REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE MOHAMED BOUDIAF - M'SILA

Faculté des Mathématiques et de l'Informatique
Département d'Informatique
N° :……..

DOMAINE : Mathématiques et Informatique
FILIERE : Informatique
OPTION : Intelligence Artificielle

Mémoire présenté pour l'obtention
Du diplôme de Master en informatique

Par :

- ALLAL Samira

- ARIOUA Haniya

Intitulé

Modèles de machine Learning pour la prédiction de la consommation en énergie électrique

Cas d'étude : La société Sonelgaz

Soutenu devant le jury composé de :

GUERNA ABDERRAHIM	Université de M'sila	Président
MEHENNI Tahar	Université de M'sila	Rapporteur
BAHACHE MOHAMMED	Université de M'sila	Examinateur

Année universitaire : 2021 / 2022

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE MOHAMED BOUDIAF - M'SILA

Faculté des Mathématiques et de l'Informatique
Département d'Informatique
N° :……..

DOMAINE : Mathématiques et Informatique
FILIERE : Informatique
OPTION : Intelligence Artificielle

Mémoire présenté pour l'obtention
Du diplôme de Master en informatique

Par :

ALLAL Samira

ARIOUA Haniya

Intitulé

Modèles de machine Learning pour la prédiction de la consommation en énergie électrique

Cas d'étude : La société Sonelgaz

Soutenu devant le jury composé de :

GUERNA ABDERRAHIM	Université de M'sila	Président
MEHENNI Tahar	Université de M'sila	Rapporteur
BAHACHE MOHAMMED	Université de M'sila	Examinateur

Année universitaire : 2021 / 2022

REMERCIEMENTS

Loué soit ALLAH, qui nous a accordé le succès et le remboursement et nous a accordé la stabilité et nous a aidé à terminer ce travail après avoir travaillé pour mettre les points sur les lettres et révéler ce qui se cache derrière le rideau de la science et de la connaissance, voici donc les fruits de notre connaissances qui ont mûri et viennent se récolter.

Nous adressons nos sincères remerciements au docteur superviseur, Mr Mehenni Taher, qui nous a accompagné tout au long de cette recherche et nous a fourni de précieux renseignements et conseils. Nous remercions également l'honorable comité d'avoir accepté d'évaluer ce travail et de nous donner ses précieux conseils.

Nous adressons également nos remerciements et notre gratitude à nos généreux parents pour leur soutien et leurs encouragements.

Enfin, nous ne manquons pas d'adresser nos salutations les plus sincères à tous ceux qui nous ont aidés de près ou de loin à accomplir cette modeste recherche.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION GÉNÉRALE

L'électricité est le pilier de l'économie nationale et ses services durables sont à la base de la vie normale de la population. La demande d'énergie électrique continue de croître rapidement, en raison de plusieurs facteurs dont les plus importants sont l'expansion démographique et développement technologique.

On sait que l'électricité ne se stocke pas, surtout en grande quantité. Ainsi, l'approvisionnement en électricité doit être en ligne avec la demande, il est donc nécessaire de prévoir avec précision la demande future d'électricité, car une meilleure prévision de la consommation des clients permet de réduire considérablement les erreurs d'ordres de production, ce qui réduit les pertes et les problèmes tels que groupes électrogènes inutiles, consommation de carburant élevée et coût d'exploitation accru Et éviter de débrancher l'électricité aux heures de pointe... etc.

La prévision des consommations d'électricité s'effectue soit en utilisant des méthodes classiques en s'appuyant notamment sur des estimations d'experts par des méthodes statistiques, ou des modèles basés sur l'intelligence artificielle tels que les réseaux de neurones récurrents (RNN), les machines à vecteurs de support (SVM), ...etc. Les méthodes traditionnelles sont faciles à utiliser, cependant elles manquent d'efficacité et de précision. Par conséquent, de nombreux travaux récents se sont orientés vers des méthodes d'intelligence artificielle pour une prédiction précise et efficace.

En raison de l'augmentation de la consommation d'électricité à des tarifs élevés et de la nécessité de déterminer le montant de la demande, nous allons essayer dans ce travail d'utiliser des modèles d'apprentissage automatique pour prédire la consommation d'énergie électrique à l'aide d'un ensemble de données historiques sur la consommation d'électricité relatives au pôle Adrar et à l'échelle nationale pour une période de deux ans et des données synthétiques avec des caractéristiques distinctives (features) qui affectent la consommation d'électricité. À travers ces données nous essayons d'atteindre des résultats précis et tangibles quant à la prédiction en énergie électrique.

Dans cette étude, nous avons discuté de quatre chapitres, de sorte que nous avons complété deux chapitres théoriques et deux chapitres pratiques. Dans le premier chapitre, nous avons défini le problème de la consommation d'électricité et les facteurs les plus importants qui l'affectent et les méthodes utilisées pour résoudre ce problème. Puis, dans

le deuxième chapitre, nous avons décrit le fonctionnement des modèles d'apprentissage automatique. Quant au troisième chapitre, nous avons mentionné les étapes les plus importantes que traverse le processus de prévision de la consommation d'électricité pour parvenir à des prévisions précises. En utilisant un ensemble de modèles d'apprentissage automatique, et enfin, dans le quatrième chapitre, nous avons analysé les résultats obtenus grâce à l'utilisation de modèles d'apprentissage automatique.

CHAPITRE 1

CONSOMMATION D'ENERGIE ELECTRIQUE

1 Introduction

L'énergie en général, et l'énergie électrique en particulier, est l'un des piliers les plus importants du développement de tout pays et le pilier de l'économie internationale, et aussi un critère qui explique le progrès ou le retard d'un pays, ce qui fait que sa consommation ou sa demande augmente constamment, que ce soit par des particuliers ou des institutions économiques. Nous allons présenter dans ce chapitre quelques notions générales sur l'énergie électrique et ses modes de production, ensuite nous aborderons les secteurs qui en consomment, de manière à éclairer l'évolution de la consommation d'électricité, en mettant en évidence les raisons et facteurs à l'origine de cette évolution.

2 Concepts généraux de l'énergie électrique

L'énergie est l'un des moteurs les plus importants d'une économie. Parmi ses formes se trouve l'électricité. C'est un bien indispensable en raison de sa grande importance dans le développement économique et industriel. C'est pourquoi elle fait l'objet d'une attention internationale.

3 Définition de l'énergie électrique

L'énergie électrique est définie comme une forme d'énergie produite à partir de sources renouvelables et non renouvelables, car elle résulte de particules chargées (électrons et ions) et est capable de dégager de la chaleur ou de la lumière. [1]

4 Méthodes de production d'énergie électrique

L'électricité est produite à partir de diverses sources qui vont des ressources renouvelables telles que le vent, le rayonnement solaire, les marées, etc., aux ressources non renouvelables, en passant par les matières fossiles telles que le gaz naturel, le pétrole et l'énergie nucléaire. Les méthodes de production d'électricité sont les suivantes : [1]

4.1 Station hydroélectrique

Une centrale hydroélectrique est un type particulier de centrale électrique qui utilise l'énergie de l'eau qui tombe ou qui coule pour produire de l'électricité. Pour ce faire, ils

dirigent l'eau sur une série de turbines qui convertissent l'énergie potentielle et cinétique de l'eau en mouvement de rotation de la turbine. La turbine est ensuite attachée à un générateur et le mouvement est utilisé pour générer de l'électricité.

4.2 Centrales à vapeur

Les centrales à vapeur sont connues (turbines à vapeur) car elles dépendent de la pression de la vapeur pour déplacer les turbines. La centrale à vapeur est une centrale électrique dans laquelle le générateur électrique fonctionne à la vapeur. La chaudière génère de la vapeur à haute pression et haute température. Les turbines à vapeur convertissent l'énergie thermique de la vapeur en énergie mécanique. Le générateur convertit ensuite l'énergie mécanique en énergie électrique.

4.3 Moteurs à combustion interne

Les installations de combustion interne sont des machines qui utilisent des combustibles liquides (mazout) car ils brûlent à l'intérieur des chambres de combustion après les avoir mélangés à de l'air dans certaines proportions. Rotation, comme dans le cas des turbines à gaz. C'est-à-dire que l'électricité est produite à partir d'installations à combustion interne par deux types des turbines :

4.3.1 Centrale électrique diesel

Les machines diesel produiront de l'électricité. Elle se caractérise par la vitesse de fonctionnement et la vitesse d'arrêt, mais il nécessite une quantité de carburant relativement élevée, et donc le coût de l'énergie produite à partir de celui-ci dépend du prix du carburant. En revanche, il n'y a pas d'unités de grande capacité. (3 mégawatts seulement). Ces générateurs sont faciles à installer et sont fréquemment utilisés dans les situations d'urgence ou pendant le pic de grossesse. Dans ce cas, un grand nombre de ces générateurs fonctionnent généralement en parallèle pour répondre aux besoins des centres de consommation.

4.3.2 Turbine à gaz

Les centrales à turbine à gaz sont relativement. Elles ont des capacités et des tailles différentes de 1 mégawatt à 250 mégawatts, et sont généralement utilisées lors des pointes de charge dans les pays où existent des centrales à vapeur ou à eau, sachant que la période de décollage et d'arrêt varie entre deux et dix minutes. On n'a pas besoin de beaucoup

d'eau pour le refroidissement. Une turbine à gaz se caractérise également par la possibilité d'utiliser de nombreux types de combustibles (pétrole brut pur - gaz naturel - gaz lourd, etc.…) et elle se caractérise également par la rapidité de fonctionnement et la vitesse d'arrêt.

Quant à ses inconvénients, elles ont le double du rendement, qui varie entre 15 et 25 %, et sa durée de vie est relativement courte, et elles consomment une plus grande quantité de combustible par rapport aux centrales thermiques à vapeur.

4.4 Station marémotrice

Énergie marémotrice, également appelée énergie marémotrice, toute forme d'énergie renouvelable dans laquelle l'action des marées dans les océans est convertie en énergie électrique.

Il existe plusieurs façons d'exploiter l'énergie marémotrice. Les systèmes de barrage marémotrice tirent parti des différences entre les marées hautes et les marées basses en utilisant un « barrage » ou un type de barrage, pour bloquer le recul de l'eau pendant les périodes de reflux. À marée basse, l'eau derrière le barrage est libérée et l'eau passe à travers une turbine qui génère de l'électricité.

4.5 Centrale nucléaire

Les centrales nucléaires produisent de l'électricité de la même manière que les centrales conventionnelles, sauf que la différence réside dans le type de combustible utilisé. Toutes les centrales produisent de l'électricité en produisant de la chaleur qui transforme l'eau en vapeur, puis la vapeur fait tourner des moteurs ou des turbines connectés à des générateurs pour produire de l'électricité. L'uranium est utilisé comme combustible pour les réacteurs nucléaires, et l'uranium est un élément radioactif naturel abondant dans la plupart des roches. Le processus de fission qui conduit à la division des atomes d'uranium en petites parties en modifiant leur structure moléculaire ; Une quantité d'énergie est générée lors de cette division, ce qui conduit à la production de suffisamment de chaleur pour produire de la vapeur, qui est utilisée par la turbine pour produire de l'électricité. Ou une tonne de charbon.

4.6 Centrale éolienne

Une éolienne transforme l'énergie éolienne en électricité en utilisant la force aérodynamique des pales du rotor, qui fonctionnent comme une aile d'avion ou une pale de rotor d'hélicoptère. Lorsque le vent traverse la pale, la pression de l'air sur un côté de la pale diminue. La différence de pression d'air entre les deux côtés de la pale crée à la fois une portance et une traînée. La force de la portance est plus forte que la traînée et cela fait tourner le rotor. Le rotor se connecte au générateur, soit directement (s'il s'agit d'une turbine à entraînement direct), soit par l'intermédiaire d'un arbre et d'une série d'engrenages (une boîte de vitesses) qui accélèrent la rotation et permettent un générateur physiquement plus petit. Cette translation de la force aérodynamique à la rotation d'un générateur crée de l'électricité.

4.7 Centrales solaires

Les technologies solaires convertissent la lumière du soleil en énergie électrique soit par des panneaux photovoltaïques (PV), soit par des miroirs qui concentrent le rayonnement solaire. Cette énergie peut être utilisée pour produire de l'électricité ou être stockée dans des batteries ou un stockage thermique.

L'image suivante montre la production d'électricité dans le monde par source d'énergie (voir Fig. 1.1) :

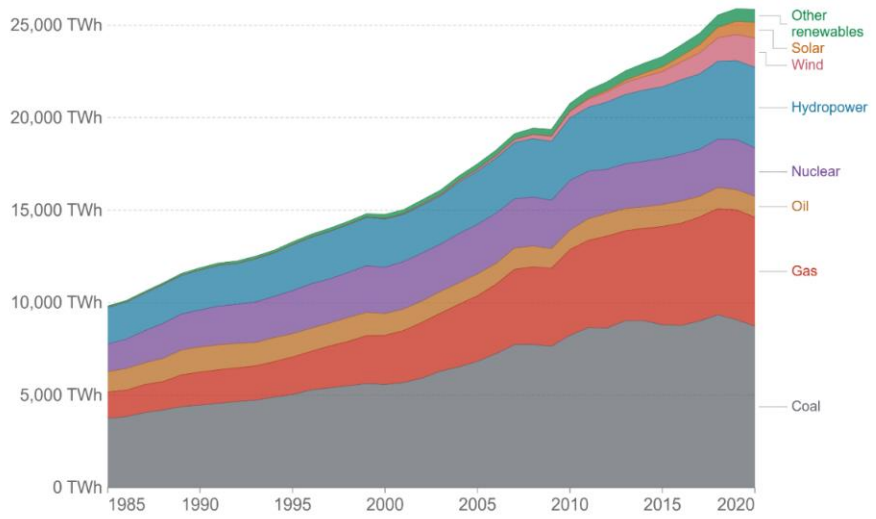

Figure 1.1 Production d'électricité dans le monde par source d'énergie 1985-2020 [2]

Il existe de nombreuses sources de production d'énergie électrique Et la recherche de moyens avancés pour le produire, Cela est dû au fait que c'est l'énergie qui est capable de fournir et de concilier les différents besoins des différents types de secteurs ce qui la place au premier rang des plus énergivores.

5 Définition de la consommation d'énergie électrique

La consommation d'énergie électrique est une forme de consommation d'énergie qui utilise de l'énergie électrique. La consommation d'énergie électrique est la quantité d'énergie ou de capacité utilisée. [3]

6 Secteurs de consommation l'énergie électrique

Plusieurs secteurs entrent dans le domaine de la consommation d'énergie électrique, et cela est dû à ses utilisations diverses et différentes. On retrouve les secteurs les plus en vue dans lesquels l'énergie électrique apparaît sont : [4]

6.1 Secteur agricole

L'énergie électrique est utilisée comme carburant pour les moyens et les machines telles que les tracteurs, les pompes à eau, etc. en tant qu'utilisations directes, tandis qu'elle est également utilisée indirectement dans la fabrication d'aliments pour le bétail et d'engrais, etc.

6.2 Secteur résidentiel

L'électricité est disponible, en particulier pour les citoyens en général, pour répondre à leurs besoins particuliers d'éclairage et de chauffage, et pour les tâches ménagères en général (comme la cuisine).

6.3 Secteur de la fonction publique

Sa part d'énergie électrique est limitée au périmètre des services publics tels que les bâtiments commerciaux, les hôpitaux, les établissements d'enseignement et le reste des secteurs (secteur de la transmission, secteur des travaux publics).

6.4 Secteur industriel et économique

Où l'électricité est utilisée dans tous les projets industriels et les produits des industries du plastique, du caoutchouc, du textile et autres.

7 L'évolution de la consommation d'énergie électrique

La consommation d'électricité reflète la croissance économique et le développement de n'importe quel pays On constate à travers la courbe qui montre l'évolution de la consommation d'énergie électrique sur une période de 1980 à 2018 que l'échelle de la consommation d'électricité a connu une évolution remarquable (voir Fig. 1.2) :

- sur la période de 1980 à 2010, elle a connu une augmentation de 18 704 TWh, ce qui correspond à un taux de croissance de 1,5 % au cours des trente dernières années.
- Alors qu'à partir de 2011, la consommation d'électricité a atteint 23 398 TWh, soit au cours des huit dernières années, le taux de croissance a doublé de 2,2 %.

Ce qui indique et témoigne du fait que l'électricité la consommation augmente à un rythme accéléré, afin de répondre aux besoins croissants des différents segments d'abonnés en fourniture d'électricité.

