



République Algérienne Démocratique et Populaire



Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université des Sciences et de la Technologie Houari Boumediene

Faculté d'Électronique et d'Informatique  
Département Informatique

Mémoire de Master

Filière : Informatique

Spécialité : MIND

---

**Thème : Développement et Implémentation d'un  
Nouveau Système de Mesure de Satisfaction Client**

---

Encadré par :

**Mr.BENCHERIK Ahmed (Djezzy)**

**Dr. BERKANI Lamia (USTHB)**

**Soutenu le : .../09/2020**

**Présenté par :**

**ELKEFIF Nassim**

**MATA Abderezak**

**Devant le jury composé de :**

**M..... Président (e)**

**M..... Membre**

**Binôme N° : 031/ 2020**

## ***Remerciements***

Nous remercions en premier le bon dieu, tout puissant de nous avoir donné le courage et la volonté pour mener ce travail.

En préambule à ce mémoire, nous souhaitons adresser nos remerciements les plus sincères aux personnes qui nous ont apportées leur aide et qui ont contribué à l'élaboration de ce mémoire.

Nous tenons à remercier nos promoteurs en l'occurrence, **Docteur Lamia BERKANI** qui nous a beaucoup aidée, **Mr Salah ABROUS et Mr Ahmed BENCHRIK** pour les efforts louables qu'ils ont fournis ainsi que pour leurs prestigieuses orientations.

Que les membres du jury soient gracieusement remerciés pour avoir accepté d'évaluer notre travail.

Nous adressons nos remerciements à nos parents, frères, sœurs et ami(e)s, pour leur soutien et encouragement tout au long de la réalisation de ce mémoire.

Merci à toutes et à tous.

## ***Dédicaces***

*Je dédie ce modeste travail à ...*

*Mes très chers parents. Qu'Allah leur procure bonne santé et longue vie.*

*A mes chers sœurs et frères pour tout leur encouragement.*

*A tous mes chers amis*

*A mon binôme Abderezak.*

*A Ceux qui me sont proches et chers de loin ou de près*

***ELKEFIF Nassim***

## ***Dédicaces***

*Je dédie ce modeste travail à ...*

*Mes très chers parents. Qu'Allah leur procure bonne santé et longue vie.*

*A mes chers sœurs et frères pour tout leur encouragement.*

*A tous mes chers amis*

*A mon binôme Nassim.*

*A Ceux qui me sont proches et chers de loin ou de près*

***Mata Abderezak***

# Résumé

La satisfaction client est une mesure de la façon dont les produits et services fournis par une entreprise répondent ou dépassent les attentes des clients. Elle est devenue un enjeu majeur pour la survie et développement des entreprises. Par conséquent, ces dernières doivent à tout prix faire de leur mieux pour pourvoir garder l'expérience de leur clientèle plus satisfaisante que jamais. Néanmoins, la compagnie doit avoir une idée sur le taux de satisfaction de plusieurs échantillons de clients afin de cibler le lieu, le temps, la manière et le profil des destinataires non satisfaits.

L'apprentissage automatique, étant une branche de l'IA et parmi les spécialisations les plus répandues et utilisées ces dernières années, est l'une des technologies que la compagnie téléphonique Djazzy vise à utiliser afin d'automatiser la mesure de satisfaction client, dans un objectif d'avoir en vue un niveau approximatif de satisfaction globale de toute sa clientèle.

Afin de répondre aux attentes de Djazzy, nous avons conçu et réalisé un système permettant de mesurer le taux de satisfaction client en tirant profit de l'apprentissage automatique. Notre système implémente un modèle de classification qui permet de classer chaque abonné par une classe NPS (Net Promoter Score). Ce modèle est, ensuite, intégré dans une application web automatisée qui représente un Dashboard avec des statistiques et des visualisations concernant les abonnés en question, ainsi que leur taux de satisfaction globale.

**Mot clés :** Satisfaction client, NPS, Djazzy, Apprentissage automatique, Machine learning, Algorithmes supervisés, classification, réseaux de neurones.

# Table des matières :

<i>Introduction générale</i> .....	<i>1</i>
<b>1. Contexte d'étude</b> .....	<b>1</b>
1.1. Introduction .....	1
1.2. Présentation de l'entreprise .....	1
1.2.1. Optimum Telecom Algérie .....	1
1.2.2. Objectifs de l'OTA .....	2
1.2.3. Organigramme de l'OTA .....	3
1.2.4. Présentation du service .....	4
1.3. Données et système actuel .....	5
1.3.1. Source de données .....	5
1.3.2. Système actuel .....	6
1.3.3. Problématique et motivations .....	8
1.3.4. Mesure de NPS (Net Promoter Score) .....	8
1.4. Satisfaction client .....	10
1.5. Conclusion .....	10
<b>2. Apprentissage automatique et Télécom</b> .....	<b>11</b>
2.1. Introduction .....	11
2.2. Apprentissage automatique .....	11
2.2.1. Définition .....	11
2.2.2. Types d'apprentissage .....	11
2.3. Algorithmes d'apprentissage automatique .....	13
2.3.1. Algorithmes d'apprentissage Supervisé .....	14
2.3.2. Algorithmes d'apprentissage Non Supervisé .....	22
2.3.3. Réseaux de neurones artificiels .....	24
2.4. Démarches de résolution d'un problème d'apprentissage automatique .....	27
2.4.1. Préparation des données .....	27
2.4.2. Choix de l'algorithme d'apprentissage .....	27
2.4.3. Entraînement du modèle .....	28
2.4.4. Test .....	28
2.5. L'apprentissage automatique et télécom .....	28
2.5.1. Customer Churn Analysis in Telecom Industry de Kiran Dahiya et Surbhi Bhatia .....	28
2.5.2. Detecting telecommunication fraud by understanding the contents of a call de Qianqian Zhao, Kai Chen, Tongxin Li, Yi Yang & XiaoFeng Wang .....	30
2.6. Conclusion .....	31
<b>3. Conception</b> .....	<b>32</b>
3.1. Introduction .....	32
3.2. Données et prétraitement .....	33
3.2.1. Collecte de données .....	33
3.2.2. Prétraitement des données .....	37
3.3. Type d'apprentissage .....	43
3.3.1. Variable cible .....	43

3.4.	Apprentissage automatique et prédiction du taux NPS.....	45
3.5.	Ingénierie du logiciel .....	46
3.5.1.	Spécification des besoins fonctionnels .....	46
3.5.2.	Spécification des besoins fonctionnels .....	47
3.6.	Conclusion.....	47
4.	Réalisation et évaluation .....	48
4.1.	Introduction .....	48
4.2.	Environnement de développement.....	48
4.2.1.	Matériels et outils de développement : .....	48
4.2.2.	Application de mesure de satisfaction client.....	51
4.3.	Évaluation du système .....	63
4.3.1.	Dataset et prétraitement.....	63
4.3.2.	Métriques d'évaluations .....	69
4.3.3.	Résultat des évaluations .....	71
3.4.	Discussion des résultats.....	75
4.4.	Conclusion.....	76
	Conclusion Générale .....	77
	Perspectives .....	78
	Perspectives à court terme.....	78
	Perspectives à moyen terme.....	78
	Perspectives à long terme.....	78
	Bibliographie.....	79

# Liste des figures :

FIGURE 1. 1 : LOGO DE L'ENTREPRISE DJEZZY .....	1
FIGURE 1. 2 : SIEGE DE L'ENTREPRISE DJEZZY DAR-EL-BEIDA .....	2
FIGURE 1. 3 : ORGANIGRAMME OTA .....	3
FIGURE 1. 4 : STRUCTURE DU SERVICE BIGDATA .....	4
FIGURE 1. 5 : SOURCES DE DONNEES DE DJEZZY .....	6
FIGURE 1. 6 : SCHEMA DU SYSTEME ACTUEL .....	6
FIGURE 1. 7 : SCHEMA SIMPLIFIE DU SYSTEME ACTUEL .....	7
FIGURE 1. 8 : CALCULE DU NPS .....	9
FIGURE 1. 9 : INTERACTIONS CLIENTS/ENTREPRISE RELATIVEMENT AUX DIFFERENTS TYPES DE QUALITE .....	10
FIGURE 2. 1: TYPE D'APPRENTISSAGE .....	12
FIGURE 2. 2: EXEMPLE DE REGRESSION .....	13
FIGURE 2. 4 : ALGORITHMES D'APPRENTISSAGE SUPERVISE .....	14
FIGURE 2. 6 : TYPES DE REGRESSION LINEAIRE .....	14
FIGURE 2. 8 : VISUALISATION REGRESSION .....	15
FIGURE 2. 10 : EXEMPLE VISUALISATION REGRESSION MULTIPLE .....	16
FIGURE 2. 12 : FORMULE REGRESSION LOGISTIQUE .....	18
FIGURE 2. 13 : SVM DE L'HYPERPLAN OPTIMAL .....	19
FIGURE 2. 15 : HYPERPLAN OPTIMAL NON LINEAIRE .....	20
FIGURE 2. 16 : ALGORITHMES D'APPRENTISSAGE NON SUPERVISE .....	22
FIGURE 2. 17 : LA PROGRESSION DU CLUSTERING AGGLOMERATIF SUR DES DONNEES DE DEUX DIMENSIONS .....	24
FIGURE 2. 18 : EXEMPLE D'UN DENDROGRAMME .....	24
FIGURE 2. 19 : ARCHITECTURE RESEAU DE NEURONE .....	25
FIGURE 2. 20 : EXEMPLE DU PLUS PETIT RESEAU DE NEURONES .....	26
FIGURE 2. 21 : ARCHITECTURE SYSTEME DU PROJET CUSTOMER CHURN ANALYSIS IN TELECOM INDUSTRY .....	29
FIGURE 2. 22 : ARCHITECTURE GLOBALE DU SYSTEME POUR LA PREDICTION DES FRAUDES PAR APPELS .....	30
FIGURE 3. 1 : SCHEMA CONCEPTUEL GLOBALE DE L'ARCHITECTURE PROPOSEE .....	32
FIGURE 3. 2 : TYPES DE DONNEES GLOBALES .....	34
FIGURE 3. 3 : PROCESSUS DES ACTIONS DE TRANSFORMATIONS .....	35
FIGURE 3. 4 : ATTRIBUTS DE DONNEES .....	36
FIGURE 3. 5 : EXEMPLES DE CAS DE PRETRAITEMENTS .....	37
FIGURE 3. 6 : COURBE DES DONNEES MANQUANTES .....	38
FIGURE 3. 7 : EXEMPLE REEL DE DONNEE MANQUANTE .....	38
FIGURE 3. 8 : EXEMPLE REEL DE DONNEES BRUTEES .....	40
FIGURE 3. 9 : EXEMPLE REEL D'UNE INTEGRATION DE DONNEES .....	41
FIGURE 3. 10 : EXEMPLE REEL DE LA GESTION DES REDONDANCES .....	42
FIGURE 3. 11 : DECOUPAGE DES DONNEES D'ENTRAINEMENT ET DE TEST .....	45
FIGURE 4. 1: LOGO DU LANGUAGE PYTHON .....	49
FIGURE 4. 2: LOGO DU L'OUTIL JUPYTER .....	51
FIGURE 4. 3: CAPTURE – ADMIN DJANGO LOGIN PAGE .....	51
FIGURE 4. 4: CAPTURE – CREATESUPERUSER .....	52
FIGURE 4. 5: CAPTURE – ADMIN DASHBOARD .....	52
FIGURE 4. 6: CAPTURE – USER LOGIN .....	52
FIGURE 4. 7: CAPTURE – USER LOGIN – INFOS .....	53
FIGURE 4. 8: CAPTURE – LANDING DASHBOARD .....	53
FIGURE 4. 9: CAPTURE - PAGE TRAINING DATA .....	54
FIGURE 4. 10: CAPTURE - PROCESSUS D'ENTRAINEMENT .....	54



FIGURE 4. 11: CAPTURE - PAGE TESTING DATA .....	55
FIGURE 4. 12: CAPTURE – EXEMPLE D’UN RESULTAT DE TESTS FORMAT PDF .....	55
FIGURE 4. 13: CAPTURE – STOCKAGE DES RESULTATS SOUS FORMAT CSV .....	56
FIGURE 4. 14: CAPTURE – DASHBOARD DES RESULTATS .....	57
FIGURE 4. 15: EXEMPLE GRAPHE – OVERVIEW OF SATISFACTION.....	58
FIGURE 4. 16: EXEMPLE GRAPHE – SATISFACTION TRENDS BY DATES .....	58
FIGURE 4. 17: EXEMPLE GRAPHE – TOP 7 WILAYA’S CUSTOMER SATISFACTION .....	59
FIGURE 4. 18: EXEMPLE TABLEAU DES RESULTATS CLASSE PAR WILAYAS .....	59
FIGURE 4. 19: CAPTURE – MENU .....	60
FIGURE 4. 20: SELECTION DES RESULTATS PAR DATE.....	60
FIGURE 4. 21: EXEMPLE D’UNE SELECTION DES RESULTATS PAR WILAYA.....	61
FIGURE 4. 22: CAPTURE – PAGE PREDICT BY PARAMETERS.....	62
FIGURE 4. 23: DISTRIBUTION DES TROIS VALEURS « MOBILE_DATA_TECHNOLOGY ».....	63
FIGURE 4. 24: EXEMPLE REEL D’UNE JOINTURE DE FICHIERS.....	64
FIGURE 4. 25: EXEMPLE REEL D’UN CHANGEMENT DE NOMINATION.....	64
FIGURE 4. 26: EXEMPLE REEL DU RESULTAT DE JOINTURE .....	64
FIGURE 4. 27: EXEMPLE REEL DE QUELQUES VALEURS QUI SIGNIFIENT L’AGE DE CHAQUE ABONNE .....	65
FIGURE 4. 28: EXEMPLE REEL D’UNE VALEUR REDONDANTE.....	65
FIGURE 4. 29: POURCENTAGE DE DONNEES MANQUANTES POUR CHAQUE VARIABLE DU FICHIER User_INFO .....	66
FIGURE 4. 30: CINQ OBSERVATIONS DE DEUX VARIABLES NUMERIQUES CONTINUES USAGE_VOICE ET USAGE_DATA.....	67
FIGURE 4. 31: CORRELATIONS ENTRE LES VARIABLES .....	67
FIGURE 4. 32: RESULTATS DU REMPLISSAGE DE DONNEES MANQUANTES .....	68
FIGURE 4. 33: EXEMPLE D’UNE VARIABLE AVEC VALEURS DOMINANTES.....	68
FIGURE 4. 34: CIBLES DE DONNEES .....	69
FIGURE 4. 35: EXEMPLE REEL DE 10 OBSERVATIONS DE L’ETAT FINAL DU DATASET .....	69
FIGURE 4. 36: GRAPHE DE DEUX METRIQUES D’EVALUATIONS – TAUX D’ERREURS .....	76
FIGURE 4. 37: PERSPECTIVES A MOYEN TERME.....	78

# Liste des tableaux :

TABLEAU 2. 1 : FONCTIONS D'ACTIVATION .....	26
TABLEAU 2. 2 : FACTEURS DES ALGORITHMES.....	27
TABLEAU 3. 1 : DIFFERENCE ENTRE VARIABLES CONTINUE SET VARIABLES DISCRETE .....	33
TABLEAU 3. 2 : EXEMPLE DE TYPES DE DONNEES .....	33
TABLEAU 3. 3 : UNE BREVE DESCRIPTION DES FICHIERS DE DONNEES .....	35
TABLEAU 3. 4 : DESCRIPTION DE QUELQUES ATTRIBUTS .....	36
TABLEAU 3. 5 : PROCESSUS DE CLASSIFICATIONS.....	45
TABLEAU 4. 2: SPECIFICATION MATERIELS .....	48
TABLEAU 4. 3: BIBLIOTHEQUE PYTHON UTILISES.....	50
TABLEAU 4. 4: TRAITEMENT DES DONNEES MANQUANTES .....	66
TABLEAU 4. 5: MATRICE DE CONFUSION .....	69
TABLEAU 4. 6: LES MESURES D'ÉVALUATIONS .....	70
TABLEAU 4. 7: ALGORITHMES TESTES EN HYPER-PARAMETRES .....	72
TABLEAU 4. 8: ÉVALUATION DE LA LOGISTIQUE REGRESSION .....	73
TABLEAU 4. 9: ÉVALUATION DE SVM .....	74
TABLEAU 4. 10: ÉVALUATION D'UN RESEAU DE NEURONES.....	75
TABLEAU 4. 11: LE RESTE DES EVALUATIONS.....	75
TABLEAU 4. 12: MEILLEUR RESULTAT OBTENU .....	75

## Liste des équations :

ÉQUATION 2. 1: ÉQUATION DE L'APPRENTISSAGE SUPERVISE .....	12
ÉQUATION 2. 2 : ÉQUATION REGRESSION LINEAIRE SIMPLE .....	15
ÉQUATION 2. 3 : ÉQUATION REGRESSION LINEAIRE MULTIPLE .....	15
ÉQUATION 2. 4 : ÉQUATION DE LA FONCTION DE PERTE .....	16
ÉQUATION 2. 5 : ÉQUATION DE LA METHODE DU GRADIENT .....	17
ÉQUATION 2. 6 : ÉQUATION DE LA FONCTION LOGISTIQUE .....	18
ÉQUATION 2. 7 : FONCTION DE PERTE ASSOCIE A LA REGRESSION LOGISTIQUE .....	18
ÉQUATION 2. 8 : FONCTION SVM.....	19
ÉQUATION 2. 9: THEOREME DE BAYES .....	20
ÉQUATION 2. 10: CLASSIFICATEUR NAÏVE BAYES.....	20
ÉQUATION 2. 11 : FORMULES DE FONCTIONS DE DISTANCE .....	21
ÉQUATION 3. 1: FORMULE DU Z-SCORE.....	40
ÉQUATION 3. 2: COEFFICIENT DE CORRELATION DE PEARSON.....	42
ÉQUATION 3. 3: FORMULE DU MIN-MAX.....	42
ÉQUATION 3. 4: FORMULE DU Z-SCORE .....	42

## Liste des abréviations :

**OTA:** Optinium Télécom Algérie

**IT:** Information Technology

**MMS:** Multimedia Messaging Service

**RBT:** Risk-based testing

**IVR:** Interactive Voice Response

**SMS:** Short Message Service

**USSD:** Unstructured Supplementary Service Data

**NPS:** Net Promoter Score

**API :** Application Programming Interface

**SVM :** Support Vector Machine

**ACP:** Analyse aux Composants Principaux

**KNN:** K-Nearest Neighbors

**KDD:** Knowledge Discovery in Databases

**IA :** Intelligence Artificiel

**CSV :** Comma-separated values

# Introduction générale

## Contexte :

Dans un environnement énormément concurrentiel qui est le domaine des télécoms, l'importance n'est plus seulement le prix des abonnements, et quelques technologies ne font presque plus la différence. La clé se trouve maintenant côté innovation, et surtout côté expérience et satisfaction client.

Aujourd'hui, la plupart des entreprises se rendent compte de l'importance de la satisfaction client, en le considérant comme étant un enjeu majeur pour leur développement et durabilité. En effet l'insatisfaction client coûte cher car d'après quelques études, 91% des clients insatisfaits ne refont généralement pas leurs abonnements auprès de leurs compagnies téléphoniques et ils se tourneront vers d'autres entreprises concurrentes, Ce qui entraîne des pertes directes sur le chiffre d'affaire de l'entreprise.

Afin de gagner un vrai avantage concurrentiel et dans le but de garder sa position tant que leader dans le marché des télécoms, L'entreprise Djazzy vise à tout prix d'assurer la fidélité de ses clients en ayant une idée sur leur niveau de satisfaction selon des critères différents, elle vise à mieux savoir ou, comment, à qui et quand perfectionner leurs services et rendre l'expérience de leurs abonnés plus satisfaisante. La mesure de satisfaction client permet d'identifier les facteurs d'insatisfaction, ainsi Djazzy pourra mettre en place des actions d'amélioration nécessaires avant que les abonnés ne quittent la marque.

## Problématique :

Étant l'un des opérateurs de téléphonie mobile algérien, Djazzy dispose déjà d'un système de mesure de satisfaction client plutôt statique qui fait l'objet d'un sondage ; Elle souhaite automatiser cette solution selon une nouvelle logique basée sur des nouvelles technologies d'apprentissage automatique et définir ainsi un modèle de mesure plus précis. En effet, le système de mesure de satisfaction actuel représente un sondage envoyé par sms à des échantillons d'abonnés durant des périodes précises de l'année ; Ce sondage n'est qu'un questionnaire contenant plus de 10 questions et visant à avoir une idée approximative sur le taux de satisfaction de l'échantillon d'abonnés en question. Cependant ce système reste statique et vu que le taux de réponses des abonnés n'est pas élevé, l'entreprise se trouve en manque de données pour l'évaluation de satisfaction de ses clients.

## Objectif :

Afin de proposer une solution innovante et qui répond à la problématique posée, nous proposons une solution à base de machine learning et développant ainsi un système de

satisfaction client plus dynamique et capable de mesurer approximativement la satisfaction d'un échantillon d'abonnés à partir de ceux qui ont déjà fait part du sondage statique.

Notre mémoire sera structuré en quatre chapitres comme suit :

- Chapitre 1, concerne le contexte de notre étude, en présentant l'entreprise avec ses objectifs ainsi que la problématique posée en décrivant le système actuel ainsi que quelques orientations sur le système qui sera proposé.
- Chapitre 2 : présente une étude sur les algorithmes d'apprentissage automatique ainsi que quelques travaux liés à leur utilisation dans le domaine du Télécom.
- Chapitre 3, concerne la conception de notre système de prédiction de la satisfaction client à base de techniques de machine learning.
- Chapitre 4 : décrit le système développé ainsi que les évaluations effectuées sur un échantillon de données suite à plusieurs prétraitements.

Finalement, une conclusion générale rappelle les contributions de notre travail avec quelques perspectives futures.

# 1. Contexte d'étude

## 1.1. Introduction

Dans ce chapitre nous allons commencer par présenter l'entreprise, ses différents services, objectifs ainsi que son schéma d'organisation, puis nous décrirons avec plus de détails notre problématique, en donnant une vue sur le système courant avec ses inconvénients. Finalement, nous allons donner quelques orientations sur notre solution qui sera détaillée dans les prochains chapitres.

## 1.2. Présentation de l'entreprise

### 1.2.1. Optimum Telecom Algérie

Djezzy, officiellement Optimum Télécom Algérie (OTA), opérateur de télécommunications algérien créé en juillet 2001, il a lancé ses activités en février 2002. Leader des technologies de communications numériques, l'entreprise fournit une vaste gamme de services tels que le prépayés, le post-payé, le Data ainsi que les services à valeur ajoutée et le Service Universel des Télécommunications.



Figure 1. 1 : Logo de l'entreprise Djezzy

En janvier 2015, Orascom Télécom Algérie est officiellement devenu Optimum Télécom Algérie. En effet, OTA n'a pas changé uniquement de nom, mais aussi de statut et d'actionnaires. La nouvelle société est détenue alors à 51% par des opérateurs Algériens et à 49% par l'opérateur international de la téléphonie mobile Russo-Norvégien Vimplecom, selon la règle controversée de 51% 49% imposée pour tout investisseur étranger

Djezzy couvre 95 % de la population à travers le territoire national et ses services 3G sont déployés dans les 48 wilayas. Djezzy a déployé ses services 4G dans 28 wilayas au 31 décembre 2017 avec une couverture de 25% de la population s'est engagée à couvrir plus de 50% de la population à l'horizon 2021.

Djezzy est engagée dans un Processus de Transformation pour devenir l'opérateur numérique de référence en Algérie.

L'entreprise est dirigée par Vincenzo Nesci Président Exécutif et Matthieu Galvani, Directeur Général.

La vision principale de l'entreprise, selon les déclarations du comité exécutif de l'OTA, est d'être l'Opérateur de Télécommunications préféré des Algériens, leader sur son marché, apportant constamment de la valeur à tous ses partenaires. Elle désire être une référence pour son orientation client et la qualité de son environnement de travail. [1]



Figure 1. 4 : Siège de l'entreprise Djezzy Dar-El-Beida

### 1.2.2. Objectifs de l'OTA

Parmi l'ensemble d'objectifs de l'OTA, on cite :

- Offrir des produits de qualité à des prix compétitifs.
- Introduire les nouvelles technologies.
- Déployer des infrastructures à la pointe de la technologie.
- Satisfaire les besoins de ses clients
- Appliquer rigoureusement sa politique environnementale.
- Créer pour ses employés le meilleur environnement de travail et d'épanouissement.
- Améliorer sans cesse son processus internes dans le respect de sa politique qualité.[1]



### 1.2.3. Organigramme de l'OTA

OTA dispose de plusieurs services, notre mission aura lieu sur le service IT.

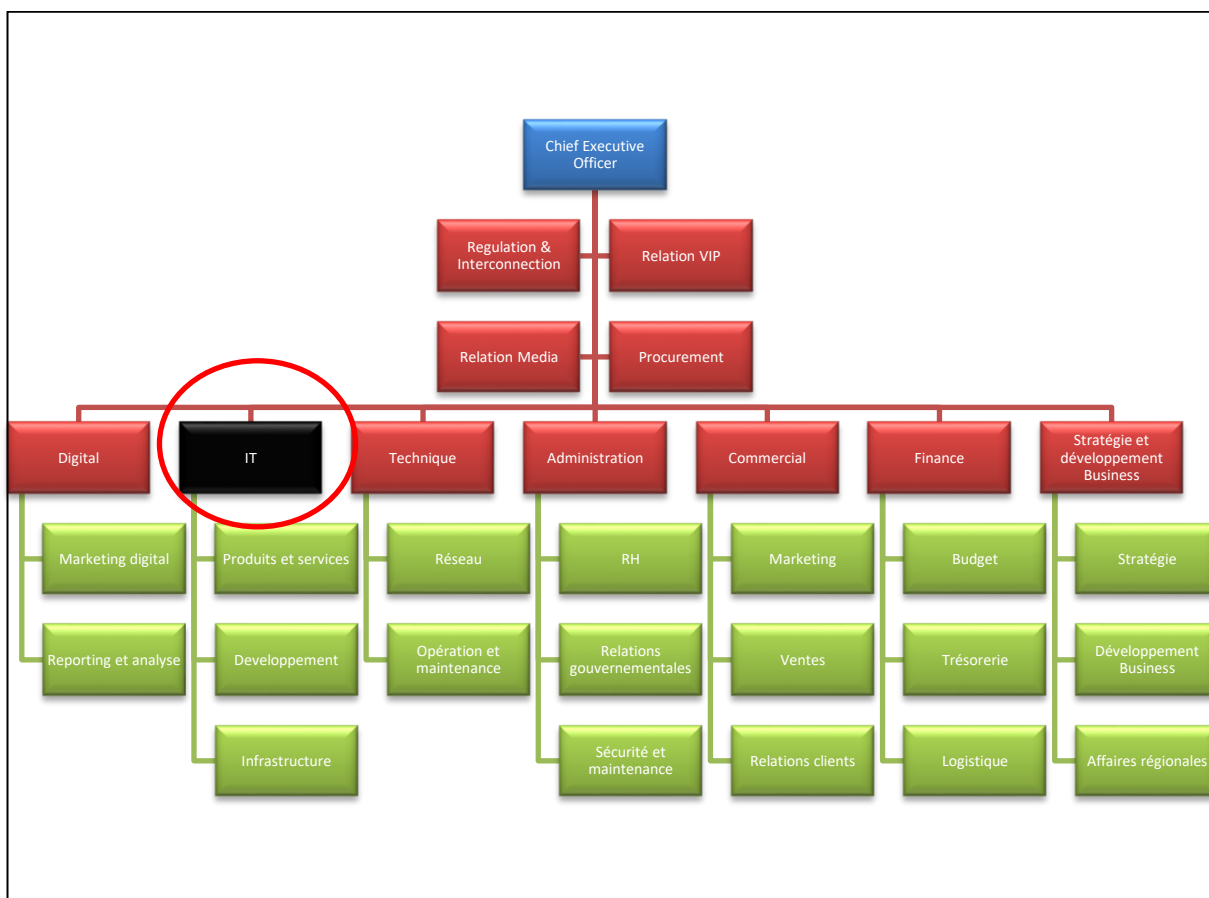


Figure 1. 7 : Organigramme OTA

## 1.2.4. Présentation du service

### 1.2.4.1. Département IT (technologies et systèmes d'informations)

Ce département se charge de développer les outils de gestions clients ainsi que les outils informatiques de l'entreprise. Il assure la maintenance des applications. Il se charge aussi du routage intelligent des appels ainsi que de la gestion multimédia (voix, web et emails).

