République Algérienne Démocratique et Populaire

Ministère de l 'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ AMAR TELIDJI-LAGHOUAT-

Faculté des Sciences

.....

THÈME

....

Noms du membre de jury

...	MA(Université de Laghouat)	Président	
...	Mc	(Université de Laghouat)	Examinateur
...	Mc	(Université de Laghouat)	Examinateur
...	Mc	(Université de Laghouat)	Rapporteur
...	Mc	(Université de Laghouat)	Co-Rapporteur

N° d' ordre : ...../2017-2018/INF

Dédicace

Je dédie ce travail A mes très chers parents,
Aucune dédicace ne saurait exprimer l'amour,
le dévouement et le respect que j'ai toujours eu pour vous.
Rien au monde ne vaut les efforts fournis jour et nuit
pour mon éducation et mon bien-être.
Ce travail est le fruit de vos sacrifices.
Puisse Allah le tout puissant vous préservez et vous accorder santé,
longue vie et bonheur. A mes frères et soeurs adorés,
Les mots ne suffisent
guère pour exprimer l'amour que j'ai pour vous.
Je vous souhaite un avenir plein de bonheur et de réussite.
Je vous exprime à travers ce travail tous mes sentiments de fraternité.
A toute ma famille pour leurs encouragements et leur soutien.
A mes chers amis, Vous avez toujours été à mes côtés
dans les bons et les mauvais moments.
A toutes la promotion de première année Master,
Je dédie ce travail enfin
à toute personne ayant contribué de près ou de loin
à l'élaboration de ce travail.

Remerciement

En premier lieu, je remercier **ALLAH**
qui m'a permis d'arriver jusque-là,
Grand merci à mon encadreur Monsieur *BOUAKAZ MUSTAPHA*,
pour l'orientation, la confiance et la patience
qui ont constitué un apport considérable,
il m'a toujours aidé et montré les meilleures voies
pour avancer sûrement dans la réalisation de ce travail.
Mes vifs remerciements vont également aux membres du jury
pour l'intérêt qu'ils ont porté à notre recherche
en acceptant d'examiner ce travail.
Mes remerciements à tous ceux
qui a contribué de près ou de loin
à l'élaboration de ce mémoire.
A tous les enseignants du département d'informatique.

arabe mola5ase

Abstract

The extraction of knowledge from data (KDD) is defined as a process of discovering certain information unknown previously and potentially valuable from the data. This process is divided into many steps : From the preparation of data (research, cleaning and coding data), data mining (search for a knowledge model), validation and interpretation of results and finally integrate the knowledge that has been learned. We can apply this process in many fields, like : biology, medicine, agriculture and another lot of fields.

The variations in climate has a tremendous affectation on agriculture production, this last is strongly connected to the alimentary security who play a primary role in sovereignty and stability socio-economic of countries.

In this context, our project "application of a data mining technique for analyzing climate data" has an objective for analyzing data about the climate in different regions, we choose ten cities in Algeria : Alger, Constantine, Tiaret, Ghardaïa, Blida, Ghilizane, EL Oued, Tbessa, Beskra, et Laghouat in order to extract relations between them with the technique of data mining based on classification.

Key words : Data mining, Climate data, Classification.

Résumé

L'extraction de connaissances à partir de données (ECD) est définie comme un processus de découverte d'informations implicites, inconnues auparavant et potentiellement utiles à partir des données. Ce processus se fait en plusieurs étapes :la préparation des données (recherche, nettoyage et codage des données), fouille des données (recherche d'un modèle de connaissances), la validation et l'interprétation du résultat et enfin l'intégration des connaissances apprises. On peut appliquer ce processus dans plusieurs domaines, tels que la biologie, la médecine, l'agriculture et autres.

Les changements climatiques ont une grande influence sur la production agricole. Cette dernière est fortement liée à la sécurité alimentaire qui joue un rôle primordial dans la souveraineté et la stabilité socio-économique des payés.

Dans ce cadre intervient notre projet de fin d'étude *Application d'une approche data mining pour l'analyse des données climatiques*, dont l'objectif est d'analyser les données climatiques issues de différentes régions. On a choisi dix willayas en Algérie qui sont : Alger, Constantine, Tiaret, Ghardaïa, Blida, Ghilizane, EL Oued, Tbessa, Beskra, et Laghouat, afin d'extraire les relations entre eux, en utilisant une des techniques de la fouille de données basée sur la classification.

Mots clé : Fouille des données, Données climatiques, Classification.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Les changements climatiques sont la question déterminante, et l'un des défis majeurs de notre époque, ils ajoutent un stress considérable à nos sociétés et à l'environnement en générale. L'évolution des conditions météorologiques ont des effets qui touchent le climat terrestre, et leur conséquence sur le mode de vie de l'humanité que ça soit au niveau sanitaire, social, agricole ...etc. Ce phénomène tient son importance du fait qu'il touche des secteurs sensibles dont les conséquences peuvent déboucher sur une catastrophe à l'échelle planétaire. Ces impacts vont amener à des adaptations, qui combineront des modifications locales des systèmes de culture ou de gestion, ainsi que déplacements géographiques des systèmes de production et sans action drastique aujourd'hui, il sera plus difficile et coûteux de s'adapter aux conséquences futures de ces changements.

L'Algérie figure parmi les pays à forts risques de changement climatiques. C'est ce que révèle le rapport de l'université des Nations unies pour l'environnement et la sécurité humaine (UNU-EHS), l'Alliance Development Works de 2014, et relayé par les Décodeurs du journal français Le Monde.[1]

Dans de nombreux domaines, il est nécessaire de prendre des décisions critiques, dans un contexte parfois difficile et en un temps limité. Par exemple, un agriculteur qui doit prendre une décision pour déterminer le type de produit qui convient au climat de sa région dans un temps **x** afin d'assurer un taux de réussite élevé et ne provoque pas de perte, fait appel à ses connaissances et expériences pour prendre sa décision. Mais il ne peut pas se souvenir de tous les dossiers qu'il a étudié depuis des années.

L'analyse de grandes quantités des données est une nécessité dans lesquels ces données génèrent des découvertes et des intuitions qui surprennent même les experts. Chaque

entreprise tire parti de la collecte et de l'analyse de ses données : les hôpitaux peuvent détecter les tendances et les anomalies dans les dossiers de leurs patients, les moteurs de recherche peuvent améliorer le classement et le positionnement des annonces... La liste continue, avec la prédiction du climat et l'analyses des données climatiques.

L'extraction de connaissances à partir des données (ECD) est un processus non trivial d'identification de structures inconnues, valides et potentiellement utiles dans les bases des données [2]. Son objectif est d'aider l'être humain à extraire les informations utiles (connaissances) à partir des données dont le volume croît très rapidement. Les étapes de ce processus sont, l'acquisition des données multiformes, la préparation des données, la fouille de données, et enfin la validation et mise en forme des connaissances.

L'étude de ce mémoire traite de l'analyse des variations climatiques observées durant les années 2017/2018, pour les différents willayas suivantes : Laghouat, Biskra, Blida, Tébessa, Tiaret, Alger, Constantine, El Oued, Ghardaïa, Relizane, dans le but de répondre à moyen terme à la problématique générale questionnant sur l'analyse des données et l'extraction des connaissance à partir des données connues, plus particulièrement l'analyse des données climatiques, comment extraire des connaissances à partir des données climatiques ? Comment représenter ces connaissances ? et sur quelles dimensions vas t-on faire notre analyse ? et quel sont les données climatiques nécessaire pour l'analyse ?

Motivation

Nous avons choisi ce sujet pour les raisons suivant :

Les raisons subjectives

— Le désir personnel de traiter un tel sujet.
— L'enrichissement des connaissances sur la technique utilisé *fouille de données*, en plus de ça, on a déjà vue ce module cette année et sa a étais intéressant.
— Développer mes capacités de programmation avec java.
— Appliquer la compréhension de l'entrepôt de donnée.

Les raisons méthodologiques

— Il s'agit d'une approche qui prend une grande occupation dans les différents domaines.

— La disponibilité des références autour de ce sujet.

Les raisons objectives

— La nouveauté du sujet dans l'Algérie.

— La propagation des changements climatiques.

— L'influence positif qu'il apporte au point de la climatologie et d'autre point d'agriculture.

— Absence de telles applications en Algérie.

— Les phénomènes climatiques qui se sont passées pendant l'année 2018 [3], tel que l'inondation du Tébessa et les pertes résultent.

Objectifs

L'objectif de notre projet de fin d'étude est de mettre en place un système qui permet à l'utilisateur de :

— Regrouper les différentes régions en classes.

— Identifier les régions les plus adapté au tel type d'agricultures.

— Une meilleure prévision climatique.

En générale il permet de mettre en classe les régions, et défini quels différents types d'agriculture a associé à chaque région, et prévoie aussi les phénomènes et les changements climatiques des willayas appartenant à la même classe.

L'intérêt du projet

On peut illustrer l'intérêt de ce sujet aux points suivants :

— Faciliter la tâche de classification des régions.

— Une meilleure identification des régions agricoles, par conséquent assurer la prise des bonnes décisions concernant les projets agricoles, et la protection de la sécurité

alimentaire.

— La possibilité de prévoir les changements climatiques en temps réel, par conséquent la prévention de l'exposition aux pertes à cause des phénomènes climatiques.

— Minimiser les effets néfastes que peuvent être imposées par les changement climatiques.

Définition de la mission

Notre mission dans le cadre de ce projet concerne le développement d'un système d'aide à la décision, issu des techniques de fouille de donnée. Il s'agit essentiellement de contribuer au développement d'un système de classification guidé par l'approche TAG. Afin permettre le regroupement de différentes régions en utilisant un algorithme de génération de graphe. Ce système permet à l'utilisateur de :

— Définir les groupes des willayas choisies.

— Identifier les régions adaptées à l'un des produits agricoles.

— La prévision des changement climatiques au temps réel.

Structure du mémoire

Premièrement, on a commencé par une introduction générale où nous avons montré la problématique et l'objectif de notre travail, puis nous avons introduit trois chapitres qui sont les suivants :

Chapitre 1 : **Climatologie et travaux connexes.**

Dans ce chapitre nous définirons la climatologie, en expliquant les notions climatiques nécessaire à l'utilisation de notre application. Puis nous présenterons l'état de l'art des travaux permettant la classification des régions au point du climat et/ou la prévision climatique.

Chapitre 2 : **Fouille de données.**

Dans ce chapitre nous présenterons les concepts de la fouille de données, où sont décrites les différentes étapes d'un processus d'extraction de connaissances à partir des données. Parmi ces étapes, on va essayer de détailler la phase de fouille de données et présenter les algorithmes nécessaire à la réalisation de notre système.

Chapitre 3 : **Implémentation et résultat.**
Dans ce chapitre nous présenterons les régions d'études et montrerons pourquoi nous les avons choisies, puis nous expliquerons le contexte de notre système, et déterminerons les principaux cas d'utilisation, et élaborerons le diagramme en schéma en étoile, et présenterons les résultats obtenus sous forme de graphe à partir des captures d'écran.

Finalement, on terminera notre mémoire avec une conclusion générale et quelques perspectives intéressantes concernant ce travail.

Chapitre 1

Climatologie et travaux connexes

1.1 Introduction

Dans ce premier chapitre nous allons présenter les différents concepts de la climatologie tel que (notion de climat, les changements climatiques, variabilité climatique... etc.), après nous terminons avec une description des travaux connexes lié à notre mémoire.

1.2 Climatologie

1.2.1 Définition

La climatologie est l'étude du climat, de ses variations et de ses extrêmes, mais aussi de ses incidences sur diverses activités, sans s'y limiter, celles qui se rapportent à la santé, à la sécurité et au bien-être. Au sens strict, le climat peut être défini comme les conditions météorologiques moyennes régnant en un lieu particulier au cours d'une certaine période de temps. Pour décrire le climat, on peut se servir de données statistiques sur les tendances centrales et la variabilité d'éléments tels que la température, les précipitations, la pression atmosphérique, l'humidité et le vent, ou alors d'un ensemble d'éléments tels que des types de temps ou des phénomènes caractéristiques d'un lieu ou d'une région, voire de l'ensemble de la planète, sur une période donnée. [4]

1.2.2 Notion de climat

Au sens étroit du terme, le climat désigne généralement le *temps moyen*, il s'agit plus précisément d'une description statistique en fonction de la moyenne et de la variabilité de grandeurs pertinentes sur des périodes variant de quelques mois à des milliers, voire à des millions d'années. Ces grandeurs sont le plus souvent des variables de surface telles que la température, les précipitations et le vent. Dans un sens plus large, le climat est la description statistique de l'état du système climatique. [5]

1.2.3 Changement climatique

Les changements climatiques désignent une variation statistiquement significative de l'état moyen du climat ou de sa variabilité persistant pendant de longues périodes (généralement, pendant des décennies ou plus). Les changements climatiques peuvent être dus à des processus internes naturels ou à des forçages externes, ou encore à des changements anthropiques persistants de la composition de l'atmosphère ou de l'affectation des terres. [5]

1.2.4 Variabilité climatique

Le système climatique terrestre se compose de l'atmosphère, la biosphère, l'hydrosphère, la cryosphère et la lithosphère (Horton et al., 2010). Ces composantes interagissent de manière très complexe et sur une grande échelle spatio-temporelle (Viner et al., 2006). [6]

La variabilité du climat se réfère à la variabilité observée dans les données climatiques quand l'état du système climatique ne montre pas de changement (Mavi et tupper, 2004). La variabilité dans une série climatique est marquée par une stabilité de la moyenne (la série est dite stationnaire), et par une fluctuation des observations autour de cette moyenne (Burroughs, 2001). A l'inverse, le changement du système climatique est caractérisé par un changement dans les moyennes des variables climatiques calculées sur une longe période d'année, qui peut être accompagner par un changement dans la distribution des fréquences des évènements rares (Salinger et al., 2000). [6]

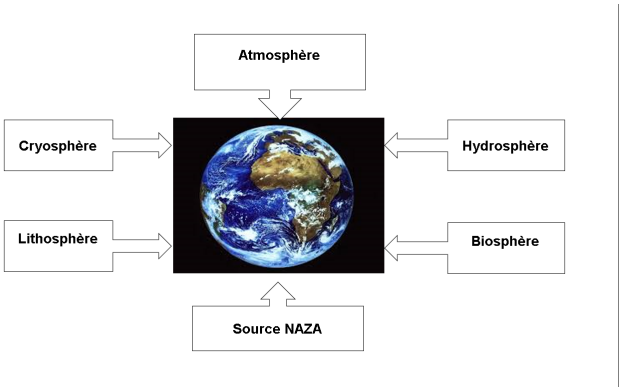

FIGURE 1.1 – *Le système climatique terrestre.*

1.2.4.1 Donnée climatique

C'est un moyen pour mesurer l'état du climat sur le long terme à travers des variables de diagnostic, tout en suivant l'écologie dans le temps et en diagnostiquant la sensibilité des activités anthropiques à l'aléa climatique.

