



This is the **published version** of the bachelor thesis:

Pugibet Martinez, Pol; Echeverria Rovira, Lluís, tut. Gestión de Portafolios Financieros Basada en Machine Learning : Predicción y Optimización Automática. 2025. (Enginyeria Informàtica)

This version is available at <https://ddd.uab.cat/record/317528>

under the terms of the  license

Gestió de Portafolis Financers Basada en Machine Learning: Predicció i Optimització Automàtica

Pol Pugibet Martínez

30 de juny de 2025

Resum– Aquest treball presenta una proposta per gestionar carteres d'inversió de manera automatitzada utilitzant tècniques de machine learning. Es combinen diferents tipus de dades financeres amb l'objectiu d'extraure informació rellevant que pugui servir per realitzar les prediccions que guiïn el procés de gestió. El flux consta de diverses etapes: primer, es busquen variables latents mitjançant la combinació de les tècniques DTW i HAC amb models FHMM. A continuació, s'entrenen models convencionals per predir el rendiment i el risc futurs. Finalment, s'utilitzen els valors SHAP d'aquests models per extreure normes i fórmules que expliquin els comportaments del mercat. El procés s'ha integrat en un sistema automatitzat (amb certes limitacions) per a la gestió intel·ligent de carteres. Aquest cobreix totes les etapes del procés de gestió de carteres: des del preprocessament de dades i l'entrenament de models predictius, fins a l'anàlisi d'explicabilitat i la generació de regles d'inversió. El flux complet ha estat implementat i validat en diferents etapes. Inicialment, sobre els actius AAPL, MSFT i SP500 de manera conjunta; i posteriorment, sobre TSLA i Ford per comprovar la generalització del sistema.

Paraules clau– Gestió automatitzada de portafolis, Machine Learning (ML), dades financeres, variables latents, DTW, HAC, FHMM, Reinforcement Learning (RL), SHAP, regles d'inversió

Abstract– This project presents a proposal for automated investment portfolio management using machine learning. Different types of financial data are combined with the aim of extracting relevant information that can be used to make predictions, which then inform the management process. The process involves several stages: first, latent variables are sought by combining DTW and HAC techniques with FHMM models; then, conventional models are trained to predict future performance and risk; finally, these models are used to extract, based on their SHAP values, rules and formulas that can explain market behaviors. The process has been integrated into an automated system (with certain limitations) for intelligent portfolio management. This system covers all stages of the portfolio management process: from data preprocessing and the training of predictive performance and risk models, to explainability analysis and the generation of investment rules. The complete workflow has been implemented and validated in different stages. Initially, it was applied jointly to AAPL, MSFT, and SP500 assets; and subsequently to TSLA and Ford to verify the system's generalization.

Keywords– Automated portfolio management, Machine Learning (ML), financial data, latent variables, DTW, HAC, FHMM, Reinforcement Learning (RL), SHAP, investment rules

1 INTRODUCCIÓ - CONTEXT DEL TREBALL

EN l'àmbit financer, un portafoli és una col·lecció d'actius gestionada per maximitzar la rendibilitat i

- E-mail de contacte: ppugibetm@gmail.com
- Menció realitzada: Computació
- Treball tutoritzat per: Lluís Echeverria Rovira
- Curs 2024/25

minimitzar el risc. La gestió de portafolis implica l'assignació i la diversificació d'actius per assolir objectius d'inversió, tot i que aquesta tasca és complexa a causa de la naturalesa imprevisible dels mercats i els factors externs que els afecten. Tradicionalment, s'han utilitzat models com el de Markowitz[1], que optimitzen l'assignació d'actius a partir de mesures estadístiques basades en dades històriques, tot i que aquestes estimacions tenen incerteses inherents. Per això, es recorre a tècniques més avançades com l'aprenentatge automàtic (en anglès *Machine Learning*, ML), que permeten processar grans volums de dades,

incloent fonts alternatives, per detectar patrons complexos i fer prediccions més precises. Aquest enfocament no només millora les estratègies d'inversió, sinó que també facilita la formulació i validació empírica de noves teories financeres que expliquin el comportament del mercat. Aquest enfocament, proposat per investigadors com Marcos López de Prado, es basa en l'ús de ML per identificar variables rellevants, i a partir d'aquestes, formular teories que puguin ser validades empíricament.

2 OBJECTIUS

L'objectiu general d'aquest treball és implantar un projecte complet de l'àmbit de la ciència de les dades per desenvolupar un sistema automatitzat per a la gestió de portafolis financers. Aquest sistema serà centrat en l'ús de tècniques de ML per millorar les estratègies d'inversió. Les etapes del projecte es basaran en el procés CRISP-DM (Cross-Industry Standard Process for Data Mining)[2], que inclou la comprensió del negoci, la comprensió de les dades, la preparació de les dades, l'entrenament i validació de models, l'avaluació final i el desplegament del model. Aquestes etapes es detallen a l'apartat 4.

A continuació es detallen els objectius que defineixen el marc d'aquest projecte:

1. Identificar com a mínim una variable latent (factor no directament observable en les dades financeres, però que pot tenir un impacte significatiu en el comportament dels mercats, com les correlacions entre actius, patrons de sentiment o factors macroeconòmics no directes com la estabilitat política)[3] a les dades financeres utilitzant tècniques de ML. Aquesta variable es considerarà vàlida si, un cop incorporada com a característica (feature) en un model de ML supervisat, n'augmenta la capacitat predictiva en comparació amb un model entrenat exclusivament amb les dades originals. Aquesta millora es quantificarà mitjançant validació creuada i mètriques com el F1-score o l'exactitud (accuracy), i s'exigirà una millora estadísticament significativa respecte al baseline.
2. Formular teories financeres basades en els descobriments realitzats, que expliquin les relacions estadísticament significatives entre les variables identificades i el comportament dels mercats, validant-les amb un nivell de confiança del 80% (és a dir, un p-valor de 0.20[4]). Aquestes teories són afirmacions o relacions específiques basades en les variables latents identificades, han de ser comprovables i han d'explicar com les variables identificades afecten el comportament del mercat.
3. **(Opcional)** Incorporar dades no tradicionals mitjançant tècniques de processament del llenguatge natural (*Natural Language Processing, NLP*), com l'anàlisi de sentiment en notícies financeres o xarxes socials i demostrar que aquestes dades milloren la precisió de les prediccions en, com a mínim, un 5% respecte dels mètodes tradicionals. Això es mesurarà mitjançant la reducció de l'error quadràtic mitjà (MSE), l'increment de la ràtio de Sharpe[5], la millora en la taxa d'encert i la comparació de la rendibilitat acumulada.

4. **(Opcional)** Optimitzar l'assignació d'actius mitjançant algorismes d'optimització bàsics per maximitzar el retorn esperat i minimitzar el risc, obtenint una ràtio Sharpe igual o superior a 1.0 en *backtesting* (el *backtesting* consisteix a provar un model predictiu utilitzant dades històriques, és a dir, realitzar una simulació en dades passades).

3 ESTAT DE L'ART

En aquesta secció es presenta resumidament el context de l'ús del ML a l'àmbit de les finances, i la seva situació actual tant en aplicacions generals com a la gestió de portafolis financers.