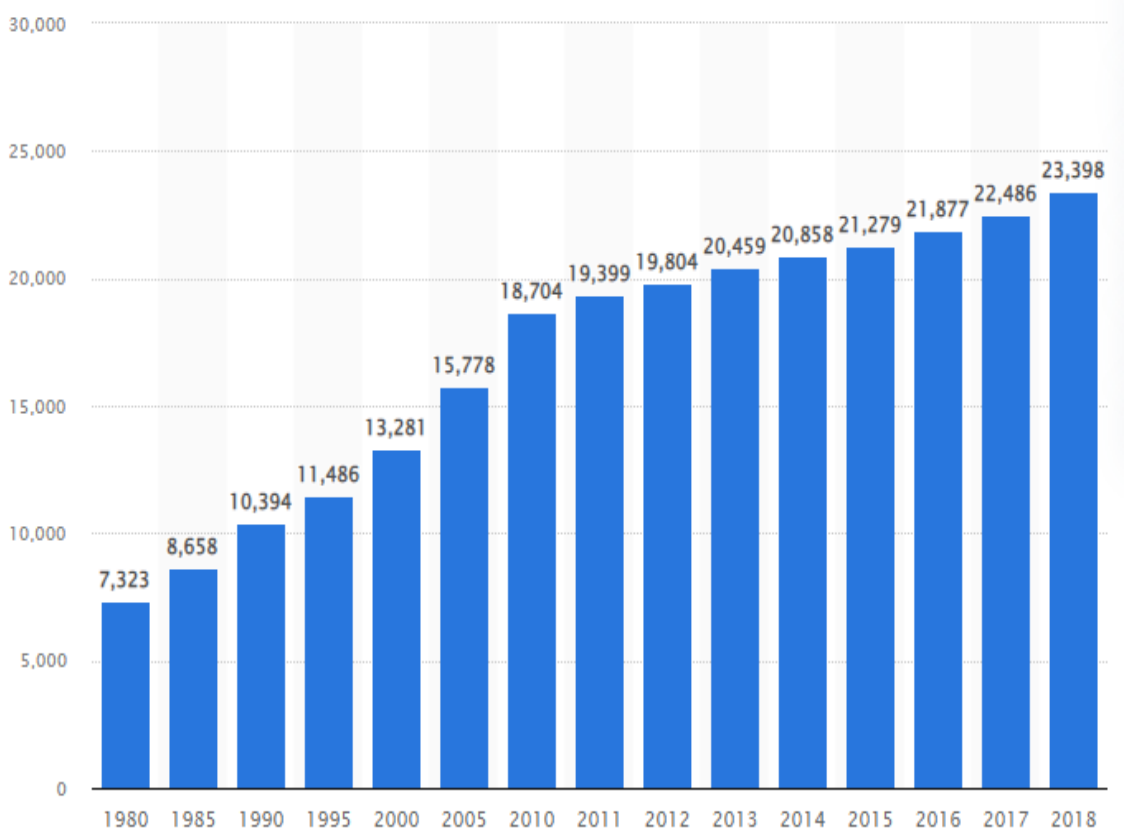

Figure 1.2 Consommation nette d'électricité au cours de certaines années de 1980 à 2018 [5]

8 Facteurs de la consommation d'énergie électrique

Nous notons que la fréquence de la quantité de consommation d'électricité augmente d'année en année et les variables et facteurs se diffèrent pour des raisons économiques, technologiques, culturelles, sociales et démographiques. Les points suivants résument les facteurs les plus importants motivant la consommation d'énergie :

- Changements climatiques : l'augmentation de la consommation d'électricité coïncide pendant les mois chauds (été) et diminue pendant les mois froids (hiver).
- Croissance démographique : au fur et à mesure que de nouveaux quartiers sont construits et construits, ce qui génère une augmentation du nombre d'abonnés aux réseaux, et donc augmente la demande d'électricité.
- Développement technologique : en est l'inclusion d'une technologie de haut niveau efficace dans la vie sociale et dans divers domaines Secteurs tels que l'industrie, l'économie, ... etc.
- Une élévation et une amélioration du niveau de vie : qui se traduit par une augmentation de l'utilisation des produits de luxe tels que les appareils électriques et autres
- Croissance économique : elle reflète les utilisations croissantes de l'énergie électrique par les agents économiques et le reste du monde Secteurs (secteur des transports, travaux publics, ... etc.) [4].

9 Outils et méthodes d'estimation de la demande d'électricité

Les grandes entreprises et les institutions économiques adoptent un ensemble d'outils et de méthodes afin d'estimer la demande de consommation d'énergie électrique, car elles dépendent, dans le processus d'estimation des attentes de consommation d'électricité, des éléments suivants :

La méthode basée sur la référence à la date de consommation d'électricité au même jour à partir de la même date des années précédentes pour faire une prévision à l'heure actuelle et calculer le taux d'augmentation, qui est calculé de cette manière [4].

Consommation l'année prochaine = Consommation de l'année dernière + x. Consommation de l'année dernière.

Où x : le Taux d'augmentation.

x. Consommation de l'année dernière : la valeur de la charge de l'année dernière.

L'augmentation du taux de croissance qui a été initialement calculée pour l'année suivante, et s'il s'avère que la consommation augmente ou diminue, elle est modifiée en fonction de la situation.

Exemple :

Calculer le taux de croissance qui sera augmenté pour les attentes de consommation pour l'année 2018 sur la base des années précédentes à partir de 2010.

C= consommation.

X.C année = la valeur de la charge de l'année

C2011=C2010+X.C2010

C11=(X+1) C10

C12=(X+1) C11=$(X+1)^2$ C10

……

C18=$(X+1)^8$ C10

$$X=\sqrt[8]{\frac{C18}{C10}}-1 \qquad (1)$$

Si le taux de croissance X=2, Il est pris en compte et dans les années à venir, le taux de croissance est augmenté de X, et jusqu'à ce moment-là, il est modifié, que ce soit en l'augmentant ou en le diminuant.

10 L'importance de rationaliser la consommation d'énergie électrique

Il reste également de la responsabilité du citoyen de suivre les consignes d'utilisation de l'électricité car il en est le premier consommateur et bénéficiaire, alors rationaliser l'usage de l'électricité aujourd'hui est le moyen de la préserver demain.

- La rationalisation de la consommation contribue à réduire l'utilisation de l'énergie électrique tout en maintenant le volume de production.
- Réduire les charges excessives sur les centrales électriques, ce qui améliore la continuité du service électrique avec l'efficacité requise.
- Contribuer à la préservation de l'environnement en réduisant la consommation de carburant.
- Il travaille à réduire la perte d'énergie électrique et à la préserver.

11 Conclusion

Dans ce chapitre, nous avons mené une étude générale sur l'état de la consommation d'énergie électrique et son orientation, à partir de laquelle nous avons conclu que le développement de l'énergie électrique s'accélère à un rythme accéléré et que les méthodes utilisées pour faire face à ce problème sont imprécises et méthodes non scientifiques, ce qui nécessite de rechercher d'autres méthodes modernes très efficaces et précises telles que les modèles d'apprentissage automatique dont nous présenterons certains de ses modèles et dont nous essaierons d'expliquer le fonctionnement dans le chapitre suivant.

CHAPITRE 2

LES MODÈLES DE PRÉVISION

1 Introduction

L'étude des modèles d'apprentissage automatique (machine Learning) a un rôle majeur dans le processus de planification et de prévision. Elle a beaucoup d'importance dans le processus d'évaluation de la croissance et du développement de certaines variables relatives au domaine étudié. Notre objectif principal est d'atteindre une prédiction qui fournit des informations futures en fonction des études précédentes. Dans ce chapitre, nous aborderons l'étude des modèles de prédiction les plus utilisées (voir Fig. 2.1) et identifierons les algorithmes les plus importants, dans l'objectif d'en choisir les plus appropriées pour notre étude.

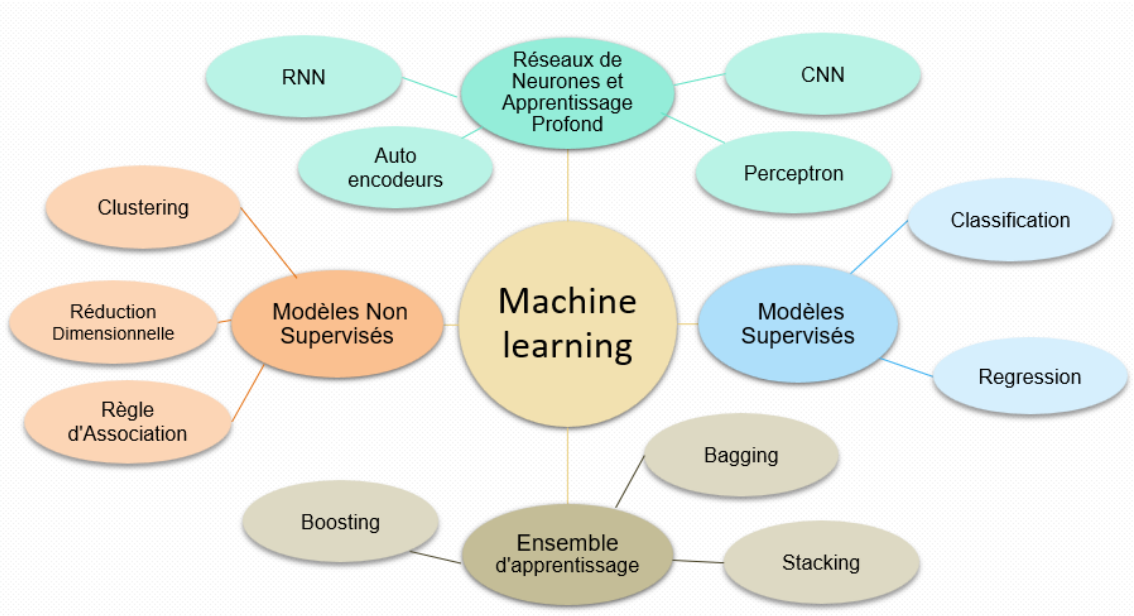

Figure 2.1 Les modèles de prévisions

2 Les modèles supervisés

Ces modèles sont utilisés pour développer un modèle prédictif basé sur des données d'entrée et de sortie. Parmi ces modèles, on peut citer les suivants :

2.1 Modèle de Régression

Son idée est qu'il existe des données liées entre elles et que de nouvelles valeurs de données sont reconnues où une distinction est faite entre les problèmes de régression et d'autres problèmes en ce que les problèmes de prédiction de la variable quantitative sont considérés comme des problèmes de régression. Parmi les applications de la régression : prix des maisons, cours de bourse, la météo, Le montant que le client achètera.

Les algorithmes utilisés pour la régression sont :

2.1.1 Régression linéaire

Ce modèle n'est qu'une fonction linéaire de l'entité en entrée. Plus généralement, un modèle linéaire (voir Fig. 2.2) effectue une prédiction en calculant simplement une somme pondérée des entités en entrée, plus une constante appelée le terme de biais (également appelé terme d'interception) et son degré = 1, comme indiqué dans Équation [6]:

$$\hat{y} = \vartheta_0 + \vartheta_1 x_1 + \vartheta_2 x_2 + \cdots \vartheta_n x_n \qquad (2)$$

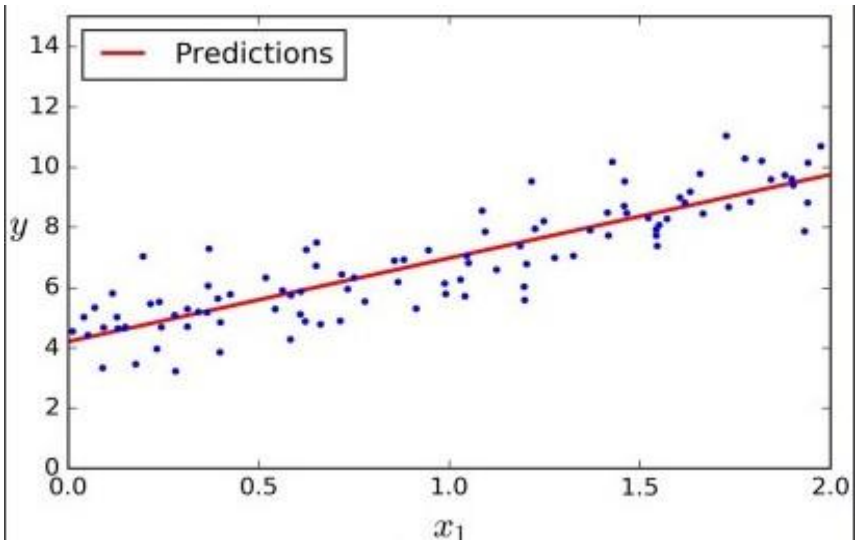

Figure 2.2 Prédictions du modèle de régression linéaire.

2.1.2 Régression Polynomiale

Les données sont en réalité plus complexes qu'une simple ligne droite. Par conséquent, on peut utiliser un modèle linéaire pour ajuster des données non linéaires (voir Fig. 2.3). Pour ce faire, on peut ajouter les puissances de chaque fonctionnalité en tant que nouvelles fonctionnalités, puis entraîner un modèle linéaire sur cet ensemble étendu de

fonctionnalités. Cette technique s'appelle la régression polynomiale et son degré supérieur à 1, comme indiqué dans Équation :

$$\widehat{y} = \vartheta_0 + \vartheta_1 x + \vartheta_2 x^2 + \vartheta_3 x^3 + \cdots \vartheta_n x^n \quad (3)\,[6].$$

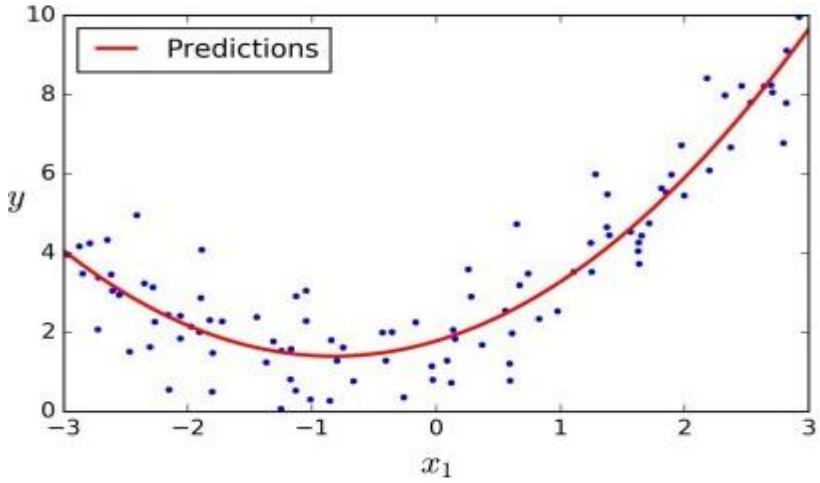

Figure 2.3 Prédictions du modèle de Régression Polynomiale.

2.1.3 Régression Ridge

C'est une version régularisée de la régression linéaire. Elle résout certains problèmes des moindres carrés ordinaires en imposant une pénalité sur la taille des coefficients. Les coefficients de crête minimisent une somme résiduelle des carrés pénalisée [7]:

$$\min_{\omega} \|\mathbb{X}\, w - y\|_2^2 + \alpha \|w\|_2^2 \quad (4)$$

2.1.4 Régression vectorielle

On utilise la méthode de Support Vector Classification pour résoudre les problèmes de régression. Le modèle produit par Support Vector Regression (SVR) ne dépend que d'un sous-ensemble des données d'apprentissage, car la fonction de coût pour la construction du modèle ignore toutes les données d'apprentissage proches de la prédiction du modèle, où le noyau (Kernel) spécifie le type à utiliser dans l'algorithme, il doit s'agir du kernel "linéaire", "poly", "rbf", "sigmoïde" ou "précalculé". Si aucun kernel n'est donné, 'rbf' sera utilisé par défaut. Le paramètre de pénalité C du terme d'erreur Epsilon spécifie la valeur à laquelle aucune pénalité n'est associée dans la fonction de perte d'entraînement avec des points prédits à une distance epsilon de la valeur réelle.

2.1.5 Régression de forêt aléatoire

L'approche de forêt aléatoire est un algorithme de classification et de régression non linéaire supervisé. La classification est un processus de classification d'un groupe d'ensembles de données dans des catégories ou des classes. L'approche de forêt aléatoire peut utiliser des techniques de classification ou de régression en fonction de l'utilisateur et de la cible ou des catégories nécessaires. Une forêt aléatoire est une collection d'arbres de décision qui spécifie les catégories avec une probabilité beaucoup plus élevée. L'utilisation d'une forêt aléatoire pour effectuer une régression où le nombre d'arbres est déterminé par la fonction *n*_estimators et la profondeur requise est max_depth.

2.2 Modèle de Classification

Son idée est qu'il existe des données qui sont liées les unes aux autres et sont divisées en groupes similaires où une distinction est faite entre les problèmes de classification et d'autres problèmes en ce que les problèmes de prédiction de la variable qualitative sont considérés comme des problèmes de classification. Parmi les applications de la classification : des images, Déterminer si le client achètera ou non.

Les algorithmes utilisés dans ce modèle sont :

2.2.1 Arbres de décision

L'objectif de DT est de créer un modèle qui prédit la valeur d'une variable cible en apprenant des règles de décision simples déduites des caractéristiques des données. Plus l'arbre est profond, plus les règles de décision sont complexes et plus le modèle est adapté. Où est utilisé critère Fonction permettant de mesurer la qualité d'un fractionnement. Les critères pris en charge sont "Gini" pour l'impureté Gini et " entropy" pour le gain d'information.

2.2.2 Naive Bayes

Hypothèse d'indépendance conditionnelle entre chaque paire d'entités compte tenu de la valeur de la variable de classe. Il existe quatre types : GaussianNB, MultinomialNB, BernoulliNB.

2.2.3 K Voisins les Plus Proches

Un point de requête se voit attribuer la classe de données qui a le la plupart des représentants dans les voisins les plus proches du point. Où la fonction de weight utilisée

dans la prédiction. Valeurs possibles : uniforme, distance et nombre de voisins pour chaque échantillon par la fonction n_neighbors.

2.2.4 Régression Logistique

La régression logistique est couramment utilisée pour estimer la probabilité qu'une instance appartienne à une classe particulière. Et c'est un classificateur binaire. fonctionne Tout comme un modèle de régression linéaire, un modèle de régression logistique calcule une somme pondérée des caractéristiques d'entrée (plus un terme de biais), mais au lieu de produire le résultat directement comme le fait le modèle de régression linéaire, il produit la logistique de ce résultat [6].

$$\hat{P} = \hbar_\theta(\chi) = \sigma(\theta^T . X) \quad (5)$$

2.2.5 Classification des vecteurs de support

Implémentation d'un modèle de support d'algorithme de machine vectorielle pour faire une classification qui est utilisée par un module svm.SVC.

2.2.6 Classification aléatoire des forêts

Utiliser une forêt aléatoire pour effectuer une classification par module RandomForestClassifier.

3 Les Modèles Non Supervisés

Ils sont utilisés pour construire un modèle basé sur la collecte et l'interprétation de données basées sur des données d'entrée uniquement. Les modèles sont les suivants :

3.1 Modèle de Clustering

Son idée est de transformer les données à partir de données publiques mixtes en sections distinctes les unes des autres selon des caractéristiques spécifiques et diviser les données en catégories distinctes en fonction de leurs caractéristiques similaires. Parmi les applications du clustering : identifier des groupes parmi les patients présentant les mêmes symptômes permet d'identifier des sous-types d'une maladie qui pourront être traités différemment. La segmentation de marché consiste à identifier des groupes d'usagers ou de clients ayant un comportement similaire. [7]

Les algorithmes qui dépendent de modèle de clustering sont :

3.1.1 K-means

Utilisation à usage général, même taille de cluster, géométrie plate, pas trop de clusters. Dépend des distances entre les points extensibles (très grands n_échantillons, n_clusters moyens avec code MiniBatch).