1.2.4.2 Caractéristique d'une donnée climatique

La donnée climatique se caractérise principalement par :

— Une mesure de variable climatique peut-être de nature directe (cause à effet) ou indirecte (par association).

— La donnée climatique comporte souvent une marge d'erreur qui peut être de nature aléatoire ou systématique.

1.2.4.3 Types des données climatiques

Distinction selon la variable climatique (Variables essentielles pour la climatologie) :

— Atmosphère.
  — Pression .
  — Température.
  — Humidité .
  — Vitesse et direction du vent.
  — Ensoleillement.
  — ...etc.
— Hydrosphère.
  — Débit.
  — Indices océaniques.
— Lithosphère.
  — Température de surface.
  — Humidité du sol.
— Cryosphère.
  — Hauteur de neige.

  - Couverture de glace.
- Biosphère.
  - Indices de végétation.

Distinction selon l'espace et le temps :

- Résolution temporelle.
  - Donnée quotidienne.
  - Donnée mensuelle.
  - Donnée saisonnière.
- Résolution spatiale.
  - Donnée ponctuelle (station).
  - Données par grille (10 [Modèle Climatique Régional] à 300-400 km [MCG]).
  - L'espace en météorologie : échelle micro (cm), échelle méso (100 m), échelle synoptique (1000 m) [7]

Distinction selon la méthode de génération :

- Données de reconstruction.
- Données indirectes :
  - Cerne des arbres.
  - Datation au carbone 14.
  - Carottage glaciaire.
  - Analyse de sédiments marins ou lacustres (Shanahan et al, 2009). [7]
- Données d'observations.
- Données de surface.
- Données de sondage atmosphériques.
- Données satellites.
- Données de bouées.
- Données aéroportées.
- Données de sorties de modèles.
  - Modèles de circulation générale (MCG).
  - Modèles climatiques régionaux (dynamique et statistique).

1.3 Choix des willayas

L'Algérie est le pays dont la plus grande partie est désertique, et les changements climatiques constituent une préoccupation majeure. L'Algérie est exposée aux effets négatifs des changements climatiques et des émissions des gaz à effet de serre, notamment les inondations, la sécheresse et les températures élevées à cause de sa position géographique.

Notre présente étude ne concerne pas toutes les régions de l'Algérie mais on a choisi de se concentré sur des différentes régions, qui sont les wilayas suivantes : Alger, Blida, Constantine, Relizane, Ghardaïa, Tébessa, Laghouat, Biskra, El Oued et Tiaret.

Le tableau 1.1 représente les willayas qu'on a choisi :

Centre	Est	Ouest	Sud
Alger	Tébessa	Tiaret	Ghardaia
Blida	Constantine	Relizane	Laghouat
	Biskra		El Oued

TABLE 1.1 – *Positionnement des willayas choisies.*

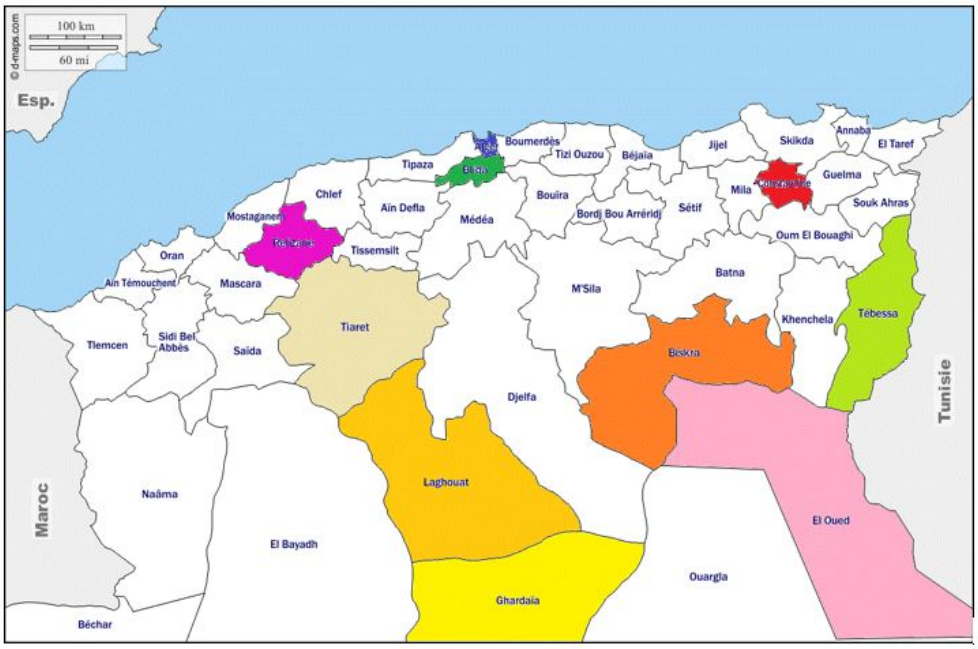

FIGURE 1.2 – *Les willayas d'Algérie sélectionner pour notre étude.*

Concernant le climat, l'Algérie se situe dans un climat de transition, entre la zone tempérée et la zone tropicale, cette position la met sous l'influence directe du climat méditerranéen au Nord où les étés sont chauds et secs et les hivers sont doux et pluvieux et parfois enneigés, et du climat désertique au Sud.

Au nord, les précipitations diminuent du nord au sud et d'est en ouest. Sur la bande littorale, le climat est tempéré, avec des hivers pluvieux ou très pluvieux, avec des moyennes pluviométriques annuelles pouvant atteindre plus de 1500 mm dans la région de Jijel et Bejaia. Cette variation dans le nord dépend de l'altitude, de la continentalité et du relief. En outre, une dissymétrie très nette existe entre les versants, ceux qui sont exposés au nord sont les mieux arrosés et ceux qui le sont au sud sont les plus secs. La moyenne des températures varie entre 8°C et 15°C en hiver et, en moyenne, 25°C en juillet et août (26,5°C à Annaba, 26°C à Bejaia). En été, le sirocco, vent sec et chaud (baptisé le Chehili localement), souffle du Sahara en direction du nord durant la saison estivale, amenant des nuages de poussières et de sable vers les régions côtières. [1]

Le climat dans la région de l'Atlas tellien est aussi tempéré, mais plus froid à cause de l'altitude. Il est caractérisé par des précipitations plus importantes. [33]

La figure 1.3 résume la pluviométrie dans le nord algérien.

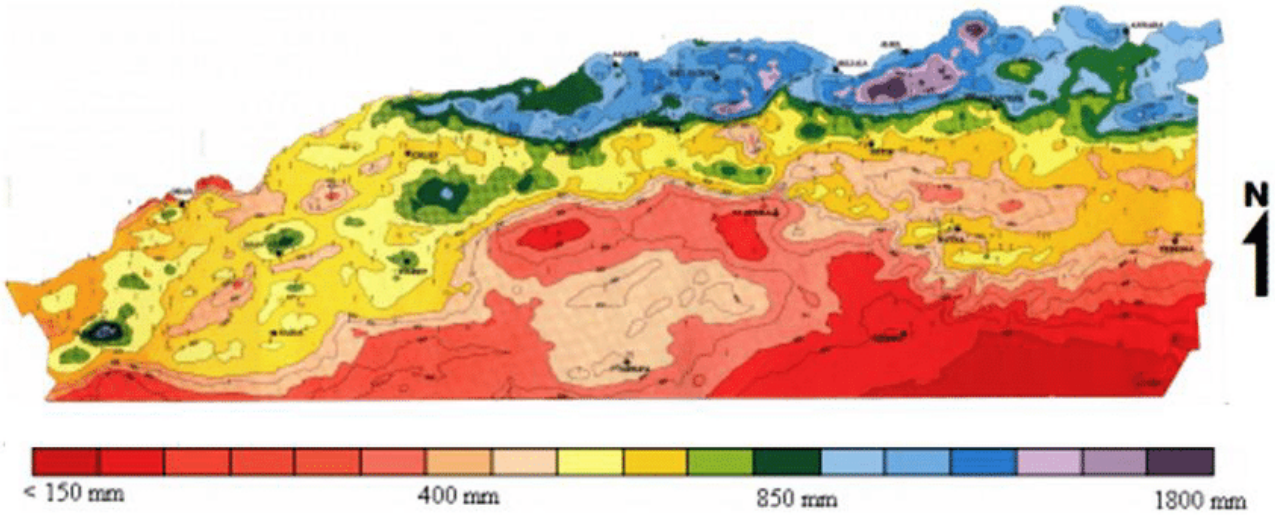

FIGURE 1.3 – *La pluviométrie dans le nord algérien.*

La figure 1.4 représente la variance des température entre les différentes willayas qu'on a choisi.

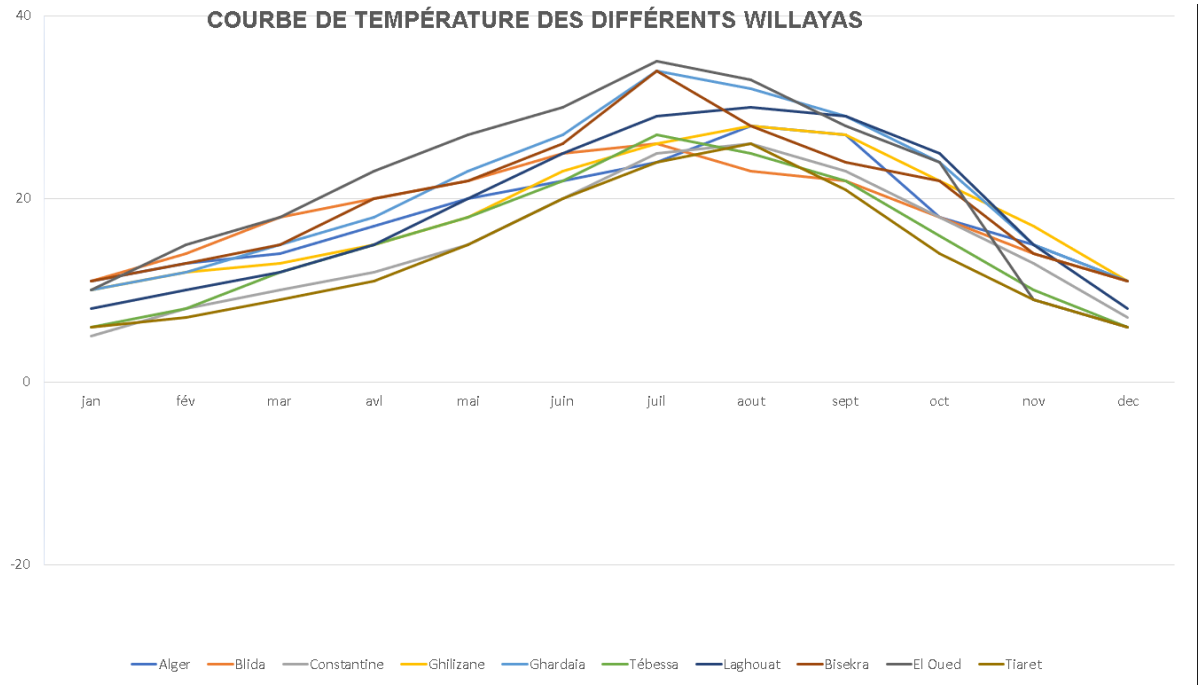

FIGURE 1.4 – *Schéma comparatif entre la température des différentes willayas.*

Alors que la demande de produits alimentaires, de fourrages, de fibres et de combustibles ne cesse d'augmenter, les changements climatiques risquent de dégrader irrémédiablement le stock de ressources naturelles dont dépend l'agriculture. La relation entre changements climatiques et agriculture est à double sens : l'agriculture contribue à maints égards aux changements climatiques, et les changements climatiques ont généralement des répercussions négatives sur l'agriculture.

