

Rapport de stage:



Prédiction de la production des stations KOCH et KRUPP

Écrit par:

- KBALA Youssef | INPT
- DAHHASSI Chaymae | ECC

Encadré par:

- Mr. FAHMI Abderrahim

REMERCIEMENTS

Au terme de notre stage, nous exprimons nos remerciements les plus sincères à Allah pour nous avoir accordé la force et la détermination nécessaires pour surmonter les défis et mener à bien ce travail.

Nous saissons cette occasion pour exprimer notre profonde gratitude envers le Groupe OCP, qui nous a offert l'opportunité exceptionnelle de réaliser notre stage au sein de leur entreprise. Ce stage a été une expérience inestimable qui nous a permis de développer nos compétences professionnelles et d'explorer de nouvelles approches dans notre domaine d'études.

Nos remerciements vont tout particulièrement à M. FAHMI Abderrahim, notre chef d'atelier Électronique au sein du Groupe OCP Benguerir, pour son soutien précieux et ses enseignements. Sa disponibilité et son partage généreux de connaissances ont grandement contribué à notre apprentissage et à notre compréhension des défis du Groupe OCP dans son secteur d'activité.

Enfin, nous adressons nos remerciements les plus sincères à tous les membres de l'équipe du Groupe OCP qui ont supervisé notre travail, pour leur temps et leur énergie investis. Leurs précieux commentaires et encouragements ont été d'une aide inestimable pour mener à bien ce stage et pour notre développement personnel et professionnel.

En conclusion, cette expérience enrichissante au sein du Groupe OCP restera gravée dans notre mémoire, et nous sommes profondément reconnaissants envers tous ceux qui ont contribué à en faire une expérience aussi fructueuse et mémorable. Nous quittons ce stage avec une reconnaissance infinie et une volonté renouvelée de poursuivre notre parcours professionnel avec détermination et succès.

Table des matières

INTRODUCTION	1
Chapitre I : Présentation de l'organisme d'accueil :.....	2
1. Brève présentation d'une envergure mondiale : l'OCP :.....	2
2. Gamme de produits :.....	4
3. Historique du Groupe OCP:.....	5
4. Mission et valeurs :.....	7
5. Activités et projets du Groupe OCP :	7
6. Présentation de la mine de Benguerir :	9
7. Processus de production des phosphates :.....	10
Chapitre II : Présentation du projet :.....	16
1. Cahier de charges :	16
a. Besoin et enjeux :	16
b. Objectifs :	16
2. Conduite du projet :.....	17
Chapitre III : Pipeline du projet :	19
I. Spécification et compréhension du problème :	19
1. Compréhension métier :.....	19
2. Technologies utilisées :	19
II. Compréhension approfondie des données.....	20
1. Séries temporelles ou Time series :.....	20
2. Aperçu des jeux de données :	20
3. Prétraitement des données :.....	21
4. Exploration et visualisation des données :.....	24
a. Line Plot :	24
b. Décomposition de séries temporelles :	25
c. Graphiques de sous-séries saisonnières ou Seasonal subseries Plots:.....	27
d. Box Plot :	28
e. Carte thermique ou Heatmap :	29
f. Statistiques glissantes ou Rolling statistics :.....	30
g. Diagramme de dispersion retardée ou le Lag Scatter Plot:.....	31
h. Histogramme :	32
III. Préparation des données :	33
1. Nettoyage des données :	33
2. Normalisation et mise à échelle :	33
3. Séparation des ensembles de données :	34

IV.	Modélisation :	35
1.	Modèle prédictif choisi: LSTMs	35
i.	Les RNNs :	35
ii.	Limitations des RNNs :	35
iii.	Les LSTMs :	36
iv.	Composants d'un LSTM :	37
v.	L'architecture d'un LSTM :	38
2.	Construction du modèle prédictif :	39
3.	Entraînement du modèle :	39
V.	Evaluation :	40
1.	Evaluation de la qualité et performance du modèle sur les données de test :	40
2.	Comparaison des modèles :	43
3.	Évaluation des modèles sur les données du test :	48
4.	Analyse des Courbes de Validation et de Prédictions :	50
5.	Prédition de valeurs de THC futures :	53
VI.	Déploiement :	55
1.	Outils et cadres de déploiement utilisés :	56
2.	Environnement de développement : Visual Studio Code	57
3.	Réalisation de la plateforme :	57
	Webographie :	61
	Annexe :	62

INTRODUCTION

C'est avec un sentiment de gratitude et d'accomplissement que nous présentons ce rapport de stage réalisé au sein du Groupe OCP Benguerir, l'un des acteurs majeurs du secteur minier et agricole au Maroc.

L'OCP, Office Chérifien des Phosphates, incarne depuis sa création une référence incontournable dans le domaine de l'exploitation minière et de la transformation des phosphates. Évoluant dans un secteur stratégique, l'entreprise joue un rôle primordial tant sur le plan économique que sur le développement du pays. De ce fait, intégrer l'OCP pour un stage a été une opportunité inestimable pour acquérir des compétences précieuses et découvrir le fonctionnement d'une entreprise d'envergure internationale.

Ce rapport se veut être le témoignage de notre immersion au sein de cette institution prestigieuse. Tout au long du stage, nous avons eu l'opportunité de collaborer avec des personnes compétentes, amicales et dynamiques qui nous ont enrichies en abordant divers sujets passionnants. Nous avons été encadrés par des professionnels chevronnés qui ont su partager leur savoir-faire avec bienveillance et enthousiasme. Cette expérience en équipe a été des plus stimulantes, nous permettant d'apprendre et de progresser dans un environnement propice à l'épanouissement personnel et professionnel.

Dans un premier temps, ce rapport dressera une présentation de l'OCP, mettant l'accent sur son histoire, son organisation et ses missions fondamentales. Ensuite, nous aborderons le projet auquel nous avons été associés, détaillant les objectifs qui nous ont été confiés et les différentes missions que nous avons accomplies. Nous partagerons également les défis auxquels nous avons été confrontés et les solutions que nous avons pu proposer pour les surmonter.

Enfin, nous conclurons ce rapport en évoquant les enseignements que nous avons tirés de cette expérience enrichissante et les perspectives qu'elle nous a ouvert.

Chapitre I : Présentation de l'organisme d'accueil :

1. Brève présentation d'une envergure mondiale : l'OCP :

Le Groupe OCP, acronyme de l'Office Chérifien des Phosphates, occupe une place prépondérante à l'échelle mondiale dans l'extraction, la transformation, la valorisation et l'exportation des phosphates. Fondé le 7 août 1920, il est devenu l'un des acteurs clés de l'industrie grâce à son envergure et à son impact significatif.



Figure 1 : Logo du Groupe OCP

Actuellement, il compte près de 20 000 collaborateurs, principalement répartis sur quatre sites miniers et deux complexes chimiques au Maroc, ainsi que sur d'autres sites internationaux.

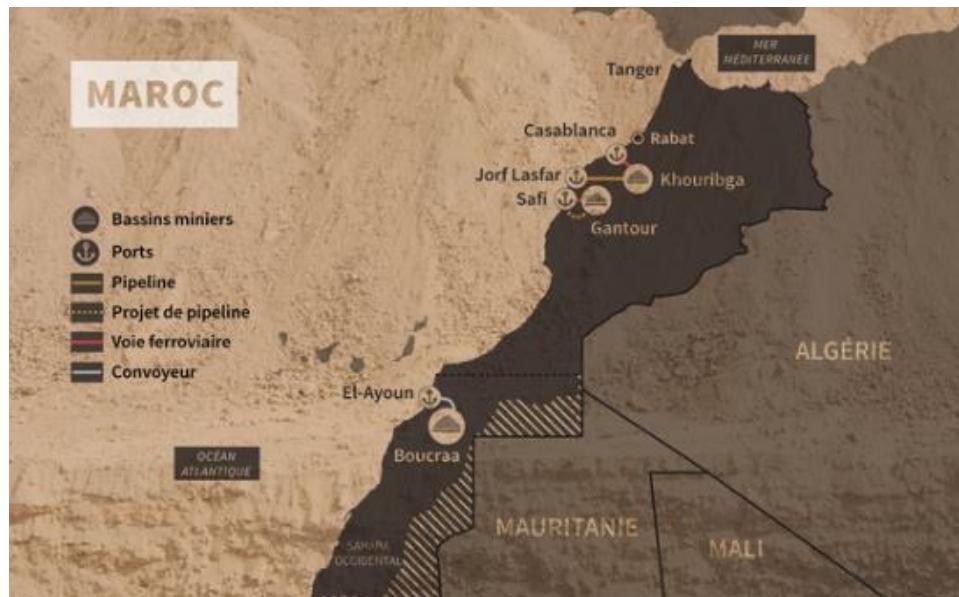


Figure 2: Carte des sites d'implantation de l'OCP au Maroc

Les principaux gisements se trouvent à :

- *GANTOUR* situé dans la région de Youssoufia Benguerir.
- *OULAD ABDOUN* situé dans la région de Khouribga.
- *BOUCRAA* situé dans la région de Boucraa Laayoune.
- *MESKALA* situé dans la région d'Essaouira, gisement non encore exploité

Les centres de transformations chimiques se trouvent à :

- *Jorf Lasfar*
- *Safi*

Ports d'embarquements :

- *Casablanca*
- *Jorf Lasfar*
- *Safi*
- *Laayoune*

En 2008, le Groupe OCP s'est transformé en une société anonyme, dénommée « OCP S.A », pour mieux répondre aux défis du marché mondial. Son capital est majoritairement détenu par l'État, à hauteur de 95 %, tandis que les 5 % restants sont détenus par la Banque centrale populaire. Cette structure lui permet de bénéficier d'une solide assise financière pour poursuivre ses activités et ses projets d'envergure.

Depuis des années, Le Maroc occupait le haut de la liste, avec des réserves géantes dépassant les **50 milliards de tonnes** (plus de **70%** des réserves mondiales), ce qui a permis à l'OCP de devenir l'un des plus grands producteurs et exportateurs de phosphates au monde.

Cependant, dernièrement, le Maroc a reculé à la deuxième place dans la liste des plus grandes réserves de phosphate au monde. Et ce, après que la Norvège a annoncé la découverte du plus grand stock de roche phosphatée au monde avec une réserve de **70 milliards de tonnes** de phosphates, détrônant ainsi le Maroc de sa position de leader en termes de réserves.

Malgré cette évolution, l'OCP continue à être l'un des géants incontestés du secteur, bénéficiant d'une solide expertise et d'une vaste expérience dans le domaine des phosphates. Grâce à sa présence internationale, à ses partenariats stratégiques et à ses efforts continus pour l'innovation et le développement durable, l'OCP maintient son

leadership en contribuant de manière significative à l'industrie des phosphates et à l'agriculture mondiale. Sa capacité à s'adapter aux nouvelles réalités du marché lui permet de relever les défis avec résilience et de saisir les opportunités pour un avenir prometteur.

2. Gamme de produits :

En tant que l'une des principales entreprises mondiales du secteur des phosphates, l'OCP propose une vaste gamme de produits liés à l'exploitation minière, à la transformation des phosphates et à l'industrie des engrains, à savoir :

- *La roche phosphatée* : utilisée comme matière première dans la fabrication d'engrais, de l'alimentation animale et dans d'autres utilisations industrielles.
- *Acide phosphorique* : fabriqué à partir de la réaction entre la roche phosphatée et l'acide sulfurique, on y distingue deux types: l'acide phosphorique marchand, utilisé dans les engrais phosphatés, et l'acide phosphorique purifié, utilisé dans l'industrie pharmaceutique, alimentaire et textile.
- *Engrais* : L'OCP fabrique et commercialise différents types d'engrais phosphatés, tels que des engrais NPK (azote, phosphore, potassium) et des engrais DAP (phosphate diammonique).
- *Engrais sur mesure* : L'OCP propose des solutions d'engrais sur mesure, adaptées aux besoins spécifiques des cultures et des sols dans différentes régions du monde.
- *Superphosphate* : engrais fabriqués à partir de la réaction entre la roche phosphatée brute et un pourcentage précis de l'acide sulfurique et l'eau.

3. Historique du Groupe OCP:

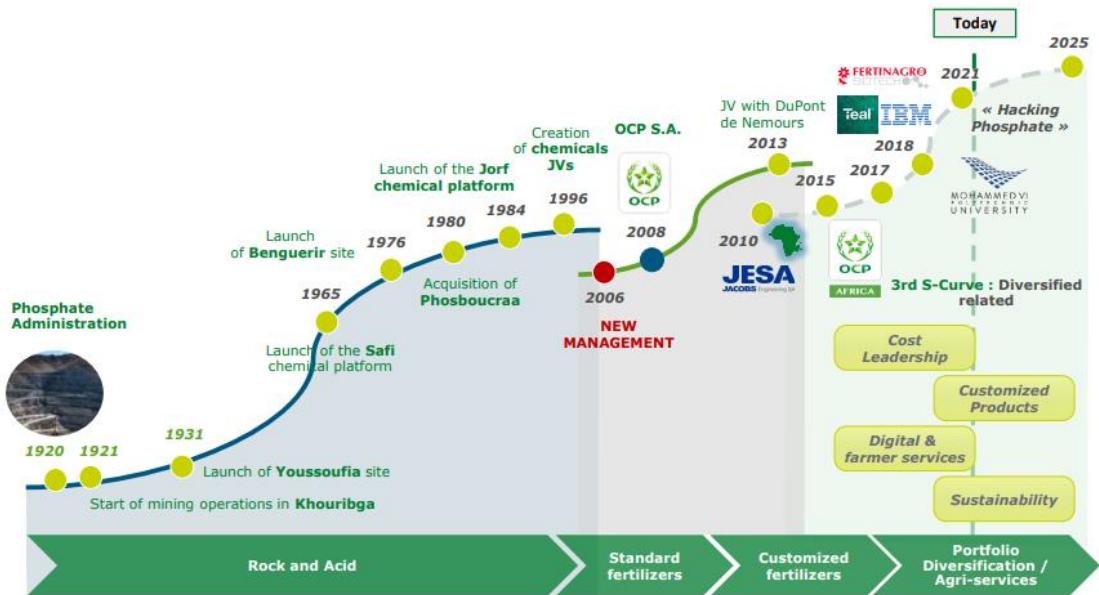


Figure 3: Évolution Historique de l'OCP jusqu'à 2025

Source : [Résultats financiers | OCP Group](#)

Date	Évènement
1921	- Début de l'extraction du phosphate à Boujniba dans la zone de Khouribga (1er Mars) - Première exportation de phosphate (27 Juillet)
1931	Ouverture d'un nouveau centre de production de phosphate : le centre de Youssoufia, connu alors sous le nom de « LOUIS-GENTIL »
1950	- Mise en œuvre de la méthode d'extraction en découverte à Khouribga (1952) - Création de l'Ecole de Maîtrise de Boujniba, en renforcement des efforts menés sur le plan de la formation professionnel (1956) - Démarrage de Maroc Chimie à Safi, pour la fabrication des dérivés phosphatés dont l'acide phosphorique et les engrais (1956) - Création d'un centre de formation professionnel à Khouribga (1958)
1960	Développement de la mécanisation du souterrain à Youssoufia

1970	- Création du groupe OCP, structure organisationnelle intégrant l'OCP et ses entreprises filiales (1975) - Intégration d'un nouveau centre miner en découverte, le centre de Phosboucraâ (1976)
1980	- Partenariat industriel en Belgique : PRAYON (1981) - Démarrage d'un nouveau site de Jorf Lasfar, avec Maroc Phosphore III-IV (1986)
1990	Exploration des nouveaux projets de partenariat industriels et de renforcement de capacités
2000	Démarrage unité de flottation de phosphate à Khouribga
2003	L'OCP est devenu le seul actionnaire de Phousboucraâ
2004	Création de la Société « Pakistan Maroc Phosphore » S.A en Joint-venture entre l'OCP et Fauji Fertilizer Bin Qasim Limited (Pakistan)
2009	Démarrage de l'exploitation de Bunge Maroc Phosphore (BMP)
2011	Coentreprise avec Jacobs Engineering (JESA)
2012	Ouverture de la mine d'Al-hallassa l'une des trois nouvelles mines sur le site de Khouribga d'étalent sur une surface de 1976 hectares d'une capacité de production 6,7 Mt/an
2013	Coentreprise avec DuPont de Nemours
2016	Création d'OCP Africa
2018	Inauguration de l'Université Mohammed VI Polytechnique et coentreprise avec IBM

4. Mission et valeurs :

Le Groupe OCP a pour vision de promouvoir une croissance durable au profit de tous, tandis que sa mission essentielle est de nourrir les sols pour nourrir la planète. Cette vision ambitieuse se déploie à travers l'ensemble de ses activités, qu'il s'agisse de ses opérations industrielles, de ses initiatives éducatives ou de ses travaux de recherche scientifique. Le champ d'action élargi du Groupe OCP lui permet de concrétiser cette vision en agissant de manière intégrée et responsable, dans le but de soutenir un avenir agricole durable et de préserver l'environnement pour les générations futures.

5. Activités et projets du Groupe OCP :

En plus de ses activités d'extraction minière, le Groupe OCP encourage également l'innovation à tous les niveaux que ce soit à travers des initiatives menées par leurs collaborateurs, d'une R&D poussée, d'initiatives de start-ups, de partenariats ou encore dans les domaines de l'éducation et du développement de compétences. Il s'agit d'une philosophie qui cherche à ouvrir la voie à de nouvelles opportunités et à des initiatives permettant de concrétiser sa vision pour un avenir durable.

a. L'engagement de l'OCP envers la durabilité : Des initiatives novatrices pour l'eau, l'énergie et les ressources naturelles :

Des programmes pour l'eau et l'énergie, visant à explorer de nouvelles manières de produire plus avec moins de ressources et à promouvoir une utilisation responsable des ressources naturelles, ont été créés par l'OCP. Dans ce sens, l'OCP dispose du plus grand site de recherche en énergie solaire en Afrique, à Benguerir : Green Energy Park.

b. Un investissement dans l'avenir : L'OCP et son engagement en faveur de l'éducation et de la recherche :

L'OCP a mis en place l'initiative "OCP School Lab" pour promouvoir l'éducation scientifique et technique auprès des jeunes élèves marocains. Ce projet vise à stimuler l'intérêt des élèves pour les sciences, l'agriculture et les technologies liées aux phosphates, contribuant ainsi à la formation de futurs talents pour le secteur. En chiffres, 80000 agriculteurs ont bénéficié de ces School Labs, presque 40000 petits agriculteurs ont bénéficié de l'initiative Agribooster (programme permettant de faciliter l'accès des agriculteurs aux financements et à l'assurance) et plus de 180 projets de R&D sont nés.

Ainsi, le Groupe OCP accompagne les acteurs de l'innovation dans le développement de nouvelles initiatives. C'est pour cette raison que l'Université Mohammed VI Polytechnique (UM6P) a été créée afin d'établir un pont entre la recherche scientifique et les activités industrielles du groupe, offrant ainsi aux chercheurs les ressources dont ils ont besoin pour tester leurs projets et accélérer leurs innovations. Le programme d'accélération IMPULSE, propulsé par l'UM6P, accompagne les start-ups des secteurs de l'Agritech et de la biotechnologie dans leur accès au marché.

c. Des partenariats stratégiques pour l'innovation et la coopération internationale :

L'OCP a toujours forgé des partenariats avec des organisations avec lesquelles il partage la même philosophie et vision. A titre d'exemple, il dispose d'un partenariat avec Plug and Play, basé aux États-Unis, ce qui leur a permis de travailler avec un vaste réseau de start-ups et en étroite collaboration avec la communauté agricole nord-américaine. Il est également associé à IBM pour créer Teal Technology Services afin d'accélérer la transformation digitale au sein du groupe, et à Fertinagro en Espagne qui se concentre pour sa part sur la production de solutions pour l'agriculture biologique. Sans oublier JESA, le plus grand groupe d'ingénierie en Afrique fournissant des services d'ingénierie innovants. Chaque partenariat noué est une opportunité pour l'échange d'expertise et le partage de connaissances. Cette approche permet au groupe de maximiser l'impact de ses projets, des programmes de formation aux joint-ventures.

Grâce à son expertise, à sa présence internationale et à ses efforts pour une croissance durable, le Groupe OCP continue de jouer un rôle clé dans l'industrie des phosphates et contribue à l'avancement de l'agriculture mondiale tout en respectant les enjeux environnementaux et sociaux.

6. Présentation de la mine de Benguerir :

a. Bassin Gantour :



Figure 4: Carte du gisement du bassin de Gantour

Le bassin de Gantour à Benguerir est un grand réservoir artificiel s'étendant sur 120km de long et 30 km de large. Il est situé près de la ville de Benguerir, au Maroc. Ce bassin fait partie intégrante du projet agricole et industriel de l'OCP dans la région. Il a été construit pour répondre aux besoins en eau de l'usine d'acide phosphorique de l'OCP.

Le projet du bassin de Gantour est une composante importante de la stratégie de développement durable de l'OCP, car il permet de sécuriser l'approvisionnement en eau nécessaire à ses activités industrielles tout en minimisant l'impact sur les ressources hydriques locales.

En plus de ses aspects industriels, le bassin de Gantour joue également un rôle dans la gestion de l'eau et la préservation de l'environnement. Il contribue à réguler les ressources hydriques de la région en fonction des besoins saisonniers, ce qui est particulièrement important dans une région semi-aride comme le Maroc.

b. Gisement de Benguerir :

Le gisement de Benguerir fait partie du plateau phosphaté de Gantour dont il occupe la partie centrale. Il est de nature sédimentaire, et consiste en une alternance de couches de

phosphate et d'intercalaires. Il est caractérisé par une abondance de niveaux phosphatés (environ 23 niveaux).

Les expéditions se constituent par des mélanges de plusieurs couches, selon les profils demandés par les clients.

7. Processus de production des phosphates :

a. L'extraction en découverte:

L'extraction du mineraï du phosphate peut se faire soit par voie **souterraine** ou **en découverte**. Actuellement, toutes les mines du groupe OCP, sont exploitées en découverte.

