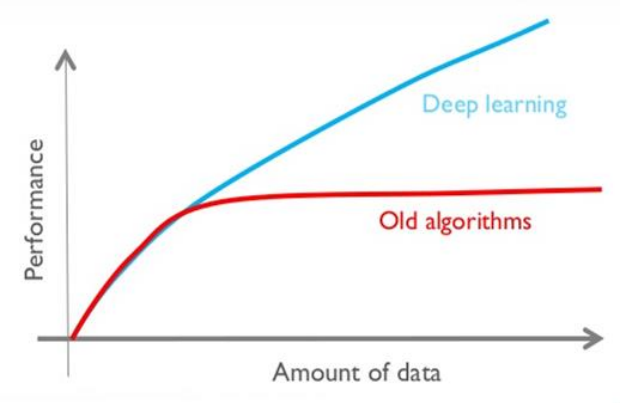
Il·lustració 7.- Resultats de l'ILSVRC des de 2010. Les tècniques d'aprenentatge profund aconsegueixen els millors resultats des del 2012.

En segon lloc, el primer punt clau que va permetre aquest increment massiu dels resultats en la competició va ser l'aparició de biblioteques i algorismes de DL, a causa de l'augment de la investigació en aquest camp. Anteriorment, les investigacions havien de fer-se amb programació de baix nivell.

No obstant això, un gran desavantatge és que amb les tècniques de ML tradicionals hi ha més flexibilitat per a l'elecció de l'extracció de característiques i els classificadors, que permeten utilitzar diferents i triar el que té els millors resultats. A més, si es comprenen les dades és possible obtenir una bona precisió amb una quantitat mínima de dades.

En canvi, els models d'DL necessiten moltes dades, la qual cosa significa també un gran cost computacional i temps per entrenar. A més, com que és una solució de caixa negra, si apareixen problemes durant l'entrenament del model, és molt difícil depurar-lo i trobar la font del problema.

Hi ha un altre enfocament basat en una combinació de les dues tècniques. Amb el DL fent l'extracció de característiques i les tècniques de ML per a la classificació, eliminant les últimes capes de la xarxa neuronal[40].



Il·lustració 8.- Augment del rendiment de DL respecte els algoritmes de ML tradicionals segons la quantitat de dades.

[40] https://en.wikipedia.org/wiki/Deep_learning#Deep_neural_networks

En tercer lloc, Un altre punt clau que va permetre el DL aconseguir una alta precisió va ser el desenvolupament de noves arquitectures informàtiques en forma de GPUs[41] que poden formar xarxes amb quantitats massives de dades. Aquestes són molt útils per a la computació paral·lela i, sobretot, per als models de DL que aprofiten enormement aquesta massiva paral·lelització de les GPU a causa de la gran quantitat de nuclis que tenen. Aquestes dispositius HW es van utilitzar històricament per a gràfics de jocs, però cada cop més, durant els darrers anys, s'han desenvolupat GPUs només per a la computació paral·lela.

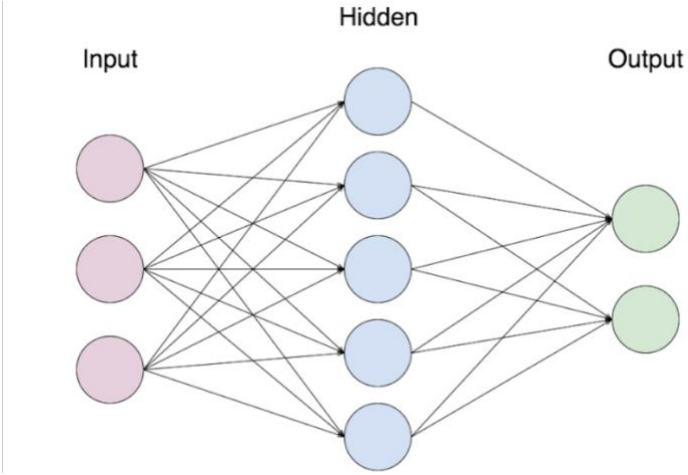
Així, des de la aparició d'aquests dispositius, el DL es pot realitzar amb més potència i rendibilitat, especialment en temps, i aconseguir grans resultats.

Finalment, hi ha un altre punt clau d'aquests resultats que va ser la disponibilitat, en els darrers anys, de quantitats de conjunts de dades en línia, per exemple dades etiquetades, per a l'entrenament de les xarxes, que necessiten una gran quantitat de dades per aconseguir bons resultats,. En part a causa de l'augment de la quantitat de persones que carreguen dades en línia a diari.

2.2.8.1 Xarxes Neuronals

El DL es basa en models de xarxes neuronals artificials. Aquestes són un model matemàtic inspirat en xarxes neuronals biològiques i la manera en què els sistemes nerviosos biològics processen la informació. Estan compostos per un gran nombre d'elements interconnectats entre ells. Aquests elements es denominen neurones i representen la unitat computacional bàsica de les xarxes neuronals biològiques.

L'arquitectura tradicional d'aquestes xarxes està composta per una capa d'entrada, a través de la qual s'alimenten les entrades, i una capa de sortida, capa que retorna les sortides esperades, i, a continuació, moltes capes ocultes entre elles. Cada neurona està connectada a totes les neurones de la següent capa i aquestes capes són tradicionalment conegudes com a capes completament connectades, però cada neurona no està connectada a les neurones de la mateixa capa i els bucles no estan permesos.



Il·lustració 9.- Il·lustració de l'arquitectura de xarxes neuronals artificials. Aquest només té una capa oculta, de manera que s'anomena una xarxa neuronal de capa oculta o una xarxa neuronal de dues capes, ja que només es compten les capes que tenen pesos.

[41] https://en.wikipedia.org/wiki/Graphics_processing_unit

La raó per la qual les neurones s'organitzen en capes és que agrupar neurones en vectors fa que el cost computacional sigui més baix que si s'utilitzen operacions de vectors i matrius. Això és molt important perquè fa que la formació sigui molt efectiva quan s'utilitza la computació paral·lela, per exemple, en GPUs com s'ha comentat. A més, un altre factor important de les xarxes neuronals és el seu poder de representació.

Per acabar, l'objectiu principal de l'entrenament de la xarxa és trobar els pesos òptims per tenir un millor rendiment, i per fer-ho, l'entrenament intenta minimitzar una funció de pèrdua, o funció de cost, que mesurar la bona relació entre les sortides previstes i les etiquetes de veritat. De certa manera quantifica quan de bo que és cada pes en el conjunt de dades.

Per tant, si la funció de pèrdua és alta, el rendiment del model serà pobre. Aquest és un problema d'optimització, concretament de com començar amb pesos aleatoris i trobar els pesos òptims que minimitzen la funció de pèrdua de manera eficient.

2.2.9 Treballant amb dades de text

Les dades de text són molt riques en contingut, però no estructurades en format i, per tant, requereixen més processament per tal que un algoritme ML pugui extreure el senyal potencial.

El repte clau consisteix a convertir el text en un format numèric per utilitzar-lo per un algoritme, mentre que expressant simultàniament la semàntica o el significat del contingut.

Hi ha diverses tècniques que capturen matisos del llenguatge que són fàcilment comprensibles per als humans que poden convertir-se en una entrada per a algorismes ML.

Les tècniques fonamentals d'extracció de trets o característiques concentren unitats semàntiques individuals, és a dir, paraules o grups breus de paraules anomenats *tokens*.

Les dades de text poden ser molt valuoses tenint en compte la quantitat d'informació que els humans comuniquen i emmagatzemen utilitzant el llenguatge natural. El conjunt divers de fonts de dades rellevants per a la inversió varia des de documents formals, com ara declaracions d'empreses, contractes i patents, fins a notícies, opinió i investigació d'analistes, i fins i tot de comentaris i diversos tipus de xarxes socials missatges i missatges.

Hi ha nombroses i diverses mostres de dades de text disponibles en línia per explorar l'ús de algorismes NLP.

La conversió de text no estructurat en un format llegible per màquina requereix un processament previ per preservar alguns valuosos aspectes semàntics de les dades.

Com els humans extreuen el significat, o comprendre el contingut del llenguatge, no està completament entès pels algorismes de NLP, de fet la millora de la comprensió lingüística per part de les màquines continua sent un àmbit de recerca molt activa.

La NLP és un repte perquè l'ús eficaç de les dades de text per a ML requereix una comprensió dels mecanismes interns del llenguatge i del coneixement del món al qual es refereix.

Els reptes als que s'enfronta aquest àmbit de cerca, inclouen els següents entre d'altres.

- Ambigüitat per polisèmia; és a dir, una paraula o frase pot tenir diferents significats depenent del context.
- Ús no estàndard i evolutiu del llenguatge, especialment en les xarxes socials.
- Ús d'expressions, com tirar la tovallola.
- Noms d'entitats complicades.
- El coneixement del món.

Moltes aplicacions de NLP aprenen a predir els resultats de la informació significativa extreta del text. L'aprenentatge supervisat requereix etiquetes per ensenyar l'algorisme la relació real entre l'entrada i la sortida. Amb les dades de text, l'establiment d'aquesta relació pot ser un repte i potser requereix un model i una recopilació de dades explícits.

Les decisions de modelització de dades inclouen com quantificar els sentiments implícits en un document de text com un correu electrònic, una entrevista transcrit o un tweet o quins aspectes d'un document de recerca o notícies per assignar a un resultat concret.

L'ús de ML amb dades de text per a la la inversió es basa en l'extracció de informació significativa en forma de funcions que prediuen directament o indirectament futurs moviments del preu. Les aplicacions van des de l'explotació de l'impacte del mercat a curt termini a notícies sobre l'anàlisi fonamental a llarg termini dels motors de valoració d'actius. Per exemple els següents.

- *Sentiment Anàlisis* de critiques de productes per avaluar la competitivitat d'una empresa o posició a indústria.
- La detecció d'anomalies en els contractes de crèdit per predir la probabilitat o l'impacte d'un valor predeterminat.
- La predicció de l'impacte de les notícies sobre les empreses o els mercats.

JP Morgan, per exemple, va desenvolupar un model predictiu basat en 250.000 informes d'analistes van superar diversos índexs de referència[42].

2.2.9.1 Sentiment anàlisis

L'anàlisi del sentiment és un dels usos més populars de la NLP i ML per a la inversió, ja que és probable que les perspectives positives o negatives sobre els actius o altres, tinguin un impacte important en el preu.

En general, els enfocaments de modelització de l'anàlisi del sentiment es basen en diccionaris, com la biblioteca *TextBlob*[43], o models formats per obtenir resultats en un domini específic. L'últim és preferible perquè permet un etiquetatge més específic; per exemple, vinculant funcions de text a canvis de preus posteriors en lloc de puntuacions de sentiment indirectes.

[42] https://www.jpmorgan.com/global/research/machine-learning
[43] https://textblob.readthedocs.io/en/dev/

2.2.10 BERT

Recentment, un investigadors de Google AI Devli et al.(2008) [2] van presentar BERT, un model de xarxa neuronal entrenat de representacions del llenguatge. És a dir, és un model de "comprensió lingüística" de propòsit general generat des d'un text ampli, i després es fa servir aquest model per a les tasques de NLP.

BERT supera els mètodes anteriors perquè és el primer sistema d'aprenentatge no supervisat i bidireccional per a la formació d'un model de xarxa neuronal entrenat de NLP. Es diu que és un model no supervisat, ja que el model BERT només s'ha format amb un text simple[44]. Aquest fet, és important perquè una gran quantitat de dades de text simple està disponible a internet de manera pública i en molts idiomes.

BERT fa ús del model *Transformer*, presentat per Vaswani et al.(2007) [3], un mecanisme d'atenció que aprèn relacions contextuals entre paraules, o sub-paraules en un text. Aquest inclou dos mecanismes separats.

- Un codificador que llegeix l'entrada de text.
- Un descodificador que produeix una predicció de la tasca.

Atès que l'objectiu de BERT és generar un model de llenguatge, només és necessari el mecanisme de codificació.

A diferència dels models direccionals, que llegeixen l'entrada de text seqüencialment (d'esquerra a dreta o de dreta a esquerra), el codificador de *Transformer* llegeix tota la seqüència de paraules alhora.

Per tant, normalment es considera bidireccional, encara d'altres creuen que seria més correcte dir que no és direccional. Aquesta característica permet al model aprendre el context d'una paraula basada en tot l'entorn (esquerra i dreta de la paraula).

La il·lustració següent és una descripció d'alt nivell del codificador de *Transformer*. L'entrada és una seqüència de paraules o *tokens*, que es passen primerament als vectors i es processen a la xarxa neuronal. La sortida és una seqüència de vectors de mida H, en la qual cada vector correspon a un *token* d'entrada amb el mateix índex.

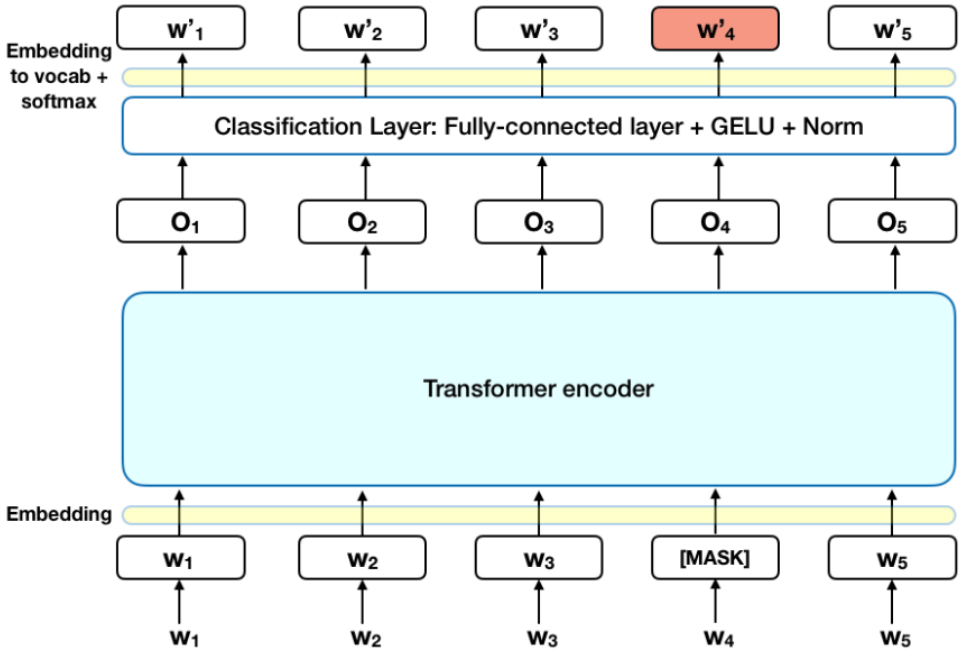
Cal fer esment, a que quan es formen models lingüístics, hi ha un repte de definir un objectiu de predicció. Molts models prediuen la següent paraula en una seqüència, per exemple, "El nen va venir a casa de ___", un enfocament direccional que limita inherentment l'aprenentatge del context. Per superar aquest repte, BERT utilitza dues estratègies de formació.

[44] https://en.wikipedia.org/wiki/Plain_text

2.2.10.1 Emmascarament

Abans d'alimentar el BERT amb les seqüències de paraules, el 15% de les paraules de cada seqüència són substituïdes per un *token*, o màscara. Aleshores, el model intenta predir el valor original de les paraules emmascarades, basant-se en el context proporcionat per les altres paraules no emmascarades de la seqüència. En termes tècnics, la predicció de les paraules de sortida requereix tres etapes:

1. Afegir una capa de classificació a la part superior de la sortida del codificador.
2. Multiplicant els vectors de sortida per una matriu, transformant-los en la dimensió del vocabulari.
3. Calcular la probabilitat de cada paraula en el vocabulari amb *softmax*[45].



Il·lustració 10.- Arquitectura del model BERT, que és sobretot una arquitectura estàndard de *Transformer.*

2.2.10.2 Predicció sentència següent

En el procés de formació del BERT, el model rep parells de sentencies com a entrada i aprèn a predir si la segona frase del parell és la frase posterior del document original. Durant l'entrenament, la meitat de les entrades són un parell en què la segona frase és la frase posterior del document original, mentre que en l'altre meitat es tria una frase aleatòria del corpus com a segona frase. La hipòtesi és que la frase aleatòria es desconnectarà de la primera frase.

Per exemple, la primera meitat esta format per sentencies com la següent.

- Sentència A: El nen va anar a la fleca.
- Sentència B: a comprar una barra de pa.
- Etiqueta: *ÉsLaSegüentSentència*.

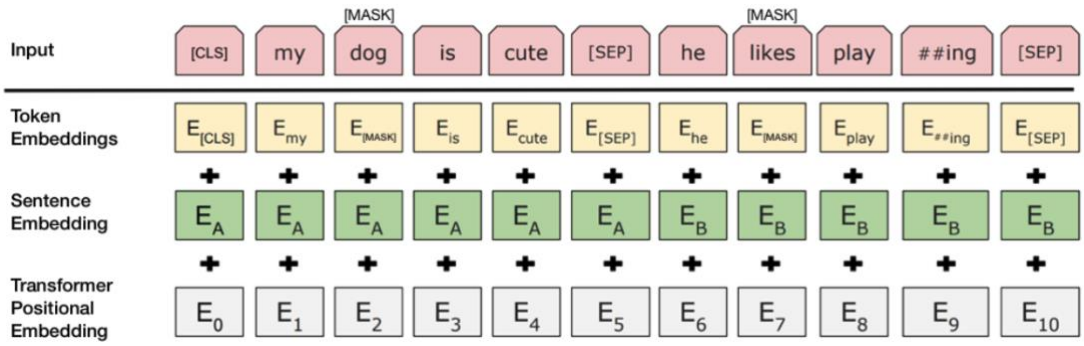
[45] https://en.wikipedia.org/wiki/Softmax_function

En canvi, l'altre esta formada per el que es mostra a continuació.

- Sentència A: El nen va anar a la fleca.
- Sentència B: tenen una pell marró groguenca.
- Etiqueta: *NoÉsLaSegüentSentència*.

Per ajudar a que el model distingeixi entre les dues frases durant l'entrenament, les dades d'entrada es processen de la manera següent abans ser introduïdes al model:

1. S'introdueix un *token* [CLS] al principi de la primera frase i un altre *token* [SEP] al final de cada frase.
2. S'afegeix una frase que indica la frase A o la frase B a cada token.
3. S'afegeix una de posició a cada token per indicar la seva posició en la seqüència. El concepte i la implementació de la inserció posicional són presenten al *paper* de *Transformer*, referenciat més amunt.



Il·lustració 11.- Representació d'entrada del BERT. Les incorporacions d'entrada són la suma de la incorporació de *tokens*, els de la classificació entre sentències i els de posició.

D'aquesta manera, per predir si la segona frase està realment connectada a la primera, es realitzen els següents passos:

1. Tota la seqüència d'entrada passa pel model de *Transformer*.
2. La sortida del testimoni [CLS] es transforma en un vector [2,1], utilitzant una capa de classificació simple, és a dir, matrius d'aprenentatge de pesos i biaixos.
3. Càlcul de la probabilitat de ÉsLaSegüentSentència amb *softmax*.

Així, quan es forma el model BERT, es formen junts l'emmascarament LM i la predicció de sentències següents, amb l'objectiu de minimitzar la funció de pèrdua combinada de les dues estratègies. En segon lloc, l'ús del model BERT es pot dividir en dues etapes. La primera és l'entrenament del model, un procés computacional molt costos, i ja elaborat pels investigadors de Google. Actualment, el model BERT només està entrenat en l'idioma anglès. Tot i que, hi ha intenció d'entrenar-lo en més idiomes. Per altra banda, la afinació o *fine-tuning*[46], és adaptar el model a una tasca en concret, fent que torni a entrenar a més del entrenament previ, ja realitzat pels investigadors. D'aquesta manera, és molt menys costos i ràpid. Per això, és un aspecte important de BERT, ja que es pot adaptar fàcilment a molts tipus de tasques de NPL.

[46] https://en.wikipedia.org/wiki/Fine-tuning

3 Estat de l'art

En aquest apartat, es resumeix les diferents tècniques de ML que s'han fet servir en el món dels mercats financers.

En primer lloc, s'han desenvolupat moltes sistemes per predir tendències en les accions. Al principi, es van fer servir mètodes de regressió clàssica per predir les tendències. Ja que les dades d'accions poden classificar-se com dades de sèries temporals no estacionàries, també s'han utilitzat tècniques no lineals d'aprenentatge automàtic. Les ANN i SVM són dos algoritmes de ML que són els més utilitzats per predir el moviment i l'índex de preus de les accions. Cada algoritme té la seva pròpia manera d'aprendre patrons. ANN emula el funcionament del nostre cervell per aprendre mitjançant la creació d'una xarxa de neurones. Hassan et al. (2007) [4] van proposar i implementar un model de fusió combinant HMM, ANN i els AG per predir el comportament del mercat financer.

Fent servir ANN, els preus de les accions diàries es van transformar en conjunts independents de valors que es van fer servir com a *input* a HMM. Wang i Leu (1996) [5] on van desenvolupar un sistema de predicció útil per pronosticar la tendència dels preus a mitjà termini en el mercat de valors de Taiwan. El seu sistema es va basar en una xarxa neuronal recurrent entrenada mitjançant l'ús de característiques extretes dels anàlisis ARIMA[47]. Els resultats empírics van mostrar que les xarxes entrenades utilitzant dades setmanals de 4 anys van ser capaços de predir fins a 6 setmanes de tendència del mercat amb una precisió acceptable.

Abraham et al. (2001) [6] van introduir tècniques de computació híbrid per a la predicció automatitzada del mercat de valors i l'anàlisi de tendències. Per analitzar la tendència dels valors de les existències previstes, van utilitzar l'índex NASDAQ-100 de les accions del mercat NASDAQ[48] amb xarxa neuronal amb un dia d'anticipació de la previsió d'estocs i un sistema *neuro-fuzzy*[49]. Els resultats de predicció i predicció de tendències utilitzant el sistema híbrid proposat van ser prometedors.

Chen, Leung i Daouk (2003) [7] van investigar la PNN per predir la direcció de l'índex després que fos entrenat amb dades històriques. Els resultats empírics van mostrar que les estratègies d'inversió basades en PNN van obtenir més retorns que altres estratègies d'inversió examinades en l'estudi, com l'estratègia de compra i retenció, així com les estratègies d'inversió guiades pels pronòstics estimats pel model de *random walk*[50] i els models paramètrics de GMM.

[47] https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average
[48] https://es.wikipedia.org/wiki/NASDAQ
[49] https://en.wikipedia.org/wiki/Neuro-fuzzy
[50] https://en.wikipedia.org/wiki/Random_walk

En segon lloc, SVM és un tipus molt específic d'algoritmes d'aprenentatge caracteritzats pel control de capacitat de la funció de decisió, l'ús de les funcions del nucli i l'escassetat de la solució. Un molt conegut desenvolupat per Vapnik (1999) [8] busca un híper-pla en una dimensió superior a classes separades. Huang, Nakamori i Wang (2005) [9] van investigar la predicció de la direcció del moviment financer amb SVM mitjançant la predicció de la direcció de moviment setmanal de l'índex NIKKEI 225. Van comparar SVM amb ADL, QDA i xarxes neuronals de *back-propagation*[51]. Els resultats de l'experiment van mostrar que els SVM van superar els altres mètodes de classificació. Kim (2003) [10] va utilitzar SVM per predir la direcció de la variació diària del preu de les accions en l'índex de preus de les accions compost de Corea (KOSPI[52]). Es van seleccionar 12 indicadors tècnics[53] per constituir els atributs inicials. Aquest estudi va comparar la SVM amb xarxes neuronals de *back-propagation* i el CBR. A partir dels resultats experimentals SVM va superar BPN i CBR.

Ou i Wang (2009) [11] van utilitzar 10 tècniques de *data mining*[54] per predir el moviment dels preus de l'índex HSI del mercat de valors de Hong Kong. L'enfoc utilitzat inclou LDA, QDA, la classificació de veïns més propers a K[55], el *naive-bayes*[56], model *Logit*[57], classificació basada en arbres de decisió, xarxa neuronal, classificació bayesiana amb procés gaussià, SVM i LS-SVM[58]. Els resultats experimentals van mostrar que el SVM i LS-SVM van generar un rendiment predictiu superior entre els altres models.

En tercer lloc, l'algoritme r*andom forest* crea n arbres de classificació usant la mostra amb reemplaçament i prediu la classe basant-se en el que prediu la majoria dels arbres. El conjunt entrenat, per tant, representa una sola hipòtesi. Aquesta, no està necessàriament continguda dins de l'espai d'hipòtesis dels models dels que es constitueix. Així, es pot demostrar que els conjunts tenen més flexibilitat en les funcions que poden representar. D'aquesta manera, en teoria permet superposar les dades d'entrenament més que un sol model, però a la pràctica, algunes tècniques de conjunt tendeixen a reduir els problemes relacionats amb l'ajust excessiu de les dades d'entrenament.

