



République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche
Scientifique



Université des Sciences et de la Technologie Houari Boumediene

Faculté d'Informatique
Département Systèmes Informatiques
Mémoire de Master
Spécialité :
Big Data Analytics

Thème :

**CONCEPTION ET RÉALISATION D'UNE
SOLUTION DÉCISIONNELLE POUR LA
PRODUCTION DE GAZ ET PREDICTIONS.
CAS D'ÉTUDES : SONATRACH**

Sujet proposé par :

Mr. BOUSSADA Redouane

Réalisé par :

GHEZALI Rayane
SEDDI Mohamed Reda

Encadré par :

Mme. AZZOUZ Mahdia

Membre du Jurys :

Mr. SELMOUNE Nazih

Mme. MEKAHLIA Fatima Zohra

Soutenu le :

03/07/2023

Remerciements

Nous remercions ALLAH le tout puissant de nous avoir guidé et de nous avoir donné la santé, le courage et la force de mettre en œuvre notre travail. Sans lui rien de tout cela n'aurait pu être.

Nous adressons notre profonde gratitude à notre promotrice **Mme. AZZOUZ Mahdia** de nous avoir encadrés et dirigés, pour les critiques enrichissantes et ses précieux conseils, pour son engagement et son dévouement, pour la confiance et le soutien qu'elle nous a accordé tout au long de ce stage.

Nous tenons aussi à remercier **Mr. BOUSSADA Redouane**, notre encadreur pour sa confiance en nos efforts, nos idées et nos propositions et l'opportunité d'apprendre et d'évoluer dans un environnement professionnel. Nous tenons également à le remercier pour sa disponibilité et son suivi, pour les efforts considérables qu'il a consacrés pour valider chaque étape réalisée.

Un merci très particulier à l'équipe **SONATRACH - Division Associations** en l'occurrence **Mr. LARACHICHE Nassim** et **BOUSTILLA Abdelkader**, ingénieurs Production, pour leur gentillesse, leur coopération, leur partage et leur précieuse aide dans l'explication de la production, ainsi que les employés que nous avons rencontré.

D'autre part, nos vives considérations vont droit au cadre administratif et pédagogique de la faculté de Mathématiques et d'Informatique pour les efforts fournis durant notre formation.

A nos chers parents qui nous ont éclairés le chemin et qui nous ont toujours tirés vers le haut. Nos plus sincères remerciements pour le soutien qu'ils nous ont apporté durant toutes ces années d'études. Qu'ils trouvent dans ce travail le fruit de leurs sacrifices.

A nos amis les plus fidèles, à nos chers camarades et à toute personne qui a participé de près ou de loin à l'accomplissement de ce projet.

Et enfin, un énorme merci pour les membres du jury d'avoir fait l'honneur d'évaluer ce modeste travail.

Table des matières

INTRODUCTION GÉNÉRALE	10
1 GÉNÉRALITÉS	12
1.1 Définition de la BI	12
1.2 Processus décisionnel général	12
1.2.1 Etape de collecte / transformation / alimentation	13
1.2.2 Etape de stockage et de modélisation	13
1.2.3 Etape de restitution / distribution	13
1.2.4 Etape d'exploitation / analyse	13
1.3 Entrepôt de données Data Warehouse	14
1.4 Magasin de données (Data Mart)	14
1.5 Tableau de bord	14
1.6 Modelisation multidimensionnelle	14
1.6.1 Définition	14
1.6.2 Concepts de base	15
1.6.3 Types de modélisation	15
1.7 Définition de Data Mining	16
1.8 Objectifs de Data Mining	16
1.9 Processus de Data Mining	16
1.9.1 Compréhension du contexte métier	17
1.9.2 Compréhension des données	17
1.9.3 Préparation des données	18
1.9.4 Modélisation	18
1.9.5 Evaluation	18
1.9.6 Déploiement	18
1.10 Approches du Data Mining	18
1.10.1 Approche descriptive	18
1.10.2 Approche Prédictive	19
1.11 Data Mining et Machine Learning	19
1.11.1 Définition de Machine Learning	19
1.11.2 Apport du Machine Learning au Data Mining	19
1.11.3 Description de certains algorithmes	19

1.12 Définition Prévision	24
1.13 Définition d'une série chronologique :	24
1.14 Les composantes une série temporelle :	24
1.15 Schéma de décomposition d'une série chronologique :	25
1.16 Séries chronologique stationnaires :	25
1.16.1 Fonctions Autocovariance et Autocorrélation :	25
1.17 Transformations des séries chronologique	26
1.18 Description de certains Modèles	26
1.18.1 Seasonal Autoregressive Integrated Moving Average (SARIMA)	26
2 PRÉDICTIONS DANS LE DOMAINE PÉTROLIER	29
2.1 Concepts de base	29
2.2 Types de Gaz	30
2.3 Prévision de la production pétrolière en utilisant Time Series Analysis	31
2.3.1 Travaux réalisés	31
2.3.2 Synthèse	32
2.4 Le ML dans le domaine pétrolier	33
2.4.1 Méthodes utilisées	33
3 ETUDE DE L'EXISTANT	35
3.1 Description de l'organisme	35
3.1.1 La Sonatrach	35
3.1.2 Organigramme de structure d'accueil de la SONATRACH	36
3.1.3 Organigramme structure d'accueil Division Association (AST)	36
3.2 Collecte d'informations et analyse des besoins	37
3.2.1 Etude des sources de données	37
3.2.2 Recueil des besoins techniques	38
3.2.3 Acteurs du système	39
3.2.4 Recueil des besoins analytiques	39
3.3 Méthode DCA utilisée par l'entreprise pour la prévision	39
3.4 Description de la solution proposée	40
3.4.1 Architecture de la solution système décisionnel	40
3.4.2 Architecture de la solution prédition	40
4 CONCEPTION	42
4.1 Architecture générale de la solution	42
4.1.1 Volet BI	43
4.1.2 Volet Prédition	43

4.2	Conception de la zone de stockage DM	44
4.2.1	Choix de l'approche de modélisation	44
4.2.2	Choix d'Indicateurs	44
4.3	Modèle multidimensionnel	44
4.3.1	Définition du grain	44
4.3.2	Définition des dimensions	45
4.3.3	Définition des niveaux et des hiérarchies	45
4.3.4	Définition de la table de fait	46
4.3.5	Schéma multidimensionnel	46
4.4	Conception de la zone d'alimentation	47
4.4.1	Etude des sources de données	47
4.4.2	Processus ETL	48
4.5	Conception de la zone de restitution	49
4.6	Conception de la solution prévisionnelle avec Time Series Analysis	49
4.6.1	Compréhension des données	49
4.6.2	Préparation des données	50
4.6.3	Modélisation	50
4.6.4	Évaluations	50
4.6.5	Déploiement	51
4.7	Conception de la solution prédictive en utilisant le Machine Learning	51
4.7.1	Compréhension des données	51
4.7.2	Préparation des données	52
4.7.3	Modélisation	52
4.7.4	Evaluation	52
4.7.5	Déploiement	53
5	IMPLÉMENTATION ET MISE EN OEUVRE	54
5.1	Présentation des outils utilisés	54
5.1.1	Outils de stockage : SQL Server	54
5.1.2	Outils de collecte : SSIS	54
5.1.3	Outils de restitution : Power BI	54
5.1.4	Langage de programmation : Python	55
5.2	Volet BI	56
5.2.1	Mise en œuvre de l'ETL	56
5.2.2	Mise en œuvre du tableau de bord	69
5.3	Volet Prédition	71
5.3.1	Prévision avec Time Series Analysis :	71
5.3.2	Solution Prédition avec le Machine Learning	88
CONCLUSION GÉNÉRALE		97

Table des figures

1.1	Processus Business Intelligence [14]	13
1.2	Schéma conceptuel en étoile [4]	15
1.3	Schéma conceptuel en flocon [4]	16
1.4	Processus CRISP-DM [44]	17
1.5	Approches de Data Mining [1]	18
1.6	Diagramme de Support Vector Regressor	20
1.7	Gradient Boosting Regressor	22
3.1	Logo de la SONATRACH	35
3.2	Organisme de Macrostructure de SONATRACH	36
3.3	Organigramme structure d'accueil Division Association (AST)	36
4.1	Schéma représentatif de la solution proposée	43
4.2	Schéma du magasin de données	46
5.1	Préavisons des groupement GTFT / GTIM Brutes	56
5.2	Previsions du groupement GSS	57
5.3	Production du Groupement Berkine	57
5.4	Contrats par périmètre	58
5.5	Production du groupement GSS	58
5.6	Exemple d'un rapport journalier GTIM	59
5.7	L'emplacement de Date dans le fichier	59
5.8	Fichier final comportant les 3 groupement	61
5.9	Schéma de la Base de données GTFT	62
5.10	Remplisage de la table Puit GTFT	63
5.11	Création des Package	64
5.12	Remplissage de la dimension Bassin	64
5.13	Conversions du type dans le fichier Excel global	65
5.14	Selection des ID Bassin Uniques	65
5.15	Remplissage Dimension Contrats	66
5.16	Recherche par correspondance	67
5.17	Jointure avec la dimension Réservoirs	67
5.18	Jointure de fusion GSS	68
5.19	Jointure de fusion GTFT et GTIM	68

TABLE DES FIGURES

5.20 Remplissage de la table Fait	69
5.21 Visualisation globale du groupement GTFT en 2020	69
5.22 Visualisation du suivi de production du puits « ABBF-2 » en 2020	70
5.23 7 janvier 2020 : Prévision journalière du Groupement GTIM atteinte, cumul mensuel prévu non atteint	70
5.24 Graphe de la Production de Gaz mensuelle de 2000 à 2022 . . .	71
5.25 Seasonal-trend décomposition	72
5.26 Comparaison entre la série originale et son estimation (Trend * Saisonnalité)	73
5.27 Box plot des résidus	73
5.28 Détection par intervalle inter-quartile	74
5.29 présentation des anomalies	74
5.30 Graphe ACF et PACF	75
5.31 Moyenne mobile et écart type mobile	75
5.32 Série ajustée	76
5.33 la série après diff (1)	77
5.34 La série temporelle stable	77
5.35 ACF et PACF de la série stationnaire	78
5.36 Modèle SARIMA Estimé	79
5.37 Graphe des résidus de l'estimation	81
5.38 Graphe ACF and PACF des résidus	81
5.39 Distribution des résidus	81
5.40 Estimation avec SARIMA	82
5.41 Estimation avec Holt-Winters Exponential Smoothing	83
5.42 Prévision du test avec SARIMA	83
5.43 Prévision du test avec Holt-Winters Exponential Smoothing . .	84
5.44 Ajustement des 2 modèles.	85
5.45 Prévisions du test	85
5.46 Prévisions du Test des deux modèles	86
5.47 Prévisions moyenne de production de gaz (Mai 2022 à Avril 2023) et intervalle d'incertitude.	87
5.48 Prévisions moyenne de production de gaz (Mai 2022 à Avril 2023)	87
5.49 Dataset de production journalière par puits	88
5.50 Pourcentage de valeurs nulles	88
5.51 Distribution des paramètres	89
5.52 Production du Puits 024 après suppression des valeurs aberrantes	89
5.53 Distribution des valeurs de production de gaz par puits	90
5.54 Production de gaz par puits	91
5.55 Distribustion des paramètres après ajustement des valeurs aberrantes	91

5.56 Matrice de corrélation globale	92
5.57 Exemple recherche des meilleurs paramètres par Grid Search . .	93
5.58 Exemple Entrainement du modèle Random Forest Regressor . .	94
5.59 Support Vector Regression Model Predictions	95
5.60 Decision Trees Regression Model Predictions	95
5.61 Random Forest Regression Model Predictions	95
5.62 Gradient Boosting Regression Model Predictions	96
5.63 Extreme Gradient Boosting Regression Model Predictions XG-BOOST	96
5.64 Previsions des groupement GTFT / GTIM finaux	99
5.65 Fichier final de Production Berkine	99
5.66 Ficher final de production par le groupement GSS	99
5.67 Données partagées par l'entreprise	100
5.68 Répertoire des rapports journaliers 2018 du groupement GTIM .	100
5.69 Fichier d'informations sur GTIM	101
5.70 Fichier final de production GTIM	101
5.71 La table CATEGORY	101
5.72 La table DPRD	102
5.73 Les Contrats du groupement GTFT	102
5.74 Suivi de production de Gaz Sec, GPL, Condensat du groupement GTFT pour l'année 2020	102
5.75 Taux de production de Gaz Sec par réservoir du groupement GTIM pour l'année 2020	103
5.76 Taux de production de Gaz Sec par périmètre du groupement GTIM pour l'année 2020	103
5.77 Liste des partenaires et taux de production de Gaz Sec par périmètre en 2020	103
5.78 Taux de production de Oil par type de puits en 2020	104
5.79 Nombre de lignes par puits	104
5.80 Production de gaz par puits	105
5.81 Nombre de lignes par puits après suppression des valeurs aberrantes	105
5.82 Aperçu des données standardisées	106

Liste des tableaux

5.1	Production mensuelle de gaz de l'année 2000	71
5.2	Résultats du test de Dickey-Fuller de la série	76
5.3	Dickey Fuller Test de la série stationnaire	77
5.4	Résultats SARIMA	80
5.5	Coefficient de détermination SARIMA	82
5.6	Coefficient de détermination Holt-Winters Exponential Smoothing	83
5.7	Calcul des mesures de SARIMA	84
5.8	Calcul des mesures de Holt-Winters Exponential Smoothing . .	84
5.9	Résultats des modèles SARIMA et Holt-Winters	85
5.10	Résultats des modèles SARIMA et DCA	86
5.11	Tableau des Prévisions de production de gaz (Mai 2022 à Avril 2023)	87
5.12	Paramètres des modèles et leurs and Valeurs	94
5.13	Model Evaluation Metrics	96

Liste des Acronymes

BI Business Intelligence

KPI Key Performance Indicator

OLAP Online Analytical Processing

ETL Extract-Transform-Load

ML Machine Learning

KDD Knowledge Discovery in Databases

SEMMA Sample, Explore, Modify, Model, Assess

CRISP-DM Cross-Industry Standard Process for Data Mining

SVM Support Vector Machine

SVR Support Vector Regression

DTR Decision Tree Regression

RFR Random Forest Regression

GBR Gradient Boosting

XGBOOST Extreme Gradient Boosting

ACF Autocorrelation Function

ARIMA Autoregressive Integrated Moving Average

SARIMA Seasonal Autoregressive Integrated Moving Average

POD Plan de développement

GOR Gas-Oil Ratio

BHP Bottom Hole Pressure

WHP Well Hole Pressure

WHT Well Hole Temperature

AIC Akaike Information Criterion

DCA Decline Curve Analysis

INTRODUCTION GÉNÉRALE

Le domaine pétrolier est un secteur économique crucial pour l'énergie mondiale. Il est basé sur l'exploration, l'extraction, la production, le raffinage et la commercialisation de pétrole brut, de gaz naturel et de produits dérivés tels que l'essence, le diesel et le kérósène. Le pétrole est une source d'énergie essentielle pour la plupart des industries et des économies modernes. Il est utilisé pour alimenter les transports, la production d'électricité, le chauffage et la fabrication de nombreux produits.

