



République Tunisienne
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université de Tunis El Manar
École Nationale d'Ingénieurs de Tunis



Département de Génie Industriel

Rapport de Stage Ingénieur

Systèmes dynamiques et météorologie

Réalisé par :

Lekehal Hammada

2^{ème} Année Génie Industriel

Encadré par :

Hasna Riahi

Réalisé à :



Année académique : 2023/2024



Dédicaces

Dédié à tous ceux dont le soutien infaillible et les précieux conseils ont été déterminants dans la concrétisation de ce projet.

À mon mentor estimé, Professeure Hasna Riahi,

dont la sagesse et l'expertise ont été des ressources inestimables tout au long de ce parcours, je vous adresse mes plus sincères remerciements.

À ma famille,

dont l'encouragement constant et la compréhension m'ont soutenu durant les moments les plus difficiles, j'exprime ma profonde gratitude.

À mes camarades et amis,

votre soutien indéfectible a été une source inépuisable de force et de motivation.

Ce travail est le fruit des efforts collectifs de tous ceux qui ont contribué à son aboutissement, et je le dédie à ceux qui m'inspirent à toujours viser l'excellence.



Remerciements

Je tiens à exprimer ma profonde gratitude à toutes les personnes qui m'ont soutenu tout au long de ce projet.

Tout d'abord, je remercie mon mentor, le **Professeur Hasna Riahi**, pour ses conseils précieux, sa disponibilité et son expertise, qui ont grandement contribué à l'avancement de mes travaux.

Je souhaite également remercier le **laboratoire LAMSIN**, pour son soutien technique et l'accès aux ressources nécessaires pour la réalisation de ce projet.

Je suis également reconnaissant envers mes camarades et amis pour leurs encouragements et leur soutien moral.

Enfin, je remercie ma famille pour leur compréhension et leur patience durant cette période de travail intense.



Résumé

Ce rapport présente une étude sur la prédiction des températures à partir de données météorologiques, en se concentrant sur différentes villes américaines, notamment New York. Nous avons appliqué des techniques de modélisation telles que la régression linéaire, les forêts aléatoires, XGBoost et les réseaux de neurones sur des données téléchargées via Kaggle.

L'objectif est d'évaluer l'efficacité de ces méthodes pour la prévision des températures, tout en soulignant l'importance des données de qualité. Les résultats révèlent des différences significatives dans la performance des modèles, ce qui met en évidence la nécessité de choisir la méthode de prédiction en fonction des caractéristiques spécifiques des données.

Cette étude contribue à la compréhension des dynamiques météorologiques et propose des perspectives pour améliorer les systèmes de prévision en intégrant des approches basées sur l'intelligence artificielle.

Table des matières

Dédicaces	i
Remerciements	ii
Résumé	iii
Liste des Abréviations	vii
Liste des figures	viii
Liste des tableaux	x
Introduction générale	1
1 Présentation de l'organisme d'accueil (ENIT-LAMSIN)	2
1.1 Introduction	2
1.2 Aperçu sur ENIT-LAMSIN	2
1.2.1 Objectifs et missions	2
1.2.2 Historique	3
1.3 Les Activités de Recherche du LAMSIN	3
1.3.1 Problèmes Inverses	3
1.3.2 Analyse Mathématique et Numérique de la Propagation des Ondes	3
1.3.3 Modélisation Stochastique et Applications	4
1.3.4 Mathématiques pour la Biologie et l'Environnement	4
1.3.5 Images, Modélisation et Géométrie (IMoGe)	4
1.3.6 Histoire et Épistémologie des Mathématiques	4
1.4 Les réalisations les plus significatives	4
1.5 Conclusion	5
2 Les systèmes dynamiques	6
2.1 Introduction	6
2.2 Généralités sur les systèmes dynamiques	6
2.2.1 Système dynamique	6
2.2.2 Exemples de systèmes dynamiques	7
2.2.3 Bref historique	10
2.2.3.1 L'analytique triomphant (jusqu'à la fin du XIX ^{ème} siècle) .	10
2.2.3.2 Premiers doutes (fin XIX ^{ème} - début XX ^{ème} siècle)	11
2.2.3.3 Avènement des concepts fondamentaux (1920 - 1970)	11
2.2.3.4 Explosion du chaos (depuis 1970)	12
2.2.4 Aspect mathématique des systèmes dynamiques	12
2.2.4.1 Système non linéaire et loi de comportement	14

2.2.4.2	Espace des phases et trajectoire	14
2.2.4.3	Portrait de phase	14
2.2.4.4	Attracteurs	16
2.3	Bifurcations	17
2.4	Chaos	18
2.4.1	Définition	18
2.4.2	Attracteur de Lorenz	18
2.5	Application en météorologie	20
2.5.1	Atmosphère	20
2.5.2	Les modèles météorologiques	20
2.5.2.1	Les bases de la modélisation	21
2.5.2.2	La prévision ensembliste et la prévision déterministe	22
2.5.2.3	Les principaux modèles météorologiques	23
2.6	Conclusion	24
3	Revue de littérature	25
3.1	Introduction	25
3.2	Aperçu des études incluses	25
3.3	Études basées sur les systèmes dynamiques	27
3.3.1	Modèles des systèmes dynamiques	27
3.3.2	Résultats des études basées sur les systèmes dynamiques	27
3.3.3	Limitations et orientations futures	27
3.4	Études utilisant ML/DL	27
3.4.1	Modèles de ML/DL	27
3.4.2	Résultats des études utilisant les modèles de ML/DL	27
3.4.3	Limitations et orientations futures	33
3.5	Conclusion	33
4	Outils et processus de l'apprentissage automatique	34
4.1	Introduction	34
4.2	Machine Learning	34
4.2.1	Définition	35
4.2.2	Les types de machine learning	35
4.2.2.1	L'apprentissage supervisé	36
4.2.2.2	L'apprentissage non supervisé	36
4.2.2.3	L'apprentissage par renforcement	37
4.3	Aperçu sur les algorithmes de machine learning	39
4.3.1	La régression linéaire	39
4.3.1.1	Définition	39
4.3.1.2	Formulation mathématique	39
4.3.2	Forêt aléatoire	40
4.3.3	XGBoost	40
4.3.4	Les réseaux de neurones artificiels (ANN)	43
4.3.4.1	Définition	43
4.3.4.2	Fonctionnement d'un réseau de neurones artificiels	45
4.3.5	Réseau Neuronal Récurrent (RNN) basé sur des séries temporelles	46
4.4	Indicateurs clés de performance (KPI)	47
4.4.1	Matrice de confusion	48

4.4.2	Exactitude (Accuracy)	48
4.4.3	Précision	49
4.4.4	Rappel	49
4.4.5	F1-Score	49
4.4.6	Spécificité	49
4.4.7	Erreur quadratique moyenne (MSE)	49
4.4.8	Erreur absolue moyenne (MAE)	50
4.4.9	Coefficient de détermination (R ²)	50
4.4.10	Erreur moyenne absolue en pourcentage (MAPE))	50
4.5	Logiciels et langages de programmation	51
4.5.1	Langages de programmation	51
4.5.1.1	Python	51
4.5.1.2	R	51
4.5.2	Logiciels	52
4.5.2.1	Power BI	52
4.5.2.2	Jupyter Notebook	52
4.6	Conclusion	53
5	Exploration des données et entraînement des modèles	54
5.1	Introduction	54
5.2	À propos du jeu de données	54
5.3	Les attributs	54
5.4	Informations supplémentaires	55
5.5	Analyse descriptive	55
5.5.1	Résumé du jeu des données	56
5.5.2	Dashboard pour l'analyse du jeu de données	58
5.5.3	Les distributions des attributs	61
5.5.4	Analyse de corrélations entre les différents attributs	65
5.6	Entraînement des modèles	66
5.6.1	La régression linéaire	66
5.6.2	Forêt aléatoire	66
5.6.3	XGBoost	67
5.6.4	Les réseaux de neurones artificiels (ANN)	68
5.6.5	Réseau Neuronal Récurrent (RNN)	69
5.6.6	Comparaison des modèles	69
5.7	Conclusion	71
Conclusion générale		72
Bibliographie		73

Liste des abréviations

- **LAMSIN** : Laboratoire de Modélisation Mathématique et Numérique en Sciences de l'Ingénieur.
- **FPU** : Fermi-Pasta-Ulam.
- **MIT** : Massachusetts Institute of Technology.
- **TU** : Temps Universel.
- **ML** : Machine Learning.
- **DL** : Deep Learning.
- **IA** : Intelligence Artificielle.
- **ANN** : Artificial Neural Networks.
- **KNN** : K-Nearest Neighbors.
- **DT** : Decision Tree.
- **MLP** : Multi-Layer Perceptron.
- **SVM** : Support Vector Machine.
- **LSTM** : Long Short-Term Memory.
- **RFR** : Random Forest Regression.
- **SVR** : Support Vector Regression.
- **XGBoost** : Extreme Gradient Boosting.
- **LASSO** : Least Absolute Shrinkage and Selection Operator.
- **GBR** : Gradient Boosting Regression.
- **MSE** : Mean Square Error.
- **RMSE** : Root Mean Square Error.
- **MAE** : Mean Absolute Error.
- **KPI** : Key Performance Indicator.
- **NOAA** : National Oceanic and Atmospheric Administration.
- **NMSA** : National Maritime Safety Association.

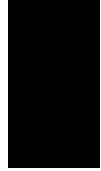
Liste des figures

1	<i>Logo ENIT-LAMSIN. [1]</i>	2
2	<i>Diagramme de décomposition.[4]</i>	7
3	<i>Pendule simple.[5]</i>	7
4	<i>Phénomène de turbulence.[6]</i>	9
5	<i>Cyclone.[7]</i>	9
6	<i>Système.[5]</i>	12
7	<i>Diagramme de décomposition des systèmes dynamiques.[5]</i>	13
8	<i>Portrait de phase d'un système dynamique d'ordre 2 non linéaire. Les points fixes (A, B, C), les orbites fermées (D), et leur stabilité, jouent un rôle clef dans la compréhension qualitative de la dynamique du système.[4]</i>	15
9	<i>Cycles limites d'un système de Van Der Pol.[4]</i>	16
10	<i>Exemple de bifurcation : flambage d'une poutre verticale lorsque la charge augmente. Le paramètre de contrôle est ici la masse et la variable dynamique la déflexion de la poutre.[10]</i>	17
11	<i>Une solution type dans l'attracteur de Lorenz lorsque $\rho = 28$, $\sigma = 10$, et $\beta = 8/3$.[16]</i>	19
12	<i>Atmosphère.[22]</i>	20
13	<i>Vue de la Terre découpée en une multitude de petites zones. Notons ici la grille horizontale et la grille verticale.[23]</i>	21
14	<i>Évaluation du modèle proposé avec jeu de données de test.[27]</i>	28
15	<i>Valeurs RMSE de quatre modèles d'apprentissage automatique existants et du modèle proposé.[27]</i>	29
16	<i>Température réelle par rapport à la température prédite par SVM.[29]</i>	31
17	<i>Température réelle par rapport à la température prévue par ANN.[29]</i>	32
18	<i>Température réelle vs température prédite par RNN.[29]</i>	32
19	<i>Relation entre l'intelligence artificielle (IA), l'apprentissage automatique (AA) et l'apprentissage profond (AP).[30]</i>	34
20	<i>Algorithmes d'apprentissage automatique et cas d'utilisation.[31]</i>	35
21	<i>Un algorithme d'apprentissage supervisé typique.[32]</i>	36
22	<i>Un algorithme d'apprentissage non supervisé.[33]</i>	37
23	<i>Cadre de l'apprentissage par renforcement.[34]</i>	38
24	<i>Régression linéaire simple.[36]</i>	39
25	<i>Fonctionnement de l'algorithme Random Forest.[37]</i>	40
26	<i>Apprentissage supervisé.[38]</i>	41
27	<i>Arbre de décision.[38]</i>	41
28	<i>Ensemble d'arbres de décision.[38]</i>	42
29	<i>Réseau de neurones biologiques.[39]</i>	43
30	<i>Réseau de neurones artificiels.[39]</i>	43
31	<i>L'architecture d'un réseau de neurones artificiel.[39]</i>	44

32	<i>Fonctionnement d'un réseau de neurones artificiels.[39]</i>	45
33	<i>Une seule couche récurrente avec deux entrées et quatre unités cachées.[40]</i>	46
34	<i>Schéma d'un RNN.[41]</i>	47
35	<i>Matrice de confusion.[42]</i>	48
36	<i>Le logo de Python.[47]</i>	51
37	<i>Le logo de R.[49]</i>	51
38	<i>Le logo de Power BI.[50]</i>	52
39	<i>Le logo de Jupyter.[51]</i>	52
40	<i>Décomposition de la variable Date_Time.</i>	57
41	<i>Les différentes valeurs de la variable Mois.</i>	57
42	<i>Les différentes villes incluses dans le jeu de données.</i>	58
43	<i>La température.</i>	59
44	<i>L'humidité.</i>	59
45	<i>La précipitation.</i>	60
46	<i>La vitesse du vent.</i>	60
47	<i>Distribution des attributs à New York.</i>	61
48	<i>Distribution des attributs à Los Angeles.</i>	61
49	<i>Distribution des attributs à Dallas.</i>	62
50	<i>Test de Kolmogorov Smirnov.</i>	62
51	<i>Region_analyser.</i>	63
52	<i>Feature_analyzer</i>	63
53	<i>Analyse des précipitations en Philadelphie.</i>	64
54	<i>Analyse de la température à New York.</i>	64
55	<i>Création du heatmap.</i>	65
56	<i>Heatmap de corrélations entre les différents attributs.</i>	65
57	<i>Linear Regression-KPI.</i>	66
58	<i>Droite de régression avec prédictions.</i>	66
59	<i>Random Forest-KPI.</i>	67
60	<i>Prédictions du modèle de forêt aléatoire.</i>	67
61	<i>XGBoost-KPI.</i>	67
62	<i>XGBoost.</i>	68
63	<i>Artificial Neural Networks-KPI.</i>	68
64	<i>Artificial Neural Networks.</i>	69
65	<i>Recurrent Neural Networks-KPI.</i>	69
66	<i>Root Mean Square Error (RMSE).</i>	70
67	<i>Mean Absolute Error (MAE).</i>	70

Liste des tableaux

1	<i>Résumé des études incluses.</i>	26
2	<i>Évaluation de la performance du modèle proposé.[27]</i>	28
3	<i>Comparaison de MAE et RMSE avec des techniques de référence utilisant les données météorologiques de l'Ouganda.[28]</i>	30
4	<i>Comparaison de l'EEM et de l'EQM avec des techniques utilisant les données météorologiques du Kenya et de la Tanzanie.[28]</i>	30
5	<i>Tableau d'exemple avec des valeurs provisoires.[29]</i>	31
6	<i>Comparaison entre les différents types d'apprentissage.[34]</i>	38
7	<i>Relation entre le réseau de neurones biologique et le réseau de neurones artificiel.[39]</i>	44
8	<i>Les premières lignes du jeu de données.</i>	55
9	<i>Statistiques descriptives.</i>	56
10	<i>La nouvelle base de données.</i>	57
11	<i>Les données spécifiques de Philadelphie.</i>	63



Introduction générale

La prévision météorologique est un enjeu de premier plan dans notre société, influençant divers secteurs tels que l'agriculture, le transport et la gestion des catastrophes naturelles. Historiquement, les prévisions météorologiques se sont appuyées sur des systèmes dynamiques et des simulations numériques, qui tentent de modéliser les interactions complexes entre les différents éléments de l'atmosphère. Ces méthodes reposent sur des équations différentielles représentant les lois physiques de la nature, offrant ainsi une base solide pour la compréhension des phénomènes météorologiques. Cependant, malgré leur robustesse, ces approches peuvent parfois s'avérer limitées en raison de la complexité intrinsèque de l'atmosphère et de la difficulté à capturer toutes les variables pertinentes.

Dans ce cadre, notre étude vise à explorer comment les techniques modernes de machine learning (ML) et de deep learning (DL) peuvent enrichir et améliorer les modèles de prévision existants. Ces méthodes, capables d'apprendre à partir de grandes quantités de données historiques, offrent la possibilité de découvrir des patterns et des relations non linéaires entre les différentes variables météorologiques, souvent difficiles à identifier avec les modèles traditionnels. En intégrant des modèles tels que les réseaux de neurones, les forêts aléatoires et les méthodes d'ensemble, nous avons cherché à évaluer leur performance comparative par rapport aux méthodes basées sur les systèmes dynamiques. Ce rapport présente les résultats de cette exploration, mettant en lumière l'impact potentiel des approches ML et DL sur l'avenir des prévisions météorologiques.

Présentation de l'organisme d'accueil (ENIT-LAMSIN)

1.1 Introduction

Dans un monde de plus en plus complexe, les sciences de l'ingénieur jouent un rôle déterminant dans le développement technologique et industriel. Le Laboratoire de Modélisation Mathématique et Numérique dans les Sciences de l'Ingénieur (LAMSIN) incarne cette dynamique en combinant les mathématiques appliquées et l'ingénierie pour aborder des défis technologiques variés. Basé à l'École nationale d'Ingénieurs de Tunis (ENIT), le LAMSIN est un centre d'excellence reconnu pour sa contribution à l'avancement de la recherche et de l'innovation. Au sein de ce laboratoire, j'ai eu l'opportunité d'effectuer mon stage d'ingénieur, ce qui m'a permis de m'immerger dans un environnement de recherche de pointe. Ce chapitre vise à présenter en détail le LAMSIN, en soulignant son rôle au sein de l'ENIT, son historique et ses réalisations les plus significatives.

1.2 Aperçu sur ENIT-LAMSIN

1.2.1 Objectifs et missions

Le LAMSIN est l'un des laboratoires les plus prestigieux de l'ENIT, dédié à la recherche et à la formation dans le domaine de la modélisation mathématique et numérique. Ce laboratoire se distingue par son approche interdisciplinaire, intégrant les mathématiques, l'informatique, et l'ingénierie pour développer des modèles et des méthodes capables de résoudre des problèmes complexes dans diverses applications industrielles et scientifiques. Les domaines d'intervention du LAMSIN couvrent un large spectre, incluant la dynamique des fluides, la mécanique des structures, l'optimisation et les systèmes dynamiques. [1]



Figure 1: Logo ENIT-LAMSIN. [1]

Le laboratoire joue également un rôle clé dans la formation des ingénieurs et des chercheurs en proposant des programmes de Master et de Doctorat, et en collaborant avec des industries pour des projets de recherche appliquée. Il est aussi impliqué dans des collaborations internationales, renforçant ainsi sa position en tant que centre de recherche de premier plan. [1]

1.2.2 Historique

Le Laboratoire de Mathématiques et de Modélisation Numérique (LAMSIN) a été constitué sous sa forme actuelle en 1999 et a été le premier laboratoire de mathématiques en Tunisie agréé par le ministère de tutelle. Depuis sa création, le LAMSIN a acquis une expérience originale, notamment grâce à la mise en place d'un groupe de recherche en modélisation mathématique et numérique, reconnu à l'international.

Cette reconnaissance est le fruit des efforts collectifs de tous les chercheurs du LAMSIN, soutenus par la bonne gouvernance de ses anciens directeurs : le Pr. Mohamed Jaoua, directeur fondateur, le Pr. Henda El Fekih et le Pr. Nabil Gmati.[2]

Depuis 2016, le LAMSIN s'est structuré autour de six équipes de recherche, chacune se concentrant sur des problématiques appliquées spécifiques. Une part significative des efforts déployés ces dernières années a été consacrée à l'organisation du laboratoire, visant à garantir sa pérennité et à renforcer son rayonnement tant au niveau national qu'international.[2]

1.3 Les Activités de Recherche du LAMSIN

Le LAMSIN (Laboratoire de Modélisation Mathématique et Numérique en Sciences de l'Ingénieur) est structuré autour de projets de recherche couvrant plusieurs domaines des mathématiques appliquées. Ces activités sont dirigées par des enseignants-chercheurs et exécutées par des équipes de recherche spécialisées. Les recherches du LAMSIN s'articulent autour de plusieurs thématiques principales, chacune ayant un impact significatif dans différents secteurs scientifiques et industriels.

1.3.1 Problèmes Inverses

L'équipe dédiée aux problèmes inverses se concentre sur l'étude théorique et le développement de méthodes numériques pour résoudre ces problèmes, souvent formulés en termes d'optimisation. Les applications concernent des domaines variés tels que la mécanique, l'hydrogéologie, et plus récemment, les sciences du vivant. Les techniques développées au LAMSIN, comme la paramétrisation adaptative pour l'estimation de paramètres, sont devenues des références dans la communauté scientifique.

1.3.2 Analyse Mathématique et Numérique de la Propagation des Ondes

Cette équipe s'intéresse aux systèmes gouvernés par des équations aux dérivées partielles, comme l'équation des ondes et celle de Schrödinger. Les recherches portent sur le

contrôle exact, les problèmes inverses, ainsi que sur les applications en ingénierie et physique, notamment la mécanique des fluides et la propagation des ondes acoustiques et électromagnétiques.

1.3.3 Modélisation Stochastique et Applications

Les recherches dans ce domaine visent à modéliser des phénomènes incertains, en particulier dans la finance, l'assurance, l'économie et la biologie. L'équipe explore des techniques de contrôle stochastique et des méthodes numériques avancées, notamment les simulations Monte Carlo et les équations différentielles stochastiques, pour résoudre des problèmes complexes liés à des systèmes dynamiques.

1.3.4 Mathématiques pour la Biologie et l'Environnement

Les préoccupations environnementales ont ouvert de nouveaux axes de recherche au LAMSIN. L'équipe s'intéresse à la modélisation et à la simulation des phénomènes liés à la pollution de l'eau et de l'air, à la gestion des ressources en eau, ainsi qu'aux énergies renouvelables. Les systèmes dynamiques sont également utilisés pour modéliser des bioprocédés dans des domaines comme l'écologie et la gestion des écosystèmes.

1.3.5 Images, Modélisation et Géométrie (IMoGe)

Cette équipe explore des problèmes issus des sciences de l'ingénieur, de la biologie et de l'industrie, en combinant la géométrie, l'analyse mathématique et les méthodes numériques. Les recherches actuelles incluent l'analyse d'images, les algorithmes pour les matrices structurées, et les méthodes d'apprentissage appliquées à l'imagerie médicale et à la modélisation géométrique.

1.3.6 Histoire et Épistémologie des Mathématiques

En parallèle aux recherches appliquées, le LAMSIN s'intéresse également à l'histoire des mathématiques, notamment à travers la reconstitution de traditions mathématiques anciennes, telles que celles liées à l'algèbre et à la géométrie dans le monde islamique médiéval. Ces études permettent une meilleure compréhension de l'évolution des concepts mathématiques.

1.4 Les réalisations les plus significatives

Le prix présidentiel du meilleur laboratoire de recherche scientifique pour l'année 2020, a été remis au laboratoire de Modélisation Mathématique et Numérique en Sciences de l'Ingénieur (LAMSIN), de l'École nationale d'Ingénieurs de Tunis (ENIT) à l'Université de Tunis El Manar (UTM), dirigé depuis 2016 par le Professeur Mourad Bellassoued.[\[2\]](#)

Le prix a été remis le lundi 04 juillet 2022, par le ministre de l'Enseignement supérieur et de la Recherche scientifique, Professeur Moncef Boukthir, en présence de Samia Charfi Kaddour, conseillère auprès de la cheffe du gouvernement, et Helmi Merdassi, directeur général par intérim de la DGRS. « Cette distinction récompense ainsi un travail de

longue haleine de toute une équipe sur plus de 25 ans », note un communiqué émis par le laboratoire.[\[2\]](#)

1.5 Conclusion

Le LAMSIN est un acteur clé de la recherche et de l'innovation à l'ENIT, réunissant des chercheurs autour de problématiques complexes grâce à une approche interdisciplinaire. Ses contributions notables, reconnues par le prix présidentiel du meilleur laboratoire en 2020, illustrent son engagement envers l'avancement des connaissances. En formant la nouvelle génération d'ingénieurs et en collaborant avec des industries, le LAMSIN renforce sa position de centre d'excellence dans le domaine de la modélisation mathématique et numérique, promettant un avenir brillant.

Les systèmes dynamiques

2.1 Introduction

Dans ce chapitre, nous abordons le contexte général de notre projet, qui vise à améliorer les techniques de prévision météorologique en combinant des modèles de systèmes dynamiques. Nous explorerons les fondements théoriques des systèmes dynamiques et leur application à la météorologie, ainsi que les innovations apportées dans ce domaine. Ce contexte servira de cadre pour comprendre les défis actuels, les opportunités et les motivations derrière notre démarche de recherche.

2.2 Généralités sur les systèmes dynamiques

2.2.1 Système dynamique

Les systèmes dynamiques sont des modèles mathématiques utilisés pour décrire l'évolution temporelle d'un système basé sur un ensemble de variables d'état. En physique, un système dynamique est souvent décrit par des équations différentielles qui représentent des lois fondamentales, telles que les lois de Newton en mécanique classique, les équations de Maxwell en électromagnétisme, ou les équations de la mécanique quantique. Ces équations définissent comment les variables d'état, telles que la position, la vitesse, l'énergie, la pression et la température, évoluent dans le temps sous l'influence de forces ou d'interactions internes et externes.^[3]

Un exemple notable de système dynamique en physique est le modèle de dynamique des fluides, où les équations de Navier-Stokes jouent un rôle central. Ces équations décrivent le mouvement des fluides, qu'il s'agisse de liquides ou de gaz, et sont essentielles pour comprendre des phénomènes complexes tels que les turbulences atmosphériques ou l'écoulement des océans. En météorologie, ces équations sont adaptées pour inclure des processus thermodynamiques et d'humidité, permettant ainsi de modéliser des phénomènes météorologiques comme les cyclones et les fronts météorologiques.^[3]

Les systèmes dynamiques peuvent être classés comme suit (figure 2) :

- **Déterministes** : L'évolution du système est entièrement déterminée par les conditions initiales et les lois de l'évolution. Par exemple, les équations de mouvement de la mécanique classique.^[4]
- **Non déterministe** : L'évolution contient des éléments aléatoires, rendant impossible la prévision exacte des états futurs à partir des seules conditions initiales. Par exemple, certains modèles de turbulence en dynamique des fluides.^[4]

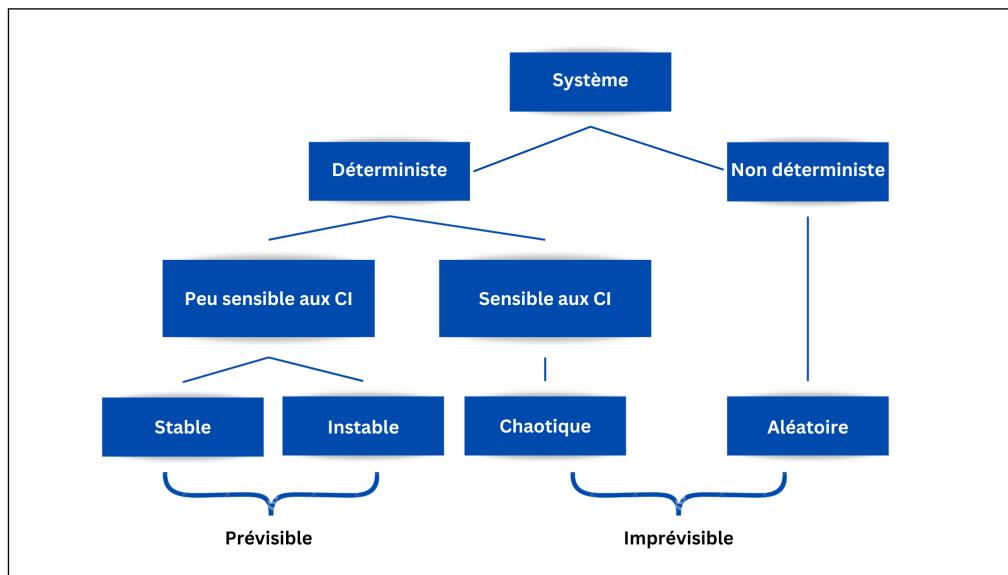


Figure 2: Diagramme de décomposition.[4]

Un système dynamique est un système qui évolue dans le temps de manière causale et déterministe.

2.2.2 Exemples de systèmes dynamiques

Mécanique Classique

Un pendule simple (figure 3), où les variables d'état sont l'angle de déviation et la vitesse angulaire, évolue selon les lois du mouvement, telles que la deuxième loi de Newton.



Figure 3: Pendule simple.[5]

Afin d'écrire les équations du mouvement, il est nécessaire d'identifier les forces agissant sur la masse. Tout d'abord, il y a la force gravitationnelle donnée par mg où g est l'accélération de la gravité. On suppose de plus que la masse est soumise à une force de résistance de friction proportionnelle à la vitesse de la masse et de coefficient de friction k . En appliquant le premier principe de la dynamique par projection sur l'axe tangentiel, on obtient l'équation différentielle du mouvement.

$$ml\ddot{\theta} = -ml \sin(\theta) - kl\dot{\theta} \quad (2.2.1)$$

À partir de ce modèle mathématique, il est possible de dériver un modèle dans l'espace d'état non linéaire en choisissant les variables d'état θ et $x_2=\dot{\theta}$.

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = -\frac{g}{l} \sin(x_1) - \frac{k}{m}x_2 \end{cases} \quad (2.2.2)$$

Dynamique des Fluides

Les équations de Navier-Stokes pour un fluide incompressible peuvent être écrites comme suit :

Équation de Navier-Stokes

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} = -\frac{1}{\rho} \nabla p + \nu \nabla^2 \mathbf{u} \quad (2.2.3)$$

où :

- $\mathbf{u} = (u, v, w)$ est le vecteur vitesse.
- ∇ est l'opérateur nabla (ou gradient).
- ∇^2 est le Laplacien.
- p est la pression du fluide.
- ρ est la densité du fluide.
- ν est la viscosité cinématique du fluide.

L'équation de continuité pour un fluide incompressible s'écrit :

Équation de Continuité

$$\nabla \cdot \mathbf{u} = 0 \quad (2.2.4)$$

où :

- \mathbf{u} est le vecteur de vitesse (u, v, w) .

Pour un fluide compressible, l'équation de continuité peut être formulée comme suit :

Équation de Continuité (Compressible)

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0 \quad (2.2.5)$$

où :

- ρ est la densité du fluide.

Ces équations sont fondamentales pour comprendre des phénomènes tels que les turbulences dans l'air (figure 4) ou l'écoulement de l'eau.