3.1 Gestió Tradicional de Portafolis

Abans de la formalització de la teoria de portafolis, la inversió es basava principalment en intuïció i experiència, sense una base científica sòlida. El punt d'inflexió va arribar amb Harry Markowitz, qui va publicar el seu article seminal "*Portfolio Selection*"[1] el 1952. Aquest treball va establir les bases de la Teoria Moderna de Portafolis.

3.1.1 La Teoria Moderna de Portafolis (1952)

Es tracta d'una formalització i extensió de la diversificació en la inversió, la idea que posseir diferents tipus d'actius financers és menys arriscat que posseir-ne només d'un tipus. La seva visió clau és que el risc i la rendibilitat d'un actiu no s'han d'avaluar per si sols, sinó per la forma en què contribueix al risc i al retorn global d'una cartera. Utilitza la variància dels preus dels actius com a substitut del risc.

3.1.2 Models clàssics de valoració d'actius i teoria multifactorial

Durant les dècades de 1960 i 1970-1980, es van establir fonaments teòrics essencials per a la valoració d'actius i la gestió de carteres. El Capital Asset Pricing Model (CAPM)[6], desenvolupat per William Sharpe, John Linter i Jan Mossin als anys 60, va proporcionar un model per calcular la rendibilitat exigida per un inversor en funció del risc assumit, considerant únicament el rendiment del mercat com a factor de risc. Posteriorment, en les dècades de 1970 i 1980, van emergir els models multifactorials que van ampliar aquesta visió incorporant diversos factors econòmics, com ara les taxes d'interès i la inflació.

3.2 Aplicacions generals del ML en Finances

El ML està transformant el sector financer mitjançant l'automatització intel·ligent de processos i la millora de decisions complexes; a part dels usos plantejats en aquest projecte (predicció de mercats, gestió de risc, optimització de portafolis...), també s'usa per a optimitzar tasques repetitives com l'anàlisi de dades o la generació d'informes, identificar patrons anòmals en transaccions o comportaments sospitosos o avaluar el risc de préstecs i targetes de crèdit, entre d'altres[7].

3.3 Ús del ML en la Gestió de Portafolis Financiers

Actualment, moltes institucions financeres i gestores d'actius estan començant a integrar tècniques de ML en les seves operacions, però encara hi ha reptes per superar, com la interpretabilitat dels models, la qualitat de les dades i la necessitat de personal especialitzat. Tot i això, el potencial del ML per millorar l'eficiència i la precisió en la gestió de portafolis és immens, i s'espera que la seva adopció continuï creixent en els pròxims anys.

Autors com Marcos López de Prado han estat pioners en aquest camp, defensant l'ús del ML no només com una eina predictiva, sinó com un mitjà per descobrir noves teories financeres. López de Prado argumenta que el ML permet identificar variables latents i patrons no lineals en les dades, que poden ser la base per a la formulació de teories amb relacions causals verificables.

A més de López de Prado, autors com Andrew Lo han contribuït significativament a aquest camp. Lo, professor del MIT i pioner en l'aplicació de tècniques quantitatives en finances, ha explorat com el ML pot millorar la comprensió dels mercats financers mitjançant l'anàlisi de grans volums de dades i la identificació de patrons complexos[8]. Un altre autor rellevant és David Easley, que ha investigat com el ML pot utilitzar-se per analitzar dades de mercat d'alta freqüència i millorar la predicció de moviments de preus[9].

3.4 Projectes recents

A continuació, es presenten alguns projectes recents que comparteixen similituds amb aquest treball sobre la gestió de portafolis financers mitjançant ML.

3.4.1 *Modern Portfolio Theory - Optimization through Machine Learning*

Aquest projecte se centra en millorar la Teoria Moderna de Portafolis (MPT) de Markowitz mitjançant tècniques de ML, com el reinforcement learning i els algorismes genètics. L'objectiu és trobar assignacions d'actius més eficients, superant les limitacions dels mètodes tradicionals.

3.4.2 *Portfolio Assets Allocation with Machine Learning*

Aquest projecte utilitza ML per optimitzar l'assignació d'actius en un portafoli, amb un enfocament especial en la predicció de rendiments i la gestió de risc. Inclou l'ús de models com les xarxes neuronals i els arbres de decisió [10].

3.4.3 *High Frequency Trading via Deep Recurrent Neural Networks*

Aquest treball de final de grau (TFG) se centra en l'aplicació de xarxes neuronals recurrents (RNN) per al trading d'alta freqüència (HFT). L'objectiu principal és desenvolupar un model predictiu que pugui anticipar moviments de preus en intervals de temps molt curts (fins a mil·lisegons), utilitzant dades de mercat d'alta freqüència[11].

3.4.4 *Bot de trading automàtic en mercados financieros*

Aquest treball de final de grau (TFG) se centra en el desenvolupament d'un bot de trading automàtic que opera en mercats financers utilitzant tècniques de ML. L'objectiu principal és crear un sistema capaç de prendre decisions de compra i venda de manera autònoma, basant-se en l'anàlisi de dades històriques i en temps real[12].

3.4.5 *Machine Learning in Portfolio Decisions*

Aquest article acadèmic explora l'ús del ML per a la presa de decisions en la gestió de portafolis financers. L'objectiu principal és millorar l'assignació d'actius i la gestió de risc mitjançant tècniques avançades de ML, com les xarxes neuronals, els arbres de decisió i els mètodes d'optimització basats en dades[13].

3.4.6 *Comparativa General*

Els projectes recents sobre l'ús del ML en la gestió de portafolis financers comparteixen l'objectiu comú de millorar les decisions d'inversió mitjançant tècniques avançades d'anàlisi de dades. No obstant això, cada projecte té un enfocament i un abast diferents.

1. **Formulació de Teories:** A diferència dels altres projectes, aquest treball no només pretén aplicar ML per a la predicció i l'optimització, sinó que també busca formular noves teories financeres basades en els resultats dels models. Aquest enfocament teòric-pràctic és una contribució única.

4 METODOLOGIA

Per al desenvolupament d'aquest projecte s'ha optat per una metodologia agile[14] que permeti adaptar-se als canvis i assegurar una execució eficient i flexible. Dins de les metodologies àgils, s'ha escollit Kanban[15], una tècnica que facilita la visualització global del projecte, la gestió de tasques i el control del flux de treball. Les tasques es classifiquen en tres estats: pendents, en curs i finalitzades. Per a la implementació d'aquesta metodologia, s'utilitzarà l'eina ClickUp[16], que permet organitzar i prioritzar les tasques de manera dinàmica.

S'utilitzarà Python com llenguatge principal per a la manipulació de dades i implementació de models, amb Jupyter Notebook com entorn interactiu per a l'execució de codi i documentació del procés. I, a part de les biblioteques específiques per cada etapa que es mencionaran més endavant, s'ha utilitzat: Pandas (per a la neteja i transformació de dades), NumPy (per a càlculs numèrics i normalització de dades), Scikit-learn (per a l'estandardització de dades i divisió en conjunts d'entrenament, validació i prova) i Matplotlib/Seaborn (per a visualitzacions que ajudin a identificar problemes com valors atípics).

