



REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE  
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique  
Université Mohamed Khider – BISKRA  
Faculté des Sciences Exactes, des Sciences de la Nature et de la Vie

## Département d'informatique

N° d'ordre : SIOD16/M2/2021

### Mémoire

Présenté pour obtenir le diplôme de master académique en

## Informatique

Parcours : Système d'Information Optimisation et Décision (SIOD)

---

# La réalisation d'un système de prédiction dans e-commerce

---

Par :

**KISRANE ALI**

Soutenu le .././.... devant le jury composé de :

|              |       |            |
|--------------|-------|------------|
| Nom Prénom   | grade | Président  |
| Chami Djazia | MAA   | Rapporteur |
| Nom Prénom   | grade | Examineur  |

Année universitaire 2020-2021

## Remerciement

بعد بسم الله الرحمن الرحيم

Je tiens premièrement à prosterner remerciant Allah le tout puissant de m'avoir donné le courage et la patience pour terminer ce travail.

Je tiens à remercier mon directeur de mémoire, madame Chami Djazia, de m'avoir encadré, orienté, aidé et conseillé.

J'adresse mes plus sincères remerciements à ma famille : Mes parents, mes frères et tous mes proches et amis, qui m'ont accompagné, aidé, soutenu et encouragé tout au long de la réalisation de ce mémoire.

## Table de matière :

|   |    |
|---|----|
| Introduction générale.....                            | 1  |
| Chapitre 1 : Prédiction (Data Mining).....            | 2  |
| 1. Introduction .....                                 | 2  |
| 2. Définition du Data Mining .....                    | 2  |
| 3. Architecture du Data Mining.....                   | 2  |
| 4. Processus du Data Mining .....                     | 3  |
| 4.1. Nettoyage des données .....                      | 4  |
| 4.2. Intégration de données.....                      | 4  |
| 4.3. Réduction de donnée .....                        | 4  |
| 4.4. Transformation des données.....                  | 4  |
| 4.5. Data Mining.....                                 | 4  |
| 4.6. Évaluation de modèle .....                       | 5  |
| 4.7. Représentation des connaissances .....           | 5  |
| 5. Types de méthode du Data Mining .....              | 5  |
| 5.1. Les méthodes descriptives.....                   | 5  |
| 5.2. Les méthodes prédictives (modélisation) .....    | 5  |
| 6. Les tâches du Data Mining .....                    | 6  |
| 6.1. La description.....                              | 6  |
| 6.2. L'estimation.....                                | 6  |
| 6.3. La prédiction.....                               | 6  |
| 6.4. La classification.....                           | 6  |
| 6.5. Le clustering .....                              | 6  |
| 6.6. L'association .....                              | 7  |
| 7. Objectif du Data Mining.....                       | 7  |
| 8. Les domaines du Data Mining.....                   | 8  |
| 8.1. Data Mining dans e-commerce.....                 | 8  |
| 8.2. Data Mining dans ressources humains .....        | 8  |
| 8.3. Data Mining dans les banques et finances.....    | 8  |
| 8.4. Data Mining dans médical et pharmaceutiques..... | 9  |
| 8.5. Data Mining dans aérospatial .....               | 9  |
| 8.6. Data Mining dans génétique humaine .....         | 9  |
| 8.7. Data Mining dans Ingénierie électrique .....     | 10 |
| 8.8. Data Mining dans assurances et santé .....       | 10 |

|                              |   |    |
|------------------------------|---|----|
| 8.9.                         | Data Mining dans Vente, distribution et Marketing ..... | 10 |
| 8.10.                        | Data Mining dans la télécommunication .....             | 10 |
| 9.                           | La prédiction.....                                      | 10 |
| 10.                          | Comparaison entre classification et prédiction.....     | 11 |
| 11.                          | Les techniques de prédiction .....                      | 11 |
| 11.1.                        | La prédiction par arbre de décision .....               | 11 |
| 11.1.1.                      | Avantages .....   | 12 |
| 11.1.2.                      | Inconvenant .....                                       | 12 |
| 11.2.                        | La prédiction par régression linéaire .....             | 13 |
| 11.2.1.                      | Avantages .....   | 13 |
| 11.2.2.                      | Inconvenant .....                                       | 14 |
| 11.3.                        | Prédiction par réseaux de neurones .....                | 14 |
| 11.3.1.                      | Avantages .....   | 14 |
| 11.3.2.                      | Inconvénients.....                                      | 15 |
| 11.4.                        | Machine à vecteurs de support .....                     | 15 |
| 11.4.1.                      | Avantages .....   | 15 |
| 11.4.2.                      | Inconvénients .....                                     | 15 |
| 11.5.                        | Prédiction avec les règles associative.....             | 15 |
| 11.5.1.                      | Avantages .....   | 16 |
| 11.5.2.                      | Inconvénients.....                                      | 16 |
| 12.                          | Conclusion.....   | 16 |
| Chapitre 2 : E-commerce..... |   | 17 |
| 1.                           | Introduction .....                                      | 17 |
| 2.                           | Définition E-commerce .....                             | 17 |
| 3.                           | Historique du commerce électronique.....                | 18 |
| 4.                           | L'évolution d'E-commerce .....                          | 19 |
| 5.                           | Les avantages d'e-commerce .....                        | 20 |
| 6.                           | Les différentes formes du e-commerce .....              | 21 |
| 6.1.                         | Business to business .....                              | 21 |
| 6.2.                         | Business to consumer .....                              | 22 |
| 6.3.                         | Business to Administration.....                         | 22 |
| 6.4.                         | Business to employer.....                               | 22 |
| 6.5.                         | Consumer to consumer .....                              | 23 |
| 7.                           | Les plus grands marchands dans e-commerce.....          | 23 |
| 7.1.                         | Alibaba .....   | 23 |
| 7.2.                         | Amazon .....  | 23 |

|   |   |    |
|---|---|----|
| 7.3.                                    | Flipkart .....  | 24 |
| 7.4.                                    | Jumia .....   | 24 |
| 8.                                      | E-commerce en Algérie.....  | 24 |
| 9.                                      | Quelle que travail utilisent la prédiction Data Mining dans E-Commerce..... | 25 |
| 9.1.                                    | Principe du système d'Amazon .....  | 25 |
| 9.2.                                    | Objectif d'utilisation Data Mining : .....                                  | 26 |
| 10.                                     | Conclusion.....   | 26 |
| Chapitre 3 : conception du système..... |   | 27 |
| 1.                                      | Introduction .....  | 27 |
| 2.                                      | Architecture générale.....  | 27 |
| 3.                                      | Architecture détaillée.....   | 29 |
| 3.1.                                    | Dataset .....   | 29 |
| 3.2.                                    | Prétraitements .....  | 29 |
| 3.2.1.                                  | Données textuelles.....   | 29 |
| 3.2.2.                                  | Données numériques .....  | 30 |
| 3.3.                                    | Entraînement .....  | 31 |
| 3.3.1.                                  | Construction l'Arbre de décision(modèle) .....                              | 31 |
| 4.                                      | Utilisation .....   | 32 |
| 5.                                      | Conclusion.....   | 32 |
| Chapitre 4 : implémentation .....       |   | 33 |
| 1.                                      | Introduction .....  | 33 |
| 2.                                      | Outils utilisés .....   | 33 |
| 2.1.                                    | Logiciel utilisé.....   | 33 |
| 2.2.                                    | Langage utilisé .....   | 35 |
| 2.3.                                    | Bibliothèques utilisées.....  | 35 |
|   | Dans éditeur Google colab .....   | 35 |
|   | Dans éditeur pycharm.....   | 38 |
| 3.                                      | Application .....   | 39 |
| 4.                                      | Conclusion.....   | 40 |
| Conclusion générale .....               |   | 41 |
| Code source .....                       |   | 42 |

## Liste des figures

|  |    |
|--|----|
| <b>Figure 1</b> Architecture du Data Mining .....  | 3  |
| <b>Figure 2</b> Processus du Data Mining .....   | 5  |
| <b>Figure 3</b> Arbre de décision .....  | 12 |
| <b>Figure 4</b> Exemple de régression linéaire .....   | 13 |
| <b>Figure 5</b> Réseaux de neurones .....  | 14 |
| <b>Figure 6</b> Le commerce électronique .....   | 18 |
| <b>Figure 7</b> Le Minitel.....  | 19 |
| <b>Figure 8</b> Chiffre d'e-commerce pour ces dernières années.....                                | 20 |
| <b>Figure 9</b> Différence formes du e-commerce.....   | 21 |
| <b>Figure 10</b> Alibaba logo.....   | 23 |
| <b>Figure 11</b> Amazon logo.....  | 24 |
| <b>Figure 12</b> Flipkart logo.....  | 24 |
| <b>Figure 13</b> Jumia logo.....   | 24 |
| <b>Figure 14</b> Architecture générale de notre système .....                                      | 27 |
| <b>Figure 15</b> diagramme de cas d'utilisation du notre système .....                             | 28 |
| <b>Figure 16</b> diagramme de séquence pour notre système.....                                     | 28 |
| <b>Figure 17</b> Encodage des valeurs textuelles.....  | 30 |
| <b>Figure 18</b> Résultat de la base obtenue .....   | 31 |
| <b>Figure 19</b> Logo du Google Colab.....   | 33 |
| <b>Figure 20</b> Logo du Pycharm .....   | 34 |
| <b>Figure 21</b> Fenêtre du logiciel Pycharm .....   | 34 |
| <b>Figure 22</b> Logo Python.....  | 35 |
| <b>Figure 23</b> Utilisation de la bibliothèque Pandas dans Google colab.....                      | 36 |
| <b>Figure 24</b> Utilisation de la bibliothèque Sklearn dans Google colab.....                     | 36 |
| <b>Figure 25</b> Utilisation de la bibliothèque Sklearn dans Google colab.....                     | 36 |
| <b>Figure 26</b> Utilisation de la bibliothèque Matplotlib dans Google colab .....                 | 37 |
| <b>Figure 27</b> visualisation des caractéristiques en utilisant Matplotlib dans Google colab..... | 37 |
| <b>Figure 28</b> Utilisation de la bibliothèque pickle dans Google colab .....                     | 38 |
| <b>Figure 29</b> Utilisation de la bibliothèque pickle dans Pycharm .....                          | 38 |
| <b>Figure 30</b> fenêtre de notre travail .....  | 39 |
| <b>Figure 31</b> capture du résultat finale .....  | 40 |

## Introduction générale

Avec le développement rapide des nouvelles technologies, internet s'est établi comme un moyen de communication incontournable. Le e-commerce s'est développé dans ce contexte, et s'est imposé comme nouveau moyen de commercialisation.

Dans ces dernières années on trouve les chiffres d'affaires mondiale du e-commerce à gravement sauté qui a dépassé 3.16 milliards de dollar en 2020, grâce à la technologie et les différentes études que ce soit dans le domaine commerciale et informatique.

Avec le développement de l'informatique et l'expansion des stratégies qui mène vers augmentation que ce soit coté finance ou côté social sous objectif de gagné et prédit toute qui concerne les biens pour e-commerce grâce à ces stratégies.

Data Mining(prédiction), le domaine utilisant pour la visualisation et la prédiction dans e-commerce par exemple prédiction pour nombre de vente ...etc.

Parmi les méthodes pour la prédiction dans le e-commerce il y a prédiction avec arbre de décision et autre, qui peut prédire le nombre de ventes et nombre de clients grâce au modèle construit d'après un entraînement avec la base de données qui convient.

Dans ce mémoire, nous allons présenter une méthode pour la prédiction dans e-commerce, qui se base sur :

- **Prétraitement de la base** : une phase faite pour bien organisé les données.
- **Arbre de décision** : l'algorithme de prédiction.
- **Teste** : phase de tests sur un système donnant les résultats.

Notre mémoire se compose de quatre chapitres :

1. Prédiction (Data Mining).
2. E-commerce.
3. Conception du système.
4. Implémentation.

# Chapitre 1 : Prédiction (Data Mining)

## 1. Introduction

Après l'énorme développement des sciences et l'ingénierie qui se base à la prévision des caractéristiques, extraire des événements et les classer et les modéliser.

Avec l'utilisation des ordinateurs et l'apparence des méthodes qui appliquent avec succès à des plusieurs domaines et gère des immenses exemples qu'elle représente des données et l'explorer.

Exploration de données est un processus qui utilise une variété d'outils d'analyse pour construire des modèles et des relations dans les données qui peuvent être utilisées pour faire des prédictions valides.

## 2. Définition du Data Mining

- Data Mining (en français la fouille de donnée), on peut même utiliser l'expression extraction de connaissances à partir de données, est apparu dans les années 1990. Gregory Piatetsky-Shapiro a inventé le « KNOWLEDGE Discovery in Data bases », ce terme est devenu populaire en apprentissage communautaire. Quant au terme « Datamining », il apparaît en 1991 et plus utilisé dans les milieux d'affaires et de presse. Aujourd'hui, les termes datamining et Knowledge Discovery sont tous deux utilisés. [1]
- Data Mining est un ensemble de méthodes qui permettent d'extraire de l'information (de façon semi-automatique) de grandes bases de données. Les informations découvertes sont des règles, des associations, des tendances inconnues (non fixées a priori), des structures particulières. Elles doivent aider à prendre des décisions. [1]
- Le Data Mining est un domaine interdisciplinaire utilisant dans le même temps des techniques d'apprentissage automatique, de reconnaissance des formes, des statistiques, des bases de données et de visualisation pour déterminer les manières d'extraction des informations de très grandes bases de données.[1]

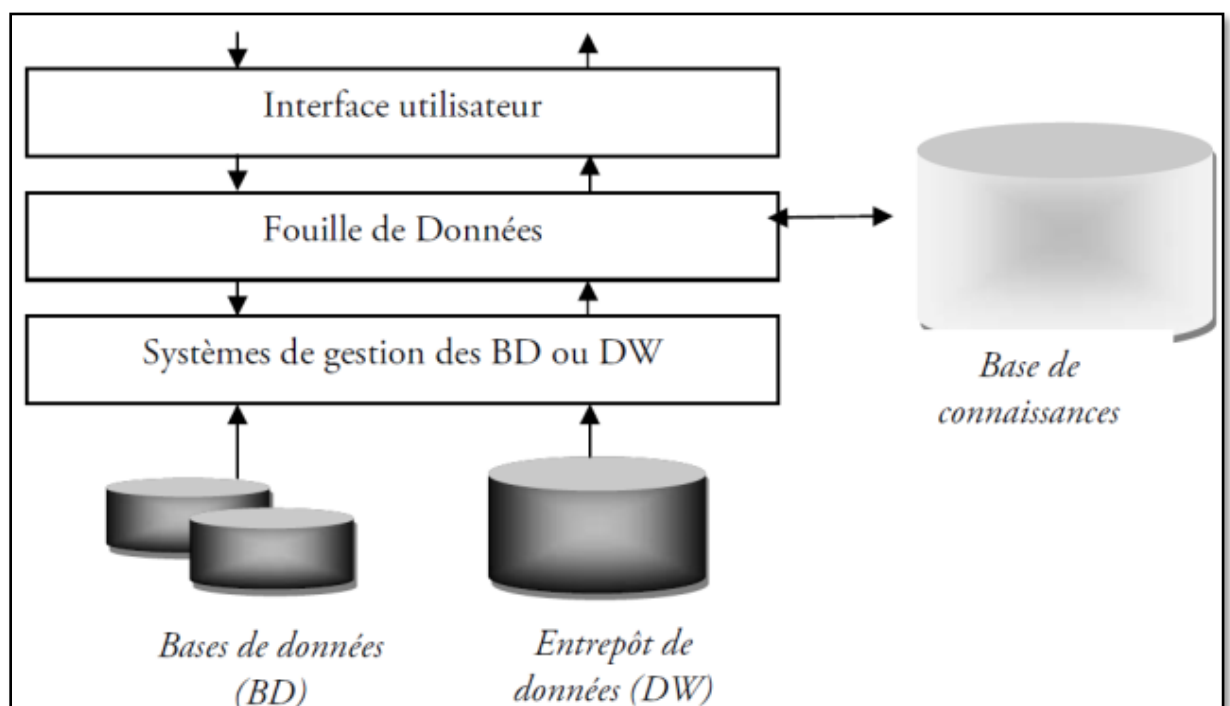
## 3. Architecture du Data Mining

Un système d'ECBD s'articule autour des composantes suivantes :

Une ou plusieurs bases de données et leurs systèmes de gestion, pour le traitement des masses de données volumineuses.