La fermeture de la dernière mine souterraine à Youssoufia date du mois de juin 2005.

L'avantage majeur de l'exploitation en découverte est de pouvoir exploiter simultanément plusieurs couches minéralisées en mettant de côté les niveaux stériles intercalaires.

La méthode d'exploitation **à ciel ouvert** dans les mines de phosphate de l'OCP consiste à enlever la découverte de 5 à 30 m par des Draglines ou des BullDozers après l'avoir forée et sautée. Le phosphate ainsi découvert est chargé sur des chargeuses ou des camions et acheminé vers les installations d'épierrage – criblage pour être chargé sur des trains.

b. Chaîne d'extraction en découverte :

La production des phosphates se fait suivant les étapes suivantes:

- **Aménagement du terrain/Prospection** : Il s'agit d'une étape préparatoire du terrain pour les étapes à venir. Elle consiste en premier lieu à faire l'opération du surfaçage qui a comme objectif d'enlever tous les éléments indésirables ou les obstacles existant sur la terre afin de rendre la surface de terrain plus appropriée pour le déplacement de la machine qui va travailler dans cette zone. Cette étape est assurée par un ensemble d'équipement constitué de : Niveleuse, Bulldozer D9 et D11 et Paydozer.
- **Foration** : Cette opération consiste en le fonçage de trous verticaux dans le sol, de diamètres soigneusement déterminés et selon une maille appropriée tenant compte des caractéristiques de la roche, de la nature de l'explosif et de la fragmentation désirée par le biais des machines de foration appelées « Sondeuses ».



Figure 5: Trous de mines



Figure 6: Sondeuse utilisée pour la foration

- **Sautage du terrain :** Le sautage est l'opération qui consiste à loger une quantité d'explosif « nitrate fuel » dans les trous de foration dans le but de fragmenter le terrain pour faciliter son enlèvement par les machines d'excavation. En effet, on s'attache à obtenir une fragmentation telle qu'on élimine, même dans les zones perturbées où les duretés varient, tout risque de voir le rendement des machines décroître et toute sollicitation anormale de leurs organes de puissance. De ce fait, pour chaque niveau à miner et pour chaque machine, on applique un dosage en explosif permettant d'obtenir la fragmentation recherchée.
- **Décapage du terrain :** Cette étape consiste à éliminer les terrains morts (stériles) fragmentés auparavant, par des Draglines et des Bulldozers, pour mettre en évidence la couche désirée du phosphate. Cette opération se réalise en deux étapes principales: la première pour enlever une grande partie du stérile fragmenté par les Draglines (Dragline PH, D11, Marion 7500M), et la deuxième, nommée le nettoyage, pour éliminer la petite partie qui reste et qui se trouve près de la couche phosphatée et ceci à l'aide des Bulldozers. Pour le nettoyage, on n'utilise pas des Draglines sinon on risque de toucher à la couche phosphatée donc dégrader la qualité de notre produit.



Figure 7: Dragline 7500M



Figure 8: Bulldozer

- **Défruitage :** Cette étape consiste à gerber le phosphate une fois l'enlèvement des intercalaires terminé. Les tas formés sont ensuite chargés puis transportés sur des camions jusqu'aux installations d'épierrage. Le défruitage est une opération délicate car non seulement il faut récupérer le phosphate de façon convenable, mais il faut aussi sauvegarder la teneur in situ du minéral.



Figure 9 : Opération de défruitage par chargeuse

- **Transport des phosphates :** Le transport du phosphate ou du stérile est assuré par des chargeuses ou des camions (camions OCP ou camions entreprises extérieures) de grande capacité de la benne vers, soit la trémie ou les décharges.

Juste après son extraction, le phosphate brut est acheminé vers des installations fixes afin d'y subir un traitement mécanique. Ces installations sont articulées sur trois axes :

- **Station d'épierrage KRUPP :** Les camions de chantier sont déchargés dans une trémie de réception alimentant un épierrage de maille 90 x 90mm destiné à l'élimination des gros blocs stériles. Ces derniers sont évacués vers une mise à terril par un convoyeur après avoir subi un recouvrement de récupération et une fragmentation dans un

concasseur à mâchoires. Le concasseur à mâchoires est conçu afin de concasser le refus des cribles et réduire les blocs à une dimension de 0 à 300mm. Le phosphate épierré est d'abord stocké dans un parc de stockage doté de deux stockeuses à translation sur rails appelées *Stackers*, avant d'être acheminé vers la station de criblage par convoyeur. Le phosphate stocké par la stacker est repris par une machine appelée *Roue-pelle*.

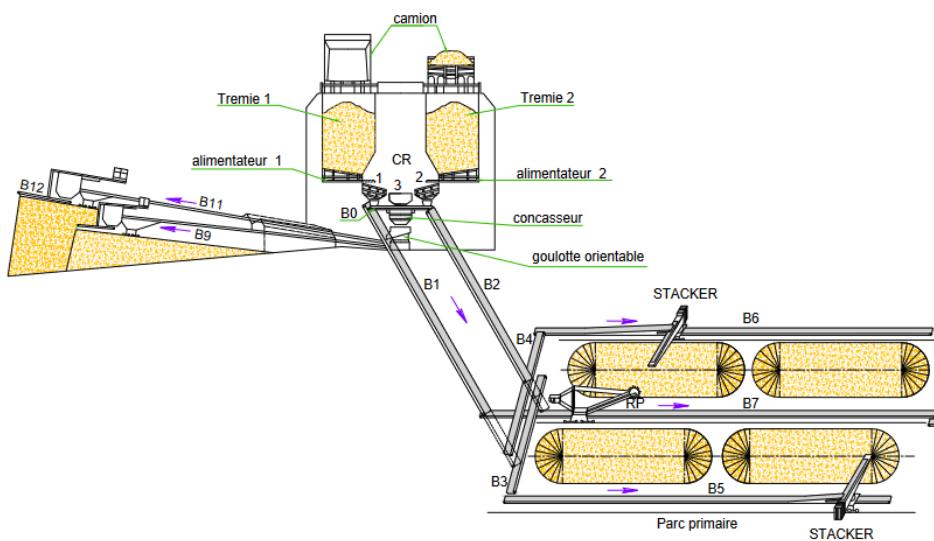


Figure 10 : Station d'épierrage KRUPP



Figure 11 : Stackér

- **Station de criblage KOCH** : consiste à faire passer le phosphate sur des cribles de mailles de 10mm. Le refus des cribles sera envoyé par des convoyeurs à la mise à terril. Le produit criblé est acheminé par des convoyeurs vers un deuxième parc de stockage pour être homogénéisé afin d'obtenir une qualité bien définie. Le phosphate

issu de cette station est donc stocké dans des zones bien spécifiques selon la qualité du produit.



Figure 12 : Opération de criblage

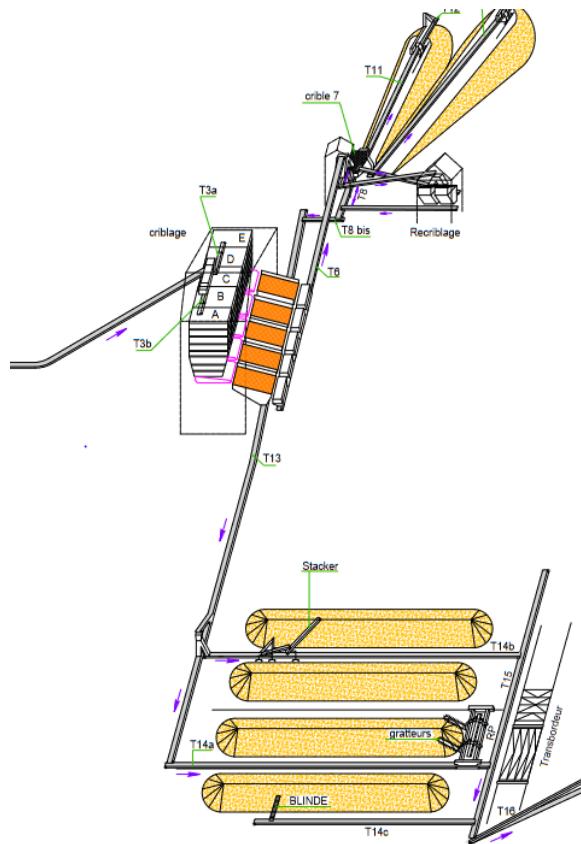


Figure 13: Station de criblage KOCH

- **Chargement :** Le phosphate stocké sera ensuite chargé dans les wagons d'une station à deux voies, et ceci à l'aide d'une roue-pelle qui permet la reprise du phosphate et

l'alimentation des convoyeurs T14a et T14b pour acheminer le phosphate vers les trémies de chargement du train, par l'intermédiaire des convoyeurs T15 et T17.



Figure 14 : Opération de chargement

Chapitre II : Présentation du projet :

1. Cahier de charges :

a. Besoin et enjeux :

Actuellement, la production minière et industrielle exige une optimisation constante pour répondre à une demande croissante tout en assurant une utilisation judicieuse des ressources. Les stations KOCH et KRUPP jouent un rôle essentiel dans le processus de transformation des phosphates à l'OCP. Cependant, la gestion de la production de ces stations peut être complexe en raison de divers facteurs tels que les fluctuations de la qualité des matières premières, les conditions opérationnelles changeantes et les défis environnementaux. Le besoin d'un modèle de prédition précis découle donc de plusieurs défis opérationnels et stratégiques auxquels l'OCP est confronté.

En anticipant ces besoins de production en prenant des décisions stratégiques, l'OCP bénéficiera d'une:

- **Efficacité opérationnelle** : en prédisant avec précision la production des deux stations, l'OCP peut optimiser l'allocation des ressources (main-d'œuvre, matières premières, équipements...) entraînant ainsi des économies au niveau des coûts liés aux stocks excessifs ou aux pénuries de produits finis.
- **Planification de la maintenance** : la maintenance prédictive aide à détecter les premiers signes de défaillance des machines en analysant leurs données historiques.
- **Gestion de la chaîne d'approvisionnement** : la prédition des niveaux de production permettra à l'OCP d'aligner son processus de production et de livraison en fonction de la demande prévue. Cette coordination entre les différentes étapes d'approvisionnement garantie que les matières premières et les composants sont disponibles en temps voulu.

b. Objectifs :

Notre mission fondamentale est d'établir une nouvelle norme pour l'excellence opérationnelle en développant un modèle de prédition pointu pour les stations d'épierrage et de criblage KOCH et KRUPP, exploitant pleinement les données fournies dans le rapport RO.

Notre démarche cherche à dévoiler un horizon d'optimisation, en anticipant avec précision les niveaux de production futurs de ces stations clés.

Nous nous engageons à concevoir un modèle qui peaufine les données, extrait les tendances cachées et, grâce à l'apprentissage automatique, sculpte des résultats qui transcendent les attentes. Le modèle ainsi créé sera une pierre angulaire pour la prise de décision éclairée, permettant d'optimiser la production, de maximiser l'efficacité opérationnelle et de créer une synergie harmonieuse entre les stations KOCH et KRUPP.

2. Conduite du projet :

Afin de gérer à bien notre projet, nous avons suivi la méthodologie CRISP-DM (Cross-Industry Standard Process for Data Mining), l'une des approches les plus structurées pour résoudre un problème qui nécessite la science des données. Elle décompose le processus d'exploration des données en six grandes phases :

❖ Compréhension métier (Business understanding)

- Objectifs et buts métier: Prédire la production de phosphates à partir des données fournies par les deux stations et élaborer un système de gestion fonctionnelle de production.
- Cadre théorique et conceptuel : comprendre les processus de production de Phosphates et les facteurs qui influencent sa production.

❖ Compréhension des données (Data understanding)

- Collecter et acquérir les données à partir des sources convenables.
- Conduire une analyse exploratoire des données (EDA)
- Comprendre la structure, la qualité et le contenu des données et identifier les problèmes de qualité s'ils existent.
- Visualiser les données.

❖ Préparation des données (Data preparation)

- Nettoyer, prétraiter et transformer les données pour l'analyse.
- Sélectionner les variables les plus pertinentes pour la prédiction de la production des phosphates.
- Gérer les valeurs manquantes ou nulles, les valeurs aberrantes et les incohérences.
- Séparer le jeu de données en données d'apprentissage et de test.

❖ Modélisation (Modeling)

- Choisir les algorithmes de modélisation appropriés pour la prédiction de la production de phosphates.

- Construire et former le modèle prédictif.
- Entraîner le modèle sur les données d'apprentissage et l'ajuster en fonction des performances.

❖ Evaluation (Evaluation)

- Évaluer la qualité et la performance du modèle sur les données de test.
- Comparer et sélectionner le(s) modèle(s) ayant les meilleures performances.

❖ Déploiement (Deployment)

- Déployer le modèle sélectionné dans l'environnement opérationnel.
- Intégrer le modèle dans les processus métier et les systèmes pour effectuer des prédictions ou des décisions.

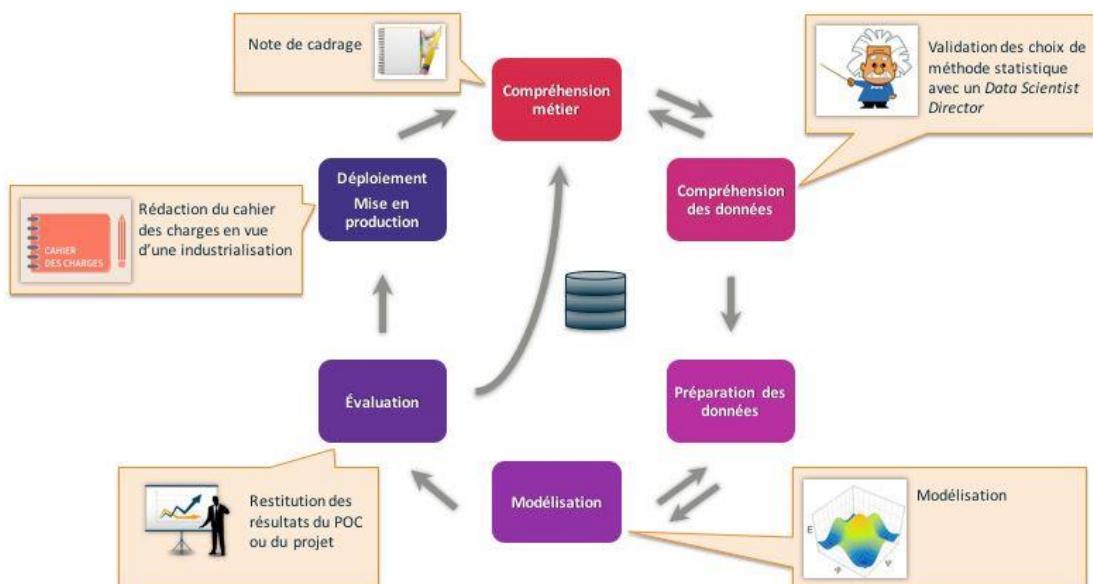


Figure 15 : Phases de la méthodologie CRISP-DM

Chapitre III : Pipeline du projet :

Après avoir désigné et organisé le domaine d'application, les connaissances préalables pertinentes obtenues auprès des experts et les objectifs finaux, je vais décortiquer les différentes phases qui forment un projet des sciences de données selon la méthodologie CRISP-DM, en commençant par la compréhension du métier jusqu'à l'évaluation du modèle.

I. Spécification et compréhension du problème :

1. Compréhension métier :

Notre but est de construire un modèle de prédiction de la production des deux installations fixes KOCH et KRUPP en se basant sur les données fournies dans le rapport R0. Il s'agit donc d'un problème d'apprentissage supervisé.

L'apprentissage supervisé est une sous-catégorie de l'apprentissage automatique (ML) et de l'intelligence artificielle (IA). Il se caractérise par l'utilisation de jeux de données étiquetés qui entraînent des algorithmes permettant de classer des données ou de prédire des résultats avec précision.

2. Technologies utilisées :

La pratique de la science des données nécessite l'utilisation d'outils analytiques, de technologies et de langages de programmation pour extraire des informations ou valeur des données, ainsi que représenter les résultats obtenus d'une manière efficace.

Dans le cadre de ce projet, nous avons opté pour l'utilisation des outils suivants :

a. Python :

Python est parmi les langages de programmation les plus utilisés pour le traitement des données (dans 83% des cas). Il dispose de beaucoup de librairies distribuées sous une licence libre. En fait, c'est un outil performant qui présente toute une série d'avantages. Comme il est open-source, il est flexible et s'améliore continuellement.

b. Jupyter Notebook:

Jupyter Notebook est un environnement de développement interactif qui permet d'écrire et d'exécuter du code Python (et d'autres langages de programmation) de manière interactive.

Il est largement utilisé dans le domaine de la science des données, de l'apprentissage automatique et de la recherche pour créer des documents mixtes contenant du code, des textes explicatifs et des visualisations.

II. Compréhension approfondie des données

1. Séries temporelles ou Time series :

Une série temporelle (ou chronologique) est une suite finie (x^1, \dots, x^n) de données enregistrées et organisées dans un ordre chronologique qui peut être selon la minute, l'heure, le jour, le mois, l'année... Le nombre n est appelé la longueur de la série. Ces données sont collectées à des intervalles réguliers ou irréguliers.

Contrairement aux données traditionnelles où chaque observation est indépendante, les séries temporelles tiennent compte du temps qui est une variable explicative incontournable en analyse de séries temporelles. C'est la variable qui évolue au fil des observations.

L'émergence de cycles est une particularité des séries temporelles. Ceux-ci peuvent être analysés en vue d'en déterminer les tendances globales et extraire toute information significative.

Les séries temporelles peuvent également être modélisées, c'est-à-dire qu'on peut utiliser des modèles mathématiques pour en capturer les motifs et variations temporelles, comme on verra prochainement. Un fois le modèle construit, il peut être utilisé pour effectuer des prévisions.

2. Aperçu des jeux de données :

Les différents jeux de données, extraits des rapports R0, sont des séries temporelles multivariées sous format Excel. Pour KRUPP, nous avons pu récupérer des données s'étalant du 20/12/2021 jusqu'au 10/07/2023 tandis que pour KOCH, nous n'avons pu avoir que presque 6 mois de données, du 01/01/2023 au 07/07/2023.

Ci-dessous les lignes de code et résultats obtenus pour les deux stations:

```
df = pd.read_excel('R0 KRUPP 2023.xlsx', sheet_name='Synthèse', header=2, skiprows=0)

# Drop rows with any empty cells (NaN values)
df = df.dropna()

df
```

	Date	poste 3	poste 1	poste 2	Journée	Poste 3.1	poste 1.1	poste 2.1	Journée.1	poste 3.2	poste 1.2	poste 2.2	Journée.2	poste 3.3	poste 1.3	poste 2.3	Journée.3
0	2021-12-20	2480.0	3840.0	5136.0	11456	2331.20	3609.60	4827.84	10768.64	2.74	3.50	5.48	11.72	850.802920	1031.314286	880.992701	977.474403
1	2021-12-21	2896.0	4856.0	5372.0	13124	2722.24	4564.64	5049.68	12336.56	2.57	4.80	6.05	13.42	1059.237354	950.966667	834.657851	977.943368
2	2021-12-22	4150.0	2851.0	2166.0	9167	3901.00	2679.94	2036.04	8616.98	3.9	2.60	2.19	8.69	1000.256410	1030.746154	929.698630	1054.890679
3	2021-12-23	5159.0	4502.0	5072.0	14733	4849.46	4231.88	4767.68	13849.02	6.17	4.94	6.60	17.71	785.974068	856.655870	722.375758	831.902880
4	2021-12-24	5271.0	4465.0	5530.0	15266	4954.74	4197.10	5198.2	14350.04	6.7	5.31	7.48	19.49	739.513433	790.414313	694.946524	783.273474
...

Figure 18: les 4 premières lignes du dataframe des données de KRUPP

```
df = pd.read_excel('R0 KOCH 2023.xlsm', sheet_name='synthèse 312', header=2, skiprows=0)

# Drop rows with any empty cells (NaN values)
df = df.dropna()

df
```

	Date	poste 3	poste 1	poste 2	Unnamed: 4	poste 3.1	poste 1.1	poste 2.1	Unnamed: 8	poste 3.2	...	poste 2.2	Unnamed: 12	poste 3.3	poste1	poste 2.3	Unnamed: 16	poste 3.4
0	2023-01-01	0	0	0	0	5424	4554	5301.0	15279.0	6.916667e+00	...	7.700000	2.055000e+01	7.26	6.05	7.77	21.08	784.192771
1	2023-01-02	0	0	0	0	5006	3779	4787.0	13572.0	6.916667e+00	...	6.616667	1.925000e+01	6.86	6.45	6.52	19.83	723.759036
2	2023-01-03	0	0	0	0	5200	3755	4481.0	13436.0	7.466667e+00	...	6.350000	1.968333e+01	10.68	5.38	7.42	23.48	696.428571
3	2023-01-04	0	0	0	0	4789	83	0.0	4872.0	7.183333e+00	...	0.000000	7.383333e+00	7.58	0.22	0.00	7.80	666.682135
4	2023-01-05	0	0	0	0	0	0	0.0	0.0	1.164153e-10	...	0.000000	1.164153e-10	0.00	0.00	0.00	0.00	0.000000
...	