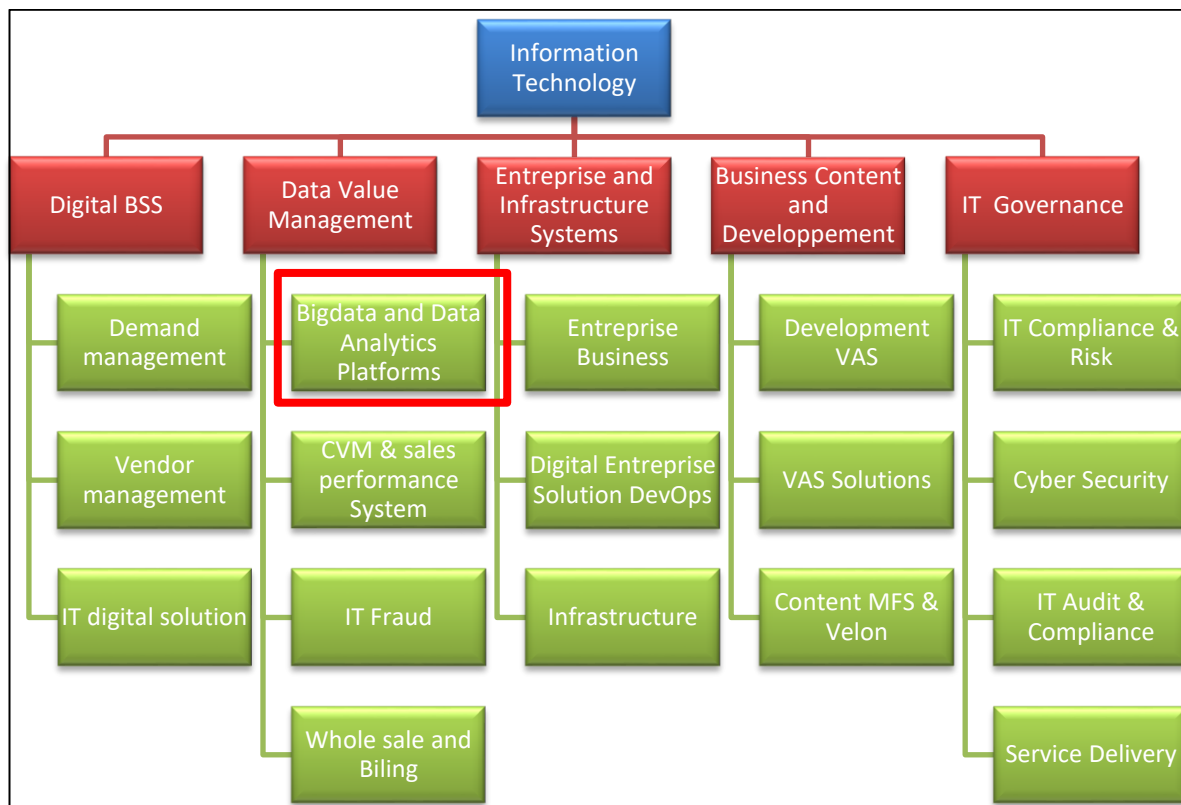


Figure 1. 10 : Structure du service BigData

Notre stage se positionne dans le département IT au sein du service **Data Value Management**.

Ce service a comme principales missions :

a) **Comprendre et cibler les abonnés** : Le département Data Value Management vise à aider Djezzy à mieux comprendre ses abonnés en collectant les données de l'entreprise pour mieux les analyser et les étudier par la suite. Comme il assure les opérations d'extraction et d'agrégation des données pour des fins décisionnelles et de pilotage.

b) **Aider les décideurs à prendre des décisions stratégiques** : Fournir un ensemble de données et d'analyses aux décideurs afin de mettre à leur disposition toutes les informations relatives à l'entreprise, au marché, aux clients, aux télécoms et à la concurrence et ce pour les aider à prendre des décisions data-driven.

c) **Optimisation des coûts** : utiliser les technologies des big data pour optimiser les coûts et les dépenses de l'entreprise.

d) **Améliorer l'expérience client** : Aider Djezzy à se rapprocher de son client et à améliorer son expérience avec l'entreprise en étant toujours à son écoute et en observant attentivement son comportement.

e) **Automatisation** : Rajouter de l'intelligence à la donnée afin d'automatiser certains processus et services de l'entreprise.

#### 1.2.4.2. Services de Djezzy

Parmi les services que Djezzy offre à ses utilisateurs :

- **La voix** : Le service primordial qui fait de Djezzy un opérateur mobile.
- **MMS** : Le service MMS fournit le contenu des SMS comprenant des images, des audio et des vidéos.
- **RBT** : appelé souvent RANATI, C'est un service qui permet de choisir la sonnerie préférée que l'utilisateur souhaite avoir lorsqu'on lui appelle.
- **IVR** : L'utilisation de l'IVR et de l'automatisation vocale permet aux requêtes des appelants d'être résolues sans qu'il soit nécessaire de mettre en file d'attente et de supporter le coût d'un agent actif.
- **Voice SMS** : Le service permet aux clients de préfixer le numéro de l'appelé souhaité avec un code d'accès prédéfini et de déposer le message dans le système de messagerie vocale.
- **Internet** : inclut les services de la 3G et 4G.
- **SMS** : Le service permet aux clients de s'échanger des messages courts
- **USSD** : Le service ressemble au service SMS sauf qu'il ne garde pas de mémoire

### 1.3. Données et système actuel

#### 1.3.1. Source de données

Djezzy a adopté une solution informatique qui lui permet de collecter les données de ses abonnés ainsi que les données relatives à son réseau et à son activité. Cette solution permet à l'opérateur de télécommunication d'améliorer la qualité de ses services, d'optimiser son efficacité opérationnelle, de réduire le taux de désabonnement et de générer des revenus. La solution comprend une surveillance et dépannage en temps réel, indépendante du fournisseur, et optimise les réseaux de bout en bout. La combinaison unique de ces produits et services fournit une optimisation automatisée, des informations géo localisées exploitables et des analyses normatives aux équipes d'opérations, réseau, Centre d'exploitation des services, Assistance clientèle et Marketing de Djezzy.

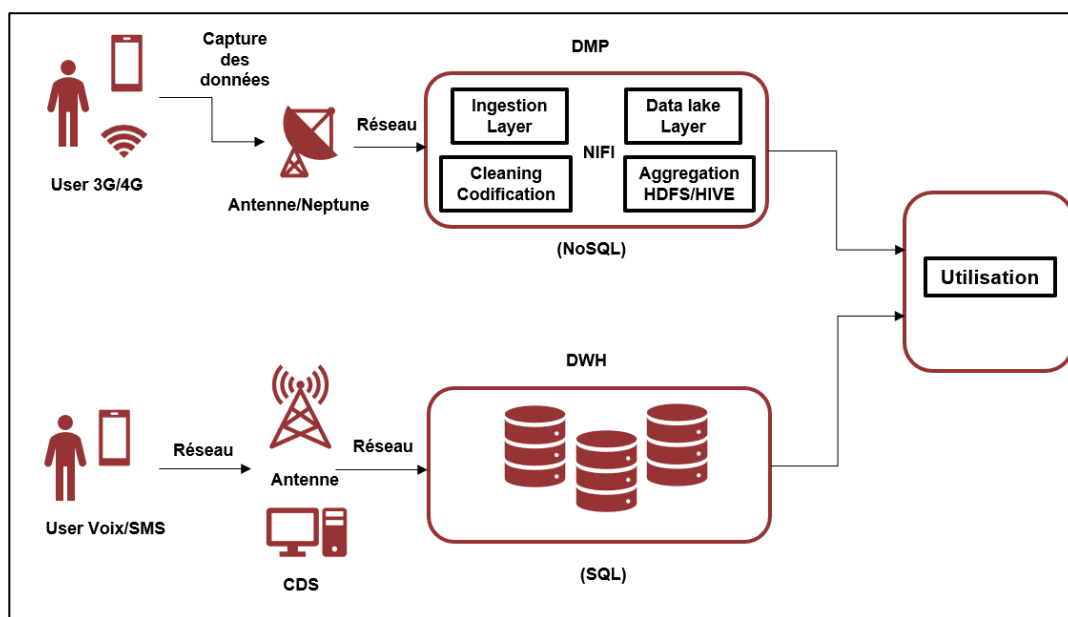


Figure 1. 13 : Sources de données de Djezzy

### 1.3.2. Système actuel

Le système actuel est un processus qui a été appliqué afin d'obtenir des données récentes prête à utiliser, on résume sur le schéma suivant :

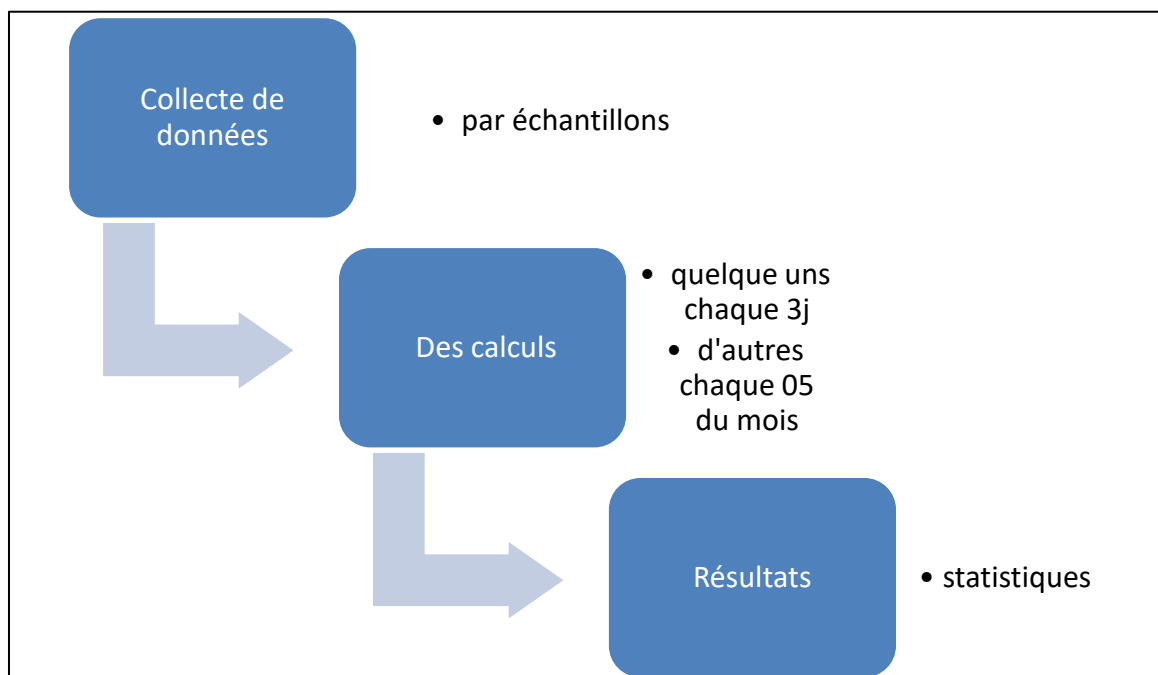


Figure 1. 16 : Schéma du système actuel

La première étape consiste à une collecte de différentes données par échantillons de clients, ces échantillons change en fonction du mois ; par la suite des calculs seront établis en deux bases :

**Base J/M :** pour les nouveaux abonnés, le calcul se fait en comptant 11 jours à partir de la première inscription, pour le reste des abonnés le calcul se fait chaque 3j principalement sur les abonnés actifs, centre d'appel, boutique

**Base M :** ce calcul va se faire une fois chaque mois (5 du mois) qui se concerne principalement les données réseaux (localisation, réseau de l'utilisation) ainsi que l'utilisation data et voix des abonnés.

Une deuxième étape consiste à récapituler les résultats et les stocker sur une base de données. Ensuite, un sondage sera envoyé aux abonnés (trouvés actifs à partir des calculs faits par Djezzy), la classe NPS de chaque abonné sera calculé et mesuré selon les réponses, et en se basant uniquement sur le résultat du sondage, des mesures seront prises en considération afin d'améliorer la satisfaction client

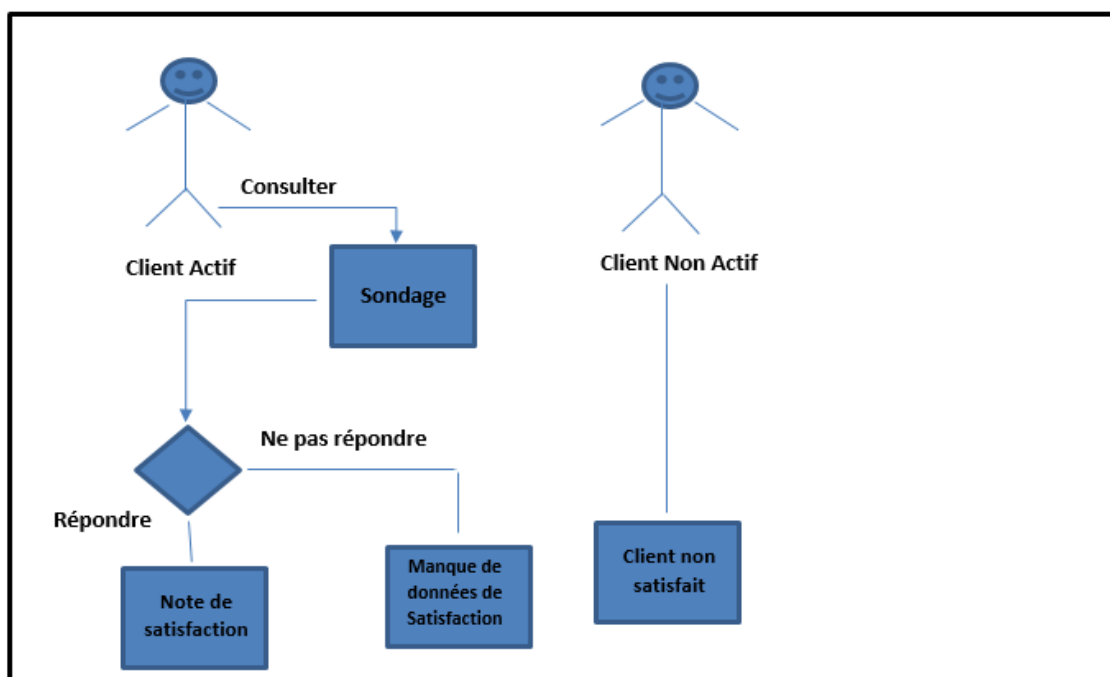


Figure 1. 19 : Schéma simplifié du système actuel

Le sondage est créé par l'organisme, les questions du sondage varient, mais généralement il y en a quelques unes qui ne changent pas pour la plupart des compagnies, la question suivante en fait partie :

A quel point vous recommandez de cette entreprise à un ami ou à un collègue ?

Pas du tout

Très d'accord

<b>0</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>	<b>10</b>
----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	-----------

### 1.3.3. Problématique et motivations

Parmi les projets à valeurs ajoutés de Djezzy, le NPS (Net Promoter Score) où il existe un système basique qui a comme rôle de collecter les avis des clients sur un ou plusieurs offres proposées par l'Entreprise, l'information sera générée à partir d'un sondage (solution de bulk SMS) ce qui permet de mesurer approximativement le degré de satisfaction d'un échantillon de clients qui répondent à ce sondage, cette proposition reste donc statique car elle n'est pas garanti à tous les coups, et la réponse des abonnés n'est pas assurée, il nécessite donc l'utilisation d'une nouvelle technologie basé sur l'apprentissage automatique afin de proposer un nouveau modèle permettant de prédire la satisfaction d'un échantillon d'abonnés qui n'ont pas pris part au sondage, et ceci en se basant sur les expériences réelles de l'échantillon des abonnés actifs qui ont répondu.

Dans une démarche de protection de données personnelles des abonnés, certaines données peuvent être garanties aux employés et stagiaires qui travaillent physiquement au sein de l'entreprise Djezzy, et encore moins de données qui peuvent être garanties aux employés et stagiaires qui travaillent par distance et les données NPS n'en font pas partie, par conséquent, la réalisation de ce projet fait face à une nouvelle problématique qui est le manque du score NPS de chaque client.

Notre mission sera donc de générer le score NPS d'une façon logique et non aléatoire pour chaque client, et de réaliser un système qui permet le ciblage des clients fiables fournis par Djezzy en se basant sur des technologies de Machine Learning afin d'obtenir un taux de satisfaction de tout l'échantillon et les valeurs des critères qui l'influencent directement ou indirectement sur les différentes régions d'Algérie et donc Djezzy pourra améliorer ce degré à travers les résultats obtenus.

### 1.3.4. Mesure de NPS (Net Promoter Score)

#### 1.3.4.1. Description de NPS

Le Net Promoter Score est un indicateur de fidélité client développé en 2003 par le consultant Fred Reichheld de Bain & Company en collaboration avec l'entreprise Satmetrix.

L'objectif était de déterminer un score uniforme et facilement interprétable pour la satisfaction client qui peut être comparé au fil du temps ou entre différentes industries.

Le NPS évalue dans quelle mesure le répondant recommande une certaine société, produit ou service à ses amis, ses proches ou ses collègues. L'idée est simple : si vous aimez utiliser un produit ou faire des affaires avec une entreprise particulière, vous voulez bien partager cette expérience avec des autres. [2]

### 1.3.4.2. Calcule du NPS

Le Net Promoter Score permet de classer les clients sur 3 positions suivant leur degré d'enthousiasme :

Promoteurs = répondants donnant un score de **9 ou 10**

Passifs = répondants donnant un score de **7 ou 8**

Détracteurs = répondants donnant un score de **0 à 6**

**Le calcul se fait avec la formule :  $NPS = \% \text{ Promoteurs} - \% \text{ Détracteurs}$**

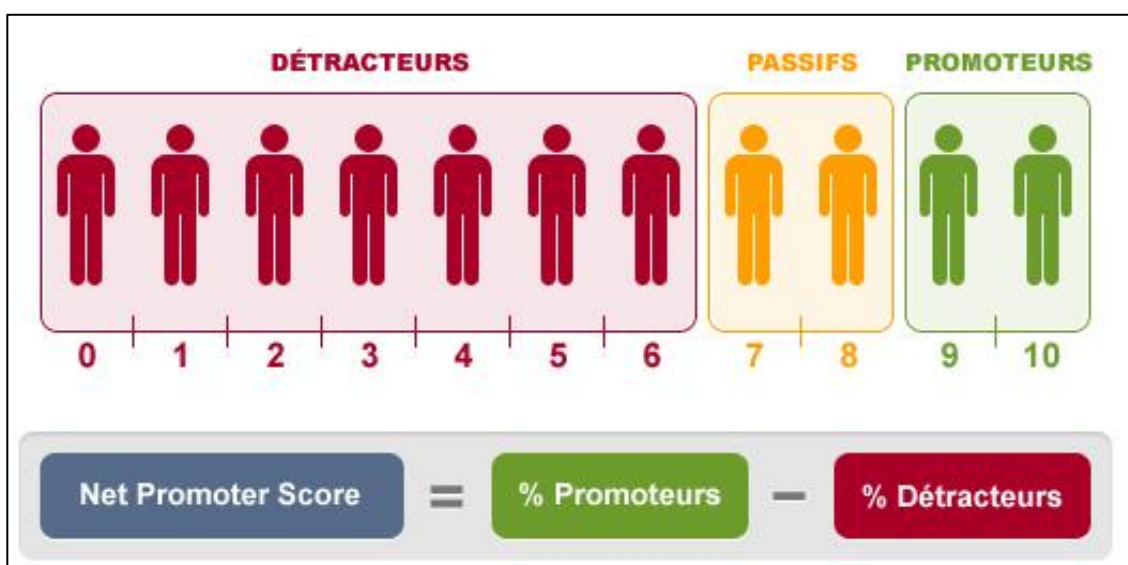


Figure 1. 22 : Calcule du NPS

Le NPS ne s'exprime pas en pourcentage, mais comme nombre absolu qui se situe entre -100 et +100.

### 1.3.4.3. Application du NPS

Le NPS est actuellement utilisé par de nombreuses grandes entreprises comme outil pour mesurer la satisfaction client, il s'agit d'un nombre unique qui est clair et facile à comprendre pour tous les employés et très utile pour les dirigeants de l'entreprise, certaines sources le considèrent comme une bonne indication du potentiel de croissance et de la fidélité client pour une entreprise ou un produit.

Pour comprendre les motivations des Promoteurs et des Détracteurs, il est recommandé d'accompagner la question NPS par une ou plusieurs questions ouvertes qui recherchent les raisons sous-jacentes pour le score donné. Cela vous permet d'effectuer les ajustements nécessaires pour augmenter le NPS dans l'avenir, soit en augmentant le pourcentage de Promoteurs, soit en réduisant la proportion de Passifs et Détracteurs (ou, mieux encore, une combinaison des deux). [2]

## 1.4. Satisfaction client

Daniel Ray explique l'interaction entre un client et une entreprise relativement aux différents types de qualité. Voyons un schéma adapté de son ouvrage *Mesurer et développer la satisfaction clients* [3] :

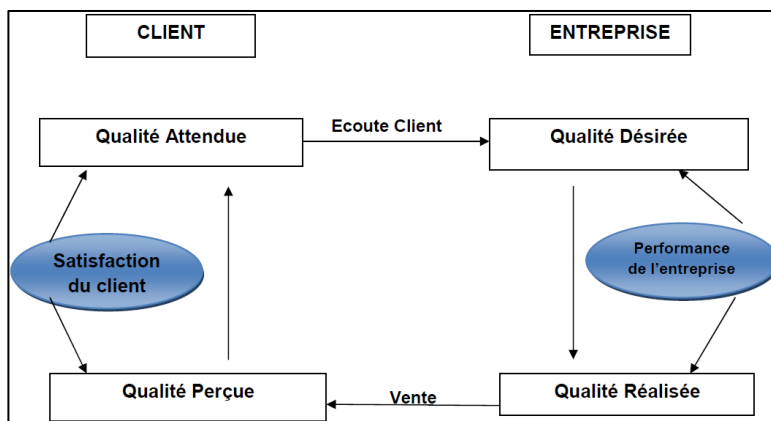


Figure 1. 25 : Interactions clients/entreprise relativement aux différents types de qualité

À l'origine, le client possède des attentes (qualité attendue) que l'entreprise écoute et interprète (qualité désirée, ce qu'elle vise à l'issue de ses processus internes pour pouvoir répondre aux attentes du client), ensuite, l'entreprise tente de répondre à cette demande en transformant cette qualité désirée en qualité réalisée.

Dans une troisième étape, celle-ci est transmise (communiquée), un processus qui permet au client de construire sa perception de la qualité perçue. Enfin, la qualité perçue est comparée avec les attentes, générant ainsi le sentiment de satisfaction du client. [3]

Les critères de satisfaction sont donc la clé pour garder les clients et nous nous intéressons justement sur un système de mesures.

## 1.5. Conclusion

Dans ce chapitre nous avons présenté l'organisme d'accueil, posé nos problématiques et notre solution et expliqué quelques notions importantes qui concerne notre sujet, le prochain chapitre traitera le côté technique de la technologie utilisé pour la gestion et la prédiction du taux de satisfaction des abonnés sur le domaine concerné.



## **2. Apprentissage automatique et Télécom**

### **2.1. Introduction**

Le secteur des télécommunications est un secteur en évolution constante. Aujourd'hui, Djezzy est en quête continue pour intégrer les avancées technologiques dans ses services et d'assurer ainsi une prestation toujours meilleure en qualité, l'apprentissage automatique est absolument une de ces avancées technologiques.

Le deuxième chapitre du mémoire consiste à définir l'apprentissage automatique ainsi que tous ses types, décrire les différents algorithmes et méthodes de chaque type, leur objectifs, quelques exemples d'algorithmes à utiliser.

### **2.2. Apprentissage automatique**

#### **2.2.1. Définition**

L'apprentissage automatique « Machine Learning » est une méthode utilisée en intelligence artificielle, il s'agit d'algorithmes qui analysent un ensemble de données afin de déduire des règles qui constituent de nouvelles connaissances permettant d'analyser de nouvelles situations. [4]

C'est aussi l'implémentation de méthodes qui permettent à une machine d'évoluer grâce à un processus d'apprentissage, et ainsi de remplir des tâches qu'il est difficile ou impossible de remplir par des moyens algorithmiques plus classiques.

#### **2.2.2. Types d'apprentissage**

Les algorithmes d'apprentissage peuvent se catégoriser selon le type d'apprentissage qu'ils emploient, nous pouvons distinguer 4 catégories :

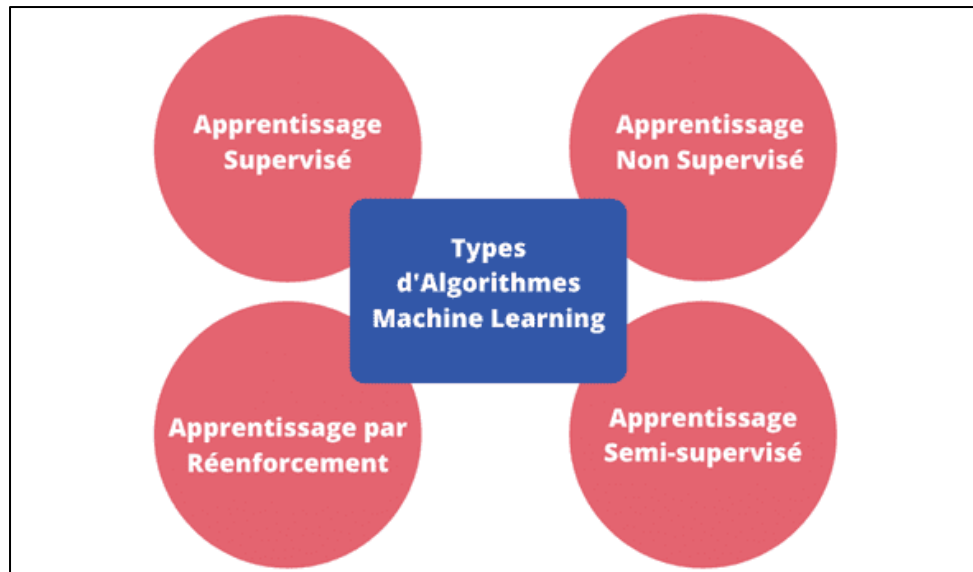


Figure 2. 2: Type d'apprentissage

### 2.2.2.1. Apprentissage Supervisé

L'apprentissage supervisé consiste en des variables d'entrée ( $X$ ) et une variable de sortie ( $Y$ ). Vous utilisez un algorithme pour apprendre la fonction, de l'entrée à la sortie.

$$Y = f(X)$$

#### Équation 2. 2: Équation de l'apprentissage supervisé

L'apprentissage supervisé est utilisé dans le cas où on dispose des connaissances préalables de ce que devraient être les valeurs de sortie de nos échantillons (valeurs cibles). Le but est d'appréhender si bien la fonction que lorsque vous avez de nouvelles données d'entrée ( $x$ ), vous pouvez prédire les variables de sortie ( $Y$ ) pour ces données.

Un algorithme d'apprentissage supervisé prend en entrée un échantillon de données connu, et une valeur cible connue, et forme un modèle pour générer des prédictions raisonnables comme réponses aux nouvelles données. L'apprentissage supervisé est généralement effectué dans le contexte de la classification et de la régression. [5]

La figure suivante représente un exemple de régression

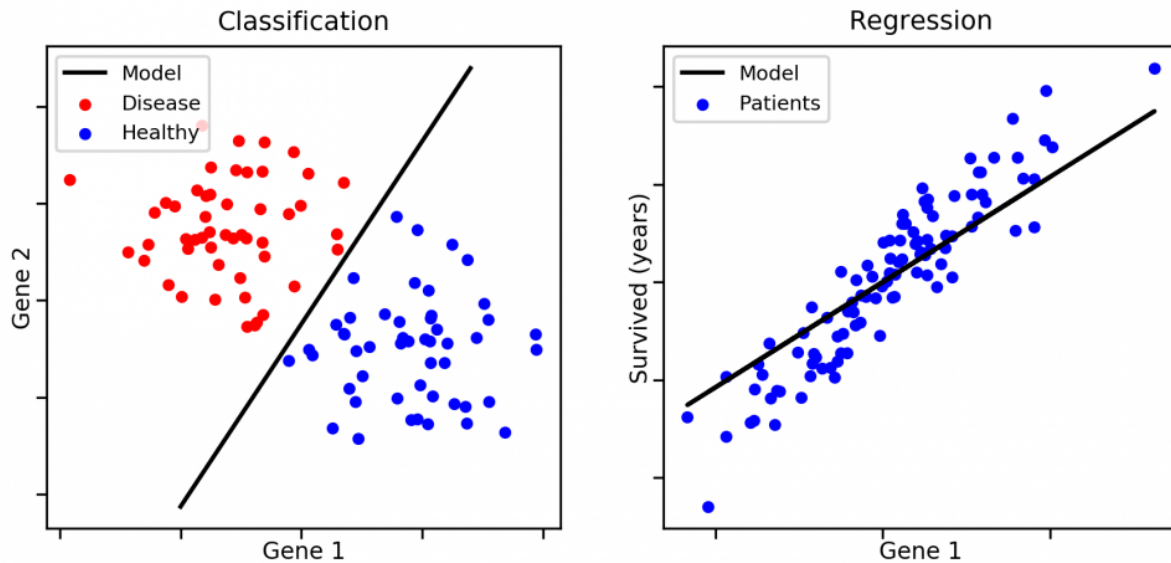


Figure 2. 5: Exemple de régression

#### 2.2.2.2. Apprentissage Non Supervisé

L'apprentissage non supervisé consiste à ne disposer que de données d'entrée ( $X$ ) et pas de variable de sortie correspondantes. On dispose d'un ensemble d'objets sans aucune valeur cible associée ; il faut apprendre un modèle capable d'extraire les régularités présentes au sein des objets pour mieux viser.