Une base de connaissances qui permet à la fois la gestion des connaissances et la résolution des problèmes liés au domaine des données. Le système utilise une base de connaissances qui est enrichie grâce aux nouvelles connaissances inférées par le système. Un système Data Mining pouvant s'appuyer sur des techniques symboliques comme l'extraction des règles d'association, la classification par treillis de Galois ou l'induction par des arbres de décision. Et une interface se chargeant des interactions avec l'analyste et de la visualisation des résultats. L'analyste et l'expert sont chargés de guider les recherches et de valider les connaissances extraites. [3]



**Figure 1** Architecture du Data Mining

#### 4. Processus du Data Mining

Le processus du Data Mining est divisé en deux parties, à savoir le prétraitement des données et l'exploration de données. Le prétraitement des données implique le nettoyage des données, l'intégration des données, la réduction des données et la transformation des données. La partie d'exploration de données effectue l'exploration de données, l'évaluation des modèles et la représentation des connaissances des données.

#### **4.1. Nettoyage des données**

Le nettoyage des données est la première étape du Data Mining. Fondamentalement, cette étape implique la suppression des données bruyantes ou incomplètes de la collection. De nombreuses méthodes qui nettoient généralement les données par elles-mêmes sont disponibles, mais elles ne sont pas robustes. [2]

#### **4.2. Intégration de données**

Lorsque plusieurs sources de données hétérogènes telles que des bases de données, des cubes de données ou des fichiers sont combinées pour l'analyse, ce processus est appelé intégration de données. Cela peut aider à améliorer la précision et la vitesse du processus d'exploration de données. Différentes bases de données ont différentes conventions de dénomination des variables, en provoquant des redondances dans les bases de données. Un nettoyage supplémentaire des données peut être effectué pour supprimer les redondances et les incohérences de l'intégration des données sans affecter la fiabilité des données. [2]

#### **4.3. Réduction de donnée**

Cette technique est appliquée pour obtenir des données pertinentes pour l'analyse à partir de la collecte de données. La taille de la représentation est beaucoup plus petite en volume tout en conservant l'intégrité. La réduction des données est effectuée à l'aide de méthodes telles que Naïve Bayes, arbres de décision, réseau neuronal, etc. [2]

#### **4.4. Transformation des données**

Dans ce processus, les données sont transformées en une forme adaptée au processus d'exploration de données. Les données sont consolidées afin que le processus d'extraction soit plus efficace et les modèles plus faciles à comprendre. La transformation des données implique le mapping de données et le processus de génération de code. [2]

#### **4.5. Data Mining**

Le Data Mining est un processus permettant d'identifier des modèles et des connaissances intéressants à partir d'une grande quantité de données. Dans ces étapes, des modèles intelligents sont appliqués pour extraire les modèles de données. Les données sont représentées sous forme de modèles et les modèles sont structurés à l'aide de techniques de classification et de regroupement. [2]

#### 4.6. Évaluation de modèle

Cette étape consiste à identifier des modèles intéressants représentant les connaissances sur la base de mesures d'intérêt. Des méthodes de synthèse et de visualisation des données sont utilisées pour rendre les données compréhensibles par l'utilisateur. [2]

#### 4.7. Représentation des connaissances

La représentation des connaissances est une étape où les outils de visualisation des données et de représentation des connaissances sont utilisés pour représenter les données extraites. Les données sont visualisées sous forme de rapports, de tableaux, etc. [2]

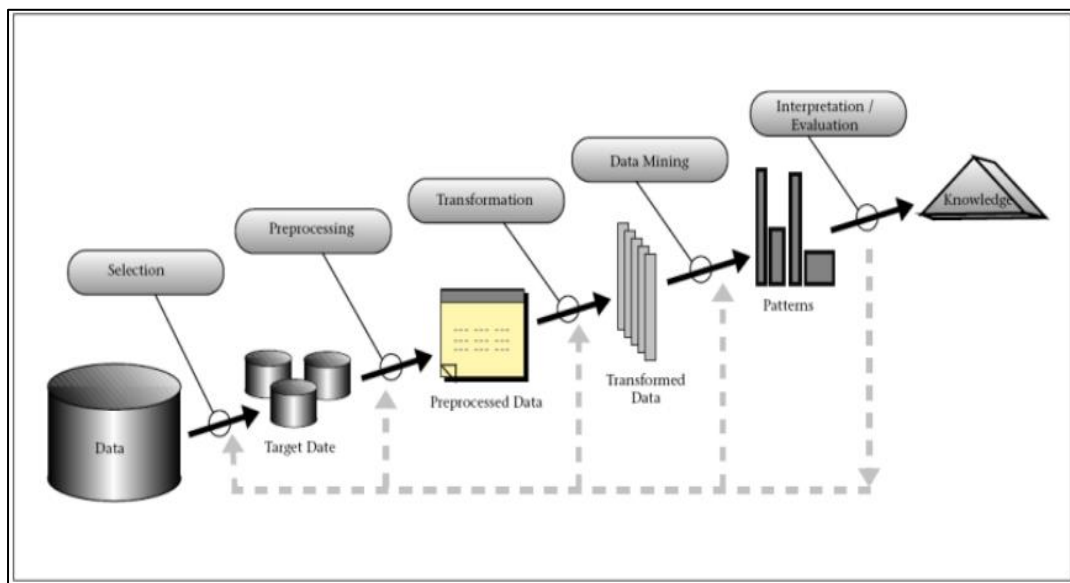


Figure 2 Processus du Data Mining

### 5. Types de méthode du Data Mining

#### 5.1. Les méthodes descriptives

- Visent à mettre en évidence des informations présentes mais cachées par le volume des données (c'est le cas des segmentations de clientèle et des recherches d'associations de produits sur les tickets de caisse).
- Réduisent, résument, synthétisent les données.
- Il n'y a pas de variable à expliquer. [3]

#### 5.2. Les méthodes prédictives (modélisation)

- Visent à extrapoler de nouvelles informations à partir des informations présentes (c'est le cas du scoring).

- Expliquez les données.
- Il y a une variable à expliquer. [3]

## **6. Les tâches du Data Mining**

### **6.1. La description :**

Parfois, les chercheurs et les analystes essaient simplement de trouver des façons de décrire des tendances cachées dans les données. Les descriptions des modèles et des tendances servent à expliquer ou vérifier un fait. Par exemple : « ceux qui ont le plus de diplômes sont les plus susceptibles d'avoir un poste à responsabilité. ». [3]

### **6.2. L'estimation**

L'estimation est similaire à la classification, sauf que la variable cible est numérique plutôt que catégorique. Les modèles sont construits en utilisant des données, qui fournissent la valeur de la variable cible, ainsi que les « prédicteurs ». Par exemple : « l'estimation de la pression artérielle d'un patient d'hôpital, basée sur son âge, son sexe, son indice de masse corporelle, et le taux de sodium. La relation entre la pression artérielle et le prédicteur variable de l'ensemble de formation nous donnerait un modèle d'estimation. Nous pouvons alors appliquer ce modèle à de nouveaux cas. [3]

### **6.3. La prédiction**

La prédiction est semblable à la classification et l'estimation, sauf que pour la prévision, les résultats se situent dans l'avenir. Exemples de tâches de prévision appliquée au marketing : « Prédire le prix d'un stock de trois mois dans le futur » [3]

### **6.4. La classification**

Supposons qu'un décideur veuille classer ses employés par tranches de revenu, ou n'importe quelle autre caractéristique associée à cette personne, comme l'âge, le sexe et la profession. Cette tâche est une tâche de classification. [3]

### **6.5. Le clustering**

Le Clustering désigne le regroupement des données, des observations ou des cas dans des classes d'objets similaires. Un cluster maximise la similarité des objets de même cluster et minimise la similarité des objets de cluster différents. En effet, il n'y a pas de variable cible pour le clustering. La tâche de clustering ne cherche pas à classer, estimer, ou prédire la valeur

d'une variable cible. Mais plutôt à segmenter l'ensemble des données en sous-groupes relativement homogènes à l'aide de mesures de distances. [3]

### **6.6. L'association**

La recherche de règles d'association est la tâche la plus intéressante du Data Mining. C'est également celle qui est la plus répandue dans le monde des affaires, notamment en marketing pour l'analyse du panier de consommation. La recherche de règles d'association cherche à découvrir les règles de quantification ou de relation entre deux ou plusieurs attributs. Les règles d'association sont de la forme « Si antécédent, puis conséquent », avec une mesure de confiance associée à la règle. La recherche de règles d'association dans une grande base de données permet de découvrir des règles cachées utiles pour la prise de décision.

Exemple de règle célèbre : lorsqu'un homme achète des couches pour bébés, il achète 2 packs de bières dans 65% des cas. Il serait alors intéressant pour le gestionnaire d'adapter ses promotions à ces nouvelles règles. [3]

## **7. Objectif du Data Mining**

Les objectifs sont par définition déterminés en fonction des données existantes et des problématiques de l'entreprise. Il est donc impossible de lister de manière exhaustive l'ensemble des applications du Data Mining, néanmoins en voici quelques-unes à titre d'exemple :

- Identifier les clients fidèles, les clients rentables, ceux à fort potentiel et inversement
- Trouver des leviers de fidélisation et modéliser de manière prédictive les départs
- Accroître la fidélité en étant capable de proposer la bonne offre, au bon client, au bon moment, par le bon canal
- Comprendre et anticiper les comportements d'achats
- Augmenter les recrutements et diminuer les coûts des campagnes de recrutement
- Développer des revenus complémentaires avec la proposition de nouvelles offres de produits ou services parfaitement appropriées à chaque segment de clients
- Augmenter le retour sur investissement des opérations marketing par un meilleur ciblage et une optimisation de l'utilisation des canaux (par exemple, accroissement des taux de réponse de 20% à 200% tout en réduisant les coûts)

- Réduire les risques d'impayés par une meilleure identification du risque tout au long du cycle de vie du client
- Optimiser l'animation de votre programme de fidélité par une meilleure compréhension des attentes et des enjeux par cible de clients
- Connaître la manière dont les facteurs démographiques ou autres influent sur les ventes
- Calculer le coût d'acquisition d'un nouveau client
- Savoir qui achète quoi [7]

## **8. Les domaines du Data Mining**

Le Datamining est une approche d'analyse de donnée, adaptée et utilisée dans un large nombre de domaines d'activités.

### **8.1. Data Mining dans e-commerce**

En e-commerce, le Data Mining est une démarche fondamentale tant en gestion de la relation client qu'en stratégie commerciale. Il s'agit de collecter toutes les informations de consultation et d'achat des clients pour comprendre comment ils se comportent sur le site. Les données servent à adapter les services proposés mais aussi à cibler les consommateurs selon leurs préférences, leurs précédentes navigations, les liens les plus consultés. On peut ainsi parvenir à une segmentation très pointue de la clientèle. Le Data Mining peut être en temps réel, c'est-à-dire que selon la façon dont navigue un consommateur, on va pouvoir personnaliser l'offre proposée. Cette pratique a fait naître un nouveau métier, celui de Datamining, chargé de concevoir les moyens de collecter l'information et de l'analyser. [9]

### **8.2. Data Mining dans ressources humaines**

Le Datamining est également utilisé dans les ressources humaines (RH) de certains ministères pour identifier les caractéristiques de leurs employés les plus performants. L'information obtenue (comme les universités fréquentées par des employés potentiels) peut contribuer aux efforts de recrutement des ressources humaines.[7]

Ces dernières années, l'exploration de données a été largement utilisée dans les domaines de la science et de l'ingénierie, tels que la bio-informatique, la génétique, la médecine, l'éducation et l'énergie électrique.

### **8.3. Data Mining dans les banques et finances**

- Détection d'usage frauduleux de cartes bancaires.