Figure 19: les 4 premières lignes du dataframe des données de KOCH

Cela étant, nous avons également supprimé certaines colonnes pour extraire nos données cibles :

	Date	Journée	Journée.1	Journée.2	Journée.3		Date	Unnamed: 8	Unnamed: 20	
0	2021-12-20	11456	10768.64	11.72	977.474403		0	2023-01-01	15279.0	746.720806
1	2021-12-21	13124	12336.56	13.42	977.943368		1	2023-01-02	13572.0	702.761556
2	2021-12-22	9167	8616.98	8.69	1054.890679		2	2023-01-03	13436.0	680.718227
3	2021-12-23	14733	13849.02	17.71	831.902880		3	2023-01-04	4872.0	360.560712
4	2021-12-24	15266	14350.04	19.49	783.273474		4	2023-01-05	0.0	0.000000

Figure 20 : Extrait du dataframe de KRUPP

Figure 21 : Extrait du dataframe de KOCH

Pour plus d'organisation, nous avons renommé les colonnes comme suit :

	THE	THC	HM	Rendement		THC	Rendement
	Date					Date	
2021-12-20	11456	10768.64	11.72	977.474403		2023-01-01	15279.0
2021-12-21	13124	12336.56	13.42	977.943368		2023-01-02	13572.0
2021-12-22	9167	8616.98	8.69	1054.890679		2023-01-03	13436.0
2021-12-23	14733	13849.02	17.71	831.902880		2023-01-04	4872.0
2021-12-24	15266	14350.04	19.49	783.273474		2023-01-05	0.0

Figure 22: Extrait du dataframe de KRUPP

Figure 23 : Extrait du dataframe de KOCH

Ainsi, nous avons effectué une **description statistique** en utilisant la fonction `.describe()`.

Cette méthode d'analyse va nous permettre de résumer et de présenter de manière concise les principales caractéristiques et tendances de nos données :

Rendement		THC Rendement	
count	309.000000	count	129.000000
mean	847.731939	mean	526.929540
std	139.366522	std	218.945788
min	467.091295	min	0.000000
25%	738.021534	25%	434.785479
50%	844.723295	50%	616.431966
75%	939.456662	75%	679.754648
max	1250.638792	max	895.216326

Figure 24 : Description statistique des données de KRUPP

Figure 25 : Description statistique des données de KOCH

Nous avons également utilisé la fonction `.info()` pour savoir les types des variables de nos données:

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 309 entries, 2021-12-20 to 2023-03-20
Data columns (total 4 columns):
 #   Column   Non-Null Count Dtype  
--- 
 0   THE      309 non-null   object 
 1   THC      309 non-null   object 
 2   HM       309 non-null   object 
 3   Rendement 309 non-null   float64
dtypes: float64(1), object(3)
memory usage: 12.1+ KB

<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 129 entries, 2023-01-01 to 2023-07-07
Data columns (total 2 columns):
 #   Column   Non-Null Count Dtype  
--- 
 0   THC      129 non-null   float64
 1   Rendement 129 non-null   float64
dtypes: float64(2)
memory usage: 3.0 KB
```

Figure 26: Types des variables des données de KRUPP

Figure 27 : Types des variables des données de KOCH

4. Exploration et visualisation des données :

La première chose à faire dans toute tâche d'analyse de données est de tracer les données.

Les graphiques permettent de visualiser de nombreuses caractéristiques des données, y compris les modèles, les observations inhabituelles, les changements dans le temps et les relations entre les variables. Les caractéristiques qui apparaissent dans ces graphiques doivent ensuite être intégrées, autant que possible, dans les méthodes de prévision à utiliser. Tout comme le type de données détermine la méthode de prévision à utiliser, il détermine également les graphiques appropriés.

a. Line Plot :

Pour les Time series, le graphique le plus évident pour commencer est le Line Plot.

Un Line Plot est un outil très pratique pour visualiser l'évolution temporelle d'une variable. Il permet d'observer les tendances à la hausse ou à la baisse et de savoir avec précision à quel moment un évènement sortant de l'ordinaire s'est produit.

Dans le contexte de notre sujet, nous avons utilisé des Line Plots pour représenter visuellement les variations de différentes variables :

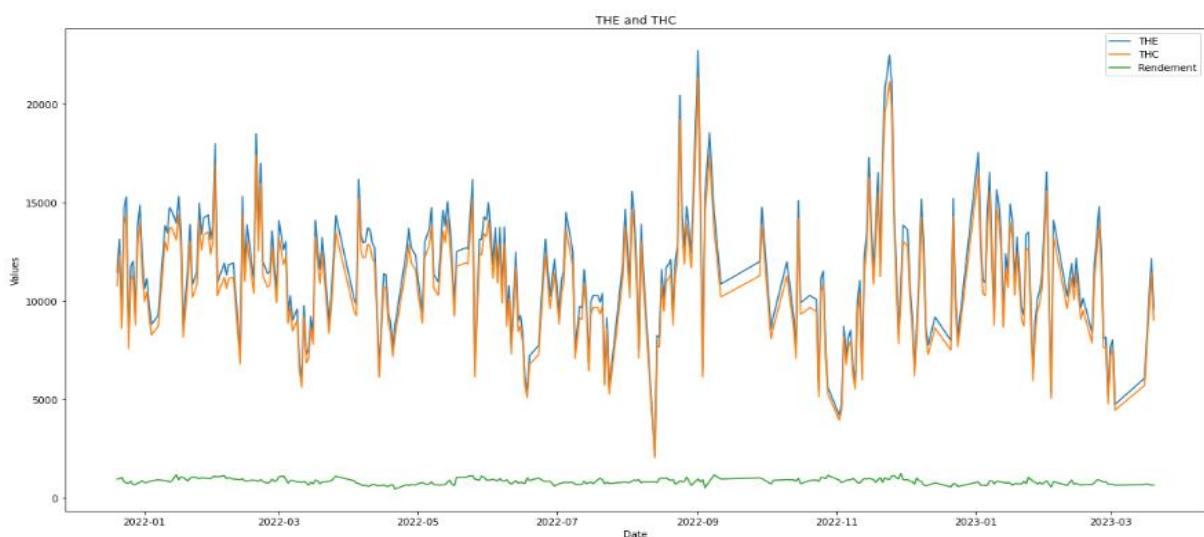


Figure 28 : Line Plot du THE, THC et le rendement de la station KRUPP

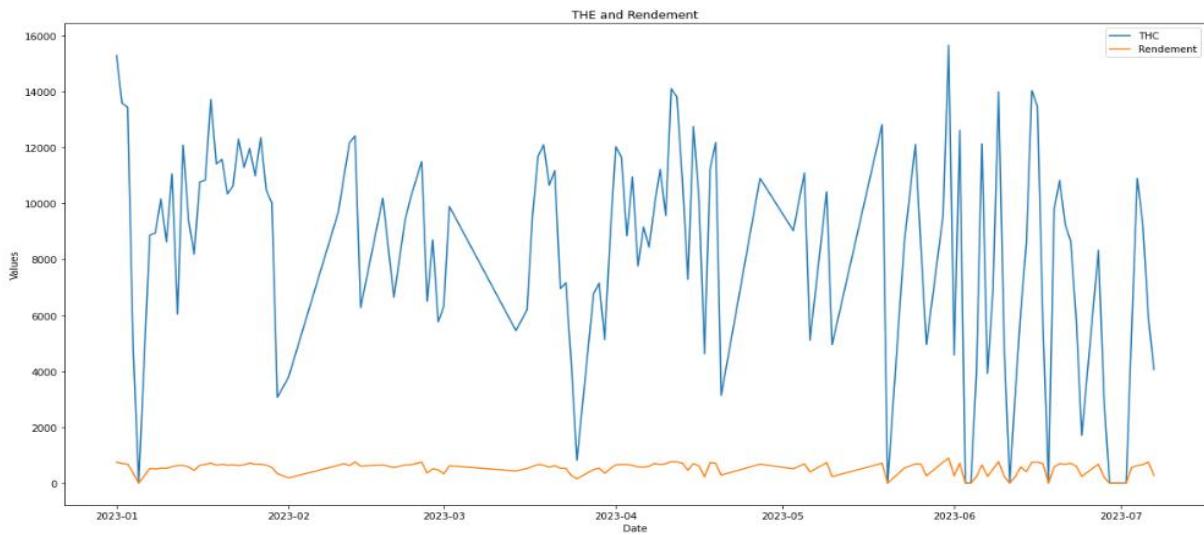


Figure 29 : Line Plot du THC et le rendement de la station KOCH

Pour KRUPP, nous avons essayé de visualiser le THE, le THC et le rendement en fonction du temps, alors que pour KOCH, nous n'avons visualisé que le THC et le rendement criblage puisque toutes les valeurs du THE sont nulles.

b. Décomposition de séries temporelles :

La décomposition de séries temporelles est une technique analytique utilisée pour comprendre les différentes composantes qui contribuent aux variations observées dans une série temporelle donnée.

- Composante de tendance (Trend component) : c'est la variation lente et régulière qui montre la direction à long terme des données. Le Trend peut être croissant, décroissant ou stable.
- Composante saisonnière (Seasonal component) : il s'agit des motifs cycliques et répétitifs à des intervalles fixes.
- Composante résiduelle (Residual component) : c'est la partie de la série temporelle non expliquée par les composantes précédentes. Elle contient généralement le bruit aléatoire ou les erreurs de mesure.

En visualisant ces composantes, nous avons obtenu des informations précieuses sur la structure temporelle des données, ce qui nous a permis de prendre des décisions plus éclairées pour des analyses et des prévisions ultérieures.

Ci-dessous est la décomposition de séries temporelles des données des deux stations KRUPP et KOCH :

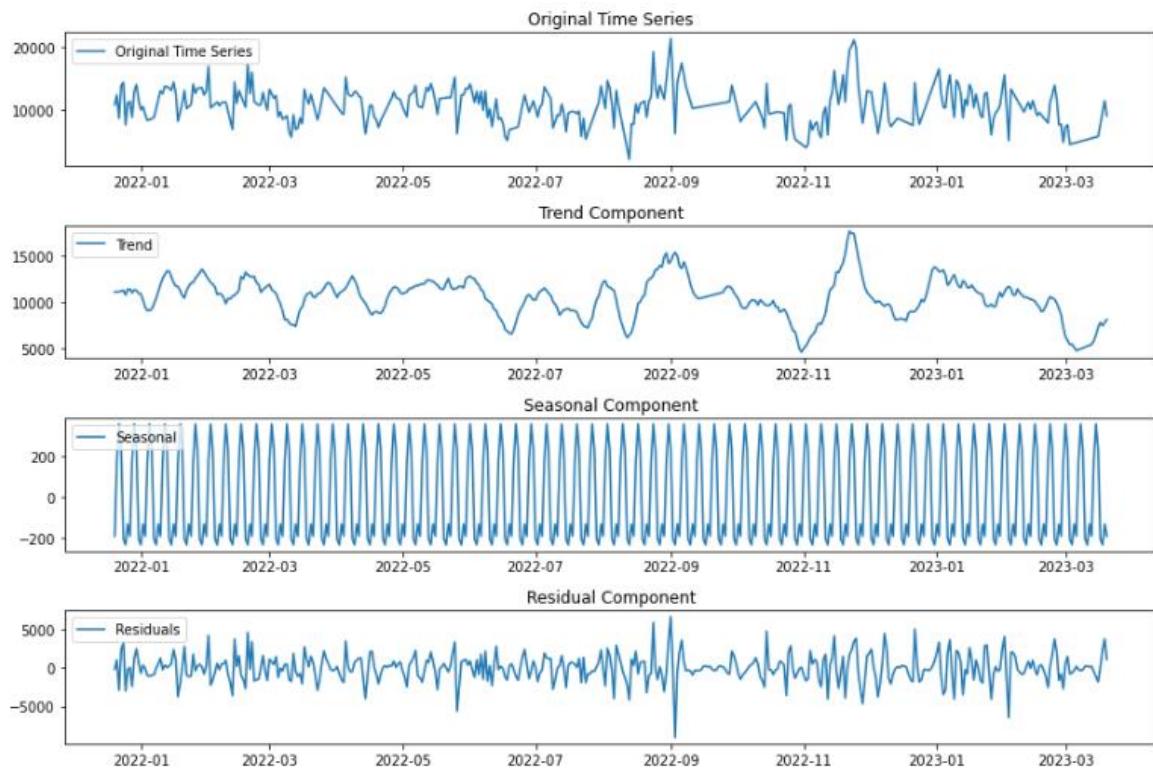


Figure 30 : Décomposition de séries temporelles des données de KRUPP

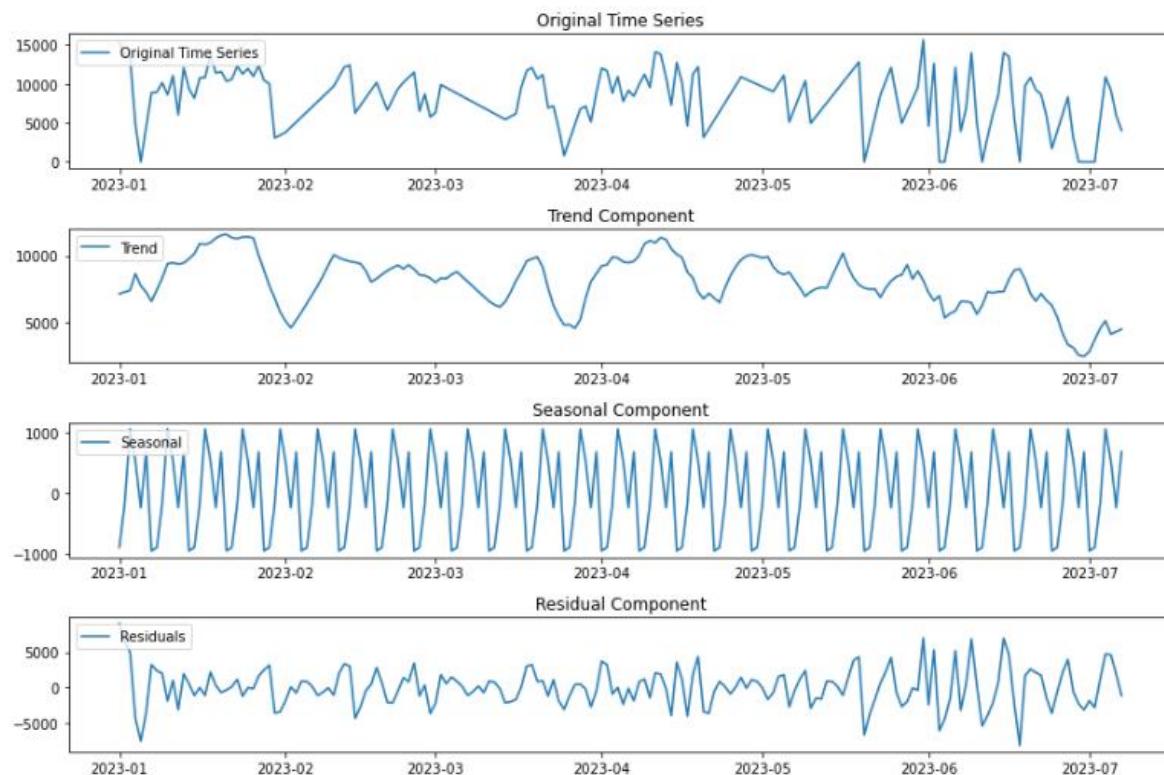
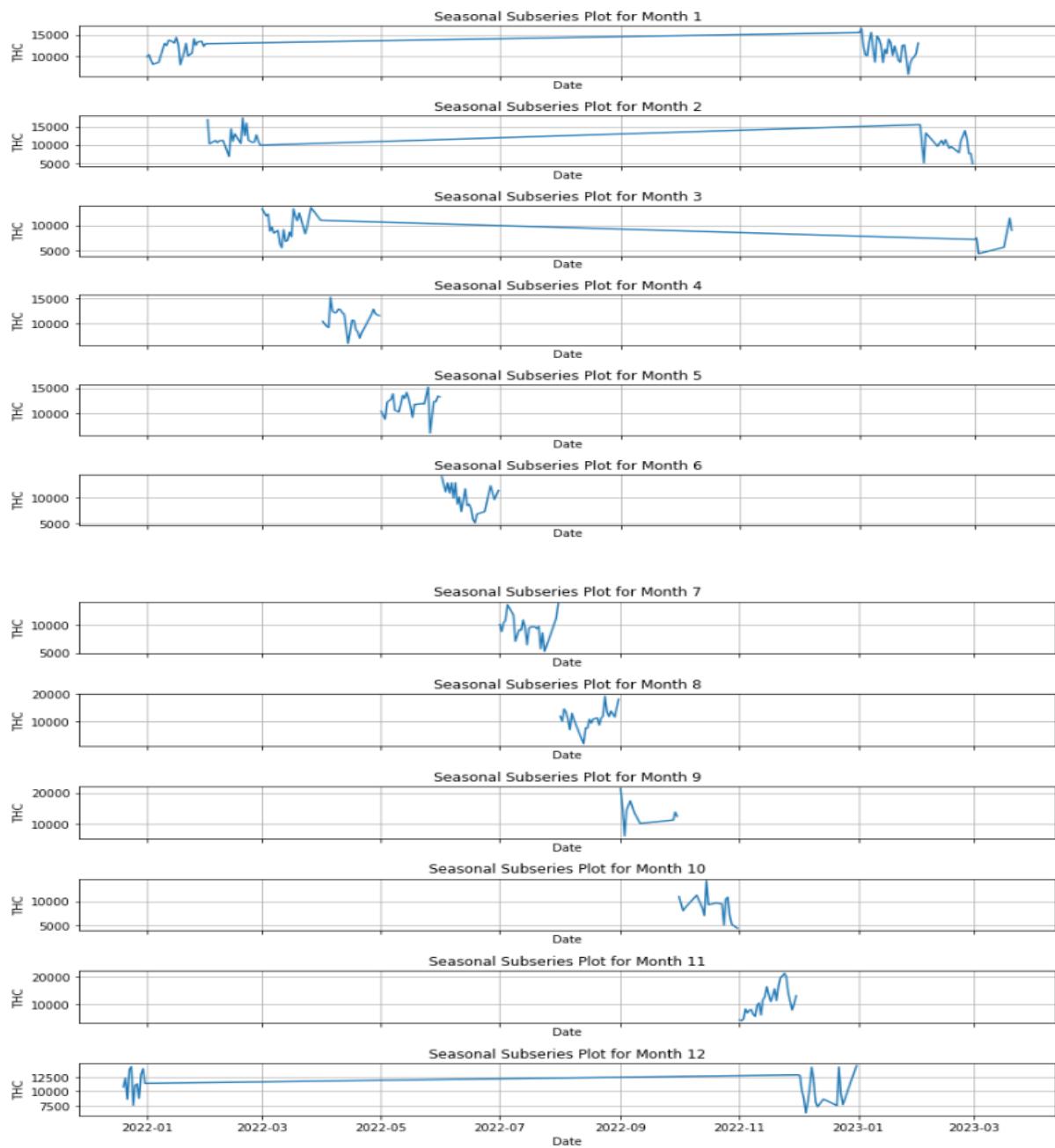


Figure 31 : Décomposition de séries temporelles des données de KOCH

c. Graphiques de sous-séries saisonnières ou
Seasonal subseries Plots:

En analyse de séries temporelles, les graphiques de sous-séries saisonnières sont utilisés pour visualiser individuellement les données de chaque saison, en divisant la série temporelle en segments correspondant à différentes saisons (dans notre cas, chaque mois) .

Cela permet d'examiner les motifs saisonniers et les fluctuations récurrentes associés à des périodes spécifiques de l'année.



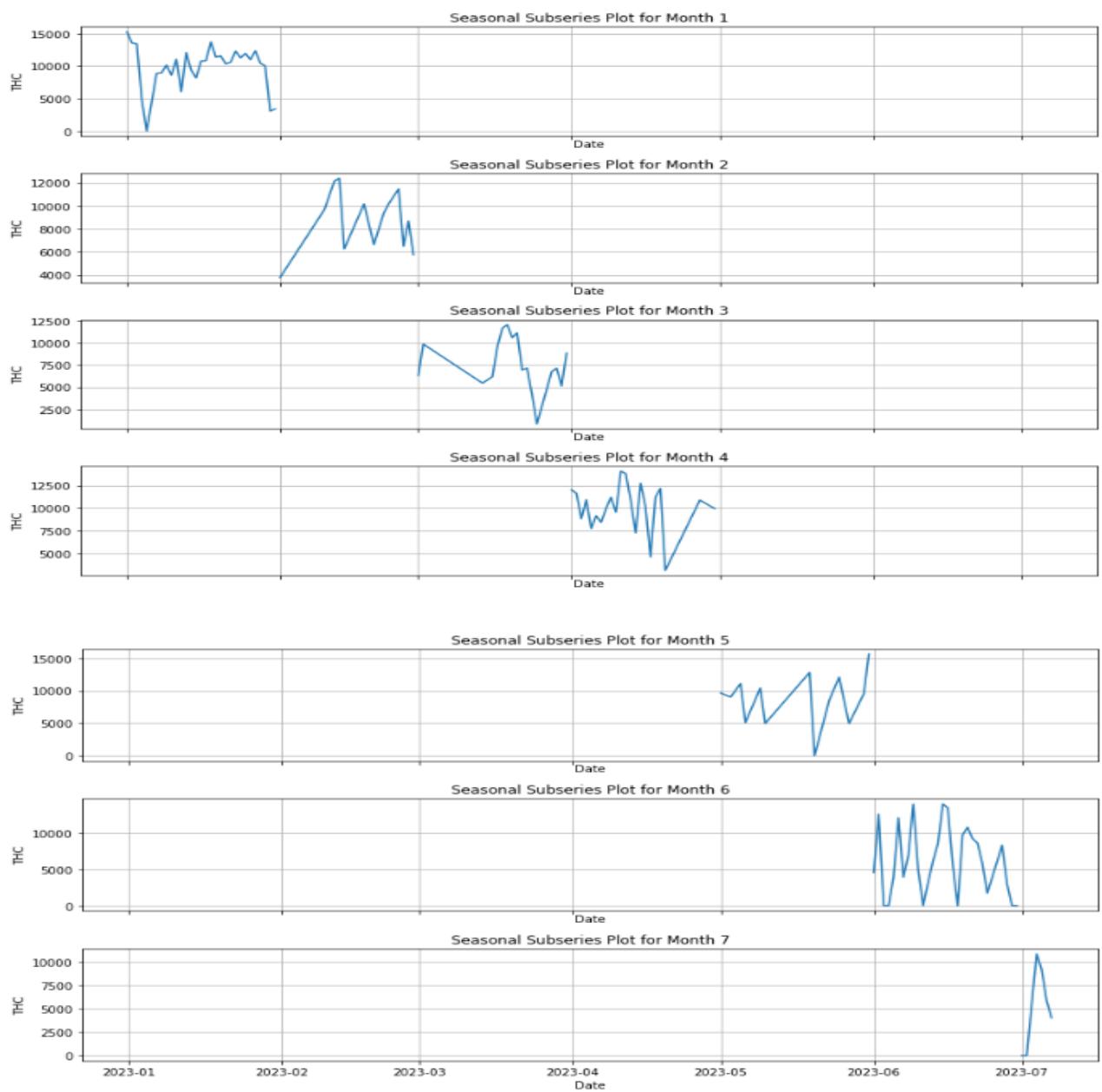
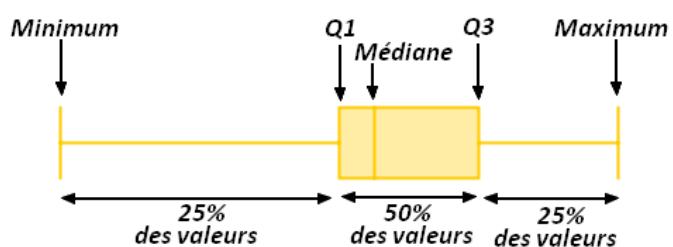


Figure 33 : Seasonal Subseries Plots des valeurs de THC de KOCH

d. Box Plot :

Un box-plot est un graphique simple composé d'un rectangle duquel deux droites sortent afin de représenter certains éléments des données.



La valeur centrale du graphique est la médiane (il existe autant de valeurs supérieures qu'inférieures à cette valeur dans l'échantillon).

Les bords du rectangle sont les quartiles (Pour le bord inférieur, un quart des observations ont des valeurs plus petites et trois quart ont des valeurs plus grandes, le bord supérieur suit le même raisonnement).

Les extrémités des moustaches sont calculées en utilisant 1.5 fois l'espace interquartile (la distance entre le 1er et le 3ème quartile).

Ce qui suit sont les Box plots des valeurs de THC des deux stations, d'épierrage et de criblage :

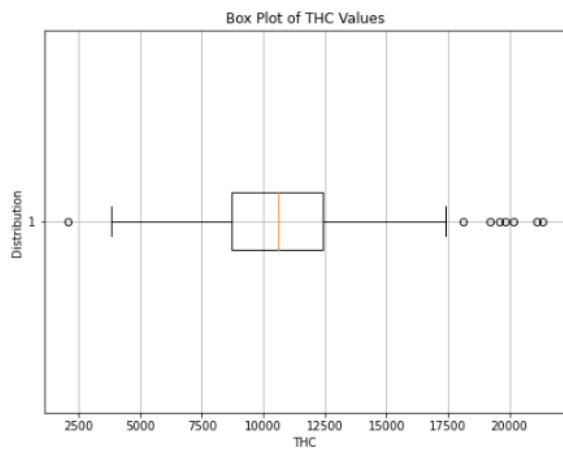


Figure 34 : Box Plot des données de THC de KOCH

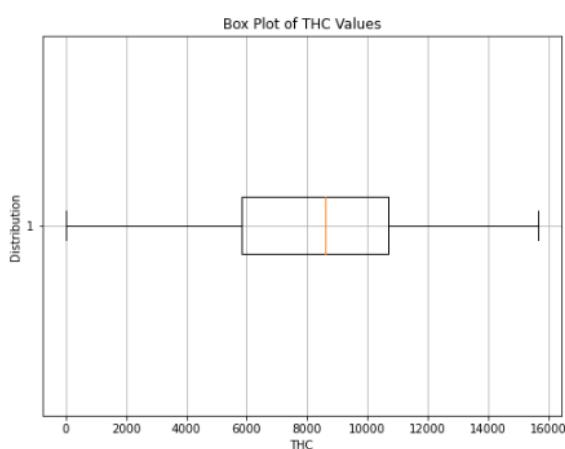


Figure 35 : Box Plot des données de THC de KRUPP

e. Carte thermique ou Heatmap :

Une heatmap est une représentation visuelle de données structurée sous la forme d'une matrice de colonnes et de lignes. Elle est très utile pour décrire la corrélation entre plusieurs variables et visualiser les patterns et les anomalies.

La carte thermique transforme la matrice de corrélation en code couleur, c'est-à-dire que les valeurs individuelles de cette matrice sont représentées par des couleurs. Chaque cellule du tableau est colorée en fonction de la valeur de la variable à l'intersection de la ligne et de la colonne correspondantes.

Dans une carte thermique, les couleurs utilisées varient du froid au chaud, d'où le terme « thermique ». Les couleurs chaudes (rouge, jaune, orange) représentent les valeurs les plus élevées tandis que les couleurs froide (vert ou bleu) représentent les valeurs les plus basses.

Les cartes thermiques pour les données de THC sont générées comme indiqué ci-dessous :

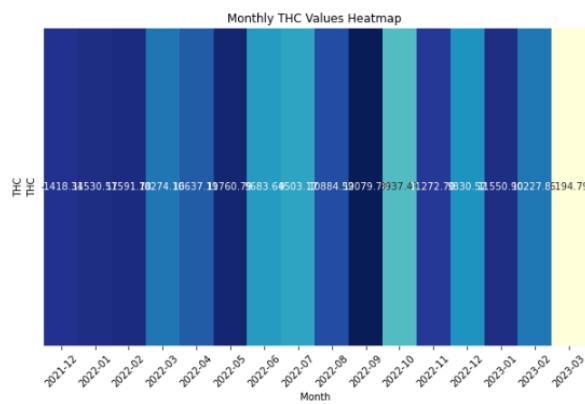


Figure 36 : Heatmap des données de THC de KRUPP

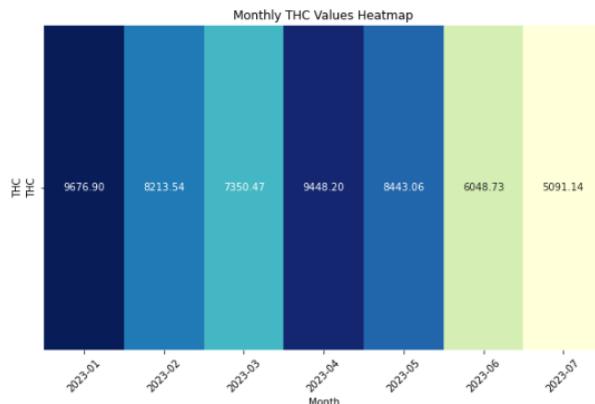


Figure 37 : Heatmap des données de THC de KOCH

f. Statistiques glissantes ou Rolling statistics :

Les statistiques glissantes font référence à des analyses statistiques effectuées sur une fenêtre mobile de données dans une série temporelle. Plutôt que de calculer des statistiques sur l'ensemble des données, les statistiques glissantes calculent ces mesures pour des sous-ensembles consécutifs et se chevauchant de la série temporelle. Cela permet de voir comment les caractéristiques statistiques évoluent au fur et à mesure que la fenêtre se déplace le long de la série. La fenêtre, dans le contexte des statistiques glissantes, a une unité, qui est généralement déterminée par le pas de temps de la série temporelle. Dans notre cas, nous avons défini une fenêtre de taille 30 jours pour les deux jeux de données.

La moyenne glissante (Rolling mean) et l'écart type glissant (Rolling standard deviation) sont inclus dans les statistiques glissantes. La moyenne glissante, également connue sous le nom de la moyenne mobile, est la moyenne de l'ensemble de valeurs dans la fenêtre se déplaçant le long de la série des données.

L'écart type glissant mesure la variation des valeurs au sein de la même fenêtre mobile.

Ces statistiques sont couramment utilisées pour lisser les données brutes, identifier les tendances, atténuer les fluctuations et éliminer le bruit dans les séries temporelles, ce qui peut rendre les données plus interprétables et faciliter l'analyse.

Les graphiques ci-dessous illustrent les statistiques glissantes calculées à partir de nos données de THC pour les deux stations:

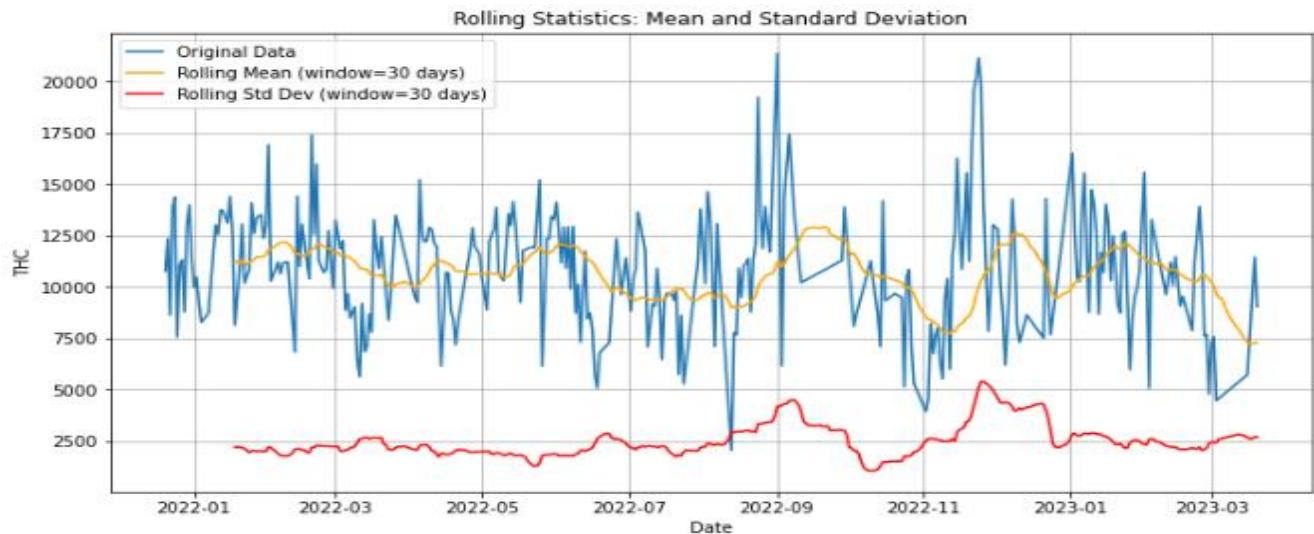


Figure 38 : Statistiques roulantes des données de KRUPP

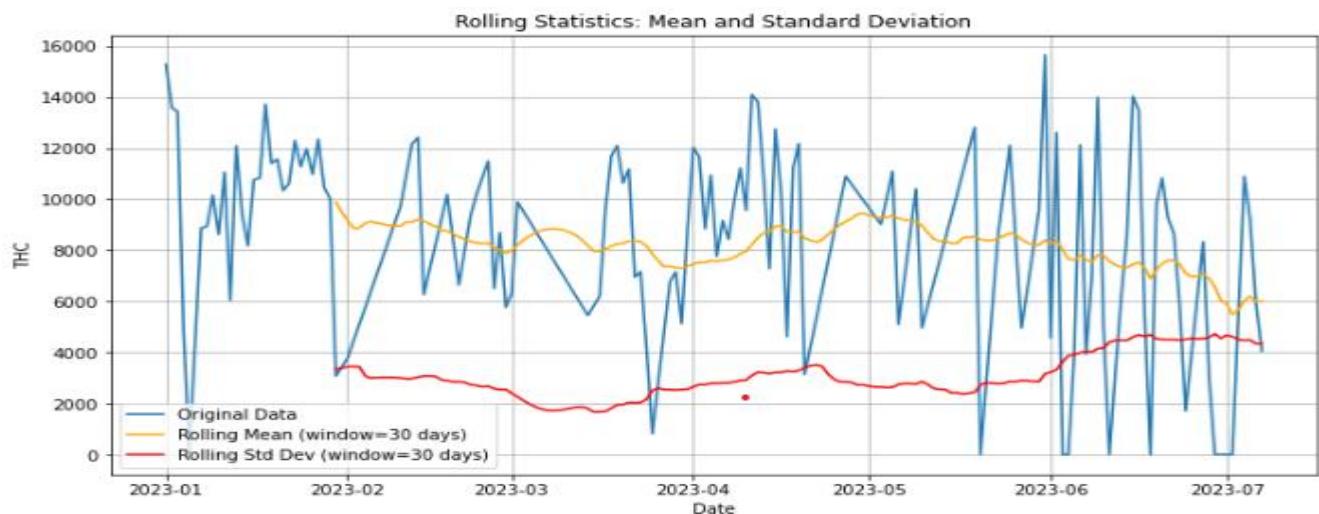


Figure 39 : Statistiques roulantes des données de KOCH

g. Diagramme de dispersion retardée ou le Lag Scatter Plot:

Un Lag Scatter plot est un type de graphique utilisé pour explorer la relation entre une valeur et ses valeurs précédentes à des retards spécifiques. En d'autres termes, il permet de visualiser la corrélation entre une observation et ses observations antérieures décalées dans le temps.

Chaque point dans le diagramme de dispersion retardé représente une paire de valeurs : la valeur actuelle et la valeur précédente à un délai donné. Cela peut aider à identifier des

tendances de corrélation à différents retards, ce qui peut être important dans l'analyse de séries temporelles pour déterminer les retards significatifs ou les relations de dépendance temporelle.

Les diagrammes de dispersion des retards pour les données de THC deux stations KRUPP et KOCH sont créés sous forme de graphiques, comme illustré ci-dessous :

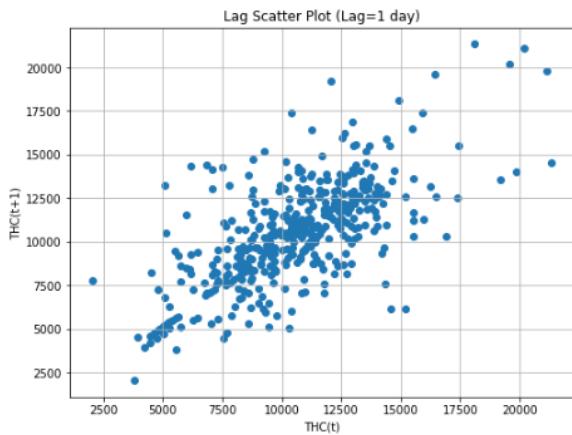


Figure 40 : Lag Scatter plot des données de KRUPP

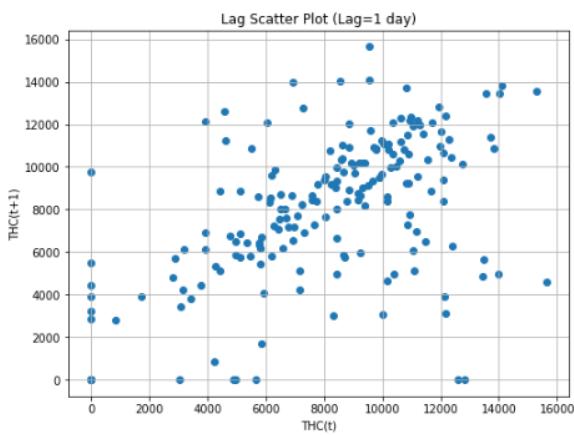


Figure 41 : Lag Scatter plot des données de KOCH

h. Histogramme :

L'histogramme est un outil de visualisation fréquemment utilisé dans l'analyse exploratoire de données. Il permet de comprendre la distribution des valeurs d'une variable quantitative. L'histogramme divise la plage des valeurs en plusieurs intervalles, appelés aussi "bins" ou "classes", et compte combien de valeurs se trouvent dans chaque intervalle. Ensuite, il trace ces comptages sous forme de barres pour représenter visuellement la fréquence de chaque intervalle.

Les histogrammes des valeurs de THC des deux stations KRUPP et KOCH sont représentés comme suit :

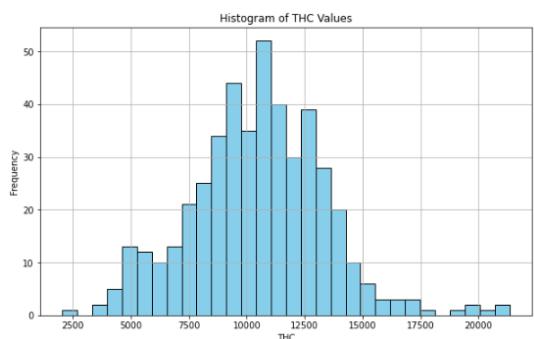


Figure 42 : Histogramme du THC de KOCH

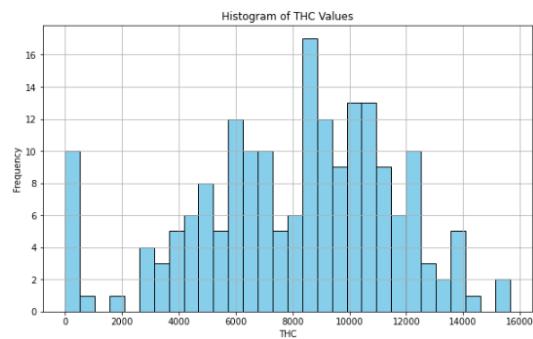


Figure 43 : Histogramme du THC de KRUPP

III. Préparation des données :

1. Nettoyage des données :

Pour les données des deux stations KRUPP et KOCH, voici la structure de leurs dataframes après avoir supprimé la colonne du THE, des HM et du rendement puisque nous nous intéressons qu'aux valeurs du THC:

THC	
2021-12-20	10768.64
2021-12-21	12336.56
2021-12-22	8616.98
2021-12-23	13849.02
2021-12-24	14350.04
...	...
2023-03-16	5724.60
2023-03-17	7710.82
2023-03-18	9572.02
2023-03-19	11433.22
2023-03-20	9030.58

Figure 44 : Dataframe des données de KRUPP

THC	
2023-01-01	15279.0
2023-01-02	13572.0
2023-01-03	13436.0
2023-01-04	4872.0
2023-01-05	0.0
...	...
2023-07-03	5502.0
2023-07-04	10891.0
2023-07-05	9249.0
2023-07-06	5940.0
2023-07-07	4056.0

Figure 45 : Dataframe des données de KOCH

2. Normalisation et mise à échelle :

Chaque dataframe est ensuite convertit en un tableau numpy puis le MinMaxScaler est appliqué pour mettre à l'échelle chaque caractéristique numérique dans le jeu de manière à ce que ses valeurs soient proportionnellement transformées pour s'inscrire dans une plage entre 0 et 1. Cette étape est cruciale pour éviter les divergences et instabilités et assurer une convergence stable lors de l'entraînement de modèles d'apprentissage automatique, notamment ceux utilisant l'erreur quadratique moyenne (MSE) comme métrique.

Ci-dessous les lignes de code illustrant ces transformations :