Figure 4: Phénomène de turbulence.[6]

Météorologie

Les systèmes météorologiques, tels que les cyclones (figure 5) et les fronts, sont décrits par des modèles dynamiques basés sur les équations de Navier-Stokes et d'autres équations de dynamique des fluides, adaptées pour inclure des processus thermodynamiques et d'humidité.

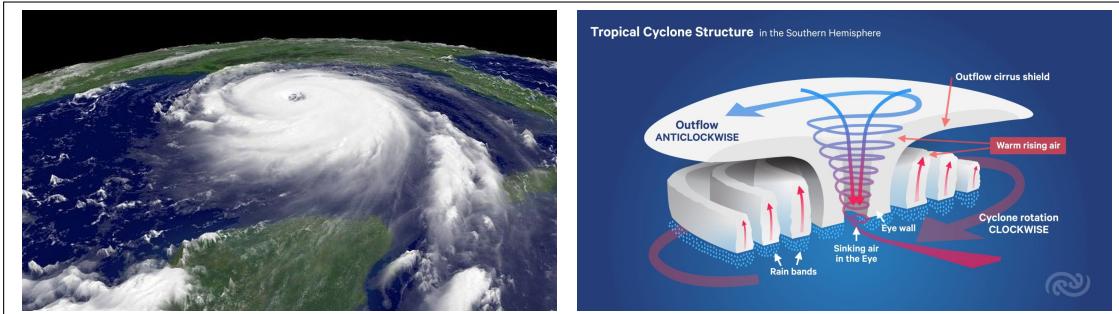


Figure 5: Cyclone.[7]

Les modèles météorologiques modernes utilisent plusieurs équations d'évolution de l'atmosphère[7]:

- **Équation du mouvement :**

$$m \frac{d^2 \mathbf{x}}{dt^2} = \mathbf{F} \quad (2.2.6)$$

- m : Masse de la particule ou de l'élément d'air considéré.
- $\frac{d^2 X}{dt^2}$: Accélération de la particule, où X est le vecteur position et t le temps.
- F : Force nette agissant sur la particule, résultant des forces telles que la gravité, la force de Coriolis, la force de pression, etc.

- **Équation de conservation de la masse totale :**

$$\frac{dM}{dt} + \nabla \cdot \mathbf{J} = 0 \quad (2.2.7)$$

- M : Masse totale de l'air dans un volume donné.
- $\nabla \cdot \mathbf{J}$: Divergence du flux de masse (\mathbf{J}) dans le volume, représentant la variation de masse entrant ou sortant du volume.
- \mathbf{J} : Vecteur densité de flux de masse, qui décrit comment la masse se déplace dans l'espace.

- **Équation d'état des gaz parfaits :**

$$PV = nRT \quad (2.2.8)$$

- P : Pression du gaz.
- V : Volume occupé par le gaz.
- n : Nombre de moles de gaz.
- R : Constante des gaz parfaits ($R = 8.314 \text{ J} \cdot \text{mol}^{-1} \cdot \text{K}^{-1}$).
- T : Température absolue du gaz en Kelvin.

- **Équation de la thermodynamique :**

$$dU = \delta Q - \delta W \quad (2.2.9)$$

- U : Énergie interne du système.
- δQ : Quantité de chaleur ajoutée au système.
- δW : Travail effectué par le système.

Les systèmes dynamiques sont utilisés pour comprendre et prédire une variété de phénomènes naturels, des mouvements planétaires à la dynamique des fluides atmosphériques, en passant par les oscillations électroniques et les systèmes chaotiques.

2.2.3 Bref historique

2.2.3.1 L'analytique triomphant (jusqu'à la fin du XIX^{ème} siècle)

Depuis l'Antiquité, la description du mouvement des astres se fait en termes de trajectoire (orbites) parfaitement régulières (cercles). Copernic (1473-1543) reprend ces idées pour le système solaire... mais le soleil est au centre, ce qui provoque un débat religieux. Kepler arrive et transforme les cercles en ellipses en utilisant les relevés de Tycho Brahe. C'est encore un **mouvement régulier et périodique**. Newton (1643-1727) développant l'intuition de Galilée (1564-1642) fonde la Science moderne. Le principe fondamental de la dynamique (PFD) conduit à l'évolution **déterministe** d'un système mécanique dans un champ de forces donné, pour des conditions initiales données. Le mouvement est régi par ces **conditions initiales**.

Le PFD avec l'attraction universelle conduit aux ellipses. Euler, Lagrange, Jacobi développent la mécanique analytique (Lagrangienne, Hamiltonienne...) ainsi que les

méthodes d'intégration qui donnent accès aux trajectoires : on cherche des solutions analytiques (intégrales premières) donnant des mouvements réguliers dans l'espace des phases. Grand succès de la mesure de la période du pendule, prédition de la position des planètes, etc. Kepler auparavant avait observé les mouvements autour du soleil. Newton explique et calcule. Cet âge d'or conduit à "la découverte de Neptune du bout de la plume" comme le dit joliment Arago après la prédition de Le Verrier en 1846. Comme souvent en Sciences, on retrouve Galilée et Newton, deux génies qui ont établi l'une des plus grandes idées de la science : le **déterminisme**. Si je sais où je suis aujourd'hui, je suis en principe capable de savoir où je serai demain. C'est incroyable quand on y pense et beaucoup de philosophes en doute encore aujourd'hui. Est-ce que le présent déterminerait l'avenir? Et pourtant ce déterminisme est tout le cœur de la Science. Il n'y a pas de Science sans déterminisme, même si ce concept a beaucoup changé au 20^{ème} siècle à cause de la théorie quantique.^[8]

2.2.3.2 Premiers doutes (fin XIX^{ème} - début XX^{ème} siècle)

Le 19^{ème} siècle était extrêmement optimiste, car les scientifiques vont résoudre tous les problèmes du monde, mais vers la fin les choses se sont compliquées lorsque l'on a étudié le mouvement de 8 ou 9 planètes autour du soleil. La machine de Newton commence à patiner et on ne savait plus très bien comment résoudre toutes ces équations différentielles. On s'aperçoit après beaucoup d'efforts que le **problème à trois corps** n'est pas soluble, au grand désespoir de certains. Burns, Poincaré (1854-1912), Painlevé parviennent à des théorèmes sur l'absence d'intégrales premières pour de très larges classes de systèmes hamiltoniens. Le problème de Kepler avec $1/r$ comme potentiel d'interaction est intégrable, mais il n'est pas du tout générique et conduit à de fausses idées sur les systèmes mécaniques en général. Poincaré développe des *méthodes géométriques* (topologie) pour analyser les propriétés qualitatives globales. Il comprend que des conditions initiales voisines peuvent conduire à des trajectoires rapidement très différentes, rendant la prédition à long terme impossible : c'est la notion de **chaos déterministe**. Pour comprendre une fonction, il est alors plus intéressant de comprendre quelques points caractéristiques : où est-elle maximum ? A-t-elle une asymptote ? Ensuite, au jugé, on trace la courbe, et on la comprend beaucoup mieux qu'avec un dessin précis. Au lieu de décrire la trajectoire de la lune par exemple, on essaie de dire si elle va rester dans le voisinage de la terre au temps long ! Ces travaux furent dans une grande mesure totalement ignorées par les physiciens pendant 50 ans, mais heureusement pas par les mathématiciens.^[8]

2.2.3.3 Avènement des concepts fondamentaux (1920 - 1970)

Étude de systèmes dynamiques variés :

- Oscillations non-linéaires en physique et applications (Van der Pol, 1927) : radar, radio, etc.
- Développement de nouvelles mathématiques (Van der Pol, Andronov, Smale, ...)
- Développement en parallèle, mais de façon largement déconnectée de méthodes géométriques proposées par Poincaré : Birkhoff, Kolmogorov, Arnold, Moser, etc.

Invention de l'ordinateur et développement de l'intuition à l'aide d'expériences numériques :

- 1955 FPU : modèle simplifié de la relaxation vers l'équilibre
- 1962 Hénon : Étude numérique en astronomie
- 1963 Lorenz : Étude numérique de la convection dans l'atmosphère. Lorenz a commencé une thèse de Maths avant la guerre, puis a été embauché dans le service de météorologie. À son retour, revenu météorologue au MIT, il a réfléchi toute sa vie (en mathématicien) au fonctionnement de l'atmosphère, pas en praticien pour faire des précisions, mais de manière théorique. Encore une fois, ces découvertes importantes n'eurent pas d'impact immédiat.[\[8\]](#)

2.2.3.4 Explosion du chaos (depuis 1970)

- 1971 Ruelle & Takens : nouvelle théorie pour le début de la turbulence.
- 1972 May : chaos en biologie. Exemples simples avec dynamique compliquée.
- 1976 Feigenbaum aux USA et Coullet en France. Concept d'universalité de la transition vers le chaos. Lien entre chaos et la théorie des transitions de phase qui venait de connaître son apogée.[\[8\]](#)

Remarques importantes :

- Les comportements génériques apparaissent déjà dans les systèmes avec peu de degrés de liberté.
- La présence de termes non-linéaires dans les équations engendrent souvent ces comportements et phénomènes, mais elle n'est pas suffisante (le problème de Kepler n'est pas linéaire), ni nécessaire (oscillateur paramétrique).

2.2.4 Aspect mathématique des systèmes dynamiques

Le système, agrégation d'éléments interconnectés, est constitué naturellement ou artificiellement afin d'accomplir une tâche prédéfinie. Son état est affecté par une ou plusieurs variables, les entrées du système (excitations). Le résultat de l'action des entrées est la réponse du système qui peut être caractérisée par le comportement d'une ou plusieurs variables de sorties.[\[5\]](#)

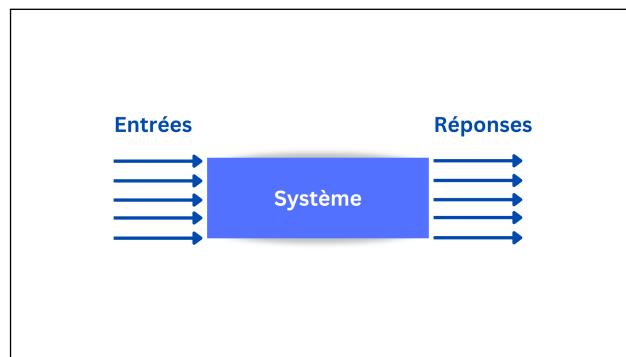


Figure 6: Système.[\[5\]](#)

L'ensemble des lois mathématiques régissant la causalité entre les entrées et les sorties du système constitue le modèle mathématique du système. La modélisation est l'étape préliminaire de l'analyse d'un système quelconque, indépendamment de sa nature physique, de sa composition et de son degré de complexité.[5]

Les systèmes dynamiques peuvent être divisés en deux grandes catégories : **systèmes discrets** et **systèmes continus**.

Systèmes discrets :

Dans les systèmes discrets, les variables d'état évoluent à des instants de temps discrets, souvent à des intervalles réguliers. Ces systèmes sont généralement modélisés par des équations aux différences. Un exemple courant est une suite définie par une relation de récurrence, comme le modèle de population de Fibonacci.[5]

Systèmes Continus :

Dans les systèmes continus, les variables d'état évoluent de manière continue dans le temps et sont décrites par des équations différentielles. Les systèmes continus peuvent être subdivisés en **systèmes linéaires** et **systèmes non linéaires** :

- **Systèmes Linéaires** : Les systèmes linéaires sont ceux où les équations différentielles sont linéaires par rapport aux variables d'état. Ils sont bien maîtrisés et de nombreuses techniques analytiques et numériques existent pour les résoudre. Cependant, les systèmes linéaires ont des limitations, notamment dans leur capacité à modéliser des phénomènes complexes du monde réel où les interactions sont souvent non linéaires.
- **Systèmes Non Linéaires** : Les systèmes non linéaires, en revanche, incluent des termes qui sont non linéaires par rapport aux variables d'état. Ces systèmes peuvent exhiber une riche variété de comportements dynamiques, y compris des oscillations périodiques, des bifurcations et le chaos. En raison de leur complexité, les systèmes non linéaires sont plus difficiles à analyser et à contrôler. Cependant, ils sont plus représentatifs des dynamiques observées dans de nombreux phénomènes naturels et industriels.[5]

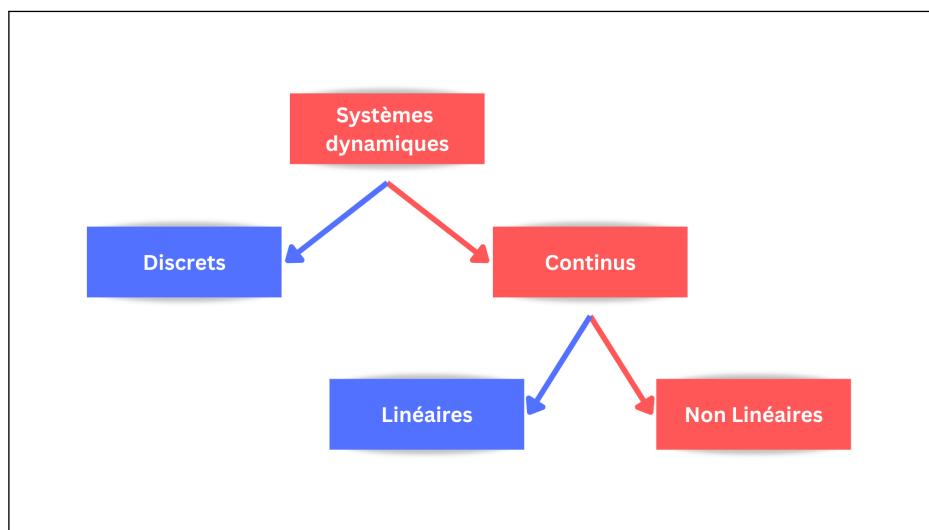


Figure 7: Diagramme de décomposition des systèmes dynamiques.[5]

Dans notre projet, nous nous concentrerons sur les systèmes non linéaires, car ils offrent une meilleure modélisation des phénomènes météorologiques complexes que nous cherchons à prévoir.

2.2.4.1 Système non linéaire et loi de comportement

Considérons un système dont l'état, à l'instant t , est représenté par n grandeurs scalaires indépendantes et réelles que l'on note $x_i(t)$ avec $i \in \{1, \dots, n\}$ et que l'on appelle **degrés de liberté** du système.

Par définition, l'évolution d'un système dynamique continu est régie par une ou plusieurs EDO du premier ordre. On parle de **loi d'évolution**. En général, ce système d'EDO est paramétré par les grandeurs réelles μ_i avec $i \in \{1, \dots, m\}$ [4]:

$$\begin{aligned} \frac{dx_1}{dt} &= f_1(x_1, \dots, x_n, \mu_1, \dots, \mu_n, t) \\ &\vdots && \text{soit} & \frac{d\vec{x}}{dt} &= f(\vec{x}, \vec{\mu}, t) \\ \frac{dx_n}{dt} &= f_n(x_1, \dots, x_n, \mu_1, \dots, \mu_n, t) \end{aligned} \tag{2.2.10}$$

où $\vec{x} = (x_1, \dots, x_n)$ est appelé vecteur d'état et $\vec{\mu} = (\mu_1, \dots, \mu_m)$ le vecteur des paramètres.

Le champ f n'est pas linéaire et lorsqu'il ne dépend pas explicitement du temps, on parle de système **autonome**. Dans le cas contraire, on parle de système **non-autonome**. On peut facilement ramener un système non-autonome à un système autonome en introduisant un nouveau degré de liberté $x_{n+1} = t$ régi par l'équation $\dot{x}_{n+1} = 1$. On se retrouve donc avec des systèmes autonomes, obéissant à l'équation d'évolution suivante, assortie de sa condition initiale :

2.2.4.2 Espace des phases et trajectoire

L'**espace des phases** est une représentation multidimensionnelle où chaque dimension correspond à une variable d'état du système. Par exemple, pour un système à deux variables d'état, l'espace des phases est un plan bidimensionnel. Chaque point dans cet espace représente un état possible du système à un instant donné.[4]

La **trajectoire** dans l'espace des phases est le chemin suivi par le point représentant l'état du système au cours du temps. Cette trajectoire est déterminée par les équations de la dynamique régissant le système. En observant ces trajectoires, on peut obtenir des informations sur la dynamique globale du système, telles que les comportements périodiques, quasi-périodiques ou chaotiques.[4]

2.2.4.3 Portrait de phase

Le portrait de phase est une représentation graphique des trajectoires dans l'espace des phases. Pour un système à deux dimensions, cela se traduit par un diagramme où chaque courbe représente l'évolution temporelle des variables d'état pour une condition

initiale donnée. Les points d'équilibre apparaissent comme des points où les trajectoires se stabilisent. Ces portraits permettent d'identifier les attracteurs, les répulseurs, et les cycles limites, facilitant l'analyse de la stabilité et du comportement global du système.[\[4\]](#)

Considérons un système dynamique de la forme :

$$\begin{cases} \dot{x}_1 = f_1(x_1, x_2) \\ \dot{x}_2 = f_2(x_1, x_2) \end{cases}$$

ou, de façon plus compacte :

$$\dot{x} = f(x)$$

où x est un point du plan de phase et \dot{x} représente la vitesse en ce point.

Dans le cas non linéaire, on ne peut généralement pas calculer les trajectoires analytiquement. De toute façon, quand cela est possible, l'expression peut être trop compliquée pour avoir une intuition du comportement du système. Nous allons donc essayer de trouver les caractéristiques qualitatives générales à partir d'une représentation graphique. La figure 8 donne un exemple de portrait de phase d'un système d'ordre 2 quelconque. Elle souligne l'importance de quelques points clefs dans la compréhension de la dynamique du système.[\[4\]](#)

La procédure générique est la suivante :

- On calcule les points fixes x^* , vérifiant $f(x^*) = 0$ (points A, B, C dans la figure 8) : ceux-ci sont des états d'équilibre stationnaires du système dynamique.
- On étudie la stabilité de ces points fixes, et on en déduit l'allure des trajectoires proches de ceux-ci. Dans l'exemple, les trajectoires sont très similaires autour des points A et C, mais le point B est qualitativement très différent, même si tous les trois sont instables.
- Ensuite, on détermine les orbites fermées, qui représentent des solutions périodiques $x(t+T) = x(t)$ (D dans notre exemple).
- De même que pour les points fixes, on étudie la stabilité de ces orbites. Le cycle limite D est par exemple stable dans la figure 8.

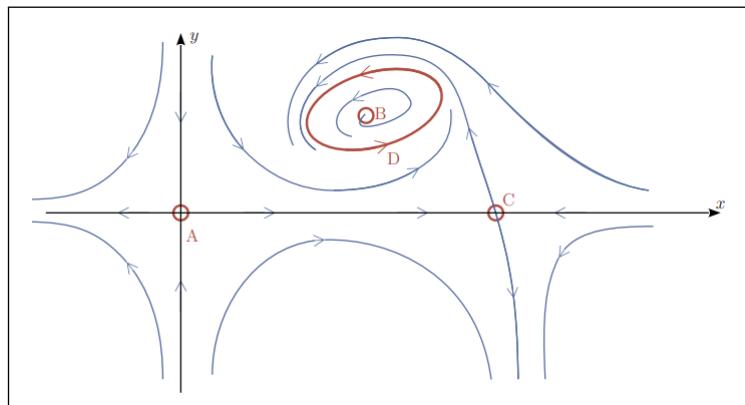


Figure 8: Portrait de phase d'un système dynamique d'ordre 2 non linéaire. Les points fixes (A, B, C), les orbites fermées (D), et leur stabilité, jouent un rôle clef dans la compréhension qualitative de la dynamique du système.[\[4\]](#)

2.2.4.4 Attracteurs

Les attracteurs sont des ensembles de points dans l'espace des phases vers lesquels les trajectoires du système convergent après un long temps. Ils représentent des comportements asymptotiques stables. Il existe plusieurs types d'attracteurs [4] :

- **Point fixe** : Un état stable où le système reste immobile. C'est donc un point \vec{x}^{eq} de l'espace des phases tel que [4] :

$$\left(\frac{d\vec{x}}{dt} \right)_{\vec{x}^{\text{eq}}} = \vec{0} \quad (2.2.11)$$

Par construction, les degrés de liberté d'un système dont l'état est représenté par un point fixe n'évoluent pas dans le temps ; le système est à l'équilibre. Compte tenu de la loi d'évolution 2.2.10, un point fixe \vec{x}^{eq} est tel que :

$$f(\vec{x}^{\text{eq}}, \vec{\mu}, t) = \vec{0} \quad (2.2.12)$$

Comme on le verra plus tard, cet état d'équilibre peut être stable (le système retourne à cet état après une perturbation) ou instable (il s'en éloigne).

- **Cycle limite** : Les cycles limites sont des objets intrinsèquement non-linéaires (ils ne peuvent pas exister dans des modèles linéaires). Ils montrent qu'un système dynamique peut avoir des oscillations périodiques sans forçage extérieur.[4]

Soit le système suivant représenté par l'équation de Van Der Pol :

$$m\ddot{x}(t) + 2c(x^2(t) - 1)\dot{x}(t) + kx(t) = 0, \quad c > 0 \quad (2.2.13)$$

où **m**, **c** et **k** sont des constantes liées au système physique.

La courbe fermée dans la figure 9 traduit un cycle limite : on retourne sur le même cycle, quelle que soit la condition initiale choisie.

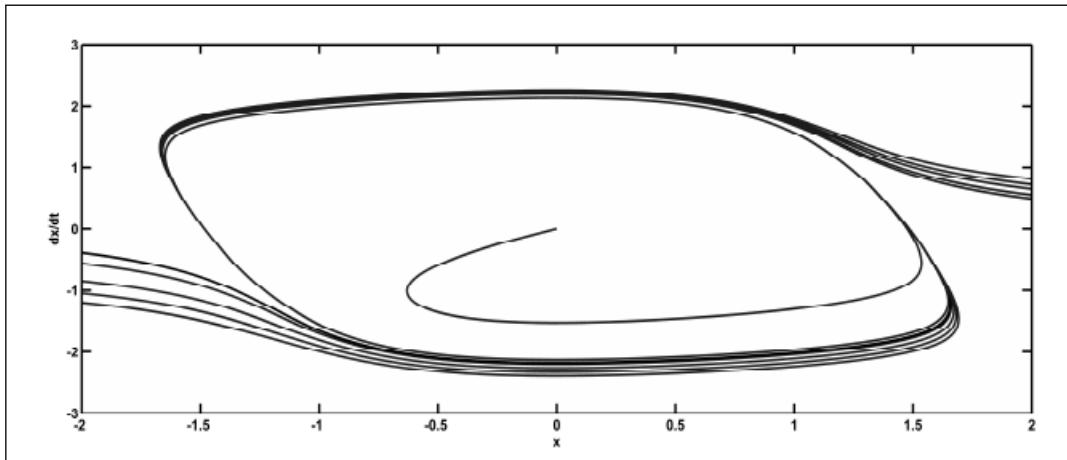


Figure 9: Cycles limites d'un système de Van Der Pol.[4]

- **Attracteur étrange :** Un attracteur est qualifié d'étrange s'il possède une structure fractale¹, c'est-à-dire s'il a une dimension d'Hausdorff² non entière. Cela est souvent le cas lorsque la dynamique qui s'y trouve est chaotique, mais des attracteurs étranges non chaotiques existent également. Si un attracteur étrange est chaotique, exhibant une sensibilité aux conditions initiales, alors deux points initiaux alternatifs, arbitrairement proches sur l'attracteur, après un certain nombre d'itérations, conduiront à des points qui sont arbitrairement éloignés (tout en restant dans les limites de l'attracteur), et après un autre nombre d'itérations, ils conduiront à des points arbitrairement proches. Ainsi, un système dynamique avec un attracteur chaotique est localement instable, mais globalement stable : une fois que certaines séquences ont pénétré l'attracteur, les points proches divergent les uns des autres, mais ne quittent jamais l'attracteur.[9]

2.3 Bifurcations

La dynamique semblait peu intéressante pour les systèmes du premier ordre, mais tout devient différent si le système dépend de paramètres externes : le système peut alors transiter entre des comportements qualitativement différents (en langage mathématique “topologiquement non équivalents”). Le système peut par exemple changer de nombre de points fixes, ou leur stabilité. Ces changements qualitatifs sont appelés bifurcations, les points correspondants points de bifurcation, et les paramètres : paramètres de contrôle. Les bifurcations sont parfois nommées transitions de phase dynamiques, par analogie avec la théorie des transitions de phase.[10]

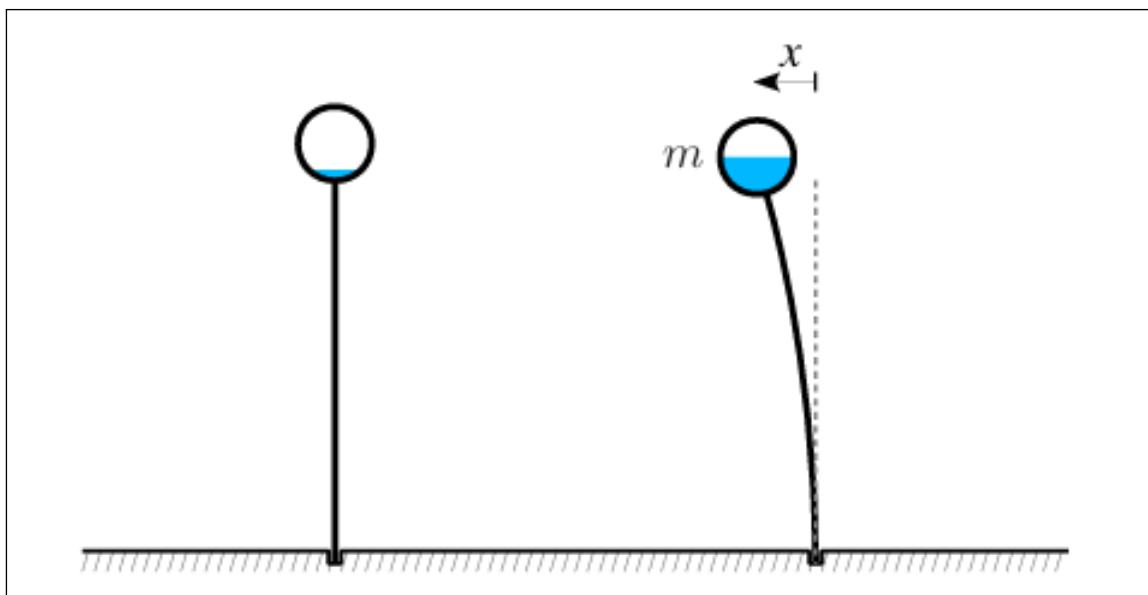


Figure 10: Exemple de bifurcation : flambage d'une poutre verticale lorsque la charge augmente. Le paramètre de contrôle est ici la masse et la variable dynamique la défexion de la poutre.[10]

¹Un fractal est une forme géométrique qui présente une structure détaillée à toutes les échelles, avec une dimension fractale souvent supérieure à sa dimension topologique.

²La dimension d'Hausdorff, introduite en 1918 par le mathématicien Felix Hausdorff, est une mesure de la rugosité ou de la dimension fractale d'un ensemble.

2.4 Chaos

2.4.1 Définition

En usage courant, le terme "chaos" signifie "un état de désordre".^[11]Cependant, dans la théorie du chaos, le terme est défini de manière plus précise. Bien qu'il n'existe pas de définition mathématique universellement acceptée du chaos, une définition couramment utilisée, formulée à l'origine par Robert L. Devaney, stipule qu'un système dynamique doit posséder les propriétés suivantes pour être classé comme chaotique [12]:

- Il doit être sensible aux conditions initiales,
- Il doit être topologiquement transitif³,
- Il doit avoir des orbites périodiques denses.

Dans certains cas, il a été démontré que les deux dernières propriétés ci-dessus impliquent en réalité la sensibilité aux conditions initiales.^{[13][14]} Dans le cas du temps discret, cela est vrai pour toutes les applications continues sur des espaces métriques.^[15]Dans ces cas, bien qu'elle soit souvent la propriété la plus significative en pratique, la "sensibilité aux conditions initiales" n'a pas besoin d'être incluse dans la définition.

2.4.2 Attracteur de Lorenz

Le système de Lorenz, étudié par Edward Lorenz, est un ensemble d'équations différentielles qui présente des solutions chaotiques pour certaines valeurs de paramètres et conditions initiales. L'attracteur de Lorenz, associé à l'effet papillon, illustre comment de petites variations initiales peuvent entraîner des évolutions imprévisibles. Bien que déterministes, ces systèmes chaotiques restent imprévisibles sur le long terme, car le chaos croît continuellement, rendant toute prédition impossible. La forme de l'attracteur de Lorenz ressemble à un papillon, symbolisant cette sensibilité aux conditions initiales.^[16]

En 1963, Edward Lorenz, avec l'aide d'Ellen Fetter, qui était responsable des simulations numériques et des figures,^[16] et de Margaret Hamilton, qui a contribué aux calculs numériques initiaux menant aux découvertes du modèle de Lorenz, ^[17]a développé un modèle mathématique simplifié pour la convection atmosphérique.^[16] Le modèle est un système de trois équations différentielles ordinaires, maintenant connues sous le nom d'équations de Lorenz :

$$\boxed{\frac{dx}{dt} = \sigma(y - x)}, \quad (2.4.1)$$

$$\boxed{\frac{dy}{dt} = x(\rho - z) - y}, \quad (2.4.2)$$

$$\boxed{\frac{dz}{dt} = xy - \beta z}. \quad (2.4.3)$$

³Un système dynamique est topologiquement transitif si, en gros, il est possible, à partir de n'importe quelle région non vide du système, d'atteindre n'importe quelle autre région non vide, après un certain temps.