Dans notre travail, on a essayé de choisi des willayas différentes par rapport à ses positionnements et aussi pour les raisons citées sur le tableau 1.2 :

Willaya	**Raisons**
Alger	Capital
Laghouat	Willaya natale
Tbessa	Inondation en 2018 Appartient au l'est
Constantine	Inondation en 2018 Appartient dans l'est
Biskra	Willaya natale
Tiaret	Possède un grand nombre des surfaces agricoles Appartient au l'ouest
Ghardaïa	Inondations en 2008 positionne dans sud d'Algérie
Blida	Variation des climat
El oued	Variation d'agriculture Positionnement sud
Ghilizane	Appartient au sud Variation des produits agriculture

TABLE 1.2 – *Les raisons de base du choix des régions de notre étude.*

1.4 Travaux liés à l'analyse de données

Dans cette section, nous passons en revue les recherches existantes sur l'entreposage de données et la technologie OLAP pour la gestion des données générées par les capteurs de station. Les données SIG sont une forme de données géo-spatiales. Le mot géo-spatial fait référence aux données ayant une composante géographique, ce qui signifie que les enregistrements d'un jeu de données sont associés à des informations de localisation, telles que des données géographiques sous forme de coordonnées, d'adresse, de ville ou de code postal. D'autres données géo-spatiales peuvent provenir de données GPS, d'images satellite et de géo-localisation. La plupart des outils SIG disponibles aujourd'hui utilisent encore les formats relationnels traditionnels. Cependant, les nouveaux ont commencé à adopter OLAP et d'autres bases de données NoSQL, mais beaucoup conservent encore

des requêtes de type SQL. [8]

Les récents développements dans les technologies sans fil ainsi que l'utilisation généralisée des capteurs ont conduit à la prévalence récente des systèmes de surveillance et de simulation de réseau. La principale fonctionnalité de ces systèmes consiste généralement à collecter et archiver des données provenant de capteurs répartis et à analyser les données archivées à des fins d'applications telles que la planification et la modélisation. Jiang et al [9]. On utilisé quatre modules pour l'infrastructure de stockage de données :

- Le module de référentiel de fichiers est utilisé pour la gestion des petits fichiers.
- Le module de base de donnée combine plusieurs bases de données et utilise à la fois la base de données NoSQL et la base de donnée relationnelle pour la gestion de donnée structurée.
- Le module de service extrait méta-données par le biais de la configuration, puis les maps à la fois sur les entités de données et les fichiers stockés dans les bases de données, ainsi que sur le référentiel de fichier en fonction des méta-données extraites, pour finalement générer le service correspondant.
- Le module de configuration des ressources prend en charge la gestion des données statiques et dynamiques en termes de méta-modèle prédéfini, ce qui permet de mettre en oeuvre un mécanisme de suppression des données tel qu'un équilibrage de charge et des préférences isolées.

Réçamement, de nombreux efforts ont été déployés dans le domaine du stockage et du traitement des données, en ce qui concerne l'Internet des objets (IoT). Des techniques de traitement de flux en adoption sont proposées pour l'intégration des données de capteur, une méthode qui prend en charge le traitement intégré flexible de grands volumes de données de capteur. Li et al [10]. fournit une méthode pour les capteurs sans fil afin de réduire la collecte de données redondantes. Li et al [11]. Crée une méthode rapide et robuste utilisant un test de compatibilité articulaire approximatif à base postérieure pour mettre en ?uvre une association de données. Wan et al [11]. Ont proposé une architecture de système à cinq couches pour intégrer les technologies de réseau de capteurs sans fil (WSN) et RFID. Afin de permettre à plusieurs utilisateurs d'effectuer la mise à jour ou la lecture de données, ces bases de données sacrifient généralement certaines fonctionnalités telles que la transaction de base de données et la cohérence pour améliorer

la disponibilité et l'évolutivité. De nombreuses bases de stockage traditionnelles reposent sur des bases de données relationnelles. Les RDMS ont été utilisés pour gérer et, dans une certaine mesure, analyser les données géo-spatiales du passé. En tant que compléments aux bases de données relationnelles, ces outils peuvent traiter efficacement des données volumineuses dans un environnement distribué. C'est pourquoi l'architecture moderne, telle que les bases de données OLAP et NoSQL, attire de plus en plus l'attention. Bien que les bases de données NoSQL fournissent un certain nombre de fonctionnalités que les bases de données relationnelles ne peuvent pas fournir, telles que l'évolutivité horizontale, la mémoire et l'index distribué, la modification dynamique du schéma de données, etc. [10].

Dans le passé, la technologie OLAP était utilisée pour gérer les données d'un système de télémétrie de ballon météo et de vol [12]. Le cadre utilise des requêtes OLAP spatiales pour analyser la trajectoire d'un ouragan en ce qui concerne la vitesse et la direction du vent. OLAP spatial permet une exploration spatio-temporelle de données volumineuses historiques avec la géo-visualisation et également une interactivité entre les capteurs . Il permet le stockage, la manipulation et la récupération de requêtes analytiques au moyen d'une expression multidimensionnelle (MDX). OLAP fournit une analyse temporelle avec des interprétations visuelles et prend également en charge la plupart des opérations ETL. De plus, il fournit à l'utilisateur un design interactif flexible pour explorer les données spatio-temporelles multidimensionnelles.

Conformément aux limitations ci-dessus, une base de données est nécessaire pour gérer ces différents ensembles de données dans le microclimat urbain. D'autre part, la base de données NoSQL manque de propriétés en termes d'atomicité, de cohérence, d'isolation, de durabilité (ACID) et de prise en charge de certaines requêtes complexes. Ainsi, dans ce travail, nous avons implémenté le cube OLAP en tant que *Plate-forme unifiée* pour stocker et récupérer les données environnementales ainsi que les données Wi-Fi.

1.5 Travaux liés à l'analyse de données climatiques

Au siècle dernier, les prévisions météorologiques constituaient l'un des problèmes les plus difficiles sur le plan scientifique et technologique dans le monde, principalement en

raison de deux facteurs principaux : l'impact direct sur les activités humaines et, deuxièmement, les phénomènes appliqués découlant de divers développements technologiques, tels que l'évolution de haute performance, Informatique (HPC ***H****igh* ***P****erformance* ***C****omputer*), technologies par satellite, etc. [8]

En général, l'extraction de connaissances est également appelée *L'extraction de connaissances à partir des bases de données* (KDD), un champ de découverte de nouvelles informations pouvant être utiles à partir de grandes quantités de données [14]. Contrairement aux méthodes traditionnelles telles que les méthodes statistiques standard, les techniques d'exploration de données recherchent des informations importantes sans pré-hypothèses, en fonction du type de modèle pouvant être détecté par les opérateurs de tâches d'extraction de connaissances. [13]

Il existe deux types de tâches d'extraction de connaissances : les tâches d'extraction de connaissances décrivant les caractéristiques générales des données actuelles et les tâches de prospection de connaissances attendues sur la base des données inférées disponibles [13]. Les méthodes les plus courantes d'extraction de connaissances sont les suivantes : algorithmes génétiques, méthodes de neurones voisins, réseaux de neurones artificiels, induction de base, logique basée sur la mémoire, régression logistique, analyse discriminante, arbres de décision, par certains chercheurs [9]. Plusieurs algorithmes sont en cours de développement pour allouer des ressources dans le système HPC et le système informatique en nuage afin de réaliser de grandes simulations afin de comprendre la variabilité climatique. [13]

Les ensembles de données en grille sur les précipitations sont utiles pour les études régionales sur le cycle hydrologique, la variabilité du climat et l'évaluation de modèles régionaux. Les données pluviométriques quotidiennes maillées haute résolution (1°x 1°lat/long) de 1951 à 2003 pour la région de l'Inde ont été utilisées ici pour évaluer les tendances d'événements extrêmes de précipitations saisonnières et annuelles et pour estimer les estimations des précipitations à partir de l'étude de cas. Cet ensemble de données a été développé par le département météorologique indien (IMD) du Centre national du climat de Pune en interpolant les données de précipitations quotidiennes de 1803 stations répar-

ties dans tout le pays [9]. Toutes les stations pluviométriques avaient au moins 90% de données disponibles au cours de la période allant de 1951 à 2003. Seules 1803 stations sur 6329 stations ont été utilisées à des fins d'interpolation afin de minimiser le risque de création d'homogénéités temporelles dans les données maillées en raison de la densité variable des stations [8]. La comparaison avec l'ensemble de données pluviométriques quadrillées globales a révélé que cet ensemble de données pluviométriques indiennes offre une meilleure représentation précise de la variation spatiale des précipitations. Lau et Wu [14] ont mené une étude similaire sur l'analyse des ensembles de données globales issus du Projet de climatologie des précipitations dans le monde (GPCP) et du projet Climat. Produit d'analyse fusionné du centre de prévision (CMAP). Bien que la variabilité interannuelle des précipitations saisonnières de la mousson estivale (juin à septembre) soit similaire dans les deux jeux de données, le jeu de données mondial sous-estime les fortes précipitations sur la côte ouest et le nord-est de l'Inde.

Les changements climatiques en Macédoine jusqu'en 2006 ont été analysés à travers leurs principaux paramètres : la température de l'air et les précipitations. Les données existantes ont été analysées entre 1971 et 2000 pour 34 stations météorologiques. Les informations sur les variations climatiques sont basées sur l'analyse comparative de deux séries sur une période de 30 ans, soit de 1961 à 1990 et de 1971 à 2000. Les études effectuées en Macédoine sont basées sur les 15 stations météorologiques sélectionnées appartenant à différents types et sous-types de climat. Des tests simples de qualité ont été appliqués[15], par exemple :

$$Tavg > Tmin, Tavg < Tmin, Tavg = (Tmin + Tmax)/2 \tag{1.1}$$

La formule 1.1 est utilisée pour détecter les erreurs de données. Il a été découvert que les données fournies comportaient souvent des erreurs, telles que l'absence du signe négatif des températures hivernales, une valeur basse ou élevée de 10°C, etc. Des données erronées peuvent fortement compromettre les tendances observées estimées, ainsi que le développement des modèles empiriques décrivant la relation entre la variabilité climatique à l'échelle mondiale et locale. En raison de problèmes avec les tendances en Macédoine, les résultats sont basés sur des simulations effectuées avec le GCM (modèle général de circulation).[14]

Les effets à long terme des changements climatiques en Macédoine sont estimés dans les zones les plus vulnérables : agriculture, foresterie, ressources en eau, bio-diversité et santé. Ces estimations sont faites en tenant compte des scénarios de changement climatique sur les sous-régions du pays. Afin d'atténuer les effets néfastes du changement climatique sur les secteurs susmentionnés, celui-ci doit respecter les priorités en matière d'adaptation dans le cadre du plan transnational.[16]

La Macédoine n'a pas obtenu de résultats significatifs liés aux conditions et aux changements climatiques. De plus, il n'y a pas d'application spéciale permettant de surveiller en permanence la situation dans le pays et ses relations avec les organisations mondiales. Jusqu'à présent, il a réalisé plusieurs projets de surveillance soutenus par le gouvernement macédonien, l'hydro-météorologie, l'institut national de la santé et plusieurs organisations internationales. Les faits mentionnés ci-dessus sont la raison principale de la mise en oeuvre d'un tel système qui pourrait fonctionner de la même manière que les applications du monde entier. La mise en place d'un tel système nécessite une bonne infrastructure technique et financière [17]

1.6 Conclusion

Ce chapitre nous a permis de détailler le cadre général de notre travail, en présentant les grands piliers de la climatologie, et on a quelque recherche qui ont été faites sur l'analyse des données climatique. Dans ce qui suit nous allons entamer la technique qu'on va suivre dans notre travail pour identifier les différentes fonctionnalités de l'application que nous allons concevoir et développer.

Chapitre 2

Fouille de données

2.1 Introduction

Dans ce deuxième chapitre nous allons présenter les concepts de fouille de données, où les différentes étapes du processus d'extraction de connaissances à partir des données seront décrites. Nous insisterons sur les différentes approches d'un modèle de fouille de données et nous allons décrire les algorithmes que nous utiliserons dans le prochain chapitre.