```
# Convert the dataframe to a numpy array
dataset = data.values

scaler = MinMaxScaler(feature_range=(0,1))
scaled_data = scaler.fit_transform(dataset)

scaled_data
```

L'output est une matrice de valeurs normalisées entre 0 et 1.

3. Séparation des ensembles de données :

Nous avons divisé les données en utilisant une répartition de 80% pour l'ensemble d'entraînement et 20% pour l'ensemble de test. Cette répartition vise à séparer les données en deux ensembles distincts pour l'entraînement et l'évaluation de notre modèle. En attribuant 80% des données à l'ensemble d'entraînement, nous permettons à notre modèle d'apprendre les motifs et les relations sous-jacentes dans les données. Les 20% restants, réservés pour l'ensemble de test, vont nous permettre de vérifier à quel point notre modèle généralise bien sur des données qu'il n'a pas vues pendant l'entraînement.

Lorsque nous entraînons notre modèle , nous fournissons à notre algorithme d'apprentissage automatique 'X_train' et 'y_train' avec 'X_train' étant l'ensemble d'entraînement des attributs et 'y_train' étant l'ensemble d'entraînement des cibles correspondantes aux sorties attendues (valeurs à prédire) pour la correspondante instance dans 'X_train'. C'est ce que notre modèle tente d'apprendre à prédire lors de l'entraînement.

Ensuite, nous utilisons 'X_test' pour générer des prédictions avec notre modèle et comparer ces prédictions à 'y_test' pour évaluer sa performance sur des données jamais vues auparavant.

Ci-dessous les tailles des training et test sets de KRUPP et KOCH :

```
Shape of X_train: (364, 30, 1)
Shape of y_train: (364, 1)
Shape of X_test: (62, 30, 1)
Shape of y_test: (62, 1)
```

Figure 46 : Pour KRUPP