4.1 Desenvolupament del projecte

Aquest projecte segueix el procés CRISP-DM (Cross-Industry Standard Process for Data Mining), que inclou les següents etapes:

1. **Comprensió del negoci:** S'ha establert els objectius, metodologia i planificació del projecte. També s'ha realitzat una revisió de la literatura sobre ML i s'ha definit els criteris d'èxit.
2. **Comprensió de les dades:** S'ha recopilat les dades de mercat (dades financeres històriques i indicadors tècnics[17]), les dades fonamentals (que inclouen indicadors macroeconòmics com el PIB o les taxes d'inflació i informació corporativa com l'informe d'ingressos o els fluxos de caixa) i alternatives (anàlisi de sentiment de notícies financeres o xarxes socials)[18] de 3 actius de dos tipus diferents: índex borsari (S&P500) i accions individuals (Apple i Microsoft).

Les dades de mercat (1986 a 2025) s'han obtingut mitjançant la biblioteca yahooquery[19]. Les sèries temporals obtingudes amb la biblioteca comprenen dades històriques com els preus d'obertura, màxim, mínim, tancament, volum, preus ajustats, dividends i splits, amb intervals diaris. A partir d'aquestes dades, s'ha calculat diversos indicadors tècnics (de tendència, volatilitat, volum i altres) per a l'anàlisi del mercat mitjançant la llibreria ta[20]. A la taula 3 de l'Apèndix es mostra una llista de les dades de mercat recopilades inicialment.

Les dades fonamentals (2004 a 2024) s'han obtingut mitjançant les APIs de FRED[21] i Eurostat[22] i realitzant *Web Scraping* a la pàgina web Macrotrends[23], degut a limitacions a les biblioteques com yahooquery, que no proporcionen dades fonamentals anteriors a l'any 2020. A les taules 4 i 5 de l'Apèndix es mostren les llistes de les dades fonamentals recopilades inicialment.

Un cop recopilats els diferents conjunts de dades, s'ha realitzat una anàlisi exploratòria de dades (EDA), que inclou estadístiques descriptives, visualitzacions de distribucions numèriques, mapes de calor de correlacions entre variables i la detecció de valors extrems o anomalies que podrien destorbar l'anàlisi.

També s'ha utilitzat diverses tècniques per comprendre les propietats de les sèries temporals financeres:

- ACF i PACF[26]: La funció d'autocorrelació (ACF) s'ha utilitzat per identificar l'impacte temporal i la dependència de les observacions passades, mentre que la PACF ajuda a discernir la correlació directa entre observacions, ajustant per l'efecte de llagues intermedis. A la figura (6) de l'Apèndix es presenta com a exemple els gràfics de l'ACF i PACF per la volatilitat d'AAPL.
- Descomposició estacional: Aquesta tècnica permet separar una sèrie temporal en tres components: tendència, estacionalitat i residuals. És útil per identificar patrons cíclics i aleatoris.

Els resultats de l'EDA realitzat han mostrat una gran similitud en el comportament estadístic dels tres actius analitzats (AAPL, MSFT i S&P 500), i s'ha pogut observar varies correlacions significatives entre ells (tant a les dades de mercat com a les fonamentals, on també

hi ha correlacions fortes amb les dades macroeconòmiques dels Estats Units). A la figura (1) es mostra la matriu de correlacions entre algunes de les variables principals de cada conjunt (mercat i fonamental), posant en evidència aquestes relacions estadístiques. També es pot veure a la figura (5) de l'Apèndix la comparació dels tancaments ajustats diaris dels tres actius.

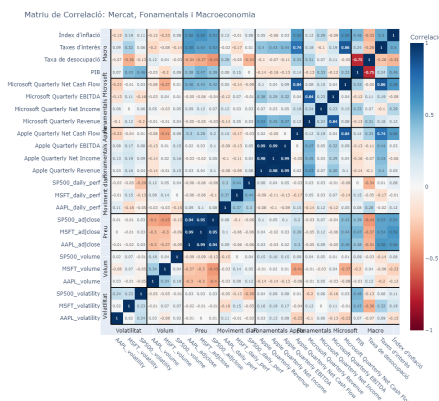


Fig. 1: Correlacions entre algunes de les variables principals

3. **Preparació de dades:** Primerament s'ha definit les variables objectiu del projecte a partir del preu de tancament (close) i altres variables de mercat. Es calculen les rendibilitats futures a curt, mitjà i llarg termini (diària, setmanal i mensual) mitjançant el canvi percentual projectat respecte al preu actual. Paral·lelament, s'estima el risc associat a cada horitzó temporal com la combinació de la volatilitat històrica i el Value at Risk (VaR)[27] calculat amb el mètode històric, utilitzant finestres mòbils de mida adequada per a cada cas. Finalment, es defineixen tres índexs de risc (risc_diari, risc_setmanal, risc_mensual) que integren aquestes dues dimensions, oferint una mesura agregada del risc potencial en cada observació.

Seguidament s'ha eliminat les files que contenen valors no numèrics (NaN) del conjunt de dades. Aquestes files corresponen únicament a les primeres observacions, ja que alguns indicadors calculats necessitaven dades de diverses observacions anteriors per poder determinar el seu valor. La presència de NaNs en aquestes observacions inicials és una conseqüència natural del procés de càlcul d'aquests indicadors basats en retards.

Un cop eliminats els valors no numèrics (NaN), també s'ha gestionat el tractament dels valors extrems (outliers) detectats en les variables numèriques del conjunt de dades. Per ajustar aquests outliers s'ha emprat el mètode basat en el rang interquartílic (IQR) amb finestres temporals de 20 dies (1 mes té aproximadament 20 dies laborables), que ajusta els valors fora dels límits superior i inferior al valor més proper dins del rang. Aquesta tècnica assegura que les dades siguin més consistents i robustes, evitant que aquests valors extrems distorsionin els resultats en etapes posteriors. Es pot observar a la figura (2) un exemple de gràfic generat per observar els outliers de les sèries.

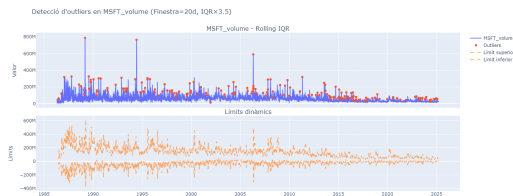


Fig. 2: Exemple de visualització dels outliers per MSFT_volume

Aquest tractament dels outliers s'ha realitzat com una mesura per facilitar la identificació de variables latents. No obstant això, de cara a l'entrenament dels models de predicció, s'ha valorat la possibilitat de preservar els outliers, ja que sovint representen esdeveniments extraordinaris que poden aportar informació valuosa. Per això, no s'ha optat per eliminar-los sistemàticament, sinó que s'ha desenvolupat un sistema automàtic de normalització que analitza individualment cada variable per determinar el tractament més adient. Aquest sistema avalua la distribució, la presència de valors extrems (outliers), l'asimetria (*skewness*) i la naturalesa dels valors (negatius, zeros, binaris o constants).

Segons aquestes característiques, s'aplica:

- *StandardScaler*, per a variables amb distribucions simètriques i sense presència significativa d'outliers.
- *RobustScaler*, per a variables amb valors extrems rellevants, ja que aquest mètode utilitza l'interval interquartílic per escalar les dades i és menys sensible als outliers.
- Una transformació logarítmica amb $\log_{1p}$ (amb desplaçament si cal) en cas de variables estrictament positives i amb alta asimetria, per reduir la distorsió causada per valors molt grans.