- Gestion du risque lié à l'attribution de prêts bancaires par le Scoring.
- Découverte de relations cachées entre les indicateurs financiers.
- Détection de règles de comportement boursier par l'analyse des données du marché. [9]

#### **8.4. Data Mining dans médical et pharmaceutiques**

- Diagnostic assisté par ordinateur (CAD) par l'apprentissage de systèmes experts.
- Explication ou prédiction de la réponse d'un patient à un traitement.
- Identification des thérapies à succès (combinaison de prescriptions).
- Étude des corrélations entre le dosage dans un traitement et l'apparition d'effets secondaires [1]

#### **8.5. Data Mining dans aérospatial**

Le Datamining est également intégré aux données spatiales. L'objectif final est de trouver des modèles dans les données relatives à la géographie. Jusqu'à présent, l'exploration de données et de systèmes d'information géographiques ont existé en tant que deux technologies distinctes, chacune avec ses propres méthodes. L'immense explosion de données géoréférencées occasionnée par l'évolution de l'informatique, la cartographie numérique, la télédétection et la diffusion mondiale des systèmes d'information géographiques mettent l'accent sur l'importance de développer une analyse et une modélisation géographique plus fines. [1]

#### **8.6. Data Mining dans génétique humaine**

Dans l'étude de la génétique humaine, le Datamining permet de répondre à l'objectif important de comprendre la relation de correspondance entre l'ADN et les maladies. En effet, il vise à savoir comment les changements dans la séquence d'ADN d'un individu affectent les risques de développer des maladies courantes telles que le cancer, qui est d'une grande importance à l'amélioration des méthodes de diagnostic, la prévention et le traitement de ces maladies. Le Data Mining peut contribuer de manière significative et avec succès à l'explication ou la prédiction de phénomènes complexes dans les domaines médical et pharmaceutique. [7]

### **8.7. Data Mining dans Ingénierie électrique :**

Dans le domaine de l'ingénierie électrique, les Datamining ont été largement utilisés pour la surveillance de l'état du matériel électrique à haute tension. Le but de surveillance de l'état est d'obtenir de précieuses informations par exemple, sur l'état de l'isolation (ou d'autres importantes des paramètres de sécurité). [1]

### **8.8. Data Mining dans assurances et santé**

- Découverte d'associations des demandes de remboursements
- Identification de clients potentiels de nouvelles polices d'assurances.
- Détection d'association de comportements pour la découverte de clients à risque.
- Détection de comportement frauduleux [7]

### **8.9. Data Mining dans Vente, distribution et Marketing**

- La gestion de la relation client (GRC ou CRM) consiste en l'ensemble des activités visant à cibler, attirer et conserver les "bons" clients.
- Détection d'associations de comportements d'achat.
- Découverte des caractéristiques de la clientèle.
- Prédiction de probabilité de réponse aux campagnes de mailing. [9]

### **8.10. Data Mining dans la télécommunication**

- Analyse des achats de services de télécommunications.
- Prédiction de modèles d'appels téléphoniques.
- Gestion des ressources et du trafic réseau.
- Automatisation de la gestion du réseau et de la maintenance en utilisant l'intelligence artificielle pour diagnostiquer et réparer les problèmes de transmission du réseau, etc. [1]

## **9. La prédiction**

La prédiction est qu'on suppose que le responsable marketing ait besoin de prédire combien un client donné dépensera lors d'une vente dans son entreprise. Dans cet exemple, nous prenons



la peine de prédire une valeur numérique. Par conséquent, la tâche d'analyse des données est un exemple de prédiction numérique. Dans ce cas, un modèle ou un prédicteur sera construit qui prédit une fonction à valeur continue ou une valeur ordonnée.

L'analyse de régression est une méthodologie statistique qui est le plus souvent utilisée pour la prédiction numérique.

La prédiction s'applique dans plusieurs domaines comme l'éducation e-commerce super markette etc. [4]

## 10. Comparaison entre classification et prédiction

- Précision - La précision du classificateur fait référence à la capacité du classificateur. Il prédit correctement l'étiquette de classe et la précision du prédicteur se réfère à la capacité d'un prédicteur donné à deviner la valeur de l'attribut prédit pour une nouvelle donnée.
- Vitesse - Il s'agit du coût de calcul lié à la génération et à l'utilisation du classificateur ou du prédicteur.
- Robustesse - Cela fait référence à la capacité du classificateur ou du prédicteur à faire des prédictions correctes à partir de données bruyantes.
- Évolutivité - L'évolutivité fait référence à la capacité de construire efficacement le classificateur ou le prédicteur, étant donné une grande quantité de données.
- Interprétabilité - Il se réfère à la mesure dans laquelle le classificateur ou le prédicteur comprend. [4]

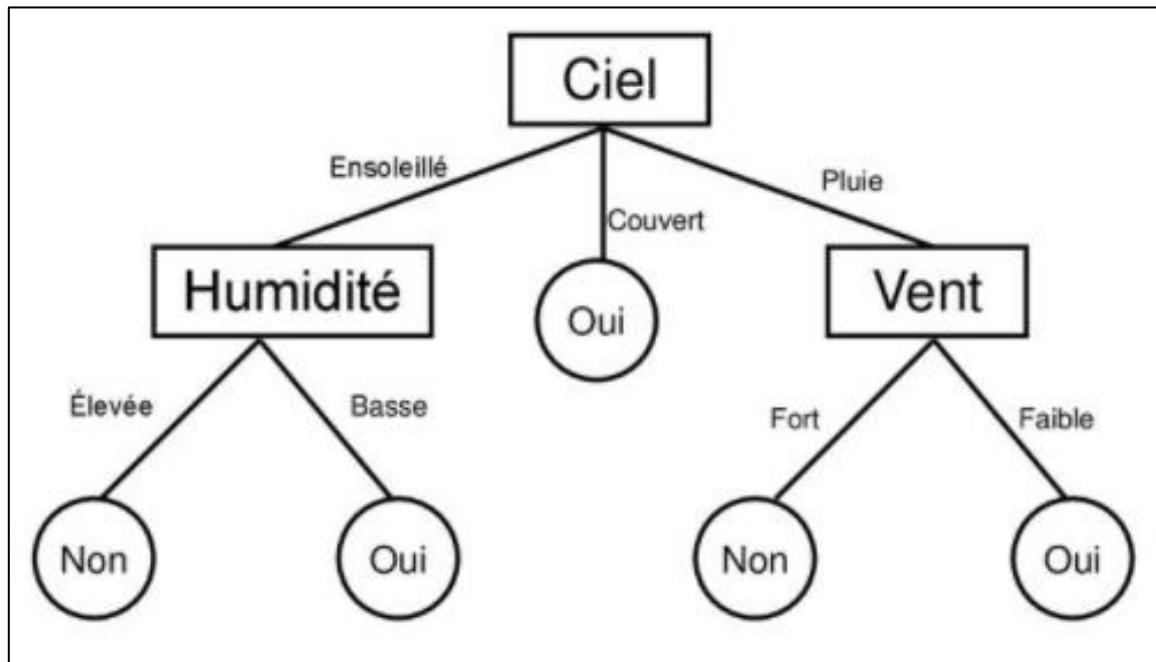
## 11. Les techniques de prédiction

### 11.1. La prédiction par arbre de décision

Est un graphe non orienté, acyclique et connexe. L'ensemble des nœuds se divise en trois catégories :

- Nœud *racine* (l'accès à l'arbre se fait par ce nœud)
- Nœuds *internes* : les nœuds qui ont des descendants
- Nœuds *terminaux* (ou *feuilles*) : nœuds qui n'ont pas de descendant

Ils emploient une représentation hiérarchique de la structure des données sous forme des séquences de décisions (tests) en vue de la prédiction d'un résultat ou d'une classe. Chaque individu (ou observation), qui doit être attribué(e) à une classe, est décrit(e) par un ensemble de variables qui sont testées dans les nœuds de l'arbre. Les tests s'effectuent dans les nœuds internes et les décisions sont prise dans les nœuds feuille. [5]



**Figure 3** Arbre de décision

#### 11.1.1. Avantages

- Les arbres de décision sont capables de produire des règles compréhensibles.
- Les arbres de décision effectuent la classification sans exiger beaucoup de calcul
- Les arbres de décision sont en mesure de manipuler à la fois les variables continues et catégorielles[8]

#### 11.1.2. Inconvénient

- Manque de performance dans le cas de plusieurs classes, les arbres deviennent très complexes et ne sont pas nécessairement optimaux.
- Demande beaucoup de temps de calcul lors de la construction.
- Moins bonnes performances concernant les prédictions portant sur des valeurs numériques [8]

## 11.2. La prédiction par régression linéaire

La régression est une mesure statistique utilisée en finance, mais aussi dans d'autres disciplines scientifiques pour tenter de déterminer la force de la relation entre une variable dépendante (habituellement désignée par Y) et une série d'autres variables changeantes (appelées variables indépendantes).

La forme générale de la régression linéaire est la suivante :  $Y = a \cdot X + b + \text{epsilon}$  avec a et b deux constantes. Y est la variable à prédire, X la variable utilisée pour prédire, a est la pente de la régression et b est l'*intercept*, c'est-à-dire la valeur de Y lorsque X est égal à zéro.

Dans le cas d'une régression linéaire multiple, il y a plusieurs variables changeantes et on écrit  $Y = a \cdot X_1 + b \cdot X_2 + \dots + z + \text{epsilon}$ .

Epsilon, supposé très petit, correspond au terme d'erreur de la régression. Il donne la possibilité de ne pas être tout à fait exact dans l'estimation. Son espérance mathématique est égale à zéro. [6]

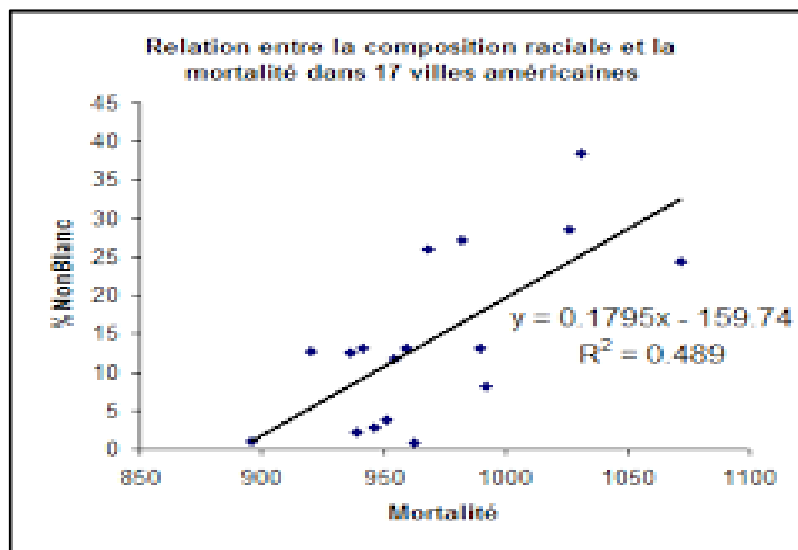


Figure 4 Exemple de régression linéaire

### 11.2.1. Avantages

L'avantage de l'algorithme de régression linéaire est sa simplicité d'interprétation et sa facilité de calcul. Par contre, le data scientist veillera à bien vérifier qu'il existe une relation linéaire entre les paramètres d'entrée et celle de sortie. [8]

### 11.2.2. Inconvenant

L'algorithme est très sensible aux valeurs aberrantes des données d'apprentissage d'où la nécessité de bien préparer ses données dès le départ. Il existe des méthodes dites de régularisation pour pallier ce problème. Les méthodes de régularisation permettent de pénaliser les valeurs trop grandes des coefficients  $a_i$  et  $b$ . [8]

### 11.3. Prédiction par réseaux de neurones

Un réseau de neurones est un modèle de calcul dont le fonctionnement vise à simuler le fonctionnement des neurones biologiques, il est constitué d'un grand nombre d'unités (neurones) ayant chacune une petite mémoire locale et interconnectées par des canaux de communication qui transportent des données numériques. Ces unités peuvent uniquement agir sur leurs données locales et sur les entrées qu'elles reçoivent par leurs connexions. Les réseaux de neurones sont capables de prédire de nouvelles observations (sur des variables spécifiques) à partir d'autres observations (soit les mêmes ou d'autres variables) après avoir exécuté un processus d'apprentissage sur des données existantes. La phase d'apprentissage d'un réseau de neurones est un processus itératif permettant de régler les poids du réseau pour optimiser la prédiction des échantillons de données sur lesquelles l'apprentissage a été fait. Après la phase d'apprentissage, le réseau de neurones devient capable de généraliser. [4]

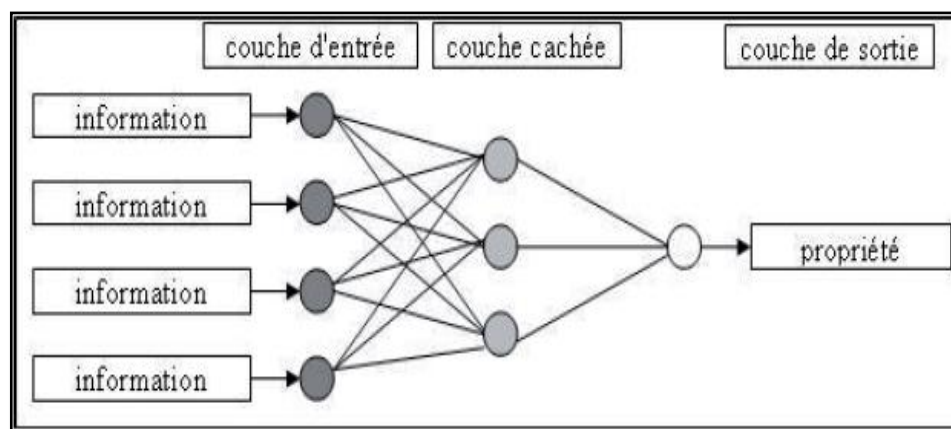


Figure 5 Réseaux de neurones

#### 11.3.1. Avantages

Les réseaux de neurones sont théoriquement capables d'approximer n'importe quelle fonction continue et ainsi le chercheur n'a pas besoin d'avoir aucune hypothèse du modèle sous-jacent. [7]

### **11.3.2. Inconvénients**

Généralement les réseaux de neurones ne sont pas souvent utilisés dans les tâches du Data Mining parce qu'ils produisent des modèles souvent incompréhensibles et demandent un long temps d'apprentissage. [7]

### **11.4. Machine à vecteurs de support**

(SVM) sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination et de régression, L'algorithme se base principalement sur 3 astuces pour obtenir de très bonnes performances tant en qualité de prédiction qu'en complexité de calcul.