Tsai et al. (2011) [12] van investigar un mètode del assemblador classificador[59] per analitzar els rendiments de les accions. El rendiment utilitzant dos tipus d'assemblador classificadors es van comparar amb els que utilitzen única línia de base classificadors, per exemple, xarxes neuronals, arbres de decisió, i la regressió logística. Els resultats van indicar que els classificadors múltiples van superar els classificadors individuals en termes d'exactitud de la predicció i els retorns de la inversió.

[51] https://en.wikipedia.org/wiki/Backpropagation
[52] https://es.wikipedia.org/wiki/KOSPI
[53] https://en.wikipedia.org/wiki/Technical_analysis
[54] https://en.wikipedia.org/wiki/Data_mining
[55] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
[56] https://en.wikipedia.org/wiki/Naive_Bayes_classifier
[57] https://en.wikipedia.org/wiki/Logit
[58] https://en.wikipedia.org/wiki/Least-squares_support-vector_machine
[59] https://en.wikipedia.org/wiki/Ensemble_learning

Sun i Li (2012) [13] van proposar el nou mètode de FDP basat en el conjunt SVM. L'algoritme va ser dissenyat tenint en compte tant el rendiment individual com l'anàlisi de diversitat. Els resultats experimentals van concloure que el conjunt SVM era significativament superior al classificador SVM individual.

En quart lloc, una arquitectura en dues etapes va ser desenvolupada per Hsu et al. (2009) [14], on van integrar el SOM[60] i la SVR per a la predicció del preu de les accions. Van examinar set importants índexs borsaris. En concret, es va utilitzar per primera vegada el SOM per descompondre tot l'espai d'entrada en regions en què s'agrupen punts de dades amb distribucions estadístiques similars per contenir i capturar la propietat no estacionària de sèries financeres. Després de la descomposició de dades heterogènies en diverses regions homogènies, SVR es va aplicar per pronosticar els índexs financers. Els resultats van suggerir que l'arquitectura de dues etapes era una alternativa prometedora per a la predicció del preu de les accions.

Per altre banda, La GP[61] i les seves variants s'han aplicat extensivament per a la modelització dels mercats de valors. Per millorar la capacitat de generalització del model, els GP s'han dissenyat models híbrids amb les seves pròpies variants GEP[62], MEP o ANN. La capacitat de generalització del model GP també pot millorar-se mitjançant una selecció apropiada del criteri de selecció del model.

Garg, Sriram i Tai (2013) [15] van treballar per analitzar l'efecte de tres criteris diferents de selecció de model a través de dues transformacions de dades sobre l'efecte de GP a la Borsa de Nova York. Es va trobar que els criteris de FPE havien mostrat un major ajust per al model GP en les dues transformacions de dades en comparació amb altres criteris de selecció de models.

Nair et al. (2011) [16] va predir el valor de tancament del dia següent de cinc índexs borsaris internacionals utilitzant un sistema de xarxes neuronals artificials adaptatives. El sistema es va adaptar tar a la dinàmica canviant del mercat amb l'ajuda de un AG que va sintonitzar els paràmetres de la xarxa neuronal al final de cada sessió de negociació.

L'estudi d'Ahmed (2008) [17] va investigar la naturalesa de les relacions causals entre els preus de les accions i les principals variables macroeconòmiques que representen el sector real i financer de l'economia de l'Índia per al període de març de 1995 a 2007 utilitzant dades trimestrals . L'estudi va revelar que el moviment dels preus de les accions no només va ser el resultat del comportament de les principals variables macroeconòmiques, sinó que també va ser una de les causes del moviment en altres dimensions macroeconòmiques.

[60] https://en.wikipedia.org/wiki/Self-organizing_map
[61] https://en.wikipedia.org/wiki/Genetic_programming
[62] https://en.wikipedia.org/wiki/Gene_expression_programming

Mantri, Gahan i Nayak (2010) [18] van calcular les volatilitats dels mercats borsaris indis utilitzant els models GARCH[63], EGARCH, GJR-GARCH, IGARCH i ANN. Aquest estudi va utilitzar catorze anys de dades de BSE Sensex[64] & NSE[65] per calcular les volatilitats. Es va concloure que no hi va haver diferències en les volatilitats de Sensex i Nifty[66] estimades sota els models citats prèviament.

Mishra, Sehgal i Bhanumurthy (2011) [19] van posar a prova la presència de dependència no lineal[67] i del caos determinista[68] en la sèrie de taxes de rendiment de sis índexs borsaris indis. El resultat de l'anàlisi va suggerir que la sèrie de retorns no va seguir un procés *random walk*. Més aviat semblava que els increments diaris en els rendiments de les accions estaven correlacionats[69] en sèrie i els exponents de Hurst[70] estimats eren indicatius de la persistència marginal en els guanys de capital.

Liu i Wang (2012) [20] van investigar i van pronosticar la fluctuació de preus per una ANN millorada de Legendre[71] assumint que els inversors van decidir les seves posicions d'inversió analitzant les dades històriques en el mercat de valors. També van introduir una funció de resistència al temps aleatòria en el model de pronòstic. Araújo i Ferreira (2013) [21] van proposar el EMRL i van realitzar una anàlisi experimental, comparant els resultats amb les xarxes de MLP i el mètode de TAEF.

Finalment, a partir de les discussions prèvies, es pot deduir que cadascun dels algoritmes pot abordar aquest problema a la seva manera. I alhora, que tenen les seves pròpies limitacions. El resultat de la predicció final no només depèn de l'algoritme de predicció utilitzat, sinó que també està influenciat per les dades usades per aquest. La identificació de característiques rellevants i un ús adequat de les mateixes poden millorar la precisió dels models de predicció.

[63] https://en.wikipedia.org/wiki/Autoregressive_conditional_heteroskedasticity
[64] https://en.wikipedia.org/wiki/BSE_SENSEX
[65] https://en.wikipedia.org/wiki/National_Stock_Exchange_of_India
[66] https://en.wikipedia.org/wiki/NIFTY_50
[67] https://en.wikipedia.org/wiki/Linear_independence
[68] https://en.wikipedia.org/wiki/Chaos_theory
[69] https://en.wikipedia.org/wiki/Correlation_and_dependence
[70] https://en.wikipedia.org/wiki/Hurst_exponent
[71] https://en.wikipedia.org/wiki/Legendre_polynomials

4 Coneixements financers

El sistema d'intel·ligència artificial que es pretén dissenyar farà servir la mateixa informació de la qual disposen la majoria d'inversors en borsa. Si bé hi ha una petita part que disposen d'informació privilegiada, i que a més, en cas que qui disposi d'aquesta tingui un gran capital per invertir, pot condicionar directament el moviment de les cotitzacions. Les dues eines més utilitzades per prendre les decisions de compra i de venda són l'anàlisi fonamental[72] i l'anàlisi tècnic[73].

- L'anàlisi fonamental és analitzar el negoci de l'empresa, el seu balanç, el compte de guanys i pèrdues, les seves perspectives de futur, les seves barreres d'entrada, etc.
- L'anàlisi tècnic es limita a estudiar les cotitzacions històriques de l'empresa, sense entrar a valorar en cap moment la marxa del negoci de l'empresa. A un analista tècnic pur li és igual que l'empresa que està analitzant vengui sabates o que sigui un banc o una petroliera, que guanyi diners o que el perdi, que porti 200 anys repartint dividend o que mai hagi tingut beneficis. Tampoc li importa si les seves perspectives són bones o dolentes, ni si té un determinat problema en aquests moments o deixa de tenir-lo. En base als gràfics que hagin dibuixat les cotitzacions passades, exclusivament, l'anàlisi tècnic intenta determinar què és el que faran les cotitzacions en el futur.

Sovint es veu a aquests dos tipus d'anàlisi, el fonamental i el tècnic, com a coses totalment oposades[74], i els defensors de tots dos discuteixen sobre la inutilitat de l'altre. Però, poden ser dues eines totalment compatibles. La utilitat d'una i d'una altra per cada inversor depèn de l'estratègia d'inversió que utilitzi aquest. A un inversor de llarg termini l'anàlisi fonamental li diu què és el que ha de comprar, i l'anàlisi tècnic l'ajuda a precisar el moment concret per comprar-lo.

En el següent apartat, s'explica amb més detall què són aquests dos tipus d'anàlisi d'empreses i quins són els indicadors més usats per la majoria. Ja que, seran els que el nostre sistema també farà servir per predir.

4.1 Anàlisi Fonamental

Recapitulant, l'anàlisi fonamental es fa servir per calcular el valor real de l'empresa basant-se en els resultats esperats, els estats financers, l'entorn general i l'específic, la capacitat dels administradors i dirigents de l'empresa, els efectes macroeconòmics, la capacitat de creixement, etc.

Principalment, hi ha dues maneres de realitzar l'anàlisi fonamental: de dalt a baix o viceversa. La primera comença analitzant la situació general, passant per l'anàlisi sectorial i acabant per analitzar l'empresa en particular. En canvi, la segona realitza justament el recorregut contrari, comença analitzant l'entorn específic de l'empresa per acabar amb el general. Doncs, s'analitza les previsions macroeconòmiques, com per exemple, PIB, IPC, despesa Pública, tipus Interès.

[72] https://www.investopedia.com/university/fundamentalanalysis/
[73] https://www.investopedia.com/university/technical/
[74] https://www.investopedia.com/university/technical/techanalysis2.asp

Els Sectors de l'Economia, per exemple, Banca, Telecomunicacions, etc. I l'anàlisi individual de l'empresa, per exemple, Microsoft.

D'aquesta manera, a l'hora d'analitzar els resultats, es dóna més importància al primer que s'ha analitzat, per tant, en el primer cas es dóna més importància a l'entorn general de l'empresa que l'específic, que per contra serà el més important en el segon tipus d'anàlisi.

Independentment de quina d'elles se li doni un major pes, l'anàlisi fonamental també realitza el següents anàlisis.

El primer és la situació de l'empresa. Per determinar aquest aspecte s'analitzen el grau de maduresa del sector i el de l'empresa, analitzant el cicle de vida dels seus productes; l'amenaça competitiva, les barreres d'entrada i els aspectes legislatius que podrien afectar. Cal destacar, que pels inversors un aspecte clau és arribar a entendre l'empresa, els seus productes i serveis, i la seva estratègia.

El segon, la situació econòmica nacional, on s'analitza la situació econòmica del país, la política econòmica i monetària i l'estabilitat política.

Finalment, la situació econòmica global: Se centra a conèixer quin és el cicle econòmic. Els factors a tenir en compte són molts, i el pes que deu tenir cadascun en l'anàlisi dependrà del tipus d'empresa, dels països on operi i de la situació global.

4.1.1 Informes Financers

No hi ha una única forma d'analitzar tot tipus d'empreses i que serveixi per a tot tipus d'estratègies d'inversió. És convenient que l'anàlisi que s'utilitzi s'adapti al tipus de empreses i a l'estratègia d'inversió.

Empreses diferents i estratègies diferents requereixen anàlisi diferents. Per exemple, no és el mateix, analitzar un blue xip[75] del qual es té un bon coneixement del seu negoci que un blue xip el qual te un negoci que no es coneix tant, o una petita empresa de creixement, o una empresa cíclica, o una empresa en reestructuració, etc.

En aquest context, per estimar si una empresa està cara o barata des del punt de vista fonamental cal mirar seus "números". Les empreses mostren la seva situació principalment mitjançant la publicació de tres documents; el Balanç de situació[76], el Compte de Pèrdues i Guanys[77], i l'Estat de Flux d'Efectiu[78].

La situació d'una empresa varia contínuament, dia a dia, i fins i tot minut a minut. Lògicament, la variació d'un minut al següent és imperceptible i insignificant, però existeix. És a dir, que el negoci d'una empresa no s'atura mai. Qualsevol document que pot veure un inversor pertany al passat, tot i que l'empresa ho acabi de publicar fa un instant.

[75] https://en.wikipedia.org/wiki/Blue_chip_(stock_market)
[76] https://www.investopedia.com/terms/b/balancesheet.asp
[77] https://www.investopedia.com/terms/i/incomestatement.asp
[78] https://www.investopedia.com/terms/c/cashflowstatement.asp

El Balanç i el Compte de Pèrdues i Guanys mostren la situació d'una empresa en un moment concret. És com si se li fes una "foto" a l'empresa en un instant per poder analitzar-la després. Tot i que, realment el concepte de "foto" és més adequat per al balanç, que mostra la situació de l'empresa en una data concreta. El compte de resultats mostra el que ha succeït al llarg d'un període de temps, normalment un trimestre o un any sencer. Per exemple, en el cas dels documents anuals, que el balanç és la "foto" que es fa a l'empresa el dia de Cap d'Any i el compte de resultats és el "vídeo" de l'any sencer.

La inversió en Borsa té part de ciència, part d'art, i molt de psicologia. A aquests coneixements cal anar afegint experiència, i no oblidar mai el factor psicològic.

L'habitual és presentar el balanç i el compte de resultats al final de cada trimestre. És a dir, quatre vegades a l'any. Els més importants, lògicament, són els que es publiquen al final de cada exercici.

La majoria de les empreses, excepte unes poques excepcions, presenten un Compte de Pèrdues i Guanys i un Balanç individuals i altres consolidats. L'individual és el que es refereix únicament a l'empresa matriu, i el consolidat el qual inclou a l'empresa matriu i a totes les seves filials. Per aquest estudi es fan servir els consolidats, ja que el que cotitza és l'empresa matriu al costat de totes les seves filials, i això és el que es compra o ven.

Les poques excepcions a això són unes poques empreses molt petites que no tenen filials, i per tant només presenten el compte de resultats i el balanç individual. No hi ha lloc a la confusió, ja que aquestes empreses no presenten ni compte de resultats ni balanç consolidats.

4.1.1.1 Balanç de Situació

El Balanç de Situació és el que té l'empresa i el que deu l'empresa. Així com d'on ha tret els diners que ha aconseguit, i què ha fet amb ell. Inclou tres conceptes:

- Actiu: És tot allò que pertany a l'empresa i té valor; fàbriques, terrenys, locals, productes, maquinàries, etc.
- Passiu: És el que deu l'empresa, els seus deutes.
- Patrimoni net: És l'actiu menys el passiu, el que té menys el que deu. Aquest és el valor comptable de l'empresa, el seu valor teòric. Són els diners dels accionistes, tant el que es va aportar inicialment en crear l'empresa com els beneficis que ha anat generant l'empresa al llarg del temps i no s'han repartit com a dividends. Aquests beneficis que en lloc de repartir-com a dividend queden dins de l'empresa són les reserves.

El valor que s'assigna als actius en el balanç és el valor comptable, que no té per què coincidir amb el valor de mercat i, de fet, el més normal és que no ho faci. Per exemple, de vegades una empresa té valorats en el seu balanç uns terrenys al preu al qual els va comprar fa bastants anys, fins i tot dècades. També pot succeir que l'empresa tingui una filial no cotitzada en borsa i la tingui valorada molt per sota del preu que tindria en cas que decidís vendre-la o treure-la a borsa. Hi ha moltes més situacions d'aquest tipus. En conclusió, el valor comptable és només una aproximació, millor o pitjor, al valor de mercat dels actius de l'empresa.

Precisament aquesta és una de les principals dificultats de l'anàlisi fonamental. Si els balanços reflectissin el valor real dels actius seria molt fàcil detectar si una empresa està sobrevalorada o infravalorada, ja que n'hi hauria prou amb comparar el seu valor comptable, que coincidiria amb el valor real, amb la cotització. Però valor comptable i valor real no coincideixen habitualment.

Per trobar aquest valor real dels actius d'una empresa i per tant el valor real d'aquesta empresa cal saber estimar el valor real de fàbriques, terrenys, locals, marques comercials, inventaris, filials, etc. La qual cosa no és gens fàcil. Per fer totes aquestes valoracions correctament cal conèixer l'empresa i el seu sector de forma molt profunda, cosa que és fora de l'abast de l'inversor mitjà.

Els deutes si estan reflectides en el balanç pel seu import real, ja que és una cosa conegut i fàcil de determinar. Qualsevol empresa sap els diners que deu i a qui l'hi deu, amb total exactitud.

Com succeeix amb totes les eines i indicadors de l'anàlisi fonamental, el grau de complexitat pot complicar fins a l'infinit. Nogensmenys, si una empresa cotitza per sota del seu valor comptable, és molt probable que estigui infravalorada i sigui una bona oportunitat d'inversió.

De l'anterior no ha de desprendre que si una empresa cotitza per sobre del seu valor comptable està cara i no s'ha d'invertir en ella. Aquesta "regla" no és certa. Hi pot haver empreses que cotitzen per sobre del seu valor comptable, fins i tot molt per sobre, i així i estiguin molt barates i siguin una gran inversió.

Hi ha molts tipus d'empreses i molts negocis. Algunes empreses necessiten molt pocs actius per generar molts beneficis, la qual cosa és un avantatge important perquè han de gastar menys diners en mantenir i reposar aquests actius.

Aquestes empreses, habitualment, cotitzen a diverses vegades la seva valor comptable (5, 10, 20, ...), i no per això estan cares. Si aquest tipus d'empreses cotitzessin per sota del seu valor comptable no estarien barates, sinó ridículament barates. Tant que a la pràctica no arriba a succeir perquè molt abans que això passi els inversors es llancen a comprar els seus accions.

Per això el valor comptable no serveix per valorar a totes les empreses. En alguns casos no té cap utilitat.

4.1.1.2 Compte de Pèrdues i Guanys

El Compte de Pèrdues i Guanys, també anomenat de resultats, mostra els beneficis, o pèrdues, d'una empresa en un període de temps determinat (un any, un semestre, un trimestre, etc.), en base als ingressos i despeses que ha tingut aquesta empresa en aquest període.

El benefici net és el que els queda als accionistes després de restar als ingressos de la empresa totes les despeses que ha tingut l'empresa en aquest període (matèries primeres, sous, bonus, lloguer d'immobles, aigua, electricitat, gas, telecomunicacions, amortitzacions, provisions, interessos pagats per del deute, impostos, etc.):

De forma molt simplificada un Compte de Pèrdues i Guanys és el següent.

$$Ingressos - Despeses \ +/- \ Resultat \ financer \ - \ Impostos \ = \ Resultat \ net$$

On

$$Despeses = \ Despeses \ fixes + Despeses \ variables + Amortitzacions \ i \ Depreciacions + \ Deterioraments$$

Per entendre com es calcula, a continuació és fa una breu explicació de cada element del Compte de Pèrdues i Guanys.

- Ingressos.

Tot comença pels ingressos, els diners que ingressa l'empresa per vendre els seus productes o serveis als seus clients. A aquests ingressos cal restar-totes les despeses, i el que sobra és el resultat de l'empresa. Si el resultat de l'empresa és positiu, que és el normal, llavors també es diu benefici net.

- Despeses fixes i variables.

Les despeses fixes són molt importants, perquè l'empresa ha de fer-los front independentment que tingui ingressos o no. La majoria de les petites empreses (les que no cotitzen en borsa) que desapareixen, ho fan per no poder fer front a les despeses fixes. Com menors siguin les despeses fixes d'una empresa, major és la capacitat de sobreviure en els mals temps. Despeses fixes són els sous dels empleats, el lloguer o manteniment dels immobles que utilitzi l'empresa, les assegurances que tingui contractats, etc.

Les despeses variables també són importants perquè es resten dels ingressos igual que els fixos, però si els ingressos de l'empresa baixen, també ho fan les despeses variables. Si les vendes d'una empresa que fabrica i ven roba baixen, l'empresa comprarà menys tela, i per tant la despesa en tela (despesa variable) disminuirà alhora que cauen els ingressos, de forma natural. Però el lloguer de els locals (despesa fixa) en on l'empresa està intentant vendre les peces que ja estan fabricades no baixa perquè baixin les vendes.

L'empresa sí pot intentar renegociar el lloguer dels locals, o traslladar-se a altres més barats, amb el que estaria reduint les seves despeses fixes i fent-se més eficient. Però no és igual de fàcil comprar menys roba si no es necessita, que és una cosa immediata, de renegociar lloguers o traslladar-a altres locals, que és un procés més llarg, i pot tenir èxit o no. El que fa mal de veritat a les empreses en els mals moments són les despeses fixes, no les despeses variables.

- Amortitzacions i Depreciacions.

Els actius físics que compra una empresa perden valor amb el pas del temps. Per exemple, una empresa compra avui un camió i, a mesura que passa el temps, aquest camió cada vegada estarà més vell i valdrà menys diners, fins que arribi un dia en què ja no compleixi la seva funció i hagi de ser substituït per un altre. La comptabilitat reconeix això, i permet a l'empresa que tots els anys reflecteixi d'alguna manera la pèrdua de valor d'aquest camió. Aquesta forma de reflectir-ho són les depreciacions.

Els béns físics es deprecien en diferents terminis de temps, que depenen de la normativa comptable. Aquesta normativa comptable és diferent d'un país a un altre, i dins d'un mateix país varia amb el temps. En general, el nombre d'anys en què es deprecien els béns depèn del tipus de béns. Els ordinadors es deprecien en menys anys que les màquines industrials, i aquestes en menys anys que els immobles, per exemple.

Les amortitzacions són igual que les depreciacions, però s'apliquen als actius intangibles, com les marques, les patents, les despeses de recerca, etc. El concepte és molt similar, però canvia el nom.

Per il·lustrar-ho, en el cas que una empresa li compra a una altra una patent informàtica. Aquesta patent cada vegada valdrà menys perquè arribarà un moment en què es farà pública i, en qualsevol cas, molt probablement serà superada per altres tecnologies noves.

De manera similar al camió s'anirà amortitzant, en lloc de depreciant, la patent informàtica. I igualment aquestes amortitzacions que es van fent al llarg dels anys no suposen que surti diners de l'empresa, ja que la patent es va pagar en el primer moment (en aquest exemple, per facilitar la comprensió, en realitat els pagaments es poden fer de diverses vegades, però el mecanisme és el mateix), igual que el camió.

Per tant ni les amortitzacions ni les depreciacions suposen sortida de diners de l'empresa. Són un reconeixement de despeses, per envelliment i obsolescència dels actius, sense que surti diners de la caixa de l'empresa, que permet rebaixar la xifra de benefici net i pagar menys impostos. Els anys en que s'amortitzi un actiu poden coincidir amb la vida útil d'aquest actiu, però no sempre és així, perquè no tots els "camions" (o edificis, o màquines, o eines, etc.) tenen la mateixa vida útil.

- Deterioraments.

Els deterioraments consisteixen que l'empresa estima que, a causa de determinades circumstàncies, algun dels seus actius ha patit una pèrdua de valor que podria ser permanent. No és una cosa que es produeixi habitualment.

Ni en reflectir el deteriorament surt diners de l'empresa ni en recuperar un deteriorament que es va realitzar en el passat entra diners a l'empresa.

- Resultat financer.

Les empreses paguen interessos pels crèdits, préstecs, que tenen, i cobren interessos pel diners que tenen en comptes remunerats, dipòsits, etc. També poden cobrar dividends de accions que tinguin en altres empreses, realitzar operacions amb derivats (opcions, futurs i similars), etc. El resultat financer és la suma de tots aquests conceptes. Generalment, el resultat financer és negatiu, perquè la majoria de les empreses tenen deute, però en alguns casos és positiu. També hi ha empreses que tenen un resultat financer positiu perquè cobren més dividends d'accions que tenen en altres empreses que interessos han de pagar pel seu deute.

L'evolució dels canvis de les divises, en el cas d'empreses que tenen negocis en diferents monedes, també es reflecteix en el resultat financer. Unes vegades pot ser positiva i altres negativa.

- Impostos.

Després de restar dels ingressos tot el que hem vist fins ara obtenim el BAI. Sobre aquesta xifra és sobre la qual es calcula l'impost (Impost de Societats) que haurà de pagar l'empresa pels beneficis obtinguts. Aquest impost varia d'un país a un altre, i fins i tot dins d'un mateix país va variant amb el temps.

- Resultat net.

És el que queda finalment als accionistes després d'haver restat als ingressos tot el anterior, inclosos els impostos. El normal és que una part del benefici net es reparteixi als accionistes com a dividend i una altra part es quedi dins de l'empresa per fer noves inversions i, d'aquesta manera, augmentar més els beneficis futurs.

Convé subratllar que el Balanç d'una empresa i el seu Compte de Pèrdues i Guanys estan totalment relacionats.

Quan una empresa obté beneficis no sol repartir-los com dividend als seus accionistes íntegrament com hem vist, sinó que en conserva una part. Aquests diners generats per l'empresa que no es lliura als accionistes s'utilitza per a una gran varietat de coses, com obrir noves fàbriques o instal·lacions, comprar altres empreses, reduir el deute que tingui l'empresa, crear noves filials o negocis, deixar en renda fixa o liquiditat[79] fins que es trobi una destinació més rendible, etc.