També s'ha creat variables sintètiques per facilitar l'anàlisi, transformant dades originals en indicadors més interpretables i estructurant informació esporàdica com dividends i splits. Entre les variables derivades de mercat destaquen: *daily_perf* (rendiment diari entre obertura i tancament), *daily_range* (volatilitat intradia mesurada pel rang màxim-mínim), *bb_width* (amplitud de les bandes de Bollinger) i *stoch_ratio* (combinació de variables estocàstiques). Per a la gestió de dividends i splits, s'han creat variables binàries que n'indiquen la presència, i s'ha calculat el *textttdividend_yield* normalitzat respecte al preu de tancament anterior.

Un cop s'ha tingut totes les dades necessàries amb la forma desitjada, s'ha procedit a la recerca de les variables latents:

- Primer s'ha separat els conjunts de dades de mercat, fonamentals (d'AAPL i MSFT) i macroeconòmiques en conjunts d'entrenament (70% de les dades), validació (15% de les dades) i test (15% de les dades). Aquesta separació ajudarà a evitar

overfitting i *data leakage*. Amb aquesta separació clara entre entrenament, validació i test, podem garantir que els models es validen de manera justa, avaluant la seva capacitat de generalitzar a dades no vistes sense influències externes.

- Seguidament s'ha combinat les tècniques de *Dynamic Time Warping* (DTW) i *Hierarchical Agglomerative Clustering* (HAC) sobre cada conjunt de dades per agrupar variables segons la seva semblança temporal. Primer es calcula una matriu de distàncies basada en DTW, que mesura la semblança entre sèries temporals desaliniades. A continuació, s'utilitza HAC per estructurar jeràrquicament aquests grups de variables, assegurant que els patrons temporals similars estiguin agrupats.
- Posteriorment, s'ha aplicat una metodologia per seleccionar les variables més representatives de cada clúster generat pel HAC. Aquest procés es basa en criteris de centralitat, identificant variables amb distàncies mínimes al centre del clúster, i de diversitat, eliminant variables massa similars entre elles. Això garanteix que les variables seleccionades encapsulin amb precisió les característiques del clúster, facilitant la reducció de dimensionalitat i l'anàlisi posterior.
- Un cop seleccionades les variables més representatives de cada conjunt, s'han entrenat diversos models Factorial Hidden Markov Models (FHMM) aplicant diferents combinacions de cadenes. Cada cadena és una seqüència d'estats ocults que evoluciona al llarg del temps i representa una font específica de variació dins les dades, i es pot entendre com una variable latent que recull una dimensió oculta i temporal. Aquestes cadenes permeten capturar diverses dinàmiques subjacents en les dades, modelant diverses perspectives dels patrons ocults i generant models alternatius que exploren les variacions en les interaccions entre les variables seleccionades.

4. **Entrenament i validació de models:** Per aquesta tasca s'ha considerat diversos algorismes:

- Models lineals (*Ridge*, *Lasso*)
- Models d'arbres (*Random Forest*[29] o *BalancedRandomForest*, *HistGradientBoosting*)
- Models basats en boosting (*XGBoost*, *LightGBM*, *CatBoost*, *Histogram-based Gradient Boosting*)[30]
- Xarxes neuronals (LSTM[31] i MLP[32])

Predicció del rendiment:

Per a la predicció del rendiment s'ha explorat dues aproximacions: regressió i classificació. Inicialment, es va intentar predir el valor exacte del rendiment futur (diari, setmanal i mensual), però els resultats van ser altament insatisfactoris per la majoria de les variables objectiu, ja que les prediccions del model de regressió es van mantenir gairebé sempre en el rang negatiu. Aquesta tendència consistent a predir un rendiment negatiu independentment del valor real o de la distribució

de la variable objectiu va evidenciar una baixa capacitat predictiva i una elevada variabilitat, fent que aquesta aproximació fos inadequada per al propòsit d'aquest projecte.

Davant d'aquesta limitació, inicialment es va abordar el problema com una tasca de classificació binària (0 per a rendiment negatiu, 1 per a positiu), una aproximació que va permetre obtenir resultats sòlids al centrar-se únicament en la direcció del moviment. Aquesta simplificació va facilitar la identificació de patrons bàsics i va millorar la robustesa de les prediccions.

No obstant això, per aprofundir en la capacitat predictiva del model, es va provar un enfocament més granular amb classificació multiclasse en sis intervals ($<-4\%$, $<-2\%$, $<0\%$, $>0\%$, $>2\%$ i $>4\%$).

Els models s'han entrenat seguint un procés de validació creuada temporal mitjançant múltiples divisions amb *TimeSeriesSplit*, per evitar problemes com el *data leakage* i assegurar una bona generalització en dades futures. Els hiperparàmetres s'han ajustat amb *Bayesian Search*[33].

Després d'entrenar els models per separat, s'ha creat un sistema d'ensemble per votació (utilitzant un *VotingClassifier*). Aquest ensemble assigna un pes a cada model individual segons la seva precisió, donant més influència als models que millor redueixen els falsos positius. L'objectiu és combinar les fortaleces de cada model per aconseguir un sistema predictiu més robust i fiable en general.

Predicció del risc:

Per abordar la predicció del risc financer, s'ha optat per un enfocament basat en regressió.

S'ha entrenat models específics per a cada combinació d'actiu i horitzó temporal (diari, setmanal i mensual). A diferència del cas del rendiment, aquí no s'ha reformulat el problema com una classificació, ja que els resultats obtinguts amb regressió han estat majoritàriament acceptables i han permès capturar adequadament la dinàmica del risc.

En el procés d'entrenament s'ha aplicat un *StandardScaler* a la variable objectiu. Aquesta normalització és especialment útil per a models com Ridge, Lasso, MLP o LSTM, que són sensibles a l'escala de les dades.

Tots els models s'han optimitzat mitjançant la mateixa estratègia de *Bayesian Search*

Aplicació d'algorismes d'optimització de carteres:

Un cop obtinguts els models de rendiment i risc, s'ha utilitzat les seves prediccions per implementar dues estratègies de gestió de portafolis en el conjunt de test, per tal de tenir una primera versió dels algorismes que s'utilitzarà més endavant en el desenvolupament del sistema automatitzat de gestió i també per tenir una altra forma d'avaluar la qualitat de les prediccions dels models.

- **Optimització Convexa**[34]: Aquesta estratègia, basada en la teoria moderna de portafolis, calcula

el percentatge del capital a dedicar a una compra en funció de la maximització del rendiment esperat ajustat pel risc del portafoli global. Quan es genera un senyal de compra per un actiu, l'optimitzador determina el pes òptim d'aquest actiu dins del portafoli sota certes restriccions.

- **Reinforcement learning**: Un agent aprèn a prendre decisions d'inversió òptimes mitjançant interacció amb l'entorn i retroalimentació basada en recompenses.

L'algorisme utilitzat és *Proximal Policy Optimization* (PPO)[35], un model actor-critic on l'actor decideix accions (comprar/mantenir/vendre) i el crític les avalua, actualitzant la política de manera controlada.

S'ha construït un entorn propi simulant un mercat de valors amb dades reals. A cada "dia", el model observa un estat que inclou el rendiment i el risc previst de cada actiu.