L'apprentissage non supervisé comprend deux catégories d'algorithmes : Algorithmes de regroupement (Clustering) et réduction de dimension.

#### 2.2.2.3. L'apprentissage semi-supervisé

Effectué de manière probabiliste ou non, il vise à faire apparaître la distribution sous-jacente des « exemples » dans leur espace de description. Il est mis en œuvre quand des données (ou « étiquettes ») manquent... Le modèle doit utiliser des exemples non-étiquetés pouvant néanmoins renseigner.

#### 2.2.2.4. L'apprentissage par renforcement

L'algorithme apprend un comportement étant donné une observation. L'action de l'algorithme sur l'environnement produit une valeur de retour qui guide l'algorithme d'apprentissage.

### 2.3. Algorithmes d'apprentissage automatique

Nous citons des exemples d'algorithmes qui concerne les deux premières catégories (supervisé et non supervisé), et qui sont majoritairement utilisés par rapport aux deux autres types.

### 2.3.1. Algorithmes d'apprentissage Supervisé

Dans ce qui suit, la majorité des algorithmes de classification et de régression seront définis

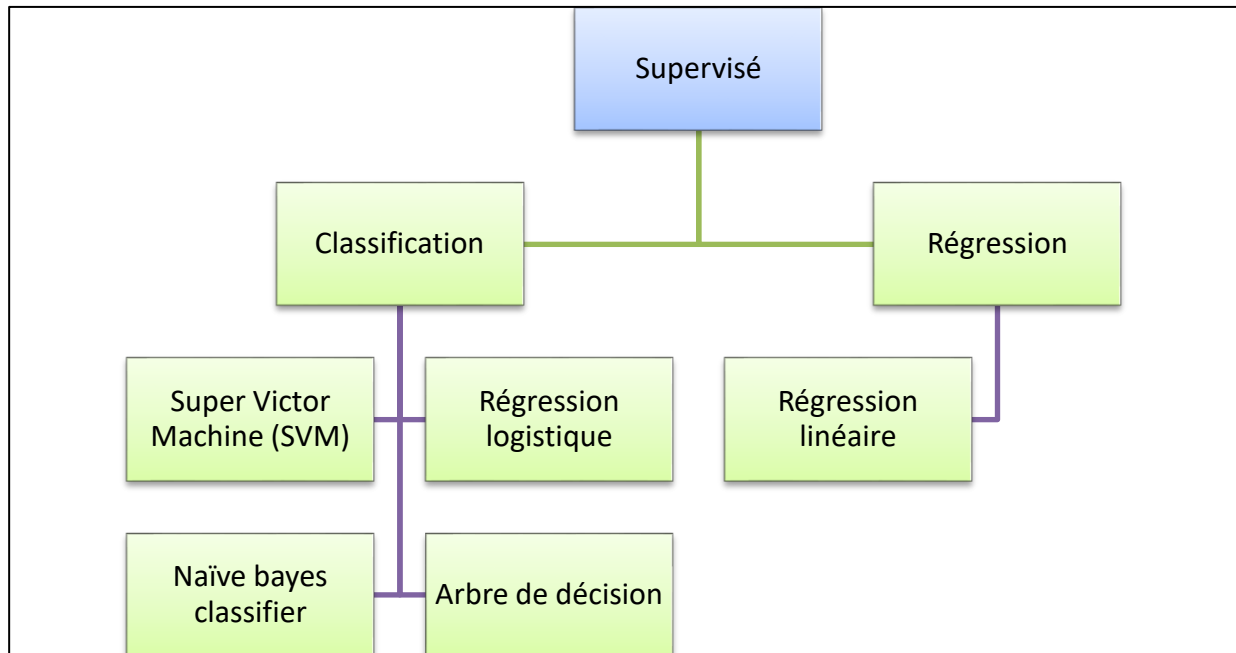


Figure 2. 8 : Algorithmes d'Apprentissage Supervisé

#### 2.3.1.1. Régression

C'est l'une des techniques les plus connues, elle fait généralement partie des premiers sujets choisis lors de l'apprentissage prédictive. Dans cette technique, la variable dépendante est continue, la ou les variables indépendantes peuvent être continues ou discrètes et la nature de la droite de régression est linéaire.

La régression linéaire établit une relation entre la variable dépendante  $Y$  et une ou plusieurs variables indépendantes  $X$  en utilisant une droite de meilleur ajustement (également appelée ligne de régression). [6]

Il existe deux types de régression linéaire, Simple et multiple :

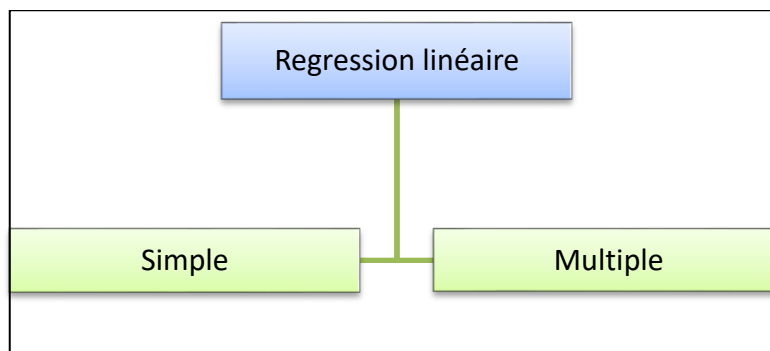


Figure 2. 11 : Types de régression linéaire

### ➤ Régression linéaire simple :

Ce type de régression linéaire ne comporte qu'une seule variable indépendante, il est représenté par l'équation mathématique suivante :

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

$$\theta_i: \text{paramètres}, i \in \{0,1\}$$

Équation 2. 5 : Équation régression linéaire simple

Où  $h_{\theta}$  de  $x$  est la variable dépendante (certaines sources la nomme  $y$ ),  $x$  est la variable indépendante,  $\theta_0$  et  $\theta_1$  sont deux variables constantes inconnues qui représentent l'intercepte et la pente de la ligne respectivement.

$\theta_0$  et  $\theta_1$  constituent la matrice de poids qui définissent le modèle.

Voici une visualisation d'un exemple qui figure la relation entre la température et la vente des glaces, avec les points de données et la ligne de régression

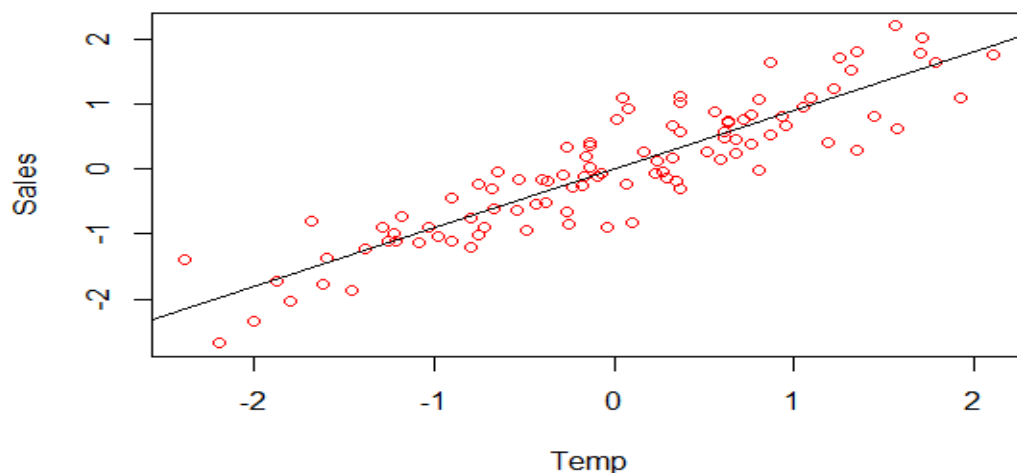


Figure 2. 14 : Visualisation régression

### ➤ Régression linéaire multiple :

Ce type de régression a plusieurs variables indépendantes, il est représenté par l'équation mathématique suivante :

$$h_{\theta}(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$$

$$\text{Paramètres: } \theta_i = \{\theta_0, \theta_1, \theta_2, \dots, \theta_n\}$$

$$\text{Variables: } x_i = \{x_0, x_1, x_2, \dots, x_n\}$$

Équation 2. 8 : Équation régression linéaire multiple

Où  $h(x)$  est toujours la variable dépendante et tous les  $x$  sont des variables indépendantes.

Voici une visualisation d'un exemple qui inclue deux variables qui affectent le résultat des ventes (sales)

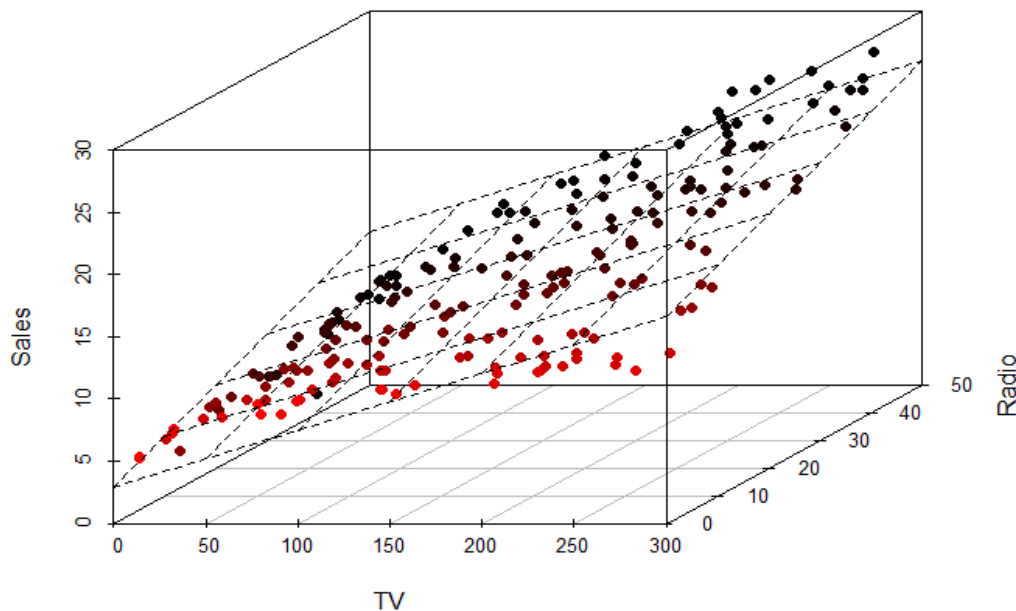


Figure 2. 17 : Exemple visualisation régression multiple

#### - Fonction de perte :

Elle peut être n'importe quelle équation qui peut nous donner une idée sur à quel point nous sommes prêts des vrais valeurs (l'hypothèse requis), plus la fonction de perte est élevée, plus on s'éloigne de la courbe requise, autrement dit, c'est une fonction qui nous permet de calculer le degré de précision d'un modèle d'apprentissage automatique.

Il existe plusieurs fonctions capables de nous permettre de calculer le taux de précision, mais généralement dans les problèmes de régression, la fonction de perte la plus populaire est la fonction de l'erreur quadratique moyenne (MEAN SQUARED ERROR), dont la formule :

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m \text{Erreur carrée des données } i$$

$(\underbrace{h_{\theta}(x^{(i)})}_{\text{Valeur prédite}}, \underbrace{y^{(i)}}_{\text{Vraie valeur}})^2$

Équation 2. 11 : Équation de la fonction de perte

Où  $J$  est la nomination de la fonction,  $m$  est le nombre de points de données,  $y$  est la valeur actuelle de la donnée  $i$ , et  $h(x)$  est la valeur prédite de la donnée  $i$ .

#### - Descente du gradient (Gradient descent)

La méthode de descente du gradient est une méthode itérative qui avance dans chaque itération avec un pas, appelé taux d'apprentissage, vers le minimum global de la fonction de perte. Une fois le minimum global est trouvé, le modèle obtient alors les meilleurs paramètres  $\theta$ .

L'équation suivante représente le pseudo code explicative de la méthode de descente du gradient :

*Répétez jusqu'à la convergence {*

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

*} tel que  $j$  représente le numéro d'index*

**Équation 2. 14 : Équation de la méthode du gradient**

Où

- ***Alpha*** est le taux d'apprentissage,
- ***J*** la fonction de perte et  $\theta$  la matrice des poids.

#### 2.3.1.2. Classification

Un problème de classification est similaire à un problème de régression, sauf que maintenant les prédictions sont des valeurs discrètes et non continues, c'est-à-dire un ensemble fini de valeurs (Vrai ou faux par exemple).

On peut noter qu'il y a deux types de classification, une classification avec seulement deux classes comme résultat, appelée binaire, et une autre avec plusieurs classes appelée multiclass.

##### 2.3.1.2.1. Régression logistique

C'est une approche statistique utilisée pour des problèmes de classification, elle est plutôt convenable aux types binaires qu'aux multi classes.

La régression logistique est connue par la fonction mathématique sigmoïde ayant un intervalle de valeurs  $[0,1]$  et définit par l'équation suivante :

$$h_{\theta}(x) = \frac{1}{1 + e^{(-\theta^T x)}}$$

[15]

Équation 2. 16 : Équation de la fonction logistique

$H(x)$  représente l'hypothèse, la probabilité qu'y soit égale à 1,  $x$  est une variable indépendante,  $\Theta$  est une variable constante inconnue qui représente un poids.

Les résultats de la fonction sigmoïde peuvent prendre n'importe quelles valeurs dans l'intervalle  $[0,1]$ , c'est pour cela qu'un seuil doit être déterminé (généralement 0.5) pour déterminer les classes auxquels appartient les données, la figure suivante montre le graphe de la fonction sigmoïde avec l'ensemble de valeurs possible pour chaque classe dans le cas où 0.5 est le seuil de la classification

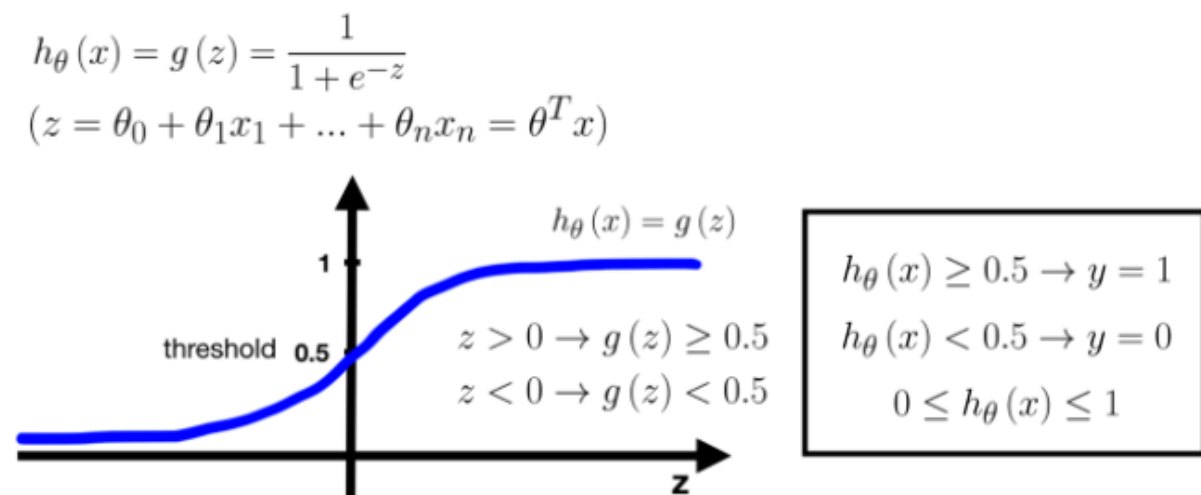


Figure 2. 20 : Formule régression logistique

La fonction de perte habituellement associée à la régression logistique est l'entropie croisée (cross entropy) :

$$J(\theta) = -\frac{1}{m} \left[ \sum_{i=1}^m y^{(i)} \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \log(1 - (h_{\theta}(x^{(i)}))) \right]$$

Équation 2. 18 : Fonction de perte associée à la régression logistique



A noter que cette forme de l'entropie croisée est seulement valable pour la classification binaire.

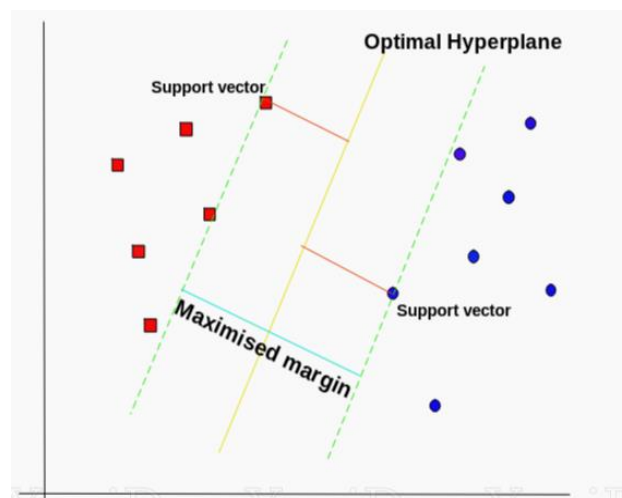
- **Optimisation de la fonction de perte :**

La méthode utilisée pour trouver le minimum global de la fonction de perte pour la régression logistique est toujours la descente du gradient comme pour la régression linéaire.

### 2.3.1.2.2. SVM (Support Vector machines) :

Machine à Vecteurs de Support (SVM) est un algorithme d'apprentissage automatique supervisé qui peut être utilisé à des fins de classification et de régression. Les SVM reposent sur l'idée de trouver un hyperplan qui divise au mieux un jeu de données en deux classes (hyperplan ou ligne optimale), similairement à la régression logistique, cet algorithme est défini par la fonction sigmoïde. [7]

On a un jeu de données et un hyperplan initial qui sépare les données en deux classes, selon l'algorithme de SVM, on trouve les points les plus proches de chaque classe, ces points sont appelés des vecteurs de support (support vectors), et on calcule la distance entre l'hyperplan les vecteurs de supports, la distance est une marge, donc le but est de maximiser cette marge, et l'hyperplan ayant la marge maximum est l'hyperplan optimale.



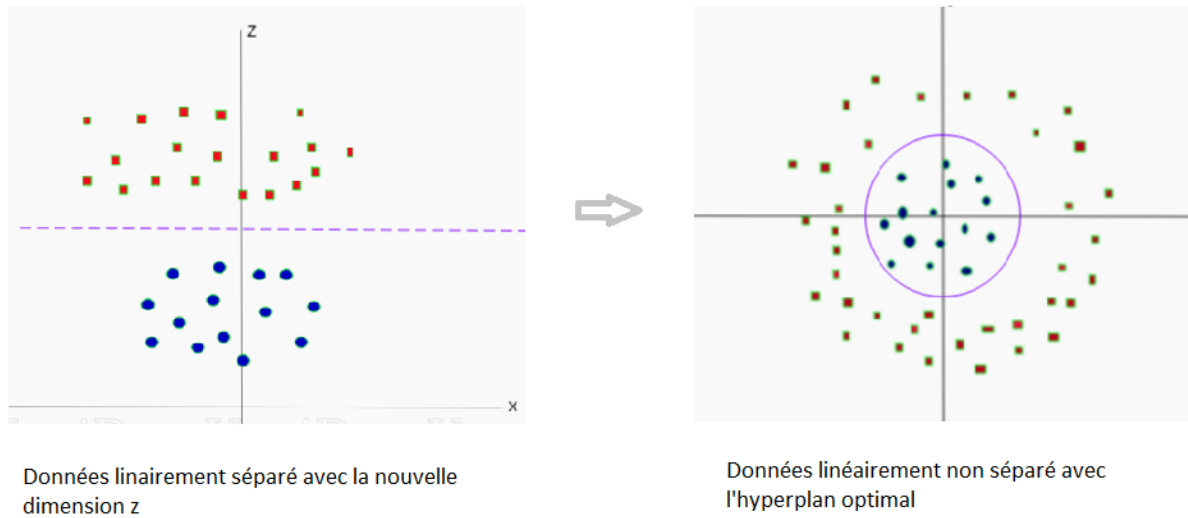
**Figure 2. 22 : SVM de l'hyperplan optimal**

Le SVM de la figure précédente a trouvé l'hyperplan optimal pour ces données qui sont bien évidemment linéairement séparables, mais est-ce que c'est faisable aussi pour les données linéairement non séparables (dont une ligne directe ne peut pas être l'hyperplan optimal) ? La réponse est oui, les données linéairement non séparables peuvent être converties en données linéairement séparables dans des dimensions plus hautes, donc une dimension supplémentaire appelé  $z$  sera ajouté selon la distance carré des anciennes dimensions :

$$z = x^2 + y^2$$

**Équation 2. 21 : Fonction SVM**

Dans ce cas, une ligne droite peut être l'hyperplan optimal, ensuite un retour vers les dimensions originales projettera l'hyperplan optimal non linéaire.



**Figure 2. 25 : Hyperplan optimal non linéaire**

### 2.3.1.2.3. Classificateur naïve bayes (Naïve bayes classifier) :

C'est un algorithme d'apprentissage supervisé utilisé pour la classification, il est gravement utile pour les problématiques de classification de texte. Cet algorithme est basé sur le théorème de Bayes créé par Thomas Bayes :

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

**Équation 2. 24: Théorème de Bayes**

Où :

- $P(A/B)$  = la probabilité de A sachant B
- $P(B/A)$  = la probabilité de b sachant A
- $P(A)$  = la probabilité de l'évènement A
- $P(B)$  = la probabilité de l'évènement B

En utilisant ces bases, la formule de l'algorithme du classificateur Naïve bayes peut s'écrire comme suivant :

$$P(y | x_1, \dots, x_j) = \frac{P(x_1, \dots, x_j | y) P(y)}{P(x_1, \dots, x_j)}$$

**Équation 2. 27: Classificateur Naïve Bayes**

Où y est classe parmi toutes les classes possibles et les x sont les variables des données.

### 3.1.2.4 K-voisins les plus proches (KNN)

C'est un des algorithmes les plus basiques qui appartient à la catégorie d'apprentissage supervisé, utilisé pour la classification, la régression et notamment le clustering.

En effet cet algorithme est qualifié comme paresseux (Lazy Learning) car il n'apprend rien pendant la phase d'entraînement. Pour prédire la classe d'une nouvelle donnée d'entrée, il va chercher ses K voisins les plus proches (en utilisant la distance euclidienne, ou autres) et choisira la classe des voisins majoritaires.

Pour appliquer cette méthode, les étapes à suivre sont les suivantes :

- On fixe le nombre de voisins k
- On détecte les k-voisins les plus proches des nouvelles données d'entrée que l'on veut classer.
- On attribue les classes correspondantes par vote majoritaire

Les types de fonctions de distance reposent sur les types de données, mais les plus fréquemment utilisées sont :

<div style="border: 1px solid #007bff; padding: 2px 5px; display: inline-block; margin-right: 10px;">Euclidean</div> $\sqrt{\sum_{i=1}^k (x_i - y_i)^2}$	<div style="border: 1px solid #007bff; padding: 2px 5px; display: inline-block; margin-right: 10px;">Manhattan</div> $\sum_{i=1}^k  x_i - y_i $
<div style="border: 1px solid #007bff; padding: 2px 5px; display: inline-block; margin-right: 10px;">Minkowski</div> $\left( \sum_{i=1}^k ( x_i - y_i )^q \right)^{1/q}$	

Équation 2. 30 : Formules de fonctions de distance

### 2.3.2. Algorithmes d'apprentissage Non Supervisé

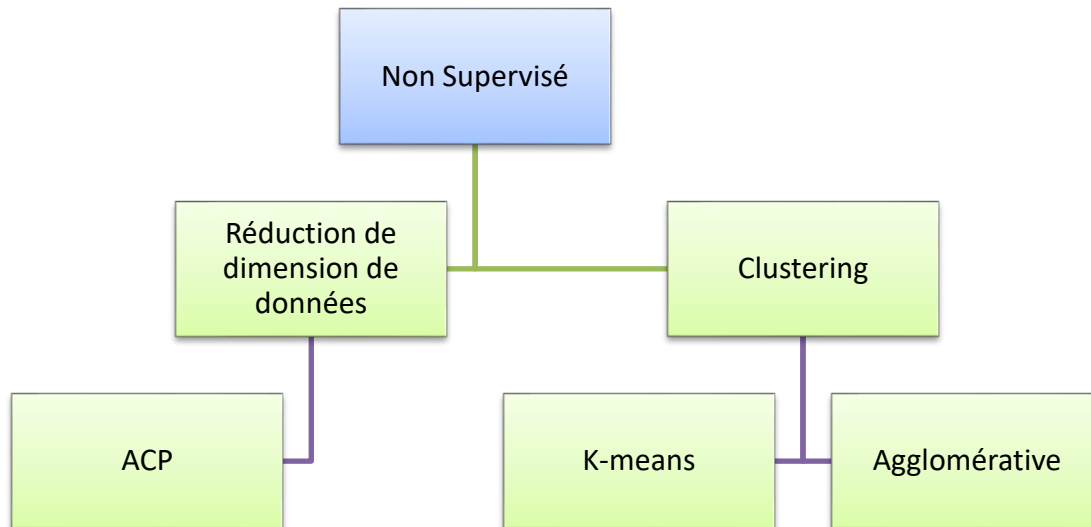


Figure 2. 28 : Algorithmes d'apprentissage Non Supervisé

#### 2.3.2.1. Réduction de dimension

Ces derniers temps, les projets du monde réel en relation avec la science de données tendent à avoir une large dimension, ces dimensions sont composées d'un large nombre de variables, dont certaines sont importantes pour l'étude et d'autres non. Cependant, les problèmes d'apprentissage automatique ayant une haute dimension ne peuvent ni être visualisés ni analysés convenablement, et en plus de ça, ces problèmes tendent à prendre un temps d'entraînement ou d'exécution énorme, c'est là où la réduction de dimension entre en scène.

C'est une méthode qui permet de compresser des larges sets de variables dans des nouveaux sets d'une dimensionnalité inférieure sans une grande perte d'information.

L'une des techniques de réduction de dimension les plus utilisées et les plus connues, c'est l'analyse aux composants principaux (PCA - ACP)

##### 3.2.1.1 L'analyse aux composants principaux (ACP)

C'est une technique de réduction de dimension généralement utilisée pour des données non supervisées. Durant cette méthode, les variables sont transformées à une nouvelle combinaison linéaire moins dense qui sont appelées les composants principaux (CP). Ensuite, ces CP sont obtenus d'une façon particulière, le premier CP inclut une variance plus haute que le reste des CP, et le deuxième aussi, ainsi de suite.

Premièrement, la matrice de variance-covariance entre les variables originales est calculée, ensuite l'ACP calcule des axes orthogonaux d'une variance maximum, ces axes sont nommés « vecteurs propres », les « valeurs propres » sont ensuite créées, ils représentent la magnitude de chaque CP. Donc pour N dimensions, on aura une matrice de variance-covariance de taille N x N, et comme résultat N vecteurs propres et N valeurs propres.

### 2.3.2.2. Clustering

Clustering (ou partitionnement des données) : c'est une méthode de classification non supervisée, rassemble un ensemble d'algorithmes d'apprentissage dont le but est de regrouper des données non étiquetées présentant des propriétés similaires, un de ses domaines d'utilisation c'est lorsque le coût d'étiqueter la donnée est élevé.

C'est néanmoins un problème mal défini mathématiquement : différentes métriques et/ou différentes représentations des données aboutiront à différents regroupements sans qu'aucun ne soit nécessairement meilleur qu'un autre. Ainsi la méthode de clustering doit être choisie avec soin en fonction du résultat attendu et de l'utilisation prévue des données. [9]

#### 2.3.2.2.1. K-MEANS

C'est l'un des algorithmes de clustering les plus répandus, il permet d'analyser un jeu de données caractérisées par un ensemble de descripteurs, afin de regrouper les données similaires en groupes (ou clusters). La similarité entre deux données peut être inférée grâce à la "distance" séparant leurs descripteurs ; ainsi deux données très similaires sont deux données dont les descripteurs sont très proches.

