République Algérienne Démocratique et Populaire
الجمهورية الجزائرية الديمقراطية الشعبية
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
وزارة التعليم العالي و البحث العلمي

جامعة امحمد بوقرة بومرداس
كلية العلوم
Université de M'hamed Bougara
Faculté des Sciences

Mémoire de fin d'études

Pour l'obtention du diplôme Master 2 en Recherche Opérationnelle

Option :Recherche Opérationnelle, Optimisation et Management Stratégique

Application technique de détection d'anomalies dans les flux de données du réseau LTE(cas de Algérie Télécom, Boumerdes)

Réalisé par :
Mlle.Hadid Fahima
Mlle.Louaidi Asma Yassinia

Encadré par :
M.R.Bellouti
(Algérie Télécom)

Soutenu le 28 Septembre 2022, Devant le jury composé de :

Mme.B.Ferdjallah :	UMBB	- Présidente-Classe(M.A.A)
M.S.Taharbouchet :	UMBB	- Examinateur-Classe(M.C.B)
Mme.W.Drici :	UMBB	- Promotrice-Classe(M.C.B)
Mme.M.Benmansour :	UMBB	- CO-Promotrice-Classe(M.A.A)

Promotion : 2021/2022

Remerciements

Nous remercions tout d'abord, **Allah** qui nous a donné la force et le courage de parvenir à élaborer ce mémoire. Sa réalisation n'aurait pas été possible sans sa bénédiction.

Nous tenons à exprimer toutes nos gratitudes à Madame **W.Drici**, d'avoir accepté de nous encadrer dans notre travail. Nous tenons à lui exprimer nos plus vifs remerciements pour sa disponibilité, son expérience, et son implication, et ses encouragements.

Nous adressons nos remerciements à notre Encadreur de l'entreprise Algérie Télécom Monsieur**Bellouti Rabeh** pour son aide et ses conseils judicieux.

Nous remercions les membres du jury de l'UMBB de Boumerdes,
Monsieur **Taharbouchet**, Madame **Ferdjallah**, et Madame **Benmansour** d'avoir l'amabilité d'accepter d'examiner notre travail.

Nous tenons à remercier nos familles , nos frères et soeurs de nous avoir soutenu. Et un spécial remerciement à mon chère Fiancé (Fahima). Nous ne serons jamais assez reconnaissants envers nos parents qui ont toujours tout mis en œuvre pour qu'on s'épanouisse dans tout ce qu'on entreprend.
Nous remercions également tous ceux qui ont contribué, de près ou de loin à la réalisation de ce modeste travail.

Dédicace

Je dédie ce modeste travail à :

Mes très chers parents sans eux rien n'aurait été possible. J'oublierai jamais, leurs amours, leurs dévouements, tous leurs sacrifices. On espère pouvoir un jour les remercier à la hauteur de ce qu'ils nous ont apportés.

A mes frères, mes sources de motivation ,

Yacine, Hakim, Nabil, Imad, Mohamed, Slimane ; en gage de ma profonde estime pour l'aide que vous m'avez apporté, soutenu, réconforté et encouragé.

Mes chères sœurs,

Razika, l'adorable grande soeur, malgré la distance qui nous sépare tu as toujours su comment me réconforté et reboosté. Hâte pour ton retour parmi nous. Lynda, ma source de joie, qui n'as cessé de me conseiller, encourager, soutenir durant toutes ces années d'études.

A mon cher fiancé Imad,

pour l'amour et l'affection qui nous unissent. Je ne saurais exprimer ma profonde reconnaissance pour le soutien continu dont il a toujours fait preuve. Son soutien m'a permis de réaliser le rêve tant attendu. Je prie Allah le tout puissant de préserver notre attachement mutuel, et d'exaucer tous nos rêves.

A ma belle famille, ma deuxième famille qui m'a soutenu et encouragé.

A mes nièces et neveux adorés Anais, Hind, Arwa, Adam , Hatem, Haitham, Yasser, Wassim, Houssem, Ayoub, Ayan, Issam, Rassim.

A mes adorables belles soeurs. Zineb et Meriem.

A mon beau frère Farid.

A mes chères amies Nour Elhouda, Melissa et Hanane, elles sont pour moi des sœurs du coeur sur qui je peux compter. En témoignage de l'amitié qui nous unit.

A Asma, binôme et sœur, partenaire des meilleurs moments. Je te remercie pour cette belle expérience vécu ensemble pendant tout ce parcours, nous avons partagé nos moments de joies et de tristesses. Merci pour ton soutien moral, ta patience et ta compréhension et surtout ta confiance. Je suis fière de toi et de nous.

"Fahima"

Dédicace

Je dédie ce modeste travail à :

Mes très chers parents sans eux rien n'aurait été possible.

Jamais on n'oubliera, leurs amours, leurs dévouements, tous leurs sacrifices. On espère pouvoir un jour les remercier à la hauteur de ce qu'ils nous ont apportés.

A mes grands frères,

Zohir et Amirouche, mes sources de motivation et d'inspiration. Ils m'ont toujours épaulé et soutenu. je les remercie pour leurs amour et confiance. Je ne saurais jamais exprimer ce que je ressens à leurs égards.

A mes petits frères,

Ghiles, Yanis et Amine, leurs présence dans ma vie est indispensable. Je n'imagine pas vivre sans eux.

A mes Belles soeurs,

Fariza et Sabrina, elles partagent la vie de mes frères, celle de notre famille, et la mienne par la même occasion. je les remercies pour les personnes exceptionnelles qu'elles sont.

A ma petite nièce adoré,

Lina Malak, mon rayon de soleil qui illumine ma vie. Elle est venue dans ce monde et apporter beaucoup de bonheur à notre famille.

A mes chères amies Nour Elhouda et Melissa, elles sont pour moi des sœurs du coeur sur qui je peux compter. En témoignage de l'amitié qui nous unit.

A Fahima, binôme et sœur, partenaire des meilleurs moments. Je te remercie pour cette belle expérience vécu ensemble pendant tout ce parcours, nous avons partagé nos moments de joies et de tristesses. Merci pour ton soutien moral, ta patience et ta compréhension et surtout ta confiance. Je suis fière de toi et de nous.

"Asma Yassinia"

Table des matières

Table des figures

Liste des tableaux

Liste des acronymes

AT	Algérie Télécom
DRT	Direction Régionales des Télécommunications
DO	Direction Opérationnelle
ADSL	Asymetric Digital Subscriber Line
DOT	Direction Opérationnelles des Télécommunications
ACTEL	Agences Commerciales des Télécommunications
PDG	Directeur Général Principal
D.IB	Direction des Institutions de Base
D.AE	Direction d'Aide a l'Emploi
D.CMI	Direction Commerciale Marketing Innovation
UMTS	Universal Mobile Telecommunication system
BTS	Base Transiver Station
3GPP	3erd Generation Partnership Project
IP	Internet Protocol
RT	Réseaux de Télécomunication
DA	Détection des Anomalies
LTE	Long Terme Evolution
GSM	Global System for Mobile
GPRS	General Packet Radio Service
EDGE	Enhanced Data Rates for GSM Evolution

FDD	Frequency Division DuplexTDDTime Division Duplex
QoS	Quality of Sercice
EPS	Evolved Packet System
QAM	Quadrature Amplitude Modulation
QPSK	Quadrature Phase-Shift Keying
SGW	Serving Gateway
PDN-GW	Pcket Data Network Gateway
PGW	Packet Gateway
GBR	Guaranteed Bit Rate
OFDMA	Orthogonal Frequency Division Multiple Access
SCFDMA	Signale-Carrier Frequency Division Multiple Access
E-UTRAN	Evolved Universal Terrestrial Radio Access Network
UE	User Equipement
eNodeB	Evolved Node B
MME	Mobility Management Entity
EPC	Evolved Packet Core
RRC	Radio Resource Control
KPI	key performance indicator
PRB	Physical Resource Block
RB	Resource BlockRSReference Signal
CPU	Central Processing Units
LBP	Lead Based Pain
PUCCH	Physical Uplink Control Channel
PDCCH	Physical Downlink Control Channel
PRACH	Physical Random Access Channel
SRS	Ressources Signal Reference Sounding

RE	Resource Element
RS	Reference Signais
RF	Fréquence Radio
OSI	Open Systems Interconnetion
PDSCH	Physical Downlink Shared Channel
RSRQ	Reference Signais Received Quality
RSRP	Reference Signais Received power
RSSI	Reference Signais strength Incdicator
SINR	Signal-to-Noise-plus-Interference Ratio
CQI	Channel Quality Indicator
QCI	Quality of Service Class Identifier
EPS	Evolved Packet System
MIMO	Multiple Input Multiple Output
BS	Base Station
PMI	Precoding Matrix Indicator
RI	Rank Indicator
AD	Anomalie Detection
IA	Intelligence Artificielle
ML	Machine Learning
IoT	Internet of Thing
KDD	Knowldge Discovery in Databases
DM	Data Mining
KNN	K-Nearest Neighbors
SVM	Support Vector Machine
DBSCAN	density-based spatial clustering of applications with noise
OPTICS	ordering points toc identify the clustering structure

Résumé

On s'intéresse dans la partie eUTRAN de réseau LTE aux ressources cellulaires, en particulier PRB le plus petit élément de données attribué par la station de base, constitué de 84 RE. Le RE est utilisé pour transmettre des données, des signalisations ou signaux physiques. Ces derniers sont mis en forme à l'émetteur afin de se propager dans de bonnes conditions (RF condition) et le traitement à mettre en œuvre au récepteur afin de les détecter correctement. Cette conception est déterminée par le canal radio, dont sa qualité peut être mesuré en liaison descendante par trois indicateurs (CQI , RI, PMI). Reconnaître les valeurs qui sont problématique parmi toute nos données, est une tâche complexe, qui est l'objet de la détection d'anomalies basé sur la surveillance des trois indicateurs précédents notamment PRBs. De nombreux facteurs influencent sur la qualité de service (débit servis aux abonnées) et chacun des facteurs aurait une sorte de corrélation avec d'autres(corrélation entre CQI et débit, utilisation de PRB avec un max nombres d'utilisateurs). CQI doit représenté l'efficacité spectrale basée sur la sélection de paramètre de modulation et de codage. Il est difficile de maintenir manuellement le système LTE. En l'automatisant, Nous trouvons les anomalies dans les données générées. L'idée principale est de démontrer et d'évaluer une approche d'apprentissage automatique pour la détection d'anomalies dans le domaine des télécommunications. Notre objectif est d'attraper automatiquement un modèle anormal, et prédire si l'abonnée a bénéficié d'une bonne QoS. Les données utilisées dans l'expérience contiennent le comportement d'une cellule à travers le temps, recueillies à partir un fournisseur de réseau cellulaire basé à Boumerdès et collecté sur un intervalle de 15jours. Nous utilisons trois algorithmes (Naïve Bayes, SVM et KNN) et python comme langage de programmation. La prédiction a été effectué à l'aide des techniques d'apprentissage automatique, à base des algorithmes de cas supervisé pour le problème de classification. Les résultats obtenus, suite à l'étape de modélisation ont démontré que le modèle SVM est le plut fiable par rapport aux autres.

Introduction Générale

Au cours des années 2000, il y a eu une énorme progression dans les communications sans fil mobiles. Commençant avec la technologie 1G qui, en très peu de temps, a été remplacée par la 2G, la 3G, la 4G. Cette progression est due à la nécessité d'une innovation parfaite en matière de transmission et à une forte augmentation du nombre de clients télécoms pour des usages quotidiens, les entreprises, le secteur de l'éducation et presque tous les autres secteurs.

Ce mémoire se focalise sur l'interface radio LTE qui se repose essentiellement sur le mode paquet, et la notion de bloc ressource. Ce système utilise le concept de canal afin d'identifier les types des données transportées sur l'interface radio, les caractéristiques de qualité de service associées, ainsi que les paramètres physiques liés à la transmission. Le contrôle de la qualité de service est essentiel pour l'opérateur afin de garantir une expérience satisfaisante à l'utilisateur. Un UE actif aura besoin de bonnes conditions radio pour une qualité de service satisfaisante et une bonne continuité de service. De plus en plus de données sont générées. Notamment, le nombre de station de bases est énorme et il y a plus de problèmes pour gérer cela. Il est très difficile de maintenir un tel système avec efforts manuels, voir même presque impossible. La seule chose que nous voulons automatiser dans le système est de trouver les anomalies dans les données générées. Nous procédons à l'apprentissage automatique afin de faire la détection d'anomalies dans KPI d'évolution à long terme. Un comportement anormal d'un KPI indique la dégradation du réseau, qui fournira la mauvaise expérience utilisateur. Notre objectif est d'attraper automatiquement un modèle anormal, le comportement est dit anormal si l'abonné ne bénéficie pas d'une bonne qualité de service. Cela aidera pour étudier plus avant la cause de la dégradation des KPI. L'idée principale est de démontrer et d'évaluer une approche d'apprentissage automatique pour la détection d'anomalies dans le domaine des télécommunications spécifique à l'évaluation à long terme (LTE) KPI. Les données utilisées dans l'expérience contiennent le comportement d'une cellule à travers le temps, recueillies à partir un fournisseur de réseau cellulaire basé à Boumerdes et collecté sur un intervalle de 15jours. Nous utilisons trois algorithmes et python comme langage de programmation. Nos résultats devront montrer que l'apprentissage automatique peut être appliqué avec succès dans les réseaux LTE pour la recherche de détection d'anomalies. L'apprentissage automatique peut réduire le temps nécessaire aux experts du domaine pour identifier les anomalies au sein du réseau. De plus, il est également utile de trouver l'analyse des causes profondes de la dégradation des KPI.

Notre mémoire est composé de quatre chapitres :

Dans le premier chapitre, nous allons présenter l'entreprise dans laquelle s'est effectué notre stage. On va définir l'organisme d'accueil Algérie Télécom et en particulier la direction opérationnelle à Boumerdes.

Le deuxième chapitre, nous allons s'intéresser aux notions de réseaux de communications, particulièrement le LTE. De même à la détection des anomalies et ses différentes techniques.

Le troisième chapitre sera consacré au Machine Learning, nous allons présenter les concepts fondamentaux de la Data Science et son processus, du processus KDD, et nous terminons avec ses domaines d'application, les types d'algorithmes appliqués et les techniques de validation.

Dans le dernier chapitre, nous allons présenter la phase finale de notre projet qui est l'application. Cette phase concerne la construction des étapes de notre modèle.

Nous finalisons notre travail par une conclusion générale dans laquelle nous exposons les résultats obtenus.

Mots-clés :
LTE : Long Terme Evolution.
KPI : Key Performance Indicator.
KDD : Knowldge Discovery in Databases.

Chapitre 1

Présentation de l'entreprise Algérie Télécom

1.1 Introduction

Dans ce premier chapitre, nous allons présenter l'entreprise dans laquelle s'est effectué notre stage. Nous allons définir l'organisme d'accueil Algérie Télécom, en particulier la direction opérationnelle à Boumerdes.
Algérie Télécom détient la grande part de marché dans le secteur de télécommunications connaissant une forte croissance tout en offrant une gamme complète de services de voix et de données aux clients résidentiels et professionnels. Cette position s'est construite par une politique d'innovation forte adaptée aux attentes des clients et orientée vers les nouveaux usages. C'est une société par actions à capitaux publics opérant sur le marché des réseaux et services de communications électroniques. Crée le 1er janvier 2003 d'une séparation des activités postales et télécommunications des anciens services de PTT. En 2003, AT comptait près de 130 000 abonnés GSM et 1,9 million de clients sur le réseau fixe. Depuis, elle s'engage dans le monde des Technologies de l'information et de la Communication. Elle a trois objectifs : rentabilité ;efficacité ;qualité de service qui sera la partie dont se portera notre cas d'étude. Son ambition est d'avoir un niveau élevé de performance technique, économique, et sociale pour se maintenir durablement leader dans son domaine, dans un environnement devenu concurrentiel. Son souci consiste, aussi, à préserver développer sa dimension internationale et participer à la promotion de la société de l'information en Algérie [48].

1.2 L'oganisation de AT

Algérie télécom est répartie le territoire national en 13 directions régionales des télécommunications(DRT), 50 Directions opérationnelles des Télécommunications (DOT) et 174 agences commerciales des télécommunications(ACTEL). À cette structure s'ajoutent trois filiales :

1.2.1 Algérie Télécom Mobile (Mobilis)

Le premier opérateur mobile en Algérie, devenu autonome en août 2003 [37]. Elle a lancé son premier réseau expérimental (UMTS) en Algérie[53].

Spécialisé dans le domaine de la téléphonie mobile. Et dispose aujourd'hui :

- De plus de 4200 Stations de Base Radio (BTS).
- De plateformes de Service des plus performantes.

Et compte :

- Plus de 7 millions d'abonnés.
- Un réseau commercial en progression dépassant les 116 Agences.
- 52 500 points de vente indirects.

1.2.2 Algérie Télécom Internet (Djaweb)

Idoom, anciennement Easy ADSL puis Djaweb. C'est un fournisseur d'accès internet algérien. Djaweb xDSL est né de la fusion de trois fournisseurs d'accès à internet filiales d'Algérie Télécom : easy adsl ; Fawri et Anis [52].

Spécialisée dans le domaine d'accès à Internet, dispose aujourd'hui de trois types d'accès :

- Accès bas débit via RTC.
- Accès direct 1515.
- Accès via cartes prépaid 1533.
- Accès haut débit par liaisons spécialisées. - Via une plateforme de 48 POP's, à raison d'un Pop par Wilaya (débit LS de 128 Kbps à 2 Mbps).
- Via une plateforme backbone international (LS de 2 Mbps à 1 Gbps)
- Accès haut débit xDSL.
- Via trois plateformes, dont deux sont mises en ouvre en partenariat avec des équipementiers étrangers.

Le 30 mars 2014 Algérie Télécom lance sa nouvelle gamme d'offres internet, baptisée « Idoom ADSL », avec des débits allant de 1 à 8 Mbit/s puis le 25 avril 2016 propose un débit allant jusqu'à 20 Mbit/s[49].

1.2.3 Algérie Télécom Satellite (RevSat)

Ayant pour principale mission de développer et de promouvoir les télécommunications par satellite. Ce qui constitue l'un des axes les plus importants de la stratégie globale du développement d'Algérie télécom[**?**].

L'organisation d'Algérie Télécom satellite comprend une direction générale autour de laquelle s'articulent 5 Directions Régionales (ALGER, ORAN, OUARGLA, BECHAR et CONSTANTINE) et de deux Antennes (SETIF et ANNABA), ainsi qu'un téléport à LAKHDARIA . spécialisée dans le domaine des solutions satellitaires, dispose aujourd'hui de :

- 2500 stations VSAT déployées
- Près de 1500 abonnés THURAYA

- Une présence nationale optimale sur 48 wilaya.

1.3 Domaines d'activités d'AT

L'entreprise AT est l'acteur majeur des télécommunications en Algérie avec cinq domaines d'activités :

- **Téléphonie fixe** : avec deux millions de lignes en service et un réseau WLL en plein expansion.
- **Téléphonie mobile** :activité au travers d'une filiale Mobilis, qui détient une part de marché de 13/100.
- **Transmission de données** : une activité de réseaux de données pour les entreprises (X25...) .
- **Accès Internet à travers** : DJAWEB, FAWRI ADSL,EASY ADSL et dernièrement IDOOM ADSL .
- **Réseau satellitaire** : des services de télécommunications s'appuyant sur VSAT, Inmarsat le réseau Thuraya.

1.4 Les structures organisationnelles de la direction d'Algerie télécom

L'organigramme de la direction générale d'Algérie télécom est composé de l'inspecteur général ;de la direction de la sécurité intérieure ;du bureau du directeur général ; de la direction de l'audit ; de la direction stratégique ;de la direction de l'appui à l'emploi ;de 52 directions opérationnelles implantées dans tous les états du pays, et agences commerciales réparties entre les départements qui lui sont affiliés à la direction opérationnelle de chaque état. Nous donnerons une brève explication de la structure organisationnelle :

1.4.1 Le Directeur Général Principal (PDG)

Il est le président du conseil d'administration, et il est le premier responsable de l'entreprise existante, il prend en charge les responsabilités avec ses assistants, entreprend la tâche d'atteindre les objectifs fixés par les départements et c'est son désir d'assurer ce qui suit :

- Maintenir les parts de marché et développer la culture d'entreprise sur le marché concurrentiel.
- Assurer la mise en œuvre des programmes approuvés et la coordination entre les départements.
- Surveiller la conduite des diverses activités de l'entreprise grâce aux rapports reçus par les divers intérêts.
- Tenir compte des suggestions soumises par les intérêts, ainsi que maintenir la bonne et normale conduite de l'entreprise.

1.4.2 La Direction de Sécurité Intérieure

Parmi ses attributions figurent :

- Il assure la sécurité intérieure de la Direction Générale.

- Considéré comme un conseiller en sécurité interne pour l'ensemble de l'entreprise.
- Supervise le travail des chefs des directions pratiques de 58 états.

1.4.3 Le Bureau du Directeur Général Exécutif

Il s'occupe de :

- Organiser le travail effectué au niveau du bureau du directeur général.
- Organiser les jours et heures des réunions et de l'accueil du Directeur Général.
- Lire les dépêches adressées au Directeur Général.

1.4.4 La Direction d'Inspection Générale

Elle est en charge de la base d'activité annuelle.

- Mise en œuvre de tâches d'inspection surprise à la demande du directeur général personnellement.
- Mener des enquêtes en cas d'atteinte à l'entreprise.
- Coordination du suivi de la surveillance des intérêts de l'inspection.

1.4.5 La Direction d'Enquête

Dans la hiérarchie administrative, le service d'enquête est directement subordonné au directeur général de l'entreprise et reçoit les ordres Directement du directeur général et ses fonctions sont :

- Procéder à une enquête sur les postes de travail et les conditions qu'ils requièrent afin que l'employé puisse exercer au mieux ses fonctions.
- Mener une enquête sur les modalités de nomination des salariés, afin que les rendements de chacun d'eux soient appropriés et suffisants pour atteindre les objectifs souhaités de l'entreprise.
- Présenter les décisions avec les solutions à suivre au directeur général.

1.4.6 La Direction Stratégique

C'est la direction directement subordonnée au directeur général de l'entreprise, parmi ses missions : Dessiner les grandes lignes de la stratégie adoptée par l'entreprise à court et à long terme.

1.4.7 La Direction de Sécurité et de Protection

Il est également subordonné au Directeur Général, et il exécute toutes les réglementations qui doivent être suivies au sein de la Direction Générale ou des directions opérationnelles, et de tous les biens de l'entreprise, afin d'assurer et de consacrer la sécurité et protection.

1.4.8 Les Directions Opérationnelles

Parmi leurs attributions figurent :

La mise en œuvre des programmes de la Direction Générale est basée sur le niveau de compétence de chacun. Il soumet des propositions ou des projets à l'approbation de la Direction générale.

- Maintenir et développer le réseau téléphonique et internet au niveau de sa juridiction.
- Soumettre tous les problèmes et lacunes qu'il peut rencontrer dans l'exercice de ses fonctions à la Direction générale afin de trouver des solutions pour cela.
- Manque de ressources humaines et de capacités financières ou matérielles.

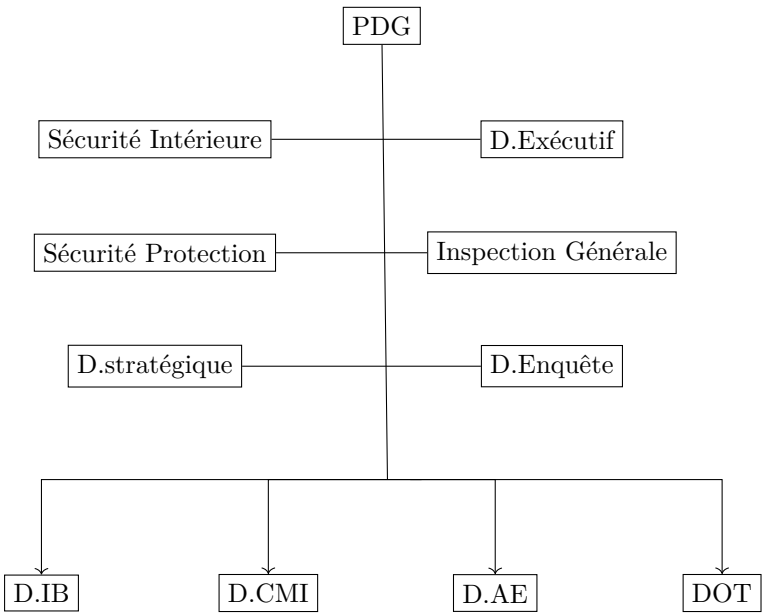
1.4.9 La Direction des Institutions de Base

Parmi ses missions figurent :

- Assure l'étude et le développement du réseau de communication pour les nouveaux projets et son expansion.
- Assure la préparation des lieux pour le placement des équipements de communication (architecture - hygiène - climatisation - protection - prévention).
- Déploiement et maintenance de l'infrastructure de base.

et nous terminons avec **la Direction d'Aide à l'Emploi** et **Direction Commerciale, Marketing et Innovation**

1.5 Organigramme Direction Générale AT



1.6 Fiche technique de la Direction Opérationnelle de la Télécommunication à Boumerdès

Dans ce sujet nous abordons la présentation de la DOT de Boumerdes , en ce qui concerne ses origines sa mission et ses objectifs , ainsi que sa structure organisationnelle.

Algérie Télécom	Entreprise
	Logo
cité 1200 logements -Boumerdes-	Adresse
+213(024)-79-15-15	Numéro de téléphone
+213(024)-79-16-16	Fax
02 B18083	Registre de commerce
00216290656936	Numéro d'identification fiscale
00021600180833716001	Numéro d'identification impôt
16293838021	Identification Statistique

Tab. 1.1: Informations relatives à la Direction des Télécommunications, Algérie, Boumerdès

1.6.1 La Mise en place direction opérationnelle à Boumerdes

Selon la décision de la Direction Générale n° 02/15 du 11 novembre 2002 portant organisation de la Direction Générale des télécommunications d'Alger, la Direction Opérationnelle des télécommunications a été crée à Boumerdès. Elle est actionnaire et a été dénommée début 2003 l'Unité opérationnelle des télécommunications jusqu'en juin 2010, date à laquelle le nom a été changé d'une unité opérationnelle à la direction opérationnelle. Cette direction est située dans le quartier des 1200 logements IBEN KHALDOUN de Boumerdes, son numéro d'immatriculation commerciale est le 18083 B 02, son numéro fiscal 002.16.290.656.936, et son numéro d'immatriculation fiscale 00021600180833716001.

Elle est subdivisée en six centres de maintenance et de production : Dellys, Borj Manaiel, Boudouaou, Khemis El Khechna, Thenia, Boumerdes, et en plus de 3 agences commerciales, un siège pour la maintenance du matériel et un centre pour les travaux publics, qui emploient environ 215 salariés, soit environ 01% du total des salariés d'Algérie Télécom.

1.6.2 Les missions et objectifs de la DOT

La Direction Opérationnelle des Télécommunications à Boumerdes exerce ses activités à travers l'utilisation de plusieurs moyens importants représentés dans certaines tâches ainsi que ses objectifs.

1.6.2.1 Tâches opérationnelles de la Direction

Les missions de la DOT de Boumerdès peuvent se résumer dans les points suivants :
1- Porter le taux de collecte des droits téléphoniques à plus de 80/100.
2- Aménagement et extension du réseau téléphonique dans l'État, augmentation du nombre d'abonnés au téléphone fixe et sans fil et augmentation du nombre d'abonnés ADSL.
3- La réparation des perturbations qui affectent les lignes d'abonnés, ainsi que le suivi quotidien du réseau de fibre optique qui s'étend sur tout l'État.
4- Fournir aux institutions publiques et aux entreprises divers services de télécommunications, tels que l'établissement de réseaux locaux (Intranet) et leur fournir des préparations servant à transmettre des données (réception et émission), telles que des lignes privées.
5- Fournir aux directions opérationnelles des statistiques hebdomadaires, mensuelles et annuelles, avec des données et informations relatives aux projets futurs.
6- Raccorder les régions isolées et les établissements scolaires au réseau.
7- Développer un nouveau système d'information qui permet de :
- Le client dispose de son propre compte au niveau de l'agence commerciale qui recueille sa demande et toutes ses informations et répond à cette demande.
- Fin des échanges de dossiers et papiers entre les services techniques de l'agence commerciale et suivi du nouveau système de gestion.
- Permettre aux clients de consulter leurs factures via Internet.
8- Offrir aux citoyens de nouveaux services, ces services facilitent et développent les conditions de vie du citoyen.
9- Soumettre des propositions à la Direction Générale en vue de l'acquisition, de l'entretien et de la location de nouveaux biens immobiliers, dans le cadre du développement d'un réseau de communication ou à d'autres fins (ouverture de nouvelles agences commerciales, ou centres d'entretien de nouvelles lignes téléphoniques).
10- Au début de chaque exercice, la DO procède à une étude préalable des dépenses (investissement et charges) qui pourraient être nécessaires au cours de l'exercice, afin de mener à bien et de consacrer les projets qui lui sont confiés.les objectifs de fonctionnement de la Direction.

1.6.2.2 Les objectifs opérationnels de la Direction

1- Augmenter l'offre de services téléphoniques et faciliter l'accès aux services de télécommunications au plus grand nombre d'utilisateurs, notamment en milieu rural.

2- Améliorer la qualité des services afin d'augmenter la compétitivité des services rendus. 3- Développer un réseau efficace connecté aux différents canaux de circulation de l'information. 4- Développer de nouveaux

services et gagner ainsi la confiance des clients. 5- Fournir des services d'assistance technique. 6- Mise en place de la convergence voix et données. 7- Améliorer la valeur des ventes.

1.7 La structure organisationnelle de la Direction Opérationnelle de la Communication à Boumerdès

La structure organisationnelle de la DO est la suivante :

1.7.1 Le Directeur Opérationnelle

Il est le principal responsable de la direction opérationnelle, où il réalise les objectifs fixés par les services concernés, et il est investi des pouvoirs de gestion de l'entreprise. Parmi ses fonctions :

- Coordination des intérêts.
- Suivi du déroulement des différentes activités de l'entreprise.
- Maintenir une bonne conduite dans l'entreprise.

1.7.2 Le chargé de la communication

Parmi ses missions figurent :

- Participer à la réalisation des opérations de communication, pour atteindre les objectifs du programme souligné.
- Il diffuse l'information dans les médias internes et externes.
- Participer à honorer et polir l'image de marque d'AT.