La Business Intelligence (BI) est devenue un outil essentiel dans de nombreux secteurs, y compris l'industrie pétrolière. Les entreprises pétrolières doivent faire face à une complexité croissante des données, provenant de différentes sources telles que les opérations d'exploration et de production, les marchés financiers, les réglementations gouvernementales et les exigences en matière d'environnement.

Les techniques de BI permettent aux entreprises pétrolières de collecter, organiser et analyser ces données afin de prendre des décisions éclairées et de maximiser leur efficacité opérationnelle et leur rentabilité. Parmi les techniques de BI utilisées dans l'industrie pétrolière, on peut citer la modélisation des données, l'analyse prédictive, la visualisation de données, la gestion de la performance et la gestion de la relation client.

L'apprentissage automatique (ML) est tout aussi important. En effet, l'exploitation pétrolière implique une quantité importante de données qui peuvent être exploitées pour optimiser les processus de production, réduire les coûts et améliorer la sécurité. Le machine learning offre la possibilité d'analyser ces données en temps réel, de détecter des tendances et de prédire des événements futurs.

Sonatrach accumule de grandes quantités de données à travers ses systèmes d'informations, mais ces données ne deviennent des informations pertinentes pour les décideurs que lorsqu'elles permettent de répondre aux besoins de l'entreprise et à son évolution. Pour être aidés dans leurs choix, les décideurs ont besoin de mesurer l'activité de production dans l'entreprise à l'aide d'indicateurs de performance (KPI) et avoir un outil de prédiction de la production pétrolière.

Notre objectif est de concevoir et implémenter un système d'aide à la décision. Cette solution comporte deux volets. Le premier volet consiste à la mise en place d'un tableau de bord qui détaille les indicateurs relatifs à l'activité de production pétrolière de Sonatrach. Le deuxième volet est relatif aux domaines de data mining et machine Learning, où nous faisons recours à ces domaines pour développer un outil de prédiction de production pétrolière.

Ce présent mémoire est structuré en deux grandes parties principales comme suit :

- Une partie théorique organisée en deux chapitres :

- **CHAPITRE 1 : Généralités** Ce chapitre est consacré à la présentation des concepts fondamentaux liés au domaine d'aide à la décision et les approches prédictives du Data Mining.
- **CHAPITRE 2 : Prédictions dans le domaine pétrolier** Cette partie traite la problématique de notre présente recherche et définit le contexte du projet pour permettre une meilleure compréhension de notre thème.
- Une partie dédiée à la pratique :
- **CHAPITRE 3 : Etude de l'existant** Ce chapitre évoque dans un premier temps, l'analyse de l'existant, les axes d'analyse à traiter dans le nouveau système et le problème exposé par l'organisme d'accueil. Et dans un second, la proposition de notre solution pour l'amélioration de la prédiction.
- **CHAPITRE 4 : Conception** Ce chapitre est dédié à l'architecture conceptuelle de la solution proposée : en premier lieu, la conception de la zone de stockage, la zone d'alimentation, la zone de restitution, pour en finir avec la prédiction.
- **CHAPITRE 5 : Réalisation** Ce dernier chapitre présente les outils utilisés et décrit de façon détaillée les différentes étapes d'implémentation de notre système d'aide à la décision et la démarche suivie pour la mise en œuvre de la prédiction par le biais des séries chronologiques Times Series et des modèles de Machine Learning.

Chapitre 1

GÉNÉRALITÉS

Introduction

L'informatique décisionnelle (Business Intelligence) s'intéresse fondamentalement au passé même s'il est très récent et se contente de fournir un état de ce qui existe ou a existé. L'Intelligence Artificielle, au contraire, est un outil qui a pour mission principale d'analyser, automatiser, optimiser et prédire.

A la frontière entre les statistiques, l'intelligence artificielle et l'informatique, le data mining vise à extraire les informations dans de grands volumes de données, découvrir des tendances et des modèles qui vont au-delà de la simple analyse, et à préparer, manipuler et analyser les données dans le but de les transformer en connaissance actionnable et en outil d'aide à la décision pour les entreprises.

Dans ce premier chapitre, nous allons exposer le processus général de la BI et ses outils, les séries chronologiques et leurs concepts de base et enfin le concept de machine learning et ses apports au data mining.

1.1 Définition de la BI

L'informatique décisionnelle (Business Intelligence) est un ensemble de moyens, méthodes, outils permettant de collecter, consolider, modéliser, restituer les données matérielles ou immatérielles d'une entreprise, en vue d'offrir une aide à la décision et de permettre aux responsables de la stratégie d'entreprise et de son opérationnalisation d'avoir une vue d'ensemble de l'activité traitée. [40]

Selon Inmon (Inmon B. , 1996) un système d'information décisionnel est défini comme étant «un regroupement de données orientées vers certains sujets, intégrées, dépendantes du temps, non volatiles, ayant pour but d'aider les gestionnaires dans leurs prises de décision ». [27]

1.2 Processus décisionnel général

Avant d'explorer les bonnes pratiques pour réussir son projet, il est important de rappeler qu'un projet de business intelligence doit respecter les étapes qui permettent de structurer la chaîne décisionnelle :

1.2.1 Etape de collecte / transformation / alimentation

Cette première étape consiste à collecter, nettoyer et consolider les données de l'entreprise venue de différentes sources de manière pertinente et cela en utilisant les outils d'ETL (Extract Transform Load). C'est à dire que ces données récupérées doivent être filtrées et adaptées en vue d'une utilisation à vocation décisionnelle. [40]

1.2.2 Etape de stockage et de modélisation

Les données résultantes de la 1ère phase sont structurées, centralisées et stockées dans le Datawarehouse (ou Data Mart qui est une version plus réduite du DW). Ce dernier doit être non volatile, orienté métier, historisé et intégré afin de préparer les données à l'analyse décisionnelle. [40]

1.2.3 Etape de restitution / distribution

Il est nécessaire de pouvoir restituer les données et d'en proposer un accès aisément en prenant en compte chaque profil et besoin métier. Cette étape inclut notamment les rapports, statistiques générés, outils de reporting ad hoc ou de masse, tableaux de bord, outils de navigation dans les cubes OLAP (ou hypercubes)...

1.2.4 Etape d'exploitation / analyse

Une fois les données collectées, nettoyées, stockées, et accessibles elles peuvent être analysées pour faire ressortir des prévisions ; ou des estimations futures en utilisant les outils du data-mining. Selon les besoins, différents types d'outils d'extraction et d'exploitation seront utilisés tels que :

- OLAP pour les analyses multidimensionnelles, notamment analyser les données.
- Le Datamining pour rechercher des corrélations.
- Les tableaux de bord présentant les indicateurs clés de l'activité.
- Le Reporting pour communiquer la performance. [40]

Figure 1.1 ci-dessous résume les étapes du processus BI :

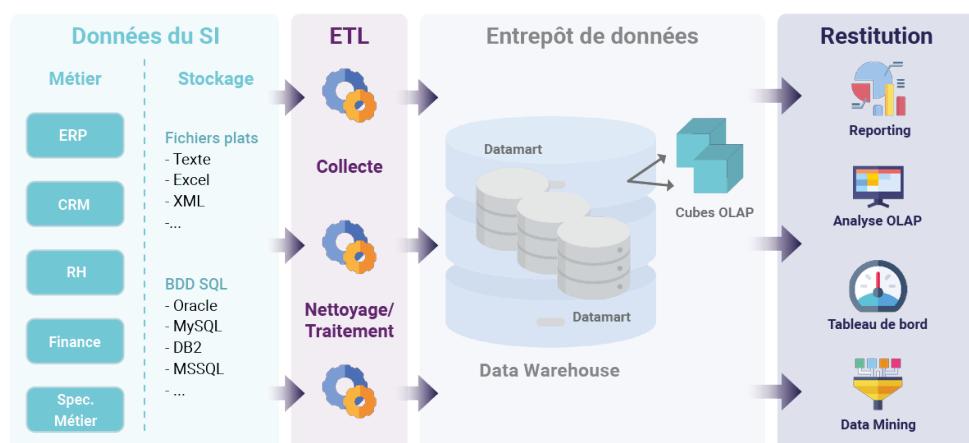


FIGURE 1.1 – Processus Business Intelligence [14]

1.3 Entrepôt de données Data Warehouse

Selon **Ralph Kimball** : « Le Data Warehouse est un système qui extrait, traite et rend conforme des données sources vers un espace de stockage multidimensionnel. Il permet ainsi la mise en oeuvre de l'interrogation et de l'analyse à des fins d'aide à la prise de décision ». [27]

On peut également le définir comme une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d'aide à la décision.

1.4 Magasin de données (Data Mart)

Selon **Ralph Kimball**, un Data Mart appelé également « magasin de données », définit comme étant : « Un sous-ensemble du Data Warehouse, constitué de tables au niveau détail et à des niveaux plus agrégés, permettant de restituer tout le spectre d'une activité métier. L'ensemble des data marts de l'entreprise constitue le Data Warehouse ». [40]

On peut le définir comme étant une version plus réduite d'un DW conçu pour répondre à un besoin métier spécifique. [27] En effet, plusieurs Data Mart peuvent être utilisés au sein d'une même entreprise : un DM marketing, DM ressources humaines, DM production,etc...

1.5 Tableau de bord

Selon **Laurent GRANGER** : « Un tableau de bord a pour fonction de permettre la visualisation, le suivi et l'exploitation facile de données pertinentes sous forme de chiffres, ratios et de graphiques. Ces indicateurs (appelés aussi KPI) sont reliés à des objectifs dans le but de prendre des décisions ». [37]

Selon **Alain FERNANDEZ**, les KPI sont classés par type d'information transmise et par l'objectif pour lequel ils ont été utilisés tels que les indicateurs d'alerte signalant un état anormal du système, les indicateurs d'équilibration qui indique l'état actuel du système et les indicateurs d'anticipation qui fournissent des informations de prévision.

Un KPI doit regrouper les caractéristiques SMART suivantes :

- **Spécifique** : Il doit être clairement défini et bien précis.
- **Mesurable** : Il doit pouvoir se traduire en une donnée quantifiée et mesurable.
- **Acceptable** : Il doit contenir des données faciles à obtenir et fiables pour des objectifs atteignables orientés action, permettant aux utilisateurs d'envisager des actions correctives.
- **Réaliste** : Il doit être pertinent et aligné avec les objectifs de l'entreprise et sa stratégie.
- **Temporellement défini** : Il est pourtant indispensable de soumettre les objectifs à un cadre temporel.

1.6 Modelisation multidimensionnelle

1.6.1 Définition

La modélisation multidimensionnelle est une technique qui vise à présenter et organiser les données sous format standardisé, de telle sorte que les applications des analyses multidimensionnelles OLAP soient performantes et efficaces. [32]

Contrairement à la modélisation relationnelle, la modélisation multidimensionnelle est plus habilitée pour répondre aux besoins d'analyse requis en matière de décision. En effet dans cette modélisation les données sont organisées de manière à mettre en évidence le sujet et les perspectives d'analyse, en les représentant avec deux concepts fondamentaux qui sont respectivement les faits et les dimensions.

