



Université Mohammed V de Rabat
Ecole Nationale Supérieure d'Informatique et d'Analyse des Systèmes

Projet du module data pre-processing

PREDICTION DU PRIX D'UNE CRYPTOMONAIE

Filière :

Ingénierie Digitale pour la Finance - 2ème Année

Réalisé par :

**TRAORE Mohamed Bourema
M'HASNI Youssef
KHALIL Benlamaalam**

Sous l'encadrement de Monsieur :

MOHAMED LAZAAR

2021-2022

Remerciement :

De prime à bord, il serait une obligation pour nous d'adresser nos sincères remerciements à tous ceux qui ont contribué de prêt ou de loin pour la réalisation de ce travail.

D'abord à notre professeur du module Dataprocessing, qui ne ménage aucun effort pour la réussite des étudiants, et aussi à l'ensemble des anciens étudiants pour leurs précieux conseils et guide.

Resumé :

Dans ce document, nous avons décrit les étapes essentielles empruntées pour la réalisation de notre projet.

L'objectif était la prédiction du prix d'une cryptomonaie en l'occurrence le Bitcoin. Le travail a été axé sur 3 point essentiels.

D'abord, nous avons en premier lieu scraper le contenu d'un site qui a constitué notre dataset. Ensuite nous sommes passés à la phase de visualisation des données et leur prétraitement. Nous avons terminé par concevoir un modèle de machine learning pour prévoir le prix de clôture du Bitcoin avec et sans réduction de dimensionnalité.

Abstract :

We describe through this document the step necessary we use to achieve our project of data processing.

The goal was to predict the close price of a crypto, the Bitcoin. It consists of 3 steps.

First of all we scrap the content of a website, which has been our dataset. Then we pass to the step of visualisation and pretreatment.

We finished by implementing an algorithm of machine learning to predict the close price of Bitcoin, with and without reduction of dimensionality

Table des matières

Introduction générale	1
1 Collecte et présentation des données	8
1.1 Introduction	8
1.2 Scraping des données	8
1.3 Description des variables du donnée	9
1.4 Conclusion	9
2 Visualisation et traitement des données	10
2.1 Introduction	10
2.2 Détermination des valeurs manquantes	10
2.3 Recherche des valeurs aberrantes	10
2.4 Type des variables	13
2.5 Conversion du type des variables	13
2.6 Conversion de la colonne Date en index et tri suivant la date	14
2.7 Visualisation des données	14
2.8 Conclusion	17
3 Prédiction du prix de clôture de Bitcoin	18
3.1 Introduction	18
3.2 Choix de l'algorithme d'apprentissage	18
3.3 Prédiction sans réduction de dimensionnalité	18
3.4 Prédiction avec réduction de dimensionnalité	20
3.5 Conclusion	21
Conclusion générale	22

Table des figures

1.1	scraping des données	8
1.2	scraping des données	9
2.1	les valeurs manquantes	10
2.2	boite à moustache pour la variable Open	11
2.3	boite à moustache pour la variable Close	11
2.4	boite à moustache pour la variable Low	11
2.5	boite à moustache pour la variable High	12
2.6	boite à moustache pour la variable Volume	12
2.7	boite à moustache après suppression de valeurs aberrantes dans variable Volume	12
2.8	le type des variables	13
2.9	Conversion type des variables	13
2.10	Conversion Date en index	14
2.11	Low	15
2.12	High	15
2.13	Open	16
2.14	Close	16
2.15	combinaison de toutes les variables	17
3.1	information sur le modèle	18
3.2	Close vs predict Close	19
3.3	Close vs predict Close en graphique	19
3.4	Matrice de corrélation	20
3.5	affichage du nouveau jeu de données après ACP	20
3.6	Closes et predicts Close après ACP	21
3.7	Closes et predicts Close sur un graphique	21

Introduction générale

Dans le présent document qui est le rapport du projet du module dataprocessing , nous allons décrire le chemin que nous avons frayé pour la mettre en œuvre.

En effet le data preprocessing est une étape très importante et primordiale dans tous les projets qui traitent des données de grandes tailles et multidimensionnelles. Ces données peuvent parfois présenter des irrégularités, des incomplétudes, des inutilités ou même des manquements nous conduisant ainsi à des modèles complètement erronés ou très peu performants. Le data preprocessing apporte une solution à ces maux afin d'avoir des données qui correspondent mieux à nos projets.

L'objectif de notre projet est de prédire le prix d'une cryptomonaie en l'occurrence le Bitcoin, à l'aide des réseaux de neurones multicouches. Pour ce faire nous allons scraper les données du site web **ADVFM**, opérer le data pre-processing sur ces données, et par la suite passer à la prédiction.