La recompensa es calcula en funció del canvi percentual del valor del portafoli: si el valor puja, la recompensa és positiva, i si baixa, negativa.

5. **Avaluació**: En aquest apartat s'explica com s'ha avaluat els resultats obtinguts, però aquests s'exposen a l'apartat 5

Avaluació de les variables latents: Per avaluar les variables latents identificades durant el projecte (1 en dades fonamentals, d'una sola dimensió; i 1 en dades de mercat, de dues dimensions), s'ha entrenat models simples que han permès examinar la rellevància i interpretabilitat de cadascuna de les variables latents. L'avaluació s'ha basat en l'ús de models com Random Forest, SVC, Logistic Regression i KNN, combinats amb validació creuada temporal (*TimeSeriesSplit*), per analitzar la millora en la capacitat predictiva dels models en incorporar-hi variables latents. Per a cada variable objectiu de rendiment, s'ha comparat el rendiment dels models amb i sense la informació latent mitjançant les mètriques accuracy i F1-score.

Avaluació dels models de rendiment:

Un cop entrenats els models base, s'han avaluat mitjançant mètriques com l'exactitud, el F1-score i la precisió. A la figura (3) es pot veure com exemple la millor matriu de confusió per la classificació binària del rendiment mensual d'AAPL, i la taula 1 presenta els resultats per totes les granularitats temporals dels millors models de classificació binària.

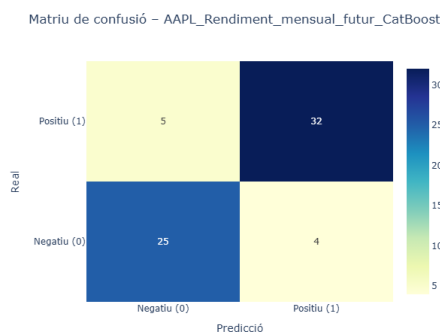


Fig. 3: Matriu de confusió pel millor model de classificació binària pel rendiment mensual d'AAPL

Avaluació dels models de risc: Els models s'han avaluat amb un conjunt de mètriques clau, seleccionades per proporcionar una comprensió completa del seu rendiment predictiu:

- **MAE (Mean Absolute Error):** S'espera un MAE baix (<0.05) donat el rang reduït de les dades (entre 0.01 i 0.15).
- **RMSE (Root Mean Squared Error):** S'espera un RMSE baix ($>MAE$) que podria ser lleugerament superior al MAE per la penalització de grans errors.
- **R2 (Coeficient de Determinació):** No s'espera un R2 alt (0.01-0.20) donada la complexitat intrínseca de predir rendiments financers.
- **Directional Accuracy:** Avaluja la capacitat del model per predir correctament la direcció del moviment de la variable objectiu (p. ex., si pujarà o baixarà, o si entrarà en un rang específic com els de classificació). S'espera una Directional Accuracy per sobre del 50% (51-60%) per demostrar un rendiment millor que l'atzar.
- **Volatilitat del MAE Mòbil:** Calcula la desviació estàndard del MAE sobre finestres de temps consecutives. S'espera un valor baix (<0.01) per indicar una fiabilitat estable del model en el temps.

Avaluació dels algorismes d'optimització de carteres:

Aquests algorismes s'han avaluat mitjançant un backtesting sobre el conjunt de dades de validació (que va des de finals de 2022 fins principis de 2025). En aquest backtesting s'ha avaluat tant el rendiment mensual mitjà com la ràtio Sharpe.

Formulació de teories financeres basades en els resultats dels models:

Amb l'objectiu d'obtenir coneixement interpretable i útil a partir dels models predictius, s'ha dut a terme un procés de formulació de teories financeres fonamentat en l'anàlisi dels valors SHAP (*SHapley Additive ex-Planations*)[36]. Aquest procés s'ha aplicat tant als models de predicció del rendiment (classificació binària) com als de predicció del risc (regressió), seguint

una metodologia comuna però amb algunes diferències clau segons el tipus de variable objectiu.

En ambdós casos, el primer pas ha estat calcular els valors SHAP per a les variables predictors dels models entrenats, permetent quantificar la contribució de cada característica a les prediccions. A continuació, es seleccionen les variables més rellevants segons la seva importància acumulada per reduir la dimensionalitat.

Les contribucions SHAP de cada variable es modelen individualment mitjançant regressió, facilitant una interpretació més detallada. Per al rendiment, s'extreuen regles interpretables del tipus "if-then" a partir dels arbres de decisió, que es filtren, fusionen i validen estadísticament per formar classificadors binaris interpretables que expliquen condicions de mercat i constitueixen teories financeres empíriques. En canvi, per al risc, les contribucions es modelen com a fórmules lineals que expliquen quantitativament l'impacte de cada variable, també validades estadísticament i combinades en una expressió global del risc esperat.

A la figura (8) de l'Apèndix es mostra les fórmules obtingudes per explicar la contribució SHAP de la variable MSFT_volume al risc de l'actiu MSFT.

6. **Desplegament del model:** Finalment, per culminar el desenvolupament pràctic del projecte, s'ha implementat una primera versió funcional del sistema automatitzat de gestió de portafolis. Aquest sistema integra totes les funcionalitats implementades al llarg del projecte: cerca de variables latents, entrenament de models predictius de rendiment i risc, explicabilitat mitjançant valors SHAP i generació de regles o fórmules interpretables.

El funcionament del sistema és totalment automatitzat: només cal especificar quins actius financers es volen gestionar i quines dades macroeconòmiques es vol afegir, i el sistema s'encarrega de tot el procés.

No obstant això, cal tenir en compte que aquesta primera versió del sistema presenta algunes limitacions:

- **Limitacions en les dades fonamentals:** El sistema no pot utilitzar dades fonamentals dels actius financers, ja que *Yahoo Finance*, el servei des del qual es recopilen les dades de mercat i fonamentals dels actius, no proporciona informació històrica anterior a l'any 2023. Això en dificulta l'aplicació en contextos on l'anàlisi fonamental és clau.
- **Limitacions en les dades macroeconòmiques:** En cas que es vulgui incorporar dades macroeconòmiques al sistema, caldrà dur a terme una recerca prèvia per seleccionar les més adients segons l'objectiu de l'anàlisi. A més, si es volen utilitzar fonts com Eurostat, serà necessari un tractament manual de les dades a causa de l'estructura complexa i poc homogènia que presenten.
- **Rendiment inferior dels models basats en regles:** S'ha creat models que fan prediccions basant-se purament en les regles i fórmules trobades. Aquests models presenten, en general, un

rendiment inferior en comparació amb els models convencionals. No obstant això, en escenaris amb granularitat temporal diària aquests models poden arribar a superar els convencionals, tenint una precisió del voltant del 60%, mentre que la dels convencionals es troba entre 50 i 55%.

A la figura (7) de l'Apèndix es pot veure un exemple de rendiment per la gestió d'un portafoli amb accions de *Tesla* i *Ford* bassant-se en les prediccions dels models convencionals, sense aplicació de regles interpretables.

5 RESULTATS

5.1 Variables latents

Per els tres actius s'ha obtingut una millora d'un 5% en l'accuracy i del 2 al 8% en la F1 de les prediccions del rendiment. En el cas de les variables de risc, l'R2 ha empitjorat significativament (fins a un decrement del 250% respecte de la predicció sense les variables latents), mentre que, en el cas del MSE, no hi ha hagut cap canvi significatiu.

