Université de Gafsa

Institut Supérieur des Sciences Appliquées et de Technologies de Gafsa

Département Informatique et Télécommunication

Projet de fin d'étude en vue de l'obtention du diplôme de Licence en Ingénierie des Systèmes Informatiques

Implémentation d'un système de prédiction de pollution atmosphérique basé sur les algorithmes de machine learning

Présenté et soutenu par :

FADWA Touati

&

RAWEN Dhieb

Sous la Direction de :

M.TELLI Mounir : Encadreur ISSAT Gafsa

M.HIDRI Raouf : Encadreur industriel (CPG)

Soutenu le :05/06/2023

Devant le jury composé de :

Président **: M.Boukhari Kabil**

Rapporteur **: M. Rekik Ahmed**

Année Universitaire : 2022/2023

Dédicaces

Du fond de mon cœur je dédie ce travail

A mon cher père et ma chère mère,

Pour leur gentillesse et leur bienveillance, le grand amour qu'ils ressentaient pour moi, les sacrifices qu'ils ont faits pour moi pendant mes études.
Puissent les résultats de mes recherches prouver ce qu'ils attendaient de moi.

A mes chères sœurs,

Même s'ils ont enduré pour satisfaire toutes mes préoccupations.

J'apprécie leur soutien et la grande joie qu'ils ressentent quand je réussis.

A tous mes amis,

Parce que leurs noms dépassent la capacité qui peut être citée sur une page,

Pour leur amitié, leur fraternité et leur soutien continu.

Puissent-ils être remplis de bonheur, de joie et de réussite.

A

Tous ceux qui de près ou de loin ont contribué à la réalisation de cette humble tâche.

Dhieb Rawen

Dédicaces

Du fond de mon cœur je dédie ce travail

A mon cher père et ma chère mère,

Pour leur gentillesse et leur bienveillance, le grand amour qu'ils ressentaient pour moi, les sacrifices qu'ils ont faits pour moi pendant mes études.
Puissent les résultats de mes recherches prouver ce qu'ils attendaient de moi.

A mes chers frères,

Même s'ils ont enduré pour satisfaire toutes mes préoccupations.

J'apprécie leur soutien et la grande joie qu'ils ressentent quand je réussis.

A tous mes amis,

Parce que leurs noms dépassent la capacité qui peut être citée sur une page,

Pour leur amitié, leur fraternité et leur soutien continu.

Puissent-ils être remplis de bonheur, de joie et de réussite.

A

Tous ceux qui de près ou de loin ont contribué à la réalisation de cette humble tâche.

Touati Fadwa

Remerciements

Au terme de ce travail, Je tiens à exprimer ma profonde gratitude ainsi que mes sincères remerciements à mon encadreur monsieur **Telli Mounir**, pour leur encadrement, leur soutien et disponibilité. Leurs conseils, suggestions de lecture, commentaires, corrections et qualités scientifiques ont été très précieux pour mener à bien ce travail.

J'adresse mes sincères remerciements à monsieur **Hidri Raouf**, mon encadreur du stage au sein de compagnie de phosphate de Gafsa qui à toujours eu le temps pour m'écouter et mes conseiller durant ma période de stage et pour m'avoir accueilli au sein de l'entreprise.

Je tiens également à remercier et exprimer mon profond respect aux membres de jury d'avoir accepté de juger ce travail.

Enfin je remercie chaleureusement mes parents, mes frères, mes sœurs pour leur soutien et leur confiance tout au long de cette épreuve.

Merci

Sommaire

Introduction générale 1
Chapitre 1 : Etat de l'art 3
Introduction 4
1. Présentation de l'organisme d'accueil 4
2. Problématique du projet 6
3. Contexte Général de la pollution atmosphérique 7
3.1. Définition de la pollution atmosphérique 8
3.2. Les différentes échelles de la pollution atmosphérique 8
3.3. Les principaux polluants : Sources et impacts 9
4. Solution proposée 9
4.1. Choix de la solution 9
4.2. Intelligence Artificielle (IA) 10
4.3. Machine Learning 11
Conclusion 11
Chapitre 2 : Analyse et spécifications des besoins 12
Introduction 13
1. Gestion du projet 13
1.1. Les méthodes classiques de gestion du projet 13
1.2. Les méthodes Agiles de gestion du projet 14
1.3. Différence entre Approche Agile et Approche Séquentielle 16
1.4. Choix de méthodologie 17
2. Spécification des besoins 18
2.1. Besoins fonctionnels 18
2.2. Besoins non fonctionnels 18
Conclusion 19
Chapitre 3 : Méthodologie Utilisée 20
Introduction. 21
1. Intelligence artificielle 21
2. Machine Learning 21
2.1. Obtention des données et pré-processing 22
2.2. Extraction des caractéristiques 22
2.3. Réalisation du modèle 22

2.4. Phase d'apprentissage 23
2.5. Phase de validation 23
2.6. Performance du modèle 24
2.7. Types de modèle 24
3. Méthodes d'architecture et algorithmes d'apprentissage 24
3.1. KNN 24
3.2. Arbre de décision (Decision trees) 25
3.3. Régression logistique (LR) 26
3.4. Support Vector Machine (SVM) 27
3.5. Random Forest 27
Conclusion 28
Chapitre 4 : Expérimentations 29
Introduction 30
1. Réalisation logicielle 30
1.1. Environnement logiciel 30
a. Google Colab 30
b. Langage Python 31
1.2. Bibliothèques pour la préparation des données 31
a. NumPy 31
b. Pandas 32
c. Seaborn 33
2. Préparation de base d'apprentissage 34
2.1. Collecte des données 34
2.2. Prétraitement des données 35
3. Structure de modèle 35
4. Résultats 36
5. Problèmes rencontrés 37
6. Chronogramme 37
Conclusion 37
Conclusion et perspectives 38
Références bibliographiques 39
Résumé 40
Abstract 40

Liste des figures

Figure 1 : Logo de l'entreprise 4
Figure 2 : Les environnements de développements du projet 7
Figure 3 : Pollution atmosphérique 8
Figure 4 : Les phases de cycle en cascade 13
Figure 5 : Les phases du projet avec la méthode RAD 17
Figure 6 : Arbre de décision 26
Figure 7 : Régression logistique 26
Figure 8 : Support Vector Machine (SVM) 27
Figure 9 : Random Forest (RF) 28
Figure 10 : Logo Google Colab 30
Figure 11 : Logo Google Python 31
Figure 12 : Logo NumPy 32
Figure 13 : Logo PIL 33
Figure 14 : Logo Seaborn 33
Figure 15 : Organigramme du modèle 36

Liste des tableaux

Tableau 1 : Coordonnées de l'entreprise 5
Tableau 2 : Différence entre Approche Agile et Approche Séquentielle 16
Tableau 3 : Terminologie des colonnes de la base d'apprentissage 34
Tableau 4 : Résultats des tests 36
Tableau 5 : Chronogramme du travail 37

Introduction générale

Depuis la révolution industrielle, la pollution atmosphérique a considérablement augmenté et est aujourd'hui reconnue comme l'un des problèmes majeurs auxquels le monde entier est confronté.

Elle est le résultat d'une variété de facteurs, y compris les activités industrielles, de transport et d'aviation, l'augmentation de la consommation d'énergie, l'incinération des machines industrielles. Par conséquent, compte tenu de leurs effets aggravants sur l'environnement et la santé humaine, la détérioration de la qualité de l'air est devenue une préoccupation majeure. La majorité des grandes villes du monde sont les plus impactées par les émissions atmosphériques, dont les effets nocifs peuvent se faire sentir pendant de longues périodes sur toute la surface de la planète.

Pour éviter d'atteindre les niveaux de pollution les plus élevés, il est nécessaire de limiter au maximum les épisodes de pollution atmosphérique. Pour diminuer la gravité de ces épisodes, des mesures urgentes anticipant toutes les éventualités sont nécessaires. Les effets de telles mesures, comme la réduction du trafic routier et de l'activité industrielle et la fertilisation sont favorables à l'environnement et aux personnes plus sensibles aux problèmes de santé. De ce fait, afin d'assurer une meilleure performance, la mise en place de ces actions anti-pollution doit être anticipée.

Dans ce cadre se présente notre projet de fin d'études réalisé au sein du société « **CPG : Compagnie des phosphates de Gafsa** », il consiste à concevoir et réaliser un modèle de classification de la pollution atmosphérique en se basant sur la « machine learning ». Cela peut être considéré comme une des pas réalisé par cette entreprise dans son projet innovant pour réduire le niveau la pollution et pour cela, nous avons divisé notre travail en quatre chapitres :

Organisation du rapport

Ce projet s'articulera autour de quatre parties :

- Une première dans laquelle on présentera l'état de l'art et l'entreprise d'accueil de notre stage.
- Une deuxième qui traitera l'analyse et les spécifications des besoins.
- En troisième lieu on s'intéresse à la méthodologie utilisée et une description détaillée de la machine learning et les algorithmes choisis pour notre solution.

- Une quatrième partie qui sera consacré à l'expérimentation et test, ainsi que l'analyse des résultats.

Finalement on clôtura le rapport par 1a conclusion et les perspectives ainsi que des annexes nécessaires pour une meilleure compréhension du contenu.

Chapitre 1 : Etat de l’art

Introduction

Dans le premier chapitre, nous décrivons d'abord l'environnement du stage en présentant la société d'accueil. Ensuite, nous présenterons le contexte de notre projet, la problématique et une étude sur la pollution atmosphérique. Par la suite les principaux objectifs et les non-objectifs du projet en respectant notre cahier des charges, de même les contraintes de base.

1. Présentation de l'organisme d'accueil

Dans cette partie nous présentons l'organisme dans lequel nous avons effectué notre stage de projet de fin d'étude.

1.1. Présentation de la société

CPG est une entreprise tunisienne d'exploitation des phosphates basée à Gafsa. La CPG figure parmi les plus importants producteurs de phosphates, occupant la cinquième place mondiale. L'activité de l'entreprise se définit en 4 grands groupes : La préparation du terrain, extraction, production et la commercialisation des phosphates.

Cette entreprise cherche toujours trois axes :

- ✓ **L'innovation :** aspire toujours à pousser plus loin les idées, les procédés et les méthodes de travail.
- ✓ **L'excellence :** vise la qualité et l'excellence via une démarche efficace par processus.
- ✓ **La diversité :** la mise en place d'un processus d'agilité, une adaptation au changement et la collaboration.