```
Shape of X_train: (150, 7, 1)
Shape of y_train: (150, 1)
Shape of X_test: (31, 7, 1)
Shape of y_test: (31, 1)
```

Figure 47 : Pour KOCH

Pour KRUPP, par exemple, après la répartition de 80% pour le training et 20% pour le test, nous avons pu avoir 364 instances de training dans l'ensemble X_train, où chaque instance a une séquence de 30 pas de temps, chaque pas de temps contient une seule caractéristique. Parallèlement, il y a 364 valeurs cibles dans y_train correspondantes à des instances dans X_train.

En ce qui concerne le test, il y a 62 instances de test dans X_test, chacune avec un pas de temps égal à 30, avec 62 valeurs cibles dans le test set y_test.

Cette même logique peut être suivie pour interpréter les résultats de KOCH ci-dessus.

IV. Modélisation :

1. Modèle prédictif choisi: LSTMs

Une fois que nous avons formulé clairement notre problème et préparé les données nécessaires, la prochaine étape de notre méthodologie CRISP consiste à choisir un modèle prévisionnel efficace. La modélisation s'articule en trois phases essentielles : la sélection du modèle, l'entraînement de ce dernier suivi de son amélioration continue, et enfin la comparaison avec d'autres modèles.

Dans notre contexte spécifique, nous avons décidé d'adopter l'architecture des réseaux LSTM (Long Short-Term Memory). Cette décision repose sur des considérations réfléchies. Pour justifier notre choix, nous allons prendre du recul et commencer par définir ce qu'est un réseau de neurones récurrent (RNN) ainsi que les limitations inhérentes qui ont conduit au développement des LSTMs.

i. Les RNNs :

Les réseaux de neurones récurrents (RNN) sont un type spécifique de réseaux de neurones artificiels utilisés pour traiter des données séquentielles ou des séries temporelles. Les RNNs, tout comme les réseaux de neurones à propagation en avant et les réseaux de neurones convolutifs (CNN), apprennent à partir de données d'entraînement. Cependant, ce qui distingue les RNNs, c'est leur capacité à incorporer une "mémoire" en utilisant les entrées passées pour influencer l'entrée et la sortie actuelles.

Contrairement aux réseaux de neurones profonds traditionnels qui supposent une indépendance entre les entrées et les sorties, les RNN tiennent compte des éléments précédents d'une séquence pour déterminer la sortie.

ii. Limitations des RNNs :

La mémoire à court terme dans les réseaux de neurones récurrents (RNN) présente des limitations pour certaines tâches en raison d'un problème bien connu appelé le Problème du Gradient Disparu.

En utilisant la rétropropagation, un réseau récurrent peut découvrir et capturer des dépendances complexes au sein des données d'entrée. Cependant, le processus de rétropropagation conduit souvent à des gradients à long terme qui deviennent soit extrêmement petits et disparaissent, soit croissent de manière exponentielle et explosent. Si un gradient est petit, par exemple, il ne sera pas possible de mettre à jour efficacement les poids et les biais des couches initiales à chaque session d'entraînement. Ces couches initiales sont essentielles pour reconnaître les éléments fondamentaux des données d'entrée. Si leurs poids et leurs biais ne sont pas correctement mis à jour, il est possible que l'ensemble du réseau puisse être inexact. En conséquence, des informations importantes destinées à être conservées en mémoire seront perdues.

iii. Les LSTMs :

Afin de surmonter le défi du gradient disparu, un sous-type des RNNs appelé Long Short Term Memory (LSTM) a été développé. Les LSTMs comportent un composant spécialisé de mémoire à court terme capable de préserver l'information pendant des périodes prolongées, d'où le nom "long-court-terme". Cela atténue le problème de perte de gradient, bien que le potentiel d'explosion de gradient subsiste.

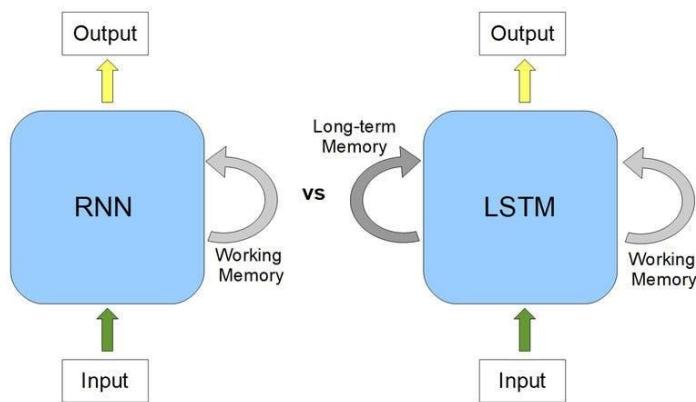


Figure 48 : Comparaison entre les RNNs et les LSTMs

Les LSTMs offrent plusieurs avantages par rapport aux RNN classiques. Tout d'abord, ils excellent dans la gestion des dépendances à long terme en conservant efficacement les informations pendant des durées prolongées.

Deuxièmement, les LSTMs présentent une vulnérabilité réduite au problème de gradient disparu en utilisant un type distinct de fonction d'activation appelée cellule LSTM, qui aide à préserver l'information à travers des séquences longues.

Enfin, les LSTMs démontrent une efficacité exceptionnelle dans la modélisation de données séquentielles complexes, car ils ont la capacité d'acquérir des représentations sophistiquées qui encapsulent la structure des données.

iv. Composants d'un LSTM :

Une cellule d'un réseau LSTM est principalement composée d'une porte d'entrée, d'une porte de sortie, d'une porte d'oubli et d'un état de cellule.

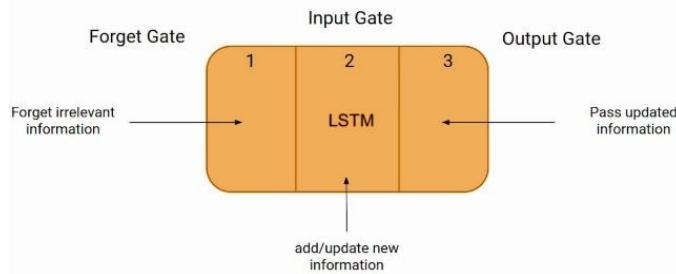


Figure 49 : Composants d'un LSTM

- La porte d'oubli (forget gate) contrôle la quantité d'information de l'étape temporelle précédente qui est conservée dans l'étape temporelle actuelle (ou en d'autres termes, combien d'informations doivent être écartées). Cette porte est entraînée à s'ouvrir lorsque l'information n'est plus importante et à se fermer lorsqu'elle l'est.
- La porte d'entrée (input gate) régit la quantité de nouvelles informations de l'étape temporelle actuelle qui est stockée dans l'état de la cellule. Elle est entraînée à s'ouvrir lorsque l'entrée est importante et à se fermer lorsque ce n'est pas le cas.
- La porte de sortie (output gate) contrôle la quantité d'information de l'état de la cellule qui est utilisée pour produire une sortie à l'étape temporelle actuelle. Elle est entraînée à s'ouvrir lorsque l'information est importante et à se fermer lorsque ce n'est pas le cas.
- L'état de la cellule (cell state) sert de vecteur qui encapsule la "mémoire" du réseau LSTM, englobant les informations à la fois des étapes temporelles précédentes et en cours.

v. L'architecture d'un LSTM :

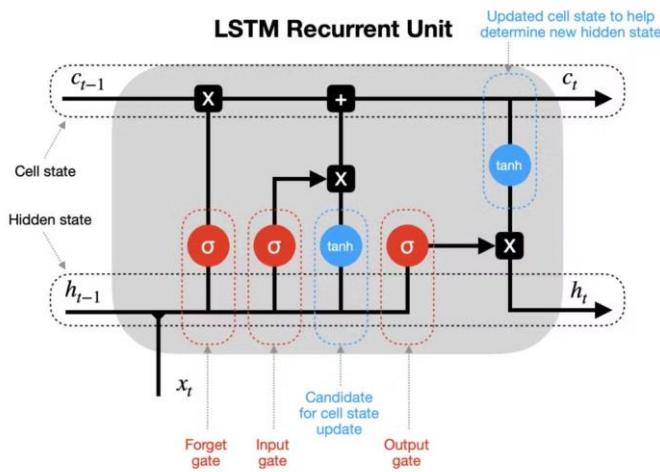


Figure 50 : Architecture d'un LSTM

Pendant chaque itération, un réseau LSTM exécute les cinq étapes suivantes :

- Identification des informations pertinentes à conserver dans l'état cellulaire qui représente la mémoire à long terme, en utilisant la porte d'oubli : À cette porte, l'entrée x_t , comme indiqué dans l'image ci-dessus, est combinée avec la sortie précédente h_{t-1} , le résultat est alimenté dans une fonction sigmoïde qui génère une valeur entre 0 et 1. La sortie est ensuite multipliée avec l'état précédent c_{t-1} .
Remarque : une sortie d'activation de 1 signifie "tout se souvenir" et une sortie d'activation de 0 signifie "tout oublier". D'un autre point de vue, un meilleur nom pour la porte d'oubli pourrait être la "porte de souvenir".
- Sélection des informations à long terme pertinentes au moyen de la porte d'entrée : La sortie de la porte d'entrée, qui est une fraction entre 0 et 1, est multipliée avec la sortie du bloc tanh qui produit les nouvelles valeurs qui doivent être ajoutées à l'état précédent c_{t-1} pour générer l'état actuel.