- On cherche l'hyperplan comme solution d'un problème d'optimisation sous-contraintes. La fonction à optimiser intègre un terme de qualité de prédiction et un terme de complexité du modèle.
- Le passage à la recherche de surfaces séparatrices non linéaires est introduit en utilisant un noyau kernel qui code une transformation non linéaire des données.
- Numériquement, toutes les équations s'obtiennent en fonction de certains produits scalaires utilisant le noyau et certains points de la base de données (ce sont les Support Vectors). [4]

#### **11.4.1. Avantages**

- Leur capacité à modéliser les phénomènes non linéaires.
- La précision dans certains cas de leurs prédictions. [7]

#### **11.4.2. Inconvénients :**

- Temps de calcul un peu long.
- Le risque de sur-apprentissage.
- Les logiciels implémentant les SVM sont rares. [7]

### **11.5. Prédiction avec les règles associatives**

Dans le domaine du Data Mining la recherche des règles d'association est une méthode populaire étudiée d'une manière approfondie dont le but est de découvrir des relations ayant un intérêt pour le statisticien entre deux ou plusieurs variables stockées dans de très importantes bases de données. Piatetsky-Shapiro présentent des règles d'association extrêmement fortes découvertes dans des bases de données en utilisant différentes mesures

d'intérêt. En se basant sur le concept de relations fortes, Rakesh Agrawal et son équipe présente des règles d'association dont le but est de découvrir des similitudes entre des produits dans des données saisies sur une grande échelle dans les systèmes informatiques des points de ventes des chaînes de supermarchés. Par exemple, une règle découverte dans les données de ventes dans un supermarché pourrait indiquer qu'un client achetant des oignons et des pommes de terre simultanément serait susceptible d'acheter un hamburger. Une telle information peut être utilisée comme base pour prendre des décisions marketing telles que par exemple des promotions ou des emplacements bien choisis pour les produits associés. En plus des exemples ci-dessus concernant le panier de la ménagère, les règles d'association sont employées aujourd'hui dans plusieurs domaines incluant celui de la fouille du web, de la détection d'intrusion et de la bio-informatique. [15]

#### **11.5.1. Avantages**

- Résultats clairs : règles faciles à interpréter.
- Simplicité de la méthode et des calculs (calculs élémentaires des fréquences d'apparition).
- Aucune hypothèse préalable (Apprentissage non supervisé).
- Méthode facile à adopter aux séries temporelles. [15]

#### **11.5.2. Inconvénients**

- La méthode est coûteuse en temps de calcul.
- Méthode inefficace pour les articles rares [15]

## **12. Conclusion**

Le Data Mining est l'extraction d'informations prédictives cachés dans de grandes bases de données. Cette technologie repose sur une diversité de techniques (intelligence artificielle, statistiques, théorie de l'information, génie logiciel, bases de données, ...) qui requièrent des compétences variées et de haut niveau. C'est ce qui rend le Data Mining la technologie la plus importante dans plusieurs domaines

## **Chapitre 2 : E-commerce**

### **1. Introduction**

De nos jours, de moins en moins de consommateurs veulent voyager pour faire leur emplette, grâce à Internet ces consommateurs peuvent faire tous leurs achats sans quitter leurs domiciles. Ce type d'achats s'intitule le commerce électronique. Celui-ci est un moyen peu coûteux qui relie des ordinateurs pour effectuer des opérations commerciales en épargnant temps et argent. Le commerce électronique s'avère utile pour les consommateurs mais il est aussi une composante importante dans les activités quotidiennes des commerçants. Il leur permet de contacter leurs clients et fournisseurs, de faire leurs publicités et même d'organiser la facturation et la distribution de leurs produits et services de façon efficace. De plus, il diminue les coûts d'exploitation de l'entreprise.

### **2. Définition E-commerce**

(Commerce électronique) d'après le terme e-commerce, on peut prendre une idée que ce terme représente les actions d'achat et autres sur internet.

L'achat peut se réaliser au travers de différents canaux et supports : ordinateurs, smartphones, tablettes, consoles, TV, avec l'évolution du e-commerce, la vente par correspondance est dorénavant devenue la vente à distance.

Le commerce électronique est la vente ou l'achat de marchandises ou de services effectués sur des réseaux informatiques au moyen de méthodes spécifiquement conçues pour la réception ou la passation de commandes. Même si les marchandises ou les services sont commandés par voie électronique, le paiement et la livraison n'ont pas besoin d'avoir lieu en ligne. Une transaction commerciale électronique peut se faire entre entreprises, ménages, particuliers, gouvernements et autres organisations publiques ou privées. Ces transactions électroniques incluent les commandes passées sur le Web, l'extranet ou l'échange électronique de données.

**[10]**



**Figure 6** Le commerce électronique

### **3. Historique du commerce électronique**

L'histoire du e-commerce est comme ce suit :

- L'e-commerce est né en France dans les années 90 grâce au minitel
- C'est aux USA que la première transaction à distance par carte bancaire s'est faite en 1994 par Phil Brandenberger qui a acheté un album de Sting à 12,48\$.
- L'arrivée du haut-débit dans les années 2000 a marqué la fulgurante ascension de l'e-commerce.
- Des acteurs américains ont été propulsés comme start-up : Amazon, AOL, Ebay.
- Dans les années 2010, avec l'émergence des smartphones, le m-commerce a fait son apparition. Les pays émergents essentiellement connectés par le mobile développent ce marché.
- La généralisation des technologies (datacenters, applications informatiques, sites web...) et une harmonisation des règles internationales permet aujourd'hui d'acheter des produits ou services à l'autre bout de la planète. Les échanges commerciaux sont plus simples et l'e-commerce ne connaît pas de frontières à son expansion.
- A ce jour, il est donc possible pour des particuliers comme pour des professionnels de vendre ou acheter des produits et services en ligne. **[10]**





**Figure 7** Le Minitel

#### **4. L'évolution d'E-commerce**

Après 25 ans, le chiffre d'affaires du e-commerce mondial ne cesse d'augmenter, en 2017, environ 2,4 milliards de personnes faisaient des achats en ligne. Au cours des douze mois suivants, le pourcentage d'internautes ayant effectué au moins un achat en ligne au cours des 12 derniers mois est passé à 93% aux États-Unis, 97% au Royaume-Uni et 92% en Chine. Le nombre total d'utilisateurs du commerce électronique dans le monde est passé à 2,7 milliards cette année-là.

Les statistiques montrent que 2019 a connu la croissance la plus impressionnante d'une année sur l'autre, le nombre de personnes faisant des achats en ligne ayant augmenté de 15% pour atteindre 3,16 milliards. La forte tendance à la hausse s'est poursuivie en 2020, le nombre d'utilisateurs de commerce électronique augmentant de 350 millions en un an.

De nombreuses plateformes de vente au détail ont connu une augmentation du trafic sans précédent l'année dernière, dépassant même la saison des vacances. Les statistiques montrent que les revenus mondiaux du commerce électronique devraient atteindre plus de 2,7 milliards de dollars en 2021. [10]



**Figure 8** Chiffre d'e-commerce pour ces dernières années

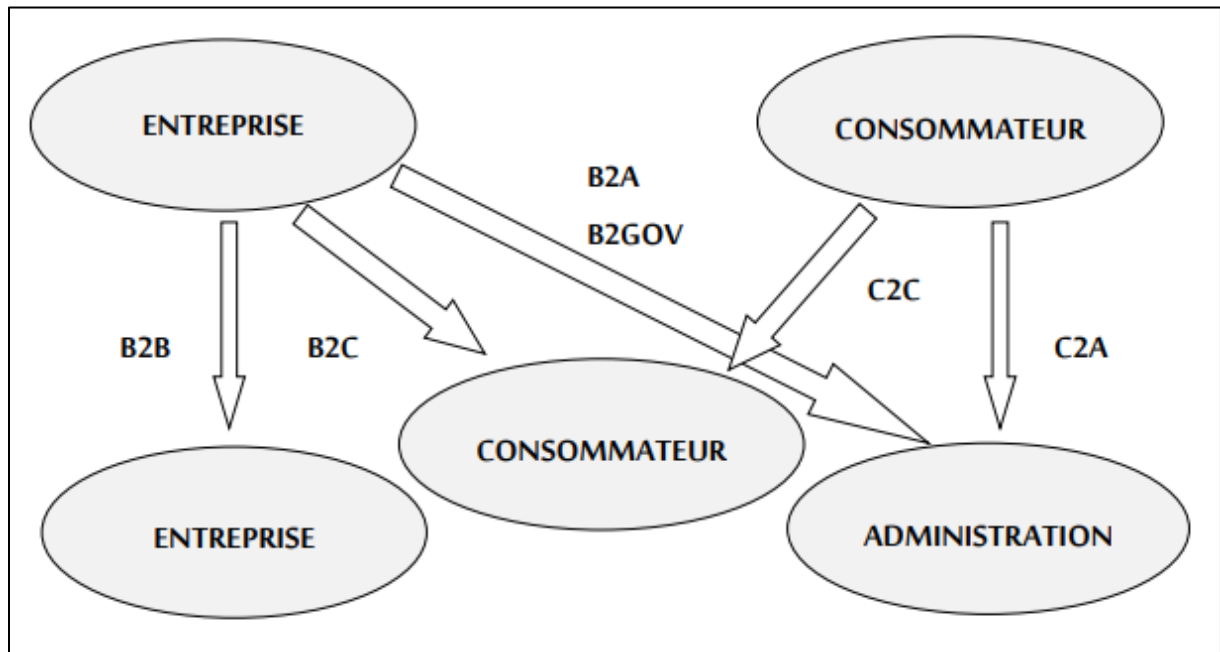
## 5. Les avantages d'e-commerce

Le commerce électronique permet de :

- D'étendre sa zone de chalandise
- De toucher une cible plus large
- D'ajouter un canal de distribution
- De vendre 24h/24
- De présenter ses produits de manière différente
- De diminuer ses coûts fixes (et donc d'améliorer sa marge)
- De profiter davantage des technologies liées à Internet
- De mieux cerner son retour sur investissement
- De développer sa relation client
- De faciliter l'achat
- De récolter différentes données sur ses clients
- De faire un repérage des produits avant d'acheter en magasin
- De trouver le meilleur produit[12]

## 6. Les différentes formes du e-commerce

Le commerce électronique regroupe tous les échanges et toutes les transactions d'une entreprise peuvent être amenées à faire au travers d'un média électronique ou d'un réseau. Cela peut bien sur prendre différentes formes



**Figure 9** Différence formes du e-commerce

### 6.1. Business to business

Désigne le commerce électronique réalisé directement entre les sociétés basées sur l'utilisation d'un support numérique pour les échanges d'informations.

La relation entre l'entreprise et son fournisseur peut se faire maintenant sans intermédiaire.

Le commerce électronique entre des entreprises telles qu'un fabricant et un grossiste ou un grossiste et un détaillant. C'est l'échange de produits, de services ou d'informations entre des entreprises plutôt qu'entre des entreprises et des consommateurs finals. Les transactions inter-entreprises représentent 9 % du commerce électronique mondial. Selon les recherches effectuées par l'International Data Corporation (IDC) établie aux États-Unis, on n'estime que le commerce électronique interentreprises au niveau mondial, notamment entre grossiste et distributeurs, représentait 12 400 milliards de dollars EU à la fin de 2012. S'il continue de se développer à un rythme aussi rapide sur les marchés développés, le commerce électronique

interentreprises et entre entreprises et consommateurs finals représentera 5 % du total des transactions interentreprises et des ventes au détail en 2017. [9] [10]

## **6.2. Business to consumer**

Désigne le commerce électronique à destination des consommateurs. C'est la forme la plus courante et la plus connue. On parle des boutiques en ligne des entreprises. Bien que le commerce électronique entre entreprises et consommateurs finals fasse l'objet d'une grande attention, les transactions interentreprises excèdent de loin les transactions entre entreprises et consommateurs. Selon l'IDC (international data corporation), les transactions mondiales entre entreprises et consommateurs étaient estimées à 1 200 milliards de dollars EU à la fin de 2012, soit 10 fois moins que les transactions interentreprises. Dans le commerce électronique entre entreprises et consommateurs, les entreprises vendent leurs produits au grand public grâce généralement à des catalogues utilisant des logiciels de panier d'achats.

Ce commerce ne représentait encore qu'une faible part de l'ensemble du commerce électronique, mais il continue de se développer. Le commerce électronique entre entreprises et consommateurs est le plus répandu en Norvège, au Danemark, en Suède, au Royaume-Uni et aux Etats Unis et porte principalement sur les produits informatiques, les vêtements et les produits numériques. Malgré la faible valeur de ses transactions, c'est lui qui a fait l'objet de la plus grande attention, en partie parce que les décideurs se sont beaucoup préoccupés de questions telles que la confiance des consommateurs et la protection des données. [9] [10]

## **6.3. Business to Administration**

Paiement d'une transaction électronique entre une entreprise et une administration.

Le commerce entreprise-gouvernement est généralement défini comme le commerce électronique entre les entreprises et le secteur public. Il désigne l'utilisation d'internet pour les marchés publics, les procédures de licence et les autres opérations liées aux pouvoirs publics. Dans ce type de commerce, le secteur public joue généralement le rôle pilote en établissant le commerce électronique afin de rendre plus efficace son système de passation des marchés. La taille du marché du commerce électronique total, car les systèmes de marchés publics électroniques restent peu développés. [9] [10]

## **6.4. Business to employer**

Est l'ensemble des échanges, services et produits qu'une entreprise destine à ses propres salariés, l'internet est le principal support des applications business to employer, s'agit aussi de

la mise à disposition de formulaires à leur intention pour la gestion de leur carrière, de leur congé ou de leur relation avec le comité d'entreprise. [9] [10]

### **6.5. Consumer to consumer**

Là, on se passe de l'entreprise, on parle des sites qui mettent en relation les consommateurs entre eux. On peut citer les sites d'annonces, eBay.

Dans ce genre, le site joue un rôle crucial (il classe les offres, génère du trafic, fournit des garanties commerciales) et la vente elle-même reste effectuée entre particuliers (du particulier à particulier). [9] [10]

## **7. Les plus grands marchands dans e-commerce**

### **7.1. Alibaba**

C'est un site de vente en ligne chinois, qui attire près de 1.8 milliards de visites par mois et notamment grâce à lui le marché chinois du E-commerce génère beaucoup de bénéfices, en 2017 il encaisse 6.4 milliards de dollars de bénéfices. [14]



**Figure 10** Alibaba logo

### **7.2. Amazon**

C'est une entreprise de commerce électronique américaine, c'est l'un des sites les plus visités dans le monde avec des chiffres d'affaires qui ne cessent d'accroître, en 2017 il encaisse un bénéfice de 3.3 milliards de dollars. [14]



**Figure 11** Amazon logo

### 7.3. Flipkart

C'est le leader indien du E-commerce avec environ 2.5 milliards de dollars de chiffres d'affaires en 2017, le site dépasse légèrement les 150 millions de visites mensuels. [14]



**Figure 12** Flipkart logo

### 7.4. Jumia

Ce site est considéré comme l'Amazon africain : apparu au Nigéria, il s'est répandu sur tout le continent pour devenir un acteur majeur de la vente en ligne. Il compte 3 millions de visites mensuelles. [14]



**Figure 13** Jumia logo

## 8. E-commerce en Algérie

Le premier site d'e-commerce algérien a été créé en 2010 Guidini.com, puis plusieurs autres sites ont été créés, eChrily.com qui fait partie des sites pionniers de ventes en ligne créé en juillet 2012. Six mois plus tard, la toile algérienne accueillait deux nouveaux sites : Tbeznyss lancé le 12 janvier 2013, et Nchrifenet lancé en 16 Janvier 2013.

Selon l'économiste M. Ould Moussa, les sites naissent comme des champignons sur le web, n'exercent pas le e-commerce, mais pratiquent une forme de commerce informel ; ce sont des

sites d'information et de publicité pour les transactions commerciales grâce à Internet. Il justifie son avis par l'absence de transparence et d'encadrement juridique de ses transactions, il poursuit le E-commerce ne peut être exercé dans un marché de bancarisé. Certes le e-commerce manque de services de livraison et de paiement en ligne en raison du manque de confiance, et malgré les entraves administratives et technologiques, les algériens ont trouvé d'autres moyens pour exercer le e-commerce, comme la création de pages facebook. [13]

## **9. Quelle que travail utilisent la prédiction Data Mining dans E-Commerce**

Avec des millions de clients à travers le monde, Amazon est la meilleure boutique en ligne de produits et services incroyables en ligne. Ils ont une merveilleuse base de données clients et l'utilisent pour établir des relations solides avec leurs clients. [11]

### **9.1. Principe du système d'Amazon**

En analysant et en résumant les informations utiles sur les clients, ils conçoivent leurs stratégies en matière de promotions et améliorations des produits. Il a incorporé le concept d'exploration de données de la chaîne d'approvisionnement à opérations marketing (Dholakia, 2013). Amazon utilise l'exploration de données (Data Mining) pour la commercialisation de leurs produits sous divers aspects pour avoir un avantage. Les clients veulent une personnalisation des entreprises qu'ils achètent principalement des entreprises en ligne en raison de l'augmentation des interventions des médias sociaux.