Les variables identificades són les següents:

- Latent fonamental: Aquesta variable latent reflecteix patrons complexos en els fluxos de caixa i la comptabilitat trimestral d'Apple i Microsoft. Permet identificar situacions on hi ha canvis significatius en les despeses operatives, ingressos i passius corrents d'Apple, alhora que mostra la relació amb els costos i ingressos nets de Microsoft. És especialment útil per captar fases on els dos gegants tecnològics mostren divergències o sinèrgies en la seva salut financera, destacant períodes d'estrès operatiu o d'inversió i moments de consolidació de resultats econòmics.
- Dimensió de mercat 1: Aquesta dimensió ens permet detectar situacions en què AAPL actua com a líder o com a valor desenganxat del consens del mercat, així com moments de transició en què aquesta força relativa canvia de mans. És útil per entendre les dinàmiques internes dins del sector tecnològic, i també per captar fases de divergència entre valors grans i l'índex general.
- Dimensió de mercat 2: Aquesta dimensió reflecteix el moment del cicle tècnic i el sentiment general, indicant si estem en fase de caiguda, estabilització o pujada conjunta.

S'ha generat la matriu de correlacions entre variables latents i variables objectius, presentada a la figura (4), que proporciona una perspectiva clara sobre les interaccions subjacents entre aquestes variables.

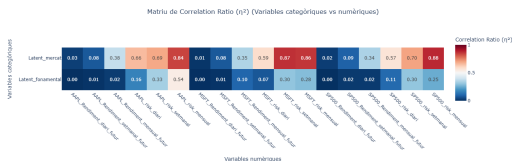


Fig. 4: Matriu de correlacions entre les variables latents (eix vertical) i les variables objectiu (eix horitzontal)

5.2 Models de predicció del rendiment

Les prediccions a nivell diari mostren un rendiment inferior en comparació amb les prediccions setmanals o mensuals. Aquesta disminució del rendiment s'explica principalment per la major volatilitat i el soroll present en les dades diàries, que dificulten la identificació de patrons consistents. A curt termini, els mercats financers estan molt influïts per factors aleatoris o notícies imprevistes, fet que redueix la capacitat dels models per generalitzar i fer prediccions precises. Així, tot i que els algorismes poden captar tendències més àmplies en agregats temporals més llargs, les prediccions diàries continuen sent un repte degut a la seva naturalesa inherentment més caòtica.

TAULA 1: MILLORS RESULTATS PER CADA ACTIU I GRANULARITAT TEMPORAL AMB MODELS DE CLASSIFICACIÓ BINÀRIA

		Accuracy	F1-Score	Precision	Model
AAPL	Diari	0.523	0.522	0.521	XGBoost
	Setmanal	0.806	0.833	0.837	BalancedRandomForest
	Mensual	0.826	0.822	0.828	CatBoost
MSFT	Diari	0.521	0.505	0.517	XGBoost
	Setmanal	0.833	0.851	0.854	BalancedRandomForest
	Mensual	0.831	0.830	0.830	XGBoost
SP500	Diari	0.543	0.518	0.539	CatBoost
	Setmanal	0.918	0.928	0.912	BalancedRandomForest
	Mensual	0.907	0.908	0.912	BalancedRandomForest

A la taula 6 de l'Apèndix es presenta la taula de les mètriques de la classificació multiclasse, la qual s'ha realitzat correctament, però s'ha decidit utilitzar els models de classificació binària pel fet que els seus resultats són força millors en aquest context.

5.3 Models de predicció del risc

A la taula 2 es presenten les mesures de les mètriques per el millor model per cada variable objectiu.

TAULA 2: RESULTATS DE L' AVALUACIÓ DELS MODELS DE PREDICCIÓ DEL RISC

Model	MAE	RMSE	R2	Directional Accuracy	Rolling MAE volatility
RandomForest	0.0105	0.013	0.0025	0.4	0.0076
HistGB	0.0108	0.0124	0.1216	0.29	0.0066
XGBoost	0.0107	0.0126	0.1823	0.51	0.0063
CatBoost	0.0114	0.0135	-0.0002	0.52	0.007

Així mateix, a l'Apèndix es pot visualitzar les comparacions entre prediccions i valors reals d'un dels models amb millors resultats, a la figura (9); i un dels models amb pitjors resultats, a la figura (10).

5.4 Algorismes d'optimització de carteres

En el cas de l'optimització convexa, aquesta estratègia ha aconseguit (en els 28 mesos del conjunt de validació) un rendiment mitjà del 1.73% mensual, mentre que el creixement dels actius ha estat, de mitjana, del 1.71% mensual. Per avaluar aquesta gestió del portafolis, s'ha calculat la ratio Sharpe, que ha donat un valor de 0.055. Aquest valor no és especialment destacable, però s'ha de tenir en compte que es tracta d'una implementació extremadament simple per tenir una altra forma d'avaluar els models finals.

Per altra banda, l'estratègia del model de Reinforcement Learning ha obtingut un rendiment del 1.77% mensual mitjà, amb una ràtio Sharpe de 0.052.

6 CONCLUSIONS

Aquest projecte ha desenvolupat i validat un sistema complet per a la gestió automatitzada de carteres d'inversió basat en tècniques de machine learning. L'objectiu principal, que era proporcionar una eina capaç de predir el rendiment i el risc, així com d'extreure informació explicable que permeti comprendre els comportaments del mercat, s'ha aconseguit amb èxit en la majoria dels aspectes abordats.

1. **Identificació de variables latents:** Mitjançant l'ús combinat de DTW i HAC, associat a models FHMM, s'ha aconseguit identificar dues variables, una de les quals ha proporcionat una millora notable en la predicció del rendiment per als tres actius, amb un increment del 5% en la precisió (accuracy) i una millora del 2% al 8% en la F1-score. En contrast, l'aplicació d'aquestes variables latents ha tingut un impacte negatiu en la predicció de les variables de risc, amb una reducció del 250% en l'R2, tot i que el MSE ha romàs pràcticament inalterat.
2. **Formulació de teories financeres explicatives:** Els models predictius entrenats han estat emprats per extreure regles i fórmules basades en els seus valors SHAP. Aquestes regles han permès construir mecanismes de predicció amb una base explicable, que identifiquen relacions estadísticament rellevants entre les variables latents i els comportaments observats al mercat.
Tot i que no s'han formulat teories financeres com a tal, els patrons obtinguts ofereixen una base empírica sòlida per a comprendre millor les dinàmiques de mercat i fonamentar futures hipòtesis.
3. **Incorporació de dades no tradicionals (opcional):** Tot i que estava previst integrar dades alternatives (com notícies financeres o el sentiment a xarxes socials) per millorar la precisió de les prediccions, aquesta part del projecte no s'ha pogut dur a terme per manca de temps.
4. **Optimització de l'assignació d'actius (opcional):** Encara que l'enfocament principal s'ha centrat en la predicció del rendiment i el risc, el sistema també proporciona una base sòlida per a futures implementacions orientades a l'optimització de l'assignació d'actius. Els resultats obtinguts indiquen que, amb l'addició de tècniques d'optimització, es podria maximitzar encara més la rendibilitat mentre es minimitza el risc.