2.2 Entrepôt de donnée(Data Warehouse)

Un entrepôt de données est une base de données conçue pour faciliter la prise de décision dans une organisation. Les données des bases de données de production sont copiées dans l'entrepôt de données afin que les requêtes puissent être traitées sans perturber les performances ni la stabilité des systèmes de production. Pour que l'exploration de données ait lieu, il est essentiel que l'entrepôt de donnée soit présent. [18]

Un entrepôt de données est généralement modélisé par une structure de données multidimensionnelle, appelée **cube de données**, dans laquelle chaque dimension correspond à un attribut ou à un ensemble d'attributs du schéma, et chaque cellule stocke la valeur d'une mesure agrégée telle que nombre ou somme (montant des ventes). Un cube de données fournit une vue multidimensionnelle des données et permet le pré-calcule et l'accès rapide aux données. [19]

2.3 Extraction des connaissances à partir des données (ECD ou KDD en anglais)

2.3.1 Définition

L'**E**xtraction de **C**onnaissances à partir des **D**onnées est un processus itératif et interactif d'analyse d'un grand ensemble de données brutes afin d'en extraire des connaissances exploitables par un utilisateur analyste qui y joue un rôle central. [20]

D'après Fayyad, un processus d'ECD est constitué de quatre phases qui sont : le nettoyage et intégration des données, le pré-traitement des données, la fouille de données et enfin l'évaluation et la présentation des connaissances. [21]

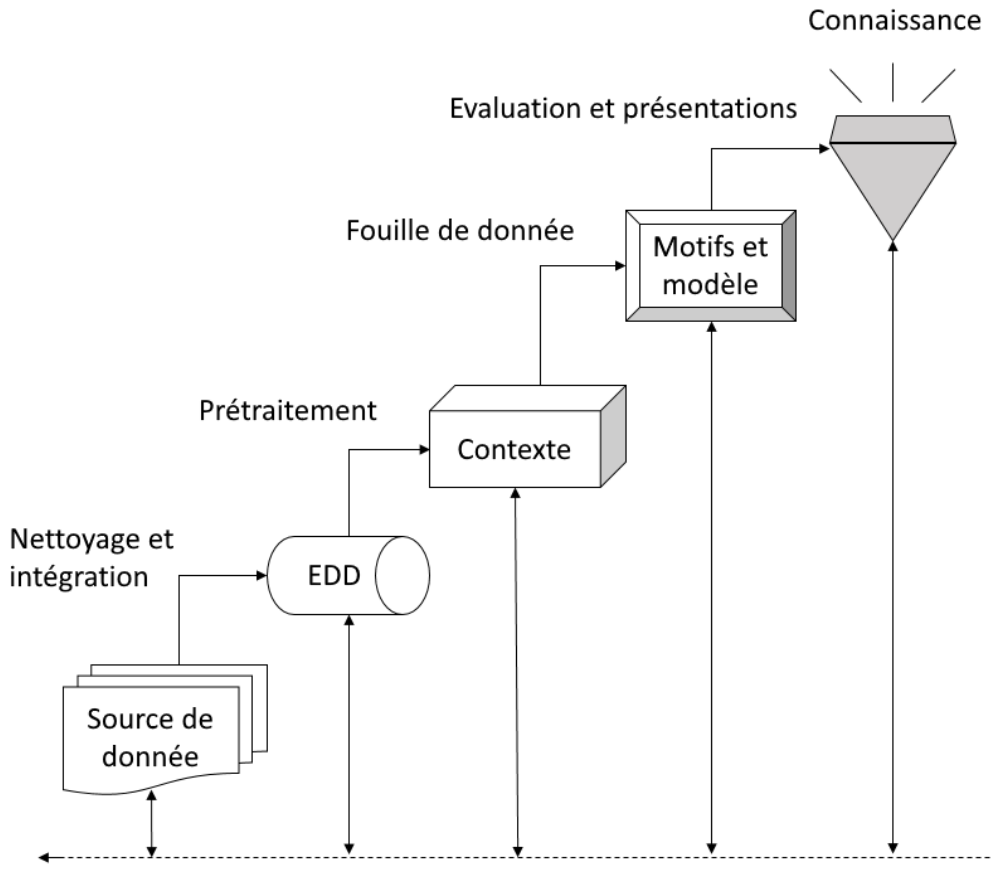

FIGURE 2.1 – *Les phases d'un ECD.*

La figure 2.1 désigne l'ensemble des phases qui permettent d'exploiter avec facilité et rapidité des données stockées massivement ainsi que les enchaînements possibles entre ces phases. Cette séparation est théorique car en pratique, ce n'est pas toujours le cas. En effet, dans de nombreux systèmes, certaines de ces étapes sont fusionnées. [22]

Le processus d'ECD peut avoir deux objectifs, soit vérifier les hypothèses d'un utilisateur, soit découvrir de nouveaux motifs. Un motif, ou schéma, est une expression dans un langage spécifique qui décrit un sous-ensemble de données ou un modèle applicable à ce sous-ensemble. [18]

2.3.2 Les étapes d'un processus d'ECD

Ce processus comporte quatre étapes principales :

1. Nettoyage et intégration des données.
2. Pré-traitement des données.
3. Fouille de données.
4. Évaluation et présentations.

2.3.2.1 Étape nettoyage et intégration des données

Le nettoyage des données consiste à traiter ces données bruitées, soit en les supprimant, soit en les modifiant de manière à tirer le meilleur profit. L'intégration est la combinaison des données provenant de plusieurs sources (base de données, sources externes, fichiers Excel, etc.). Le but de ces deux opérations est de générer des entrepôts de données et/ou des magasins de données spécialisés contenant les données traitées pour faciliter leurs exploitations futures. [23]

2.3.3 Étape pré-traitement des données

Il peut arriver parfois que les bases de données contiennent à ce niveau un certain nombre de données incomplètes et/ou bruitées. Ces données erronées, manquantes ou inconsistantes doivent être traitées si cela n'a pas été fait précédemment. Dans le cas contraire, durant l'étape précédente, les données sont stockées dans un entrepôt. Cette étape permet de sélectionner et transformer des données de manière à les rendre exploitables par un outil de fouille de données. Cette seconde étape du processus d'ECD permet d'affiner les données. Si l'entrepôt de données est bien construit, le pré-traitement de données peut permettre d'améliorer les résultats lors de l'interrogation dans la phase de fouille de données. [23]

2.3.4 Étape fouille de données

La fouille de données (data mining en anglais), est le coeur du processus ECD. Il s'agit à ce niveau de trouver des pépites de connaissances à partir des données. Tout le

travail consiste à appliquer des méthodes intelligentes dans le but d'extraire cette connaissance. Il est possible de définir la qualité d'un modèle en fonction de critères comme les performances obtenues, la fiabilité, la compréhensibilité, la rapidité de construction et d'utilisation et enfin l'évolutivité. Tout le problème de la fouille de données réside dans le choix de la méthode adéquate à un problème donné. Il est possible de combiner plusieurs méthodes pour essayer d'obtenir une solution optimale globale. Nous ne détaillerons pas d'avantage la fouille de données dans ce paragraphe car elle fera l'objet d'une section complète. [23]

2.3.5 Étape évaluation et présentations

Cette phase est constituée de l'évaluation, qui mesure l'intérêt des motifs extraits, et de la présentation des résultats à l'utilisateur grâce à différentes techniques de visualisation. Cette étape est dépendante de la tâche de fouille de données employée. En effet, bien que l'interaction avec l'expert soit importante quelle que soit cette tâche, les techniques ne sont pas les mêmes. Ce n'est qu'à partir de la phase de présentation que l'on peut employer le terme de connaissance à condition que ces motifs soient validés par les experts du domaine. Il y a principalement deux techniques de validation qui sont la validation statistique et la validation par expertise.

La validation statistique : Consiste à utiliser des méthodes de base de statistique descriptive. L'objectif est d'obtenir des informations qui permettront de juger le résultat obtenu, ou d'estimer la qualité ou les biais des données d'apprentissage. Cette validation peut être obtenue par :

— Le calcul des moyennes et variances des attributs.
— Si possible, le calcul de la corrélation entre certains champs.
— Ou la détermination de la classe majoritaire dans le cas de la classification.

La validation par expertise : Est réalisée par un expert du domaine qui jugera la pertinence des résultats produits. Par exemple pour la recherche des règles d'association, c'est l'expert du domaine qui jugera la pertinence des règles. Pour certains domaines d'application (Le diagnostic médical, par exemple), le modèle présenté doit être compré-

hensible. Une première validation doit être effectuée par un expert qui juge la compréhensibilité du modèle. Cette validation peut être, éventuellement, accompagnée par une technique statistique.

Grâce aux techniques d'extraction de connaissances, les bases de données volumineuses sont devenues des sources riches et fiables pour la génération et la validation de connaissances. La fouille de données n'est qu'une phase du processus d'ECD et consiste à appliquer des algorithmes d'apprentissage sur les données afin d'en extraire des modèles (motifs). L'extraction de connaissances à partir des données se situe à l'intersection de nombreuses discipline, comme l'apprentissage automatique, la reconnaissance de formes, les bases de données, les statistiques, la représentation des connaissances, l'intelligence artificielle, les systèmes experts, etc. (Figure 2.2). [23]

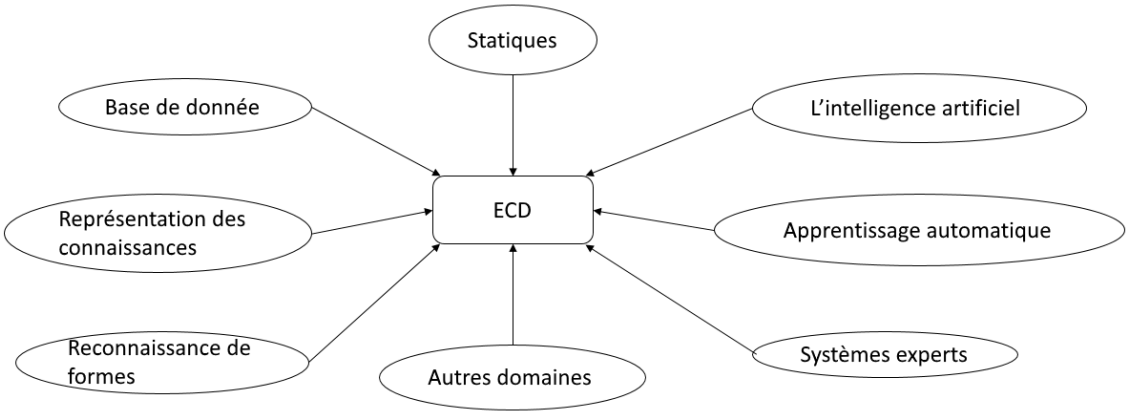

FIGURE 2.2 – *L'ECD à la confluence de nombreux domaines.*

On peut utiliser les connaissances pour modifient le comportement de l'agent qui les possèdent et inconnues auparavant. Transmissibilité vers les humains (compréhensibilité), vers d'autres systèmes fondés sur les connaissances ou relative indépendance par rapport à l'agent qui les a produites. [24]

2.3.6 Domaines d'application de l'ECD

— Activités commerciales tel que les grandes distributions, les ventes par correspondances, les banques et les assurances.

  - Segmentation de la clientèle.
  - Détermination du profil du consommateur.
  - Analyse du ”panier de la ménagère”.
  - Mise au point des stratégies de rétention de la clientèle.
  - Prédiction des ventes.
  - Détection des fraudes.
  - Identification des clients à risque.
- Activités financières
  - Recherche des corrélations ”cachées” entre les indicateurs financiers.
  - Prédiction de l'évolution de ces indicateurs.
- Activités de gestion des ressources humaines.
  - Prévision du plan de carrière.
  - Aide au recrutement.
- Activités industrielles
  - Détection et diagnostic des pannes et des défauts.
  - Réglage des équipements.
- Activités scientifiques
  - Diagnostic médical.
  - Étude du génome.
  - Analyses chimiques et pharmaceutiques.
  - Exploitation des données astronomiques.
  - Reconnaissance de la parole. [24]

2.4 Fouille de données : (Data Mining)

Les concepts de fouille de données et d'extraction de connaissances à partir de données sont parfois confondus et considérés comme synonymes. Mais, formellement on considère la fouille de données comme une étape centrale du processus d'extraction de connaissances des bases de données. [25]

2.4.1 Historique

L'expression "data mining" est apparue vers le début des années 1960 et avait, à cette époque, un sens péjoratif. En effet, les ordinateurs étaient de plus en plus utilisés pour toutes sortes de calculs qu'il n'était pas envisageable d'effectuer manuellement jusque-là. Certains chercheurs ont commencé à traiter sans a priori statistique les tableaux de données relatifs à des enquêtes ou des expériences dont ils disposaient. Comme ils constataient que les résultats obtenus, loin d'être aberrants, étaient tout au contraire prometteurs, ils furent incités à systématiser cette approche opportuniste. Les statisticiens officiels considéraient toutefois cette démarche comme peu scientifique et utilisèrent alors les termes "data mining" ou "data fishing" pour les critiquer. Cette attitude opportuniste face aux données coïncida avec la diffusion dans le grand public de l'analyse de données dont les promoteurs, comme Jean-Paul Benzecri [26], ont également dû subir dans les premiers temps les critiques venant des membres de la communauté des statisticiens. Le succès de cette démarche empirique ne s'est pas démenti malgré tout. L'analyse des données s'est développée et son intérêt grandissait en même temps que la taille des bases de données. Vers la fin des années 1980, des chercheurs en base de données, tel que Rakesh Agrawal [27], ont commencé à travailler sur l'exploitation du contenu des bases de données volumineuses comme par exemple celles des tickets de caisses de grandes surfaces, convaincus de pouvoir valoriser ces masses de données dormantes. Ils utilisèrent l'expression "database mining" mais, celle-ci étant déjà déposée par une entreprise (Database mining workstation), ce fut "data mining" qui s'imposa. En mars 1989, Shapiro Piatetski [28] proposa le terme "knowledge discovery" à l'occasion d'un atelier sur la découverte des connaissances dans les bases de données. Actuellement, les termes data mining et knowledge discovery in data bases (KDD, ou ECD en français) sont utilisés plus ou moins indifféremment. Nous emploierons par conséquent l'expression "data mining", celle ci étant la plus fréquemment employée dans la littérature. La communauté de "data mining " a initié sa première conférence en 1995 à la suite de nombreux atelier (workshops) sur le KDD entre 1989 et 1994. La première revue du domaine " Data mining and knowledge discovery journal " publiée par "Kluwers " a été lancée en 1997.

2.4.2 Définition

La fouille de données, est l'ensemble des méthodes et techniques destinées à l'exploration et l'analyse de bases de données informatiques (souvent grandes), de façon automatique ou semi-automatique, en vue de détecter dans ces données des règles, des associations, des tendances inconnues ou cachées, des structures particulières restituant l'essentiel de l'information utile tout en réduisant la quantité de données [29].