Fournir les informations ciblées au représentant du service à la clientèle traitant un client particulier est une opportunité nécessaire pour être développé. Si les employés ont les bons outils pour accéder aux informations requises lors de la négociation avec le client, cela gagnera beaucoup de temps et fera une bonne impression sur le client. Chez Amazon les représentants des clients ont accès à compléter les données clients et analyser les problèmes et discuter dès que vous recevez un appel de client. Les clients se sentiront à l'aise et détendus qu'ils connaissent son problème. Intelligent les détaillants comme Amazon utilisent efficacement les données recueillies auprès de sources efficaces et utiliser les résultats plus raisonnablement. Aussi les clients ont le contrôle sur les informations qu'ils souhaitent partager ou pas. Cela leur donne un sentiment d'appartenance et de contrôle.

Les données clients sont devenues un moyen de construire des marques fortes et de la clientèle grâce à l'extraction de données efficace. Chaque client est traité à titre individuel et prioritaire. Ils se sentent à l'aise avec le service interagissant avec eux s'ils obtiennent le contrôle

et la transparence. Les spécialistes du marketing obtiennent des informations à chaque clic. Le l'historique des achats des clients sur Amazon les aide pour identifier les préférences et les choix des clients (Clufia, Bunzel et Snnuggs.2014). Ils ont ensuite mené la campagne publicitaire selon le choix du client. Le mouvement du social les données sont beaucoup plus rapides que les données manuelles et sont téléchargées automatiquement. Les modèles de comportement sont étudiés pour identifier les canaux marketing et élaborer des stratégies marketing par conséquent. [11]

### **9.2. Objectif d'utilisation Data Mining :**

- Collecté plus de client
- Gagné leur confiance
- Connaître la pensée des clients et l'attirez
- Faire beaucoup d'offre
- Avoir prévision a les favoris des clients
- Vente rapide
- Savoir les defaults [11]

## **10. Conclusion**

L'exploration de données offre de nombreux avantages aux entreprises de commerce électronique et leur permet de planifier les marchandises, d'analyser les comportements d'achat des clients et de prévoir leurs ventes, ce qui les placerait au-dessus d'autres entreprises et générerait plus de revenus. D'autre part, il existe certains défis de l'exploration de données dans le domaine du commerce électronique tels que l'identification des araignées, la transformation des données, l'évolutivité des algorithmes d'exploration de données, rendre le modèle d'exploration de données compréhensible pour les utilisateurs professionnels, prendre en charge les dimensions à évolution lente et rendre la transformation des données.

L'exploration de données joue un rôle important dans la fourniture de services orientés client pour augmenter la satisfaction client. Il est devenu évident que l'utilisation d'outils d'exploration de données est une nécessité pour les entreprises de commerce électronique dans cet environnement concurrentiel mondial



## Chapitre 3 : conception du système

### 1. Introduction

Dans ce chapitre nous allons voir une description générale de mon système, voire aussi le côté conceptuel du prétraitement qui constitue des étapes avant la mise en œuvre du système, ensuite détaillé chaque phase et voire les algorithmes et techniques utilisées.

### 2. Architecture générale

Mon système se base sur l'utilisation de la prédiction dans le e-commerce. Le système prend en entrée une base brute d'un site e-commerce et la transforme en une base utilisable pour la prédiction. Cette transformation signifie pré-traitement, dans le pré-traitement on effectue des opérations de nettoyage et encodage, La base obtenue je vais la subdivisée en trois parties, une pour l'entraînement et une pour le test de validation et une pour le test, La partie d'entraînement utilise la base d'entraînement et un algorithme de prédiction pour fournir un modèle de décision qui est appliqué sur la base de test. Si le modèle est accepté, c-à-d a pu atteindre un taux de reconnaissance acceptable, il sera conservé et utilisé par le module d'utilisation et l'entraînement se termine.

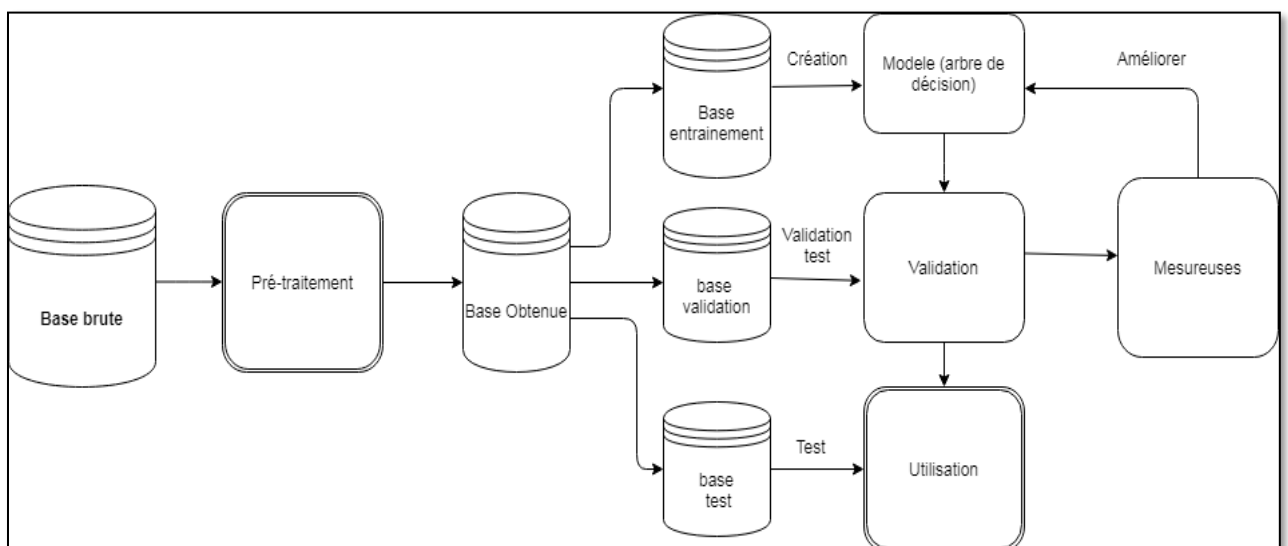
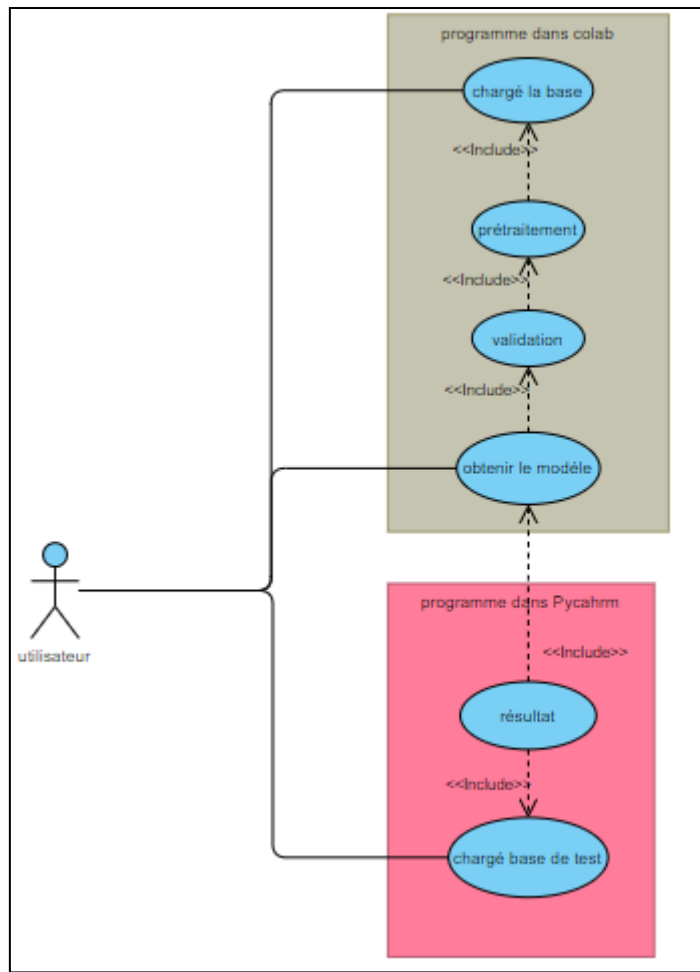
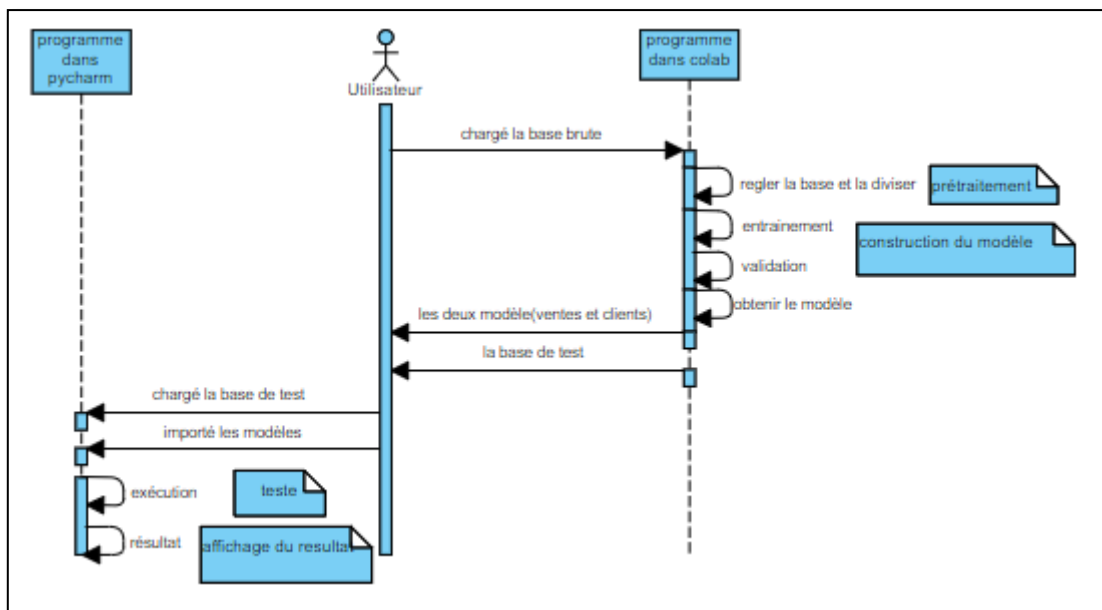


Figure 14 Architecture générale de notre système



**Figure 15** diagramme de cas d'utilisation du notre système



**Figure 16** diagramme de séquence pour notre système

### 3. Architecture détaillée

Dans cette partie je vais expliquer chaque phase du système :

#### 3.1. Dataset

Comme n'importe quel système, nous avons besoin d'un dataset pour l'entraîner sur les données en utilisant ses algorithmes pour prédire le nombre des ventes et aussi le nombre des consommateurs de la base de test.

Le dataset brute que j'ai utilisé et préparé, chaque information a été enregistrée dans un fichier CSV (plus de 800.000 lignes). Le fichier est ensuite subdivisé en trois parties, une consacrée à l'entraînement et au test et au validation. Après la validation, nous sommes passées à l'étape de l'utilisation qui consiste à prédire le nombre des ventes et aussi le nombre des consommateurs.

Le dataset je l'ai extrait depuis le site d'internet kaggle.com

#### 3.2. Prétraitements

Le prétraitement est d'effectuer des opérations a les données du dataset brute qui sont subdivisées en deux catégories : les données textuelles et les données numériques, pour chaque catégorie :

##### 3.2.1. Données textuelles

Représentent les valeurs des variables de type chaîne de caractère, on l'effectue les opérations suivantes :

**Nettoyage** consiste à éliminer les mots vides ou on trouve (Nan) et toute information non utile.

**Encodage** cette opération consiste à transformer les données textuelles à une valeur numérique, dans la base j'ai utilisé des simples énumérations dans les variables (**StateHoliday, StoreType, assortment**) comme ce suit :

Dans StateHoliday la variable doit être soit (0, a, b, c) alors la transformation va être soit (0, 1, 2, 3)

Dans StoreType la variable doit être soit (a, b, c, d) alors la transformation va être soit (0, 1, 2, 3)

Dans assortment la même chose on a la variable doit être soit (a, b, c) alors la transformation va être soit (0, 1, 2)

| <b>Variable</b> | <b>Encodage</b> |
|-----------------|-----------------|
| StateHoliday    | <b>0 à 0</b>    |
|                 | <b>A à 1</b>    |
|                 | <b>B à 2</b>    |
| StoreType       | <b>C à 3</b>    |
|                 | <b>A à 0</b>    |
|                 | <b>B à 1</b>    |
|                 | <b>C à 2</b>    |
| assortment      | <b>D à 3</b>    |
|                 | <b>A à 0</b>    |
|                 | <b>B à 1</b>    |
|                 | <b>C à 2</b>    |

**Figure 17** Encodage des valeurs textuelles

### 3.2.2. Données numériques

Concernant les données numériques comme la date qu'est valeur numérique j'ai seulement découpé en trois valeurs uniques : jour, mois et année. J'ai effectué aussi l'opération de nettoyage qui fait :

#### **Nettoyage**

Cette opération consiste à éliminer toute information non utile.

Voici résultats de la base obtenue :

|   | Date       | Store | DayOfWeek | Sales | Customers | Open | Promo | StateHoliday | SchoolHoliday | Year | Month | Day | WeekOfYear | StoreType | Assortment |
|---|------------|-------|-----------|-------|-----------|------|-------|--------------|---------------|------|-------|-----|------------|-----------|------------|
| 0 | 2015-07-31 | 1     | 5         | 5263  | 555       | 1    | 1     | 0            | 1             | 2015 | 7     | 31  | 31         | 2         | 0          |
| 1 | 2015-07-31 | 2     | 5         | 6064  | 625       | 1    | 1     | 0            | 1             | 2015 | 7     | 31  | 31         | 0         | 0          |
| 2 | 2015-07-31 | 3     | 5         | 8314  | 821       | 1    | 1     | 0            | 1             | 2015 | 7     | 31  | 31         | 0         | 0          |
| 3 | 2015-07-31 | 4     | 5         | 13995 | 1498      | 1    | 1     | 0            | 1             | 2015 | 7     | 31  | 31         | 2         | 2          |
| 4 | 2015-07-31 | 5     | 5         | 4822  | 559       | 1    | 1     | 0            | 1             | 2015 | 7     | 31  | 31         | 0         | 0          |

**Figure 18** Résultat de la base obtenue

### 3.3. Entraînement

Pour entraîner notre modèle, nous avons choisi l’algorithme que nous avons déjà présenté dans le chapitre précédent, **Les arbres de décision (dans le cas régression)**. Les arbres de décision (Regression Tree) permettent de prédire une quantité réelle, une valeur numérique (par exemple, le prix d’une maison ou la durée de séjour d’un patient dans un hôpital).