3.1.2 Mean Shift

Utilise de nombreux clusters, taille de cluster inégale, géométrie non-plate Dépend des distances entre les points Non-évolutif avec n_samples.

3.1.3 DBSCAN

Utilise une géométrie non-plate, des tailles de cluster inégales Dépend des distances entre les points les plus proches extensibles (très grands n_échantillons, moyens n_clusters).

3.1.4 Agglomération

Appelé aussi la classification hiérarchique, cet algorithme utilise ee nombreux clusters, éventuellement des contraintes de connectivité, non-Distances Euclidienne, dépend de n'importe quelle distance extensible par paires (grands n_échantillons et n_clusters).

3.2 Modèle de Réduction Dimensionnelle

C'est une technique utilisée pour annuler un certain nombre de caractéristiques afin d'accélérer le processus de traitement des données et son but n'est pas de réduire l'espace utilisé dans la mémoire du disque dur mais réduisez l'espace de RAM utilisée dans le processeur, Accélérez également le processus GD (Gradient Descent) Les caractéristiques qui sont supprimées doivent être liées à d'autres caractéristiques Sinon, nous ne pourrons pas l'annuler [6].

Les algorithmes basés sur le modèle de réduction de dimensionnalité sont :

3.2.1 Analyse des composants principaux

PCA est utilisée pour décomposer un ensemble de données multivariées en un ensemble de composantes orthogonales successives qui expliquent un maximum montant de l'écart. L'objet PCA est très utile, mais présente certaines limitations pour les grands ensembles de données. La plus grande limitation est que PCA ne prend en charge que le

traitement par lots, ce qui signifie que toutes les données à traiter doivent tenir dans la mémoire principale.

3.2.2 Analyse discriminante linéaire

LDA est un modèle probabiliste génératif pour les collections d'ensembles de données discrets tels que les corpus de texte. C'est également un modèle de sujet utilisé pour découvrir des sujets abstraits à partir d'une collection de documents.

3.2.3 Analyse sémantique latente

L'analyse sémantique latente LSA est une technique permettant de créer une représentation vectorielle d'un document. Avoir une représentation vectorielle d'un document donne un moyen de comparer les documents pour leur similitude en calculant la distance entre les vecteurs. Cela signifie à son tour peut faire des choses pratiques comme classer les documents

3.2.4 Décomposition en valeurs singulières

SVD est une Technique de réduction de dimensionnalité linéaire, très similaire à l'PCA mais elle ne centre pas les données avant de calculer la décomposition en valeurs singulières. Cela signifie qu'il peut fonctionner efficacement avec des matrices peu nombreuses.

3.2.5 Intégration de voisins stochastiques distribués en t

T-SNE Réduit la dimensionnalité tout en essayant de garder des instances similaires proches et des instances différentes séparées. Il est principalement utilisé pour la visualisation, dans notamment pour visualiser des clusters d'instances dans un espace de grande dimension (par exemple, pour visualiser les images MNIST en 2D.

3.3 Modèle de Règles d'Association

Il s'agit d'une méthode d'apprentissage basée sur des règles pour découvrir des relations intéressantes entre les variables dans de grandes bases de données. Son objectif est d'identifier les règles fortes découvertes dans les bases de données à l'aide de certaines métriques et d'en générer de nouvelles. Parmi les applications de règle d'association : Prendre des décisions concernant les activités de marketing telles que la tarification ou le placement de produit, exploration Web, systèmes de détection d'intrusion et correction continue . [6]

Des algorithmes de modèle de règle d'association sont :

3.3.1 Apriori

L'algorithme Apriori est utilisé pour extraire des ensembles d'éléments fréquents et concevoir des règles d'association à partir d'une base de données transactionnelle. Les paramètres « support » et « confiance » sont utilisés. Le support fait référence à la fréquence d'occurrence des éléments ; la confiance est une probabilité conditionnelle.

3.3.2 Eclat

L'algorithme Eclat est un algorithme d'exploration de données utilisé pour trouver des éléments fréquents. Certains algorithmes d'extraction de règles d'association utilisent un format de données horizontal et certains d'entre eux utilisent un format de données vertical pour la génération d'ensembles d'éléments fréquents. Eclat ne peut pas utiliser de base de données horizontale. S'il existe une base de données horizontale, vous devez la convertir en base de données verticale.

3.3.3 FP-Growth

L'algorithme de croissance FP est une amélioration de l'algorithme a priori. Algorithme de croissance FP utilisé pour trouver des ensembles d'éléments fréquents dans une base de données de transactions sans génération de candidats. La croissance FP représente des éléments fréquents dans des arbres de modèles fréquents ou FP-tree.

4 Réseaux de Neurones et Apprentissage Profond

Réseaux de neurones artificiels (ANN) ou réseaux de neurones simulés (SNN), Leur nom et leur structure sont inspirés du cerveau humain, imitant la façon dont les neurones biologiques se signalent les uns aux autres, les réseaux de neurones artificiels (ANN) sont constitués d'une couche de nœuds, contenant une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie. Chaque nœud, ou neurone artificiel, se connecte à un autre et possède un poids et un seuil associés. Si la sortie d'un nœud individuel est supérieure à la valeur seuil spécifiée, ce nœud est activé, envoyant des données à la couche suivante du réseau. Sinon, aucune donnée n'est transmise à la couche suivante du réseau, Il existe de nombreux modèles de réseaux de neurones sont [6]:

4.1 Modèle de Réseaux de Neurones Récurrents

RNN fonctionne sur le principe de sauvegarder la sortie d'une couche particulière et de la renvoyer à l'entrée afin de prédire la sortie de la couche. Un RNN peut gérer des données séquentielles, accepter les données d'entrée actuelles et les entrées précédemment reçues, les RNN peuvent mémoriser les entrées précédentes en raison de leur mémoire interne.

Les algorithmes de réseaux de neurones récurrents sont :

4.1.1 Mémoire longue à court terme

LSTM est une cellule composée de trois « portes » : ce sont des zones de calculs qui régulent le flot d'informations (en réalisant des actions spécifiques). La cellule de mémoire contient des paramètres de pondération pour l'entrée, la sortie et l'état interne qui sont déterminés par l'exposition au pas de temps d'entrée. Où est utilisé la fonction d'activation pour la couche cachée (identité, logistique, tanh, relu) et le solveur pour l'optimisation du poids (lbfgs, sgd, adam) [7].

4.1.2 Unité récurrente fermée

GRU est un type de réseau neuronal récurrent. Il est similaire à un LSTM, mais n'a que deux portes - une porte de réinitialisation et une porte de mise à jour - et manque notamment d'une porte de sortie. Moins de paramètres signifie que les GRU sont généralement plus faciles/rapides à former que leurs homologues LSTM [6].

4.1.3 Fusion structurée en journal

est une structure de données avec des caractéristiques de performance qui la rendent attrayante pour fournir un accès indexé aux fichiers avec un volume d'insertion élevé, tels que les données de journal transactionnel. Les arbres LSM, comme les autres arbres de recherche, conservent des paires clé-valeur. Les arborescences LSM conservent les données dans deux structures distinctes ou plus, chacune étant optimisée pour son support de stockage sous-jacent respectif ; les données sont synchronisées efficacement entre les deux structures [6].

4.2 Modèle de Réseaux de Neurones Convolutifs

Tout comme les réseaux de neurones ordinaires, ils se composent de neurones avec des poids apprenables et des constantes de biais. Chaque neurone reçoit une entrée et effectue des calculs de produits ponctuels, et la sortie est le résultat de chaque classe et Un réseau neuronal convolutif se compose de plusieurs couches, les entrées sont 3D et la sortie est 3D, certaines couches ont des paramètres et certaines couches ne nécessitent pas de paramètres [6].

Les algorithmes de réseaux de neurones convolutifs sont :

4.2.1 Réseaux de neurones à convolution profonde

DCNN est un modèle pour les données structurées en graphes. Grâce à l'introduction d'une opération de diffusion-convolution, les représentations basées sur la diffusion peuvent être apprises à partir de données structurées en graphes et utilisées comme base efficace pour la classification des nœuds

4.3 Modèle de Perceptron

Le Perceptron est l'une des architectures ANN les plus simples, il est basé sur un neurone artificiel légèrement différent appelé unité de seuil linéaire (LTU) : les entrées et sorties sont maintenant des nombres (au lieu de valeurs binaires on/off) et chaque connexion d'entrée est associée à un poids. La LTU calcule une somme pondérée de ses entrées ($z = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n = w^T.x$), applique ensuite une fonction échelon à cette somme et affiche le résultat : $h_w(x) = \text{step}(z) = \text{step}(w^T.x)$ (6) [6].

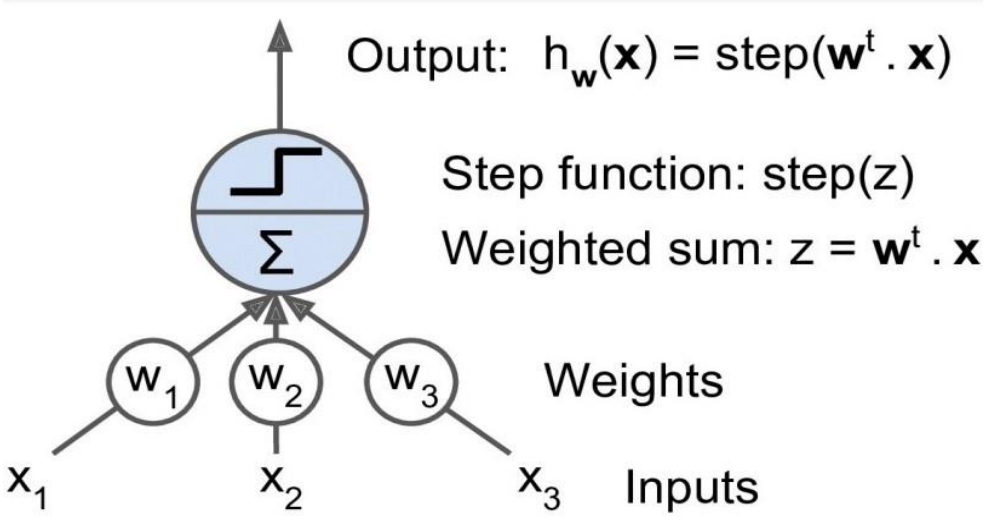

Figure 2.4 Unité de seuil linéaire.

4.4 Modèle d'Auto-Encodeurs

Les auto-encodeurs sont des réseaux de neurones artificiels capables d'apprendre des représentations efficaces des données d'entrée, appelés codages, sans aucune supervision (c'est-à-dire que l'ensemble d'apprentissage n'est pas étiqueté). Ces codages ont généralement une dimensionnalité beaucoup plus faible que les données d'entrée, ce qui rend les auto-encodeurs utiles pour la réduction de la dimensionnalité les auto-encodeurs agissent comme de puissants détecteurs de caractéristiques et peuvent être utilisés pour un pré-entraînement non supervisé de réseaux de neurones profonds.

Ils sont capables de générer de manière aléatoire de nouvelles données qui ressemblent beaucoup aux données d'apprentissage ; c'est ce qu'on appelle un modèle génératif, En bref, le codage est un sous-produit des auto-encodeurs qui tentent d'apprendre la fonction d'identité sous certaines limitations, Parmi ses applications Entraînement d'un encodeur facial automatique pour générer de nouveaux visages [6].

5 Les modèles d'ensembles

Le but des méthodes d'ensemble est de combiner les prédictions de plusieurs estimateurs de base construits avec un algorithme d'apprentissage donné afin d'améliorer la généralisabilité/robustesse sur un seul estimateur.Les modèles sont les suivants : [6]

5.1 Modèle de Boosting

Le boosting est une technique de modélisation d'ensemble qui tente de construire un classificateur fort à partir du nombre de classificateurs faibles. Cela se fait en construisant un modèle en utilisant des modèles faibles en série. Tout d'abord, un modèle est construit à partir des données d'entraînement. Ensuite, le deuxième modèle est construit qui essaie de corriger les erreurs présentes dans le premier modèle. Cette procédure se poursuit et les modèles sont ajoutés jusqu'à ce que l'ensemble de données d'entraînement complet soit correctement prédit ou que le nombre maximal de modèles soit ajouté.

Quelques algorithmes de modèle de Modèle de Boosting :

5.1.1 AdaBoost

Est un méta-estimateur qui commence par ajuster un classificateur sur l'ensemble de données d'origine, puis ajuste des copies supplémentaires du classificateur sur le même

ensemble de données, mais où les poids des instances mal classées sont ajustés de sorte que les classificateurs suivants se concentrent davantage sur les cas difficile [6].

5.1.2 CatBoost

CatBoost est un algorithme d'apprentissage automatique d'ensemble basé sur prend en charge les fonctionnalités numériques, catégorielles et textuelles, a une bonne technique de gestion des données catégorielles. L'algorithme CatBoost dispose d'un certain nombre de paramètres pour ajuster les fonctionnalités de l'étape de traitement.

5.1.3 XGBoost

XGBoost est un algorithme d'apprentissage automatique d'ensemble basé sur un arbre de décision qui utilise un cadre d'amplification de gradient. En ce qui concerne les données structurées/tabulaires petites à moyennes, les algorithmes basés sur l'arbre de décision sont actuellement considérés comme les meilleurs de leur catégorie.

5.2 Modèles de Bagging

C’est une technique d'apprentissage d'ensemble qui permet d'améliorer les performances et la précision des algorithmes d'apprentissage automatique. Il est utilisé pour traiter les compromis biais-variance et réduit la variance d'un modèle de prédiction. L'ensachage évite le surajustement des données et est utilisé à la fois pour les modèles de régression et de classification, en particulier pour les algorithmes d'arbre de décision.

Il existe un algorithme dans le modèle de remplissage qui est :

5.2.1 Forêt aléatoire

Dans les forêts aléatoires, chaque arbre de l'ensemble est construit à partir d'un échantillon tiré avec remplacement de l'ensemble d'apprentissage. et lors de la division de chaque nœud lors de la construction d'un arbre, la meilleure division est trouvée soit à partir de toutes les caractéristiques d'entrée, soit d'un sous-ensemble aléatoire de taille max_features.

5.3 Modèle de Stacking

L'empilement ou la généralisation empilée est un algorithme d'apprentissage automatique d'ensemble Il utilise un algorithme de méta-apprentissage pour apprendre à

combiner au mieux les prédictions de deux algorithmes d'apprentissage machine de base ou plus.

L'avantage de l'empilement est qu'il peut exploiter les capacités d'une gamme de modèles performants sur une tâche de classification ou de régression et faire des prédictions qui ont de meilleures performances que n'importe quel modèle unique dans l'ensemble.

6 Métriques d'évaluation pour les modèles de prédiction

Ce sont les fonctions utilisées pour calculer le taux d'erreur des modèles de prédiction et elles sont les suivantes :

6.1 Métriques d'évaluation de regression

Les métriques utilisées dans la régression sont :

6.1.1 Score de variance expliqué

EVS calcule le score de régression de la variance expliquée. Si $\hat{y}$ est la sortie cible estimée, y la sortie cible correspondante (correcte) et Var est la variance, le carré de la écart-type, alors la variance expliquée est estimée comme suit [6]:

$$explained\_variance\ (y, \hat{y}) = 1 - \frac{Var(y - \hat{y})}{Var(y)} \quad (7)$$

6.1.2 Erreur absolue moyenne

La fonction MAE calcule l'erreur absolue moyenne, une métrique de risque correspondant à l'espérance valeur de la perte d'erreur absolue ou perte de norme $l1$.Si $\hat{y}_i$ est la valeur prédite du i-ème échantillon, et y_i est la valeur vraie correspondante, alors l'erreur absolue moyenne (MAE) estimée sur $n_{\text{échantillons}}$ est définie comme [6]:

$$MAE\ (y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} |y_i - \hat{y}_i| \quad (8)$$

6.1.3 Erreur quadratique moyenne

La fonction MSE calcule l'erreur quadratique moyenne, une mesure de risque correspondant à la valeur attendue de l'erreur ou de la perte au carré (quadratique).Si $\hat{y}_i$ est la valeur prédite du i-ème échantillon, et y_i est la valeur vraie correspondante, alors MSE estimée sur n échantillons est définie comme [6]:

$$MSE\ (y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2 \quad (9)$$

6.1.4 Erreur absolue médiane

Le MAE est particulièrement intéressant car il est robuste aux valeurs aberrantes. La perte est calculée en prenant la médiane de toutes les différences absolues entre la cible et la prédiction. Si $\hat{y}_i$ est la valeur prédite du i-ème échantillon et y_i est la valeur vraie correspondante, puis MedAE estimée sur $n_{\text{échantillons}}$ est définie comme :

$$MedAE\,(y, \hat{y}) = median\left(|y_1 - \hat{y}_1|, |y_2 - \hat{y}_2|, \ldots, |y_n - \hat{y}_n|\right) \quad (10)$$

6.1.5 R au carré

La fonction R^2_score calcule le coefficient de détermination, Il représente la proportion de la variance (de y) qui a été expliquée par les variables indépendantes du modèle. Il fournit une indication de la qualité de l'ajustement et donc une mesure de la manière dont les échantillons invisibles sont susceptibles d'être prédits par le modèle, Si $\hat{y}_i$ est la valeur prédite du i-ème échantillon et y_i est la valeur vraie correspondante pour le total de $n_{\text{échantillons}}$, le R^2 estimé est défini comme suit:

$$R^2\,(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (11)$$

Où $$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad (12)$$

6.2 Métriques d'évaluation de classification

Les métriques utilisées dans la classification sont : [6].