1.6.2 Concepts de base

Fait

Un fait est un centre d'intérêt décisionnel. Il regroupe un ensemble d'attributs numériques représentant les mesures d'activité. La table de fait : est la table primaire qui modélise le sujet de l'analyse dans un modèle dimensionnel et où les mesures de performance correspondant aux informations de l'activité analysée sont stockées. Elle est ainsi formée des clés vers une liste de dimensions définissant le grain de la table. [17]

- Fait additif
- Fait semi-additif
- Fait non-additif

Dimension

Les tables de dimension représentent les axes d'analyse selon lesquels sont visualisées les mesures d'activité d'un sujet d'analyse. Elles contiennent des informations sur l'entreprise et ses activités et comportent de nombreuses colonnes et attributs. Ces attributs décrivent les lignes de la table de dimension. [40]

Les attributs/membres d'une dimension sont organisés suivant des hiérarchies : chaque membre appartient à un niveau hiérarchique (ou niveau de granularité) particulier.

1.6.3 Types de modélisation

- **En étoile :** Représentation dénormalisée initiée par **Ralph Kimball**. Elle est représentée par une table de fait central, connectée à plusieurs tables de dimensions, ces dernières n'ont aucune liaison entre elles, permettant d'analyser les faits. Sa force réside dans sa lisibilité et sa capacité d'assurer un haut niveau de performance des requêtes, même sur de gros volume de données. Cependant ce modèle présente une grande redondance des données. [40] La Figure 1.2 ci-après représente le schéma en étoile :

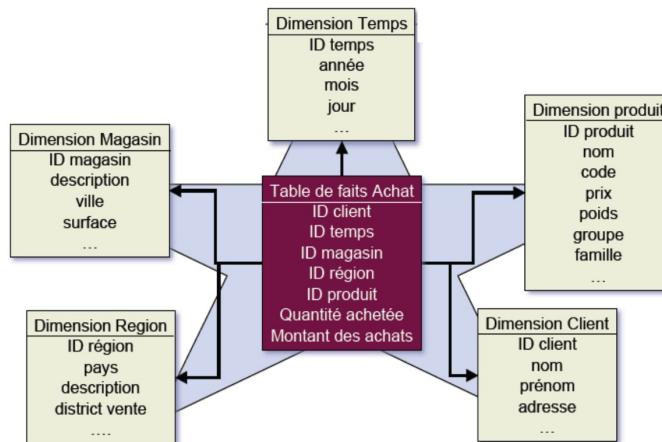


FIGURE 1.2 – Schéma conceptuel en étoile [4]

- **En Flocon de neige :** C'est un dérivé du modèle en étoile, il vient pour réduire les redondances du modèle en étoile. Toutefois les tables dimensions sont normalisées, les redondances sont éliminées, ce qui permet d'économiser l'espace de stockage. Mais il est très couteux en termes de performance et moins lisible. [40]

La Figure 1.3 ci-après représente le schéma en flocon :

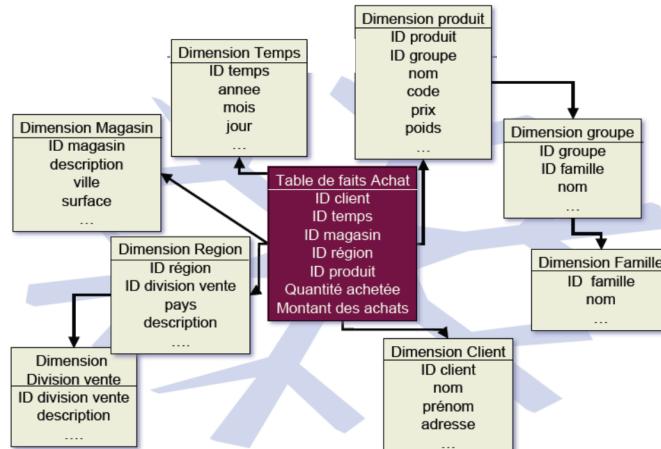


FIGURE 1.3 – Schéma conceptuel en flocon [4]

- **En constellation de faits :** c'est une série de schéma en étoile, il représente une fusion de modèles en étoile ayant des dimensions communes. Il contient par conséquent un ensemble de tables de faits et des tables de dimensions parfois communes.

1.7 Définition de Data Mining

Le Data Mining se traduit par fouille de données ou forage de données, se définit selon **Vercellis Carlo** comme étant : « un processus itératif visant l'analyse de grandes bases de données, dans le but d'extraire des informations et des connaissances qui peuvent s'avérer exactes et potentiellement utiles pour les travailleurs du savoir engagés dans la prise de décision et la résolution de problèmes ». [46]

1.8 Objectifs de Data Mining

Les objectifs de Data Mining sont résumés en :

- L'exploratation autant que possible l'essentiel d'information disponible et faire sortir les informations dissimulées. [18]
- Permettre de mieux comprendre les liens entre les données, découvrir les tendances et anticiper l'avenir. [42]

1.9 Processus de Data Mining

Le premier processus de base (**Knowledge Discovery in Databases -KDD**) a été proposé par **Fayyad**. [48] Ce processus est constitué de plusieurs étapes qui s'exécutent de manière séquentielle. Chaque étape commence une fois que les précédentes sont bien finalisées, parce qu'elle utilise les sorties des étapes précédentes et dans le cas où les résultats ne sont pas satisfaisants, certaines étapes peuvent être refaites. [41]

En partant du modèle de Fayyad, les efforts ont été orientés pour trouver d'autres processus de Data Mining, par exemple, ceux de **SEMMA** [5] et **CRISP-DM** [22].

CRISP-DM (Cross-Industry Standard Process for Data Mining) [22], formé par les compagnies NCR, SPSS, et Daimler-Benz est un moyen éprouvé par l'industrie pour guider l'exploration des données.

CRISP-DM ne propose pas un chemin linéaire unique entre le démarrage du projet et le déploiement, cette méthodologie a la particularité d'adopter une démarche cyclique et itérative, semblable à celle du modèle Agile. L'ensemble du cycle du projet est itératif, et la marche arrière est non seulement autorisée, mais recommandée : les ajustements en cours de route sont les bienvenus, puisqu'ils permettent de garder le modèle efficace. [22]

Figure 1.4 résume les différentes étapes du Crisp-DM :

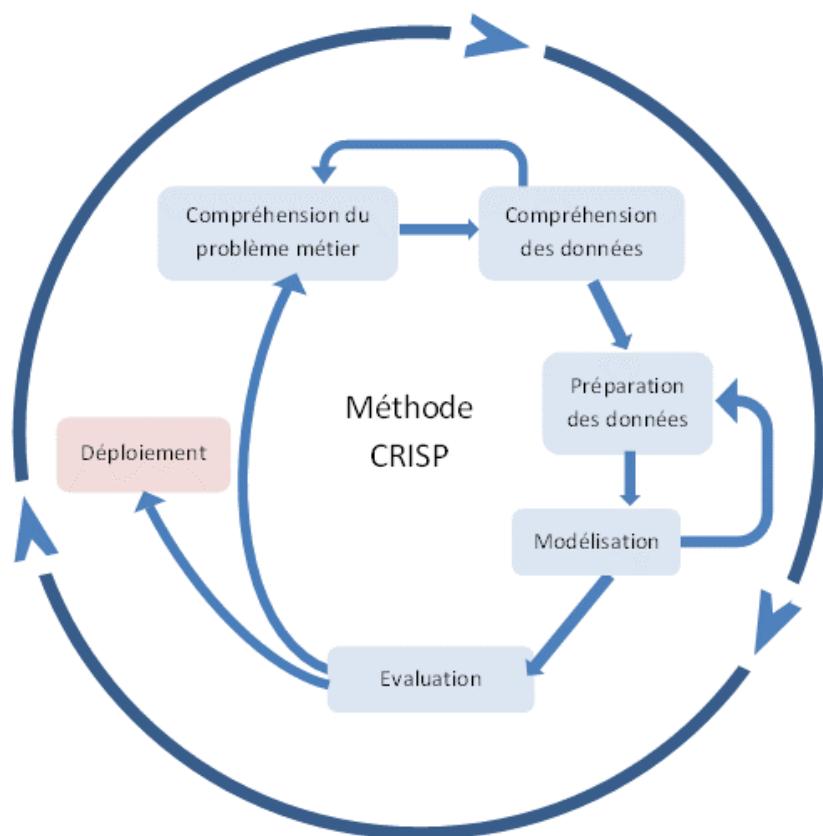


FIGURE 1.4 – Processus CRISP-DM [44]

1.9.1 Compréhension du contexte métier

Cette étape implique la définition des besoins de l'entreprise, les objectifs à atteindre à la fin du projet et la mise en place du plan de la solution.

1.9.2 Compréhension des données

La deuxième étape consiste à collecter, analyser et comprendre les données.

1.9.3 Préparation des données

Une fois les données collectées et analysées, cette étape consiste à nettoyer, transformer, remplacer les valeurs manquantes afin de les préparer pour l'étape de modélisation.

1.9.4 Modélisation

En utilisant les données préparées, cette étape consiste à sélectionner des techniques de modélisation et générer une conception de test. Il est souvent nécessaire de revenir à l'étape de préparation des données car certaines techniques nécessitent des formats de données spécifiques.

1.9.5 Evaluation

Cette étape consiste à évaluer les différents modèles en se basant sur des mesures de performance et les comparer entre eux afin de procéder au déploiement final du modèle le plus performant et être certain qu'il atteint correctement les objectifs de l'entreprise.

1.9.6 Déploiement

Cette dernière étape du processus se résume à présenter et organiser les résultats obtenus sous une forme adaptée aux utilisateurs finaux dans l'environnement opérationnel afin de maintenir son utilisation continue.

1.10 Approches du Data Mining

Les approches descriptives et prédictives sont utilisées dans le Data Mining pour extraire les types de modèles à utiliser. Figure 1.5 ci-dessous démontre les différentes approches du Data Mining :

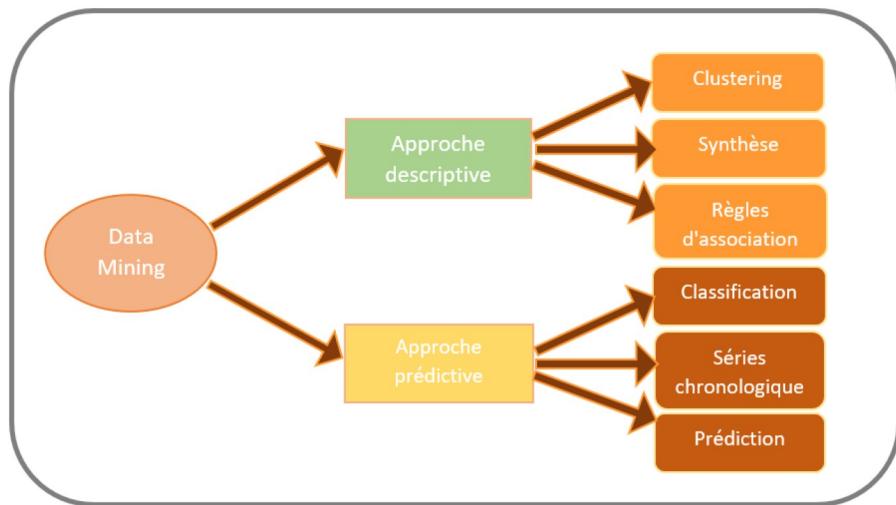


FIGURE 1.5 – Approches de Data Mining [1]

1.10.1 Approche descriptive

L'analyse descriptive ou les statistiques décrivent ou résument des données brutes qui décrivent des événements qui se sont produits dans le passé. Les analyses descriptives sont utiles car elles nous permettent d'apprendre des comportements passés et de comprendre comment

ils pourraient influencer les résultats futurs [22].

Les tâches du Data Mining descriptif peuvent être divisées en trois types : **Clustering, Synthèse et Règles d'association**.

1.10.2 Approche Prédictive

L'analyse prédictive a pour but de prédire ce qui pourrait arriver à l'avenir. Malgré le fait qu'aucun algorithme statistique ne peut prédire l'avenir avec une certitude à 100%, L'analyse prédictive fournit donc des estimations de la probabilité d'un résultat futur.

Les tâches du Data Mining prédictif peuvent être divisées en trois types : **la classification, les séries chronologiques et la prédition**. [22]

1.11 Data Mining et Machine Learning

D'après S. Kelly : « Le datamining est un processus qui applique les techniques de l'intelligence artificielle dans le but de découvrir des modèles au sein des données ».

1.11.1 Définition de Machine Learning

Le Machine Learning est la science qui permet aux ordinateurs d'apprendre et d'agir comme les humains et d'améliorer leur apprentissage au fil du temps de manière autonome, en leur fournissant des données et des informations sous forme d'observations et d'interactions réelles [22].

L'apprentissage automatique peut être supervisé ou non supervisé. L'apprentissage supervisé repose sur des ensembles de données étiquetés, tandis que l'apprentissage non-supervisé s'effectue à l'aide d'ensembles de données non étiquetés.

1.11.2 Apport du Machine Learning au Data Mining

Le Machine learning est essentiellement un processus d'analyse de données destiné à améliorer les performances d'un système, il représente donc, une opportunité pour les entreprises pour permettre de mieux utiliser de très grandes bases de données, de les visualiser en contexte et de prendre de meilleures décisions [22].

D'une façon plus scientifique, le concept de Machine learning est, par définition, un ensemble d'algorithmes capables de construire des modèles sans les paramétriser. Il est défini comme une amélioration des capacités d'analyse grâce à l'apprentissage automatique. Alimenté par une multitude de données, le Machine learning effectue des analyses statistiques qui permettent de prédire un résultat. En d'autres termes, il est capable d'apprendre à partir d'informations existantes grâce aux algorithmes et aux modélisations [22].