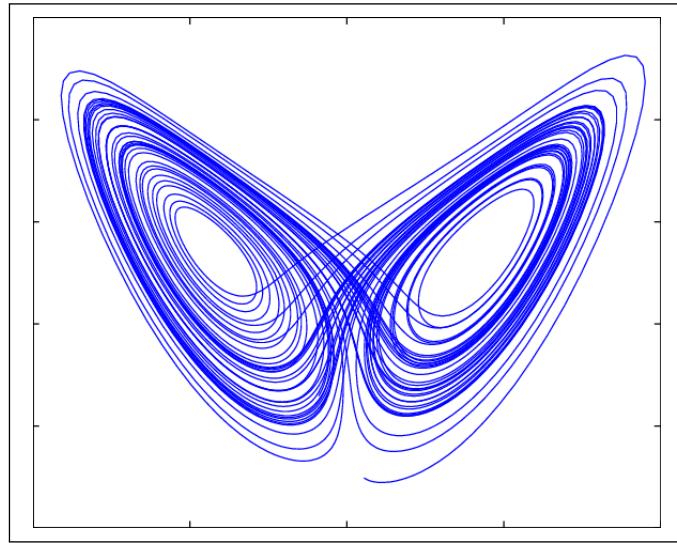


Figure 11: Une solution type dans l'attracteur de Lorenz lorsque $\rho = 28$, $\sigma = 10$, et $\beta = 8/3$. [16]

Ces équations décrivent les propriétés d'une couche de fluide bidimensionnelle uniformément chauffée par le bas et refroidie par le haut. Elles spécifient en particulier le taux de changement de trois quantités par rapport au temps : x est proportionnel au taux de convection, y à la variation de température horizontale et z à la variation de température verticale. [18] Les constantes σ , ρ et β sont des paramètres du système proportionnels au nombre de Prandtl⁴, au nombre de Rayleigh⁵, et à certaines dimensions physiques de la couche elle-même.[18]

Au cours des dernières années, une série d'articles sur les modèles de Lorenz en haute dimension ont conduit à un modèle de Lorenz généralisé[19], qui peut être simplifié en un modèle de Lorenz classique à trois variables d'état ou en le modèle de Lorenz à cinq dimensions suivant pour cinq variables d'état[20] :

$$\boxed{\frac{dx}{dt} = \sigma(y - x)}, \quad (2.4.4)$$

$$\boxed{\frac{dy}{dt} = x(\rho - z) - y}, \quad (2.4.5)$$

$$\boxed{\frac{dz}{dt} = xy - xy_1 - \beta z}, \quad (2.4.6)$$

$$\boxed{\frac{dy_1}{dt} = xz - 2xz_1 - d_0y_1}, \quad (2.4.7)$$

$$\boxed{\frac{dz_1}{dt} = 2xy_1 - 4\beta z_1}. \quad (2.4.8)$$

⁴Le nombre de Prandtl (Pr) est un nombre sans dimension, défini comme le rapport entre la diffusivité de la quantité de mouvement et la diffusivité thermique, et est nommé d'après le physicien allemand Ludwig Prandtl.

⁵Le nombre de Rayleigh (Ra) est un nombre sans dimension qui caractérise l'écoulement d'un fluide en convection naturelle, indiquant si l'écoulement est laminaire ou turbulent.

2.5 Application en météorologie

2.5.1 Atmosphère

L'atmosphère est la couche de gaz entourant une planète ou un autre corps céleste. Pour la Terre, l'atmosphère est principalement composée d'azote (78%) et d'oxygène (21%), avec des traces d'autres gaz tels que l'argon, le dioxyde de carbone et la vapeur d'eau. Elle s'étend de la surface de la planète jusqu'aux confins de l'espace, en s'amenuisant progressivement avec l'altitude.[\[21\]](#)

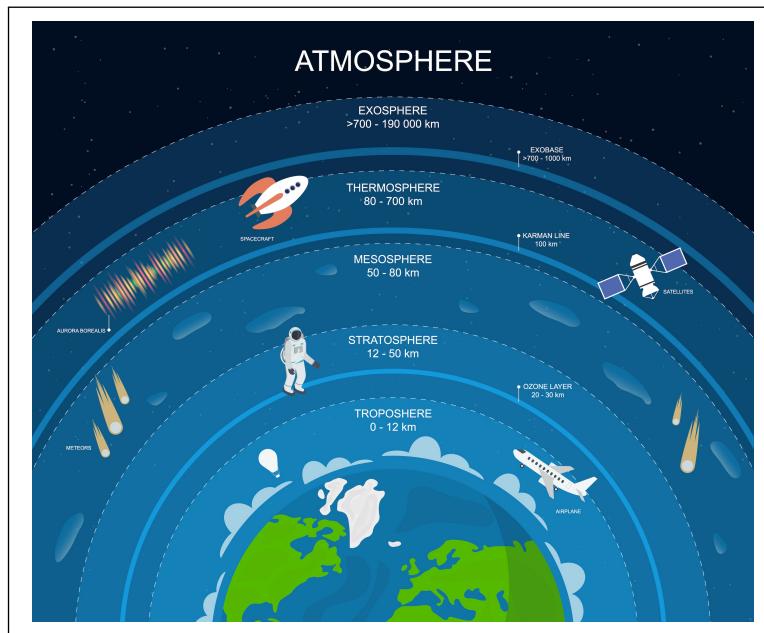


Figure 12: Atmosphère.[\[22\]](#)

L'atmosphère est divisée en plusieurs couches en fonction de la température et de la composition : la troposphère (où se produisent les phénomènes météorologiques), la stratosphère (contenant la couche d'ozone), la mésosphère, la thermosphère et l'exosphère. Elle joue un rôle crucial dans le soutien de la vie en fournissant de l'oxygène, en protégeant contre les radiations solaires nocives et en régulant la température grâce à l'effet de serre.[\[22\]](#)

L'atmosphère peut également être considérée comme un système dynamique, caractérisé par des changements constants et des interactions entre ses composants. Étudier son comportement implique de comprendre les interactions complexes entre la température, la pression, l'humidité et les courants d'air, ce qui est essentiel pour la prévision météorologique et la modélisation climatique.[\[22\]](#)

2.5.2 Les modèles météorologiques

La prévision météorologique repose sur des modèles de simulation de l'atmosphère, nécessaires pour prédire les conditions futures. Pour prévoir ces aspects, un modèle mathématique approximatif est créé à partir de lois d'évolution, simulant le comportement du système à différentes échéances. L'atmosphère, étant complexe, nécessite des ordinateurs pour réaliser les nombreuses opérations arithmétiques requises. Depuis les années 50, les

ordinateurs ont rendu possible la modélisation de l'atmosphère. Aujourd'hui, les modèles météorologiques informatiques, intégrant divers paramètres comme les températures, l'humidité et le vent, sont essentiels pour toute prévision, et les supercalculateurs créent des cartes de modélisation numérique à partir de ces données.[21]

2.5.2.1 Les bases de la modélisation

L'atmosphère est régie par des lois exprimées par des systèmes d'équations différentielles, dont les solutions exactes sont mathématiquement inaccessibles. Pour contourner cela, des solutions approchées sont obtenues via des modèles numériques. La numérisation de l'atmosphère consiste à discréteriser l'espace et le temps en un réseau de points (grille) pour simplifier les calculs. La précision des prévisions dépend de cette grille et de la condition initiale, qui détermine le scénario simulé. Une erreur dans cette condition, fausse les résultats de la modélisation.[23]

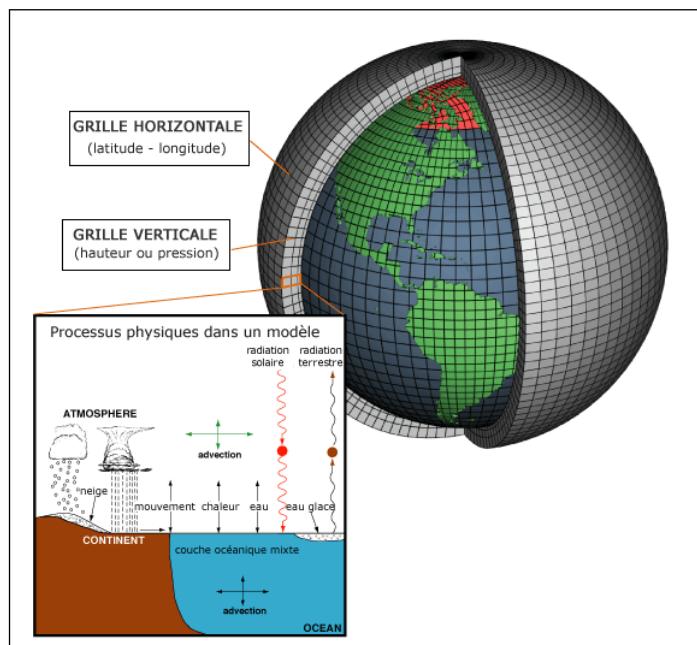


Figure 13: Vue de la Terre découpée en une multitude de petites zones. Notons ici la grille horizontale et la grille verticale.[23]

A. Processus d'assimilation des données

Cette première étape consiste à recueillir toutes les données d'observations (manuelles ou automatiques) issues des stations météorologiques terrestres et maritimes, des satellites, des radars de précipitations, des radiosondages, etc., qui parviennent au centre de calcul.

Dans la mesure où les observations météorologiques sont continues, varient sans cesse, il est nécessaire de faire tourner plusieurs fois par jour les modèles numériques afin d'avoir des scénarios toujours plus proches de la réalité. Il est ainsi possible de faire jouer « observations » et « ébauche » selon un cycle d'assimilation. À chaque heure d'observation synoptique principale (00 TU, 06 TU, 12 TU, 18 TU)⁶ on réalise l'analyse.[23]

⁶Les heures synoptiques principales (00 TU, 06 TU, 12 TU, 18 TU) sont des moments standardisés pour la collecte de données météorologiques, permettant une intégration cohérente dans les modèles numériques.

B. Résolution des équations et modélisation

Dans un second temps, on passe à la modélisation. Les modèles météorologiques modernes utilisent plusieurs équations d'évolution de l'atmosphère :

- Équation du mouvement
- Équation de conservation de la masse totale
- Équation d'état des gaz parfaits
- Équation de la thermodynamique

À partir de ces équations et de l'état initial décrit ci-dessus, le calculateur est chargé d'effectuer un ensemble d'opérations afin de modéliser à des échéances successives les paramètres évoluant dans une portion d'atmosphère (ou maille). L'avancée dans le temps se fait de façon itérative. En l'occurrence, le calculateur reprend toujours l'échéance précédente pour modéliser la prévision brute suivante. Ces échéances dépendent du pas de temps caractéristique du modèle et donc de sa maille. Sur les modèles de petites mailles, les données brutes peuvent être calculées toutes les heures (grandes précisions) tandis tandis que pour les grosses mailles, la prévision brute est calculée par pas de 24 heures.[\[23\]](#)

À l'issue de ces deux étapes, il y a création d'un run. Échéance après échéance, le modèle sort des données brutes qu'il est possible de retranscrire au moyen de l'outil informatique sur des cartes.

2.5.2.2 La prévision ensembliste et la prévision déterministe

La prévision météorologique utilise deux approches principales :

- **Prévision déterministe** : est utilisée pour des périodes allant de plusieurs heures à environ 3 ou 4 jours. Elle fournit une seule estimation basée sur les conditions initiales et les modèles de prévision.
- **Prévision d'ensemble** : concerne les prévisions au-delà de 3 ou 4 jours et génère plusieurs scénarios possibles. Chaque scénario est une variante de la prévision initiale, obtenue en introduisant de petites perturbations dans les conditions de départ.

Les erreurs dans le processus d'assimilation des données et l'incertitude croissante sont des défis majeurs en prévision météorologique. En introduisant des perturbations légères dans les conditions initiales, on peut générer différents scénarios possibles. Ces scénarios permettent de comparer les prévisions et d'évaluer leur fiabilité. Plus les scénarios sont similaires, plus la prévision est fiable ; inversement, plus les scénarios varient, plus la fiabilité diminue.[\[23\]](#) Grâce à ces scénarios, on peut aussi obtenir des informations statistiques, comme la probabilité de dépasser un certain seuil (par exemple, « il y a 40% de chance que les cumuls de pluie dépassent 15 mm le jeudi 20 septembre »).[\[23\]](#)

"TU" signifie Temps Universel.

2.5.2.3 Les principaux modèles météorologiques

- **Le modèle GFS (américain) :** Le modèle GFS (Global Forecast System) est produit par le National Centers for Environmental Prediction (NCEP). Cet organisme est un regroupement de plusieurs centres nationaux de prévisions météorologiques aux États-Unis. Il fait également partie du National Weather Service (NWS). Parmi les neuf centres nationaux, l'Environmental Modeling Center (EMC) développe particulièrement le modèle GFS.

Le Global Forecast System est initialisé quatre fois par jour : run 00z – run 06z – run 12z – run 18z. Les calculs de prévisions brutes vont jusqu'à 384h (16 jours). Sa résolution horizontale est de 27 km jusqu'à 192h et 70 km de 192 à 384 h. À noter que le GFS est un modèle libre et gratuit.[\[23\]](#)

- **Le modèle ECMWF - CEPMMT (européen) :** Le modèle CEPMMT (Centre européen pour les prévisions météorologiques à moyen terme) est un modèle utilisé pour la prévision allant jusqu'à 10 jours. Contrairement au modèle GFS, une grande partie des paramètres du modèle CEPMMT ne sont pas accessibles gratuitement. Le modèle CEPMMT est initialisé deux fois par jour : run 00z – run 12z.

- **Le modèle WRF (américain) :** Le modèle WRF (Weather Research and Forecasting) est un modèle météo utilisé par le National Weather Service des États-Unis et pour la recherche en simulation de l'atmosphère. C'est un modèle dit de méso-échelle avec une résolution horizontale entre 2 et 15 km. À noter que WRF est un modèle libre et gratuit.

Le Weather Research and Forecasting est initialisé quatre fois par jour : run 00z – run 06z – run 12z – run 18z.[\[23\]](#)

- **Le modèle ARPEGE (monde) :** Le modèle ARPEGE (Action de Recherche Petite Échelle Grande Échelle) est un modèle qui couvre l'ensemble de la planète avec une maille fluctuante selon les zones géographiques (7.5 km en moyenne pour l'Europe). L'échéance de la prévision est de quatre jours.

Le modèle ARPEGE est initialisé quatre fois par jour : run 00z – run 06z – run 12z – run 18z.[\[23\]](#)

- **Le modèle AROME (français) :** Le modèle AROME (Application of Research to Operations at MEsoscale) est un modèle avec une maille très fine (maille de 1.3 km) pour la prévision en France. L'échéance de la prévision est limitée à 36 heures. Ce modèle développé par la météo nationale en France appartient à la dernière génération de modèles. Grâce à sa maille très fine, il permet de mieux appréhender les phénomènes convectifs tels que les orages, et ce grâce à l'intégration de nouvelles données d'observation ou encore la prise en compte de la topographie, des villes, des cours d'eau, de la végétation, etc.

Le modèle AROME est initialisé quatre fois par jour : run 00z – run 06z – run 12z – run 18z.[\[23\]](#)

- **Le modèle CFS (américain) :** Le modèle CFS (Seasonal Climate Forecast) est un modèle saisonnier développé par le National Centers for Environmental Prediction (NCEP) et la NOAA (National Oceanic and Atmospheric Administration). Il prend en compte les situations du passé et les statistiques d'évolution, El Nino,

La Nina, l'Oscillation Nord Atlantique ou encore l'évolution des masses d'air des dernières semaines. Grâce à toutes ces données, les grandes tendances à six mois sont proposées.[\[23\]](#)

La prévision météorologique est confrontée à des défis majeurs en raison de la complexité et du caractère chaotique de l'atmosphère. Les systèmes dynamiques qui régissent les conditions météorologiques sont intrinsèquement non linéaires et sensibles aux conditions initiales, ce qui signifie que de petites erreurs dans les observations peuvent conduire à des divergences significatives dans les prévisions à long terme. Bien que les modèles numériques actuels tentent d'améliorer la précision des prévisions en utilisant des équations différentielles pour simuler l'évolution de l'atmosphère, les limites de la résolution numérique et les erreurs d'assimilation des données persistent. Les efforts pour perfectionner ces modèles ont conduit à l'utilisation de supercalculateurs pour traiter des grilles de plus en plus fines, mais les incertitudes restent importantes, surtout pour les prévisions à long terme.

Pour surmonter ces limitations, le monde de la météorologie commence à explorer des approches innovantes telles que le machine learning et le deep learning. Ces techniques offrent un potentiel prometteur pour développer des modèles plus robustes, capables de traiter des quantités massives de données et d'améliorer la fiabilité des prévisions. L'apprentissage automatique permet de détecter des patterns complexes et de mieux intégrer les données d'observation, tandis que l'apprentissage profond peut affiner la résolution des problèmes de prévision en exploitant les capacités de calcul avancées. En combinant ces nouvelles approches avec les méthodes traditionnelles, il est possible d'améliorer la visualisation des données météorologiques et de fournir des prévisions plus précises et fiables.

2.6 Conclusion

Dans ce chapitre, nous avons présenté les systèmes dynamiques et leur importance dans notre projet de prévision météorologique. Nous avons discuté des concepts fondamentaux et des innovations récentes qui permettent de modéliser des phénomènes complexes en météorologie. Nous avons également identifié les défis et les opportunités liés à l'application de ces modèles pour améliorer la précision des prévisions. Ce cadre théorique est essentiel pour notre recherche et ouvre la voie à une exploration des méthodes spécifiques dans les chapitres suivants, visant à enrichir notre compréhension des phénomènes météorologiques et à répondre aux besoins d'une prévision fiable.

Revue de littérature

3.1 Introduction

Les prévisions météorologiques sont cruciales pour de nombreux secteurs tels que l'agriculture, l'aviation et la gestion des catastrophes naturelles. Cette étude explore deux approches principales pour la prévision météorologique : les systèmes dynamiques et les techniques de machine learning (ML) et deep learning (DL). Nous comparerons l'efficacité de ces approches et discuterons leurs avantages, leurs limitations et les orientations futures.

3.2 Aperçu des études incluses

Une recherche systématique a été réalisée pour identifier les études consacrées à la prévision météorologique, classées en fonction des outils utilisés et des problématiques spécifiques abordées. Les études identifiées couvraient divers aspects de la prévision météorologique, y compris l'impact du changement climatique sur les précipitations à court terme, ainsi que les défis liés à la prédiction de paramètres tels que la température, l'humidité, les précipitations et la vitesse du vent. Il est à noter que les outils employés dans ces études incluaient principalement des approches statistiques, des algorithmes d'apprentissage automatique, et la théorie des systèmes dynamiques. Nous avons porté une attention particulière aux études se concentrant sur la prédiction des différents paramètres.

La majorité des études étaient de nature observationnelle et ont été publiées essentiellement après 2010. Les résultats indiquent que les algorithmes d'intelligence artificielle et d'apprentissage automatique (IA/ML) ont démontré des capacités prédictives supérieures aux modèles théoriques basés sur les systèmes dynamiques.

Cependant, l'efficacité des algorithmes IA/ML en prévision météorologique dépend encore fortement de l'accès à des bases de données volumineuses et de leur interprétation par des experts en climatologie. À l'avenir, l'intégration de l'IA pourrait améliorer significativement la précision des prévisions météorologiques, contribuant ainsi à une meilleure anticipation des catastrophes naturelles et des phénomènes météorologiques extrêmes, et par conséquent à l'amélioration de la qualité de vie.

Le tableau 1 présente un aperçu des études clés dans le domaine des prévisions météorologiques, en détaillant les méthodologies employées, les résultats obtenus et leurs domaines d'application spécifiques. Ces études couvrent un large éventail de techniques statistiques et d'IA/ML, telles que les modèles de régression linéaire, les réseaux de neurones, et les algorithmes de forêts aléatoires. Leur objectif commun est d'identifier des schémas prédictifs pertinents pour divers paramètres météorologiques. En utilisant des méthodologies computationnelles avancées, ces études visent à fournir des informations cruciales pour améliorer la précision

des prévisions météorologiques et affiner l'anticipation des phénomènes climatiques extrêmes.

Table 1: Résumé des études incluses.

Référence	Contribution	Techniques	Domaine d'application
ECMWF[24]	Amélioration du modèle IFS (Integrated Forecasting System) pour des prévisions précises à 10 jours	Modélisation numérique à échelle globale, Systèmes dynamiques	Prévisions à court et moyen terme, Gestion des risques naturels
Warner[25]	Introduction aux modèles de méso-échelle et leur application dans la météorologie moderne	Modèles de méso-échelle, Systèmes dynamiques	Prévisions météorologiques locales et régionales
Baboo et Shereef[26]	Une approche est capable de déterminer la relation non linéaire qui existe entre les données historiques (température, vitesse du vent, humidité, etc.)	ANN	Prédiction de la température
D. Endalie, G. Haile, and W. Taye[27]	Un modèle de prédiction des précipitations basé sur LSTM est proposé pour Jimma, dans l'ouest de l'Oromia, en Éthiopie	KNN, DT, MLP, SVM, LSTM	La prédiction des précipitations quotidiennes
Tumusiiime, Eyobu et Mugume[28]	Analyse rigoureuse des algorithmes d'apprentissage automatique pour la prédiction des précipitations à court terme	RFR, NNR, SVR, XGBoost, LASSO, GBR	Prédiction de quantités de précipitations à court terme
Singh, al.[29]	Prouver que l'utilisation des séries temporelles avec les RNN est une meilleure méthode pour la prévision météorologique	SVM, ANN, Time Series RNN	La prévision météorologique

3.3 Études basées sur les systèmes dynamiques

3.3.1 Modèles des systèmes dynamiques

Les systèmes dynamiques ont longtemps été utilisés pour la modélisation des phénomènes météorologiques. Ces modèles se basent sur des équations différentielles pour représenter les interactions complexes entre différentes variables atmosphériques. Les approches traditionnelles incluent les modèles de circulation générale (GCM), qui simulent les processus physiques de l'atmosphère, de l'océan et des terres.

3.3.2 Résultats des études basées sur les systèmes dynamiques

Les résultats obtenus à partir des modèles dynamiques montrent qu'ils sont capables de fournir des prévisions globales de qualité pour de longues périodes, mais leur précision diminue lorsqu'il s'agit de prévisions à court terme ou de phénomènes localisés comme les précipitations intenses. En outre, ces modèles nécessitent des ressources de calcul importantes et sont souvent limités par la qualité des données d'entrée.

3.3.3 Limitations et orientations futures

Les principales limitations des modèles dynamiques incluent la complexité inhérente des processus atmosphériques qu'ils cherchent à modéliser, ainsi que leur dépendance à des conditions initiales extrêmement précises. L'exécution de ces modèles nécessite des superordinateurs puissants, indispensables pour la recherche avancée. Cependant, les ressources limitées des superordinateurs posent un défi majeur, car elles restreignent la capacité des chercheurs à effectuer des simulations de grande envergure, ce qui peut entraver la précision et l'utilité des prévisions obtenues. À l'avenir, l'intégration de l'IA avec les modèles dynamiques pourrait offrir une amélioration significative de la précision des prévisions en exploitant les points forts des deux approches.

3.4 Études utilisant ML/DL

3.4.1 Modèles de ML/DL

Les méthodes employées dans les études sont variées et illustrent la diversité des approches d'analyse et de modélisation prédictive appliquées à la prévision météorologique avec l'IA. Ces études, fondées sur des données réelles et de haute qualité, ont permis aux chercheurs d'explorer différentes techniques telles que la régression linéaire, les réseaux de neurones artificiels (ANN), les forêts aléatoires, ainsi que les réseaux de neurones récurrents (RNN) pour les séries temporelles.

3.4.2 Résultats des études utilisant les modèles de ML/DL

Lt. Dr. S. Santhosh Baboo et I. Kadar Shereef [26] présentent une application de réseaux de neurones à rétro propagation pour la prédiction de la température. La méthodologie comprenait la collecte et l'analyse d'un ensemble de données contenant des observations météorologiques sur une année complète, notamment la température, le point de rosée, l'humidité, la pression au niveau de la mer, la visibilité, la vitesse du vent, la

vitesse des rafales et les précipitations. Un réseau de neurones à rétro propagation a été entraîné à l'aide de ces données, avec différents paramètres tels que le nombre de couches cachées, le nombre de neurones par couche et le taux d'apprentissage. Les résultats clés montrent que le modèle développé peut prédire la température avec une erreur minimale. Les écarts entre les valeurs exactes et prédictives pour les jours non vus ont été présentés, démontrant la capacité de généralisation du réseau. Le modèle a prouvé son efficacité pour capturer les relations complexes entre les différentes variables atmosphériques et a montré une amélioration notable par rapport aux prévisions traditionnelles.

Les auteurs de [27] prédisent les précipitations quotidiennes en Éthiopie à l'aide d'algorithmes de régression et de classification tels que les arbres de décision (DT), les machines à vecteurs de support (SVM), les K plus proches voisins (KNN), la perception multicouche (MLP) et la mémoire à long terme (LSTM). Leurs résultats montrent que LSTM surpassé significativement les algorithmes de régression dans la prédiction des précipitations quotidiennes.

Table 2: Évaluation de la performance du modèle proposé.[27]

Predictive model	Evaluation metrics					
	NRMSE	MAPE	RMSE	MAE	R² (%)	NSE
LSTM	0.018	0.4786	0.010	0.0082	99.72	0.81

La figure 14 présente les résultats de l'évaluation du modèle proposé sur une période de 60 jours, démontrant que le modèle proposé (LSTM) a pu prédire les précipitations avec une grande précision.

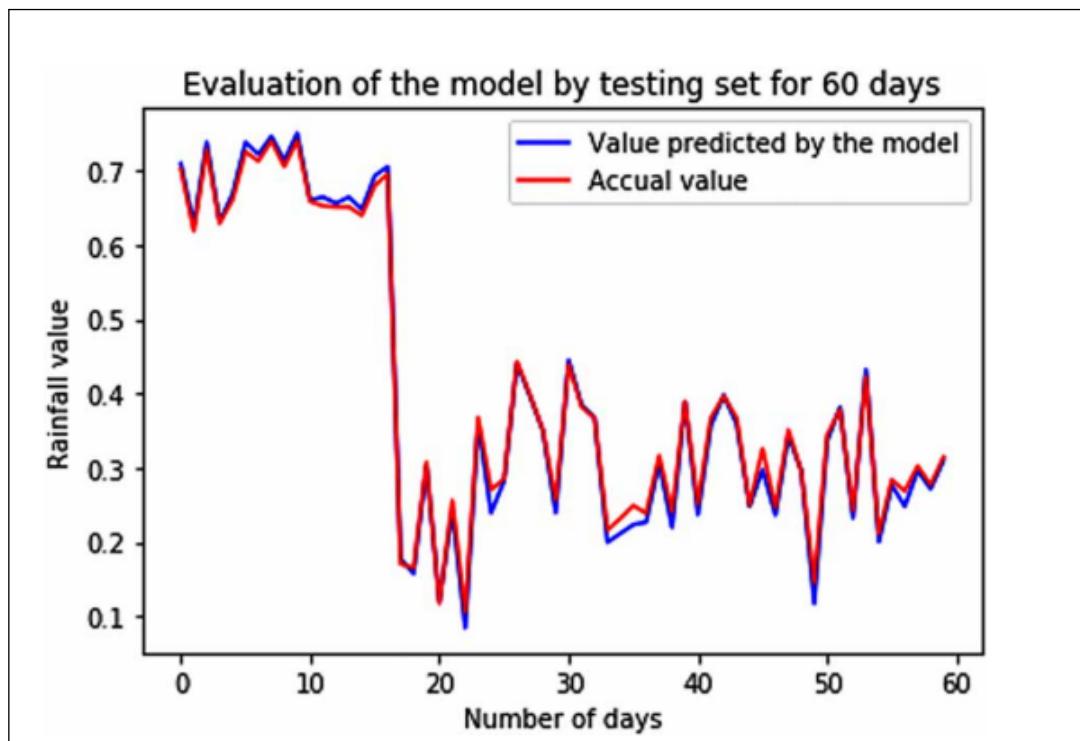


Figure 14: Évaluation du modèle proposé avec jeu de données de test.[27]

Ils ont comparé le modèle proposé (LSTM) avec MLP, SVM, KNN et d'autres méthodes sur l'ensemble de données de la NMSA (National Maritime Safety Association), comme le montre la figure 15 en termes de RMSE (Root Mean Square Error).

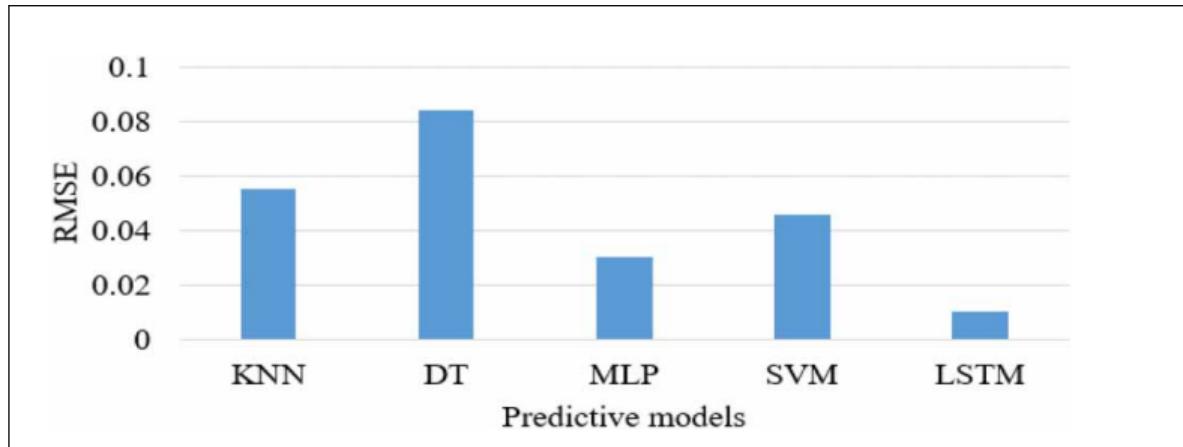


Figure 15: Valeurs RMSE de quatre modèles d'apprentissage automatique existants et du modèle proposé.[27]

Le modèle proposé (LSTM) a systématiquement surpassé les techniques d'apprentissage automatique étudiées, réduisant le RMSE de 4,5 %, 7,4 %, 2 % et 3,6 % par rapport à celles de KNN, DT, MLP et SVM, respectivement.[27]

Une étude [28] récente datée du 27 mai 2024, soutenue par l'Université de Makerere via le Fonds pour la recherche et l'innovation (RIF) sous l'égide du gouvernement ougandais, a examiné l'impact du changement climatique sur la prévision de la quantité de précipitations à court terme. Cette étude met en lumière les limites des modèles de prévision numérique actuels, notamment en raison de leurs besoins de calcul élevés et des erreurs fréquentes qui en découlent.

Le jeu de données météorologiques du bassin du lac Victoria, couvrant les pays de l'Ouganda, du Kenya et de la Tanzanie, a été utilisé pour mener une analyse rigoureuse de divers algorithmes de régression d'apprentissage automatique, y compris la régression par forêt aléatoire, la régression par vecteurs de support, la régression par réseau de neurones, la régression LASSO, la régression par gradient boosting, et la régression par Extreme Gradient Boosting.