Figure 1 : Logo de l'entreprise

1.2. Historique et coordonnées

Le tableau suivant résume l'historique de la CPG :

Tableau 1 : Coordonnées de l'entreprise

Fondation	1897
Capital	267.935 Mille Dinars
Collaborateurs	(à fin 12/2016) 6619
Type	Société multinational
Contact Web	http://www.cpg.com.tn/

C'était en avril 1885, lors d'une prospection dans la région de Metlaoui, partie occidentale du sud du pays, que Philippe THOMAS, géologue amateur français, a découvert des couches puissantes de phosphates de calcium sur le versant Nord de JEBEL THELJA. D'autres prospections géologiques et des explorations de grande envergure ont suivi cette découverte décisive. Celles-ci ont révélé l'existence d'importants gisements de phosphate au sud et au Nord de l'île de Kasserine.

A partir de 1896, date de création de la Compagnie de Phosphate et de Chemin de Fer de Gafsa, une nouvelle activité industrielle des phosphates a vu le jour dans le pays. Les premières excavations ont commencé dans la région de Metlaoui et vers 1900, la production de phosphate marchand a atteint un niveau de 200,000 tonnes.

Après ces débuts, la Compagnie de Phosphate et de Chemin de Fer de Gafsa a connu tout au long de sa longue histoire une série de changements structurels avant d'acquérir son statut actuel et de devenir en janvier 1976, la Compagnie des Phosphates de Gafsa - CPG.

Avec une expérience centenaire dans l'exploitation et la commercialisation des phosphates tunisiens, la CPG figure parmi les plus gros producteurs de phosphate dans le monde. Elle occupe le cinquième rang à l'échelle mondiale avec une production actuelle excédant 8 millions de tonnes de phosphate marchand (année 2007).

Les dates clés depuis la découverte du phosphate en Tunisie sont :

- 1885 : Découverte de gisements phosphatés par Philippe Thomas sur le versant Nord de Jbel Thelja près de Metlaoui.

- 1897 : Fondation de la Compagnie des Phosphates et de Chemins de Fer de Gafsa et mise en route du premier chantier d'extraction à Metlaoui. En même temps, démarrage de la construction de la ligne de chemins de fer reliant Metlaoui au port de Sfax.
- 1899 : Ouverture de la première mine souterraine (à Metaloui)
- 1905 : Fondation de la Société d'exploitation des Phosphates STEPHOS
- 1920 : Fondation de la Compagnie Tunisienne Des Phosphates De Jbel M'dhilla
- 1969 : Fusion de la Compagnie Tunisienne Des Phosphates De Jbel M'dhilla et la Compagnie Des Phosphates et De Chemins De Fer De Gafsa
- 1976 : La STEPHOS fusionne à son tour avec la Compagnie des Phosphates et de Chemins de Fer de Gafsa et, à partir de cette date, toutes les compagnies se regroupent pour former la Compagnie des Phosphates de Gafsa - CPG,
- 1978 : Entrée en exploitation de la première mine à ciel ouvert (Kef Schfaier) Création du Centre de Recherches
- 1996 : Fusion des deux structures commerciales de la CPG et du GCT
- 2006 : Fermeture de la dernière mine souterraine (Redeyef)

2. Problématique du projet

Pour aboutir à un système intelligent qui répond aux besoins le compagnie des phosphates de Gafsa, il est important de se focaliser en premier lieu sur les problématiques du projet pour pouvoir s'organiser.

Donc, on va déterminer le périmètre d'action et de faisabilité de ce projet. Comme il est illustré à la figure 2.

Le projet passe par plusieurs environnements, cela commence sur la machine du développeur, en passant par une phase de réalisation matérielle, vers l'environnement de test réel pour arriver finalement à l'environnent de production.

La migration et le déploiement du projet d'un environnement vers un autre ne se fait pas de la même façon et ne présente pas les mêmes degrés de complexité, puisque sur l'environnement de développement, une simple compilation suffira, mais le déploiement de la solution sur les autres environnements surtout la réalisation matérielle présente plusieurs contraintes et difficultés de plus si nous voulons transférer seulement un composant ou une fonctionnalité bien spécifique, sans toucher aux autres composants ce qui rend cette manipulation délicate et pénible d'où le besoin de concevoir une approche qui permet de réaliser cette tâche.

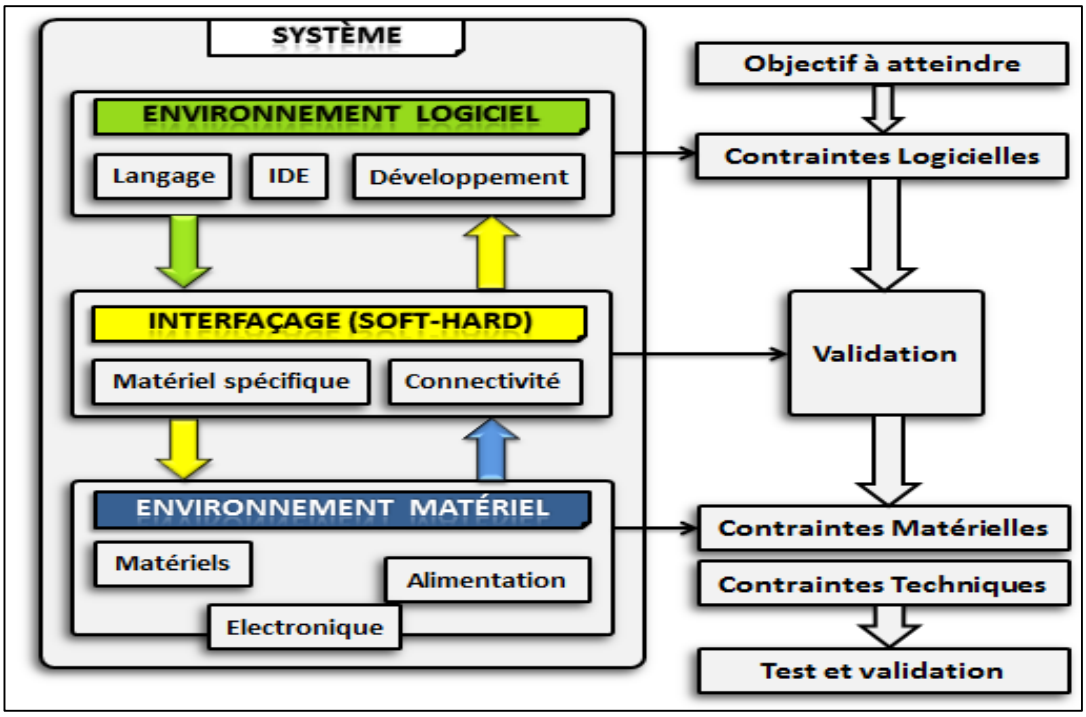

Figure 2 : Les environnements de développements du projet

De même le système à réaliser doit répondre aux besoins suivants :

- Les besoins spécifiques en termes de fonctionnalités.
- Une marge d'évolutivité assez grande.
- Prouver la possibilité de fusionner l'intelligence des cartes Arduino avec le milieu quotidien.

Les besoins spécifiques de ce projet se résume dont l'implémentation d'un modèle d'intelligence artificielle qui permet de prédire la pollution atmosphérique, ce dernier doit :

- Assurer des fonctions de surveillance de la pollution atmosphérique.
- Reconnaitre d'après les données d'entrée la prédiction dans le futur de niveau de la pollution pour que l'entreprise prend en charge un système de fertilisation ou tout autre système qui permet de réduire ce niveau de pollution.

3. Contexte Général de la pollution atmosphérique

L'atmosphère représente une fine couche gazeuse qui entoure notre planète. Cette couche mince joue un rôle primordial en tant que filtre du rayonnement solaire. L'atmosphère est composée de 78,09% d'azote, 20,95% d'oxygène, 0,93% d'argon ainsi qu'une variété de gaz en traces. Suite à différents mécanismes d'échauffement et de refroidissement, la température de l'atmosphère évolue en fonction de l'altitude.

3.1. Définition de la pollution atmosphérique

Selon la Loi sur l'air et l'utilisation rationnelle de l'énergie de 1996, la pollution atmosphérique est définie par la présence de substances (gazeux ou particules) dans l'atmosphère ayant des effets nocifs sur l'environnement et sur la santé humaine. Cette pollution peut être soit d'origine naturelle (volcanisme, érosion, embruns, océans, incendies et feux de forêt…), soit d'origine anthropique liée à des activités humaines (circulation automobile, processus industrielles, production d'énergie, combustion, incinération …). Depuis quelques années, différentes recherches ont montré un lien entre la présence de ces polluants dans l'atmosphère et la dégradation de l'environnement et de la santé humaine.

Figure 3 : Pollution atmosphérique

3.2. Les différentes échelles de la pollution atmosphérique

Le phénomène de la pollution atmosphérique relie trois composantes principales : les sources d'émission, le milieu (l'atmosphère) et les récepteurs (l'homme, l'animal, le végétal).

L'étude de ce phénomène est distinguée en trois niveaux spatiotemporels. Ces échelles dépendent du transport des espèces chimiques (de leur durée de vie) qui dépend de la stabilité des polluants.

- **Échelle locale** (10 m à 10 km des sources d'émission de polluants) : Les polluants atmosphériques proviennent des effets directs du chauffage individuel, des industries et du trafic automobile. Ces polluants affectent directement la santé humaine et les animaux, la végétation ainsi que les matériaux (Coman, 2008).
- ***Échelle régionale*** (environ 100 km des sources d'émission de polluants) : Dans ce cas, des phénomènes physico-chimiques variés et complexes interviennent. Cette échelle concerne les zones où on observe des phénomènes secondaires, tels que les pluies cides ayant un impact non négligeable sur les forêts, les écosystèmes aquatiques ou la production d'ozone dans les basses couches atmosphériques. Cette pollution dépend beaucoup des conditions météorologiques (Coman, 2008).

- **Échelle globale** (environ 1000 km) : A cette échelle, les études couvrent de très larges régions où les effets des polluants les plus stables chimiquement agissent sur l'ensemble de la planète : réduction de la couche d'ozone à haute altitude ou encore augmentation de l'effet de serre qui pourrait provoquer des changements climatiques importants (UNG, 2003).

