Plan de Projet

Analyse et Prévision du Marché Boursier

Réalisé par :

Chaimaa EL AFFAS & Youssef AIT BAHSSIN

Encadré par :

Pr. KHADIJA BOUZAACHANE

Cadre :

Projet de fin de module Data Mining

Décembre 2025

Table des matières

1 Contexte et Problématique

1.1 Contexte

Le marché boursier constitue un environnement économique dynamique où les prix des actifs évoluent en fonction d'une multitude de facteurs interdépendants : indicateurs macroéconomiques, résultats financiers des entreprises, événements géopolitiques, sentiment des investisseurs et comportements collectifs. Cette complexité rend la prévision des cours boursiers particulièrement difficile, mais aussi stratégiquement cruciale pour les décideurs financiers.

L'émergence du Data Mining et de l'apprentissage automatique a révolutionné l'analyse financière en permettant d'exploiter d'importants volumes de données historiques pour détecter des patterns non évidents et modéliser des relations non-linéaires. Les algorithmes de Machine Learning et de Deep Learning offrent aujourd'hui des capacités prédictives supérieures aux méthodes statistiques traditionnelles, ouvrant de nouvelles perspectives pour l'aide à la décision en finance.

1.2 Problématique

Dans un contexte de volatilité accrue des marchés financiers et de disponibilité massive de données, comment peut-on exploiter les techniques avancées de Data Mining pour :

— Modéliser efficacement l'évolution temporelle des prix d'actions ?

— Identifier les indicateurs techniques les plus prédictifs ?

— Comparer rigoureusement différentes approches d'apprentissage automatique ?

— Construire un système de prévision robuste et généralisable ?

1.3 Objectifs

— Analyser les tendances et patterns historiques du marché boursier sur la période 2020-2025

— Développer et comparer six approches de modélisation (Machine Learning et Deep Learning)

— Évaluer les performances prédictives selon des métriques quantitatives rigoureuses

— Identifier les variables et indicateurs techniques les plus influents dans la prévision

— Sélectionner le modèle optimal offrant le meilleur compromis performance-robustesse

2 Données

— **Source :** Yahoo Finance API (yfinance)

— **Période :** Janvier 2020 - Décembre 2025 (6 ans)

— **Actions :** S&P 500, Apple (AAPL), Microsoft (MSFT), Tesla (TSLA)

— **Variables :** Date, Open, Close, High, Low, Volume, Adj Close

— **Volume :** $\sim$1 512 observations par action

3 Processus de Traitement des Données

3.1 Exploration des Données

— Statistiques descriptives (moyenne, écart-type, quartiles)
— Détection des valeurs manquantes et outliers (méthode IQR)
— Analyse de corrélation entre variables
— Visualisations : graphiques temporels, chandeliers, heatmap, histogrammes

3.2 Préparation des Données

Nettoyage des données :

— Traitement des valeurs manquantes par interpolation linéaire
— Détection et analyse contextuelle des outliers (méthode IQR)
— Vérification de la cohérence temporelle des séries

Feature Engineering :

Indicateurs techniques :

— Moyennes mobiles (court, moyen et long terme)
— RSI (Relative Strength Index) pour le momentum
— MACD pour la détection des changements de tendance
— Bandes de Bollinger pour la volatilité
— Volume relatif normalisé

Variables dérivées :

— Rendement quotidien
— Volatilité glissante
— Range quotidien
— Gap d'ouverture

Features temporelles :

— Jour de la semaine (encodage cyclique)
— Mois et trimestre
— Indicateurs de début/fin de mois

Transformation :

— Normalisation des variables pour homogénéiser les échelles
— Création de séquences temporelles pour modèles Deep Learning
— Création de features tabulaires avec lags pour modèles Machine Learning

Division du dataset :

— Entraînement : 70% (2020-2023)

— Validation : 15% (2024)

— Test : 15% (2025)

Note : Division chronologique respectée pour éviter le data leakage.

4 Approches de Modélisation

4.1 Approche 1 : LSTM (Long Short-Term Memory)

Principe : Réseau de neurones récurrent capable de capturer les dépendances temporelles à long terme grâce à sa structure de mémoire cellulaire.

Caractéristiques : Architecture empilée avec régularisation dropout, optimisation par descente de gradient adaptative.

4.2 Approche 2 : GRU-Attention

Principe : Réseau GRU bidirectionnel combiné avec un mécanisme d'attention qui pondère l'importance des différents pas de temps dans la séquence.

Caractéristiques : Traitement bidirectionnel (passé et futur), normalisation par batch, mécanisme d'attention pour l'interprétabilité.

4.3 Approche 3 : XGBoost

Principe : Algorithme d'ensemble basé sur le gradient boosting qui construit séquentiellement des arbres de décision en corrigeant les erreurs des arbres précédents.

Caractéristiques : Régularisation L1/L2, échantillonnage aléatoire, optimisation des splits, importance des features calculée automatiquement.

4.4 Approche 4 : CNN-LSTM Hybride

Principe : Architecture hybride combinant des couches convolutionnelles pour l'extraction automatique de patterns locaux et des couches LSTM pour la modélisation temporelle.

Caractéristiques : Extraction hiérarchique de features, réduction de dimensionnalité par pooling, capture multi-échelle des patterns.

4.5 Approche 5 : Random Forest

Principe : Méthode d'ensemble construisant plusieurs arbres de décision sur des échantillons bootstrap et agrégeant leurs prédictions par moyenne.

Caractéristiques : Robustesse au bruit et outliers, gestion native des interactions non-linéaires, estimation de l'importance des variables.

4.6 Approche 6 : Ensemble Hybride (Voting)

Principe : Méta-modèle combinant les prédictions de plusieurs modèles hétérogènes (LSTM, XGBoost, Random Forest) par moyenne pondérée.

Caractéristiques : Exploitation de la complémentarité entre ML et DL, réduction de la variance, poids optimisés sur ensemble de validation.

5 Évaluation

— **Métriques :** RMSE, MAE, MAPE, R^2

— **Critères :** Précision, temps d'entraînement, stabilité, robustesse, interprétabilité

— **Validation :** Cross-validation temporelle, test sur différentes actions et périodes de volatilité

6 Planning (5 Semaines)

— **Semaine 1 :** Collecte et exploration des données

— **Semaine 2 :** Préparation et feature engineering

— **Semaine 3 :** Modèles LSTM, XGBoost, Random Forest

— **Semaine 4 :** Modèles GRU-Attention, CNN-LSTM, Ensemble Hybride

— **Semaine 5 :** Évaluation finale, comparaison et rapport

7 Livrables

— Rapport technique complet

— Code source (Jupyter Notebook)

— Présentation

— Modèles entraînés sauvegardés

— Dataset préparé