1.7.3 Service sûreté

Il est chargé de :

- Gestion et organisation de réunions.
- Enregistrer et sauvegarder le courrier entrant et sortant.
- Classement des dossiers, courrier entrant et sortant.

1.7.4 La sous direction des fonctions

C'est celle qui exerce les différents métiers liés à l'entreprise et est divisée en :

1.7.4.1 Direction des Affaires Juridiques

Ses missions consistent notamment à :

- Suivi des contrats.
- Suivi des affaires soulevées au niveau judiciaire.

- Dépôt de plaintes.
- Fournir des conseils juridiques à divers intérêts (commerciaux, ressources humaines, technologie...).

1.7.4.2 Service support de système d'information

- Installer et entretenir tous les appareils électroniques au niveau de la DO.

1.7.4.3 Département de patrimoine et moyens

Qui se divise en :

Le Département des médias publics Il fait :

- Se donner les moyens du bon fonctionnement de l'entreprise.
- Suivre l'entretien des différentes structures de l'entreprise, et suivre les nouveaux programmes en matière de construction.

Service Inventaire et Assurances :

Il s'occupe de :

- Faire le suivi des nouvelles propriétés, et faire le suivi des statistiques annuelles de l'entreprise.
- Comptage de toutes les propriétés de l'entreprise au niveau de l'état, et suivi des documents ou dossiers administratifs des propriétés.
- S'assurer que les biens de l'entreprise sont sécurisés et signaler les accidents quotidiens.
- Responsable des archives et de la documentation.
- Suivi des archives pour certains intérêts (commerciaux, techniques, ressources humaines....).

1.7.4.4 Département des Finances et de la Comptabilité

Ses missions consistent à :

- Gestion des comptes financiers.
- Approbation des opérations de trésorerie.
- Élaboration du schéma prévisionnel du Trésor.
- Assure la formation du personnel et le développement des services.
- Service Comptabilité : Parmi ses missions :

1- Assurer les registres et la comptabilité.

2- Compléter le budget annuel.

3- Maintient les dossiers et documents juridiques.

1.7.4.5 Département des Achats et des Services Logistiques

Qui se décompose en :

Service Achats : il effectue :

- Le suivi et entretien des transports, et l'acquisition des besoins de la direction en matière de fournitures de

bureau...etc.

• L'acquisition des besoins de la Direction en matière de fournitures de maintenance.

Prestations logistiques : Il s'agit de :

• Suivi des factures d'électricité, d'eau et de gaz pour l'entreprise.

Responsable d'entrepôt • Suivre l'entrée et la sortie des approvisionnements de l'entreprise au niveau de l'entrepôt.

1.7.5 La Sous-Direction Commercial

Elle est divisée en :

1.7.5.1 Département Générale de Ventes

Il s'occupe de :

• Assurer l'exactitude en fournissant les moyens nécessaires pour atteindre les objectifs des unités commerciales.

• Réalisation d'enquêtes pour analyser le comportement des clients par rapport à la force de vente dans les unités commerciales.

• Détermine les moyens nécessaires pour atteindre les objectifs commerciaux de ses filiales.

• C'est un support des unités commerciales pour améliorer la force de vente.

• Former et qualifier les employés du département.

1.7.5.2 Direction des Relations et des Grandes Institutions

Qui s'occupe de :

• Coordination avec les institutions externes (tant publiques que privées).

• Facturation de la valeur des services accordés aux établissements susmentionnés.

1.7.5.3 Département Support Commercial

Qui s'occupe de :

• Suivi de l'agence commerciale.

• Préparer un plan d'affaires qui définit les objectifs du processus de vente dans chaque unité d'affaires.

• Suivre la collecte annuelle.

• Soumettre les factures, collecter et traiter les commandes.

• Déterminer les budgets de facturation et assurer la fiabilité des informations afin de contribuer à la satisfaction des clients.

• Elle tient à fournir l'interface technique et commerciale générale pour communiquer ses objectifs commerciaux.

• Application du système d'information GAIA/BELING ADSL, pour identifier les appels, internet et payer les factures.

1.7.6 Sous Direction Technique

Elle est divisée en :

1.7.6.1 Département Planification et Suivi

Il s'occupe de :

- Élaborer des études et des plans.
- Suivi des projets.
- Assurer l'application des normes techniques à tous les projets complétés.
- Développer une stratégie pour suivre le rythme de la modernité.

1.7.6.2 Département Réseau d'Accès au Service

Parmi ses missions • Assurer l'exploitation et la maintenance du réseau d'accès.

- Assurer le déploiement des accès réseau.
- Assurer la qualité de service et la maintenance des réseaux.

1.7.6.3 Département Réseau des Transports

Il fait :

- Assurer la qualité de service et la maintenance du réseau.
- Superviser la réparation des pannes techniques 24h/24.

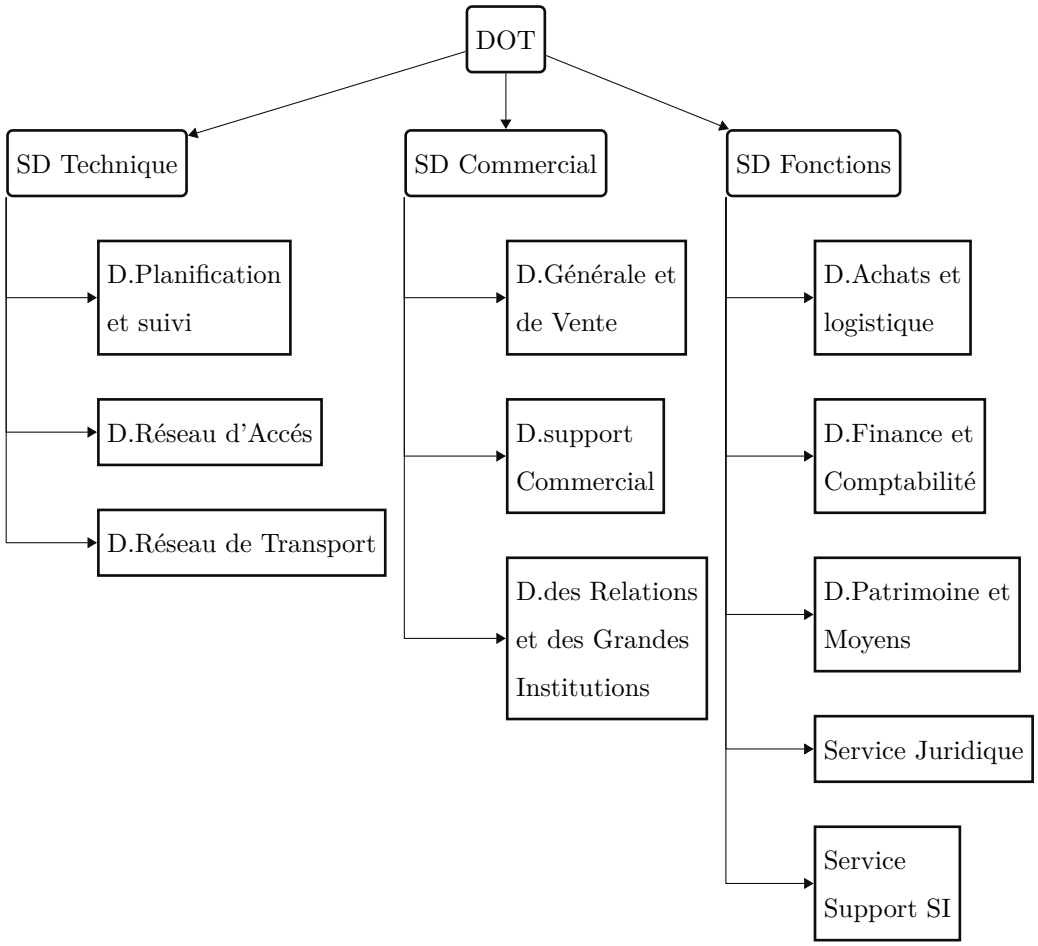
1.8 Le Département Réseau d'Accès (Département d'Accueil)

Il fait : • Assurer l'exploitation et la maintenance du réseau d'accès.

- Assurer le déploiement des accès réseau.
- Assurer la qualité de service et la maintenance des réseaux.

C'est notre département d'accueil au sein de AT et fait l'objet de notre étude dans le chapitre 2.

1.9 Organigramme de la DOT



1.10 Conclusion

L'objectif de cette présentation est de définir la place de l'entreprise par rapport au secteur d'activité de la télécommunication, et surtout l'atout de l'entreprise en tant que lieu de stage pour une formation pratique . AT engagé dans le monde des technologies, en guise de l'accroissement de l'offre de ses services et la qualité service offerte. En outre le développement d'un réseau national fiable et connecté aux autoroutes de l'information. En ce qui suit, nous allons entamer notre sujet de thème avec lequel nous avons eu recours à de différents domaines d'application (Télécommunications ; Détection des Anomalies ; Machine Learning ...).

Chapitre 2

Etat de l'art sur les réseaux de télécommunication et la détection des anomalies

2.1 Introduction

Au cours de ce chapitre, nous allons s'intéresser aux notions de réseaux de télécommunications, particulièrement le LTE. De même à la détection des anomalies et ses différentes techniques. La quatrième génération des réseaux mobiles apporte un véritable tournant dans le foisonnement et la disparité des générations précédentes. Cependant, ils doivent garantir les exigences de disponibilité ; de continuité et de qualité de service. L'optimisation de ce dernier conduit à la satisfaction des abonnées. En effet, l'évolution de QoS est un exercice de mesure visant à rendre compte la diversité des expériences des utilisateurs dans les conditions d'usages les plus répandues comme le transfert des données. La détection des anomalies, en particulier les défauts de qualité est une fonction qui intéresse de nombreux secteurs, nous allons traiter le cas sur de notre réseau LTE dans le prochain chapitre d'application. Ce domaine de recherche est très sollicité depuis quelque années. Il s'agit des problèmes assez difficiles pour lesquels les données normales sont majoritairement présentes, tandis que les données anormale sont rarement étiquetées .Le but initial de contrôle de qualité est la recherche d'une prévision de défaillance, en vue d'une maintenance prédictive.

2.2 État de l'art sur les réseaux de Télécommunication

2.2.1 Présentation de 3GPP

Le 3GPP est un consortium créé en 1998 à l'initiative de l'ETSI (European Telecommunications Standards Institute). Le 3GPP a pour objectif de définir des spécifications permettant l'interfonctionnement d'équipements

de constructeurs différents. Contrairement à ce que son nom suggère, le champ d'activités du 3GPP ne se limite pas à la normalisation de systèmes 3G. Son rôle consiste à maintenir et développer les spécifications des systèmes :

- GSM/GPRS/EDGE.
- UMTS (FDD et TDD).
- LTE, ainsi que celles du réseau cœur EPC.

Le 3GPP est composé d'un groupe de coordination appelé PCG (Project Coordination Group) et de différents groupes de spécifications techniques appelés TSG (Technical Specification Groups). Il convient d'indiquer que le 3GPP n'est pas un organisme de normalisation en tant que tel. Il définit des spécifications techniques qui sont par la suite approuvées et publiées par des organismes de normalisation régionaux, propres à un pays ou une région du monde. Les modifications des spécifications approuvées par les groupes de travail sont associées à une release (version). Une release correspond à un ensemble de nouvelles fonctionnalités introduites dans la norme par les groupes du 3GPP dans une période de temps donnée et représente un palier significatif dans l'évolution des systèmes. Le 3GPP a défini plusieurs releases, nous allons se focaliser sur la 8ème release qui est l'introduction des évolutions HSPA+ CPC et DC-HSDPA, et la première release du réseau d'accès LTE et du réseau cœur EPC[16].

2.2.2 Évolution des réseaux mobiles

La technologie dans le domaine de la téléphonie mobile n'a pas cessé de se développer depuis ces dernières années. En effet, en quelques décennies, la qualité des signaux a connu de grands changements en passant de la 1ère génération à la 5ème Génération, La première génération de ces systèmes sans fil a été introduite dans les années 70. Dans cette section, Nous présentons ces différentes générations de réseaux mobiles [18] :

- **La 1ère G** : Ce réseau qui fonctionne sur un système de communication analogique n'a pas connu le succès espéré à cause de certains problèmes de communication et de la qualité des téléphones mobiles de l'époque.
- **Le GSM 2G** : (global system for mobile communication) ne permet que d'échanger par voix.
- **Le GPRS 2.5G** : (General packer radio service) permet d'échange des données SMS ,data, appels , son débit théorique maximale est de 171.2 kb/s.
- **Le EDGE 2.75** : (Enhaced Data Rates For GSM Evolution) son débit peut atteindre les 384kb/s « pré-3G ».
- **La 3G** : la 3ème Génération, le débit proposer d'échange est à 1.9 Mb/s.
- **La 3G+** : appelé « HSDPA », permet de monter le débit d'échange de données théorique à 14.4Mb/s.
- **Le H+** : appelé « Dual carrier » ou « HSPA+ », le débit est à 42Mb/s il s'agit d'un réseau approchant de 4G.
- **La 4G** : la 4ème Génération Nommé LTE (Tong Terme Evolution) l'échange de données peut dépasse les 100Még/s.
- **Le 4G+** : nommé LTE advenced propose un débit rapide que la 4G avec un débit théorique compris entre 100 et 150 Mb/s.
- **La 4.9G** : nommé LTE Advenced Pro son débit est à 3Gbits/s.
- **La 5G** : elle ambitionne en effet d'offrir aux usagers l'ultra haut débit mobile, avec des débits dépassant les 10 Gbit/s.

2.2.3 Architecture du réseau LTE

Les réseaux LTE sont des réseaux cellulaires constitués de milliers de cellules radio qui utilisent les mêmes fréquences hertziennes, grâce aux codages radio. Dans le sens descendant le mécanisme OFDMA est une combinaison de technique de modulation et de technique d'accès multiple répartit la bande passante en N multiples sous-porteuses orthogonales qui sont partagées par de plusieurs utilisateurs. Chaque sous-porteuse est modulée indépendamment en utilisant des modulations numériques : QPSK, QAM-16, QAM-64. Et dans le sens montant le mécanisme SC-FDMA est basé sur le même principe qu'OFDMA, mais il a été choisi car son taux de PAPR « Peak-to-Average Power Ratio », est inférieur à celui de l'OFDMA. Plus ce taux est haut, plus le prix et la consommation d'énergie du terminal augmentent.
Le réseau LTE est constitué d'une partie radio E-UTRAN [24].

2.2.3.1 Réseau d'accés Radio e-UTRAN

Le réseau d'accès E-UTRAN comprend un seul type d'entité, la station radioélectrique eNodeB à laquelle se connecte le mobile (UE)[38].

- **LTE-Uu avec le mobile UE** : Cette interface est utilisée pour la connexion du mobile à l'entité eNodeB. Elle supporte le trafic du mobile et la signalisation échangée entre le mobile et l'entité eNodeB. Cette signalisation supporte la signalisation échangée entre le mobile et l'entité MME du cœur de réseau.
- **X2 avec les autres entités eNodeB** : Cette interface est utilisée pour la mobilité intra E-UTRAN et pour l'échange d'information de charge de la cellule. Elle supporte le trafic du mobile et la signalisation échangée entre deux entités eNodeB.
- **S1-MME avec l'entité MME du réseau cœur** : Cette interface est utilisée pour l'établissement du support (porteur) radioélectrique, pour le paging et pour la gestion de la mobilité. Elle supporte la signalisation échangée entre l'entité MME et l'entité eNodeB. Cette signalisation porte la signalisation échangée entre le mobile et l'entité MME.
- **S1-U avec l'entité SGW du réseau cœur** : Cette interface supporte uniquement le trafic du mobile.

2.2.3.2 Le réseau cœur EPC (Evolved Packet Core)

Le cœur EPC est un cadre normalisé dans la version 8 du 3GPP pour fournir des données et une voix convergente sur un réseau basé sur LTE. Evolved Packet Core est basé sur une connexion réseau constante ou une connexion permanente. Il aide à combiner la voix et les données sur une architecture de service de protocole Internet. Cela aide les opérateurs de services dans les opérations ainsi que le déploiement d'un réseau de paquets pour LTE ainsi les générations précédentes ou un accès fixe[38].
Evolved Packet Core est considéré comme le composant clé de l'évolution de l'architecture de service. Les composants clés d'Evolved Packet Core sont :

• **L'entité MME (Mobility Management Entity)** : le MME Aide à authentifier et à suivre les utilisateurs du réseau ainsi qu'à gérer les états de session. Il fournit les informations nécessaires à l'identification de l'usager au moment de son authentification dans le système.

• **L'entité SGW** : La passerelle de service SGW, est un élément du plan de données au sein de la LTE. Aide au routage des paquets de données sur le réseau ainsi d'acheminer les données entre la partie accès et le PDN-GW.

• **L'entité PGW (PDN Gateway)** : L'entité PGW est le routeur de passerelle, Aide à gérer la qualité du service fourni et également à l'inspection approfondie des paquets.

• **L'entité HSS (Home Subscriber Server)** : Le HSS est un composant de base essentiel dont un opérateur mobile a besoin pour fournir des services mobiles sur le réseau LTE. Il permet de stocker des informations d'abonnement pouvant servir au contrôle des appels et à la gestion de session des utilisateurs réalisé par le MME. Il entrepose, pour l'identification des utilisateurs, la numérotation et le profil des services aux quels ils sont abonnés.

• **L'entité PCRF (Policy and Charging Rules Function)** : est une entité qui exécute principalement deux grandes tâches. La première est de gérer la qualité de service que requiert le réseau, et alloue en conséquence les porteuses bearer appropriées. La deuxième tâche se rapporte principalement à la tarification.

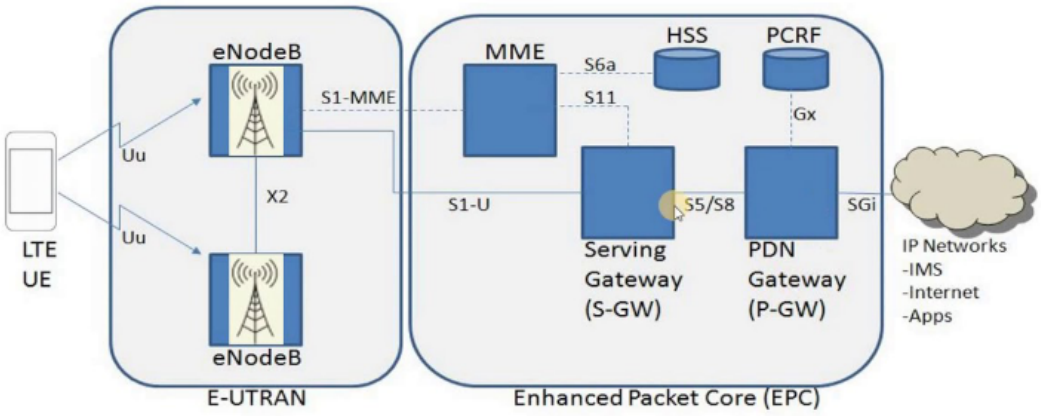


Fig. 2.1: Architecture du réseau LTE

2.2.4 Les Ressources radio du réseau LTE

Les eNodeBs ont des systèmes complexes qui consomment plusieurs types de ressources pendant le fonctionnement. Ils sont constitués de plusieurs modules de traitement, avec des limitations dans leur utilisation de la mémoire et du processeur, mais consomment également d'autres ressources, telles celles associées à l'utilisation de la radio spectre par antennes . Nous allons présenter ici brièvement ceux qui sont mentionnés dans ce mémoire.

2.2.4.1 Les ressources eNodeBs

Nous allons présenter ici brièvement quelques unes [23] :

- **Licence du nombre d'utilisateurs :** connectés La licence du nombre d'utilisateurs connectés spécifie le nombre maximal autorisé d'utilisateurs en mode RRC CONNECTED. Chaque utilisateur connecté consomme les ressources radio et les ressources de transport. S'il y a trop d'utilisateurs connectés, ils ne peuvent pas être bien servis par eNodeB et les nouveaux services ne peuvent pas être admis.
- **Ressources radio messagerie :**si le nombre réel de messages d'appel dépasse la capacité d'échange, la eNodeB sera incapable de traiter tous les messages d'appel.
- **Processeur principal de commande :**si la charge de commande du processeur principal dépasse sa capacité de traitement, la détérioration des indicateurs clés de performance (KPIs) se produira.
- **CPU LBP :** si la charge sur une unité de traitement en bande de base LTE (LBP) excède la capacité de traitement du processeur, la détérioration des indicateurs de performance se produit.
- **Ressource groupes de transport :** les groupes de ressources de transport portent chacun un ensemble de flux de données. Ils sont situés à la couche de liaison du modèle TCP / IP. Si la charge sur un groupe de ressources de transport dépasse la bande passante configurée pour le groupe, des exceptions peuvent se produire (par exemple, les paquets peuvent être perdus), affectant l'utilisateur.

2.2.4.2 Ressources cellulaires

Nous allons présenter ici brièvement quelques unes [23] :

- **PRBs :**l'utilisation de blocs de ressources physiques reflète la bande passante de liaison montante et descendante consommée sur l'interface air (interface Radio).
- **Ressources PUCCH :**Ressources du canal de commande de liaison montante physique (PUCCH) insuffisant ont des effets négatifs sur les éléments suivants :

-Admission de nouveaux services et handover.

- Nombre de UEs qui peuvent être prévu.

- **SRS Ressources signal de référence Sounding (SRS) :**sont attribués aux UEs pour l'accès au réseau. Si un LBPd est utilisé, UEs peuvent accéder au réseau, même lorsque les ressources SRS ne sont pas attribuées. Si les ressources SRS sont insuffisantes, l'eNodeB ne peut pas obtenir des informations des mesures précises, ce qui affecte alors l'utilisation efficace des ressources radio.
- **Ressources PRACH :**Ressources canal d'accès aléatoire physiques(PH), il s'agit de préambules d'accès aléatoires effectués sur le PRACH. Si les ressources PRACH sont insuffisantes pour traiter toutes les tentatives d'accès, on enregistre des retards d'accès et même des échecs de tentatives d'accès.
- **Ressources PDCCH :** si les ressources de canal de commande de liaison descendante physique (PDCCH) sont limitées, les retards de planification sont longs ainsi que le nombre d'utilisateur servis sera limité d'où une expérience utilisateur insatisfaisante voir médiocre surtout durant les heures de pointe.

2.2.5 Bloc de ressources et élément de ressource (RB,RE)

En LTE, l'espace temps/fréquence est divisé en PRB (Physical Resource Blocks). Un PRB est le plus petit élément d'allocation des ressources affectées par le planificateur de station de base. Il est constitué en domaine fréquentiel de 12 sous-porteuses, chacune de largeur 15 KHz, en tout 180 KHz, et d'un Time Slot dans le domaine temporel, autrement dit 6 ou 7 symboles selon la taille du préfixe cyclique(PC). Un élément de ressource RE (Resource Element) est formé par une seule sous-porteuse et un seul symbole dans le domaine temporel, d'où il ne peut contenir qu'un seul symbole de modulation (QPSK, 16QAM, 64QAM). Ce dernier est utiliser soit pour transmettre des données ,soit pour transmettre les signalisation ou des signalisations physiques. Le nombre total de sous-porteuses disponibles dépend de la largeur de bande de transmission globale du système. Les spécifications LTE définissent les paramètres de bande passante de système à partir de 1,25 MHz à 20 MHz. Le nombre de sous-porteuses pour chaque ressource bloc et le nombre de symboles par ressources block varient en fonction de la longueur de préfixe cyclique et de l'espacement de sous-porteuse pour les deux voie [17].

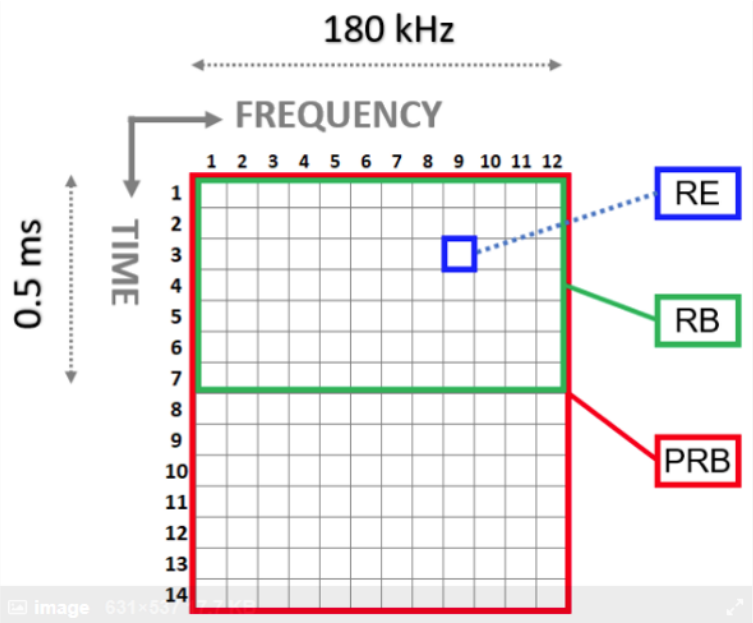


FIG. 2.2: La différence entre ressources block et physical ressources block en LTE

Rappelons que selon la bande allouée au LTE qui s'étend de 1.4 MHz minimum à 20 MHz, le nombre de PRB est le suivant :

1,4 MHz $\rightarrow 6PRBs$

$3MHz \rightarrow 15PRBs$

$5MHz \rightarrow 25PRB$

$10MHz \rightarrow 50PRBs$

$15MHz \rightarrow 75PRBs$

$20MHz \rightarrow 100PRBs$

2.2.6 Le débit en LTE

Le débit est la quantité de données qui transite sur un réseau pendant une durée déterminée.

2.2.6.1 Voie de transmissions

Une voie de transmission est un ensemble des moyens nécessaires pour assurer une transmission de signaux dans un seul sens entre deux points. Plusieurs voies de transmission peuvent partager un support commun ; par exemple, dans les multiplex à répartition en fréquence ou les multiplex temporels, chaque voie dispose d'une bande de fréquences particulière ou d'un intervalle de temps particulier répété périodiquement. Une voie de transmission peut être qualifiée par la nature des signaux qu'elle transmet, par sa largeur de bande ou par son débit binaire. Exemples : voie téléphonique, voie télégraphique, voie de données, voie de 10 MHz, voie à 34 Mbit/s.
La transmission de l'information sur la voie radio dans les systèmes mobiles s'effectue soit depuis une station de base vers un mobile (liaison descendante ou "downlink"), soit depuis un mobile vers la station de base (liaison montante ou "uplink") comme illustré par la figure 2.3.

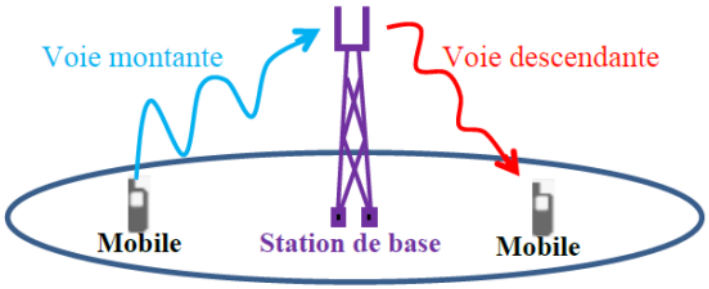


FIG. 2.3: Les voies de transmission dans un réseau mobile

2.2.6.2 Calcul du débit en DOWNLINK et UPLINK

En LTE pour 20 MHZ, on a 100 PRB, et chaque PRB contient 12 sous-porteuses x 7 symboles (Normal CP) soit 84 symboles. Un PRB contient 84 symboles dans un slot de 0.5 ms. On parle de mbs/s donc dans 1ms, on a 84x2 = 168 symboles/1ms Donc dans 100 PRB on a 168 x 100 = 16800 symboles/ms =16800000 symboles/s=16.8Msps Si on utilise une modulation 64 QAM, chaque symbole est codé sur 6 bits, on aura un débit de 16.8 Msps x 6 = 100.8 Mbps. Pour un système LTE avec MIMO 4x4, le débit sera multiplié par 4 soit 403.2 Mbps si on utilise une modulation 64 QAM, chaque symbole est codé sur 6 bits, on aura un débit de 16.8 Msps x 6 = 100.8 Mbps. Pour un système LTE avec MIMO 4x4, le débit sera multiplié par 4 soit 403.2 Mbps.

2.2.7 Le modèle Open Systems Interconnetion(OSI)

Est une norme de communication, en réseau, de tous les systèmes informatiques. Ce modèle a tenté de définir un standard de développement par couche/niveau/protocole. Même si ce modèle n'a pas vraiment réussi à s'imposer réellement, les concepts développés sont utilisés de manière assez universelle au niveau des 3 premières couches (couche physique, couche de liaison de données et couche réseau).
En LTE, La couche physique (niveau 1) sert à recevoir/transmettre via la radio toutes les informations provenant de la couche MAC (niveau 2) en mappant les canaux de transport aux canaux physiques de l'interface air. La couche MAC (Medium Access Control, niveau 2) sert à recevoir/transmettre toutes les informations provenant de la couche RLC (Radio Link Control, niveau 2) en mappant les canaux logiques aux canaux de transports. La couche RRC (Radio Resource Control, niveau 3) reçoit/transmet les messages de contrôle pour gérer les appels au niveau du réseau d'accès. La couche NAS (Non Access Stratum, niveau 3) reçoit/transmet les messages de contrôle sur lequel le réseau d'accès n'a aucune action mais sont échangés par son biais entre le mobile et le cœur de réseau .