6.2.1 Accuracy

La fonction de Accuracy calcule la précision, soit la fraction (par défaut) soit le nombre (normaliser=Faux) des prédictions correctes. Dans la classification multiétiquette, la fonction renvoie la précision du sous-ensemble. Si l'ensemble complet d'étiquettes prédites pour un échantillon correspond strictement à l'ensemble réel d'étiquettes, la précision du sous-ensemble est de 1,0 ; sinon c'est 0.0. Si $\hat{y}_i$ est la valeur prédite du i-ème échantillon et y_i est la valeur vrai correspondante, alors la fraction de prédictions correctes sur $n_{\text{échantillons}}$ est définie comme:

$$accuracy\,(y, \hat{y}) = \frac{1}{n_{samples}}\sum_{i=0}^{n_{samples}-1} 1(\hat{y}_i = y_i) \quad (13)$$

6.2.2 Confusion matrix

La fonction évalue la précision de la classification en calculant la matrice de confusion avec chaque ligne correspondant à la vraie classe. Par définition, l'entrée i, j dans une matrice de confusion est le nombre d'observations réellement dans le groupe i, mais prédites comme étant dans le groupe j. Est définie comme :

	Actual class (obervation)	
Predicted class (expectation)	tp (true positive) correct result	fp (false positive) unexpected result
	fn (false negative) missing result	tn (true negative) correct absence of result

Tableau 2.1 Confusion matrix

6.2.3 Précision et rappel

La précision est la capacité du classificateur à ne pas étiqueter comme positif un échantillon négatif, et le rappel est la capacité du classificateur à trouver tous les échantillons positifs.est définie comme :

$$precision = \frac{tp}{tp+fn} \quad recall = \frac{tp}{tp+fp} \quad (14)$$

6.2.4 F- Mesure

Les mesures F_β et $F1$ peuvent être interprétées comme une moyenne harmonique pondérée de la précision et du rappel. Une mesure F_β atteint sa meilleure valeur à 1 et son pire score à 0. Avec $\beta = 1$, F_β et $F1$ sont équivalents, et le rappel et la précision sont tout aussi importants. Est définie comme :

$$\mathcal{F}_\beta = (1 + \beta^2) \frac{precision \times recall}{\beta^2\ precision + recall} \quad (15)$$

7 Conclusion

Dans ce chapitre, nous avons mené une étude théorique des modèles d'apprentissage automatique. Dans lequel nous avons discuté les méthodes prédictives les plus importantes ainsi que leur fonctionnement. Nous avons montré la grande importance de ces méthodes dans la connaissance des valeurs futures. Cela se traduit par

l'accompagnement des grandes et moyennes entreprises afin de trouver des solutions appropriées à leurs problèmes. Maintenant, après avoir identifié les modèles de machine learning, nous allons appliquer quelques modèles au problème de la consommation d'électricité après être passé par quelques étapes, que nous aborderons dans le troisième chapitre.

CHAPITRE 3
UTILISATION DES MODÈLES DE PRÉVISION POUR LA CONSOMMATION D'ÉNERGIE ÉLECTRIQUE

1 Introduction

Après avoir donné un aperçu des modèles d'apprentissage automatique, nous devons faire des prédictions. Il est spécifiquement lié aux modèles de régression et aux modèles de classification, où nous détaillerons comment préparer les données pour appliquer les modèles et comment les utiliser.

Nous commencerons d'abord dans ce chapitre par présenter quelques études antérieures similaires à ce sujet, puis nous ferons une introduction à la Société nationale algérienne de l'électricité et du gaz, puis nous appliquerons des modèles de régression à ses données, qui sont la consommation d'électricité en mégawatts par heure pour le pôle d'Adrar ainsi que le territoire national sur une période de deux ans. En plus des données réelles obtenues, nous appliquerons les algorithmes de prévision sur des données synthétiques pour montrer l'utilité de certaines fonctionnalités (features) et leur influence sur la consommation en énergie.

2 Travaux Reliés

Dans cette section nous résumons un ensemble de travaux récents reliés à notre problème (5 articles) :

Article 1: Comparison of Linear Rregressions and Neural Networks for Forecasting Electricity Consumption

Cette étude utilise 2 ensembles de données relatives à la consommation d'électricité. Le modèle de régression linéaire et de réseau de neurones a été utilisé. Cette étude veut prouver qu'il existe une relation non linéaire dans l'ensemble de données de consommation d'électricité utilisé dans cette étude, et comparer les meilleures performances entre l'utilisation de la régression linéaire et les réseaux de neurones.

Où L'utilisation des réseaux de neurones montre une valeur RMSE plus petite par rapport à l'utilisation de linéaire régressions. Cela montre que le problème qui est prouvé dans le jeu de données de consommation d'électricité A et l'ensemble de données de consommation d'électricité B est qu'il a une relation non linéaire, elle peut être surmontée par les réseaux de neurones. Pour que les réseaux de neurones puissent améliorer les performances mieux que les régressions linéaires [8].

Article 2: Evaluation and Improvement of Energy Consumption prediction models using principal component analysis based feature reduction

Dans cet article, une méthode de prédiction hybride basée sur l' PCA est proposée pour prévoir la consommation énergétique du bâtiment au fil du temps. L'objectif de cette étude étant donné les grands ensembles de données en constante augmentation, des algorithmes étendus doivent être utilisés pour améliorer la prédiction. La méthode PCA est introduite en tant que méthode de réduction des données et le prétraitement des données est effectué pour cinq modèles de prédiction, y compris la régression linéaire arbre de régression, forêt aléatoire et KNN. De plus, quatre types d'ensembles de données (quatre modèles de consommation d'énergie) sont rassemblés pour étudier l'effet de la méthode de prétraitement sur les performances des modèles de prédiction. Les résultats indiquent que la méthode PCA peut être utile et les résultats confirment que cette approche pourrait être appliquée à tout autre modèle de prédiction d'énergie avec de grands ensembles de données, résultant en une prédiction précise avec un temps d'exécution considérablement réduit [9].

Article 3: Forecast electricity demand in commercial building with machine learning models to enable demand response programs

Cette étude vise à la création de modèles avec un ajustement fort et précis pour traiter les caractéristiques non linéaires. Sur la base de l'ensemble de données de charge réelle, où cette étude compare les deux principales techniques de prévision de charge à court terme. Avant de parler sur les expériences menées, cet article énumère d'abord les méthodes courantes de prévision de la charge à court terme et explique les principes des réseaux à mémoire longue à court terme (LSTM) et des machines à vecteurs de support (SVM) utilisés. En fonction des caractéristiques de l'ensemble de données de charge électrique, du prétraitement des données et de la fonctionnalité la sélection a lieu.

Cet article décrit les résultats d'une expérience contrôlée pour étudier l'importance de la sélection des fonctionnalités. Les modèles LSTM et SVM sont appliqués à la prévision de charge d'une heure à l'avance et prévision de la charge de pointe et de vallée à un jour à l'avance. La précision prédictive de ces modèles est calculée sur la base de l'erreur calculée entre les charges réelles et prévues, et le temps d'exécution du modèle est enregistré. Les résultats montrent que le modèle LSTM a une précision de prédiction plus élevée lorsque les données de charge sont suffisantes. Cependant, les performances globales du modèle SVM sont meilleures lorsque les données de charge utilisées pour former le modèle sont insuffisantes et le coût du temps est prioritaire [10].

Article 4: Research on Electricity Consumption Forecasting model based on wavelet transform and multi-layer LSTM model

Visant les caractéristiques des séries chronologiques des données de consommation d'électricité, cet article propose un modèle de prévision basé sur la combinaison de la transformée en ondelettes et de multiples LSTM. Grâce à la formation et à la prédiction d'échantillons de données et à la comparaison horizontale avec les algorithmes traditionnels LSTM et Bi-LSTM, les résultats expérimentaux montrent directement que la méthode de cet article a considérablement amélioré la précision de la prédiction de la consommation quotidienne d'électricité, et le traitement de réduction du bruit de WT peut être utilisé dans une certaine mesure. Améliorer la stabilité et la précision du modèle. Cette méthode peut mieux prédire la consommation d'électricité quotidienne, aider à formuler des plans raisonnables de production et de transmission d'électricité et prévenir efficacement le gaspillage des ressources électriques [11].

Article 5 : Optimization of industrial Energy Consumption for sustainability using time-series regression and gradient descent algorithm based on historical Electricity Consumption data

La méthodologie appliquée dans cette étude a été développée en utilisant une technique d'optimisation de descente de gradient combinée à une analyse de régression pour minimiser le gaspillage d'électricité dans les industries. En utilisant des données historiques de quatre ans d'une industrie pétrolière pour la période de 2015 à 2018, l'erreur quadratique moyenne a été utilisée comme fonction de coût dans le processus d'optimisation. Dans une étude de cas, l'algorithme en déterminant les meilleurs points de

fonctionnement des variables indépendantes contribuant au coût mensuel total de l'électricité [12].

3 Présentation de la société Sonelgaz

SONELGAZ est l'une des plus anciennes fondations connues en Algérie, c'est une entreprise publique d'électricité et de gaz qui contribue efficacement au développement économique et industriel du pays. Afin de présenter davantage cette entreprise, nous discuterons de sa création et de son développement, des produits qu'elle propose à ses clients, ainsi que de ses tâches et fonctions. A la fin, nous présentons les objectifs auxquels l'entreprise transcende.

3.1 Origine et développement de la société Sonelgaz

Sonelgaz est considérée comme l'une des plus importantes grandes entreprises d'Algérie. Pendant cinquante ans, elle a dessiné les plus belles pages du développement économique et social dans la production d'énergie électrique ainsi que la production et la distribution. Ses pouvoirs se sont étendus à la vente, l'installation et la maintenance.

La Fondation Sonelgaz est passée par plusieurs étapes dans sa création et son développement, comme suit :

Année 1947 : Construction de l'électricité et du gaz en Algérie EGA N° 47-1002 du 5 juin 1947.

EGA : L'ensemble des plus anciennes sociétés de production et de distribution d'électricité relevant de la loi de nationalisation N° 46-628 du 04/08/1946 promulguée par l'autorité française.

Année 1969 : Remplacement de l'EGA par la création de Sonelgaz (création de la Société nationale de l'électricité et du gaz) par ordonnance N° 69-59 du 28 Juillet 1969 et une nouvelle ère avec un effectif de 6000 agents Sonelgaz d'une taille respectable.

Année 1983 : Le premier processus de restructuration de Sonelgaz, qui a entraîné avec lui six établissements de service public, comme suit :

KAHRIF (électrification rurale).

KAHRAKIB (infrastructures et installations électriques).

KANAGHAZ (réalisation des réseaux gaz).

INERGA (Génie Civil).

ETTERKIB (montage industriel).

AMC (fabrication des compteurs et appareils de mesure et de contrôle).

Année 1995 : Sonelgaz (EPIC) en 1995, elle devient un établissement public à caractère industriel et commercial par le décret exécutif N° 95-280 du 17 septembre 1995

Année 2002 : Sonelgaz devient une Société par Actions (SPA). Par le Décret présidentiel n° 02-195 du 01 Juin 2002, Elle est régie par les dispositions de la loi relative à l'électricité et à la distribution du gaz par canalisations et par les dispositions du code de commerce

Ce statut lui donne la possibilité d'élargir ses activités à d'autres domaines relevant du secteur de l'énergie et aussi d'intervenir à l'international

Année 2004 : Elle adopte une organisation de Groupe industriel par la transformation en filiales de ses entités en charge des métiers de base :

Production d'Electricité (SPE).

Transport d'Electricité (GRTE).

Conduite du Système Electrique (OS).

Transport du Gaz (GRTG).

Distribution de l'Electricité et du Gaz d'Alger (SDA) : du Centre (SDC), de l'Est (SDE) et enfin de l'Ouest (SDO)

Année 2009 : Sonelgaz adopte une nouvelle organisation. Celle-ci aboutit à un Groupe comptant 33 filiales et 6 Sociétés en participation directe, Avec l'ouverture de l'Institut de Formation en Electricité et Gaz (IFEG) en 2007, ainsi que la création des sociétés d'engineering, des systèmes d'information et de la gestion immobilière (CEEG, ELIT & SOPIEG) et l'intégration de la Société Rouïba Eclairage en 2009

A partir de 2011, Sonelgaz devient une holding, puis elle et ses filiales forment ce qu'on appelle le Groupe Sonelgaz (voir Fig. 3.1), puis se tournent vers le domaine des énergies renouvelables et montent des projets à dimension internationale [13].

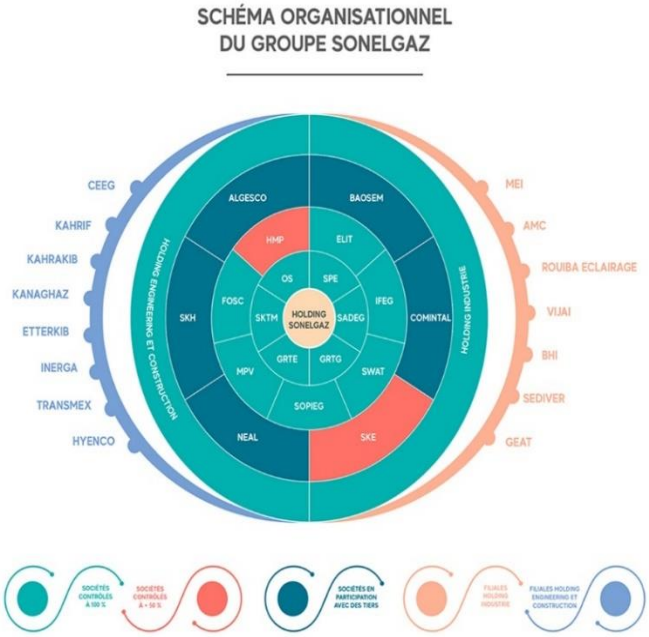

Figure 3.1 Image montrant le groupe Sonelgaz actuel

3.2 Produits Sonelgaz

Sonelgaz, comme toute autre entreprise, fournit des produits à ses clients et Les produits proposés par Sonelgaz sont les suivants :

3.2.1 Production du gaz naturel

Sonelgaz produit du gaz naturel, qui est transporté et distribué sur tout le territoire national. Les statistiques suivantes représentent des informations sur le réseau gazier :

- Taux de pénétration du Gaz Naturel : 65%
- Réseau de Transport : 22 623 km
- Réseau de Distribution : 126 952 km
- Longueur du réseau Gaz : 149 575 km
- Nombre de clients : 6 450 538

3.2.2 Production de l'énergie électrique

Où l'électricité est-elle produite, transportée et distribuée sur le territoire national, Les figures suivantes donnent des informations sur le réseau électrique :

- Taux d’électrification : 98 %
- Capacité installée : 22 979 MW
- Réseau de Transport : 31 164km

- Réseau de Distribution : 357 185km
- Longueur du réseau Electrique : 388 349 km
- Nombre de clients : 10 494 465

3.2.3 L'énergie renouvelable

Cette énergie est produite par les centrales de types :

- Centrale photovoltaïque : 356,1 MW
- Centrale éolienne : 10,2 MW [13]

3.3 Emplois de l'entreprise

La société Sonelgaz offre à ses produits plusieurs fonctions dont :

- Assurer la qualité de la production, du transport et de la distribution de l'énergie électrique ainsi que de la distribution du gaz dans le cadre du respect des conditions de protection et de sécurité.
- Installation, réparation et maintenance de centres de production, transport et distribution d'électricité, en plus des centres de distribution de gaz.
- Élaborer des études et des plans annuels pour assurer le bon déroulement des programmes.
- Assurer la sécurité d'approvisionnement du réseau électrique et la continuité de service et d'approvisionnement est nécessaire.
- Œuvrer pour consolider et représenter la position de Sonelgaz en tant qu'entreprise nationale de premier plan.
- De manière générale, la société Sonelgaz garantit la réalisation des investissements appartenant à l'établissement et la maîtrise de l'énergie qui contribue au développement économique et industriel du pays [13].

3.4 Les objectifs de l'entreprise

La société Sonelgaz s'efforce, par les missions et fonctions qu'elle exerce, d'atteindre un certain nombre d'objectifs et de résultats qui se sont fixés comme suit :

- Agir pour répondre à la demande toujours croissante en énergie électrique par l'utilisation optimale des ressources de base et en tirer profit tout en préservant l'environnement.

- Les sociétés du Groupe Sonelgaz envisagent de développer les activités de production, de transport et de distribution d'électricité, ainsi que les activités de transport et de distribution de gaz.
- Développement et amélioration dans le domaine des services énergétiques pour développer et diversifier ses produits.
- La poursuite du développement des équipements de maintenance et d'exploitation et des projets de modernisation de la gestion et de l'exploitation pour participer aux réalisations industrielles et commerciales dans le monde pour atteindre le client final.
- De manière générale, Sonelgaz Electricité et Gas Compagnie est une entreprise à caractère industriel et commercial qui cherche toujours à participer à la compétitivité mondiale pour obtenir une part du marché mondial [13].

4 Étapes de la prévision à l'aide de modèles d'apprentissage automatique

Le problème auquel nous allons faire face est de projeter la consommation d'électricité en Algérie (territoire national) et en particulier le pôle industriel d'Adrar à partir des données de consommation d'électricité des deux dernières années, c'est un problème supervisé. Il est modéré car nous avons les caractéristiques et les objectifs que nous voulons prédire.

Tout d'abord, nous allons construire un modèle de régression pour prédire la consommation d'électricité en utilisant 4 algorithmes pour les données Adrar et les données à l'échelle nationale de l'Algérie à l'heure, au jour et à la semaine. Ensuite, nous allons effectuer des prédictions sur des données synthétiques dans le but de montrer l'influence de certaines caractéristiques sur la prédiction de la consommation en énergie électrique. Ces données ont été générées en raison de l'absence d'un ensemble de données réelles qui montre, outre l'heure, d'autres attributs liés à la consommation d'énergie comme la température, le niveau de vie de la population, etc. Nous appliquons des modèles de classification sur ces données afin de montrer l'efficacité et l'utilité de la prédiction, puis évaluer les performances des modèles utilisés et visualiser les résultats (voir Fig. 3.2).