Cette définition permet de formuler le problème de partitionnement des données comme la recherche de K "données prototypes", autour desquelles peuvent être regroupées les autres données. Ces données prototypes sont appelés centroïdes : en pratique l'algorithme associe chaque donnée à son centroïde le plus proche, afin de créer des clusters. Après avoir initialisé ses centroïdes en prenant des données au hasard dans le jeu de données, K-MEANS alterne plusieurs fois ces deux étapes pour optimiser les centroïdes et leurs groupes :

1. Regrouper chaque objet autour du centroïde le plus proche.
2. Remplacer chaque centroïde selon la moyenne des descripteurs de son groupe.

Après quelques itérations, l'algorithme trouve un découpage stable du jeu de données : on dit que l'algorithme a convergé. [10]

#### 2.3.2.2.2. Algorithme hiérarchique Agglomératif

C'est un algorithme de clustering, qui commence par déclarer chaque point comme un cluster, c'est-à-dire les N points seront initialisés comme N clusters, ensuite il fusionne les deux clusters les plus similaires jusqu'à ce qu'une certaine condition d'arrêt soit satisfaite (il y a plusieurs conditions d'arrêts, mais généralement la plus utilisée par défaut c'est le nombre de clusters voulu), la figure suivante visualise la progression du clustering agglomératif sur des données de deux dimensions :

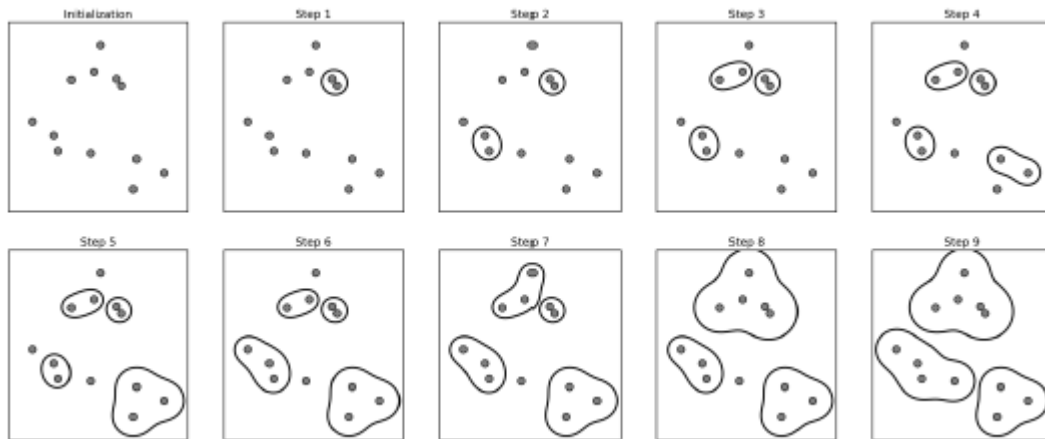


Figure 2. 31 : la progression du clustering agglomératif sur des données de deux dimensions

Le clustering agglomératif est un type de clustering hiérarchique, il procède par des itérations, et chaque point commence en étant un cluster unique et termine en faisant partie d'un cluster plus grand, la figure suivante représente un outil de visualisation qui s'appelle le dendrogramme, et qui montre un clustering agglomératif qui commence avec 12 clusters et en résulte 2 :

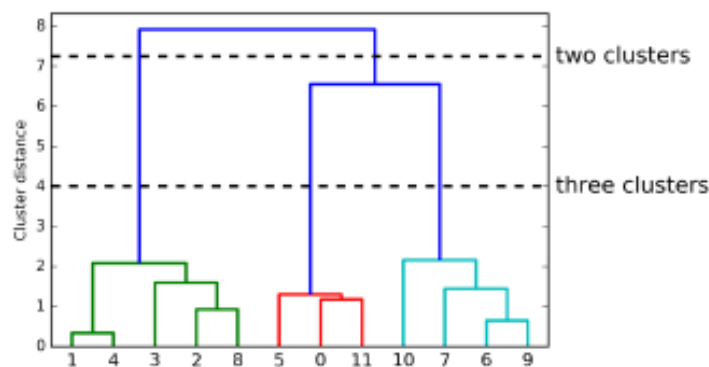


Figure 2. 34 : exemple d'un dendrogramme

### 2.3.3. Réseaux de neurones artificiels

Un réseau de neurones artificiels (Artificial Neural Network – ANN) est un paradigme de traitement de l'information inspiré par les systèmes nerveux biologiques et composé d'un nombre d'éléments hautement interconnectés (neurones) qui travaillent en collaboration pour résoudre des problèmes spécifiques, les neurones sont organisés en 3 types de couches :

- **La couche d'entrée** : Contient les données d'entrée.
- **Les couches cachées** : Elles contiennent des valeurs intermédiaires calculées lors de l'entraînement.
- **La couche de sortie** : contient un ou plusieurs neurones ayant comme valeur le résultat de l'opération.

L'architecture du réseau de neurones peut être en partie définie par les éléments suivants : le nombre de couches, nombre de neurones, les types de connexions entre les couches. Le réseau de neurones le plus connu est le réseau de neurones multicouche à propagation-avant (feed forward multilayer neural network). Il comporte une couche d'entrée, une ou plusieurs couches cachées et une seule couche de sortie. Chaque couche peut avoir un nombre différent de neurones [11]

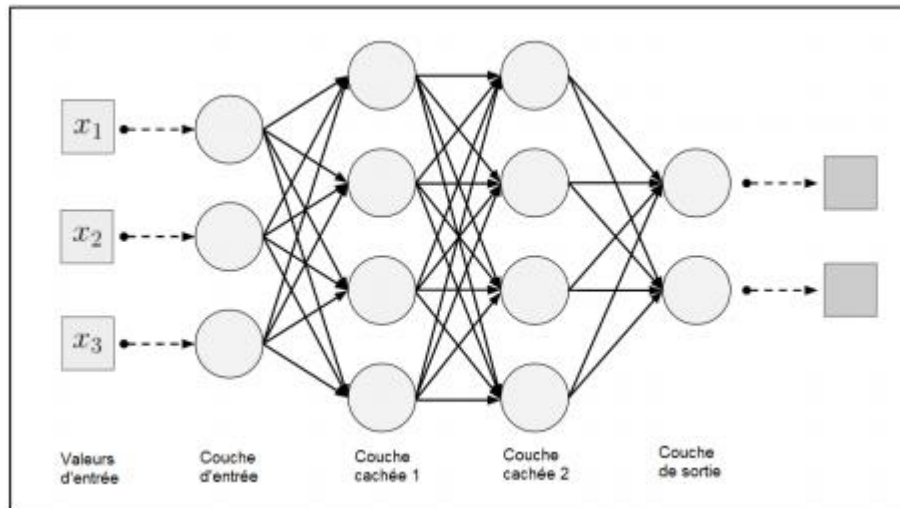


Figure 2. 37 : Architecture réseau de neurone

Chaque couche inclut un jeu de paramètres  $W$  dont la qualité va définir la précision et la qualité finale du modèle, et une fonction d'activation  $f$  (peut être appelé  $g$ ) :

- Le biais  $w_0$  (peut être appelé  $b$ ) : une valeur scalaire ajoutée à l'entrée pour garantir que le modèle essaiera toujours de nouvelles interprétations à chaque itération, dans plusieurs cas le terme biais est égal à 1.
- Les poids  $w_1 \dots w_n$  : ils sont des coefficients associés à chaque couche et à chaque neurone (nœud), ils sont estimés et optimisés pendant l'apprentissage.
- La fonction d'activation : diffère selon l'objectif de la couche, elle transforme la combinaison des entrées, des poids et des biais et est transmise à la couche suivante. [17]

### 2.3.3.1. Le perceptron linéaire :

Pour bien comprendre comment les réseaux de neurones fonctionnent et utilisent leur paramètre, il faut comprendre comment fonctionne un perceptron linéaire, qui est le plus petit réseau de neurones qui puisse exister

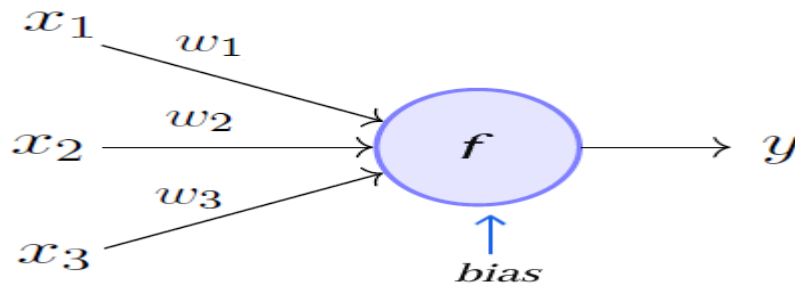


Figure 2. 40 : exemple du plus petit réseau de neurones

Étant donné un set de données d'entrée  $X_i = [x_1, x_2, \dots, x_n]$ , chaque connexion qui les relie au neurone suivant est associé avec des poids  $W_i = [w_1, w_2, \dots, w_n]$ . Ces poids multiplient les entrées  $X_i$  et sont ajoutées au biais avant d'être passées par une fonction d'activation  $f$  présente au niveau du neurone pour obtenir  $y$ . Puisque nous sommes dans le cas linéaire, la donnée de sortie du réseau peut donc être représentée comme  $y = f(x_1.w_1 + x_2.w_2 + x_3.w_3 + \text{biais})$ , et le cas générale  $y = f(w_1.w_1 + x_2.w_2 + \dots + x_n.w_n)$ .

### 2.3.3.2. Les fonctions d'activation

Lorsqu'un neurone transmet une valeur non nulle à un autre neurone, on dit qu'il est activé, et ceci à partir d'une fonction d'activation  $f$ , les fonctions d'activation les plus utilisées sont classé dans le tableau suivant :

<b>Fonction linéaire</b>	$F(x) = w.x + b$	<b>Fonction sigmoïde</b>	$F(x) = 1 / (1 + \exp(x))$
<b>Fonction Tanh</b>	$F(x) = \sinh(x) / \cosh(x)$	<b>Fonction Softmax</b>	Une généralisation de la sigmoïde
<b>Fonction ReLu</b>	$F(x) = \max(0, x)$	-	-

Tableau 2. 2 : Fonctions d'activation

### 2.3.3.3. L'entraînement du réseau

Il comporte deux étapes principales :

Premièrement l'algorithme du feedforward, qui est tout simplement le fonctionnement du perceptron comme décrit plus haut est répliqué à l'échelle de dizaines de couches et de milliers de neurones depuis la couche d'entrée jusqu'à la couche de sortie, Ensuite à la fin du réseau, une fonction de perte va permettre de calculer la différence entre le résultat obtenu par le réseau et le résultat correct. [12]

Deuxièmement, la Rétro propagation du gradient (BackPropagation). Dans le but de faire des prédictions plus précises, il faut être en mesure de diminuer le résultat de la fonction de perte, donc elle est dérivée pour en trouver le minimum. Le but est de trouver le signe de la dérivée afin de savoir si les poids des connexions (les inconnues) doivent être augmentés ou



diminués afin de minimiser la perte et donc de faire des prédictions plus exactes, ensuite la perte est répercutée en changement de tous les poids du réseau, mais cette fois ci de la dernière couche à la première.

## 2.4. Démarches de résolution d'un problème d'apprentissage automatique

La résolution d'un problème d'apprentissage automatique consiste à construire des modèles à partir des jeux de données (données d'entraînement). Par conséquent ce modèle fera la prédiction avec une certaine précision, sur de nouvelles données (données de test). Pour aboutir à un tel modèle, les praticiens suivent les étapes suivantes

- Préparation des données
- Choix de l'algorithme d'apprentissage
- Entraînement du modèle
- Evaluation

### 2.4.1. Préparation des données

Elle consiste premièrement à collecter les données (télécharger une base de données open source, collecter les données une à une à partir de différentes sources et puis les joindre ou bien simplement les récupérer depuis l'organisme propriétaire du projet), ensuite un prétraitement et nettoyage des données pour s'assurer de leur qualité et de leur convenabilité.

### 2.4.2. Choix de l'algorithme d'apprentissage

Le choix de l'algorithme d'apprentissage automatique dépend des facteurs suivants :

<b>Catégorisation depuis les données d'entrée</b>	<b>Données étiquetée</b>	Apprentissage supervisé
	<b>Données Non étiquetée avec l'objectif de trouver une structure</b>	Apprentissage non supervisé
<b>Catégorisation depuis les données de sortie</b>	<b>La variable dépendante est numérique continue</b>	Un problème de régression
	<b>La variable dépendante est numérique discrète</b>	Peut-être régression ou classification, ça dépend la limite des valeurs de la variable
	<b>La variable dépendante est Catégorique</b>	Un problème de classification

Tableau 2. 3 : Facteurs des Algorithmes

Après l'identification du type du problème, l'étape suivante sera d'identifier les algorithmes applicables et pratiques pour la résolution du problème, ensuite, le mieux sera d'implémenter un système de comparaison entre chaque algorithme parmi les algorithmes choisis, et ceci pour seulement un sous-groupe du jeu de données (disant 20%), puis en observant les valeurs des critères d'évaluation choisis, on aboutit à un choix d'un algorithme d'apprentissage automatique final.

### 2.4.3. Entraînement du modèle

Durant cette phase, les praticiens ajustent les paramètres initiaux du modèle (L'intercepte et la pente de la ligne de régression pour la régression linéaire, la variable  $\theta$  pour la régression logistique, le  $K$  et la distance à utiliser pour KNN, le nombre  $K$  de clusters pour un algorithme de clustering, les poids initiaux et le biais dans le cas du réseau de neurones, ...), puis le modèle est entraîné en plusieurs reprises sur des entrées des exemples appelées données d'entraînement, et la fonction de perte du modèle atteint son minimum.

### 2.4.4. Test

Une fois le modèle entraîné et la fonction de perte a atteint son minimum, l'étape suivante sera de tester le modèle, c'est-à-dire, ce dernier va recevoir des entrées différentes de celles des exemples qu'il a appris durant l'entraînement, ces nouveaux exemples sont appelés données de test. Durant la phase de test, les praticiens ont tendance d'ajuster les hyperparamètres (par exemple le nombre de couches et de neurones d'un réseau de neurones, ou bien la fonction de perte à utiliser, ...).

## 2.5. L'apprentissage automatique et télécom

L'apprentissage automatique comme dit précédemment joue un rôle dans presque toutes les sciences et techniques, particulièrement dans le domaine de Télécom, l'un des problèmes majeurs qui fait appel à l'IA c'est le problème de churn, il est très important de prédire les clients qui risquent de quitter l'entreprise, afin d'appliquer certains changements qui permettent de garder ces derniers, indirectement connaître les critères qui aident à rétablir la satisfaction clients. Un autre problème qu'on peut trouver aussi c'est la détection de fraudes sur certaines transactions notamment sur le contenu des appels. Nous citons ci-dessus quelques projets qui utilisent l'apprentissage automatique sur le domaine de Télécom en expliquant en résumé l'architecture de chaque système.

### 2.5.1. Customer Churn Analysis in Telecom Industry de Kiran Dahiya et Surbhi Bhatia

[13]

Il s'agit d'un problème de classification, c'est-à-dire de classer chaque abonné comme cherner potentiel ou non cherner potentiel (le terme cherner désigne un client quittant l'entreprise), le schéma présenté ci-dessous est basé sur le processus des données de découverte des connaissances (KDD), il est composé de cinq modules qui sont les suivants :

**Acquisition de données :** l'acquisition de données auprès de l'industrie du téléset est une tâche difficile en raison de la crainte d'en abuser, l'ensemble de données de cette étude acquis à la KDD Cup 2009, il permet d'analyser la tendance marketing des clients à partir des grandes bases de données de la société de télécommunications française Orange [SOURCE].

**Préparation des données :** étant donné que l'ensemble de données acquis ne peut pas être appliqué directement aux modèles de prédiction du taux de désabonnement, l'agrégation des données est donc requise lorsque de nouvelles variables sont ajoutées aux variables existantes en visualisant le comportement d'utilisation périodique des clients. Ces variables sont très importantes pour prédire à l'avance le comportement des clients car elles contiennent des informations critiques utilisées par les modèles de prédiction.

**Prétraitement des données :** le prétraitement des données est la phase la plus importante des modèles de prédiction car les données sont constituées d'ambiguïtés, d'erreurs, de redondance qui doivent être nettoyées au préalable. Les données recueillies à partir de plusieurs sources sont d'abord agrégées, puis nettoyées, car les données complètes collectées ne conviennent pas à des fins de modélisation.

Les enregistrements avec des valeurs uniques n'ont aucune signification car ils ne contribuent pas beaucoup à la modélisation prédictive. Les champs avec trop de valeurs nulles doivent également être ignorés.

**Extraction de données :** les attributs sont identifiés pour le processus de classification. Dans notre travail, nous avons travaillé avec des valeurs numériques et catégoriques.

**Décision :** L'ensemble de règles permettra aux abonnés d'identifier et de classer dans les différentes catégories de churners et de non churners en définissant une valeur seuil particulière.

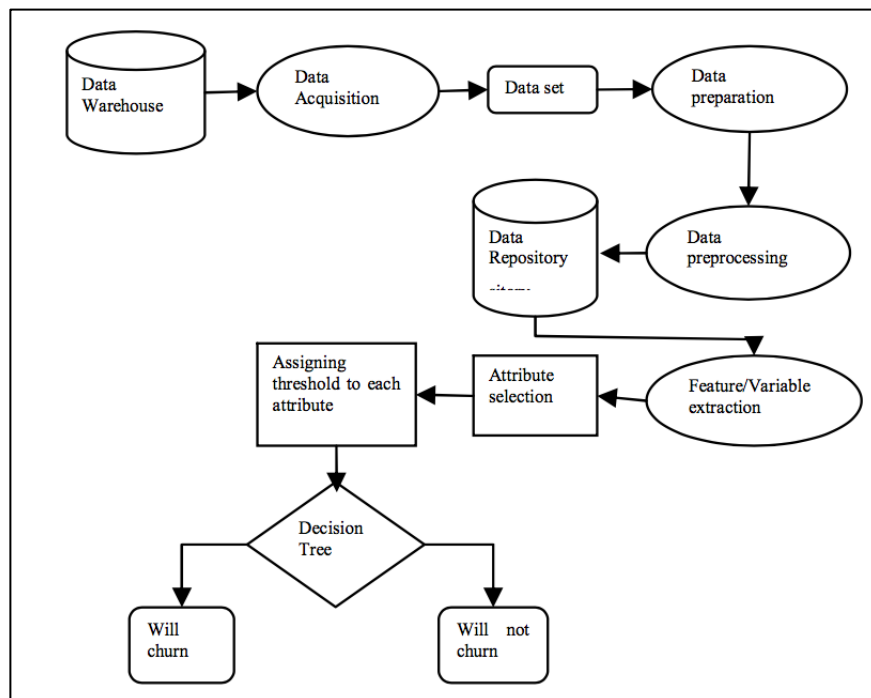


Figure 2. 43 : Architecture système du projet Customer Churn Analysis in Telecom Industry

De nombreuses approches ont été appliquées pour prédire le taux de désabonnement dans les entreprises de télécommunications. La plupart de ces approches ont utilisé l'apprentissage automatique et l'exploration de données. La majorité des travaux connexes se sont concentrés sur l'application d'une seule méthode d'exploration de données pour extraire des connaissances, et les autres se sont concentrés sur la comparaison de plusieurs stratégies pour prédire le taux de désabonnement.

### 2.5.2. Detecting telecommunication fraud by understanding the contents of a call de Qianqian Zhao, Kai Chen, Tongxin Li, Yi Yang & XiaoFeng Wang

[14]

Il s'agit d'un apprentissage automatique appliqué sur les appels afin de détecter les fautes en utilisant les différentes techniques de l'IA, le schéma suivant résume le travail :

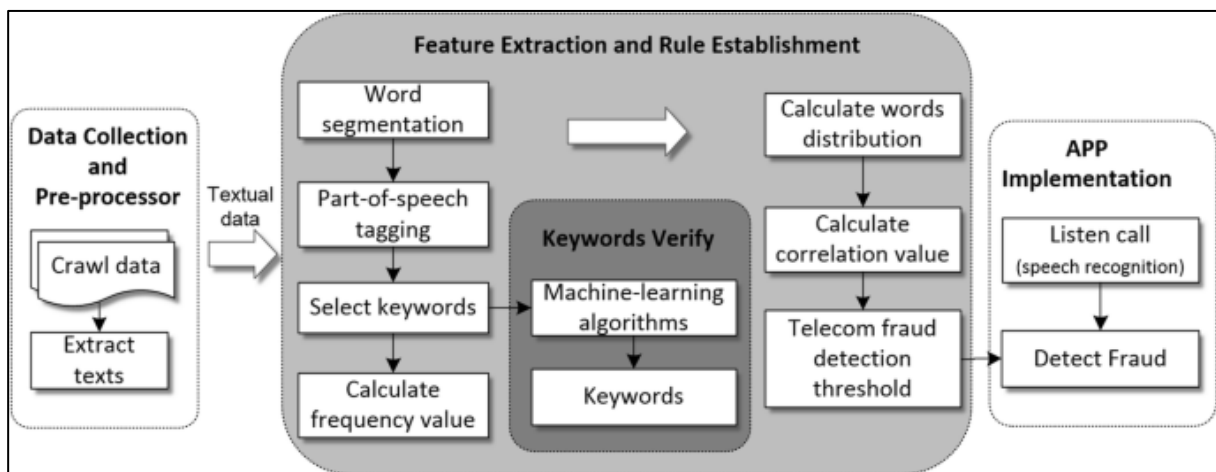


Figure 2. 46 : Architecture globale du système pour la prédiction des fraudes par appels

La première étape est la collecte de données sur la fraude afin d'analyser les caractéristiques et les modes de fraude aux télécommunications, commençant par les données d'exploration puis les extraire sous format textes, les données cibles comprennent le cas des appels frauduleux, le langage de description de la fraude aux télécommunications et les informations sur les médias.

Dans le processus de collecte de données, la technologie des robots d'exploration Web est utilisée pour collecter des données et les moteurs de recherche (tels que Baidu, etc.) sont aidés à collecter des données textuelles sur la fraude aux télécommunications sur Internet.

La deuxième étape est l'extraction des fonctionnalités et la création de règles, il est important d'extraire les fonctionnalités et de construire des règles de détection des fraudes aux télécommunications.

Cette recherche utilise la technologie de traitement du langage naturel pour extraire des caractéristiques qui sont des mots-clés du texte de fraude. Et en utilisant des algorithmes d'apprentissage automatique pour prouver la pertinence des données textuelles sur les données

collectées et la validité des mots-clés extraites. Ensuite, en fonction des caractéristiques extraites du texte, cette recherche construit les règles de détection de la fraude aux télécommunications.

La dernière partie est la mise en œuvre de la détection des fraudes aux télécommunications. Il s'agit d'un développement d'une application d'alerte de fraude par télécommunication sur la plateforme Android. En détail, l'application commence par surveiller l'appel entrant lorsqu'un appel arrive sur le téléphone des utilisateurs, ensuite elle utilise la technologie de reconnaissance vocale pour convertir la voix de l'appelant en texte. Après cela, l'application utilise les règles de détection intégrées à l'étape précédente pour déterminer s'il s'agit d'un appel frauduleux ou non. Si l'application indique qu'il s'agit d'un appel frauduleux, une information d'avertissement apparaîtra sur l'écran du smartphone pour inviter l'utilisateur à faire attention à cet appel.

## **2.6. Conclusion**

Dans ce chapitre nous avons défini les différentes notions de l'apprentissage automatique, décrit les différents algorithmes et méthodes de chaque type, leur objectif, ainsi que quelques exemples d'algorithmes à utiliser à la fin nous avons présenté deux exemplaires de projets sur le thème de la télécommunication, le prochain chapitre sera consacré sur la conception de notre travail.

## 3. Conception

### 3.1. Introduction

Le présent chapitre, présente notre contribution pour le développement d'un système permettant la prédiction automatique de la satisfaction client en utilisant des techniques de machine learning. Nous présentons les différentes étapes de notre conception, à savoir le prétraitement de données et la définition des attributs importants la segmentation selon un nombre de classes donné et l'apprentissage automatique sur ces données.

Le schéma ci-dessous représente l'architecture globale de la solution proposée :

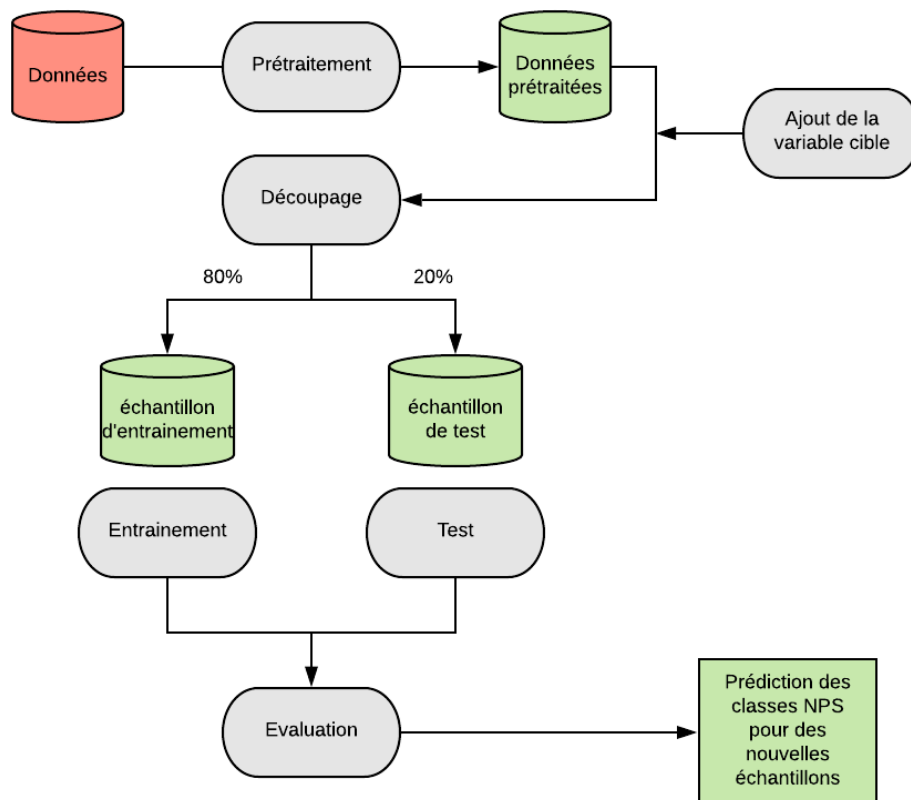


Figure 3. 2 : Schéma conceptuel globale de l'architecture proposée

## 3.2. Données et prétraitement

Les données utilisées en générale dans notre système concernent les clients, initialement nous disposons de certaines données brutes qui concerne notre travail, ces dernières doivent passer par un certain processus afin qu'elles soient prêtes à utiliser, le schéma suivant représente les étapes de traitement des données :

### 3.2.1. Collecte de données

#### 3.2.1.1. Généralités sur les données

Les données sont définies comme une collection de faits et de statistiques, collectées à partir des observations pour des besoins analytiques. Les données généralement sont classées selon deux types, certaines sont qualitatives, autrement dit catégorielles et d'autres sont quantitative, aussi appelées numériques.

- **Données quantitatives**

- Ce type de données est traité en nombres et en objets mesurables, et peut être exploité mathématiquement. Il y a deux types de données numériques, les données continues et discrètes :
- **Données continues** : peuvent prendre n'importe quelle valeur entière ou une infinité de décimale.
- **Données discrètes** : peuvent prendre que des valeurs entières.

Variables continues	Variables discrète
- La vitesse d'une voiture.	- Le nombre de personnes dans une salle
- Le poids d'une personne.	- Le nombre d'enfants dans une famille

Tableau 3. 2 : Différence entre variables continue set variables discrète

- **Données qualitatives**

- Ce type de données ne s'exprime en aucun cas par une valeur numérique, mais plutôt par des catégories.
- L'exploitation mathématique de ce type de données est impossible.
- Il y a deux types de données quantitatives, les données nominales et ordinales.
- **Données nominales** : ne peuvent pas être hiérarchisées, aucune valeur n'est supérieure à l'autre.
- **Données ordinales** : peuvent être classées les uns par rapport aux autres.
- **Données binaires** : peuvent prendre seulement deux valeurs possibles, 0 ou 1.