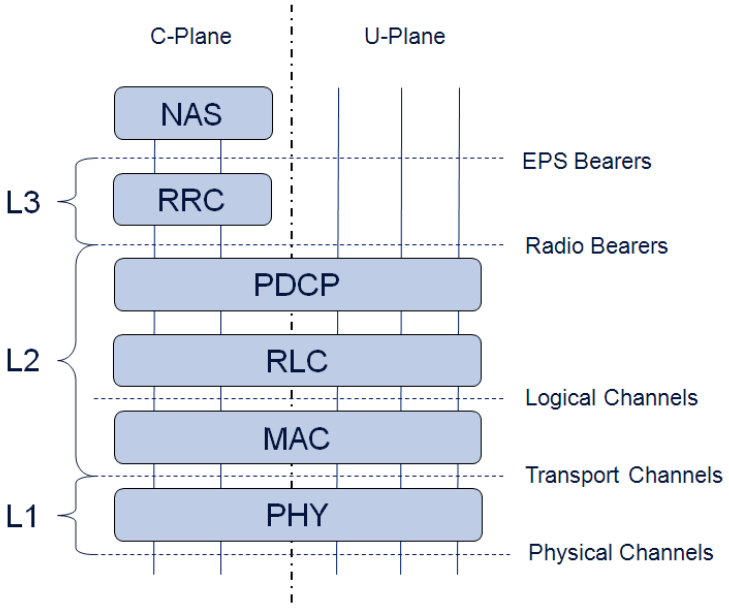


Fig. 2.4: Protocol stack LTE

2.2.8 Les signaux physiques

Les signaux physiques correspondent également à des éléments de ressource et sont associés à des paramètres de transmission physiques prédéfinis. On distingue deux grands types de signaux physiques : Les signaux de référence ou Reference Signais (RS) et les signaux de synchronisation.
nous allons donner des détails sur les signaux de références.

2.2.8.1 Signaux de référence

Les signaux de référence (RS, pour Reference Signals), aussi appelés pilotes, sont des signaux connus à l'avance du récepteur qui permettent à l'UE d'estimer son canal et plus généralement, d'effectuer les différentes mesures définies au niveau de la couche physique. Parmi ces dernières, on distingue en particulier la puissance reçue sur les signaux de référence (RSRP, pour Reference Signais Received Power) et la qualité du signal reçu sur les signaux de référence (RSRQ, pour Reference Signais Received Quality). Les signaux de référence ne portent pas d'information et occupent des éléments de ressource qu'il n'est pas possible de réutiliser pour transmettre des données. Leur définition doit ainsi répondre à un compromis entre les performances qu'ils apportent et la réduction du nombre de ressources utiles qu'ils entraînent.
La figure 2.5 représente la position des signaux de références au niveau de la couche physique.

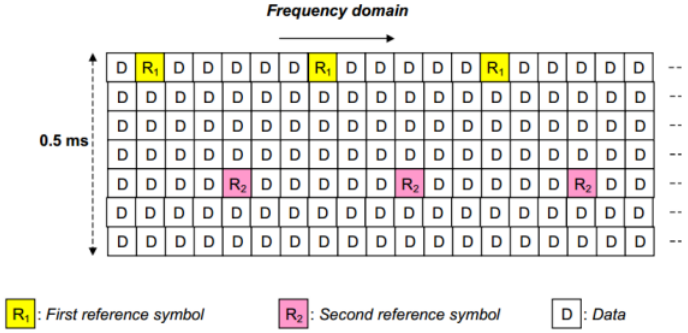


FIG. 2.5: La position des signaux de références

2.2.8.2 Les conditions RF (Fréquence Radio)

En LTE, comme tout système sans fil, les performances ont une relation directe avec les conditions RF du moment. Pour faciliter l'analyse des performances dans LTE, il est courant de définir certaines plages de mesures RF qui correspondent à certaines conditions RF typiques dans lesquelles on peut se trouver[6].

- S : indique la puissance des signaux utilisables mesurés. Les signaux de référence (RS) et les canaux physiques partagés de liaison descendante (PDSCH) sont principalement impliqués.
- I : indique la puissance d'interférence moyenne - la puissance des signaux mesurés ou des signaux d'interférence de canal provenant d'autres cellules du système actuel.
- N : indique le bruit de fond, qui est lié aux largeurs de bande de mesure et aux coefficients de bruit du récepteur.

Toutes les quantités sont mesurées sur la même largeur de bande et normalisées à une largeur de bande de sous-porteuse.

- **RSSI :** (indication du niveau du signal reçu) de la connexion LTE. La valeur est mesurée en dBm (dBm).

$$RSSI = S_{tot} + I_{tot} + N_{tot}$$

où l'indice 'tot' indique que la puissance est mesurée sur les 12 sous-porteuses NRE de la largeur de bande de mesure. La puissance totale reçue par la cellule de desserte dépend du nombre de sous-porteuses émises dans le symbole OFDM transportant R0, et du nombre d'antennes d'émission. Nous pouvons modéliser cet impact en utilisant le facteur d'activité de sous-porteuse par antenne x et définir : $S_{tot} = x \times 12 \times N_{prb} \times RSRP$ Tel que $x = RE/RB$.

- **RSRP :** (puissance du signal de référence reçu) - puissance moyenne des signaux pilotes reçus (signal de référence) ou niveau du signal reçu de la station de base. La valeur RSRP est mesurée en dBm (dBm).
- **RSRQ :** (Qualité du signal de référence reçu) - caractérise la qualité des signaux pilotes reçus. La valeur RSRQ est mesurée en dB. Voici la formule de calcule tel que $N_{prb} = RB_s$.

$$RSRQ = N_{prb} \times (RSRP/RSSI)$$

- **SINR :** (rapport signal sur interférence + bruit), également appelé rapport CINR (rapport porteuse à interférence + bruit), est le rapport entre le niveau du signal et le niveau de bruit (ou simplement le rapport signal sur bruit). La valeur SINR est mesurée en dB (dB). C'est simple : plus la valeur est élevée, meilleure est la qualité du signal. Aux valeurs SINR inférieures à 0, la vitesse de connexion sera très faible, car cela signifie qu'il y a plus de bruit dans le signal reçu que dans la partie utile et que la probabilité de perdre une connexion LTE existe également. en ce qui suite, nous donnons sa formule mathématiques Tel que $I_{tot} + N_{tot} = RSSI - S_{tot}$

$$SINR = S/(I + N)$$

Le SINR est également une mesure de la qualité du signal, mais il n'est pas défini dans les spécifications 3GPP mais défini par le fournisseur UE. Il n'est pas signalé au réseau. SINR est beaucoup utilisé par les opérateurs, et l'industrie LTE en général, car il quantifie mieux la relation entre les conditions RF et le débit . Les UE LTE utilisent généralement le SINR pour calculer le CQI (Channel Quality Indicator) qu'ils rapportent au réseau. Il est courant d'utiliser le rapport signal sur interférence (SINR) comme indicateur de la qualité du réseau. Il convient toutefois de noter que les spécifications 3GPP ne définissent pas le SINR et que, par conséquent, l'UE ne signale pas le SINR au réseau. Le SINR est toujours mesuré en interne par la plupart des UE et enregistré par les outils de test de conduite. La figure 2.6 montre les différentes valeurs de ces paramètres, qui correspondent à une qualité de signal LTE très mauvaise (Cell Edge), médiocre (Mid Cell), bonne (Bonne) et très bonne (Excellent). Une bonne classification des conditions RF par rapport aux KPI LTE.

Signal quality:	RSRP (dBm)	RSRQ (dB)	SINR/CINR (dB)
very good	>= -80	>= -10	>= 20
good	from -80 to -90	from -10 to -15	from 13 to 20
bad	from -90 to -100	from -15 to -20	from 0 to 13
very bad	<= -100	< -20	<= 0

FIG. 2.6: Les 3 valeurs des paramètres de qualité de signal

2.2.9 Les KPI (les indicateurs clés de performances)

Dans le processus d'optimisation des performances du réseau radio, nous devons surveiller la valeur du KPI afin de fournir une meilleure qualité d'abonné ou d'obtenir une meilleure utilisation des ressources réseau installées. Pour mesurer les performances du réseau, nous avons des catégories de KPI et des nombres de KPI de chaque catégorie[30].

2.2.9.1 Accessibilité KPI

Ils sont utilisés pour mesurer correctement si les services demandés par les utilisateurs peuvent être accessibles dans des conditions données,et se réfère également à la qualité d'être disponible lorsque les utilisateurs en ont besoin. Par exemple : la demande de l'utilisateur pour accéder au réseau, accéder à l'appel vocal, à l'appel de données[30].

2.2.9.2 Rétention KPI

Sont utilisés pour mesurer comment le réseau conserve la possession de l'utilisateur ou est capable de détenir et de fournir les services aux utilisateurs[30].

2.2.9.3 Mobilité KPI

Sont utilisés pour mesurer les performances du réseau qui peut gérer le mouvement des utilisateurs tout en conservant le service pour l'utilisateur, comme le transfert des données[30].

2.2.9.4 Intégrité KPI

Sont utilisés pour mesurer le caractère ou l'honnêteté du réseau vis-à-vis de son utilisateur, tels que le débit, la latence auxquels les utilisateurs ont été servis.
Ces KPI's nous renseignent sur les indicateurs de l'état du canal et de la latence de délivrance des données mesurés par l'UE en voie montante comme CQI, RI et des KPI's sur le HARQ qui est un mécanisme de retransmission des blocs de transport reçus de manière erronée.

$$DLthroughput_{QCI=x} = DRB \times IPthroughput_{QCI=x}DL$$
$$DLthroughput_{QCI=x} = DRB \times IPthroughput_{QCI=x}UL$$

On utilisera dans notre étude CQI. Ce KPI indique la valeur moyenne de la qualité du signal radio, la valeur de CQI devrait être supérieur à 7, une valeur autour de 7 peut être acceptable, une valeur autour de 8 peut être considérée comme bonne, une valeur supérieur a 9 est considérée comme excellente ; nous allons donner plus de détails prochainement [30].

$$IPthroughputinDL = ThpVOIDI/ThptimeDI(Kbits/s)$$

Débit IP E-UTRAN : Uu KPI qui montre comment E-UTRAN impacte la qualité de service fournie à un utilisateur final.

Latence IP E-UTRAN : une mesure qui montre l'impact de l'E-UTRAN sur le retard subi par un utilisateur final.Temps entre la réception du paquet IP et la transmission du premier paquet sur l'Uu.

2.2.9.5 KPI de disponibilité

Sont utilisés pour mesurer la disponibilité du réseau, adapté ou prêt pour les utilisateurs à utiliser les services[30].

Disponibilité des cellules E-UTRAN(Availability) : Un KPI qui montre la disponibilité de la celluleE-UTRAN. Pourcentage de temps pendant lequel la cellule est considérée comme disponible[30].

$$Availability = \frac{Le\ temps\ de\ ladisponibilit\ de\ lacellule}{la\ mesure\ du\ temps} \times 100\%$$

Quant à la définition de la cellule comme disponible, elle doit être considérée comme disponible lorsque l'eNodeB peut fournir le service E-RAB dans la cellule.

2.2.9.6 KPI d'utilisation

Sont utilisés pour mesurer l'utilisation du réseau, si la capacité du réseau est atteinte sa ressource.

Utilisation moyenne du porteur EPS dédié actif(Mean Active Dedicated EPS Bearer Utilization) : cet indicateur de performance clé décrit le rapport entre le nombre moyen de supports EPS dédiés actifs et le nombre maximal de supports EPS dédiés actifs fournis par le réseau EPC, et il est utilisé pour évaluer les performances d'utilisation du réseau EPC. Ce KPI est obtenu par le nombre moyen de supports EPS dédiés en mode actif divisé par la capacité du système.

$$MADRBU = \frac{Le\ NB\ moyen\ supports\ EPS\ ddis}{la\ capacit\ du\ systme} \times 100\%$$

2.2.10 La modulation

La modulation utilisée dans le LTE est une modulation adaptative qui varie en fonction de la distance qui sépare l'abonné de l'eNodeB. Chaque sous-porteuse est modulée à l'aide de différents niveaux de modulation : QPSK (Quadrature Phase Shift Keying) ou (4-QAM), 16-QAM et 64-QAM (Quadrature Amplitude Modulation). Par exemple, si les modulations disponibles sont le QPSK et le 16-QAM, dans le cas où le canal est marqué

comme bon, on utilisera la modulation 16-QAM, qui offre un meilleur débit mais une plus faible robustesse. Par contre, si le canal est marqué comme dégradé, on utilisera la modulation QPSK, permettant un débit plus faible, mais plus robuste (moins sensible aux interférences). La modulation d'amplitude en quadrature (QAM) permet de doubler l'efficacité de la modulation d'impulsion en amplitude (PAM) en modulant les amplitudes des composants sinus et cosinus de la porteuse. Le signal produit consiste en deux trains d'impulsions PAM en quadrature de phases.

Modulations Downlink : QPSK, 16QAM et 64QAM.

Modulations Uplink : QPSK et 16QAM.

Modulation en quadrature de phase (QPSK) :

Souvent connue sous le nom de 4-PSK ou QPSK (Quadrature Phase-Shift Keying), cette modulation utilise un diagramme de constellation à quatre points, à équidistance autour d'un cercle. Avec quatre phases, QPSK peut coder deux bits par symbole. Deux signaux en quadrature sont générés à partir d'un oscillateur local à la fréquence quadruple. Le train de donnée binaire est séparé en deux "sous trains" appelés I et Q. La paire de valeur, constitue ce que l'on appelle un symbole.

Modulation 16-QAM :

La modulation d'amplitude en quadrature (en anglais, quadrature amplitude modulation QAM) est une forme de modulation d'une porteuse par modification de l'amplitude de la porteuse elle-même et d'une onde en quadrature (une onde déphasée de 90° avec la porteuse) selon l'information transportée par deux signaux d'entrée. L'amplitude et la phase de la porteuse sont simultanément modifiées en fonction de l'information à transmettre. La constellation, qui est en conséquence le nombre de bits pouvant être transmis une seule fois, peut être augmentée pour un meilleur débit binaire, ou diminuée pour améliorer la fiabilité de la transmission en générant moins d'erreurs binaires.

Modulation 64-QAM :

Dans le cas d'une modulation, six bits sont mappés dans le symbole complexe. La constellation est constituée de 64 symboles.

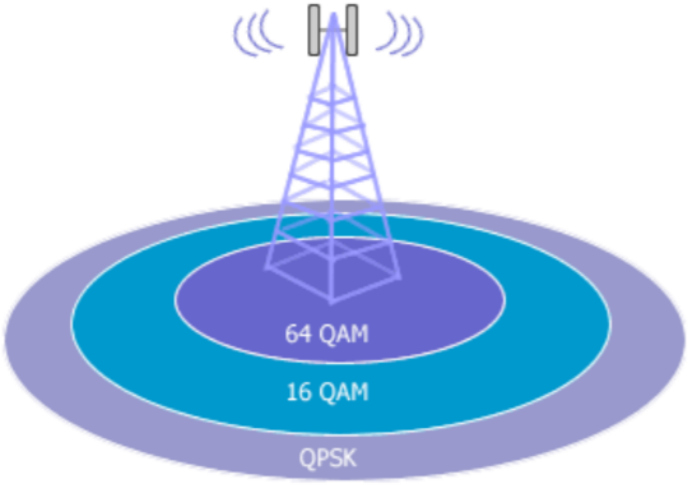


Fig. 2.7: Types de modulation en DL

2.2.11 La technologie MIMO (Multiple Input Multiple Output)

La technologie MIMO consiste en l'utilisation de plusieurs antennes à l'émission et à la réception. Le but de la technique MIMO était d'améliorer le débit, d'augmenter l'efficacité spectrale, diminuer la probabilité de coupure du lien radio, etc. Le principe de la technologie MIMO consiste à émettre dans un même canal des signaux transmis sur des antennes différents. A la réception aussi et avec un certain nombre d'antennes et des traitements adéquats, il s'agit de simuler cette réception dans une même bande de n canaux différents.
La technologie MIMO profite de ces différents canaux pour améliorer la rapidité de transmission des données. On peut considérer trois catégories de MIMO :

- **La diversité spatiale MIMO :** on transmet simultanément un même message sur différentes antennes à l'émission. Les signaux reçus sur chacune des antennes de réception sont ensuite remis en phase et sommés de façon cohérente.
- **Le multiplexage spatial MIMO :** chaque message est découpé en sous message. On transmet simultanément les sous-messages différents sur chacune des antennes d'émission. Les signaux reçus sur les antennes de réception sont réassembles pour reformer le message entier d'origine.
- **Le MIMO – Beamforming (formation de Faisceau :)**le réseau d'antennes MIMO est utilisé pour orienter et contrôle le faisceau d'onde radio (amplitude et phase du faisceau). On peut ainsi créer des lobes constructifs / destructifs et optimiser une transmission entre l'émetteur et la cible. Les techniques de beamforming permettent à la fois d'atteindre une couverture radio (d'une station de base ou d'un point d'accès par exemple) et de limiter les interférences.

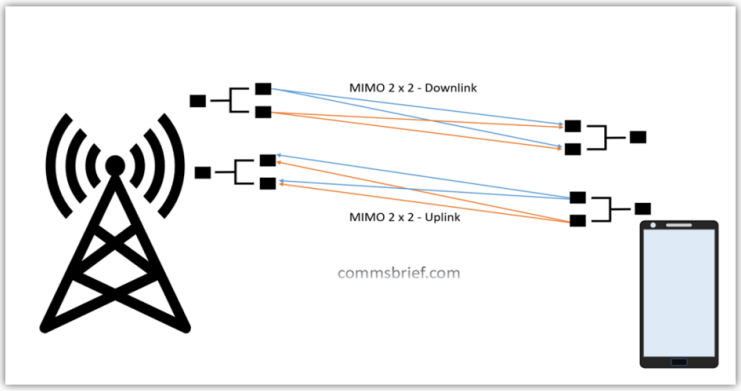


Fig. 2.8: La technique MIMO en LTE

2.2.12 La qualité de service (QoS)

2.2.12.1 Définition

La qualité de service est la capacité de transmission dans de bonnes conditions d'un certain nombre de paquets dans une connexion entre un émetteur et un récepteur. Elle peut être présentée sous plusieurs termes tels que la disponibilité, le débit, les délais de transmission, la gigue, le taux de perte de paquets, etc.
Elle regroupe un ensemble de technologies mise en œuvre pour assurer des débits suffisants et constants sur tous les types de réseaux[47].

2.2.12.2 Le poteur(bearer) EPS

Dans les réseaux LTE, la notion de canaux logiques EPS est introduite afin de définir plusieurs classes de QoS supportées par l'E-UTRAN et l'EPC. Les données véhiculées dans un même canal logique EPS subissent le même traitement et requirent la même qualité de service. Néanmoins, ce traitement diffère d'un canal à un autre par les priorités accordées aux paramètres d'EPC. Le porteur EPS est un canal logique établit entre l'équipement utilisateur et la passerelle PGW, il permet de véhiculer tout le trafic de l'équipement utilisateur à travers le même canal physique ou le canal radio. Ainsi, plusieurs porteurs doivent être créés afin de distinguer entre les méthodes de traitement des différentes transmissions. Pour différencier les bearer, les flux sont identifiés par deux critères (QCI ,ARP) .
Différents porteurs(Bearer) physique : Chaque porteur est identifié par l'identifiant de tunnel TEID (Tunnel Endpoint ID) sur chacune des interfaces. Evidemment, les paramètres CQI/ARP sont identiques sur chaque porteur mis en place pour une EPS session donnée. Ainsi que l'EPS session se charge de gérer les flux sur chaque équipement, autrement dit gère les Bearer entre :

l'UE - eNb - SGW - PGW.

L'utilisateur pouvant lancer plusieurs applications simultanément, plusieurs EPS bearer peuvent être établis pour un même utilisateur. Chaque EPS bearer est identifié par l'EPS bearer ID, lequel est alloué par le MME.

• [UE] - [eNodeB] (Data Radio Bearer (DRB)) :EPS bearer est établi sur l'interface LTE-Uu. Le trafic utilisateur (IP packet) est délivré dans le DRB. Differents DRBs sont identifiés par le DRB ID alloués par le eNodeB.
• [eNB] - [S-GW] (S1 bearer) :EPS bearer établi sur l »interface sur l'interface S1-U interface. Le trafic utilisateur est délivré via un tunnel GTP (GTP-U) Différents S1 bearers sont identifiés par le TEID, qui est alloué par les équipements périphérique (eNB et S-GW).
• [S-GW] - [P-GW] (S5 bearer) :EPS bearer est établi sur l'interface S5.. Le trafic utilisateur est délivré via un tunnel GTP (GTP-U) Différents S5 bearers sont identifiés par le TEID, qui est alloué par les équipements périphérique (S-GW et P-GW).
• [UE] - [S-GW] (E-RAB bearer) :E-RAB est un bearer logique entre l'UEet le S-GW. Il est constitué du DRB et du S1 bearer[42].
Il existe deux types de porteurs EPS peuvent coexister dans un réseau LTE :
Porteurs par défaut :Est établi avec les paramètres QCI et ARP fournis par le MME. Ces valeurs sont définies par l'abonnement de l'utilisateur dont les données de souscriptions sont sauvegardées dans le HSS. Le bearer par défaut fourni une connectivité IP, le débit n'est pas garanti[42].
Porteurs dédié : sont des porteurs établis à n'importe quel moment après la procédure d'enregistrement pour que l'utilisateur puisse profiter de services nécessitant de la QoS spécifique (latence, débit, ...) et sur d'autres PDN. Les valeurs de QoS sont reçues au niveau du P-GW par le PCRF et transférées ensuite au S-GW. Enfin, le MME transfère les valeurs reçues par le S-GW vers le eNodeB[42].

2.2.12.3 Les Paramètres de la qualité de service pour le porteur EPS

Le profil QoS d'un porteur EPS comprend principalement l'identificateur de classe de qualité de service et la priorité d'allocation et de rétention :
La priorité d'allocation ou de rétention (ARP) :les ressources dans le réseau sont limitées, une demande d'établissement ou de modification de porteur peut être rejetée.Également le paramètre ARP est introduit pour faciliter la prise de décision.
• Les niveaux de priorités d'ARP sont utilisés pour assurer que les requêtes émanant d'un porteur de haute priorité sont privilégiées par rapport aux autres demandes provenant des porteurs moins prioritaires, ce qui offre la possibilité au réseau de choisir le porteur de faible priorité à préempter et par conséquent, libérer les ressources nécessaires.
• Le choix du porteur est basé sur la valeur de la capacité de préemption qui définit si un porteur donné est dans la mesure d'être préempté.
• la valeur de la capacité de vulnérabilité définie si un porteur est susceptible d'être préempté y compris le porteur de plus haute priorité ARP. Il convient de noter qu'après la mise en place du porteur, la valeur d'ARP n'a aucun effet sur le traitement de la transmission des paquets.

2.2.12.4 Le QCI (QoS class identifier)

Le QCI est un paramètre défini pour les réseaux LTE/EPC afin de différencier les qualités de services entre les flux de services.
Une fois les porteurs sont établis à l'aide des mécanismes de contrôle d'accès fournis par l'ARP, l'eNodeB doit encore savoir comment traiter les paquets de chaque porteur et allouer les ressources nécessaires. Ainsi, il est indispensable de définir un second paramètre pour accomplir la tâche de gestion de ressources. La technologie LTE définit une valeur scalaire désignée QCI afin de déterminer un groupe de paramètres de la QoS permettant de traiter les paquets à acheminer de chaque porteur. Ce traitement est effectué grâce à l'ordonnancement qui permet d'allouer les ressources radio à chaque porteur[42].

Le 3GPP a identifié neuf QCI dont chacune est caractérisée par :

Le type de ressource allouée (GBR/non-GBR) : Le type de ressource GBR pour un EPS signifie que la bande passante du support est garantie. Un support EPS de type GBR a un "débit binaire" garanti associé. Seul un support EPS dédié peut être un support de type GBR et aucun support EPS par défaut ne peut être de type GBR. Le QCI d'un support EPS de type GBR peut aller de 1 à 4.
Avoir un type de ressource non-GBR pour un support EPS signifie que le support est un support de type best-effort et que sa bande passante n'est pas garantie. Alors qu'un support EPS dédié peut être GBR ou non-GBR. Le QCI d'un support EPS non-GBR peut aller de 5 à 9[5].

La priorité des paquets : Afin d'arbitrer entre les différents modes de paquets, une gestion de priorité est installée soit au sein du réseau, soit à ses extrémités, tout dépend du type de paquet à traiter[5].

La latence : Il existe deux types de latence, la latence du plan de contrôle et la latence du plan usager.
La latence du plan de contrôle correspond au temps nécessaire à un UE de passer d'un état inactif à un état actif, par contre la latence du plan usager est définie par le temps moyen écoulé entre l'envoie d'un élément de données par un utilisateur, une requête ou une page web à charger et le moment ou il reçoit la réponse du serveur.
La latence se mesure donc par le temps d'aller-retour du terminal au serveur ensuite de serveur au terminal.
Le calcul de la latence du plan de contrôle se fait par la sommation des durées des différentes étapes pour qu'un UE passe d'un état de repos à un état actif[5].

Taux d'erreur résiduel (PELR) : désigne le rapport entre le nombre de données traitées au sein du réseau et le nombre de données reçu de l'UE dans la voie descendante, ça s'applique ainsi dans le sens montant. Le taux d'erreur entre l'eNodeB et PGW est négligeable, cependant, le PELR s'applique essentiellement à la partie radio du réseau[5].

Les différents paramètres associés à la classe de la qualité de service sont résumés dans le tableau 2.1.

QCI	Type de ressource	Priorité	Latence	Taux d'erreur résiduel	Exemple d'utilisation
1	GBR	2	100ms	10^{-2}	Voix
2	GBR	4	150ms	10^{-3}	TV,Streaming vidéo
3	GBR	3	50ms	10^{-3}	Jeu interactif
4	GBR	5	300ms	10^{-6}	Vidéo à la demande
5	Non GBR	1	100ms	10^{-6}	Signalisation IMS Vidéo à la demande , service basés sur TCP
6	Non GBR	6	300ms	10^{-6}	Signalisation IMS Vidéo à la demande , service basés sur TCP
7	Non GBR	7	100ms	10^{-6}	Voix ,streaming vidéo, jeu interactif
8	Non GBR	8	300ms	10^{-3}	Porteur EPS pour les abonnés premium
9	Non GBR	9	300ms	10^{-6}	Porteur EPS pour les abonnés non premium

TAB. 2.1: Les paramètres associés aux classes du QCI [5].

2.2.12.5 Objectif de l'étude de QoS

La qualité de service correspond à la manipulation du trafic de sorte qu'un équipement réseau, tel qu'un routeur ou un commutateur, puisse transférer ce trafic conformément aux comportements requis de la part des applications à l'origine de ce trafic. L'objectif des réseaux LTE est de fournir un accès haut débit aux utilisateurs dans une large zone, pour cela, des exigences applicables à ces réseaux ont été définies afin d'optimiser la qualité de service.

La QoS permet à un équipement réseau de différencier le trafic et de lui appliquer différents comportements. Selon les types de service envisagés, la qualité pourra résider.

1. Le débit (téléchargement ou diffusion vidéo) : Efficacité spectrale élevée pour offrir des débits de 300 Mbps pour un accès à faible mobilité.

2. Le délai de transit fiable (pour les applications ou la téléphonie).

3. La disponibilité (accès à un service partagé).

4. Le taux de pertes de paquets[5].

2.2.13 Le canal radio

En communications, le canal de transmission représente toutes les transformations subies par le signal entre l'émetteur et le récepteur, de par sa propagation dans le milieu de transmission, ainsi que dans les équipements d'émission et de réception. Le canal de transmission détermine la manière dont les données doivent être mises en forme à l'émetteur afin de se propager dans de bonnes conditions dans le milieu, ainsi que les traitements à mettre en œuvre au récepteur afin de les détecter correctement. Le canal de transmission est donc d'une importance clé, car il détermine une grande partie de la conception d'un système de communication[5].

2.2.13.1 Qualité de canal radio

La qualité du signal reçu, aussi appelée la qualité du canal, est caractérisée par le rapport signal sur interférence et bruit (Signal to Interference and Noise Ratio, SINR), défini comme suit :

$$SINR = \frac{Puissance\ du\ signal\ utile}{Puissance\ de\ \ l'interfrence + Puissance\ du\ bruit}$$

Dans cette équation, les différentes puissances mises en jeu sont mesurées au niveau symbole, en sortie des divers traitements de réduction d'interférence du récepteur (notamment de l'égaliseur), mais avant le décodage de canal. Le débit pouvant être offert à un UE dépend directement de son SINR. Sous l'hypothèse d'un canal fixe et d'une interférence gaussienne, le débit maximal pouvant être atteint pour un SINR donné est donné par la formule de Shannon, où B est la largeur de bande de la transmission (en Hz) :

$$C(SINR, B) = B \times log_2(1 + SINR) \text{ en (bits/s).}$$

Ce débit maximal est appelé la capacité du canal. La formule précédente est relative à la transmission d'un seul bloc de données. Il existe d'autres formules plus détaillées donnant la capacité du canal pour des scénarios de transmission particuliers, notamment MIMO où plusieurs blocs de données sont transmis sur les mêmes ressources. On pourra trouver ces formules par exemple dans [Tse, Viswanath, 2005]. Dans la pratique, le débit de la transmission est adapté en réglant le type de modulation et de codage de manière à s'approcher au plus près de la capacité du canal, avec une certaine probabilité d'erreur sur le paquet transmis. La formule de Shannon, si elle reste théorique, donne néanmoins les grandes tendances de l'évolution du débit en fonction du SINR.

2.2.13.2 Les indicateurs de qualité de canal du réseau LTE

Pour mesurer la qualité de transmission de canal en liaison descendante (downlink) ; la norme LTE définit trois indicateurs qu'on abordera par la suite .

- Le User Equipment (UE) peut les mesurer tous les trois et les retransmettre en liaison montante (uplink) à la station de base (BS).

- La station de base adapte alors la transmission du signal en downlink.

Les propriétés statistiques du canal doivent rester constantes durant le temps de la notification d'un indicateur qualité à la BS et la transmission modifiée. Cela pour parvenir à une véritable amélioration de la transmission par une modification en downlink .

Channel Quality Indicator (CQI)

Le CQI indique le schéma de modulation le plus élevé possible ainsi que le taux de codage, pour lesquels le taux d'erreur de bloc (BLER) dans le canal ne dépasse pas 100%. Il peut adopter une valeur discrète de 0 à 15. L'indice 0 indique que l'UE n'a reçu aucun signal LTE exploitable et que le canal est par conséquent inutilisable. Il existe de nombreuses possibilités de configuration pour le rapport CQI de l'UE. L'UE peut par exemple envoyer la valeur CQI à la BS via la liaison montante de deux manières différentes :

- Périodique, via les canaux PUCCH ou PUSCH.
- Apériodique via le canal PUSCH (dans ce cas, la BS demande explicitement à l'UE d'envoyer un rapport CQI).

De plus, la résolution du rapport CQI peut varier dans la gamme de fréquence : outre le Wideband-CQI pour la totalité de la largeur de bande de canal, il existe différentes SubbandCQI qui sont en charge de la qualité de transmission dans une sous-gamme de fréquence spécifique.