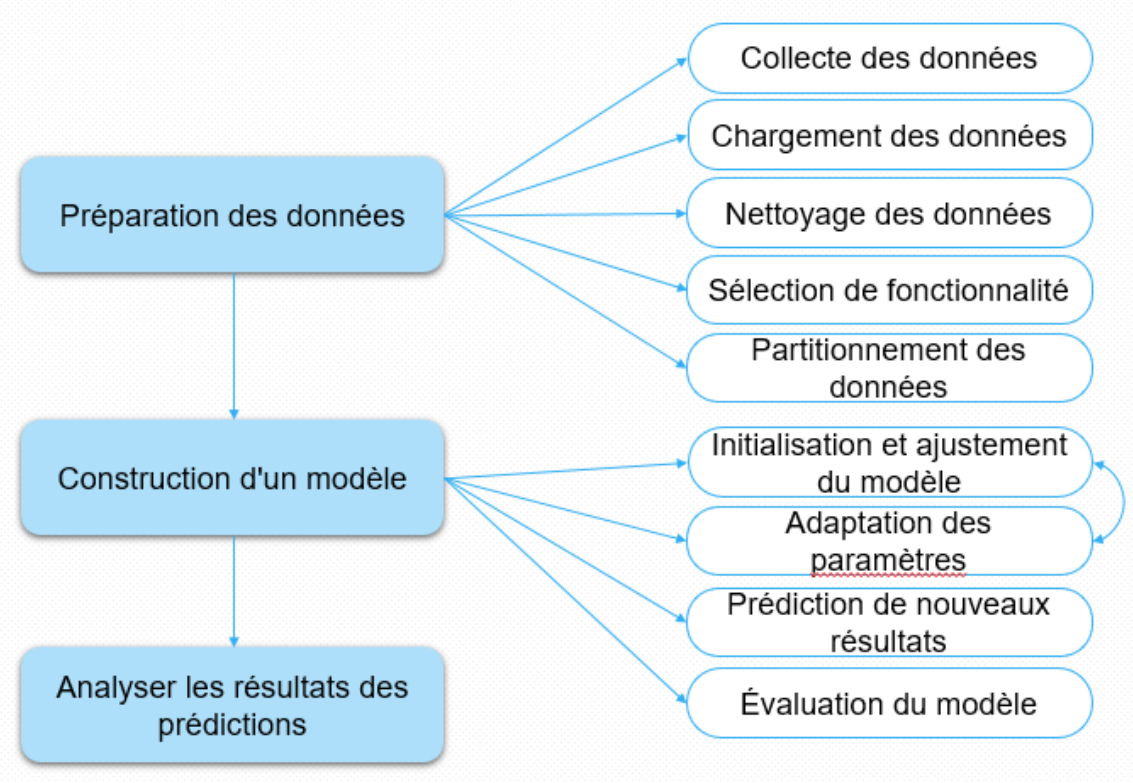

Figure 3.2 Étapes de la prévision à l'aide de modèles d'apprentissage automatique

4.1 Préparation des données

4.1.1 Collecte des données

A- Données réelles

Après avoir fait plusieurs tentatives pour obtenir des données de consommation d'électricité pour l'état de M'sila ou l'une des cités de M'sila auprès de Sonelgaz, nous n'avons pas pu les obtenir. Nous avons par la suite préparé 2 ensembles de données pour l'expérimentation via le site de l'opérateur du système électrique https://www.os.dz/.

Le site Internet de l'opérateur du système électrique contient des données sur la consommation d'électricité en Algérie et au Pôle Adrar pour chaque jour de l'année pendant chaque quart d'heure.

- **Ensemble de données du Pôle Adrar :**

Nous avons pris la consommation d'électricité à Adrar pendant deux ans 2018-2019, où nous l'avons organisée dans un tableau Excel au format CSV (voir tableau 3.1), créé une colonne de temps et une colonne d'heure, et mis les valeurs de consommation en mégawatt/heures (MWh) dans la dernière colonne.

	Date	Hour	Consumption
0	2018.01.01	0:00:00	20.7500
1	2018.01.01	1:00:00	18.7500
2	2018.01.01	2:00:00	18.2500
3	2018.01.01	3:00:00	17.5000
4	2018.01.01	4:00:00	17.6875
5	2018.01.01	5:00:00	17.7500
6	2018.01.01	6:00:00	19.6250
7	2018.01.01	7:00:00	22.9375
8	2018.01.01	8:00:00	23.4375
9	2018.01.01	9:00:00	24.1875

Tableau 3.1 Ensemble de données de pôle Adrar 2018-2019

- **Ensemble de données du territoire national**

Nous avons pris la consommation d'électricité du territoire national (Algérie) pour deux ans 2018-2019, où nous l'avons organisée dans un tableau Excel au format CSV (voir tableau 3.2), créé une colonne de temps et une colonne d'heure, et mis les valeurs de consommation en mégawatt-heures (MWh) dans la dernière colonne, comme nous l'avons fait avec les données du pôle d'Adrar.

	Date	Hour	Consumption
0	2018.01.01	0:00:00	1663.5000
1	2018.01.01	1:00:00	1589.0625
2	2018.01.01	2:00:00	1544.6875
3	2018.01.01	3:00:00	1525.0000
4	2018.01.01	4:00:00	1523.3125
5	2018.01.01	5:00:00	1544.0000
6	2018.01.01	6:00:00	1633.3750
7	2018.01.01	7:00:00	1707.3750
8	2018.01.01	8:00:00	1739.5000
9	2018.01.01	9:00:00	1806.0000

Tableau 3.2 Ensemble de données national (Algérie) 2018-2019

B- Données Synthétiques

Les données synthétiques ont été générées à l'aide d'un algorithme en Python, puis converties en fichier Excel au format CSV (voir tableau 3.3). La raison de ces données est :

- Manque de données historiques réelles de taille suffisante (plus de cinq ans au moins)

- Nous avons besoin de plus de caractéristiques qui affectent la consommation d'électricité afin de garantir l'exactitude et la précision des prévisions futures.

Ces données ont les caractéristiques (features) suivantes :

Date et heure : c'est la date et l'heure enregistrées lors de la prise de la consommation. Les durées varient entre le 01.01.2018 et 31.12.2019.

Population : Cet attribut donne le nombre des habitants de la région concernée par l'étude. Nous avons donné des valeurs entières entre 20 000 et 53 000, suite à plusieurs contacts et recherche sur l'évolution de la population durant les années 2018 et 2019 de la région concernée par l'étude.

Capacité du transformateur : c'est l'attribut qui donne une information sur le transformateur utilisé, notamment sa capacité. Il est important de savoir quelle quantité d'électricité un transformateur peut supporter, nous leur avons donné des valeurs de type entier compris entre 50 et 200.

Urbanisation : Nous avons besoin de savoir quel taux d'urbanisation relatif à la région étudiée. Ceci donne une idée sur l'expansion de l'urbanisation de la région et ainsi sur la demande en énergie électrique. Nous lui avons donné des ratios flottants artificiels entre 1,5 et 6,5.

Holiday : Cet attribut permet de déterminer si le jour de la date est un jour férié (ou de vacances) ou non. Cette information est intéressante car elle permet de savoir si la demande en énergie est en hausse durant ce jour ou non. Nous avons donné une valeur booléenne : 1 si le jour est Holiday, 0 sinon.

Température : C'est l'attribut qui permet de donner la température enregistrée durant la journée. Nous l'avons divisée en 3 catégories : haute, moyenne et basse.

Consommation : c'est la colonne qui donne la consommation en heure de l'électricité. Pour les modèles de classification, nous avons réparti la consommation en trois classes : basse (Inférieure à 23MWh), moyenne (entre 23 et 43 MWh) et haute (supérieure à 43 MWh).

	DateHour	Population	c_transfo	Urbanization	n_holiday	temperature	Consumption
0	2018-01-01 00:00:00	28827	139	4.9	True	low	1663.5000
1	2018-01-01 01:00:00	20440	100	6.4	True	low	1589.0625
2	2018-01-01 02:00:00	23249	116	4.3	True	low	1544.6875
3	2018-01-01 03:00:00	23080	181	1.5	False	low	1525.0000
4	2018-01-01 04:00:00	42610	153	2.0	True	low	1523.3125
5	2018-01-01 05:00:00	45101	98	1.9	True	low	1544.0000
6	2018-01-01 06:00:00	21073	134	1.7	True	low	1633.3750
7	2018-01-01 07:00:00	32177	169	2.0	False	low	1707.3750
8	2018-01-01 08:00:00	30005	92	5.2	False	low	1739.5000
9	2018-01-01 09:00:00	46446	176	3.0	True	low	1806.0000

Tableau 3.3 Ensemble de données synthétiques 2018-2019

4.2.2 Chargement des données

Nous commençons par importer la bibliothèque pandas, puis nous lirons le contenu des données des fichiers CSV à l'aide de la fonction read_csv() en Python.

4.2.3 Nettoyage des données

Le nettoyage des données est le processus de correction ou de suppression des données corrompues, incorrectes ou inutiles d'un ensemble de données avant l'analyse des données.

A- Traitement des valeurs manquantes

Les données contiennent souvent beaucoup de valeurs manquantes, pour plusieurs raisons. Nous citons par exemple :

- Les données n'existent pas.
- Données non collectées en raison d'une erreur humaine.
- Données supprimées accidentellement.

Plusieurs techniques sont utilisées pour traiter les valeurs manquantes. Nous en citons les plus utilisées dans ce qui suit :

❖ **Techniques de Suppression**

Supprimer les valeurs manquantes d'un jeu de données. Les suppressions sont de deux types :

- Supprimer les lignes contenant des valeurs manquantes.
- Supprimer les colonnes contenant des valeurs manquantes.

❖ **Techniques de remplissage**

- Utiliser la méthode fillna () :

La fonction fillna () nous permet de travailler avec des cellules vides en entrant une nouvelle valeur sans avoir à supprimer les valeurs. Souvent utilisé pour les données de séries temporelles. On peut remplacer les valeurs nulles par :

- Remplir les lignes manquantes avec des valeurs à l'aide de ffill : Cette méthode remplit chaque ligne manquante avec la valeur de la plus proche au-dessus.
- Remplir les lignes manquantes avec des valeurs à l'aide de bfill qui remplit chaque ligne manquante avec la valeur la plus proche en dessous.
- Remplir les valeurs manquantes avec la moyenne, la médiane ou le mode : les cellules vides sont remplacées en calculant la valeur moyenne, la médiane ou le mode de la colonne ou certaines lignes spécifiques de la colonne si les valeurs sont numériques.

- **Utiliser la méthode SimpleImputer ()**

La fonction SimpleImputer () fonctionne avec des séries de données non temporelles pour remplacer les cellules vides par une nouvelle valeur. Vous pouvez remplacer les valeurs nulles par :

- Imputation à valeur constante.
- Imputation à partir des statistiques (moyenne, médiane ou la plus fréquente) de chaque colonne dans laquelle se situent les valeurs manquantes.

- **Utiliser la méthode d'interpolation linéaire**

L'interpolation linéaire signifie simplement estimer une valeur manquante en reliant des points en ligne droite dans un ordre croissant. En bref, il estime la valeur inconnue dans le même ordre croissant à partir des valeurs précédentes. La méthode par défaut utilisée par Interpolation est Linéaire, donc lors de son application, nous n'avons pas eu besoin de la spécifier.

- **Utiliser la méthode K-Nearest Neighbor**

Les valeurs manquantes de chaque échantillon sont imputées à l'aide de la valeur moyenne de n_neighbors les plus proches voisins trouvés dans l'ensemble d'apprentissage. Deux échantillons sont proches si les caractéristiques qui ne manquent pas sont proches.

B- Suppression des doublons

La suppression des doublons signifie la suppression de toutes les valeurs en double. Il n'y a pas besoin de valeurs en double dans l'analyse des données. Ces valeurs n'affectent que la précision et l'efficacité du résultat de l'analyse, nous utilisons donc la fonction drop_duplicates () pour supprimer les colonnes ou les lignes en double

4.2.4 Standardisation des données

Les données réalistes ont des caractéristiques qui varient en taille, en unités et en plage. La mise à l'échelle des fonctionnalités aide essentiellement à normaliser les données dans une certaine plage. Nous devons donc importer StandardScalar () de la bibliothèque sklearn et l'appliquer à notre ensemble de données.

4.2.5 Sélection des fonctionnalités

L'objectif de la sélection des caractéristiques est de trouver le meilleur ensemble de caractéristiques dont nous avons besoin dans le processus de prédiction. Afin d'augmenter l'efficacité du modèle, nous sélectionnons les caractéristiques d'entrée et les caractéristiques de sortie.

4.2.6 Partitionnement des données

Après avoir terminé le traitement des données, nous divisons les données en deux parties, une partie est un ensemble de traitement à 70 %. Pendant la formation, nous laissons le modèle "voir" les réponses, afin qu'il puisse comprendre comment prédire la consommation d'électricité de la caractéristique. Et l'autre partie est une suite de tests à 30 %. L'ensemble de traitement est utilisé pour traiter le modèle et l'ensemble de test est utilisé pour tester la précision du modèle en comparant ces prédictions avec la valeur réelle. La plupart des données sont utilisées pour le traitement et une petite partie des données est utilisée pour les tests.

4.2 Construction du modèle de prédiction

Nous avons appliqué deux types de modèles supervisés de prédiction sur les données réelles et synthétiques que nous avons obtenues, à savoir, les modèles de régression et les modèles de classification (voir Fig. 3.3).

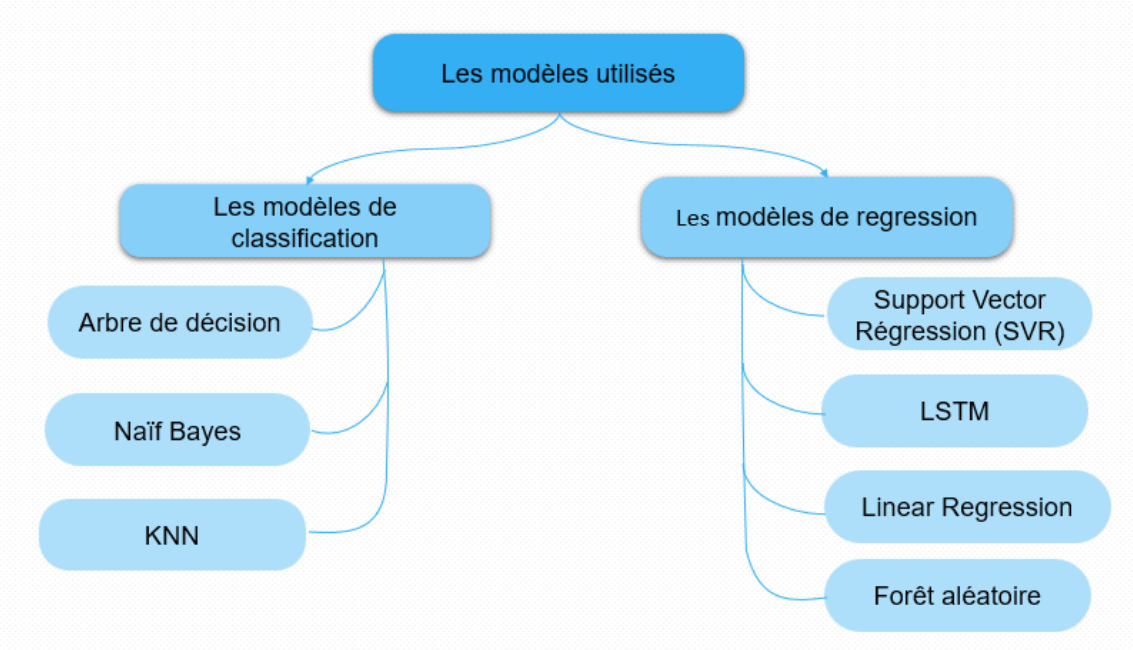

Figure 3.3 Les Modèles de prévision utilisés

4.2.1 Les modèles de régression

A- Modèle SVR

La régression vectorielle de support est une sorte de machine à vecteur de support qui prend en charge la régression linéaire et non linéaire. La première ligne importe la fonction SVR () de "sklearn.svm". Ensuite, la fonction SVR () est affectée à une variable. Le paramètre le plus important de SVR est le type de noyau. Il peut être linéaire, polynomial ou gaussien SVR. On a une condition non linéaire donc on peut définir un polynôme ou une gaussienne mais ici on choisit un noyau RBF (de type gaussien) On modifie aussi le paramètre de régularisation C et epsilon.

B- Modèle des réseaux de neurones LSTM

Les réseaux de neurones récurrents (RNN) sont des types de réseaux de neurones conçus pour utiliser des données séquentielles telles que des séries chronologiques. Dans les RNN, la mémoire à long terme (LSTM) est la nouvelle génération de réseau de neurones récurrents (RNN), nous allons donc aujourd'hui appliquer ce modèle à nos

données. Tout d'abord, nous initialisons le modèle en tant que Sequential (), puis nous avons un opérateur caché qui formate les données d'entrée comme input_shape Et nous mettons 10 blocs LSTM ou neurones, puis nous avons une couche de sortie avec un bloc LSTM pour faire une prédiction d'une valeur. Et nous avons utilisé la fonction d'activation relu.

Deuxièmement, nous ajoutons la fonction de compilation en définissant Adam comme une échelle améliorée pour calculer la perte afin de recycler le modèle sur les poids du modèle et d'utiliser l'erreur quadratique moyenne.

Enfin, nous entraînerons notre modèle à des intervalles (itérations) de 30 fois et une taille de lot de 5.

C- Modèle de combinaison de LinearRegression et PolynomialFeature

L'algorithme de régression linéaire simple est un algorithme qui ne traite que lorsque la relation entre les données est linéaire et puisque nos données sont des données non linéaires, la régression linéaire ne pourra pas tracer la meilleure ligne appropriée, pour cela nous avons utilisé la régression polynomiale qui est un forme de régression linéaire de En convertissant les variables d'entrée en termes polynomiaux en utilisant un certain degré et dans notre exemple, nous avons utilisé 3 degrés, puis nous l'entraînons.