Chapitre 1

Collecte et présentation des données

1.1 Introduction

Dans ce chapitre, il sera question d'extraire le contenu d'un site web, qui va constituer notre data. Les données ont été scrapées à l'adresse suivante **ADVFN**. Il s'agit des mesures sur la fluctuation du cour du Bitcoin, de la période du 12/02/2019 au 28/01/2022

1.2 Scraping des données

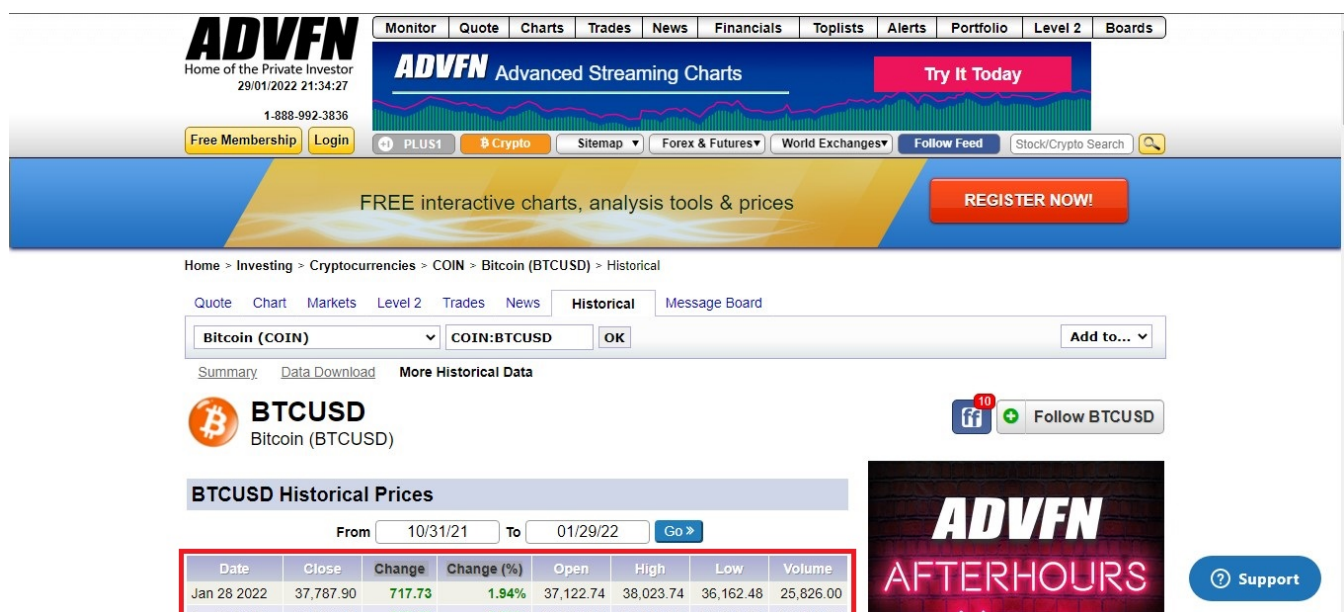
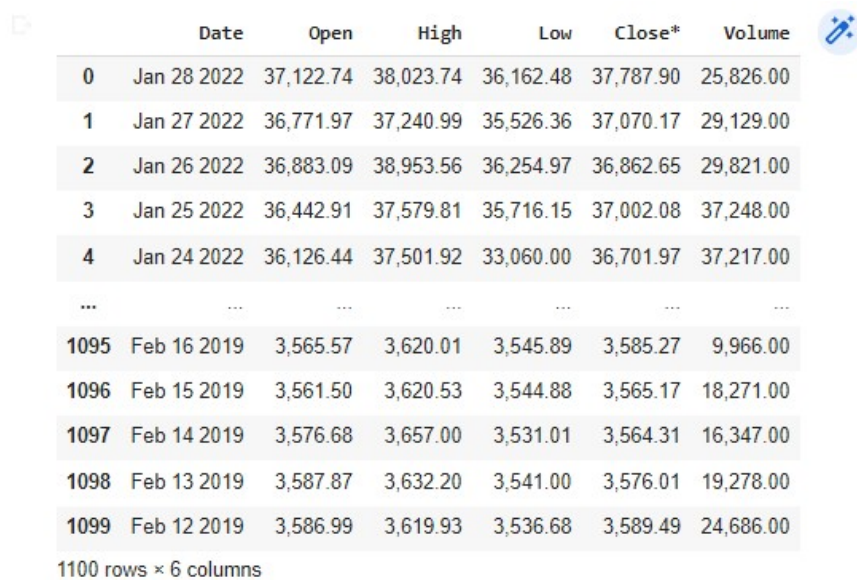


FIGURE 1.1 – scraping des données

Cette image correspond à la page où nous allons scraper les données. La section cadrée en rouge correspond aux données que nous allons scraper. Les variables **Change**, **Changes(%)** seront pas pris en compte, sinon les données présenteront des redondances



	Date	Open	High	Low	Close*	Volume
0	Jan 28 2022	37,122.74	38,023.74	36,162.48	37,787.90	25,826.00
1	Jan 27 2022	36,771.97	37,240.99	35,526.36	37,070.17	29,129.00
2	Jan 26 2022	36,883.09	38,953.56	36,254.97	36,862.65	29,821.00
3	Jan 25 2022	36,442.91	37,579.81	35,716.15	37,002.08	37,248.00
4	Jan 24 2022	36,126.44	37,501.92	33,060.00	36,701.97	37,217.00
...
1095	Feb 16 2019	3,565.57	3,620.01	3,545.89	3,585.27	9,966.00
1096	Feb 15 2019	3,561.50	3,620.53	3,544.88	3,565.17	18,271.00
1097	Feb 14 2019	3,576.68	3,657.00	3,531.01	3,564.31	16,347.00
1098	Feb 13 2019	3,587.87	3,632.20	3,541.00	3,576.01	19,278.00
1099	Feb 12 2019	3,586.99	3,619.93	3,536.68	3,589.49	24,686.00

1100 rows x 6 columns

FIGURE 1.2 – scraping des données

Après scraping, nous avons les données brutes présentées dans l'image ci-dessus. Il s'agit de 1100 lignes sur 6 colonnes.