D'après Haddad [30], la définition la plus communément admise de Data Mining est celle de U. Fayyad [31] : *La fouille de donnée est un processus non trivial qui consiste à identifier, dans des données, des schémas nouveaux, valides, potentiellement utiles et surtout compréhensibles et utilisables.* En bref, la fouille de donnée est l'art d'extraire des informations (ou même des connaissances) à partir des données. [32]

2.4.3 Les techniques de la fouille de données

En général, Les tâches peuvent être classées en deux catégories descriptive et prédictive. Les tâches descriptives caractérisent les propriétés des données d'un ensemble de données cible. Les tâches prédictives effectuent une induction sur les données actuelles afin d'effectuer des prévisions. [31]

La fouille de données n'est pas une méthode d'attaque des données, au contraire, c'est une façon de s'associer à partir des données et d'utiliser ensuite ces informations. Pour cette raison, nous avons besoin d'un nouvel état d'esprit en fouille de données. Nous devons être ouverts à la recherche de relations et de modèles que nous n'avions jamais imaginés. Il existe quatre catégories de techniques/outils d'exploration de données [32] :

— Prédiction
— Classification
— Regroupement (ou Segmentation)
— Découverte des règles d'association

Chaque technique a ses propres outils :

Outils de prévision :

Ce sont les méthodes dérivées de la prévision statistique traditionnelle pour prédire la valeur d'une variable. Les applications les plus courantes et les plus importantes dans la fouille de donnée implique la prédiction. Cette technique implique des statistiques traditionnelles telles que l'analyse de régression, l'analyse à 10 discriminants multiples, etc. Méthodes non traditionnelles utilisées dans les outils de prévision sont l'intelligence artificielle et l'apprentissage automatique.

Outils de classification :

Les plus couramment utilisés dans la fouille de données. Les outils de classification tentent de distinguer différentes classes d'objets ou d'actions. Par exemple, dans le cas d'une carte de crédit transaction, ces outils pourraient le classer comme l'un ou l'autre. Cela permettra à la société émettrice de cartes de crédit d'économiser des sommes considérables.

Outils d'analyse de regroupement :

Ce sont des outils très puissants pour regrouper des produits en groupes qui tombent naturellement ensemble. Ces groupes sont identifiés par le programme et non par les chercheurs. La plupart des clusters découverts n'auront peut-être aucune utilité dans la prise de décision. Cependant, un ou deux découverts peuvent être extrêmement importants et peuvent être pris à l'entreprise un avantage face aux ses concurrents. L'utilisation la plus courante des outils de classification est probablement ce que les économistes appellent la *segmentation du marché* .

Découverte des règles d'association :

Les outils de fouille de données découvrent ici les associations, par exemple, quels types de livres lus par certains groupes de personnes, quels films regards par certains groupes de personnes, etc. Les entreprises peuvent utiliser ces informations pour cibler leurs marchés. Les détaillants en ligne comme Netflix et Amazon utilisent ces outils assez intensément. Par exemple, Netflix recommande des films basés sur ceux que les gens ont visionnés et notés dans le passé. Amazon fait quelque chose de similaire en recommandant des livres

lorsque vous revisitez leur site web. [31]

2.4.4 Les méthodes de fouille de donnée

Aujourd'hui, énormément de problèmes, de méthodes et d'algorithmes de fouille de données existent dans la littérature. Chaque domaine possède une panoplie de problèmes d'extraction de connaissances, chaque problème détient plusieurs méthodes et chaque méthode dispose de plusieurs algorithmes, que l'on choisira en fonction de :

— La tâche à résoudre.

— La nature et de la disponibilité des données.

— L'ensemble des connaissances et des compétences disponibles.

— La finalité du modèle construit.

— L'environnement social, technique, philosophique de l'entreprise.

Nous allons présenter dans les sous sections suivantes, les méthodes de la fouille de données, nous détaillerons par domaine les méthodes et algorithmes nécessaires pour la compréhension de ce travail.

2.4.4.1 La fouille de données classique

Nous appelons fouille de données classique, l'ensemble des problèmes de fouille de données qui ne considèrent pas les relations spatiales. Les objets étudiés sont souvent des tuples de bases de données ne contenant pas de dimension spatiale (localisation absolue). Les tuples peuvent être définis comme des ensembles de valeurs d'attributs ordonnées relatifs à un enregistrement (observation). Ces tuples sont organisés sous forme de structures de données en table qui sont reliées par des relations logiques. Pour ce qui est des relations spatiales, elles vont être définies dans la section 4.4.2.

Les associations

Les associations sont des problèmes non supervisés de fouille de données permettant d'extraire des relations d'implication au sein d'un même événement ou entre des séquences d'événements ordonnés dans le temps, on a :

— **Les Règles d'association :** Les règles d'association sont des règles d'occurrence

mettant en exergue les relations cachées par le volume important des bases de données. Elles sont basées sur la découverte de motifs fréquents qui sont des ensembles d'items apparaissant souvent ensemble dans les bases de données. Il y a toute une batterie d'algorithmes d'extraction de règles d'association qui existent comme Apriori, FP-growth, TreeProjection. Ces algorithmes différents du point de vue des performances (temps et espace requis d'exécution).

— **Les Motifs séquentiels :** Le problème de fouille de motifs séquentiels a été introduit pour la première fois par Srikant et Agrawal (Srikant et Agrawal 1996). La recherche de motifs séquentiels peut être vue comme une extension de la notion de règles d'association, intégrant des contraintes temporelles. Cette recherche met en évidence des associations entre les transactions alors que la recherche des règles d'association détermine les liens au sein d'une même transaction. Les motifs séquentiels sont extraits à partir de séquences d'événements ordonnées et souvent sauvegardés dans des bases de données transactionnelles. Plusieurs algorithmes d'extraction de motifs séquentiels sont proposés dans la littérature, nous pouvons citer SPADE (Zaki 2001), GPS (Srikant et Agrawal 1996) et PrefixSpan (Pei et al., 2001). Chacun des algorithmes cités précédemment, représente une approche d'extraction particulière. [30]

Le classement et prédiction

Selon S. Tufféry *le classement estime la valeur d'une variable à expliquer par d'autres variables du même individu appelées cibles ou explicatives. Si la variable à expliquer est qualitative alors la technique est appelée Classement et si elle est continue, elle est appelée Prédiction.* Le classement consiste à analyser de nouvelles données et à les affecter à des classes prédéfinies ou modélisées au préalable. Le classement et prédiction sont des problèmes supervisés d'analyse de données qui sont souvent utilisés pour prédire des valeurs ou des libellés de classes. Plusieurs techniques, approches et méthodes de classement et de prédiction existent, comme les arbres de décision, les réseaux bayésiens, le raisonnement à base de règles, les algorithmes génétiques, la régression linéaire/non linéaire et le support vecteur machine.

La segmentation (Clustering)

La segmentation est utilisée depuis toujours dans le subconscient humain pour distinguer les différents éléments qui composent le monde qui l'entoure. Plusieurs techniques peuvent être utilisées pour la découverte de clusters. Ces techniques sont de plusieurs catégories :

— Méthodes par partitionnement (K-means, K-medoids, CLARANS, EM, etc.),
— Méthodes hiérarchiques (DIANA, BIRCH, CURE, etc.),
— Méthodes de densité (DBSCAN, OPTICS, etc.),
— Méthodes de grilles (STING, WaveClusters, Clique),
— Méthodes par modèle (Réseaux de neurones, etc.).

Les séries chronologiques

Les séries chronologiques permettent d'identifier les séquences similaires à une portion de données, de prévoir et de déterminer les causalités à partir des bases de données de séries temporelles. Ces bases de données peuvent êtres des bases de données séquentielles ou de valeurs (mesures) obtenues pour des intervalles de temps, comme le cas des mesures de température. L'objectif des Séries temporelles est de chercher dans de grandes quantités de données, des motifs similaires, réguliers, cycliques, les comportements (tendance), les impulsions, etc. Cela permet de modéliser le mouvement en le décomposant en une série de mouvements basiques (tendance, saisonnalité, cyclicité, irrégularités, horizontalité, etc.) et faire des prédictions de mouvements futures.

Analyse d'aberrations

Les aberrations appelées aussi outliers (terme anglais) sont des données qui ne suivent pas le comportement ou le modèle général. Les outliers doivent être manipulés avec prudence car ils peuvent décrire une erreur ou une variabilité importante dans le comportement du système étudié. Ces outliers sont soit supprimés pour qu'ils n'influencent pas les tendances globales, ou au contraire mis en évidence s'ils véhiculent une information utile à un domaine d'application particulier. Il y a énormément de domaines qui s'intéressent à

ce problème de détection d'aberrations. Dans cet objectif, différentes méthodes informatiques ont été proposées dans la littérature. Ces méthodes utilisent différentes approches comme les statistiques, les mesures de distances, la déviation et la densité

2.4.4.2 La fouille de données spatiales

Le domaine de la fouille de données spatiale est un domaine à part entière qui a été amorcé par les premiers travaux de Koperski, J. Han (Han et al., 1997) et Ester (Ester et al., 1997). Ce domaine s'intéresse à la découverte de modèles dans une base de données spatiales (Committee 2003). La principale caractéristique de ce domaine est sa prise en compte de la dimension spatiale et des relations entre les objets (Chelghoum et Zeitouni 2004). Les objets étudiés sont des thèmes rassemblant les objets de même type. Ces thèmes ne sont rien d'autre que des tables avec un attribut de localisation où les interactions entre les objets sont représentées par des prédicats et des tables de distances. Les méthodes de fouille de données spatiales sont la plupart du temps scindées en deux types (Aufaure et al., 2000) : Les méthodes exploratoires ou mono thématiques qui s'appliquent à un seul thème géographique et permettent d'identifier les écarts et/ou les similarités entre les objets, et Les méthodes décisionnelles ou multi-thématiques qui s'appliquent à plusieurs thèmes géographiques dans le but d'expliquer les écarts et les caractéristiques des groupements. [32]

2.4.4.3 La fouille de données d'objets mobiles

L'application de la fouille de données aux historiques de positions d'objets mobiles ouvre de nouvelles perspectives intéressantes. En effet, cela va permettre d'identifier les comportements normaux et anormaux d'objets en mouvement et faire des prédictions ou des classements. D'un autre point de vue, le déplacement pose des problèmes aux modèles de données géo-spatiales et à la fouille de données spatiales. En effet, il est difficile d'identifier et de catégoriser les objets mobiles sur des trajectoires (Committee 2003). [29]

2.4.4.4 La fouille de données du trafic de mobiles

Pour pouvoir distinguer ces deux derniers domaines, à savoir la fouille de données d'objets mobiles et la fouille de données du trafic, il faut observer l'objet sur lequel porte l'analyse. Dans la fouille de données d'objets mobiles, l'analyse porte sur le mobile et ses

déplacements. Dans la fouille de données du trafic, l'objectif est d'analyser les flux d'objets mobiles qui passent par des segments de réseaux (routier, ferroviaire, etc.). La différence peut être cernée aussi dans les données de capteurs, étant donné que la fouille de données d'objets mobiles utilise des données individuelles issues de capteurs embarqués alors que la fouille de données du trafic, utilise des données du trafic mesuré en différentes parties du réseau.

2.4.5 Notre contribution

Plusieurs travaux ont été proposés dans la littérature dans les téchniques de fouille de données, dans notre étude nous avons adopté un algorithme proposé récemment [33] qui sert à agrégé les données d'étude afin d'extraire des connaissances utiles pour les décideurs. Cet algorithme est nommé TAG (m) que nous l'avons utilisé pour agréger des données climatiques à la place des données textuelles.

L'approche TAG prend en entrée un ensemble de documents D, il utilise les fonctions ExtractKeywords(D) pour extraire les mots-clés des faits qui sont regroupés dans une matrice FreMat qui correspond aux fréquences des mots clés dans les documents du corpus, puis à l'aide d'une fonction qui calcule l'affinité entre ces mots clés, une matrice carrée AffiMat est obtenue. Enfin nous construisons le graphe d'affinité AffiGraphe afin de trouver les mots clés les plus représentatifs à partir du circuit le plus pertinent. Ce dernier nous permet d'obtenir les agrégats représentatifs du corpus. Le pseudo-code de cette approche est présenté dans l'algorithme TAG.

Algorithme TAG

1. **Entrées**
2. Un corpus de document D = {D1,D2,D3...,Dn}
3. **Sorties**
4. Une matrice de fréquence FreMat
5. Une matrice d'affinité AffMat
6. Un graphe d'affinité AffiGraph
7. Un circuit d'agrégats
8. Liste des agrégats ListAgg
9. **Début**

   //Extraction de l'ensemble des mots clé et calcul leurs fréquences
10. Kw = ExtractKeywords(D) ;
11. NombreDesMotsClés = |Kw|
12. pour chaque Di à Dn faire
13. pour chaque Kwj à Kwm faire
14. FreMat(i,j) = FreqKwInD(Di, Kwj) ;

    // Calcul de la somme des fréquences de chaque mot clé
15. pour chaque Kwj à Kwm faire
16. pour i =1 ; i<= |D| ; i++ faire
17. SomFreqKw(Kwj) ; MettreàJour(VFreqKw(j)) ;

    // Construction de la matrice d'affinité AffiMat de Kw
18. pour chaque (Kwi,Kwj) à VKw faire
19. Calculer Affinité (Kwi,Kwj) ;
20. MettreàJour (AffiMat(I,j)) ;

    // Construction de graphe d'affinité AffiGraph
21. i= randmoise(NombreDesMotsClés) ; count =0 ;
22. Tant que count < NombreDesClés faire
23. Indice-Sommet = MaxAffiMat(Kwi) ;

24. Valeur-de-lien = AffiMat(Kwi, Indice-Sommet) ;
25. MettreàJour-AffiGraph (Kwl, Indice-Sommet, Valeur-de-lien) ;
26. Si (il existe un circuit) alors
27. MettreàJour-Liste-des-circuits(circuit, Poids) ;
28. count++ ;
    // Sélection de circuit pertinent
29. circuit-pertinent = MaxPoids(Liste-des-circuits) ;
30. ListAgg.add(circuit-pertinent) ;
31. **Fin.**

2.5 Conclusion

Dans ce deuxième chapitre, nous avons présenté les principaux concepts de fouille de données, les processus, les tâches et les méthodes les plus utilisés en fouille de donnée. Dans notre travail nous nous intéressons aux techniques de la classification automatique, ou segmentation (clustering). Et on a détailler sur l'approche TAG car elle sera utilisée dans notre système.