La régression consiste à produire un modèle qui permet de prédire ou d’expliquer les valeurs d’une variable à prédire continue (endogène) à partir des valeurs d’une série de variables prédictives (exogènes), continues ou discrètes. La régression linéaire multiple est certainement l’approche la plus connue, mais d’autres méthodes, moins connues en économétrie mais plus populaires dans la communauté de l’apprentissage automatique, permettent de remplir cette tâche.

#### 3.3.1. Construction l’Arbre de décision(modèle)

##### Définir le problème

L’étape suivante consiste à désigner la variable à prédire et les variables prédictives. Dans ce cas on va faire la prédiction dans les ventes et le nombre des consoleurs pour chaque saison à partir des autres variables de la base traitée.

##### Définir la profondeur maximale de l’arbre

La profondeur maximale théorique qu’un arbre de décision peut atteindre est un de moins que le nombre d’échantillons d’apprentissage.

Dans notre cas, la profondeur maximale de l’arbre. Si aucun, les nœuds sont développés jusqu’à ce que toutes les feuilles soient pures ou jusqu’à ce que toutes les feuilles contiennent moins d’échantillons. [16]

#### **4. Utilisation**

C'est la dernière phase et la plus importante dans notre système. Après être arrivé au meilleur taux de reconnaissance, ou après avoir construit le bon modèle dans la phase précédente, je dois l'utiliser sur des nouvelles informations, le modèle nous permet de prédire le nombre de ventes et le nombre de clients.

#### **5. Conclusion**

Ce chapitre a décrit la conception de mon système et il a présenté la démarche de ces différentes phases. Dans le chapitre suivant, nous allons décrire le fonctionnement de mon application mettant en œuvre le système proposé.

## Chapitre 4 : implémentation

### 1. Introduction

L'objectif de ce chapitre est de présenter les outils (les logiciels, les langages, les bibliothèques et les données utilisés dans mon système).

### 2. Outils utilisés

Les outils utilisés dans le travail sont :

#### 2.1. Logiciel utilisé

Dans mon travail j'ai utilisé google collaboratory pour la phase de pré-traitement et pour l'entraînement et la validation, et pour application je l'ai créé avec Pycharm.

#### Google collaboratory

Colaboratory, raccourci en "Colab", nous permet d'écrire et d'exécuter du code Python dans votre navigateur. Il offre les avantages suivants :

- Aucune configuration requise
- Accès gratuit aux GPU
- Partage facile



Figure 19 Logo du Google Colab

## Pycharm

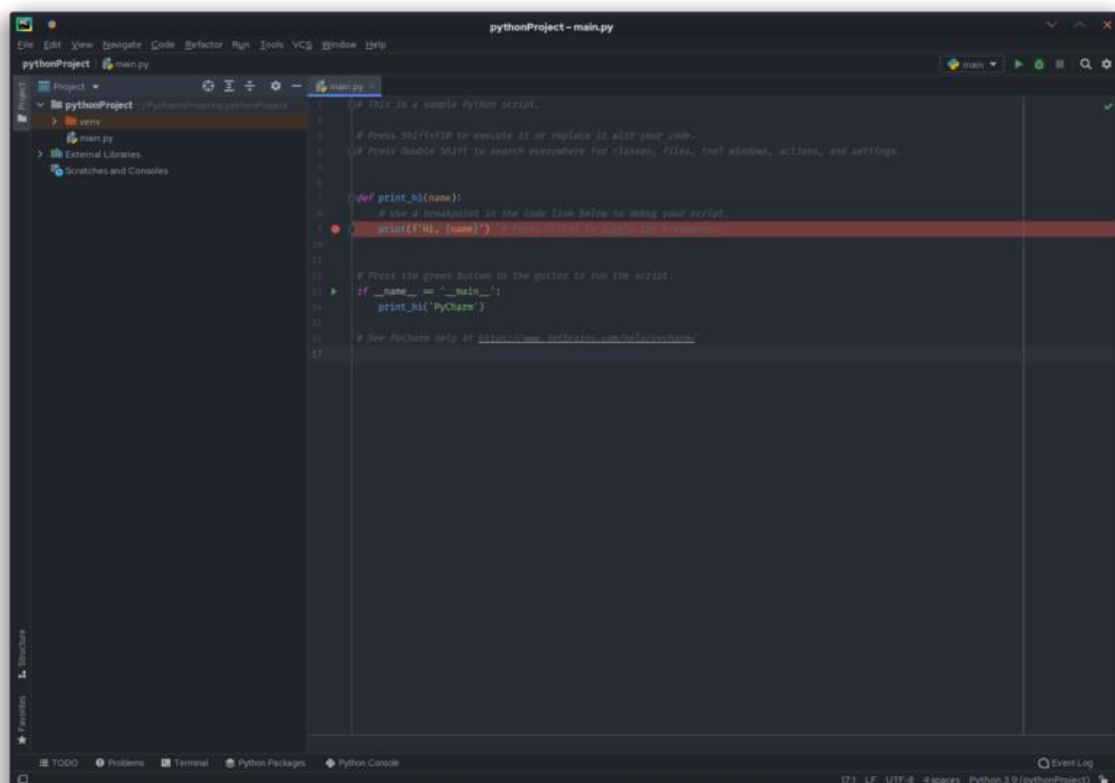
Est un environnement de développement intégré utilisé pour programmer en Python.

Il permet l'analyse de code et contient un débogueur graphique. Il permet également la gestion des tests unitaires, l'intégration de logiciel de gestion de versions, et supporte le développement web avec Django.

Développé par l'entreprise tchèque JetBrains, c'est un logiciel multi-plateforme qui fonctionne sous Windows, Mac OS X et Linux.



**Figure 20** Logo du Pycharm



**Figure 21** Fenêtre du logiciel Pycharm



## 2.2. Langage utilisé

Dans mon travail j'ai utilisé le langage Python (version = 3.8) dans toutes les phases de ce projet, que ce soit dans le prétraitement, l'entraînement, la validation et l'application ainsi que pour le traitement.

### Python

Est un langage de programmation interprété, multiparadigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet. Il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire par ramasse-miettes et d'un système de gestion d'exceptions ; il est ainsi similaire à Perl, Ruby, Scheme, Smalltalk et Tcl.

Le langage Python est placé sous une licence libre proche de la licence BSD 5 et fonctionne sur la plupart des plates-formes informatiques, des smartphones aux ordinateurs centraux , de Windows à Unix avec notamment GNU/Linux en passant par macOS, ou encore Android, iOS, et peut aussi être traduit en Java ou .NET. Il est conçu pour optimiser la productivité des programmeurs en offrant des outils de haut niveau et une syntaxe simple à utiliser.



**Figure 22** Logo Python

## 2.3. Bibliothèques utilisées

### Dans éditeur Google colab

On a utilisé les bibliothèques suivantes :

#### Pandas

Bibliothèque pour la manipulation des données, dans notre cas, nous l'utilisons pour lire les fichiers d'entrée (la base d'entraînement).

```
[6] train_data=pd.read_csv("/content/drive/MyDrive/AliKisrane/Data/train.csv",parse_dates = True, index_col = 'Date')
store_data = pd.read_csv("/content/drive/MyDrive/AliKisrane/Data/store.csv")
train_data.head()
```

**Figure 23** Utilisation de la bibliothèque Pandas dans Google colab

## Sklearn

Scikit-learn est une bibliothèque libre Python destinée à l'apprentissage automatique. Elle est développée par de nombreux contributeurs notamment dans le monde académique par des instituts français d'enseignement supérieur et de recherche comme Inria.

Cette bibliothèque est utilisée pour diviser la base en deux parties, partie d'entraînement du modèle d'apprentissage et l'autre pour validation.

```
from sklearn.model_selection import train_test_split
xd_train,xd_test,yd_train,yd_test = train_test_split(X,z,test_size=0.2, random_state=1)
```

**Figure 24** Utilisation de la bibliothèque Sklearn dans Google colab

D'autre part, nous avons utilisé cette bibliothèque pour importer le modèle de décision (Pour importer le modèle arbre de décision).

```
#####prediction sur le Nombre des Clients (Arbre de decision)#####
from sklearn import tree
model = tree.DecisionTreeRegressor()
model.fit(xd_train, yd_train)
yd_test_predicted = model.predict(xd_test)
model.score(xd_test, yd_test)
```

**Figure 25** Utilisation de la bibliothèque Sklearn dans Google colab

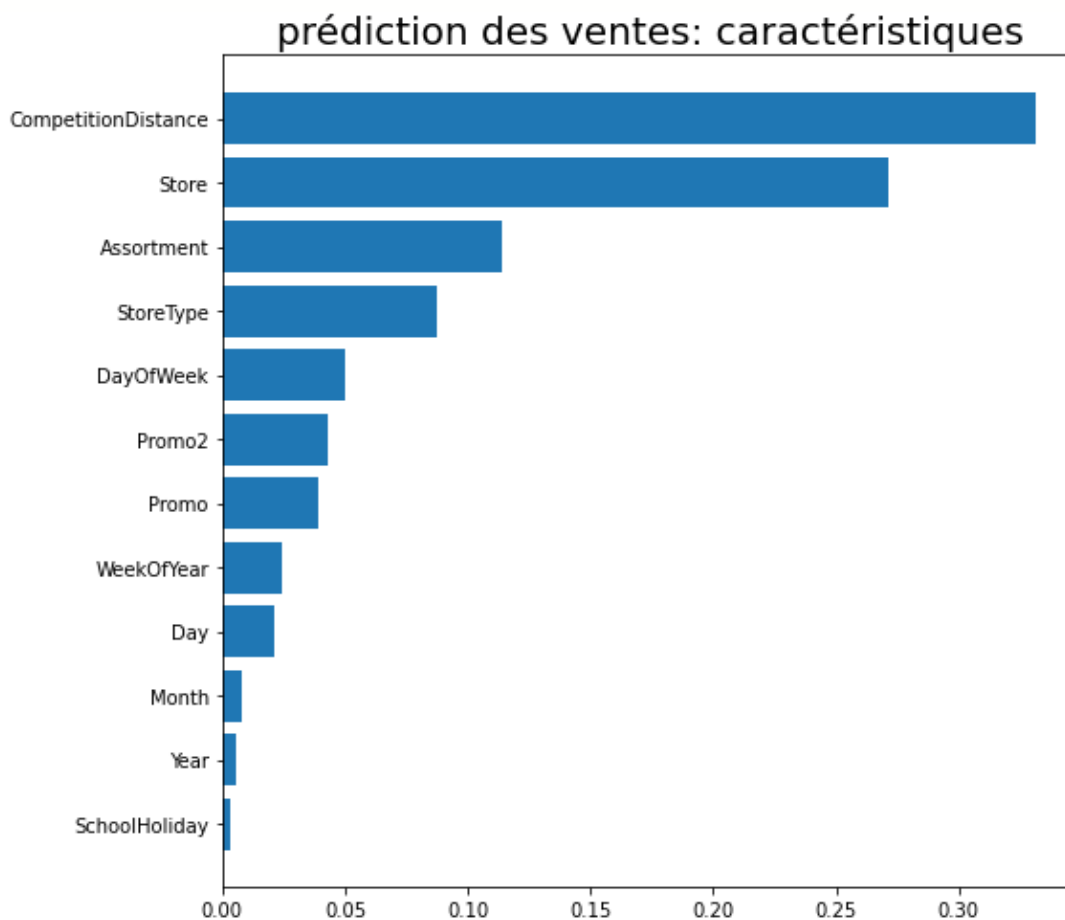
## Matplotlib

Matplotlib est une bibliothèque complète permettant de créer des visualisations statiques, animées et interactives en Python.

Dans notre cas, nous avons utilisé cette bibliothèque pour extraire les caractéristiques (les variables de base d'entraînement) sur lesquelles se base le modèle de prédiction (la prédiction des ventes et la prédiction des nombres de clients).

```
[23] features = xd_train.columns
importances = model.feature_importances_
indices = np.argsort(importances)
plt.figure(figsize=(8,8))
plt.title('prédiction des ventes: caractéristiques', fontsize=20)
plt.barh(range(len(indices)), importances[indices], align='center')
plt.yticks(range(len(indices)), features[indices])
plt.xlabel('caractéristiques')
```

**Figure 26** Utilisation de la bibliothèque Matplotlib dans Google colab



**Figure 27** visualisation des caractéristiques en utilisant Matplotlib dans Google colab

## Pickle

Bibliothèque utilisée pour sauvegarder le modèle de décision créé dans google colab.

```
[ ] import pickle
filename = '/content/drive/MyDrive/AliKisrane/Models/sales_model_arbre.sav'
pickle.dump(model, open(filename, 'wb'))
import pickle
filename = '/content/drive/MyDrive/AliKisrane/Models/costumers_model_arbre.sav'
pickle.dump(model1, open(filename, 'wb'))
```

**Figure 28** Utilisation de la bibliothèque pickle dans Google colab

### Dans éditeur pycharm

On a utilisé les bibliothèques suivantes :

#### OS

Dans notre cas, la bibliothèque est utilisée pour importer la base de test.

#### Pandas

Dans notre cas, nous l'utilisons pour lire les fichiers d'entrée (la base de test) qui déjà importer par fonction **OS.PATH**.