- Intégration des informations choisies dans l'état cellulaire : La cellule mémoire met à jour son état en fonction de la porte d'entrée, de la porte d'oubli et des données d'entrée actuelles. Elle combine les informations pour modifier sa représentation interne, lui permettant de capturer les motifs séquentiels et les dépendances à long terme dans les données.
- Reconnaissance des informations à court terme importantes dans l'état cellulaire.

- Génération de l'état caché mis à jour en utilisant la porte de sortie : À la porte de sortie, l'entrée et l'état précédent sont à nouveau régulés pour générer une autre fraction de mise à l'échelle qui est combinée avec la sortie du bloc tanh qui transmet l'état actuel c_t . Cette sortie est ensuite produite. La sortie h_t et l'état c_t sont renvoyés dans le bloc LSTM.

2. Construction du modèle prédictif :

Après avoir choisi les LSTMs comme modèle de prédiction de la production des deux stations KRUPP et KOCH, et puisque nous n'avons pas beaucoup de données, l'architecture de notre LSTM n'était pas très compliquée, formée d'abord d'une couche LSTM avec un nombre spécifié d'unités (50, 65, 80, 95, 110, 128) et une fonction d'activation Relu. Cette couche prend en entrée des séquences de données avec une forme de (time_steps, 1) où "time_steps" représente le nombre de pas de temps dans la séquence et "1" la dimension des caractéristiques à chaque pas de temps. Puis, on a ajouté une couche Dense avec une seule unité. C'est cette couche qui produira la sortie prévue.

Ci-dessous est un schéma simplifié de l'architecture que nous avons adoptée :

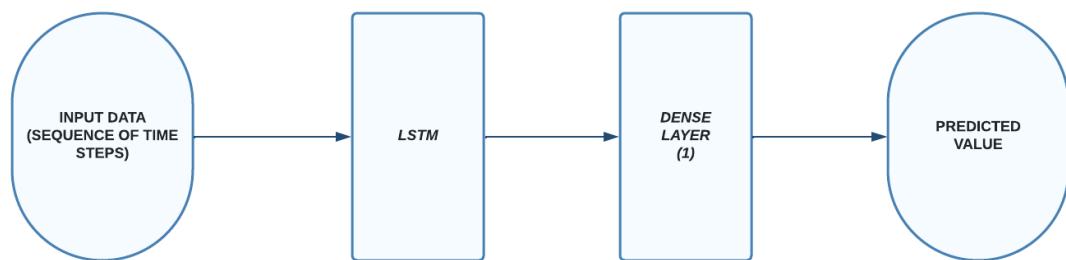


Figure 51 : Architecture du modèle choisi

3. Entraînement du modèle :

Nous avons effectué une série d'entraînements de notre modèle en utilisant différentes configurations d'unités pour les réseaux LSTM ainsi que des nombres d'époques variés. Notre approche consistait à explorer différentes combinaisons pour déterminer comment les performances du modèle évoluent en fonction de ces paramètres. Plus précisément, nous avons testé des unités de LSTM allant de 50 à 128 et des nombres d'époques de 50, 100 et 200. Pour chaque combinaison d'unités et d'époques, nous avons formé le modèle en

utilisant les données d'entraînement, puis évalué ses performances sur les données de test. Nous avons sauvegardé les modèles entraînés et les historiques d'entraînement pour chaque configuration, ce qui nous a permis de comparer et d'analyser les résultats obtenus. En ajustant ces paramètres, nous avons cherché à obtenir une meilleure compréhension de la manière dont les performances du modèle LSTM sont influencées par les choix d'architecture et de durée d'entraînement.

V. Evaluation :

1. Evaluation de la qualité et performance du modèle sur les données de test :

Dans l'étape d'évaluation de la méthodologie CRISP, nous avons choisi la Mean Squared Error (MSE) comme métrique de performance pour évaluer l'efficacité du modèle LSTM. La MSE mesure la qualité des prédictions du modèle en calculant la moyenne des carrés des différences entre les valeurs prédites et les valeurs réelles.

Pour explorer davantage la performance du modèle, nous avons suivi le processus d'entraînement en surveillant les courbes de pertes d'entraînement et de validation. Nous avons observé les résultats pour différentes durées d'entraînement, notamment avec des époques de **50, 100 et 200**. Les courbes de pertes (loss) d'entraînement et de validation ci-dessous fournissent une vision détaillée de la convergence du modèle pour les deux stations au fil des époques, et elles nous ont permis d'analyser comment le modèle s'améliore ou peut présenter des signes de surajustement. Nous avons étudié l'évolution des pertes pour chaque configuration d'époques afin de mieux comprendre la dynamique de l'apprentissage et d'identifier le point optimal où le modèle atteint un équilibre entre l'apprentissage et la généralisation. Ces résultats ont été essentiels pour prendre des décisions éclairées concernant la sélection finale de la configuration du modèle.

Pour KRUPP, les courbes de pertes des modèles de différentes unités et époques sont les suivantes :

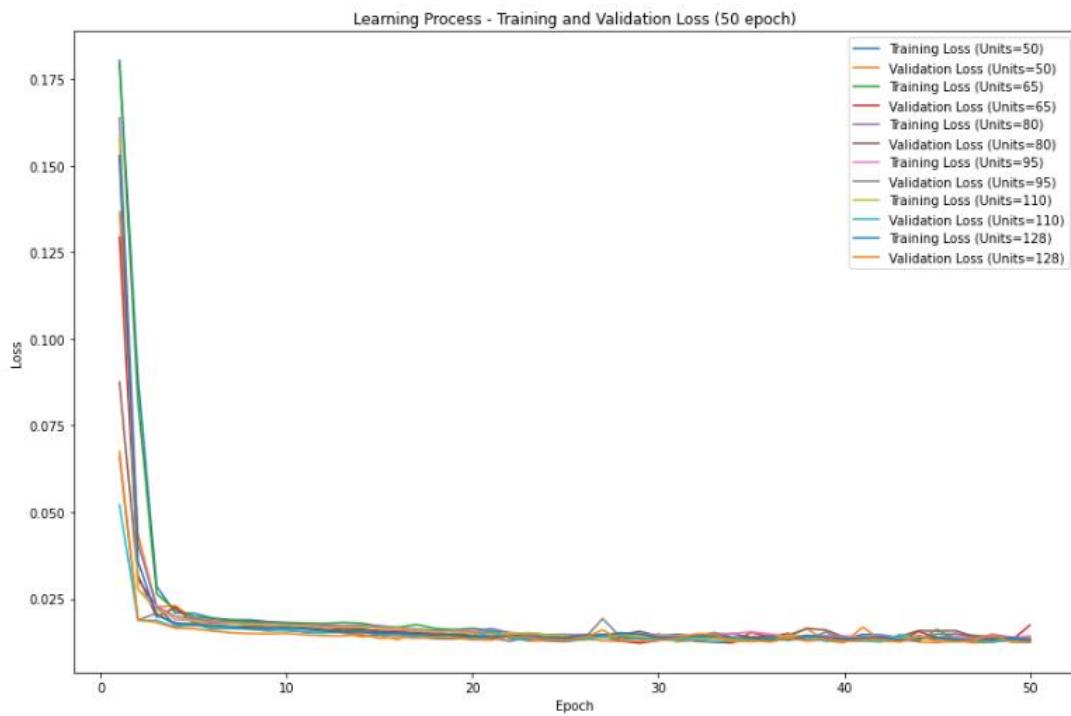


Figure 52 : Courbes de Perte d'Entraînement et de Validation pour 50 Époques pour les différentes unités de LSTM

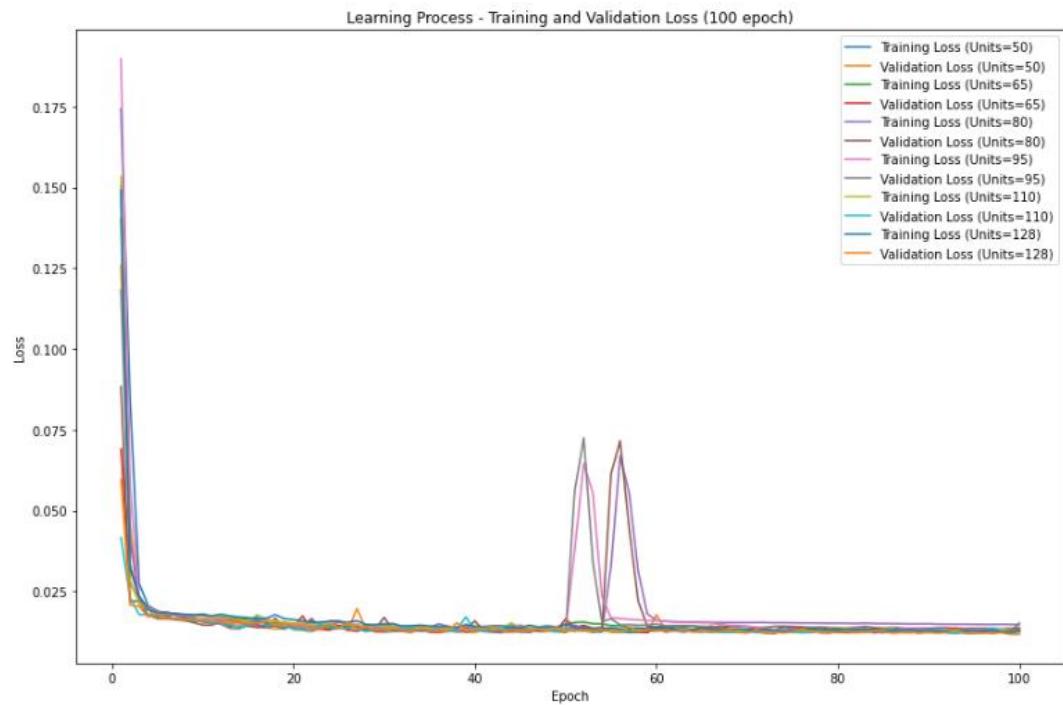


Figure 53 : Courbes de Perte d'Entraînement et de Validation pour 100 Époques pour les différentes unités de LSTM

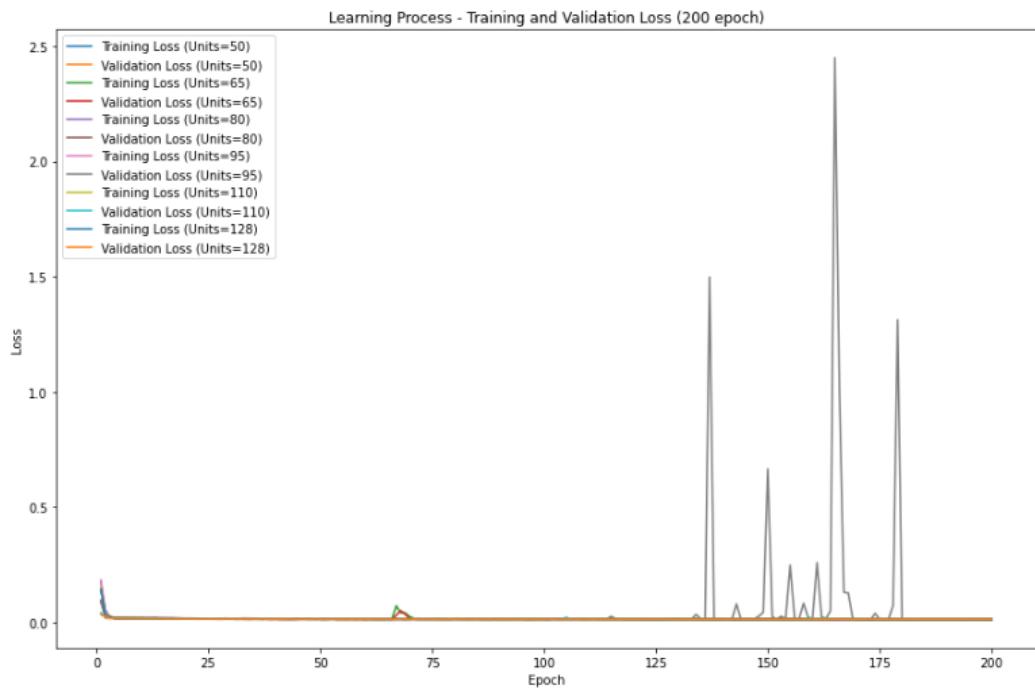


Figure 54 : Courbes de Perte d'Entraînement et de Validation pour 200 Époques pour les différentes unités de LSTM

Et pour KOCH, la station de criblage :

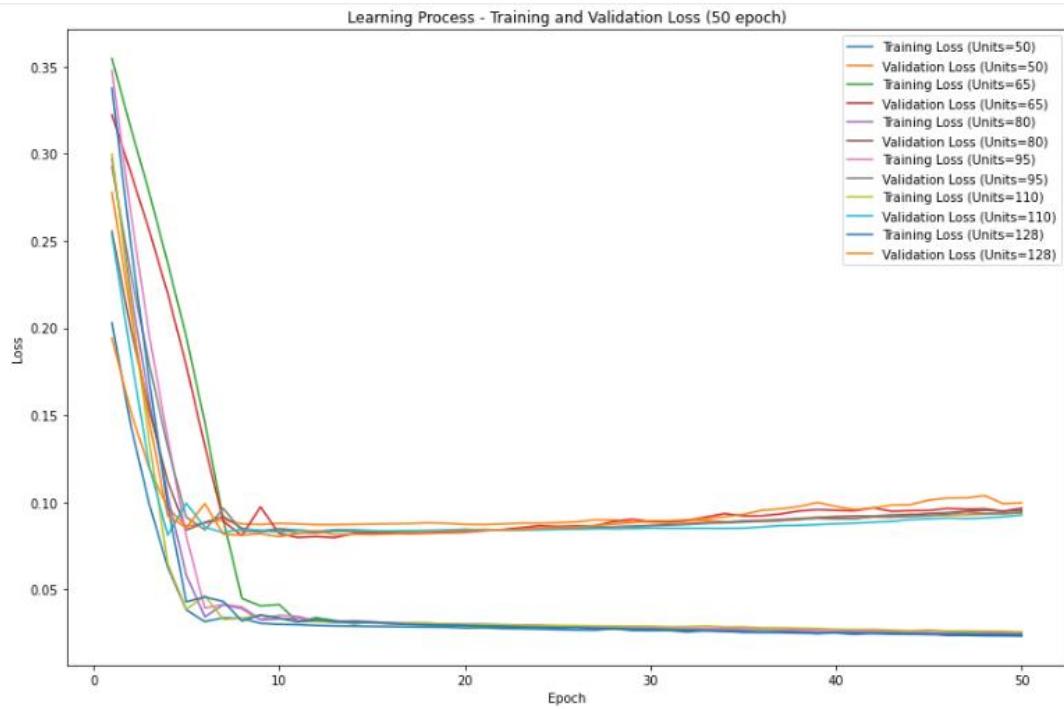


Figure 55 : Courbes de Perte d'Entraînement et de Validation pour 50 Époques pour les différentes unités de LSTM

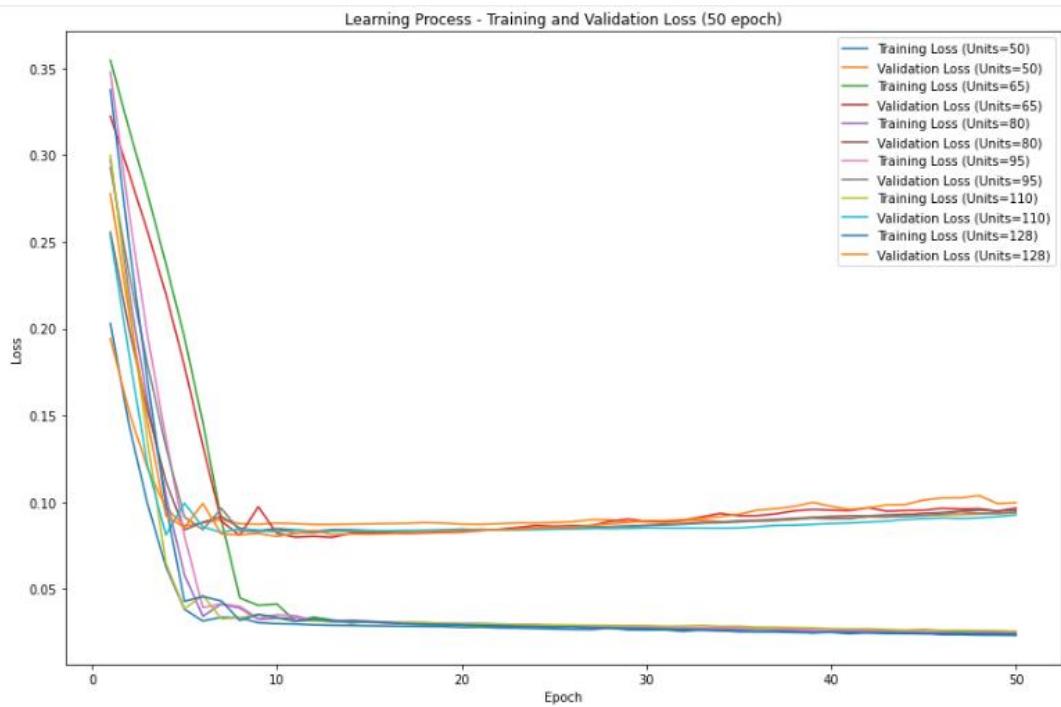


Figure 56 : Courbes de Perte d'Entraînement et de Validation pour 100 Époques pour les différentes unités de LSTM

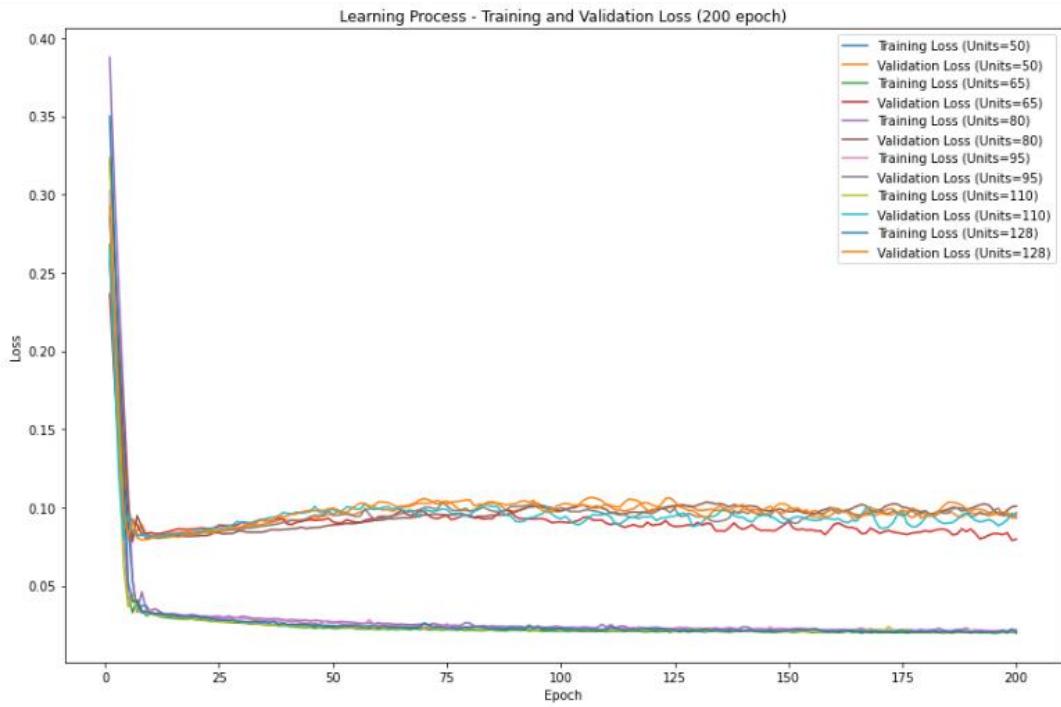


Figure 57 : Courbes de Perte d'Entraînement et de Validation pour 200 Époques pour les différentes unités de LSTM

2. Comparaison des modèles :

a. Modèles prédictifs pour KRUPP :

D'après les courbes précédentes, nous avons remarqué que, pour KRUPP, les meilleurs modèles prédictifs étaient:

Nom du modèle	Unités LSTM	Epoques	loss
<i>Model 1 (KRUPP)</i>	128	50	0.0124
<i>Model 2 (KRUPP)</i>	128	100	0.0118
<i>Model 3 (KRUPP)</i>	50	200	0.0112

Les graphes ci-dessous illustrent ces résultats :

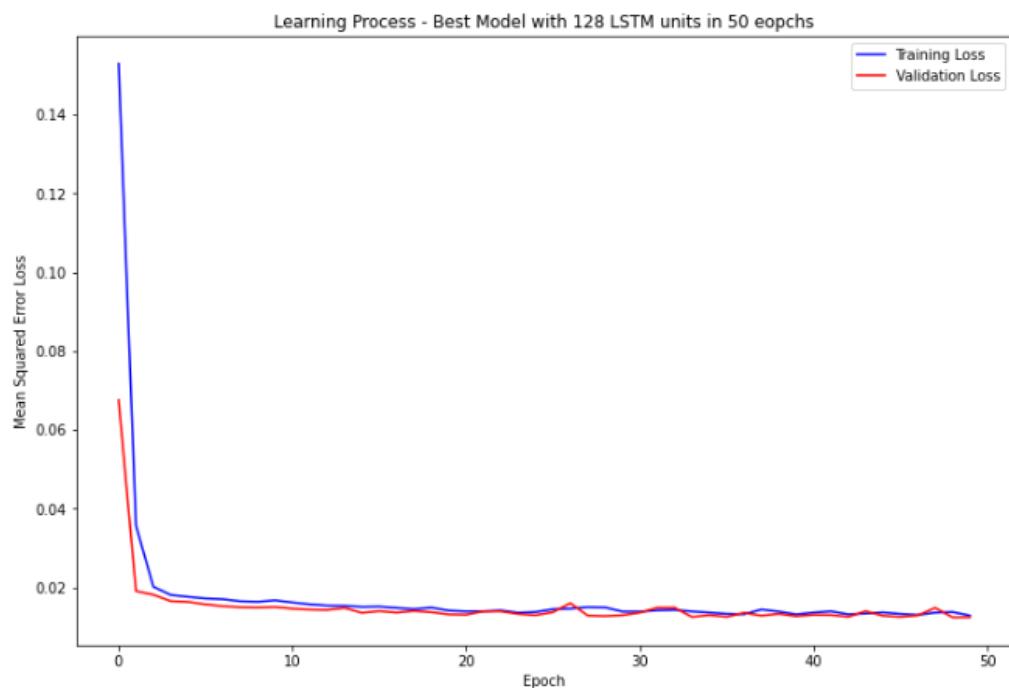


Figure 58 : Le meilleur modèle avec 128 unités entraînés dans 50 époques

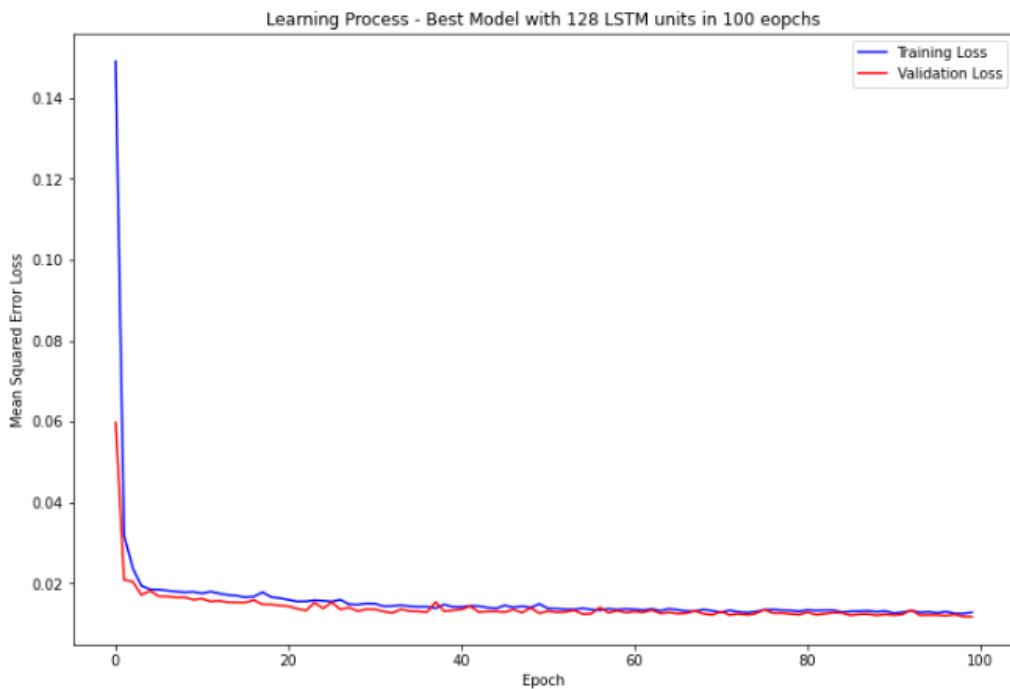


Figure 59 : Le meilleur modèle avec 128 unités entraînés dans 100 époques

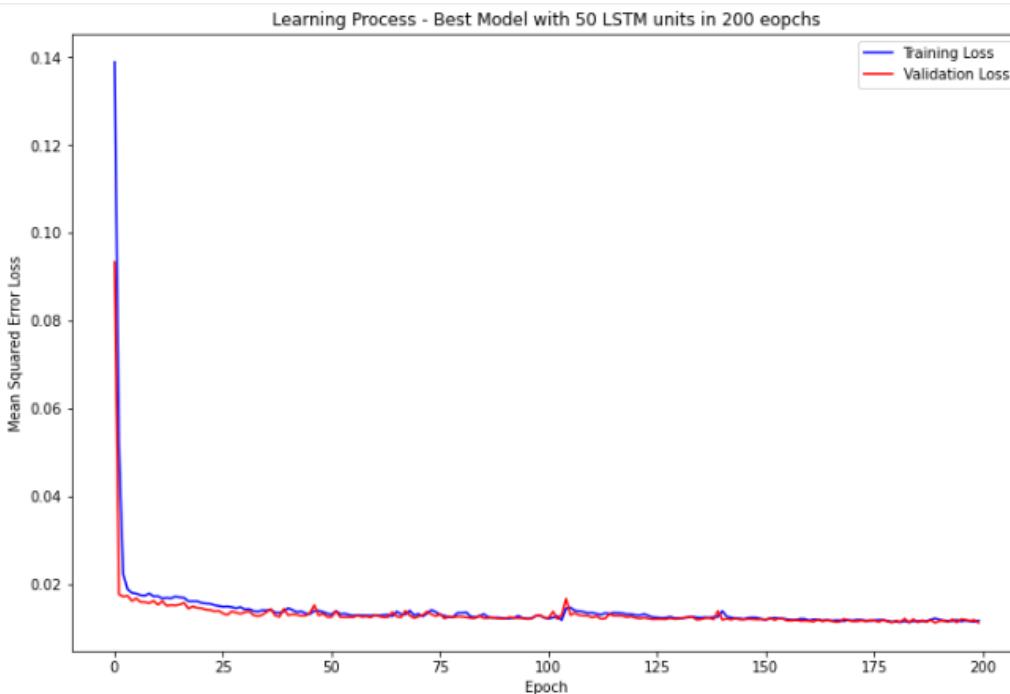


Figure 60 : Le meilleur modèle avec 50 unités entraînés dans 200 époques

Remarque :

Dans les trois modèles adoptés pour les données de KRUPP, la MSE diminue au fil des époques d'entraînement, ce qui indique que leurs prédictions s'ajustent progressivement aux données réelles, ce qui est une indication positive de leur capacité à apprendre et à

généraliser les patterns des données d'entraînement. La MSE sur les jeux de validation diminue également, ce qui nous assure que les modèles ne s'adaptent pas aux données d'entraînement et par conséquent, on n'a pas le problème d'overfitting.

b. Modèles prédictifs pour KOCH :

Pour KOCH, les modèles les plus performants accompagnés de leurs graphes de perte sont représentés ci-dessous:

Nom du modèle	Unités LSTM	Epoques	loss
<i>Model 1 (KOCH)</i>	110	50	0.0926
<i>Model 2 (KOCH)</i>	110	100	0.0926
<i>Model 3 (KOCH)</i>	65	200	0.0797

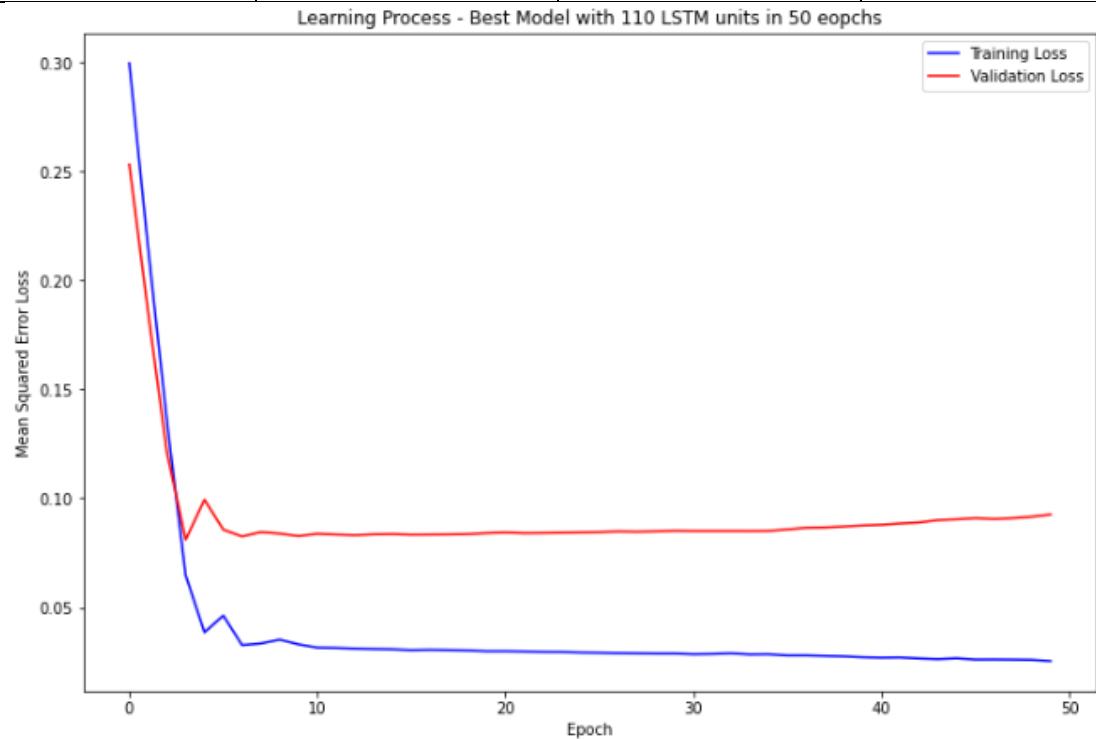


Figure 61 : Meilleur modèle avec 110 unités entraîné dans 50 époques

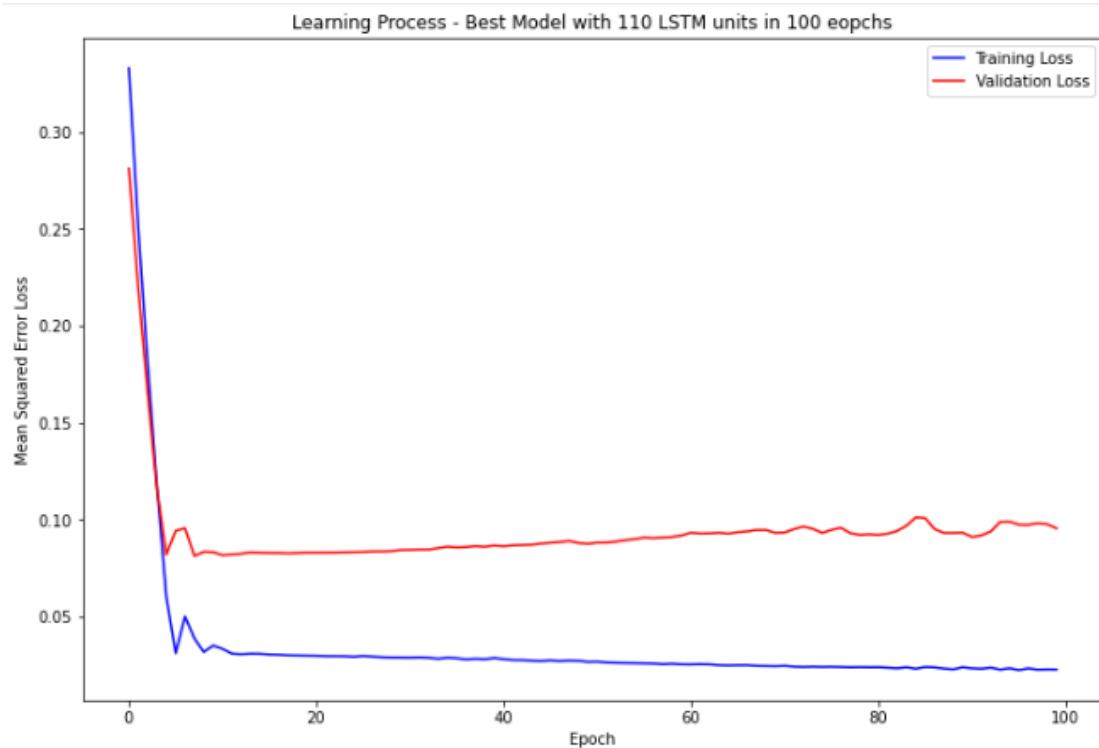


Figure 62 : Meilleur modèle avec 110 unités entraîné dans 100 époques

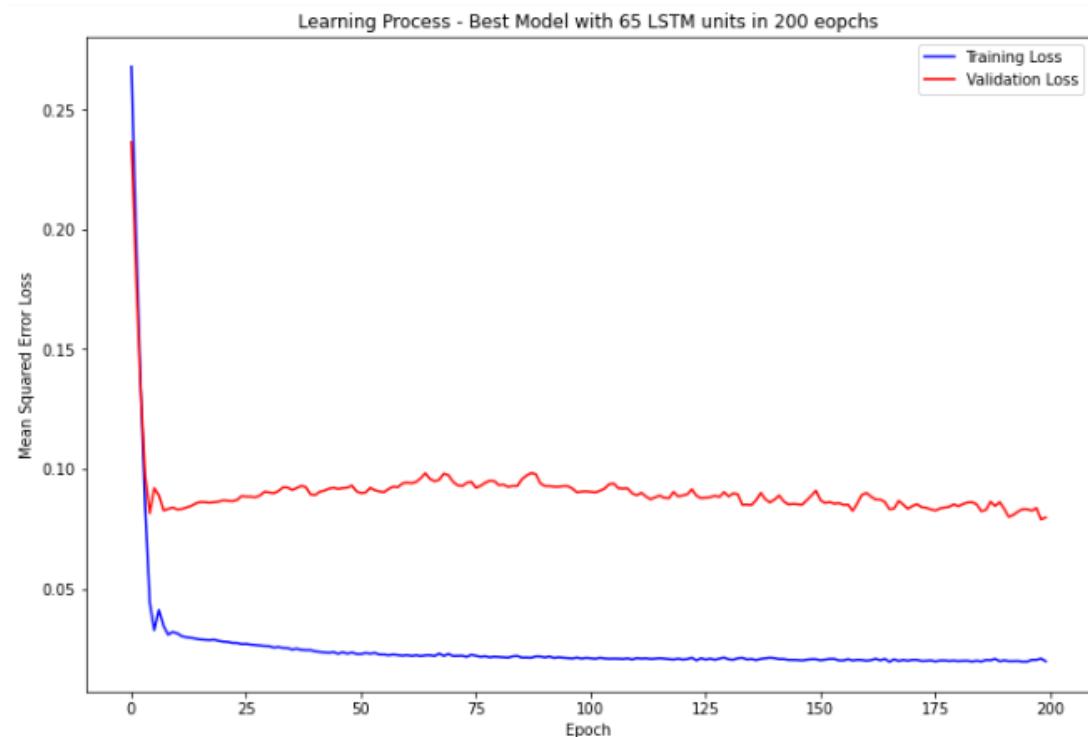


Figure 63 : Meilleur modèle avec 65 unités entraîné dans 200 époques

Remarque :

Dans les trois modèles adoptés pour les données de KOCH, nous remarquons que la MSE sur l'ensemble d'entraînement tend vers 0, ce qui indique que les modèles sont en train de