1.3 Description des variables du donnée

Les caractéristiques qui définissent nos données sont :

- **Date** : la date de la mesure des variables
- **Open** : le cour d'ouverture
- **High** : le point maximal atteint que le cour atteignit
- **Low** : le point minimal atteint que le cour atteignit
- **Close** : cour de fermeture
- **Volume** : la valeur totale de l'échange effectuée en milliard de Dollard

1.4 Conclusion

L'objectif de ce chapitre était d'extraire le contenu d'un site web et de décrire ses caractéristiques. Dans le chapitre suivant nous allons passer au prétraitement de nos données.

Chapitre 2

Visualisation et traitement des données

2.1 Introduction

Dans ce chapitre il sera question de nettoyer nos données. c'est une étape essentielle pour avoir des données adaptées avant de passer à la prédiction

2.2 Determination des valeurs manquantes

Nous allons verifier si une variable contient des valeurs manquantes.

number of missing value	
Date	0
Open	0
High	0
Low	0
Close*	0
Volume(T)	0

FIGURE 2.1 – les valeurs manquantes

Nous constatons qu'aucune des variables ne presentent des valeurs manquantes.

2.3 Recherche des valeurs abbérantes

Pour déterminer les valeurs abbérantes des variables, nous allons afficher les boites à moustaches de chacune d'elles.