1.11.3 Description de certains algorithmes

Parmi les algorithmes de régression les plus populaires et les plus utilisés, on cite :

Support Vector Regressor SVR

Support Vector Machine (SVM) est une méthode de classification supervisée, qui a été initialement identifiée par Cortes et al. [15].

A la différence de ce problème, où les valeurs de sortie sont des étiquettes de classe (valeurs -1 ou 1), SVR est une extension de SVM qui lui permet de gérer les problèmes de régression(Figure 1.6).

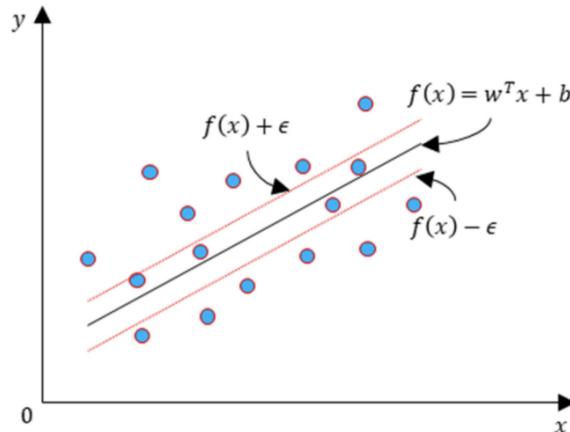


FIGURE 1.6 – Diagramme de Support Vector Regressor

Les valeurs de sortie dans ce cas sont des valeurs réelles. C'est un problème plus difficile, car il faut apprendre à prédire les valeurs d'une fonction à partir d'un nombre fini d'échantillons.

Decision Tree Regressor DTR

Comme son nom l'indique, un arbre de décision [38] aide à prendre une décision grâce à une série de questions.[42]

Cette méthode est choisie pour sa simplicité d'implémentation et son faible coût de calcul. Néanmoins, l'arbre de régression est un peu plus complexe que l'arbre de classification. La régression de l'arbre de décision divise de façon optimale les données en sections appelées feuilles en utilisant la valeur du seuil pour répondre à une série de questions.

Chaque question correspond à un noeud, c'est-à-dire à un endroit où une branche se sépare en deux branches. En fonction de la réponse à chaque question, nous allons nous orienter vers telle ou telle branche de l'arbre pour finalement arriver sur une feuille de l'arbre (ou extrémité) qui contiendra la réponse à notre question.

Enfin, l'algorithme prend la moyenne des points de feuilles du terminal (Y) dans chaque split, donc quand un nouveau point arrive (x_1, x_2, \dots) le modèle prédit ses résultats avec la valeur moyenne de (Y) avec l'équation suivante :

$$\hat{y} = \sum_{i=1}^M c_i * Id \quad (1.1)$$

où :

N nombre total de feuilles dans l'arbre

I fonction indicatrice

Random Forest Regressor RFR

Les Random Forest [8] sont composées de plusieurs sous-ensembles (arbres) aléatoirement constitués d'échantillons (d'où le « Random » dans Random Forest). [16]

Le nombre d'arbres est ajusté généralement par validation croisée (cross-validation en anglais) pour entraîner et tester le modèle sur des morceaux du dataset de départ.

Chaque arbre est entraîné sur un sous-ensemble du dataset et donne une réponse. Les résultats de tous les arbres de décision sont alors combinés pour donner une réponse finale. C'est ce que l'on appelle une méthode de **bagging** :

Le Bagging est une technique en intelligence artificielle qui consiste à assembler un grand nombre d'algorithmes avec de faibles performances individuelles pour en créer un beaucoup plus efficace. Les algorithmes de faible performance sont appelés les « weak learners » et le résultat obtenu « strong learner ». [21] L'idée derrière cet algorithme est que plusieurs petits algorithmes peuvent être plus performants qu'un seul grand algorithme.

Les « weak learners » peuvent être de différentes natures et avoir des performances variées, mais ils doivent être indépendants les uns des autres. L'assemblage des « weak learners » (arbustes) en « strong learner » (forêt) se fait par votation. C'est-à-dire que chaque « weak learner » va émettre une réponse (un vote), et la prédiction du « strong learner » sera la moyenne de toutes les réponses émises. De cette manière on construit un modèle robuste à partir de plusieurs modèles qui ne sont pas forcément aussi robustes.

Dans les tâches de régression, la prévision aléatoire des forêts est générée en utilisant la prédiction moyenne $f_i(x)$ des N arbres de régression. L'équation de Random Forest Regression s'écrit comme suit :

$$\hat{y} = 1/N \sum_{i=1}^N (f_i(x)) \quad (1.2)$$

où :

N nombre total d'arbres du modèle

$f_i(x)$ prédiction de l'arbre i pour l'observation x

Gradient Boosting Regressor GBR

Les algorithmes de Boosting se basent sur le même principe que ceux de Bagging. La différence apparaît lors de la création des « weak learner ». [22]

Pour le Boosting, les algorithmes ne sont plus indépendants. Au contraire, chaque « weak learner » est entraîné pour corriger les erreurs des « weak learner » précédents. Ainsi, si l'on reprend l'image de la forêt, ici chaque arbre a été soigneusement planté pour rendre la forêt plus résistante aux intempéries.

L'algorithme de Gradient Boosting [23] est un ensemble de « weak learners », créés les uns après les autres, formant un strong learner. De plus, chaque « weak learner » est entraîné pour corriger les erreurs des « weak learners » précédents. Néanmoins, contrairement à Adaboost, les « weak learners » ont tous autant de poids les uns que les autres, peu importe leur performance.

Le premier « weak learner » (w_1) est très basique, il s'agit tout simplement de la moyenne des observations. Il est donc très peu efficace, mais il va servir de base au reste de l'algorithme.

Par la suite, c'est l'écart entre cette moyenne et la réalité que nous appelons premier résidu. De manière générale, on appellera ici résidu l'écart entre la prédiction de l'algorithme en cours de création et la réalité.

La particularité de Gradient Boosting est qu'il essaye de prédire à chaque étape non pas les données elles-mêmes mais les résidus. Ainsi, le second « weak learner » est entraîné pour prédire le premier résidu. Figure 1.7 ci-dessous explique le fonctionnement de Gradient Boosting Regressor : Les prédictions du second weak learner sont ensuite multipliées par un facteur

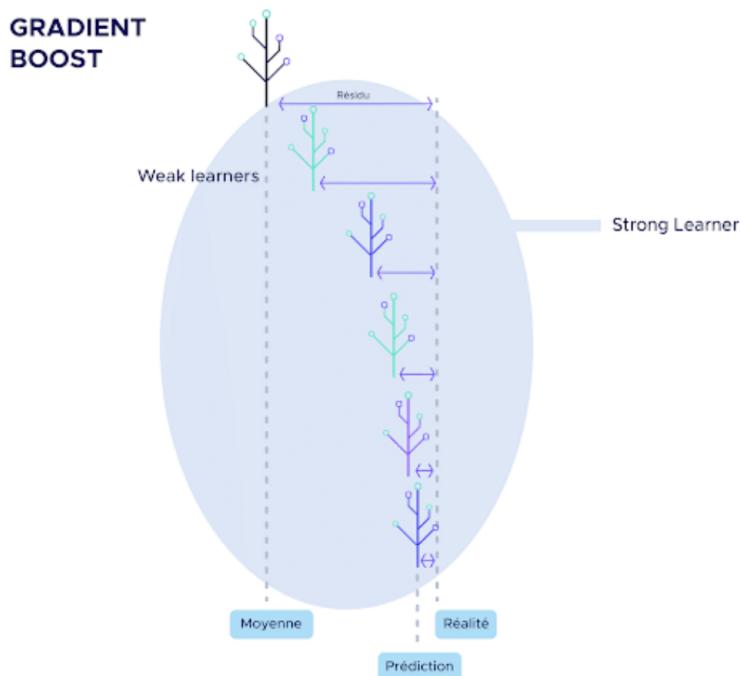


FIGURE 1.7 – Gradient Boosting Regressor

inférieur à 1. L'idée derrière cette multiplication est que plusieurs petits pas sont plus précis que quelques grands pas. La multiplication réduit donc la taille des « pas » pour augmenter la précision. L'objectif étant « d'écartier » petit à petit les prédictions du modèle de la moyenne, pour les rapprocher de la réalité.

Pour demander la prédiction de Gradient Boosting sur une observation, il suffit d'interroger chaque « weak learner ». La prédiction du « strong learner » sera la somme de toutes les réponses des « weak learner ».

Extreme Gradient Boosting Regressor XGBOOST

XGBoost [10] est une sorte d'arbre de décision boosté par gradient (GBM) qui est informatiquement optimisé pour rendre les différents calculs nécessaires à l'application d'un Gradient Boosting rapides.

Le modèle est enseigné dans une approche incrémentale. En effet, XGBoost fonctionne de manière similaire à un arbre de décision en ce qu'il crée un nombre spécifique d'arbres en fonction des questions, mais il le fait un par un, chaque arbre suivant utilisant les connaissances

obtenues par l'arbre précédent pour l'améliorer. Autrement dit, tout nouvel arbre corrigera les erreurs causées par l'arbre précédent. [22]

Après chaque étape de croissance de l'arbre, le rétrécissement ajuste les poids récemment ajoutés par un ratio. La réduction, comme un taux d'apprentissage dans l'optimisation stochastique, diminue l'influence de chaque arbre tout en permettant aux arbres futurs d'améliorer le modèle.

Les arbres qui ne sont pas assez bons sont “élagués”, c'est à dire qu'on leur coupe des branches, jusqu'à ce qu'ils soient suffisamment performants. Sinon ils sont complètement supprimés. Cette méthode est appellée le « **pruning** » (élagage). Ainsi, XGBoost s'assure de ne conserver que de bons weak learners.

Lorsque le nombre de fonctionnalités de l'ensemble de formation est inférieur au nombre d'observations de l'ensemble de formation, ou si l'ensemble de données contient exclusivement des fonctionnalités numériques, XGBoost est utilisé.

Synthèse

En résumé, **Support Vector Regression (SVR)** peut gérer efficacement les données non linéaires grâce à l'utilisation de noyaux et donne de bons résultats avec un petit nombre d'échantillons. Cependant, la sélection du bon noyau et des hyperparamètres peut être difficile et le temps d'exécution risque d'être lent pour des ensembles de données de grande taille. [19]

Decision Tree Regressor est facile à comprendre et à interpréter et peut gérer à la fois des données numériques et catégoriques. Néanmoins, sa performance peut être limitée par rapport à des modèles plus avancés.

Random Forest Regressor peut gérer efficacement les données avec un grand nombre de variables et d'observations et contrairement aux arbres de décision individuels, il est moins susceptible de surajuster (overfitting) de l'ensemble d'entraînement. [19]

Cependant, il peut être plus difficile à interpréter qu'un simple arbre de décision et peut être moins précis et plus lent que certaines méthodes d'apprentissage plus avancées.

Le **Gradient Boosting Regressor (GBR)** quant à lui peut gérer les données manquantes sans prétraitement supplémentaire et peut capturer des relations complexes entre les variables grâce à la combinaison d'arbres de décision faibles. [19]

Néanmoins, son temps d'entraînement peut être plus long que les autres algorithmes et nécessite un ajustement minutieux des hyperparamètres pour obtenir de bonnes performances.

XGBoost (Extreme Gradient Boosting), reconnu pour être l'algorithme gagnant dans tous les domaines, est rapide, précis et efficace, permettant une souplesse de manœuvre inédite sur le Gradient Boosting. [19]

En effet, XGBoost propose un panel d'hyperparamètres très important. Il est ainsi possible grâce à cette diversité de paramètres, d'avoir un contrôle total sur l'implémentation du Gradient Boosting. Toutefois, il peut nécessiter plus de ressources computationnelles que certains autres algorithmes.

Le point commun entre ces algorithmes est qu'ils peuvent être sensibles aux valeurs aberrantes et aux valeurs extrêmes dans les données d'entraînement. La performance des algorithmes peuvent varier en fonction de la nature des données et de l'application spécifique. Il faut tester plusieurs algorithmes et de sélectionner celui qui donne les meilleurs résultats pour une tâche donnée. [19]

1.12 Définition Prévision

Il est difficile de faire des prédictions, surtout concernant l'avenir.

NEILS BOHR, Physicien Danois

Une prévision consiste à prédire un ou plusieurs événements futurs. Comme l'a suggéré Niels Bohr, faire de bonnes prévisions n'est pas toujours facile.

Les prévisions jouent un rôle crucial dans de nombreux domaines, mais les défis varient selon l'horizon temporel considéré. Les méthodes statistiques sont couramment utilisées pour les prévisions à court et moyen terme, en tirant parti des données historiques.

La plupart des problèmes de prévision impliquent l'utilisation de données de séries temporelles. Elles regroupent des ensembles de données organisées chronologiquement, où chaque observation est associée à un moment précis.