Variables nominales	Variables ordinales	Variables Binaires
- Les nationalités	- Les niveaux des difficultés	- Une personne P
- Les couleurs	- La remarque d'un enseignant à son élève	adulte ou pas
		- Compte actif ou pas

Tableau 3. 5 : Exemple de types de données

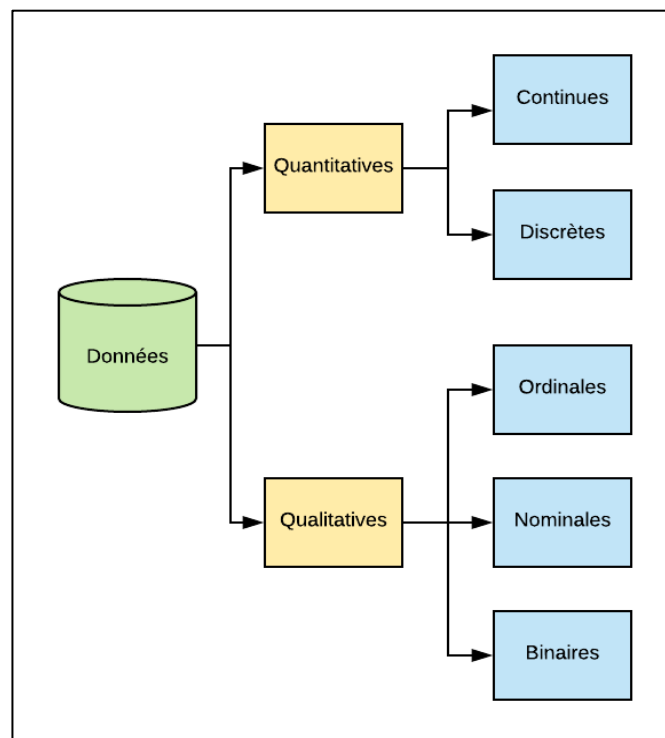


Figure 3. 5 : Types de données globales

### 3.2.1.2. Format et forme des fichiers de données

Le format et la forme des données sont des aspects très importantes dans l'étude, car en observant ces 2 aspects qu'on peut réfléchir, chercher et aboutir à une manière de traitement qui satisfera les objectifs du projet.

Les Fichiers de données initiales fournis par l'organisme d'accueil sont en forme de plusieurs fichiers en format TXT, et leur importance pour l'objectif de l'étude n'est pas assurée, cependant, pour la bonne conduite du projet, l'état actuel de ces 2 aspects sera difficilement compatible aux traitements à venir, donc un ensemble d'actions ou de transformations est envisageable pour l'amélioration de cohérence et de la comptabilité par rapport à ces futurs traitements.

Cet ensemble d'action est résumé comme suite :

- Sectionnement des fichiers importants pour l'étude : Consiste à choisir seulement les fichiers dont l'importance de leurs attributs est grande.
- Un changement de format : Consiste à transformer le format de données au format **CSV**
- Jointure : Joindre les fichiers sélectionnés pour former un seul fichier cohérent, compatible et utilisable pour nos traitements

La figure suivante explique le processus qui inclut les actions et les transformations utilisés :



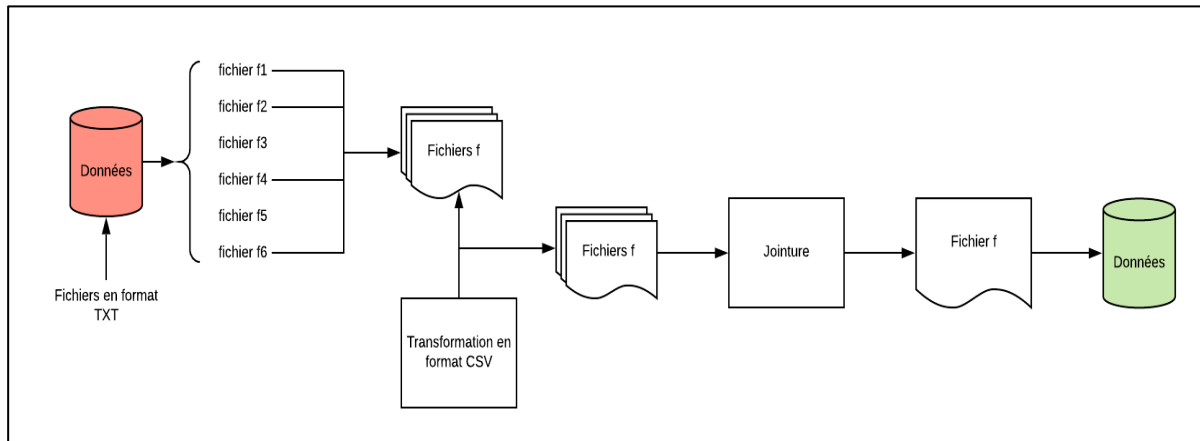


Figure 3. 8 : Processus des actions de transformations

### 3.2.1.3. Description des fichiers de données

Les fichiers de données fournis par Djezzy sont de la même structure les uns par rapport aux autres mais incluent des attributs différents, le tableau suivant contient une brève description de ces fichiers :

Nom du fichier	Description du fichier	Forme du fichier (nombre de lignes, nombre de colonnes)
User_info	Ce fichier contient quelques informations sur les clients comme leur identifiant, wilaya de résidence, âge, ...	(5000, 22)
Use_join_sample	Contient des données indiquant la date du premier appel et du dernier appel, avec le statut d'activité	(5000, 8)
Handset_data	Contient les informations et quelques fonctionnalités des dispositifs de quelques clients	(2015, 6)
Subscriber_usage	Contient des données indiquant les usages des voix et des data des clients	(4790, 6)
Wilaya région	Contient des données sur les 48 wilayas de l'Algérie tel que le code et la région	(48, 3)
Charge_minutes_calls	Contient des informations sur les appels, les minutes et crédit de chaque jour, nuit et soirée	(5000, 9)

Tableau 3. 8 : une brève description des fichiers de données

### 3.2.1.4. Description des attributs de données

Les fichiers de données fournis par l'établissement d'accueil et utilisées pour la mesure de satisfaction client sont constitués d'une variété de significations des attributs de données, ainsi que des types différents. En prenant par exemple le fichier User\_info, selon la figure suivante, il contient 5000 colonnes dont certaines ne contiennent aucune valeur nulle (5000/5000), d'autres ont des valeurs manquantes (moins de 5000) et une colonne TARIFF\_PROFILE\_POST\_PREP qui ne contient aucune valeur (0/5000)

Data columns (total 22 columns):		
Subs_Id	5000 non-null	int64
BIRTH	5000 non-null	int64
CUSTOMER_AGE_CLASS	5000 non-null	int64
AGE_GROUP	4584 non-null	object
GENDER	5000 non-null	object
segmentation_category	4986 non-null	object
B2B_SEGMENT_TYPE	1 non-null	object
TARIFF_PLAN	5000 non-null	object
TARIFF_PROFILE_POST_PREP	0 non-null	float64
PREPAID_IND	5000 non-null	int64
SUBSCRIPTION_TYPE	5000 non-null	object
CUSTOMER_TYPE	5000 non-null	object
TARIFF_PROFILE	5000 non-null	object
LANGUAGE_INDICATOR	5000 non-null	object
RESRVATION_CODE	4938 non-null	object
WILAYA	4999 non-null	object
Region_Name	4859 non-null	object
REGION_ID	4859 non-null	float64
Code_Wilaya	4859 non-null	float64
DESCRIPTION	2565 non-null	object
OWNERD	2565 non-null	object
COORDINATOR_NAME	6 non-null	object

Figure 3. 11 : Attributs de données

Le tableau suivant montre la description de quelques attributs utilisés :

Attribut	Type	Signification	Exemple
<b>AGE_GROUP</b>	Qualitative ordinale	La tranche d'âge englobant l'âge de plusieurs individus	'18-25', '66-75', ...
<b>TARIFF_PLAN</b>	Qualitative nominale	Le plan DJEZZY actuel utilisé par le client	Hayla Bezzef Prepaid_4G, Djezzy SPECIAL PREPAID_4G, ...
<b>CODE_WILAYA</b>	Quantitative discrète	Le code unique de chaque wilaya de l'Algérie	16, 09, 13, 27, 15 ...
<b>USAGE_DATA</b>	Quantitative continue	Quantité de données mobile utilisée par le client depuis l'activation	0.00, 234.27, ...

Tableau 3. 11 : Description de quelques attributs

### 3.2.2. Prétraitement des données

Beaucoup de facteurs peuvent affecter les performances d'un algorithme d'apprentissage automatique, et certainement la plus importante de ces facteurs c'est la qualité et l'état de ces données.

Un échantillon de données réelles contient souvent des données incomplètes, c'est à dire qu'il peut y avoir des valeurs manquantes, données simplifiées, bruitées qui contient des erreurs et des exceptions ou incohérences (nommage, codage) c'est pour cette raison qu'il est indispensable de réaliser un prétraitement ou autrement dit un nettoyage, pour qu'on puisse avoir un résultat de traitement performant et une précision satisfaisante.

La figure ci-dessous résume quelques cas du prétraitement des données.

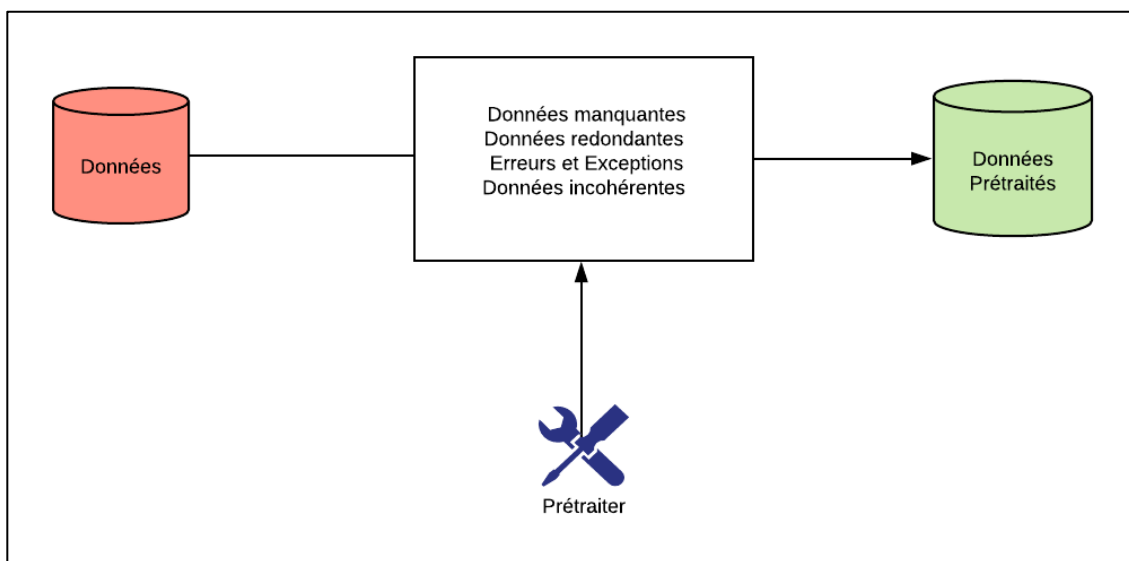


Figure 3. 14 : Exemples de cas de prétraitements

#### 3.2.2.1. Données manquantes

Ce cas est fréquent dans le prétraitement des données, ce sont les données non disponibles, certains attributs n'ont pas de valeur, parmi les causes de cette anomalie :

- Mauvais fonctionnement de l'équipement
- Incohérences avec d'autres données et donc supprimées
- Non saisies car elles sont non ou mal comprises
- Considérées peu importantes au moment de la saisie

La figure suivante représente une courbe qui représente les pourcentages des données manquantes pour plusieurs variables, et ceci pour l'état initial du fichier User\_info, cette figure informe que plusieurs de nos variables n'ont presque aucune donnée manquante (0%), et d'autres n'ont presque aucune valeur (99% - 100%)

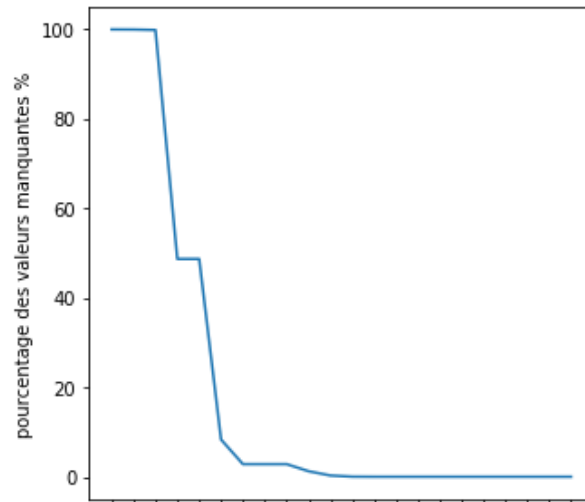


Figure 3. 17 : Courbe des données manquantes

La donnée manquante est généralement désignée par la chaîne de caractère 'NaN', ce qui veut dire 'Not a Number' (pas un numéro), Exemple réel de nos données :

	Subs_Id	BIRTH	WILAYA	AGE_GROUP	GENDER	LANGUAGE_INDICATOR	TARIFF_PLAN
685	143551266	24	NaN	18-25	male	ARD	B2C H'BALPrepaid_4G
1274	146160883	26	ILLIZI	26-35	female	ARD	Hayla Bezzef Prepaid_4G
2742	147850885	25	ILLIZI	18-25	male	ARD	B2C H'BALPrepaid_4G

Figure 3. 20 : Exemple réel de donnée manquante

Pour remplir les trous, il existe plusieurs façons nous citons :

- Ignorer le tuples : peu efficace quand le pourcentage de valeurs manquantes est élevé
- Compléter manuellement les données : laborieux ou infaisable
- Utiliser une constante globale : ex : « inconnue », qui devient une nouvelle catégorie
- Utiliser la moyenne de l'attribut
- Utiliser la moyenne de l'attribut pour la même classe : mieux
- Utiliser la valeur la plus probable : formule Bayésienne ou arbre de décision

Cependant, pour le remplacement des données manquantes dans notre cas, il y a quelques conditions à respecter, ces conditions font partie de l'algorithme suivant :

---

**Algorithme 1 : Remplissage de données manquantes**


---

**Entrée :** \$dataset, \$donnée;**Sortie :** \$dataset ;

\$i, \$long : entiers

**Début :**

```

    $type_variable = type ($donnée) ;
    /* Le type de la variable qu'on veut traiter */
$variable_corrélée = corrélation ($dataset, $valeurs_variable) ;
    /* retourne la variable la plus corrélée avec la variable qu'on traiter*/
$pourcentage_manque = pourcentage des données manquantes ($valeurs_variable) ;
    /* Retourne le pourcentage des données manquantes de la variable qu'on veut traiter*/
    Si $pourcentage_manque > 40% Alors :
        Supprimer ($dataset, $donnée) ;
        /* supprimer la donnée du dataset */
    Sinon :
        Si $type_variable == 'chaîne de caractère ' Alors :
            $groupe = Grouper ($donnée, $variable_corrélée) ;
            /* avoir une distribution de la variable groupé par les valeurs de la variable la
            plus corrélée avec elle */
            $mode_groupe = mode ($groupe) ;
            /* retourner les modes de la donnée par rapport aux valeurs de la variable
            corrélée */
            $donnée = Remplacer les manque ($donnée, $mode_groupe) ;
            /* remplacer les manque par rapport au groupement de chaque valeur */
        Sinon :
            $groupe = Grouper ($donnée, $variable_corrélée) ;
            /* avoir une distribution de la variable groupé par les valeurs de la
            variable la plus corrélée avec elle */
            $moyenne_groupe = moyenne ($groupe) ;
            /* retourner les moyennes de la donnée par rapport aux valeurs de la
            variable corrélée*/
            $donnée = Remplacer les manque ($donnée, $moyenne_groupe) ;
            /* remplacer les manque par rapport au groupement de chaque valeur */

```

**Fin****3.2.2.2. Données bruitées**

Représente une erreur ou variance aléatoire d'une variable mesurée, parmi ses causes :

- Instrument de mesure défectueux
- Problème de saisie
- Problème de transmission
- Limitation technologique
- Incohérence dans les conventions de nommage

Exemple réel de nos données :

	Subs_Id	BIRTH	WILAYA	AGE_GROUP	GENDER	LANGUAGE_INDICATOR	TARIFF_PLAN
5	147889854	50	W-PREACT	46-55	male	ARD	B2C H'BALPrepaid_4G
10	148028825	50	W-PREACT	46-55	male	ARD	B2C H'BALPrepaid_4G
15	147818421	50	W-PREACT	46-55	male	ARD	B2C H'BALPrepaid_4G
30	147202034	50	W-PREACT	46-55	male	ARD	Hayla Maxi Prepaid_4G
43	86049253	120	DUMMY	NaN	male	ARD	Flexy revendeur_2G
48	9716716	45	BISKRA	36-45	male	ARD	Good prepaid_2G
51	123079701	120	DUMMY	NaN	male	ARD	VIP-PARTNER-DISTRIBUTOR CARTE_2G

Figure 3. 23 : Exemple réel de données bruitées

La correction des données bruitées peut se faire avec plusieurs solutions nous citons :

- Par partitionnement (binning) :
  - Trier et partitionner les données
  - Lisser les partitions par la moyenne, la médiane, les bornes...
- Clustering : Détecter et supprimer les exceptions (**OUTLIERS**)
- Inspection humaine et informatique combinée : Détection des valeurs suspectes et vérification humaine
- Régression : Lisser les données par des fonctions de régression

Comme remarque, les données bruitées et les outliers sont deux aspects qui se ressemblent beaucoup mais pas exactement la même chose, la seule différence qu'on peut définir c'est le fait que les outliers sont des données hors de portée qu'on ne compte pas trouver dans les bornes de l'étude, quant aux données bruitées, ils sont fausses, non voulu et doivent être corrigé ou supprimé.

Il existe quelques façons pour détecter les données de bruits (ou les outliers) :

Premièrement le Z-score, pour un point de donnée, il exprime l'écart par rapport à la valeur moyenne, en déviation standard, pendant l'application du z-score pour chaque point de données, il existe un seuil à spécifier qui est généralement égale à 2.5 ou 3 (Threshold en anglais), les point ayant un z-score supérieur au seuil précisé sommes considérés comme outliers, sa formule est :

$$z = \frac{x - \mu}{\sigma}$$

Équation 3. 1: Formule du Z-score

Il y a aussi le BoxPlot, c'est un graphique tout simple qui permet de résumer une variable de manière simple et visuel, d'identifier les valeurs extrêmes et de comprendre la répartition des observations.

**Autres problèmes :**

- Enregistrement duplique
- Données incomplètes
- Données incohérentes

**3.2.2.3. Intégration et transformation des données****3.2.2.3.1. Intégration des données**

C'est la combinaison de différentes sources de données (des fichiers distincts) en une seule en utilisant un attribut commun ayant les mêmes valeurs.

Cependant, Il existe un problème de nommage qui consiste à identifier les différents noms des mêmes données réelles, ex : num\_client et client\_id

**Causes :** représentation différentes, échelles différentes, ...

Exemple réel de nos données :

Entrée [29]:

1 Network\_Sample.head(3)

Out[29]:

	<u>ID</u>	USE_QUOTA_CLUSTER	USE_NW_QUOTA_SEGM	HANDSET_TYPE	HANDSET_BRAND	targetted
0	22617449	1.0	1	3	1.0	1
1	109581158	1.0	1	1	2.0	1
2	641673	1.0	1	1	2.0	1

Entrée [36]:

1 full\_handset\_sub.head(3)

Out[36]:

	TAC	Devicetype	LTE	<u>Subscriber_ID</u>
0	35896710	1.0	0.0	139598004
0	35896710	1.0	0.0	27653849
1	35565805	2.0	0.0	144703084

**Figure 3. 26 : Exemple réel d'une intégration de données**

Dans cet exemple, on souhaite combiner deux tables selon l'attribut **ID** de la première table et l'attribut **Subscriber\_ID** de la deuxième, Mais avant, en doit traiter le problème de nommage, deux solutions s'offrent :

- Renommer une des deux attributs.
- Joindre directement en précisant les noms des deux attributs, et un nouveau nom pour la nouvelle table.

➤ **Gestion de la redondance**

C'est le cas où au moins deux attributs peuvent avoir les mêmes informations portant des noms différents, ce cas est fréquent lors de l'intégration de plusieurs sources de données.

Exemple réel de nos données :

	Subs_Id	BIRTH	CUSTOMER_AGE_CLASS	WILAYA	AGE_GROUP	GENDER	LANGUAGE_INDICATOR	TARIFF_PLAN
1274	146160883	26	26	ILLIZI	26-35	female	ARD	Hayla Bezzef Prepaid_4G
2742	147850885	25	25	ILLIZI	18-25	male	ARD	B2C H'BALPrepaid_4G
3026	145098606	37	37	TAMANRASSET	36-45	male	ARD	B2C H'BALPrepaid_4G

Figure 3. 29 : Exemple réel de la gestion des redondances

Pour détecter ce genre de problème, une analyse du coefficient de corrélation de Pearson entre les deux variables fera l'affaire, car le résultat sera 1 ou 0.9999.

#### ➤ Coefficient de corrélation

Le coefficient de Pearson est un indice reflétant une relation linéaire entre deux variables continues. Le coefficient de corrélation varie entre -1 et +1, 0 reflétant une relation nulle entre les deux variables, une valeur négative (corrélation négative) signifiant que lorsqu'une des variables augmente, l'autre diminue ; tandis qu'une valeur positive (corrélation positive) indique que les deux variables varient ensemble dans le même sens, voici la formule pour calculer le coefficient de corrélation de Pearson :

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

Équation 3. 2: Coefficient de corrélation de Pearson

#### 3.2.2.3.2. Transformation des données

##### ➤ Normalisation

Mise à l'échelle pour avoir un petit intervalle spécifié, ce qui permet de simplifier le problème d'apprentissage en s'affranchissant de ces deux paramètres.

##### ▪ Min max

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A$$

Équation 3. 3: Formule du min-max

##### ▪ Z score normalisation

$$z = \frac{x - \mu}{\sigma}$$

Équation 3. 4: Formule du z-score

##### ➤ Avantage de la transformation des données

- La transformation des données affine les métadonnées pour qu'on puisse plus facilement organiser et comprendre le contenu de votre ensemble de données



- Le processus de transformation des données peut réduire ou éliminer les problèmes liés à la qualité, comme les incohérences.

### 3.3. Type d'apprentissage

Le but du projet est d'implémenter un système capable de mesurer le taux de satisfaction client à partir d'un échantillon d'abonnés Djezzy, comme mentionné précédemment, ce taux est calculé à partir de trois classes où l'abonné peut appartenir. Ces classes sont mentionnées sur la figure suivante, détracteur, passif et promoteur. Par conséquent, tous les abonnés d'un échantillon doivent être classifiés selon l'un de ces trois classes afin que le système soit en mesure de mesurer le taux global de satisfaction NPS de l'échantillon. Cependant le système doit être entraîné à partir d'autres échantillons d'abonnés annotés (dont le score NPS est présent). Ceci dit, le type de prédiction à faire se résume en un problème d'apprentissage supervisé de classification

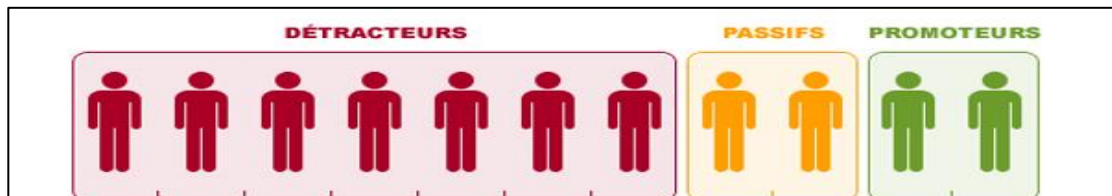


Figure 3.2: Catégories des clients

#### 3.3.1. Variable cible

La données cible c'est la variable sur laquelle le modèle d'apprentissage automatique est entraîné, et c'est la variable dont les valeurs sont prédites.

Pour notre cas, la donnée cible est d'une signification des classes NPS : Promoteur, Passifs et détracteurs.

Le problème qui s'est posé à la conduite de ce projet c'est le manque de la donnée cible, pour régler ce problème, et pour ne pas générer la variable cible aléatoirement (ce qui diminuera les performances), un simple algorithme a été conçu :

**Algorithme 2 : Génération de la variable cible.****Entrée :** \$dataset, \$borne\_voice, \$borne\_data;      **Sortie :** \$dataset ;

\$i, \$long : entiers

**Début :**

```
$Moyenne_voice = moyenne ($dataset ['usage_voice']) ; /* la moyenne de l'ensemble des valeurs de
la variable qui concerne l'utilisation de voix de l'abonné */
```

```
$Moyenne_data = moyenne ($dataset ['usage_data']) ; /* la moyenne de l'ensemble des valeurs de la
variable qui concerne l'utilisation de data de l'abonné */
```

```
$Long = longueur ($dataset) ;
```

```
/* la moyenne de l'ensemble des valeurs de la variable qui concerne l'utilisation de voix
de l'abonné */
```

```
$dataset = ajout nouvelle variable dans le dataset appelé satisfaction ($dataset, 'Satisfaction') ;
```

```
/* ajouter une nouvelle colonne au dataset qui doit contenir les valeurs de la donnée
cible */
```

**Pour \$i allant de 0 à \$Long faire :**

```
    Si $dataset [$i] [$usage_voice] < Moyenne_voice - $borne_voice alors :
```

```
        Si $dataset [$i] [$usage_data] < Moyenne_data - $borne_data alors :
```

```
            $X = 0 ;
```

```
        Sinon
```

```
            Si $dataset [$i] [$usage_data] > Moyenne_data + $borne_data alors :
```

```
                $X = 1 ;
```

```
            Sinon
```

```
                $X = 1 ;
```

```
            Fin Si
```

```
        Fin Si
```

```
    Sinon
```

```
        Si $dataset [$i] [$usage_voice] > Moyenne_voice + $borne_data alors :
```

```
            Si $dataset [$i] [$usage_voice] < Moyenne_voice - $borne_voice alors :
```

```
                $X = 1 ;
```

```
            Sinon
```

```
                Si $dataset [$i] [$usage_data] > Moyenne_data + $borne_data alors :
```

```
                    $X = 2 ;
```

```
                Sinon
```

```
                    $X = 1 ;
```

```
                Fin Si
```

```
    Fin si
```

```
    Si non
```

```
Si $dataset [$i] [$usage_voice] < Moyenne_voice - $borne_voice alors :
```

```
        $X = 1 ;
```

```
    Sinon
```

```
        Si $dataset [$i] [$usage_data] > Moyenne_data + $borne_data
```

```
alors :
```

```
            $X = 2 ;
```

```
        Sinon
```

```
            $X = 1 ;
```

```
        Fin Si
```

```
    Fin si
```

```
    Fin si
```

```
        $dataset [$i] ['satisfaction'] = ajout nouvelle valeur ($X) ;
```

```
Fin Pour
```

```
Retourner $dataset ;
```

```
Fin
```

### 3.4. Apprentissage automatique et prédiction du taux NPS

Avant d'entamer la classification, le dataset doit être découpé en deux sous datasets, dataset d'entraînement et dataset de test. Les pourcentages de ce découpage peuvent varier et dépendent totalement du dataset en question mais généralement le dataset d'entraînement est de 80%, et 20% pour le dataset de test. Bien entendu le dataset de test contient seulement les données non annotées, en d'autres termes les données qui ne contiennent pas la variable cible NPS.



Figure 3. 32 : Découpage des données d'entraînement et de test

Après que les données soient prêtes à utiliser (prétraitées et contenant la donnée cible), l'étape qui reste est l'utilisation d'un algorithme de classification pour l'entraînement et pour la prédiction des données. Plusieurs algorithmes de classification mentionnés dans le chapitre 2 peuvent être entraînés sur l'échantillon, testés et évalués, et le choix du meilleur algorithme sera fait par rapport à la meilleure précision.