L'indice CQI, que l'UE transmet à la BS, est dérivé de la qualité du signal en liaison descendante. Contrairement à d'autres systèmes sans fil, comme par exemple en HSDPA, l'indice CQI en LTE ne dépend plus directement du rapport signal/bruit. Il est en effet également influencé par le traitement du signal dans l'UE : dans un canal identique, un UE doté d'un algorithme de traitement du signal plus puissant peut transmettre à la BS un indice CQI plus élevé qu'un UE moins performant.

Pour résumer, on peut dire que la procédure des rapports CQI est une caractéristique fondamentale des réseaux LTE car elle permet l'estimation de la qualité de la voie descendante à la BS. Chaque CQI est calculé comme une mesure quantifiée du SINR (Signal to Interférence Noise Ratio). chaque UE décode les signaux de référence, calcule le CQI, et le renvoie à la BS. Puis La BS utilise les informations CQI pour les décisions d'affectation et remplit un « masque de répartition » de RB [19].

Precoding Matrix Indicator (PMI)

La Precoding Matrix détermine la façon dont les flux de données individuels (désignés dans LTE par « Layer ») sont reproduits sur les antennes. En choisissant judicieusement cette matrice, il est possible de maximiser le nombre de bits de données susceptibles d'être reçus par l'UE via l'ensemble des Layers. Cela nécessite toutefois la connaissance de la qualité du canal pour chaque antenne du downlink que l'UE peut identifier par le biais des mesures. S'il connaît les matrices de précodage admissibles, l'UE peut transmettre un rapport PMI à la BS et proposer une matrice appropriée.

Rank Indicator (RI)

Le rang (« Rank ») du canal indique le nombre de Layers et donc le nombre de flux de signaux différents transmis en liaison descendante. Un seul Layer est utilisé dans une configuration SIMO (Single Input Multiple Output) ou de diversité de transmission alors que deux le sont en MIMO 2×2 (Multiple Input Multiple Output) avec

CQI	Modulation	Rendement du code (approché)	Efficacité spectrale (approchée) (bit/symbole)
0	Hors de portée		
1	QPSK	0.076	0.15
2	QPSK	0.12	0.23
3	QPSK	0.19	0.38
4	QPSK	0.30	0.60
5	QPSK	0.44	0.88
6	QPSK	0.59	1.18
7	16QAM	0.37	1.48
8	16QAM	0.48	1.91
9	16QAM	0.60	2.41
10	64QAM	0.46	2.73
11	64QAM	0.55	3.32
12	64QAM	0.65	3.90
13	64QAM	0.75	4.52
14	64QAM	0.85	5.12
15	64QAM	0.93	5.55

FIG. 2.9: L'efficacité spectrale basée sur la sélection des paramètres la modulation et codage

multiplexage. L'objectif d'un RI optimisé est de maximiser la capacité du canal sur l'ensemble de la bande passante disponible dans la liaison descendante en exploitant chaque rang de canal.
Le RI du LTE n'est pas la seule mesure de l'état du canal. De plus, il est également toujours tenu compte du CQI et du PMI étant donné que la valeur du RI influence aussi les valeurs des matrices de précodage admissibles et du CQI. En revanche, la BS ne peut utiliser que le rapport CQI pour adapter le canal downlink (à condition que le RI ne change pas, comme par exemple en mode SIMO pur). Si la BS n'est tenue ni de réagir au feed-back du UE ni de modifier en conséquence le signal en liaison descendante, cela est cependant souvent utile pour réduire le taux d'erreur et augmenter le débit de données. Mais un feed-back erroné du UE sur l'état du canal peut provoquer exactement l'effet inverse. C'est la raison pour laquelle il faut s'assurer que le UE reproduise correctement l'état du canal via les paramètres CQI, PMI et RI.

2.3 État de l'art sur la détection des anomalies

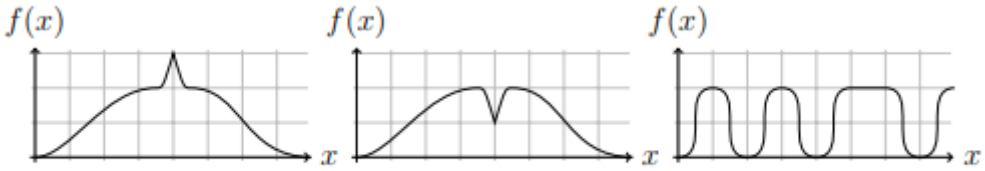
La détection d'anomalies (outliers) est un enjeu ancien et majeur des applications industrielles de la statistique notamment pour la détection d'une défaillance ou défaut de fabrication. Historiquement très présente dans les services de suivi de la qualité par contrôle statistique des procédés. Plusieurs méthodes ont été proposées pour la détection d'anomalies et chaque méthode à ses forces et ses faiblesses[41]. Patcha et Park (2007) ont fait une revue des méthodes utilisées pour la détection d'intrusion. Une revue plus générale des techniques existantes couvrant plusieurs approches est proposée dans Aggarwal (2017) et Chandola et al.(2009). Gupta et al.(2014) fait l'état de l'art des méthodes en fonction du type de données considérées : les données temporelles telles que les séries temporelles, les données spatio-temporelles et les flux de données. Salehi et Rashidi (2018), Souiden et al.(2016), Thakkar et al.(2016), Tellis et D'Souza (2018) présentent également des méthodes applicables aux flux de données[20].

2.3.1 Anomalie

Une anomalie est une observation (portant sur un objet, une donné,..) inattendue au regard d'un ensemble d'autres observations préétablies considérées comme normales.
Plus formellement, dans un ensemble Dcontenant n observations notées x_i, alors X(p A sera considérée comme anormale si elle diffère, par ses caractéristiques, des autres observations, c'est-à-dire de celles contenues dans l'ensemble A par rapport à x_p.
La définition du terme anomalie est spécifique au cas d'usage. La plus courante dans le domaine de la détection est une observation qui se distingue des autres par sa singularité : elle pourrait résulter d'un ensemble de règles différentes par rapport aux autres observations[20].

2.3.1.1 Les types d'anomalies

Il existe trois types d'anomalies. Pour donner une meilleure interprétation à chacune de ces dernières, prenons l'ensemble de départ D contenant n observations X_i avec $0 <$ i $\leq n$.
Anomalie globale : Un objet est considéré comme une anomalie globale s'il dévie fortement des autres objets. Dans notre ensemble de départ D, les observations xi sont des entiers compris entre 1 et 5. Si une nouvelle observation x_{n+1} déroge à la règle en ayant une valeur entière de 12, elle sera identifiée dans ce cas de figure comme une anomalie globale par rapport à l'ensemble A. Son comportement diffère des autres données contenues dans cet ensemble. Un exemple d'anomalie globale est illustré sur la Figure 2.18.a.[20].
Anomalie contextuelle : Une anomalie est contextuelle si elle dépend de la situation où elle se trouve. De manière conditionnelle, on ne peut pas décider sans avoir d'information sur le contexte.
Une anomalie contextuelle se définit à une échelle locale, en tenant compte du contexte de son voisinage et des instances qui l'entourent, si bien qu'à l'échelle globale elle n'est pas caractérisée comme une anomalie ainsi présenté sur la figure 2.18.b Pour ce nouvel exemple, les observations xi de D avec un indice i impair ont une valeur entière strictement négative tel que $x_1 = 3, x_3 = 5$, et les observations avec un indice i pair possèdent des valeurs entières strictement positives tel que $x_2 = 2$, $x_4 = 5$, etc. Si une nouvelle observation à $x_n + 1$ avec n+1 impair a une valeur positive tel que $x_{n+1} = 3$, elle est dans ce cas une anomalie contextuelle. Elle aurait pu ne pas être considérée comme anormale si nous n'avions pas tenu compte de la parité de l'indice i ou si elle avait été observée dans un autre contexte, ici avec un indice pair. Cette observation se distingue des autres par la nature même de son contexte, ici la parité de l'indice. Ce type d'anomalie est donc très spécifique aux données temporelles telles que les séries temporelles ou les flux de données[20].
Anomalie collective : Une observation seule ne peut être une anomalie collective. Ce type d'anomalie est un ensemble de plusieurs observations normales comme présenté sur la figure 2.18.c En analysant de façon collective un ensemble d'observations normales et en tenant compte des liens entre celles-ci, elles reflètent un comportement anormal par rapport à l'ensemble des données[20].



(a) Anomalie globale (b) Anomalie locale (c) Anomalie collective

Fig. 2.10: Les 3 types d'anomalies

[20].

2.3.2 Détection d'anomalies

La détection d'anomalies est définie comme la recherche de structures dans un jeu de données qui ne correspondent pas au comportement attendu. Cette dernière permet d'améliorer la qualité des données par suppression ou remplacement des données anormales.

2.3.3 Les difficultés de détection d'anomalie

La frontière entre ce qui est normal et ce qui ne l'est pas est souvent floue. Cela peut être dû à une dépendance directe au contexte défini localement et non pris en compte lors de l'étude de l', ce qui peut rendre le travail de détection d'autant plus difficile. Si l'anomalie est parfois clairement identifiable, sa détection peut être plus délicate dans d'autres situations. On distingue cinq difficultés, également appelé challenges pour la détection d'anomalie :

. Choix du seuil de décision.

. Identification des anomalies.

. L'évolution de la définition de l'anomalie.

. Données bruitées.

. Généralisation compliquée[33].

2.3.4 Jeux de données

Les jeux de données sont utilisés en machine learning. Ils regroupent un ensemble de données cohérents qui peuvent se présenter sous différents formats (textes, chiffres, images, vidéos etc....). Ils peuvent être représentés sous différents types, que ce soient des tableaux, des graphes, des arbres ou autres. Chaque valeur présente dans un jeu de donné est associée à un attribut et à une observation. On trouve plusieurs jeux de données ayant des caractéristiques déférentes et apportant de nouveaux challenges dans la détection d'anomalies[12].

2.3.4.1 Types de jeux de données

On distingue les catégories suivante :

Attributs discrets ou continus

Un attribut discret possède un ensemble fini ou infini dénombrable de valeurs, qui peuvent être ou non des entiers. Les couleurs, les formes géométriques sont des attributs de description qui peuvent être considérés comme des attributs discrets. Dans la catégorie des ensembles infinis dénombrables, ce sont la plupart du temps des entiers permettant d'identifier un état ou de comptabiliser un effectif. Dans le cas contraire, si un attribut n'est pas dénombrable, alors il est continu. Ces attributs sont généralement représentés par des nombres réelles[20].

Données textuelles

Il existe une large variété de type de données représentées sous forme de texte. Lorsque des données textuelles sont utilisées, elles sont généralement considérées comme des données discrètes. Les données textuelles sont composées de mots, eux-mêmes composés de lettres, et le tout est porteur de sens. La variabilité du type d'information est large, ce qui rend le domaine d'autant plus intéressant qu'il propose de nombreuses méthodes spécifiques aux données discrètes[20].

Données visuelles

Les images occupent une place de plus en plus importante grâce au développement du numérique et des puissances de calculs toujours plus grandes. Des entreprises comme Google stockent de plus en plus de photos pour alimenter leurs bases de données et améliorer leur système de reconnaissance d'images. Les domaines d'application des méthodes d'apprentissage automatique sont nombreux, et peuvent aller du domaine médical à la conduite autonome, en passant par la vidéo-surveillance. Les images peuvent être une représentation de ce que nous connaissons, comme ce que nous voyons et que nous immortalisons à l'aide d'un appareil photo ou d'une caméra. Au format numérique, la photographie est une matrice de pixels (où chaque pixel représente une couleur).

Réseau et graphe

Avec le développement d'Internet, les réseaux de communications sont toujours plus grands. Même au sein d'une entreprise, le réseau interne est parfois très complexe et peut transporter des données sensibles. Les graphes sont des structures de données qui illustrent les relations entre les entités et sont par conséquent universellement applicables. Par exemple, une machine permettant l'émission et la réception de données est décrite comme un nœud, et les arêtes entre les différents nœuds directement interconnectés symbolisent les connexions entre ces machines. Suivant le type d'anomalie que l'on souhaite détecter, l'analyse peut s'effectuer sous différentes granularités.

Données temporelles

Ces données comportent une information complémentaire : le temps, il permet de contextualiser les données, notamment les unes par rapport aux autres. C'est pourquoi, la détection d'anomalies dans les données temporelles va principalement se focaliser sur les anomalies contextuelles, mais aussi détecter des ruptures dans les données, c'est-à-dire des anomalies globales.[20]

Notons l'ensemble D de n données contenant deux attributs, dont l'attribut temporel t :

$$D = (x1, x2, ..., x^n) \text{et } x_i \text{ un tuple noté } (a_i, t_i)$$

sachant $0 < \text{i} \leq n$.

avec $a_i$représentant la valeur observée à l'instant ti . Les données peuvent aussi posséder plusieurs attributs en

plus de l'attribut temporel. Dans ce cas, on notera :

$$x_i = (a_{i,1}, a_{i,2}, .., a_{i,d}, t_i) \text{ avec } x_i \text{ possèdant d attributs.}$$

Les attributs temporels peuvent aussi représenter une durée (par exemple 30 secondes, 1 heure, 3 jours et 12 heures, . . .), un horaire (c'est-à-dire une durée modulo un jour), ou un moment précis appelé horodatage. L'horodatage indique la date et l'heure de l'observation. Dans la représentation matricielle multivariée de l'ensemble D, définissons le vecteur t, tel que pour tout i tel que $0 < i \leq n$ on obtient l'attribut temporel t_i de l'observation xi indiqué à la i ème colonne de la matrice de l'ensemble D :

$$D = \begin{pmatrix} x_{1,1} & x_{2,1} & ... & ... & x_{n,1} \\ x_{1,2} & x_{2,2} & ... & ... & x_{n,2} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{1,d} & x_{2,d} & ... & x_{n,d} & \end{pmatrix}$$

et $t = (t_1, t2, ..., t_n)$.

Nous pouvons constater qu'il existe une séquentialité dans les données permettant de définir pour chaque observation une notion de voisinage temporel avec les autres observations, grâce à cette notion d'ordre intrinsèque aux données temporelles. La problématique de prévision est très proche de la détection d'anomalies temporelles, car les anomalies recherchées à l'état courant sont souvent définies comme une déviation de ce qui est attendu en comparaison à ce qui a été observé précédemment. La temporalité de ces données permet de les lier les unes aux autres en prenant en considération leurs voisinages respectifs. De plus, une pratique assez courante consiste à agréger durant l'analyse deux observations très proches ou égales temporellement, pour ne faire qu'une seule et même observation. Par exemple l'observation de la valeur de $x_{1,1}$ généré à $t_{1,1}$ le 2 juin à 16h l'observation de la valeur de $x_{1,2}$ a été générée à $t_{1,2}$ le 2 juin à 21h. Si nous posons t1 représentant l'attribut temporel du 2 juin alors l'observation multivariée x_1 contient les deux valeurs $x_{1,1}$ et $x_{1,2}$. Mais encore, il est possible de tenir compte de l'attribut a1 de x_1 généré à l'instant t_1 pour l'analyse de la pertinence de l'attribut a2 de x2 généré à l'instant t_2. Par exemple, l'ouverture d'une fenêtre quelques minutes après une augmentation du chauffage dans une même pièce sont deux actions indépendantes et assez normales, mais l'ordre et la proximité temporelle de ces deux actions renvoient à un comportement anormal. C'est aussi le cas pour une suite de commandes système, considérée comme une anomalie collective si une suite de plusieurs commandes s'avère anormale.
comprenant une répétition de tentative de connexion, sera annotée comme anormale. Les données temporelles sont définies sous différentes caractéristiques. Elles peuvent être synchrones ou asynchrones, et dans ce dernier cas le décalage temporel entre différentes sources de données peut compliquer l'analyse. Les données temporelles peuvent aussi être périodiques car elles sont parfois la représentation d'observations répétées à intervalles réguliers (quotidien, mensuel, . . .) et dans le cas contraire, elles sont apériodiques. Certaines méthodes sont

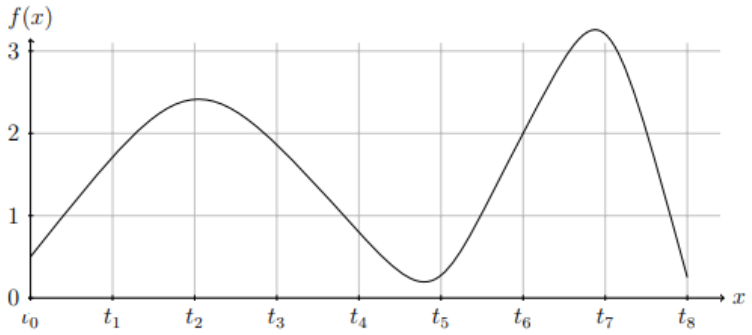
spécialisées dans l'apprentissage de cette caractéristique, telle que ARIMA. Nous pouvons alors analyser les données temporelles point par point, ou bien par séquence de valeurs temporelles appelée "série temporelle". Les données temporelles peuvent être aussi bien univariées (même si la temporalité peut finalement être vue comme une variable particulière) ou multivariées, comme expliqué plus haut dans ce paragraphe. Il existe aussi certains cas où les données sont récoltées en continu, sans interruption, formant ainsi un flux de données. Nous allons à présent nous intéresser à deux catégories de données temporelles, les séries temporelles et les flux de données (communément appelé data stream dans la littérature).[20]

1. **Séries temporelles :** Les séries temporelles sont des séquences de valeurs ordonnées dans le temps.le terme"série temporelle" est utilisé sans qu'il y ait réellement de notion temporelle dans les données brutes elles peuvent représenter différents types de données telles que le cours d'une valeur boursière, la consommation électrique. Les séries temporelles sont généralement vues comme des données uni-variées, dans lesquelles chaque valeur temporelle ti possède un attribut ai représentant le même type de données.[20]
Pour une série temporelle S de longueur fixe w :

$$S = [(a_1, t_1), (a_2, t_2), ..., (a_w, t_w)]$$

Mais les séries temporelles peuvent aussi être multivariées et contenir des types de données différents pour chacun des attributs. Dans ce cas, si plusieurs attributs sont observés durant une même valeur temporelle ti, l'ensemble des attributs D est noté dans un vecteur colonne et la série temporelle S est sous forme d'une matrice d × w :

$$S = \begin{pmatrix} (a_{1,1}, t_1) & (a_{2,1}, t_2) & ... & ... & (a_{w,1}, t_w) \\ (a_{1,2}, t_1) & (a_{2,2}, t_2) & ... & ... & (a_{w,d}, t_w) \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ (a1, d, t1) & (a2, d, t2) & ... & ... & (a_{w,d}, t_w) \end{pmatrix}$$



(a)

Temps	t1	t2	t3	t4	t5	t6	t7	t8
Couleurs	Bleu	Vert	Rouge	Rouge	Bleu	Rouge	Vert	Bleu

(b)

Figure 2.19 : Un exemple de série temporelle multivariée de longueur 8.La figure (a) représente la courbe des valeurs du premier attribut, et la figure (b) représente les valeurs du second attribut. Par exemple, la série temporelle multivariée à l'instant t1 retourne les valeurs 38 et "bleu".[20]

Lors de l'analyse classique de plusieurs séries temporelles, ces dernières possèdent la même temporalité, c'est-à-dire qu'elles possèdent les mêmes attributs temporels, et contiennent le même nombre de valeur. En règle générale, les méthodes d'analyse et de détection utilisent des mesures de distance entre les séries temporelles d'une même base de séries. Ces méthodes permettent par exemple de les trier par similarité, ou retourner celles qui sont les plus aberrantes. Si les séries temporelles à comparer ont des longueurs différentes, et donc un nombre d'observations différent, il est envisageable d'effectuer un pré traitement sur celles-ci.
Durant la comparaison de séries temporelles dont les créneaux temporels se chevauchent mais qui ne sont pas identiques, il est envisageable de confronter entre elles les sous-séquences qui ont les mêmes valeurs temporelles, pour éviter la méthode de force brute, la plus coûteuse, qui consiste à comparer les valeurs deux à deux. Cependant, la distance obtenue est relative au nombre d'observations temporellement identiques entre deux séquences, et est difficilement comparable avec les autres distances obtenues entre les autres séries temporelles dont la taille d'intersection temporelle des observations est différente.[20]

2. **Flux de données :** Un flux de données consiste en une série d'éléments de données ordonnés dans le temps. Les données représentent un « événement » ou un changement d'état qui s'est produit dans l'entreprise et qu'il est utile pour l'entreprise de connaître et d'analyser, souvent en temps réel.
Dans le domaine de télécommunications, on parle de « Data Flow » ou « Data Stream ». Ce terme flux est désigné par « stream ». À l'origine, les flux étaient considérés comme des canaux.[20]
Notre étude se portera sur cette sous catégorie de données temporelles.
Nous donnons plus de détails prochainement.

2.3.5 Relation entre KPI et détection des anomalies

Les KPI sont généralement utilisés pour surveiller et optimiser les performances du réseau cellulaire. Par conséquent, les KPI peuvent également être bien adaptés aux tâches de détection d'anomalies. KPI métriques telles que CQI, puissance reçue du signal de référence (RSRP), signal de référence la qualité reçue (RSRQ), CQI et SINR sont tous des candidats qui peuvent fournir des indices et des informations destinées à la détection d'anomalies et au suivi de la salubrité du réseau. Parmi celles-ci, les algorithmes de détection qu'on abordera se cantonneront à deux métriques, CQI et SINR, pour leur coût de calcul relativement faible. CQI transporte des informations sur la qualité du canal de la liaison de communication et SINR est essentiellement les rapport

signal/interférence plus bruit mesuré sur un UE. Les métriques KPI de chaque UE sont signalées à l'OSS via les stations de base. Dans l'OSS, CQI et les valeurs SINR de chaque UE doivent être surveillées pendant un certain intervalle de temps.

2.3.6 Techniques de détection des anomalies

2.3.6.1 Techniques Basés sur les statistiques et la théorie des probabilités

Le modèle de comportement normal d'un système est décrit par des statistiques ou un modèle de probabilité représentatif des données collectées sur le système et le décrivant. Tiresias est un système de prédiction de défaillance en ligne pour des systèmes distribués. Il fonctionne selon une approche boîte noire (c'est-à-dire en ne connaissant que les spécifications fonctionnelles d'un système). Il suppose qu'une défaillance est précédée d'un comportement instable du système.

1. Basées sur les statistiques

Tiresias détecte des comportements instables dans les observations de monitoring du système en construisant des séries temporelles analysées grâce à des seuils de détection et par la technique de DFT (Dispersion Frame Technique). Des séries temporelles de performances système sont de plus analysées dans par le biais de relations calculées deux à deux entre chaque métrique d'une observation. Une anomalie est détectée lorsque certaines de ces relations sont cassées (i.e. elles ne sont plus constatées à un instant donné) et localisée en analysant les métriques dont les relations sont cassées. Ces relations sont définies par un modèle auto-régressif (dans lequel la valeur d'une série temporelle est expliquée par ses valeurs passées) avec des variables exogènes supposant que les métriques sont indépendantes entre elles[31].

Ces techniques peuvent être paramétriques ou non paramétriques :

Paramétriques

L'approche paramétrique suppose une connaissance a priori de la distribution des données. Les méthodes statistiques construisent un modèle avec un intervalle de confiance à partir des données existantes. Les nouvelles données qui ne correspondent pas à ce modèle seront considérées anormales (Desforges et al. (1998) ; Aggarwal (2017))[31].

Non paramétriques

Cette approche ne suppose généralement pas la connaissance de la distribution sous-jacente et basée également sur la construction du modèle de distribution[31].

2. Basées sur les probabilités

Les travaux dans opèrent une détection d'anomalies en utilisant deux distributions de probabilité sur des traces d'appels système UNIX afin d'évaluer par des tests statistiques si une suite de symboles contenus dans une trace correspond à une anomalie ou non. Des symptômes d'anomalies sont détectés de manière probabiliste dans les travaux de [Shen 2009]. Pour cela, des chutes de performance du système étudié sont prédites en comparant l'exécution d'un système à celle d'un autre système servant de modèle (un benchmark par exemple). L'exécution d'un système est définie sur la base de mesures telles que le débit de réponse à des requêtes ou le

nombre d'entrées/sorties concurrentes dans un système.

2.3.6.2 Techniques basées sur l'approximation

Elles regroupent celles basées sur les plus proches voisins et celles basées sur le clustering.

1. Basées sur le plus proche voisin

Déterminent pour une observation o ses k plus proches voisins à travers le calcul de la distance entre toutes les observations du jeu de données. Ces méthodes nécessitent un calcul préalable, et de ce fait, elles sont coûteuses en temps d'exécution. Il existe deux approches de méthodes basées sur les plus proches voisins : l'approche basée sur la distance (Angiulli et Pizzuti (2002) ; Yamanishi et al. (2004)) et l'approche basée sur la densité (Breunig et al. (2000))[50].

2. Clustering

Elles ont pour objectif principal de diviser le jeu de données en clusters contenant les données qui ont des comportements similaires. On distingue deux approches dans ces techniques : l'approche basée sur la distance selon laquelle le cluster le plus éloigné représente une anomalie et l'approche basée sur la densité définie l'anomalie par le cluster qui contient le moins de données.

La but des algorithmes de clustering est de donner un sens aux données et d'extraire de la valeur à partir de grandes quantités de données structurées et non structurées. Ces algorithmes vous permettent de séparer les données en fonction de leurs propriétés ou fonctionnalités et de les regrouper dans différents clusters en fonction de leurs similitudes. Et ils ont plusieurs utilisations dans différents secteurs.

2.3.6.3 Techniques basées sur le deep learning

Représentent une classe d'algorithmes d'apprentissage automatique supervisé ou non supervisé basés sur l'utilisation de plusieurs couches d'unité de traitement non linéaire. Parmi ces méthodes on cite les auto-encoders (AE) et One-Class Neural Networks (OCNN) (Chalapathy et Chawla (2019))[50].

2.3.6.4 Techniques basées les supports vecteurs machines

Souvent traduit par l'appellation de Séparateur à Vaste Marge (SVM) sont une classe d'algorithmes d'apprentissage initialement définis pour la discrimination c'est-à-dire la prévision d'une variable qualitative binaire. Ils ont été ensuite généralisés à la prévision d'une variable quantitative. Dans le cas de la discrimination d'une variable dichotomique, ils sont basés sur la recherche de l'hyperplan de marge optimale qui, lorsque c'est possible, classe ou sépare correctement les données tout en étant le plus éloigné possible de toutes les observations. Le principe est donc de trouver un classifieur, ou une fonction de discrimination, dont la capacité de généralisation (qualité de prévision) est la plus grande possible. Cette approche découle directement des travaux de Vapnik en théorie de l'apprentissage à partir de 1995. Elle s'est focalisée sur les propriétés de généralisation (ou prévision) d'un modèle en contrôlant sa complexité.

Le principe fondateur des SVM est justement d'intégrer à l'estimation le contrôle de la complexité c'est-à-dire le nombre de paramètres qui est associé dans ce cas au nombre de vecteurs supports.

Il existe d'autre techniques que nous allons aborder dans le prochain chapitre de Machine Learning.

2.3.7 La problématique

La supervision des réseaux de télécommunication et plus particulièrement la détection d'anomalies représente un aspect important de la qualité de service. Une anomalie peut être définie comme une observation inattendue au regard d'un ensemble d'autres observation préétablie considérés comme normale, la détection de ces anomalies est un défi que nous trouverons dans de nombreux domaines d'applications, ainsi en recherche opérationnelle. Ce qui fait varier les problématiques et les types de données à traiter et influence fortement sur le choix de technique à utiliser. L'objectif de ce mémoire, est donc de repérer un comportement anormale des KPIs. Les obstacles résident dans la volumétrie des données à analyser et le fait qu'elle doit se faire au plus tôt c'est-à-dire en temps réel.
Comme nous l'avons déjà mentionné durant le deuxième chapitre, ces flux de données se différencient des autres types de données par le fait qu'ils n'ont pas de longueur fixe, la spécifié des flux de données est qu'ils arrivent en continu sans durées déterminé. Il n'y a pas de fin de flux, ni de temps strictement défini entre l'arrivée de deux données. Chaque donnée contient un ou plusieurs attributs.
La détection des anomalies dans les flux de données de réseaux LTE est basée sur la surveillance de l'ensemble des indicateurs clés de performances. Elle peut se faire avec des techniques de Machine Learning qui offrent des aperçus centrés sur les données du problème.
Cependant, Quelle est la meilleure approche à proposer ? Sera-t-elle efficace dans notre étude ? Quel est l'algorithme de Machine Learning le plus fiable à notre problématique ?

2.4 Conclusion

Nous avons vu dans ce chapitre, les notions sur les réseaux de télécommunication et son évolution durant ces dernières années. Nous avons présenté l'architecture de réseau LTE, plus particulièrement le réseau d'accès radio qui est la partie la plus importante des réseaux de télécommunication. Le LTE nommé 4G est une solution adopté par le groupe 3GPP pour résoudre les problèmes de limite de capacité du réseau et de zone de mauvaise couverture afin de pouvoir bien suivre l'amélioration de cette technologie, nous sommes passé à l'étape de détection des anomalies et préciser les types de ce derniers ainsi les causes et citer les différentes technique de détections et pour compléter le travail qui nous a été confier. Nous avons terminé ce chapitre en mentionnant quelques méthodes de détection d'anomalies appliquaient auparavant.
Dans ce qui suit, nous allons nous concentrer sur la détection des anomalies avec les méthodes de machine Learning en vue d'élargir nos compétences dans ce domaine d'intelligence artificielle.