L'analyse des séries temporelles permet de découvrir des tendances, d'anticiper les comportements futurs et de prendre des décisions éclairées en exploitant les modèles et les structures sous-jacentes aux données chronologiques. [33]

1.13 Définition d'une série chronologique :

Une série temporelle est une séquence d'observations prises de manière séquentielle dans le temps. Ces ensembles de données, tels qu'une quantité mensuelle de marchandises expédiées par une usine, le nombre hebdomadaire d'accidents de la route ou les précipitations quotidiennes, représentent des séries temporelles. Une caractéristique essentielle des séries temporelles réside dans la dépendance entre les observations adjacentes. L'analyse des séries temporelles se concentre sur les techniques permettant de comprendre cette dépendance. [7]

La valeur actuelle à l'instant t de la chronique est notée y_t . Ici, t représente le temps, qui varie entre 1 et n , où n correspond au nombre total d'observations de la chronique. Le nombre de points ou de valeurs à prévoir dans la chronique est désigné par h . L'horizon de prévision désigne la prévision de la série temporelle, c'est-à-dire de $(n + 1)$ à $(n + h)$, en se basant sur l'historique de y_1 à y_n . [33]

1.14 Les composantes une série temporelle :

Les méthodes traditionnelles de traitement des séries chronologiques opèrent par décomposition suivie de reconstitution de la série temporelle afin de réaliser des prévisions. Cette approche suppose que la structure de la série temporelle peut être décomposée en éléments simples (pouvant être modélisés) et, par conséquent, plus facilement prévisibles, pour ensuite être reconstituée et donner la prévision de la série temporelle. [6]

Les techniques traditionnelles de traitement des chroniques procèdent par décomposition, on peut considérer de manière classique trois grandes composantes :

- **La tendance**, également appelée **Trend** et notée N_t , qui est censée décrire le mouvement à long terme, fondamental ou structurel du phénomène. Ce mouvement est traditionnellement représenté par des formes analytiques simples : polynomiales, logarithmiques, exponentielles.
- **La composante saisonnière**, notée S_t , qui est une composante cyclique relativement régulière de période annuelle et qui correspond souvent à des phénomènes de mode, de coutumes, de climat...
- **La composante résiduelle**, notée R_t , qui regroupe tout ce que les autres éléments n'ont pas pu expliquer du phénomène observé. Elle contient donc de nombreuses fluctuations, notamment accidentelles, dont la nature est exceptionnelle et imprévisible, telles que les catastrophes naturelles, les grèves, les guerres... Étant donné que ces événements sont supposés être corrigés, le résidu présente généralement une allure aléatoire, plus ou moins stable autour de sa moyenne. [6]

1.15 Schéma de décomposition d'une série chronologique :

Dans le domaine des entreprises, les composantes sont maintenues constantes, mais elles peuvent parfois varier (par exemple, de manière hebdomadaire). La technique de décomposition-recomposition repose, bien entendu, sur un modèle qui la justifie. Ce modèle est connu sous le nom de schéma de décomposition. [6]

Il existe principalement trois grands types de schéma :

- **Le schéma additif**, qui suppose l'orthogonalité (l'indépendance) des différentes composantes. Il s'exprime ainsi :

$$y_t = N_t + S_t + R_t \quad (1.3)$$

- **Le schéma multiplicatif** :

$$y_t = N_t \times S_t + R_t \quad (1.4)$$

Dans lequel la composante saisonnière est liée à la composante extra-saisonnière (avec une saisonnalité souple, variant en amplitude au fil du temps). [6].

1.16 Séries chronologique stationnaires :

Un type extrêmement important de série chronologique est la série chronologique stationnaire. On considère qu'une série chronologique est strictement stationnaire lorsque ses caractéristiques ne sont pas affectées par un changement de l'origine temporelle. En d'autres termes, une série chronologique est considérée comme strictement stationnaire si les distributions de probabilité des observations $y_t, y_{t+1}, \dots, y_{t+n}$ sont exactement identiques aux distributions des observations $y_{t+k}, y_{t+k+1}, \dots, y_{t+k+n}$. La stationnarité crée un équilibre statistique et une stabilité particulière dans les données, assurant ainsi une moyenne et une variance constantes [33].

1.16.1 Fonctions Autocovariance et Autocorrélation :

Si une série temporelle est stationnaire, cela implique que la distribution de probabilité conjointe de deux observations, à savoir y_t et y_{t+k} , demeure identique pour deux périodes t et $t + k$ séparées par le même intervalle k . Des informations précieuses concernant cette distribution conjointe, ainsi que la nature de la série temporelle, peuvent être obtenues en

traçant un diagramme de dispersion regroupant toutes les paires de données y_t et y_{t+k} séparées par le même intervalle k . Ce laps de temps est communément appelé "lag". [33] La covariance entre y_t et sa valeur à une autre période, soit y_{t+k} , est dénommée autocovariance avec un décalage de k , définie par

$$\gamma_k = Cov(y_t, y_{t+k}) \quad (1.5)$$

. L'ensemble des valeurs de $\gamma_k, k = 0, 1, 2, \dots$ est désigné sous le terme de **fonction d'autocovariance**. Il est important de noter que l'autocovariance avec un retard $k = 0$ correspond simplement à la variance de la série temporelle, laquelle reste constante pour une série temporelle stationnaire. [33]

Le coefficient d'autocorrélation avec un retard k pour une série temporelle stationnaire est quant à lui donné par

$$\rho_k = \frac{Cov(y_t, y_{t+k})}{Var(y_t)} = \frac{\gamma_k}{\gamma_0} \quad (1.6)$$

L'ensemble des valeurs de $\rho_k, k = 0, 1, 2, \dots$ est connu sous le nom de **fonction d'autocorrélation (ACF)**. Il convient de souligner que, par définition, $\rho_0 = 1$. De plus, la fonction d'autocorrélation est indépendante de l'échelle de mesure de la série temporelle, ce qui en fait une grandeur sans dimension. De plus, $\rho_k = \rho_{-k}$, ce qui signifie que l'ACF est symétrique par rapport à zéro. Par conséquent, il est seulement nécessaire de calculer la moitié positive (ou négative). [33]

Il est indispensable d'estimer l'autocovariance et les ACF à partir d'une série temporelle de longueur finie, notée y_1, y_2, \dots, y_T . Généralement, il est recommandé d'avoir au moins $T/4$ observations pour obtenir une estimation fiable de l'ACF, avec T le nombre d'observations total. [33]

1.17 Transformations des séries chronologique

Il existe plusieurs types d'ajustements utiles à la modélisation et à la prévision des séries temporelles pour la rendre stationnaire. L'une des plus utilisées est la différenciation. [33] La différenciation est une méthode utilisée pour éliminer la tendance dans une série chronologique. Elle consiste à calculer la différence entre une observation et l'observation précédente. Par exemple, si on a une série chronologique y_t , la première différence est calculée en soustrayant y_{t-1} de y_t . Cela permet de capturer la variation d'une période à l'autre.

$$x_t = y_t - y_{t-1} = \nabla y_t \quad (1.7)$$

où ∇ est l'opérateur de différence (à reculons). Une autre façon d'écrire l'opération de différenciation est en utilisant l'opérateur de décalage arrière B , défini comme $By_t = y_{t-1}$, donc :

$$x_t = (1 - B)y_t = \nabla y_t = y_t - y_{t-1} \quad (1.8)$$

La différenciation peut être effectuée plusieurs fois si nécessaire. Par exemple, la deuxième différence est obtenue en prenant la différence entre les différences successives. Cela permet de mettre en évidence les changements de pente dans la série chronologique. [33]

1.18 Description de certains Modèles

1.18.1 Seasonal Autoregressive Integrated Moving Average (SARIMA)

AutoRegressive Integrated Moving Average (ARIMA) [7] est un Modèle statistique utilisant des variations et des régressions de données statistiques pour trouver des motifs en vue de

prédictions futures. Il s'agit d'un modèle de série chronologique dynamique où les estimations futures sont expliquées par les données passées et non par des variables indépendantes.

Le modèle ARIMA identifie les coefficients et le nombre de régressions utilisées, ce qui le rend sensible à la précision avec laquelle ses coefficients sont déterminés.

ARIMA s'exprime sous la forme (p, d, q) , où les paramètres p , d et q sont des entiers non négatifs qui indiquent l'ordre des différentes composantes du modèle. Le paramètre p indique l'ordre de la partie autorégressive, d indique le degré de la différenciation impliquée pour rendre la série stationnaire si nécessaire, et q indique l'ordre de la partie de la moyenne mobile (est utilisée pour modéliser la composante de moyenne mobile de la série temporelle). Cette composante représente les erreurs aléatoires ou les chocs imprévisibles qui ne peuvent pas être expliqués par les termes autoregressifs ou les différences dans le modèle) [7].

Par conséquent, lorsque l'un des trois paramètres est nul, il est courant d'omettre les lettres correspondantes de l'acronyme, où AR représente la composante autorégressive, I la composante intégrée et MA la moyenne mobile. Par exemple, ARIMA(1, 0, 0) peut être exprimé comme AR(1) et ARIMA(0, 0, 1) comme MA(1). [7]

Le modèle ARIMA (p, d, q) peut être représenté comme suit :

$$Y_t = -(\Delta^d Y_t - Y_t) + \varphi_0 + \sum_{i=1}^p \varphi_i \Delta^d Y_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \varepsilon_t \quad (1.9)$$

où d correspond aux différences nécessaires pour convertir la série originale en une série stationnaire, $\varphi_1, \dots, \varphi_p$ sont les paramètres appartenant à la partie **Autoregressive** du modèle, $\theta_1, \dots, \theta_q$ sont les paramètres appartenant à la partie **Moving Average** du modèle, φ_0 est une constante et ε_t est le terme d'erreur (également appelé perturbation stochastique, ce dernier étant associé plus aux modèles économétriques à équations simples ou multiples). [7]

SARIMA (Seasonal Autoregressive Integrated Moving Average) [7] est une généralisation du modèle pour prendre en compte l'effet de la saisonnalité. Pour la méthode SARIMA, il faut configurer quatre paramètres saisonniers supplémentaires, en plus des trois éléments de la méthode ARIMA(p, d, q) .

Le paramètre P est l'ordre autorégressif saisonnier, D est l'ordre de différence saisonnière, Q représente l'ordre de la moyenne mobile saisonnière et m est le nombre d'étapes temporelles pour une période saisonnière unique. La notation du modèle SARIMA est spécifiée comme SARIMA(p, d, q)(P, D, Q, m) :

$$Y_t = -(\Delta^d Y_t - Y_t) + \phi_0 + \sum_{i=1}^p \phi_i \Delta^d Y_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i} - (\Delta^D Y_t - Y_t) + \sum_{i=1}^P n_i \Delta^D Y_{t-m_{t-i}} + \sum_{i=1}^Q w_i \varepsilon_{t-m_{t-i}} + \varepsilon_t \quad (1.10)$$

Où D est le nombre de différenciations saisonnières, n_1, \dots, n_P ont les coefficients des termes autorégressifs saisonniers, w_1, \dots, w_Q sont les coefficients des termes de moyenne mobile saisonniers et m est le l'ordre de saisonnalité [35].

Holt-Winters Exponential Smoothing

Le modèle de lissage exponentiel saisonnier Holt-Winters [9] est une méthode couramment utilisée pour analyser et prévoir les séries temporelles saisonnières. Il permet de capturer la

tendance générale, les variations saisonnières et les fluctuations aléatoires des données. Le modèle se compose de trois composantes principales : la tendance, la composante saisonnière et la composante d'erreur.

La composante de tendance est utilisée pour modéliser la tendance générale des données, tandis que la composante saisonnière capture les variations périodiques régulières. La composante d'erreur quantifie les fluctuations aléatoires non expliquées par la tendance et la saisonnalité [33].

L'équation pour prévoir la valeur de la série temporelle à l'instant $t + 1$ en utilisant le modèle de lissage exponentiel Holt-Winters multiplicatif est la suivante :

$$\hat{y}_{t+1} = (l_t + b_t) \cdot s_{t+1-m} \quad (1.11)$$

où :

- y_{t+1} est la prédiction de la valeur de la série temporelle à l'instant $t + 1$.
- Le niveau estimé (l_t) représente la valeur de la tendance générale des données à l'instant t .
- La pente estimée (b_t) représente la variation de la tendance entre deux instants consécutifs.
- s_{t+1-m} est la composante saisonnière à l'instant $t + 1 - m$.
- ϵ_t quantifie les fluctuations aléatoires non expliquées par la tendance, la saisonnalité et la pente.

Cette équation combine la tendance, la pente et la composante saisonnière pour estimer la valeur future de la série temporelle. [33].

Conclusion

Dans ce chapitre, nous avons rappelé certaines notions fondamentales de la Business Intelligence et le Data Mining. Notamment les différentes étapes à suivre et processus de chacun de ces deux domaines pour mener à bien notre projet. Dans le chapitre suivant, nous présenterons une vue d'ensemble sur le domaine pétrolier ainsi que les différentes approches de prévision et prédiction utilisées dans ce dernier.

Chapitre 2

PRÉDICTIONS DANS LE DOMAINE PÉTROLIER

Introduction

Nous aborderons dans ce chapitre, une explication des principales caractéristiques et termes liées à l'industrie pétrolière et gazière. Il décrit également les approches de prédition dans le domaine pétrolier.

2.1 Concepts de base

- **Bassin sédimentaire :** Est une dépression en forme de cuvette évasée due à un affaissement lent et progressif (subsidence), où se sont empilés pendant une longue période (10 à 100 Ma.) et sur une grande épaisseur des sédiments variés. (Pour notre cas d'études nous aurons : Bassin de Berkine, Bassin de Illizi, Bassin de Zarzaitine et Bassin de Timimoune).

Dans un bassin pétrolier, les quantités d'hydrocarbures ne sont pas dispersées dans le sous-sol de façon homogène. Celles-ci sont concentrées dans des structures géologiques particulières appelées « Bloc ».