Tot i haver obtingut resultats raonables, el sistema mostra una dependència estructural dels actius utilitzats durant l'entrenament. Això podria limitar la seva capacitat per detectar desequilibris reals o ineficiències al mercat.

Es plantegen diverses línies de millora per reforçar-ne la robustesa i flexibilitat:

- Replantegar el mecanisme de compra-venda amb accions més agressives o contràries.
- Afegir penalitzacions per correlació elevada amb índexs de referència amb l'objectiu d'incentivar estratègies diferenciades que aportin valor independent.

- Ampliar l'univers d'actius amb opcions menys correlacionades (renda fixa, commodities, mercats emergents...), per diversificar les situacions analitzades i millorar la capacitat de generalització del sistema.
- Incorporar tècniques d'explicabilitat i quantificació de la confiança predictiva, per adaptar el nivell d'exposició al risc segons la certesa del model i evitar errors en situacions d'alta incertesa.
- Combinar granularitats temporals per fer la gestió del portafolis, la qual cosa permetria al sistema operar amb una visió més completa i adaptable a diferents horitzons d'inversió.

REFERÈNCIES

- [1] H. Markowitz, "Portfolio Selection," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, Mar. 1952, doi: <https://doi.org/10.1111/j.1540-6261.1952.tb01525.x>
- [2] C. Shearer, "The CRISP-DM model: the new blueprint for data mining," *Journal of Data Warehousing*, vol. 5, no. 4, pp. 13–22, Fall 2000
- [3] J. Kmenta, "Latent variables in econometrics," *Statistica Neerlandica*, vol. 45, no. 2, pp. 73–84, Jun. 1991, doi: <https://doi.org/10.1111/j.1467-9574.1991.tb01295.x>
- [4] "P-VALUES SIMPLIFIED," Western Michigan University. <https://wmedu.edu/sites/default/files/P-VALUES%20SIMPLIFIED.pdf>
- [5] W. F. Sharpe, "The Sharpe Ratio," *Journal of Portfolio Management*, vol. 21, (1), pp. 49, 1994. Available: <https://www.proquest.com/scholarly-journals/sharpe-ratio/docview/195581284/se-2>
- [6] W. F. Sharpe, "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk," *The Journal of Finance*, vol. 19, no. 3, pp. 425–442, 1964, Available: <https://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.1964.tb02865.x>
- [7] P. Tierno, "Artificial Intelligence and Machine Learning in Financial Services," Apr. 03, 2024. <https://crsreports.congress.gov/product/details?prodcode=R47997>
- [8] MIT CSAIL, "MIT Professor on How AI LLMs are Shaping Financial Advice, Analysis, Risk Management: Part 1," YouTube, Sep. 18, 2024. <https://www.youtube.com/watch?v=ui7kRlJMqjM&list=PLcwbNON2hqP5DlwQeXn8gflpMOxhLs0Mr&index=9>
- [9] D. Easley, M. López de Prado, M. O'Hara, and Z. Zhang, "Microstructure in the Machine Age," *The Review of Financial Studies*, Jul. 2020, doi: <https://doi.org/10.1093/rfs/hhaa078>

- [10] R. Marino, "Portfolio Assets Allocation with Machine Learning," Quantitative Finance Algo Trading Blog by QuantInsti, Dec. 19, 2022. <https://blog.quantinsti.com/portfolio-assets-allocation-machine-learning-project-raimondo-marino/>
- [11] R. Morral, "High Frequency Trading via Deep Recurrent Neural Networks," Upc.edu, May 2020, doi: <http://hdl.handle.net/2117/332062>
- [12] B. F. Pablo, "Bot de trading automático en mercados financieros | Archivo Digital UPM," Oa.upm.es, Jun. 2021, doi: <https://oa.upm.es/67989/>
- [13] M. Guidolin, G. Panzeri, and M. Pedio, "Machine Learning in Portfolio Decisions," 2024, doi: <https://doi.org/10.2139/ssrn.4988124>
- [14] "What is Agile Software Development?," Agile Alliance l, Jun. 29, 2015. <https://www.agilealliance.org/agile101/#:~:text=Agile%20is%20the%20ability%20to,an%20uncertain%20and%20turbulent%20environment.>
- [15] K. McDonald, "What is Kanban?," Agile Alliance l, Apr. 04, 2017. <https://www.agilealliance.org/glossary/kanban/>
- [16] "ClickUpTM," Clickup.com, 2025. <https://clickup.com/>
- [17] C. Team, "Technical Indicator," Corporate Finance Institute, Sep. 19, 2024. <https://corporatefinanceinstitute.com/resources/career-map/sell-side/capital-markets/technical-indicator/>
- [18] Lopez De Prado, Marcos. *Advances in Financial Machine Learning*. New Jersey, Wiley, 2018, pp. 23–25
- [19] D. Guthrie, "yahooquery," Dpguthrie.com, 2025. <https://yahooquery.dpguthrie.com/>
- [20] "Welcome to Technical Analysis Library in Python's documentation! — Technical Analysis Library in Python 0.1.4 documentation," Readthedocs.io, 2022. <https://technical-analysis-library-in-python.readthedocs.io/en/latest/>
- [21] "St. Louis Fed Web Services: FRED® API," Stlouisfed.org, 2025. <https://fred.stlouisfed.org/docs/api/fred/>
- [22] "API - Getting started - Eurostat Online Help for Data Browser - EC Public Wiki," Europa.eu, 2024. <https://wikis.ec.europa.eu/display/EUROSTATHELP/API+-+Getting+started>
- [23] "Macrotrends | The Long Term Perspective on Markets," Macrotrends.net, 2025. <https://www.macrotrends.net/>
- [24] "reddit.com: api documentation," Reddit.com, 2025. <https://www.reddit.com/dev/api/>
- [25] "pytrends," PyPI, Apr. 13, 2023. <https://pypi.org/project/pytrends/>
- [26] iamleonie, "Time Series: Interpreting ACF and PACF," Kaggle.com, Mar. 15, 2022. <https://www.kaggle.com/code/iamleonie/time-series-interpreting-acf-and-pacf>
- [27] C. Team, "Value at Risk (VaR)," Corporate Finance Institute, Oct. 04, 2023. <https://corporatefinanceinstitute.com/resources/career-map/sell-side/risk-management/value-at-risk-var/>
- [28] "1.3.5.16. Kolmogorov-Smirnov Goodness-of-Fit Test," Nist.gov, 2025. <https://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm>
- [29] "RandomForestClassifier," scikit-learn, 2025. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
- [30] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, no. 3, Aug. 2020, doi: <https://doi.org/10.1007/s10462-020-09896-5>.
- [31] "Introducción a la memoria a corto-largo plazo (LSTM)," Mathworks.com, 2025. <https://la.mathworks.com/discovery/lstm.html>
- [32] "1.17. Neural network models (supervised)," scikit-learn, 2025. https://scikit-learn.org/stable/modules/neural_networks_supervised.html#multi-layer-perceptron
- [33] "skopt.BayesSearchCV — scikit-optimize 0.8.1 documentation," Github.io, 2017. <https://scikit-optimize.github.io/stable/modules/generated/skopt.BayesSearchCV.html>
- [34] S. Boyd and L. Vandenberghe, "Convex Optimization," 2014. https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf
- [35] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Openai, "Proximal Policy Optimization Algorithms," Aug. 2017. <https://arxiv.org/pdf/1707.06347>
- [36] "Welcome to the SHAP documentation — SHAP latest documentation," Readthedocs.io, 2018. <https://shap.readthedocs.io/en/latest/>