3.3. Les principaux polluants : Sources et impacts

Un polluant atmosphérique est un corps d'origine anthropique ou naturel, émis dans l'atmosphère sous forme gazeux ou particulaire par diverses sources puis ce dernier est transporté et/ou transformé ce compartiment pour ensuite être éliminé par dépôts secs ou humide (PIOT, 2011). Les polluants atmosphériques sont classés en deux grandes catégories : Les polluants primaires qui sont ceux issus des sources d'émission telle que : les dioxydes de soufre (SO2), les oxydes d'azote (NOX), les particules en suspension (PS), le monoxyde de carbone (CO)…Les polluants secondaires, l'ozone et les particules secondaires dits photo-oxydants, qui se forment par réaction chimique ou photochimique à partir des polluants précurseurs.

4. Solution proposée

4.1. Choix de la solution

Dans le but de limiter l'exposition de la population aux polluants atmosphériques des recommandations de l'organisation mondiale pour la santé OMS existent. Les valeurs limites fixées par l'OMS sont des lignes directrices pour diriger les moyens de réductions de la pollution de l'air. Ces valeurs sont fixées à partir des études épidémiologiques et toxicologiques. Dans un pays comme la Tunisie où l'environnement est fragile et sensible aux rejets atmosphériques, il est important de définir des règles de bonne conduite permettant de préparer le pays et de le rendre capable de relever les défis environnementaux. Afin d'évaluer les effets des polluants atmosphériques sur la santé et l'environnement, il est nécessaire d'en connaître les concentrations dans l'air ambiant et de suivre leurs évolutions dans l'espace et dans le temps. Ces valeurs de concentrations sont comparées à des normes de référence de la qualité de l'air. Pour une série de polluants, des objectifs de qualité de l'air ont été déterminés par la réglementation tunisienne qui a fixé des valeurs limites et des valeurs seuils. ϖ Les valeurs limites : Ces valeurs ont été fixées afin d'éviter, de prévenir ou de réduire les effets nocifs des polluants sur la santé humaine oui sur l'environnement dans son ensemble. Ces valeurs sont à atteindre dans un délai donné et ne pas dépasser. ϖ Le seuil d'alerte : Est un niveau au-delà duquel une exposition même de courte durée présente un risque pour la santé de l'ensemble de

population et à partir duquel les pouvoirs publics doivent immédiatement prendre des mesures d'urgence.

En Tunisie, le décret du 18 mai 2018 fixe les valeurs limites et les seuils d'alerte des concentrations des polluants dans l'air ambiant, en vue de la protection de la santé et de l'environnement. (JORT, 2018) Pour les NO2, les valeurs limites en moyenne horaire sont fixées à 200 ug/m3 et en moyenne annuelles elles sont fixées à 40 ug/m3 (Pour des conditions de température et de pression respectivement 293 K et 101,3 kPa). Les seuils d'alertes sont définis de 400 ug/m3 en moyenne horaire. La valeur limite pour les SO2 est fixée à 350 ug/m3 maximum pour une moyenne horaire (à ne pas dépasser plus de 24 heures par an) et elle est fixée à 125 ug/m3 pour une moyenne journalière (avec 3 jours de dépassement autorisé par année). Les seuils d'alertes sont définis de 500 ug/m3 en moyenne horaire. Pour les PM10, les valeurs limites en moyenne journalière sur l'année sont définies à 50 ug/m3 et en moyenne annuelles elles sont définies à 40 ug/m3. Les seuils d'alertes sont définis de 150 ug/m3 en moyenne journalière.

Delà la CPG va chercher à respecter le maximum possible l'être humain avant tous et par la suite les décrets nationaux pour aboutir à un bon environnement de vie.

Examiner et protéger la qualité de l'air dans la région de CPG est devenu l'une des activités essentielles pour chaque être humain dans ses nombreuses zones industrielles et urbaines aujourd'hui. Les facteurs météorologiques et de trafic, la combustion de combustibles fossiles et les paramètres industriels jouent un rôle important dans la pollution de l'air. Avec ça augmentation de la pollution de l'air, nous devons mettre en œuvre des modèles qui enregistreront des informations sur concentrations de polluants atmosphériques. Le dépôt de ces gaz nocifs dans l'air affecte la qualité de vie des personnes en altérant leur santé, notamment en milieu urbain. Notre solution aide en contribuant pour une faible part à la diminution de la pollution par la régulation de ses niveaux.

4.2. Intelligence Artificielle (IA)

Les neurosciences et l'exploration du cerveau humain font régulièrement la une des journaux. Leurs progrès soulèvent une question pleine d'espoir et de peur : serait-il jamais possible de reproduire tout le cerveau humain ? Les ordinateurs lui sont déjà supérieurs aujourd'hui en matière de puissance de calcul, même si le cerveau humain à un degré de complexité beaucoup plus élevé : la marée sera-t-elle bientôt inversée ?

Ces doutes soulèvent la question de l'intelligence artificielle (en abrégé IA ou AI, pour intelligence artificielle en anglais). La recherche en intelligence artificielle consiste à traves

l'informatique, la neurologie et la psychologie, recréer les fonctions techniques du cerveau. L'approche de l'intelligence artificielle remet profondément en question notre conception de l'humanité et de ce que nous appelons l'intelligence.

L'intelligence artificielle avec sa propre volonté autonome est toujours le domaine de la fiction. Pourtant, les technologies visionnaires jouent un rôle vital dans de nombreux aspects de nos vies, mais nous ne le réalisons pas toujours. Beaucoup de gens ignorent ce qu'est réellement l'intelligence artificielle et son fonctionnement. Les médecins l'utilisent pour les diagnostics et pour planifier les traitements, les prévisions de marché sont plus efficaces avec l'IA et les algorithmes de recherche de Google sont également plus dynamiques. L'IA se trouve derrière chaque assistant tel que Cortana ou Siri, aide les voitures à être autonomes ou peut aider à sélectionner de nouveaux employés. Aux Etats-Unis, des lois sont déjà créées avec l'aide de l'intelligence artificielle. La recherche a fait de nombreux progrès ces dernières années en ce qui concerne les lotissements.

4.3. Machine Learning

L'apprentissage automatique communément appelé en anglais Machine learning, est un ensemble d'algorithmes d'apprentissage automatique qui tentent d'apprendre à plusieurs niveaux, correspondant à différents niveaux d'abstraction. Il a la capacité d'extraire des caractéristiques à partir des données brutes grâces aux multiples couches de traitement composé de multiples transformations linéaires et non linéaires et apprendre sur ces caractéristiques petites à petit à travers chaque couche avec une intervention humaine minime.

Conclusion

Dans ce chapitre, nous avons passé en revue les différentes notions nécessaires à la compréhension de notre projet en chiffrant la voile sur beaucoup des données qui vont être les bases de notre futur système et généralement celle de la pollution atmosphérique et de la machine learning.

Chapitre 2 : Analyse et spécifications des besoins

Introduction

Nous allons, dans ce chapitre, présenter la phase d'analyse et de spécification des besoins. En effet, nous commençons ce chapitre avec une description de la gestion de notre projet en citant les différents types des méthodes de gestion puis nous allons identifier et préciser les besoins à satisfaire. Ces besoins représentent les fonctionnalités à réaliser dans notre projet.

Le choix d'une méthode agile est évident et après une comparaison entre les principales méthodes agiles, nous allons choisir la méthodologie la plus adéquate pour réaliser ce projet.

1. Gestion du projet

Notre mission dans ce projet est très enrichissante mais aussi très complexe car ce dernier est amené à évoluer dans différents environnements qui sont en constantes évolutions. Nous devons donc maîtriser les techniques de gestion de projet et de comprendre les spécificités du projet.

L'objectif du est de pouvoir mener notre projet à terme, en respectant les délais et le budget alloué. Pour atteindre ces objectifs, il doit prendre en compte les différentes contraintes citées dans le chapitre précédant en utilisant une méthode de gestion de projet rigoureuse.

1.1. Les méthodes classiques de gestion du projet

Depuis toujours, les projets sont gérés avec la méthode dite « classique » qui se caractérise par recueillir les besoins, définir le produit, le développer et le tester avant de le livrer. On parle alors ici d'une approche prédictive « cycle en cascade ».

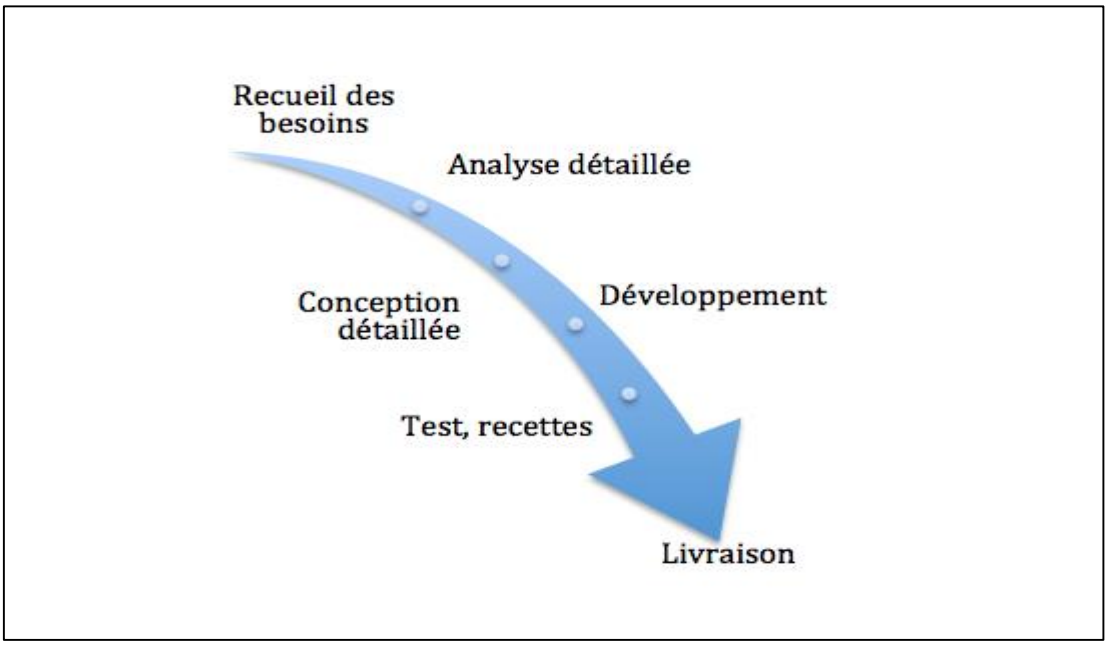

Figure 4 : Les phases de cycle en cascade

Comme son nom l'indique, il s'agit ici de prévoir des phases séquentielles où il faut valider l'étape précédente pour passer à la suivante. On doit alors s'engager sur un planning précis de réalisation du projet.