D- Modèle de forêt aléatoire

Une forêt aléatoire est un ensemble d'algorithmes d'arbre de décision. Nous formons un algorithme de forêt aléatoire pour résoudre un problème de régression en utilisant la classe RandomForestRegressor de la bibliothèque sklearn.ensemble . Le paramètre le plus important de la classe RandomForestRegressor est le paramètre n_estimators. Ce paramètre précise le nombre d'arbres dans la forêt aléatoire, on va commencer avec n_estimator = 10 et la profondeur maximale de l'arbre avec 4.

4.2.2 Les modèles de classification

La classification a été utilisée pour prédire des données aléatoires en divisant la consommation en 3 catégories : élevée, moyenne et faible.

A- Modèle de classificateur d'arbre de décision

L'arbre de décision est un algorithme qui est une structure graphique en forme d'arbre (voir Fig. 3.4) qui utilise divers paramètres affinés pour prédire les résultats. Parmi ces

paramètres, nous avons un critère qui est une fonction pour mesurer la qualité de la division. Où nous avons choisi "gini" pour l'indice de Gini.

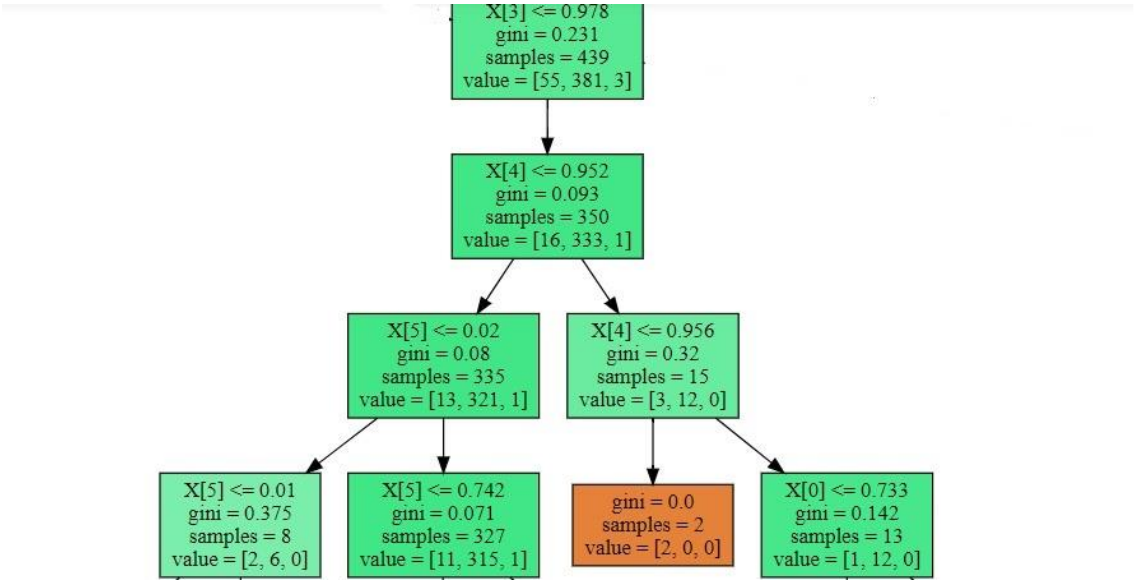

Figure 3.4 Une partie de l'arbre de décision

B- Modèle Naïve Bayes

Naive Bayes est un modèle de classification probabiliste simple, nous avons formé notre modèle avec un classificateur Gaussian Naive Bayes importé de SKlearn GaussianNB. Ce classificateur est utilisé lorsque les valeurs des prédicteurs sont de nature continue et qu'il est supposé qu'elles suivent une distribution gaussienne.

C- Classificateur de K voisins

Afin de construire un modèle de classificateur KNN, nous importons d'abord le module KNeighborsClassifier et créons un objet classificateur KNN en passant un argument nombre de voisins dans la fonction proche KNeighborsClassifier () que nous définissons sur 5 voisins et utilisons la distance dans le paramètre weights.

4.2.3 Test du modèle de prédiction

Maintenant que le modèle a été construit et formé pour connaître les relations entre les caractéristiques et les objectifs, nous allons l'appliquer pour faire des prédictions sur l'ensemble de test via la fonction predict () dans l'ensemble de test, puis comparer les prédictions avec les valeurs réelles (voir tableau 3.4) [14].

	Real Values	Predicted Values
0	2705.9375	2722.459638
1	1850.1875	1712.357861
2	2364.6875	2381.232551
3	1545.7500	1403.673643
4	2743.6250	2684.702664
5	1911.1250	2118.238283
6	2015.8750	2042.705321
7	2111.1875	2164.648631
8	2000.5625	1946.962765
9	2250.9375	2268.226193

Tableau 3.4 Comparaison entre les valeurs réelles des données nationales et les valeurs prédites

4.2.4 Evaluation du modèle de prédiction

L'étape suivante consiste à déterminer l'efficacité du modèle pour mesurer l'efficacité de différents algorithmes basés sur des métriques et des ensembles de données. Les métriques varient pour différents algorithmes d'apprentissage automatique. Pour les modèles de régression, nous pouvons utiliser des mesures de performance de régression telles que MSE (erreur quadratique moyenne), R2_score et MAE (erreur absolue moyenne) pour calculer l'efficacité du modèle. Quant au modèle de classification, nous utiliserons accuracy_score et la matrice de confusion.

4.2.5 Visualisation du résultat

Nous avons utilisé la visualisation de données, où tous les résultats seront convertis en un modèle graphique et analysées pour comprendre et comparer les résultats obtenus à l'aide de bibliothèques Python telles que Matplotlib et Seaborn.

5 Conclusion

Dans ce chapitre, nous avons présenté les grands axes de notre étude qui concerne l'utilisation des modèles de machine learning pour la prédiction de la consommation d'énergie.

Après avoir donné un aperçu sur l'état de l'art de quelques travaux antérieurs sur le domaine de la prédiction d'énergie, nous avons présenté la société Sonelgaz responsable sur la production et l'offre de l'énergie électrique en Algérie. Ensuite, nous décrivons la

démarche de notre étude, depuis la préparation des données jusqu'à la visualisation des résultats.

Le chapitre suivant sera consacré à la présentation des résultats de notre étude, ainsi que la comparaison des différents modèles utilisés.

CHAPITRE 4
RÉSULTATS DES TESTS ET VALIDATION

1 Introduction

Dans ce chapitre, une évaluation des performances des algorithmes d'apprentissage automatique pour la prédiction, représentée par des modèles de régression et des modèles de classification, est présentée et nous comparerons ces modèles pour trouver le modèle le plus efficace pour aider à prendre la meilleure décision de prédiction. La capacité prédictive d'un algorithme de classification ou de régression est généralement mesurée par sa précision prédictive ou son taux d'erreur, dans des exemples de test. Nous afficherons les résultats sous forme de tableaux et de graphiques pour une observation précise.

2 Analyse des résultats de prédiction

2.1 Pôle Adrar

Nous présentons dans cette section les résultats de la prédiction de la valeur de la consommation d'électricité pour les données du pôle Adrar à l'aide d'un modèle de régression d'apprentissage automatique avec quatre modèles différents à travers le nuage de points, qui est un graphique dans lequel les valeurs de deux variables sont tracées sur deux axes. Les valeurs des deux variables sont représentées dans les valeurs réelles des données de test où l'on ne prend qu'une petite partie et leurs valeurs attendues :

2.1.1 Régression par Forêt aléatoire

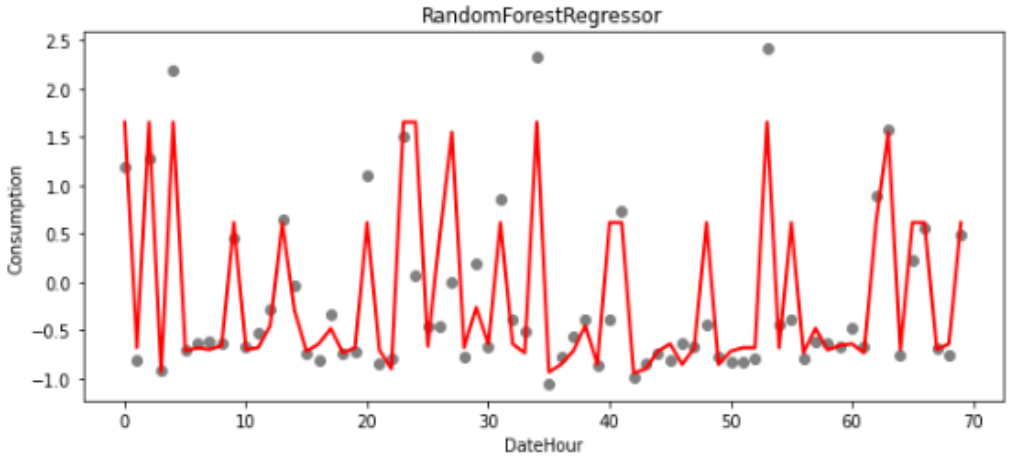

Figure 4.1 Prédiction de données Adrar par forêt aléatoire.

2.1.2 Régression à vecteurs de support

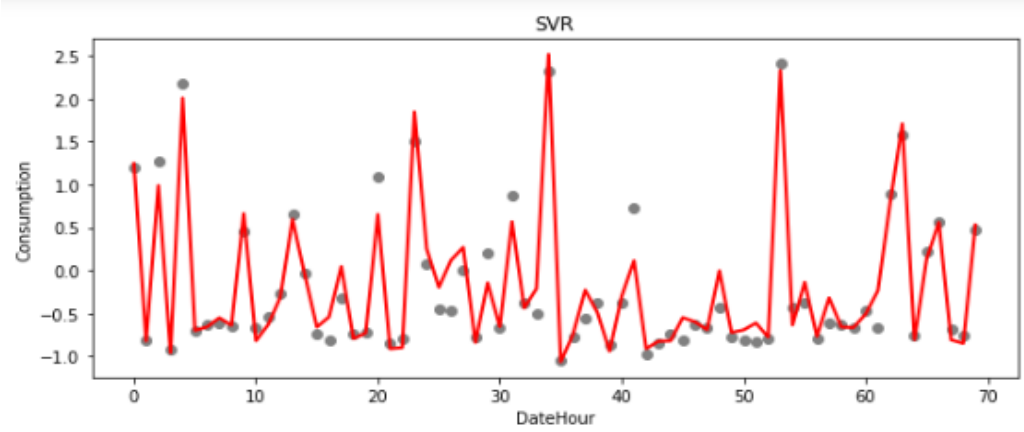

Figure 4.2 Prédiction de données adrar par modèle SVR.

2.1.3 Modèle de combinaison de LinearRegression et PolynomialFeatures

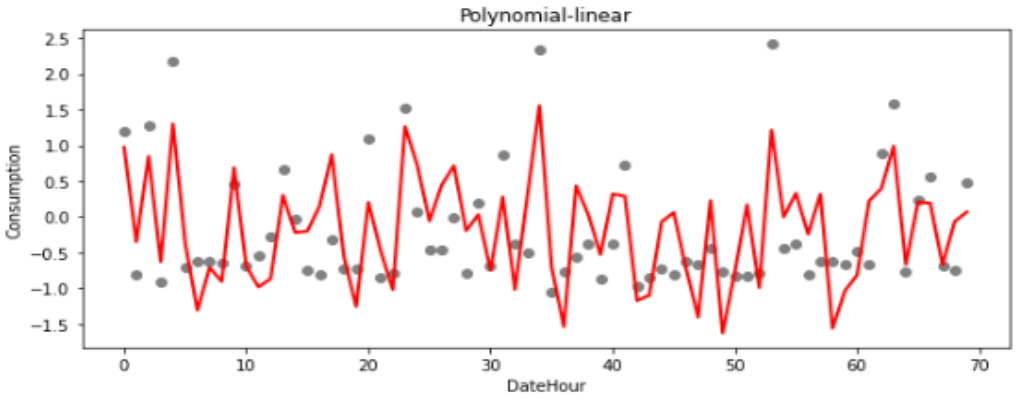

Figure 4.3 Prédiction de données adrar par LinearRegression et PolynomialFeatures

2.1.4 Modèle LSTM

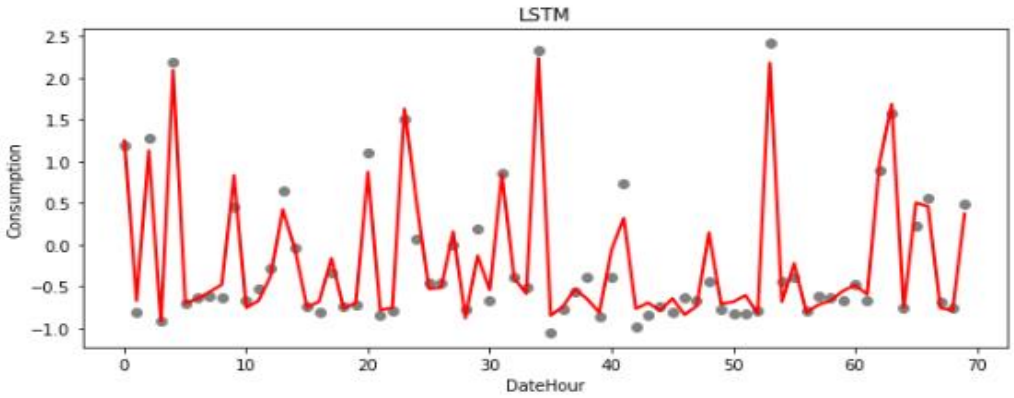

Figure 4.4 Prédiction de données adrar par modèle LSTM.

2.1.5 Synthèse des résultats obtenus

Après avoir présenté le schéma de diffusion pour tous les modèles (voir tableau 4.1) nous présentons quelques observations. Le schéma de diffusion pour le modèle de forêt aléatoire montre qu'il a fourni un bon résultat, mais il est faible avec quelques valeurs aberrantes. Quant au modèle SVR, nous notons qu'il a une très bonne capacité à capturer les valeurs aberrantes. Nous avons ensuite présenté le modèle combinaison de LinearRegression et PolynomialFeatures. Il ne capte pas la plupart des points et il est très faible en valeurs aberrantes, contrairement au modèle LSTM qui donne un excellent résultat et il capte la plupart des points.

Nous allons maintenant afficher ces modèles avec différentes échelles au fur et à mesure que nous les avons rassemblés dans un tableau. Nous notons que le modèle LSTM était le plus précis parmi les modèles, il a la valeur d'erreur la plus faible, MSE = 0,0324, puis le modèle SVR montre une bonne prédiction. avec un taux d'erreur de MSE = 0,0555. Le modèle de forêt aléatoire a également donné une petite valeur d'erreur avec MSE = 0,1404, tandis que le modèle composite pour la régression linéaire et les caractéristiques polynomiales est resté au dernier rang, la valeur d'erreur étant légèrement élevée avec MSE = 0,3638.

regression for Pole Adrar dataset per hour

	MAE	MSE	R2
RandomForestRegressor	0.244888	0.140466	0.859042
SVR	0.160874	0.055560	0.944245
Polynomial-linear	0.497608	0.363838	0.634887
LSTM	0.133424	0.032421	0.967466

Tableau 4.1 Résultats de prédiction des modèles de régression pour les données du pôle Adrar

2.2 Territoire National

Nous verrons dans cette section les résultats de la prédiction de la valeur de la consommation d'électricité pour les données nationales (Algérie) à l'aide d'un modèle d'apprentissage automatique de régression avec quatre modèles différents au moyen d'un nuage de points, qui est un graphique dans lequel les valeurs de deux variables sont tracées sur deux axes qui font partie des données de test et de leurs valeurs attendues :

2.2.1 Forêt aléatoire

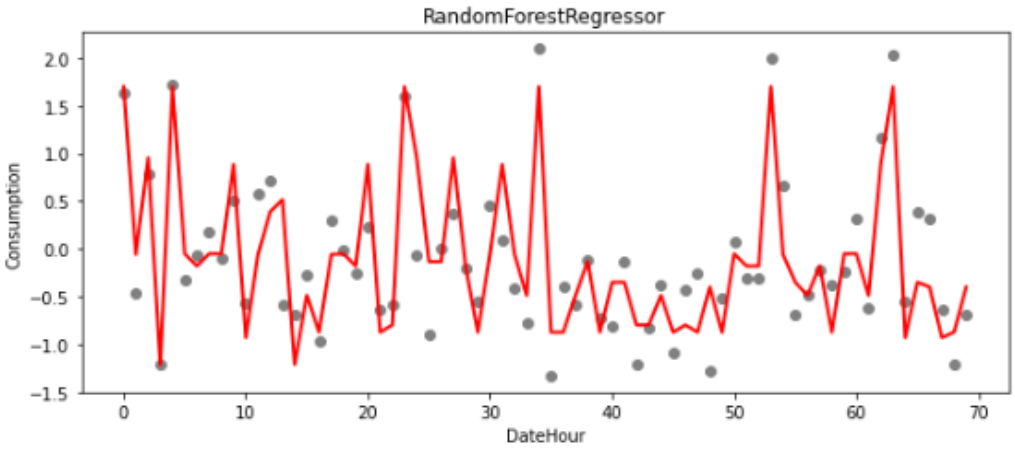

Figure 4.5 Prédiction de données nationales par modèle forêt aléatoire.