s'ajuster très étroitement aux données d'entraînement, voire même de les mémoriser. Cependant, la MSE sur l'ensemble de validation reste relativement constante à un niveau plus élevé (à environ 0.1), ce qui peut suggérer que les modèles ne généralisent pas aussi bien aux nouvelles données alors qu'ils ont réussi à capturer les détails spécifiques des données d'entraînement, ce qui est un signe classique d'overfitting.

3. Évaluation des modèles sur les données du test :

Pour les deux stations, nous avons évalué les trois modèles sur les données de test. Après avoir comparé les valeurs réelles de la production de THC avec celles prédites par les modèles, nous avons obtenu les graphes suivants :

a. Pour KRUPP :

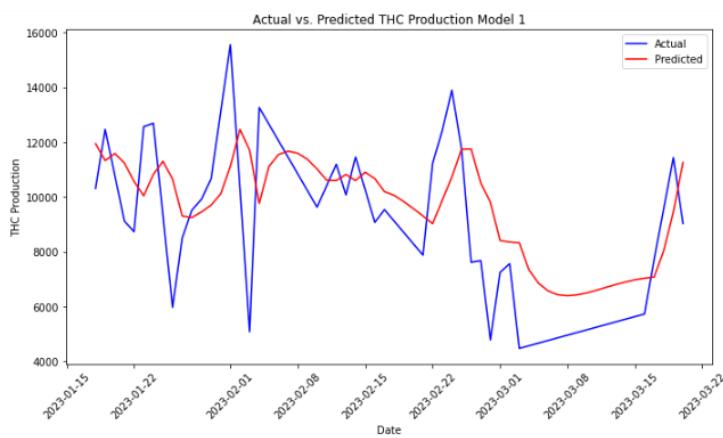


Figure 64 : Différence entre les valeurs de THC réelles et prédites pour Model 1

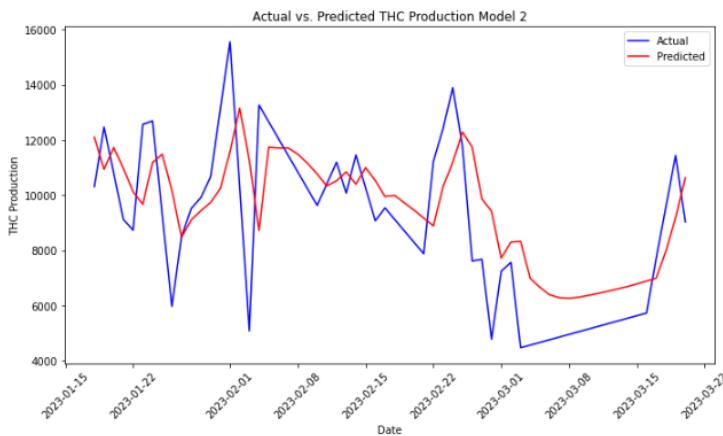


Figure 65 : Différence entre les valeurs de THC réelles et prédites pour Model 2

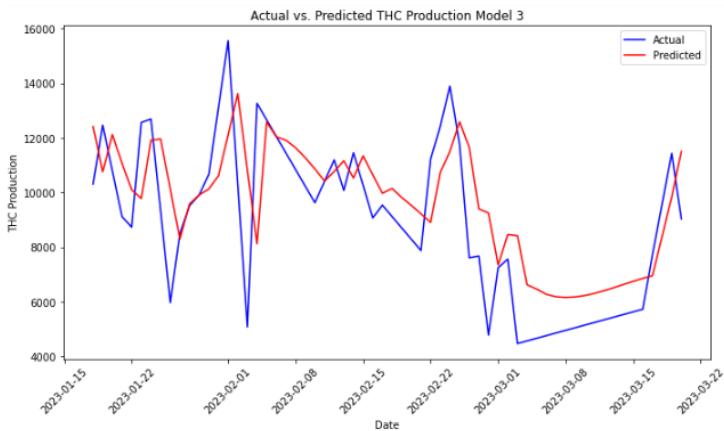


Figure 66 : Différence entre les valeurs de THC réelles et prédictes pour Model 3

Remarque :

D'après ces graphiques, il est clair que *Model 3* affiche la meilleure performance.

b. Pour KOCH :

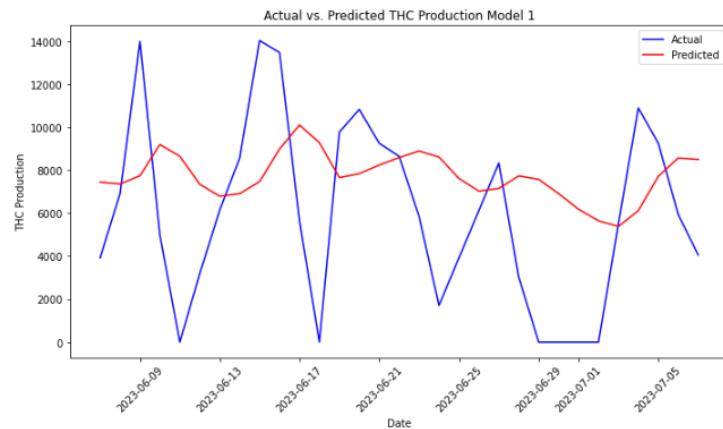


Figure 67 : Différence entre les valeurs de THC réelles et prédictes pour Model 1

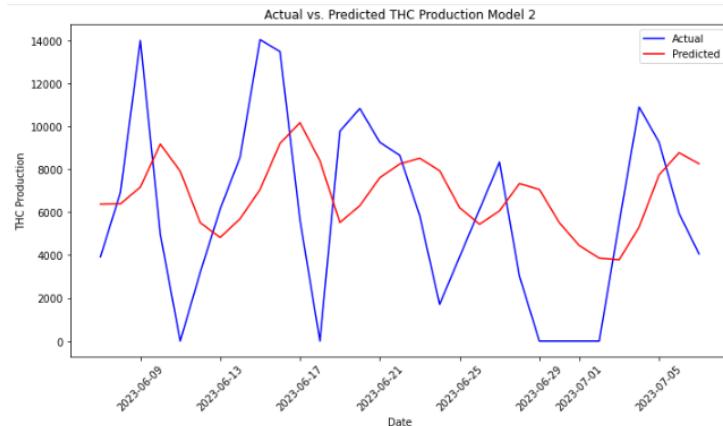


Figure 68 : Différence entre les valeurs de THC réelles et prédictes pour Model 2

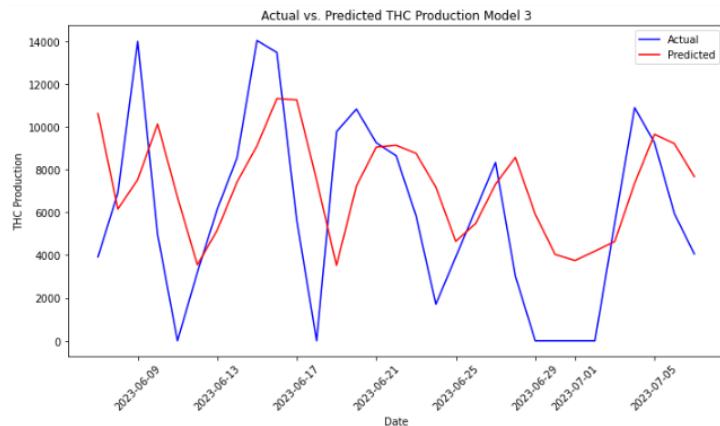


Figure 69 : Différence entre les valeurs de THC réelles et prédites pour Model 3

Remarque :

D'après ces graphiques, il est clair que *Model 3* affiche la meilleure performance.

4. Analyse des Courbes de Validation et de Prédictions :

a. Pour les données de KRUPP :

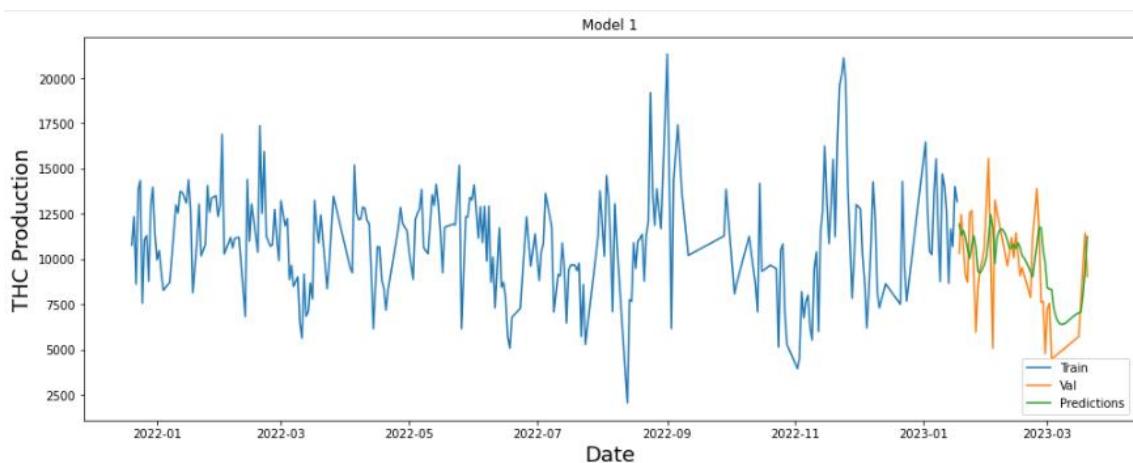


Figure 70 : Graphique illustrant la différence entre les valeurs de THC de l'ensemble de validation et celles prédites par Model 1

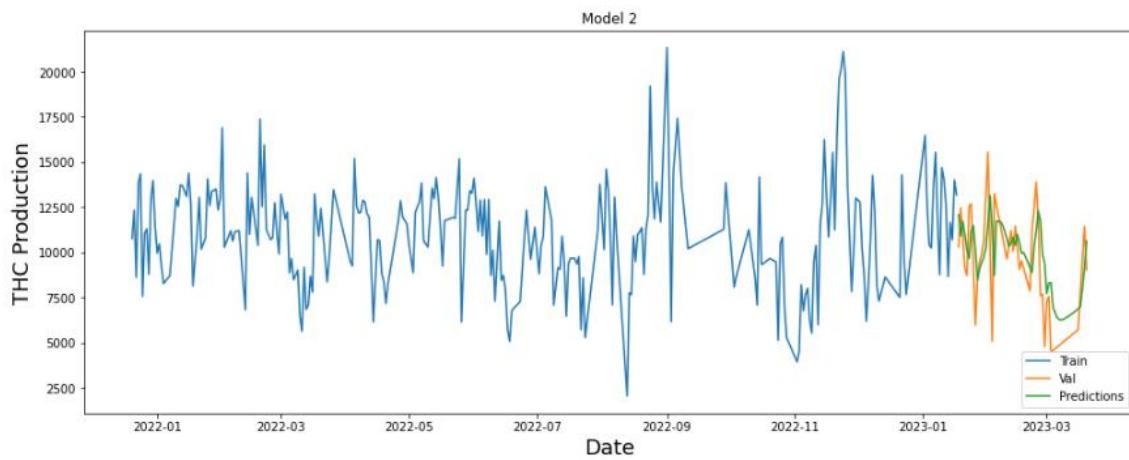


Figure 71 : Graphique illustrant la différence entre les valeurs de THC de l'ensemble de validation et celles prédites par Model 2

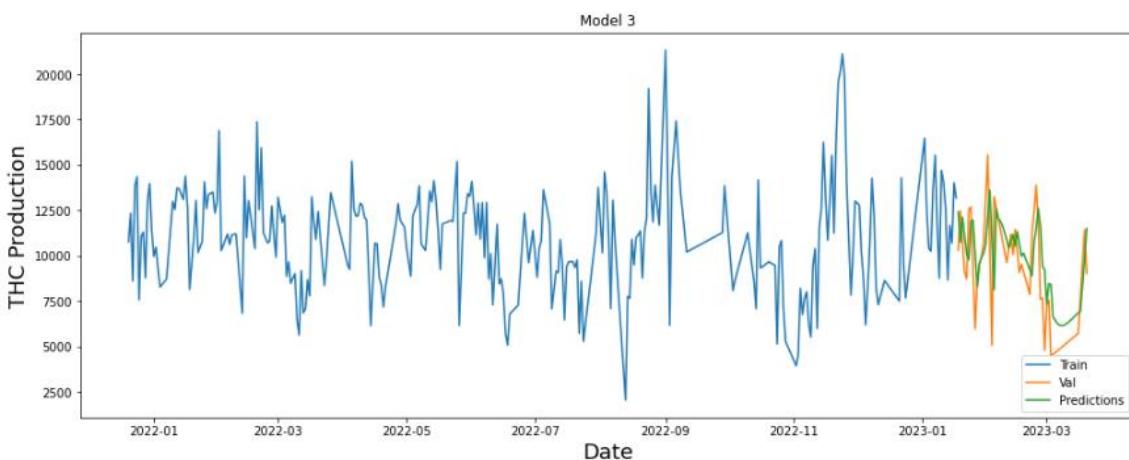


Figure 72 : Graphique illustrant la différence entre les valeurs de THC de l'ensemble de validation et celles prédites par Model 3

Remarque :

D'après ces graphiques, il est évident que *Model 3* se distingue par sa performance supérieure.

b. Pour les données de KOCH :

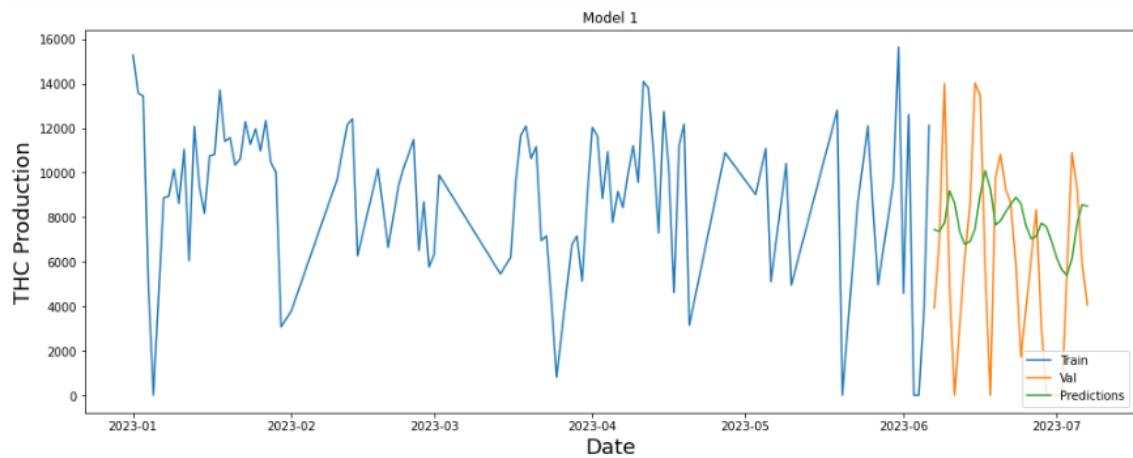


Figure 73 : Graphique illustrant la différence entre les valeurs de THC de l'ensemble de validation et celles prédites par Model 1

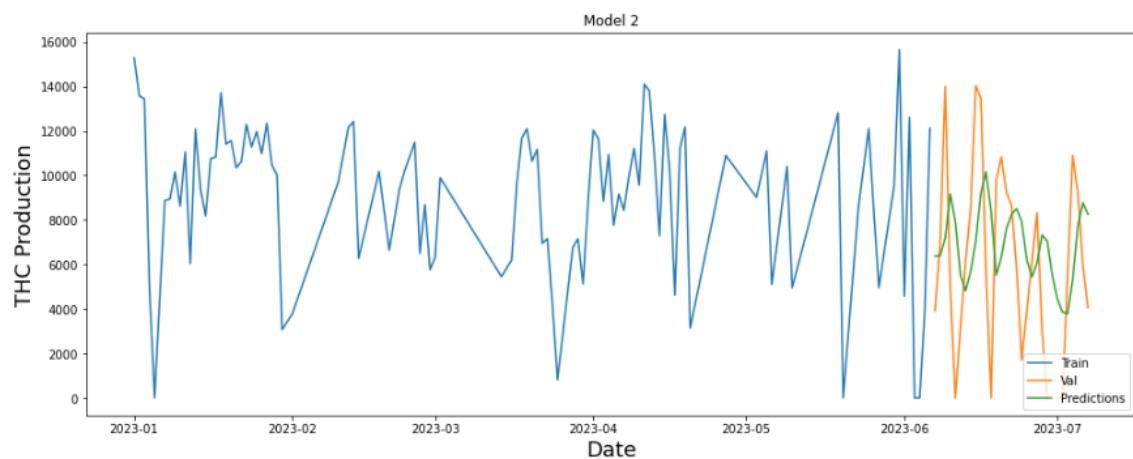


Figure 74 : Graphique illustrant la différence entre les valeurs de THC de l'ensemble de validation et celles prédites par Model 2

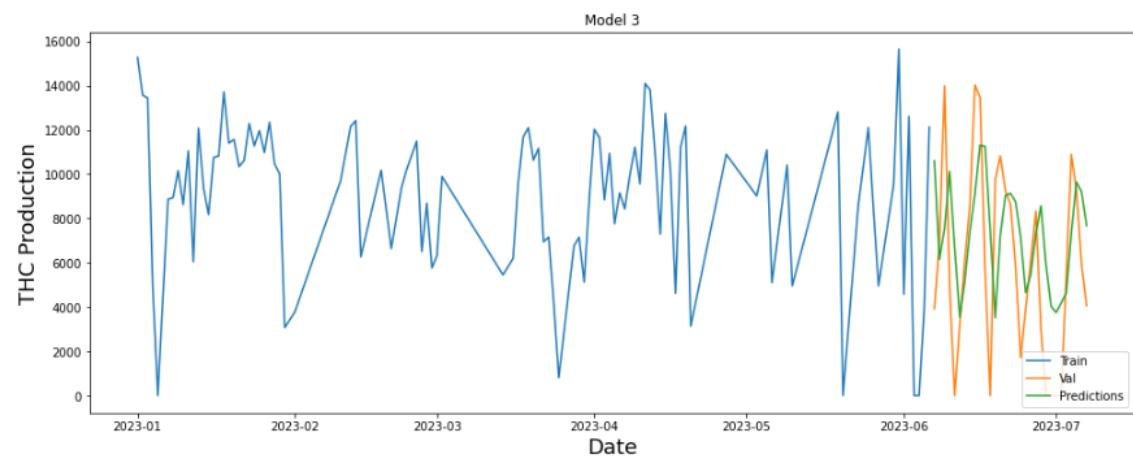


Figure 75 : Graphique illustrant la différence entre les valeurs de THC de l'ensemble de validation et celles prédites par Model 3

Remarque :

D'après ces graphiques, il est évident que *Model 3* se distingue par sa performance supérieure.

5. Prédiction de valeurs de THC futures :

Pour choisir les modèles les plus performants pour la prédiction de la production des deux stations KRUPP et KOCH, nous avons essayé de prédire, grâce aux modèles précédents, des valeurs de THC pour des nouveaux temps futurs, des valeurs qui n'existent même pas dans l'ensemble de test. Les résultats de ces prédictions sont présentés par la suite :

a. Pour KRUPP :

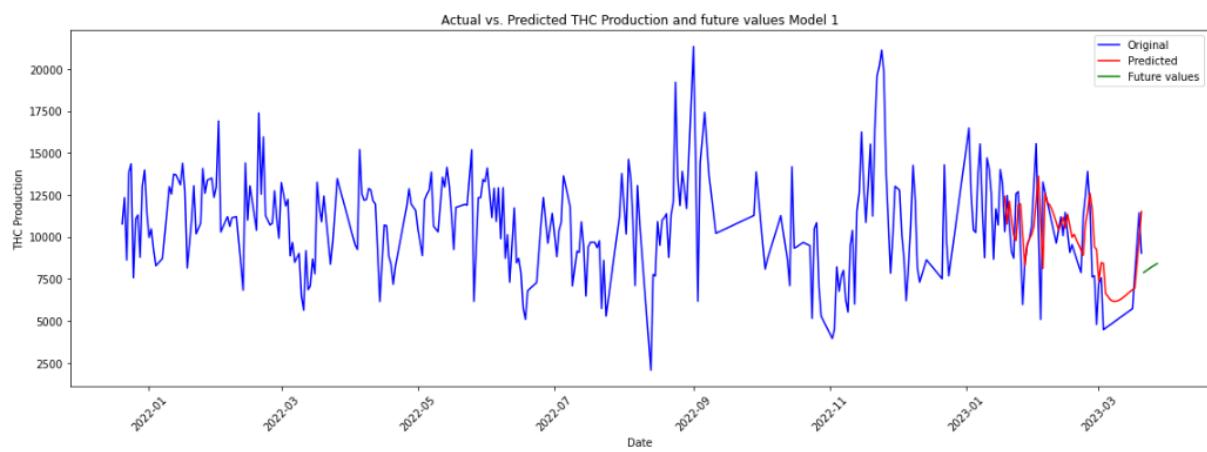


Figure 76 : Prédiction de futures valeurs de THC par Model 1

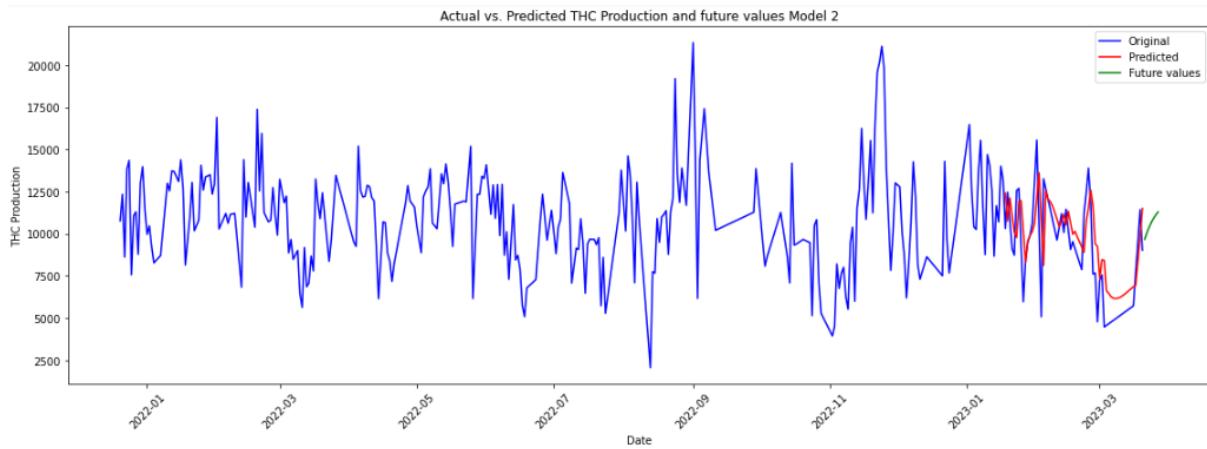


Figure 77 : Prédiction de futures valeurs de THC par Model 2

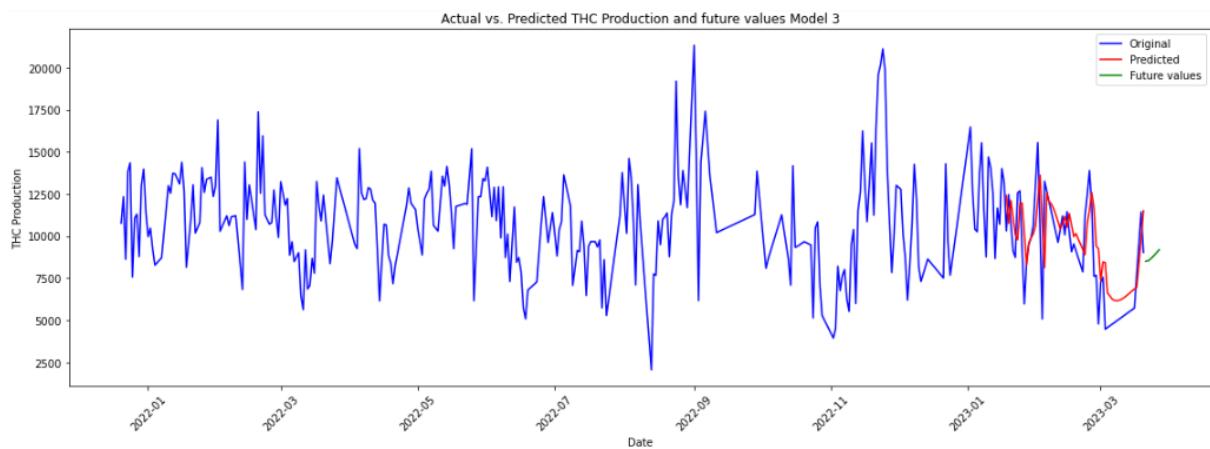


Figure 78 : Prédiction de futures valeurs de THC par Model 3

Remarque :

Les trois graphiques concordent pour montrer que *model 3* se démarque par ses performances supérieures.

b. Pour KOCH :

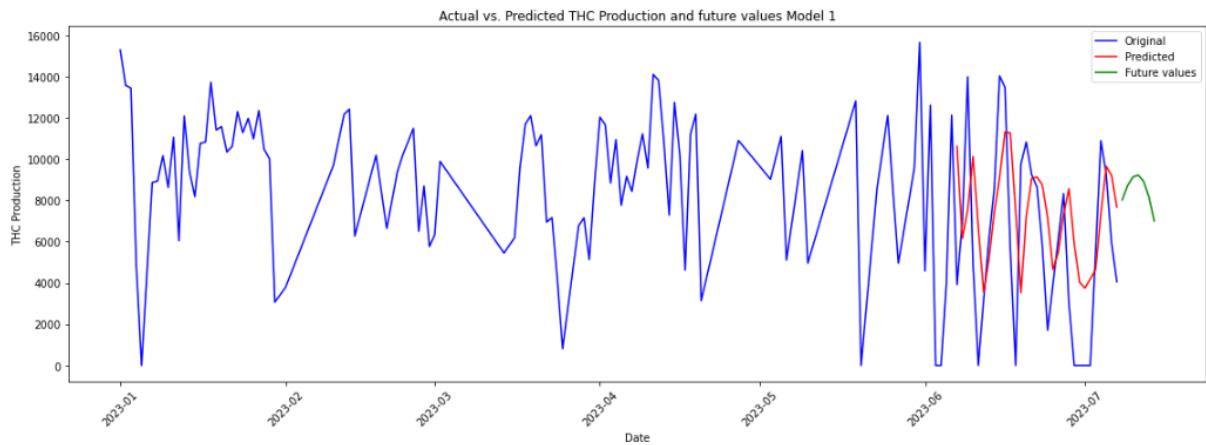


Figure 79 : Prédiction de futures valeurs de THC par Model 1

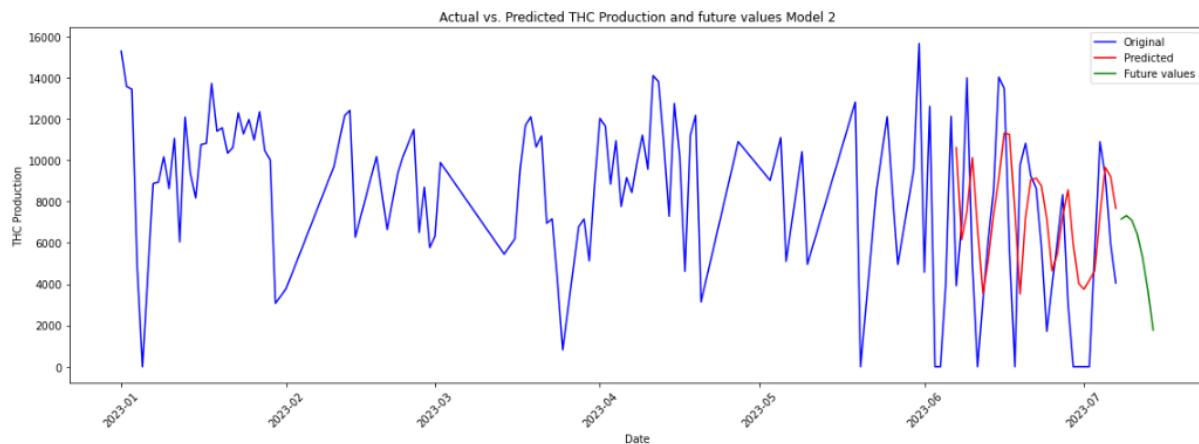


Figure 80 : Prédiction de futures valeurs de THC par Model 2

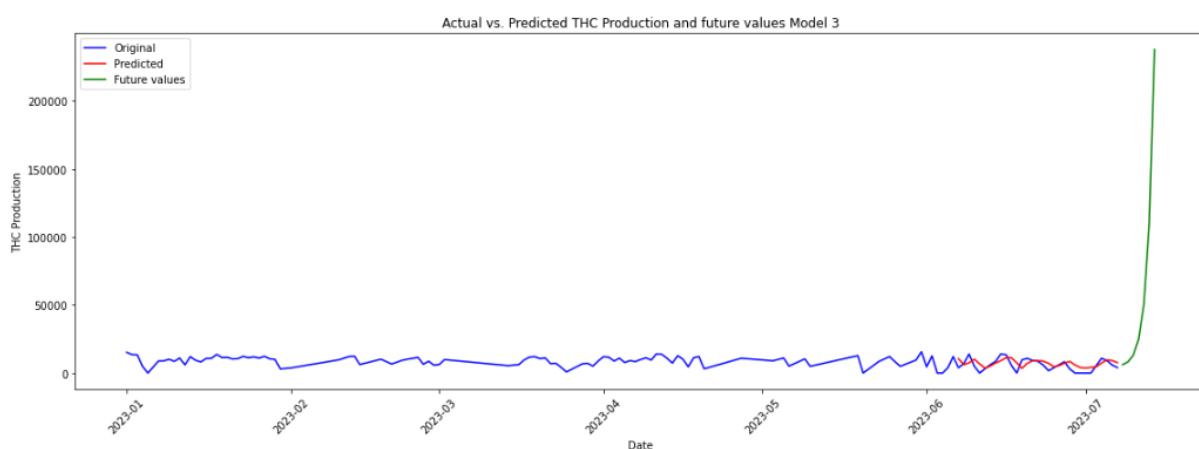


Figure 81 : Prédiction de futures valeurs de THC par Model 3

Remarque :

Les trois graphiques concordent pour montrer que *Model 3* se démarque par ses performances supérieures.

VI. Déploiement :

Arrivant à la dernière étape de la méthodologie CRISP-DM (Cross-Industry Standard Process for Data Mining) ; le déploiement.

Le déploiement est défini comme un processus par lequel un modèle d'apprentissage automatique est intégré dans un environnement de production existant pour obtenir des décisions commerciales efficaces basées sur des données. C'est l'une des dernières étapes du cycle de vie de l'apprentissage automatique.

1. Outils et cadres de déploiement utilisés :

a. Le Framework Flask :



Figure 82 : Logo de Flask

Flask est un micro Framework open-source de développement web en Python. Il est très léger et garde un noyau simple mais extensible, ce qui en fait un choix populaire pour la création d'applications web simples et rapides. Il offre les fonctionnalités essentielles au développement web, comme la gestion des requêtes HTTP, le serveur web ou la gestion des cookies. Il se base sur la flexibilité et la lisibilité du langage Python. Il est facile à prendre en main et optimise le processus de développement.

b. Bootstrap :



Figure 83 : Logo de Bootstrap

Bootstrap est un Framework front-end open source largement utilisé pour le développement web. Il a été créé par Twitter et est maintenant maintenu par la communauté open source. Bootstrap simplifie la création d'interfaces utilisateur réactives, modernes et esthétiquement agréables en fournissant des composants prêts à l'emploi tels que des boutons, des formulaires, des barres de navigation, des modales, des grilles de mise en page, et bien plus encore. De plus, Bootstrap est basé sur HTML, CSS et JavaScript, ce qui en fait un choix populaire pour la conception de sites web et d'applications web.

c. Autres langages utilisés :

HTML , CSS, JavaScript

2. Environnement de développement : Visual Studio Code

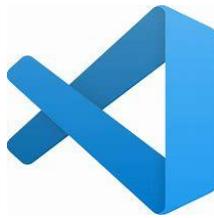


Figure 84 : Logo de Visual Studio Code

Visual Studio Code est un éditeur de code extensible développé par Microsoft pour Windows, Linux et macOS. Les fonctionnalités incluent la prise en charge du débogage, la mise en évidence de la syntaxe, la complétion intelligente du code, les snippets, la refactorisation du code et Git intégré. Les utilisateurs peuvent modifier le thème, les raccourcis clavier, les préférences et installer des extensions qui ajoutent des fonctionnalités supplémentaires.

3. Réalisation de la plateforme :

Sur la page d'accueil de notre application web, nous avons inclus une brève description de notre projet. En haut à droite, vous trouverez l'année et la période d'exécution du projet pendant notre stage. En haut à gauche, le logo du Groupe OCP est affiché, et en cliquant dessus, vous serez dirigé vers le site officiel du groupe. En bas de la page, nos noms, en tant que deux stagiaires ayant travaillé sur le projet, sont répertoriés. En cliquant sur chaque nom, vous serez dirigé vers le profil LinkedIn respectif du stagiaire.