#### Pickle

Bibliothèque utiliser pour importer le modèle de décision créé à partir google colab et prédire les nombres des ventes et les nombre de clients dans chaque saison.

```
if value == "Prédiction des Vents":
    train_data = pd.read_csv("" + str(fName) + "")
    X = train_data.drop(['Date', 'Sales', 'Open', 'StateHoliday', 'Customers'], axis=1)
    y = train_data.Sales
    loaded_model = pickle.load(open('sales_model_random.sav', 'rb'))
    result = loaded_model.predict(X)
    X['preds'] = result
    df_out = pd.merge(train_data, X[['preds']], how='left', left_index=True, right_index=True)
    print(df_out[['preds']])
elif value == "Prédiction Nbr Clients":
    train_data = pd.read_csv("" + str(fName) + "")
    X = train_data.drop(['Date', 'Sales', 'Open', 'StateHoliday', 'Customers'], axis=1)
    y = train_data.Customers
    loaded_model = pickle.load(open('costumers_model_random.sav', 'rb'))
    result = loaded_model.predict(X)
    X['preds'] = result
    df_out = pd.merge(train_data, X[['preds']], how='left', left_index=True, right_index=True)
    print(df_out[['preds']])
```

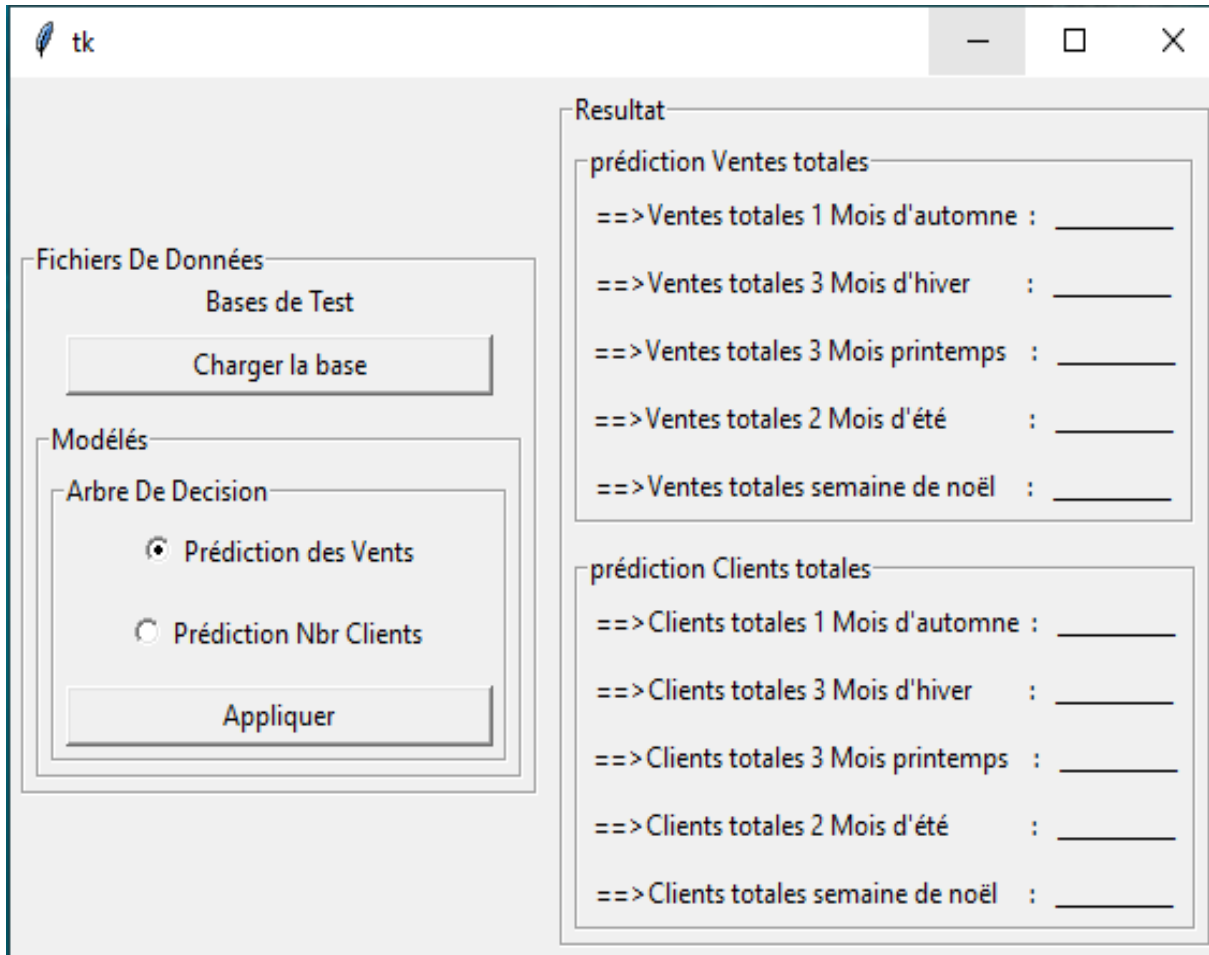
**Figure 29** Utilisation de la bibliothèque pickle dans Pycharm

### Tkinter

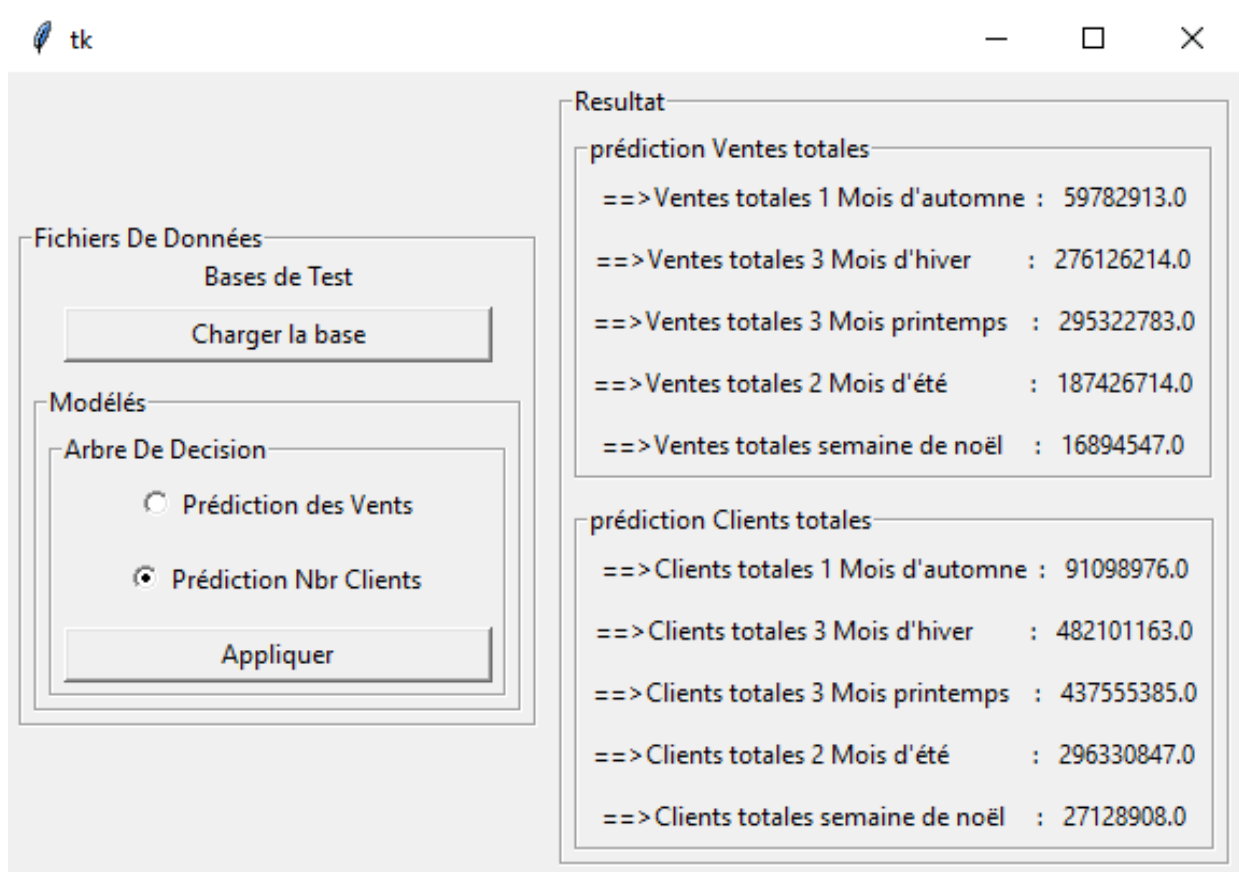
Bibliothèque utilisée pour modéliser l'interface graphique.

### 3. Application

C'est une fenêtre qui permet de tester des bases après avoir fourni le modèle lors de l'entraînement, le télécharger et le mettre dans le package du travail dans Pycharm et enfin appliquer le test et voir les résultats de la prédiction des ventes et clients pour chaque saison.



**Figure 30** fenêtre du notre travail



**Figure 31** capture du résultat finale

#### 4. Conclusion

Ce chapitre a décrit l'implémentation de notre système et il a présenté les différents outils utilisés pour la mise en œuvre.

## **Conclusion générale**

Ce mémoire présente une méthode de prédiction dans le e-commerce en utilisant l'arbre de décision, en essayant d'avoir des meilleurs résultats pour la prédiction des nombres de ventes et du nombre de clients.

Nous avons commencé par étudier la prédiction de la tâche du Data Mining.

Nous avons aussi présenté e-commerce, leur historique, évolution, objectif ...etc.

Nous avons ensuite conçu et implémenté sur un *Dataset* une solution qui se base sur le prétraitement, c'est pour nettoyer et encode les données brutes, après faire le modèle(arbre) et le valider, finalement appliqué et testé dans l'application.

## Code source

### Dans google colab

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")

train_data=pd.read_csv("/content/drive/MyDrive/AliKisrane/Data/train.csv",parse_dates = True, index_col = 'Date')
store_data = pd.read_csv("/content/drive/MyDrive/AliKisrane/Data/store.csv")
train_data.head()

train_data["Year"]=train_data.index.year
train_data["Month"]=train_data.index.month
train_data["Day"]=train_data.index.day
train_data["WeekOfYear"]=train_data.index.week
train_data = train_data.reset_index()

store_data = store_data.drop(['CompetitionOpenSinceMonth', 'CompetitionOpenSinceYear', 'Promo2SinceWeek',
                              'Promo2SinceYear', 'PromoInterval'], axis=1)

# remplacer les values null avec le median dans competitionDistance
store_data.CompetitionDistance.fillna(store_data.CompetitionDistance.median(), inplace=True)

df = pd.merge(train_data, store_data, how='left', on='Store')

# nombre des magazine ouvert avec 0 de vente
print("Number of stores with zero sales:",df[df.Sales == 0].shape[0])

# voir les variables nominale
set(df.StateHoliday), set(df.StoreType), set(df.Assortment)
df.loc[df.StateHoliday == 0, 'StateHoliday'] = df.loc[df.StateHoliday == 0, 'StateHoliday'].astype(str)
set(df.StateHoliday)

# modifier les caractères en numérique
```



```
df['StateHoliday'] = df.StateHoliday.map({'0':0, 'a':1, 'b' : 2, 'c': 3}
)
set(df.StateHoliday)
```

```
df['StoreType'] = df.StoreType.map({'a':0, 'b':1, 'c' : 2, 'd': 3})
df['Assortment'] = df.Assortment.map({'a':0, 'b':1, 'c' : 2})
set(df.StoreType)
set(df.Assortment)
```

```
##### entrainment #####
df1 = df.iloc[236382:844338, :]
##### test #####
df2 = df.iloc[0:236381, :]
df2.to_csv('/content/drive/MyDrive/AliKisrane/Testing_data.csv')
```

```
X = df1.drop(['Date', 'Sales', 'Open', 'StateHoliday', 'Customers'], axis =
1)
y = df1.Sales
z = df1.Customers
```

```
from sklearn.model_selection import train_test_split
xd_train, xd_test, yd_train, yd_test = train_test_split(X, z, test_size=0.2,
random_state=1)
```

```
#####prediction sur le Nombre des Clients (Arbre de decision)#
#####
from sklearn import tree
model = tree.DecisionTreeRegressor()
model.fit(xd_train, yd_train)
yd_test_predicted = model.predict(xd_test)
model.score(xd_test, yd_test)
```

```
features = xd_train.columns
importances = model.feature_importances_
indices = np.argsort(importances)
plt.figure(figsize=(8,8))
plt.title('prédiction des ventes: caractéristiques', fontsize=20)
plt.barh(range(len(indices)), importances[indices], align='center')
plt.yticks(range(len(indices)), features[indices])
plt.xlabel('caractéristiques')
```

```
#####prediction sur les Vents (Arbre de decision)#####
```

```

xd_train1,xd_test1,yd_train1,yd_test1 = train_test_split(X,y,test_size=
0.2, random_state=1)
modell = tree.DecisionTreeRegressor()
modell.fit(xd_train1, yd_train1)
yd_test_predicted = model.predict(xd_test1)
modell.score(xd_test1, yd_test1)

features = xd_train.columns
importances = modell.feature_importances_
indices = np.argsort(importances)
plt.figure(figsize=(8,8))
plt.title('prédiction nombre de Costomurs: caractéristiques', fontsize=
20)
plt.barh(range(len(indices)), importances[indices], align='center')
plt.yticks(range(len(indices)), features[indices])
plt.xlabel('caractéristiques')

##### sauvegarde le modèle

import pickle
filename = '/content/drive/MyDrive/AliKisrane/Models/sales_model_arbre.
sav'
pickle.dump(model, open(filename, 'wb'))
import pickle
filename = '/content/drive/MyDrive/AliKisrane/Models/costumers_model_ar
bre.sav'
pickle.dump(modell, open(filename, 'wb'))

```

## dans pycharm

```

import pandas as pd
import pickle
import tkinter as tk
import tkinter.ttk as ttk
from os import path
from tkinter import messagebox as msg
from tkinter import filedialog as fd
import warnings
warnings.filterwarnings("ignore")
global ext
global fName
"""-----charge de la base-----"""
def LoadTrainingDataSet():
    global ext
    global extn
    global fName
    try:
        extn = 0
        fDir = path.dirname(__file__)
        fName = fd.askopenfilename(parent=win, initialdir=fDir)
        if fName.endswith("Testing_data.csv"):
            ext = list(fName)

```

```

        else:
            msg.showerror('Erreur', 'Erreur')
    except:
        msg.showerror('Erreur', 'Erreur chargement des données')
"""-----"""
"""-----"""
"""-----"""
def validateArbre():
    try:
        value = radVar.get()
        if value == "Prédiction des Vents":
            train_data = pd.read_csv("" + str(fName) + "")
            X = train_data.drop(['Date', 'Sales', 'Open', 'StateHoliday',
'Customers'], axis=1)
            y = train_data.Sales
            loaded_model = pickle.load(open('sales_model_arbre.sav', 'rb'))
            result = loaded_model.predict(X)
            X['preds'] = result
            df_out = pd.merge(train_data, X[['preds']], how='left',
left_index=True, right_index=True)
            '''-----hiver-----'''
            df_out['Date'] = pd.to_datetime(train_data['Date'])
            mask = (df_out['Date'] >= '2014-12-01') & (df_out['Date'] <=
'2015-02-28')
            ssx=df_out.loc[mask].sum()
            label1.config(text=="=>Ventes totales 3 Mois d'hiver      :
" + str(ssx['preds']))
            '''-----automne-----'''
            df_out['Date'] = pd.to_datetime(train_data['Date'])
            mask = (df_out['Date'] < '2014-12-01')
            ssx = df_out.loc[mask].sum()
            label2.config(text=="=>Ventes totales 1 Mois d'automne :  " +
str(ssx['preds']))
            '''-----printemps-----'''
            df_out['Date'] = pd.to_datetime(train_data['Date'])
            mask = (df_out['Date'] >= '2015-03-01') & (df_out['Date'] <=
'2015-05-31')
            ssx = df_out.loc[mask].sum()
            label3.config(text=="=>Ventes totales 3 Mois printemps      :  "
+ str(ssx['preds']))
            '''-----ete-----'''
            df_out['Date'] = pd.to_datetime(train_data['Date'])
            mask = (df_out['Date'] >= '2015-06-01')
            ssx = df_out.loc[mask].sum()
            label4.config(text=="=>Ventes totales 2 Mois d'été
:  " + str(ssx['preds']))
            '''-----la semaine de
noel-----'''
            df_out['Date'] = pd.to_datetime(train_data['Date'])
            mask = (df_out['Date'] >= '2014-12-24') & (df_out['Date'] <=
'2015-01-01')
            ssx = df_out.loc[mask].sum()
            label5.config(text=="=>Ventes totales semaine de Noël      :  "
+ str(ssx['preds']))
            '''-----'''
            '''-----'''

```