Chapitre 3

Implémentation et résultat

3.1 Introduction

Dans cette partie, on va décrire les étapes qui nous ont permis la réalisation de notre système pour l'analyse des données climatiques, on va utiliser les techniques de fouille de donné, et Java comme langage de programmation. Pour développer cette application, on a utilisé la base de données MYSQL pour le stockage des données climatiques. Puis on a présenté les résultats obtenus sous formes des graphes pour avoir une vision générale sur le fonctionnement du système.

3.2 Localisation géographique

Dans cette section nous allons présenter et citer les régions sélectionner pour faire notre étude, puis nous décrirons les raisons pour lesquels nous les avant choisi.

L'Algérie est située au centre du continent Nord-africain, elle est le plus grand pays en Afrique. Avec près de 1200 Km de côte sur la mer Méditerranée, elle est bordée à l'est par la Tunisie, au Sud Est par la Libye, au Sud par le Niger et le Mali, au Sud-Ouest par la Mauritanie et à l'ouest par le Sahara Occidental et le Maroc (voir figure 3.1).
[34]

La figure 3.1, présente la position géographique de l'Algérie dans l'Afrique du nord.

FIGURE 3.1 – *Location géographique d'Algérie.*

3.3 L'agriculture en Algérie

La superficie agricole totale, représentant 3% de la superficie totale de l'Algérie, Elle comprend principalement :

— Les cultures herbacées : 3,8 millions ha.

— Les terres au repos (jachères) : 3,7 millions ha.

— Les plantations fruitières : 576 990 ha.

— Les vignobles : 81 550 ha.

— Les prairies naturelles : 23 640 ha. [33]

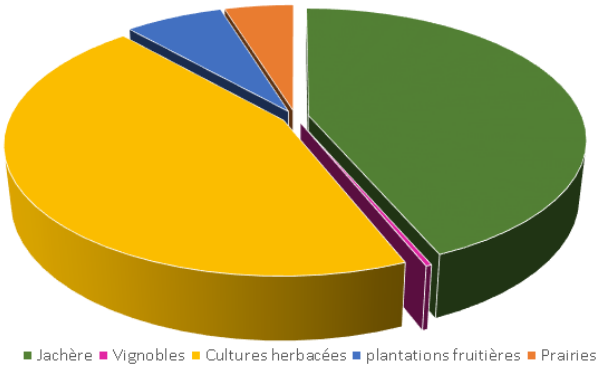

FIGURE 3.2 – *Occupation des terres agricoles en Algérie.*

La figure 3.2 représente les différentes occupations des terres agricoles en Algérie.

3.4 Création de l'entrepôt de données

3.4.1 Schéma en étoile

Afin de refléter les besoins décisionnels, nous utilisons le modèle schéma en étoile qui permet l'analyse du contenu des données climatiques pour faciliter les prises de décisions. Un schéma en étoile textuel E est défini par E = (TF, TD) où TF est la table de fait, et TD = {D1, ..., Dn} est un ensemble de dimensions. Un fait F est défini par F = (CF, MF) où CF = {CF1, ..., CFq} est un ensemble de clés étrangères qui associe la table du fait F aux dimensions Di. Une dimension D est définie par D = (AD) où AD = {AD1, ...,

ADu} est un ensemble d'attributs (paramètres et attributs faibles). Une mesure MF = {M1, ..., Mn} est un ensemble de mesures. La figure 3.7 représente notre conception de l'entrepôt de données, comme en peut le voir dans la figure 3.3, la table climat représente la table de fait de l'entrepôt, et les autres tables (temps, willaya, station, climat-variable, et meusure) sont les dimensions de la table de fait.

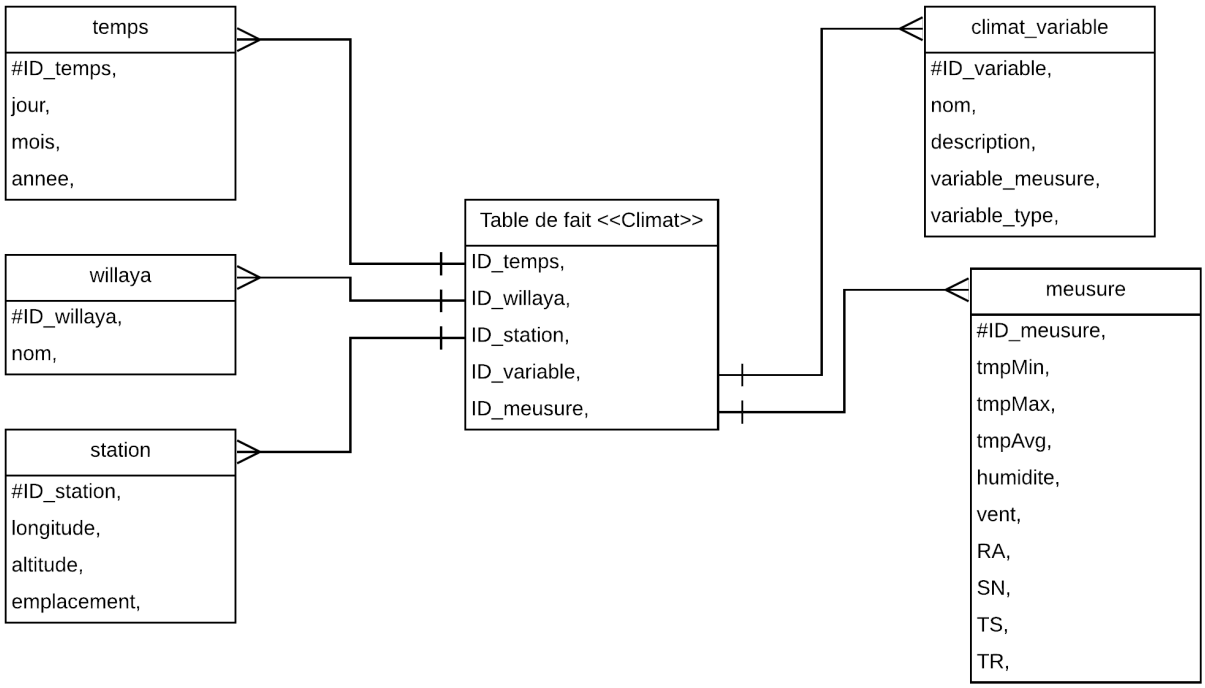

FIGURE 3.3 – *Schéma en étoile de notre entrepôt de données.*

La table *Climat* contient que les clés étrangères des tables de dimensions, mais elle ne possède aucune mesure, dans notre cas on n'a pas besoin c'est la raison pour ne pas créer aucune mesure.

La table *station*, représente la station qui a fait la capture des données, cette table a été créer pour connaitre quelle station et connaitre aussi son emplacement, d'où les variable longitude et altitude, et pour la variable emplacement c'est pour savoir dans quelle zone se trouve-t-elle.

La table *meusure* contient les différentes données climatiques nécessaire à notre étude, tel que :

— **tmpMin** pour représente la température minimale atteint dans la journée,
— **tmpMax** pour la température maximale atteint dans la journée
— Et, **tmpAvg** pour la température moyenne.
— **RA** signifier indicateur d'occurrences de pluie ou bruine
— **SN** signifier indicateur d'occurrences de neige ou pluie verglaçante.
— **TS** signifier indicateur d'occurrences de tonnerre.
— **TR** signifier indicateur d'occurrences de tornade ou nuage en entonnoir.

La table *climat-variable* représente les phénomènes climatiques tel que (inondations, séismes, sécheresse et les tempêtes côtières...etc.) Avec leur (nom, description, variable mesure, variable type).

3.4.2 Traitement des données

Cette étape est la plus importantes dans notre travail, car tous les résultats qui seront obtenus plus tard dépendront d'elle. On a essayé de travailler avec des données réelles mais, malheureusement, tous ce qu'on a réussi à trouver était payantes, On a trouvé des sites relatifs continents les données climatiques qui était trop chères, donc on a créé un programme en java qui génère des données aléatoires de façons semblable à la réalité pour une durée de 2 ans (2017/2018).(voire l'annexe)

Comment les données sont générées

— **Temps :** On a créé un petit programme java qui gêner des dates comprises entre le 01/01/2017 et le 31/12/2018 et les stocks dans un fichier, ce fichier est utilisé ensuite pour remplir notre EDD.
— **Température** : On a fait des recherches en consultant plusieurs sites web qui détiennes les données climatiques d'Algérie, puis on a extrait les températures maximales et minimales durant les années 2017 et 2018, pour chaque mois de chaque willaya choisie, on a généré la température nécessaire pour notre travail de façons proche à la réalité. Suivant les conditions suivantes :

$$Tmoy \in [Tmin, Tmax] \tag{3.1}$$

$$Tmin' \in [Tmin, Tmoy] \tag{3.2}$$

$$Tmax' \in [Tmoy, Tmax] \tag{3.3}$$

tel que

— **Tmoy :** Représente la température moyenne pendant une journée.

— **Tmin' :** Représente la température minimale atteinte pendant une journée.

— **Tmax' :** Représente la température maximale atteinte pendant une journée.

L'exemple figurant sur le tableau 3.3, représente la température maximale et minimale atteintes dans Laghouat pendant toute l'année 2017. Ce que nous a permis de générer notre base d'étude.

N°du mois	**01**	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**
La température minimale(°C)	2.7	3.4	8	8.9	13.5	18.4	21.6	20.6	17.1	11.4	5.8	3.4
La température maximale(°C)	12.9	15.5	18.1	22.8	26.8	32.6	36.3	35.3	30	23.5	18.1	13.4

TABLE 3.1 – *Tableau climatique de Laghouat.*

3.5 Notre système

L'objectif de notre contribution est de proposer une approche qui est une variante de l'algorithme TAG (*c.à.d, que TAG travail avec des données textuelle et notre système à été adapté pour travailler avec des données numériques*) décrits précédemment dans le chapitre 2, section 2.5. Et en utilisant les graphes pour présenter les résultats obtenus La démarche se décompose en quatre phases principales. Tout d'abord, à partir des données climatiques de chaque willaya, un pré-traitement nécessaire est appliqué à ces données pour pouvoir les utilisés, car cette approche ne prend pas des documents textuels comme paramètre d'entrée comme pour l'algorithme TAG. Dans la deuxième phase, nous modélisons ces données à l'aide de différentes matrices. Un algorithme de construction du graphe est ensuite appliqué. Enfin, l'extraction des agrégats est faite par la sélection de circuit le plus pertinent dans le graphe.

3.5.1 Formalisation par graphe

Arrêtes et sommets : la structure du graphe permet de représenter la relation et la dépendance entre plusieurs éléments, même si les éléments sont très nombreux. Les éléments qui composent un graphe sont généralement appelés sommets ou noeuds (vertices). Les liens entre ces éléments se divisent en deux types, liens orientés et liens non orientés selon les propriétés structurelles des graphes. Dans la modélisation du graphe, les liens orientés sont appelés "arêtes" pour les distinguer de ceux non orientés qui sont appelés *arc*. Un graphe est dit non orienté si tous les arcs sont symétriques, c'est-à-dire : il existe une relation binaire entre ces sommets. [34]

Théoriquement un graphe **G = (V, E)** est défini par deux ensembles, l'ensemble fini **V = {v1, v2, ..., vn}** dont les éléments sont appelés sommets (Vertices en anglais), et par l'ensemble fini **E = {e1, e2, ..., em}** dont les éléments sont appelés arêtes (Edges en anglais). Une arête e de l'ensemble **E** est définie par une paire non ordonnée de sommets, appelés les extrémités de **e**. Si l'arête e relie les sommets **a** et **b**, on dira que ces sommets sont adjacents, ou incidents avec **e**, ou bien que l'arête e est incidente avec les sommets **a** et **b**. On appelle ordre d'un graphe le nombre de sommets n de ce graphe. Pour la représentation graphique, les graphes sont des modélisations topologiques sans géométrie particulière. Ils tirent leur nom du fait qu'on peut les représenter par des dessins. À chaque sommet de **G**, on fait correspondre un point distinct du plan et on relie les points correspondant aux extrémités de chaque arête. Il existe donc une infinité de représentations d'un graphe.