Il faut tout faire bien du premier coup car elle ne peut pas permettre de retours en arrière. Une décision ou un problème rencontré dans une phase peuvent remettre en cause partiellement ou totalement les phases précédentes validées.

Dans un cycle « en cascade » les risques sont détectés tardivement puisqu'il faut attendre la fin du développement pour effectuer la phase de test. Plus le projet avance, plus l'impact des risques augmente : il sera toujours plus difficile et coûteux de revenir en arrière lorsqu'on découvre une anomalie tardivement.

Afin d'anticiper au mieux ces risques il est nécessaire de produire des documents très détaillés en amont (recueil des besoins, cahier des charges…). Néanmoins, ces documents restent théoriques et conceptuels jusqu'à ce que le dispositif soit testé dans des conditions réelles.

1.2. Les méthodes Agiles de gestion du projet

Les méthodes agiles utilisent un principe de développement itératif qui consiste à découper le projet en plusieurs étapes qu'on appelle « itérations ». Ces itérations sont en fait des mini-projets définis en détaillant les différentes fonctionnalités qui seront développées en fonction de leur priorité.

Le but est d'assumer le fait que l'on ne peut pas tout connaître et anticiper quel que soit notre expérience. On découpe alors le projet en itérations plutôt que de tout prévoir et planifier en sachant que des imprévus arriveront en cours de route. On cite ici les principales méthodes Agiles. **(Tabaka, 2009)**

Les principales méthodes Agiles

- **Scrum :**

Scrum (qui signifie mêlée au rugby) est aujourd'hui la méthode agile la plus populaire. Elle se caractérise par itérations (appelées sprints) assez courts (maximum 1 mois) et un formalisme réduit : rôles (Product Owner, ScrumMaster, équipe), timeboxes (planification de release, planification de sprint, scrum quotidien, revue de sprint, introspection) et artéfacts (backlog de produit, plan de produit, plan de sprint, burdown/burnup de release, burdown/burnup de sprint)

- **EXtreme Programming (XP)**

L'objectif principal de cette méthode est de réduire les coûts du changement. Elle met l'accent sur la revue de code (faite en permanence par un binôme), sur les tests (ils sont faits systématiquement avant chaque développement), la conception continue, la simplicité, la traduction des besoins en métaphores.

- **Rational Unified Process (RUP)**

Cette méthode qui peut être considérée comme la moins agile des méthodes présentées ici, est un mélange des pratiques issues des méthodes traditionnelles et des méthodes agiles. Le principe est de parcourir un cycle de vie (inspection, élaboration, construction, transition) durant une itération. Chaque phase du cycle de vie est très précisément détaillée.
Son approche assez lourde et le coût d'investissement de cette méthode la réserve à des projets de grande ou moyenne taille.

- **Feature Driven Development (FDD)**

Moins connue que les 2 méthodes précédentes, FDD est essentiellement axé sur le design et le développement. Pour cela elle s'appuie sur une formalisation du modèle objet à l'aide de diagrammes UML, un découpage par fonctions qui seront développées par des petites équipes responsables d'une ou deux fonctions. Elle accorde un aspect très important à la qualité du produit fini, et s'aide d'outils pour suivre le déroulement du projet.

- **Rapid Application Development (RAD)**

C'est la méthode agile la plus ancienne et celle qui a été la première à être en rupture avec les méthodes traditionnelles. Elle a introduit les notions d'itération et d'incrément. Elle vise à adopter la solution la plus stratégique (en termes de délais), la moins risquée, la plus fiable et la moins coûteuse. Son cycle de développement est simple : cadrage, design, construction et finalisation dans le respect absolu d'une durée comprise entre 90 et 120 jours.

- **Dynamic systems development method (DSDM)**

DSDM est méthode agile développée en Angleterre au milieu des années 90. Elle reprend les principes déjà vus dans les autres méthodes (implication des utilisateurs, autonomie de l'équipe, visibilité et adéquation du résultat, développement itératif et incrémental, réversibilité des modifications, tests continus, coopération des acteurs).

1.3. Différence entre Approche Agile et Approche Séquentielle

Voici un tableau récapitulatif des différences entre les deux méthodes.

Tableau 2 : Différence entre Approche Agile et Approche Séquentielle

Thème	**Approche Séquentielle**	**Approche agile**
Cycle de vie	En cascade ou en V, sans rétroaction possible, phases séquentielles.	Itératif et incrémental.
Planification	Prédictive, caractérisée par des plans plus ou moins détaillés sur la base d'un périmètre et d'exigences définies au début du projet.	Adaptative avec plusieurs niveaux de planification avec ajustements si nécessaires au fil de l'eau en fonction des changements survenus.
Qualité	Contrôle qualité à la fin du cycle de développement. Le client découvre le produit fini.	Un contrôle qualité précoce et permanent, au niveau du produit et du processus. Le client visualise les résultats tôt et fréquemment.
Changement	Résistance voire opposition au changement. Processus lourds de gestion des changements acceptés.	Accueil favorable au changement inéluctable, intégré dans le processus.
Suivi de l'avancement	Mesure de la conformité aux plans initiaux. Analyse des écarts.	Un seul indicateur d'avancement : le nombre de fonctionnalités implémentées et le travail restant affaire.
Gestion des risques	Processus distinct, rigoureux, de gestion des risques.	Gestion des risques intégrée dans le processus global, avec responsabilisation de chacun dans l'identification et la résolution des risques. Pilotage par les risques.
Mesureur succès	Respect des engagements initiaux en termes de coûts, de budget et de niveau de qualité.	Satisfaction client par la livraison de valeur ajoutée.

Tant que nous connaissons mieux les différences majeures entre approches traditionnelles et approches agiles à travers la comparaison faite ci-dessus nous avons opté pour une approche agile afin de gérer notre projet.

1.4. Choix de méthodologie

Puisque nous avons choisie d'adopter une approche agile pour gérer notre projet nous allons maintenant, d'après l'étude précédente sur les principales méthodes Agiles, choisir de travailler avec la méthode **Rapid Application Development (RAD).**

La méthode RAD structure le cycle de vie du projet en 5 phases (dont 3 systématiques) :

- L'initialisation prépare l'organisation, puis détermine le périmètre et le plan de communication ;
- Le CADRAGE définit un espace d'objectifs, de solutions et de moyens ;
- Le DESIGN modélise la solution et valide sa cohérence systémique ;
- La CONSTRUCTION réalise en prototypage actif (validation permanente) ;
- La finalisation est réduite à un contrôle final de qualité en site pilote.

La figure suivante présente le mieux possible ces phases en précisant une estimation du temps nécessaire au bon déroulement.

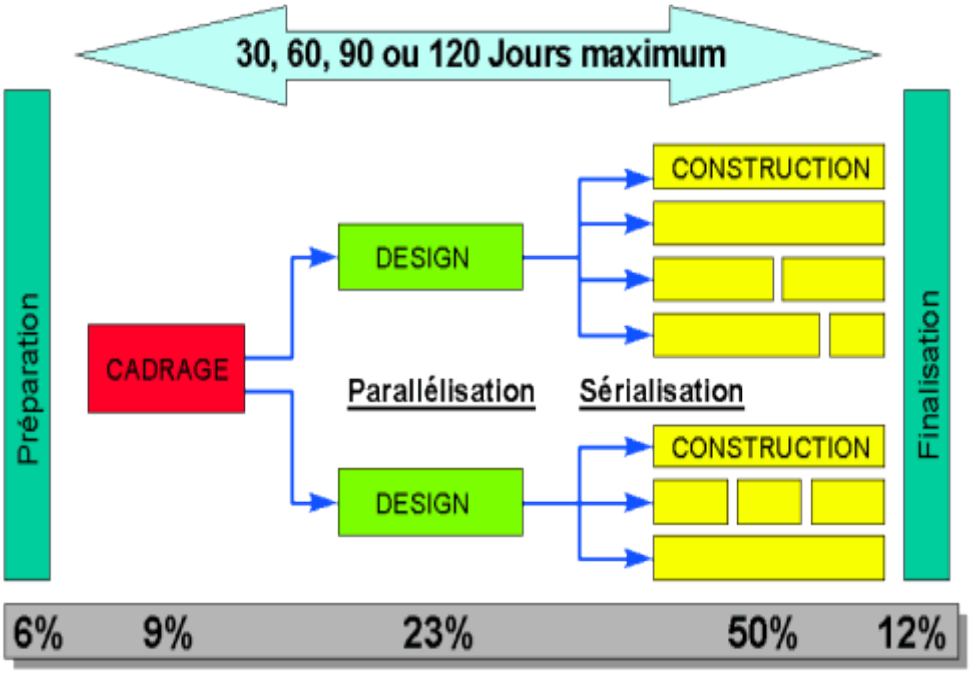

Figure 5 : Les phases du projet avec la méthode RAD

.

2. Spécification des besoins

Les besoins de notre futur système peuvent se diviser en deux grandes catégories qui sont les besoins fonctionnels qui englobent les cas d'utilisation principaux du système et les besoins non fonctionnels qui présentent les caractéristiques de base qui mènent au bon fonctionnement du système afin d'être acceptable par l'utilisateur.

2.1. Besoins fonctionnels

La phase de spécification des besoins fonctionnels est indispensable pour que les résultats de la réalisation de notre application soient conformes aux attentes du « Product owner ». Ainsi les différentes fonctionnalités que nous envisageons de mettre en place dans le cadre de ce projet, peuvent être regroupées dans les points suivants :

- Identification du jeu de données approprié pour notre énoncé de problème.
- Application des techniques de prétraitement sur Dataset.
- Étudier le jeu de données ainsi que différents types d'algorithmes.
- Implémentation d'une poignée d'algorithmes ci-dessus à notre ensemble de données.
- Analyser les inférences à partir des résultats de nos algorithmes et en tirer une conclusion.
- Visualisation des résultats de notre ensemble de données à l'aide d'un logiciel tiers.