L'étude a révélé que le taux de précipitations est un facteur crucial pour déterminer les quantités de pluie tombant sur une base horaire, ce qui est essentiel pour des secteurs tels que l'agriculture, le tourisme, l'aviation, l'éducation et l'ingénierie.

Bien que les variables environnementales aient montré une faible corrélation avec la variable cible selon le coefficient de corrélation de Pearson, l'algorithme Extreme Gradient Boosting a démontré la meilleure performance, avec les valeurs de MAE (Mean Absolute Error) les plus faibles, soit 0,006, 0,018 et 0,005 pour les données météorologiques du bassin du lac Victoria en Ouganda (Table 3), au Kenya (Table 4) et en Tanzanie (Table 4) respectivement.

Table 3: Comparaison de MAE et RMSE avec des techniques de référence utilisant les données météorologiques de l'Ouganda.[28]

Method	Test		Ranking	
	MAE	RMSE	MAE	RMSE
RFR	0.12183	0.42925	5	5
SVR	0.05888	0.12302	3	3
NNR	0.01018	0.05952	2	2
LASSO	0.06610	0.26626	4	4
GBR	0.12624	0.43931	6	6
XGBoost	0.00616	0.04439	1	1

Table 4: Comparaison de l'EEM et de l'EQM avec des techniques utilisant les données météorologiques du Kenya et de la Tanzanie.[28]

Stations	Model	Performance Metrics	
		MAE	RMSE
Kenya	RFR	0.478242	0.702832
	SVR	0.143649	0.313051
	NNR	0.426446	0.623602
	LASSO	0.424491	0.592471
	GBR	0.315397	0.472994
	XGBoost	0.018452	0.227948
Tanzania	RFR	0.233441	0.842882
	SVR	0.081293	0.487932
	NNR	0.144132	0.702039
	LASSO	0.233208	0.759692
	GBR	0.055811	0.462857
	XGBoost	0.005862	0.213369

Cette étude suggère que l'Extreme Gradient Boosting est un algorithme mieux adapté à la prédiction de la quantité de précipitations sur la base de caractéristiques environnementales sélectionnées.

L'étude [29] explore l'application de techniques d'apprentissage automatique pour la prévision météorologique, en mettant l'accent sur l'évolution des méthodes, depuis les approches manuelles traditionnelles jusqu'aux modèles informatisés modernes. Historiquement, la prévision du temps était basée sur des observations manuelles telles que la pression barométrique, les conditions météorologiques actuelles et la couverture nuageuse. Cependant, avec la découverte de la non-linéarité dans les données météorologiques, l'attention s'est portée sur des modèles de prédiction non linéaires.

L'étude a comparé les performances de différents modèles d'apprentissage automatique sur des données recueillies auprès de stations météorologiques d'aéroports de plusieurs villes. Les modèles étudiés incluent les machines à vecteurs de support (SVM), les réseaux de neurones artificiels (ANN) et les réseaux de neurones récurrents pour les séries temporelles (RNN).

Les résultats de l'analyse montrent que le RNN pour les séries temporelles offre une meilleure précision de prédiction par rapport au SVM et à l'ANN. De plus, l'étude a mis en évidence l'importance de la taille de la fenêtre de prédiction, notant que des fenêtres de prédiction plus grandes entraînent une erreur plus élevée. Le tableau 5 présente la comparaison entre les trois modèles.

Table 5: Tableau d'exemple avec des valeurs provisoires.[29]

Model	Prediction window	MRS error
SVM	8 weeks	6.67
ANN	8 weeks	3.1
Time Series RNN	8 weeks	1.41

La figure 16 montre la température sur l'axe des y et le nombre de séquences des données d'essai sur l'axe des x, la température prédite par la SVM est représentée par la couleur bleue et la température réelle est représentée par l'orange.

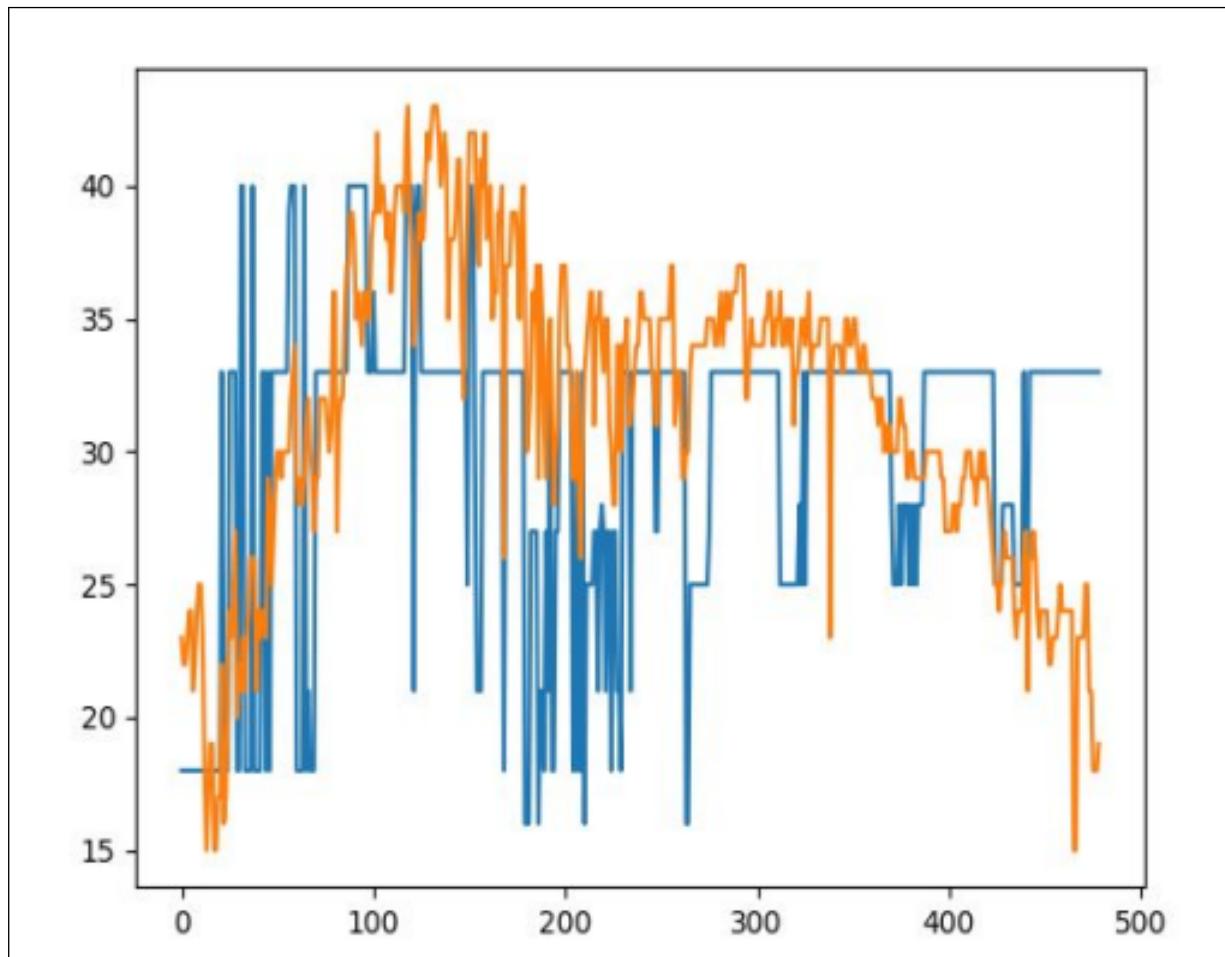


Figure 16: Température réelle par rapport à la température prédite par SVM.[29]

La figure 17 montre la température sur l'axe des y et le nombre de séquences des données de test sur l'axe des x, la température prédite par ANN est représentée par la couleur rouge et la température réelle est représentée par la couleur bleue.

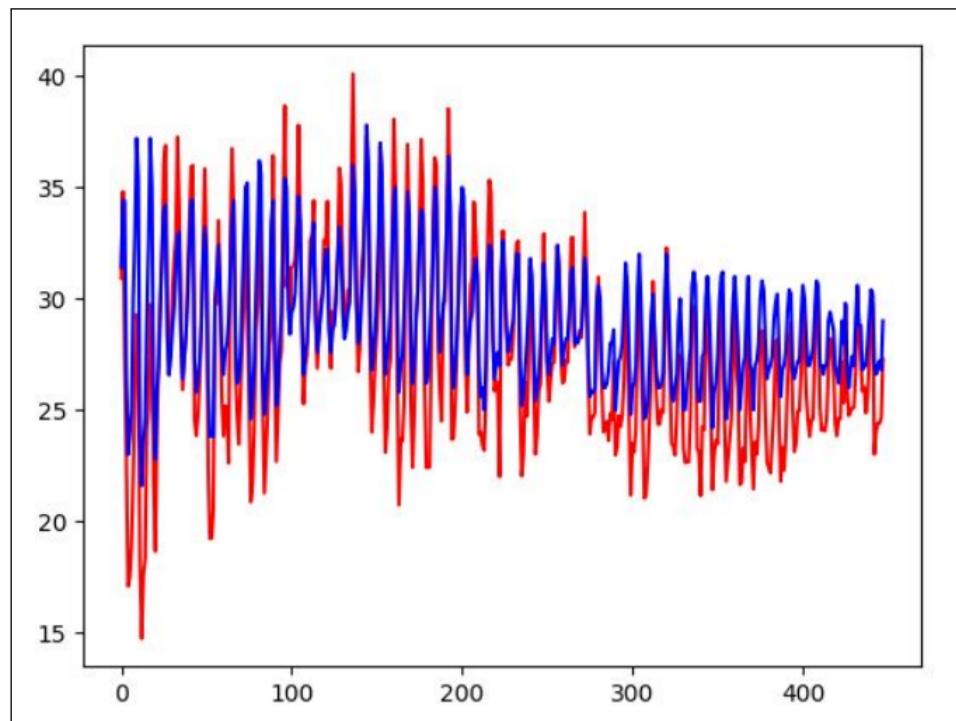


Figure 17: Température réelle par rapport à la température prévue par ANN.[29]

La figure 18 montre la température sur l’axe des y et le nombre de séquences des données de test sur l’axe des x, la température prédite par RNN est représentée par la couleur verte et la température réelle est représentée par le bleu.

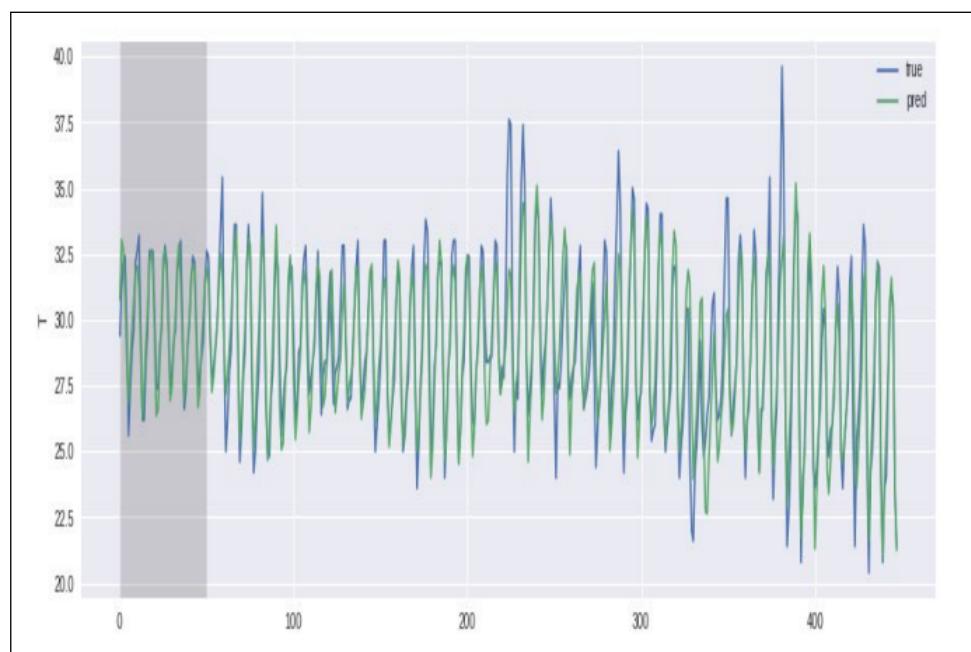


Figure 18: Température réelle vs température prédite par RNN.[29]

En utilisant l’erreur quadratique moyenne (RMSE) comme métrique d’évaluation, l’étude conclut que l’utilisation d’un RNN pour les séries temporelles est la méthode la plus efficace pour la prévision météorologique parmi les modèles testés. Ces résultats suggèrent que le RNN pour les séries temporelles est un outil prometteur pour améliorer la précision

des prévisions météorologiques et aider à protéger la vie et les biens en fournissant des avertissements météorologiques plus précis.

3.4.3 Limitations et orientations futures

Les modèles ML/DL sont prometteurs pour la prévision météorologique en raison de leur capacité à traiter de grandes quantités de données et à capturer des relations non linéaires complexes. Cependant, ces modèles peuvent être limités par la qualité des données d'entrée et nécessitent des ressources computationnelles importantes. L'intégration des connaissances des systèmes dynamiques pourrait améliorer encore leur performance. Les études menées jusqu'à présent se concentrent sur la compétence moyenne des prévisions, sans traitement particulier des événements extrêmes. Même si quelques études de cas ont été menées, par exemple le suivi des cyclones et les températures extrêmes de surface, celles-ci ne permettent pas une évaluation juste de la capacité des modèles basés sur les données à prévoir les extrêmes météorologiques à l'échelle mondiale.

3.5 Conclusion

Dans ce chapitre, nous avons examiné les approches dynamiques et les techniques de machine learning (ML) et deep learning (DL) pour la prévision météorologique. Les systèmes dynamiques, bien qu'efficaces pour les prévisions à long terme, souffrent de limitations en précision à court terme et nécessitent des ressources computationnelles importantes. En revanche, les approches ML/DL excellent dans le traitement de grandes quantités de données et la capture de relations complexes, offrant de meilleures performances pour les prévisions à court terme, notamment pour des paramètres spécifiques.

Outils et processus de l'apprentissage automatique

4.1 Introduction

L'apprentissage automatique (ML) est devenu une composante essentielle du développement de l'intelligence artificielle (IA), offrant des outils puissants pour analyser des données complexes et automatiser des tâches autrefois réservées à l'intelligence humaine. Ce chapitre explore les fondements du ML, les différentes approches d'apprentissage, ainsi que les outils et méthodes utilisés pour construire des modèles performants. En examinant les algorithmes, la matrice de confusion, et les langages de programmation, nous offrirons une vue d'ensemble des processus qui sous-tendent l'apprentissage automatique et son application dans divers domaines.

4.2 Machine Learning

L'intelligence artificielle (IA) vise à imiter le comportement humain et combine des principes de l'informatique, des statistiques, des algorithmes, de l'apprentissage automatique (ML) et de la science des données. L'apprentissage automatique, en particulier, est efficace pour détecter des motifs complexes dans de grandes quantités de données, offrant des insights précieux pour divers domaines.[\[30\]](#)

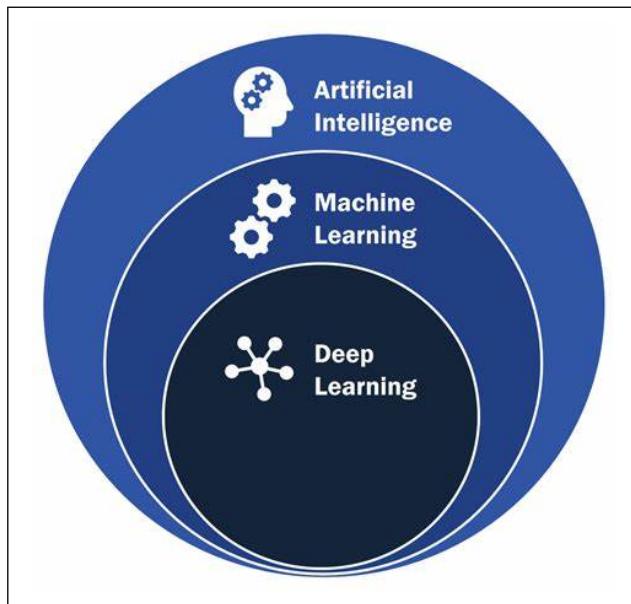


Figure 19: Relation entre l'intelligence artificielle (IA), l'apprentissage automatique (AA) et l'apprentissage profond (AP).[\[30\]](#)

Les premières recherches en IA datent des années 1940, avec des développements notables dans les années 1950, mais les années 1960 ont révélé les limites des premiers systèmes, entraînant un ralentissement des investissements. Les années 1990 ont vu l'émergence du ML comme discipline distincte, facilitée par l'accès à des données massives et la collaboration interdisciplinaire. Aujourd'hui, le ML évolue rapidement, bénéficiant des avancées en Big Data et en puissance de calcul, et joue un rôle clé dans des domaines tels que la santé, la finance et les technologies de l'information.[31]

4.2.1 Définition

L'apprentissage automatique (ML) est une branche de l'intelligence artificielle (IA) dédiée à la création de systèmes capables de reproduire des comportements intelligents ou de simuler cette intelligence. Pour que les machines apprennent et s'adaptent automatiquement, elles doivent stocker les connaissances acquises de leurs expériences. Ces connaissances sont utilisées pour prédire des événements futurs ou générer de nouvelles données. ML s'appuie sur des concepts mathématiques et statistiques, tels que la régression, pour modéliser les données et choisir les paramètres optimaux pour les décrire. Les techniques d'apprentissage automatique s'inspirent également d'autres disciplines telles que la physique, la biologie, l'imagerie et l'ingénierie pour développer des modèles et des méthodes d'apprentissage avancés, repoussant ainsi les frontières de l'interdisciplinarité.[31]

4.2.2 Les types de machine learning

Les techniques d'apprentissage automatique (ML) utilisées pour construire des applications d'intelligence artificielle se classent en trois grandes familles : l'apprentissage supervisé (SL), l'apprentissage non supervisé (UL) et l'apprentissage par renforcement (RL). L'apprentissage supervisé et l'apprentissage non supervisé sont brièvement décrits ci-dessous, bien que les détails techniques dépassent le cadre de ce rapport. La plupart des applications de l'apprentissage par renforcement se trouvent dans les domaines des jeux de plateau et des jeux vidéos, et sont hors du champ de ce document.[31]

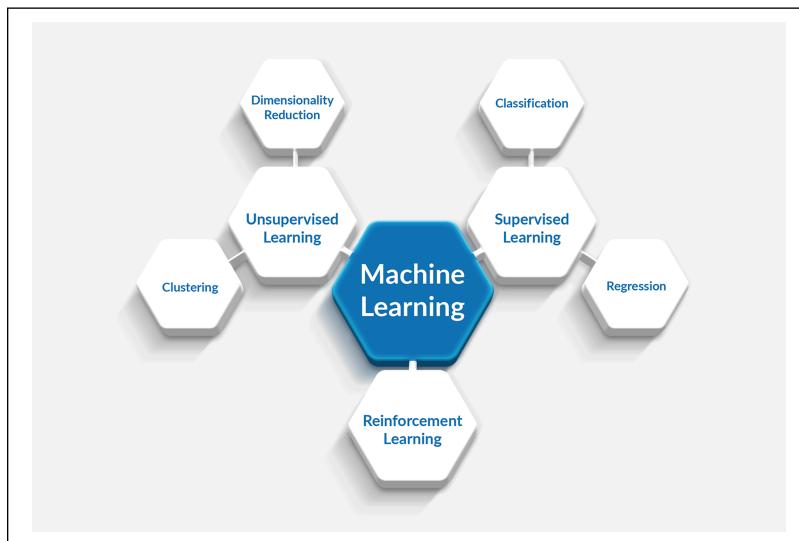


Figure 20: Algorithmes d'apprentissage automatique et cas d'utilisation.[31]

4.2.2.1 L'apprentissage supervisé

L'apprentissage supervisé est une approche de l'apprentissage automatique caractérisée par l'utilisation de jeux de données étiquetés. Ces jeux de données sont conçus pour entraîner ou guider les algorithmes afin de classer les données avec précision ou de prédire des résultats. En utilisant des entrées et des sorties étiquetées, le modèle peut évaluer sa précision et affiner ses performances au fil du temps.[32]

Cette approche peut être classée en deux grands types de problèmes dans le domaine de l'exploration de données : la **classification** et la **régression**.

- **Classification :** Les tâches de classification consistent à utiliser des algorithmes pour attribuer avec précision des données d'essai à des catégories distinctes, telles que distinguer différents types de fruits ou filtrer les emails indésirables des emails légitimes. Les algorithmes de classification courants comprennent les classificateurs linéaires, les machines à vecteurs de support, les arbres de décision et les forêts aléatoires.
- **Régression :** En revanche, la régression est une méthode d'apprentissage supervisé qui utilise des algorithmes pour comprendre la relation entre les variables dépendantes et indépendantes. Les modèles de régression sont précieux pour prédire des valeurs numériques basées sur divers points de données, comme prévoir les revenus de ventes pour une entreprise particulière. Les algorithmes de régression populaires incluent la régression linéaire, la régression logistique et la régression polynomiale.

Comme montré dans la Figure 21, l'algorithme d'apprentissage supervisé utilise un jeu de données étiqueté pour générer un modèle. Par la suite, ce modèle peut être utilisé avec de nouvelles données pour évaluer sa précision ou déployé en temps réel avec des données en direct.[32]

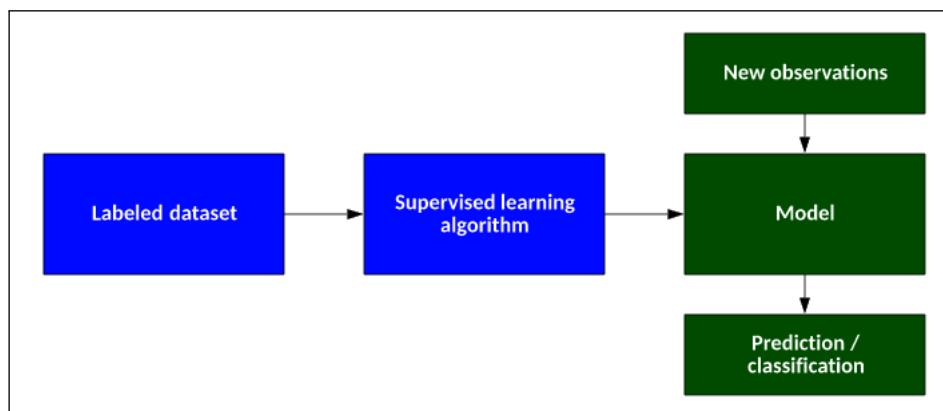


Figure 21: Un algorithme d'apprentissage supervisé typique.[32]

4.2.2.2 L'apprentissage non supervisé

L'apprentissage non supervisé, en revanche, utilise des algorithmes d'apprentissage automatique pour analyser et regrouper des jeux de données non étiquetés. Ces algorithmes découvrent de manière autonome des motifs cachés dans les données sans intervention humaine, d'où le terme "**non supervisé**."

Les modèles d'apprentissage non supervisé servent trois objectifs principaux : le clustering, l'association et la réduction de dimensionnalité [33] :

- **Clustering** : Il consiste à regrouper des données non étiquetées en fonction de leurs similitudes ou de leurs dissemblances. Par exemple, les algorithmes de clustering K-means regroupent des points de données similaires, la valeur de K déterminant la taille et la granularité des groupes. Cette technique est utile pour des tâches telles que la segmentation de marché et la compression d'images.
- **Association** : Une autre méthode d'apprentissage non supervisé utilise diverses règles pour identifier les relations entre les variables au sein d'un jeu de données. Ces méthodes sont couramment utilisées dans l'analyse des paniers d'achat et les systèmes de recommandation, facilitant des suggestions telles que « Les clients ayant acheté cet article ont également acheté ».
- **Réduction de dimensionnalité** : est utilisée lorsque les jeux de données contiennent un nombre excessif de caractéristiques ou de dimensions. Cette technique réduit le nombre d'entrées de données à une taille gérable tout en préservant l'intégrité des données. Souvent utilisée dans le prétraitement des données, les méthodes de réduction de dimensionnalité comme les autoencodeurs éliminent le bruit des données visuelles pour améliorer la qualité des images.

Les algorithmes d'apprentissage non supervisé organisent les données au sein d'un jeu de données non étiqueté en fonction de caractéristiques sous-jacentes inhérentes (voir Figure 22). Étant donné qu'il n'y a pas de labels prédéfinis, évaluer le résultat est difficile, ce qui constitue une distinction notable par rapport aux algorithmes d'apprentissage supervisé.[33]

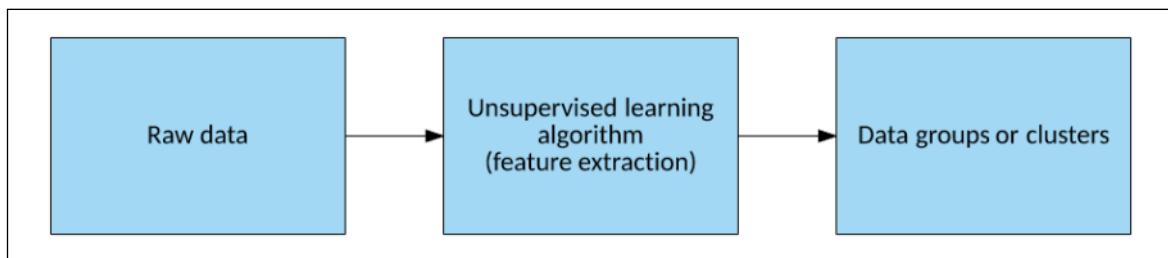


Figure 22: Un algorithme d'apprentissage non supervisé.[33]

4.2.2.3 L'apprentissage par renforcement

L'apprentissage par renforcement est une approche de l'apprentissage automatique où un agent apprend à prendre des décisions en interagissant avec un environnement donné. L'objectif est que l'agent adopte des actions qui maximisent les récompenses qu'il reçoit ou minimisent les pénalités.

En interagissant avec l'environnement, l'agent reçoit des signaux de récompense ou de pénalité en fonction de ses actions. Ces signaux guident l'agent dans l'apprentissage d'une stratégie de décision optimale, une politique qui lui permet de choisir des actions qui maximisent les récompenses cumulatives au fil du temps.[34]

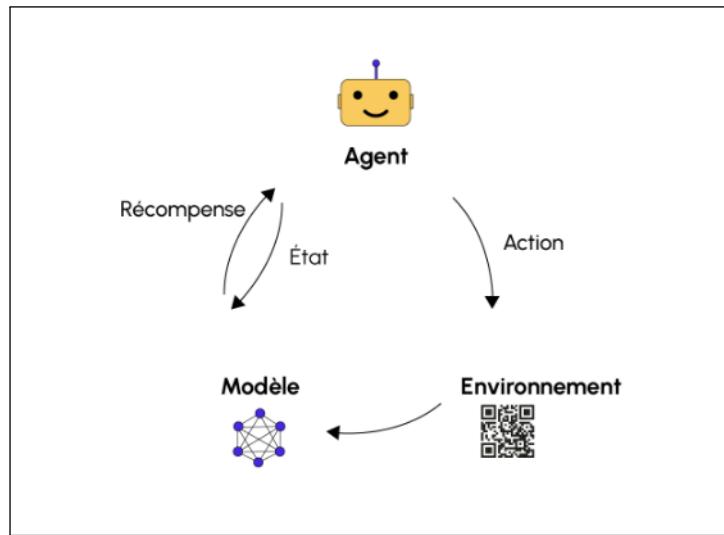


Figure 23: Cadre de l'apprentissage par renforcement.[34]

Le tableau 6 compare trois approches de l'apprentissage automatique : l'apprentissage supervisé, qui utilise des données étiquetées pour prédire ou classifier ; l'apprentissage non supervisé, qui découvre des motifs dans des données non étiquetées ; et l'apprentissage par renforcement, où l'algorithme apprend en interagissant avec son environnement et en ajustant ses actions en fonction des récompenses reçues.

Table 6: Comparaison entre les différents types d'apprentissage.[34]

	Apprentissage supervisé	Apprentissage non supervisé	Apprentissage par renforcement
Définition	L'algorithme apprend à partir de données étiquetées.	L'algorithme est entraîné sur des données non étiquetées sans indications spécifiques.	L'algorithme interagit avec son environnement en prenant des actions et en apprenant de ses erreurs et de ses succès.
Types de problèmes	Régression et classification.	Association et clustering.	Basé sur un système de récompenses.
Types de données	Données étiquetées.	Données non étiquetées.	Aucune donnée fournie au préalable.
Approche	Étudie les relations sous-jacentes entre les données d'entrée et les étiquettes.	Découvre des motifs communs au sein des données d'entrée.	Apprend une stratégie de comportement basée sur les expériences et les récompenses

4.3 Aperçu sur les algorithmes de machine learning

4.3.1 La régression linéaire

4.3.1.1 Définition

Le modèle de régression linéaire est un algorithme d'apprentissage supervisé qui permet de prédire une variable cible continue (variable dépendante) grâce à une ou plusieurs variables explicatives (variables indépendantes ou prédictives). En d'autres termes, il s'agit d'établir les relations entre deux ou plusieurs variables.[\[35\]](#)

Lorsqu'il n'y a qu'une variable explicative, on parle de régression linéaire simple. En revanche, s'il y en a plusieurs, on parle de régression linéaire multiple.[\[35\]](#)

4.3.1.2 Formulation mathématique

L'équation mathématique de la régression linéaire se traduit comme suit :

$$Y = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$$

Dans cette équation :

- Y correspond à la variable cible.
- θ correspond au terme de biais ou vecteur de paramètres.
- x_1, x_2, \dots, x_n correspondent aux variables explicatives.