Chapitre 3

Machine Learning

3.1 Introduction

Le Machine Learning ou l'Apprentissage Automatique appartient au champ de l'Intelligence Artificielle (IA), son but est de permettre à une machine d'apprendre de façon automatique à partir d'un jeu de données, pour ensuite réaliser une tâche sans avoir été explicitement programmée à cet effet.
Les algorithmes d'apprentissage reposent sur des modèles qui peuvent être de nature différente. Chaque algorithme d'apprentissage a ses propres spécificités et est plus ou moins efficace selon la nature des tâches qu'il doit accomplir. Récemment, le Deep Learning, qui repose sur un modèle de réseaux de neurones, a permis de nombreuses avancées dans le domaine de l'apprentissage automatique. Cette évolution est une des causes principales des progrès attribués à l'IA ces dernières années. Au-delà des types de modèles utilisés, Il existe divers modes d'apprentissage en fonction des données dont on dispose pour entraîner l'intelligence artificielle et de la réponse souhaitée ainsi que des usages envisagé.
Ils permettent aux ordinateurs de s'entraîner sur les entrées de données et d'utiliser l'analyse statistique afin de générer des valeurs qui se situent dans une plage spécifique. Pour cette raison, l'Apprentissage Automatique permet aux ordinateurs de créer des modèles à partir d'échantillons de données afin d'automatiser les processus de prise de décision basés sur les entrées de données. Le Machine Learning est au cœur des Data Sciences et s'applique à une multitude de domaines , et aujourd'hui nous pouvons dire que tout utilisateur de technologie bénéficie de l'Apprentissage Automatique ; la technologie de reconnaissance faciale permet aux plateformes de médias sociaux d'aider les utilisateurs à marquer et à partager des photos d'amis, la technologie de reconnaissance optique de caractères (OCR) convertit les images de texte en caractères mobiles, Les moteurs de recommandation alimentés par l'Apprentissage Automatique suggèrent quels films ou émissions de télévision à regarder ensuite en fonction des préférences de l'utilisateur, Les voitures autonomes qui reposent sur l'Apprentissage Automatique pour naviguer pourraient bientôt être disponibles pour les consommateurs...etc.
Dans ce chapitre sur le Machine Learning, nous allons présenter les concepts fondamentaux de la Data Science et son processus, du processus KDD (Knowldge Discovery in Databases), et de Machine Learning ainsi que ses

domaines d'application, les types d'algorithmes appliqués et les techniques de validation existantes.

3.2 L'intelligence artificielle (IA)

En termes simples, l'intelligence artificielle fait référence à des systèmes ou des machines qui imitent l'intelligence humaine pour effectuer des tâches et qui peuvent s'améliorer en fonction des informations collectées grâce à l'itération. IA a couvert plusieurs domaines ,elle a accompli un travail remarquable dans le domaine de télécommunications. Cette derrière a optimisée leurs investissements, et elle a réduit les coûts et l'efficacité de l'exploitation et de la maintenance...etc[2].

3.3 Data Science

3.3.1 Définition

Le terme data Science est apparu en 2002 avec la publication du Data Science journal, crée par l'international Council for science (commitee on Data for science and Technology). La Data Science est une science interdisciplinaire s'appuyant sur des méthodes scientifiques, des algorithmes, des processus et autres systèmes afin d'exploiter des grands ensembles de données. La Data Science fait appel aux mathématiques, aux statistiques et à l'informatique et intègre des techniques telles que l'Apprentissage Automatique, l'analyse topologique, elle provient aussi du croisement des domaines de l'extraction de données appelé aussi forage de données(DataMining). En 2008, le titre de data scientist a fait son apparition. La mission principale du data scientist est d'élaborer des stratégies d'analyse de données, mais également de préparer ses données pour leur analyse, puis d'explorer et analyser ces informations. Le data scientist doit ensuite créer des modèles avec ces données, en s'appuyant sur des langages de programmation afin de déployer ces modèles dans des applications[13].

3.3.2 Processus de la Science des Données

En Data Science, il existe une manière de procéder utiliser par les data scientist. Cette dernière est essentiellement agile et itérative permettant de fournir des solutions d'analyse prédictive et des applications intelligente. Elle procède par un raisonnement inductif qui consiste à partir des données de produire de la connaissance. L'approche se construit par étapes en posant d'abord des hypothèses, puis en validant ces hypothèses à l'aide d'algorithmes statistiques et/ou Machine Learning[29].
Le processus de Science des Données passe par six étapes capitales :
La collecte de données : est la première étape du traitement. Il est important que les sources de données disponibles soient fiables et correctement structurées pour que les données importées soient de la meilleure qualité possible[29].
La préparation de données : Après la collecte des données, suit la préparation des données appelée « prétraitement », dans cette étape les données brutes sont filtrées, nettoyées et structurées, en vue de l'étape de

traitement des données[29].

L'importation de données : Les données propres générées de l'étape précédente sont ensuite importées dans leur emplacement de destination et converties vers un format supporté par cette destination. L'importation des données est la première étape au cours de laquelle les données commencent à se transformer en informations exploitables[29].

L'exploration et analyse de données : Une fois que les données sont prêtes à être utilisées, et avant de se lancer dans l'Intelligence Artificielle et l'Apprentissage Automatique, nous devons examiner ces données. Nous devons d'abord inspecter les données et leurs propriétés. Ensuite, l'étape suivante consiste à calculer des statistiques descriptives pour extraire des fonctionnalités et tester des variables significatives, le test de ce dernier est souvent effectué avec la corrélation[29].

Enfin, nous utiliserons la visualisation des données pour nous aider à identifier des modèles et des tendances significatives dans nos données. Nous pouvons obtenir une meilleure image grâce à des graphiques simples comme des graphiques linéaires ou des graphiques à barres pour nous aider à comprendre l'importance des données[29].

Construction de modèles de données : Les étapes de nettoyage et d'exploration sont cruciales pour créer des modèles utiles. Dans cette étape, le processus de création de modèle proprement dit démarre. Ici, le Scientifique des Données distribue les données en un ensemble de formation et un ensemble de test. Des techniques telles que l'association, la classification et le regroupement sont appliquées à l'ensemble de données d'apprentissage. Le modèle une fois préparé est testé par rapport à l'ensemble de données de test[29].

Communication des résultats de l'analyse : À ce stade, les principaux résultats sont communiqués à toutes les parties prenantes. Cela va aider à décider si les résultats du projet sont un succès ou un échec en fonction des entrées du modèle[29].

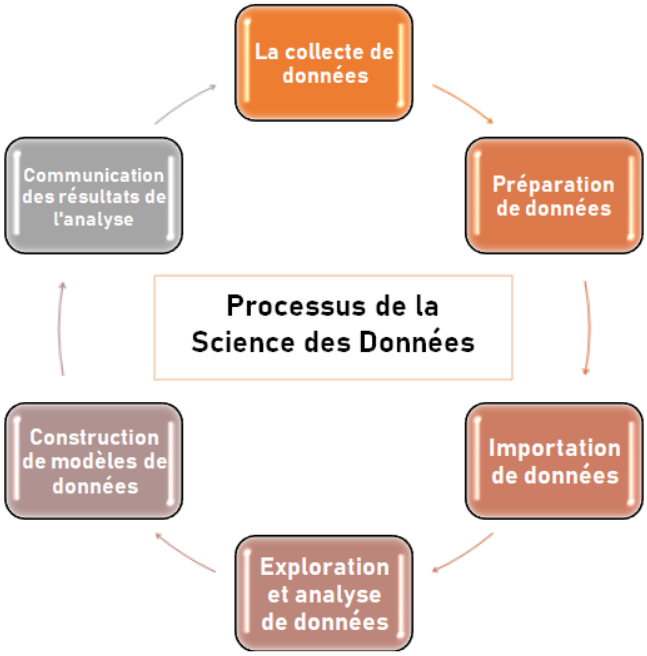


FIG. 3.1: Schéma du processus de la Science des Données[29].

3.3.3 les différentes technologies de science des données

Les praticiens de la science des données travaillent avec des technologies complexes telles que les suivantes :

1. **Intelligence artificielle** : des modèles de machine learning et les logiciels associés sont utilisés pour l'analyse prédictive et prescriptive.

2. **Cloud computing** : les technologies cloud ont doté les scientifiques des données de la flexibilité et de la puissance de traitement nécessaires à l'analytique des données avancée.

3. **Internet des objets** : l'IoT fait référence à divers appareils qui peuvent se connecter automatiquement à Internet. Ces appareils collectent des données pour les initiatives de science des données. Ils génèrent des données massives pouvant être utilisées pour l'exploration et l'extraction de données.

4. **Informatique quantique** : les ordinateurs quantiques peuvent effectuer des calculs complexes à haut débit. Les scientifiques des données qualifiés les utilisent pour créer des algorithmes quantitatifs complexes.

3.4 knowledge Discovery Database (KDD)

3.4.1 Définition

L'approche moderne d'extraction de connaissances à partir des Données (KDD) est un domaine de recherche récent par rapport à ses domaines parents, elle se définit comme « l'acquisition de connaissances nouvelles et l'extraction non triviale de données implicites, intelligibles et potentiellement utiles à partir de faits cachés au

sein de grandes quantités de données ». Elle est caractérisée par le fait qu'elle extrait les connaissances les plus pertinentes et intelligibles possibles à l'usager. Elles doivent être validées, mises en forme et agencées. Pour ce faire le KDD ne se limite pas aux outils d'analyse les plus récents et incorpore explicitement des méthodes pour la préparation des données, pour l'analyse et pour la validation des connaissances produites, ces méthodes proviennent en majorité de la statistique, de l'analyse des données, de l'Apprentissage Automatique et de la reconnaissance de formes, nous allons détailler toutes ces notions et les situer dans le processus général de KDD. Enfin, l'usage des nouvelles connaissances découvertes peut être une contribution dans la réalisation des systèmes experts en concevant des programmes informatiques capables d'apprendre et de découvrir leurs propres connaissances. Cependant, une confusion subsiste encore entre le Data Mining, que nous appelons en français « Fouille de Données », le Data Mining (DM) est l'un des maillons de la chaîne de traitement pour la découverte des connaissances à partir des données, nous pourrions dire que KDD est un véhicule dont le Data Mining est le moteur, La différenciation entre ces deux désignations réside dans le type d'approche utilisée : l'Intelligence Artificielle pour KDD avec utilisation d'heuristiques provenant de l'apprentissage symbolique, statistique pour le DM considéré comme une industrialisation des techniques d'analyse des données.
Pour certains auteurs les outils de Data Mining se résument aux Réseaux de Neurones et aux Arbres de Décision autorisant la prédiction d'une variable qualitative (Arbres de Classification) ou quantitative (Arbre de Régression). Quoi qu'il en soit, le KDD et le Data Mining ont en commun l'utilisation de méga bases ou entrepôts de données. Le Machine Learning (ML) et le KDD ont un lien très fort, ils reconnaissent tous les deux l'importance de l'induction comme mode de pensée normale, parmi les méthodes de l'induction : la régression, les réseaux de neurones simples ou multicouches ou les graphes d'induction, tandis que d'autres domaines scientifiques hésitent à l'accepter, c'est le moins que nous puissions dire.

3.4.2 Le processus KDD

Le processus d'Extraction de Connaissances dans les Bases de Données (KDD) est un processus itératif et interactif complexe partiellement automatique, où l'interaction de l'homme est primordiale. Il est itératif, dans le sens où l'utilisateur peut à tout moment revenir à l'une des étapes. Le processus du KDD est défini comme un processus non trivial qui construit un modèle valide, nouveau, potentiellement utile et au final compréhensible, à partir de données. Il s'effectue sur cinq étapes, détaillons dans les points suivants, les cinq importantes phases de ce processus, à savoir la sélection, le prétraitement (Data preparation), la transformation, la fouille de données (Data Mining) qui est l'étape centrale de KDD et l'évaluation. Par la suite nous donnons des détails sur les neuf étapes de schémas de ce processus.[28]

La sélection :L'approche générale de KDD recommande comme est de coutume dans la conception des systèmes d'information de commencer le projet par une étude préalable avec une identification claire des objectifs. L'étape de sélection vise ainsi à cibler, même de façon grossière, l'espace des données qui va être exploré, l'analyste agit ainsi un peu à l'image du géologue qui définit des zones de prospection, étant persuadé que certaines régions seront probablement vite abandonnées car elles ne recèlent aucun ou peu de minerais. Ce processus n'est pas linéaire car il arrive aussi que le nous revenions, après analyse, rechercher de nouvelles données. Cette

phase peut passer par les moteurs de requêtes des Bases de Données comme le langage SQL, la sélection peut aussi se faire à travers des outils de requêtes plus spécifiques aux données non structurées comme les données textuelles, les images ou le web, faisant pour cela appel à des moteurs de recherche d'informations et d'images auxquelles ils accèdent par le contenu. La phase de sélection sert généralement à nettoyer les données qui sont rapatriées, Par exemple, si l'un des attributs retenus s'avère au moment du rapatriement peu ou mal renseigné, nous pouvons le laisser tomber tout de suite, nous pouvons également explicitement chercher à limiter le nombre d'enregistrements que nous souhaitons traiter.[28]

Le prétraitement (préparation du data) : Le prétraitement est un acte de modélisation d'expert il consiste à s'intéresser à l'examen de la qualité des données collectées, ces données seront traités pour faire face à des problèmes courants tels que les doublons, les erreurs de saisie, l'intégrité de données et le problème des valeurs manquantes. De nombreuses solutions sont proposées, comme le remplacement dans le cas des données numériques continues de toute donnée manquante par le mode de la distribution statistique (la valeur la plus fréquente) de l'attribut concerné, si ce mode existe, nous pouvons également cher- cher à estimer ces valeurs manquantes par des méthodes d'induction comme la régression, les réseaux de neurones simples ou multicouches, ou les graphes d'induction.

Pour le traitement des données aberrantes, il faut d'abord repérer ces dernières au moyen d'une règle préétablie, par exemple, toutes les données numériques dont la valeur sur un attribut donné s'écarte de la valeur moyenne plus deux fois l'écart-type, pourraient être considérées comme des données possiblement aberrantes et qu'il conviendrait à traiter.[28]

La fouille de données (Data Mining) : La fouille de données concerne le Data Mining dans son sens restreint, elle constitue véritablement le cœur du processus KDD, elle est souvent difficile à mettre en œuvre et coûteuse, cette étape fait appel principalement à des multiples méthodes de l'Intelligence Artificielle, de la statistique et de l'analyse de données. Elle se décompose en trois catégories de méthodes :

- Les méthodes de visualisation et de description.
- Les méthodes de classification et de structuration.
- Les méthodes d'explication et de prédiction.

L'objectif de la mise en œuvre des méthodes de fouille de données est de découvrir ce que contiennent les données comme informations ou modèles utiles, chacune de ces familles de méthodes comporte plusieurs techniques appropriées aux différents types de tableaux de don- nées ; certaines sont mieux adaptées à des données numériques continues alors que d'autres sont plus généralement dédiées aux traitements de tableaux de données qualitatives. Le problème du Data Mining réside dans le choix de la méthode adéquate à un problème donné, le choix se fera en fonction de la tâche à résoudre, de la nature des données ou encore de l'environnement de l'entreprise, de plus il est souhaitable de combiner différentes techniques afin de les comparer et d'en retenir une ou plusieurs combinées.[28]

L'évaluation : L'étape d'évaluation ou de validation consiste à valider les modèles extraits, à les rendre intelligibles. Ainsi, l'utilisateur décidera d'appliquer ou non le modèle de prédiction en connaissance des risques qu'il prend. Deux techniques de validation sont utilisées :

Une qui est fondée sur des mesures statistiques utilisant des méthodes de base de statistique descriptive, son objectif est d'obtenir des informations qui permettront de juger le résultat obtenu, ou d'estimer la qualité des données d'apprentissage. Cette validation est obtenue par :

- Le calcul des moyennes et variances des attributs.
- Le calcul de la corrélation entre certain champs.
- La détermination de la classe majoritaire dans le cas de la classification.

Alors que la deuxième technique est par expertise, elle est réalisée par un expert du domaine qui jugera la pertinence des résultats produits. Par exemple, pour les problèmes de segmentation la validation est essentiellement du ressort de l'expert, qui juge de la qualité et la pertinence des groupes constitués par le système. Pour certains domaines d'application (le diagnostic médical par exemple), le modèle présenté doit être compréhensible, une première validation doit être effectuée par un expert qui juge la compréhensibilité du modèle, cette validation peut être éventuellement accompagnée par une technique statistique, la matrice de confusion et la validation croisée sont d'autres techniques de validation couramment utilisées.

La figure 3.2 représente l'enchaînement des étapes du processus d'extraction de connaissances (Processus du KDD) telle que l'entrée de chaque étape est la sortie de la précédente), de manière itérative (les analystes appliquent des boucles de rétroaction si nécessaire) et interactive (les auteurs de ce modèle ont déclaré sa mode itérative, mais ils n'ont donné aucun détail précis) . Nous avons définie chaque étape présentée sur la figure :

1. **Développer et comprendre le domaine d'application (Understanding goal) :** acquérir des connaissances préalables pertinentes, identifier les objectifs de l'utilisateur final (entrée : problème à résoudre/notre objectif, sortie : compréhension du problème/domaine/objectif)[43].

2. **Création d'un ensemble de données cible (Selection) :** sélection (interrogation) de l'ensemble de données, identification des variables du sous-ensemble (attributs de données) et création d'échantillons de données pour le KDP (sortie : données cibles/ensemble de données)[43].

3. **Nettoyage et prétraitement des données (Data cleaning and preprocessing) :** traiter les valeurs aberrantes et la suppression du bruit, gérer les données manquantes, collecter des données sur des séquences temporelles et identifier les modifications connues des données (sortie : données prétraitées)[43].

4. **Réduction et projection des données (Transformation) :**il s'agit de trouver des fonctionnalités utiles qui représentent les données (selon l'objectif), y compris les réductions et les transformations de dimension (sortie : données transformées)[43].

5. **Sélection de la tâche d'exploration de données (Data minig task selection) :** la décision sur les méthodes à appliquer pour la classification , le regroupement, la régression ou une autre tâche (résultat : méthode(s) sélectionnée(s))[43].

6. **Sélection d'algorithmes d'exploration de données (Data minig algorithm selection) :** sélectionner la méthode de recherche de modèles, décider des modèles appropriés et de leurs paramètres, et faire correspondre les méthodes avec l'objectif du processus (résultat : algorithmes sélectionnés)[43].

7. **Exploration de données (Data mining) :** recherche de modèles d'intérêt sous une forme spécifique comme des règles de classification, des arbres de décision, des modèles de régression, des tendances, des clusters

et des associations (sortie : modèles)[43].

8. **Interprétation des modèles extraits (Interpretation) :** compréhension et visualisations des modèles basés sur les modèles extraits (sortie : modèles interprétés)[43].

9. **Consolidation des connaissances découvertes (Consolidation) :**utilisation des modèles découverts dans un système analysé par le processus KDD, documentation et communication des connaissances aux utilisateurs finaux, vérification et résolution des conflits si nécessaire (résultats : connaissances, actions/décisions basées sur les résultats [43].

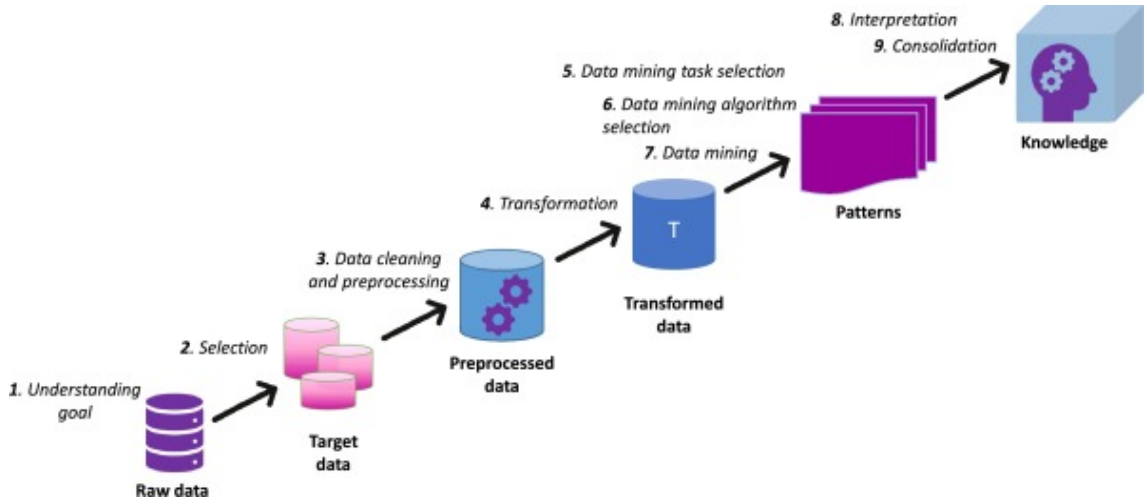


FIG. 3.2: Les étapes du processus KDD

3.5 Machine Learning

3.5.1 Définition

L'Apprentissage automatique (machine learning) est une branche de l'Intelligence Artificielle (IA) et de l'informatique qui utilise principalement des données et des algorithmes pour imiter la manière dont les être humains apprennent, en améliorant progressivement sa précision. Cette technologie donc sert à donner à une machine la capacité d'apprendre sans la programmer d'une façon explicite.

Le machine learning a été appliquée dans divers domaines comme : l'industrie, la finance, l'agriculture, l'énergie, les médias, les télécommunications ...etc. Le Machine Learning ou l'Apprentissage Automatique consiste à laisser des algorithmes auto-découvrir des 'patterns' à savoir 'des motifs récurrents' dans des ensembles de données c'est à dire, lui faire entrer un jeu de données afin qu'il puisse s'entraîner et s'améliorer d'où le mot apprentissage, et cela pour accomplir la tâche qui lui est demandée (prédiction, identification...etc). Pour résoudre les problèmes avec le Machine Learning, il faut :

Premièrement, disposer des données nécessaires car ils constituent littéralement le nerf de notre machine.

Plus nous maîtrisons et nous avons une bonne compréhension des données, plus nous serons en mesure de les

utiliser facilement et sûrement lors de l'entraînement des algorithmes.

Les types de données rencontrées en Machine Learning sont principalement :

- Les Bases de Données : c'est la source principale de récupération de données.
- Les données brutes : il s'agit de données sous leur forme source, sans préparation préalable pour le ML.
- Le texte : tous les types de texte (texte libre écrit à la main, livres, messages, etc.).
- Les images et les vidéos.
- IOT (Internet of Things).

Deuxièmement, préciser la tâche à accomplir ; qui est une tâche spécifique qui correspond au problème que nous cherchons à résoudre. Chaque tâche est traduite de manière différente et nécessitera un bon choix d'algorithmes.

Troisièmement, choisir des algorithmes en fonction du type de tâche que nous souhaitons accomplir et du type de données dont nous disposons. C'est la façon dont nous allons paramétrer notre modèle statistique en utilisant des jeux de données, le choix de l'algorithme change en fonction de la tâche à accomplir et des données dont nous disposons.

En revanche, il faut déterminer une mesure principale (l'analyse d'erreur) spécifique à la tâche à accomplir. Le choix de cette mesure est très important pour être sûr de mesurer correctement la pertinence et la qualité des algorithmes utilisés.

Un algorithme d'Apprentissage Automatique est un processus de calcul qui utilise des données d'entrée pour réaliser une tâche souhaitée sans être littéralement programmée pour produire un résultat particulier, ces algorithmes sont en un sens « codés en douceur », en ce sens ils modifient ou adaptent automatiquement leur architecture par la répétition (l'expérience) afin de devenir de mieux en mieux pour accomplir la tâche souhaitée. Le processus d'adaptation est appelé « Apprentissage », dans lequel des échantillons de données d'entrée sont fournis ainsi que les résultats souhaités. L'algorithme se configure alors de manière optimale pour qu'il puisse non seulement produire le résultat souhaité lorsqu'il est présenté avec les intrants de formation, mais peut généraliser pour produire le résultat souhaité à partir de nouvelles données. Cette formation est la partie « Apprentissage » de l'Apprentissage Automatique.[29]

Il existe trois catégories de Machine Learning : Machine Learning avec supervision, Machine Learning sans supervision et le Machine Learning par renforcement (Deep Learning).[27]

3.5.2 Machine Learning Supervisé

3.5.2.1 Définition

L'apprentissage supervisé est une méthode d'apprentissage automatique dans laquelle les modèles sont entraînés à l'aide de données étiquetées. Dans l'apprentissage supervisé, les modèles doivent trouver la fonction de mappage pour mapper la variable d'entrée (X) avec la variable de sortie (Y) avec :

$$Y=f(X)$$

3.5.2.2 Prétraitement

Avant de pouvoir entraîner les modèles, un prétraitement des données est nécessaire afin d'éviter de diminuer les performances des modèles[3]. En générale, le prétraitement des données signifie les opérations de sélection, nettoyage, transformation et la sélection des facteurs les plus significatifs appliqués aux données brutes avant leur traitement, ce prétraitement a un rôle primordial dans le processus de sciences de données, il est nécessaire afin d'éviter de ruser les performances des modèles, il facilite la détection des erreurs avant le traitement des données, il permet d'obtenir des données de qualité ce qui permet de les traiter facilement ce qui conduit à prendre des décisions plus étudiées. Le prétraitement peut être réalisé à l'aide de plusieurs techniques de sélection des facteurs les plus pertinents dans une étude, l'une des techniques les plus intéressantes sont la Matrice de Corrélation, la méthode Stepwise et les méthodes Wrapper.

Matrice de Corrélation : La Matrice de Corrélation est simplement un tableau qui affiche les coefficients de corrélation pour différentes variables[15]. la Matrice de Corrélation se compose de lignes et de colonnes qui affichent les variables, Chaque cellule de la matrice contient le coefficient de corrélation entre deux variables[15]. Le coefficient de corrélation est une mesure statistique de la force de la relation entre les mouvements relatifs de deux variables, ces valeurs sont comprises entre -1et 1. Un nombre calculé supérieur à 1 ou inférieur à -1 signifie qu'il y a eu une erreur dans la mesure de corrélation. Une corrélation de -1 montre une corrélation négative parfaite, tandis qu'une corrélation de 1 montre une corrélation positive parfaite, ainsi une corrélation proche à 1ou -1 montre une grande corrélation positive ou négative respectivement. Une corrélation de 0 ne montre aucune relation linéaire entre le mouvement des deux variables et ainsi une corrélation proche à 0 montre une faible corrélation. L'intérêt de l'utilisation de cette matrice est d'éliminer les facteurs hautement corrélés afin d'éliminer le bruit et d'assurer une meilleure qualité des données résultantes.

Méthode Stepwise : La méthode Stepwise ou la Régression Pas-à-Pas est un outil automatisé qui permet dans les phases exploratoires de l'élaboration d'un modèle d'identifier un sous ensemble utile de prédicteurs afin d'éliminer les prédicteurs non significatifs dans le but d'améliorer la qualité des résultats[35].
La méthode d'élimination vers l'arrière (Stepwise) commence par un modèle complet chargé de plusieurs variables, puis à chaque itération supprime ou ajoute une (ou plusieurs) variable pour tester son importance et son influence sur résultats globaux.
La régression pas à pas peut être obtenue soit en essayant une variable indépendante à la fois et en l'incluant dans le modèle de régression si elle est statistiquement significative, soit en incluant toutes les variables indépendantes potentielles dans le modèle et en éliminant celles qui ne sont pas statistiquement significatives, mais en général nous utilisons une combinaison de ces deux méthodes, et par conséquent, il existe trois approches de la régression pas à pas[25] :
- La sélection avant (Forward selection) : commence sans aucune variable dans le modèle, teste chaque variable au fur et à mesure qu'elle est ajoutée au modèle, puis conserve celles qui sont considérées comme les statisti-

quement plus significatives, ce processus est répété jusqu'à ce que les résultats soient optimaux.
- L'élimination en arrière (Backward elimination) : commence par un ensemble de variables indépendantes, en supprimant une à la fois, puis en testant pour voir si la variable supprimée est statistiquement significative.
- L'élimination bidirectionnelle (Bidirectional elimination) est une combinaison des deux premières méthodes qui testent les variables à inclure ou à exclure.

Méthodes Wrapper : Les méthodes Wrapper mesurent l'utilité des prédicteurs (facteurs ou caractéristiques) en fonction des performances du classificateur[45]. Ces méthodes sont appelées algorithmes Gloutons (Greedy Algorithms), car elles visent à trouver la meilleure combinaison possible de fonctionnalités (prédicteurs) qui aboutissent au modèle le plus performant.
Les méthodes wrapper fonctionnent de la manière suivante[36] :
1. Rechercher un sous-ensemble de fonctionnalités : à l'aide d'une méthode de recherche, nous sélectionnons un sous-ensemble de fonctionnalités parmi celles disponibles.
2. L'algorithme ML choisi est entraîné sur le sous-ensemble de fonctionnalités précédemment sélectionné.
3. Évaluer les performances du modèle : dans cette étape, nous évaluons le modèle ML nouvellement formé avec une métrique choisie.
4. Répéter : L'ensemble du processus recommence avec un nouveau sous-ensemble de fonctionnalités, un nouveau modèle ML formé, etc.
Nous nous arrêtons quand la condition souhaitée soit remplie, puis nous choisissons le meilleur sous-ensemble avec le meilleur résultat dans la phase d'évaluation.

3.5.2.3 Les algorithmes d'Apprentissage Supervisé

L'apprentissage supervisé peut être classé dans les problèmes de classification et de régression. Il comprend divers algorithmes tels que la régression linéaire, la régression logistique, la machine à vecteurs de support, KNN, Naïve Bayes, les réseaux de neurones artificiels… etc.

a) Régression linéaire

La régression linéaire est l'un des algorithmes d'apprentissage supervisé les plus populaires. Il est aussi simple et parmi les mieux compris en statistique et en apprentissage automatique. Cet algorithme est un type d'analyse prédictive de base.
Nous pouvons les représenter les relations entre une variable dépendante et une ou plusieurs variables indépendantes par une forme simple :

$$y = ax + b + \epsilon \text{ où } F(x) = ax + b$$

, avec y la variable cible, aléatoire dépendante, a et b les coefficients (pente et ordonnée à l'origine) à estimer, x la variable explicative et indépendante, ϵ une variable aléatoire qui représente l'erreur. Dans le cas de modèle linéaire simple, ou multiple dans ce qui suit [4] :

$$y = ax_1 + bx_2 + cx_3 + \ + K + \epsilon \text{ où } F(X) = aX + b$$

y la variable cible, aléatoire dépendante et a,...,K les coefficients (pente et ordonnée à l'origine) à estimer,$X = (x_1, , x_q)$ la variable explicative, indépendante, ϵ une variable aléatoire qui représente l'erreur.
Pour juger la qualité d'une Régression Linéaire. Nous avons le coefficient de corrélation noté R^2, il mesure l'adéquation entre le modèle et les données observées ou encore à quel point l'équation de régression est adaptée pour décrire la distribution des points, ainsi le coefficient de détermination est calculé avec l'équation :

$$R^2 = 1 \text{ -(SCR/SCT)}$$

avec SCT la somme des carrés totaux et SCR la somme des carrés des résidus. Plus R^2est proche de 1, plus la qualité de la prédiction par le modèle de Régression Linéaire est bonne (le nuage de points est resserré autour de la droite). l'inverse, plus R^2 est proche de 0, plus la qualité de la prédiction est mauvaise. Un R^2 égal à 1 signifie que la prédiction du modèle est parfaite [1].
La figure 3.3 représente la droite de régression linéaire ou la droite des moindres carrés de Y en X représente la droite d'ajustement linéaire, celle qui résume le mieux la structure du nuage de points pendant la phase d'apprentissage.