- **Bloc :** Partie délimitée du domaine minier relative aux activités pétrolières, il est composé d'une ou de plusieurs parcelle(s), limitée(s). (TFT, Timimoune, Zarzaitine et El merk).
- **Périmètre :** Partie délimitée du domaine minier relative aux activités pétrolières, il est composé d'une ou de plusieurs parcelle(s), limitée(s), délimitant un ou plusieurs gisement(s).
- **Gisement (Champ) :** Le champ pétrolier peut être considéré comme une aire géographique élémentaire, supposée indivisible dont le sol ou le sous-sol renferment des Hydrocarbures et qui peut être composée d'une ou plusieurs entités géologiques que l'on peut exploiter par le forage des puits.

Les quantités d'hydrocarbures contenues dans un gisement ne peuvent jamais être intégralement extraites du sous-sol. C'est pourquoi pour un même gisement, on parle de « réserves » pour les quantités d'hydrocarbures qui sont ou seront récupérables.

- **Réservoir** : horizons géologiques qui renferment des Hydrocarbures entre deux formations imperméables. Ces réserves pétrolières diminuent lorsque le champ est exploité, et augmentent si les techniques de production permettent d'augmenter le taux de récupération du gisement.
- **Plan de développement (POD)** : est un Document qui englobe un Programmes de travaux relatifs aux opérations de développement, de la mise en production, et d'exploitation des Hydrocarbures, d'abandon et de remise en état du site. Il établit le nombre de puits à forer, le type de récupération envisagé, les débits de fluides, le coût des installations annexes (ex : séparation, traitement). Les recettes prévisionnelles sont évaluées selon les prix du baril estimatifs, les conditions de l'accord de partage avec le pays propriétaire, etc.
- **Manifold** : il a pour fonction de collecter les arrivées d'effluent de tous les puits et de pouvoir les diriger soit vers le séparateur PRODUCTION soit vers le séparateur TEST par le biais d'un jeu de vannes.
- **Séparateur TEST** : Il est utilisé comme son nom l'indique lors de tests soit pour un suivi de la production (quantifier chaque phase de l'effluent (gaz, huile et eau) pour chaque puits) ou pour une analyse suite à une anomalie. Les mesures effectuées sur le séparateur de test ne sont utilisables que si le puits est stable durant le test.
- **Séparateur PRODUCTION** : Lorsque les puits sont en production, il faut les monitorer : c'est-à-dire qu'il faut suivre leur évolution afin de détecter les éventuels manques à produire (différence entre ce que le puits est censé produire et ce qu'il produit réellement).

2.2 Types de Gaz

Le constituant principal des gisements de gaz naturel est le méthane. Il est aussi composé d'éthane, propane, butane et pentane à très faibles quantités. La composition peut varier en fonction de la source et de la forme du gaz, par exemple, certains types de gaz ont une proportion plus élevée de propane et de butane. Il existe quatre exemples principaux de sources et de formes de gaz :

- Gaz associé : il s'agit du gaz naturel qui se trouve associé au pétrole dans un réservoir. Ce gaz peut être brûlé dans des torchères ou traité pour être utilisé comme produit.
- Gaz non associé : provient des réservoirs ne contenant que du gaz naturel et pas de pétrole.
- Gaz riche : type de gaz naturel qui contient des hydrocarbures plus lourds qu'un gaz pauvre, c'est-à-dire des concentrations plus élevées de propane et de butane.
- Liquides de Gaz naturel : il s'agit des composantes du gaz naturel qui sont dissociées de l'état gazeux sous forme de liquides dans une usine de traitement du gaz par absorption ou condensation. Ils sont classés en fonction de leur pression de vapeur :
 - Faible = condensat
 - Intermédiaire = gaz naturel
 - élevée = gaz de pétrole liquéfié, c'est-à-dire le propane ou butane.

2.3 Prévision de la production pétrolière en utilisant Time Series Analysis

La prévision précise de la production de gaz et d'huile est cruciale dans l'industrie pétrolière. Cependant, cela requiert d'importants investissements en argent, temps et technologie pour estimer avec précision les réserves de pétrole. [31] [20]

Les données temporelles pétrolières sont perturbées par du bruit, des défauts et des anomalies, et peuvent avoir une dimensionnalité élevée [34] [30]. De plus, ces données ne sont pas stationnaires et présentent des tendances variables, ce qui rend les estimations plus difficiles. [28]

La production pétrolière dépend de paramètres dynamiques tels que la saturation en fluide et la pression, ainsi que de paramètres statiques comme la porosité et la perméabilité [11]. Cependant, la disponibilité de ces paramètres est souvent limitée, ce qui réduit la précision globale des prévisions. [30]

2.3.1 Travaux réalisés

Plusieurs approches ont été développées pour surmonter les défis susmentionnés de prévision des séries temporelles pétrolières, cependant, la clé d'une prévision réussie réside dans le choix de la bonne représentation. [28] Ces approches peuvent être classées en deux catégories principales, à savoir **les approches statistiques et les approches de réseaux de neurones**.

— Approches Statistiques

1. L'une des méthodes statistiques traditionnelles les plus courantes est Autoregressive Integrated Moving Average (ARIMA) [26]. ARIMA et ses variantes (Seasonal Autoregressive Integrated Moving Average (SARIMA)) peuvent être utilisées pour réaliser diverses activités de prévision dans l'industrie pétrolière telles que les prix, les niveaux de consommation et la production des réservoirs. [13]
 - En 2015, Choi, J., Roberts, D. et Lee, E ont utilisé le modèle SARIMA pour prévoir la production pétrolière future au Dakota du Nord. Ils ont analysé les données historiques et ont appliqué le modèle SARIMA pour estimer la quantité de pétrole qui pourrait être extraite dans la région dans les années à venir.
 - En 2022, M.Sánchez et F.ANGUIANO ont réalisé l'application d'ARIMA et de SARIMA dans une étude de cas sur la production de pétrole et de gaz au Mexique. L'étude a utilisé à la fois des méthodes non saisonnières(moyenne mobile simple, lissage exponentiel simple et lissage exponentiel double de Holt) et saisonnières (SARIMA, Holt-Winters Exponential Smoothing). [12]
2. Une autre méthode mathématique connue est l'analyse des courbes de déclin (DCA), elle a été largement utilisée dans l'industrie pétrolière, notamment dans les scénarios décrivant le déclin de la production pétrolière. [13]

Ces approches se basent sur l'analyse de données subjectives et ajustent les paramètres du modèle de simulation numérique pour obtenir des valeurs raisonnables et des interprétations des champs pétroliers. [2]

Cependant, la nature non linéaire et hétérogène de la géologie des champs pétroliers ainsi

que des propriétés des fluides représente un défi majeur pour ces méthodes. [20] [28]

— Réseaux de neurones

Depuis la dernière décennie, des efforts sincères ont clairement été publiés dans la littérature présentant l'utilisation des réseaux de neurones pour réaliser différentes activités de prévision dans un certain nombre d'applications d'ingénierie pétrolière.

1. En 2013, Chakra et al. ont présenté un modèle innovant de réseau neuronal d'ordre supérieur axé sur la prévision de la production cumulée de pétrole à partir d'un réservoir pétrolier situé au Gujarat, en Inde [11].
2. En 2016, Aizenberg et al. ont présenté un réseau neuronal multicouche avec des neurones à valeurs multiples capable de réaliser une prévision de séries temporelles de production pétrolière. [2]
3. En 2019, Alaa Sagheera et Mostafa Kotbb ont proposés une architecture de mémoire à long terme récurrente profonde (DLSTM), en tant qu'extension du réseau neuronal récurrent traditionnel. [12]

Ces approches offrent des solutions plus adaptées pour modéliser des données complexes et bruitées, ce qui peut conduire à une meilleure estimation de la production future de pétrole.

2.3.2 Synthèse

L'utilisation des modèles SARIMA (Seasonal Autoregressive Integrated Moving Average) et Holt-Winters Exponential Smoothing présente efficacité considérable pour prévoir la production de gaz à court terme.

La production de gaz peut présenter des schémas saisonniers, avec des fluctuations régulières à court terme. Les modèles SARIMA et Holt-Winters Exponential Smoothing sont spécifiquement conçus pour capturer ces variations saisonnières. Ils utilisent des termes d'autorégression et de moyenne mobile saisonnière pour modéliser et prédire efficacement ces variations.

D'autre part, le choix de SARIMA et Holt-Winters Exponential Smoothing réside dans leur simplicité par rapport aux réseaux de neurones. Ces modèles nécessitent moins de paramètres à ajuster et sont plus rapides à entraîner.

Cette caractéristique peut être bénéfique lorsque les données sont limitées ou que les ressources informatiques sont restreintes. Ainsi, les modèles SARIMA et Holt-Winters peuvent être adaptés pour tenir compte des particularités propres aux données pétrolières, telles que les tendances non linéaires, les changements abrupts ou les variations cycliques.

Le choix entre SARIMA, Holt-Winters Exponential Smoothing et les réseaux de neurones dépendra des caractéristiques spécifiques des données et des objectifs de prévision.

2.4 Le ML dans le domaine pétrolier

Il y a plusieurs années de nombreux chercheurs de l'industrie pétrolière ont adopté l'apprentissage automatique et l'apprentissage en profondeur pour prédire l'avenir avec des connaissances antérieures (apprentissage supervisé), telles que la correspondance de l'historique des réservoirs, les prévisions de production de pétrole, de gaz et d'eau, la reconnaissance de formes dans les analyses de test de puits etc... [36]

Le ML peut également aider les entreprises pétrolières à prévoir les pannes et à planifier les opérations de maintenance. Les capteurs peuvent être utilisés pour collecter des données sur l'état des équipements, et ces données peuvent être analysées à l'aide de modèles de ML pour détecter les signes précurseurs de pannes. Cela permet aux entreprises pétrolières de planifier les opérations de maintenance de manière proactive, réduisant ainsi les temps d'arrêt non planifiés et améliorant la sécurité.

Faisant l'objet principal de notre étude, nous exposons ci-après les travaux liés à l'application du ML dans la prévision de la production de gaz. La prédition de production de gaz peut se baser sur différents critères, mais l'historique de production est l'un des principaux éléments utilisés pour effectuer ces prédictions.

2.4.1 Méthodes utilisées

Liao et al. ont utilisé la forêt aléatoire Random Forest, Extreme Gradient Boosting (XGBoost) et Light Gradient Boosting Machine (LGBM) pour construire un modèle d'empilage et ont déterminé que la longueur stimulée, le nombre total d'étages du réservoir, le proppant pompé par étage, les fluides pompés par longueur et le taux d'injection sont les facteurs les plus importants pour la formation de gaz de réservoirs étanches Wapiti-Montney. Toutefois, le modèle ne comprenait pas de paramètres de réservoir comme la teneur en carbone organique totale (TOC). [29]

Hirschmiller et al ont utilisé recursive feature elimination pour sélectionner les caractéristiques et ont utilisé la forêt aléatoire pour prédire et optimiser les performances du puits. [25] Sheikhi et al ont utilisé la régression linéaire, la forêt aléatoire Random Forest, Gradient Boosting, XGBoost, Bagging, ExtraTrees et le réseau de neurones pour évaluer la performance de compléition : elle fait référence à l'évaluation des résultats obtenus liés aux opérations finales réalisées lors du forage d'un puits pétrolier ou gazier. [43]

Wang et al a utilisé Random Forest, Adaptive Boosting, Support Vector Machine et Réseaux de neurones pour estimer la performance des puits. Il a été conclu que la forêt aléatoire a la meilleure performance et ces modèles ont été utiles pour concevoir des traitements de fracture hydraulique. [49]

Liang et al a utilisé Multi-objective Random Forest pour prévoir les données de production dynamiques. [49]

Cependant, il reste un défi à relever pour choisir la bonne méthode de prévision de la production. En général, la régression linéaire multiple ne peut que décrire la relation linéaire. L'arbre de régression nécessite beaucoup de données pour bien fonctionner. La construction du réseau neuronal nécessite un traitement lourd et un temps d'exécution très long. [47]

En comparaison au Deep Learning, l'arbre de régression, la régression linéaire multiple MLR, Support Vector Regression SVR et Gaussian Process Regression GPR ont une meilleure performance pour les petits ensembles de données [24] [16].

Conclusion

Faisant l'objet principal de notre projet d'études, ce chapitre permet une meilleure compréhension du domaine pétrolier et du fonctionnement des puits. Nous avons également décrit les accomplissements et réalisations des chercheurs dans l'élaboration des solutions prévisionnelles à l'aide des séries chronologique, et le rôle important joué par le ML dans le domaine pétrolier. Dans le chapitre suivant, nous exposerons les besoins de l'organisme d'accueil ainsi que la solution proposée.

Chapitre 3

ETUDE DE L'EXISTANT

Ce chapitre fait l'objet en premier lieu de la présentation de l'organisme où s'est déroulé notre stage. Nous exposerons par la suite, les besoins analytiques et les exigences des utilisateurs du système, ainsi que la démarche suivie afin d'élaborer la solution proposée.

3.1 Description de l'organisme

3.1.1 La Sonatrach

La SONATRACH (Société Nationale pour la Recherche, la Production, le Transport, la Transformation, et la Commercialisation des Hydrocarbures) est une entreprise publique algérienne spécialisée dans l'exploration, la production, le transport et la commercialisation de pétrole et de gaz naturel. Fondée en 1963, la SONATRACH est devenue un acteur majeur du marché mondial de l'énergie, avec des activités dans plus de 20 pays et des partenariats avec les plus grandes compagnies pétrolières et gazières du monde.