Au moment où l'algorithme le plus précis est connu et est intégré dans le système, de nouveaux échantillons peuvent être classifiés et le processus résulte les trois classes attendues, afin de finalement aboutir au score NPS global. Le schéma suivant explique la façon dont ce processus se déroulera :

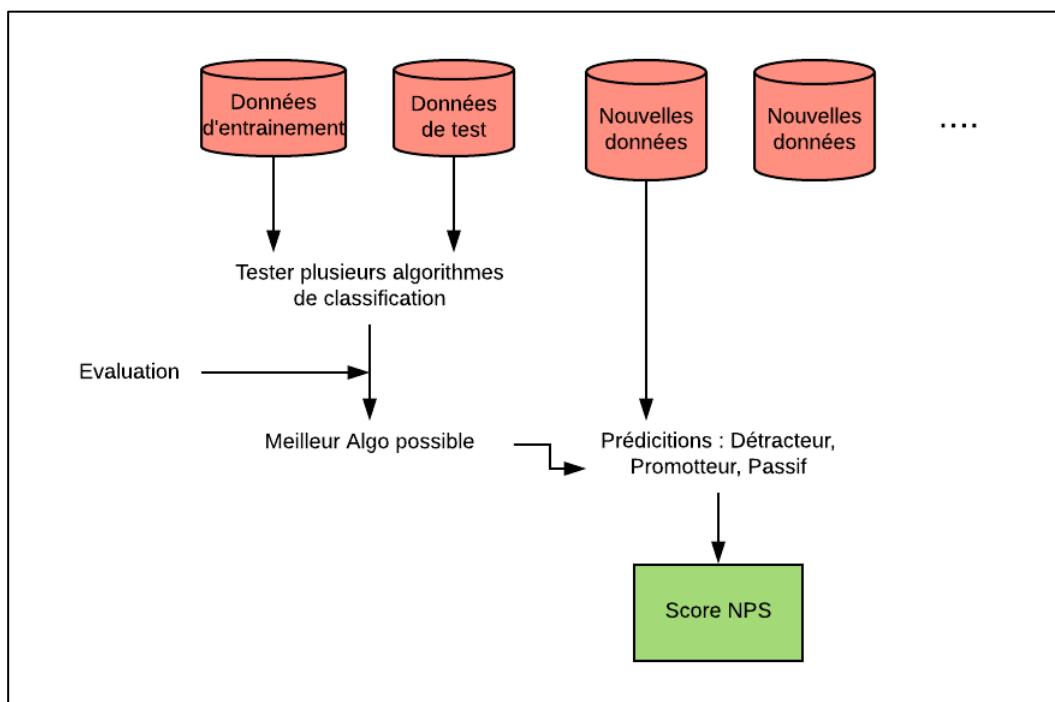


Figure 3. 14 : Processus de classifications

### 3.5. Ingénierie du logiciel

Dans cette section nous allons présenter une brève étude conceptuelle pour le développement de notre système. Notre objectif est de comprendre le contexte du système, il s'agit d'identifier les besoins fonctionnels et non fonctionnels de notre application, préciser les acteurs et identifier le cas d'utilisation initial

#### 3.5.1. Spécification des besoins fonctionnels

Les besoins fonctionnels sont les actions que le système doit exécuter, ils sont spécifiés par l'analyseur (particulièrement par le développeur pour notre application, car il s'agit d'une application de test). Ces fonctionnalités sont non négligeables. On cite ci-dessous les besoins fonctionnels de notre application du côté Utilisateur et côté Administrateur :

##### *Administrateur :*

- L'authentification
- La gestion des utilisateurs :  
(Ajout, modification, suppression des utilisateurs)

##### *Utilisateur :*

- L'authentification
- Importation des fichiers de tests pour prédire le résultat
- Génération d'un taux de satisfaction à partir des paramètres donnés
- Afficher les statistiques globales des tests, particulière des données de tests

Ainsi, notre application fait intervenir deux acteurs principaux : l'utilisateur et l'administrateur

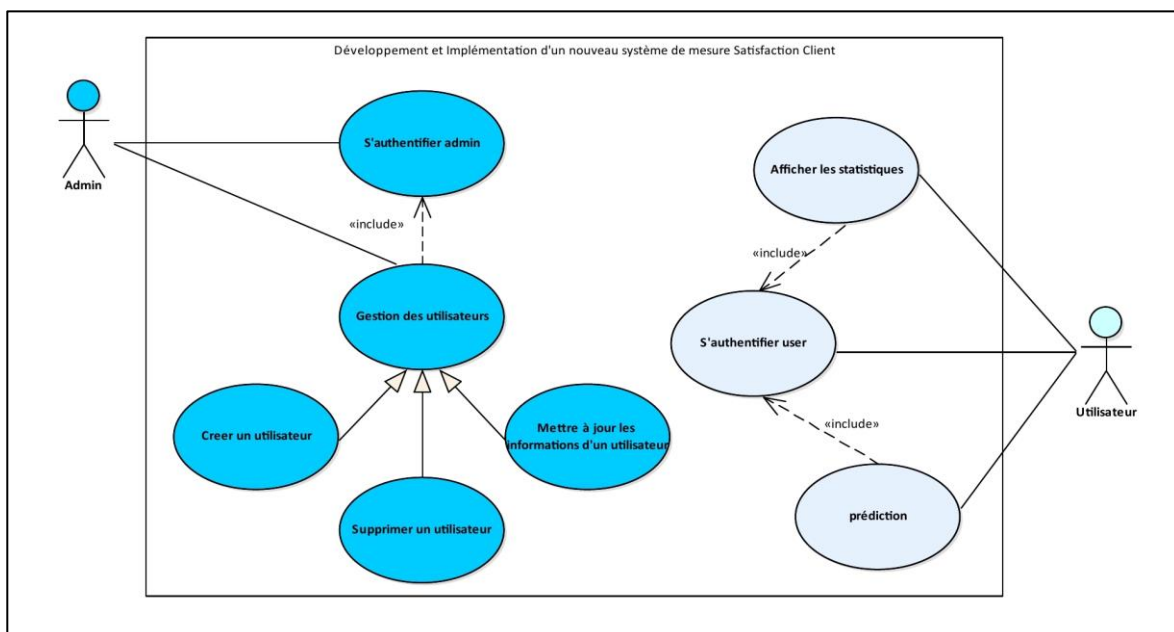


Figure 3.15 : Diagramme cas d'utilisation globale de l'application

### 3.5.2. Spécification des besoins non fonctionnels

La spécification des besoins non fonctionnels a pour but de fournir une application performante satisfaisante qui fait face aux risques de pannes et non fonctionnement. On cite ci-dessous les besoins non fonctionnels de notre application web :

- Disponibilité : l'application doit être simplement utilisée par les utilisateurs d'une façon intuitive.
- Sécurité : L'application doit prendre en compte la confidentialité des informations des utilisateurs
- Garantir la cohérence et l'intégrité des données lors des mises à jour et des insertions.
- L'ergonomie de l'interface graphique : l'application doit fournir une interface conviviale et facile pour l'utilisateur.
- Performance : le temps de réponse de l'application doit être très court.
- La modularité de l'application : le code doit être clair et facile pour permettre la maintenance et l'amélioration en cas de besoins.
- Possibilité de retour à la page d'Accueil à partir de n'importe quelle fenêtre.

Extensibilité : l'application doit permettre d'ajouter d'autres modules à tout moment.

### 3.6. Conclusion

Dans ce chapitre nous avons proposé une architecture globale de notre travail puis on l'a détaillé en deux parties, le premier côté données, nous avons expliqué le prétraitement en décrivant les différentes techniques avec quelques exemples de notre cas, ensuite nous avons cité une phase importante qui est la variable cible en justifiant le fait de la générer, et enfin nous avons présenté la phase de classification et l'évaluation des modèles, et à la fin nous avons présenté une brève description qui concerne notre application de test, le prochain chapitre sera consacré sur la réalisation de notre travail.

## 4. Réalisation et évaluation

### 4.1. Introduction

Dans le chapitre précédent, nous avons présenté nos solutions et le principe de fonctionnement. Dans ce chapitre nous allons voir la solution d’un point de vue pratique, nous exposerons les détails de l’implémentation, les outils utilisés, l’environnement matériel et logiciel de développement des phases de ce projet et de test

### 4.2. Environnement de développement

Nous allons décrire dans cette section les spécifications matérielles utilisées pour le développement et test en premier. Ce détail est important car le temps de calcul est un critère important pour la comparaison des solutions. En effet, une bonne solution est une solution qui n’exige pas beaucoup de ressources. Ensuite, nous exposerons le langage et les bibliothèques utilisés pour concrétiser la solution, avec la justification de chaque choix

#### 4.2.1. Matériels et outils de développement :

##### 4.2.1.1. Description du matériel de développement

En ce qui concerne l’environnement matériel, au début on était censé travailler sur les machines qu’on nous a offert à l’entreprise mais malheureusement et vu le covid-19 nous étions obligés d’utiliser nos propres machines qui sont les suivant :

	Machine 1 (MacbookAir)	Machine 2 (DELL)
Processeur	Intel Core i5 (1.7GHz)	Intel Core i7-5500U 2.4GHz
Mémoire	4 GO 1600 MHz DDR3	8 GO DDR
Stockage	64 GO SSD	1 TO HDD
Graphisme	IntelHD Graphics 4000	NVIDIA GeoForce

Tableau 4. 1: Spécification matériels

#### 4.2.1.2. Description des langages et logiciels

Pour l'implémentation de ce projet, le langage utilisé est **Python**, ainsi que ses différentes librairies hautement utilisées pour l'apprentissage automatique et le notamment l'apprentissage profond

##### 4.2.1.2.1. Choix du langage

Python n'est probablement pas le meilleur langage en termes de vitesse d'exécution, les bienfaits de ce langage sont néanmoins nombreux, surtout par rapport à notre cas d'étude.

Premièrement python est Human-friendly, la syntaxe est très proche de ce que l'humain pense et réfléchit et avec des variables dynamiques qui n'ont pas besoin d'être déclarées et des lignes de codes moins longues comparant à d'autres langages ;

En plus, et avec le langage **R**, python est un des langages les plus populaires et le plus performants dans les domaines de BigData, apprentissage machine et science de données, et ceci est grâce à ses riches standards de bibliothèques et frameworks pour des buts différents.



Figure 4. 1: Logo du langage PYTHON

#### 4.2.1.2.2. Bibliothèque utilisée :

Le tableau suivant résume les bibliothèques python utilisé pour la réalisation de ce projet :






Bibliothèque	Description	Quelques utilités	Version utilisée	Lien
	Bibliothèque open-source fondamentale du calcul scientifique	Traitement des tableaux multidimensionnels, algèbre linéaire, génération des nombres aléatoire, ...etc.	1.18.4	<a href="https://numpy.org/">https://numpy.org/</a>
	Bibliothèque open-source pour l'analyse et le traitement des données	Principalement conçu pour l'utilisation des structures DataFrame, analyse des données, importation des fichiers...etc.	0.24.2	<a href="https://pandas.pydata.org/">https://pandas.pydata.org/</a>
	Bibliothèque open-source pour la visualisation des données	Affichage des images et visualisation des différents types de graphs	3.0.3	<a href="https://matplotlib.org/">https://matplotlib.org/</a>
	Bibliothèque open-source créé à partir de Matplotlib pour la visualisation des données statistiques	Affichage des images et visualisation des différents types de graphs	0.9.0	<a href="https://seaborn.pydata.org/">https://seaborn.pydata.org/</a>
	Bibliothèque open-source d'apprentissage automatique	Outils simples et efficaces pour les modèles d'apprentissage automatique et d'analyse de données	0.22	<a href="https://scikit-learn.org/">https://scikit-learn.org/</a>

Tableau 4. 2: Bibliothèque Python utilisés

En plus de ces bibliothèques mentionnées ci-dessous, nous avons utilisé le Package **Keras** pour un modèle de réseaux de neurones et **Django** pour concrétiser les résultats et les statistiques dans une plateforme web.

- **Keras** est une librairie de réseaux de neurones de haut niveau, écrite en Python et ineffaçable avec TensorFlow, CNTK ou Theano. Elle a été développée avec pour objectif de permettre des expérimentations rapides. Être capable d'aller de l'idée au résultat avec le plus faible délai possible étant la clef d'une recherche efficace [16]

- **Django** est un framework Python de haut niveau, permettant un développement rapide de sites internet, sécurisés, et maintenables.



#### 4.2.1.2.3. Outil de développement

L'outil de développement utilisé pour la réalisation de ce projet est **Jupyter notebook**, c'est un environnement interactif pour l'exécution du code sur un navigateur web, c'est un outil très agréable et compatible pour l'exploration des données et l'analyse, et hautement utilisé par les praticiens de la science de données. Jupyter notebook inclut le support de plusieurs langages de programmation, cependant on va seulement l'utiliser pour l'exécution du code python.

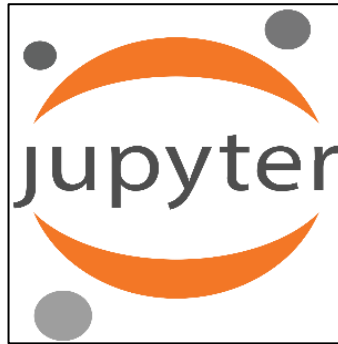


Figure 4. 2: Logo du l'outil Jupyter

#### 4.2.2. Application de mesure de satisfaction client

En ce qui concerne l'application, nous avons mis en place un Dashboard en utilisant le Framework Django, système d'authentification géré par le SUPERUSER (Administrateur) les figures ci-dessous montre l'interface Dashboard de l'administrateur :

##### 1. Administration :

##### 1.1. Page d'Authentification :



Figure 4. 3: Capture – admin Django Login Page

L'administrateur doit s'authentifier pour pouvoir accéder à l'administration, la création d'un SuperUser se fait à partir de la console (la première fois) en utilisant la commande de Django :

```
(base) MacBook-Air-2:SatisfactionClient_Djezzy FILIN0$ python manage.py createsuperuser
Username: username
Email address: username@mail.com
Password: ?
```

**Figure 4. 4: Capture – createsuperuser**

Ensuite, la gestion des utilisateurs sera faite directement à partir de l'interface admin de Django.

### 1.2.Interface Administrateur Django :

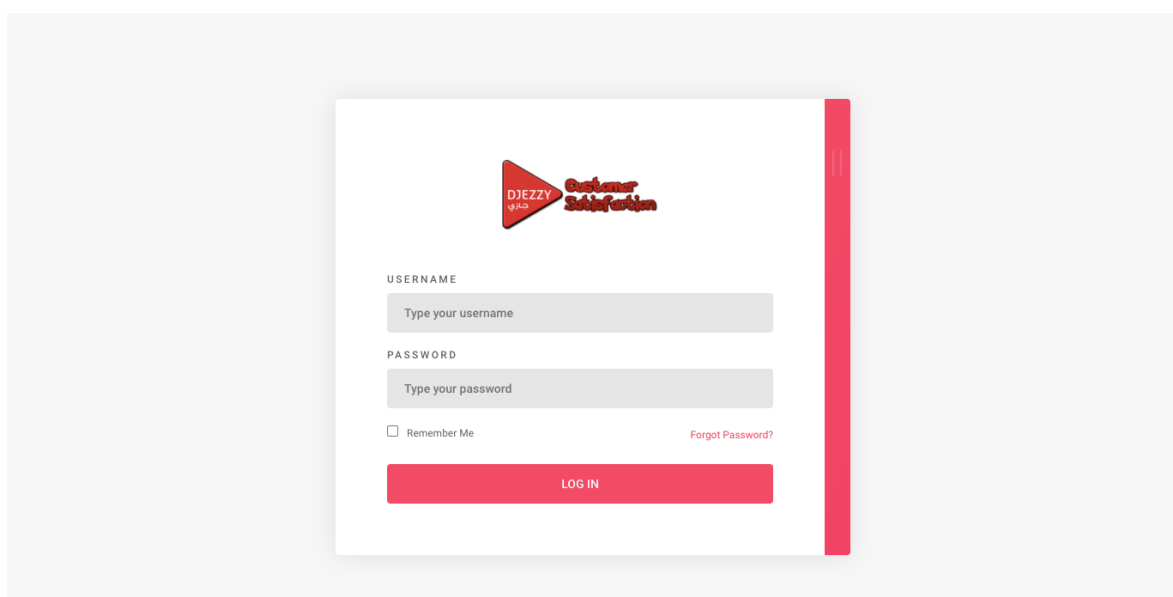


**Figure 4. 5: Capture – admin dashboard**

L'administrateur possède les autorisations suivantes :

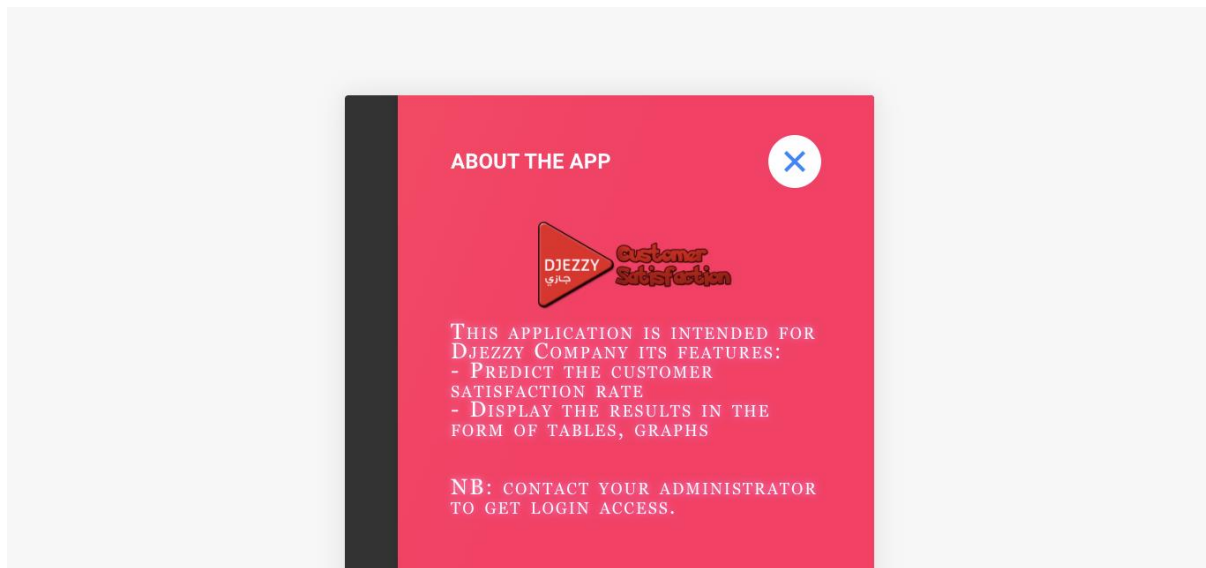
- 1 Création des comptes utilisateurs
- 2 Création des groupes
- 3 La gestion des utilisateurs

Une autre page d'authentification pour les utilisateurs a été mise en place la figure suivante montre son interface :



**Figure 4. 6: Capture – user login**

User Name et Mot de passe seront offerte par l'administration qui est le gestionnaire de l'application.

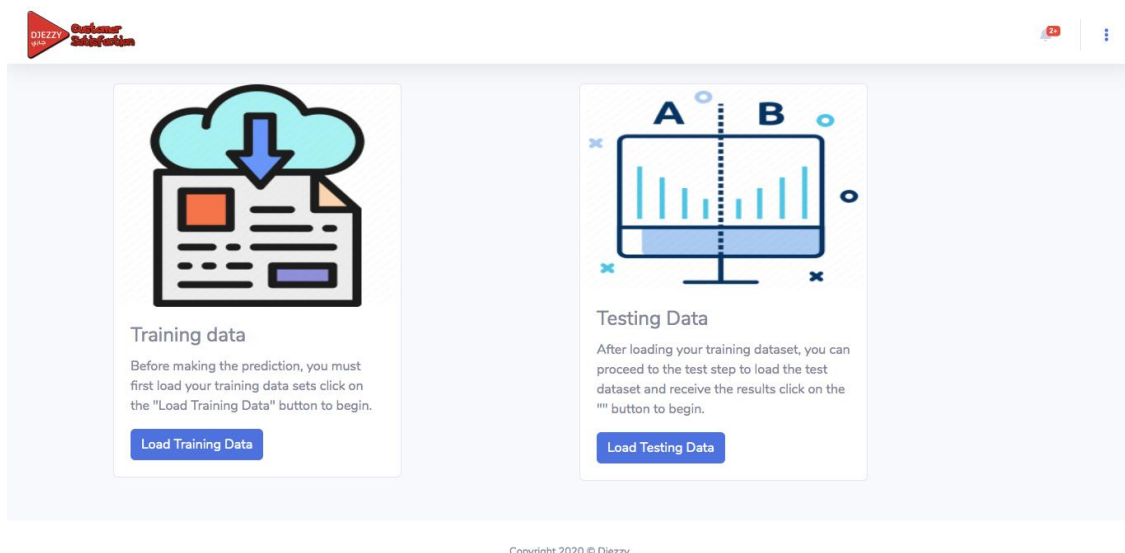


**Figure 4. 7: Capture – user login – infos**

Si l'utilisateur n'arrive pas à accéder en utilisant ses identifiants, il doit contacter l'administrateur pour corriger le problème.

## 2. Dashboard d'entraînements et de tests :

Ce dashboard concerne l'entraînement des données, avant d'avoir les résultats il est nécessaire d'entraîner un jeu de données et faire les tests par la suite la figure suivante montre la page « Landing » qui s'affichera lors de la première utilisation de l'application :



**Figure 4. 8: Capture – landing dashboard**

En cliquant sur le bouton « Load Training Data » la page d'entraînement va s'ouvrir, la figure suivante montre la page « Training Data »

- Page « Training Data » :

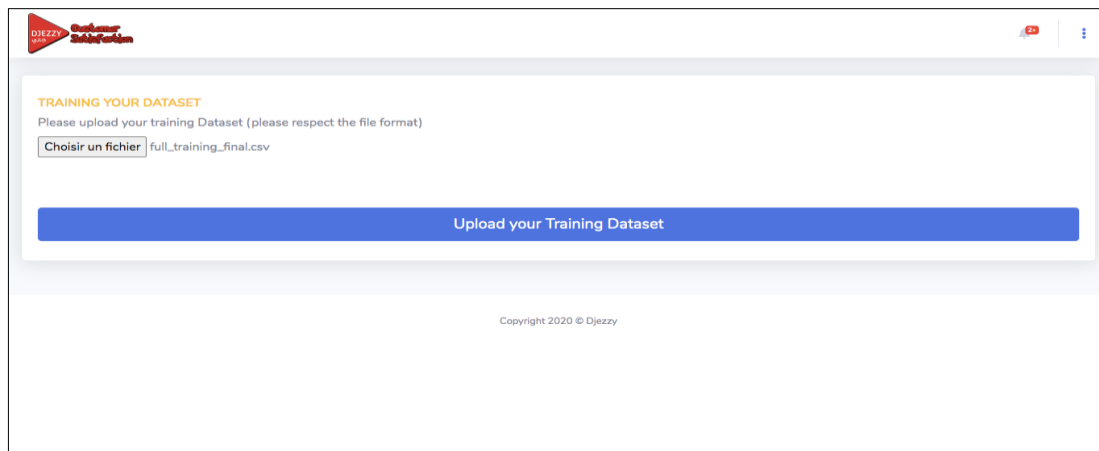


Figure 4. 9: Capture - page Training Data

L'utilisateur doit sélectionner un fichier csv contenant les données d'entrainements (y compris la satisfaction pour chaque client), en cliquant sur « **Upload your Training Data** » l'entrainement de données commencera, le résultat sera sauvegardé dans un dossier afin de pouvoir l'utiliser ultérieurement, la figure suivante montre le processus d'entrainement via la console « Terminal »

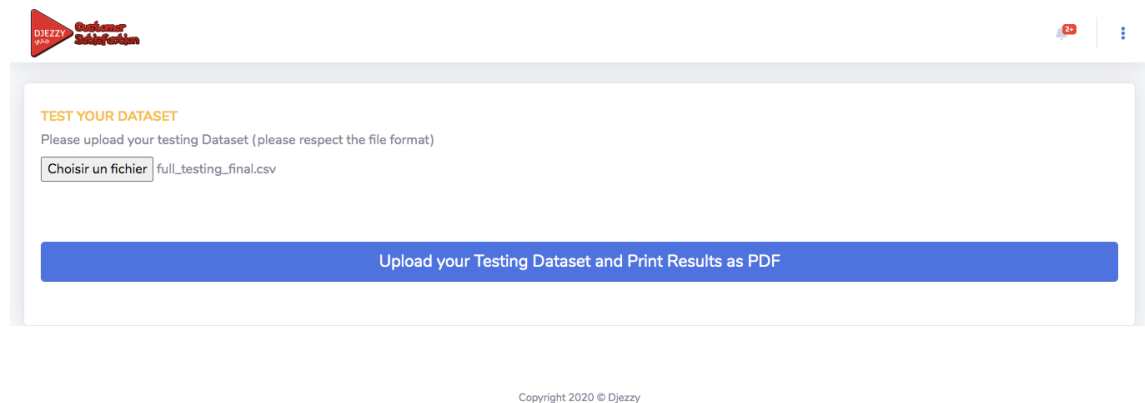
**Remarque** : le fichier doit respecter le format exigé sinon l'opération ne va pas s'effectuer.

```
63/63 [=====] - 0s 8ms/step - loss: 0.2878 - accuracy: 0.8857
Epoch 26/40
63/63 [=====] - 1s 8ms/step - loss: 0.2762 - accuracy: 0.8890
Epoch 27/40
63/63 [=====] - 0s 8ms/step - loss: 0.2805 - accuracy: 0.8825
Epoch 28/40
63/63 [=====] - 0s 8ms/step - loss: 0.2671 - accuracy: 0.8880
Epoch 29/40
63/63 [=====] - 1s 8ms/step - loss: 0.2495 - accuracy: 0.8995
Epoch 30/40
63/63 [=====] - 0s 8ms/step - loss: 0.2568 - accuracy: 0.8892
Epoch 31/40
63/63 [=====] - 1s 9ms/step - loss: 0.2471 - accuracy: 0.8945
Epoch 32/40
63/63 [=====] - 0s 8ms/step - loss: 0.2425 - accuracy: 0.9005
Epoch 33/40
63/63 [=====] - 1s 8ms/step - loss: 0.3045 - accuracy: 0.8913
Epoch 34/40
63/63 [=====] - 0s 8ms/step - loss: 0.2675 - accuracy: 0.8932
Epoch 35/40
63/63 [=====] - 0s 8ms/step - loss: 0.2499 - accuracy: 0.8955
Epoch 36/40
63/63 [=====] - 1s 8ms/step - loss: 0.2380 - accuracy: 0.9050
Epoch 37/40
63/63 [=====] - 0s 8ms/step - loss: 0.2432 - accuracy: 0.9020
Epoch 38/40
63/63 [=====] - 1s 8ms/step - loss: 0.2317 - accuracy: 0.9015
Epoch 39/40
63/63 [=====] - 0s 8ms/step - loss: 0.2469 - accuracy: 0.8990
Epoch 40/40
63/63 [=====] - 1s 8ms/step - loss: 0.2301 - accuracy: 0.9038
The model is not saved
The file has been trained successfully, the training data is saved on /api/train
[05/Sep/2020 19:07:41] "POST /training/ HTTP/1.1" 200 11467
[05/Sep/2020 19:07:41] "GET /static/js/forms.js HTTP/1.1" 304 0
```

Figure 4. 10: Capture - Processus d'entrainement

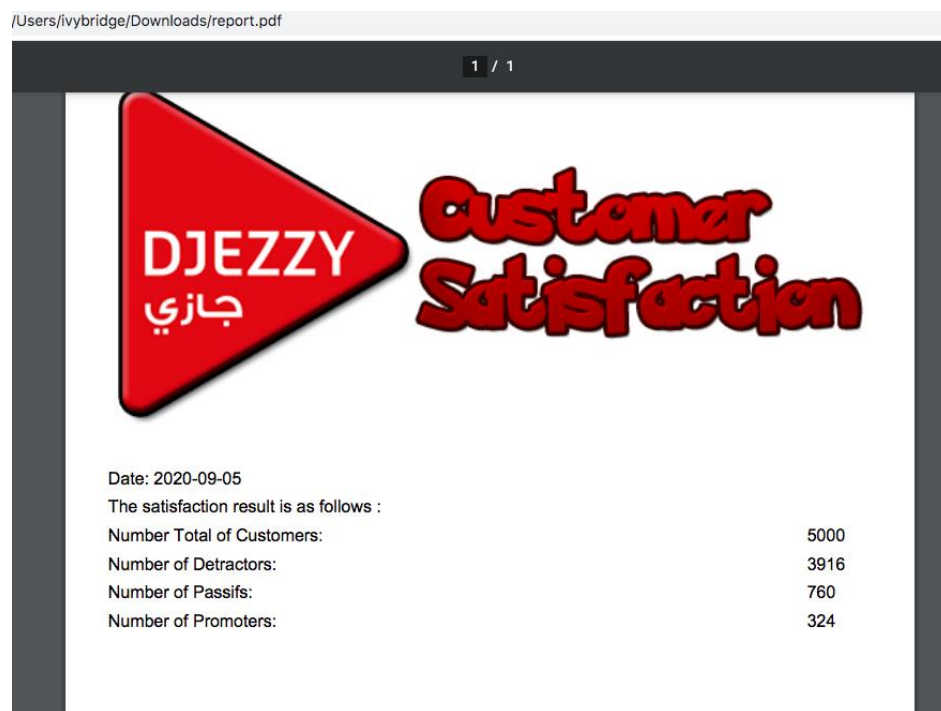
- Page « Testing Data » :

À partir de la page « Landing » (ou bien le menu), on peut accéder à la page de test, la figure suivante le montre :



**Figure 4. 11: Capture - Page Testing Data**

L'utilisateur doit charger le dataset qui contient les données de tests, en cliquant sur le bouton « Upload your Testing Dataset and Print Results as PDF », les données seront traitées, la satisfaction sera générée à partir du modèle entraîné précédemment et pour bien éclairer les choses, un PDF contenant un rapport des résultats sera téléchargé automatiquement la figure suivante montre un exemple des résultats :



**Figure 4. 12: Capture – exemple d'un résultat de tests format pdf**

Les résultats seront sauvegardés également sous format CSV afin qu'utilisateur pourra revoir les résultats d'une façon détaillé en consultant le « Dashboard des résultats », la figure suivante montre un exemple comment les données sont stockés sur l'application :

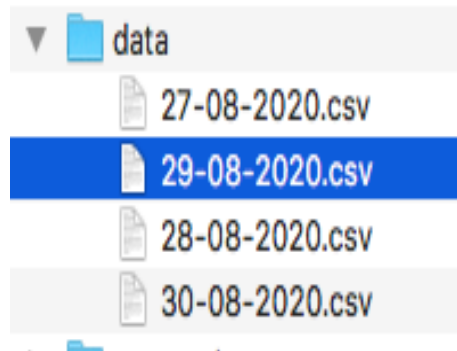


Figure 4. 13: capture – stockage des résultats sous format CSV

### 3. Dashboard des résultats :

Un Dashboard a été mis en place pour visualiser les résultats des tests, l'authentification des utilisateurs est nécessaire pour y accéder.