2.2. Besoins non fonctionnels

Les besoins non fonctionnels peuvent être considérés comme des besoins fonctionnels spéciaux. Parfois, ils ne sont pas rattachés à un cas d'utilisation particulier, mais ils caractérisent tout le système (l'architecture, la sécurité, le temps de réponse, etc.) en vue de faciliter l'utilisation et améliorer les performances.

a. La performance :

Le temps de réponse du système doit présenter la rapidité afin de le mieux gérer.

b. Extensibilité

Le système doit être souple pour une extension future (ajouter des nouvelles fonctionnalités).

c. Utilisation et efficacité

L'interface utilisateur doit être simple et facile à comprendre pour que l'utilisateur puisse bénéficier des fonctionnalités du système. En fait, le système doit préserver une bonne qualité en termes de gestion d'erreur.

d. Ergonomie

L'application de contrôle doit respecter les standards de l'interfaçage Homme-Machine, en offrant à l'utilisateur une interface ergonomique et une bonne expérience d'utilisation.

L’apparence de cette interface est principalement caractérisée par des composants, des formes, des couleurs et la disposition des éléments.

e. Implémentation

Pour l’implémentation on doit suivre une méthodologie de développement orienté objet, permettant ainsi la maintenance du système d’une manière simple.

f. Contraintes techniques et matérielles

La partie technique et matérielle doit être adaptée aux besoins du projet et doit être totalement contrôlable et gérable via la partie logicielle et d’une façon transparente à l’utilisateur.

g. Maintenabilité

Les différents modules développés du système doivent être faciles à maintenir. Pour cela, le code doit être lisible et bien structuré. Nous devons respecter les standards de codage concernant par exemple les noms des attributs et des variables, les noms des méthodes ainsi que la disposition des commentaires.

Conclusion

Dans ce chapitre, nous avons détaillé les besoins fonctionnels du futur système ainsi que le choix de la méthode de travail en donnant les objectifs attendus de ce dernier.

Le chapitre suivant est consacré à la description de la méthodologie adopté en machine learning pour atteindre notre objectif.

Chapitre 3 : Méthodologie Utilisée

Introduction.

Ce chapitre sera consacré à une présentation générale de ces différentes approches ainsi que notre solution proposée en termes de modèle, sa structure et les différents composants.

1. Intelligence artificielle

L'intelligence artificielle (ou AI en anglais pour Artificiel Intelligence) consiste à mettre en œuvre un certain nombre de technologies visant à permettre à des machines de simuler une forme d'intelligence réelle.

De Google à Microsoft en passant par Apple, IBM ou Facebook, toutes les grandes entreprises du monde informatique travaillent aujourd'hui sur les problématiques de l'IA en essayant de l'appliquer à quelques domaines précis. Ainsi chacun a créé des réseaux de neurones artificiels constitués de serveurs et a permis de traiter des calculs lourds au sein de bases de données géantes.

2. Machine Learning

Le domaine du machine Learning inclut la construction d'un modèle à partir de données grâce à l'utilisation d'un algorithme. Ce modèle va au mieux généraliser, en représentant ou en approximant les données.

Il permet, selon les données qu'on lui donne en input, de prédire celles inconnues ainsi que de mieux comprendre celles existantes. Le domaine d'application du machine Learning est très varié : la prédiction de valeurs financières, la détection d'intrusion dans le domaine de la sécurité informatique, le moteur de recherche influençable par le pro de l'utilisateur, la détection de vols de machine, l'implémentation d'un anti-virus et la cryptanalyse. Le cycle de vie d'une implémentation du machine Learning est la suivante :

1. Obtention et nettoyage des données

2. Réalisation du modèle

3. Phase d'apprentissage

4. Phase de validation

5. Phase d'exécution

2.1. Obtention des données et pré-processing

La première étape à réaliser est donc l'obtention de données en suffisance, représentatives du problème à résoudre. Ceci n'est pas toujours aisé. Certaines informations sont plus coûteuses à obtenir que d'autres. Par exemple, un header d'un paquet réseau est plus simple à obtenir qu'une information dans la partie data quand celle-ci est chiffrée.

La deuxième étape est le nettoyage, appelé aussi pré-processing de la donnée récoltée, c'est-à-dire une réduction de ce qui est strictement intéressant, ainsi que leur traduction. Le but de cette étape est une meilleure précision du modèle, une optimisation de son temps d'exécution et de son apprentissage ainsi que de sa taille.

Voici quelques exemples de nettoyage :

- Transposer un ensemble de nombres vers un range.
- Transposer un ensemble de réels vers un ensemble de naturels
- Ajouter des valeurs qui ont été calculées à partir des données récoltées
- Sélectionner un résumé des informations
- Enlever les informations inutiles représentant du bruit

Ce nettoyage est souvent très compliqué à mettre en œuvre et demande une bonne connaissance des données à traiter. C'est pourquoi des techniques automatiques ont fait l'objet de recherches : ce sont les techniques de feature sélection.

2.2. Extraction des caractéristiques

Il existe essentiellement 3 types de méthodes de « ***feature sélection*** ». On retrouve d'abord la méthode de filtrage, qui ne se basent que sur l'utilité d'une variable sans tenir compte de son impact dans le modèle. Ensuite on a les méthodes qui tiennent compte de l'algorithme d'apprentissage pour déduire l'apport des variables. Enfin, on distingue aussi les méthodes qui sont spécifiques à un modèle et sont exécutées lors de la procédure d'apprentissage.

Ces méthodes peuvent être combinées pour obtenir de meilleurs résultats. Il est conseillé de ne pas utiliser les mêmes données pour les phases de sélection et d'évaluation, pour éviter un biais au niveau des performances estimées du système. Après avoir obtenu les données nettes du problème, la prochaine étape est la réalisation du modèle.

2.3. Réalisation du modèle

Elle consiste en une recherche de la meilleure structure ainsi que l'ensemble des paramètres à initialiser dedans. La complexité, et plus précisément la qualité, du modèle aura

une influence directe sur la précision de la généralisation des données. Plusieurs types de modèles vont être présentés dans la suite.

Après avoir construit le modèle, il est nécessaire de le configurer selon le problème à traiter grâce à la phase d'apprentissage.

2.4. Phase d'apprentissage

Un sous-ensemble de l'ensemble des informations nettoyées forme les données d'entrainement, permettant d'exécuter la phase d'apprentissage du modèle. Ceci permet d'ajuster les paramètres du modèle. Il existe plusieurs sortes d'algorithmes d'apprentissage. Certains de ces algorithmes sont supervisés et d'autres non supervisés.

Un algorithme supervisé est un algorithme à qui on présente l'entrée et la sortie (ou la cible) désirée en supposant qu'il y a une relation inconnue mais réelle entre les deux. Il devra minimiser l'erreur entre la sortie désirée et celle qu'il produit. Ils sont souvent utilisés pour des problèmes de reconnaissance. Un algorithme non supervisé est un algorithme à qui on présente l'entrée mais dont la sortie est inconnue. Ce type d'algorithmes est souvent utilisé pour des problèmes de partitionnement où le nombre et la nature des partitions ne sont pas connus a priori. Néanmoins, ce dernier ne donne aucun résultat si les données ne contiennent pas de partitions. Après avoir entraîné le modèle, il est important de le valider pour éviter le sur-apprentissage.

2.5. Phase de validation

Durant cette phase, on va tester et valider le modèle et ses paramètres selon des critères se basant sur ses résultats. Il permet d'obtenir le meilleur modèle généralisant les données obtenues lors de la phase d'apprentissage. Pour cela, on a un ensemble d'exemples pour l'apprentissage et un autre pour les tests. Voici quelques méthodes pour les tests :

- Hold-out : On coupe aléatoirement l'ensemble des informations en deux groupes : groupe d'apprentissage et groupe de tests.
- Leave-one-out : Cette méthode sort de l'ensemble des informations une donnée en particulier e la laisse de côté, puis construit le modèle avec celles restantes et enfin évalue la structure avec l'exemple laissé de côté. On répète le processus pour chacune des données de l'ensemble de données. Ainsi, on peut avoir une moyenne globale de la précision du modèle.
- Cross-validation : Cette méthode réalise un partitionnement des données de manière aléatoire en groupes. On utilise une partition comme un ensemble de test et le reste pour

former celui d'entraînement. Comme précédemment, on applique un algorithme à l'ensemble d'entraînement et on évalue le modèle résultant sur celui de tests. On répète ce processus pour chaque partition et on regarde l'erreur moyenne.

D'autres méthodes existent telles que les boots rap. Enfin, après avoir validé le modèle, il reste à quantifier ses performances en pratique.

2.6. Performance du modèle

Après avoir choisi et évalué notre modèle, il est intéressant de pouvoir quantifier ses performances. Pour cela, on compare plusieurs modèles sur un même jeu de données. En effet, certains sont plus adaptés pour certains problèmes. Lors de cette phase, il est primordial de ne pas enlever de données pour ne pas biaiser l'évaluation. Pour cette phase, on s'intéresse au pourcentage de vrais positifs, de vrais négatifs, de faux négatifs et enfin de faux positifs. Néanmoins, ça n'inclut pas la taille du modèle ni son temps d'apprentissage ou son temps d'exécution. La technique « receveur operating characteristic (ROC) » est largement utilisé et permet de tester une structure. Concrètement, un ROC est une courbe d'un modèle représentant ses vrais positifs par rapport à ses faux positifs. L'aire sous cette courbe représente la performance du modèle.

2.7. Types de modèle

Il existe plusieurs types de modèles. Outre le fait qu'ils sont différenciables par leur côté supervisé ou non supervisé, ils le sont aussi par leur côté classification ou régression. Le premier classifie les données et le deuxième prédit des valeurs pour chaque donnée.

3. Méthodes d'architecture et algorithmes d'apprentissage

3.1. KNN

L'algorithme K-Nearest Neighbors, également connu sous le nom de KNN ou k-NN, est un classificateur d'apprentissage supervisé non paramétrique, qui utilise la proximité pour faire classifications ou prédictions sur le regroupement d'un point de données individuel. Alors qu'il peut être utilisé pour les problèmes de régression ou de classification, il est généralement utilisé comme algorithme de classification, partir de l'hypothèse que des points similaires peuvent être trouvés à proximité les uns des autres.