7 APÈNDIX

TAULA 3: COLUMNES DE LES DADES DE MERCAT RECOPILADES (PER CADA ACTIU)

Open
High
Low
Close
Volume
Adjclose
Dividends
Splits
Volatility
RSI
MACD_hist
BB_upper
BB_lower
Stoch
ADX
Williams_%R
CCI
ATR
OBV

TAULA 4: COLUMNES DE LES DADES FONAMENTALS RECOPILADES (PER AAPL I MSFT)

Quarterly Revenue
Quarterly Cost of Goods Sold
Quarterly Gross Profit
Quarterly Operating Expenses
Quarterly Operating Income
Quarterly Pre-Tax Income
Quarterly Net Income
Quarterly EBITDA
Quarterly EPS
Quarterly Total Current Assets
Quarterly Total Assets
Quarterly Total Current Liabilities
Quarterly Total Liabilities
Quarterly Share Holder Equity
Quarterly Cash Flow Ops
Quarterly Net Income/Loss
Quarterly Cash Flow Investing
Quarterly Cash Flow Financial
Quarterly Net Cash Flow

TAULA 5: COLUMNES DE LES DADES FONAMENTALS RECOPILADES (PER MACROECONÒMIQUES USA)

GDP
Unemployment Rate
Federal Funds Effective Rate
Consumer Price Index for All Urban Consumers
Trade Balance
Industrial Production: Total Index
University of Michigan: Consumer Sentiment
Advance Retail Sales: Retail Trade
Federal Debt: Total Public Debt

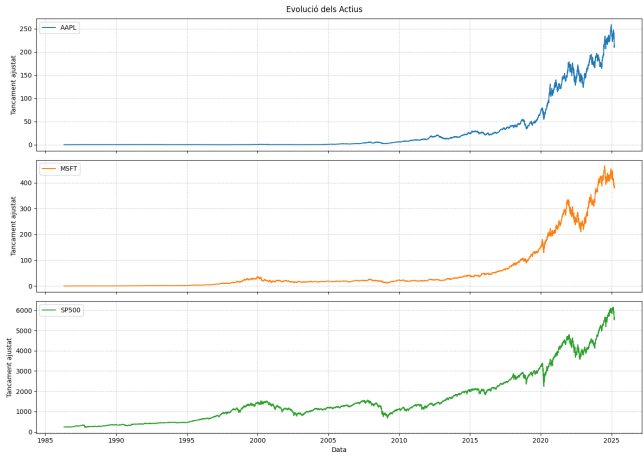


Fig. 5: Comparació dels tancaments ajustats dels tres actius

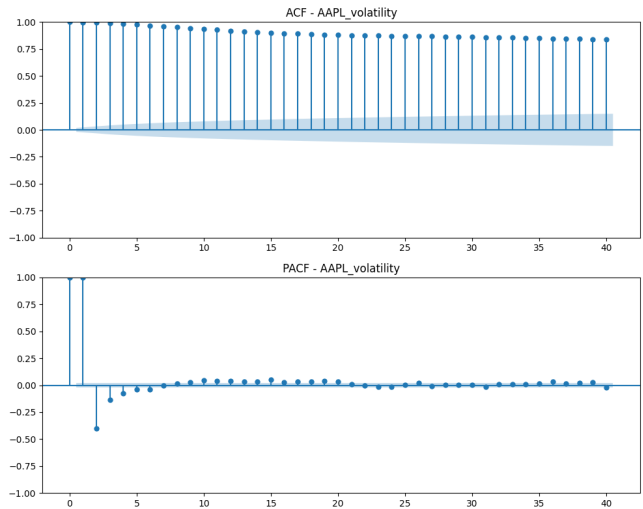


Fig. 6: Exemple de gràfic d'ACF i PACF (volatilitat d'AAPL)

TAULA 6: MILLORS RESULTATS PER CADA ACTIU I GRANULARITAT TEMPORAL AMB MODELS DE CLASSIFICACIÓ MULTICLASSE

		Accuracy	F1-Score	Precision	Model
AAPL	Diari	0.271	0.285	0.368	BalancedRandomForest
	Setmanal	0.483	0.419	0.457	XGBoost
	Mensual	0.413	0.421	0.563	CatBoost
MSFT	Diari	0.384	0.344	0.395	XGBoost
	Setmanal	0.433	0.391	0.401	BalancedRandomForest
	Mensual	0.652	0.544	0.484	BalancedRandomForest
SP500	Diari	0.458	0.458	0.471	BalancedRandomForest
	Setmanal	0.658	0.654	0.667	XGBoost
	Mensual	0.521	0.422	0.388	XGBoost



Fig. 7: Gràfic de comparació de rendiment dels preus de Tesla i Ford i rendiment del portafoli amb els models entrenats (no basats en teories)

```

SHAP_MSFT_volume = 0.0000 + 0.0009*MSFT_risk_diari_lag1 + 0.0010*MSFT_risk_diari_lag5 + 0.0008*MSFT_adjclose + 0.0002*MSFT_AIR +
0.0001*SP500_daily_perf + 0.0005*SP500_daily_range + 0.0012*SP500_adjclose + -0.0007*AAPL_adjclose + -0.0001*MSFT_ADX +
-0.0002*industrial_production_us

```

Fig. 8: Exemple de fórmules per explicar contribució SHAP en les prediccions de risc (MSFT_volume)

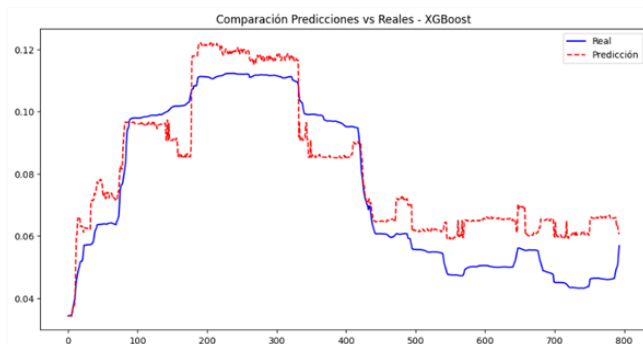


Fig. 9: Gràfic de comparació entre valors predits i reals - Model amb bon rendiment

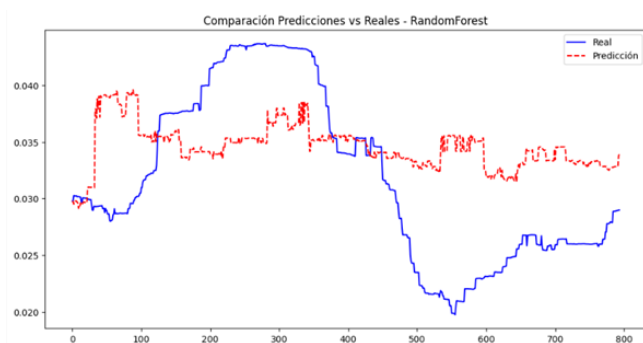


Fig. 10: Gràfic de comparació entre valors predits i reals - Model amb mal rendiment