2.2.2 Régression à vecteurs de support

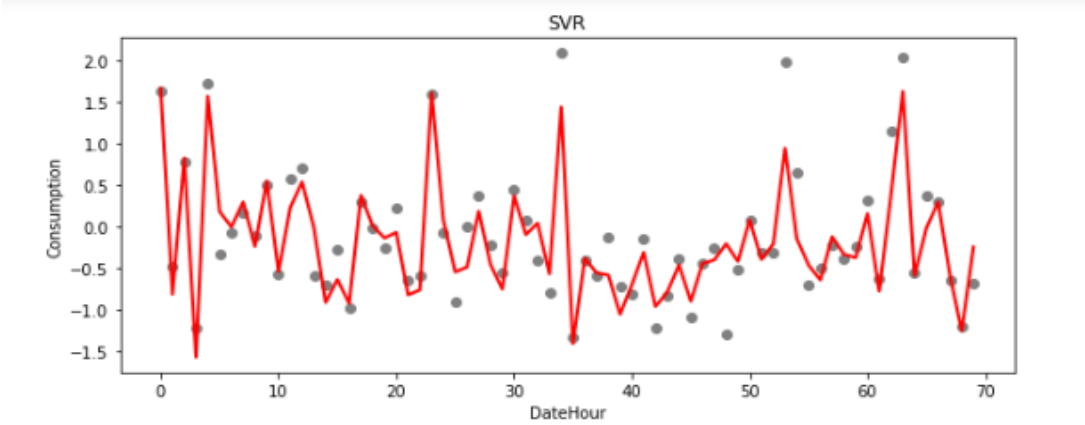

Figure 4.6 Prédiction des données nationales par modèle SVR.

2.2.3 Modèle de combinaison de LinearRegression et PolynomialFeatures

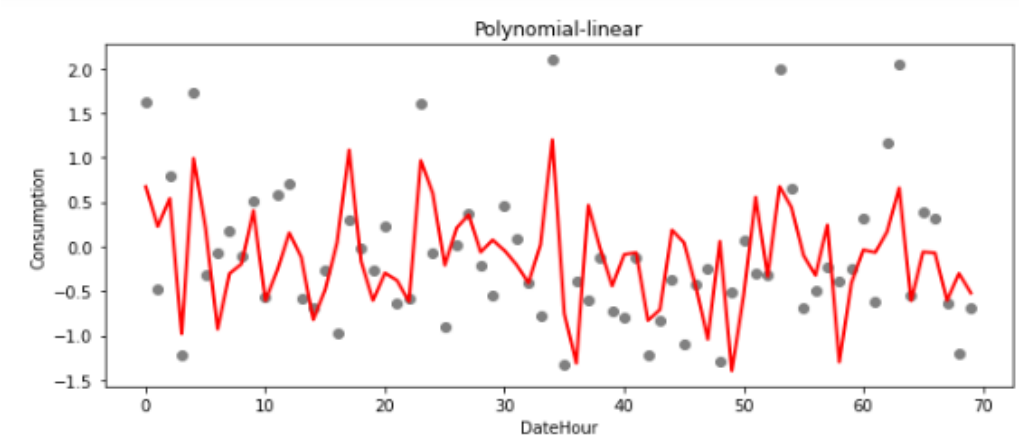

Figure 4.7 Prédiction des données nationales par LinearRegression et PolynomialFeatures

2.2.4 Modèle LSTM

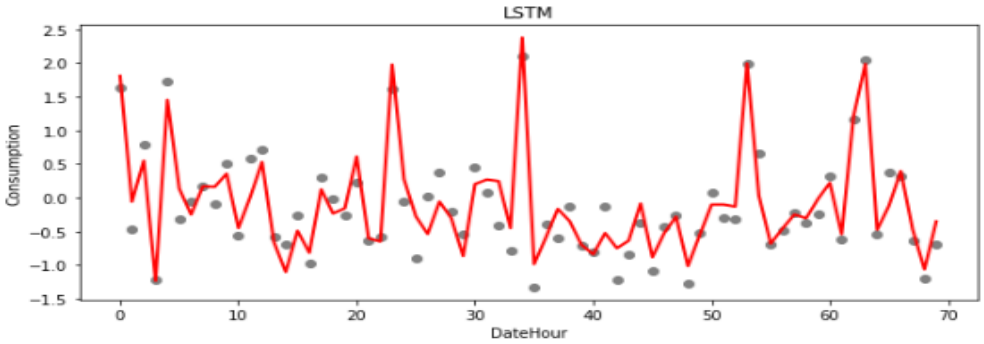

Figure 4.8 Prédiction des données nationales par modèle LSTM

2.2.5 Synthèse des résultats

En utilisant les résultats du nuage de points (voir tableau 4.2), nous ferons quelques observations. Nous avons d'abord commencé avec un modèle de forêt aléatoire. L'image montre qu'elle a fourni un bon résultat, mais qu'elle n'a pas bien fonctionné avec certaines valeurs aberrantes. Pour le modèle SVR, il a donné un bon résultat avec sa capacité à capturer certaines valeurs aberrantes, puis nous avons introduit une combinaison de régression linéaire et un modèle polynomial. Nous avons remarqué qu'il n'a pas capturé le plus de points et en bonté nous notons que le modèle LSTM a donné une valeur très élevée et a obtenu le plus de points.

Nous allons maintenant afficher ces modèles à différentes échelles en les rassemblant dans un tableau. Nous notons que le modèle LSTM dans cet exemple était également le plus précis parmi les modèles, et qu'il a la valeur d'erreur la plus faible de MSE = 0,1140, alors le modèle SVR montre une bonne prédiction. Taux d'erreur MSE = 0,2437. Le modèle de forêt aléatoire a également donné une petite valeur d'erreur avec MSE = 0,2819, tandis que le modèle composite pour la régression linéaire et les caractéristiques polynomiales est resté en dernière position mais n'a pas donné un résultat satisfaisant car il s'agissait d'une valeur d'erreur légèrement élevée avec MSE = 0,6124.

regression for National dataset per hour

	MAE	MSE	R2
RandomForestRegressor	0.386947	0.281985	0.718120
SVR	0.338469	0.243713	0.756377
Polynomial-linear	0.612835	0.612487	0.387739
LSTM	0.263255	0.114047	0.885995

Tableau 4.2 Résultats de prédiction des modèles de régression pour un données national.

2.3 Ensemble de données Synthétiques

Pour les données synthétiques que nous avons générées (voir chapitre 3, section 4.1), nous utiliserons deux types de prédiction, la régression, où l'on prédit une valeur, et la classification, où l'on prédit le type de la consommation. Pour ce dernier cas, nous avons transformé les données de l'ensemble d'apprentissage de la colonne consommation en valeurs catégoriques. Les valeurs de consommation ont été classées en trois catégories de consommation élevée, faible ou moyenne.

2.3.1 Régression par Forêt aléatoire

Figure 4.9 Prédiction de données synthétiques par modèle Forêt aléatoire

2.3.2 Régression par SVR

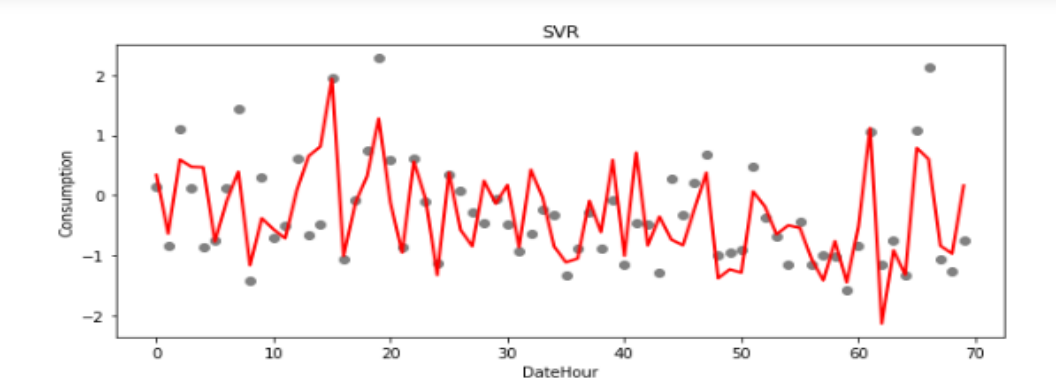

Figure 4.10 Prédiction de données synthétiques par modèle SVR

2.3.3 Modèle de combinaison de LinearRegression et PolynomialFeatures

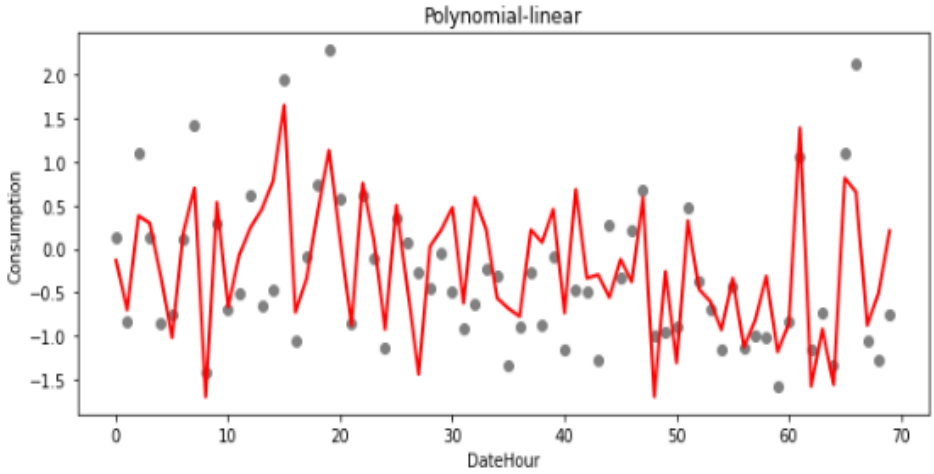

Figure 4.11 Prédiction de données synthétiques par LinearRegression et PolynomialFeatures

2.3.4 Modèle LSTM

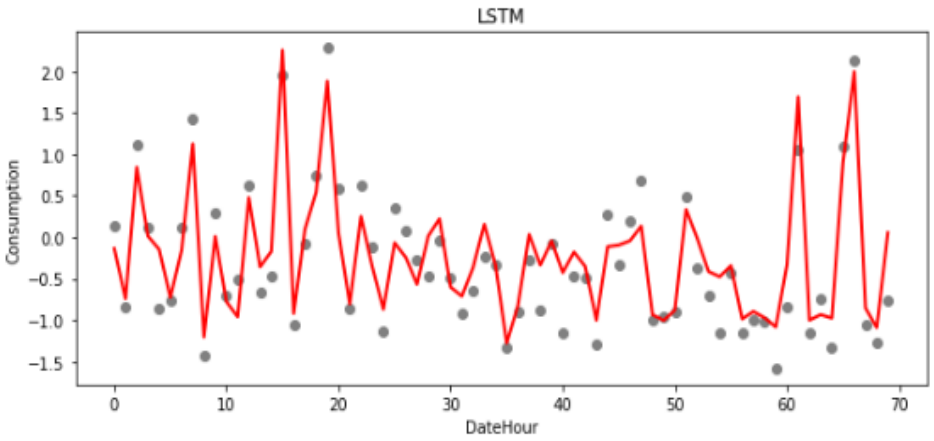

Figure 4.12 Prédiction de données synthétiques par modèle LSTM

2.3.5 Compaison des résultéts des modèles de Régression

En comparant les performances prédictives dans les diagrammes de dispersion pour tous les modèles (voir tableau 4.3), il a été constaté que le modèle de forêt aléatoire fournissait un bon résultat avec son incapacité à capturer les valeurs aberrantes par rapport au modèle SVR, il fournissait un résultat moyen avec sa capacité à capturer certaines valeurs aberrantes.

Nous avons présenté le modèle de combinaison de LinearRegression et PolynomialFeatures. On note qu'il n'a pas capté la plupart des points et qu'il est très faible avec des valeurs aberrantes. Quant au modèle LSTM, contrairement aux autres modèles, il a donné un excellent résultat et a pu prévoir la plupart des valeurs. Nous allons maintenant afficher ces modèles à différentes échelles en les rassemblant dans un tableau.

Nous notons que le modèle LSTM dans cet exemple était également le plus précis parmi les modèles, et qu'il a la valeur d'erreur la plus faible de MSE = 0,1354, alors le modèle de forêt aléatoire montre une bonne prédiction. Taux d'erreur MSE = 0,2710. Le modèle SVR a également donné une petite valeur d'erreur avec MSE = 0,3513 tandis que le modèle composite pour la régression linéaire et les caractéristiques polynomiales est resté en dernière position mais n'a pas donné un résultat satisfaisant car il s'agissait d'une valeur d'erreur légèrement élevée avec MSE = 0,408.

	MAE	MSE	R2
RandomForestRegressor	0.384676	0.271007	0.720373
SVR	0.438863	0.351339	0.637486
Polynomial-linear	0.494185	0.408882	0.578113
LSTM	0.280075	0.135453	0.860239

Tableau 4.3 Résultats de prédiction des modèles de régression sur les données synthétiques

2.3.6 Classification par arbre de décision

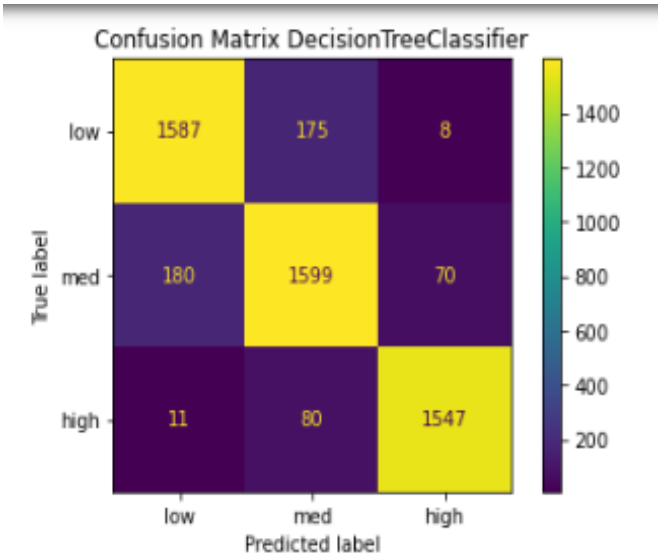

Figure 4.13 Matrice de confusion pour données synthétiques avec arbre de décision

2.3.7 Classification Bayésienne Naïve

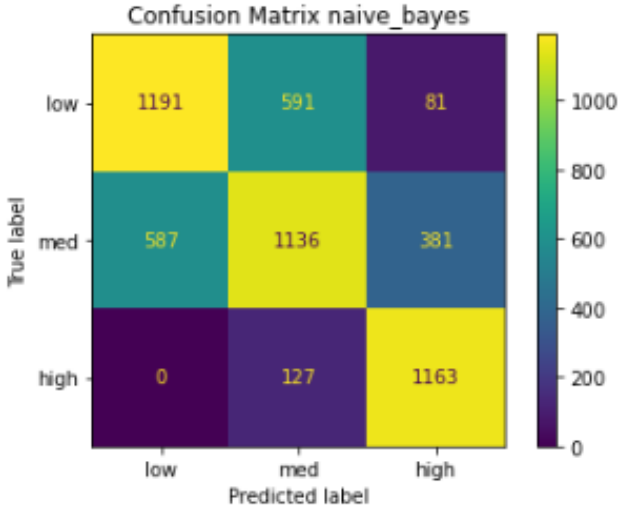

Figure 4.14 Matrice de confusion pour données synthétiques avec modèle Naïve Bayes

2.3.8 Classification par le modèle des k voisins les plus proches

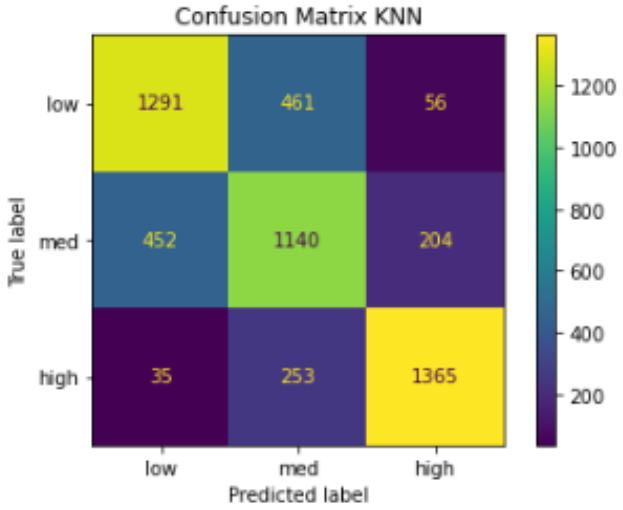

Figure 4.15 Matrice de confusion pour données synthétiques avec modèle KNN

2.3.9 Comparaison des résultats de classification

Dans cette section, nous comparons la précision entre les moèles : Arbre de décision, Naive Bayes, et KNN (voir tableau 4.4).Nous analyserons les résultats présentés dans la matrice de confusion.

Le modèle d'arbre de décision a montré un excellent résultat car les valeurs correctes ont été estimées high = 1547 sur 1625 valeurs et medium=1599 sur 1854 et low = 1587. Ainsi, les résultats ont montré dans le tableau le taux de réussite (Accuracy) du modèle 90.03%.

Quant au modèle KNN, il a également donné un bon résultat, et les valeurs correctes pour les valeurs hautes, moyennes et basses ont été estimées à 1365, 1140 et 1291, respectivement, où le taux de réussite du modèle était de 71,80%.

Enfin, nous avons le modèle Naive Bayes, et les valeurs correctes étaient pour haut à 1163, moyen avec 1136 et bas avec 1191, avec un taux de réussite de 66,38%.

Grâce à ces résultats, nous avons constaté que le modèle d'arbre de décision fonctionnait avec la meilleure performance, car il a devancé les deux autres modèles, puis le modèle KNN en deuxième position, ensuite vient le modèle Naïve Bayes en troisième position.

Classification données synthétiques par hour

	Accuracy
DecisionTreeClassifier	0.900323
KNN	0.722085
naive_bayes	0.663877

Tableau 4.4 Résultats de prédiction des modèles de classification sur les données synthétiques

3 Synthèse globale des résultats

La précision des prévisions varie légèrement selon les régions en raison de divers facteurs. La comparaison des résultats pour différents modèles d'apprentissage automatique peut être vue dans le graphique (voir Fig. 4.16).