Pour les problèmes de classification, une étiquette de classe est attribuée sur la base d'un vote majoritaire, c'est-à-dire l'étiquette qui est le plus fréquemment représenté autour d'un point de données donné est utilisé. Bien que cela soit techniquement Considéré comme « vote à la pluralité », le terme « vote à la majorité » est plus couramment utilisé dans la littérature. La

distinction entre ces terminologies est que le « vote à la majorité » nécessite techniquement une majorité de supérieur à 50 %, ce qui fonctionne principalement lorsqu'il n'y a que deux catégories. Lorsque vous avez plusieurs classes - par exemple, quatre catégories, vous n'avez pas nécessairement besoin de 50% des voix pour faire une conclusion sur une classe ; vous pouvez attribuer une étiquette de classe avec un vote supérieur à 25 %.

3.2. Arbre de décision (Decision trees)

L'arbre de décision est une technique d'apprentissage supervisé qui peut être utilisée à la fois pour problèmes de classification et de régression, mais la plupart du temps, il est préféré pour résoudre Problèmes de classement.

Dans un arbre de décision, il y a deux nœuds, qui sont le nœud de décision et la feuille Nœud. Les nœuds de décision sont utilisés pour prendre n'importe quelle décision et ont plusieurs branches, alors que les nœuds feuilles sont la sortie de ces décisions et ne contiennent aucun d'autres succursales.

Les étapes algorithmiques pour le regroupement d'arbres de décision :

1. Commencez l'arbre avec le nœud racine, dit S, qui contient la base de données.
2. Trouvez le meilleur attribut dans l'ensemble de données à l'aide de la mesure de sélection d'attributs (ASM)
3. Divisez le S en sous-ensembles qui contiennent des valeurs possibles pour les meilleurs attributs.
4. Générez le nœud de l'arbre de décision, qui contient le meilleur attribut.
5. Créer récursivement de nouveaux arbres de décision en utilisant les sous-ensembles de l'ensemble de données créé à l'étape -3
6. Continuez ce processus jusqu'à ce qu'un stade soit atteint où vous ne pouvez plus classer les nœuds et appeler le nœud final en tant que nœud feuille

Voici un exemple d’arbre de décision :

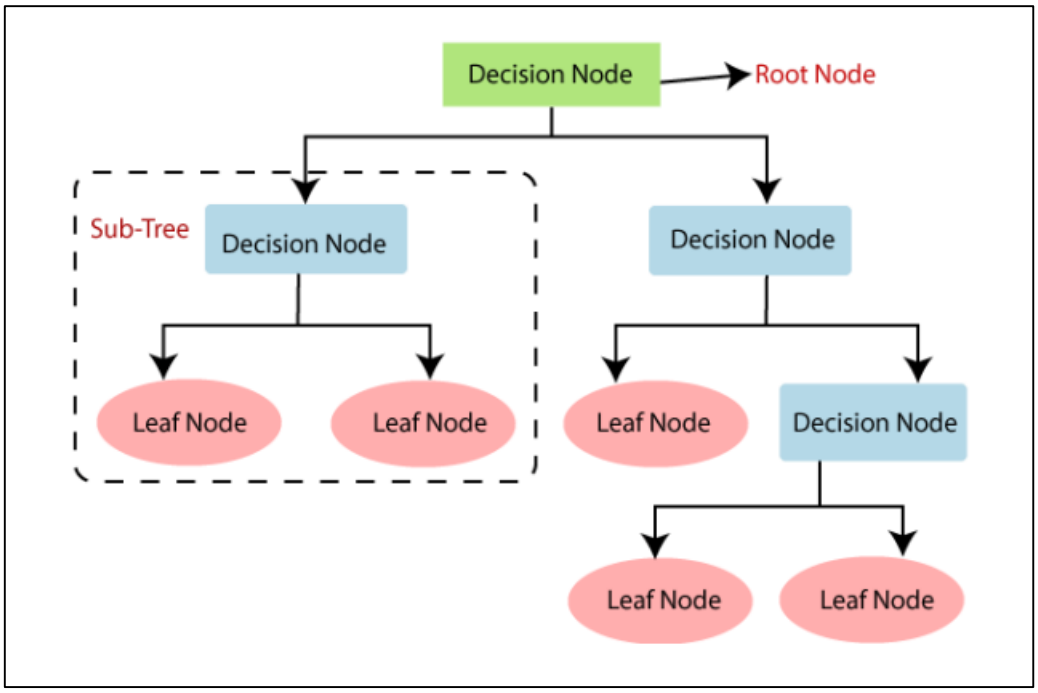

Figure 6 : Arbre de décision

3.3. Régression logistique (LR)

La régression logistique est un algorithme d'apprentissage automatique statistique populaire pour problèmes de classification, la prédiction de la sortie est effectuée à l'aide d'une fonction non linéaire telle que fonctions sigmoïdes et logit. Cet algorithme est appliqué sur la variable dépendante catégorielle. La figure suivante présente l'algorithme Logistic regression.

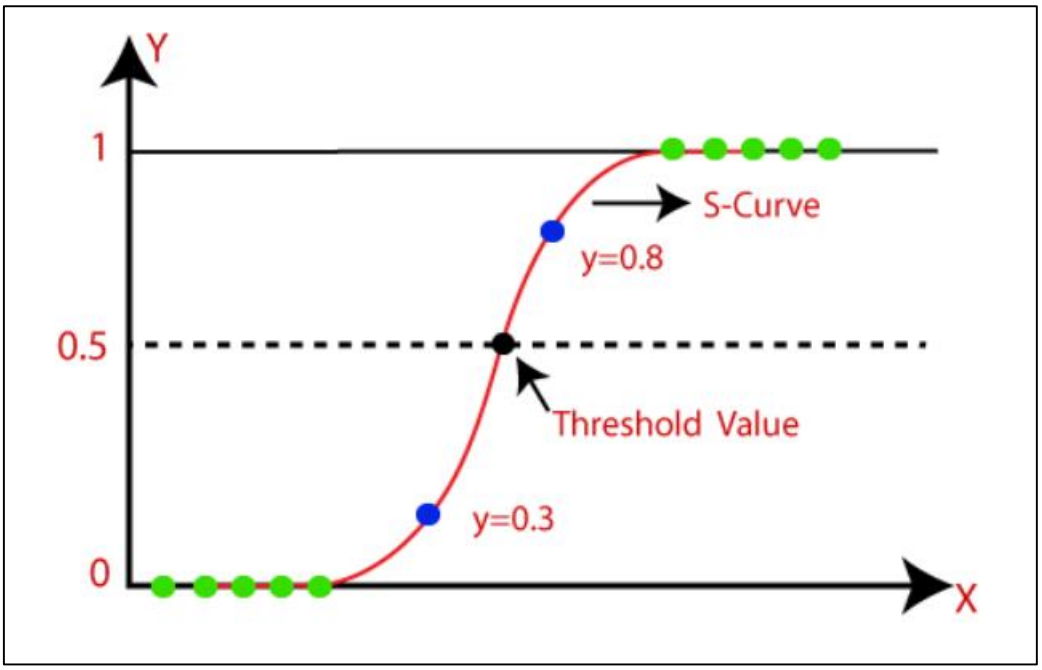

Figure 7 : Régression logistique

3.4. Support Vector Machine (SVM)

Support Vector Machine (SVM) est une machine bien connue algorithme d'apprentissage à des fins de classification et de prédiction. SVM étiquette chaque instance à un certain et une classe cible donnée, en en faisant un classificateur linéaire binaire non probabiliste. Le modèle va concentrez-vous sur les instances à la périphérie de chaque cluster et utilisez le point médian entre les clusters comme seuil, puis allouer chaque nouvelle instance en fonction de sa distance au seuil. La figure suivante présente le SVM :

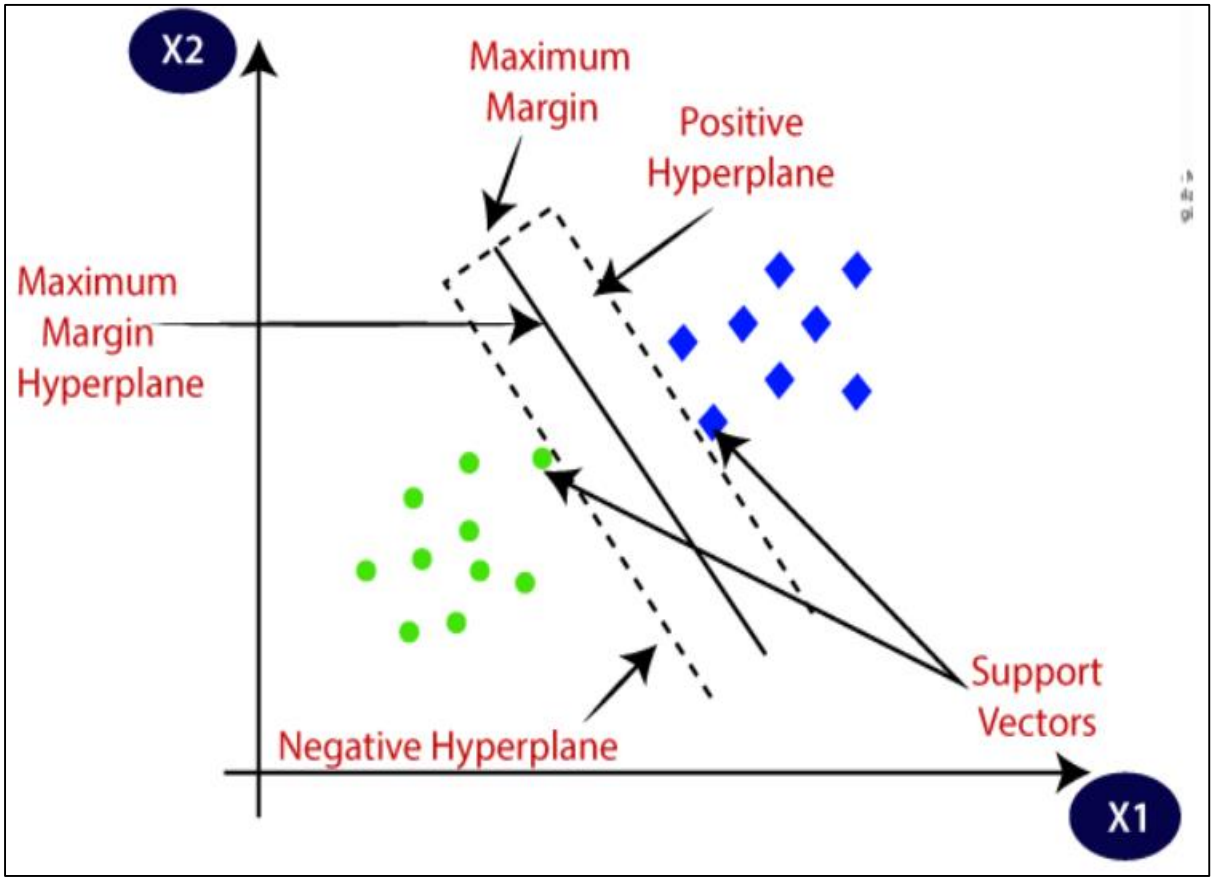

Figure 8 : Support Vector Machine (SVM)

3.5. Random Forest

Random Forest (RF) est l'un des algorithmes d'apprentissage automatique les plus populaires pour tâches de régression et de classification. RF crée un certain nombre d'arbres de décision appelés arbres forestiers pour améliorer le processus de prédiction et produire une plus grande précision. La construction de l'arborescence RF est similaire à la décision arbre (DT) en utilisant le gain d'information ou d'autres mesures. Puisque RF est un ensemble de DT; chaque arbre obtient un certaines sorties et RF choisiront la sortie majoritaire produite par les DT ou la

moyenne en cas de problème de régression. RF est utilisé sur DT en raison de sa capacité à gérer les éléments manquants et à résoudre les problèmes de surchauffe.