Algunes d'aquestes alternatives incrementen els actius de l'empresa (obertura de fàbriques, compra d'empreses, etc.) i altres redueixen el passiu (reducció del deute). És a dir, els diners que procedeix del compte de resultats (els beneficis retinguts) modifica el balanç (l'actiu i / o el passiu).

Si una empresa té un balanç sanejat podrà utilitzar aquests recursos per obrir nous negocis o reorganitzar els que ja té. La qual cosa es reflectirà en el futur en el compte de resultats amb un augment dels beneficis, una part dels quals es repartiran com a dividend i una altra part passarà a engrossir novament el balanç, etc.

En una situació ideal, el balanç fa augmentar els beneficis i els beneficis augmenten el balanç, de forma indefinida. La conseqüència de tot això és un augment continu del valor de l'empresa.

En el cas contrari, si una empresa té pèrdues en el seu compte de resultats haurà afrontar-les augmentant el seu deute (augmentant el passiu) o venent algunes fàbriques, terrenys, filials, etc (Reduint el seu actiu). El resultat és una disminució del patrimoni net, i per tant del valor de l'empresa.

[79] https://www.investopedia.com/terms/l/liquidity.asp

La relació entre el balanç i el compte de resultats és continua i constant. És important distingir què pertany al balanç i què al compte de resultats.

4.1.1.3 Estat de Flux d'Efectiu

A més del compte de resultats i el balanç ha un tercer document, l'estat de fluxos d'efectiu. Aquest document, igual que els balanços i els comptes de resultats, es pot trobar a les webs de les empreses, a l'apartat per inversors. Té tres dades principals:

- Fluxos d'efectiu de les activitats d'explotació, o *cash-flow* d'explotació.

El flux d'explotació, o *cash-flow* d'explotació, són els diners que entren i surten realment de les comptes de l'empresa pel desenvolupament de la seva activitat ordinària.

El compte de resultats de les empreses es fa amb el criteri de meritació. És a dir, els ingressos i les despeses es comptabilitzen en el moment en què es comprometen, però no quan es realitzen físicament. El criteri de caixa és el que utilitzen els *cash-flows* o fluxos de caixa, i comptabilitzen les despeses (o pagaments) quan els diners surten realment dels comptes de l'empresa.

Així, com més gran sigui el flux de caixa d'explotació, millor. L'ideal és que en el futur el *cash-flow* d'explotació pugi tot el possible.

- Fluxos d'efectiu de les activitats d'inversió, o *cash-flow* d'inversió.

El *cash-flow* d'inversió és la diferència entre els diners que han sortit de l'empresa per realitzar inversions i els diners que han entrat a l'empresa per les desinversions que ha realitzat. El normal és que sigui una xifra negativa, perquè l'habitual és que les empreses inverteixin més del que desinverteixin. Això, és el que les fa créixer més en el futur, invertir en el present. A més de que, tota empresa necessita invertir diners tots els anys per seguir funcionant.

Altrament, quan una empresa té un deute excessiu i ven actius per reduir-la, aquest flux pot ser positiu (en aquest cas l'empresa desinverteixin més del que inverteix), i en aquest cas serà beneficiós perquè li permetrà reduir el seu deute a nivells més manejables, que és el que necessita aquesta empresa en aquestes circumstàncies concretes. Però una empresa no pot mantenir de forma indefinida un *cash-flow* d'inversió positiu, excepte en el cas que l'empresa estigui en liquidació, i es venguin tots els seus actius fins tancar-la.

- Fluxos d'efectiu de les activitats de finançament, o *cash-flow* de finançament.

El *cash-flow* de finançament mostra els diners que entren o surten de l'empresa per l'emissió de deute, pagament de deute, pagament de dividends, etc. Si és positiu, l'empresa ha augmentat el seu deute i/o la seva liquiditat. En canvi, si és negatiu el deute es redueix, i/o la liquiditat disminueix.

Cal pensar, que quan una empresa contreu un deute entra diners a l'empresa, els diners que acaba de rebre prestat, i per això aquest *cash-flow* de finançament és positiu. Quan l'empresa torna aquest deute els diners surt de la empresa (per pagar el deute), i per això el flux de caixa de finançament és negatiu.

A més, les ampliacions de capital són entrades de diners, que augmenten el flux de caixa de finançament, mentre que els pagaments de dividends són sortides de diners, que redueixen el flux de caixa de finançament.

Normalment, el *cash-flow* de finançament és positiu quan l'empresa s'està endeutant, i és negatiu quan va retornant els deutes contrets en el passat. Però es pot donar la situació que el flux de caixa de finançament sigui negatiu i que alhora deute augmenti, perquè els dividends pagats per l'empresa no s'hagin pagat amb el *cash-flow* d'explotació, que és el més convenient, sinó amb un augment del deute. Pagar els dividends contraient més deute pot succeir algun any per un "desquadrament" temporal entre cobraments i pagaments que es vagi a solucionar a curt termini, però no és una cosa que es pugui mantenir en el temps. En un cas així, la empresa podria reduir el seu dividend de manera temporal fins que les condicions milloressin.

L'ideal és que el flux de caixa d'explotació sigui suficient per cobrir les inversions, *cash-flow* de inversió, pagar els dividends i reduir el deute (o augmentar la liquiditat de l'empresa). Però, no sempre es compleix l'ideal.

4.1.2 Indicadors Fonamentals

Durant l'anàlisi fonamental, s'obtenen diferents indicadors que transmeten informació resumida o no, a l'inversor. La majoria es poden trobar als principals llocs webs d'informació financera i sobre cotitzacions, i en el última instancia, sempre hi seran en els informes financers que les empreses pengen a la pagina web a l'apartar d'inversors. Alguns no apareixen sempre calculats en tots aquests informes, però es poden calcular a partir de altres que sí es publiquen sempre.

El nombre de indicadors, ràtios i variables que es poden calcular amb els informes financers, mencionats prèviament, d'una empresa és pràcticament infinit.

Però, a més del cost de calcular-los tots, al inversor mitja no l'ajudaria a l'hora d'invertir, sinó que l'ompliria de dubtes per l'excés de informació. Per tant, l'important de la informació no és la quantitat, sinó la qualitat. L'excés d'informació pot resultar perjudicial, i porta a cometre errors que no es cometrien si es seleccionés millor la informació que s'analitza.

D'aquesta manera, hi ha una gran varietat d'indicadors i cap demostració de que uns siguin més rellevants que altres. Tot i així, hi ha uns més usats pels inversors.

Com que l'objectiu de l'estudi no es entendre els indicadors, no és fa èmfasi en l'explicació d'aquests i la definició de criteris per escollir els més rellevants. Així és que, per aquest estudi s'ha optat per fer servir tots els indicadors que s'han obtingut i no discriminar-ne cap.

Les dades s'han obtingut de la pagina web *stockpup*[80], que ofereix gratuïtament dades provinents de més de 20 anys de declaracions de 10-Q[81] i 10-K[82] realitzades per empreses públiques amb la Comissió de Borsa de Valors dels EUA (SEC[83]).

[80] http://www.stockpup.com
[81] https://www.investopedia.com/terms/1/10q.asp
[82] https://www.investopedia.com/terms/1/10-k.asp

L'informe 10-Q, un formulari de la SEC, és un informe complet del rendiment d'una empresa que ha de ser presentat trimestralment per totes les empreses públiques a la SEC. En el 10-Q, les empreses han de revelar informació rellevant sobre la seva situació financera. No hi ha presentació després del quart trimestre, ja que és quan es presenta el 10-K.

Per altra banda, l'informe 10-K és un informe complet presentat anualment per una empresa cotitzada en borsa sobre el seu rendiment financer i és requerit per la SEC. L'informe conté molt més detalls que l'informe anual[84] d'una empresa, que s'envia als seus accionistes abans d'una reunió anual per elegir els directors d'empresa.

Algunes de les informacions que una empresa necessita per documentar el 10-K inclouen la seva història, estructura organitzativa, estats financers, benefici per acció, filials, compensació executiva i qualsevol altra dada rellevant.

La SEC exigeix que aquest informe mantingui els inversors conscients de la situació financera d'una empresa i que els permeti disposar d'informació suficient abans de comprar o vendre accions de la corporació o abans d'invertir en els bons corporatius de l'empresa.

Per a cada empresa hi ha una sèrie d'indicadors, que a continuació es llisten amb una breu explicació, agrupats per categories.

4.1.2.1 Indicadors D'accions

- Accions(*Shares*[85]): Nombre total d'accions ordinàries en circulació al final d'un trimestre determinat, incloses totes les classes d'accions.
- Accions dividides ajustades(*Shares split adjusted*): El nombre d'accions que l'empresa tenia al final d'un trimestre determinat, ajustada a la divisió per ser comparable a les accions actuals.
- Factor de divisió(*Split factor*[86]): Si un inversor va començar amb 1 acció al final d'un determinat trimestre, el factor de divisió per a aquest trimestre indica quantes accions posseiria l'inversor avui com a conseqüència de les divisió de les accions.

4.1.2.2 Dades del balanç

- Actius corrents(*Current assets*[87]): Actius corrents al final d'un trimestre.
- Actius(*Assets*[88]): Actius totals al final d'un trimestre.
- Passius corrents(*Current liabilities*[89]): Passius corrents al final d'un trimestre.
- Passius(*Liabilities*[90]): Total passius al final del trimestre.
- Patrimoni net(*Shareholders equity*[91]): Total dels fons propis al final del trimestre, que inclouen els accionistes comuns i els preferits.

---

[83] https://www.investopedia.com/terms/s/sec.asp
[84] https://www.investopedia.com/terms/a/annualreport.asp
[85] https://www.investopedia.com/terms/s/shares.asp
[86] https://www.investopedia.com/terms/s/stocksplit.asp
[87] https://www.investopedia.com/terms/c/currentassets.asp
[88] https://www.investopedia.com/terms/a/asset.asp
[89] https://www.investopedia.com/terms/c/currentliabilities.asp
[90] https://www.investopedia.com/terms/l/liability.asp
[91] https://www.investopedia.com/terms/s/shareholdersequity.asp

- Interessos minoritaris(*Non-controlling interest*[92]): Interessos minoritaris, si n'hi ha, exclosos del patrimoni net.
- Patrimoni preferent(*Preferred equity*): Patrimoni net preferent, si escau, inclòs en el patrimoni net.
- Fons de comerç i intangibles(*Goodwill*[93] *& intangibles*[94]): Fons de comerç total i tots els altres actius intangibles, si n'hi ha.
- Deute a llarg termini(*Long-term debt*[95]): Tot el deute a llarg termini incloent obligacions de lloguer de capitals.

4.1.2.3 Dades del compte de resultats

- Ingressos(*Revenue*[96]): Ingressos totals d'un trimestre determinat.
- Guanys(*Earnings*[97]): Guanys o ingressos nets per a un trimestre determinat.
- Guanys disponibles per als accionistes habituals(*Earnings available for common stockholders*[98]): Benefici net menys els ingressos que s'han de distribuir als accionistes preferents. Depenent de l'empresa es pot ometre aquesta dada.
- EPS bàsic(*EPS basic*[99]): Beneficis bàsics per acció per un trimestre determinat.
- EPS diluït(*EPS diluted*[100]): Beneficis diluïts per acció.
- Dividend per acció(*Dividend per share*[101]): Dividends d'accions comuns pagats durant un trimestre per acció, incloent-hi tots els dividends i distribucions regulars i especials a accionistes comuns.

4.1.2.4 Dades dels estats financers

- Efectiu de les activitats operatives(*Cash from operating activities*[102]): Efectiu produït per activitats d'explotació durant un trimestre determinat, incloses les operacions continuades i discontinuades.
- Efectiu de les activitats d'inversió(*Cash from investing activities*[103]): Efectiu produït per activitats d'inversió durant un trimestre determinat, incloses les operacions continuades i discontinuades.
- Efectiu de les activitats de finançament(*Cash from financing activities*[104]): Efectiu produït per activitats de finançament durant un trimestre determinat, incloses les operacions continuades i discontinuades.
- Canvi de diners en efectiu durant el període(*Cash change during period*): Canvi en efectiu i equivalents d'efectiu durant un trimestre determinat, incloent-hi l'efecte dels moviments del tipus de canvi i altres ajustaments de canvi d'efectiu, si escau.

[92] https://www.investopedia.com/terms/n/noncontrolling_interest.asp
[93] https://www.investopedia.com/terms/g/goodwill.asp
[94] https://www.investopedia.com/terms/i/intangibleasset.asp
[95] https://www.investopedia.com/terms/l/longtermdebt.asp
[96] https://www.investopedia.com/terms/r/revenue.asp
[97] https://www.investopedia.com/terms/e/earnings.asp
[98] https://finance.zacks.com/formula-calculating-earnings-available-common-stockholders-4704.html
[99] https://www.investopedia.com/terms/b/basic-earnings-per-share.asp
[100] https://www.investopedia.com/terms/d/dilutedeps.asp
[101] https://www.investopedia.com/terms/d/dividend-per-share.asp
[102] https://www.investopedia.com/terms/c/cash-flow-from-operating-activities.asp
[103] https://www.investopedia.com/terms/c/cashflowfinvestingactivities.asp
[104] https://www.investopedia.com/terms/c/cashflowfromfinancing.asp

- Diners en efectiu al final del període(*Cash at end of period*): Efectiu i equivalents d'efectiu al final d'un trimestre, incloses les operacions continuades i interrompudes.
- Les despeses de capital(Capital expenditures[105]): Les despeses de capital són les sortides d'efectiu per a actius productius a llarg termini, nets d'efectiu procedents de la venda de béns de capital.

4.1.2.5 Dades del mercat

- Preu(*Price*): El preu mitjà per acció de les accions comunes de la companyia durant un trimestre determinat.
- Preu alt(*Price High*): El preu més alt per acció de les accions comunes de la companyia durant un trimestre determinat.
- Preu baix(*Price Low*): El preu més baix de les accions comunes de la companyia durant un trimestre.

4.1.2.6 Indicadors derivats

- ROE(*ROE*[106]): El rendiment del patrimoni net és l'indicador de guanys (disponibles per als accionistes habituals) TTM (al llarg dels dotze mesos finals) al patrimoni net mitjà normal de TTM.
- ROA(*ROA[107]*): Rendiment dels actius és la relació entre els ingressos totals TTM i els actius mitjans TTM.
- Valor comptable dels fons propis per acció(*Book value of equity per share[108]*): Fons propis ordinaris dels accionistes per acció, també coneguts com a BVPS.
- Ràtio P/B(*P/B ratio[109]*): La relació entre el valor i el valor del llibre de l'equitat per acció al trimestre anterior.
- Relació P/E(P/E ratio[110]): La ràtio de preu a TTM diluïda en EPS a partir del trimestre anterior.
- Dividends acumulats per acció(*Cumulative dividends per share[111]*): L'import agregat dels dividends pagats per la quota d'acords ordinaris dividits des del primer trimestre d'informe disponible fins a un trimestre determinat.
- Ràtio de pagament del dividend(*Dividend payout ratio[112]*): La proporció de dividends TTM amb guanys (disponibles per als accionistes habituals) TTM.
- Ràtio de deute a patrimoni a llarg termini(*Long-term debt to equity ratio[113]*): La proporció del deute a llarg termini amb el patrimoni net comú (fons propis de participació menys fons propis preferents).
- Relació entre el patrimoni net i els actius(*Equity to assets ratio*): La proporció de fons propis comuns (patrimoni net menys fons propis preferents) a actius.
- Ràtio actual(*Current ratio[114]*): La relació entre l'actiu corrent i el passiu corrent.

---

[105] https://www.investopedia.com/terms/c/capitalexpenditure.asp
[106] https://www.investopedia.com/terms/r/returnonequity.asp
[107] https://www.investopedia.com/terms/r/returnonassets.asp
[108] https://www.investopedia.com/terms/b/bvps.asp
[109] https://www.investopedia.com/terms/p/price-to-bookratio.asp
[110] https://www.investopedia.com/terms/p/price-earningsratio.asp
[111] https://www.investopedia.com/terms/c/cumulativedividend.asp
[112] https://www.investopedia.com/terms/d/dividendpayoutratio.asp
[113] https://www.investopedia.com/terms/d/debtequityratio.asp

- Marge net(*Net margin*[115]): La ràtio de guanys (disponible per als accionistes habituals) TTM a Ingressos TTM.
- Facturació d'actius(*Asset turnover*[116]): La proporció entre els actius mitjans TTM i els ingressos de TTM.
- Flux de caixa lliure per acció(*Free cash flow per share*[117]): Efectiu de les activitats operatives menys les despeses de capital durant un trimestre.

4.2 Anàlisi Tècnic

Com s'ha avançat prèviament, l'anàlisi tècnic es limita a estudiar les cotitzacions històriques de l'empresa, sense entrar a valorar en cap moment la marxa del negoci de l'empresa. En base als gràfics que hagin dibuixat les cotitzacions passades, exclusivament, l'anàlisi tècnic intenta determinar què és el que faran les cotitzacions en el futur.

Similar a aquest tipus d'anàlisi hi ha l'anàlisi gràfic[118], que és en realitat un estudi del comportament humà en el passat, amb la intenció de fer estimacions sobre quin serà, probablement, aquest comportament humà en el futur, i obtenir un rendiment econòmic amb això.

Un gràfic representa l'oferta i la demanda sobre un actiu. I l'oferta i la demanda es basen també, i molt, en els fonamentals d'aquest actiu. En un gràfic no està només el comportament de la gent que mira els gràfics, sinó el comportament d'absolutament tot el món que ha operat sobre aquest actiu en el passat, incloent aquells inversors que només miren els fonamentals.

L'anàlisi gràfic (i el tècnic) considera que tota la informació sobre un actiu (acció, índex, matèria primera, etc.) està continguda en el gràfic de preus, incloses totes les notícies sobre el negoci de la empresa, els seus resultats, el seu balanç, la informació privilegiada, etc. se suposa que tota aquesta informació està, repartida, en poder de totes les persones que operen sobre aquest actiu, i per tant el preu d'aquest actiu ja incorpora tota la informació existent sobre el mateix. Perquè tots els que han decidit comprar i vendre (i al preu que ho hagi fet cada un), o no fer res, van prendre totes aquestes decisions (milions de decisions) utilitzant tota la informació, del tipus que sigui, disponible al respecte. Cap persona té tota la informació, però el preu sí que té incorporada tota la informació, ja que el preu és el resultat final, o "resum", de totes les decisions que es han pres utilitzant tota la informació disponible al món sobre dit actiu.

Però, aquesta premissa de l'anàlisi tècnic ("el preu el incorpora tot ") seria certa en el cas d'un actiu molt líquid en el que la informació estigués molt repartida entre totes les persones que compren i venen aquest actiu, i cap dels intervinents en aquest mercat tingués un poder molt superior als altres per influir en la cotització. En cas que no es compleixin aquestes condicions, i el normal és que no es compleixin, potser aquesta premissa no sigui totalment certa, sinó parcialment certa.

---

[114] https://www.investopedia.com/terms/c/currentratio.asp
[115] https://www.investopedia.com/video/play/net-margin/
[116] https://www.investopedia.com/video/play/asset-turnover-ratio/
[117] https://www.investopedia.com/terms/f/freecashflowpershare.asp
[118] https://www.investopedia.com/university/technical/techanalysis8.asp

També pot succeir que algú tingui una gran capacitat d'influència sobre la cotització, conegui l'anàlisi tècnic i decideixi dibuixar determinades figures en els gràfics per induir a la resta d'inversors i operadors a operar de manera contrària a la que ho farien si tinguessin aquesta informació de qualitat que sí que té la persona que ha dibuixat aquestes figures gràfiques (és a dir, induir a la gent a vendre, per poder comprar més barat, o a comprar, per poder vendre més car). Infinitat de qüestions tècniques com arbitratges, garanties de préstecs que s'executen per circumstàncies personals d'un gran inversor, productes derivats, etc. També poden tenir una gran influència en la cotització d'un actiu, sense ser informació pròpia d'aquest actiu.

Cal destacar que, totes, o gairebé totes, les persones reaccionen d'una forma molt semblant davant els esdeveniments. Això és el que s'anomena un patró de comportament. Es pot dir que, en analitzar gràfics de cotitzacions, en realitat el que s'analitza és el comportament humà en el passat sobre aquesta empresa, índex, matèria primera, etc. I podem buscar i detectar certs patrons, i estimar en moltes ocasions amb una alta probabilitat (Relativament, això no és infal·lible) què farà la cotització en el futur.

Per altra banda, l'anàlisi tècnic és un intent de objectivar l'anàlisi gràfic. L'anàlisi tècnic pur consisteix en una sèrie d'indicadors i oscil·ladors[119] que es representen de forma gràfica. En aquest cas tots els analistes veuen el mateix. És a dir, hi ha multitud d'indicadors i oscil·ladors per triar, i no tots els analistes utilitzen els mateixos. Però tots els analistes que utilitzin l'indicador RSI, per exemple, veuen el mateix en la seva pantalla en un mateix moment. En canvi, en l'anàlisi gràfic les ratlles les pinta cada analista, i per tant en un mateix gràfic analistes diferents poden veure coses diferents.

Així doncs, ja que l'anàlisi gràfic és molt subjectiu i difícil de comprovar, i com resultat, no es presta bé a ser informatitzat, es descarta aquest anàlisi i l'estudi es centra en l'anàlisi tècnic amb els indicadors i oscil·ladors, que es detallen més endavant, i que seran els *inputs* del sistema.

En l'anàlisi tècnic és important el volum amb què es produeixen els moviments. El volum no es mesura en diners, sinó en nombre d'accions. Per què el que importa en realitat és el percentatge del capital social que està sent negociat en cada moment, i això es mesura amb el nombre d'accions, no amb el valor d'aquestes accions. Si es mesurés en diners, quan les cotitzacions estan altes el volum sempre (o gairebé sempre) seria superior a quan les cotitzacions estan baixes, però això no ens donaria una imatge real de l'activitat que està havent en el mercat en cada moment.

Un volum alt o baix reforça o debilita qualsevol figura de l'anàlisi tècnic. En general, una figura que es formi amb un volum alt, té més probabilitats de desenvolupar-tal com s'espera d'ella. I a l'inrevés, si una figura es forma amb un volum relativament baix, la seva fiabilitat és menor, encara que això no vol dir que no pugui comportar tal com s'espera.

Un volum alt indica que molts inversors (millor dit, molts diners, sigui de pocs o de molts inversors) està donant suport aquest moviment. Aquests diners pot estar equivocat o no, però

[119] https://www.investopedia.com/terms/o/oscillator.asp

quan és molt té moltes probabilitats de moure les cotitzacions a favor seu, encara que estigui equivocat. Per exemple, està equivocat quan ven una empresa que està barata, o quan compra una empresa que està cara, en els dos casos respecte al seu valor fonamental.

Quan el volum és baix, és que hi ha molts diners (i moltes accions) que està indecís, i que per tant en qualsevol moment podria decidir-se, en un sentit o un altre. Per això els moviments amb volum baix són menys fiables, perquè en qualsevol moment els diners (o les accions) que estan esperant podrien entrar en la direcció contrària a la que ha tingut la cotització fins a aquest moment, i contrarestar aquest moviment que s'ha produït prèviament amb un volum baix.

4.2.1 Indicadors Tècnics

Recapitulant, els indicadors tècnics són un intent de representar l'anàlisi gràfic com alguna cosa objectiu. Els indicadors tècnics utilitzen la mateixa informació que el anàlisi gràfic, les cotitzacions i els volums de contractació, però la mostren d'una forma diferent.

Amb les mateixes dades que l'anàlisi gràfic mostra un gràfic de cotitzacions, l'anàlisi tècnic mostra una sèrie de corbes (els indicadors tècnics). En estudiar les figures gràfiques vèiem que és diferent un Triangle "gran" de un Triangle "petit", per exemple, i que s'han d'interpretar de diferent forma. Però també vèiem que no hi ha una frontera clara que delimiti els Triangles "grans" dels "petits", de manera que un mateix Triangle per un analista pot ser "gran", mentre que per a un altre pot ser "petit".

Aquestes ambigüitats són les que intenten resoldre els indicadors tècnics. Si un indicador tècnic està en 27,89, per exemple, està en 27,89 per a tots els analistes, i no és possible que uns analistes interpreten que està en 27,89 mentre uns altres creuen que està en 82,15, i altres pensen que més aviat està en 67,54.

De tota manera, això no vol dir que resolguin totalment el problema de la ambigüitat en la interpretació de les anàlisis, perquè també hi ha coses en els indicadors que són interpretables. Els indicadors tècnics són una bona ajuda per a l'anàlisi dels gràfics, però el conjunt de l'anàlisi, que és el realment important, segueix sent una cosa interpretable, que té part de ciència, i part d'art i d'experiència.

Tots els indicadors tècnics es calculen amb dades (cotitzacions i/o volums) del passat i, per tant, no són, en si mateixos, una predicció del futur. Es pot dir que cada un d'ells és un petit resum del passat, i que cada un d'aquests resums es realitza de diferent manera. Però sempre amb dades del passat, mai amb dades del futur, que evidentment són impossibles de conèixer.