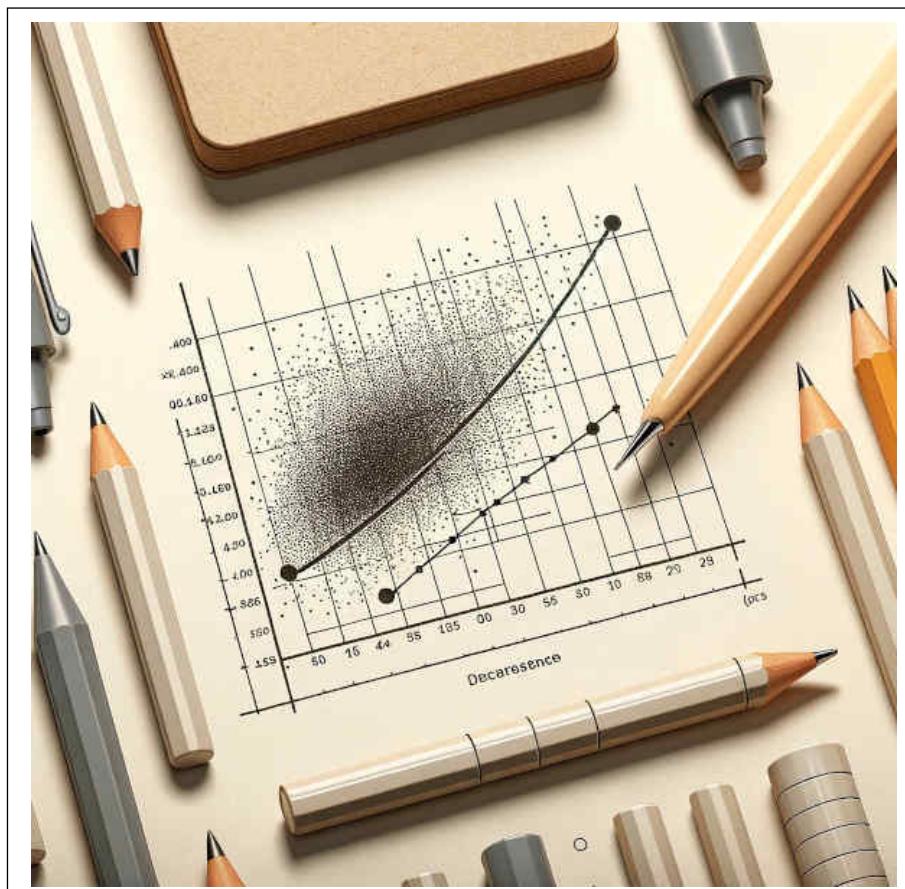


Figure 24: Régression linéaire simple.[\[36\]](#)

4.3.2 Forêt aléatoire

La forêt aléatoire (Random Forest) est un algorithme d'apprentissage automatique très populaire qui fait partie des techniques d'apprentissage supervisé. Il peut être utilisé pour des problèmes de classification et de régression en machine learning. Il repose sur le concept d'apprentissage par ensemble (ensemble learning), qui consiste à combiner plusieurs classificateurs pour résoudre un problème complexe et améliorer les performances du modèle.^[37]

Comme son nom l'indique, une Forêt Aléatoire est un classificateur qui contient un certain nombre d'arbres de décision appliqués à divers sous-ensembles du jeu de données donné et qui prend la moyenne pour améliorer la précision prédictive de ce jeu de données. Au lieu de se fier à un seul arbre de décision, la forêt aléatoire prend la prédiction de chaque arbre et, sur la base de la majorité des votes des prédictions, elle prédit le résultat final.^[37]

La figure 25 explique le fonctionnement de l'algorithme Random Forest.

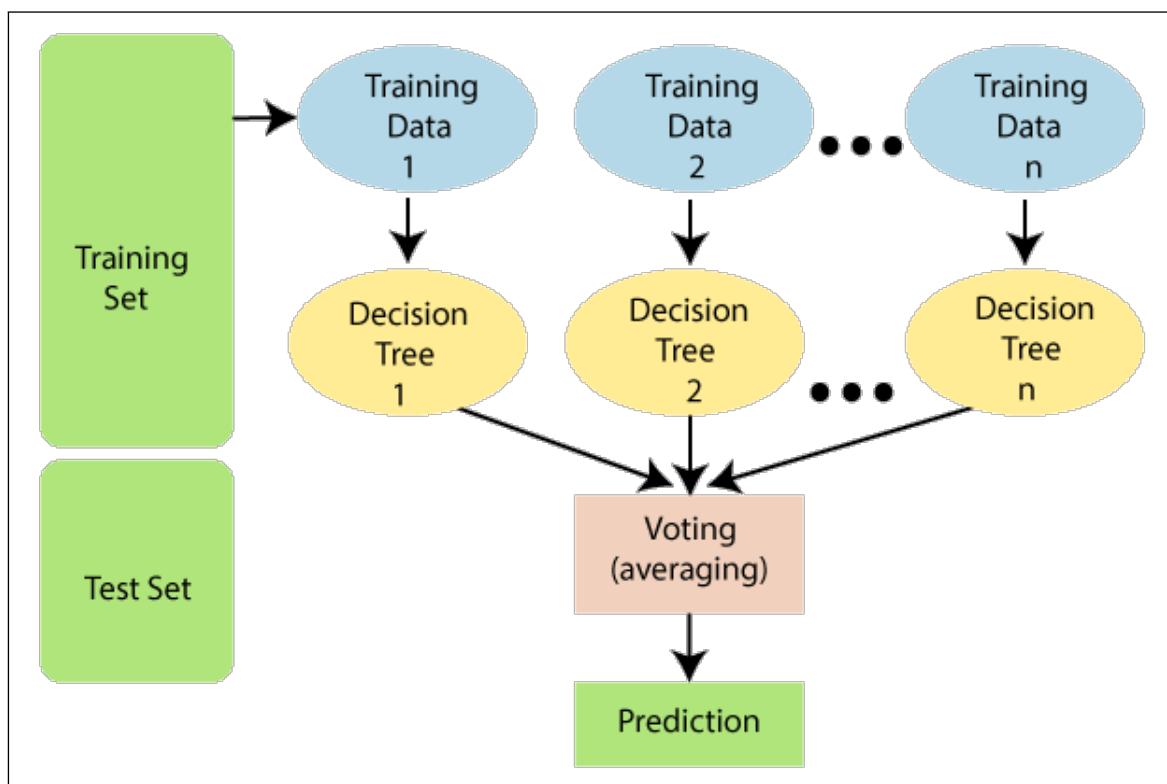


Figure 25: Fonctionnement de l'algorithme Random Forest.^[37]

4.3.3 XGBoost

Pour bien comprendre XGBoost, il est essentiel de saisir d'abord les concepts et algorithmes de machine learning sur lesquels XGBoost repose : l'apprentissage supervisé, les arbres de décision, l'apprentissage ensembliste et le gradient boosting.

L'apprentissage supervisé utilise des algorithmes pour entraîner un modèle à identifier des schémas dans un ensemble de données comportant des étiquettes et des caractéristiques, puis utilise le modèle entraîné pour prédire les étiquettes sur les caractéristiques d'un nouvel ensemble de données.^[38]

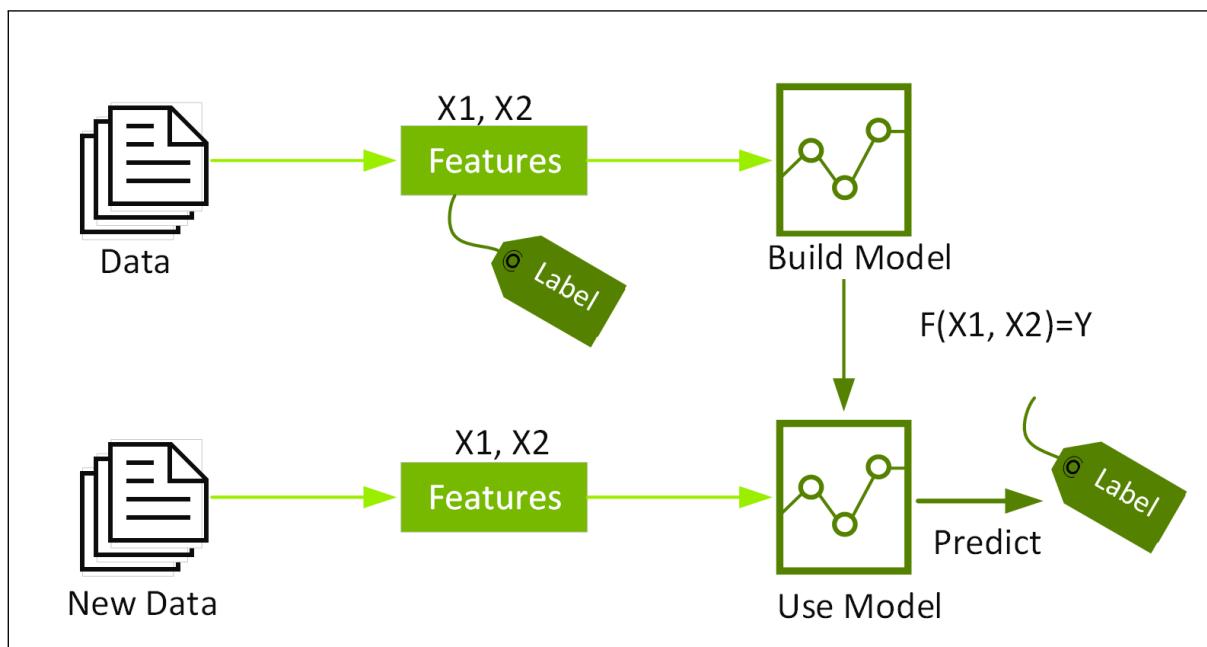


Figure 26: Apprentissage supervisé.[38]

Les arbres de décision créent un modèle qui prédit l'étiquette en évaluant un arbre de questions basées sur des caractéristiques, vrai/faux, et en estimant le nombre minimal de questions nécessaires pour évaluer la probabilité de prendre une décision correcte. Les arbres de décision peuvent être utilisés pour la classification afin de prédire une catégorie, ou pour la régression afin de prédire une valeur numérique continue. Dans l'exemple simple ci-dessous, un arbre de décision est utilisé pour estimer le prix d'une maison (l'étiquette) en fonction de la taille et du nombre de chambres (les caractéristiques).[38]

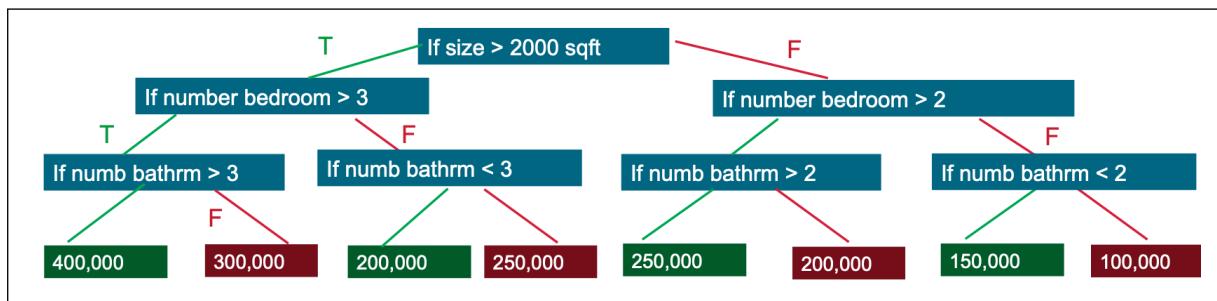


Figure 27: Arbre de décision.[38]

Les Gradients Boosting Decision Trees (GBDT) sont un algorithme d'apprentissage par ensemble d'arbres de décision, similaire à la forêt aléatoire, utilisé pour la classification et la régression. Les algorithmes d'apprentissage par ensemble combinent plusieurs algorithmes d'apprentissage automatique pour obtenir un meilleur modèle.[38]

Tant la forêt aléatoire que les GBDT construisent un modèle composé de plusieurs arbres de décision. La différence réside dans la manière dont les arbres sont construits et combinés.[38]

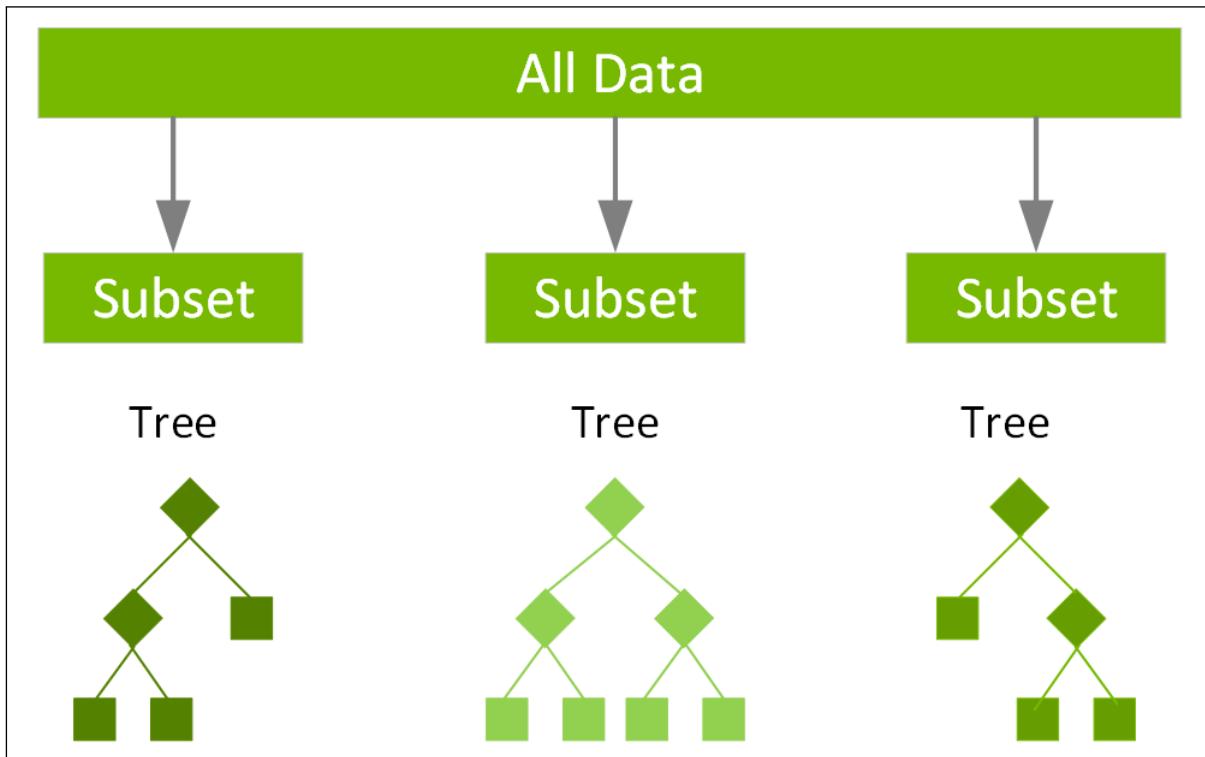


Figure 28: Ensemble d'arbres de décision.[38]

La forêt aléatoire utilise une technique appelée **bagging** pour construire des arbres de décision complets en parallèle à partir d'échantillons bootstrap aléatoires du jeu de données. La prédiction finale est une moyenne de toutes les prédictions des arbres de décision.

Le terme **gradient boosting** vient de l'idée de **boosting** ou d'amélioration d'un modèle faible en le combinant avec plusieurs autres modèles faibles pour générer un modèle collectivement fort. Le gradient boosting est une extension du boosting où le processus de génération additive de modèles faibles est formalisé comme un algorithme de descente de gradient sur une fonction objectif. Le gradient boosting fixe des résultats cibles pour le modèle suivant afin de minimiser les erreurs. Les résultats cibles pour chaque cas sont basés sur le gradient de l'erreur (d'où le nom gradient boosting) par rapport à la prédiction.[38]

Les GBDT entraînent itérativement un ensemble d'arbres de décision peu profonds, chaque itération utilisant les résidus d'erreur du modèle précédent pour ajuster le modèle suivant. La prédiction finale est une somme pondérée de toutes les prédictions des arbres. Le "bagging" des forêts aléatoires minimise la variance et le surajustement, tandis que le "boosting" des GBDT minimise le biais et le sous-apprentissage.[38]

XGBoost est une implémentation évolutive et très précise du gradient boosting qui pousse les limites de la puissance de calcul pour les algorithmes d'arbres boostés, étant principalement conçu pour dynamiser la performance des modèles d'apprentissage automatique et la vitesse de calcul. Avec XGBoost, les arbres sont construits en parallèle, au lieu d'être séquentiels comme dans les GBDT. Il suit une stratégie de niveau, en parcourant les valeurs du gradient et en utilisant les sommes partielles pour évaluer la qualité des divisions à chaque possible division dans le jeu de données d'entraînement.[38]

4.3.4 Les réseaux de neurones artificiels (ANN)

4.3.4.1 Définition

Le terme **réseau de neurones artificiels** est dérivé des réseaux de neurones biologiques qui forment la structure du cerveau humain. De manière similaire au cerveau humain, qui possède des neurones interconnectés les uns aux autres, les réseaux de neurones artificiels ont également des neurones interconnectés entre eux dans différentes couches du réseau. Ces neurones sont appelés des noeuds[39].

La figure 29 illustre le schéma typique d'un réseau de neurones biologiques.

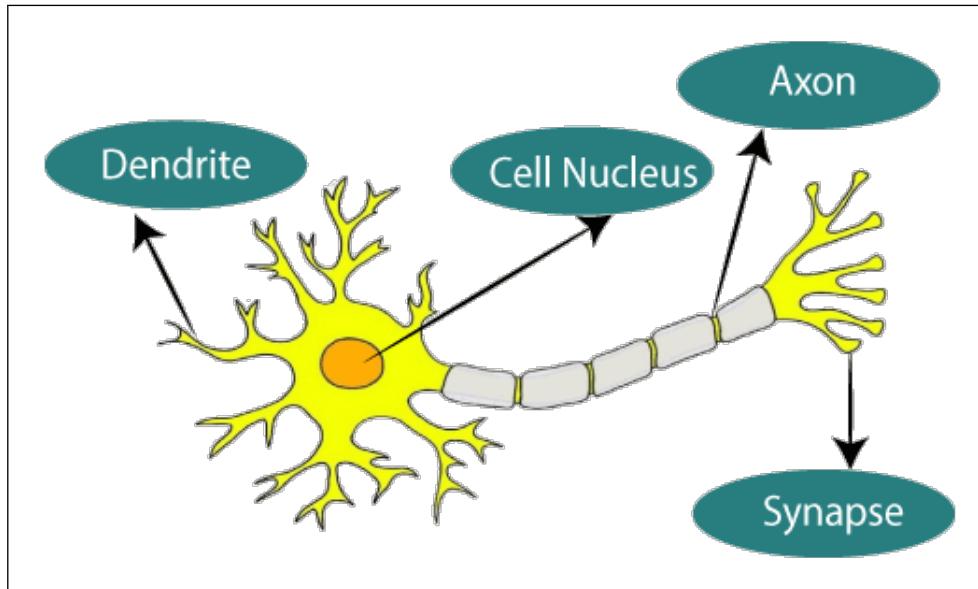


Figure 29: Réseau de neurones biologiques.[39]

Le réseau de neurones artificiels typique ressemble à quelque chose comme la figure 30.

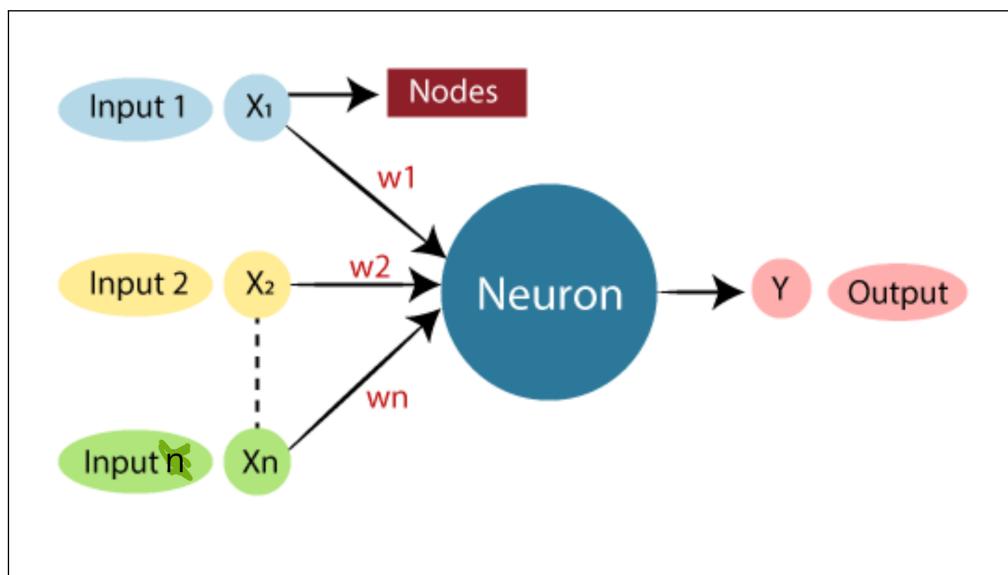


Figure 30: Réseau de neurones artificiels.[39]

Les dendrites des réseaux de neurones biologiques représentent les entrées dans les réseaux de neurones artificiels, le noyau cellulaire représente les nœuds, la synapse représente les poids, et l'axon représente la sortie.

Le tableau 7 résume la relation entre le réseau de neurones biologique et le réseau de neurones artificiel.

Table 7: Relation entre le réseau de neurones biologique et le réseau de neurones artificiel.[39]

Réseau de neurones biologique	Réseau de neurones artificiel
Dendrites	Entrées
Noyau cellulaire	Nœuds
Synapse	Poids
Axone 1	Sortie

Pour comprendre le concept de l'architecture d'un réseau de neurones artificiel, nous devons comprendre de quoi se compose un réseau de neurones. Un réseau de neurones est constitué d'un grand nombre de neurones artificiels, appelés unités, disposés en une séquence de couches.[39]

Examinons les différents types de couches disponibles dans un réseau de neurones artificiel.

Le réseau de neurones artificiel se compose principalement de trois couches, comme illustré dans la figure 31.

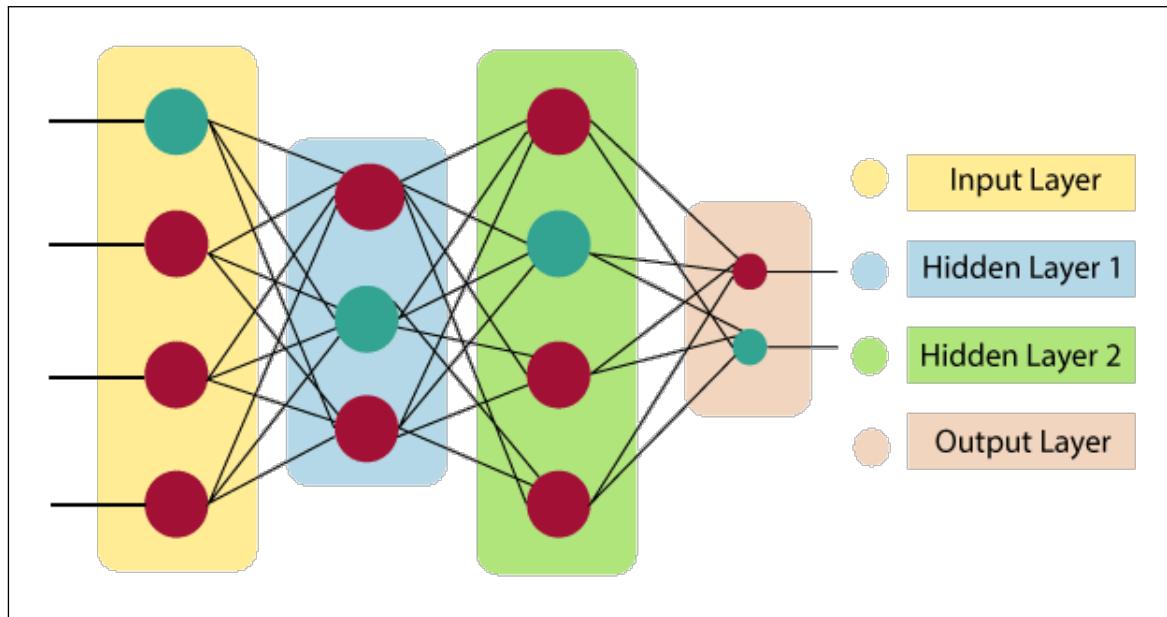


Figure 31: L'architecture d'un réseau de neurones artificiel.[39]

- **Couche d'entrée (Input Layer) :** Comme son nom l'indique, elle accepte des entrées sous plusieurs formats différents fournis par le programmeur.

- **Couche cachée (Hidden Layer)** : La couche cachée se situe entre les couches d'entrée et de sortie. Elle effectue tous les calculs pour trouver des caractéristiques et des motifs cachés.
- **Couche de sortie (Output Layer)** : Les données d'entrée passent par une série de transformations à l'aide de la couche cachée, ce qui donne finalement un résultat transmis par cette couche.

Le réseau de neurones artificiel prend des entrées et calcule la somme pondérée des entrées, en incluant un biais. Ce calcul est représenté sous la forme d'une fonction de transfert.

$$\sum_{i=1}^n w_i \cdot x_i + b$$

Il détermine que le total pondéré est passé en tant qu'entrée à une fonction d'activation pour produire la sortie. Les fonctions d'activation décident si un nœud doit s'activer ou non. Seuls ceux qui s'activent passent à la couche de sortie. Il existe différentes fonctions d'activation disponibles qui peuvent être appliquées en fonction du type de tâche que nous effectuons.

4.3.4.2 Fonctionnement d'un réseau de neurones artificiels

Un réseau de neurones artificiels peut être mieux représenté sous forme de graphe dirigé pondéré, où les neurones artificiels forment les noeuds. L'association entre les sorties des neurones et les entrées des neurones peut être vue comme des arêtes dirigées avec des poids. Le réseau de neurones artificiels reçoit le signal d'entrée de la source externe sous forme de vecteur. Ces entrées sont ensuite assignées mathématiquement par les notations x_n pour chaque nombre n d'entrées.[39]

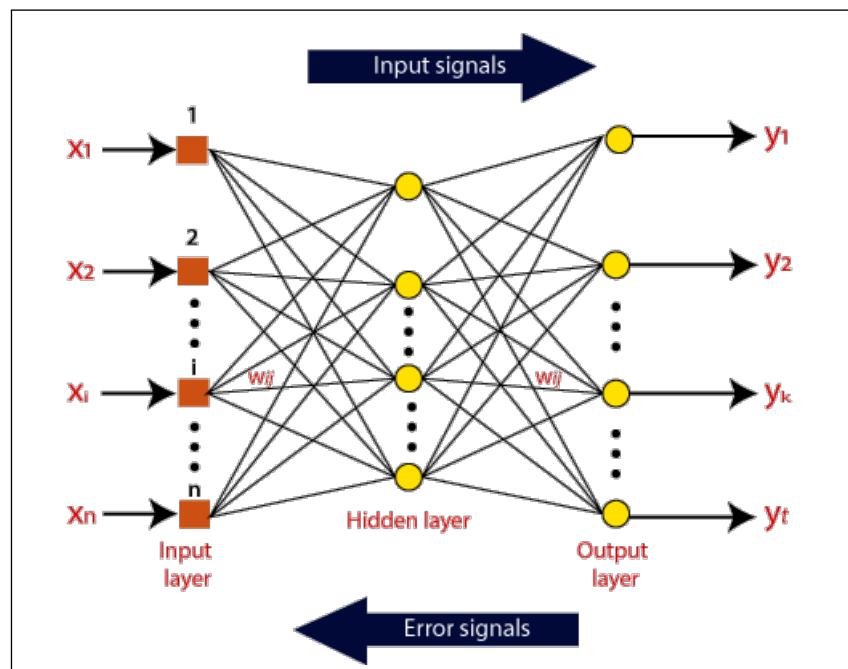


Figure 32: Fonctionnement d'un réseau de neurones artificiels.[39]

Ensuite, chacune des entrées est multipliée par ses poids correspondants (ces poids sont les paramètres utilisés par les réseaux de neurones artificiels pour résoudre un problème spécifique). En termes généraux, ces poids représentent généralement la force de l'interconnexion entre les neurones au sein du réseau de neurones artificiels. Toutes les entrées pondérées sont ensuite résumées dans l'unité de calcul.[39]

Si la somme pondérée est égale à zéro, un biais est ajouté pour que la sortie ne soit pas nulle ou pour ajuster la réponse du système. Le biais a la même entrée et un poids égal à 1. Ici, le total des entrées pondérées peut varier de 0 à l'infini positif. Pour maintenir la réponse dans les limites de la valeur souhaitée, une certaine valeur maximale est établie comme référence, et le total des entrées pondérées est ensuite passé à travers la fonction d'activation.[39]

4.3.5 Réseau Neuronal Récurrent (RNN) basé sur des séries temporelles

Les réseaux neuronaux récurrents (RNN) sont un type de réseau neuronal profond où les données d'entrée et les états cachés antérieurs sont introduits dans les couches du réseau, donnant au réseau un état et donc de la mémoire. Les RNN sont couramment utilisés pour les données basées sur la séquence ou le temps. Pendant l'entraînement, les données d'entrée sont transmises au réseau avec une certaine taille de mini-lot (le nombre de séquences de données dans le mini-lot) et la longueur de la séquence (le nombre de pas de temps dans chaque séquence). Comme le montre la figure 33, chaque couche d'un RNN est composée d'unités récurrentes, dont le nombre est la taille cachée de la couche.[40]

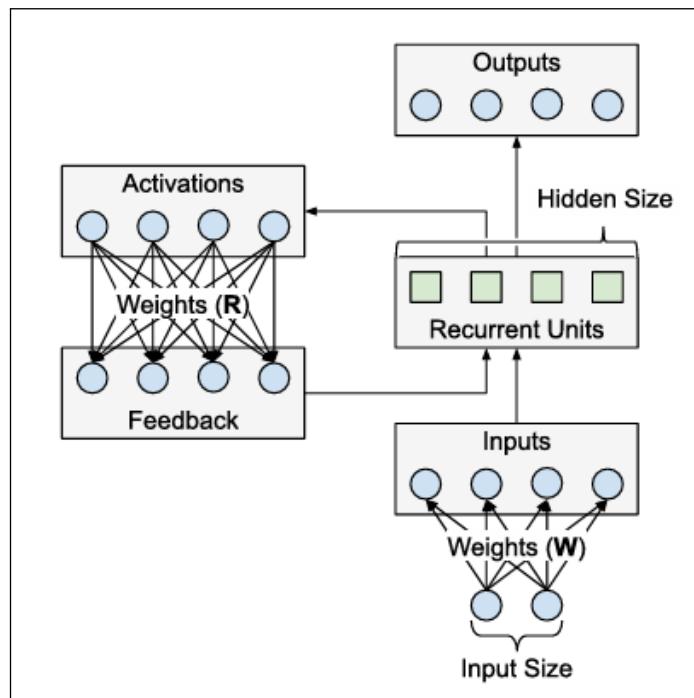


Figure 33: Une seule couche récurrente avec deux entrées et quatre unités cachées.[40]

Les RNN sont constitués de neurones : des noeuds de traitement de données qui travaillent ensemble pour effectuer des tâches complexes. Les neurones sont organisés

en couches d'entrée, de sortie et cachées. La couche d'entrée reçoit les informations à traiter et la couche de sortie fournit le résultat. Le traitement, l'analyse et la prévision des données ont lieu dans la couche cachée. La figure 34 montre un schéma d'un RNN.[41]

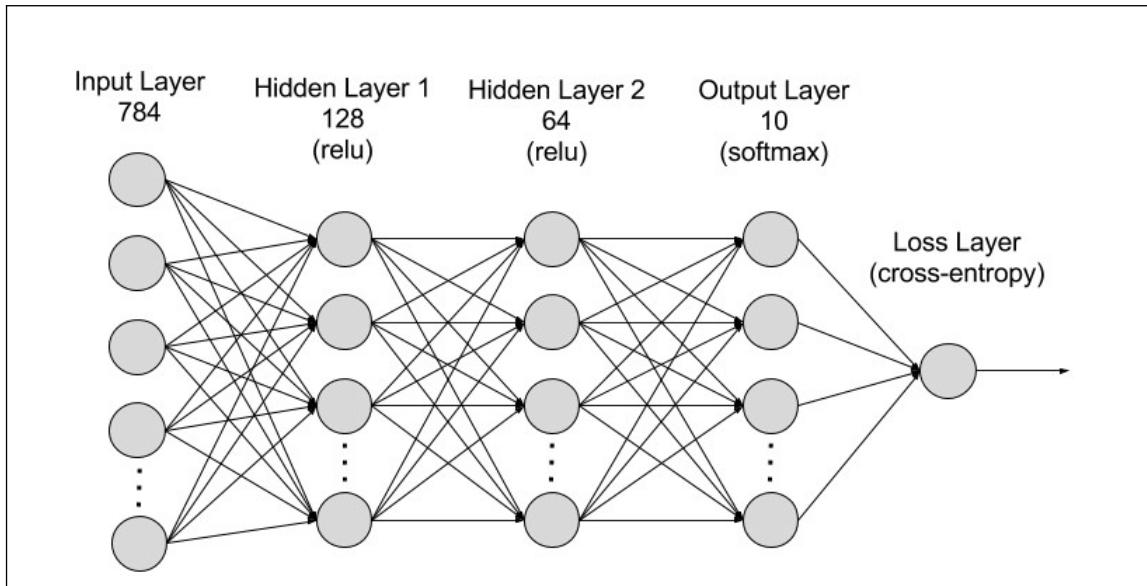


Figure 34: Schéma d'un RNN.[41]

Les RNN sont souvent caractérisés par une architecture biunivoque : une séquence d'entrée est associée à une sortie. Cependant, vous pouvez les ajuster de manière flexible dans différentes configurations à des fins spécifiques. Voici plusieurs types de RNN courants.[41]

- **Un à plusieurs :** Ce type RNN canalise une entrée vers plusieurs sorties. Il permet des applications linguistiques telles que le sous-titrage d'images en générant une phrase à partir d'un seul mot clé.[41]
- **Plusieurs à plusieurs :** Le modèle utilise plusieurs entrées pour prévoir plusieurs sorties. Par exemple, vous pouvez créer un traducteur de langue avec un RNN, qui analyse une phrase et structure correctement les mots dans une autre langue.[41]
- **Plusieurs à un :** Plusieurs entrées sont mappées vers une sortie. Cela est utile dans des applications telles que l'analyse des sentiments, où le modèle prédit les sentiments positifs, négatifs et neutres des clients à partir des témoignages entrants.[41]

4.4 Indicateurs clés de performance (KPI)

Pour comparer les modèles, nous devons utiliser la même unité de mesure pour évaluer chaque modèle et sélectionner le plus efficace pour le contexte correspondant. En apprentissage automatique, on utilise des indicateurs tels que la précision, l'exactitude, le rappel (la sensibilité), la spécificité et le F1-score pour évaluer la performance du modèle. De plus, on peut également utiliser des indicateurs comme l'erreur quadratique moyenne (MSE), l'erreur absolue moyenne (MAE), le coefficient de détermination (R^2) et l'erreur absolue pourcentage moyenne (MAGE).