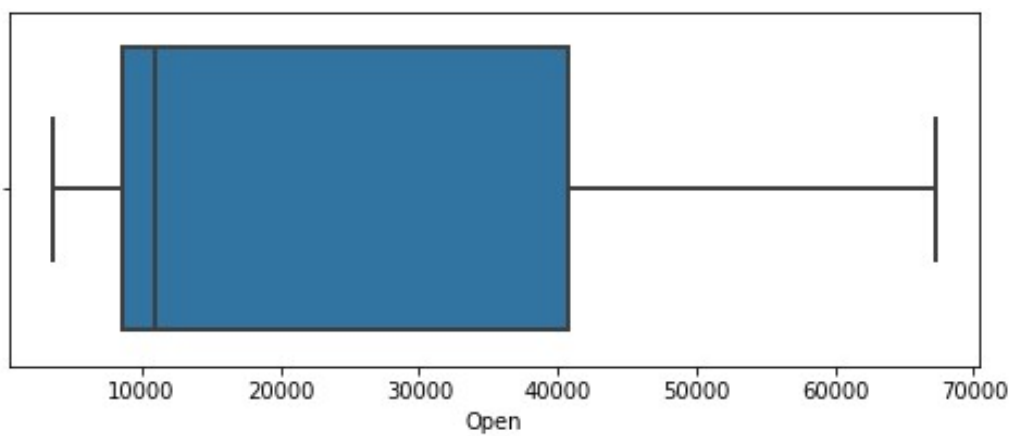


FIGURE 2.2 – boîte à moustache pour la variable Open

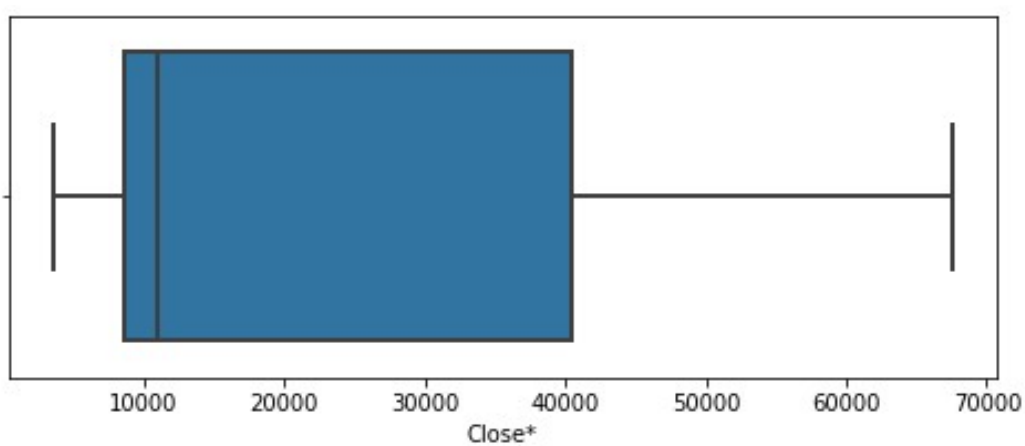


FIGURE 2.3 – boîte à moustache pour la variable Close

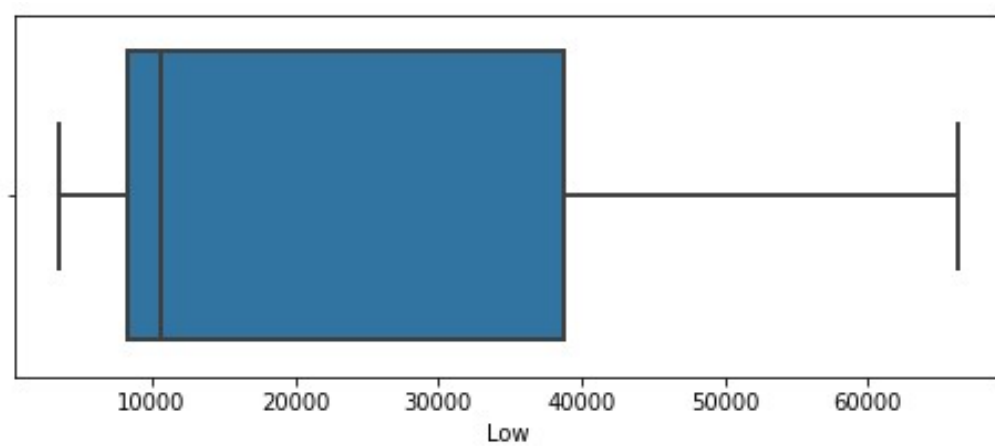


FIGURE 2.4 – boîte à moustache pour la variable Low

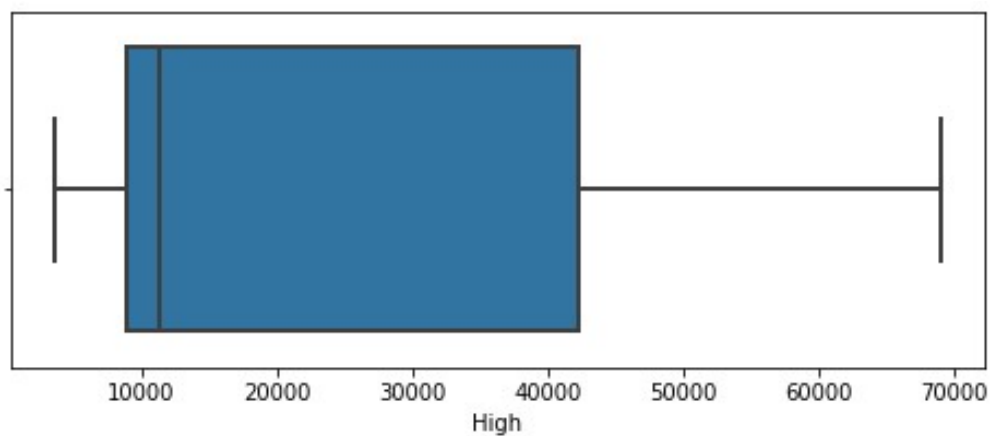


FIGURE 2.5 – boîte à moustache pour la variable High

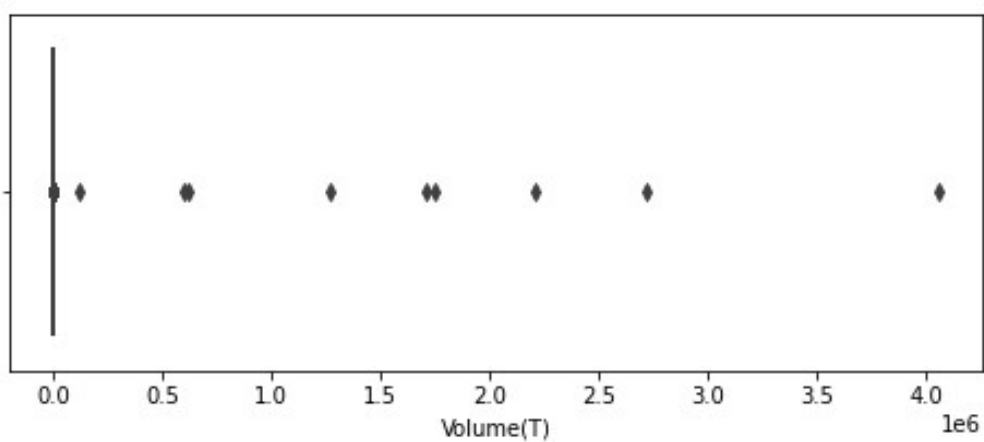


FIGURE 2.6 – boîte à moustache pour la variable Volume

Nous remarquons que nous avons des valeurs aberrantes dans la variable Volume. Nous allons les supprimer

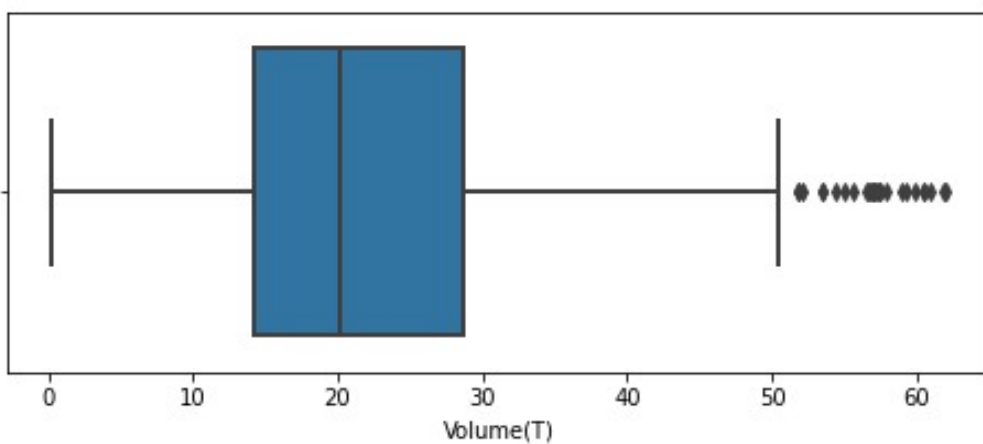
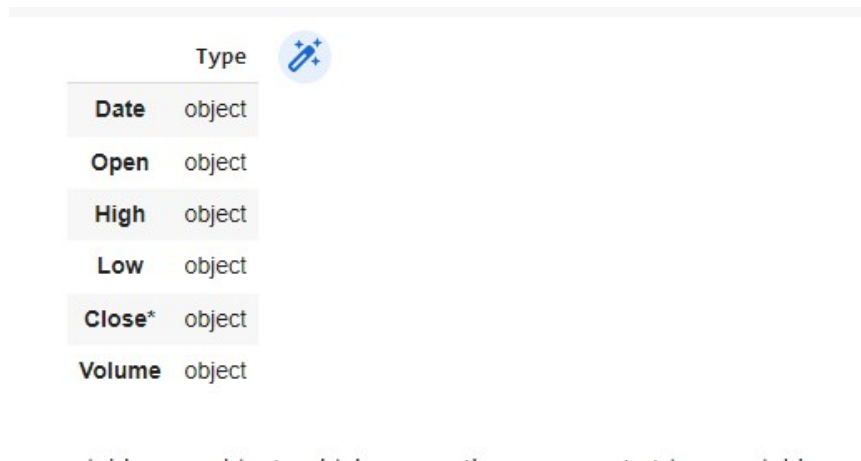