Encara que, per descomptat, per molt alt que sigui el percentatge de probabilitats no es tracta d'encertar sempre, sinó de millorar el resultat global, respecte al resultat global si no utilitzéssim totes aquestes eines tècniques.

Cada indicador utilitza una fórmula matemàtica diferent, i no hi ha límit per al nombre de fórmules que es poden utilitzar. Cada inversor podria inventar-un, o més, indicadors tècnics. No és normal que l'inversor mitjà es posi a inventar-se indicadors tècnics, però sí que és habitual que vagin apareixent nous indicadors tècnics que aconsegueixen certa popularitat.

Bàsicament hi ha dos tipus d'indicadors, els indicadors de tendència i els oscil·ladors. Els indicadors de tendència[120], com per exemple las mitjanes mòbils o el MACD, són els que funcionen millor quan les cotitzacions estan en tendència alcista o baixista, i els oscil·ladors, com per exemple el RSI, són els que funcionen millor quan les cotitzacions estan en tendència lateral.

Dit d'una altra manera, els indicadors de tendència donen molts senyals falses quan la tendència és lateral, i els oscil·ladors donen molts senyals falses quan la tendència és alcista o baixista. A continuació, es mostren els indicadors tècnics que es faran en aquest estudi servir com a inputs del sistema de predicció, i una breu explicació dels mateixos.

4.2.1.1 Mitjanes Mòbils

Una mitjana mòbil[121] és un indicador àmpliament utilitzat en l'anàlisi tècnic que ajuda a suavitzar l'acció de preus filtrant el soroll de les fluctuacions aleatòries dels preus a curt termini. És un indicador que segueix la tendència o que té un retard, ja que es basa en preus anteriors.

Les dues mitjanes mòbils bàsiques i comunament utilitzades són la SMA, que és la mitjana simple d'una cotització en un determinat nombre de períodes de temps i la EMA) la qual cosa dóna més pes als preus més recents.

Les aplicacions més habituals de les mitjanes mòbils són identificar la direcció de la tendència i determinar els nivells de suport i resistència. Tot i que les mitjanes mòbils són prou útils per si soles, també constitueixen la base per a altres indicadors tècnics, com ara la MACD.

- SMA.

Així doncs, les mitjanes mòbils s'utilitzen per suavitzar les dades d'una matriu per ajudar a eliminar el soroll i identificar tendències. El SMA[122], literalment, és la forma més simple d'una mitjana mòbil. Cada valor de sortida és la mitjana dels valors anteriors n.

En una mitjana mòbil simple, cada valor en el període de temps té el mateix pes i els valors fora del període de temps no estan inclosos a la mitjana. Això fa que sigui menys sensible als canvis recents en les dades, que poden ser útils per filtrar aquests canvis.

Formula:

$$SMA = \frac{\sum_1^n preu}{n}$$

On *n* és el nombre de períodes totals.

[120] https://www.investopedia.com/terms/t/trendanalysis.asp
[121] https://www.investopedia.com/terms/m/movingaverage.asp
[122] https://www.investopedia.com/terms/s/sma.asp

- EMA

La EMA[123] és un element bàsic de l'anàlisi tècnic i s'utilitza en innombrables indicadors tècnics. En una SMA, cada valor en el període de temps té el mateix pes i els valors fora del període de temps no estan inclosos a la mitjana. No obstant això, la EMA és un càlcul acumulat, incloses totes les dades. Els valors passats tenen una contribució decreixent a la mitjana, mentre que els valors més recents tenen una contribució més gran. Aquest mètode permet que la mitjana mòbil sigui més sensible als canvis de dades.

Formula:

$$EMA = EMA_{-1} + K * (preu - EMA_{-1})$$

$$K = \frac{2}{(n+1)}$$

On *n* és el nombre de períodes totals.

- MACD

La MACD[124] és la diferència entre dos mitjanes mòbils exponencials. La línia de senyal és una mitjana mòbil exponencial del MACD.

El MACD assenyala canvis de tendència i indica l'inici de la nova direcció de tendència. Els valors elevats indiquen condicions de sobrecompra, els valors baixos indiquen condicions de sobrevenda. La divergència amb el preu indica el final de la tendència actual, especialment si el MACD té valors extrems o baixos. Quan la línia MACD creua per sobre de la línia de senyal, es genera un senyal de compra. Quan el MACD creua per sota de la línia de senyal, es genera un senyal de venda. Per confirmar el senyal, el MACD hauria d'estar per sobre de zero per a una compra i per sota de zero per a una venda.

Els períodes de temps per al MACD sovint es donen com a 26 i 12. Per altra banda, per crear un indicador similar amb períodes de temps diferents dels integrats al MACD, existeix la funció *Price Oscillator*[125].

Formula:

$$MACD = 12\ period\ EMA - 26\ period\ EMA$$

[123] https://www.investopedia.com/terms/e/ema.asp
[124] https://www.investopedia.com/terms/m/macd.asp
[125] https://www.investopedia.com/terms/p/ppo.asp

4.2.1.2 Oscil·ladors

Els oscil·ladors són indicadors que s'utilitzen quan es visualitzen gràfics que no són de tendència. Les mitjanes mòbils i les tendències són primordials quan s'estudia la direcció d'una acció. Un analista tècnic utilitzarà oscil·ladors quan els gràfics no mostrin una tendència definitiva en cap de les direccions.

Els oscil·ladors són, per tant, més beneficiosos quan les accions d'una empresa es troben en un patró de negociació horitzontal o lateral o no han estat capaços d'establir una tendència definitiva en un mercat agitat[126].

Quan la cotització es troba en una situació de sobrecompra o sobrevenda, s'exposa el valor real de l'oscil·lador. Amb els oscil·ladors, un pot veure quan la cotització arriba al punt en què es mou a una situació de sobrecompra. Això significa simplement que el volum de compres ha disminuït durant diversos dies de negociació, cosa que significa que els comerciants començaran a vendre les seves accions. Per contra, quan una acció ha estat venuda per un nombre més gran d'inversors durant un període de temps consistent entre un i sis mesos o més, la cotització entrarà en una situació de sobrevenda.

Hi ha una gran varietat d'Oscil·ladors i a continuació es detalla els que es fan servir per aquest estudi.

- STOCH

L'STOCH mesura quan el tancament té relació amb l'interval de negociació recent. Els valors oscil·len entre zero i 100. Els valors %D superiors a 75 indiquen una condició de sobrecompra; els valors inferiors a 25 indiquen una condició de sobrevenda. Quan el %D ràpid es creua per sobre del %D lent, és un senyal de compra; quan es creua per sota, és un senyal de venda. El Raw %K es considera generalment massa erràtic per utilitzar-lo per a senyals de creuament.

Formula:

$$\%K = 100 * \frac{close - LowestLow_{last\ n\ periods}}{HighestHigh_{last\ n\ periods} - LowestLow_{last\ n\ periods}}$$

On *close* el preu de tancament més recent, *HighestHigh* és el preu més alt negociat de les *n* sessions anteriors, *LowestLow és* el preu més baix negociat durant el mateix període *n*, i %K el valor actual de l'indicador estocàstic.

$$\%D = MovingAverage(\%K)$$

On %D sempre es refereix a un %K suavitzat (independentment si el propi %K es suavitza o no).

[126] https://www.investopedia.com/terms/c/choppymarket.asp

- RSI

L'RSI[127] va ser desenvolupat per J. Welles Wilder[128] i és un dels oscil·ladors més usats i coneguts pels inversors d'anàlisi tècnic. Calcula una relació entre els moviments recents de preus a l'augment i el moviment dels preus absoluts. L'RSI oscil·la entre 0 i 100 i s'interpreta com un indicador de sobrecompra, quan el valor és superior a 70, i sobrevenda per sota de 30.

Formula:

$$RSI = 100 - \frac{100}{1 + RSI}$$

On

$$RS = \frac{EMA_n(Up)}{EMA_n(Down)}$$

On

$$Up = \max(preu_n - preu_{n-1}, 0)$$

$$Down = \max(preu_{n-1} - preu_n, 0)$$

- WILL %R

El Williams[129] és un oscil·lador % R similar a l'STOCH sense suavitzar. Els valors oscil·len entre zero i 100, i es representen a escala inversa, és a dir, amb zero a la part superior i 100 a la part inferior. Reflecteix el nivell del preu de tancament en relació amb el preu més elevat per al període de temps analitzat, tradicionalment de 14 períodes. En contrast, l'oscil·lador estocàstic reflecteix el nivell del preu de tancament en relació amb el preu més baix (baix més baix). Els valors per sota de 20 indiquen una condició de sobrecompra i es genera un senyal de venda quan creua la línia 20. Els valors superiors a 80 indiquen una condició de sobrevenda i es genera un senyal de compra quan creua la línia 80.

Formula:

$$WILL\ \%R = 100 * \frac{HighestHigh_{last\ n\ periods} - close}{HighestHigh_{last\ n\ periods} - LowestLow_{last\ n\ periods}}$$

On *close* el preu de tancament més recent, *HighestHigh* és el preu més alt negociat de les *n* sessions anteriors, *LowestLow és* el preu més baix negociat durant el mateix període *n*, i %R el valor actual de l'indicador.

[127] https://www.investopedia.com/articles/active-trading/042114/overbought-or-oversold-use-relative-strength-index-find-out.asp
[128] https://www.amazon.com/exec/obidos/ASIN/0894590278/fmlabs/102-2215040-6704167?_encoding=UTF8&camp=1789&link_code=xm2
[129] https://www.investopedia.com/terms/w/williamsr.asp

- CCI

El CCI[130] està dissenyat per detectar les tendències del mercat inicial i final. El rang de 100 a -100 és el rang normal. Els valors de CCI fora d'aquest rang indiquen condicions de sobrecomrpa o sobrevenda. També es pot cercar divergències de preus fent servir el CCI. Si el preu està produint nous màxims, i el CCI no ho és, és probable que hi hagi una correcció de preus.

Formula:

$$CCI = \frac{TP - EMA_n(PT)}{0{,}015 * MeanDeviation}$$

On

$$TP = \frac{High_n + Low_n + close}{3}$$

On *close* el preu de tancament més recent, *High* és el preu més alt negociat de les *n* sessions anteriors, *Low és* el preu més baix negociat durant el mateix període *n*, tradicionalment de 20 períodes. Finalment, CCI és el valor actual de l'indicador.

- MOM

El *Momentum*[131] és una mesura de l'acceleració i la desacceleració dels preus. Indica si els preus estan augmentant a un ritme creixent o disminueixen a un ritme decreixent. La funció MOM es pot aplicar al preu o a qualsevol altra sèrie de dades.

Formula:

$$MOM = preu - preu_{(-n)}$$

- ROC

EL ROC[132] mesura la taxa de canvi en relació amb períodes anteriors i s 'utilitza per determinar amb quina rapidesa canvien les dades. El factor sol ser de 100 i s'utilitza simplement per facilitar la interpretació o la representació gràfica dels números. La funció es pot utilitzar per mesurar la taxa de canvi de qualsevol sèrie de dades, com ara el preu o un altre indicador.

Formula:

$$ROC = factor * \frac{preu}{preu_{(-n)}}$$

[130] https://www.investopedia.com/terms/c/commoditychannelindex.asp
[131] https://www.investopedia.com/terms/m/momentum.asp
[132] https://www.investopedia.com/terms/p/pricerateofchange.asp

- OBV

L'OBV[133] és un indicador tècnic que utilitza el flux de volum per predir canvis en la cotització i va ser desenvolupat per Joseph Granville i es descriu en el seu llibre[134] de 1963. Quan el tancament és superior al tancament anterior, el volum s'afegeix al total en execució i, quan el tancament és inferior al tancament anterior, el volum es resta del total en execució. Per interpretar-lo, es busca que l'OVV es mogui amb el preu o precedeixi els moviments de preus.

Formula:

Si preu de tancament és superior al dia anterior, llavors:

$$OBV = OBD_{-1} + volum$$

En canvi, si preu de tancament és inferior al dia anterior, llavors:

$$OBV = OBD_{-1} - volum$$

Finalment, altrament:

$$OBV = OBD_{-1}$$

- AD

L'oscil·lador Chaikin AD[135] (Acumulació/distribució) rep el nom del seu creador Marc Chaikin i mesura la línia d'acumulació-distribució del MACD. La línia d'acumulació / distribució és similar a la del OBV, que suma els temps de volum + 1 / -1 en funció de si el tancament és superior al tancament anterior. L'indicador AD, no obstant això, multiplica el volum pel valor de localització proper (CLV). El CLV es basa en el moviment de la cotització dins d'una sola barra i pot ser +1, -1 o zero.

La línia AD s'interpreta buscant una divergència en la direcció de l'indicador relativa al preu. Si té tendència a l'augment, indica que preu pot seguir-la. A més, si es torna plana mentre el preu segueix pujant (o caient), sol indicar que el preu es mantindrà en lateral.

Formula:

$$AD = AD_{-1} + CLV * volum$$

On

$$CLV = \frac{(close - low) - (\ high - close)}{(high - low)}$$

On *close* el preu de tancament més recent, *high* és el preu més alt de la sessió, *low* és el preu més baix de la sessió.

[133] https://www.investopedia.com/terms/o/onbalancevolume.asp
[134] https://www.amazon.com/exec/obidos/ASIN/0138509336/fmlabs/102-2215040-6704167?_encoding=UTF8&camp=1789&link_code=xm2
[135] https://www.investopedia.com/terms/c/chaikinoscillator.asp

- ADX

L'ADX[136] és un indicador d'anàlisi tècnic utilitzat per alguns analistes per determinar la fortalesa d'una tendència, desenvolupat per J. Welles Wilder i es descriu en el seu llibre[137] de 1978. La tendència pot ser cap amunt o cap avall, i això es mostra a través de dos indicadors acompanyants: l'indicador direccional negatiu (-DI) i l'indicador direccional positiu (+DI).

Per tant, ADX normalment inclou tres línies separades i s'utilitzen per ajudar a avaluar si un comerç ha de ser llarg o curt, o bé si s'hauria de fer un intercanvi. Els valors oscil·len entre 0 i 100, però poques vegades superen els 60. Per interpretar l'ADX, considerem un nombre elevat com una tendència forta i un nombre baix, una tendència feble.

Formula:

$$ADX = \frac{ADX_{-1} * (n-1) + DX}{n}$$

On DX normalment es suavitza amb una mitjana mòbil. Els valors oscil·len entre 0 i 100, però poques vegades superen els 60. Per interpretar el DX, es considera un nombre elevat com una tendència forta i un nombre baix, una tendència dèbil.

$$DX = \frac{(+DI) - (-DI)}{(+DI) + (-DI)}$$

- AROON

L'indicador AROON[138] intenta mostrar quan apareix una nova tendència i fou desenvolupat per Tushar S. Chande i descrit per primera vegada el setembre de 1995. L'indicador consta de dues línies, amunt(Up) i avall(Down), que mesuren el temps que ha transcorregut entre el màxim i mínims absoluts, produïts dins d'un interval de *n* períodes.

Quan l'AROON Up es manté entre els 70 i els 100, indica una tendència ascendent. Quan l'AROON Down es manté entre els 70 i els 100, això indica una tendència baixista. Una forta tendència a l'alça s'indica quan l'AROON Up és superior a 70 mentre que l'AROON Down està per sota de 30. Així mateix, s'indica una forta tendència a la baixa quan l'AROON Down està per sobre de 70 mentre que l'AROON Up està per sota de 30. També són considerats els creuaments. Quan l'AROON Down creua per sobre de l'AROON Up, indica un debilitament de la tendència ascendent (i viceversa).

Formula:

$$AROON\ Up = 100 * \frac{n - PeriodsSinceHighestHigh}{n}$$

$$AROON\ Down = 100 * \frac{n - PeriodsSinceLowestLow}{n}$$

[136] https://www.investopedia.com/terms/a/adx.asp
[137] https://www.amazon.com/exec/obidos/ASIN/0894590278/fmlabs/102-2215040-6704167?%5Fencoding=UTF8&camp=1789&link%5Fcode=xm2
[138] https://www.investopedia.com/terms/a/aroon.asp

- BBANDS

Les bandes de Bollinger[139] estan formades per tres línies. La banda mitjana és una SMA, de generalment 20 períodes, del preu típic (TP). Les bandes superior i inferior són desviacions estàndard *F* (generalment 2) per sobre i per sota de la banda central. Les bandes s'amplien i es redueixen quan la volatilitat del preu és superior o inferior, respectivament.

Les bandes de Bollinger no generen, en si mateixes, senyals de compra o venda; són un indicador de condicions de sobrecompra o sobrevenda. Quan el preu estigui a prop de la banda superior o inferior, indica que pot ser que la reversió sigui imminent. La banda central es converteix en un nivell de suport[140] o resistència[141]. Les bandes superior i inferior també es poden interpretar com a objectius de preus. Quan el preu rebota de la banda inferior i creua la banda central, la banda superior es converteix en el preu objectiu.

Formula:

$$UpperBand = MidBand + F * \sigma(TP)$$

$$LowerBand = MidBand - F * \sigma(TP)$$

On

$$MidBand = SMA(TP)$$

On

$$TP = \frac{high + low + close}{3}$$

[139] https://www.investopedia.com/terms/b/bollingerbands.asp
[140] https://www.investopedia.com/terms/s/support.asp
[141] https://www.investopedia.com/terms/r/resistance.asp

5 Arquitectura de l'estudi

En aquest capítol es desenvolupa l'estratègia usada per aquest estudi i les eines que s'han fet servir. Es detalla el per què de les decisions preses i s'acompanya d'il·lustracions. Cal remarcar, que hi ha moltes maneres de afrontar un estudi com aquest i que el resultat pot variar significativament entre les diferents estratègies.

5.1 Python

Abans de començar amb el desenvolupament del estudi, cal destacar que aquest s'ha fet íntegrament fent servir Python[142]. Ja que, és llenguatge de programació molt usat per la majoria de analistes de dades, i disposa d'una gran comunitat. Fet que facilita molt poder trobar solucions als problemes que van apareixen al llarg del estudi.

A més és un llenguatge de programació interpretat, el qual te la filosofia de posar l'accent en una sintaxi que afavoreixi un codi llegible. Un altre avantatge és que existeixen una gran quantitat de llibreries de ML, que resulten molt útils per un estudi com aquest. De fet, s'han fet servir moltes llibreries i cal destacar-ne les més rellevants, com per exemple, Pandas[143], NumPy[144], SciPy[145].

Així, NumPy és una extensió de Python, que li agrega major suport per vectors i matrius, constituint una biblioteca de funcions matemàtiques d'alt nivell per operar amb aquests vectors o matrius.

Per altra banda, Pandas és una biblioteca de programari escrita com extensió de NumPy per la manipulació i anàlisi de dades per al llenguatge de programació Python. En particular, ofereix estructures de dades i operacions per manipular taules numèriques i sèries temporals.

Com també, SciPy és una biblioteca lliure i de codi obert per a Python i es compon d'eines i algoritmes matemàtics.

Finalment, l'entorn de programació és mitjançant un *software* obert Jupyter[146], que permet treballar amb *notebooks*[147], que tenen una alta flexibilitat i provar lliurement parts del codi.

5.2 Arquitectura de sistema

En segon lloc, és de gran ajut dissenyar l'arquitectura del sistema abans de començar a desenvolupar. Això, pot evitar errors en el futur i ajuda a clarificar que s'està dissenyant. Al llarg de la memòria, el context del sistema s'ha descrit, així com els objectius.

Però tota arquitectura, té uns requeriments que ha de satisfer. A continuació, es detalla els d'aquest estudi pels dos casos.

---

[142] https://www.python.org/
[143] https://pandas.pydata.org/
[144] https://www.numpy.org/
[145] https://www.scipy.org/
[146] https://jupyter.org/
[147] https://ipython.org/notebook.html

5.2.1 Primer cas

El primer cas pretén estudiar i dissenyar un sistema capaç de predir el moviment del preu de les accions de les empreses d'un dels principals índex borsaris del mon, el DJIA, basat en el històric de preu d'aquestes, i també en indicadors extrets d'anàlisi fonamental i tècnic.

5.2.1.1 Requeriments

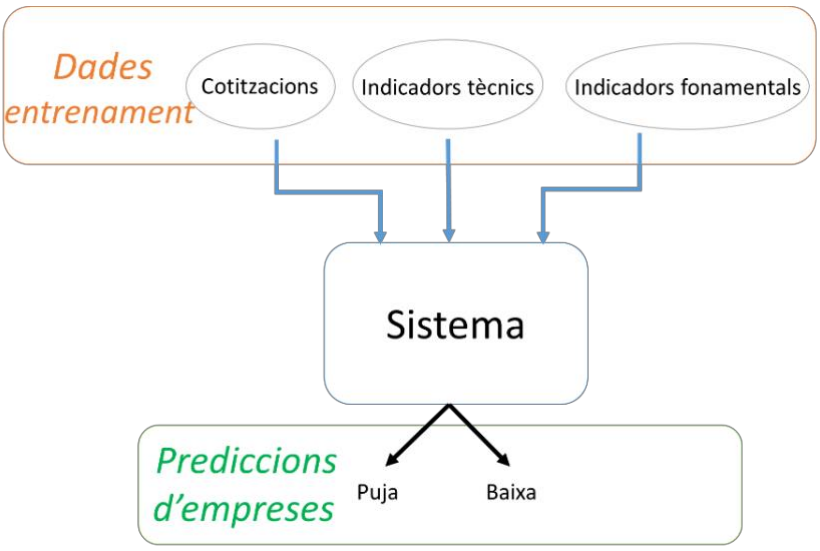
Els requeriments són les funcionalitats i limitacions del sistema, escrites de tal manera que independent del lector, la comprensió d'aquests no variï. Així, segons l'abast i els objectius definits per aquest estudi, s'estableixen una sèrie de requeriments a complir pel sistema. Es poden resumir en els següents.

- El sistema està conformat per un model de ML.
- L'objectiu és predir si el preu de la cotització de les accions de les empreses pujarà o baixarà el dia següent.
- El model ha de ser escalable i replicable, és a dir, fàcilment usable si s'afegeixen més dades d'entrada o aquestes canvien per altres accions d'empreses.
- El model ha de obtenir les dades mitjançant les APIs de les webs.
- Ha de ser capaç de funcionar en un ordinador corrent, per tant no pot tenir un cost computacional molt elevat.

5.2.1.2 Arquitectura

Recapitulant, una arquitectura de sistemes és un model conceptual que defineix l'estructura, el comportament i més vistes d'un sistema. Una descripció d'arquitectura és una descripció i representació formal d'un sistema, organitzat de manera que doni suport al raonament sobre les estructures i els comportaments del sistema. Una arquitectura del sistema pot consistir en components del sistema i en els subsistemes desenvolupats, que treballaran junts per implementar el sistema global.

Per aquest estudi s'ha dissenyat una arquitectura bàsica que satisfà els requeriments mencionats prèviament.



Il·lustració 12.- Arquitectura del primer sistema.

5.2.2 Segon cas

Per acabar, el segon cas pretén estudiar i dissenyar un sistema capaç de predir el moviment del preu d'un dels principals índex borsaris del mon, com per exemple el DJIA, basat en notícies d'un conegut diari digital.

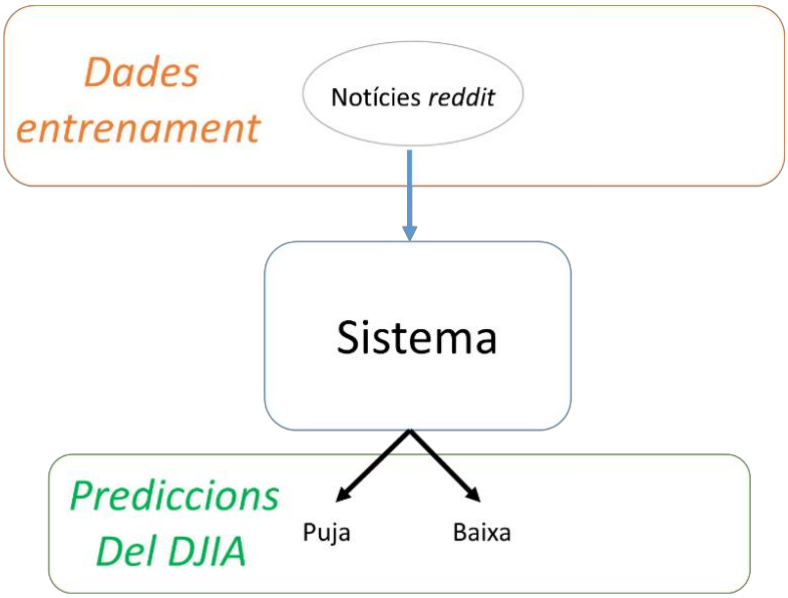
5.2.2.1 Requeriments

Aquest segon cas, també té una sèrie de requeriments a complir pel sistema. Es poden resumir en els següents, molt similars al primer.