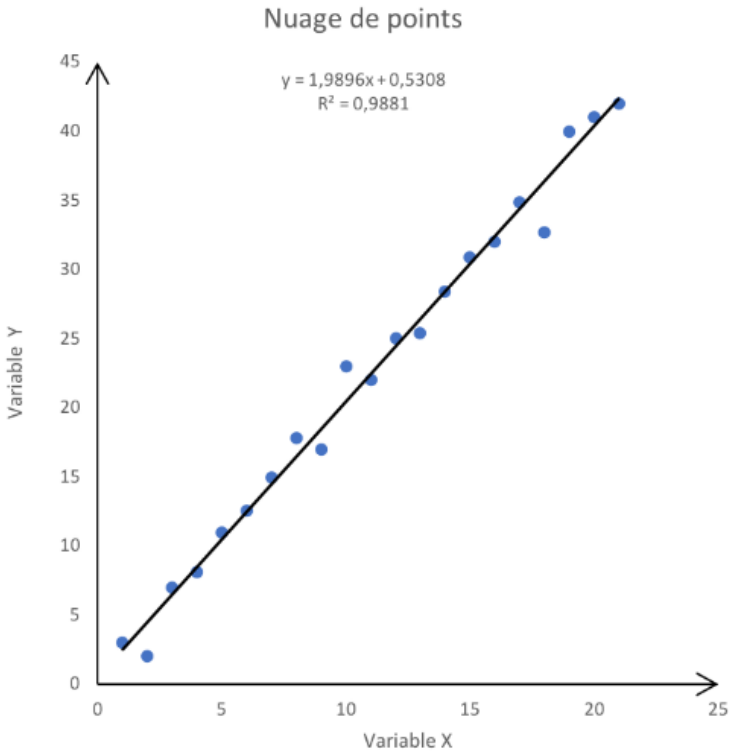


FIG. 3.3: Exemple de graphe de Régression Linéaire simple [4].

b) Régression Logistique

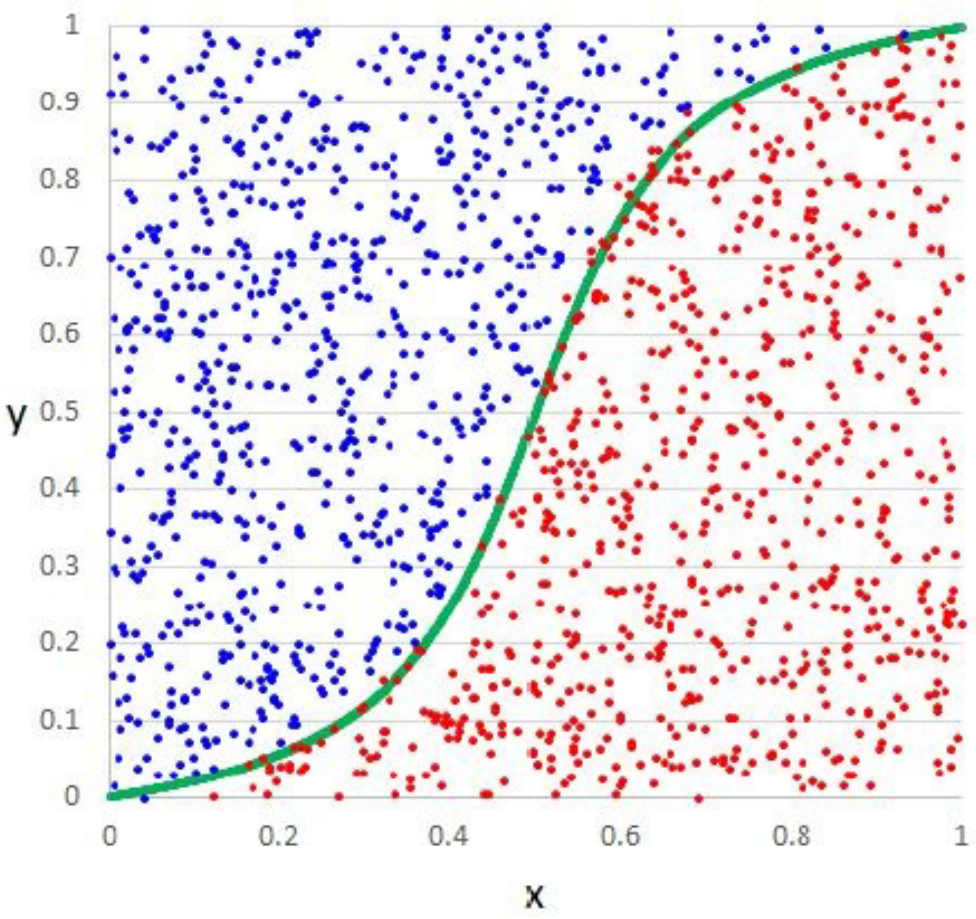


Fig. 3.4: Exemple de graphe de Régression Logistique [51]

La régression logistique est utilisée pour prédire la probabilité d'une variable cible, c'est l'un des algorithmes les plus couramment utilisés pour les problèmes de classification binaire propose le résultat sous forme de probabilités de la classe par défaut. Le résultat x appartient donc à l'intervalle [0:1]. C'est-à-dire qu'il est compris entre 0 et 1, vu qu'il s'agit d'une probabilité.

Un seuil est ensuite appliqué sur x pour forcer cette probabilité dans une classification binaire à l'aide d'un seuil, ainsi la variable dépendante est de nature binaire avec des données codées soit 1 (succès) ou 0 (échec) , généralement le seuil pris est égale à 0,5.

Mathématiquement, un modèle de Régression Logistique prédit P(x) en fonction de x. C'est l'un des algorithmes ML les plus simples qui peuvent être utilisés pour divers problèmes de classification.

On vous présente dans la figure 3.4 un exemple de graphe de Régression Logistique, tels que la courbe en vert représente une frontière entre anomalies représentés avec des points en rouge et les non anomalies avec des points en bleu [51].

c) K-NN

L'algorithme K-Nearest Neighbors (KNN) ou K-plus proches voisins est un type d'algorithme de ML Supervisé qui peut être utilisé à la fois pour les problèmes de classification et de prédiction de régression. Cependant, il est principalement utilisé pour les problèmes prédictifs de classification dans l'industrie. Cet algorithme est considéré à la fois non paramétrique et une exemple d'apprentissage paresseux [40].

- Non paramétrique signifie qu'il ne fait aucune hypothèse. Le modèle est entièrement constitué de données qui lui sont données.

- L'apprentissage paresseux signifie que l'algorithme ne fait aucune génération, cela signifie qu'il y a peu de formation impliquée lors de l'utilisation de cette méthode, pour cette raison, toutes les données d'entraînement sont également utilisées dans les tests lors de l'utilisation de KNN.

L'algorithme K-plus proches voisins (KNN) utilise la « similarité des caractéristiques » pour prédire les valeurs des nouveaux points de données, ce qui signifie que les nouveaux points de données se verra attribuer une valeur en fonction de sa correspondance avec les points de l'ensemble d'apprentissage. KNN classe un nouveau point de données en fonction de sa similitude.

Dans KNN, K est le nombre de voisins les plus proches. Le nombre de voisins est le principal facteur décisif.

La figure 3.5 illustre un exemple d'application de l'algorithme KNN, ainsi que l'importance du choix du facteur k. La figure montre les différents points de données qui sont les rouges qui appartiennent à la classe A, les verts qui appartiennent à la classe B, et un point de données jaune qui sera classé parmi ces deux classes (le nouvel exemple à classifier).

Lorsque k = 3, deux parmi les trois voisins les plus proches sont de la classe B et un est de la classe A, donc le nouvel exemple sera classifié dans la classe B, alors que lorsque k=7, trois parmi les sept voisins les plus proches sont de la classe B et quatre sont de la classe A, donc le nouvel exemple sera classifié dans la classe A.

Écriture algorithmique

Nous allons schématiser le fonctionnement de K-NN en l'écrivant en pseudo-code suivant :

Début Algorithme

Données en entrée :

- Un ensemble de données D.
- Une fonction de définition distance d.
- Un nombre entier K.

Pour une nouvelle observation X dont on veut prédire sa variable de sortie y Faire :

1. Calculer toutes les distances de cette observation X avec les autres observations du jeu de données D.

2. Retenir les K observations du jeu de données D les proches de X en utilisation le fonction de calcul de distance d.

3. Prendre les valeurs de y des K observations retenues :

- Si on effectue une régression, calculer la moyenne (ou la médiane) de y retenues.
- Si on effectue une classification , calculer le mode de y retenues.

4. Retourner la valeur calculée dans l'étape 3 comme étant la valeur qui a été prédite par K-NN pour l'observation X.

Fin Algorithme.

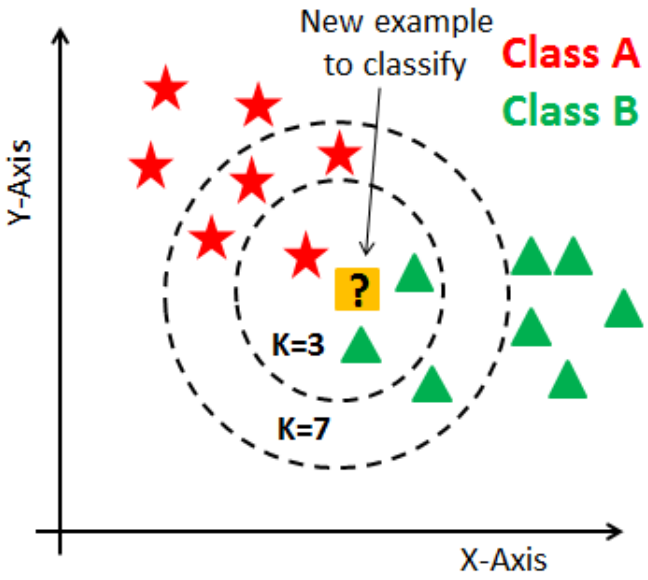


FIG. 3.5: Exemple de graphe de l'algorithme KNN.

d) Naïve Bayes

L'algorithme Naïve Bayes est un algorithme d'apprentissage supervisé, basé sur le théorème de Bayes et utilisé pour résoudre des problèmes de classification. Il est principalement utilisé dans la classification de texte qui inclut un ensemble de données d'apprentissage de grande dimension. Naïve Bayes est l'un des algorithmes de classification simples et les plus efficaces qui aident à créer des modèles d'apprentissage automatique rapides capables de faire des prédictions rapides. Aussi est un classificateur probabiliste, ce qui signifie qu'il prédit sur la base de la probabilité d'un objet. Certains exemples populaires de l'algorithme Naïve Bayes sont la filtration du spam, détection d'anomalie (anomalie , non anomalie) et la classification des articles. Naive Bayes dépend du principe du théorème de Bayes suivant :

P(A) = (P(B)P(A))/(P(B)) avec, P(A|B) est Probabilité a posteriori : Probabilité de l'hypothèse A sur l'événement observé B.

P(B|A) est la probabilité de vraisemblance : probabilité de la preuve étant donnée que la probabilité d'une hypothèse est vraie.

P(A) est la probabilité a priori : probabilité d'hypothèse avant d'observer la preuve.

P(B) est la probabilité marginale : probabilité de preuve.

On vous présente dans la figure suivante un exemple de graphe de Naive Bayes.

Écriture algorithmique

Nous allons schématiser le fonctionnement de Naive Bayes en l'écrivant en pseudo-code suivant :

Début Algorithme

Données d'entrée :

Ensemble de données d'entraînement T, F=(f1,f2,f3,...,fn) //valeur de la variable prédictive dans l'ensemble de données de test.

Données de sorties : A classe d'ensemble de données de test.

Les étapes :

1. Lire l'ensemble de données d'entraînement T ;
2. Calculer la moyenne et l'écart type des variables prédictives dans chaque classe ;
3. Répéter

Calculer la probabilité de fi en utilisant l'équation la densité de Gauss dans chaque classe ;

Jusqu'à ce que la probabilité de toutes les variables prédictives (f1,f2,f3,...,fn) ait été calculée ;

4. Calculer la probabilité pour chaque classe ;
5. Obtenir la plus grande probabilité ;

Fin de l'algorithme.

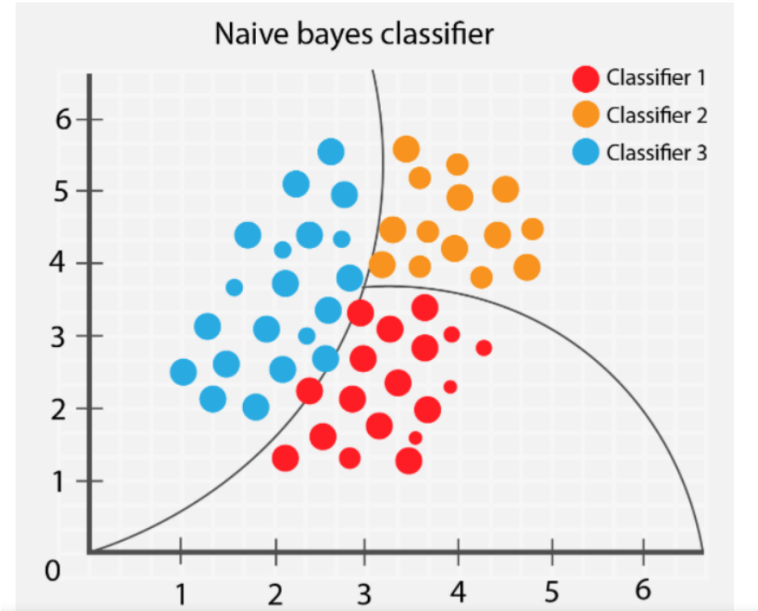


Fig. 3.6: Exemples de graphe de Naive Bayes

e) Les Réseaux de Neurones Artificiels

Les Réseaux de Neurones Artificiels peuvent être décrits comme des systèmes composés d'au moins deux couches de neurones ; une couche d'entrée et une couche de sortie et comprenant généralement des couches intermédiaires. Plus le problème à résoudre est complexe, plus le Réseau de Neurones Artificiels doit comporter plus de couches, chaque couche contient un grand nombre de neurones artificiels spécialisés, si le nombre de couches intermédiaire est supérieur à 1 nous classons le réseau de neurones comme algorithme d'Apprentissage Profond. Au sein d'un Réseau de Neurones Artificiels, le traitement de l'information suit toujours la même séquence [26] ; les informations sont transmises sous la forme de signaux aux neurones de la couche d'entrée, où elles sont traitées. À chaque neurone est attribué un « poids » particulier, et donc une importance différente. Associé à la fonction dite de transfert, le poids permet de déterminer quelles informations peuvent entrer dans le système. À l'étape suivante, une fonction dite d'activation associée à une valeur seuil calcule et pondère la valeur de sortie du neurone. En fonction de cette valeur, un nombre plus ou moins grand de neurones sont connectés et activés. Cette connexion et cette pondération dessinent un algorithme qui fait correspondre un résultat à chaque entrée. Chaque nouvelle itération permet d'ajuster la pondération et donc l'algorithme de façon à ce que le réseau donne à chaque fois un résultat plus précis et fiable [26]. La figure 3.7 représente les différents types de couches qui constituent un Réseau de Neurones Artificiels.

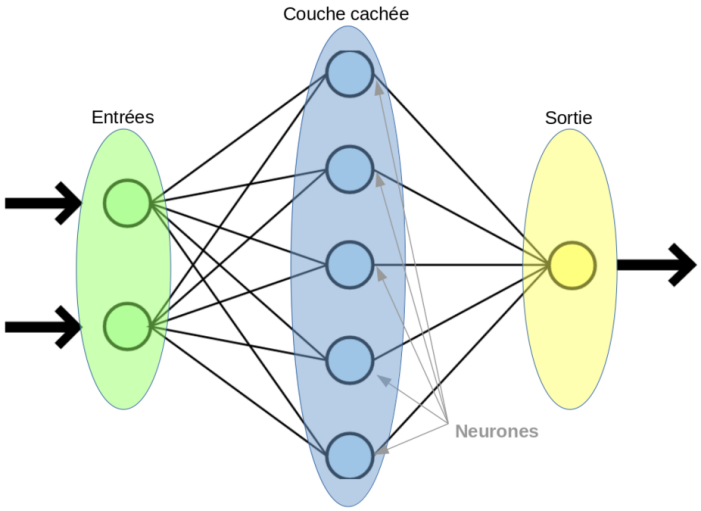


Fig. 3.7: le graphe de Réseaux de Neurones Artificiels [26]

f) Les machines à vecteurs de support SVM

Les Machines à Vecteurs de Support ou Séparateurs à Vastes Marges (SVM) sont des algorithmes qui séparent les données en classes. Pendant l'entraînement, un SVM trouve une ligne qui sépare les données d'un jeu en classes spécifiques et maximise les marges (les distances entre les frontières de séparation et les échantillons les plus proches) de chaque classe [40], après avoir appris les lignes de classification, le modèle peut ensuite les appliquer aux nouvelles données. Les spécialistes placent le SVM dans la catégorie des « classificateurs linéaires » ; l'algorithme est idéal pour identifier des classes simples qu'il sépare par des vecteurs nommés « hyperplans ». Il est également possible de programmer l'algorithme pour des données non linéaires, que nous ne pouvons pas séparer clairement par des vecteurs.
L'objectif principal du SVM est de diviser les ensembles de données en classes pour trouver un hyperplan marginal maximal (MMH) , et cela peut être fait par deux étapes : le SVM générera des hyperplans de manière itérative qui séparent les classes de la meilleure façon, ensuite, il choisira l'hyperplan qui sépare correctement les classes, plus précisément dans l'algorithme SVM, nous traçons chaque élément de données comme un point dans un espace à n dimensions (où n est le nombre d'entités dont vous disposez), la valeur de chaque entité étant la valeur d'une coordonnée particulière. Ensuite, nous effectuons la classification en trouvant l'hyperplan qui différencie les deux classes de la meilleure façon [9]. La figure 3.8 illustre un exemple de l'application de l'algorithme SVM, avec un hyperplan marginal optimal (HMM) qui sépare les classes de la meilleure façon.

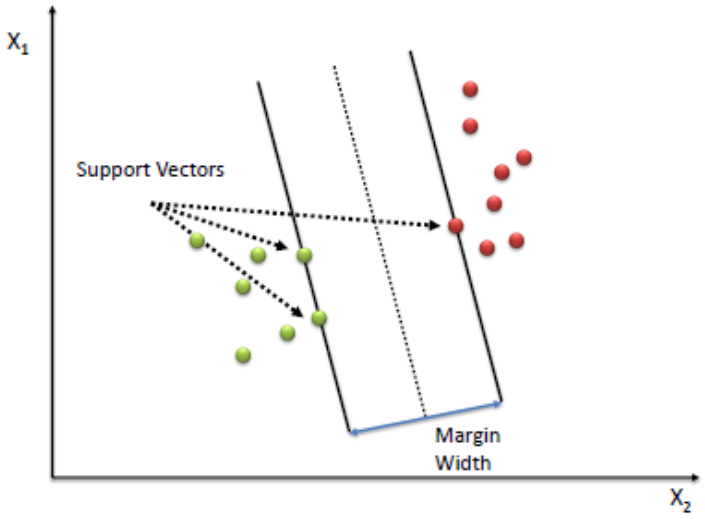


FIG. 3.8: Graphe de l'algorithme SVM [9]

3.5.3 Machine Learning Non-Supervisé

3.5.3.1 Définition

L'Apprentissage Non Supervisé ou Unsupervised Learning est une méthode d'analyse des données dans le domaine de l'Intelligence Artificielle dans laquelle les algorithmes s'appuient sur les similitudes entre les différentes valeurs d'entrée.
Contrairement à l'Apprentissage Supervisé, dans l'Apprentissage Non Supervisé, l'ordinateur essaie d'identifier par lui-même des modèles et des structures au sein des valeurs saisies, ainsi il n'y a pas de réponse correcte ni d'enseignant [27]. Dans cette méthode, les développeurs gardent un contrôle total et détaillent le but de l'apprentissage au préalable. Différents processus interviennent dans ce cadre, il consiste à apprendre à un algorithme des informations qui ne sont ni classées, ni étiquetées, et à permettre à cet algorithme de réagir à ces informations sans supervision.
Il existe de nombreux exemples pratiques de l'Apprentissage Non Supervisé.

3.5.3.2 Les algorithmes d'Apprentissage Non-Supervisé

Les algorithmes d'Apprentissage Non Supervisé peuvent exécuter des tâches de traitement plus complexes que les systèmes d'Apprentissage Supervisé, mais ils peuvent aussi être plus imprévisibles. Même si un système d'IA d'Apprentissage Non Supervisé parvient tout seul, par exemple, à faire le tri entre des catégories A et des catégories B, il peut aussi ajouter des catégories inattendues et non désirées pour y classer des races inhabituelles, créant la confusion au lieu de mettre de l'ordre, nous distinguons :

a)DBSCAN(density-based spatial clustering of applications with noise) :

Les méthodes de Clustering (ou méthodes de regroupement) utilisent la même approche, cette approche consiste à calculer d'abord les similitudes, puis les utiliser pour regrouper les points de données en groupes ou lots.

L'algorithme DBSCAN est un algorithme qui utilise une méthode basée sur la notion intuitive de « clusters » et de « bruit ». L'idée clé de cet algorithme est de diviser les points en k groupes appelés clusters, homogènes et compacts , il nécessite la définition de deux paramètres [34] :

- Epsilon : il définit le voisinage autour d'un point de données, c'est-à-dire que si la distance entre deux points est inférieure ou égale à « epsilon », alors ils sont considérés comme voisins. Le choix de ce paramètre est très important car si la valeur epsilon est choisie trop petite, une grande partie des données sera considérée comme des valeurs aberrantes et si elle est choisie très grande, les clusters se fusionnent.
- MinPts : est le nombre minimum de voisins dans le rayon epsilon. Plus le jeu de données est plus grand, une valeur plus élevée de MinPts doit être choisie.

Le DBSCAN fonctionne de la manière suivante :

1. Le DBSCAN commence par un point de données de départ arbitraire qui n'a pas été visité. Le voisinage de ce point est extrait en utilisant une distance epsilon.
2. S'il y a un nombre suffisant de points (selon les MinPts) dans ce voisinage, le processus de mise en cluster démarre et le point de données actuel devient le premier point du nouveau cluster. Sinon, le point sera étiqueté comme bruit (plus tard, ce point bruyant pourrait devenir la partie du cluster). Dans les deux cas, ce point est marqué comme « visité ».
3. Pour ce premier point du nouveau cluster, les points situés dans son voisinage à distance se joignent également au même cluster. Cette procédure est ensuite répétée pour tous les nouveaux points qui viennent d'être ajoutés au groupe de cluster.
4. Ce processus des étapes 2 et 3 est répété jusqu'à ce que tous les points du cluster soient dé- terminés, c'est-à-dire que tous les points à proximité du voisinage du cluster ont été visités et étiquetés.
5. Une fois terminé avec le cluster actuel, un nouveau point non visité est récupéré et traité, ce qui permet de découvrir un nouveau cluster ou du bruit. Ce processus se répète jusqu'à ce que tous les points soient marqués comme étant visités. A la fin chacun des points visités a été marqué comme appartenant à un cluster ou comme étant du bruit.

La figure 3.9 illustre un exemple d'application de l'algorithme DBSCAN sur un ensemble de données, les données d'origine à gauche et les clusters identifiés par l'algorithme DBSCAN à droite. Pour le DBSCAN cluster, les grands points colorés représentent les membres principaux du cluster, les petits points colorés représentent les membres du bord du cluster et les petits points noirs représentent les valeurs aberrantes.[10]

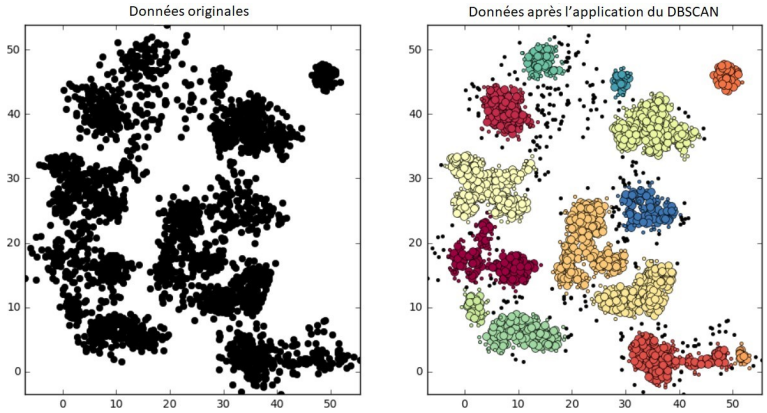


FIG. 3.9: exemple d'application de l'algorithme DBSCAN [10]

b)OPTICS algorithm (Ordering Points Toc Identify the Clustering Structure)

L'algorithme OP- TICS s'inspire de l'algorithme de Clustering DBSCAN, et il ajoute deux autres termes aux concepts de Clustering DBSCAN [22] :

1. **Distance centrale (ou core distance) :** Il s'agit de la valeur minimale du rayon requise pour classer un point donné en tant que point central. Si le point donné n'est pas un point central, sa distance centrale n'est pas définie.

2. **Distance d'accessibilité (ou reachability distance) :** elle est définie par rapport à un autre point de données q. La distance d'accessibilité entre un point p et q est le maximum de la distance centrale de p et de la distance euclidienne (ou d'une autre) entre p et q, la distance d'accessibilité n'est pas définie dans le cas où le point p n'est pas un point central.

Cette technique de Clustering est différente des autres techniques de Clustering car cette tech- nique ne segmente pas explicitement les données en clusters mais elle produit une visualisation des distances d'accessibilité et utilise cette visualisation pour regrouper les données.

La figure 3.10 représente un exemple d'application du principe de l'algorithme OPTICS.

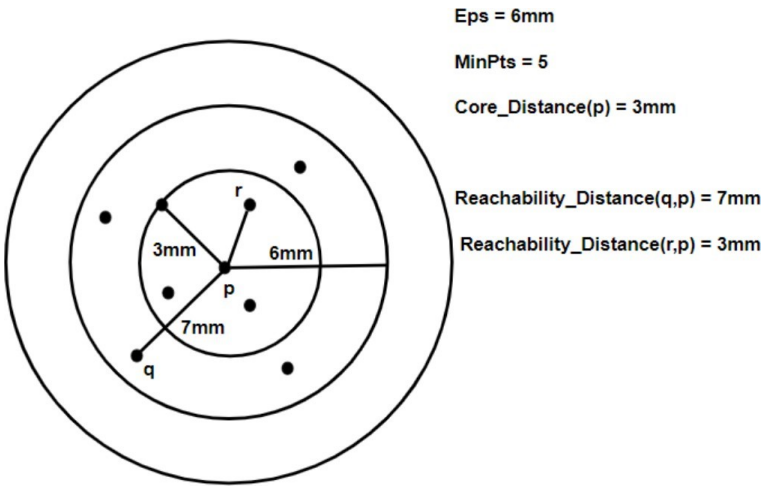


Fig. 3.10: Exemple d'application de l'algorithme OPTICS [22]

3.5.4 Machine Learning renforcé

Le Deep Learning (DL) ou Apprentissage Profond est un ensemble de méthodes d'Apprentissage Automatique conçues sur la base de Réseaux de Neurones Artificiels Profonds, visant à mimer la profondeur des couches d'un cerveau [21]. Les différentes couches du réseau de neurones permettent au modèle d'apprendre progressivement afin de trouver les caractéristiques les plus pertinentes parmi des données qui lui sont fournies (effectuer une sélection des facteurs les plus intéressants)[32].Le Deep Learning a permis des résultats bien plus intelligents que ceux rendus à l'origine par le Machine Learning. Pour faire simple, cette méthode consiste à laisser l'algorithme apprendre de ses propres erreurs. Afin d'apprendre à prendre les bonnes décisions, l'Intelligence Artificielle se retrouve directement confrontée à des choix. Si elle se trompe, elle est "pénalisée". Au contraire, si elle prend la bonne décision, elle est "récompensée". Afin d'obtenir toujours plus de récompenses, l'IA va donc faire de son mieux pour optimiser sa prise de décisions[32].
La technologie commencera par prendre des décisions totalement aléatoires et de récompense en développera sa propre méthode pour accomplir systématiquement la tâche qui lui est confiée. Ainsi, cette méthode d'apprentissage présente l'avantage fascinant de laisser la machine faire preuve de "créativité".

3.5.5 Techniques de validation des modèles

Les techniques de validation en Apprentissage Automatique sont utilisées pour obtenir le taux d'erreur du modèle ML, qui peut être considéré comme proche du taux d'erreur réel de la population. Si le volume de données est suffisamment important pour être représentatif de la population, nous n'aurons peut-être pas besoin des techniques de validation.
Cependant, dans des scénarios du monde réel, nous travaillons avec des échantillons de données qui peuvent ne pas être un véritable représentant de la population. C'est là que les techniques de validation entrent en jeu.
Bien que le processus de validation ne puisse pas trouver directement ce qui ne va pas, le processus peut parfois

nous montrer qu'il y a un problème avec la stabilité du modèle. Le choix de la bonne méthode de validation est très important, Cependant, il existe différents types de techniques de validation à suivre, tout en s'assurant qu'elles conviennent au modèle de ML et qu'elles assurent un travail transparent et impartial, rendant le modèle de ML complète- ment fiable et acceptable dans le monde de l'IA. Parmi les techniques de validation nous nous intéressons aux suivantes :

3.5.5.1 Matrice de Confusion

Une Matrice de Confusion (ou Matrice de Contingence ou Confusion Matrix) est une matrice N x N utilisée pour évaluer les performances d'un modèle de Machine Learning, où N est le nombre de classes cibles. La matrice compare les valeurs cibles réelles avec celles prédites par le modèle d'Apprentissage Automatique [46]. Elle mesure la performance d'un modèle de classification d'Apprentissage Automatique où la sortie peut être deux classes ou plus. C'est un tableau avec quatre combinaisons différentes de valeurs prévues et réelles.

la figure 3.11 représente une Matrice de Confusion composée des quatre combinaisons possibles de valeurs prévues et réelles : Vrai Positif, Vrai Négatif, Faux Positif et Faux Négatif.