Elle est reconnue pour sa maîtrise technologique, son engagement sur les marchés internationaux et sa participation active dans la résolution des problèmes énergétiques à l'échelle mondiale. Avec deux filiales clés, à savoir la « SONATRACH petroleum corporation (SPC) » basée à Londres et « SONATRACH trading » basée à Amsterdam, la SONATRACH s'impose fortement sur le marché mondial de l'énergie. En tant que membre actif de l'OPEP et d'autres organisations, la SONATRACH joue un rôle crucial dans la stabilisation des prix du pétrole et la recherche de solutions pour les problèmes énergétiques mondiaux. Grâce à cette expertise, la SONATRACH figure actuellement parmi les 11 plus grandes entreprises pétrolières dans le monde.



FIGURE 3.1 – Logo de la SONATRACH

3.1.2 Organigramme de structure d'accueil de la SONATRACH

L'entreprise possède une organisation hiérarchique détaillée ci-dessous :

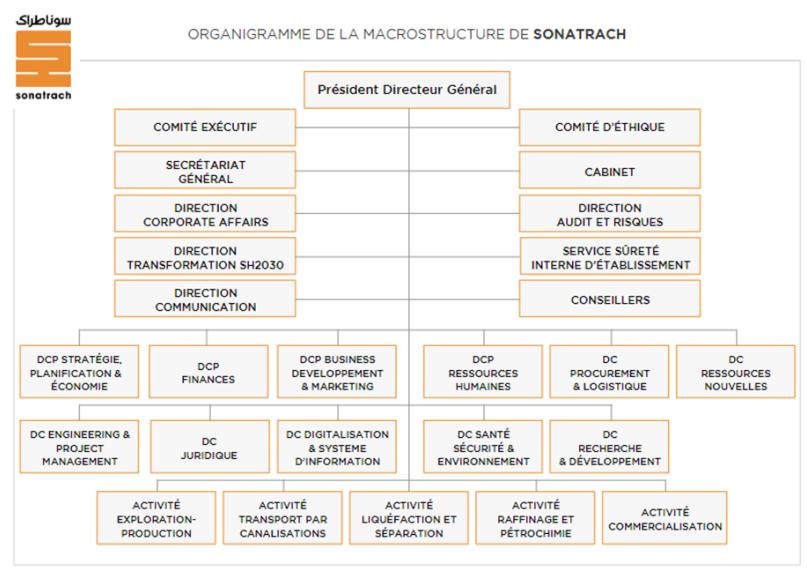


FIGURE 3.2 – Organisme de Macrostructure de SONATRACH

3.1.3 Organigramme structure d'accueil Division Association (AST)

Le département banque de données et système d'information de la Division Association est organisé en trois services, à savoir le Service Réseaux et Télécoms, le Service Système Informatique et le Service Systèmes d'Information.

Le Service Système d'Information a pour principales missions :

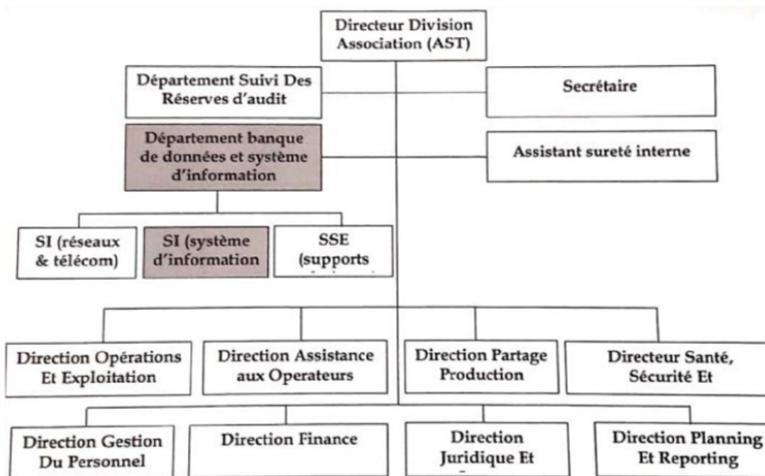


FIGURE 3.3 – Organigramme structure d'accueil Division Association (AST)

Le développement et la maintenance du système d'information spécifique à la Division Association.

- La mise en place d'applications et de banques de données métiers en collaboration avec toutes les structures concernées de la division.
- L'administration de l'ensemble des données de la division.
- La normalisation et la standardisation des "Applications Métiers" destinées aux opérateurs en association, telles que la gestion des données pétrolières, la maintenance, les stocks, les finances et la GED, conformément à la politique informatique de SONATRACH.
- Le suivi des projets liés au volet Système d'Information au niveau des opérateurs en Association.

3.2 Collecte d'informations et analyse des besoins

Tout projet nécessite en premier lieu la récolte d'informations pour une bonne compréhension du fonctionnement décisionnel de l'organisme et dans le but d'identifier les besoins des utilisateurs finaux.

Nous avons eu recours à de la recherche et documentation, dans le but de nous familiariser avec le domaine pétrolier et le fonctionnement des puits. Des documents nous ont été livrés et des séances de questionnaires avec les personnes concernées par la production d'huile/gaz ont été organisées afin d'avoir un maximum d'informations

Pour bien clarifier l'objectif de notre travail, nous avons effectués des séances de travail hebdomadaires avec notre encadreur. Sur les 17 groupements présents à la SONATRACH, il a été décidé de travailler sur les données de 4 groupements uniquement en raison du manque de temps dont 2 groupements de production d'huile et Gaz Sec, 1 de Gaz Sec uniquement et 1 groupement produisant différents fluides(GPL, Condensat), en plus du Gaz Sec.

3.2.1 Etude des sources de données

Pour des raisons de confidentialité, l'entreprise nous a procurés un accès limité aux données.

1. Données de prévision par groupement :

- Prévisions mensuelles de production de chaque fluide pour les années de 2020 à 2022 des groupements GTIM et GTFT sous format Excel.
- Prévisions annuelles de production d'huile de 2022 à 2046 du groupement GSS sous format Excel.
- Prévisions du groupement Berkine indisponibles.

2. Données de production journalière par puits :

- **Groupement GTIM :** dossier contenant des fichiers de rapports journaliers sous format Excel de 2018 à 2021

- **Groupement GTFT** : base données Access de 1999 à 2021
- **Groupement BERKINE** : fichier CSV de 2013 à 2021
- **Groupement GSS** : fichier Excel de 1999 à 2021

3. Autres données :

- **Groupement Berkine** : fichier Excel contenant pour chaque périmètre, la liste des partenaires et le type de contrat.
- **Groupement GTFT et GTIM** : un fichier Excel contenant sur chaque feuille la liste des partenaires par périmètre de chaque groupement, le type de contrat ainsi que le bassin sédimentaire.

3.2.2 Recueil des besoins techniques

Les données du groupement GTIM nous ont été transmises dans un dossier contenant des fichiers Excel sous forme de rapports journaliers nommés par date. Le principal besoin identifié est de pouvoir regrouper et fusionner toutes les données dans un seul même fichier de sortie consolidé.

Les données des groupements étant toutes en provenance de sources différentes, il a également été demandé de concevoir un système décisionnel permettant aux utilisateurs d'avoir une vue globale sur le suivi de production de tous les groupements réunis dans un tableau de bord capable de calculer automatiquement les indicateurs de performance pour suivre et anticiper le fonctionnement des puits.

Le tableau de bord doit notamment attirer l'attention sur l'écart entre la production journalière prévue et la production journalière réalisée, ainsi que l'écart entre les productions mensuelles réalisées et les cumuls de production mensuels prévus afin de prendre les dispositions nécessaires pour augmenter le débit de production.

La méthode utilisée pour la prévision est particulièrement utilisée afin de détecter un éventuel déclin prévu dans les mois qui suivent. De plus que cette méthode s'applique uniquement sur une production stable non affectée par aucun autre moteur artificiel qui déclencherait une hausse de production tels que les groupements **GTIM et GTFT**. Par ailleurs, elle se base sur l'historique de production cumulée dans tout le groupement.

En effet, si l'on souhaite avoir les prévisions journalières exactes de chaque puits, ou encore les prédictions d'un groupement affecté par les moteurs externes, le groupement **Berkine** par exemple, ou même le cas d'un nouveau puits foré dont l'historique est indisponible, cette méthode ne peut être utilisée.

Afin de répondre à cette problématique, il a été demandé de proposer une solution plus générale et plus adaptée pouvant d'une part fournir les prévisions mensuelles du groupement pour l'année suivante et d'autre part, prédire la production journalière exacte des puits d'un groupement où les méthodes d'injection sont pratiquées.

3.2.3 Acteurs du système

Les utilisateurs finaux du système décisionnel et de la solution mise en place sont :

- Ingénieurs Production.
- Chef département Production.
- Directeur général de l'organisme.

3.2.4 Recueil des besoins analytiques

Les questionnaires, entretiens, et réunions nous ont amenés à résumer leurs besoins analytiques sur ce qui suit :

- Comment se porte la production de gaz/huile sur chaque champ, réservoir, puit ?
- Quelle est la production journalière de chaque puit ?
- Le cumul mensuel prévu de production a t'il été atteint ?
- Comment se déroulera la production de gaz sur chaque champ dans les mois qui suivent ?
- Peut-on prévoir la production des puits affectés par l'injection ?
- Peut-on prédir la production journalière exacte de chaque puits ?

3.3 Méthode DCA utilisée par l'entreprise pour la prévision

La méthode utilisée par la division association pour la prévision de la production de gaz et de l'huile du groupement GTFT et GTIM est l'analyse de courbes de déclin (ou "**Decline Curves Analysis**" en anglais). Cette méthode est couramment utilisée dans l'industrie pétrolière et gazière pour prédire la production future en se basant sur les données de production passées.

L'analyse de courbes de déclin consiste à ajuster une courbe mathématique à la courbe de production historique. Cette courbe mathématique est généralement choisie parmi une famille de courbes de déclin existantes, telles que la courbe de déclin **exponentielle, hyperbolique, harmonique**.

En ce qui concerne les puits ayant un long historique de production et produisant à une pression d'écoulement constante au fond du trou (BHP), **Arps (1945)** [3] a dérivé trois types de déclins de production par la relation taux (production cumulée) -temps, y compris le déclin exponentiel, le déclin hyperbolique et le déclin harmonique. Cette méthode est simple, sans considération des paramètres du réservoir ou du puits, et peut être appliquée à différents types de réservoirs. Cependant, cette méthode présente certaines limites :

- Dépendance des données historiques : Cette méthode repose sur des données historiques de production de gaz. Si ces données sont peu fiables, incomplètes ou ne couvrent pas une période de temps suffisamment longue, les prévisions basées sur ces données peuvent

être moins précises.

- Hypothèses simplificatrices : L'analyse de la courbe de déclin utilise des hypothèses simplifiées sur la stabilité des paramètres de déclin et l'absence de facteurs externes. Dans la réalité, ces hypothèses peuvent ne pas être totalement exactes, ce qui peut affecter la précision des prévisions.
- Variabilité des courbes de déclin : Les réservoirs de gaz peuvent présenter une variabilité naturelle des courbes de déclin en raison de la complexité géologique et des conditions de production. Une seule courbe de déclin ne peut pas représenter toutes les caractéristiques du réservoir, ce qui peut conduire à des prévisions moins précises.
- Sensibilité aux changements opérationnels : Les modifications des pratiques d'exploitation, comme l'utilisation de nouvelles technologies ou l'ajustement des paramètres de production, peuvent avoir un impact sur les courbes de déclin. Si ces changements ne sont pas pris en compte dans l'analyse, les prévisions peuvent être décalées.
- Incertitude des résultats : Comme toute méthode de prévision, l'analyse de la courbe de déclin comporte une certaine marge d'erreur. Les résultats obtenus ne sont que des estimations basées sur les données et les hypothèses utilisées, et ils peuvent différer des résultats réels. Il est donc important de considérer cette incertitude lors de l'interprétation des prévisions. [45]

3.4 Description de la solution proposée

Après analyse des besoins analytiques des utilisateurs finaux, nous avons en premier lieu mis en place une solution décisionnelle afin de répondre à leurs besoins de consulter plusieurs sources de données hétérogènes sur une seule plateforme. Par ailleurs, nous avons proposé une amélioration de leur méthode de prévision, en utilisant une approche plus précise et plus générale. Nous avons également conçu un modèle pouvant prédire les productions des puits en se basant sur les paramètres mesurés et caractéristiques de chacun.

3.4.1 Architecture de la solution système décisionnel

La solution décisionnelle permettant de regrouper et consolider les sources de données hétérogènes consiste à concevoir un entrepôt de données alimenté directement depuis les bases de données opérationnelles ACCESS, SQL Server et les autres sources (fichiers Excel + csv) via un moteur ETL.

Pour l'exploitation, la restitution et la visualisation des données, nous avons mis en place des tableaux de bord dotés de KPIs pouvant répondre à leurs questions d'analyse.

3.4.2 Architecture de la solution prédition

Pour notre cas, le besoin principal est d'améliorer la précision des prévisions mensuelles par groupement et d'effectuer la prédition de la production de gaz d'un puits individuellement, ou d'un nouveau puits foré n'ayant pas d'historique, et seulement en se basant sur ses caractéristiques.