Connexité et parcours : Un graphe non orienté est donc un ensemble de noeuds reliés par des arcs, on le considère comme un graphe simple s'il y a au plus un arc entre deux noeuds d'une part et s'il n'y a pas de boucle réflexive d'un noeud sur lui-même. Un graphe se caractérise par son ordre et sa taille : le nombre de ses noeuds représente l'ordre du graphe et le nombre de ses arcs en est la taille. Un graphe est connexe si à partir d'un noeud **x** on peut atteindre un autre noeud **y** par quelques sauts consécutifs. Un noeud se caractérise par son degré qui représente le nombre des arcs qui l'ont pour extrémité. Si un noeud n'a aucune arrête/arc qui le prend comme un point de départ en l'appel un noeud

isolé, c'est à dire il n'a aucun voisin. [34]

3.5.2 Description de système

On a développé un système avec le langage de programmation JAVA, basé sur les techniques de fouille de donnée qui est la classification, en suivant une variante de l'algorithme TAG, décrite sur la section 2.4.5. Tout d'abord on a commencé par programmé une fonction permettant de calculer la température moyenne pour chaque mois de chaque willaya, pendant la période janvier 2017 / décembre 2018 en appliquant la formule 3.4 :

$$TmpMoy(x) = \sum_{1 \leq w \leq 10; 1 \leq m \leq 12; 2017 \leq a \leq 2018} \frac{tmpMoy(w, j, m, a)}{(nombredejourparmois)} \tag{3.4}$$

Puis on a affiché le résultat dans un matrice (24 x10) qui contient dans les colonnes les mois de Janvier 2017 jusqu'à Décembre 2018, et les dix willayas dans les lignes (voir le tableau 3.2).

N°du mois	**01**	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**
Laghouat	7.0	9.1	12.5	15.7	19.8	26.1	28.8	28.9	23.4	17.9	12.3	7.9
Biskra	10.9	12.2	16.4	20.7	24.6	30.9	33.7	32.6	28.9	21.6	17.0	12.0
Blida	10.4	11.7	13.3	15.2	18.9	22.2	25.4	26.1	24.1	19.3	15.2	11.6
Tebessa	5.7	7.5	9.8	13.4	18.2	24.1	27.2	26.4	23.1	15.3	11.0	7.1
Tiaret	5.9	6.2	8.7	11.4	16.1	20.1	25.9	25.7	22.4	15.9	10.4	6.1
Alger	11.0	12.0	13.8	15.1	10.0	21.7	24.3	25.6	23.9	19.6	14.7	11.9
Constantine	7.3	7.9	10.6	11.8	16.2	21.3	24.4	26.5	22.8	16.6	12.6	7.8
El Oued	10.3	12.2	16.7	20.1	26.9	29.5	32.6	31.7	30.1	21.9	14.7	11.6
Ghardaia	9.6	11.3	15.3	18.3	24.4	30.1	32.3	33.3	27.0	21.2	14.1	11.4
Relizane	10.4	11.2	13.3	16.7	19.9	23.8	27.6	29.0	24.7	19.6	14.9	10.8

TABLE 3.2 – *Tableau des températures moyennes pour l'an 2017.*

Puis on a utilisé le tableau précédent pour calculer la matrice d'affinité en utilisant la

fonction 3.5 :

$$Aff[i,j] = \begin{cases} \Sigma \frac{|Tmp(k,j) - Tmp(k,i)|}{24} & , \ i \neq j \\ \Sigma \frac{T(k,i)}{48} & , \ i = j \end{cases} \tag{3.5}$$

On a obtenu la matrice d'affinité représenté sur le tableau 3.3.

Finalement on a classé les willayas selon le résultat obtenu, après l'application de la formule d'affinité 3.5 et l'utilisation d'un graphe d'affinité, on a choisi une des willayas de façons aléatoire, puis on chercher a trouver les willayas qui ont une similitude de température selon leur poids, pour les mettre dans une seule et même classe.

Graphe d'affinité : un graphe d'affinité (AffiGraph) est un graphe pondéré où les arrêtes sont étiquetées et dont toutes les étiquettes sont des nombres réels positifs ou nuls. Ces nombres sont les poids des liaisons entre les sommets. Le poids d'un circuit dans un graphe d'affinité est la somme des poids des arêtes qui constituent le circuit. [34]

	LAGHOUT	*BISKRA*	*BLIDA*	*TBESSA*	*TIARET*	*ALGER*	*CONSTA*
LAGHOUT							
BISKRA							
BLIDA							
TBESSA							
TIARET							
ALGER							
CONSTANTINE							
ALOUED							
GHARDAIA							
RILIZANE							

TABLE 3.3 – *Tableau des températures moyennes pour l'an 2017.*

3.6 Résultat

La figure 3.4 représente notre graphe d'affinité obtenu après l'application de notre l'algorithme, en a commencer d'abord par sélectionner un sommet au hasards, le 2éme sommet correspond a la valeur **Min** du 1er sommet, en continu comme ça jusqu'à que toutes les willayas sont pris. les valeurs sont prises en utilisant la matrice d'affinité résultant (voir le tableau 3.3)

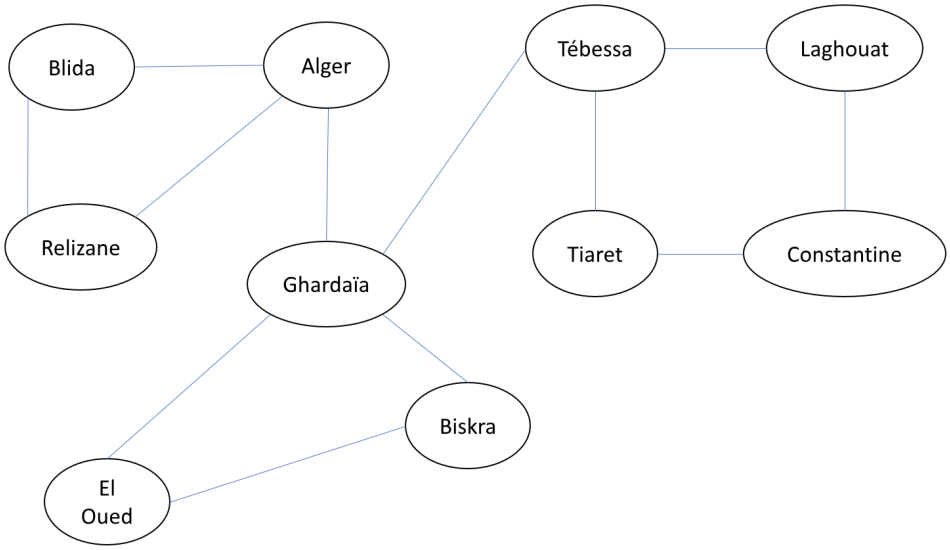

FIGURE 3.4 – *Graphe d'affinité obtenu.*

En utilisant le graphe d'affinité, on a cherché les circuits pertinents dans ce graphe, pour finalement obtenir les classes représenter sur la figure 3.5

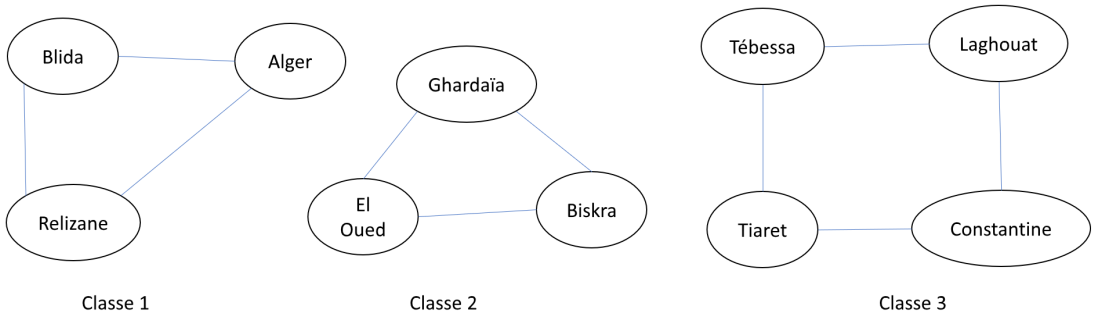

FIGURE 3.5 – *Classification des willayas*

Nous avons ensuite représenté ses classes dans un histogramme pour mieux observée

la température atteinte durant les 2 saisons, celle d'hiver et d'été. (voire la Figure 3.7)

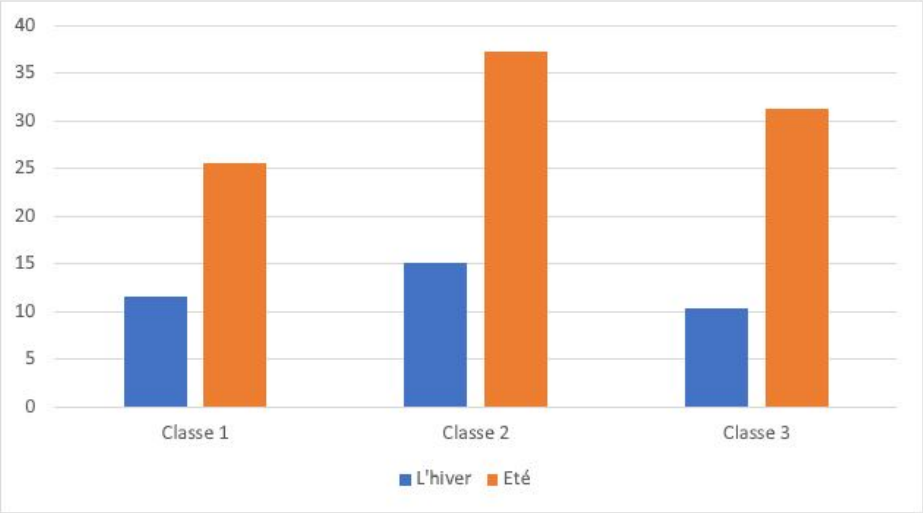

FIGURE 3.6 – *La température atteint des classes durant les deux saisons l'hiver et l'été.*

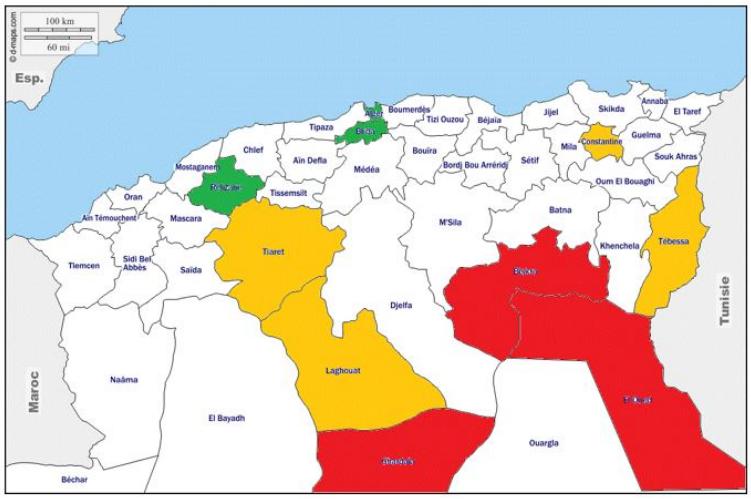

FIGURE 3.7 – *Représentation géographique des classes obtenu.*

3.7 Conclusion

Nous avons consacré ce chapitre " Implémentation et résultat" qui est la dernière partie dans notre travail pour expliquer l'algorithme qu'on a utilisé pour réaliser notre système, et présenter les résultats de notre projet.

Conclusion générale

Notre mémoire a pour objectif principal d'utiliser une technique de fouille de données pour l'analyser les données climatiques en Algérie. Jusqu'à nos jours aucune étude n ?est faite dans ce sens. Pour cela on a essayé d'initier une première étude qui prend en charge l'historique des changements climatiques en Algérie à fin d'extraire des connaissances utiles à l'aide d'une technique de fouille de données nommée TAG. Cette approche que nous avons implémenté en JAVA sert à regrouper les Wilaya d'Algérie en classe selon leurs similarité climatiques.

Comme perspectives, on pourrait penser aux fonctionnalités suivantes :

— Comparer les résultats obtenus par TAG avec d'autres approches.

— Le développement d'un système avec une interface graphique qui offert à l'utilisateur plusieurs fonctionnalités, tel que la recherche des régions similaires selon un produit agricole.

Annexe

Abréviation

ECD : **E**xtraction de **C**onnaissance à partir des **D**onnées.

EDD : **E**ntrepôt **D**e **D**onnées.

TAG : **A**grégation **T**extuelle

MCG : **M**odèle **C**limatique **R**égional.

OLAP : **O**n**L**ine **A**nalytical **P**rocessing.

SIG : **S**ystème d'**I**nformation **G**éographique.

SQL : **S**tandard **Q**uerry **L**anguage.

NoSQL : **N**ot **O**nly **SQL**.

GPS : **G**lobal **P**ositioning **S**ystem.

MDX : **M**ultiDimensional E**X**pressions.

ETL : **E**xtract **T**ransform **L**oad.

ACID : **A**tomic **C**ohérence **I**solation **D**urabilité.

HPC : High **P**erformance **C**omputer.