```

elif value == "Prédiction Nbr Clients":
    train_data = pd.read_csv("" + str(fName) + "")
    X = train_data.drop(['Date', 'Sales', 'Open', 'StateHoliday',
'Customers'], axis=1)
    y = train_data.Customers
    loaded_model = pickle.load(open('costumers_model_arbre.sav',
'rb'))

    result = loaded_model.predict(X)
    X['preds'] = result
    df_out = pd.merge(train_data, X[['preds']], how='left',
left_index=True, right_index=True)
    '''-----WINTER-----
-----'''
    df_out['Date'] = pd.to_datetime(train_data['Date'])
    mask = (df_out['Date'] >= '2014-12-01') & (df_out['Date'] <=
'2015-02-28')
    ssx = df_out.loc[mask].sum()
    label6.config(text=="=>Clients totales 3 Mois d'hiver      :
" + str(ssx['preds']))
    '''-----Autumn-----
-----'''
    df_out['Date'] = pd.to_datetime(train_data['Date'])
    mask = (df_out['Date'] < '2014-12-01')
    ssx = df_out.loc[mask].sum()
    label7.config(text=="=>Clients totales 1 Mois d'automne : "
+ str(ssx['preds']))
    '''-----Spring-----
-----'''
    df_out['Date'] = pd.to_datetime(train_data['Date'])
    mask = (df_out['Date'] >= '2015-03-01') & (df_out['Date'] <=
'2015-05-31')
    ssx = df_out.loc[mask].sum()
    label8.config(text=="=>Clients totales 3 Mois printemps      :
" + str(ssx['preds']))
    '''-----Summer-----
-----'''
    df_out['Date'] = pd.to_datetime(train_data['Date'])
    mask = (df_out['Date'] >= '2015-06-01')
    ssx = df_out.loc[mask].sum()
    label9.config(text=="=>Clients totales 2 Mois d'été
: " + str(ssx['preds']))
    '''-----christmas week-----
-----'''
    df_out['Date'] = pd.to_datetime(train_data['Date'])
    mask = (df_out['Date'] >= '2014-12-24') & (df_out['Date'] <=
'2015-01-01')
    ssx = df_out.loc[mask].sum()
    label10.config(text=="=>Clients totales semaine de Noël      :
" + str(ssx['preds']))
    '''-----'''

except:
    msg.showerror('Erreur', 'SVP!! charger les données')
"""-----"""

def validateRandom():
    try:
        value = radVar1.get()
        if value == "Prédiction des Vents":
            train_data = pd.read_csv("" + str(fName) + "")

```

```

        X = train_data.drop(['Date', 'Sales', 'Open', 'StateHoliday',
'Customers'], axis=1)
        y = train_data.Sales
        loaded_model = pickle.load(open('sales_model_random.sav',
'rb'))

        result = loaded_model.predict(X)
        X['preds'] = result
        df_out = pd.merge(train_data, X[['preds']], how='left',
left_index=True, right_index=True)
        print(df_out['preds'])
        elif value == "Prédiction Nbr Clients":
            train_data = pd.read_csv("" + str(fName) + "")
            X = train_data.drop(['Date', 'Sales', 'Open', 'StateHoliday',
'Customers'], axis=1)
            y = train_data.Customers
            loaded_model = pickle.load(open('costumers_model_random.sav',
'rb'))

            result = loaded_model.predict(X)
            X['preds'] = result
            df_out = pd.merge(train_data, X[['preds']], how='left',
left_index=True, right_index=True)
            print(df_out['preds'])

    except:
        msg.showerror('Erreur', 'SVP!! charger les données')
'''
-----'''
win = tk.Tk()
groupbox = tk.LabelFrame(win, text='Fichiers De Données')
groupbox.grid(column=0, row=0, padx=5, pady=5)
ttk.Label(groupbox, text="Bases de Test").grid(column=0, row=0)
action = tk.Button(groupbox, text="Charger la base", width=27,
command=LoadTrainingDataSet)
action.grid(column=0, row=1, padx=5, pady=5)

groupbox1 = tk.LabelFrame(groupbox, text='Modélés')
groupbox1.grid(column=0, row=2, padx=5, pady=5)
ttk.Label(groupbox1, text="Choisir Une Modélé :").grid(column=0, row=0)
groupbox2 = tk.LabelFrame(groupbox1, text='Arbre De Decision')
groupbox2.grid(column=0, row=0, padx=5, pady=5)
radVar=tk.StringVar()
rad = tk.Radiobutton(groupbox2, text='Prédiction des
Vents', value="Prédiction des Vents", variable=radVar)
rad.grid(column=0, row=1, padx=5, pady=5)
rad1 = tk.Radiobutton(groupbox2, text='Prédiction Nbr
Clients', value="Prédiction Nbr Clients", variable=radVar)
rad1.grid(column=0, row=2, padx=5, pady=5)
action1 = tk.Button(groupbox2, text="Appliquer", width=27,
command=validateArbre)
action1.grid(column=0, row=3, padx=5, pady=5)

groupbox4 = tk.LabelFrame(win, text='Resultat')
groupbox4.grid(column=1, row=0, padx=5, pady=5)
groupbox55 = tk.LabelFrame(groupbox4, text='prédiction Ventes totales')
groupbox55.grid(column=0, row=0, padx=5, pady=5)
label1=ttk.Label(groupbox55, text=="=>Ventes totales 3 Mois d'hiver
:
_____")
label1.grid(column=0, row=1, padx=5, pady=5)
label2=ttk.Label(groupbox55, text=="=>Ventes totales 1 Mois d'automne :
_____")

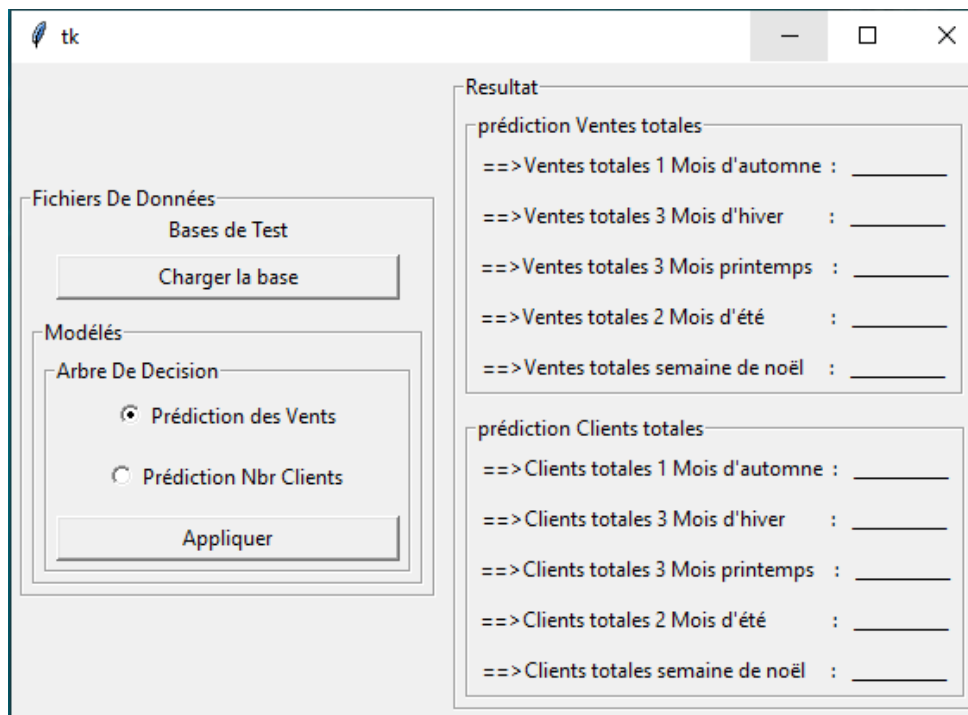
```

```

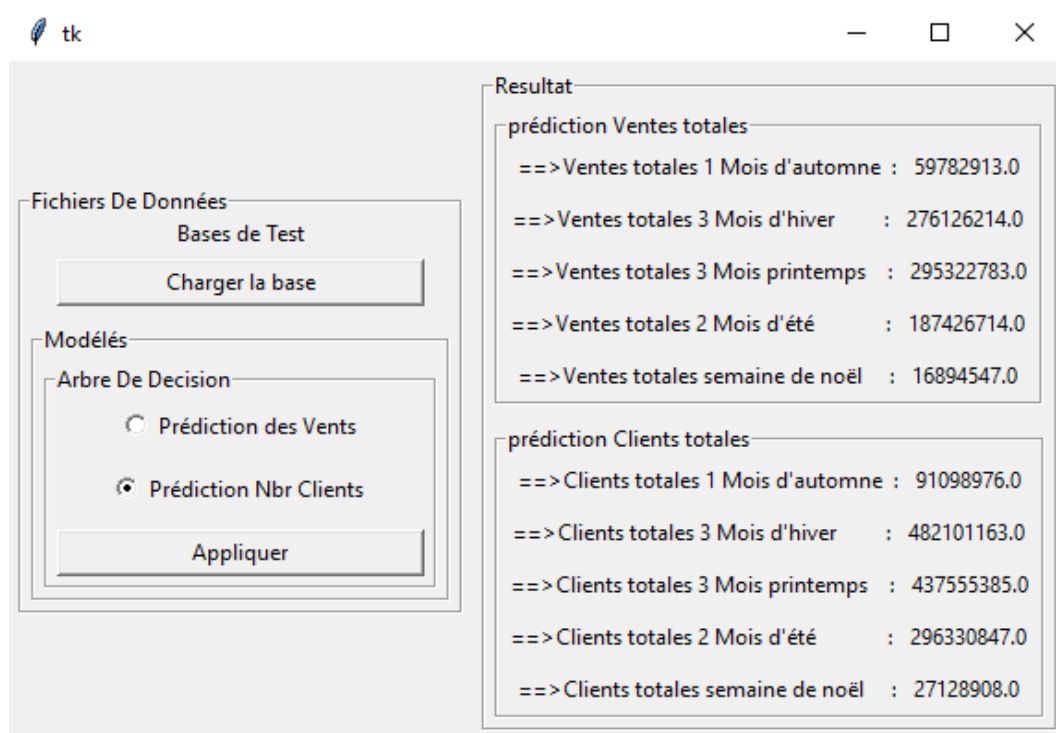
label2.grid(column=0, row=0, padx=5, pady=5)
label3=ttk.Label(groupbox55, text=="=>Ventes totales 3 Mois printemps :
")
label3.grid(column=0, row=2, padx=5, pady=5)
label4=ttk.Label(groupbox55, text=="=>Ventes totales 2 Mois d'été
:
")
label4.grid(column=0, row=3, padx=5, pady=5)
label5=ttk.Label(groupbox55, text=="=>Ventes totales semaine de Noël :
")
label5.grid(column=0, row=4, padx=5, pady=5)

groupbox6 = tk.LabelFrame(groupbox4, text='prédiction Clients totales')
groupbox6.grid(column=0, row=1, padx=5, pady=5)
label6=ttk.Label(groupbox6, text=="=>Clients totales 3 Mois d'hiver
:
")
label6.grid(column=0, row=1, padx=5, pady=5)
label7=ttk.Label(groupbox6, text=="=>Clients totales 1 Mois d'automne :
")
label7.grid(column=0, row=0, padx=5, pady=5)
label8=ttk.Label(groupbox6, text=="=>Clients totales 3 Mois printemps :
")
label8.grid(column=0, row=2, padx=5, pady=5)
label9=ttk.Label(groupbox6, text=="=>Clients totales 2 Mois d'été
:
")
label9.grid(column=0, row=3, padx=5, pady=5)
label10=ttk.Label(groupbox6, text=="=>Clients totales semaine de Noël :
")
label10.grid(column=0, row=4, padx=5, pady=5)
win.mainloop()

```



Copie Figure 32 fenêtre du notre travail



**Copie Figure 33** capture du résultat finale

## **bibliography**

- [1] DATA MINING Concepts, Models, Methods, and Algorithms's 3<sup>rd</sup> Edition, Mehmed Kantardzic, 2019.
- [2] Data Mining: Practical Machine Learning Tools and Techniques (Morgan Kaufmann Series in Data Management Systems) 4th Edition, Ian H. Witten , Eibe Frank, Mark A. Hall , Christopher J. Pal, 1 October 2016.
- [3] Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems) 3rd Edition Jiawei Han, Jian Pei, Micheline Kamber, 6 July 2011
- [4] Predictive Data Mining Models Second Edition David L. Olson, Desheng Wu, 7 August 2019
- [5] Rokach, Lior, Maimon, *Data mining with decision trees: theory and applications.*, 2nd Ed, World Scientific Pub Co Inc., 2015.
- [6] Statistique appliquée aux sciences de la vie, Valentin ROUSSON, 02/2017
- [7] Data mining et statistique décisionnelle, La science des données, Stéphane Tuffery, 2017
- [8] Modélisation prédictive et apprentissage statistique avec R, Stéphane Tuffery, 2010, revue 2017
- [9] Business Intelligence et management, Alphonse Carlier, 2013
- [10] E-commerce : les bonnes pratiques pour réussir, demien djacob, 2012
- [11] Amazon : La boutique à tout vendre, brade stone, 2013
- [12] LEARNING FOR BUSINESS – RETAIL, crossknowledge, 2016
- [13] <https://web-rocket.dz/guide-algerien-du-e-commerce/> 07/04/2021.
- [14] <https://www.blogdumoderateur.com/chiffres-cles-e-commerce-2020/#:~:text=Le%20chiffre%20d'affaires%20du,6%20%25%20par%20rapport%20%C3%A0%202018> 08/04/2021.
- [15] R. GILLERON, M. TOMMASI, Découverte de connaissances à partir de données, 2000.
- [16] [https://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr\\_Tanagra\\_Regression\\_Tree.pdf](https://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Regression_Tree.pdf)