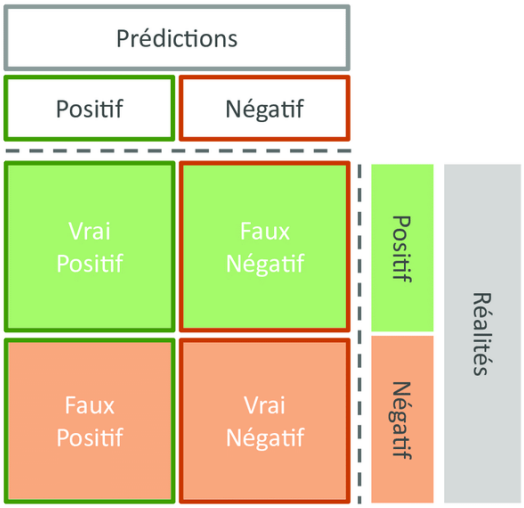


FIG. 3.11: Matrice de confusion

Vrai Positif : nous avons prédit positif et c'est vrai.

Vrai Négatif : nous avons prédit négatif et c'est vrai.

Faux Positif (Erreur de type 1) : nous avons prédit positif et c'est faux.

Faux Négatif : (Erreur de type 2) : nous avons prédit négatif et c'est faux.

À partir de la Matrice de Confusion, nous pouvons calculer plusieurs métriques mesurant la validité de notre modèle, les plus importantes sont :

L'Exactitude (Accuracy) : Calcule sur toutes les classes, combien classes le modèle a prédit correctement[46].

$$Accuracy = TP + TN/(TP + TN + FP + FN)$$

La Précision (Precision) : Calcule parmi toutes les classes positives que le modèle a prédit, combien sont réellement positives.

$$Precision = TP/(TP + FP)$$

Le Rappel (Recall) : Il calcule combien de vrais positifs ont été trouvés parmi toutes les classes qui sont réellement positives.

$$Recall = TP/(TP + FN)$$

La Mesure-F (F-measure or F-score) : Il est difficile de comparer deux modèles avec une faible Précision et un Rappel élevé ou vice versa. Donc, pour les rendre comparables, nous utilisons F-Score. Le F-score aide à mesurer le Rappel et la Précision en même temps. Il utilise la moyenne harmonique à la place de la moyenne arithmétique en punissant davantage les valeurs extrêmes.

$$Fmeasure = (2RecallPrecision)/(Recall + Precision)$$

Recall est une des métriques utiliser dans la Matrice de Confusion pour mesurer la validité d'un modèle. Le rappel (recall) est littéralement combien de vrais positifs (True Positive) ont été rappelés (trouvés) parmi les résultats qui sont réellement positifs (True Positive + False Negative).

3.6 Conclusion

Dans ce chapitre, nous nous somme familiariser avec les différentes notions liées à l'apprentissage automatique, et ses différents aspects. En expliquant la façon de le mettre en œuvre dès la phase de collecte de données jusqu'à la modélisation et la validation des modèles.

A travers ce chapitre, nous pouvons entamer la partie application de notre sujet de thème qui est la détection des anomalies sur les réseaux cellulaires (LTE) avec les outils de machine learning supervisé.

Chapitre 4

Application

4.1 Introduction

Dans ce dernier chapitre, nous allons présenter la phase finale de notre projet qui est l'application. Cette phase concerne la construction des étapes de notre modèle. Comme nous avons cité dans le chapitre précèdent, la classification des anomalies est un problème d'apprentissage automatique qui relève de la catégorie de l'apprentissage supervisé. Nous allons définir les différentes étapes qui composent la méthodologie proposée ainsi que leur enchaînement, les étapes étant : la collecte et la préparation de données ; la modélisation en utilisant les classificateurs à savoir, SVM, KNN et naïve bayes pour classifier un ensemble de flux de donnés en anomalie ou non anomalie (les indicateurs de performances) en surveillant leurs comportements ; le choix du meilleur modèle.

4.2 Technologie Utilisée

4.2.1 Python

Python est l'un des langages de programmation les plus populaires au monde ces dernières années. Ce dernier est un langage de programmation informatique souvent utilisé pour créer des sites Web et des logiciels, automatiser des tâches et effectuer des analyses de données. Python est un langage à usage général, ce qui signifie qu'il peut être utilisé pour créer une variété de programmes différents et qu'il n'est pas spécialisé pour des problèmes spécifiques.

4.2.1.1 Jupyter Notebook

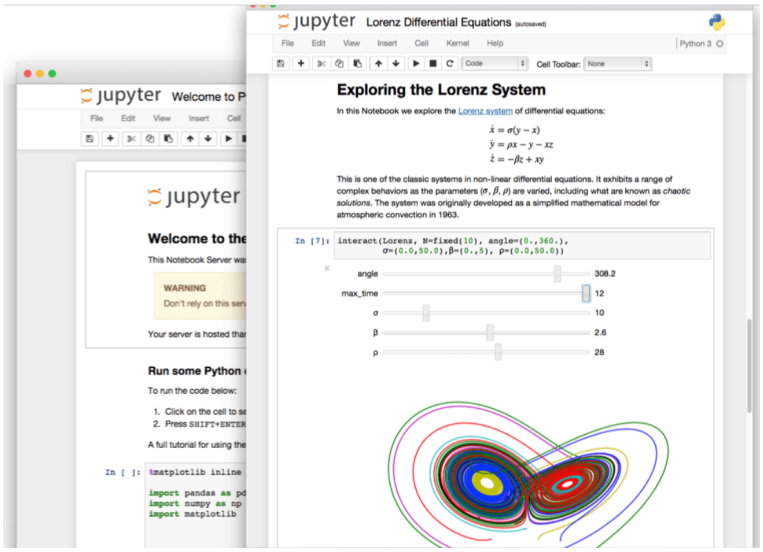


FIG. 4.1: Interface de Jupyter Notebook

Jupyter Notebook est une application Web Open Source permettant de créer et de partager des documents contenant du code (exécutable directement dans le document), des équations, des images et du texte. Avec cette application il est possible de faire du traitement de données, de la modélisation statistique, de la visualisation de données, du Machine Learning, etc. Elle est disponible par défaut dans la distribution Anaconda[44].

4.2.1.2 Google Colab

Google Colaboratory

Google Colab ou Colaboratory est un service cloud, offert par Google (gratuit), basé sur Jupyter Notebook et destiné à la formation et à la recherche dans l'apprentissage automatique. Cette plateforme permet d'entraîner des modèles de Machine Learning directement dans le cloud. Sans donc avoir besoin d'installer quoi que ce soit sur notre ordinateur à l'exception d'un navigateur. C'est un outil complet pour entraîner rapidement et tester rapidement des modèles d'apprentissage automatique sans avoir de contrainte matérielle. Colab permet à n'importe qui d'écrire et d'exécuter le code Python de son choix par le biais du navigateur[44].

- **Les blocs notes Colab :**ils permettent de combiner un code exécutable et du texte enrichi dans un seul document, ainsi que des images , HTML, LaTex et plus encore. Lorsque vous créez vos propres blocs-notes Colab, ils sont stockés dans votre compte Google Drive. Vous pouvez facilement partager vos blocs-notes Colab avec des collègues ou des amis, leur permettant de commenter vos blocs-notes ou même de les modifier. vous pouvez utiliser Google Colabs comme les notebooks Jupyter. Ils sont vraiment pratiques car Google Colab les héberge. Nous pouvez également partager ces blocs-notes afin que d'autres personnes puissent exécuter notre

code, le tout dans un environnement standard car il ne dépend pas de nos propres machines locales. Cependant, vous devriez peut-être installer certaines bibliothèques dans notre environnement lors de l'initialisation.
C'est la manière dont nous avons procéder pour effectuer notre projet de data-science pour la détection des anomalies. Cette fonctionnalité nous a permet de travailler sur des notebooks communs.

- **Colab permet :**

1- D'améliorer vos compétences de codage en langage de programmation Python.
2- De développer des applications en Deep Learning en utilisant des bibliothèques Python populaires telles que Keras, TensorFlow, PyTorch et OpenCV.
3- D'utiliser un environnement de développement (Jupyter Notebook) qui ne nécessite aucune configuration.

4.3 Package Utilisés

Une des grandes forces du langage de programmation Python est l'énorme communauté de développeurs qui développent et maintiennent un grand nombre de librairies. Ces librairies permettent d'utiliser des nouveaux types d'objets spécialisés, aux applications très variées, dont par exemple l'import/export/le traitement/la visualisation de données spécifiques[8].
Dans le domaine de l'analyse de données en particulier, nous avons La figure 4.2 qui illustrent les librairies existantes :

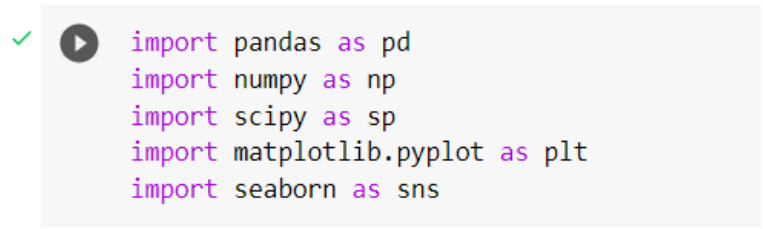


Fig. 4.2: Importation des librairies sur Colab

- **Seaborn :**est une librairie qui vient s'ajouter à Matplotlib, remplace certains réglages par défaut et fonctions, et lui ajoute de nouvelles fonctionnalités. Seaborn vient corriger trois défauts de Matplotlib :

1- Matplotlib, surtout dans les versions avant la 2.0, ne génère pas des graphiques d'une grande qualité esthétique.
2- Matplotlib ne possède pas de fonctions permettant de créer facilement des analyses statistiques sophistiquées.
3- Les fonctions de Matplotlib ne sont pas faites pour interagir avec les Dataframes de Panda (que nous verrons au chapitre suivant).

- **Pandas :** est une librairie écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques(Dataframe) et de séries temporelles[8].

- **NumPy :** est une librairie pour langage de programmation Python, destinée à manipuler des matrices ou tableaux multidimensionnels ainsi que des fonctions mathématiques opérant sur ces tableaux[8].
- **Scipy :** est une librairie open source utilisée pour résoudre des problèmes mathématiques, scientifiques, d'ingénierie et techniques. Il permet aux utilisateurs de manipuler les données et de les visualiser à l'aide d'un large éventail de commandes Python de haut niveau. SciPy est construit sur l'extension Python NumPy[8].
- **Matplotlib :** est une libraire du langage de programmation Python destinée à tracer et visualiser des données sous formes de graphiques5. Elle peut être combinée avec les bibliothèques python de calcul scientifique NumPy et SciPy[8].

Pour importer un package, il faut tout d'abord qu'il soit installé sur la machine (ce qui est déjà fait la plupart du temps dans Google Colab). Ensuite, il faut utiliser les mots-clés suivants :

import package **as** alias

Où :

from package import subpackage **as** alias

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import classification_report,confusion_matrix
from matplotlib.colors import ListedColormap
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
from sklearn.svm import SVC
from sklearn.model_selection import cross_val_score
```

FIG. 4.3: Exemples de package utilisé

On termine notre présentation de package par la librairie Scikit-learn.

Scikit-learn : Scikit-learn, encore appelé sklearn, est la bibliothèque la plus puissante et la plus robuste pour le machine learning en Python. Elle fournit une sélection d'outils efficaces pour l'apprentissage automatique et la modélisation statistique, notamment la classification, la régression et le clustering via une interface cohérente en Python. Cette bibliothèque, qui est en grande partie écrite en Python, s'appuie sur NumPy, SciPy et Matplotlib. Scikit-learn couvre la plupart des algorithmes d'apprentissage automatique (SVM,KNN,Naive Bayes,RNN ...Etc)[39].

4.4 Maintenance Prédictive

La maintenance prédictive consiste à anticiper les défaillances à venir sur un équipement, un objet, un système, etc. Concrètement, il s'agit d'aller au-devant d'une panne ou d'un dysfonctionnement grâce au cumul d'un ensemble de données, en se basant sur une surveillance méthodique et une analyse précise de l'évolution d'une machine ou d'un composant. Elle repose sur la collecte et l'analyse de données, via des capteurs communicants, qui font partie de l'Internet of Thing (IoT) [11]qui fait référence aux appareils physiques qui reçoivent et transfèrent des données sur des réseaux sans fil, avec une intervention humaine limitée. Cette technologie repose sur l'intégration d'un système informatique à toutes sortes d'objets. On l'associe à l'intelligence artificielle car ce sont les outils analytiques qui permettent de détecter les anomalies annonciatrices de pannes dans les données relevées par les capteurs. L'intelligence artificielle est ainsi essentielle au bon fonctionnement de la maintenance prédictive, tel Le machine learning la branche de IA qui permet, grâce à des algorithmes d'apprentissage automatique, d'analyser des données et de diagnostiquer des pannes à un stade précoce. Il s'agit de la technologie d'IA utilisée dans la maintenance prédictive. Il existe d'autre types de maintenance. La première est la maintenance curative, aussi appelée maintenance corrective ou réactive, qui consistait à intervenir sur une machine une fois qu'une panne survenait. Ensuite a été adoptée la maintenance préventive, dont l'objectif était de changer les pièces avant qu'une panne ne puisse intervenir. Avec la maintenance prédictive, la panne est anticipée dès que des signes avant-coureurs se font ressentir sur la machine, ce qui permet de changer les pièces au bon moment et de réduire les coûts en changeant des pièces inutilement. Établir un protocole de maintenance prédictive par l'entreprise AT, permet de détecter et analyser finement les pertes de qualité du réseau LTE. Cette analyse et ces données peuvent ensuite être utilisées pour prédire les comportements du réseau. Algérie télecom peut alors mettre en place des correctifs et outils nécessaires. Ainsi, il garantit à ses clients une qualité de réseau améliorée. Son Objectif est de proposer et garantir un réseau fixe et mobile de qualité est un enjeu majeur pour séduire de nouveaux clients et fidéliser les abonnés. Bien utilisée, l'analyse prédictive peut fournir en temps réel de précieuses informations pour améliorer la qualité de service.

4.4.1 Operational Support System (OSS)

C'est l'ensemble des composants opérationnels ou les systèmes informatiques utilisés par un opérateur de télécommunications pour la gestion de son réseau. Elle permet la maintenance opérationnelle du réseau de télécommunications déployé par un opérateur pour fournir ses services de télécommunications (téléphonie, accès internet, télévision, transmission de données, réseau privé virtuel).
L'OSS est un système informatique interfacé au réseau de télécommunications jusqu'au niveau des équipements de réseau pour assurer le maintien des processus d'exploitation tels que la maintenance du réseau. Le maintien des processus pour l'exploitation et la maintenance du réseau est assuré par des composants logiciels back-office qui travaillent en interaction les uns avec les autres et qui sont utilisés dans différents domaines :

- Le recensement et la gestion de l'inventaire réseau.
- L'installation et la configuration des composants réseau.

- Le Service provisioning ou mise en œuvre des services pour le client.
- La gestion des incidents réseaux.
- La performance et qualité de service perçues par le client.
- La gestion de la sécurité réseau en particulier protection contre les intrusions et les attaques externes [54].

4.5 Collecte et Préparation de Données

4.5.1 Collecte de données

A partir des facteurs jugés pertinents cité dans les chapitres précédents, nous avons choisi les attributs suivants :PRB_75,Usernum_d'utilisateurs , CQI_LOW ,CQI_ MEDIUM,CQI_EXCELENT, RANG1, RANG2, RANG3, RANGE 4 ,RANG5 et Le RI.

Les données collectées proviennent de CMP de Algérie Télécom de la wilaya de Boumerdes. La figure 4.4 illustre la Dataset utilisé dans notre projet .

	PRB_75	User_num	CQI_LOW	CQI_MEDIUM	CQI_EXELEN	RANG1	RANG2	RANG3	RANG4
2	99	87	39	44	17	32.83%(0_to_1500m)	28.45%(1500_to_3000m)	13.22%(3000_to_5000m)	25.5%(5000_to_13000m)
3	99	112	24	49	28	28.08%(0_to_1500m)	17.11%(1500_to_3000m)	25.3%(3000_to_5000m)	29.52%(5000_to_13000m)
4	54	37	10	44	47	90.69%(0_to_1500m)	5.72%(1500_to_3000m)	2.39%(3000_to_5000m)	1.01%(5000_to_13000m)
5	27	14	29	42	29	55.6%(0_to_1500m)	29.07%(1500_to_3000m)	15.07%(3000_to_5000m)	0.26%(5000_to_13000m)
6	98	79	28	43	29	56.97%(0_to_1500m)	19.69%(1500_to_3000m)	22.11%(3000_to_5000m)	1.19%(5000_to_13000m)
7	77	58	19	48	33	85.37%(0_to_1500m)	9.34%(1500_to_3000m)	4.94%(3000_to_5000m)	0.35%(5000_to_13000m)
8	18	9	20	53	27	80.74%(0_to_1500m)	0.08%(1500_to_3000m)	0.01%(3000_to_5000m)	19.17%(5000_to_13000m)
9	27	24	17	45	39	96.42%(0_to_1500m)	0.18%(1500_to_3000m)	0.31%(3000_to_5000m)	3.09%(5000_to_13000m)
10	58	51	13	41	47	54.68%(0_to_1500m)	11.35%(1500_to_3000m)	30.49%(3000_to_5000m)	3.49%(5000_to_13000m)
11	65	34	34	48	18	44.64%(0_to_1500m)	49.24%(1500_to_3000m)	5.59%(3000_to_5000m)	0.53%(5000_to_13000m)
12	88	70	24	50	27	52.14%(0_to_1500m)	31.7%(1500_to_3000m)	6.87%(3000_to_5000m)	9.17%(5000_to_13000m)
13	94	51	31	42	28	56.07%(0_to_1500m)	32.48%(1500_to_3000m)	6.94%(3000_to_5000m)	4.51%(5000_to_13000m)
14	99	105	21	43	37	6.54%(0_to_1500m)	11.35%(1500_to_3000m)	45.4%(3000_to_5000m)	36.71%(5000_to_13000m)
15	84	58	19	39	42	42.39%(0_to_1500m)	8.16%(1500_to_3000m)	20.09%(3000_to_5000m)	29.36%(5000_to_13000m)

	RANG5	RI	POWER	Adresse	RSI
2	0.0%(13000_t	RI	40W		624
3	0.0%(13000_t	69.08	40W		632
4	0.2%(13000_t	44.71	40W		640
5	0.0%(13000_t	68.91	40W		312
6	0.04%(13000_	71.35	40W		320
7	0.0%(13000_t	66.84	40W		328
8	0.0%(13000_t	70.8	40W		768
9	0.0%(13000_t	53.65	40W		776
10	0.0%(13000_t	50.71	40W		784
11	0.0%(13000_t	82.72	40W		792
12	0.11%(13000_	61.26	40W		800
13	0.0%(13000_t	73.12	40W		808
14	0.0%(13000_t	99.94	40W		312
15	0.0%(13000_t	99.95	40W		320

Fig. 4.4: Échantillon du Dataset

4.5.2 Préparation de Données

Suite à la collecte de données, la préparation de données nommé pré-traitement, qui consiste à ajouter une classe de décision appelée ”target” comme une dernière colonne qui permet d'identifier la présence d'une anomalie. Les valeurs de la colonne ”target” sont discrètes : 1 si la qualité de service est mauvaise et 0 sinon(la qualité de service est bonne). Nous avons suivis dans cette étape une légende de saturation obtenue par notre entreprise AT, illustrée dans la figure 4.5.

		Valeur	Légende
1		1	PRB >= 80 ET USER >= 90
2		2	PRB >= 80 ET USER < 90 ET USER >= 20
3		3	PRB >= 80 ET USER < 20
4		4	PRB < 80 ET PRB >= 50 ET USER >= 90
5		5	PRB < 80 ET PRB >= 50 ET USER < 90 and USER >= 20
6		6	PRB < 80 ET PRB >= 50 ET USER < 20
7		7	PRB < 50 ET USER >= 90
8		8	PRB < 50 ET USER < 90 ET USER >= 20
9		9	PRB < 50 ET USER< 20

Fig. 4.5: Légende de saturation

Cette dernière nous a permet de fixer les conditions suivantes pour mettre en oeuvre le code qui nous permet de faire le remplissage des valeurs de notre classe ”target”. La figure 4.6 illustre le code utilisé.

$$A= \begin{cases} \textbf{Si } PRB \geq 80 \& User\_num \geq 90 & \Rightarrow \quad \text{Mauvaise QoS (1)} \\ \textbf{Sinon} & \Rightarrow \quad \text{Bonne QoS (0)} \end{cases}$$

$$B= \begin{cases} \textbf{Si } CQI\_LOW > a & \Rightarrow \quad \text{Mauvaise QoS (1)} \\ \textbf{Sinon} & \Rightarrow \quad \text{Bonne QoS (0)} \end{cases}$$

$$C= \begin{cases} \textbf{Si } RANG1 < RANG2 & \Rightarrow \quad \text{Mauvaise QoS (1)} \\ \textbf{Sinon} & \Rightarrow \quad \text{Bonne QoS (0)} \end{cases}$$

$$D= \begin{cases} \textbf{Si } RANG1 < RANG3 & \Rightarrow \quad \text{Mauvaise QoS (1)} \\ \textbf{Sinon} & \Rightarrow \quad \text{Bonne QoS (0)} \end{cases}$$

$$E= \begin{cases} \textbf{Si } RANG2 < RANG3 & \Rightarrow \quad \text{Mauvaise QoS (1)} \\ \textbf{Sinon} & \Rightarrow \quad \text{Bonne QoS (0)} \end{cases}$$

Avec :

$a = CQI\_MEDIUM + CQI\_EXCELENT$

Ajouter une classe de décision appelé "target"

```
[6]  #l'importation des packages
     import pandas as pd
     import numpy as np
```

```
[10] #chargement de data
     df = pd.read_excel('data3.xlsx')
     df
```

```
[11] #l'addition des deux cases ("CQI_EXELENT" et "CQI_EXELENT")
     a = df['CQI_EXELENT'] + df['CQI_MEDIUM']
```

```
[12] #les conditions
     conditions = [
         (df['PRB_75']>=80)& (df['User_num']>= 90),
         (df['CQI_LOW'] > a ),
         (df['RANG1']<df['RANG2']),
         (df['RANG1']<df['RANG3']),
         (df['RANG2']<df['RANG3'])
         ] ##
     choices = ['1','1','1','1','1']
     df ['target'] = np.select(conditions,choices,default= '0')
```

FIG. 4.6: Code de remplissage de la classe ”target”

Le nettoyage des données est une étape indispensable en machine learning. Pour assurer la qualité des données, nous avons procédé par 3 phases essentiels .

la première phase consiste à éliminer les colonnes vides et invalides qui contiennent des zéros existantes dans notre Dataset(l'existence de ces colonnes vides dans la Base de Données génère des erreurs et diminue l'efficacité de cette dernière). Nous avons utiliser les outils de logiciel Excel pour effectuer cette première tache.

La deuxième phase, consiste à la suppression des valeurs aberrantes, les chaînes de caractères et les pourcentages apparentes sur la figure 4.4. Nous avons effectué ce nettoyage avec un code sur google colab qui est illustré sur la figure 4.7.

Nettoyage des données

```
[151] #l'importation des packages
      import pandas as pd
      import numpy as np
      import math
```

```
[150] #lecture du data
      df = pd.read_excel("data.xlsx")
      df
```

```
[147] #supression des caractéres invalides
      for column in df.columns :
       df[column] = df[column].astype(str).str.replace(r"\(.*\)" ,"").astype(str).str.replace('\%','')
      print(df)
```

```
#suppression des colonnes vides et invalide
df.drop(['POWER','Adresse','RSI'], axis = 'columns',inplace=True)
df
```

FIG. 4.7: Nettoyage des données

La dernière phase, consiste à faire la conversion de valeurs des attributs de 2 jusqu'à 9, pour exprimer le pourcentage de ces derniers. En suite on a exporter dataset illustré sur la figure 4.8.

```
[89] #affichage des features
     df.columns=[x.upper() for x in df.columns]
     df.columns
```

```
[84] #multiplication des features (CQI_LOW,CQI_MEDIUM,CQI_EXCELENT,RANG1,RANG2,RANG3,RANG4,RANG5)avec 0.01
     df['CQI_LOW']= df['CQI_LOW']*0.01
     df['CQI_MEDIUM']= df['CQI_MEDIUM']*0.01
     df['CQI_EXELENT']= df['CQI_EXELENT']*0.01
     df['RANG1']= df['RANG1']*0.01
     df['RANG2']= df['RANG2']*0.01
     df['RANG3']= df['RANG3']*0.01
     df['RANG4']= df['RANG4']*0.01
     df['RANG5']= df['RANG5']*0.01
     df
```

```
[85] datatoexcel = pd.ExcelWriter('data (1).xlsx')
```

```
[86] df.to_excel(datatoexcel)
```

```
[87] datatoexcel.save()
```

```
print('DataFrame is Written to Excel File successfully.')

DataFrame is Written to Excel File successfully.
```

FIG. 4.8: Nettoyage des données(la suite)

La transformation de dataset est illustré sur la figure 4.9.

PRB_75	User_num	CQI_LOW	QI_MEDIU	QI_EXELEN	RANG1	RANG2	RANG3	RANG4	RANG5	RI	target
99	112	0,24	0,49	0,28	0,2808	0,1711	0,253	0,2952	0	69,08	1
54	37	0,1	0,44	0,47	0,9069	0,0572	0,0239	0,0101	0,002	44,71	0
27	14	0,29	0,42	0,29	0,556	0,2907	0,1507	0,0026	0	68,91	0
98	79	0,28	0,43	0,29	0,5697	0,1969	0,2211	0,0119	0,0004	71,35	1
77	58	0,19	0,48	0,33	0,8537	0,0934	0,0494	0,0035	0	66,84	0
18	9	0,2	0,53	0,27	0,8074	0,0008	0,0001	0,1917	0	70,8	0
27	24	0,17	0,45	0,39	0,9642	0,0018	0,0031	0,0309	0	53,65	1
58	51	0,13	0,41	0,47	0,5468	0,1135	0,3049	0,0349	0	50,71	0
65	34	0,34	0,48	0,18	0,4464	0,4924	0,0559	0,0053	0	82,72	0
88	70	0,24	0,5	0,27	0,5214	0,317	0,0687	0,0917	0,0011	61,26	0
94	51	0,31	0,42	0,28	0,5607	0,3248	0,0694	0,0451	0	73,12	0
99	105	0,21	0,43	0,37	0,0654	0,1135	0,454	0,3671	0	99,94	1
84	58	0,19	0,39	0,42	0,4239	0,0816	0,2009	0,2936	0	99,95	1
46	41	0,06	0,27	0,67	0,4877	0,4837	0,0269	0,0017	0	99,98	0

Fig. 4.9: Échantillon du Dataset préparé

4.6 Sélection des Facteurs

La sélection des caractéristiques (ou facteurs) est l'une des phases de pré-traitement des données les plus intéressantes, c'est une étape cruciale pour sélectionner les meilleures fonctionnalités avant de construire un modèle prédictif. Les modèles de sélection de facteurs tentent de réduire un jeu de données en supprimant les facteurs non pertinents ou redondants. Le processus de sélection des caractéristiques cherche à obtenir un ensemble minimal d'attributs, de sorte que les résultats des techniques d'exploration de données appliquées sur l'ensemble de données réduit soient aussi proches que possible (voire mieux) des résultats obtenus en utilisant tous les attributs. Cette réduction facilite la compréhension des motifs extraits et augmente la vitesse des étapes d'apprentissage postérieures et surtout la qualité des résultats. Ainsi ces techniques de sélection des facteurs peuvent être appliquées afin d'éliminer les facteurs qui faussent les résultats et garder que les facteurs les plus pertinents qui optimisent les résultats des algorithmes appliqués.

Le processus de sélection des facteurs pertinents est appliqué uniquement sur Dataset préparé à l'aide de la Matrice de Corrélation. les figures 4.10 et 4.11 et 4.12 représentent la code utilisé pour notre sélection.

```
[106] #importation des bibloithéques
      import pandas as pd
      import numpy as np
      import matplotlib.pyplot as plt

[107] #chargement de data
      df = pd.read_excel('data.xlsx')
      df

[108] feature_names = 'PRB_75', 'User_num', 'CQI_LOW' ,'CQI MIDIUM' ,'CQI EXCELENT' ,'RANG1' ,'RANG2', 'RANG3' ,'RANG4', 'RANG5','RI'
      target_name = 'target'

[109] x = df.drop("target",axis=1)
      y= df["target"]

      #séparation de data en ensemble de test et un ensemble d'entrainement
      from sklearn.model_selection import train_test_split
      x_train, x_test, y_train, y_test = train_test_split(
      x,
      y,
      test_size=0.3,
      random_state=0)
      x_train.shape, x_test.shape

      ((124, 11), (54, 11))
```

Fig. 4.10: Première partie de code de sélection

```
[95] #Avec la fonction suivante, nous pouvons sélectionner des fonctionnalités hautement corrélées,
     #cela supprimera la première fonctionnalité quiest corréllée avec toute autre fonctionnalité
     def correlation(dataset, threshold):
       col_corr = set()#ensemble de tous les noms de colonnes corrélées
       corr_matrix = dataset.corr()
       for i in range (len(corr_matrix.columns)):
         for j in range(i):
           if abs(corr_matrix.iloc[i, j]) > threshold:#on s'intéresse à la valeur du coeff
             colname = corr_matrix.columns[i]#obtenir le nom de la colonne
             col_corr.add(colname)
             return col_corr

[111] #Nombre des facteurs corrélés 7 avec un seuil 0.8
      corr_features = correlation(x_train , 0.8)
      len(set(corr_features))

      1

[101] #affichage du facteur hautement corrélé
      print('correlated_features are:' , corr_features)

      correlated_features are: {'CQI_EXELENT'}

[66] x_train.shape

     (124, 11)
```

Fig. 4.11: Deuxième partie de code de sélection

```
[65] x_train.shape

     (124, 11)

[105] x_train_noncorr = x_train.drop(corr_features,axis=1)

[68] x_train_noncorr.shape

     (124, 10)

[104] x_test_noncorr=x_test.drop(corr_features,axis=1)#la suppression de la colonne "CQI

     x_test_noncorr.shape

     (54, 10)
```

Fig. 4.12: Troisième partie de code de sélection

La Matrice de Corrélation illustré sur la figure 4.13 définie la dépendance entre chaque deux facteurs (attributs), à partir de cette matrice nous détectons les facteurs qui sont fortement corrélés, deux facteurs sont hautement corrélés si leur coefficient de corrélation est supérieur à un certain seuil (paramètre choisi). D'après notre matrice de corrélation, le facteur est le CQI_EXCELENT, suite au choix de la valeur 0.8 comme un seuil de corrélation.