Code source

```
public class GenerateurDonnée {

    public static void main(String[] args) throws IOException {
        File temps = new File("C:\\Users\\DELL\\Dropbox\\Mémoire AM\\BDD\\temps.txt");
        BufferedWriter bw = new BufferedWriter(new FileWriter(temps, true));
        bw.write("INSERT INTO temps (jour,mois,annee) values ");
        bw.write("\r\n");
        String date = "";
        Calendar c = new GregorianCalendar(2017, Calendar.JANUARY, 1);
        int j = 0;
        int m = 0;
        int a = 0;
        while (a != 2019) {
            j = c.getTime().getDate();
            m = c.getTime().getMonth() + 1;
            a = c.getTime().getYear() + 1900;
            date = "(" + j + "," + m + "," + a + "),";
            bw.write(date);
            bw.write("\r\n");
            c.add(Calendar.DAY_OF_MONTH,1);
            a = c.getTime().getYear() + 1900;
        }
        bw.close();
    }
}
```

```
public static void genererTemps(float x, float y, int a, int b, int c, int k) throws IOException{
    float tmpMinM = x;
    float tmpMaxM = y;
    File mesure = new File("/home/arthas/Bureau/achour/mesure.txt");
    File climat = new File("/home/arthas/Bureau/achour/climat.txt");
    BufferedWriter bwMesure = new BufferedWriter((new FileWriter(mesure, true)));
    BufferedWriter bwClimat = new BufferedWriter((new FileWriter(climat, true)));
    float tmpAvg,tmpMin,tmpMax ;
    Random r = new Random();
    int idTemps = a;
    int idMesure = b;
    int idWillaya =c;
    for (int i = 0; i < k ; i++) {
        tmpAvg = r.nextFloat();
        tmpAvg = tmpAvg*(tmpMaxM - tmpMinM )+ tmpMinM ;
        tmpAvg = Math.round(tmpAvg*10);
        tmpAvg = tmpAvg/10;
        tmpMin = r.nextFloat();
        tmpMin = tmpMin*(tmpAvg - tmpMinM )+ tmpMinM ;
        tmpMin = Math.round(tmpMin*10);
        tmpMin = tmpMin/10;
        tmpMax = r.nextFloat();
        tmpMax = tmpMax*(tmpMaxM - tmpAvg )+ tmpAvg ;
        tmpMax = Math.round(tmpMax*10);
        tmpMax = tmpMax/10;
        bwMesure.write("("+tmpMin+","+tmpMax+","+tmpAvg+","+0+","+0+","+0+","+0+","+0+","+0+"),");
        bwMesure.write("\r\n");
        bwClimat.write("("+idTemps+","+idWillaya+",1,1,"+idMesure+"),");
        bwClimat.write("\r\n");
        idTemps++;
        idMesure++;
```

```
class Willaya {

    public String nom;
    public int idWillaya ;
    public float tmpMoy [];
    public String marke ;
    public int pred ;
    public int poids;

    public Willaya (String nom, int idWillaya){
        this.nom = nom;
        this.idWillaya = idWillaya;
        tmpMoy = new float[24];
        marke = "blanc";
        pred = -1;
    }

    public boolean estMarke(){
        if(marke.equals("blanc")){
            return false;
        }
        return true ;
    }

    public static void genererGraphe(Willaya w[], int x){
        System.out.println("Le Graphe generer est : ");
        while (w[x].marke.equals("gris")){
                System.out.print(" "+w[x].nom);
                w[x].marke = "rouge";
                x = w[x].pred;
        }
```

```
public void markeWillaya(int pred, Willaya w[]){
    if(marke.equals("blanc")){
        this.pred = pred;
        marke= "gris";
    } else if (marke.equals("gris")) {
        int ancienPred = this.pred;
        this.pred = pred;
        while (ancienPred != -1){
            w[ancienPred].marke = "blanc";
            int x = w[ancienPred].pred;
            w[ancienPred].pred = -1;
            ancienPred = x ;
        }
        genererGraphe(w, this.pred);
    }
}

public static boolean verifierColeur(Willaya w[]){
    boolean b = false;
    int i = 0;
    while(w[i].marke.equals("rouge") && i < w.length){
        i++;
    }
    if(i == w.length){
        b = true ;
    }
    return b;
}



public void calculerTmpMoyX(float [] t, int x){
    tmpMoy[x]= 0;
    for(int i=0; i<t.length ; i++){
        tmpMoy[x] += t[i];
    }
    tmpMoy[x] = tmpMoy[x]/t.length;
    tmpMoy[x] = Math.round(tmpMoy[x] *10);
    tmpMoy[x] = tmpMoy[x] /10;
}

public void afficherTmpMoy(){
    System.out.println("La temperature pour la willaya de "+nom);
    for (int i = 0 ; i < tmpMoy.length ; i++){
        System.out.print("  "+tmpMoy[i]);
    }
    System.out.println("");
}
```

```java
import java.sql.Connection;
import java.sql.DriverManager;
import java.sql.SQLException;

public class MysqlCon {

    private static String url = "jdbc:mysql://localhost:3306/climat";
    private static String driverName = "com.mysql.jdbc.Driver";
    private static String username = "Jaina";
    private static String password = "*******";
    private static Connection con;

    public static Connection getConnection() {
        try {
            Class.forName(driverName);
            try {
                con = DriverManager.getConnection(url, username, password);
            } catch (SQLException e) {
                System.out.println("Failed to create the database connection.");
                System.out.println(e.getMessage());
            }
        } catch (ClassNotFoundException ex) {
            System.out.println("Driver not found.");
        }
        return con;
    }
}
```

```java
private static Connection con;
private static PreparedStatement ps;
private static ResultSet rs;
private static float tmpMoy[][] = new float[10][24];
private static float aff [][] = new float[10][10];
private static Willaya willaya[] ;

private static void init() {
    con = MysqlCon.getConnection();
    String sql = "Select count(ID_willaya) from willaya";
    try {
        ps = con.prepareStatement(sql);
        rs = ps.executeQuery();
        if(rs.next()){
            int l = rs.getInt(1);
            willaya = new Willaya[l];
        }
        sql = "Select * from willaya";
        ps = con.prepareStatement(sql);
        rs = ps.executeQuery();
        while(rs.next()){
            int i = rs.getInt(1);
            willaya[i-1] = new Willaya(rs.getString(2),i);
            System.out.println("i = "+i +"  nom = "+willaya[i-1].nom );
        }
    } catch (SQLException ex) {
        Logger.getLogger(Classification.class.getName()).log(Level.SEVERE, n
    }
```

```
private static void getData(Willaya a) {
    for (int annee = 2017; annee <= 2018; annee++) {
        for (int mois = 1; mois <= 12; mois++) {
            try {
                String sql = "Select tmpAvg from willaya w, temps t, meusure m,
                ps = con.prepareStatement(sql);
                ps.setString(1, a.nom);
                ps.setInt(2, mois);
                ps.setInt(3, annee);
                rs = ps.executeQuery();
                int l = getLength(mois);
                float t [] = new float[l];
                int i=0;
                while(rs.next()){
                    t[i] = Float.parseFloat(rs.getString(1));
                    i++;
                }
                int x;
                if(annee == 2017){
                    x = mois-1;
                }else{
                    x = mois +11;
                }
                a.calculerTmpMoyX(t, x);
            } catch (SQLException ex) {
                Logger.getLogger(Classification.class.getName()).log(Level.SEVER
            }

        }
    }
```

```
public static float calculerWW(int i , int j ){
    float w=0;
    for(int k=0 ; k<tmpMoy[i].length ; k++){
        w += Math.abs(tmpMoy[i][k] - tmpMoy[j][k]);
    }
    w = w/tmpMoy[i].length;
    w = Math.round(w*10);
    w = w/10;
    return w;
}

public static void remplireTmpMoy(float t[], int i ){
    for (int j = 0; j < t.length ; j++) {
        tmpMoy[i][j] = t[j] ;
    }
}
public static float calulerWWDiagonal (int i) {
    float w = 0;
    for (int j =0; j<tmpMoy[i].length ; j++){
        w += tmpMoy[i][j];
    }
    w = w/(tmpMoy[i].length*2);
    w = Math.round(w*10);
    w = w/10;
    return w;
}
```

Bibliographie

[1] Article de presse ń l ?Algérie est un pays à fort risque de changement climatique ż, www.liberte-algerie.com, consulté le 12/02/2019.

[2] HAN, Jiawei, KAMBER, Micheline, et PEI, Jian. Data mining concepts and techniques third edition. Morgan Kaufmann, 2011.

[3] Article de presse ń Intempéries à Tébessa : le centre-ville et plusieurs quartiers inondés, un enfant de 5 ans décédé ż, http ://www.radioalgerie.dz, consulté le 16/06/2019.

[4] CRUCIFIX, Michel. La climatologie aujourd ?hui. Revue des Questions Scientifiques, 2011, vol. 182, no 1, p. 3-32.

[5] MOUNIA, TOUHAMI. Régionalisation et variabilité pluviométrique dans le Nord Centre-Ouest algérien (Approche statistique. 2017.

[6] BILAN, G. I. E. C. des changements climatiques : les éléments scientifiques. 2001.

[7] KARL, Thomas R., NICHOLLS, Neville, et GHAZI, Anver. Clivar/GCOS/WMO workshop on indices and indicators for climate extremes workshop summary. In : Weather and Climate Extremes. Springer, Dordrecht, 1999. p. 3-7.

[8] Casas DM, Gonzalez AT, Rodrigue JEA, Pet JV. Using data mining for short-term rainfall forecasting. Notes in Computer Science. 2009.

[9] NOM, Prénom, PRIÉ, Yannick, et BLANCHARD, Julien. Fouille interactive instantanée de motifs évolutifs pour l ?exploration de données d ?activité.

[10] BESSE, Philippe, GARIVIER, Aurélien, et LOUBES, Jean-Michel. Big Data Analytics-Retour vers le Futur 3 ; De Statisticien Data Scientist. arXiv preprint arXiv :1403.3758, 2014.

[11] Bregman, J.I., Mackenthun K.M., Environmental Impact. Statements, Chelsea : MI Lewis Publication. 2006.

[12] ZIGHED, Djamel A. et RAKOTOMALALA, Ricco. Graphes d ?induction : apprentissage et data mining. Paris : Hermes, 2000.

[13] Randall O and Bluestein M, Analysis of break and active monsoon spells. Current Science 91 : 296 ?306. 2005.

[14] BOVALO, Christophe, BARTHE, Christelle, et BÈGUE, Nelson. A lightning climatology of the South-West Indian Ocean. Natural Hazards and Earth System Sciences, 2012, vol. 12, no 8, p. 2659-2670.

[15] Folorunsha Olaiya, Adesesan Barnabas Adeyemo, Application of Data Mining Techniques In Weather Prediction And Climate Changes Studies, International Journal of Information Engineering and Electronic Buisness, pp.51-59,2012.

[16] GOULDEN, Marisa, CONWAY, Declan, et PERSECHINO, Aurelie. Adaptation to climate change in international river basins in Africa : à review/Adaptation au changement climatique dans les bassins fluviaux internationaux en Afrique : une revue. Hydrological Sciences Journal, 2009, vol. 54, no 5, p. 805-828.

[17] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth P, From data mining to knowledge discovery in databases, advices in knowledge discovery and data mining, MIT Press, vol. 1, pp 136, 1998.

[18] NOM, Prénom, PRIÉ, Yannick, et BLANCHARD, Julien. Fouille interactive instantanée de motifs évolutifs pour l ?exploration de données d ?activité.

[19] The Morgan Kaufmann Series in Data Management Systems Jiawei Han Micheline-Kamber Jian Pe Data Mining. Concepts and Techniques 3rd Edition MorganKaufmann, 2011.

[20] BOUATTANE, O., CHERRADI, B., et MAROC, ENSET Bd Hassan II Mohammedia. ALGORITHME PARALLELE DE CLASSIFICATION APPLICATION A LA SEGMENTATION D ?IMAGES IRM CEREBRALES.

[21] LENCA, Philippe, MEYER, Patrick, VAILLANT, Benoît, et al.Evaluation et analyse multicritère des mesures de qualité des règles d ?association. Revue des Nouvelles Technologies de l ?Information (Mesures de Qualité pour la Fouille de Données), 2004, p. 219-246.

[22] FAYYAD, Usama M., PIATETSKY-SHAPIRO, Gregory, SMYTH, Padhraic, et al. Advances in knowledge discovery and data mining. 1996.

[23] Article de presse ń Extraction de connaissances à partir de données incomplètes et imprécises ż, these.univ-msila.dz, consulté le 06 février 2019.

[24] litis.univ-lehavre.fr/ fournier/SRO/dess0607, consulté le 06 février 2019

[25] J. Lieber. Fortement mais librement inspire du cours d'amedeo napoli. Fouille de données : notes de cours. 2007.

[26] ZIGHED, Djamel A. et RAKOTOMALALA, Ricco. Graphes d'induction : apprentissage et data mining. Paris : Hermes, 2000.

[27] AGRAWAL, Rakesh, IMIELI ?SKI, Tomasz, et SWAMI, Arun. Mining association rules between sets of items in large databases. In : Acm sigmod record. ACM, 1993. p. 207-216.

[28] HAN, Jiawei, FU, Yongjian, WANG, Wei, et al. DMQL : A data mining query language for relational databases. In : Proc. 1996 SiGMOD. 1996. p. 27-34.

[29] KANTARDZIC, Mehmed. Data mining : concepts, models, methods, and algorithms. John Wiley et Sons, 2011.

[30] BOVALO, Christophe, BARTHE, Christelle, et BÈGUE, Nelson. A lightning climatology of the South-West Indian Ocean. Natural Hazards and Earth System Sciences, 2012, vol. 12, no 8, p. 2659-2670.

[31] FAYYAD, Usama M., PIATETSKY-SHAPIRO, Gregory, SMYTH, Padhraic, et al. Advances in knowledge discovery and data mining. 1996.

[32] NOM, Prénom, PRIÉ, Yannick, et BLANCHARD, Julien. Fouille interactive instantanée de motifs évolutifs pour l'exploration de données d'activité.

[33] The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-MichelineKamber-Jian-Pei-Data-Mining. -Concepts-and-Techniques-3rd-Edition-MorganKaufmann-2011 pp 10

[34] BOUAKKAZ, Mustapha, LOUDCHER, Sabine, et OUINTEN, Youcef. OLAP textual aggregation approach using the Google similarity distance. IJBIDM, 2016, vol. 11, no 1, p. 31-48.