- El sistema està conformat per un model de ML.
- L'objectiu és predir si el preu de la cotització de l'índex borsari DJIA pujarà o baixarà el dia següent.
- El model ha de ser escalable i replicable.
- Ha de ser capaç de funcionar en un ordinador corrent, per tant no pot tenir un cost computacional molt elevat.

5.2.2.2 Arquitectura

La arquitectura del segon estudi és la següent, satisfent els requisits mencionats.



Il·lustració 13.- Arquitectura del segon sistema.

6 Dataset

6.1 Primer cas

El ML necessita dades per funcionar i el primer pas és la obtenció d'aquestes. Les dades són un element important i la seva elecció resulta crucial. La quantitat i la qualitat de les mateixes té un efecte directe en el resultat.

Hi ha diverses empreses que ofereixen dades dels mercats de valors, però la majoria no ofereixen un servei gratuït. Com que quedava fora de l'abast d'aquest estudi, s'ha buscat per bases de dades gratuïtes. Fet que ha limitat el nombre de dades disponibles, ja que les bases de dades amb més quantitat i varietat són de pagament.

En aquest context, es va fer servir per obtenir les dades una empresa anomenada *Alphavantage*[148], creada per una comunitat d'investigadors, enginyers i professionals de l'empresa, i és un proveïdor web d'APIs gratuïtes per a dades històriques i en temps real sobre accions, divises i criptomonedes.

Les dades és divideixen en tres tipus: valors dels principals mercats de valors d'accions d'empreses, indicadors d'anàlisi fonamental, i indicadors d'anàlisi tècnic.

6.1.1 Cotitzacions

Una cotització és el preu que adquireix una acció, lletra, o qualsevol altre títol valor o valor mobiliari al mercat borsari (Borsa). La cotització és el valor que assigna un mercat a un bé o servei.

En el cas de la cotització de la borsa, fa referència a la publicació de la llista dels preus que tenen els diferents valors que es comercialitzen en ella. Per exemple, el DJIA[149], o simplement el *Dow*, és un índex borsari que indica el valor de 30 grans empreses de propietat pública amb seu als Estats . Aquestes 30 empreses també estan incloses a l'índex S&P 500[150].

El valor del *Dow* no és una mitjana aritmètica ponderada i no representa la capitalització borsària de les empreses que el componen, sinó la suma del preu d'una acció borsària per a cada empresa component. La suma es corregeix per un factor que canvia quan una de les accions del component té una divisió d'accions o un dividend en accions, de manera que es generi un valor consistent per a l'índex.

És el segon índex més antic del mercat nord-americà després del DJTA, creat per l'editor de *Wall Street Journal* i cofundador de *Dow Jones & Company*, Charles *Dow*. La mitjana industrial es va calcular per primera vegada el 26 de maig de 1896.

[148] https://www.alphavantage.co/
[149] https://www.investopedia.com/terms/d/djia.asp
[150] https://en.wikipedia.org/wiki/S%26P_500_Index

Actualment, està conformat per les següents empreses.

Companyia	Borsa	Símbol	Industria
3M	NYSE	MMM	Diversificada
American Express	NYSE	AXP	Serveis financers
Apple	NASDAQ	AAPL	Informàtica
Boeing	NYSE	BA	Aeroespacial i armament
Caterpillar	NYSE	CAT	Construcció i Mineria
Chevron	NYSE	CVX	Oli i gas
Cisco Systems	NASDAQ	CSCO	Informàtica
Coca-Cola	NYSE	KO	Alimentaria
Dow Inc.	NYSE	DOW	Química
ExxonMobil	NYSE	XOM	Oli i gas
Goldman Sachs	NYSE	GS	Serveis financers
The Home Depot	NYSE	HD	Minorista
IBM	NYSE	IBM	Informàtica
Intel	NASDAQ	INTC	Informàtica
Johnson & Johnson	NYSE	JNJ	Farmacèutica
JPMorgan Chase	NYSE	JPM	Serveis financers
McDonald's	NYSE	MCD	Alimentaria
Merck & Company	NYSE	MRK	Farmacèutica
Microsoft	NASDAQ	MSFT	Informàtica
Nike	NYSE	NKE	Tèxtil
Pfizer	NYSE	PFE	Farmacèutica
Procter & Gamble	NYSE	PG	Bens de consum
Travelers	NYSE	TRV	Insurance
UnitedHealth Group	NYSE	UNH	Salud
United Technologies	NYSE	UTX	Aeroespacial
Verizon	NYSE	VZ	Telecomunicacions
Visa	NYSE	V	Serveis financers
Walmart	NYSE	WMT	Minorista
Walgreens Boots Alliance	NASDAQ	WBA	Minorista
Walt Disney	NYSE	DIS	Entreteniment

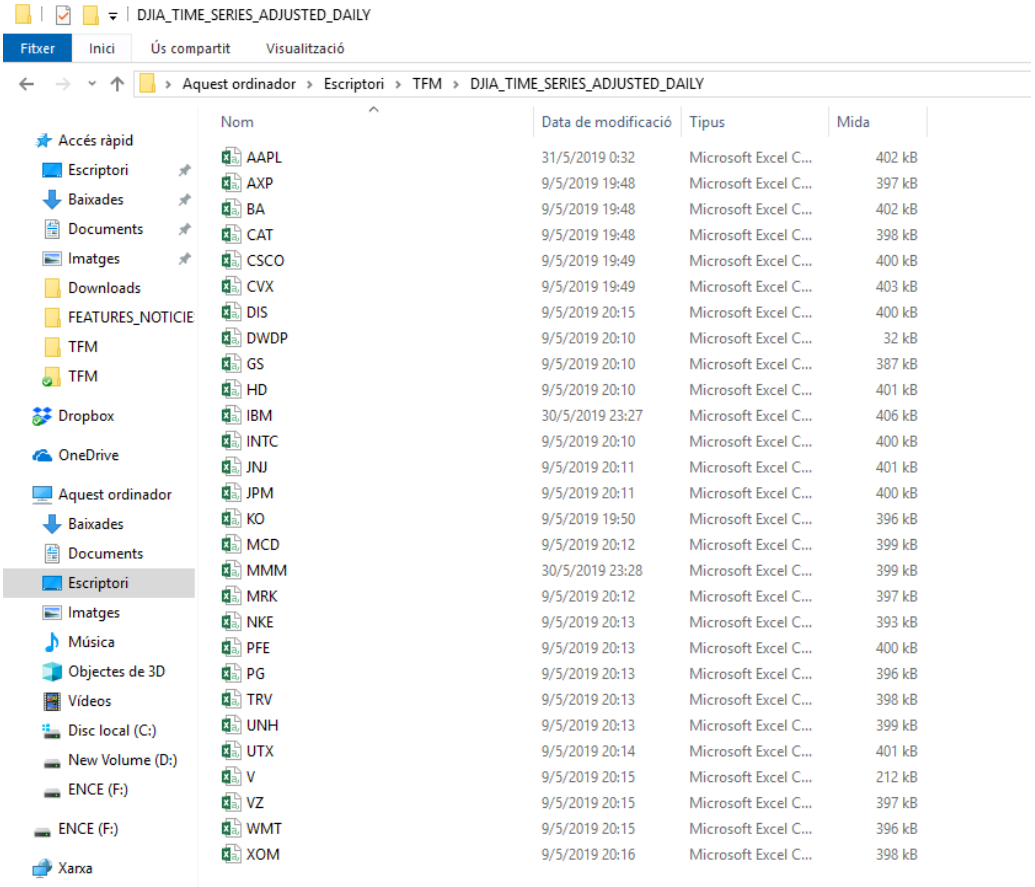
D'aquestes empreses, s'obté a través de l'API de *Alphavantage* un històric amb el valor de la cotització, amb les dades de tancament(close), obertura(open), mínim(low) i màxim(high). El període és diari i per tant, hi ha un valor per cada dia en que hi ha activitat a la borsa.

Cal destacar que per aquest estudi és fan servir dades de cotitzacions diàries, per diversos factors. El primer és que els models de ML necessiten moltes dades per poder entrenar, i si s'agafen dades amb un període de temps major com setmanes o mesos, poder resultar massa poques dades. Aquest fet és deu a que moltes empreses que avui cotitzen al DJIA poden ser recents i d'altres amb més antiguitat que hagin deixat de cotitzar a l'índex.

Per altra banda, les dades interdiàries, que són les que habitualment és fan servir, no només per la gran quantitat sinó perquè és on prefereixen invertir els inversors que fan servir aquest tipus de models, són molt poc accessibles i moltes només a canvi de pagaments.

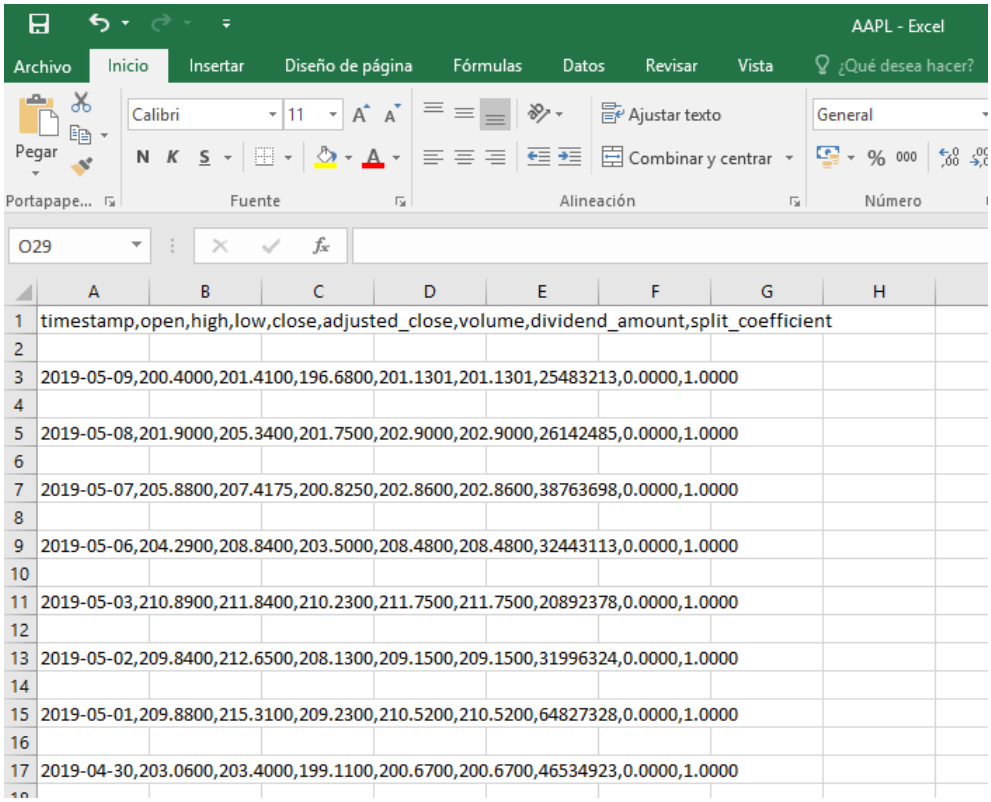
Per tot això, és va escollir treballar amb els preus diaris de les accions. Així doncs, per obtenir les dades, es fa un *python notebook*, on l'únic que cal especificar com a paràmetre és el següent.

- Les empreses desitjades, mitjançant el símbol que tenen assignat per la borsa on cotitzen. Es fa servir les del DJIA.
- Tipus de període de cotització, per exemple per aquest estudi es fa servir les cotitzacions diàries. Per tant, cal assignar el paràmetre *FUNCTION* igual a *'TIME_SERIES_DAILY_ADJUSTED*'. Cal destacar, que es fa servir els preus ajustats.
- El període temporal de cotització, que en aquest cas és des de l'inici de la cotització de l'empresa en qüestió. Ja que pot variar segons l'empresa. Per fer-ho, el paràmetre *SIZE* ha de ser igual a *'full'*.
- També cal registrar-se[151] a l'API per obtenir una clau d'accés, que s'ha de assignar al paràmetre API_KEY.
- Per últim, indicar que és vol les dades en format .csv.



Il·lustració 14.- Llistat de .csv amb les cotitzacions de les empreses del DJIA.

[151] https://www.alphavantage.co/support/#api-key



Il·lustració 15.- Exemple de .csv amb la cotització de la empresa AAPL. Notis que te les columnes *timestamp,open,high,low,close,adjusted_close,volume,dividend_amount i split_coefficient*.

6.1.2 Indicadors d'anàlisi fonamental

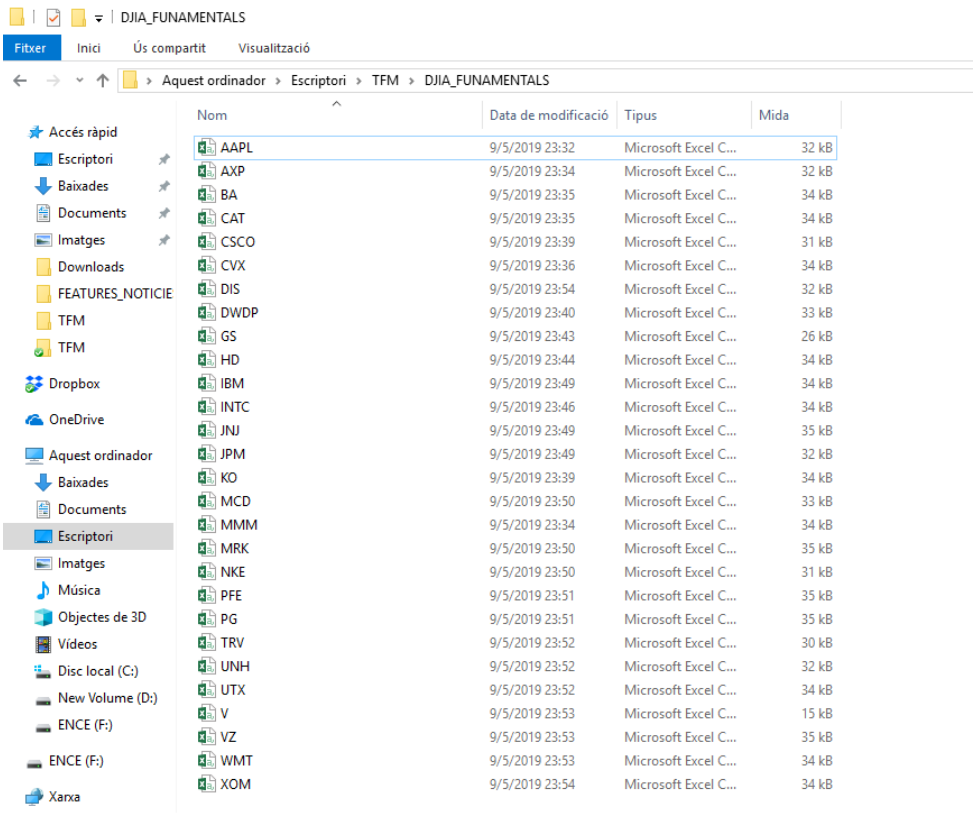
A més de les dades amb els valors de les cotitzacions, també s'obtenen les dades relacionades amb els indicadors d'anàlisi fonamental, vist en capítols anteriors.

A diferencia de la resta de dades aquestes s'obtenen manualment de la web *stockpup*[152]. El principal motiu és que aquesta no disposava de una API per poder obtenir-les, i per tant hagués resultat molt costos en temps i esforç fer-ho.

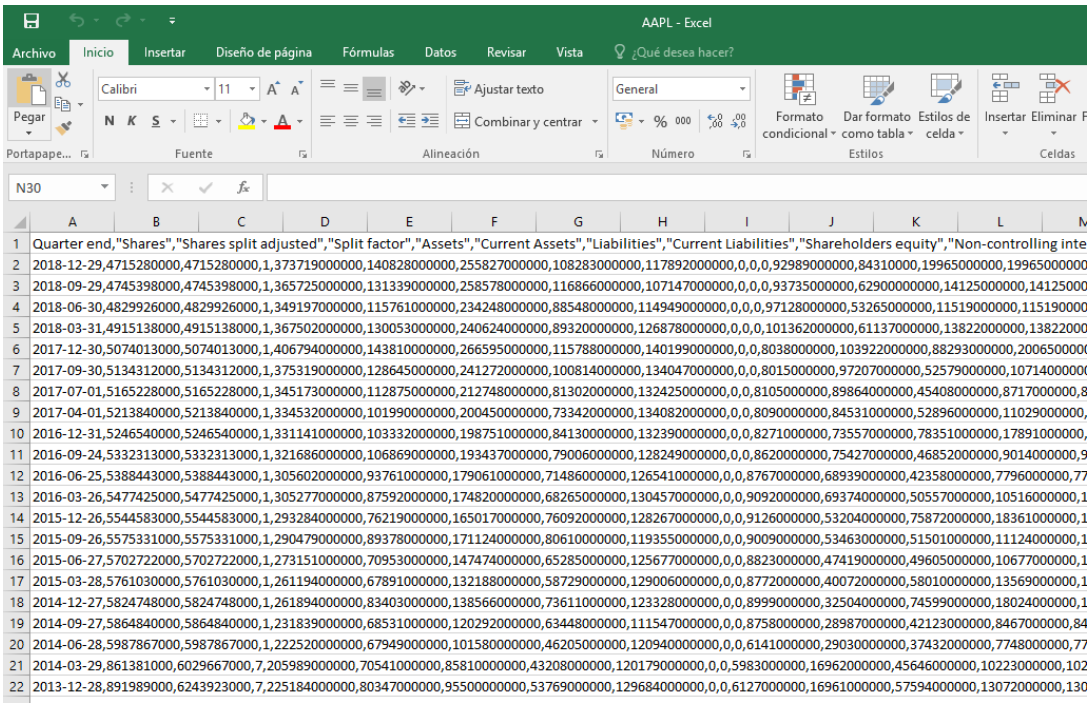
Per això, es va optar per descarregar-les manualment una a una, totes les empreses que conformen l'índex DJIA. A continuació es mostra la llista d'arxius i un en concret.

Cal recordar, que en el capítol 4.1.2 s'ha detallat quins indicadors fonamentals s'obtenen, per a cada empresa, amb les seves respectives explicacions.

[152] http://www.stockpup.com/



Il·lustració 16.- Llistat d'arxius .csv amb les dades fonamentals per a cada empresa del DJIA.

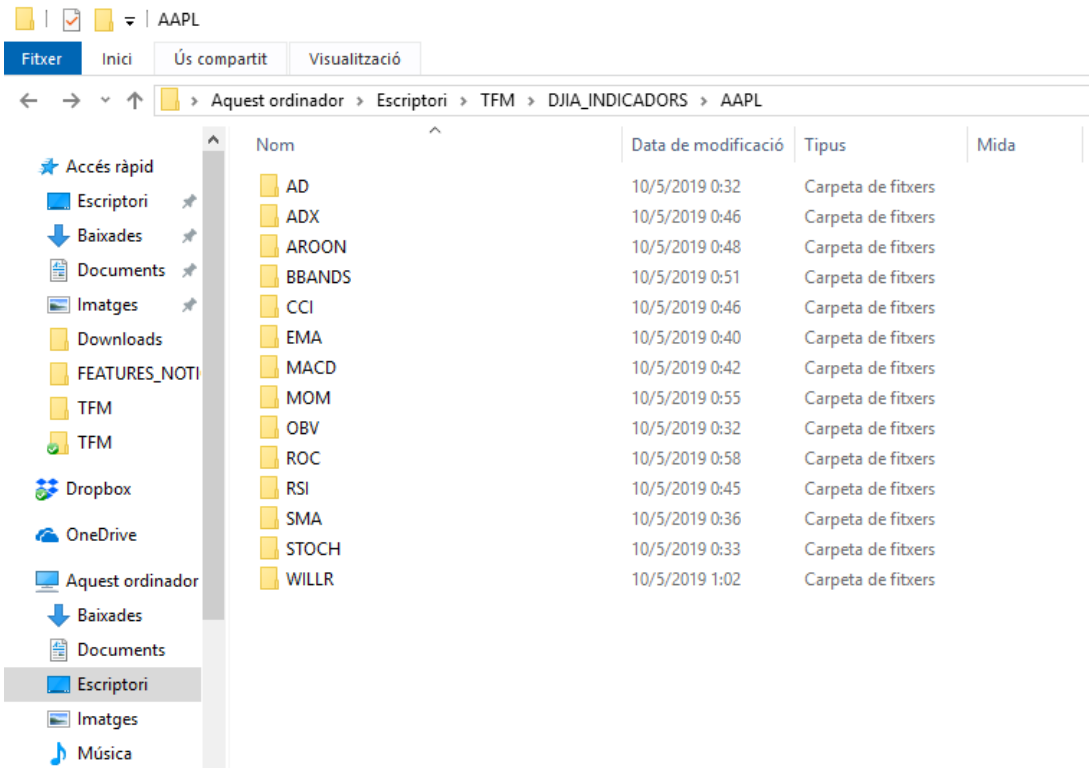


Il·lustració 17.- Exemple d'indicadors fonamentals per l'empresa AAPL. Notis que les columnes contenen els següents indicadors *Quarter end,"Shares","Shares split adjusted","Split factor","Assets","Current Assets","Liabilities","Current Liabilities","Shareholders equity","Non-controlling interest","Preferred equity","Goodwill & intangibles","Long-term debt","Revenue","Earnings","Earnings available for common stockholders","EPS basic","EPS diluted","Dividend per share","Cash from operating activities","Cash from investing activities","Cash from financing activities","Cash change during period","Cash at end of period","Capital expenditures","Price","Price high","Price low","ROE","ROA","Book value of equity per share","P/B ratio","P/E ratio","Cumulative dividends per share".*

6.1.3 Indicadors d'anàlisi tècnic

Per últim, també s'obtenen les dades relacionades amb els indicadors d'anàlisi tècnic, també desrits en capítols anteriors.

Per fer-ho es crea un altre *notebook* que fa servir l'API de *Alphavantage* per descarregar les dades i crear un *.csv*, per cada indicador de cada empresa del DJIA respectivament.



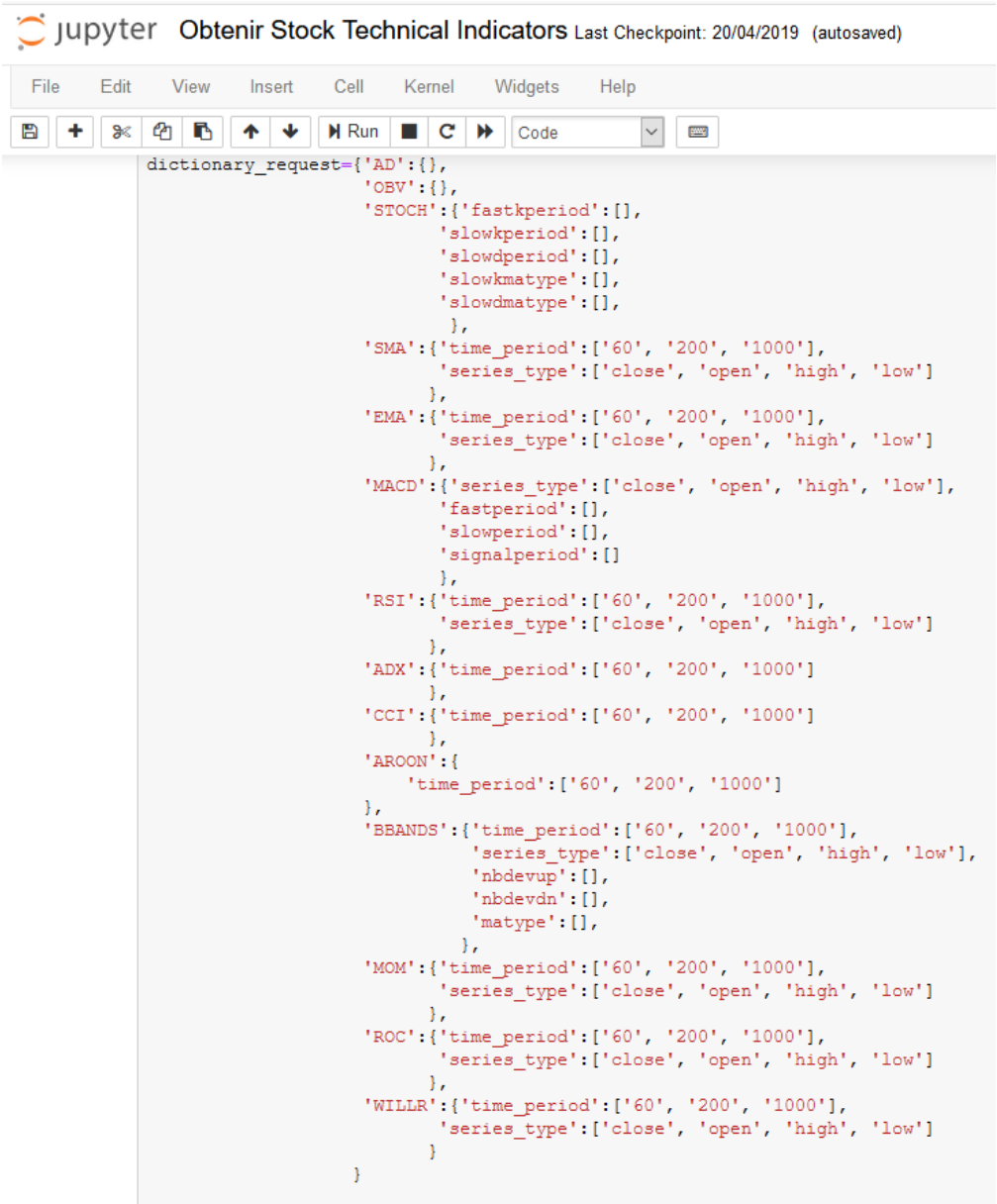
Il·lustració 18.- En aquesta il·lustració es mostra els indicadors tècnics per una empresa, AAPL.