FIGURE 2.7 – boîte à moustache après suppression de valeurs aberrantes dans variable Volume

2.4 Type des variables



	Type
Date	object
Open	object
High	object
Low	object
Close*	object
Volume	object

FIGURE 2.8 – le type des variables

Toutes les variables sont de types objects. Nous allons donc les convertir en de type numériques.

2.5 Conversion du type des variables

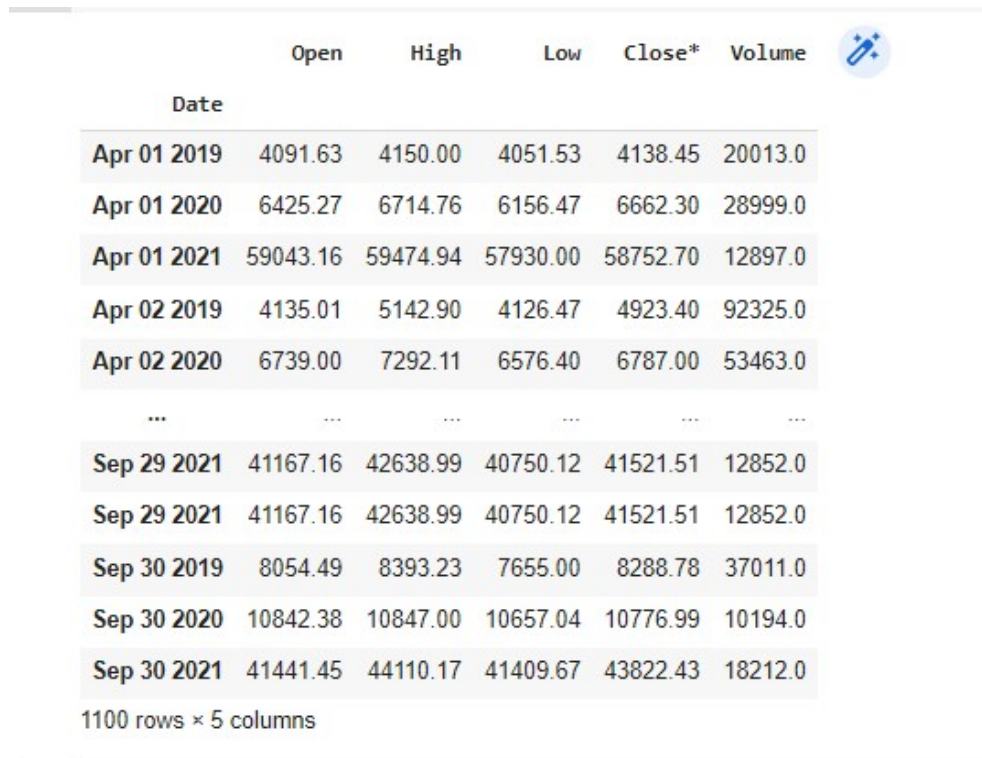


	Type
Date	object
Open	float64
High	float64
Low	float64
Close*	float64
Volume	float64

FIGURE 2.9 – Conversion type des variables

Toutes les variables maintenant à part Date sont de type numérique.

2.6 Conversion de la colonne Date en index et tri suivant la date



Date	Open	High	Low	Close*	Volume
Apr 01 2019	4091.63	4150.00	4051.53	4138.45	20013.0
Apr 01 2020	6425.27	6714.76	6156.47	6662.30	28999.0
Apr 01 2021	59043.16	59474.94	57930.00	58752.70	12897.0
Apr 02 2019	4135.01	5142.90	4126.47	4923.40	92325.0
Apr 02 2020	6739.00	7292.11	6576.40	6787.00	53463.0
...
Sep 29 2021	41167.16	42638.99	40750.12	41521.51	12852.0
Sep 29 2021	41167.16	42638.99	40750.12	41521.51	12852.0
Sep 30 2019	8054.49	8393.23	7655.00	8288.78	37011.0
Sep 30 2020	10842.38	10847.00	10657.04	10776.99	10194.0
Sep 30 2021	41441.45	44110.17	41409.67	43822.43	18212.0

1100 rows × 5 columns

FIGURE 2.10 – Conversion Date en index

Cette conversion aura pour avantage de prendre la date comme unité de mesure des autres variables puisqu'elle n'influe pas sur le modèle qui sera conçu.