Ceci est sa structure :

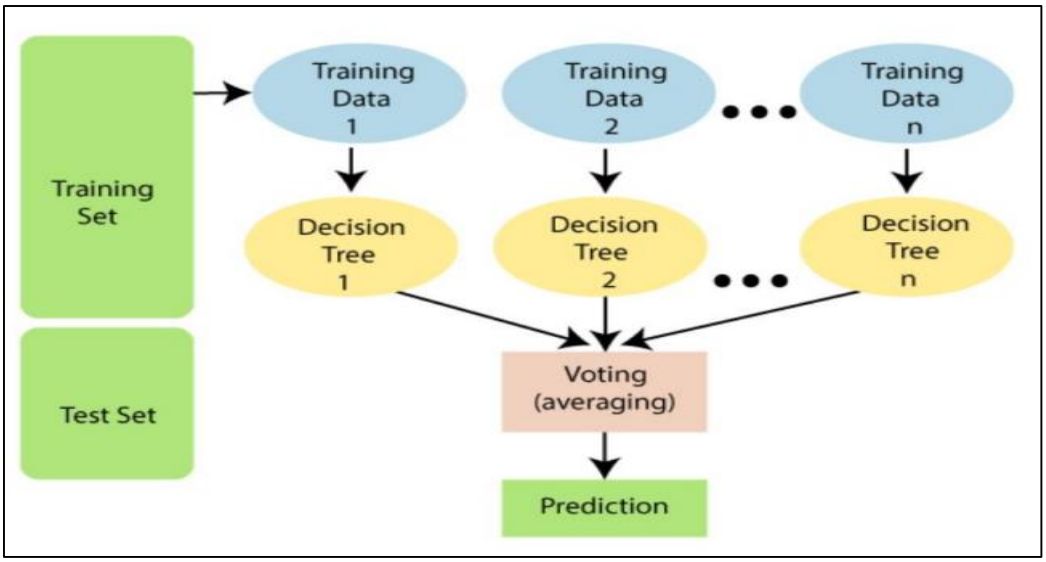

Figure 9 : Random Forest (RF)

Conclusion

Dans ce chapitre, on a déterminé les algorithmes à entreprendre pour la réalisation de notre solution. On a aussi détaillé les différentes astuces de ces derniers, en élaborant les images explicatifs. Dans le prochain chapitre, on passera à une description de la partie expérimentations du projet.

Chapitre 4 : Expérimentations

Introduction

Dans ce chapitre, on donne les détails de réalisation de notre système de prédiction de pollution atmosphérique, qui est la phase la plus essentielle pour profiter mieux qu'on puisse, de l'analyse et les méthodes proposées dans ce projet, en donnant les résultats obtenus.

La réalisation de ce projet se fait par la préparation de base d'apprentissage, l'application nos modèles K-NN, Arbre de décision et Régression linéaire ainsi que le SVM et Random Forest de l'apprentissage automatique puis un test et évaluation pour clôturer le chapitre.

1. Réalisation logicielle

1.1. Environnement logiciel

a. Google Colab

Google Colaboratory, également connu sous le nom de "Google Colab", est un environnement de développement de machine learning basé sur le cloud et gratuit. Il permet aux utilisateurs d'écrire et d'exécuter du code Python dans un environnement Jupyter Notebook, sans avoir à installer de logiciel sur leur ordinateur local. Colab utilise des machines virtuelles hébergées par Google qui offrent un accès gratuit à des ressources de calcul, notamment des processeurs, des GPU et des TPU, pour des tâches de machine learning et d'apprentissage profond. Les utilisateurs peuvent également partager leurs notebooks Colab avec d'autres utilisateurs pour la collaboration en temps réel.

Figure 10 : Logo Google Colab

b. Langage Python

Python est un langage de programmation interprété, conçu pour être facile à lire et à écrire. Il est utilisé pour développer des applications dans une grande variété de domaines, tels que la science des données, l'apprentissage automatique, la visualisation de données, la création de sites Web, l'automatisation de tâches, etc.

Python est maintenant l'un des langages de programmation les plus populaires au monde, grâce à sa simplicité et sa polyvalence. Il est apprécié pour sa syntaxe concise et claire, qui facilite la lecture et la maintenance du code, ainsi que pour sa grande variété de bibliothèques et de frameworks disponibles pour faciliter le développement de projets.

Python est un langage interprété, ce qui signifie que le code source est exécuté directement par l'interpréteur Python, sans nécessiter de compilation préalable. Cela rend le processus de développement plus rapide et plus souple, car les développeurs peuvent tester leur code en temps réel et effectuer des modifications rapidement.

Python est également connu pour sa communauté active et engagée, qui contribue à la création et à la maintenance de nombreuses bibliothèques et frameworks open source, tels que NumPy, Pandas, Scikit-learn, TensorFlow, PyTorch et bien d'autres. Ces bibliothèques ont grandement contribué à faire de Python un langage de choix pour le développement d'applications de science des données et d'apprentissage automatique.

Figure 11 : Logo Google Python

1.2.Bibliothèques pour la préparation des données

a. NumPy

NumPy est une bibliothèque open-source de calcul numérique pour Python. Elle fournit des objets de tableaux multidimensionnels hautement performants et des fonctions pour travailler avec ces tableaux. Les tableaux NumPy permettent de représenter et de manipuler facilement

des données numériques telles que des images, des sons et des données scientifiques. Les tableaux NumPy sont souvent utilisés en conjonction avec des bibliothèques de visualisation de données telles que Matplotlib et des bibliothèques d'apprentissage automatique telles que TensorFlow et PyTorch.

NumPy est largement utilisé dans le domaine du calcul scientifique et de l'apprentissage automatique en raison de sa rapidité, de sa facilité d'utilisation et de sa capacité à effectuer des calculs sur des tableaux de grandes dimensions.

Figure 12 : Logo NumPy

b. Pandas

Pandas est une bibliothèque open-source permettant la manipulation et l'analyse de données de manière simple et intuitive en Python. Elle a été développée par Wes McKinney en 2008 alors qu'il travaillait chez AQR Capital Management. À la fin de l'année 2009, elle a été mise en open source et est aujourd'hui activement utilisée dans le domaine de la Big data et de la data science car celle-ci offre des performances et une productivité élevée à ces utilisateurs.
L'une des forces de Panda est qu'il se base sur la très populaire bibliothèque NumPy. Elle fournit diverses structures de données et opérations pour le traitement de données numériques et de séries chronologiques.

Figure 13 : Logo PIL

c. Seaborn

Seaborn est une bibliothèque qui offre la possibilité de résumer et de visualiser des données. Elle permet de créer de jolis graphiques statistiques en Python. Cette bibliothèque apporte des fonctionnalités inédites qui favorisent l'exploration et la compréhension des données. Son interface utilise de fonctions intuitives qui assurent notamment la cartographie sémantique et aident à la conversion des données en graphiques statistiques à visualiser.

Il convient de voir Seaborn comme un complément de la bibliothèque principale de visualisation de données en Python et non comme un remplaçant. Et pour cause, bien souvent, Matplotlib est toujours utilisé pour des graphiques simples et un certain niveau de connaissances de cette bibliothèque est nécessaire pour modifier des tracés effectués avec Seaborn

Figure 14 : Logo Seaborn

2. Préparation de base d'apprentissage

2.1. Collecte des données

Dans le système proposé, l'ensemble de données sur la qualité de l'air est téléchargé, qui est disponible au format CSV.

Le format de données de valeurs séparées par des virgules peut facilement être traité et analysé rapidement à l'aide d'un ordinateur et les données utilisées à diverses fins. Il est importé dans le projet à l'aide d'un package ***pandas*** disponible dans ***google colab*** et le logiciel ***Jupyter***.

Les données contiennent 10 attributs importants qui aident à la prévision de la qualité de l'air. Initialement, ces données sont prétraités avec techniques appropriées pour supprimer les données valorisées incohérentes et manquantes, et les fonctionnalités nécessaires de l'ensemble de données sont sélectionnés pour de meilleurs résultats.

L'ensemble de données sur la qualité de l'air pour ce projet est collecté à partir du référentiel UCI. Le jeu de données est disponible au format CSV. Il est téléchargé et importé dans le projet en mentionnant l'emplacement des données téléchargés à l'aide du package pandas disponible dans Colab. Le l'ensemble de données contient des données de réponses horaires moyennes de différents éléments dans l'air.

Elle est sur ce lien https://www.kaggle.com/datasets/shrutibhargava94/india-air-quality-data.

Le tableau suivant résume les colonnes de la base d'apprentissage :

Tableau 3 : Terminologie des colonnes de la base d'apprentissage

Colonne	Description de la colonne
STN_CODE	Code de série pour l'emplacement
Simpling Date	Date
State	Nom de l'État
City	Nom de la ville
Location	Le type de zone où les données sur la pollution ont été enregistrées
{Pollutant} SO_2	Valeur de ce polluant (SO_2)
{Pollutant} NO_2	Valeur de ce polluant (NO_2)
{Pollutant} RSPM	Valeur de ce polluant (RSPM)
{Pollutant} SPM	Valeur de ce polluant (SPM)
Date	Date

Dans le Dataset les polluants : SO_2, NO_2,RSPM,SPM (PM est une matière particulaire) ont été mentionné de diverses régions d'États du sous-continent indien. À ce sujet, le prétraitement a été fait pour préparer les données pour l'analyse.