Per obtenir les dades, es fa un *python notebook*, on l'únic que cal especificar com a paràmetre és el següent.

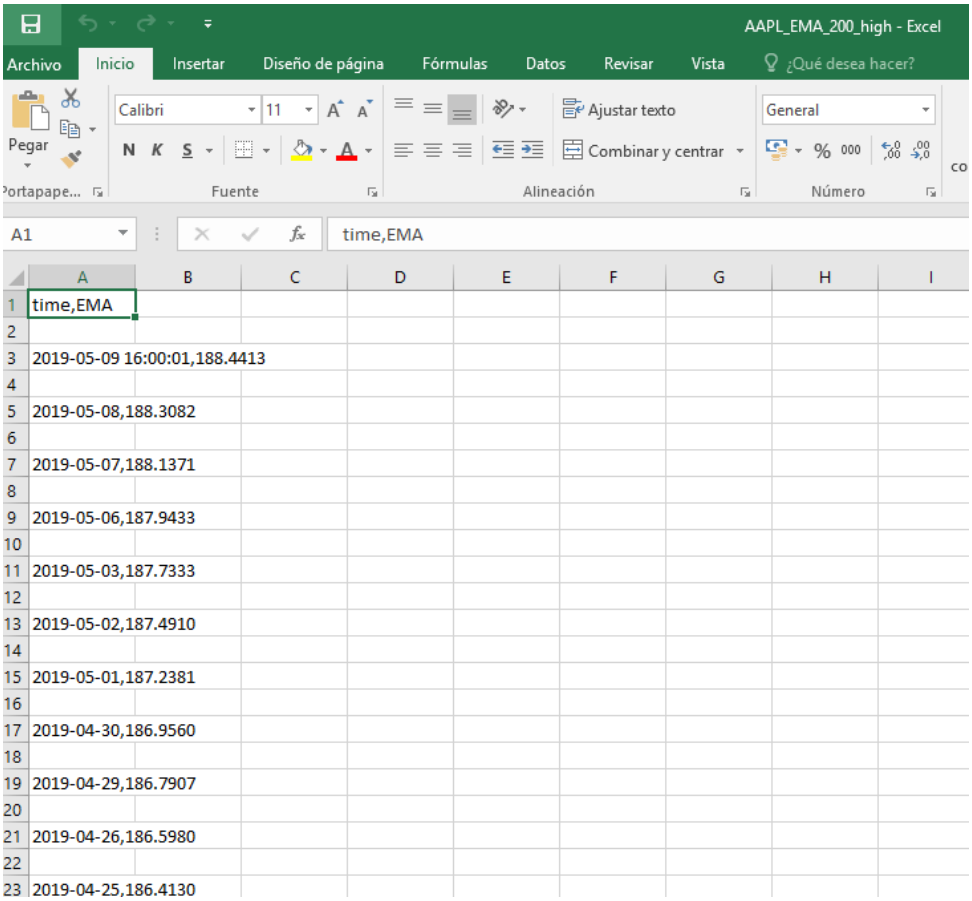
- Les empreses desitjades, mitjançant el símbol que tenen assignat per la borsa on cotitzen. Es fa servir les del DJIA, com més amunt.
- Els símbols dels indicadors. En el capítol 4.2.1, s'ha detallat quins són els indicadors tècnics més destacats i usats per aquest estudi.
- Finalment, alguns indicadors disposen de més d'un tipus de dada segons el període o el tipus de preu. Per exemple, es pot indicar que es vol la SMA de 100 dies amb el preu de tancament (*close*).

Val a dir, que per obtenir més informació dels indicadors, aquells indicadors en el que es pot obtenir diferents períodes o preus, s'ha demanat varies mostres.

Aquesta informació serà utilitzada pel model i a priori és molt difícil saber quin d'ells pot resultar més útil. Per això, s'agafen diverses mostres.



Il·lustració 19.- Imatge del *python notebook* per obtenir els indicadors tècnics. Es veu clarament quin son els indicadors, i si s'escau, quins periodes i preus, s'han obtingut.

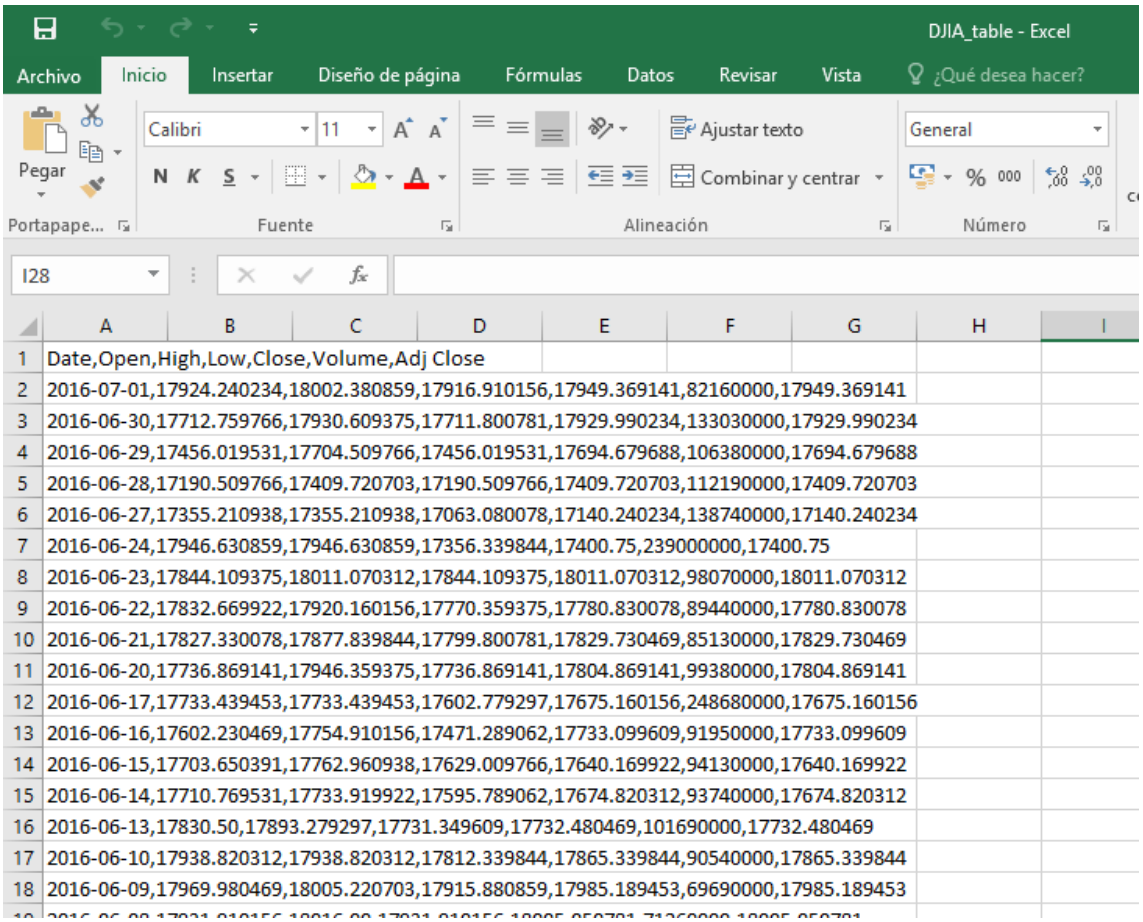


Il·lustració 20.- Exemple d'un fitxer .csv d'un indicador tècnic, de l'empresa AAPL. Notis, que aquest indicador és el EMA per un període de 200 dies i amb el preu màxim diari (*high*).

6.2 Segon cas

6.2.1 Dades històriques del DJIA

En el segon cas, es va obtenir de la mateixa forma que les dades històriques de les cotitzacions, les dades històriques del DJIA. Tot i que, com es veu a continuació, només es va poder obtenir notícies des de l'any 2008 fins al 2016, per tant només seran aquestes dates les que cal conèixer el preu de l'índex.



	A	B	C	D	E	F	G	H	I
1	Date,Open,High,Low,Close,Volume,Adj Close								
2	2016-07-01,17924.240234,18002.380859,17916.910156,17949.369141,82160000,17949.369141								
3	2016-06-30,17712.759766,17930.609375,17711.800781,17929.990234,133030000,17929.990234								
4	2016-06-29,17456.019531,17704.509766,17456.019531,17694.679688,106380000,17694.679688								
5	2016-06-28,17190.509766,17409.720703,17190.509766,17409.720703,112190000,17409.720703								
6	2016-06-27,17355.210938,17355.210938,17063.080078,17140.240234,138740000,17140.240234								
7	2016-06-24,17946.630859,17946.630859,17356.339844,17400.75,239000000,17400.75								
8	2016-06-23,17844.109375,18011.070312,17844.109375,18011.070312,98070000,18011.070312								
9	2016-06-22,17832.669922,17920.160156,17770.359375,17780.830078,89440000,17780.830078								
10	2016-06-21,17827.330078,17877.839844,17799.800781,17829.730469,85130000,17829.730469								
11	2016-06-20,17736.869141,17946.359375,17736.869141,17804.869141,99380000,17804.869141								
12	2016-06-17,17733.439453,17733.439453,17602.779297,17675.160156,248680000,17675.160156								
13	2016-06-16,17602.230469,17754.910156,17471.289062,17733.099609,91950000,17733.099609								
14	2016-06-15,17703.650391,17762.960938,17629.009766,17640.169922,94130000,17640.169922								
15	2016-06-14,17710.769531,17733.919922,17595.789062,17674.820312,93740000,17674.820312								
16	2016-06-13,17830.50,17893.279297,17731.349609,17732.480469,101690000,17732.480469								
17	2016-06-10,17938.820312,17938.820312,17812.339844,17865.339844,90540000,17865.339844								
18	2016-06-09,17969.980469,18005.220703,17915.880859,17985.189453,69690000,17985.189453								

Il·lustració 21.- Mostra de les dades històriques del valor de la cotització de l'índex borsari DJIA.

6.2.2 *Reddit WorldNews* top 25 Notícies

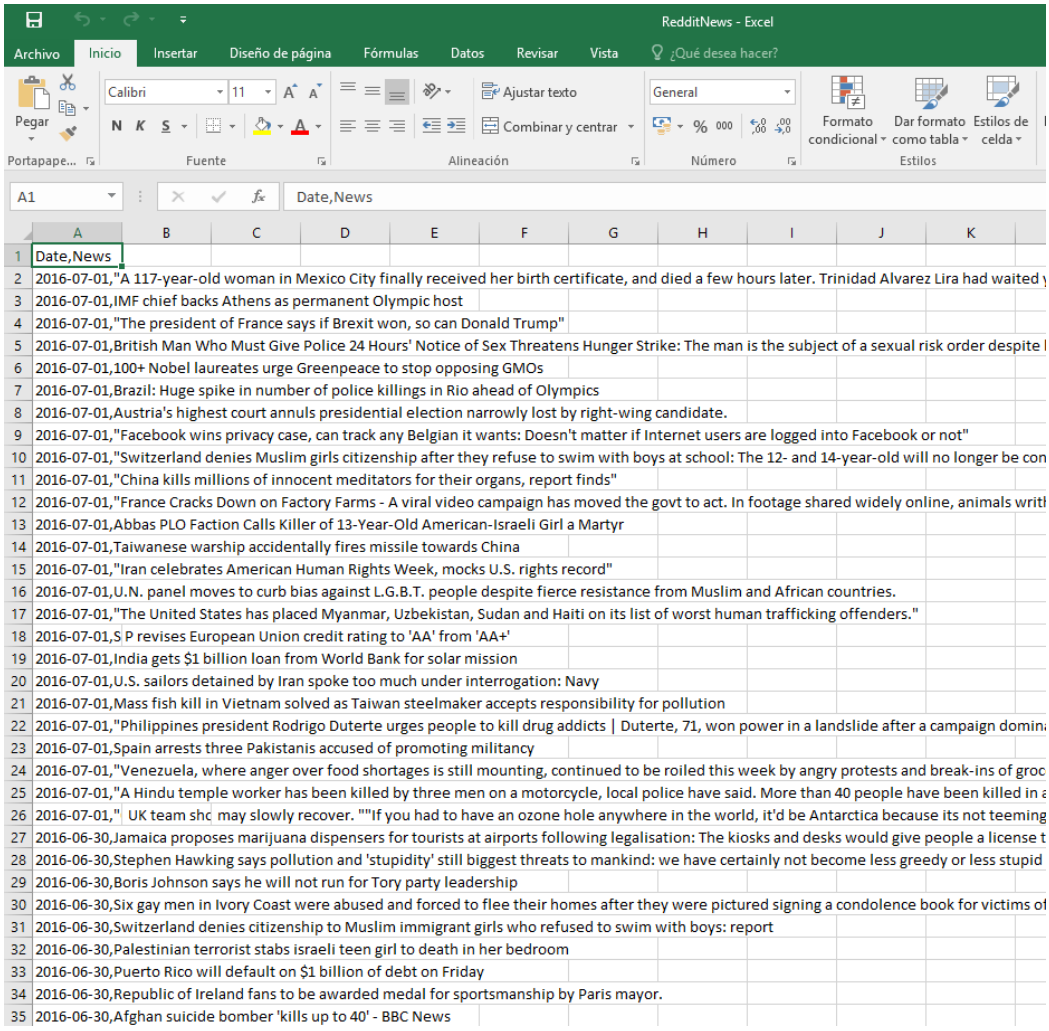
Per altra banda, per a les dades de notícies no s'ha desenvolupat cap script i s'ha buscat *datasets* disponibles i gratuïts per internet. Fet que té un impacte en el resultat i en l'abast d'aquest projecte, com es veu més endavant. Inicialment, es va pensar fer un *web scraping*[153] capaç de obtenir notícies relacionades amb les empreses a analitzar. Malauradament, va quedar fora de l'abast d'aquest estudi pel cost que tenia desenvolupar-lo.

Així, es va buscar per dades de notícies relacionades amb empreses que cotitzessin a borsa, fet que va acabar marcant el fet de treballar amb empreses que cotitzen a les borses del estats units, ja que són les empreses de les que hi ha més informació.

[153] https://en.wikipedia.org/wiki/Web_scraping

Unes dades com les que volia trobar, no són gaire accessibles si no s'està disposat a pagar per elles. Tot i així, es van trobar diversos *datasets*, especialment a *kaggle*[154], i es van comparar entre ells.

Finalment, després de veure'ls tots, i malauradament no poder obtenir d'altres, es va triar per un *dataset* que conté les 25 notícies més destacades de la pagina de *reddit*. És a dir, titulars de notícies del canal *WorldNews* de *reddit*[155]. Les notícies es classifiquen segons els vots dels usuaris de *reddit* i només es consideren els 25 primers titulars per a una data única. Les dades van des de el 2008-06-08 fins al 2016-07-01.



	A
1	Date,News
2	2016-07-01,"A 117-year-old woman in Mexico City finally received her birth certificate, and died a few hours later. Trinidad Alvarez Lira had waited
3	2016-07-01,IMF chief backs Athens as permanent Olympic host
4	2016-07-01,"The president of France says if Brexit won, so can Donald Trump"
5	2016-07-01,British Man Who Must Give Police 24 Hours' Notice of Sex Threatens Hunger Strike: The man is the subject of a sexual risk order despite
6	2016-07-01,100+ Nobel laureates urge Greenpeace to stop opposing GMOs
7	2016-07-01,Brazil: Huge spike in number of police killings in Rio ahead of Olympics
8	2016-07-01,Austria's highest court annuls presidential election narrowly lost by right-wing candidate.
9	2016-07-01,"Facebook wins privacy case, can track any Belgian it wants: Doesn't matter if Internet users are logged into Facebook or not"
10	2016-07-01,"Switzerland denies Muslim girls citizenship after they refuse to swim with boys at school: The 12- and 14-year-old will no longer be con
11	2016-07-01,"China kills millions of innocent meditators for their organs, report finds"
12	2016-07-01,"France Cracks Down on Factory Farms - A viral video campaign has moved the govt to act. In footage shared widely online, animals writ
13	2016-07-01,Abbas PLO Faction Calls Killer of 13-Year-Old American-Israeli Girl a Martyr
14	2016-07-01,Taiwanese warship accidentally fires missile towards China
15	2016-07-01,"Iran celebrates American Human Rights Week, mocks U.S. rights record"
16	2016-07-01,U.N. panel moves to curb bias against L.G.B.T. people despite fierce resistance from Muslim and African countries.
17	2016-07-01,"The United States has placed Myanmar, Uzbekistan, Sudan and Haiti on its list of worst human trafficking offenders."
18	2016-07-01,S P revises European Union credit rating to 'AA' from 'AA+'
19	2016-07-01,India gets $1 billion loan from World Bank for solar mission
20	2016-07-01,U.S. sailors detained by Iran spoke too much under interrogation: Navy
21	2016-07-01,Mass fish kill in Vietnam solved as Taiwan steelmaker accepts responsibility for pollution
22	2016-07-01,"Philippines president Rodrigo Duterte urges people to kill drug addicts | Duterte, 71, won power in a landslide after a campaign domin
23	2016-07-01,Spain arrests three Pakistanis accused of promoting militancy
24	2016-07-01,"Venezuela, where anger over food shortages is still mounting, continued to be roiled this week by angry protests and break-ins of groc
25	2016-07-01,"A Hindu temple worker has been killed by three men on a motorcycle, local police have said. More than 40 people have been killed in
26	2016-07-01," UK team shc may slowly recover. ""If you had to have an ozone hole anywhere in the world, it'd be Antarctica because its not teeming
27	2016-06-30,Jamaica proposes marijuana dispensers for tourists at airports following legalisation: The kiosks and desks would give people a license t
28	2016-06-30,Stephen Hawking says pollution and 'stupidity' still biggest threats to mankind: we have certainly not become less greedy or less stupid
29	2016-06-30,Boris Johnson says he will not run for Tory party leadership
30	2016-06-30,Six gay men in Ivory Coast were abused and forced to flee their homes after they were pictured signing a condolence book for victims of
31	2016-06-30,Switzerland denies citizenship to Muslim immigrant girls who refused to swim with boys: report
32	2016-06-30,Palestinian terrorist stabs israeli teen girl to death in her bedroom
33	2016-06-30,Puerto Rico will default on $1 billion of debt on Friday
34	2016-06-30,Republic of Ireland fans to be awarded medal for sportsmanship by Paris mayor.
35	2016-06-30,Afghan suicide bomber 'kills up to 40' - BBC News

Il·lustració 22.- Imatge del *dataset* de notícies de *reddit on apareix les 25 notícies més destacades de la pagina*.

Cal destacar que com es pot veure a la il·lustració prèvia, les notícies poden no estar relacionades amb les empreses que cotitzen al DJIA. Per tant, no tindria sentit classificar les notícies per empreses.

Per aquest motiu, es fa servir per predir el DJIA en el conjunt, ja que moltes notícies són globals i afecten a totes les empreses que cotitzen al índex borsari.

[154] https://www.kaggle.com
[155] https://www.reddit.com/r/worldnews/

7 Desenvolupament

7.1 Primer cas

Un cop ja s'han obtingut les diferents dades, mitjançant els respectius *notebooks*, aquestes s'han de tractar per poder-les utilitzar com a input del model de ML.

Per una banda s'ha fet servir la llibreria *Pandas*. Com s'ha vist prèviament, és una llibreria de programació escrita per al llenguatge de programació *Python* per a la manipulació i anàlisi de dades. En particular, ofereix estructures de dades i operacions per manipular taules numèriques i sèries temporals. Per això, és ideal per l'objectiu d'agrupar i classificar les dades. A més, és una llibreria amb una corba ràpida d'aprenentatge i molt potent.

7.1.1 Pandas

Així, primerament es comença ajuntant les dades dels preus de les cotitzacions, amb els indicadors tècnics i els fonamentals. S'agrupa tot en una taula agrupada per empreses i organitzada per dies. Abans, però s'ha de llegir les dades dels fitxers csv. que s'han obtingut.

En primer lloc, el llegeix els fitxers que contenen els preus de les empreses del DJIA. Com es va mencionar prèviament, es fa servir el preu ajustat. A més, el preu ajustat, a partir d'ara, *adjusted_close* és el preu de tancament. I amb aquest preu és a partir del qual és fa les prediccions, i per tant, la *label*.

Per això, per calcular la *label* abans primer es resta la diferencia del *adjusted_close* d'un dia respecte del anterior. Aquesta, s'inclou en una nova columna anomenada *diff*, com es mostra a la taula següent.

	NOM_EMPRESA	timestamp	old_adjusted_close	open	high	low	close	adjusted_close	volume	dividend_amount	split_coefficient	diff
0	AAPL	1998-01-02	NaN	13.63	16.25	13.50	16.25	0.5103	6411700	0.0	1.0	NaN
1	AAPL	1998-01-05	0.5103	16.50	16.56	15.19	15.88	0.4987	5820300	0.0	1.0	-2.326048
2	AAPL	1998-01-06	0.4987	15.94	20.00	14.75	18.94	0.5948	16182800	0.0	1.0	16.156691
3	AAPL	1998-01-07	0.5948	18.81	19.00	17.31	17.50	0.5496	9300200	0.0	1.0	-8.224163
4	AAPL	1998-01-08	0.5496	17.44	18.62	16.94	18.19	0.5713	6910900	0.0	1.0	3.798355
5	AAPL	1998-01-09	0.5713	18.12	19.37	17.50	18.19	0.5713	7915600	0.0	1.0	0.000000
6	AAPL	1998-01-12	0.5713	17.44	18.62	17.12	18.25	0.5731	4610700	0.0	1.0	0.314081
7	AAPL	1998-01-13	0.5731	18.62	19.62	18.50	19.50	0.6124	5686200	0.0	1.0	6.417374
8	AAPL	1998-01-14	0.6124	19.87	19.94	19.25	19.75	0.6203	5261300	0.0	1.0	1.273577
9	AAPL	1998-01-15	0.6203	19.19	19.75	18.62	19.19	0.6027	4993500	0.0	1.0	-2.920192

Il·lustració 23.- Taula amb les dades històriques del preu ajustat de les accions del DJIA.

En segon lloc, es fa el mateix procediment per obtenir les dades dels indicadors fonamentals i tècnics. Després, s'ajunten en una mateixa taula, agrupats per accions i per la data.

```
In [22]: print(df_preus_indicadors.NOM_EMPRESA.unique())
         print(len(df_preus_indicadors.NOM_EMPRESA.unique()))

['AAPL' 'AXP' 'BA' 'CAT' 'CSCO' 'CVX' 'DIS' 'DWDP' 'GS' 'HD' 'IBM' 'INTC'
 'JNJ' 'JPM' 'KO' 'MCD' 'MMM' 'MRK' 'NKE' 'PFE' 'PG' 'TRV' 'UNH' 'UTX' 'V'
 'VZ' 'WMT' 'XOM']
28
```

Il·lustració 24.- Imatge en que es mostra les empreses de les que s'han afegit dades a la taula final.