L'utilisateur a la possibilité de consulter les tests (ou prédiction) effectué précédemment en utilisant plusieurs options :

- **En sélectionnant le fichier de test :**

Les fichiers de test sont classés par date, il peut sélectionner n'importe quel fichier afin de visualiser les résultats qui lui correspond.

- **En sélectionnant une région/Wilaya :**

Si l'utilisateur souhaite spécifier les résultats qui concernent uniquement la Wilaya désiré, il a qu'à sélectionner la Wilaya et la page va afficher uniquement les résultats concernés.

- **Impression de la page :**

L'utilisateur peut imprimer les résultats en cliquant sur le bouton tout en bas « Print this page » afin d'avoir une version papier des résultats concerné.

- L'utilisateur a le droit de visualiser les résultats avec plusieurs options :

#### **Globalement :**

Par défaut, le résultat affiché concerne la totalité des tests, plusieurs graphes en été mis en place

La figure ci-dessous représente la page dashboard des résultats qui concerne la totalité des tests.

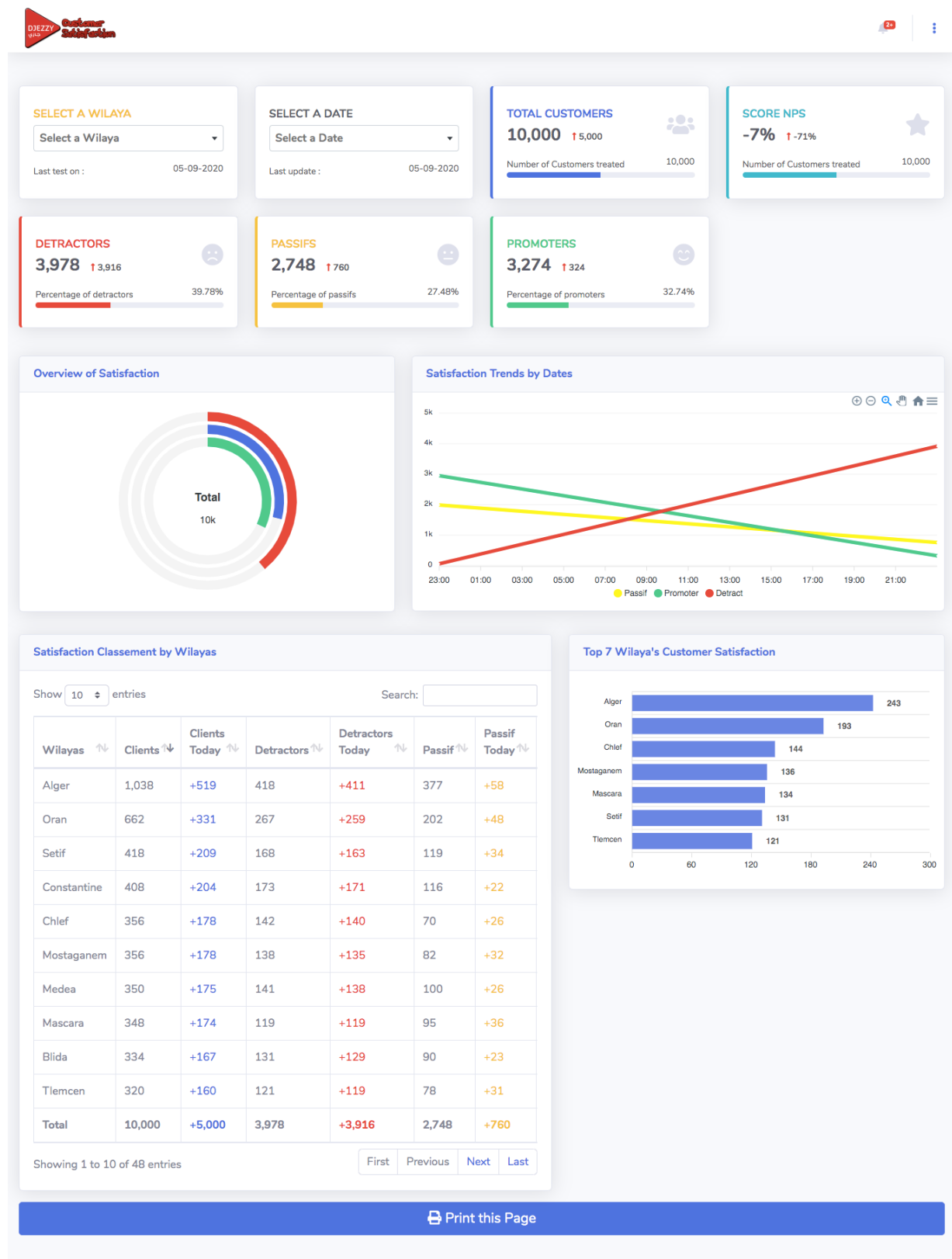


Figure 4. 14: Capture – dashboard des résultats

Le graphe suivant montre le taux global de satisfaction de nos tests :



Figure 4. 15: Exemple graphe – Overview of Satisfaction

Le graphe suivant montre un exemple du taux de satisfaction sur une période préfini :

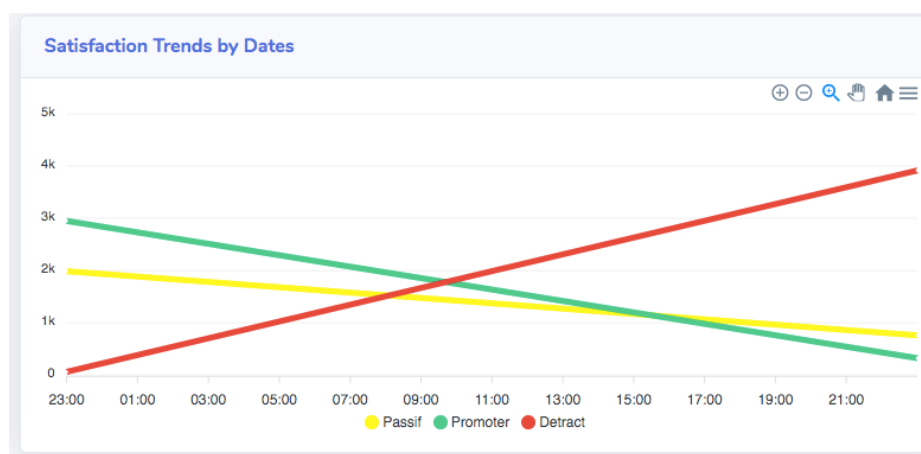
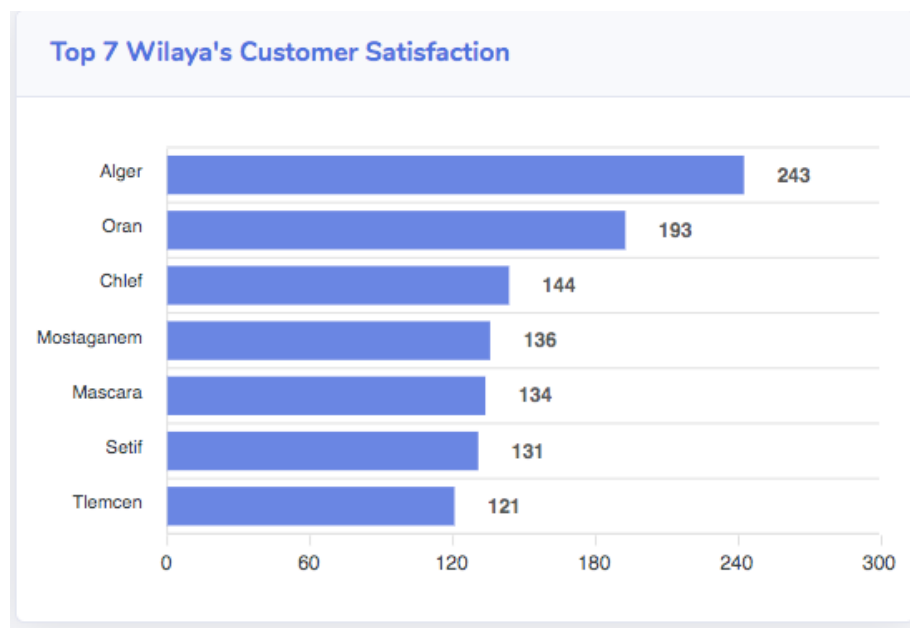


Figure 4. 16: Exemple graphe – Satisfaction Trends by Dates



Le graphe suivant montre un exemple des 7 Wilayas qui ont le taux de satisfaction le plus élevé :



**Figure 4. 17: Exemple graphe – Top 7 Wilaya's Customer Satisfaction**

La figure suivante montre un exemple d'un tableau de classement des résultats par Wilaya, avec la possibilité de faire une recherche par n'importe quel paramètre des résultats

Satisfaction Classement by Wilayas

Show  entries Search:

Wilayas	Clients	Clients Today	Detractors	Detractors Today	Passif	Passif Today
Medea	13,045	+2,180	4,419	+770	4,338	+703
Relizane	12,954	+2,096	4,271	+694	4,320	+699
Batna	12,918	+2,200	4,258	+726	4,390	+754
Ain Temouchent	12,875	+2,158	4,278	+719	4,308	+727
Ouargla	12,867	+2,091	4,250	+692	4,370	+706
Tiaret	12,861	+2,134	4,262	+700	4,276	+716
Setif	12,846	+2,157	4,282	+714	4,372	+743
Skikda	12,844	+2,103	4,217	+702	4,293	+691
Illizi	12,840	+2,098	4,245	+714	4,330	+689
Naama	12,829	+2,191	4,225	+700	4,373	+776
Total	610,000	+100,000	203,269	+33,417	203,363	+33,243

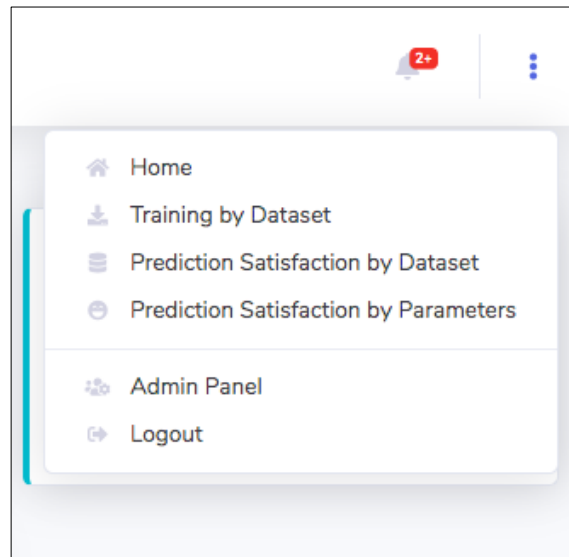
Showing 1 to 10 of 48 entries

First Previous Next Last

**Figure 4. 18: Exemple Tableau des résultats classé par Wilayas**

L'interface du Dashboard est dynamique et offre la possibilité d'afficher les résultats par Wilayas et Région :

Un menu a été mis à la disposition de l'utilisateur afin de parcourir facilement les pages de l'application.



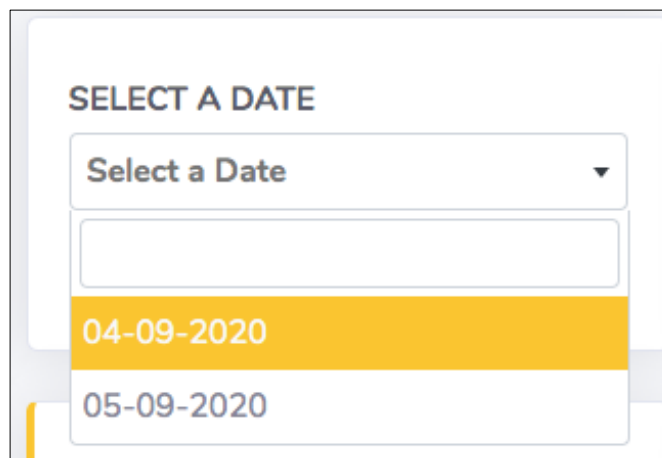
**Figure 4. 19: Capture – menu**

**En plus de détail :**

L'utilisateur peut également obtenir des résultats en utilisant deux options :

- **Par Date :**

L'utilisateur peut spécifier les résultats par date la figure suivante montre un exemple :



**Figure 4. 20: Sélection des résultats par date**

- **Par Wilaya :**

L'utilisateur peut spécifier une Wilaya afin d'obtenir uniquement les résultats qui concerne la Wilaya sélectionné, la figure suivante montre un exemple :

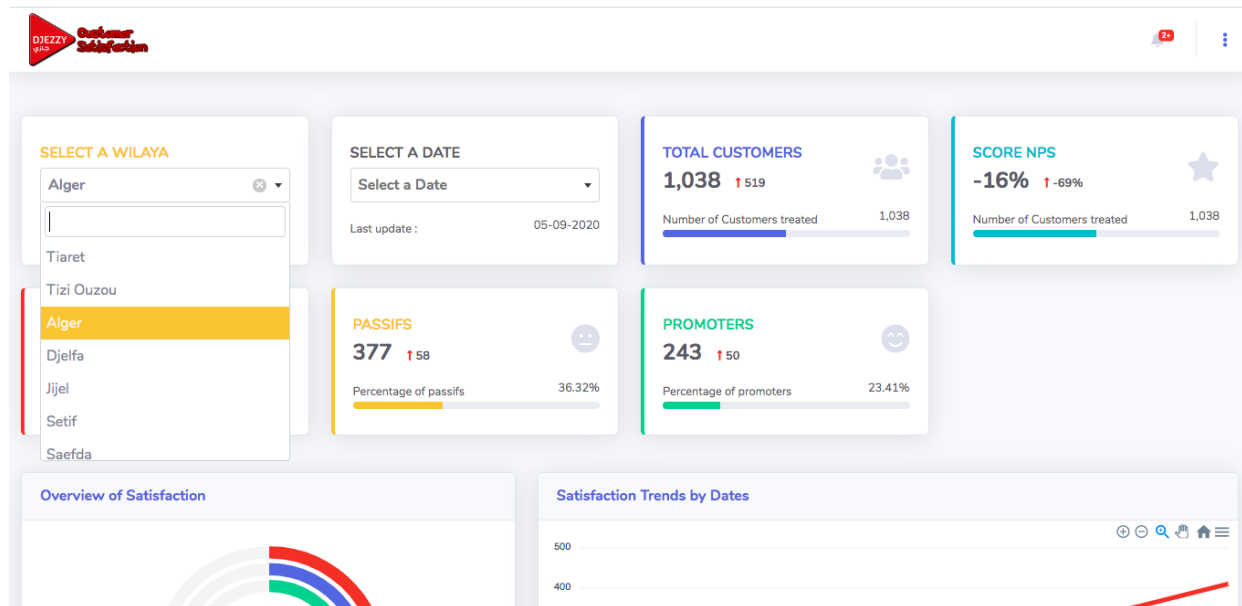
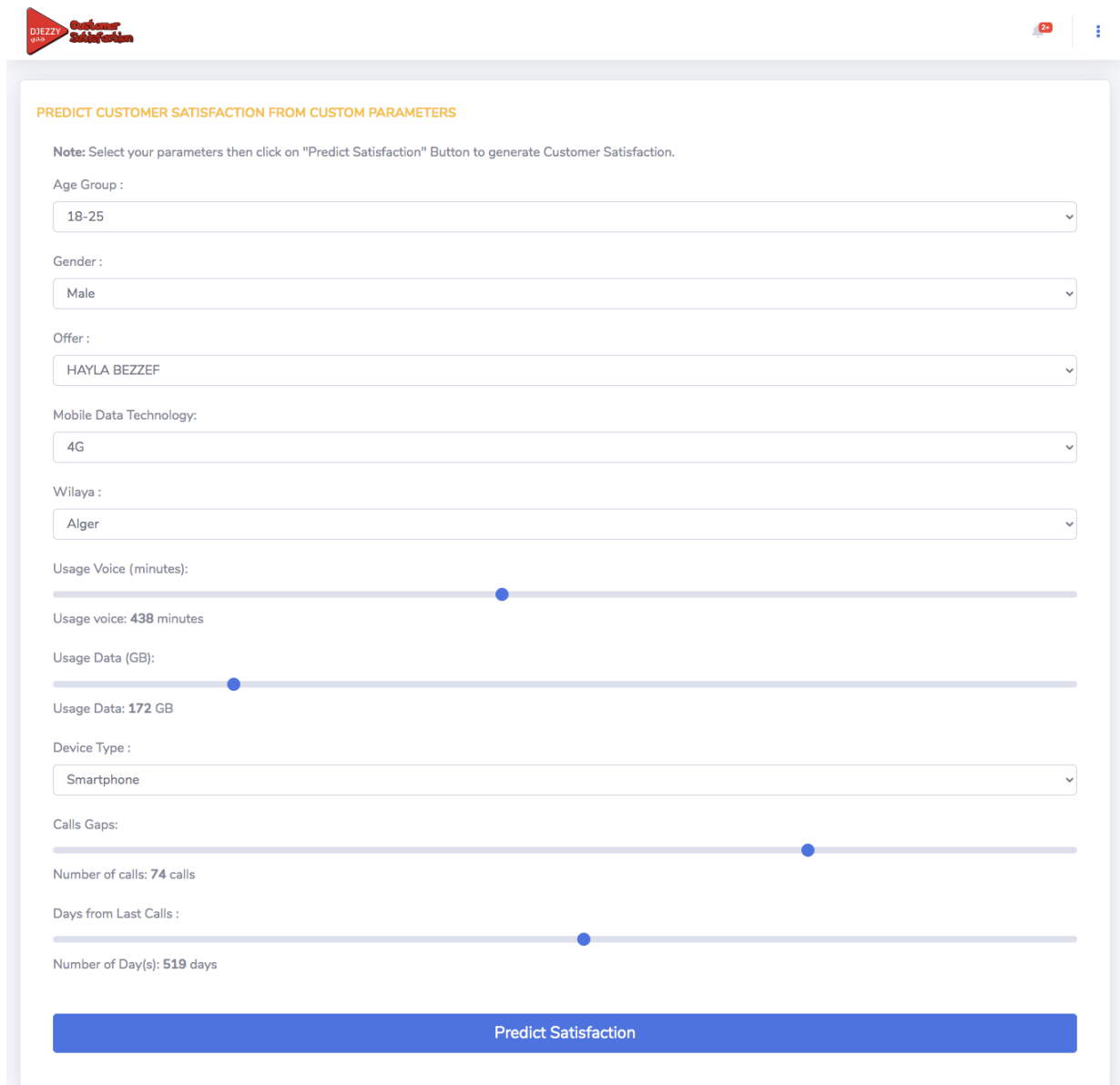


Figure 4. 21: Exemple d'une sélection des résultats par Wilaya

#### 4. Page Predict Satisfaction By Parameters

Cette page concerne la génération du taux de Satisfaction à partir des paramètres saisis par l'utilisateur, cette option est très importante pour le décideur afin d'estimer quelle sont les paramètres sensibles de la Satisfaction Client, en cliquant sur « Predict Satisfaction », un résultat sera affiché instantanément qui prédit la satisfaction du client concerné, la figure suivante montre une capture de la page « Predict Satisfaction By Parameters »



The screenshot shows a web application interface for predicting customer satisfaction. At the top left is the 'DJEZZY' logo. The main heading is 'PREDICT CUSTOMER SATISFACTION FROM CUSTOM PARAMETERS'. A note states: 'Note: Select your parameters then click on "Predict Satisfaction" Button to generate Customer Satisfaction.' The form contains several input fields and sliders:

- Age Group :** A dropdown menu with '18-25' selected.
- Gender :** A dropdown menu with 'Male' selected.
- Offer :** A dropdown menu with 'HAYLA BEZZEF' selected.
- Mobile Data Technology:** A dropdown menu with '4G' selected.
- Wilaya :** A dropdown menu with 'Alger' selected.
- Usage Voice (minutes):** A horizontal slider with a blue dot in the middle. Below it, the text 'Usage voice: 438 minutes' is displayed.
- Usage Data (GB):** A horizontal slider with a blue dot on the left side. Below it, the text 'Usage Data: 172 GB' is displayed.
- Device Type :** A dropdown menu with 'Smartphone' selected.
- Calls Gaps:** A horizontal slider with a blue dot on the right side. Below it, the text 'Number of calls: 74 calls' is displayed.
- Days from Last Calls :** A horizontal slider with a blue dot in the middle. Below it, the text 'Number of Day(s): 519 days' is displayed.

At the bottom of the form is a large blue button labeled 'Predict Satisfaction'.

Copyright 2020 © Djezzzy

**Figure 4. 22: Capture – page Predict by Parameters**

Le Dashboard offre également la possibilité d’afficher les résultats selon les caractéristiques désirés y compris le sexe des clients, le groupe d’âge, l’offre mobile choisi, la technologie du réseau utilisé ainsi que d’autres paramètres ce qui offre à l’analyste une décision bien précise afin d’améliorer les services offerts.

### 4.3. Évaluation du système

#### 4.3.1. Dataset et prétraitement

Les données de mauvaise qualité influencent gravement les résultats et performances du modèle d'apprentissage automatique, pour espérer garantir un bon score et une précision acceptable, un prétraitement est exigé.

Comme expliqué précédemment, le prétraitement est fait par rapport aux problèmes que les données subissent.

Pour notre cas, les données fournis par l'établissement d'accueil ont besoin d'un prétraitement avant de passer à l'évaluation de chaque.

##### 4.3.1.1. Exploit des données catégoriques

Pour un modèle d'apprentissage automatique supervisé ayant pour but une classification ou régression, les données dont les valeurs sont catégoriques ne peuvent pas être exploitées, donc un remplacement est exigé, prenons l'exemple de la variable **MOBILE\_DATA\_TECHNOLOGY**, qui représente la technologie utilisé par l'abonné, elle a comme valeurs '2G', '3G' et '4G', qui sont des chaînes de caractères, le graphe du côté droit de la figure suivante représente la distribution des trois valeurs de ce cas.

Maintenant que les valeurs sont identifiées, le changement des chaînes de caractère par des valeurs numérique est possible et les nouvelles valeurs de la variable sont donc numériques : 2,3 et 4

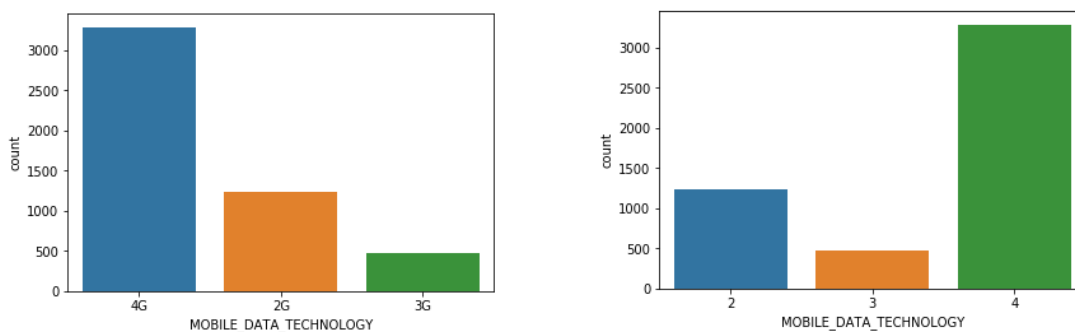


Figure 4. 23: Distribution des trois valeurs « MOBILE\_DATA\_TECHNOLOGY »

##### 4.3.1.2. Jointure des fichiers

Les données sont fournies en forme de plusieurs fichiers de type texte, séparés les uns des autres, mais contenant une variable commune qui nous permet de les joindre convenablement en passant par une fonction de jointure.

Pour la plupart des fichiers utilisés, la variable en commun c'est l'identificateur de l'abonné.

Cette action paraît simple, mais dans plusieurs cas il y a un problème de nomination des variables en communs comme l'exemple si dessous, on remarque la différence des nominations malgré l'indifférence de la signification, comme l'exemple ci-dessous :

Entrée [29]: 1 Network\_Sample.head(3)

Out[29]:

	ID	USE_QUOTA_CLUSTER	USE_NW_QUOTA_SEGM	HANDSET_TYPE	HANDSET_BRAND	targetted
0	22617449	1.0	1	3	1.0	1
1	109581158	1.0	1	1	2.0	1
2	641673	1.0	1	1	2.0	1

Entrée [36]: 1 full\_handset\_sub.head(3)

Out[36]:

	TAC	Devicetype	LTE	Subscriber_ID
0	35896710	1.0	0.0	139598004
0	35896710	1.0	0.0	27653849
1	35565805	2.0	0.0	144703084

Figure 4. 24: Exemple réel d'une jointure de fichiers

Pour y remédier, il suffit de changer la nomination d'une des deux variables ou de les changer les deux si la nomination et la signification ne coïncide pas :

ID	Subscriber_ID
0 22617449	0 22617449
1 109581158	1 109581158
2 641673	2 641673
3 129290187	3 129290187
4 146939651	4 146939651

Figure 4. 25: Exemple réel d'un changement de nomination

En changeant la nomination, la jointure est donc possible, et pour l'exemple du problème posé en conception, le résultat sera comme suite :

	TAC	Devicetype	LTE	Subscriber_ID	USE_QUOTA_CLUSTER	USE_NW_QUOTA_SEGM	HANDSET_TYPE	HANDSET_BRAND	targetted
1147	35495109	2.0	1.0	147563685	2.0	2.0	1.0	1.0	0.0
1815	35705708	1.0	0.0	109225074	1.0	1.0	3.0	1.0	0.0
532	35719109	2.0	1.0	19735253	1.0	1.0	1.0	2.0	1.0

Figure 4. 26: Exemple réel du résultat de jointure

Concernant le reste des fichiers fournis, ceci est fait de la même façon seulement pour ceux qui contient une variable commune par rapport aux autres.

#### 4.3.1.3. Gestion des redondances

Maintenant que les fichiers de données ont bien été joints, on commence par des traitements concernant les valeurs des données.

Commençant par gérer les données redondantes. Une donnée ou variable redondante est détectée par une analyse de corrélation, cette dernière doit être d'une valeur 1 ou presque pour deux variables afin de conclure que la variable est redondante, la figure suivante représente quelques valeurs de deux variables qui signifient l'âge de chaque abonné :

BIRTH	CUSTOMER_AGE_CLASS
43	43
67	67
22	22
53	53
28	28
82	82

**Figure 4. 27: Exemple réel de quelques valeurs qui signifient l'âge de chaque abonné**

En analysant la corrélation de ces deux variables, on trouve qu'elle est bel et bien égale à 1, donc c'est une variable redondante.

	BIRTH	CUSTOMER_AGE_CLASS
BIRTH	1.000000	1.000000
CUSTOMER_AGE_CLASS	1.000000	1.000000

**Figure 4. 28: Exemple réel d'une valeur redondante**

### 3.1.4. Gestion des données manquantes

Les données manquantes sont un phénomène qui réduit le plus les performances d'un modèle d'apprentissage automatique, voir même l'impossibilité de réaliser l'entraînement sous sa présence.

Certains fichiers de données relatifs à ce projet contiennent des données manquantes dans leur structure. Sur quelques fichiers ce phénomène est presque partout, et sur quelques d'autres il ne l'est pas.

Certaines variables ont un pourcentage très élevé, et pour d'autres il est très bas, donc la façon avec laquelle on doit procéder pour les éliminer diffère d'une variable à l'autre mais comment les identifier ?

Les données manquantes de toutes les variables de chaque fichier utilisé sont détectées à partir d'une étude de chacune des variables et par rapport à des pourcentages calculés et affichés en langage Python.