```
[91] #Création de la matrice de corrélation
     import seaborn as sns
     plt.figure(figsize=(12,10))
     cor = x_train.corr()
     sns.heatmap(cor, annot=True, cmap=plt.cm.CMRmap_r)
     plt.show()
```

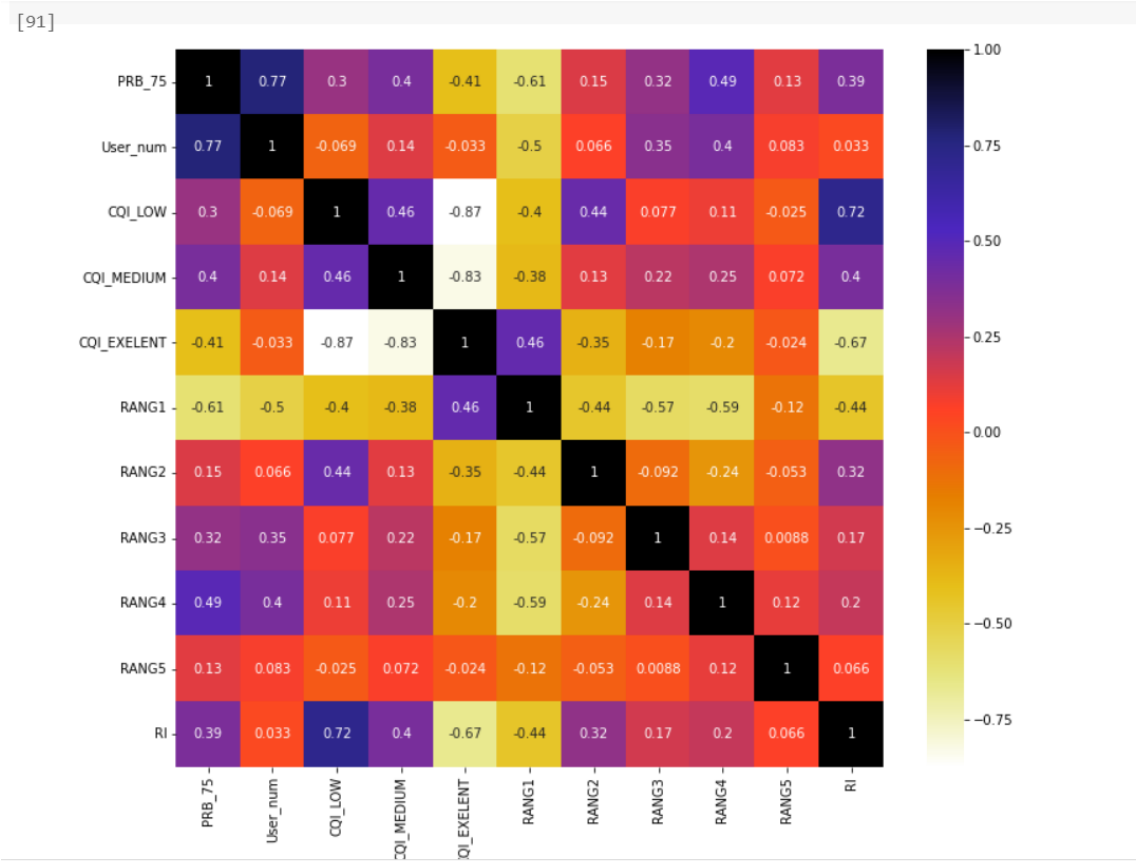


Fig. 4.13: La Matrice de Corrélation

4.7 Identification de type d'Apprentissage Supervisé

Nous avons déjà mentionné dans le chapitre Machine learning que les problèmes d'apprentissage supervisé peuvent être regroupés en problèmes de régression et de classification. Ses deux problèmes ont pour objectif la construction d'un modèle succinct qui peut prédire la valeur de l'attribut dépendant à partir des variables d'attribut.

En premier lieu, Un problème de régression se produit lorsque la variable de sortie est une valeur réelle ou continue.

En deuxième lieu, les problèmes de classifications se produisent lorsque la variable de sortie est une catégorie. Un modèle de classification tente de tirer des conclusions à partir des valeurs observées. Étant donné une ou plusieurs entrées, un modèle de classification tentera de prédire la valeur d'un ou plusieurs résultats.

Suite au Dataset préparé dans l'étape de prétraitement, vu le type de sortie que l'on attend du modèle est une valeur Discrète (catégorie) c'est à dire anomalie (l'abonnée bénéficie d'une mauvaise QoS) ou non anomalie (l'abonnée bénéficie d'une bonne QoS). Nous allons utilisé l'apprentissage automatique pour le problème de classification.

4.8 Modélisation

Dans la partie modélisation nous avons choisi 3 algorithmes de classifications,parmi ceux cités dans les chapitres précédents qui sont : SVM , Naïve Bayes, KNN. Ces algorithmes sont appliqués sur la Dataset résultante de la sélection des facteurs.

4.8.1 Naïve Bayes

Naïve Bayes est un algorithme de classification pour les problèmes de classification binaire (à deux classes) et multiclasses. On l'appelle Bayes naïf ou Bayes idiot parce que les calculs des probabilités pour chaque classe sont simplifiés pour rendre leurs calculs traitables. plutôt que d'essayer de calculer les probabilités de chaque valeur d'attribut, elles sont supposées être conditionnellement indépendantes compte tenu de la valeur de classe. Il s'agit d'une hypothèse très forte qui est très improbable dans des données réelles, c'est-à-dire que les attributs n'interagissent pas. Néanmoins, l'approche fonctionne étonnamment bien sur des données où cette hypothèse ne tient pas.

Avantage de modèle Naïve bayes : c'est l'algorithme de classification le plus simple et le plus rapide. Les calculs de probabilités ne sont pas coûteux d'où sa rapidité. Ce modèle d'apprentissage ne nécessite qu'un nombre faible d'échantillons d'entraînement par rapport à d'autres modèles, pour effectuer une classification efficace.

Il convient aussi aux larges ensembles de données.

Inconvénient de modèle Naive Bayes : il suppose que les variables sont indépendantes, ce qui n'est toujours pas vrai dans les cas réels.

En effet, si la corrélation entre les caractéristiques est grande, ce modèle d'apprentissage va donner une mauvaise performance.

En raison de l'hypothèse de la distribution normale, nous avons utilisé Gaussian Naive Bayes est utilisé car toutes nos caractéristiques sont continues. Dans la dataset, les caractéristiques sont L'utilisation de PRBs dans la bande passante, qualité de canal, la distribution d'abonnés , l'utilisation de MIMO. Ainsi, ses caractéristiques peuvent avoir des valeurs différentes dans l'ensemble de données car elles varient. Nous ne pouvons pas représenter les entités en fonction de leurs occurrences. Cela signifie que les données sont continues. Par conséquent, nous utilisons ici Gaussian Naive Bayes.

La figure 4.14 illustre le code de cette algorithme.

```
[282] #Charger l'ensemble de données
      df = pd.read_excel("data.xlsx")
      df

[283] #Diviser l'ensemble de données en valeurs d'entrées
      x=df.drop('target',axis=1)
      x

[284] #Diviser l'ensemble de données en valeurs de sorties
      y=df['target']
      y

[285] #Diviser les données en un ensemble d'entrainement et de test
      from sklearn.model_selection import train_test_split
      x_train, x_test, y_train, y_test = train_test_split(x ,y ,test_size=0.3,random_state=0)

[286] #Normalisation de données
      from sklearn.preprocessing import StandardScaler
      sc_x = StandardScaler()
      x_train = sc_x.fit_transform(x_train)
      x_test = sc_x.fit_transform(x_test)

[287] #Appeler le classificateur Naive Bayes
      from sklearn.naive_bayes import GaussianNB
      #Ajuster le modèle
      model = GaussianNB ()
      model.fit(x_train,y_train)

      GaussianNB()

[288] #Prediction sur l'ensemble de données de test
      y_pred = model.predict(x_test)

[289] y_pred

[290] y_test

[292] #import le module sklearn metrics pour le calcule de précision
      from sklearn.model_selection import cross_val_score
      cross_val_score(GaussianNB(), x_train, y_train,cv=7)

      array([0.77777778, 0.77777778, 0.66666667, 0.77777778, 0.83333333,
```

```
[33] #import le module sklearn metrics pour le calcule de précision
     from sklearn.model_selection import cross_val_score
     cross_val_score(GaussianNB(), x_train, y_train,cv=7)

     array([0.72222222, 0.72222222, 0.66666667, 0.77777778, 0.83333333,
            0.76470588, 0.64705882])

[34] from sklearn.metrics import accuracy_score
     accuracy_score(y_test, y_pred)

     0.5740740740740741

[35] #création de la matrice de confusion
     from sklearn.metrics import confusion_matrix
     cm= confusion_matrix(y_test,y_pred)

[36] cm

     array([[ 8, 23],
            [ 0, 23]])

[37] #Métrique d'évaluation
     print(cm)

     [[ 8 23]
      [ 0 23]]

[38] Precision = (cm[0,0]/(cm[0,0]+cm[0,1]))*100
     Precision

     25.806451612903224

[39] Rappel = (cm[1,1]/(cm[1,1]+cm[0,1]))*100
     Rappel

     50.0

[40] Fmesure =((2*Precision*Rappel)/(Rappel+Precision))
     Fmesure

     34.042553191489354
```

Fig. 4.14: Code Naïve Bayes

4.8.2 SVM

Les machines à vecteurs de supports sont une classe de méthodes d'apprentissage statistique basées sur le principe de la maximisation de la marge (séparation des classes). Il existe plusieurs formulations (linéaires, versions à noyaux) qui peuvent s'appliquer sur des données séparables (linéairement) mais aussi sur des données non séparables. La figure 4.15 illustre le code de cette algorithme [7].

Avantages de modèle SVM :

1. SVM fonctionne relativement bien lorsqu'il existe une marge de séparation claire entre les classes.
2. SVM est plus efficace dans les espaces de grande dimension.
3. SVM est efficace dans les cas où le nombre de dimensions est supérieur au nombre d'échantillons.
4. SVM est relativement économe en mémoire.

```
[4] #Charger l'ensemble des données
    df = pd.read_excel("data.xlsx")
    df
```

```
[5] #valeurs des caractéristiques
    x=df.drop('target',axis=1)
    x
```

```
[6] y=df['target']
    y
```

```
[7] len(df)
```

```
[8] #Diviser en données d'entrainement et de test
    from sklearn.model_selection import train_test_split
```

```
[9] x_train, x_test, y_train, y_test = train_test_split(x, y, test_size =0.30, random_state = 0)
```

```
[10] #Normalisation des données
     from sklearn.preprocessing import StandardScaler
     sc_x = StandardScaler()
     x_train = sc_x.fit_transform(x_train)
     x_test = sc_x.transform(x_test)
```

```
[11] #Appeler le classificateur SVM et Ajuster le modele
     from sklearn.svm import SVC
     clf = SVC(kernel= 'rbf', random_state = 0)
     clf.fit(x_train, y_train)

     SVC(random_state=0)
```

```
[12] #Prediction sur l'ensemble des données de test
     y_pred = clf.predict(x_test)
```

```
[13] y_pred

     array([0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1,
            0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0,
            1, 0, 0, 0, 0, 1, 0, 0, 0, 0])
```

```
[14] y_test
```

Fig. 4.15: Code SVM

4.8.3 KNN

L'algorithme des k plus proches voisins est une méthode d'apprentissage supervisé. Qui est utilisé pour la classification. Afin de faire une prédiction, cet algorithme ne construit pas de modèle prédictif. Cependant, il n'y a pas de phases d'apprentissage puis qu'il classifie directement des points dont la classe est inconnue en fonction de leurs distances par rapport à des points appartenant à une classe connue auparavant. Cet algorithme est ainsi classé dans la catégorie de lazy learning[55].

Dans la classification, KNN est basé sur le vote majoritaire des voisins. Un objet est classé par un vote majoritaire de ses voisins, l'objet étant attribué à la classe la plus commune, parmi ses k voisins les plus proches, où k est le numéro du voisin de l'objet. Le choix du paramètre k est très crucial dans cet algorithme dont le meilleur choix dépend des données. En général, des valeurs plus élevées de k réduisent l'influence du bruit sur la classification.

Suite au principe de choix de k, nous avons choisi K=7 comme illustré sur le code de la figure 4.16.

Avantage de modèle KNN :

1. Temps de calcul rapide.

2. Polyvalent – utile pour la régression et la classification.

3. Haute précision.

La figure 4.16 illustre le code de cette algorithme.

```
[111] #Charger les données
      df = pd.read_excel('data.xlsx')
      df

[112] #Appeler le classificateur KNeighbors
      from sklearn.neighbors import KNeighborsClassifier

[81] from sklearn.model_selection import train_test_split

[103] #Diviser l'ensemble de données en valeurs d'entrées
      x= df.iloc[:,[0,1,2,3,4,5,6,7,8,9]]
      print(x)

[104] #Diviser l'ensemble de données en valeurs de sortie
      y=df.iloc[:,10]
      print(y)

[105] #Diviser les données a un ensemble d'entrainement et de test
      x_train, x_test, y_train, y_test = train_test_split(x,y,random_state=0,test_size=0.30)

[113] #Appliquer le classificateur KNN avec un K=7
      knn = KNeighborsClassifier(n_neighbors=7, metric='euclidean')

[114] #Entrainer le modèle à l'aide des ensembles s'apprentissage
      model=knn.fit(x_train,y_train)

[87] print(model)
0s

      KNeighborsClassifier(metric='euclidean', n_neighbors=7)

[115] #Prédire la répense pour l'ensemble de données de test
      y_pred = knn.predict(x_test)

      y_pred
0s
```

```
[109] #importer le module sklearn metrics pour le calcule de précision
      from sklearn.metrics import accuracy_score
      ac = accuracy_score(y_test, y_pred)*100

[92] print(knn.score(x_test, y_test)*100)

     74.07407407407408

[93] #Création de la matrice de confusion
     from sklearn.metrics import confusion_matrix

[94] cm = confusion_matrix(y_pred,y_test)

[97] cm

     array([[27, 10],
            [ 4, 13]])

[91] #Métriques d'évaluation
     print(ac)

     74.07407407407408

[98] Precision=(cm[0,0]/(cm[0,0]+cm[0,1]))*100
     Precision

     72.97297297297297

[99] Rappel = (cm[1,1]/(cm[1,1]+cm[0,1]))*100
     Rappel

     56.52173913043478

[102] Fmesure=((2*Precision*Rappel)/(Rappel+Precision))
      Fmesure

      63.70235934664247
```

Fig. 4.16: Code KNN

4.9 Choix du Meilleur Modèle

Dans cette partie, nous allons commencer par définir une métrique d'évaluation.
Une métrique d'évaluation quantifie la performance d'un modèle prédictif. Le choix de la bonne métrique est donc crucial lors de l'évaluation des modèles de Machine Learning, et la qualité d'un modèle de classification dépend directement de la métrique utilisée pour l'évaluer. Pour les problèmes de classification, les métriques consistent globalement à comparer les classes réelles aux classes prédites par le modèle. Elles peuvent également permettre d'interpréter les probabilités prédites pour ces classes. L'un des concepts clés de performance pour la classification est la matrice de confusion dont nous avons donné les détails dans le chapitre précèdent, qui est une visualisation, sous forme de tableau, des prédictions du modèle par rapport aux vrais labels. Chaque ligne de la matrice de confusion représente les instances d'une classe réelles et chaque colonne représente les instances d'une classe prédite [14].

Ainsi pour une classe donnée [14] :

- Une précision et un rappel élevé ⇒ La classe a été bien gérée par le modèle. C'est le cas de l'algorithme SVM (Précision 78.37% et Rappel 65.21%).
- Une précision élevée et un rappel faible ⇒ La classe n'est pas bien détectée mais lorsqu'elle l'est, le modèle est très fiable. C'est le cas de l'algorithme de KNN (Précision 72.37% et Rappel 56.52%).
- Une précision faible et un rappel élevé ⇒ La classe est bien détectée, mais inclut également des observations d'autres classes. c'est le cas de l'algorithme Naïve Bayes (Rappel 50% et Précision 25.80%).
- Une précision et un rappel faible ⇒ La classe n'a pas du tout été bien gérée. Aucun des trois algorithme ne correspond.

Le tableau 4.1 représente les résultats des métriques d'évaluation de chaque modèle appliqué sur notre dataset. A partir de ce tableau nous constatons que le meilleur modèle est le machine à vecteurs de support (SVM) appliqué sur notre dataset qui est le résultat de l'application de la sélection des facteurs à l'aide de la matrice de corrélation avec un seuil égale à 0.8. Nous avons adopter le principe de dominance pour comparer entre les résultats des métriques d'évaluation des modèles précédents comme suit :

- L'exactitude du modèle SVM **domine strictement** celle de Naive Bayes et KNN

Car : (81.48% > 57.40% **et** 81.48% >74.07%).

- Précision du modèle SVM **domine strictement** celle de Naive Bayes et KNN

Car : (78.37%>25.80% **et** 78.37%>72.97%).

- Rappel du modèle SVM **domine strictement** celle de Naive Bayes et KNN

Car : (65.21%>50% **et** 65.21%>56.52%).

- Fmesure du modèle SVM **domine strictement** celle de Naive Bayes et KNN

Car : (71.19%>34.03% **et** 71.19% >63.70%).

Les résultats de ces derniers ont démontré que le modèle SVM est le plus fiable par rapport aux autres

Modèle / Métrique d'évaluation	Naive Bayes	S V M	K N N
Exactitude	57.40%	81.48%	74.07%
Précision	25.80%	78.37%	72.97%
Rappel	50%	65.21%	56.52%
Fmesure	34.04%	71.19%	63.70%

Tab. 4.1: Métriques d'évaluations des modèles utilisés sur la dataset

4.10 Conclusion

Tout travail doit être terminé par un fruit et doit atteindre un objectif, notre projet était de prédire si l'abonné a bénéficié d'une bonne qualité de service ou non. Tout en respectant les conditions de surveillance de comportement sur l'ensemble des indicateurs clés de performances(Features de dataset).
La prédiction a été effectué à l'aide des techniques d'apprentissage automatique, à base des algorithmes de cas supervisé pour le problème de classification. Nous avons appliqué ces trois modèles (Naïve Bayes,SVM, KNN). Les résultats de ces derniers ont démontré que le modèle SVM est le plut fiable par rapport aux autres.

Conclusion Générale

Au cours de ce travail au sein de l'entreprise Algérie Télécom, le département réseau d'accès nous a confié d'aborder la problématique de détection des anomalies sur le réseau cellulaire LTE pour une période de temps de 15 jours. Le problème étudié a porté sur la surveillance des indicateurs de performance de réseau LTE.
Pour ce faire, **le premier chapitre** est consisté pour une présentation de l'entreprise AT, pour définir la place de cette dernière par rapport au secteur d'activité de la télécommunication et surtout l'atout de l'entreprise en tant que lieu de stage pour une formation pratique.
Le deuxième chapitre a été consacré intégralement aux notions sur les réseaux de Télécommunication, particulièrement le LTE, ainsi à la détection des anomalies et ses différentes techniques appliqué auparavant.
Ensuite, **le troisième chapitre** nous a permis d'approfondir nos connaissances sur le domaine de machine learning et ses différents aspects. En expliquant la façon de le mettre en œuvre dès la phase de collecte de données jusqu'à la modélisation et la validation des modèles.
Enfin, **le quatrième chapitre** où nous avons présenté la phase finale de notre projet qui est l'application qui était de prédire si l'abonné a bénéficié d'une bonne qualité de service ou non. Tout en respectant les conditions de surveillance de comportement sur l'ensemble des indicateurs de performances (Features de dataset). La prédiction a été effectué à l'aide des techniques d'apprentissage automatique, à base des algorithmes de cas supervisé pour le problème de classification. Nous avons appliqué ces trois modèles (Naïve Bayes,SVM, KNN). Les résultats de ces derniers ont démontré que le modèle SVM est le plut fiable par rapport aux autres.
Les compétences acquises lors de ce travail apparaissent sur le développement de nos compétences en terme de langage de programmation en machine learning , et mieux se familiariser avec le langage python qui est un environnement de programmation connaissant un grand succès auprès des programmeurs.
Au final, nous espérons, par ce travail, avoir apporter des résultats satisfaisants pour l'entreprise Algérie Télécom.

Bibliographie

[1] 123calculus. Linear regression calculator. https://www.123calculus.com/en/linear-regression-page-1-50-140.html, 2022.

[2] Algeria, O. Qu'est-ce que l'ia ? en savoir plus sur l'intelligence artificielle. https://www.oracle.com/dz/artificial-intelligence/what-is-ai/#:~:text=l'intelligence%20artificielle-,Intelligence%20artificielle%20(IA)%20%2D%20Explication,des%20informations%20qu'ils%20recueillent., 2022.

[3] Amine, S., Dartigues-Pallez, C., and Gaetan, R. *CLASSIFICATION SUPERVISÉE DE DONNÉES PÉDAGOGIQUES POUR LA RÉUSSITE DANS L'ENSEIGNEMENT SUPÉRIEUR*. PhD thesis, I3S, Université Côte d'Azur, 2020.

[4] Blohorn, A. Concepts mathématiques derrière le machine learning : la régression linéaire. https://www.actuia.com/tutoriel/concepts-mathematiques-derriere-le-machine-learning-la-regression-lineaire/, 2019.

[5] Bouguen, Y., Hardouin, É., and Wolff, F.-X. réseaux 4g.

[6] CableFree. Lte rsrq à sinr -cablefree. https://www-cablefree-net.translate.goog/wirelesstechnology/4glte/lte-rsrq-sinr/?_x_tr_sl=auto&_x_tr_tl=fr&_x_tr_hl=fr&_x_tr_pto=wapp, 2020.

[7] Cedric.Cenam. Travaux pratiques - svm linéaires. https://cedric.cnam.fr/vertigo/cours/ml2/tpSVMLineaires.html, 2022.

[8] CETIC, M. T. Les librairies python pour l'analyse de données. https://colab.research.google.com/github/titsitits/Python_Data_Science/blob/master/5_Python_packages.ipynb#scrollTo=DfQ-d2onvooo, 2022.

[9] chahar (Analytics Vidhya), J. Support vector machine(svm) : A complete guide for beginners. https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/, 2021.

[10] CHRISWERNST. how dbscan can be implemented using python. https://github.com/chriswernst/dbscan-python, 2017.

[11] CLERC, L. Maintenance prédictive : une obligation pour l'industrie 4.0. https://larevuedestransitions.fr/2022/05/03/maintenance-predictive-une-obligation-pour-lindustrie-4-0/#:~:text=La%20maintenance%20prédictive%20consiste%20à,à%20l%27Internet%20des%20Objets., 2022.

[12] COBBAÏ. Vocabulaire de l'intelligence artificielle | les mots à connaître. https://www.cobbai.com/blog/vocabulaire-de-lintelligence-artificielle, 2019.

[13] COHERIS. Qu'est-ce la data science ? https://ia-data-analytics.fr/datascience/definition/, 2022.

[14] DATASCIENTEST. Comment gérer les problèmes de classification déséquilibrée ? partie ì. https://datascientest.com/comment-gerer-les-problemes-de-classification-desequilibree-partie-i?fbclid=IwAR3aG_sgQ-Z7mxgAhD4WOIMsvuZh4XbHxVq70jluNUCusDp4_1dI-F3H3Tw, 2022.

[15] DISPLAYR. What is a correlation matrix ? https://www.displayr.com/what-is-a-correlation-matrix/, 2022.

[16] ELECTRONICS-NOTE. 3gpp specification release numbers electronics notes. https://www.electronics-notes.com/articles/connectivity/3gpp/standards-releases.php.

[17] FATMA, D., AND FAIZA, O. *Planification et optimisation d'un réseau 4G LTE*. PhD thesis, Université Mouloud Mammeri, 2017.

[18] FAURE, L. L'évolution des réseaux mobiles : de la 2g à la 5g. https://itsocial.fr/actualites/levolution-reseaux-mobiles-de-2g-a-5g/, 2018.

[19] FIL, T. S. Le système de test rs ts8980 vérifie les indicateurs qualité lte cqi, pmi et ri. https://cdn.rohde-schwarz.com/magazine/pdfs_1/article/203/NEWS_203_french_TS8980.pdf.

[20] FOULON, L. *Détection d'anomalies dans les flux de données par structure d'indexation et approximation : Application à l'analyse en continu des flux de messages du système d'information de la SNCF*. PhD thesis, Université de Lyon, 2020.

[21] FUTURA. Deep learning : qu'est-ce que c'est ? https://www.futura-sciences.com/tech/definitions/intelligence-artificielle-deep-learning-17262/, 2021.

[22] GEEKSFORGEEKS. Ml | optics clustering explanation. https://www.geeksforgeeks.org/ml-optics-clustering-explanation/, 2019.

[23] GHILAS, B., AND YOUVA, F. *Etude et optimisation des ressources d'un réseau LTE*. PhD thesis, Université Mouloud Mammeri, 2018.

[24] Hakim, Z. *Les performances de la 4G dans la transmission de la voix sur IP*. PhD thesis, Université Mouloud Mammeri, 2017.

[25] HAYES, A. Stepwise regression. https://www.investopedia.com/terms/s/stepwise-regression.asp, 2022.

[26] IONOS, D. G. Qu'est-ce qu'un réseau de neurones artificiels ? https://www.ionos.fr/digitalguide/web-marketing/search-engine-marketing/quest-ce-quun-reseau-neuronal-artificiel/, 2020.

[27] ISMAILI, Z. Aprentissage supervisé vs. non supervisé - analytics insights. https://analyticsinsights.io/apprentissage-supervise-vs-non-supervise/, 2022.

[28] javatpoint. Kdd- knowledge discovery in databases. https://www.javatpoint.com/kdd-process-in-data-mining, 2021.

[29] johnson, D. Data science tutorial for beginners : Learn basics in 3 days. https://www.guru99.com/data-science-tutorial.html, 2022.

[30] Knowledge, T. Lte kpi. https://telecom-knowledge.blogspot.com/2016/09/lte-kpi.html, 2016.

[31] Kraiem, I. B. *Détection d'anomalies multiples par apprentissage automatique de règles dans les séries temporelles*. PhD thesis, Université de Toulouse-Jean Jaurès, 2021.

[32] L, B. Reinforcement learning : qu'est-ce que l'apprentissage par renforcement ? https://www.lebigdata.fr/reinforcement-learning-definition, 2021.

[33] LeanBI. Les 5 plus grandes difficultés de la détection d'anomalies. https://leanbi.ch/fr/blog/5-grandes-difficultes-de-la-detection-danomalies/, 2017.

[34] Mbengue, M. Clustering avec l'algorithme dbscan. https://penseeartificielle.fr/clustering-avec-lalgorithme-dbscan/, 2019.

[35] Minitab. Using stepwise regression and best subsets regression. https://support.minitab.com/en-us/minitab/21/help-and-how-to/statistical-modeling/regression/supporting-topics/basics/using-stepwise-regression-and-best-subsets-regression/, 2021.

[36] Mishra), K. A. Feature-selection. https://www.kaggle.com/getting-started/170842, 2020.

[37] Mobilis, A. Mobilis. https://www.mobilis.dz/apropos.php#:~:text=Filiale%20du%20Groupe%20Télécom%20Algérie,devenu%20autonome%20en%20août%202003., 2022.

[38] Netmanias. Architecture réseau lte. https://www.netmanias.com/en/post/techdocs/5904/lte-network-architecture/lte-network-architecture-basic, 2013.

[39] Numérique, D. T. Scikit-learn : guide de démarrage rapide en machine learning avec python. https://www.data-transitionnumerique.com/scikit-learn-python/, 2022.

[40] Osisanwo, F. Y., AKINSOLA, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., and Akinjobi. Supervised machine learning algorithms : Classification and comparison. https://ir.tech-u.edu.ng/344/, 2017.

[41] Rosner, B. Percentage points for a generalized esd many-outlier procedure. *Technometrics 25*, 2 (1983), 165–172.

[42] Salem, M. Eps bearer in lte. https://mobilepacketcore.com/eps-bearer-lte/, 2018.

[43] ScienceDirect. Knowledge discovery in database. https://www.sciencedirect.com/topics/computer-science/knowledge-discovery-in-database, 2022.

[44] Scientist, D. Google colab : Le guide ultime. https://ledatascientist.com/google-colab-le-guide-ultime/, 2022.

[45] sebastianraschka. What is the difference between filter, wrapper, and embedded methods for feature selection ? https://sebastianraschka.com/faq/docs/feature_sele_categories.html, 2013.

[46] Team, D. S. Confusion matrix. https://datascience.eu/machine-learning/confusion-matrix/, 2020.

[47] Techno-Science. Qos - définition et explications. https://www.techno-science.net/definition/11442.html, 2022.

[48] Telecom, A. Présentation du groupe. https://www.algerietelecom.dz/fr/page/le-groupe-p2.

[49] Telecom, A. Idoom adsl. https://www.algerietelecom.dz/fr/particuliers/idoom-adsl-prod3, 2022.

[50] Togbe, M. U., Boly, A., Chiky, R., et al. Etude comparative des méthodes de détection d'anomalies. *Revue des Nouvelles Technologies de l'Information* (2020).

[51] Valentin(Datascientist). La régression logistique, qu'est-ce que c'est ? https://datascientest.com/regression-logistique-quest-ce-que-cest, 2022.

[52] Wikipedia. Djaweb-wikipedia. https://fr.wikipedia.org/wiki/Djaweb, 2022.

[53] Wikipedia. Mobilis-wikipedia. https://fr.wikipedia.org/wiki/Mobilis, 2022.

[54] Wikipedia. Télécommunication-wikipedia. https://fr.wikipedia.org/wiki/Djaweb, 2022.

[55] Ysance. Algorithme n°5 - comprendre la méthode des "k-plus proches voisins" en 5 min. https://blog.ysance.com/algorithme-n5-comprendre-la-methode-des-k-plus-proches-voisins-en-5-min, 2022.