	Quarter end	Shares	Shares split adjusted	Split factor	Assets	Current Assets	Liabilities	Current Liabilities	Shareholders equity	Non-controlling interest	Preferred equity	Goodwill & intangibles	
0	2018-12-29	4715280000	4715280000	1.0	373719000000	140828000000	255827000000	108283000000	117892000000	0	0	0	9
1	2018-09-29	4745398000	4745398000	1.0	365725000000	131339000000	258578000000	116866000000	107147000000	0	0	0	9
2	2018-06-30	4829926000	4829926000	1.0	349197000000	115761000000	234248000000	88548000000	114949000000	0	0	0	9
3	2018-03-31	4915138000	4915138000	1.0	367502000000	130053000000	240624000000	89320000000	126878000000	0	0	0	10
4	2017-12-30	5074013000	5074013000	1.0	406794000000	143810000000	266595000000	115788000000	140199000000	0	0	8038000000	10
5	2017-09-30	5134312000	5134312000	1.0	375319000000	128645000000	241272000000	100814000000	134047000000	0	0	8015000000	9
6	2017-07-01	5165228000	5165228000	1.0	345173000000	112875000000	212748000000	81302000000	132425000000	0	0	8105000000	8
7	2017-04-01	5213840000	5213840000	1.0	334532000000	104000000000	200450000000	73242000000	134082000000	0	0	8000000000	8

Il·lustració 25.- Taula amb les dades fonamentals.

	time	Chaikin A/D_	ADX_1000	ADX_200	ADX_60	Aroon Down_1000	Aroon Up_1000	Aroon Up_200	Aroon Down_200	Aroon Down_60	Aroon Up_60	Real Lower Band_1000_close	Real Middle Band_1000_close	Re Band_10(
0	2019-05-08	3.662993e+09	6.4663	6.3415	13.2430	24.9	85.2	57.0	26.0	0.0000	91.6667	72.9553	146.0890	2
1	2019-05-07	3.672387e+09	6.4674	6.3674	13.3337	25.0	85.3	57.5	26.5	0.0000	93.3333	72.9642	146.0163	2
2	2019-05-06	3.687219e+09	6.4684	6.3935	13.4259	25.1	85.4	58.0	27.0	1.6667	95.0000	72.9706	145.9422	2
3	2019-05-03	3.659151e+09	6.4695	6.4162	13.4818	25.2	85.5	58.5	27.5	3.3333	96.6667	72.9906	145.8627	2
4	2019-05-02	3.640594e+09	6.4709	6.4301	13.4375	25.3	85.6	59.0	28.0	5.0000	98.3333	73.0134	145.7770	2
5	2019-05-01	3.658149e+09	6.4723	6.4441	13.3925	25.4	85.7	59.5	28.5	0.0000	100.0000	73.0299	145.6937	2
6	2019-04-30	3.695468e+09	6.4738	6.4568	13.3297	25.5	85.8	60.0	29.0	0.0000	93.3333	73.0513	145.6095	2
7	2019-04-29	3.708159e+09	6.4747	6.4845	13.4080	25.6	85.9	60.5	29.5	0.0000	95.0000	73.0531	145.5364	2
8	2019-04-26	3.714579e+09	6.4758	6.5061	13.4116	25.7	86.0	61.0	30.0	0.0000	96.6667	73.0589	145.4571	2
9	2019-04-25	3.704995e+09	6.4769	6.5290	13.4275	25.8	86.1	61.5	30.5	0.0000	98.3333	73.0640	145.3778	2

Il·lustració 26.- Taula amb les dades dels indicadors tècnics.

7.1.2 LABEL

Un cop afegides les empreses a la mateixa taula, el següent pas afegir la *label*. Així, el que es fa es comprovar si la diferencia, a la columna *diff*, és positiva o negativa, i canviar-la per 2 o 0 respectivament, en una nova columna anomenada *label.*

WILLR_60_low	WILLR_60_open	old_adjusted_close	open	high	low	close	adjusted_close	volume	dividend_amount	split_coefficient	diff	LABEL
0.0000	0.0000	202.9000	200.4000	201.4100	196.6800	201.1301	201.1301	25483213	0.00	1.0	-0.879978	0
-27.0194	-27.0194	202.8600	201.9000	205.3400	201.7500	202.9000	202.9000	26142485	0.00	1.0	0.019714	2
-27.0300	-27.0300	208.4800	205.8800	207.4175	200.8250	202.8600	202.8600	38763698	0.00	1.0	-2.770384	0
-14.5660	-14.5660	211.7500	204.2900	208.8400	203.5000	208.4800	208.4800	32443113	0.00	1.0	-1.568496	0
-7.5922	-7.5922	209.1500	210.8900	211.8400	210.2300	211.7500	211.7500	20892378	0.00	1.0	1.227863	2
-13.1371	-13.1371	210.5200	209.8400	212.6500	208.1300	209.1500	209.1500	31996324	0.00	1.0	-0.655032	0
-10.2154	-10.2154	200.6700	209.8800	215.3100	209.2300	210.5200	210.5200	64827328	0.00	1.0	4.678890	2
-18.9563	-18.9563	204.6100	203.0600	203.4000	199.1100	200.6700	200.6700	46534923	0.00	1.0	-1.963423	0
-9.0952	-9.0952	204.3000	204.4000	205.9700	203.8600	204.6100	204.6100	22204716	0.00	1.0	0.151508	2

Il·lustració 27.- Taula amb el preu, els indicadors tècnics i fonamentals, i la columna label.

7.1.3 Train, Validation i Test

Per últim, cal separar les dades en entrenament, validació i test. Les primeres les fa servir l'algorisme per entrenar i reduir la funció de pèrdua del conjunt d'entrenament. Alhora, el de validació hauria de disminuir paral·lelament i la seva precisió hauria d'augmentar.

En cas de que es continuï entrenant, pot arribar un moment en que els resultats de l'entrenament i de validació difereixin, la precisió de l'entrenament continuarà augmentant mentre que la validació començarà a disminuir lleugerament. En aquest moment està passant el sobre entrenament, és a dir, el model ha aprés massa detall de les dades d'entrenament i no es capaç de generalitzar a altres dades amb bons resultats.

Un cop acabat l'entrenament, és el moment de fer l'optimització de paràmetres i un cop escollit uns, predir amb les dades de test per última vegada.

És importat saber que les dades de test només és poden fer servir al final, ja que si no, es pot estar manipulant el resultat final del model.

Per separar el *dataset* en *train, validation,* i *test*, es fa una separació tal que així. Com que no es disposa d'una gran quantitat de dades, és dedica quasi el 90% per entrenar. I la resta es divideix en dos parts per validar i testejar un cop al final. Per fer-ho es separa les dades des de l'inici fins 2017-01-01, des de 2017-01-01 fins a 2018-03-01, i des de 2018-03-01 fins les meves recents.

```
In [42]: print('train: %.2f'% float(df_preus_indicadors_train.shape[0]*100/df_preus_indicadors.shape[0]), '%')
         print('val: %.2f'% float(df_preus_indicadors_val.shape[0]*100/df_preus_indicadors.shape[0]), '%')
         print('test: %.2f'% float(df_preus_indicadors_test.shape[0]*100/df_preus_indicadors.shape[0]), '%')

         train: 88.79 %
         val: 5.53 %
         test: 5.68 %
```

Il·lustració 28.- Representació del *dataset* dividit en les tres categories.

```
In [43]: print(Counter(df_preus_indicadors_train['LABEL']))
         print(Counter(df_preus_indicadors_val['LABEL']))
         print(Counter(df_preus_indicadors_test['LABEL']))

         Counter({1.0: 65376, 0.0: 60755})
         Counter({1.0: 4286, 0.0: 3571})
         Counter({1.0: 4290, 0.0: 3780})
```

Il·lustració 29.- Divisió de les dades en les tres categories segons la *label*.

```
In [71]: df_preus_indicadors_train.shape
Out[71]: (126131, 174)
```

Il·lustració 30.- Nombre de dades per entrenar en files i columnes. Les files representen els dies disponibles i les columnes els indicadors.

7.2 Segon cas

7.2.1 Notícies

Primerament, es llegeixen les notícies i de la mateixa manera que abans es posen en una taula fent servir la llibreria *Pandas*.

	Date	News
0	2016-07-01	A 117-year-old woman in Mexico City finally re...
1	2016-07-01	IMF chief backs Athens as permanent Olympic host
2	2016-07-01	The president of France says if Brexit won, so...
3	2016-07-01	British Man Who Must Give Police 24 Hours Noti...
4	2016-07-01	100+ Nobel laureates urge Greenpeace to stop o...

Il·lustració 31.- Exemple de les 5 primeres notícies de la taula que conté les notícies importades del fitxer .csv.

Com es veia més amunt, les notícies s'agrupen per dies i cada un conté les 25 notícies més destacades de la pagina de *reddit* per aquell dia. D'aquesta manera, queda una taula amb 25 notícies per dia.

```
In [8]: df_news[df_news['Date'] == '2009-09-15']
```

Out[8]:

	Date	News
61989	2009-09-15	The Church of Scientology won\t be dissolved i...
61990	2009-09-15	New virus from rats can kill 80 per cent of hu...
61991	2009-09-15	The gruesome spectacle of dolphins being slaug...
61992	2009-09-15	The End of Innocence in Afghanistan: \The Germ...
61993	2009-09-15	France approves Internet piracy bill
61994	2009-09-15	The Rural Doctors Association says right now t...
61995	2009-09-15	Al Jazeera English - Africa - Shabab to avenge...
61996	2009-09-15	How Sri Lanka governs through detentions - Sri...
61997	2009-09-15	Two months after the Pakistani Army wrested co...
61998	2009-09-15	N. Korean cargo ship repels pirates off Somali...
61999	2009-09-15	Canada: Stephen Harper attempts to prop up reg...
62000	2009-09-15	New geological data provides hope for claims b...
62001	2009-09-15	Ukraine set to block Elton John adoption becau...
62002	2009-09-15	German plane makes emergency landing
62003	2009-09-15	Street artist catches chief of the Israeli arm...
62004	2009-09-15	In an equine echo of the controversy surroundi...
62005	2009-09-15	UPDATE: 5-New York homes raided in terrorism p...
62006	2009-09-15	Population Growth Impeding Progress on the MDGs?
62007	2009-09-15	Global Population to Reach 7 Billion by 2011
62008	2009-09-15	Government Funded Feminist Porn
62009	2009-09-15	Can someone enlighten me re:Holy Land disputes?
62010	2009-09-15	Human Rights Watch official suspended for coll...

Il·lustració 32.- Exemple de les notícies pel dia "2009-09-15".

7.2.2 DJIA

També s'ha de llegir i obtenir les dades de preu històriques de l'índex DJIA, per fer la predicció amb noticies. Del fitxer .csv s'obtenen aquestes dades i es guarden en una taula, fent servir la llibreria *Pandas*.

De la mateixa manera que en el primer cas, es calcula el label amb el preu *adjusted_close.* Però en aquest cas, s'identifica com que ha pujat respecte al dia anterior amb un *True* a la columna *Label*, o un *False*, en cas contrari.

Finalment, s'esborren els preus ja que no es tenen en compte per entrenar l'algorisme de ML i s'adjunta a la taula de notícies, fent una agregació per la data.

	Date	News	LABEL
0	2008-08-08	150 Russian tanks have entered South Ossetia w...	True
1	2008-08-08	Afghan children raped with impunity, U.N. offi...	True
2	2008-08-08	Al-Qaeda Faces Islamist Backlash	True
3	2008-08-08	Announcing:Class Action Lawsuit on Behalf of A...	True
4	2008-08-08	BREAKING: Musharraf to be impeached.	True
5	2008-08-08	Breaking: Georgia invades South Ossetia, Russi...	True
6	2008-08-08	Caucasus in crisis: Georgia invades South Ossetia	True
7	2008-08-08	China tells Bush to stay out of other countrie...	True
8	2008-08-08	Condoleezza Rice: The US would not act to prev...	True
9	2008-08-08	Did World War III start today?	True
10	2008-08-08	Did the U.S. Prep Georgia for War with Russia?	True

Il·lustració 33.- Mostra de les noticies un cop afegida la columna Label, extreta de les dades històriques de preus del DJIA.

7.2.3 BERT

Per altra banda, amb el *dataset* de notícies s'ha fet servir el BERT com a codificador de paraules de text, a dades que poden ser usades per algorismes de ML. Aquest converteix cada noticia en 511 codis, en la sortida de la seva xarxa neuronal. Així, a la taula s'afegeix a cada notícia els seus respectius codis en columnes, juntament a la columnes de data i *Label.*

	Date	News	0	1	2	3	4	5	6	7	...	502	503	504	505	
0	2016-07-01	A 117-year-old woman in Mexico City finally re...	0.009438	0.023999	-0.056785	0.077329	0.040650	0.015022	-0.080607	0.049994	...	0.046310	0.048117	0.043167	-0.094162	0.03
1	2016-07-01	IMF chief backs Athens as permanent Olympic host	-0.026954	-0.035928	-0.040790	0.028364	-0.017530	0.013404	-0.059126	0.016241	...	-0.053508	-0.000010	0.005058	-0.068459	-0.00
2	2016-07-01	The president of France says if Brexit won, so...	-0.050344	0.015471	0.002197	0.012018	0.087000	0.045529	-0.032218	0.017832	...	-0.036459	0.020173	0.057131	-0.064132	-0.07

Il·lustració 34.- Mostra de notícies amb la seves respectives codificacions a la sortida del BERT.

	Date	News	LABEL	0	1	2	3	4	5	6	...	502	503	504	
0	2008-08-08	150 Russian tanks have entered South Ossetia w...	True	-0.010008	-0.049946	-0.039040	0.044493	-0.051375	-0.033865	-0.010494	...	0.028300	0.028869	-0.023994	-0.07
1	2008-08-08	Afghan children raped with impunity, U.N. offi...	True	0.057858	-0.023948	-0.052333	0.074174	0.002603	0.008861	0.012562	...	0.020418	-0.002361	0.025968	-0.07
2	2008-08-08	Al-Qaeda Faces Islamist Backlash	True	0.032738	0.004100	-0.028606	0.068931	-0.030323	0.013643	0.039414	...	0.014873	-0.054470	0.031471	-0.08
3	2008-08-08	Announcing:Class Action Lawsuit on Behalf of A...	True	0.019158	-0.036425	-0.007986	-0.008051	-0.005715	-0.007488	0.005220	...	-0.022368	0.023669	0.064597	-0.11
4	2008-08-08	BREAKING: Musharraf to be impeached.	True	0.071786	-0.017450	-0.055387	0.030300	-0.070433	0.046062	-0.024634	...	-0.066330	0.005785	0.057112	-0.06

Il·lustració 35.- Mostra de notícies amb la seves respectives codificacions a la sortida del BERT, afegida la columna Label.

8 Anàlisi dels resultats

En aquest capítol s'analitza i es comenta els resultats obtinguts. Tots els resultats han anat variant considerablement al llarg del desenvolupament del projecte, i s'han provat diferents algorismes i estratègies que aquí s'exposen.

8.1 Primer cas

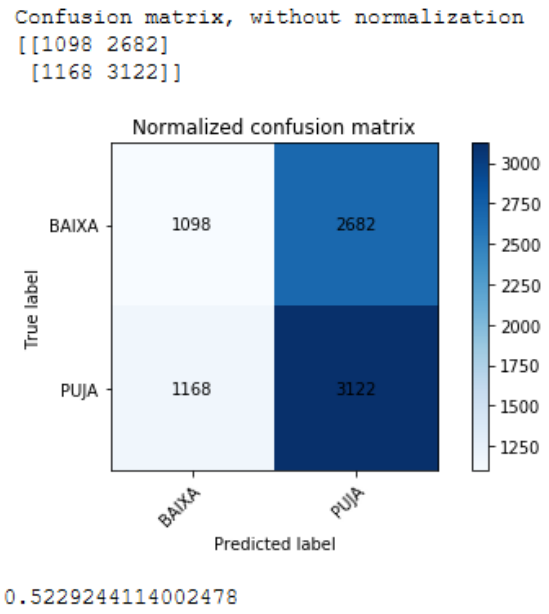
En primer lloc, cal preparar les dades abans de ser usades per un algorisme de ML. La majoria de algorismes, requereixen una normalització de les dades per ser tractades. Per aquest estudi, s'ha decidit utilitzar un XGBOOST per a fer la predicció.

El primer motiu és que al tenir moltes dades, provinents d'anàlisis força diferents entre ells, i aquest algorisme es força flexible amb les dades que li entren per funcionar. El segon motiu i més important és els bons resultats d'aquest algorisme en els darrers anys, com es veu en el capítol de revisió de l'estat de l'art, especialment quan es disposa de una quantitat no molt gran de dades, com es el cas. El tercer és que l'XGBOOST és un algorisme que pot córrer sense un cost computacional molt elevat.

Finalment abans de passar les dades dels indicadors al XGBOOST, per alguns es va fer una normalització per facilitar les coses. Per exemple, alguns indicadors estan formats per valors entre 0 i 100, en canvi d'altres, estan sumant o multiplicant per valors com el preu i el volum. Doncs, aquests es dividien per les dades corresponents per tenir dades normalitzades.

8.1.1 XBOOST

Després de fer moltes proves i optimitzar paràmetres, el resultat va ser el següent, on es veu la matriu de confusió resultat i el valor de l'*accuracy* (consultar taula 2). Es fa evident que el resultat es dolent i fins i tot es pot dir que el XGBOOST no ha sigut capaç de trobar cap relació entre el moviment de preu de les accions i els indicadors tècnics i fonamentals.



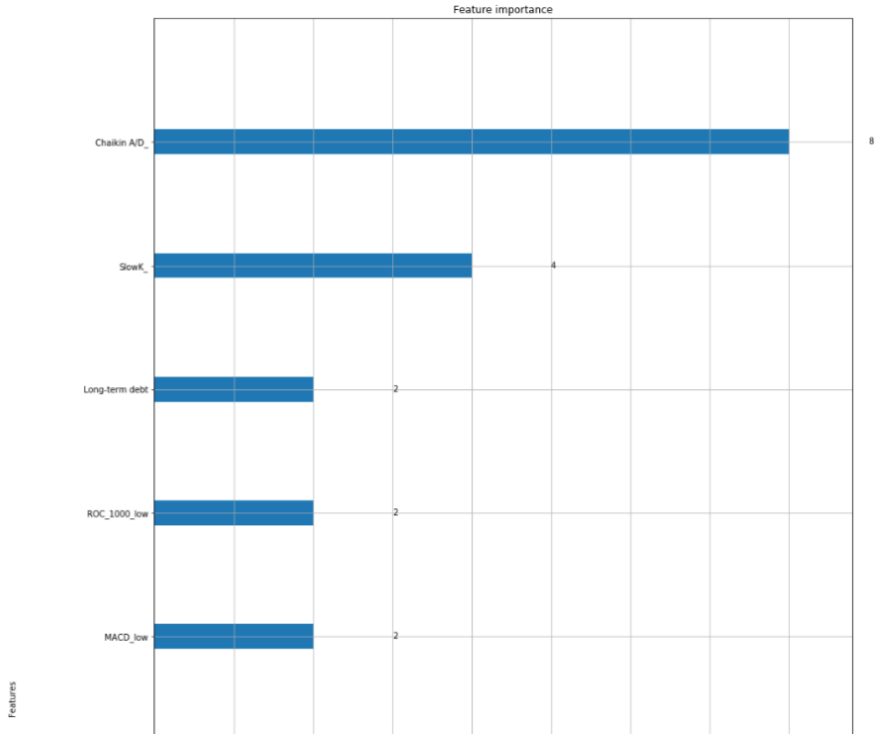
Il·lustració 36.- Resultat de la predicció per les dates de test. Es mostra una matriu de confusió i el valor de predicció, igual a 52,29%.

En el següent capítol, es justifica aquest resultat i es comenta els diversos escenaris i problemes que han sorgit durant el desenvolupament del cas. Tot i així, cal destacar dos factors.

El primer és que no té gaire sentit fer una gran optimització dels paràmetres del algorisme, ja que aquesta només és capaç de millorar o empitjorar "lleugerament" el resultat, però no de fer-lo canviar de forma significativa. I en aquest cas, seria necessari revisar els paràmetres d'entrada més que optimitzar, per que el que cal es canviar el resultat significativament.

En segon lloc, no es deixa que el algorisme entreni durant gaires iteracions, ja que te tendència a sobre entrenar i els resultats empitjoren ràpidament.

Els arbres de decisió no només es poden visualitzar per inspeccionar el camí de decisió d'una característica determinada, també es pot mostrar un resum de la contribució de cada *feature* al model, adequat a les dades d'entrenament.



Il·lustració 37.- Mostra la importància de les *features* del XGBOOST per a la predicció. La més important resulta ser l'indicador tècnic *Chaikin A/D.*

8.2 Segon cas

En el segon cas, els resultats han sigut una mica superior al primer. Com abans, en el següent capítol es comenta i justifica el resultat, així com els problemes que han anat apareixen durant el desenvolupament del cas.

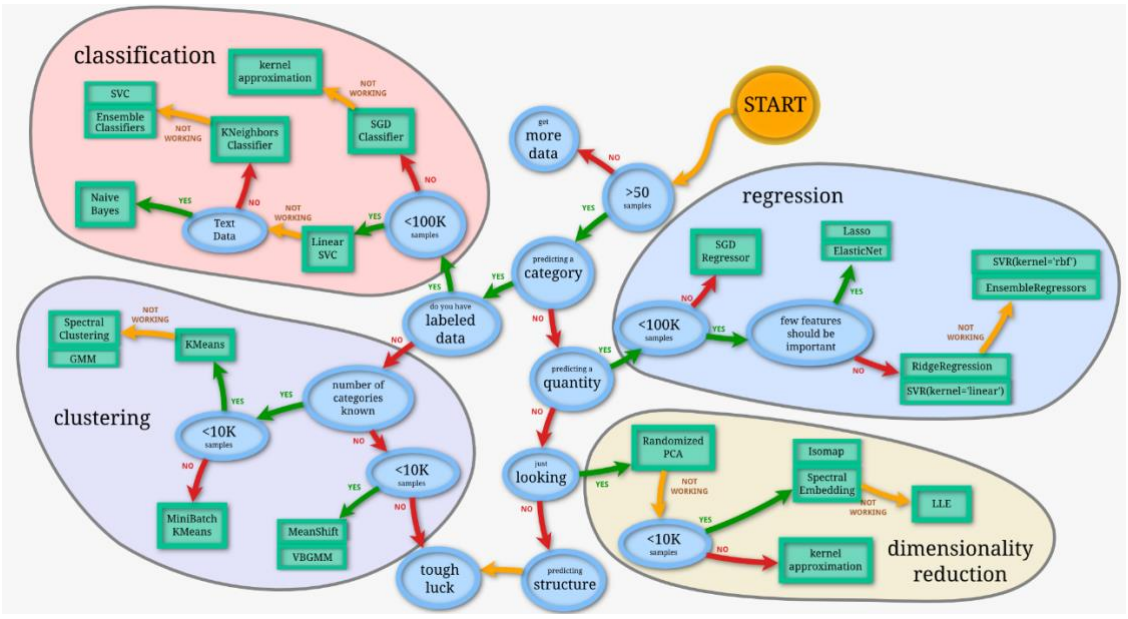
8.2.1 SVM

Així, el resultat predit l'índex DJIA amb notícies dona un lleuger millor resultat. Per predir, es fa ser un *C-Support Vector Classification*[156].

Sovint, la part més difícil de resoldre un problema de ML pot ser trobar l'algorisme adequat per al treball. Els diferents algorismes són més adequats per a diferents tipus de dades i per a diferents problemes.

El diagrama de flux següent està extret de la llibreria *scikit-learn*[157] i és una petita guia sobre com abordar els problemes pel que fa al la elecció d'algorismes de ML.

Seguint-la es va trobar que el que millor podria funcionar per aquest estudi, serien els algorismes classificadors *SVM*[158].



Il·lustració 38.- Diagrama de flux dissenyat per donar als usuaris de la llibreria *scikit-learn* una petita guia aproximada sobre com abordar els problemes pel que fa als algorisme de ML.

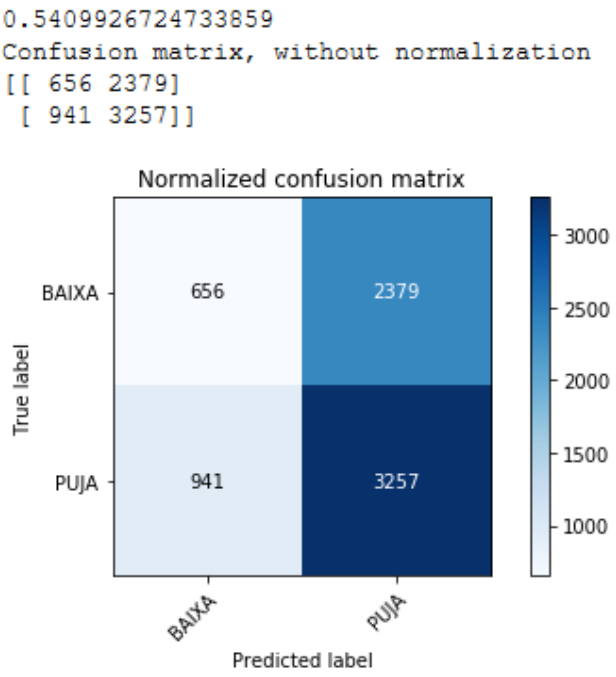
Doncs, es van provar diversos algorismes SVM i els resultats van ser similirs, tot i així el que millor resultat va donar va ser el mencionat prèviament, un *C-Support Vector Classification*.

[156] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
[157] https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html
[158] https://en.wikipedia.org/wiki/Support-vector_machine

Cal destacar, que en el següent capítol, problemes observats, es fa menció a que la primera intenció va ser la de fer servir una xarxa neuronal per fer aquesta predicció. El motiu és que al fer servir el BERT com a codificador, que és una xarxa neuronal, les dades resultants d'aquest estaven normalitzades per fer-les servir com a entrada d'una xarxa neuronal.