2.7 Visualisation des données

Visualisation des variables Close Price, Low, Hight, Adjusted close, Price par la date de la période du 12/02/2019 au 28/01/2022

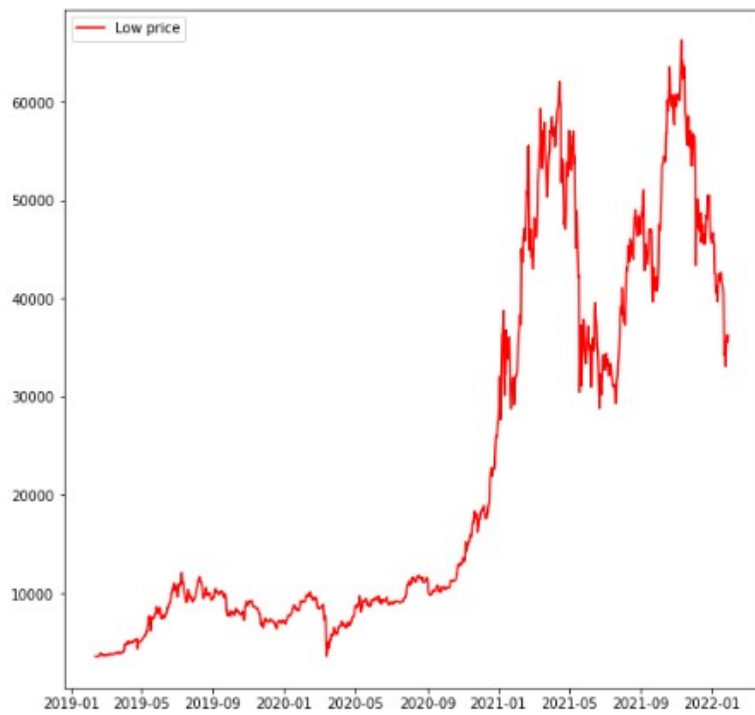


FIGURE 2.11 – Low

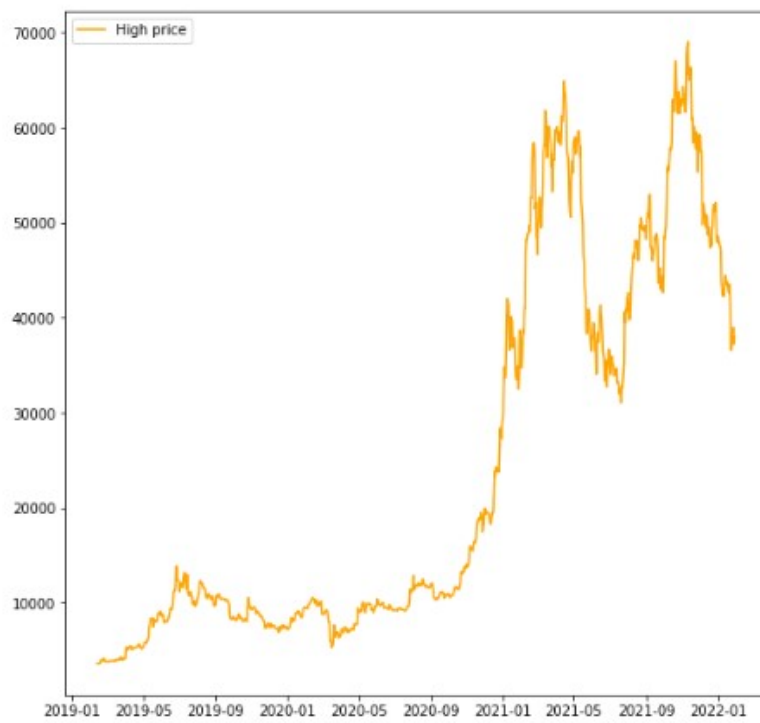


FIGURE 2.12 – High

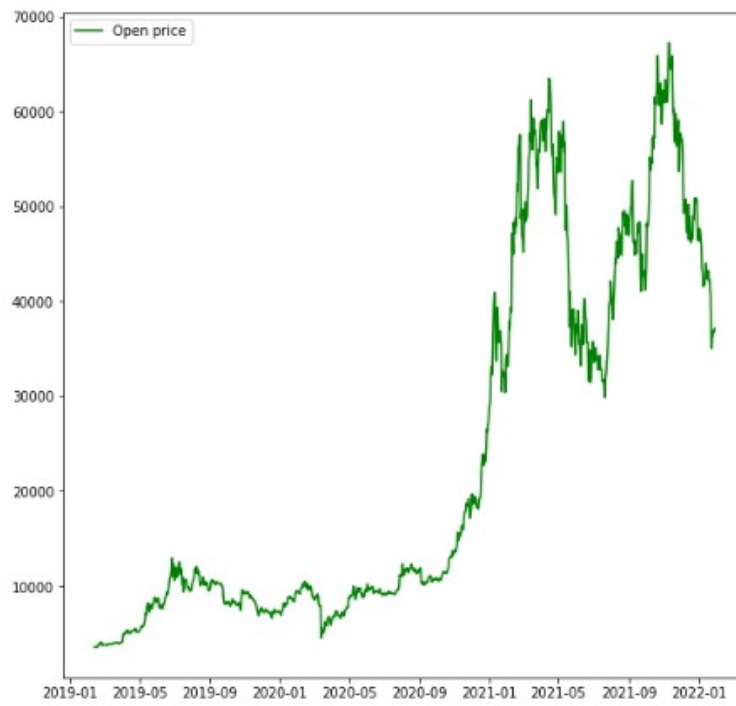


FIGURE 2.13 – Open

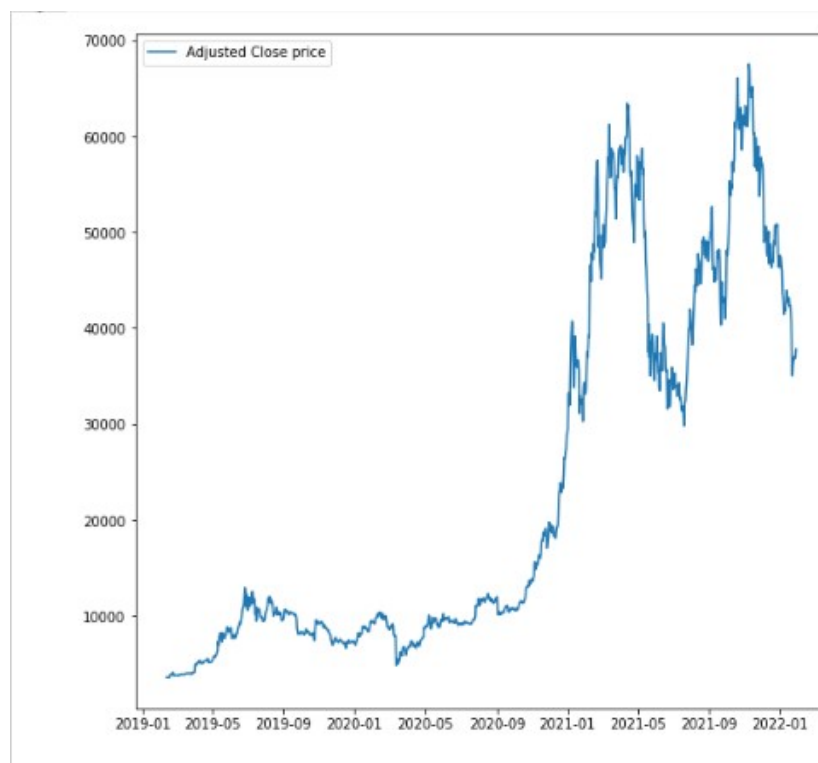


FIGURE 2.14 – Close



FIGURE 2.15 – combinaison de toutes les variables

2.8 Conclusion

Nous sommes enfin arrivés à bout de ce chapitre. Après nettoyage de nos données, dans le chapitre suivant nous allons utiliser ces données pour la prédiction.

Chapitre 3

Prédiction du prix de cloture de Bitcoin

3.1 Introduction

Nous sommes arriver à la phase prédiction. Ce qui fera l'objet de ce chapitre. Nous allons le faire, avant et après la réduction de dimensionnalité.

3.2 Choix de l'algorithme d'apprentissage

L'algorithme que nous allons implementer est le réseau de neurone multi-couche. Ce choix est justifié par le fait que nous faisons face à un problème de prédiction et les réseaux de neurones multi-couches peuvent mieux résoudre ce genre de problème.

3.3 Prédiction sans réduction de dimensionnalté

Dans cette partie, nous allons implementer l'algorithme pour prédire sans réduction de dimensionnalité. Dabord nous allons diviser notre données en les données d'entrainement et les données de tes, on prend 80% pour l'entrainement et 20% pour le test. Nous allons maintenant créer notre modèle.

Layer (type)	Output Shape	Param #
dense_29 (Dense)	(None, 120)	480
dense_30 (Dense)	(None, 120)	14520