La figure suivante représente les pourcentages des données manquantes pour chaque variable du fichier User\_info

TARIFF_PROFILE_POST_PREP	100.00
B2B_SEGMENT_TYPE	99.98
COORDINATOR_NAME	99.88
DESCRIPTION	48.70
OWNERD	48.70
AGE_GROUP	8.32
Code_Wilaya	2.82
REGION_ID	2.82
Region_Name	2.82
RESRVATION_CODE	1.24
segmentation_category	0.28
WILAYA	0.02
BIRTH	0.00
CUSTOMER_AGE_CLASS	0.00
GENDER	0.00
SUBSCRIPTION_TYPE	0.00
TARIFF_PLAN	0.00
PREPAID_IND	0.00
CUSTOMER_TYPE	0.00
TARIFF_PROFILE	0.00
LANGUAGE_INDICATOR	0.00
Subs_Id	0.00

**Figure 4. 29: Pourcentage de données manquantes pour chaque variable du fichier User\_info**

On voit qu'il y a plusieurs variables ayant un pourcentage supérieur à 40%, d'autres inférieurs à 40% et même quelques-unes vides, le tableau suivant résume les deux actions à faire pour chaque cas :

Cas	Action
Plus de 40 % des données manquantes	Variable supprimée
Moins de 40% des données manquantes	Variable gardée et remplie pour chaque cas convenablement

**Tableau 4. 3: Traitement des données manquantes**

Pour le premier cas, les variables ont été supprimées, quant au deuxième, ils ont été gardés et remplis, et ceci à une façon particulière :

Les données manquantes des variables catégoriques et numérique discrète ont été remplies par le mode, et les données numérique continue par la moyenne de la variable.

La figure suivante représente cinq observations de deux variables numériques continues, usage\_voice et usage\_data, et une variable originellement catégorique, deviceType, ayant un nombre de données manquantes



	Subs_Id	usage_voice	usage_data	Devicetype	MOBILE_DATA_TECHNOLOGY
4806	76455011	NaN	NaN	NaN	2
4858	147431219	NaN	NaN	NaN	4
4929	74920212	NaN	NaN	NaN	2
4975	27866802	NaN	NaN	NaN	2
4987	82078352	NaN	NaN	NaN	3

Figure 4. 30: Cinq observations de deux variables numériques continues usage\_voice et usage\_data

Usage\_voice et Usage\_data vont être rempli par une moyenne, et deviceType par rapport au mode de la variable, mais certainement pas directement.

En faisant une étude de corrélation entre la variable, la plus forte variable corrélée avec chacune des trois est utilisée pour grouper les valeurs et remplir les manques pour chaque groupe :

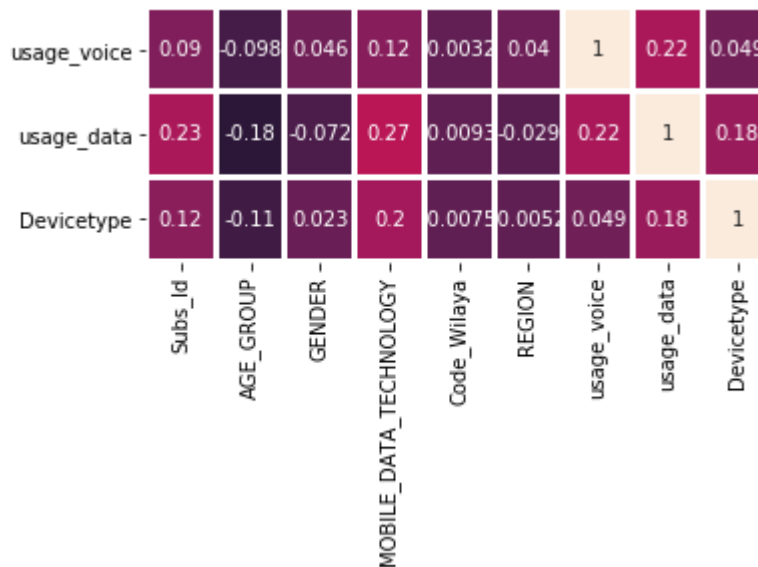


Figure 4. 31: Corrélations entre les variables

En analysant la figure de corrélation, on trouve que la variable la plus corrélée avec usage\_voice est usage\_data, mais cette dernière a aussi un manque et en plus elle est continue donc les groupe ne seront totalement pas récurrent, dans ce cas on prend la deuxième variable la plus corrélée qui est MOBILE\_DATA\_TECHNOLOGY avec un résultat positif de 0.12 (pas fort).

Ensuite, la variable la plus corrélée avec usage\_data et avec deviceType est aussi MOBILE\_DATA\_TECHNOLOGY avec un résultat de 0.27 et 0.2 respectivement, c'est des résultats toujours pas forts.

Donc en groupant MOBILE\_DATA\_TECHNOLOGY, qui est la plus corrélée avec chacune des trois, les résultats du remplissage de données manquantes pour chaque groupe de 2,3 et 4 est comme suivant :

	Subs_Id	usage_voice	usage_data	Devicetype	MOBILE_DATA_TECHNOLOGY
4806	76455011	297.424459	10.202464	1.0	2
4858	147431219	499.961987	3894.290489	2.0	4
4929	74920212	297.424459	10.202464	1.0	2
4975	27866802	297.424459	10.202464	1.0	2
4987	82078352	440.638004	1162.287792	1.0	3

Figure 4. 32: Résultats du remplissage de données manquantes

#### 4.3.1.4. Variables avec valeurs dominantes

Dans notre cas, certaines variables ont des valeurs très dominantes par rapport aux autres valeurs, c'est-à-dire que pour la plupart des observations, les cases de cette variable ont presque seulement une seule valeur.

Les variables ayant ce cas-là n'apportent aucune information de valeur et aucun plus pour l'étude et pour la précision du modèle, et doivent être supprimés

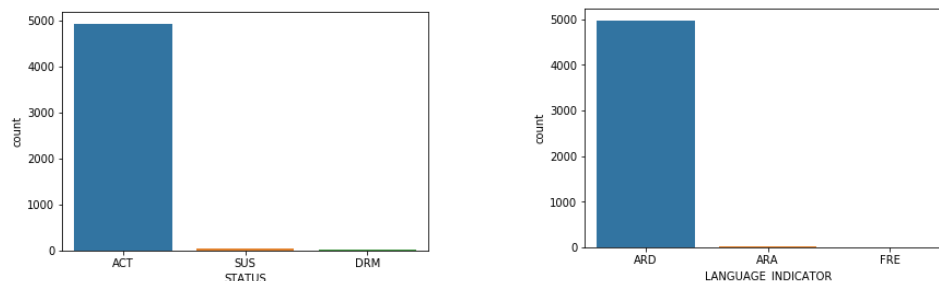


Figure 4. 33: Exemple d'une variable avec valeurs dominantes

#### 4.3.1.5. Génération de la variable cible

La génération de la cible est faite d'une façon particulière et pas au hasard, elle est générée par rapport à des intervalles de variables continues et selon l'algorithme mentionné.

Comme décrit précédemment, le score NPS est décrit et calculé à partir de 3 classes, Promoteurs, Passifs et Détracteurs, et les classes de la variable cible générée sont 2,1 et 0, et ils sont attribués de la façon suivante :

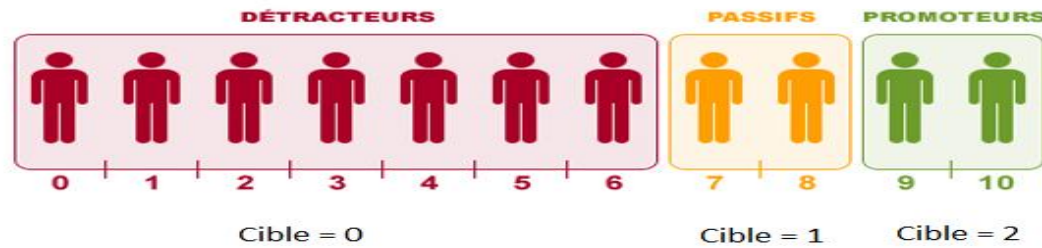


Figure 4. 34: Cibles de données

La variable cible générée est nommée 'Satisfaction', voici un exemple de 10 observations de l'état final du Dataset après tout type de prétraitement :

AGE_GROUP	GENDER	OFFER	MOBILE_DATA_TECH	Code_Wilaya	REGION	usage_voice	usage_data	Devicetype	calls_gap	lastcall_since	Satisfaction
3	0	11	2	21	2	65.960000	0.370000	1.0	4224	79	0
2	0	11	2	19	2	43.730000	0.000000	1.0	2357	75	0
6	0	2	2	35	3	280.080000	0.000000	1.0	4519	74	1
5	0	5	4	46	4	1824.230000	27537.620000	3.0	173	74	2
2	0	6	4	6	3	499.961987	3894.290489	2.0	14	74	1
3	0	8	4	13	4	293.280000	0.060000	1.0	771	74	1
6	1	5	2	19	2	176.410000	5.890000	3.0	4880	74	0
3	0	5	4	27	4	213.320000	977.570000	3.0	165	74	1
4	0	7	4	42	3	33.480000	0.000000	3.0	4321	75	0
4	0	5	4	19	2	555.690000	0.000000	2.0	4822	74	1

Figure 4. 35: Exemple réel de 10 observations de l'état final du Dataset

### 4.3.2. Métriques d'évaluations

L'évaluation des modèles c'est un processus qui consiste à comprendre à quelle précision ils donnent la classification correcte.

Pour choisir le meilleur algorithme possible à utiliser, et pour évaluer le modèle de classification final, on utilise une certaine matrice appelée **matrice de confusion**.

#### 4.3.2.1. Matrice de confusion

Une matrice de confusion (ou tableau de contingence) est un outil permettant de mesurer les performances d'un modèle de machine learning en vérifiant notamment à quelle fréquence ses prédictions sont exactes par rapport à la réalité dans des problèmes de classification.

	P' (Prédite)	N' (Prédite)
P (Réel)	Vraie Positive (VP)	Faux Négative (FN)
N (Réel)	Faux Positive (FP)	Vraie Négative (VN)

Tableau 4. 4: Matrice de confusion

Pour bien comprendre le fonctionnement d'une matrice de confusion, il convient de **bien comprendre les quatre terminologies principales : VP, FN, FP, VN**. Voici la définition précise de chacun de ces termes

- **VP** : les cas où la prédiction est positive, et où la valeur réelle est effectivement positive.
- **VN** : les cas où la prédiction est négative, et où la valeur réelle est effectivement négative
- **FP** : les cas où la prédiction est positive, mais où la valeur réelle est négative.
- **FN** : les cas où la prédiction est négative, mais où la valeur réelle est positive.

Les mesures d'évaluations qu'on peut tirer de la matrice de confusion sont :

Nom	Formule	Brève explication
<b>Sensitivité</b>	$VP / (VP + FN)$	Pour la classification binaire
<b>Spécificité</b>	$VN / (VN + FP)$	Pour la classification binaire
<b>Accuracy</b>	$(VP + VN) / (VP + VN + FP + FN)$	Le degré de précision
<b>Précision</b>	$VP / (VP + FP)$	Le degré dont lequel les mesures répétées donnent le même résultat
<b>Rappel</b>	$VP / (VP + FN)$	Pour la classification binaire
<b>F1 score</b>	$2VP / (2VP + FP + FN)$	La précision d'un modèle de classification binaire

**Tableau 4. 5: Les mesures d'évaluations**

En plus des mesures calculées à partir de la matrice de confusion, il y a aussi le **taux d'erreur**

#### 4.3.2.2. Fonction d'erreur

Un taux d'erreur est la métrique la plus simple de classification, il reflète le taux de perte des modèles utilisés, il est généré à partir d'une fonction d'erreur.

La fonction d'erreur de la classification multiclassées que nous considérons utiliser dans nos tests est la fonction entropie croisée catégorique (Categorical cross entropy ou Softmax loss)

$$\text{Loss} = - \sum_{i=1}^{\text{output size}} y_i \cdot \log \hat{y}_i$$

**Équation 3. 3 : Fonction d'erreur**

Où  $\hat{y}$  est la  $i$  ème classe prédite et  $y$  est sa vraie valeur correspondante de la variable cible.

### 4.3.3. Résultat des évaluations

Après avoir détaillé nos données et le prétraitement de chaque cas, l'étape qui suit est l'entraînement et test,

L'entraînement des données pour chaque algorithme sera fait en utilisant la bibliothèque Scikit-Learn, et qui offre un tas de fonctions et de possibilités d'application d'un modèle d'apprentissage automatique.

Durant cette phase, les différents paramètres et hyperparamètres seront présentés pour chaque algorithme.

#### 4.3.3.1. Paramètres et hyper-paramètres

Les hyper-paramètres sont tout paramètres de configuration pouvant être choisi par l'utilisateur et susceptible d'affecter les performances. Ils sont les variables qui régissent le processus d'entraînement lui-même, Par exemple, une partie de la configuration d'un réseau de neurones consiste à décider combien de couches de nœuds cachées seront utilisées entre la couche d'entrée et la couche de sortie, ce nombre-là est un des hyper-paramètres d'un réseau de neurones.

Il y a une différence entre les paramètres du modèle et les hyper-paramètres, les paramètres sont les variables que la technique de machine learning sélectionnée utilise pour s'adapter aux données, Par exemple, un réseau de neurones est composé de nœuds de traitement (neurones), Lorsque le réseau est entraîné, chaque nœud possède une valeur de pondération (Poids) qui indique au modèle l'impact de ce nœud sur la prédiction finale. Le nombre de nœuds, les valeurs des poids et les classes des prédictions finales sont des paramètres.

Le réglage des hyper-paramètres c'est d'exécuter plusieurs essais au cours d'un même entraînement, chaque essai correspond à une exécution complète d'entraînement.

#### 4.3.3.2. Algorithmes testés et hyper-paramètres

Le problème de ce projet est un problème supervisé de classification multi-classes, et les algorithmes testés pour sa résolution accompagnés par quelques-unes de leurs hyper-paramètres les plus importants sont les suivants :

Algorithme	Hyper-paramètres	Valeurs
<b>Logistique régression</b>	Solver	L'algorithme à utiliser pour l'optimisation
	Max_iter	Maximum nombre d'itérations pour que le Solver converge vers l'optimum
<b>SVM</b>	Kernel	Pour Spécifier le type du Kernel utilisé dans l'algorithme (linéaire, polynomial, ...)
	gamma	Le degré de la fonction du kernel polynomial
	C	Le paramètre de régularisation
<b>Classificateur Naïve Bayes</b>	/	/
<b>Réseau de neurones artificiels</b>	activation	Désigne la fonction d'activation, qui est généralement Softmax pour les classifications multiclassées
	dropout	Un mécanisme avec lequel pour chaque itération d'entraînement, un sous-groupe de neurones ne sera pas pris en charge dans le calcul
	Hidden layers	Nombre de couches cachées
	Cost function	Fonction d'erreur désignée
	epochs	Nombre d'itération d'entraînement
	Optimizer	Fonction d'optimisation de l'erreur
	Learning rate	Le taux d'apprentissage, c'est un paramètre de la fonction de d'optimisation qui détermine la taille du pas à chaque itération

Tableau 4. 6: Algorithmes testés en Hyper-paramètres

### 4.3.3.3. Évaluation de la logistique régression

Solver	Max_iter	Précision
SAG	100	71,1%
	1000	79%
	10000	85.3%
SAGA	100	69.4%
	1000	78.1%
	10000	83.7%
LBFGS	100	85.6%
	1000	86.1%
	10000	86.6%
NEWTON_CG	100	87.5%
	1000	87.6%
	10000	87.6%

Tableau 4. 7: Évaluation de la logistique régression

### 3.3.4. Évaluation de SVM

Kernel	Gamma	C	Degré	Précision
RBF	$1^e - 3$	1	/	76.9%
		10		77.1%
		100		77.1%
		1000		77.1%
	$1^e - 4$	1		83.4%
		10		84.5%
		100		84.5%
		1000		84.5%
Polynomial	/	/	1	83%
			2	75%
			3	70.6%
			4	66.6%
			5	60%

Tableau 4. 8: Évaluation de SVM

### 3.3.5. Évaluation du Classificateur Naïve bayes :

Pour un modèle créé avec Sklearn de l'algorithme classificateur naïve bayes, aucun hyper-paramètres n'a été spécifié pour notre cas, donc une seule tentative d'entraînement pour cet algorithme et la précision est la suivante : **70.1%**

### 3.3.6. Evaluation d'un réseau de neurones artificiels

Pour la réalisation d'un modèle de réseau de neurones artificiels, la bibliothèque Keras a été utilisée, grâce à ses fonctions et utilités, elle a facilité l'implémentation et le traitement du réseau. Avant le test on a fixé les valeurs suivantes :



Activation	Nombre de couches cachées	Fonction de perte	Dropout
'relu' pour les couches cachées 'Softmax' pour la couche de sortie	2	Cross entropy catégoriques	0.2

Tableau 4. 9: Évaluation d'un réseau de neurones

Au lieu de l'optimisateur **SGD**, **Adam** a aussi été fixé, à partir des études, c'est l'algorithme le plus rapide et utilisé dans l'optimisation des réseaux de neurones, il combine les avantages de 2 extensions de SGD et la moyenne quadratique.

Et puis le reste des variantes ont été évaluées de la façon suivante :

Epochs	Learning rate	Précision	Taux d'erreur
20	0.01	88.2%	0.305
	0.02	79.2%	0.427
	0.03	75.4%	0.595
	0.1	44.6%	1.01
30	0.01	86%	0.512
	0.02	75.4%	0.829
	0.03	58%	0.886

Tableau 4. 10: Le reste des évaluations

### 3.4. Discussion des résultats

D'après les évaluations des tests de chaque modèle, le meilleur résultat obtenu est :

Optimizer	Learning rate	Epoch	Précision	Taux d'erreur
Adam	0.01	20	88.2%	0.305

Tableau 4. 11: Meilleur résultat obtenu

Les graphes suivants représentent deux le graphe de deux métriques d'évaluations qui sont la précision et le taux d'erreur

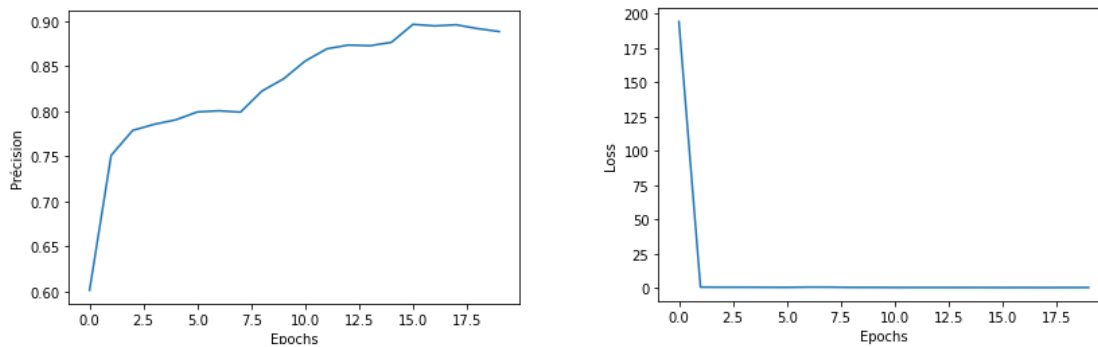


Figure 4. 36: Graphe de deux métriques d'évaluations – taux d'erreurs

On a abouti à un meilleur algorithme en comparant les deux métriques d'évaluations des uns aux autres et de chaque valeur des paramètres et hyper paramètres, ce qui nous a conduit à l'intégration de ce modèle dans le système de mesure de satisfaction client.

#### 4.4. Conclusion

Dans ce chapitre nous avons décrit notre environnement de travail, les outils utilisés ainsi que le prétraitement des données pour améliorer les performances du modèle ensuite, nous avons présenté le fonctionnement de l'application web destiné pour Djezzy et qui montre des statistiques et des visualisations sur la satisfaction clientèle par rapport à des critères d'affichage différents, et enfin nous évalué les algorithmes utilisés pour pouvoir aboutir à un meilleur modèle.

# Conclusion Générale

Djezzy est une entreprise pionnière dans le secteur de télécommunication. Durant ces dernières années, Djezzy a connu plusieurs changements pour devenir l'entité qu'elle représente aujourd'hui. Avec l'arrivée des nouvelles technologies dont l'apprentissage automatique, l'apprentissage profond et les BigData, Djezzy cherche à maintenir sa position parmi les leaders du domaine en exploitant à la meilleure façon possible ces outils.

L'apprentissage automatique ne cesse d'évoluer, cette sous branche de l'intelligence artificielle qui est utilisée dans beaucoup de domaines notamment la sécurité, le marketing digitale, le domaine médical, mais aussi les télécommunications, est très utile pour notre cas d'études qui est la satisfaction clientèle pour un leader des télécoms en Algérie.

Nous nous sommes intéressés dans le cadre de ce projet de fin d'études de master à proposer et développer une solution innovante et pratique permettant de mesurer le taux de satisfaction pour un échantillon d'abonnés d'une façon plus dynamique. Afin de résoudre le problème de la solution statique disponible actuellement, nous avons proposé un système basé sur les techniques de machine learning et d'une approche supervisée en utilisant des algorithmes de classification. En prenant en considération toutes les problématiques secondaires qui ont fait face à la réalisation de ce projet, nous avons pu proposer un concept basé sur un algorithme de classification qui aboutit à des prédictions de classes NPS pour chaque abonné de l'échantillon. Ces classes vont être utilisées pour le calcul du score de satisfaction globale. Le modèle de classification qui a été choisi est un réseau de neurones implémenté avec la bibliothèque Keras ; il a été choisi parmi d'autres algorithmes en se basant certainement sur des évaluations et des tests fait sur quelques algorithmes de classification. Nous notons que, le modèle implémenté a été intégré dans une application qui représente un Dashboard et qui visualise des statistiques et des graphes sur la satisfaction des échantillons testés. Il permet aussi le test de nouveaux cas que l'entreprise n'a aucune idée sur leurs classes NPS et leur niveau de satisfaction.

## Perspectives

Quel que soit l'ampleur de ce projet, comme tout autre projet, on a toujours plusieurs axes d'amélioration ou plusieurs aspects qu'on aurait souhaité utiliser, ainsi nos perspectives pour ce projet sont classées comme suit :

### Perspectives à court termes

Les aspects à court terme qu'on peut améliorer c'est de perfectionner le taux de précision de la mesure de satisfaction client, avec des algorithmes d'apprentissage profond. Cependant ceci est relié avec la nécessité de procurer plus de données relatives aux abonnés Djezzy y compris la donnée cible avec laquelle le résultat de satisfaction globale est directement dépendant.

### Perspectives à moyen termes

Les données Djezzy sont intégrées dans un entrepôt de données, incluant des quotas de données énormes. En ayant accès à cet entrepôt ainsi qu'aux serveurs de l'entreprise, une architecture Big Data pourrait être très utile. Ainsi, un software Apache NiFi, un logiciel de gestion et d'automatisation des flux de données entre plusieurs systèmes, nous permettra de transmettre les données de l'entrepôt vers Apache Spark. Ce dernier facilitera plusieurs tâches mais surtout l'utilisation de l'apprentissage automatique avec la bibliothèque MLlib et le langage PySpark (l'API de python pour Spark).

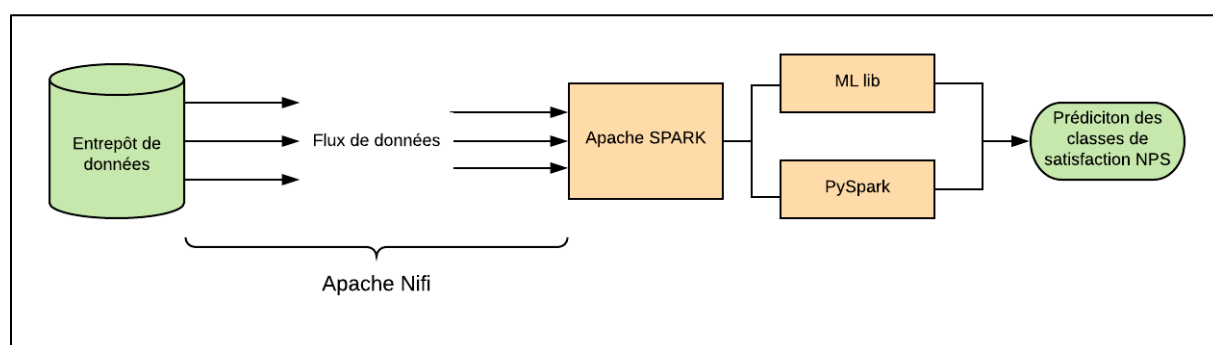


Figure 4. 37: Perspectives à moyen terme

### Perspectives à long termes

Le système qu'on a implémenté contient toujours des failles, il dépend toujours des réponses des abonnés car si on suppose que dorénavant les abonnés Djezzy ne répondront jamais au sondage de satisfaction, notre système sera mis en échec car il n'y aura pas de nouvelles données que le modèle d'apprentissage peut s'entraîner dessus, et avec chaque nouvel échantillon de test les prédictions seront de moins en moins précises.

Sachant que Djezzy dispose d'une application mobile, il serait plus intéressant que le sondage soit intégré dans l'application. Ainsi les notifications du sondage envoyé par sms se transformeront en notification push de l'application ce qui serait déjà plus agréable pour l'utilisateur, mais aussi pour le système de mesure de satisfaction de la globalité des clients.

# Bibliographie

- [1] A PROPOS DE DJEZZY. (s. d.). Djezzy. Consulté 2 février 2020, à l'adresse <http://www.djezzy.dz/djezzy/nous-connaître/a-propos-de-djezzy/>
- [2] Net Promoter Score - calcul et application. Consulté 29 juin 2020. CheckMarket. <https://fr.checkmarket.com/blog/votre-net-promoter-score/>
- [3] Adapté de l'ouvrage suivant : Daniel Ray, op. cit. , p. 34-35
- [4] Définition : Apprentissage automatique. Consulté 20 juin 2020. Psychomédia. <http://www.psychomedia.qc.ca/lexique/definition/apprentissage-automatique>
- [5] Ismaili, Z. (2019, 12 novembre). Apprentissage Supervisé Vs. Non Supervisé. Consulté 20 juin 2020. Le DataScientist. <https://le-datascientist.fr/apprentissage-supervise-vs-non-supervise>
- [6] Issarane, H. (2019, 20 février). La Régression Linéaire. Consulté 20 juin 2020. Le DataScientist. <https://le-datascientist.fr/top-5-des-types-de-regression>
- [7] Issarane, H. (2019b, mars 2). Les SVM, Support Vector Machine. Le DataScientist. Consulté 20 juin 2020. <https://le-datascientist.fr/les-svm-support-vector-machine>
- [8] Benzaki, Y. (2017, 22 août). 9 Algorithmes de Machine Learning que chaque Data Scientist doit connaître. Mr. Mint : Apprendre le Machine Learning de A à Z. Consulté 20 juin 2020. <https://mrmint.fr/9-algorithmes-de-machine-learning-que-chaque-data-scientist-doit-connaître>
- [9] Clustering. (2018, 13 décembre). Data Analytics Post. Consulté 20 juin 2020. <https://dataanalyticspost.com/Lexique/clustering>
- [10] K-means (ou K-moyennes). (2018, 5 décembre). Data Analytics Post. <https://dataanalyticspost.com/Lexique/k-means-ou-k-moyennes>
- [11] Réseau de neurones artificiels : qu'est-ce que c'est et à quoi ça sert ? .(2019, 5 avril). Le Big Data. <https://www.lebigdata.fr/reseau-de-neurones-artificiels-definition>
- [12] Kalipe, G. K. (2020, 3 mars). Introduction au Deep Learning et aux réseaux de neurones pour les nuls. Medium. <https://medium.com/@godsonkkalipe/introduction-au-deep-learning-et-aux-r%C3%A9seaux-de-neurones-pour-les-nuls-de6f8a7b7a35>
- [13] K. Dahiya and S. Bhatia, "Customer churn analysis in telecom industry," 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Noida, 2015, pp. 1-6, doi: 10.1109/ICRITO.2015.7359318.
- [14] Zhao, Q., Chen, K., Li, T. et al. Detecting telecommunication fraud by understanding the contents of a call. Cybersecur 1, 8 (2018). <https://doi.org/10.1186/s42400-018-0008-5>
- [15] Medium. (s. d.). Medium Get smarter about what matters to you. Consulté 20 août 2020, à l'adresse <https://medium.com/>
- [16] Débuter avec Keras - Documentation en français. (2018, 14 mai). Actu IA. <https://www.actuia.com/keras/>
- [17] Patterson, J., Gibson, A. (2017). Deep Learning: A Practitioner's Approach. Beijing: O'Reilly.