4.4.1 Matrice de confusion

Pour comprendre le calcul de ces indicateurs, nous devons d'abord examiner une matrice de confusion.

		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

Figure 35: Matrice de confusion.[42]

Une matrice de confusion représente la performance prédictive d'un modèle sur un ensemble de données. La matrice de confusion comporte quatre composants essentiels :

- **Vrais Positifs (TP)** : Nombre d'échantillons correctement prédits comme positifs.
- **Faux Positifs (FP)** : Nombre d'échantillons incorrectement prédits comme positifs.
- **Vrais Négatifs (TN)** : Nombre d'échantillons correctement prédits comme négatifs.
- **Faux Négatifs (FN)** : Nombre d'échantillons incorrectement prédits comme négatifs.

En utilisant les composants de la matrice de confusion, nous pouvons définir les différentes métriques utilisées pour évaluer les classificateurs : l'exactitude, la précision, le rappel (sensibilité), la spécificité et le F1_score.[42]

4.4.2 Exactitude (Accuracy)

Il s'agit de la mesure des échantillons correctement classifiés par rapport au nombre total d'échantillons de test. Moyenné pour chaque classe, cela fournit une mesure de l'exactitude de l'ensemble du classificateur.[43]

$$\text{Exactitude} = \frac{\text{Vrais Positifs} + \text{Vrais Négatifs}}{\text{Total(Positifs} + \text{Négatifs)}}$$

4.4.3 Précision

C'est la capacité du classificateur à ne pas étiqueter un échantillon comme positif s'il devrait être négatif. Il est calculé comme le rapport entre les vrais positifs et la somme des vrais positifs et des faux positifs.[43]

$$\text{Précision} = \frac{\text{Vrais Positifs}}{\text{Vrais Positifs} + \text{Faux Positifs}}$$

4.4.4 Rappel

Le rappel (recall) représente le pourcentage d'échantillons positifs qui ont été correctement étiquetés comme positifs.[43]

$$\text{Rappel} = \frac{\text{Vrais Positifs}}{\text{Vrais Positifs} + \text{Faux Négatifs}}$$

4.4.5 F1-Score

est calculé comme une moyenne harmonique pondérée de la précision et du rappel du classificateur (F1=1 étant le meilleur).[43]

$$\text{F1-Score} = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

4.4.6 Spécificité

La spécificité est la proportion de prédictions négatives correctes parmi toutes les instances négatives réelles. Elle complète le rappel en se concentrant sur la capacité du modèle à identifier correctement les cas négatifs.[43]

$$\text{Spécificité} = \frac{\text{Vrais Négatifs}}{\text{Vrais Négatifs} + \text{Faux Positifs}}$$

4.4.7 Erreur quadratique moyenne (MSE)

L'erreur quadratique moyenne est une métrique utilisée pour déterminer la différence carrée suggérée entre les valeurs anticipées générées par le modèle et les valeurs réelles trouvées dans l'ensemble de données. Elle mesure essentiellement le degré auquel la prédiction s'écarte de la réalité. Statistiquement, l'EQM est calculée par la différence carrée entre chaque valeur anticipée et sa valeur réelle correspondante dans l'ensemble de données complet.[44]

$$\text{MSE} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$$

où y_i est la valeur réelle, \hat{y}_i la valeur prédictive et n le nombre total d'observations.

4.4.8 Erreur absolue moyenne (MAE)

L'erreur absolue moyenne est l'une des métriques les plus simples, qui mesure la différence absolue entre les valeurs réelles et les valeurs prédictives, où "absolue" signifie prendre un nombre comme positif.[45]

Pour comprendre l'EAM, prenons un exemple de régression linéaire, où le modèle trace une ligne de meilleur ajustement entre les variables dépendantes et indépendantes. Pour mesurer l'EAM ou l'erreur de prédiction, nous devons calculer la différence entre les valeurs réelles et les valeurs prédictives. Mais pour trouver l'erreur absolue pour l'ensemble du jeu de données, nous devons calculer la moyenne absolue de l'ensemble des données.

La formule ci-dessous est utilisée pour calculer l'EAM :

$$\text{MAE} = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}$$

où y_i est la valeur réelle, \hat{y}_i la valeur prédictive et n le nombre total d'observations.

4.4.9 Coefficient de détermination (R2)

L'erreur R carré est également connue sous le nom de coefficient de détermination, qui est une autre métrique populaire utilisée pour l'évaluation des modèles de régression. La métrique R carré nous permet de comparer notre modèle avec une base constante afin de déterminer la performance du modèle. Pour choisir la base constante, nous devons prendre la moyenne des données et tracer la ligne à la moyenne.[45]

Le score R carré sera toujours inférieur ou égal à 1, peu importe si les valeurs sont trop grandes ou trop petites.[45]

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

4.4.10 Erreur moyenne absolue en pourcentage (MAPE))

L'Erreur moyenne absolue en pourcentage est une métrique d'évaluation utilisée pour mesurer l'exactitude des prédictions dans divers secteurs, tels que la finance et les prévisions économiques. La MAPE est souvent utilisée comme fonction de perte dans les problèmes de régression et les modèles de prévision en raison de son interprétation intuitive en termes d'erreur relative pour l'évaluation. Également connue sous le nom de déviation absolue pourcentage moyenne (DAPM), La MAPE est définie comme la différence moyenne absolue en pourcentage entre les valeurs prédictives et les valeurs réelles.[46]

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

4.5 Logiciels et langages de programmation

4.5.1 Langages de programmation

4.5.1.1 Python

Python se distingue comme le langage de programmation informatique le plus populaire et le plus utilisé, en particulier dans les domaines de la science des données et de l'apprentissage automatique. De plus, Python est un langage multiplateforme, fonctionnant sans problème sur divers systèmes d'exploitation tels que Windows, macOS et Linux, ce qui en fait un choix idéal pour les développeurs travaillant dans différents environnements.[\[47\]](#)



Figure 36: Le logo de Python.[\[47\]](#)

Python facilite le scripting de commandes système pour automatiser les tâches informatiques. De plus, Python règne en maître en tant que langage de prédilection pour le traitement des Big Data, les calculs mathématiques et l'apprentissage automatique, ce qui en fait le langage préféré pour les projets de science des données.[\[47\]](#)

4.5.1.2 R

Le langage de programmation R est devenu un choix populaire pour les applications d'apprentissage automatique et de science des données en raison de sa large gamme de packages, de sa polyvalence et de sa facilité d'utilisation. R offre une variété de fonctions, de méthodes et d'outils qui simplifient le processus d'implémentation des algorithmes d'apprentissage automatique et d'analyse des données.[\[48\]](#)



Figure 37: Le logo de R.[\[49\]](#)

4.5.2 Logiciels

4.5.2.1 Power BI

Power BI est un ensemble de services logiciels, d'applications et de connecteurs qui œuvrent ensemble pour transformer des sources de données disparates en insights cohérents, visuellement immersifs et interactifs. Les données peuvent être sous forme de feuille de calcul Excel ou de collection, d'entrepôts de données hybrides, locaux ou sur le cloud. Power BI permet de se connecter facilement aux sources de données, de visualiser et de découvrir ce qui est important, et de partager ces informations.[\[50\]](#)

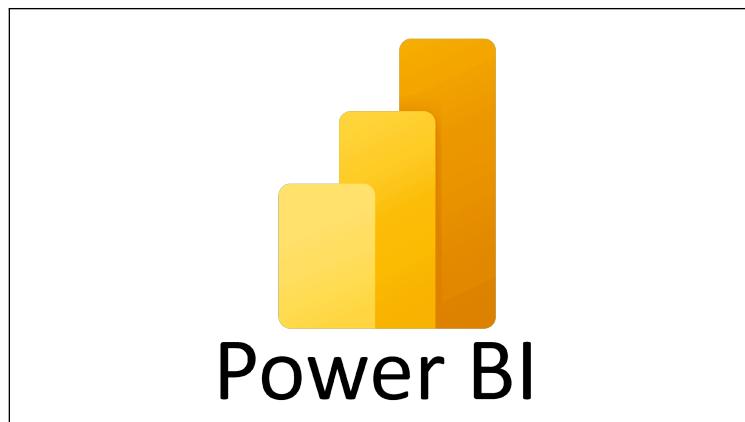


Figure 38: Le logo de Power BI.[\[50\]](#)

Nous allons utiliser ce logiciel afin de créer un tableau de bord visuel pour analyser nos données.

4.5.2.2 Jupyter Notebook

Le Jupyter Notebook est l'application web originale pour créer et partager des documents computationnels. Il offre une expérience simple, épurée et centrée sur le document.[\[51\]](#)



Figure 39: Le logo de Jupyter.[\[51\]](#)

Nous allons utiliser Python sur Jupyter Notebook pour analyser nos données et entraîner nos modèles.

4.6 Conclusion

L'apprentissage automatique (ML) est essentiel pour l'intelligence artificielle moderne, apportant des outils puissants pour analyser des problèmes complexes, comme la prédition météorologique. Ce chapitre a exploré les fondements du ML, en examinant les types d'apprentissage, les algorithmes principaux, les indicateurs de performance et les langages de programmation clés. Les concepts d'apprentissage supervisé, non supervisé et par renforcement, combinés à des indicateurs de performance tels que la précision et le F1-score, permettent de créer des modèles précis et robustes. Avec la croissance exponentielle des données, le ML est de plus en plus vital pour extraire des informations pertinentes, comme dans le domaine de la météorologie, où il aide à automatiser les processus et à améliorer les prédictions, stimulant une innovation continue pour rester à la pointe des technologies météorologiques.

Exploration des données et entraînement des modèles

5.1 Introduction

Dans ce chapitre, nous explorerons les données météorologiques et entraînerons des modèles prédictifs pour la température. En utilisant des données synthétiques de grandes villes américaines, nous analyserons les tendances climatiques et comparerons l'efficacité de différents modèles, tels que les méthodes de régression, les forêts aléatoires et les réseaux de neurones, afin d'identifier celui qui offre les prévisions les plus précises.

5.2 À propos du jeu de données

Ce jeu de données contient des données météorologiques synthétiques générées pour dix grandes villes américaines, notamment **New York**, **Los Angeles**, **Chicago**, **Houston**, **Phoenix**, **Philadelphie**, **San Antonio**, **San Diego**, **Dallas** et **San José**. Chaque ville est associée à un large éventail de paramètres climatiques, tels que **la température**, **l'humidité**, **les précipitations** et **la vitesse du vent**, offrant une vue d'ensemble des variations atmosphériques locales. En tout, un million de points de données ont été générés pour chaque paramètre, permettant ainsi une analyse approfondie et robuste des conditions météorologiques dans ces différents endroits, facilitant les études de modélisation et de prédiction.

5.3 Les attributs

Les données météorologiques analysées sont structurées autour de plusieurs attributs essentiels, chacun jouant un rôle spécifique dans la description des conditions climatiques à un moment et un emplacement donnés. Ces attributs fournissent des informations détaillées sur des variables telles que la température, l'humidité, les précipitations et la vitesse du vent, qui sont cruciales pour comprendre les phénomènes météorologiques et effectuer des analyses prédictives.

- **Location** : La ville où les données météorologiques ont été simulées.
- **Date_Time** : La date et l'heure à laquelle les données météorologiques ont été enregistrées.
- **Temperature_C** : La température en **Celsius** à l'emplacement et au moment donnés.

- **Humidity_pct** : L'humidité en **pourcentage** à l'emplacement et au moment donnés.
- **Precipitation_mm** : Les précipitations en **millimètres** à l'emplacement et au moment donnés.
- **Wind_Speed_kmh** : La vitesse du vent en **kilomètres par heure** à l'emplacement et au moment donnés.

5.4 Informations supplémentaires

- **Variabilité et complexité** : Le jeu de données intègre la variabilité et la complexité pour simuler des modèles météorologiques réalistes. Par exemple, des ajustements ont été effectués sur la température et les précipitations en fonction des variations saisonnières observées dans certaines régions. À New York, des températures et des précipitations plus élevées sont simulées pendant les mois d'été, tandis qu'à Phoenix, des températures plus basses et des précipitations accrues sont simulées durant les mois d'hiver.
- **Méthode de génération des données** : Le jeu de données a été généré à l'aide de la bibliothèque Faker de Python pour créer des données météorologiques synthétiques pour chaque emplacement. Des valeurs aléatoires dans des plages réalistes ont été générées pour la température, l'humidité, les précipitations et la vitesse du vent, avec des ajustements effectués pour refléter les variations saisonnières.

La table 8 présente les premières lignes du jeu de données, offrant ainsi un aperçu de sa structure.

Table 8: Les premières lignes du jeu de données.

	Location	Date_Time	Temperature_C	Humidity_pct	Precipitation_mm	Wind_Speed_kmh
0	San Diego	2024-01-14 21:12:46	10.683001	41.195754	4.020119	8.233540
1	San Diego	2024-05-17 15:22:10	8.734140	58.319107	9.111623	27.715161
2	San Diego	2024-05-11 09:30:59	11.632436	38.820175	4.607511	28.732951
3	Philadelphia	2024-02-26 17:32:39	-8.628976	54.074474	3.183720	26.367303
4	San Antonio	2024-04-29 13:23:51	39.808213	72.899908	9.598282	29.898622

5.5 Analyse descriptive

Cette section se concentre sur l'analyse exploratoire des données pour mieux les comprendre et s'y familiariser. Elle fournit des informations statistiques telles que l'écart type, les valeurs maximales et minimales, ainsi que les distributions de chaque variable. De plus, elle inclut une analyse des corrélations entre les variables, ce qui nous aide à sélectionner les modèles à entraîner et à déterminer la méthodologie à suivre. En parallèle, une analyse des corrélations entre les variables est effectuée afin de mieux comprendre les relations sous-jacentes. Cette exploration des interdépendances entre les paramètres

météorologiques est essentielle pour guider la sélection des modèles à entraîner et affiner la méthodologie d'apprentissage automatique, en identifiant les variables les plus pertinentes et en évaluant les interactions qui pourraient affecter la performance prédictive des modèles.

5.5.1 Résumé du jeu des données

La table 9 présente des statistiques descriptives pour le jeu de données météorologiques comprenant un million d'observations.

Ces statistiques fournissent un aperçu général de la répartition des valeurs pour chaque variable, ce qui est essentiel pour comprendre les tendances climatiques et évaluer les conditions météorologiques sur une période étendue.

Table 9: Statistiques descriptives.

	Temperature_C	Humidity_pct	Precipitation_mm	Wind_Speed_kmh
count	1000000.000000	1000000.000000	1000000.000000	1000000.000000
mean	14.779705	60.021830	5.109639	14.997598
std	14.482558	17.324022	2.947997	8.663556
min	-19.969311	30.000009	0.000009	0.000051
25%	2.269631	45.008500	2.580694	7.490101
50%	14.778002	60.018708	5.109917	14.993777
75%	27.270489	75.043818	7.613750	22.514110
max	39.999801	89.999977	14.971583	29.999973

Les statistiques descriptives révèlent plusieurs informations clés. La température moyenne est de 14,78°C, avec une large gamme de valeurs allant de -19,97°C à 39,99°C, indiquant une grande variabilité climatique dans l'échantillon. L'humidité varie de 30% à près de 90%, avec une médiane de 60%, reflétant des conditions allant de modérément sèches à très humides. Les précipitations moyennes sont de 5,11 mm, avec des variations notables, et la vitesse du vent oscille principalement entre 0 et 30 km/h, avec une moyenne de 15 km/h. Ces données montrent la diversité des conditions météorologiques capturées et soulignent l'importance de ces variables dans l'analyse climatologique.

Maintenant, nous allons décomposer la variable Date_Time en plusieurs composantes distinctes : l'année, le mois, le jour, l'heure, les minutes et les secondes. Cette décomposition est essentielle pour permettre une analyse plus granulaire des données temporelles, en offrant la possibilité d'examiner les effets saisonniers, les variations intra-journalières et les cycles annuels. En segmentant la variable temporelle, nous pouvons plus facilement identifier des motifs récurrents, comme des fluctuations de température en fonction des saisons ou des changements dans la vitesse du vent à des heures spécifiques de la journée. Cette étape permet également de créer des variables supplémentaires qui peuvent être utilisées comme des caractéristiques importantes dans les modèles prédictifs. Par exemple, l'extraction de l'heure ou du mois pourrait révéler des corrélations entre certains événements météorologiques et des périodes particulières, facilitant ainsi une analyse plus détaillée des tendances et des schémas cycliques au fil du temps.

```

1 BD['Date_Time'] = pd.to_datetime(BD['Date_Time'], utc=True)
2 BD1 = BD.copy()
3 BD1['Year'] = BD1['Date_Time'].dt.year
4 BD1['Month'] = BD1['Date_Time'].dt.month
5 BD1['Day'] = BD1['Date_Time'].dt.day
6 BD1['Hour'] = BD1['Date_Time'].dt.hour
7 BD1['Minute'] = BD1['Date_Time'].dt.minute
8 BD1['Second'] = BD1['Date_Time'].dt.second
9 BD1.drop(columns=['Date_Time'], inplace=True)
10 nw_columns=['Year', 'Month', 'Day', 'Hour', 'Minute', 'Second']
11 column_order = nw_columns + [col for col in BD1.columns if col not in nw_columns]
12 BD1 = BD1[column_order]
13 BD1.head()

```

Figure 40: Décomposition de la variable Date_Time.

Table 10: La nouvelle base de données.

	Year	Month	Day	Hour	Minute	Second	Location	Temperature_C	Humidity_pct	Precipitation_mm	Wind_Speed_kmh
0	2024	1	14	21	12	46	San Diego	10.683001	41.195754	4.020119	8.233540
1	2024	5	17	15	22	10	San Diego	8.734140	58.319107	9.111623	27.715161
2	2024	5	11	9	30	59	San Diego	11.632436	38.820175	4.607511	28.732951
3	2024	2	26	17	32	39	Philadelphia	-8.628976	54.074474	3.183720	26.367303
4	2024	4	29	13	23	51	San Antonio	39.808213	72.899908	9.598282	29.898622

Poursuivons l'exploration de notre jeu de données pour approfondir l'analyse.

```

1 BD1['Month'].value_counts()

1 223290
3 223072
4 216804
2 208478
5 128356
Name: Month, dtype: int64

```

Figure 41: Les différentes valeurs de la variable Mois.

Le jeu de données s'étend donc sur une période de cinq mois, commençant au mois de janvier et se poursuivant jusqu'au mois de mai. Cette période permet d'analyser les variations saisonnières et d'identifier les tendances sur plusieurs mois consécutifs. L'étendue de ces enregistrements offre une base solide pour des analyses approfondies et des modèles prédictifs couvrant une large gamme de conditions temporelles.

Passons maintenant à l'exploration des différentes villes représentées dans notre jeu de données. Cette étape nous permettra de mieux comprendre la répartition géographique des données et d'identifier les spécificités climatiques propres à chaque localité. L'analyse des données par ville peut révéler des variations régionales significatives et offrir des perspectives supplémentaires sur les tendances météorologiques observées.

1	BD1['Location'].value_counts()
	Phoenix 100209
	Chicago 100164
	Philadelphia 100122
	Houston 100076
	New York 99972
	San Antonio 99962
	Dallas 99936
	Los Angeles 99922
	San Jose 99863
	San Diego 99774
	Name: Location, dtype: int64

Figure 42: Les différentes villes incluses dans le jeu de données.

Notre jeu de données inclut des enregistrements provenant de dix villes (Phoenix, Chicago, Philadelphie, Houston, New York, San Antonio, Dallas, Los Angeles, San José, et San Diego). Comme mentionné dans la description de la base de données, le nombre d'observations est presque le même pour chacune de ces villes.

5.5.2 Dashboard pour l'analyse du jeu de données

Nous allons à présent nous concentrer sur l'analyse détaillée des différentes variables présentes dans le jeu de données, en suivant leur évolution au fil du temps, avec une ventilation par ville et par mois. Cette approche nous permettra d'identifier des tendances temporelles et géographiques dans les phénomènes météorologiques. Avant de procéder à une analyse approfondie des données avec Python, nous allons d'abord concevoir un tableau de bord interactif à l'aide de Power BI. Ce tableau de bord servira à fournir une vue d'ensemble intuitive et visuelle des tendances observées pour chaque variable.

Le tableau de bord sera structuré en quatre feuilles d'analyse distinctes, chacune dédiée à une variable météorologique clé : la température, l'humidité, les précipitations et la vitesse du vent. Chaque feuille présentera l'évolution de la variable correspondante sous forme de graphiques interactifs, permettant de filtrer et d'explorer les données par ville, par période, ou par plage temporelle spécifique. Une barre de navigation intuitive facilitera la transition entre les différentes feuilles d'analyse, permettant ainsi une exploration fluide et complète des données. En intégrant des éléments interactifs comme des curseurs de temps, des graphiques dynamiques, et des options de filtrage personnalisées, ce tableau de bord offrira une visualisation claire des tendances et des variations climatiques, constituant une étape préliminaire essentielle pour orienter l'analyse plus poussée avec des méthodes de machine learning.

La figure 43 montre l'évolution de la température moyenne à Houston au cours du mois de janvier. On observe que la température moyenne fluctue autour de 15 degrés Celsius. La température maximale enregistrée atteint 40 degrés Celsius, tandis que la température minimale descend jusqu'à -10 degrés Celsius.

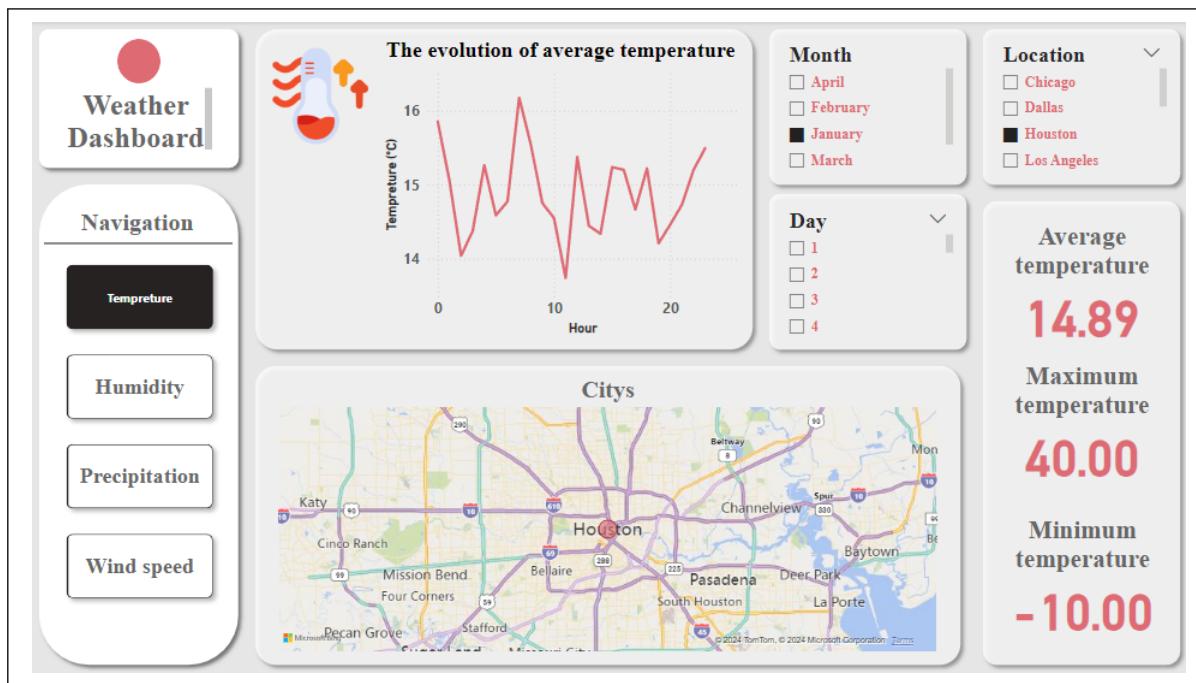


Figure 43: La température.

La figure 44 illustre l'évolution de l'humidité à Los Angeles pour le mois de février. Entre 2h et 8h, l'humidité présente de légères fluctuations avec une tendance générale à la hausse, indiquant une accumulation de l'humidité nocturne. De 9h à 20h, l'humidité oscille autour d'une valeur moyenne de 59,68%, reflétant des conditions relativement stables pendant la journée. En début de nuit, entre 20h et 23h, on observe une augmentation presque linéaire de l'humidité, potentiellement due à la baisse des températures. La valeur maximale d'humidité enregistrée à Los Angeles en février est de 89,99%, tandis que la valeur minimale est de 30%.

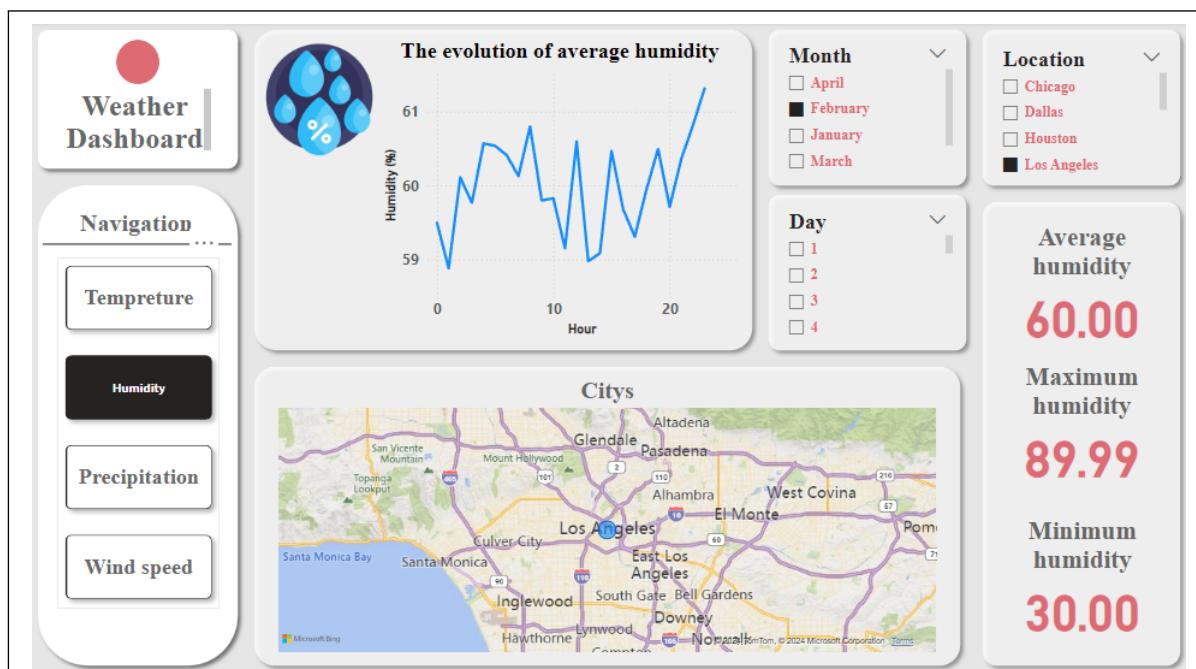


Figure 44: L'humidité.