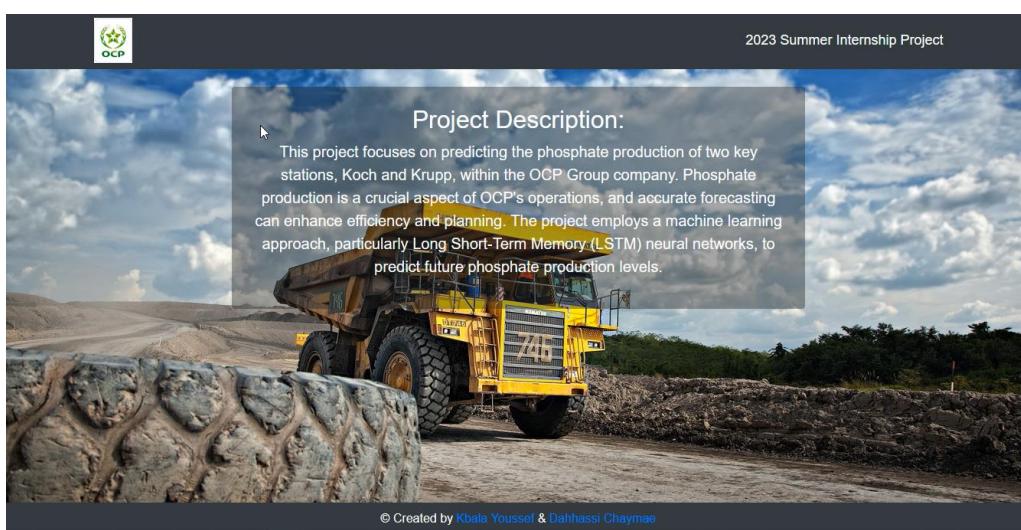


Figure 85 : Page d'accueil de la page web

Sur la même page, vers le bas, vous trouverez deux boutons distincts, chacun d'entre eux vous dirigeant vers l'une des stations, soit KOCH, soit KRUPP.



Figure 86 : Page d'accueil de la page web

Lorsque vous cliquez sur KRUPP, par exemple, vous avez la possibilité de sélectionner la source de vos données : vous pouvez choisir entre les données déjà présentes sur la plateforme pour l'année 2023, ou vous avez la possibilité d'importer d'autres données que vous souhaitez télécharger.

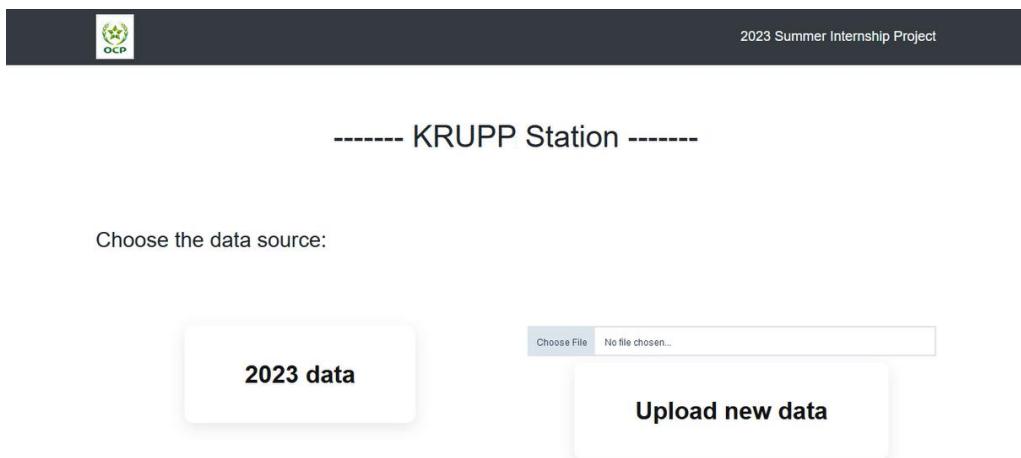


Figure 87 : Page de choix de la source de données

Pour les données de 2023 de la même station par exemple, nous commencerons par examiner un aperçu du jeu de données :

Overview of the data source file:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1			THE			THC								Rendement			
2	Date	poste 3	poste 1	poste 2	Journée	poste 3	poste 1	poste 2	Journée	poste 3	poste 1	poste 2	Journée	poste 3	poste 1	poste 2	Journée
3																	
4	12/20/2021	2480	3840	5136	11456	2331	3410	4828	10769	2.74	3.5	5.48	11.72	851	1031.314	881	977
5	12/21/2021	2896	4854	5372	13124	2722	4545	5050	12337	2.57	4.8	6.05	13.42	1059	950.7447	835	978
6	12/22/2021	4150	2851	2166	9187	3701	2480	2034	8617	3.7	2.4	2.19	6.67	1000	1030.746	930	1.055
7	12/23/2021	5159	4802	5072	14733	4849	4232	4768	13849	6.17	4.94	6.6	17.71	786	856.6559	722	832
8	12/24/2021	5271	4465	5530	15264	4755	4197	5198	14350	6.7	5.31	7.48	19.49	740	790.4143	695	783
9	12/25/2021	4271	461	3318	8050	4015	433	3119	7567	5.25	0.67	4.71	10.63	745	646.7761	642	757

Figure 88 : Aperçu des données

Ensuite, nous présenterons des visualisations graphiques dans l'ordre suivant :

- Un graphique exploratoire des données.
- Une représentation graphique de la décomposition en séries temporelles.
- Un graphique affichant des statistiques glissantes.
- Une illustration de la performance du modèle sélectionné sur l'ensemble d'entraînement.
- Un graphique présentant les prédictions des valeurs futures du THC

Conclusion générale et perspectives :

Notre mission durant ce stage de formation au sein de l'OCP SA a consisté à développer un modèle d'apprentissage automatique pour la prédiction de la production des stations KRUPP et KOCH en exploitant des séries temporelles.

Alors, pour mener à bien cette mission, j'ai suivi la méthodologie CRISP-DM avec ses six étapes permettant d'explorer et de traiter les données destinées pour l'entraînement du modèle. Ainsi, une étude comparative des algorithmes de prédiction des séries temporelles a été effectuée pour y aboutir à la fin au choix final qui se base sur l'implémentation des réseaux de neurones LSTM.

Finalement, nous avons décidé de déployer les modèles sous forme d'une application web en utilisant les Frameworks Flask et Bootstrap.

Par ailleurs, ce stage fut une expérience très enrichissante pour nous sur les deux plans personnel et professionnel, renforçant nos compétences en travail d'équipe, notre gestion de temps et prise d'initiative, notre capacité à communiquer efficacement, ainsi que notre aptitude à résoudre des problèmes complexes et à nous adapter rapidement à de nouvelles situations. C'était aussi une occasion de découvrir le dynamisme et l'enthousiasme qui caractérisent les équipes du Groupe OCP SA.

Enfin, nous espérons que cette expérience sera une préparation à une meilleure insertion dans le domaine professionnel.

Webographie :

<https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/>

https://en.wikipedia.org/wiki/Long_short-term_memory

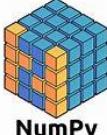
<https://datascientest.com/recurrent-neural-network>

<https://datavalue-consulting.com/deep-learning-reseaux-neurones-recurrents-rnn/>

<https://towardsdatascience.com/exploring-the-lstm-neural-network-model-for-time-series-8b7685aa8cf#:~:text=One%20of%20the%20most%20advanced,more%20parameters%20to%20be%20learned.C-model-for-time-series->

Annexe :

Les bibliothèques Python essentielles utilisées dans le projet

Nom de la bibliothèque	Logo	Définition et utilité dans le projet
Pandas		Bibliothèque dédiée à la manipulation et l'analyse des données. Elle propose notamment des structures de données et des opérations pour la manipulation de tableaux numériques et de séries temporelles.
Matplotlib		Bibliothèque python destinée à tracer et visualiser des données sous forme de graphiques.
Numpy		C'est une extension de Python, destinée à manipuler des matrices ou tableaux multidimensionnels ainsi que des fonctions mathématiques opérant sur ces tableaux.
Seaborn		Bibliothèque pour la réalisation de graphiques statistiques en Python.
Keras		Utilisée pour développer et évaluer des modèles d'apprentissage approfondi. Dans notre cas, nous l'avons utilisé pour le développement des réseaux de neurones LSTM.
Scipy		Bibliothèque Python utilisée pour le calcul scientifique et le calcul technique.
statsmodels		Statsmodels est un module Python qui fournit des classes et des fonctions pour l'estimation de nombreux modèles statistiques différents.

Parmi les fonctionnalités PYTHON utilisées au cours de l'analyse exploratoire des données en exploitant ces bibliothèques, on trouve :

- ✚ Pandas : `dataframe.describe()`

La bibliothèque pandas de Python offre la fonction `describe()` qui renseigne sur les informations générales du fichier de données, ceci par colonnes:

Count = Le nombre d'enregistrements pour l'attribut(=variable=colonne);

Mean = La moyenne arithmétique des valeurs de la colonne;

std = L'écart-type des valeurs de l'attribut par rapport à la moyenne;

min = La valeur minimale sur la colonne (aussi appelé quartile 0);

25% = Le 1er quartile qui sépare les 25 % inférieurs des données;

50% = Le 1er quartile qui sépare les 50 % inférieurs des données;

75% = Le 1er quartile qui sépare les 75 % inférieurs des données;

max = La valeur maximale sur la colonne (aussi appelé quartile 4);

Matplotlib : Carte de chaleur (heatmap)

C'est une représentation graphique de données statistiques qui fait correspondre à l'intensité d'une grandeur variable une gamme de tons ou un nuancier de couleurs sur une matrice à deux dimensions. Ce procédé permet de donner aux données un aspect visuel plus facile à saisir qu'un tableau de statistiques.

Elle montre ainsi la corrélation entre les différentes variables d'une trame de données quelconque.

La valeur d'un coefficient de corrélation peut varier de -1 à +1. Un -1 indique une corrélation négative parfaite, tandis qu'un +1 indique une corrélation positive parfaite. Une corrélation de 0 signifie qu'il n'y a pas de relation entre les deux variables. Lorsqu'il y a une corrélation négative entre deux variables, la valeur d'une variable augmente, celle de l'autre diminue, et vice versa. En d'autres termes, pour une corrélation négative, les variables fonctionnent en sens inverse. Lorsqu'il y a une corrélation positive entre deux variables, la valeur d'une variable augmente, la valeur de l'autre variable augmente également. Les variables se déplacent ensemble.