On note qu'avec des données différentes, le modèle LSTM reste le plus fort pour résoudre ce problème. Quant au modèle SVR, il a fourni d'excellents résultats avec son avantage à supporter les valeurs aberrantes. Il a fait ses preuves dans les données du pôle Adrar et National, contrairement aux données synthétiques qui l'ont affecté par les caractéristiques distinctives qui contiennent des valeurs générées. Le modèle de forêt aléatoire avait une bonne précision avec de légères différences dans les résultats pour les trois ensembles de données.

Enfin nous avons le modèle de combinaison de LinearRegression et PolynomialFeatures qui était toujours en dernière position car les données non linéaires ont des résultats insatisfaisants, surtout avec les données aléatoires. En ce qui concerne les modèles de classification (voir Fig. 4.17), le modèle d'arbre de décision a été distingué comme le meilleur modèle par rapport au modèle KNN puis Naïve Bayes.

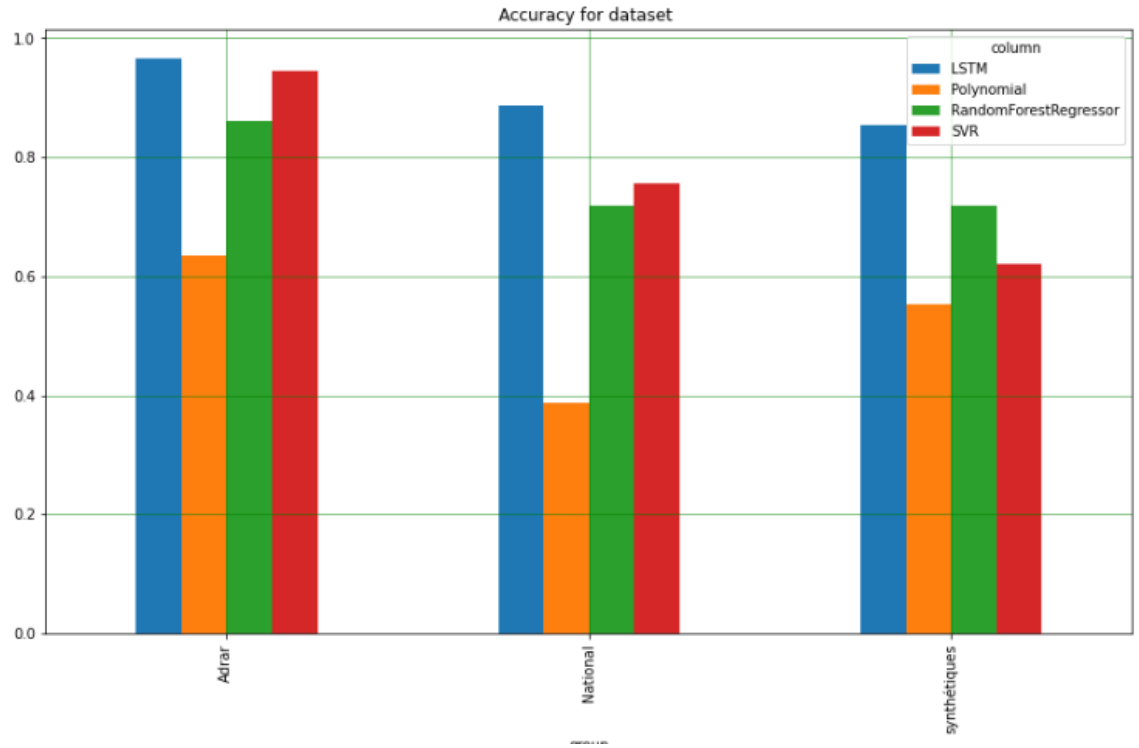

Figure 4.16 Globale résultats pour le régression

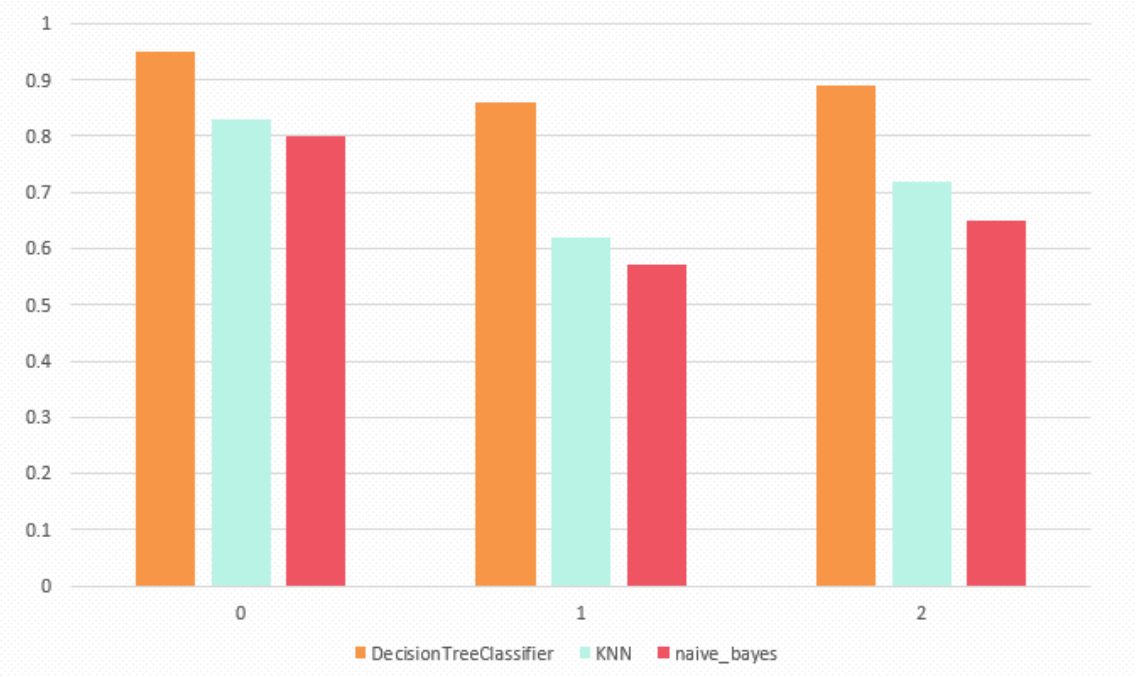

Figure 4.17 Globale résultats pour classification de données synthétiques

4 Conclusion

Dans ce chapitre, nous avons présenté une analyse comparative des différents modèles de prédiction de la consommation d'électricité sur trois ensemble de données, à savoir les données du pôle d'Adrar, les données du territoire national, et enfin les données synthétiques générées qui contiennent non seulement la date et la valeur de la consommation , mais également d'autres caractéristiques (features) qui ont une certaine influence sur la consommation en énergie électrique.

Deux types de modèles de prédiction ont été utilisés : les modèles de régression qui prévoient une valeur de la consommation, et les modèles de classification qui prévoient le type de consommation parmi trois catégories : consommation élevée, moyenne ou faible.

Les modèles de régression appliqués sont : SVR, LSTM, Forêt aléatoire et le modèle combinaison de LinearRegression et PolynomialFeatures. Les modèles de classification utilisées sont : Arbre de décision, KNN et Naïve Bayes.

Concernant la prédiction par régression, nous avons constaté que le modèle LSTM le meilleur parmi les modèles étudiés qui a donné des prédictions très proches des valeurs réelles (R^2=0.86 et MAE=0.13), vient ensuite le modèle SVR, qui a une capacité très élevée à donner des résultats proches de la réalité valeurs, ensuite le modèle de forêt aléatoire, qui a une bonne efficacité prédictive, et le dernier est le modèle PolynomialFeatures. Cela indique finalement qu'il s'agit du modèle le plus faible d'entre eux.

Pour les modèles de classification, le modèle d'arbre de décision a donné les meilleures performances et prédiction de la plupart des valeurs (Accuracy=90%), ensuite vient le modèle KNN (Accuracy=71%) et enfin le modèle Naïve Bayes en dernier lieu (Accuracy=66%).

CONCLUSION GÉNÉRALE

La demande d'électricité augmente parallèlement au développement rapide de la technologie et de l'urbanisation, ce qui générera un défaut dans l'incapacité de fournir de l'énergie électrique dans la quantité requise et entraînera donc une crise pour le producteur et le consommateur car l'électricité est un phénomène contrôlant tous les aspects de la vie. Par conséquent, la prévision précise de la demande d'énergie électrique est un enjeu très important pour les décideurs car elle permet une meilleure planification et gestion des ressources énergétiques.

Dans cette étude, nous avons discuté de la possibilité de prédictions de haute précision de la consommation d'électricité en appliquant certains modèles d'apprentissage automatique représentés par les modèles de régression pour prédire la valeur de consommation, tels que LSTM, forêt aléatoire, régression polynomiale linéaire et SVR. Nous avons également appliqué des modèles de classification pour prédire le type de la consommation d'électricité, où nous déterminerons si la consommation sera élevée, moyenne ou faible en utilisant l'arbre de décision, KNN et Naïf Bayes.

Tous les modèles de prédiction ont été appliqués sur trois types d'ensembles de données : Données issues du pôle Adrar, données issues du territoire national et enfin des données générées artificiellement.

Après comparaison des résultats des modèles de régression, nous avons constaté que le modèle basé sur les réseaux de neurones (LSTM) est nettement en avance par rapport au reste des modèles pour toutes les données, mais il est également possible de s'appuyer sur le modèle SVR pour prédire les valeurs futures de consommation d'électricité

En ce qui concerne les modèles de classification, l'arbre de décision a montré son efficacité en tant que modèle puissant pour prédire les classes de consommation.

Quant aux perspectives de recherche, nous pouvons dire que la prédiction de consommation en valeur peut être réalisée en utilisant LSTM ou SVR. Par contre, pour la prédiction par classification, il est judicieux d'étudier plus de caractéristiques (features) pouvant être des facteurs influents sur la consommation et de déterminer le meilleur modèle de classification sur des données réelles.

BIBLIOGREPHIE

[1] «mafahem,» 13 02 2021. [En ligne]. Available: https://mafahem.com/%D8%AA%D8%B9%D8%B1%D9%8A%D9%81-%D8%A7%D9%84%D8%B7%D8%A7%D9%82%D8%A9-%D8%A7%D9%84%D9%83%D9%87%D8%B1%D8%A8%D8%A7%D8%A6%D9%8A%D8%A9.

[2] «our world in data,» [En ligne]. Available: https://ourworldindata.org/energy-production-consumption.

[3] «wikipedia,» [En ligne]. Available: https://ar.wikipedia.org/wiki/%D8%A7%D8%B3%D8%AA%D9%87%D9%84%D8%A7%D9%83_%D8%A7%D9%84%D8%B7%D8%A7%D9%82%D8%A9?fbclid=IwAR3CkaCcSxhsawCebZdq_78N1HZKKlZCEvSsw-6WI220LUdCSzQd8mM9w60#cite_note-1. [Accès le 16 03 2022].

[4] *Direction Planification - Sonelgaz,* M'sila.

[5] «statista,» [En ligne]. Available: https://www.statista.com/statistics/280704/world-power-consumption/.

[6] A. Géron, Hands-On Machine Learning with Scikit-Learn.https://www.academia.edu/es/37865470/Hands_on_Machine_Learning_with_Scikit_Learn_and_Tensorflow, United States of America, 2017.

[7] scikit-learn user guide.https://scikit-learn.org/0.21/_downloads/scikit-learn-docs.pdf, Jul 29, 2019.

[8] F. Tyas Setiyorini, «COMPARISON OF LINEAR REGRESSIONS AND NEURAL NETWORKS FOR FORECASTING ELECTRICITY CONSUMPTION,» *PILAR Nusa Mandrir Vol,* n° %16, 2020.

[9] T. Parhizka, «Evaluation and improvement of energy consumption prediction,» *Journal of Cleaner Production,* n° %118, 2021.

[10] F. Pallonetto, «Forecast electricity demand in commercial building with machine learning,» *Energy and AI,* n° %113, 2022.

[11] C. Dianwei, «Research on electricity consumption forecasting model based on,» *Energy Reports,* n° %19, pp. 220-228, 2022.

[12] K. Nkrumah, «Optimization of industrial energy consumption for sustainability using,» *Sustainability Analytics and Modeling,* n° %115, 2022.

[13] «sonelgaz,» [En ligne]. Available: https://www.sonelgaz.dz/. [Accès le 2 Avril 2022].

ANNEXE

1 Business Planner

1.1 Segments de clientèle

Les clients auxquels le service s'adresse sont des marchés divers, et ils sont les suivants :

- Sociétés Sonelgaz (client industrial).
- Institutions (éducation comme les universités, les écoles et les instituts, santé comme les hôpitaux et les centres de santé, ... etc.).
- Entreprises (gouvernementales, comme la compagnie des eaux, et privées, comme une société de commercialisation).
- Les usines telles que les unités de fabrication, les usines et les moulins (parfumeries, ... etc.), et des centres de stockage tels que des entrepôts, des entrepôts frigorifiques et des unités de stockage de céréales.
- Les centres commerciaux comme Bureaux, hôtels...etc.

Le projet cible plusieurs états du pays.

1.2 Relation avec les clients

Construire la relation avec les clients à travers:

- Impliquer et contribuer aux clients dans le développement et la modernisation des services.
- Assurer la confidentialité de leurs données et ne pas être exploitées et divulguées.
- Service après-vente.
- Répondre à leurs questions.
- Rencontrez-les lors d'événements et de rassemblements dans le domaine de l'énergie.

1.3 Canaux de distribution

- Offrir une période d'essai pour évaluer le service.
- Participation à des forums et conférences dans le domaine de l'énergie.
- Communication directe avec les officiels.
- Les réseaux sociaux.
- Une plateforme numérique qui contient des numéros de téléphone et des e-mails.
- Réception du service via des rapports par e-mail.

- Paiement par comptes bancaires (ccp).

1.4 Valeur fournie

- Prévision de la consommation d'énergie électrique.
- Ajuster la production électrique.
- Eviter les coupures d'électricité aux heures de pointe.
- Réduire les dangers et les problèmes des groupes électrogènes inutiles, la consommation élevée de carburant, l'augmentation des coûts.
- Optimisation des coûts grâce à la détection précoce des défauts.
- Etude de préparation budgétaire.
- Améliorer l'efficacité énergétique et optimiser et rationaliser la consommation d'énergie (par une meilleure gestion de la production et une optimisation des ressources en plus de la compréhension du comportement des consommateurs).
- Facile à utiliser pour économiser la charge de travail en étudiant de nombreux fichiers et des mises en page approfondies.
- Modèle intelligent basé sur **Artificial intelligence & Data Mining**.
- Haute précision et excellente efficacité.
- Partenariats avec des institutions nationales (compagnie des eaux, spe, ose).
- Prix bas, tenant compte des normes de qualité Internationales

1.5 Activités principales

- La collecte des données.
- Préparation et nettoyage des données.
- Conception d'algorithmes.
- Développer et améliorer l'efficacité du modèle.
- Utilisez le modèle.
- Commercialisation.
- Bénéficiez de l'évaluation client.

1.6 Ressources principales

- L'ordinateur.
- Les données de la compagnie.
- Développeurs.
- Centre de données.

1.7 Partenaires principales

- Sociétés SPE.
- L'opérateur du système électrique -Sonelgaz.
- Etablissements.
- Des usines.
- Entreprises.
- Centres commerciaux.
- Le Fonds National de Soutien aux Startups.

1.8 Sources de revenus

- Vendre les services fournis par l'outil.
- Frais supplémentaires.
- Frais d'abonnement.

1.9 Frais

- Coûts directs: ordinateur, centre de données.
- Coûts fixes: factures d'électricité, Internet, lieu de travail, frais de bureau, salaires des employés.
- Coûts variables: nombre d'employés.

RÉSUMÉ

Une prévision précise de la consommation d'électricité permet de surmonter les problèmes auxquels sont confrontées les compagnies d'électricité. Le déroulement de cette étude a porté sur le développement d'un modèle prédictif utilisant le langage Python. Où un ensemble de données de Sonelgaz a été analysé pour la consommation d'électricité, par machine Learning. Nous avons comparé ces modèles en termes de performances de prédiction et avons constaté que le modèle DT avait la meilleure précision dans la méthode de classification. La régression a donné à LSTM le meilleur résultat dans toutes nos données. Il peut être invoqué dans le processus d'estimation de la quantité de consommation d'électricité.

Mots clés : énergie électricité ; consommation ; optimisation ; Sonelgaz ; prédiction ; classification ; régression ; Apprentissage automatique.

نبذة مختصرة

يلعب التنبؤ الدقيق لاستهلاك الكهرباء بالمساعدة على تخطي المشاكل التي تواجه شركات الكهرباء. ركز مسار هذه الدراسة على تطوير نموذج تنبؤي باستخدام لغة بايثون. حيث تم تحليل مجموعة من بيانات سونلغاز لاستهلاك الكهرباء، بواسطة التعلم الالي. قمنا بالمقارنة بين هذه النماذج من حيث أداء التنبؤ ووجدنا ان نموذج DT قد توصل الى أحسن دقة في طريقة التصنيف. والانحدار اعطى LSTM أفضل نتيجة في كل البيانات المتوفرة لدينا. ويمكن الاعتماد عليه في عملية تقدير كمية استهلاك الكهرباء.

الكلمات المفتاحية: الطاقة الكهربائية؛ استهلاك ;تحسين ; سونلغاز ؛ تنبؤ ؛ تصنيف؛ الانحدار ؛ التعلم الالي.

.ABSTRACT

Accurate forecasting of electricity consumption helps to overcome the problems facing electricity companies. The aim of this study focused on developing a predictive model using the Python language. Where a dataset from Sonelgaz data was analyzed for electricity consumption, by machine learning. We compared these models in terms of prediction performance and found that DT model had the best accuracy in the classification method. The regression gave LSTM the best result in all our data. It can be relied upon in the process of estimating the amount of electricity consumption.

Keywords: energy electricity; consumption ; optimisation ; Sonelgaz ; prediction ; classification ; regression ; Machine Learning.