Però, es va provar amb diverses mides i xarxes neuronals diferents, i sempre van aparèixer problemes tècnics relacionats amb la memòria disponible del ordinador. Per tant, es va haver de descartar aquest tipus de algorismes, tot i que podrien donar un millor resultat. Una solució, com es veu més endavant podria ser fer servir serveis de pagament que et permeten fer anar el model fent servir servidors amb la capacitat desitjada.

Per últim, a continuació és mostren els resultats obtinguts, en una matriu de confusió on també es pot veure el *accuracy* (consultar taula 2).



Il·lustració 39.- Resultat de la predicció per les dates de test. Es mostra una matriu de confusió i el valor de predicció, igual a 54,09%.

9 Problemes observats

En aquest capítol es descriu els obstacles i problemes sorgits durant el plantejament i el desenvolupament de l'estudi. Es poden classificar segons diversos factors, com per exemple, segons l'impacte en el resultat o per la quantitat d'hores que ha costat trobar-ne la solució, si s'escau. Però, es fa un llistat dels problemes i cada un es desenvolupa per separat, explicant l'impacte, els efectes i la, o les, possibles solucions.

9.1 Component humà

El primer obstacle per a la predicció de cotitzacions en borsa és sens dubte el component humà i psicològics. Els moviments de la borsa no són deguts únicament a variables econòmiques o financeres, de l'entorn de l'empresa o de la pròpia, si no que es veuen afectades per moltíssims factors. D'aquí la naturalesa tan complicada de l'objectiu d'aquest estudi.

El valor exacte d'una cotització és el valor d'una sèrie d'accions de compra i venta que es produeixen i defineixen el valor d'aquesta en un determinat instant. Per tant, per intentar conèixer quin valor tindrà una determinada acció caldria conèixer que decidiran fer els inversors en cada moment, i per tant, conèixer el que coneixen els inversors. Evidentment, això no és possible.

Hi ha economistes que defenses que la economia no es pot predir, ja que el futur està obert a totes les possibilitats creatives. Defensen que les prediccions concretes sobre el futur són impossibles en economia, perquè el que passi demà dependrà d'un coneixement que encara no ha estat creat avui. El futur és sempre incert, en el sentit que encara està per fer, i les persones només tenen idees, imaginacions o expectatives que esperen fer realitat mitjançant la seva acció personal i interacció amb altres. De fet, els indicadors tècnics mai han pogut ser demostrats matemàticament, si no que moltes vegades acaben encertant degut a que molts inversors els prenen com a certesa i manipulen el resultat per que així sigui, amb la acció conjunta de tots ells.

Avui dia, grans corporacions, i sobretot bancs d'inversió, inverteixen una gran quantitat de recursos per contractar els millors estudiants per a que desenvolupin models matemàtics així com de ML per predir la borsa. Aquest tipus de transaccions ocupen ja més del 30% del total de les transaccions[159] i tot i que aconsegueixen treure un rendiment que supera al mercat, no ho fan amb gaire marge.

Aquestes firmes operen amb grans quantitats de volums que els fa ser poc àgils en les operacions. Tot i que degut a aquest precisament, els és possible manipular el mercat a la seva voluntat. A més, la majoria es gasten recursos en tenir els servidors el més a prop possible a les borses perquè les seves ordres s'executin immediatament.

[159] https://www.eleconomista.es/mercados-cotizaciones/noticias/8380454/05/17/Quants-los-expertos-en-informacion-cuantitativa-que-manejan-Wall-Street.html

Així, moltes transaccions que es fan mitjançant models predictius es produeixen en el que es coneix com *fast trading o High-frequency trading*, on les ordres s'executen en menys d'un segon.

9.2 Limitacions tècniques

En segon lloc, i un cop entès el punt anterior, cal destacar que un estudi com aquest té un abast que no és comparable al d'aquestes firmes, que aconsegueixen superar el mercat per poc. Amb un ordinador que no disposi de components amb una gran capacitat la limitació tècnica a l'hora d'operar amb grans volums de dades.

Un clar exemple és el algorisme BERT, que al fer-lo córrer bloquejava completament l'ordinador en el qual s'ha desenvolupat aquest estudi. A més, un cop amb les dades resultats del BERT es va intentar fer servir una xarxa neuronal com a model predictiu, però no va ser possible ja que ni la més petita de les xarxes era capaç de córrer sense que l'ordinador acabes reiniciant-se.

Una possible solució pot ser la fer servir serveis, per exemple com els que ofereix Google, que permet fer anar servidors a través del *cloud* amb la capacitat que es vulgui. Però, no queda dins de l'abast d'aquest estudi haver de gastar recursos econòmics.

9.3 Interpretació dels indicadors

En tercer lloc, un problema que va sorgir va ser la interpretació dels indicadors, especialment els tècnics. Ja que, si es fa l'esforç de posar-se en el cap d'un inversor que fa servir aquests indicadors, es veuria que no els fa servir de la mateixa manera que el model rep aquestes dades. Moltes vegades, i depenen del indicador, l'important pot ser la tendència, la pendent, la duració, etc.

Per això, es va fer *feature engineering*[160] per extreure característiques de les dades dels indicadors tècnics, que poguessin ser d'utilitat per al algorisme de ML. El problema va ser que degut a les limitacions tècniques, comentades al punt anterior, de cada indicador s'extreien centenars de característiques i el resultat final no cabia a la memòria de l'ordinador. Com es comenta més endavant, una solució pot ser fer un processament de cada indicador i afegir característiques d'aquest, però moltes necessiten d'uns coneixements financers que queden fora de l'abast d'aquest estudi.

9.4 Manipulació dels resultats

En quart lloc, va haver un aspecte important a destacar és la facilitat per manipular els resultats de la predicció del algorisme XGBOOST. Durant el desenvolupament del projecte, en diversos moments deguts a errors es va intentar predir de forma errònia, per exemple, passant a l'XGBOOST dades del futur o calculant el *label* amb dades del passat en comptes del futur, i en tots ells el resultat obtingut va ser d'una *accuracy* molt elevat. Aquest fet, immediatament va fer saltar les alarmes i finalment es va acabar trobant els errors.

[160] https://en.wikipedia.org/wiki/Feature_engineering

En canvi, hi ha altres estudis, usats com a referències per aquest, que s'han intentat contrastar, ja que fan servir dades molt similars a les d'aquest estudi i un model XGBOOST per predir les cotitzacions i tenen uns resultats molt bons. Un dels qual[161] el resultat és quasi el 100% d'encert, i per tant, a primera vista és sospitós. Segons s'ha vist en els comentaris deixats en aquest estudi per part d'altres usuaris, hi ha un error en les dades que fa servir per predir, i li passa dades de futur al model.

Per altra banda, hi ha un altre[162] que obté un bon resultat fent servir una metodologia i dades molt similars a aquest estudi. Aquest fet va provocar que es descarregues el codi en qüestió per provar-ho amb les dades utilitzades per aquest estudi i es va veure que el resultat canviava completament, i s'obtenia el mateix que en aquest. Després, es va investigar quina era la causa i es va trobar que el *label* es calculava amb el preu d'obertura de la cotització i per predir-ho li passava al model el preu de tancament del dia anterior. A més, així també es va descobrir que el fet de calcular els indicadors manualment mitjançant les respectives formules matemàtiques, no influïa en el resultat.

9.5 Dataset

Un altre obstacle va ser la obtenció de dades per entrenar amb dades públiques. Moltes dades són restringides per a APIs de pagament, fet limitant per aquest estudi, ja que es volia gastar recursos econòmics, sinó fer servir dades públiques. Tant amb les dades de les cotitzacions com especialment amb les dades de les noticies va resultat un obstacle aquest fet. Amb especial menció a aquestes últimes, que van resultat difícils de obtenir.

Inicialment es pretenia fer servir notícies relacionades directament amb les empreses o l'índex borsari però degut a aquest obstacle, es va optar per noticies relaciones de manera més general.

9.6 Cross-validation

Finalment, després de veure els resultats es va aplicar una validació creuada[163], per consolidar els resultats. A més, també pot resultat útil per evitar el sobreentrenament del model.

Es va escollir una *k-fold cross-validation* per una *k* igual a 10. I els resultats van acabar sent molt similars als presentats. Per tant es va poder comprovar que en aquest cas, fer una validació creuada no ajudava a millorar el resultat, consolidant així els mateixos.

[161] https://www.kaggle.com/shreyams/stock-price-prediction-94-xgboost
[162] https://medium.com/@hsahu/stock-prediction-with-xgboost-a-technical-indicators-approach-5f7e5940e9e3
[163] https://en.wikipedia.org/wiki/Cross-validation_(statistics)

10 Conclusions

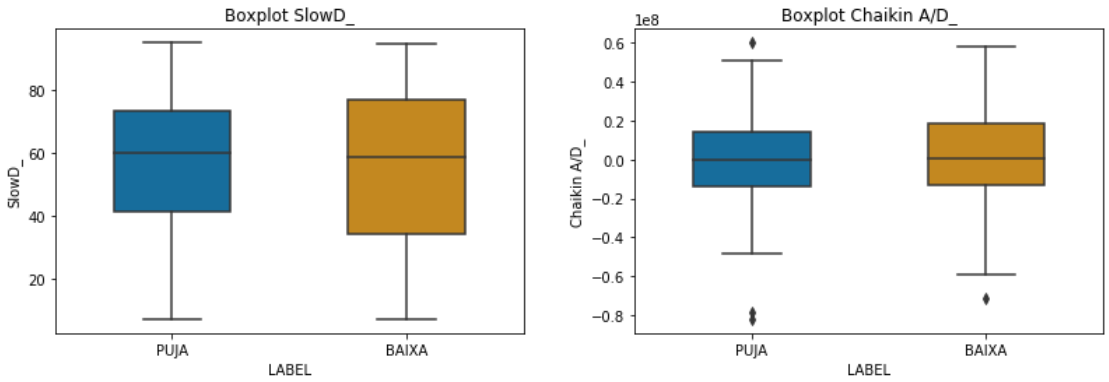
En primer lloc cal esmentar que els resultats d'aquest estudi resulta evident que no són atractius i que no s'ha pogut replicar els resultats de referència trobat en altres estudis similars, degut a la manipulació d'aquests, com hem pogut comprovar.

Resulta evident, que quan és fa una estudi s'intenta demostrar una hipòtesi i en aquest cas ha resultat no ser certa. Han sigut diversos els factors, així com els obstacles que han sorgit, juntament amb la naturalesa difícil d'aquest els que han portat a aquest resultat. A continuació se'n fa un recull i s'expliquen les línies de futur i possibles millores.

Predir la cotització de les accions de la borsa és un repte tan atractiu com difícil. Molts són els que hi dediquen molts recursos i esforços en aconseguir-ho, i no hi ha un resultat clar que sobresurti. A més, donat el cas seria quasi impossible coneixeu-ho. En el capítol anterior, s'ha vist com el factor psicològic i el component humà fa d'aquest estudi un repte quasi impossible. Alhora, que el temps és un factor determinant en el resultat. Per altra banda, mirant els resultats obtinguts es poden extreure algunes conclusions. La primera és que segurament cal fer un estudi més detallat sobre quin és el rang de temps en que pot ser més fàcil predir els moviments de les cotitzacions. Per exemple, en els moviments interdiaris és on la majoria de grans firmes executen aquest tipus de ordres basades en models de ML. Com s'ha comentat, la distancia a les borses i per tant la rapidesa d'aquestes resulta determinant en el resultat. Així com el volum, ja que grans operacions poden moure els moviments de les cotitzacions.

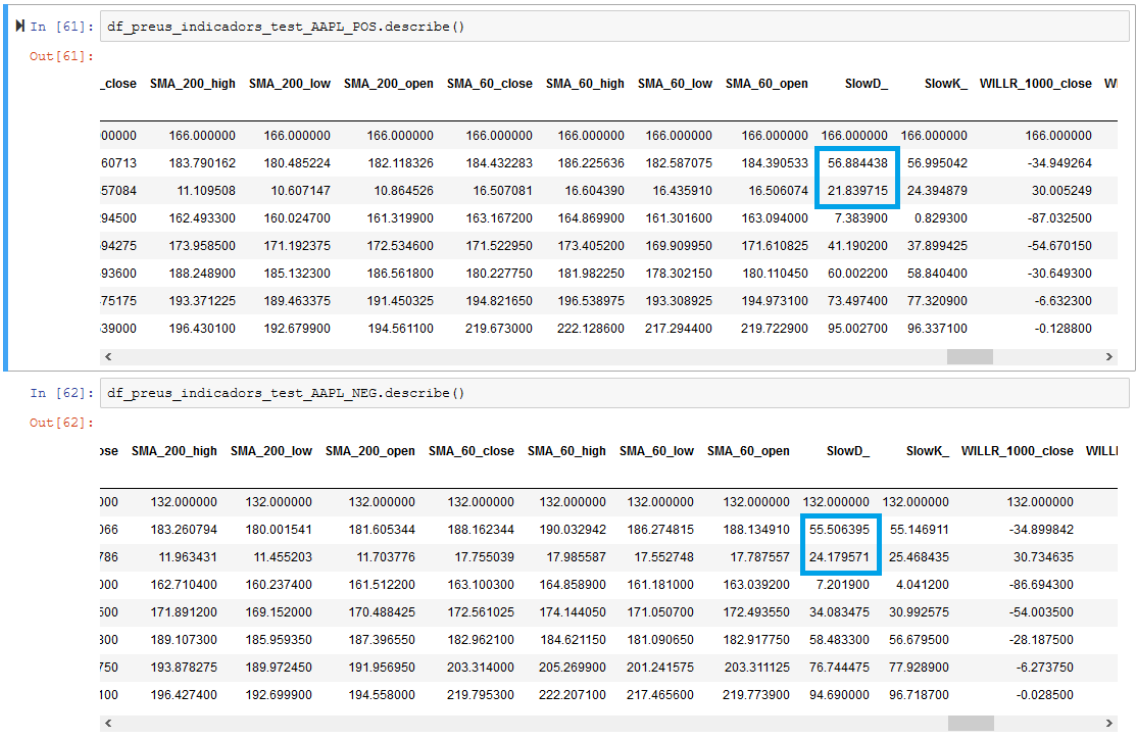
La segona conclusió és que el model de ML no ha sigut capaç de trobar relacions entre les dades d'entrenament. Fet destacable, ja que s'ha vist que quan sí que hi és, el resultat de la predicció augmenta significativament i de forma immediata. Això, convida a pensar que les dades d'entrenament són similars en el cas de que la cotització pugi o baixi.

Per comprovar-ho, es va agafar els indicadors d'una empresa, en aquest cas, AAPL i es va fer dos grups. Per una banda tots els indicadors que tenien una *label* de puja, i per l'altra, la resta d'indicadors que tenen la contraria, baixa. Així, es va fer evident que les dades eren estadísticament quasi idèntiques, i per tant, molt difícils de separar pel model predictor.



Il·lustració 40.- *Boxplot* de la distribució dels dos indicadors amb més pes pel XGBOOST.

Quasi tots els indicadors tenien la mateixa distribució o molt similar. En molts casos, com a mostra en la següent il·lustració, la mitjana de les dades d'un indicador era molt similar en els dos grups, puja o baixa. I a més, la desviació estàndard és superior, o molt superior, a la diferència entre mitjanes. Per tot això, es pot dir que el model de ML pateix a l'hora de trobar relacions quan la cotització del dia següent puja i quan baixa.



```
In [61]: df_preus_indicadors_test_AAPL_POS.describe()
```

Out[61]:

_close	SMA_200_high	SMA_200_low	SMA_200_open	SMA_60_close	SMA_60_high	SMA_60_low	SMA_60_open	SlowD_	SlowK_	WILLR_1000_close	W
00000	166.000000	166.000000	166.000000	166.000000	166.000000	166.000000	166.000000	166.000000	166.000000	166.000000	
60713	183.790162	180.485224	182.118326	184.432283	186.225636	182.587075	184.390533	56.884438	56.995042	-34.949264	
57084	11.109508	10.607147	10.864526	16.507081	16.604390	16.435910	16.506074	21.839715	24.394879	30.005249	
94500	162.493300	160.024700	161.319900	163.167200	164.869900	161.301600	163.094000	7.383900	0.829300	-87.032500	
94275	173.958500	171.192375	172.534600	171.522950	173.405200	169.909950	171.610825	41.190200	37.899425	-54.670150	
93600	188.248900	185.132300	186.561800	180.227750	181.982250	178.302150	180.110450	60.002200	58.840400	-30.649300	
75175	193.371225	189.463375	191.450325	194.821650	196.538975	193.308925	194.973100	73.497400	77.320900	-6.632300	
39000	196.430100	192.679900	194.561100	219.673000	222.128600	217.294400	219.722900	95.002700	96.337100	-0.128800	

```
In [62]: df_preus_indicadors_test_AAPL_NEG.describe()
```

Out[62]:

ose	SMA_200_high	SMA_200_low	SMA_200_open	SMA_60_close	SMA_60_high	SMA_60_low	SMA_60_open	SlowD_	SlowK_	WILLR_1000_close	WILLI
000	132.000000	132.000000	132.000000	132.000000	132.000000	132.000000	132.000000	132.000000	132.000000	132.000000	
066	183.260794	180.001541	181.605344	188.162344	190.032942	186.274815	188.134910	55.506395	55.146911	-34.899842	
786	11.963431	11.455203	11.703776	17.755039	17.985587	17.552748	17.787557	24.179571	25.468435	30.734635	
000	162.710400	160.237400	161.512200	163.100300	164.858900	161.181000	163.039200	7.201900	4.041200	-86.694300	
500	171.891200	169.152000	170.488425	172.561025	174.144050	171.050700	172.493550	34.083475	30.992575	-54.003500	
300	189.107300	185.959350	187.396550	182.962100	184.621150	181.090650	182.917750	58.483300	56.679500	-28.187500	
750	193.878275	189.972450	191.956950	203.314000	205.269900	201.241575	203.311125	76.744475	77.928900	-6.273750	
100	196.427400	192.699900	194.558000	219.795300	222.207100	217.465600	219.773900	94.690000	96.718700	-0.028500	

Il·lustració 41.- Exemple del indicador SlowD on es mostra la mitjana (primer element) i la desviació estàndard (segon element). Notis que són valors quasi idèntics.

També, a la il·lustració anterior es veu com la resta d'indicadors també tenen una distribució molt similar. Així, la conclusió anomenada al principi de que el algorisme de ML no es capaç d'aprendre relacions entre les dades d'entrenament i la *label* queda palès amb aquests exemples.

Per altra banda, es va calcular el p valor a través del t test, com es mostra a la figura següent, per els dos indicadors més rellevants, SlowD i Chaikin A/D. Es veu com la correlació no es gaire elevada, però el més destacat és els valors de p valor[164] molt elevats, que indiquen que la correlació indicada pot no ser deguda al indicador en qüestió.

```
In [75]: from scipy.stats import ttest_ind
         for col in ['Chaikin A/D_', 'SlowD_']:
             d_puja = data.where(data.LABEL== 'PUJA').dropna()[col]
             d_baixa = data.where(data.LABEL== 'BAIXA').dropna()[col]
             r = ttest_ind(d_puja,d_baixa)
             print(col, r)

Chaikin A/D_ Ttest_indResult(statistic=-0.3002507239566887, pvalue=0.7641966754371444)
SlowD_ Ttest_indResult(statistic=0.637074176442862, pvalue=0.5245689403728456)
```

Il·lustració 42.- Resultat de *t test* per els dos indicadors més rellevants.

[164] https://en.wikipedia.org/wiki/P-value

Per últim, hi ha altres motius que expliquen aquests resultats i els hem vist en el capítol previ, però a continuació es fa una explicació de les possibles línies de futur així com es presenten propostes que es creu que poden millorar el resultat, però que han quedat fora de l'abast d'aquest estudi.

10.1 Línies de futur

Abans de descriure les possibles solucions i millores que es poden fer, cal destacar que la manera de encarar i definir l'estratègia per resoldre un estudi d'aquest tipus pot fer variar el resultat significativament. Tot i així, hi ha unes consideracions generals i més en concret sobre aquest que es fan a continuació, per a futures línies d'investigació.

La primera proposta és la de dedicar més recursos per poder obtenir unes dades més amplies, i en el cas de les notícies, més especifiques que segur tenen més relació amb el moviment de les cotitzacions.

El fet de disposar de més dades, com per exemple la dels preus intradiaris, permet dissenyar models de ML que altrament no són possibles, per la poca quantitat de dades, com per exemple, xarxes neuronals. A més, amb més recursos es pot fer servir serveis, ja mencionats, que no limitin l'estudi per les limitacions dels equips. Aquestes dues avantatges tindrien segur un gran impacte en el resultat.

En segon lloc, es pot estudiar la possibilitat de fer un processament dels indicadors i fer una traducció per al model. Per fer-ho caldria uns coneixements financers i d'inversió en borsa, fet que queda fora de l'abast i els objectius d'aquest estudi, però així el model de ML disposaria de dades similars a les que un inversor ho fa quan inverteix. Tot i així, el model hauria de ser capaç de trobar les relacions entre les dades, i amb el *feature engineering* es pot extreure moltes característiques que serien útils pel model.

En tercer lloc, fer un estudi sobre el temps sobre el qual es vol fer la predicció. Per aquest estudi, s'ha decidit intentar predir el preu del dia següent, però podria incrementar el resultat si es fa la predicció per temps menors, com per exemple la següent hora o el tancament del mateix dia. Ja que, la majoria de grans firmes d'inversió que fan servir aquestes models per predir el moviment de les cotitzacions ho fan en la franja intradiaria.

En quart lloc, una bona idea seria fer servir com a dades d'entrenament indicadors de la economia global, com per exemple l'índex *US Equity Market Uncertainty*[165], i/o dades d'altres mercats a diferents parts del món. Per exemple, es podria intentar predir el moviment de la borsa americana fent servir les dades de la borsa del Japó.

Finalment, com ja s'ha comentat varies vegades, inicialment es volia fer servir notícies relacionades directament amb les empreses i/o amb l'índex en qüestió. Però, com s'ha vist aquestes dades són difícils de trobar de manera pública. Una solució, sense invertir recursos econòmics, seria la de desenvolupar un *web scaper* que busques les empreses i/o l'índex en un portal de notícies econòmiques i les obtingues. Només caldria una mica de processament per fer-ho servir pel algorisme BERT. Per acabar, després de passar-les per aquest, una bona idea

[165] http://www.policyuncertainty.com/equity_uncert.html

seria fer servir una xarxa neuronal, ja que les dades ja queden normalitzades al sortir d'una xarxa com el BERT.

11 Referències

[1] A. Krizhevsky, I. Sutskever, and G. Hinton (2012). *ImageNet classification with deep convolutional neural networks*.

[2] Jacob Devli et al.(2008), *BERT: Pre-training of Deep Bidirectional Transformers forLanguage Understanding.*

[3] Vaswani et al.(2007), *Attention Is All You Need*.

[4] Hassan et al. (2007), *A fusion model of HMM, ANN and GA for stock market forecasting*.

[5] J.H. Wang i J.Y. Leu. (1996) *Stock market trend prediction using arimabased neural networks*.

[6] Abraham et al. (2001), *Hybrid Intelligent Systems for Stock Market Analysis.*

[7] Chen et al. (2003), *Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index*.

[8] Vapnik V. (1999), *Statistical Learning Theory*.

[9] Huang, Nakamori i Wang (2005), *Forecasting stock market movement direction with support vector machine*.

[10] Kim (2003), *Financial time series forecasting using support vector machines*.

[11] Ou i Wang (2009), *Prediction of Stock Market Index Movement by Ten Data Mining Techniques*.

[12] Tsai et al. (2011), *Predicting stock returns by classifier ensembles*.

[13] Sun i Li (2012), *Financial distress prediction using support vector machines: Ensemble vs. individual*.

[14] Hsu et al. (2009), *A Two-stage Architecture for Stock Price Forecasting by Integrating Self-Organizing Map and Support Vector Regression*.

[15] Garg, Sriram i Tai (2013), *Empirical analysis of model selection criteria for genetic programming in modeling of time series system*.

[16] Nair et al. (2011), *A GA-Artificial Neural Network Hybrid System for Financial Time Series Forecasting*.

[17] Ahmed (2008), *Aggregate Economic Variables and Stock Markets in India*.

[18] Mantri, Gahan i Nayak (2010), *Artificial Neural Networks – An Application to stock market volatility*.

[19] Mishra, Sehgal i Bhanumurthy (2011), *A search for long-range dependence and chaotic structure in Indian stock market*.

[20] Liu i Wang (2012), [16] *Fluctuation prediction of stock market index by Legendre neural network with random time strength function*.

[21] Araújo i Ferreira (2013) [17], *A Morphological-Rank-Linear evolutionary method for stock market prediction*.