dense_31 (Dense)	(None, 120)	14520
dense_32 (Dense)	(None, 1)	121
Total params: 29,641		
Trainable params: 29,641		
Non-trainable params: 0		

FIGURE 3.1 – information sur le modèle

Cette image montre les informations de notre modèle, nombre de couches, de neurones par couche, et le nombre des parametres.

On choisi **Adam** gradient pour optimiser error et la fonction **mean squared error** comme fonction d'erreur qui sera optimisé.

Après l'entraînement de notre modèle, on trouve une erreur : **mean squared_error = 554.6103644938019**

Nous allons maintenant comparer les valeurs prédites avec notre modèle contre les vraies valeurs dans les données de test.

	Close*	predicted Close
Date		
Nov 19 2021	58031.22	57851.566406
Nov 20 2019	8074.09	8373.978516
Nov 20 2020	18668.30	18653.996094
Nov 20 2021	59850.00	59624.816406
Nov 21 2019	7614.15	7944.830078
...
Sep 29 2021	41521.51	42465.796875
Sep 30 2019	8288.78	8231.642578
Sep 30 2020	10776.99	10936.791016
Sep 30 2021	43822.43	43904.367188
Sep 30 2021	43822.43	43904.367188

220 rows × 2 columns

FIGURE 3.2 – Close vs predict Close

Dans l'image on constate que les valeurs predictes et les vraies valeurs sont plus ou moins proches.



FIGURE 3.3 – Close vs predict Close en graphique

Cette image illustre mieux notre conclusion, les deux courbes se suivent plus ou moins, on conclut que la qualité de la prediction est assez bonne.

3.4 Prédiction avec réduction de dimensionnalité

Dans cette seconde partie, la prédiction sera faite après avoir fait la réduction de dimensionnalité sur les données.

Nous allons utiliser ACP pour faire la réduction de dimensions.