La figure 45 illustre l'évolution des précipitations pour le 2 février à Dallas. Les précipitations présentent de fortes fluctuations autour de 5 mm. La valeur maximale enregistrée est de 9,9 mm, tandis que la valeur minimale est de 0,03 mm.

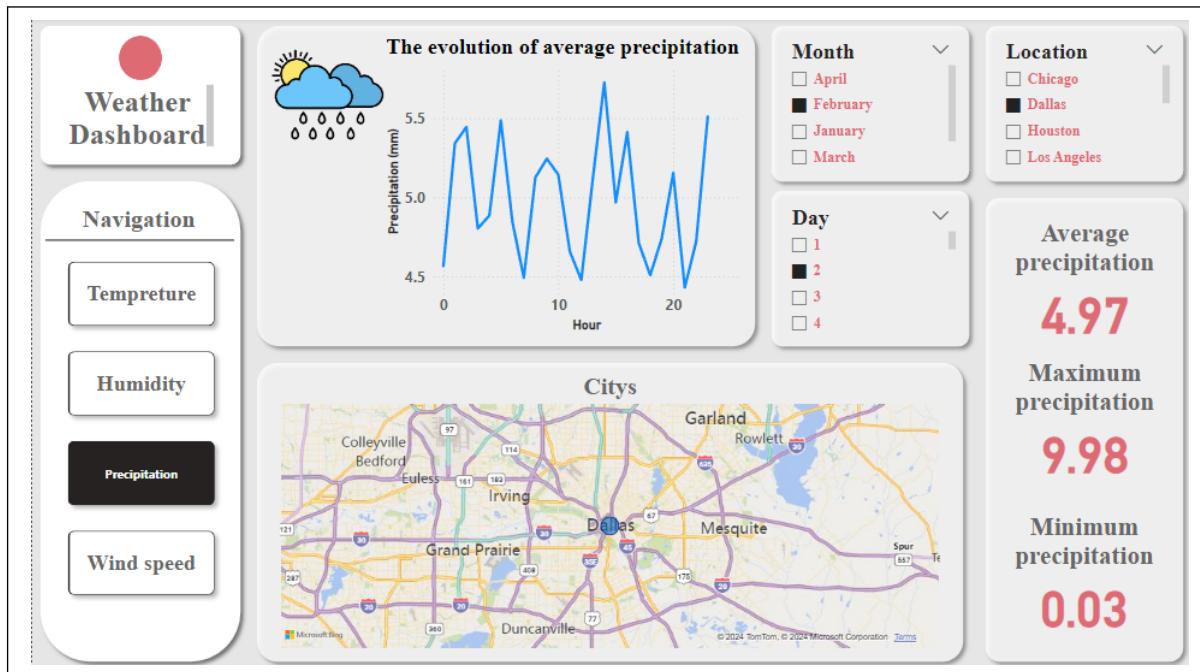


Figure 45: La précipitation.

La figure 46 illustre l'évolution de la vitesse moyenne du vent pour le 2 avril à Houston. La vitesse du vent fluctue tout au long de la journée, avec une valeur minimale de 10,29 km/h enregistrée vers 10h du matin.

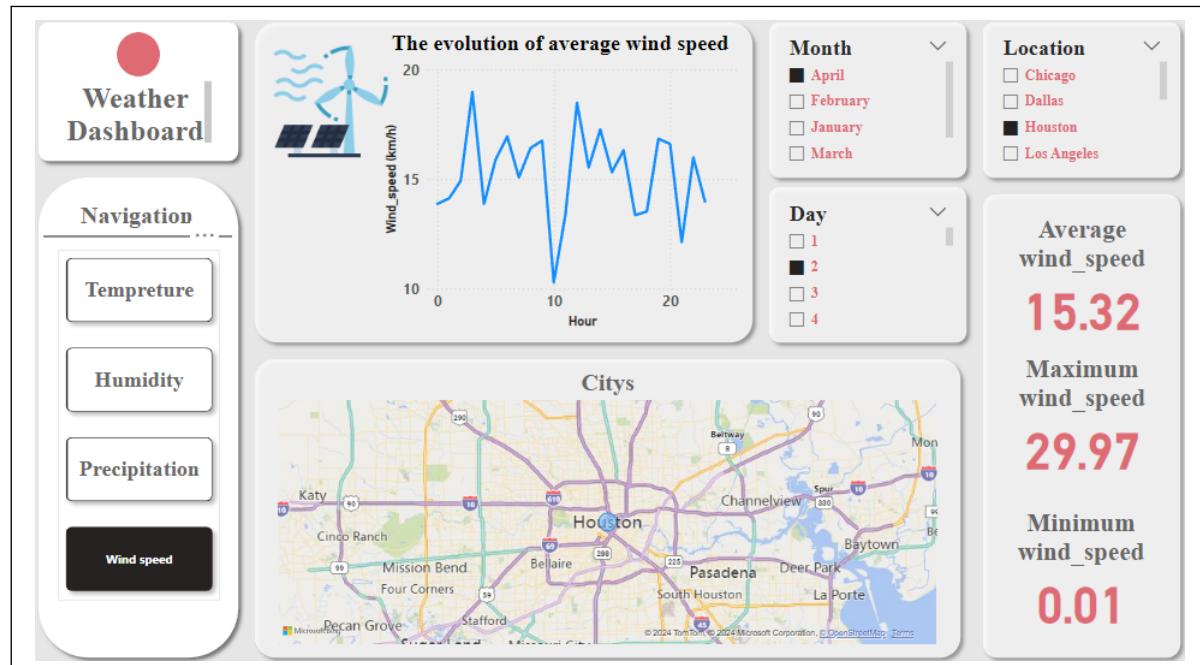


Figure 46: La vitesse du vent.

5.5.3 Les distributions des attributs

Nous allons examiner les différentes distributions afin d'obtenir une meilleure compréhension des variations de ces attributs.

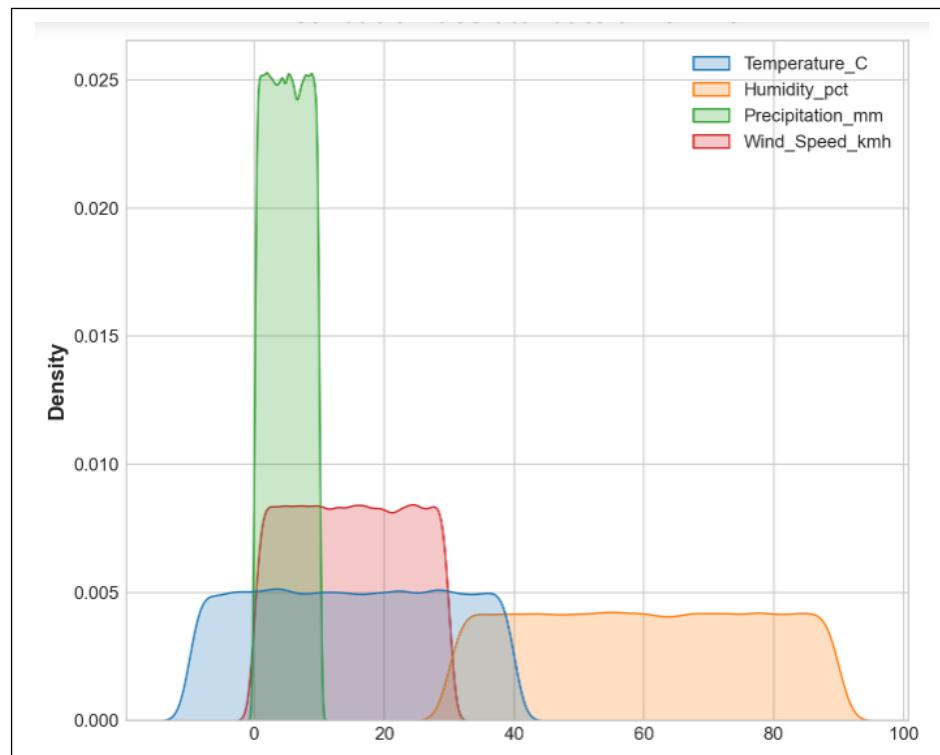


Figure 47: Distribution des attributs à New York.

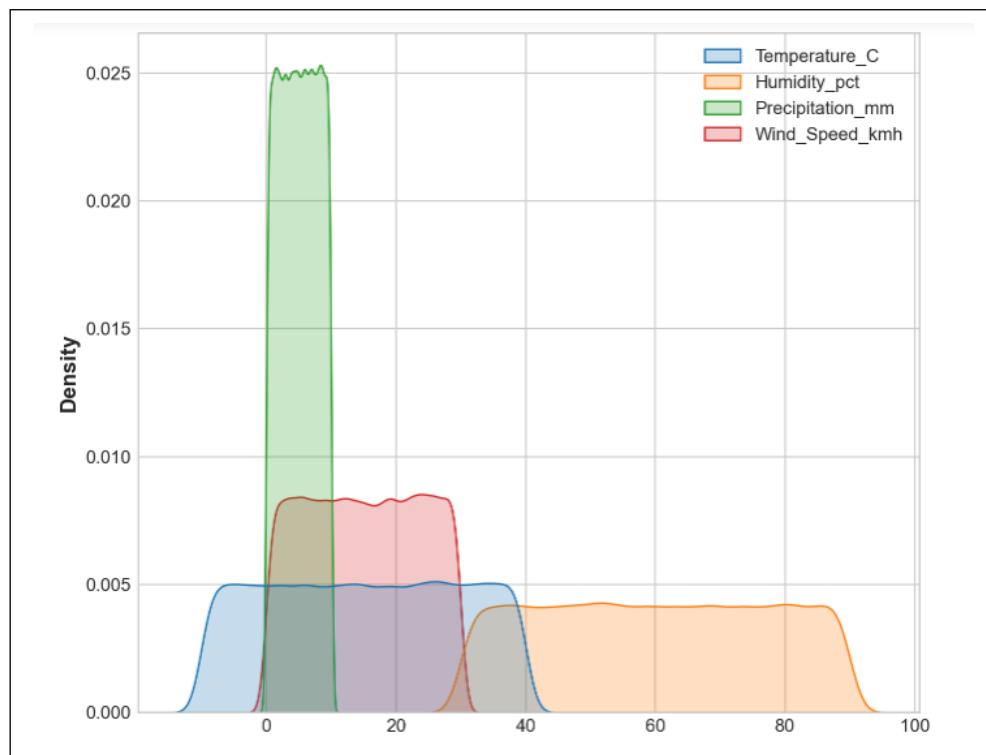


Figure 48: Distribution des attributs à Los Angeles.

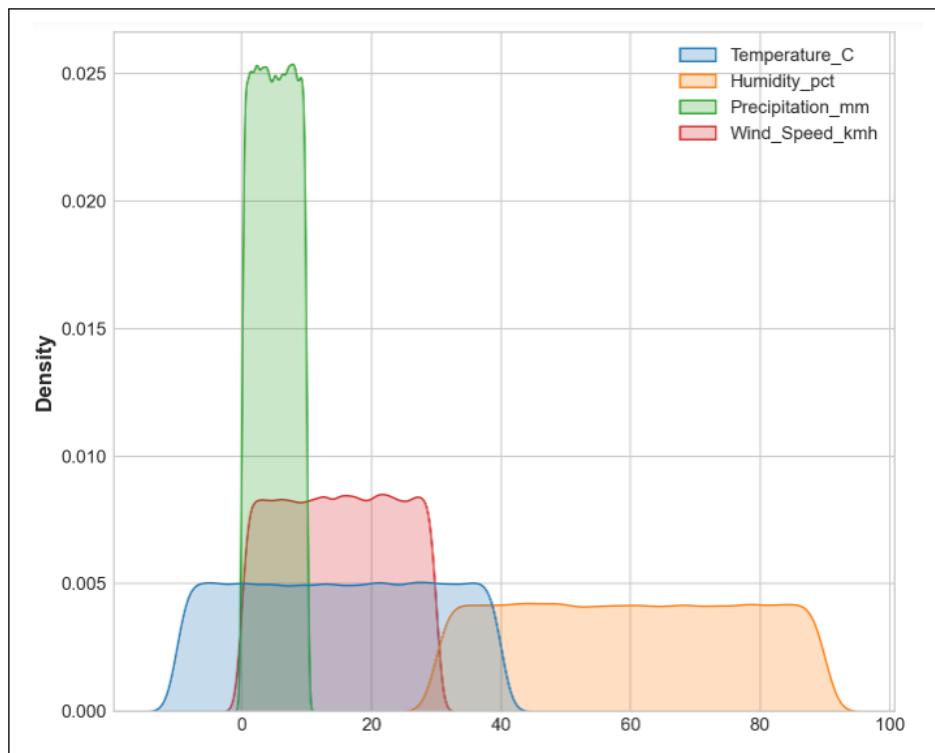


Figure 49: Distribution des attributs à Dallas.

Nous pouvons constater que les attributs suivent des distributions presque similaires dans les différentes régions explorées. Cette homogénéité des distributions pourrait indiquer des facteurs environnementaux ou climatiques communs qui influencent les attributs mesurés. Par exemple, si les variations d'humidité sont semblables dans plusieurs zones, cela pourrait suggérer que ces régions partagent des caractéristiques géographiques, telles que la proximité de plans d'eau ou des conditions climatiques similaires, qui impactent l'humidité ambiante. De plus, cette similitude peut également refléter des comportements humains semblables, tels que l'utilisation de l'eau ou les pratiques agricoles, qui pourraient uniformiser les attributs observés.

Examinons si la distribution de la température à New York suit une loi normale en utilisant le test de Kolmogorov Smirnov.

```

1 from scipy.stats import kstest
2 ks_statistic, p_value = kstest( region_analyser('New York')['Temperature_C'], 'norm')
3 print(f"{'Temperature_C'} - KS Statistic: {ks_statistic:.4f}, P-value: {p_value:.4f}")

Temperature_C - KS Statistic: 0.7451, P-value: 0.0000

```

Figure 50: Test de Kolmogorov Smirnov.

Statistique KS (0.7451) : Cette valeur indique la distance maximale entre la distribution empirique des données et la distribution normale théorique. Une valeur élevée comme 0.7451 suggère une différence significative entre les deux distributions.

P-valeur (0.0000) : La p-valeur étant très proche de zéro, cela signifie que nous rejetons l'hypothèse nulle selon laquelle les données suivent une distribution normale.

Maintenant, nous allons approfondir l'analyse pour chaque région individuellement. Pour ce faire, nous allons créer une fonction permettant d'isoler les données de chaque région.

```

1 def region_analyser(region):
2     data=BD[df1['Location']==region]
3     data=data.sort_values('Date_Time')
4     data=data.reset_index()
5     data=data.drop('index',axis=1)
6     data=data.drop('Location',axis=1)
7     data=data.set_index('Date_Time')
8     return data

```

Figure 51: Region_analyser.

Appliquons cette fonction à Philadelphie pour examiner les données spécifiques à cette région.

Table 11: Les données spécifiques de Philadelphie.

	region_analyser('Philadelphia')			
	Temperature_C	Humidity_pct	Precipitation_mm	Wind_Speed_kmh
Date_Time				
2024-01-01 00:02:01+00:00	14.053927	36.819607	7.674568	26.373158
2024-01-01 00:04:08+00:00	20.338468	83.716826	0.433529	6.721690
2024-01-01 00:04:17+00:00	33.151358	43.907956	2.468438	2.805796
2024-01-01 00:07:11+00:00	27.492741	62.202252	7.471210	14.933386
2024-01-01 00:07:27+00:00	5.981641	53.821122	6.865934	20.463696
...
2024-05-18 19:41:01+00:00	7.870642	73.829501	7.753614	0.281414
2024-05-18 19:41:39+00:00	29.127268	77.020368	8.510493	16.058369
2024-05-18 19:42:21+00:00	15.650388	32.674892	6.435619	7.985720
2024-05-18 19:42:44+00:00	2.868892	79.067609	6.838498	1.300156
2024-05-18 19:43:50+00:00	8.597967	88.073644	5.137802	21.671935

100122 rows x 4 columns

Nous créons aussi une fonction pour voir les variations des attributs dans chaque région.

```

1 def feature_analyzer(region,feature,rows):
2     data=region_analyser(region)
3
4     plt.figure(figsize=(15, 6))
5     plt.subplot(2,1,1)
6     data[feature].head(rows).plot()
7     plt.axhline(data[feature].mean(),color='k',linestyle='--',linewidth=2.5,label='mean')
8     plt.axhline(data[feature].median(),color='g',linestyle='--',linewidth=2.5,label='median')
9     plt.axhline(data[feature].mode()[0],color='m',linestyle='--',linewidth=3,label='mode')
10    plt.legend(loc='best')
11    plt.grid(True)
12
13    plt.figure(figsize=(10, 6))
14    plt.subplot(2,1,2)
15    sns.kdeplot(data[feature].head(rows))
16    plt.axvline(data[feature].mean(),color='k',linestyle='--',linewidth=2.5,label='mean')
17    plt.axvline(data[feature].median(),color='g',linestyle='--',linewidth=2.5,label='median')
18    plt.axvline(data[feature].mode()[0],color='m',linestyle='--',linewidth=3,label='mode')
19    plt.legend(loc='best')
20    plt.grid(True)
21
22
23    plt.tight_layout()
24    plt.show()

```

Figure 52: Feature_analyzer

Appliquons cette fonction à Philadelphie et New York pour visualiser l'évolution et la distribution des précipitations et de la température, respectivement. Nous allons nous concentrer uniquement sur les 150 premières lignes de notre jeu de données, étant donné que celui-ci est volumineux.

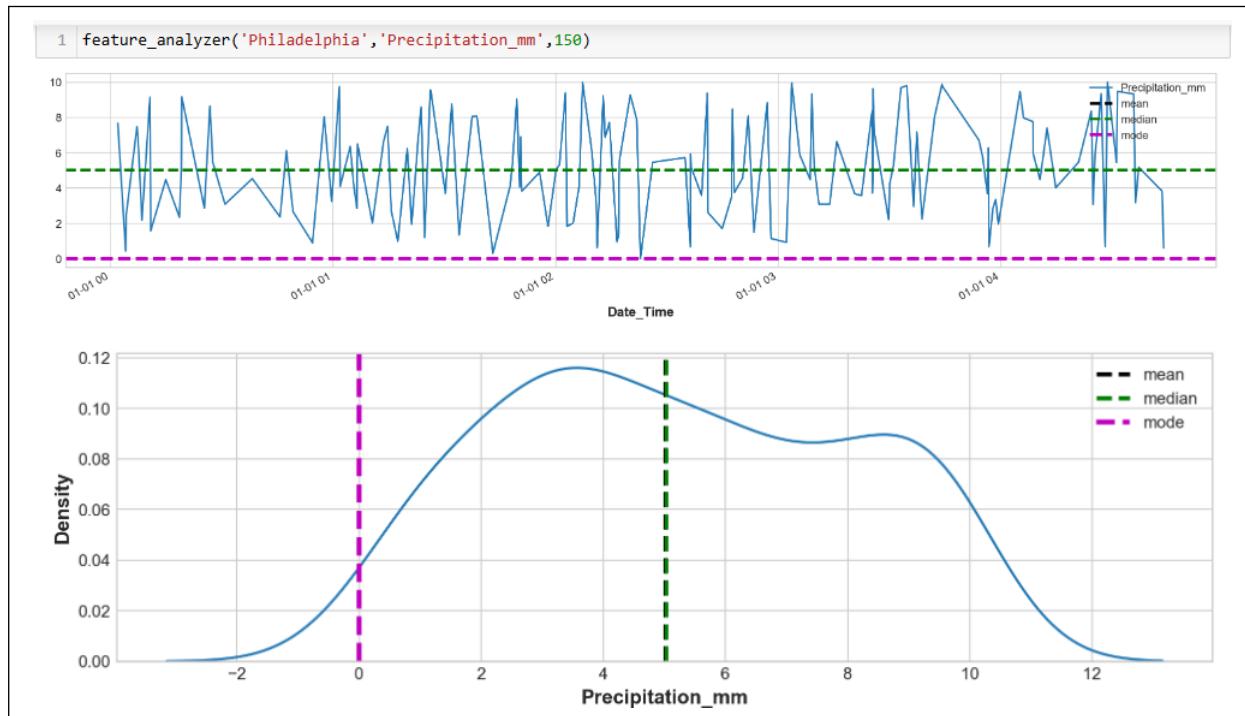


Figure 53: Analyse des précipitations en Philadelphie.

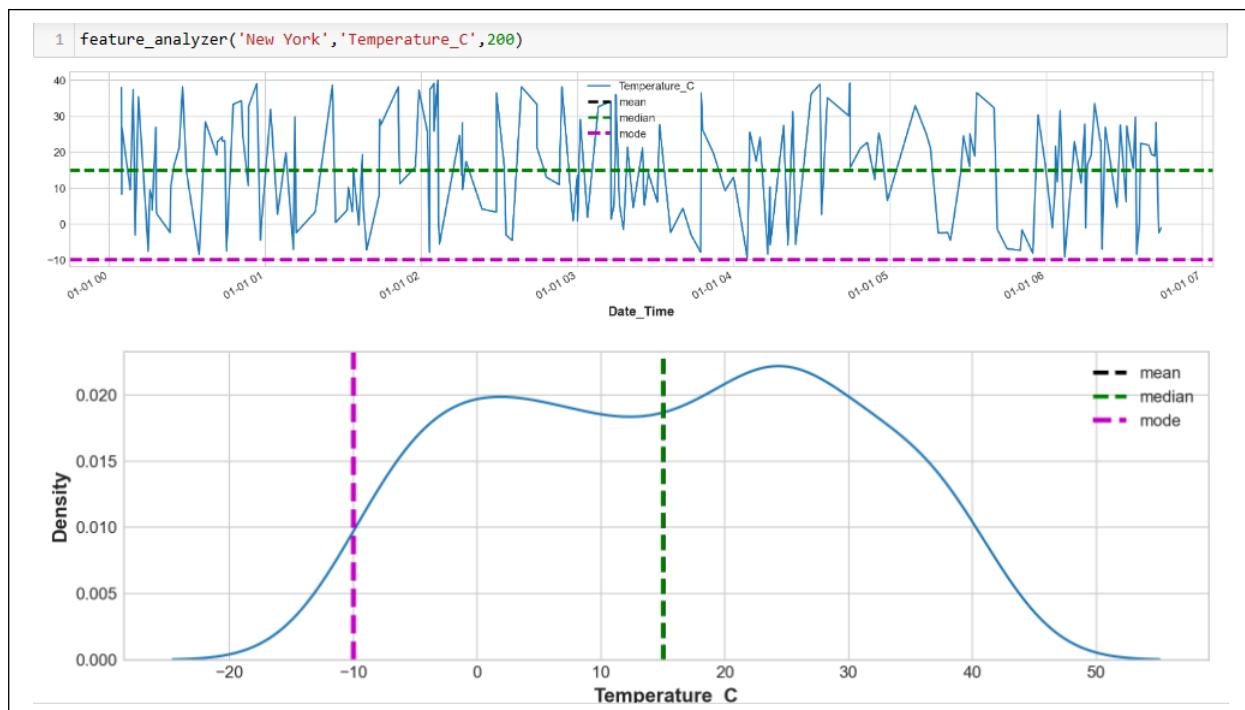


Figure 54: Analyse de la température à New York.

5.5.4 Analyse de corrélations entre les différents attributs

Dans cette section, nous allons explorer les corrélations entre les différents attributs de notre jeu de données, tels que la température, l'humidité, les précipitations et la vitesse du vent. L'analyse de corrélation permet de comprendre les relations linéaires entre ces variables, ce qui peut offrir des insights précieux sur leur interaction. Pour visualiser ces corrélations, nous utiliserons un heatmap, qui illustrera de manière claire et intuitive les forces et les directions des relations entre les attributs étudiés.

```

1 df_subset = BD[['Temperature_C', 'Humidity_pct', 'Precipitation_mm', 'Wind_Speed_kmh']]

1 correlation_matrix = df_subset.corr()

1 plt.figure(figsize=(7, 5)) # Taille de la figure
2 sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', vmin=-1, vmax=1)
3 #plt.title('Heatmap de corrélations entre les variables météorologiques')
4 plt.show()

```

Figure 55: Création du heatmap.

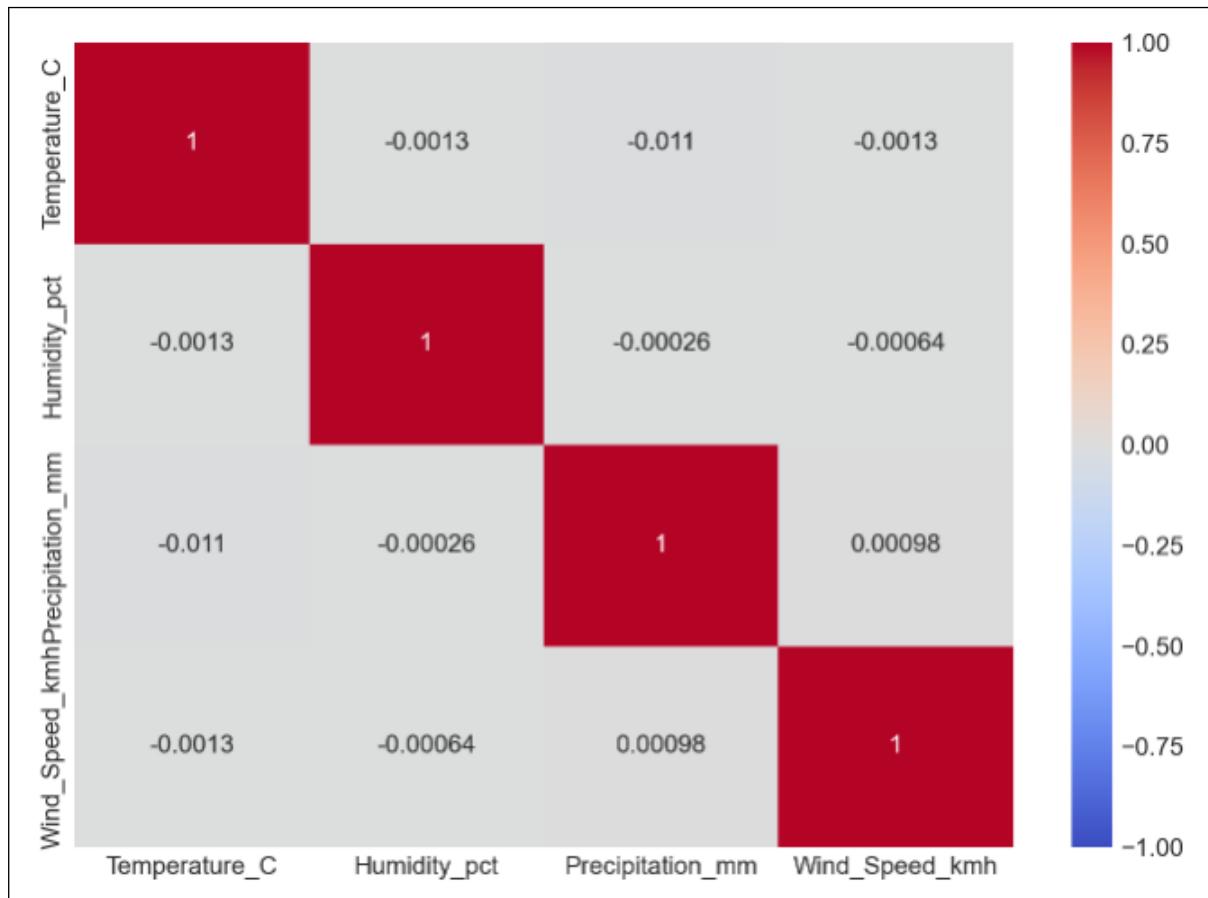


Figure 56: Heatmap de corrélations entre les différents attributs.

Les valeurs des corrélations sont toutes très proches de zéro, ce qui indique une absence de relation linéaire significative entre les variables. En d'autres termes, ces variables semblent être indépendantes les unes des autres, ne présentant pas de corrélations notables.

5.6 Entraînement des modèles

Dans cette section, nous nous concentrerons exclusivement sur la prédiction de la température, étant donné que le processus est similaire pour les autres paramètres. De plus, nous limiterons notre analyse à la zone de New York, en nous focalisant sur la prédiction de la température moyenne journalière.

5.6.1 La régression linéaire

Le modèle de régression linéaire a montré une performance modérée (figure 57) dans la prédiction de la température moyenne journalière pour la zone de New York.

Mean Squared Error (MSE): 0.3414048099589692
RMSE: 0.5842985623454581
Mean Absolute Error (MAE): 0.47210107477012414

Figure 57: Linear Regression-KPI.

Le RMSE obtenu est de 0.58, ce qui indique une erreur moyenne relativement faible ($\pm 0.58^{\circ}\text{C}$) entre les valeurs prédites et les valeurs réelles. Quant au MAE, il est de 0.47, suggérant que la différence moyenne absolue entre les prédictions et les observations est limitée, mais laisse encore place à des améliorations.

Dans la figure 58, nous visualisons la droite de régression avec les prédictions.