2.2. Prétraitement des données

Le prétraitement des données peut faire référence à la manipulation ou à la suppression de données avant qu'elles ne soient utilisées afin de assurer ou améliorer les performances, et constitue une étape importante dans le processus d'exploration de données.

La phrase "garbage in, garbage out" s'applique particulièrement à l'exploration de données et à l'apprentissage automatique projets. Les méthodes de collecte de données sont souvent mal contrôlées, ce qui entraîne des valeurs hors plage (par exemple, Revenu : −100), des combinaisons de données impossibles (par exemple, Sexe : Homme, Enceinte : Oui) et valeurs manquantes, etc. L'analyse de données qui n'ont pas été soigneusement examinées pour de tels problèmes peut produisent des résultats trompeurs. Ainsi, la représentation et la qualité des données sont avant tout avant de lancer toute analyse. Souvent, le prétraitement des données est la phase la plus importante d'un projet d'apprentissage.

S'il y a beaucoup d'informations non pertinentes et redondantes présentes ou des données bruyantes et peu fiables, alors la découverte des connaissances pendant la phase de formation est plus difficile. La préparation et filtrage des données peuvent prendre un temps de traitement considérable. Exemples de prétraitement des données inclure le nettoyage, la sélection d'instance, la normalisation, un encodage à chaud, la transformation, la fonctionnalité extraction et sélection, etc. Le produit du prétraitement des données est l'ensemble d'apprentissage final.

Le prétraitement des données peut affecter la manière dont les résultats du traitement final des données peuvent être interprété.

3. Structure de modèle

La figure ci-dessous englobe la structure de notre modèle : La partie gauche de l'organigramme montre le fonctionnement de l'apprentissage automatique non paramétrique, c'est-à-dire, K-Nearest Neighbors (KNN). Alors que la partie droite du schéma du modèle représente le fonctionnement d'algorithmes d'apprentissage automatique paramétriques utilisés dans l'analyse de la qualité de l'air, à savoir, Arbre de Décision, Support Vector Machine (SVM), Random Forest et Régression Logistique.

Tous ces algorithmes sont faits pour fonctionner ensemble et leurs précisions dans l'analyse de la qualité de l'air a été bien mémorisées.

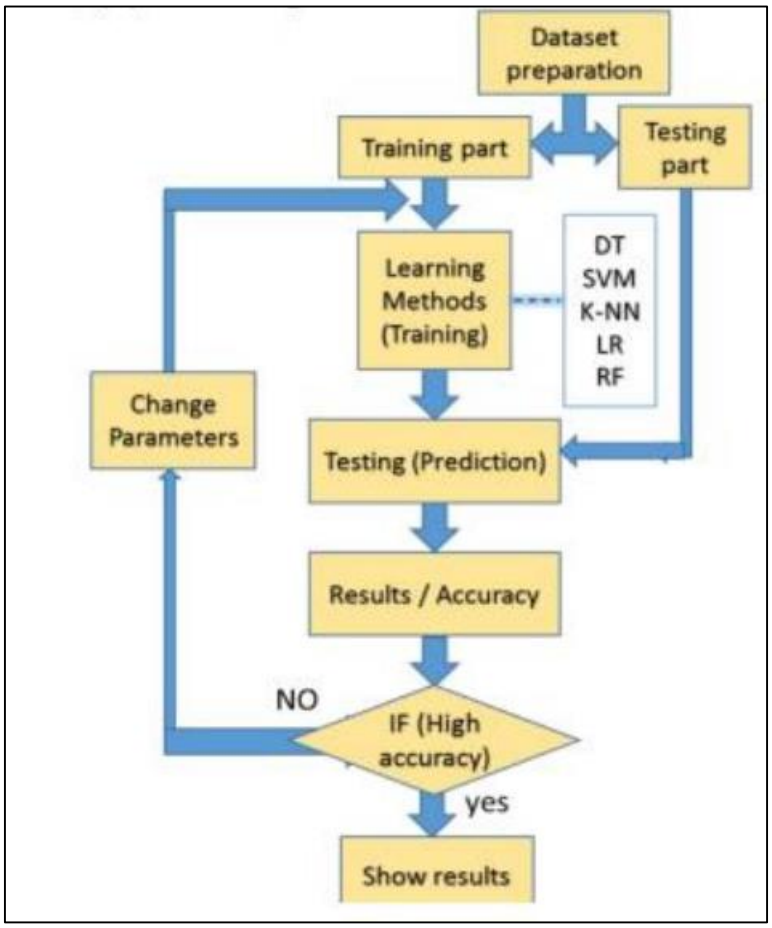

Figure 15 : Organigramme du modèle

4. Résultats

Le tableau suivant donne les résultats obtenues :

Tableau 4 : Résultats des tests

Algorithme	**Accuracy**
Decision Tree	99.97%
Support Vector machine (SVM)	99.72%
K-Nearest Neighbours (KNN)	99.67%
Random Forest	99.98%
Logistic Regression	82.71%

Les tests et la validation ont été effectués sur divers algorithmes d'apprentissage automatique, comme mentionné ci-dessus. Le algorithme avec la plus grande précision de 99,98 %, Random Forest a été choisi en conséquence comme le plus performant.

5. Problèmes rencontrés

Nous citons ici les différents problèmes rencontrés tout au long de son réalisation :

- La recherche des dataset et le retard des réponses des Courriel.
- Le problème de synchronisation entre la partie gestion du projet et la partie pratique.

6. Chronogramme

La réalisation de ce projet a été distribuée au niveau de plusieurs environnements en parallèle et sur une période de 120 jours. Ce tableau affiche le chronogramme de notre travail :

Tableau 5 : Chronogramme du travail

<table>
<tr><th colspan="4">Février</th><th colspan="4">Mars</th><th colspan="4">Avril</th><th colspan="4">Mai</th></tr>
<tr><th>1</th><th>7</th><th>15</th><th>28</th><th>1</th><th>7</th><th>15</th><th>31</th><th>1</th><th>7</th><th>15</th><th>30</th><th>1</th><th>7</th><th>15</th><th>31</th></tr>
<tr><td colspan="14">Collecte d’information (machine learning) et recherche de la base d’apprentissage adéquate</td><td></td><td></td></tr>
<tr><td></td><td></td><td colspan="14">Développement des parties de codage</td></tr>
<tr><td></td><td></td><td></td><td colspan="3"></td><td></td><td colspan="10">Test des algorithmes d’apprentissage</td></tr>
<tr><td></td><td></td><td colspan="5">Recherche sur la pollution atmosphérique</td><td></td><td></td><td colspan="8">Rédaction du rapport</td></tr>
</table>

Conclusion

Nous avons essayé au cours de ce dernier chapitre de mettre l’accent sur l’importance de la tests et analyse pour une modèle d’apprentissage automatique. En fait c’est une étape indispensable que nous ne pouvons pas ignorer vu le rôle qu’elle joue pour faciliter la tâche des utilisateurs et gérer leur dialogue avec le système.

Conclusion et perspectives

Étant donné que notre modèle peut prédire les données actuelles avec une précision de 99 %, il peut prédire avec succès l'air avec un indice de qualité de toute donnée spécifique dans une région donnée. Avec ce modèle, nous pouvons prévoir l'IQA et alerter les régions appropriées du pays. Parce qu'il s'agit d'un modèle d'apprentissage progressif, il est capable de remonter à l'emplacement spécifique qui nécessite une attention si les données de séries chronologiques de toutes les régions possibles est disponible. Les données sur la qualité de l'air utilisées dans cet article proviennent des tests et de l'enquête sur la qualité de l'air en Inde, et il comprend le problème quotidien normal de particules fines (PM2,5), le problème de particules inhalables (PM10), fixation de l'ozone (O3), du CO, du SO_2, du NO_2 et de la qualité de l'air (AQI). Les aspects les plus importants à considérer lors de la mesure du foyer de poison sont ses différentes sources ainsi que les facteurs qui influent sur sa fixation.

Il y a beaucoup de choses qui peuvent être faites dans ce domaine du système de prévision de la pollution atmosphérique comme le choix de travailler avec une dataset tunisienne.

Augmenter l'ensemble de données et l'ajout de plus de données peuvent donner au modèle une formation pour prédire des valeurs plus déviantes et plus vives qui sont requis. Il peut également être intégré au matériel et être une application de l'IoT (Internet des objets) par lequel il peut prédire la qualité de l'air en temps réel d'un lieu particulier.

Références bibliographiques

1. https://mrmint.fr/introduction-machine-learning
2. https://www.talend.com/fr/resources/what-is-machine-learning/
3. https://machinelearnia.com/apprentissage-supervise-4-etapes/
4. https://www.ibm.com/docs/fr/spss-modeler/saas?topic=models-how-svm-works
5. https://cedric.cnam.fr/vertigo/cours/ml2/tpSVMLineaires.html#:~:text=Les%20avantages%20des%20SVM%20%3A,(les%20vecteurs%20de%20s
6. https://www.journaldunet.fr/web-tech/guide-de-l-intelligence-artificielle/1501309-apprentissage-non-supervise/
7. https://www.talend.com/fr/resources/what-is-machine-learning/#:~:text=Les%20principaux%20algorithmes%20du%20machine,logistique%20et%20boos

Résumé

Le but de ce projet est de proposer système de prédiction et classification de la pollution au sein de la CPG pour son projet de fertilisation, afin de chercher les zones les plus polluées par le déchet de phosphates pour les considérer comme primordiale dans ce projet.

Le travail consiste à utiliser une Base d'apprentissage pré réalisée et faire le prétraitement de cette base pour l'utiliser comme entrée dans un modèle de machine learning permettant de prédire la pollution atmosphérique. L'utilisation des algorithmes SVM est basique puis une comparaison des résultats avec des autres type d'algorithme peut être enrichissante pour ce travail.

Mots clés : Machine learning, pollution atmosphèrique, SVM, Python

Abstract

The purpose of this project is to propose a pollution prediction and classification system within the CPG for its fertilization project, in order to seek the most polluted areas by phosphate waste to consider them as essential in this project.

The work consists of using a pre-made learning base and pre-processing this base to use it as input in a machine learning model to predict air pollution. The use of SVM algorithms is basic then a comparison of the results with other types of algorithm can be enriching for this work.

Keywords : Machine learning, air pollution, SVM, Python