L'application d'ACP sur notre données se fera avec choix de 2 axes factoriel car ce nombre d'axe donne 99.99% d'inertie et pour un seul axe on trouve 80% d'inertie.

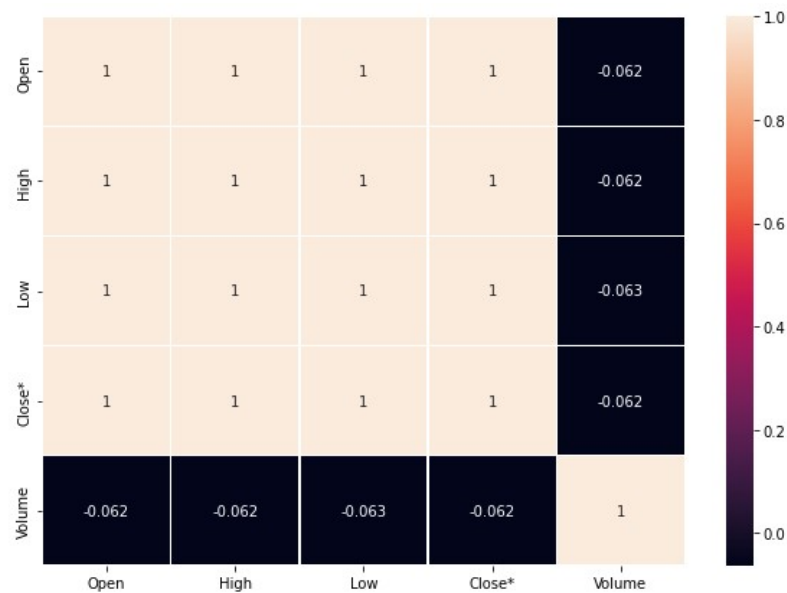


FIGURE 3.4 – Matrice de corrélation

	PC1	PC2
0	-2.034961	-0.159357
1	-2.034860	-0.159382
2	-2.035097	-0.159408
3	-2.036024	-0.159436
4	-2.035376	-0.159455
...
1095	1.340214	-0.019222
1096	1.430084	-0.015530
1097	1.487438	-0.013187
1098	1.426719	-0.015719
1099	1.491740	-0.013034

1100 rows × 2 columns

FIGURE 3.5 – affichage du nouveau jeu de données après ACP

Maintenant on entraîne notre modèle avec ces nouvelles données .

D'abord on divise notre nouveau jeu de données en deux parties, l'une pour entraînement et

l'autre pour le test.

Après avoir conçu le modèle on trouve une erreur **mean squared_error = 3119.06601765526**

	Close*	predicted Close
Date		
Nov 19 2021	58031.22	62978.031250
Nov 20 2019	8074.09	8622.345703
Nov 20 2020	18668.30	19932.697266
Nov 20 2021	59850.00	64758.984375
Nov 21 2019	7614.15	8288.044922
...
Sep 29 2021	41521.51	45656.218750
Sep 30 2019	8288.78	8590.117188
Sep 30 2020	10776.99	11555.553711
Sep 30 2021	43822.43	46947.121094
Sep 30 2021	43822.43	46947.121094

220 rows x 2 columns

FIGURE 3.6 – Closes et predicts Close après ACP

On constate une difference entre les valeurs des Closes et predicts Close.



FIGURE 3.7 – Closes et predicts Close sur un graphique

Avec le graphique, on constate mieux la difference.

3.5 Conclusion

Ce chapitre s'achève à présent. Nous avons implementé notre modèle avec et sans réduction de dimensionnalité. Les résultats selon les deux cas diffèrent.

Conclusion générale

L'objectif de ce chapitre était la prédiction du prix d'une cryptomonaie en l'occurrence le Bitcoin, par un modèle de machine learning. Le travail a été réparti sur 3 chapitres.

Dans le premier chapitre qui a fait l'objet de collecte des données. Nous avons scrapé le contenu d'un site web. Ces données étaient décrits par les variables suivantes : Date, Close, Change, Change(%), Open, High, Low, Volume. Les deux variables Change, Change(%) ont été écartées puisqu'elles ne donnaient pas d'informations supplémentaires.

Dans le second chapitre nous avons traitées ces données. Le but de ce chapitre a été de visualiser les données et de les traiter afin d'avoir des données adaptées qui reflète mieux notre problème avant de passer à la prédiction.

Dans le troisième chapitre, nous sommes passés la conception du modèle proprement dit. Elle a été faite en 2 étapes : le modèle avec et sans réduction de dimensionnalité. L'algorithme implémenté est le réseau de neurone multi-couche. Le modèle sans réduction de dimensionnalité à présenter mean square error largement inférieur à celui du modèle avec réduction de dimensionnalité, nous pouvons donc conclure que le premier est plus efficace. Cette affirmation a été mieux illustré avec le graphique qui a opposé la valeur Close et predict Close pour les deux modèles. .

Bibliographie

- [1] <https://www.youtube.com/watch?v=DxzkmNkUU-w>
- [2] <https://www.advn.com/stock-market/COIN/BTCUSD/historical/more-historical-data?current=0Date1=10/01/19Date2=01/29/22>
- [3] <https://docs.streamlit.io/knowledge-base/tutorials>
- [4] <https://medium.com/analytics-vidhya/stock-price-prediction-ade10ac8ce08>
- [5] <https://medium.com/codex/stock-price-prediction-a-modified-approach-8d63ea6726a7>