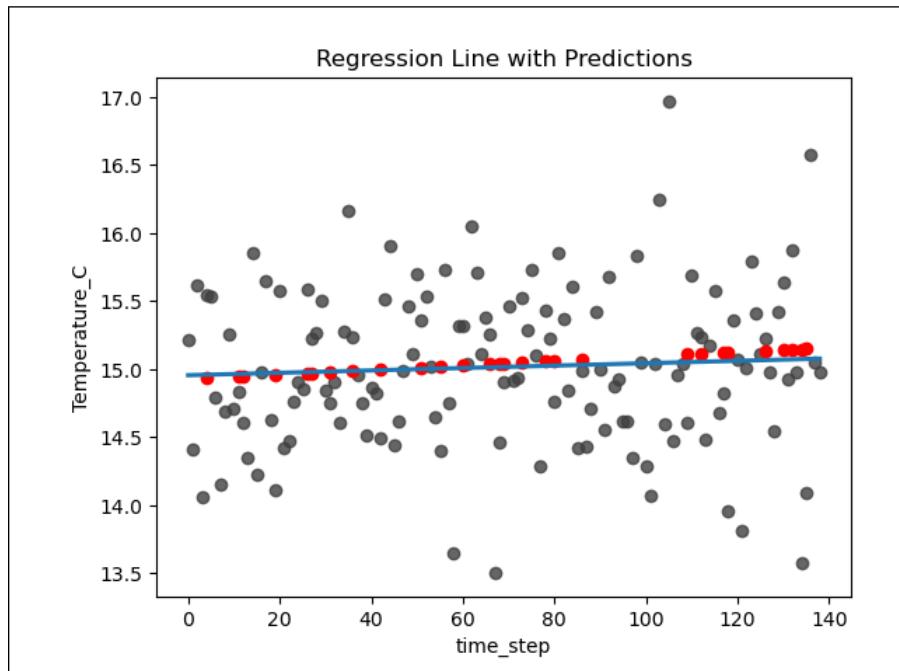


Figure 58: Droite de régression avec prédictions.

5.6.2 Forêt aléatoire

Le modèle de forêt aléatoire a montré une performance prometteuse (figure 59) dans la prédiction de la température moyenne journalière pour la zone de New York. Avec un

RMSE de 0.72, il indique une erreur moyenne relativement faible entre les valeurs prédictes et les valeurs réelles, ce qui suggère une bonne capacité d'ajustement aux données. Par ailleurs, le MAE enregistré est de 0.53, ce qui souligne que la différence moyenne absolue entre les prédictions et les observations est limitée. Ces résultats témoignent de l'efficacité du modèle de forêt aléatoire dans la capture des tendances sous-jacentes des données, tout en suggérant un potentiel d'amélioration continue pour atteindre une précision optimale.

Mean Squared Error (MSE): 0.524460664948251
RMSE: 0.7241965651314918
Mean Absolute Error (MAE): 0.5353171565194094

Figure 59: Random Forest-KPI.

Dans la figure 60, nous visualisons les prédictions du modèle de forêt aléatoire.

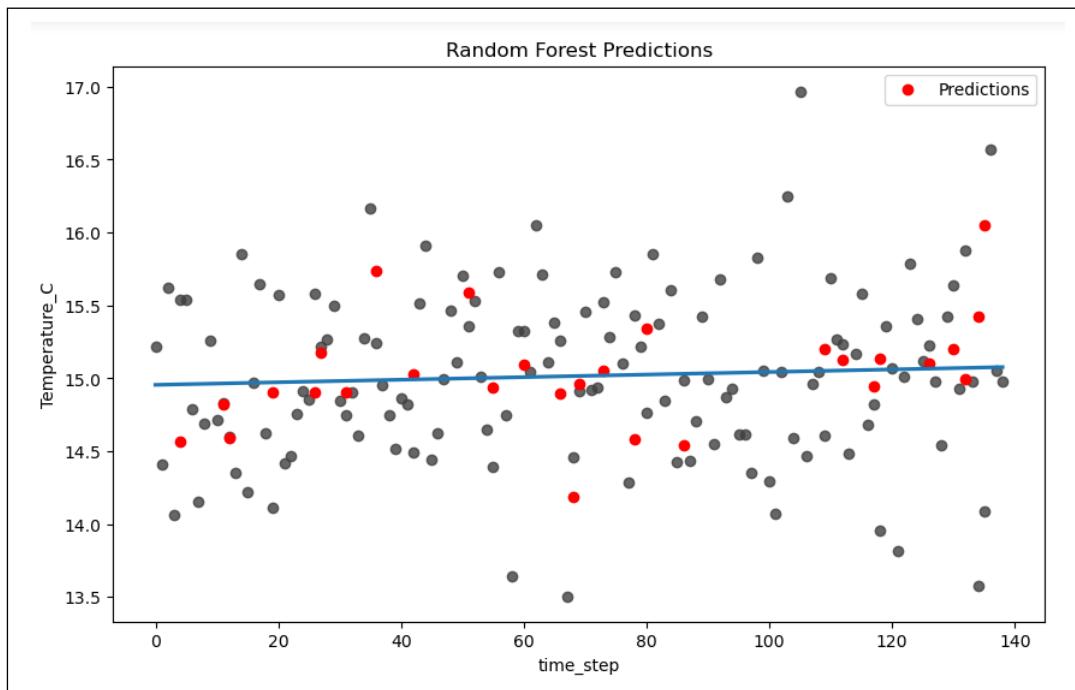


Figure 60: Prédiction du modèle de forêt aléatoire.

5.6.3 XGBoost

Le modèle XGBoost a révélé une capacité impressionnante (figure 61) à prédire la température moyenne journalière pour la zone de New York. Avec un RMSE de 0.6, ce modèle montre une gestion efficace des erreurs, ce qui indique que les prévisions sont généralement très proches des valeurs réelles. En outre, le MAE obtenu est de 0.47, ce qui témoigne d'une différence moyenne absolue relativement faible entre les prédictions et les observations.

Mean Squared Error (MSE): 0.3697499689629767
RMSE: 0.6080706940504341
Mean Absolute Error (MAE): 0.47097662332841594

Figure 61: XGBoost-KPI.

La figure 62 présente les prédictions du modèle XGBoost comparées aux valeurs réelles.

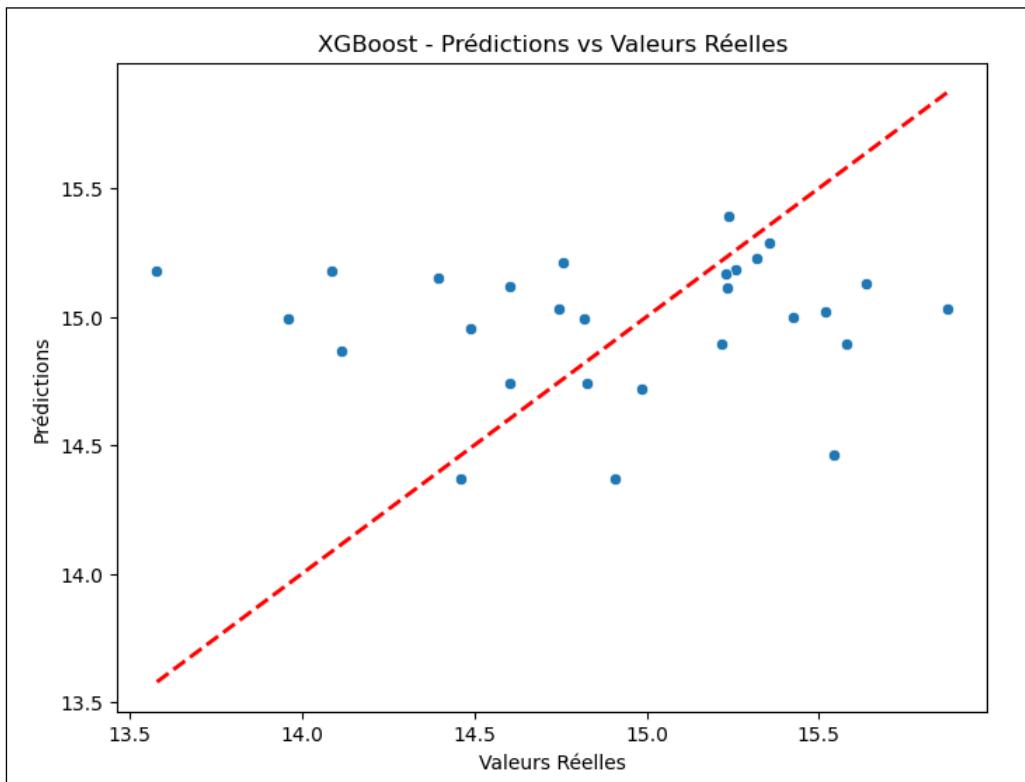


Figure 62: XGBoost.

5.6.4 Les réseaux de neurones artificiels (ANN)

Le modèle de réseau de neurones artificiels (ANN) a démontré une performance remarquable (figure 63) dans la prédiction de la température moyenne journalière pour la zone de New York. Avec un RMSE de 0.65, ce modèle illustre une gestion efficace des erreurs, indiquant que les prévisions sont généralement très proches des valeurs réelles. De plus, le MAE obtenu est de 0.49, ce qui souligne une différence moyenne absolue relativement faible entre les prédictions et les observations. Ces résultats renforcent l'idée que l'ANN est capable de capturer les tendances sous-jacentes des données et d'offrir des prévisions précises.

Mean Squared Error (MSE): 0.4342910770448029
RMSE: 0.6590076456649064
Mean Absolute Error (MAE): 0.49387545481766654

Figure 63: Artificial Neural Networks-KPI.

La figure 64 illustre la comparaison entre les prédictions générées par le modèle de réseau de neurones artificiels (ANN) et les valeurs réelles de la température moyenne journalière pour la zone de New York. Cette visualisation permet d'évaluer l'exactitude des prévisions du modèle en mettant en évidence les points de correspondance et les divergences éventuelles par rapport aux observations réelles. L'analyse de cette figure fournit des insights précieux sur la performance globale du modèle et son efficacité à capturer les variations de la température.

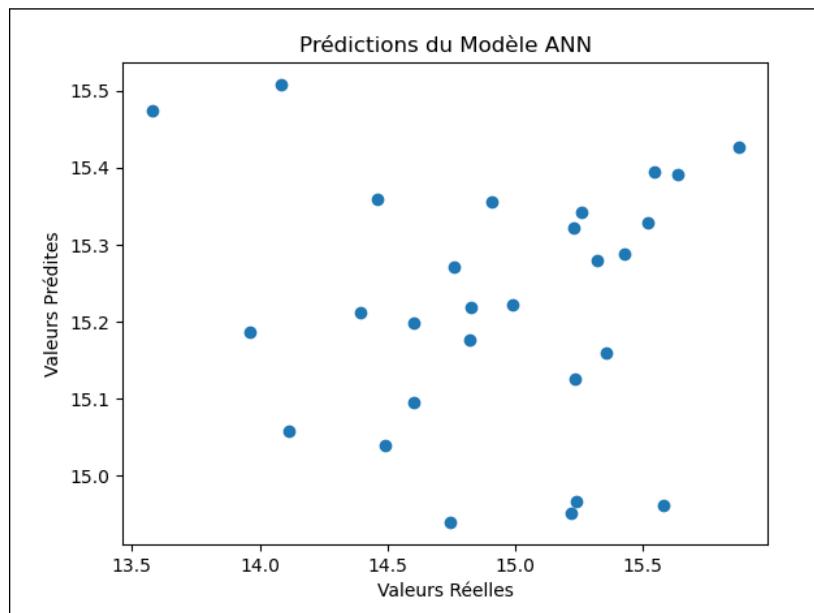


Figure 64: Artificial Neural Networks.

5.6.5 Réseau Neuronal Récurrent (RNN)

Le modèle de réseau de neurones récurrents (RNN) a affiché une performance impressionnante (figure 65) dans la prévision de la température moyenne journalière pour la région de New York. Avec un RMSE de 0.69, ce modèle démontre une gestion des erreurs particulièrement efficace, suggérant que ses prévisions se rapprochent étroitement des valeurs observées. Par ailleurs, le MAE calculé est de 0.54, ce qui met en évidence une faible différence moyenne absolue entre les prédictions et les valeurs réelles. Ces résultats témoignent de la capacité du RNN à modéliser les dynamiques temporelles des données, renforçant ainsi sa pertinence pour des applications de prévision météorologique.

MSE: 0.48768659141629506
RMSE: 0.6983456102935673
MAE: 0.5499435764589692

Figure 65: Recurrent Neural Networks-KPI.

5.6.6 Comparaison des modèles

Après avoir évalué nos modèles, il nous reste simplement à les comparer afin de sélectionner le meilleur modèle en termes de performance. Nous allons donc comparer le RMSE et le MAE pour déterminer lequel des modèles offre les prévisions les plus précises et fiables, en tenant compte de la gestion des erreurs et des écarts par rapport aux valeurs réelles.

Les figures suivantes (figure 66 et figure 67) illustrent cette comparaison : la première présente les valeurs de RMSE pour chaque modèle, tandis que la seconde met en évidence les valeurs de MAE. Cette analyse visuelle nous permettra d'identifier rapidement les modèles les plus performants et de tirer des conclusions sur leur efficacité respective.

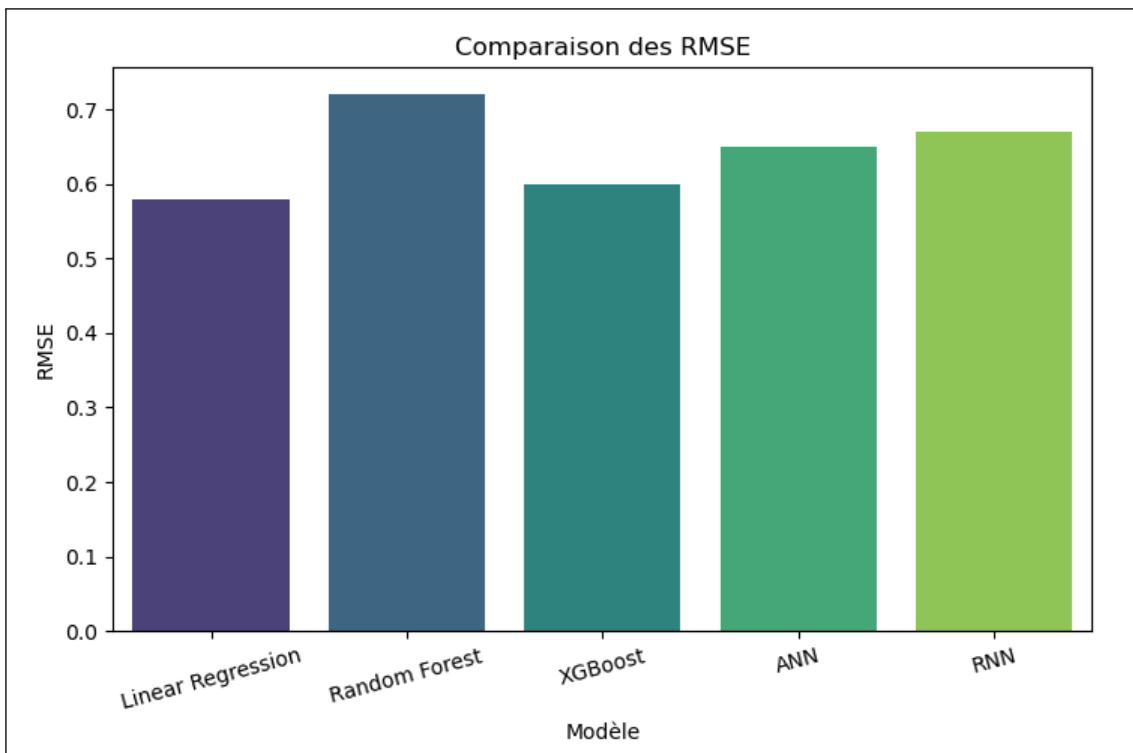


Figure 66: Root Mean Square Error (RMSE).

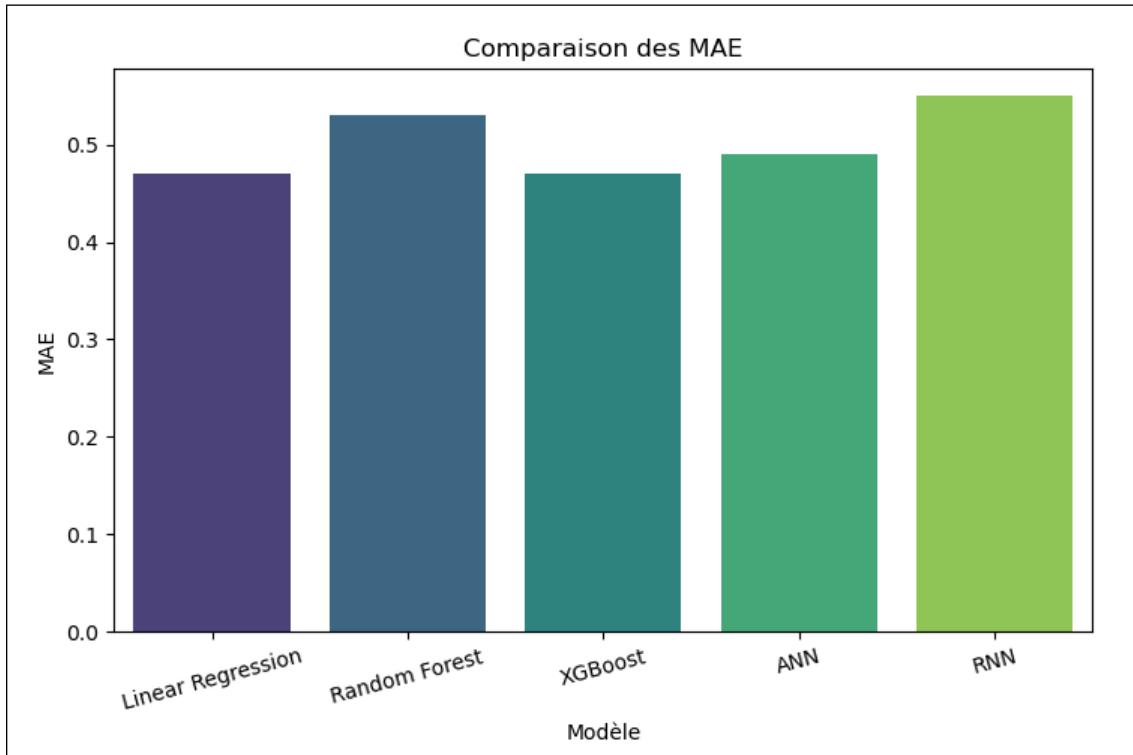


Figure 67: Mean Absolute Error (MAE).

Les cinq modèles montrent des performances impressionnantes et comparables. Cependant, le modèle de régression linéaire présente les RMSE et MAE les plus faibles, ce qui en fait le meilleur modèle parmi ceux évalués.

5.7 Conclusion

Ce chapitre a présenté une analyse approfondie de la température moyenne journalière dans la région de New York à l'aide de plusieurs modèles de machine learning. Après avoir isolé et analysé les données spécifiques à chaque région, nous avons évalué la performance de la régression linéaire, de la forêt aléatoire, de XGBoost, des réseaux de neurones artificiels (ANN) et des réseaux de neurones récurrents (RNN).

Les résultats indiquent que tous les modèles offrent des performances comparables, mais la régression linéaire se distingue par ses RMSE et MAE les plus faibles, la positionnant comme le modèle le plus efficace pour cette étude.

Conclusion générale

Ce rapport a permis d'étudier l'évolution des modèles de prévision météorologique, marquant un tournant significatif dans l'approche traditionnelle, qui privilégiait les systèmes dynamiques et la simulation numérique. Nous avons constaté que, bien que ces méthodes demeurent essentielles pour une compréhension théorique des phénomènes atmosphériques, l'intégration des techniques de machine learning et de deep learning ouvre de nouvelles perspectives pour améliorer la précision des prévisions. Les résultats obtenus révèlent que, parmi les différents modèles analysés, la régression linéaire a affiché une performance remarquable, avec des valeurs de RMSE et de MAE inférieures à celles des autres approches. Cette étude a également mis en lumière l'importance des systèmes dynamiques dans la modélisation des comportements complexes des données météorologiques, offrant ainsi un cadre théorique robuste pour comprendre les interactions entre les différents paramètres. Par ailleurs, cette expérience m'a permis de renforcer mes compétences en traitement de données et en analyse statistique, tout en approfondissant ma compréhension des concepts fondamentaux de la modélisation prédictive. Je suis convaincu que les connaissances acquises lors de ce stage seront un atout précieux pour mes futures activités académiques et professionnelles, et j'aspire à poursuivre mes recherches dans le domaine de l'intelligence artificielle appliquée à la météorologie et aux systèmes dynamiques.

Bibliographie

- [1] *Modélisation Mathématique et Numérique dans les Sciences de l'Ingénieur.* École nationale d'Ingénieurs de Tunis. http://www.edsti.enit.rnu.tn/fr/structures_02.php?id=3&code=LR-99-ES20, n.d. Consulté le 18 août 2024.
- [2] *Le prix présidentiel du meilleur laboratoire de recherche scientifique 2020 remis à LAMSIN.* <https://www.businessnews.com.tn/le-prix-presidentiel-du-meilleur-laboratoire-de-recherche-scientifique-2020-remis-a-lamsin,520,120994,3>. Consulté le 28 août 2024.
- [3] Steven H. Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering.* Consulté le 6 août 2024. Perseus Books Publishing, 1994.
- [4] O. Bonnefoy. *Systèmes Dynamiques et instabilités hydrodynamiques.* <http://www.emse.fr/~bonnefoy/Public/SD-EMSE.pdf>, 2021. Consulté le 6 août 2024.
- [5] E. Nechadi. *Systèmes non linéaires.* n.d. Consulté le 6 août 2024.
- [6] *Qu'est-ce que la turbulence atmosphérique ?* <https://parlonssciences.ca/ressources-pedagogiques/les-stim-expliquees/quest-ce-que-la-turbulence-atmospherique>. Consulté le 7 août 2024.
- [7] *Les modèles météo.* <https://www.meteocontact.fr/pour-aller-plus-loin/les-modeles-meteo>. Consulté le 8 août 2024.
- [8] S. Joubaud. *Systèmes Dynamiques et Chaos (Master Science de la matière 2020–2021).* Consulté le 8 août 2024.
- [9] Celso Grebogi, Edward Ott **and** James A. Yorke. “Chaos, Strange Attractors, and Fractal Basin Boundaries in Nonlinear Dynamics”. *in* *Science*: 238.4827 (1987). Consulté le 11 août 2024, **pages** 632–638. DOI: [10.1126/science.238.4827.632](https://doi.org/10.1126/science.238.4827.632).
- [10] *Master Science de la matière : Systèmes Dynamiques et Chaos, 2020–2021.* Consulté le 17 août 2024.
- [11] *Definition of chaos / Dictionary.com.* www.dictionary.com. Consulté le 18 août 2024.
- [12] Boris Hasselblatt **and** Anatole Katok. *A First Course in Dynamics: With a Panorama of Recent Developments.* Cambridge University Press, 2003. ISBN: 978-0-521-58750-1, Consulté le 18 août 2024.
- [13] Saber N. Elaydi. *Discrete Chaos.* Chapman & Hall/CRC, 1999. ISBN: 978-1-58488-002-8, Consulté le 14 août 2024.
- [14] William F. Basener. *Topology and its Applications.* Wiley, 2006, **page** 42. ISBN: 978-0-471-68755-9, Consulté le 13 août 2024.
- [15] J. Banks **and** al. “On Devaney’s definition of chaos”. *in* *The American Mathematical Monthly*: 99.4 (1992). Consulté le 11 août 2024, **pages** 332–334. DOI: [10.1080/00029890.1992.11995856](https://doi.org/10.1080/00029890.1992.11995856).

- [16] Edward N. Lorenz. “Deterministic nonperiodic flow”. in *Journal of the Atmospheric Sciences*: 20.2 (1963). Consulté le 23 août 2024, pages 130–141. DOI: [10.1175/1520-0469\(1963\)020<0130:DNF>2.0.CO;2](https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2).
- [17] Edward N. Lorenz. *The statistical prediction of solutions of dynamic equations*. https://www.jstage.jst.go.jp/article/jmsj1965/39/3/39_3_160/_pdf. Symposium on Numerical Weather Prediction in Tokyo.1960. Consulté le 17 août 2024.
- [18] Colin Sparrow. *The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors*. Consulté le 25 août 2024. Springer, 1982.
- [19] Bo-Wen Shen. “Aggregated Negative Feedback in a Generalized Lorenz Model”. in *International Journal of Bifurcation and Chaos*: 29.3 (2019). Consulté le 25 août 2024, pages 1950037–1950091. DOI: [10.1142/S0218127419500378](https://doi.org/10.1142/S0218127419500378).
- [20] Bo-Wen Shen. “Nonlinear Feedback in a Five-Dimensional Lorenz Model”. in *Journal of the Atmospheric Sciences*: 71.5 (2014). Consulté le 26 août 2024, pages 1701–1723. DOI: [10.1175/jas-d-13-0223.1](https://doi.org/10.1175/jas-d-13-0223.1).
- [21] *Composition de l'atmosphère terrestre, Climat et Météo*. <https://www.nasa.gov/earth/atmosphere>. Consulté le 29 août 2024.
- [22] *Atmosphère*. <https://www.britannica.com/science/atmosphere>. Consulté le 18 août 2024.
- [23] *Les modèles météo*. <https://www.meteocontact.fr/pour-aller-plus-loin/les-modeles-meteo> n.d. consulté le 19 août 2024.
- [24] *The IFS Documentation*. <https://www.ecmwf.int/en/forecasts/documentation-and-support/technical-references/ifs-documentation>, 2010. Consulté le 10 août 2024.
- [25] Thomas T. Warner. “Numerical Weather and Climate Prediction”. in *Applied Meteorology and Climatology*: 52.1 (2011). Consulté le 6 août 2024, pages 5–27. DOI: [10.1175/2010BAMS3052.1](https://doi.org/10.1175/2010BAMS3052.1).
- [26] S. S. Baboo **and** I. K. Shereef. “An efficient weather forecasting system using artificial neural network”. in *International Journal of Environmental Science and Development*: 1.4 (2010). Consulté le 18 août 2024, pages 321–324. DOI: [10.7763/IJESD.2010.V1.64](https://doi.org/10.7763/IJESD.2010.V1.64).
- [27] D. Endalie, G. Haile **and** W. Taye. “Deep learning model for daily rainfall prediction: Case study of Jimma, Ethiopia”. in *Water Supply*: 22.3 (2022). Consulté le 17 août 2024, pages 3448–3461. DOI: [10.2166/ws.2021.391](https://doi.org/10.2166/ws.2021.391).
- [28] A. G. Tumusiiime, O. S. Eyobu **and** I. Mugume. “Analysis of Machine Learning Algorithms for Prediction of Short-Term Rainfall Amounts Using Uganda’s Lake Victoria Basin Weather Dataset”. in *IEEE Access*: (2024). Consulté le 20 août 2024. DOI: [10.1109/ACCESS.2024.3396695](https://doi.org/10.1109/ACCESS.2024.3396695). URL: <https://doi.org/10.1109/ACCESS.2024.3396695>.
- [29] Siddharth Singh, Mayank Kaushik **and** al. “Weather Forecasting Using Machine Learning Techniques”. in *Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE) 2019*: Consulté le 20 août 2024. 2019. DOI: [10.2139/ssrn.3350281](https://doi.org/10.2139/ssrn.3350281). URL: <https://ssrn.com/abstract=3350281>.

- [30] S. Chaudhuri, A. Long **and** H. et al. Zhang. “Artificial intelligence enabled applications in kidney disease”. *inSeminars in Dialysis*: 34 (2021). Consulté le 20 août 2024, **pages** 5–16.
- [31] Quantmetry. *Une petite histoire du Machine Learning*. <https://www.quantmetry.com/blog/une-petite-histoire-du-machine-learning/>. Quantmetry part of Capgemini Invent. 2015. Consulté le 21 août 2024.
- [32] M. Tim Jones. *Supervised Learning Models*. Artificial Intelligence. 2018. Consulté le 15 août 2024.
- [33] M. Tim Jones. *Unsupervised Learning for Data Classification*. Artificial Intelligence. 2017. Consulté le 18 août 2024.
- [34] Jérémie Robert. *Reinforcement Learning: Définition et application*. DataScientest, Consulté le 15 août 2024. 2020. URL: <https://dataScientest.com/reinforcement-learning-definition-application>.
- [35] DataScientest. *Régression linéaire en Python*. <https://dataScientest.com/regression-lineaire-python> n.d. Consulté le 18 août 2024.
- [36] DataScientest. *Régression linéaire : Tout savoir*. <https://dataScientest.com/regression-lineaire-tout-savoir> n.d. Consulté le 16 août 2024.
- [37] JavaTpoint. *Machine Learning - Random Forest Algorithm*. <https://www.javatpoint.com/machine-learning-random-forest-algorithm>. Consulté le 16 août 2024.
- [38] NVIDIA. *What is XGBoost?* <https://www.nvidia.com/en-us/glossary/xgboost/> n.d. Consulté le 21 août 2024.
- [39] Javatpoint. *Artificial Neural Network*. <https://www.javatpoint.com/artificial-neural-network>. Consulté le 18 août 2024.
- [40] NVIDIA. *Optimizing Recurrent Layers User Guide*. <https://docs.nvidia.com/deeplearning/performance/pdf/Optimizing-Recurrent-Layers-User-Guide.pdf>. Consulté le 22 août 2024.
- [41] Inc. Amazon Web Services. *What is a recurrent neural network?* <https://aws.amazon.com/fr/what-is/recurrent-neural-network/>. Consulté le 19 août 2024.
- [42] Jean-Christophe Chouinard. *What is a confusion matrix in python (scikit-learn example)*. Consulté le 21 août 2024. 2023.
- [43] Anja Thieme, Danielle Belgrave **and** Gavin Doherty. “Machine Learning in Mental Health: A Systematic Review of the HCI Literature to Support the Development of Effective and Implementable ML Systems”. *inACM Transactions on Computer-Human Interaction*: 27.5 (2020). Consulté le 22 août 2024.
- [44] JavaTpoint. *MSE and bias-variance decomposition*. <https://www.javatpoint.com/mse-and-bias-variance-decomposition>. Consulté le 24 août 2024.
- [45] JavaTpoint. *Performance metrics in machine learning*. <https://www.javatpoint.com/performance-metrics-in-machine-learning>. Consulté le 27 août 2024.
- [46] Arize AI. *Mean absolute percentage error (MAPE) : What you need to know*. <https://arize.com/blog-course/mean-absolute-percentage-error-mape-what-you-need-to-know/>. Consulté le 12 août 2024.
- [47] DataScientest. *Python*. https://dataScientest.com/?s=python&id=15238&post_type=post Consulté le 12 août 2024.

-
- [48] Study Smarter. *Machine learning using R programming.* R Programming Language. Consulté le 13 août 2024.
 - [49] Working Nation. *Logo de R.* https://workingnation.com/wp-content/uploads/2018/05/R_logo.svg_.png Consulté le 17 août 2024.
 - [50] Microsoft. *Power BI overview.* <https://learn.microsoft.com/fr-fr/power-bi/fundamentals/power-bi-overview>. Consulté le 23 août 2024.
 - [51] Project Jupyter. *Jupyter.* <https://jupyter.org/>. Consulté le 2 septembre 2024.