Dans un premier temps, il est possible d'utiliser l'analyse des séries temporelles Time Series pour effectuer les prévisions mensuelles du groupement. Cette méthode permet de modéliser et de prévoir les tendances et les motifs de la production de gaz à partir de données historiques, en prenant en compte les variations saisonnières. En complément de l'analyse de la courbe de déclin DCA, l'analyse des séries temporelles fournit des prévisions plus robustes et plus fiables.

D'autre part, les modèles prédictifs de Machine Learning sont tout aussi intéressants, dans le cas où l'on souhaite avoir les prédictions exactes pour chaque puits et ce, en se basant sur les paramètres impactant sur la production.

Conclusion

Dans ce chapitre, nous avons exposé l'organisme d'accueil, leur méthode de travail ainsi que leurs besoins à savoir : avoir un suivi global de production de tous les groupements et les prédictions futures. Nous avons également présenté une vue générale de l'architecture de la solution proposée à la fois analytique, prévisionnelle et prédictive. Dans le chapitre suivant, ces solutions seront décrites plus en détail de leur conception à leur mise en œuvre.

Chapitre 4

CONCEPTION

Introduction

Rappelons que l'objectif de notre travail est de développer une solution décisionnelle qui permet la visualisation rapide et automatique du suivi de production et une solution prédictive qui aiderait les décideurs à avoir une meilleure maîtrise sur les puits.

Ce chapitre a pour objectif l'élaboration de la solution finale qui répond au mieux à ces besoins décisionnels et analytiques. Nous présenterons dans un premier temps les étapes de la mise en place de l'entrepôt de données, i.e la conception de la zone de stockage, la conception de la zone d'alimentation et enfin la conception de la zone de restitution. Dans ce qui suit, nous expliquerons également la démarche suivie pour la réalisation de la solution à la fois prédictive et prévisionnelle, à savoir : l'analyse approfondie des données sources, les transformation requises pour la modélisation des algorithmes et enfin l'évaluation des modèles pour en tirer le plus performant.

4.1 Architecture générale de la solution

Comme expliqué précédemment, l'entreprise souhaite avoir un tableau de bord qui permet de visualiser le suivi de production des groupements ainsi qu'un système qui fournit les futures valeurs de production pour mieux contrôler le fonctionnement de chaque puits individuellement et anticiper les risques. Notre solution comporte deux volets.

La (Figure 4.1) ci-après, permet une meilleure compréhension de la solution proposée :

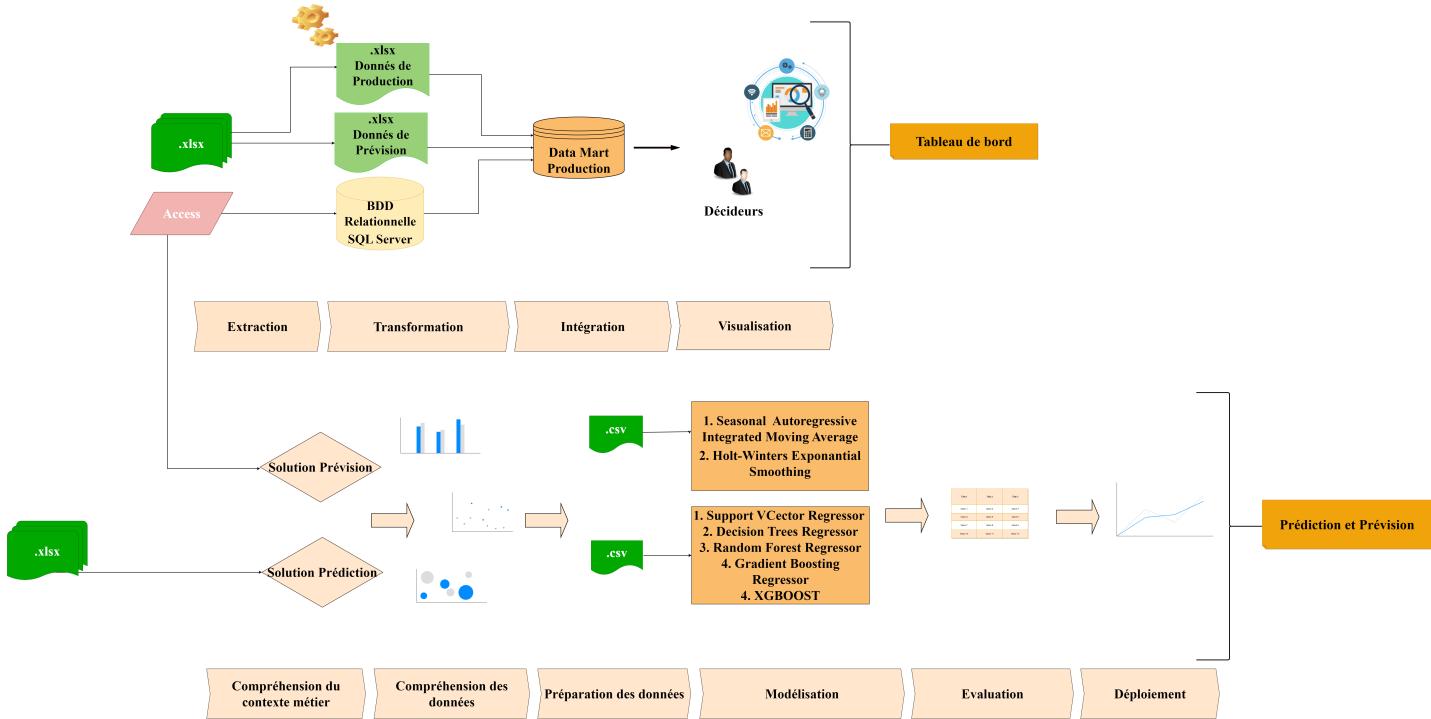


FIGURE 4.1 – Schéma représentatif de la solution proposée

4.1.1 Volet BI

Comme indiqué dans la (Figure 4.1) ci-dessus, la solution BI vise en premier lieu à extraire les données de production des années précédentes en provenance de sources hétérogènes (**ACCESS, Excel**), les traiter, nettoyer et transformer, pour ensuite les charger dans le magasin de données et les présenter enfin sous forme de tableau de bord pour répondre à leurs besoins d'analyse de suivi.

La modélisation du magasin de données repose sur :

- Un fait contenant les mesures à analyser (KPIs)
- Des dimensions contenant les paramètres de l'analyse (temps, lieu, ...)

Pour chaque dimension, les paramètres sont hiérarchisés selon des niveaux de détail (temps : années, mois, jour ...).

4.1.2 Volet Prédiction

Cette solution à la fois prévisionnelle et prédictive, consiste dans un premier temps à généraliser la méthode prévisionnelle appliquée par l'entreprise et ce, en se basant sur l'historique des données de production cumulée mensuellement des puits d'un groupement. La méthode appliquée pour ce cas est la modélisation des séries chronologiques en utilisant Time Series.

Par ailleurs, cette méthode basée sur l'historique de production de tout le groupement présente des inconvénients dans le cas du forage d'un nouveau puits ou dans le cas où l'on souhaite avoir la production exacte par puits. Afin de permettre à l'entreprise une meilleure maîtrise sur les puits individuellement, la solution prédictive utilisant les techniques de ML

fournissent une prédition journalière des valeurs de production de chacun des puits en se basant sur les paramètres de production. Pour le volet prédition nous avons suivi le processus CRISP-DM.

4.2 Conception de la zone de stockage DM

4.2.1 Choix de l'approche de modélisation

Etant donné que notre projet se concentre uniquement sur un sujet d'analyse spécifique, à savoir : la production, l'approche ascendante de **Kimball** offre une méthode structurée pour la mise en place d'un Datamart dédié à ce sujet en permettant une évolutivité pour prendre en compte leurs futurs besoins d'analyse.

A partir du fait et des dimensions proposées, il est possible d'établir une structure de données selon plusieurs formats : conception en étoile, en flocon de neige ou hybride. Le modèle en étoile constitué d'un fait central entourée de ses dimensions présente quelques inconvénients tels que la redondance dans les dimensions, l'alimentation complexe et la non-normalisation.

La modélisation en flocon de neige est une émanation de la modélisation en Etoile dans laquelle le fait est conservé et les dimensions sont éclatées en sous hiérarchies. Pour notre cas, il a été convenu d'opter pour une modélisation hybride étant donné le gros volume de données des différentes hiérarchies de la dimension « Puits ».

4.2.2 Choix d'Indicateurs

Il est important d'établir des indicateurs corrects pour les activités que l'on veut étudier. Comme cité dans le chapitre précédent, l'activité étudiée est la production journalière propre à chaque puits. En effet, chaque ligne de la table de fait doit représenter le taux de production de chaque fluide et le débit journalier prévu par puits à une date donnée.

Afin de pouvoir analyser le suivi de production sur une période donnée ou en fonction d'un axe d'analyse fixé, nous avons identifiés les indicateurs de performance ci-dessous :

- **Volume d'huile produit** : Le cumul d'huile réalisé.
- **Volume de Gaz Sec produit** : Le cumul de Gaz Sec réalisé.
- **Volume de Gaz GPL produit** : Le cumul de GPL réalisé.
- **Volume de Condensat produit** : Le cumul de Condensat réalisé.
- **Volume d'huile prévu** : Le seuil minimum fixé d'huile à produire.
- **Volume de Gaz Sec prévu** : Le seuil minimum fixé de Gaz Sec à produire.
- **Volume de Gaz GPL prévu** : Le seuil minimum fixé de GPL à produire.
- **Volume de Condensat prévu** : Le seuil minimum fixé de Condensat à produire.

4.3 Modèle multidimensionnel

4.3.1 Définition du grain

Le grain ou le niveau de granularité décrit le niveau de détail de représentation. Son choix fait objet de compromis entre avoir une plus grande précision pour les analyses et garder une taille raisonnable de l'entrepôt.

Ainsi, le fait ou le sujet d'analyse est représenté par une table qui englobe plusieurs mesures. Chaque ligne de la table de fait « **Production** » représente le taux d'huile ou de gaz produit pendant une période donnée.

Exemple : Temps : une journée.

4.3.2 Définition des dimensions

Après une analyse détaillée des différentes sources de données transmises et afin de répondre aux questions posées, nous avons pu dégager les différentes dimensions et hiérarchies :

- **La dimension Puits :**

Cette dimension regroupe tous les puits de tous les groupements confondus. Elle est générée à l'aide d'un processus ETL qui se charge de l'extraction à partir des données sources.

- **La dimension Réservoirs :**

Elle regroupe les réservoirs de tous les groupements confondus. Il est à noter qu'un puit exploite un seul réservoir. Elle est générée de la même manière que la dimension Puits.

- **La dimension Gisements :**

Cette dimension regroupe les gisements de chaque groupement. De la même manière, les instances de celle-ci sont extraites à partir des systèmes sources.

- **La dimension Périmètres :**

En général un périmètre est lui-même le gisement, mais dans certains cas, un périmètre peut se diviser en plusieurs gisements. Elle est remplie à partir des données sources.

- **La dimension Contrats :**

On retrouve dans cette dimension les différents contrats ainsi que leurs types. Les contrats se font par périmètre.

- **La dimension Bloc :**

Elle regroupe les blocs de tous les groupements confondus. Elle est générée de la même manière que les dimensions citées plus-haut.

- **La dimension Bassin :** Dimension spatiale utile pour l'analyse de la production des puits selon leur lieu géographique.

- **La dimension Groupement :**

Elle constitue le plus bas niveau de la hiérarchie. Le groupement englobe plusieurs bassins. Dans notre cas, on étudie un seul bassin sédimentaire par groupement.

- **La dimension Temps :**

Cette dimension assure l'historisation. Contrairement aux autres dimensions, le remplissage de cette dernière s'effectue à l'aide d'un script exécuté lors du chargement initial de notre DW. Le script génère toutes les dates à partir d'une date de début à une date de fin, tous deux initialisés au départ, y compris celles ne synchronisant pas avec la table de fait.

4.3.3 Définition des niveaux et des hiérarchies

Les hiérarchies sont les chemins d'accès dans les données (drill-down paths). On retrouve dans notre cas :

- **Hiérarchie 1 :** Dimension Temps : Jour → Mois → Trimestre → Année.

- **Hiérarchie 2 :** Dimension Puits : Puits → Réservoirs → Gisements → Périmètres → Bloc → Bassin → Groupements.

- **Hiérarchie 3 :** Dimension Puits : Puits → Réservoirs → Gisements → Périmètres → Contrats.

Comme la dimension Puits a été éclatée, chaque niveau de hiérarchie représente en fait une dimension à part entière avec une clé étrangère la reliant au niveau plus bas.

4.3.4 Définition de la table de fait

La conception de la table de faits permet d'avoir une vision consistante et globale des données de l'entreprise. A l'aide de la jointure entre les dimensions, on peut ainsi avoir les mesures recherchées sur différents axes.

Dans notre cas, un enregistrement d'une table de fait représente une production. La production d'un fluide a lieu dans un puit exploitant un réservoir bien précis, d'un gisement appartenant à un périmètre, inclus dans un bloc situé dans une zone géographique bien précise appelée « bassin sédimentaire » appartenant à un groupement précis dans un temps bien précis.

Exemple : Un puit a produit tant de Gaz Sec, GPL, Condensat et huile dans un jour donné.

4.3.5 Schéma multidimensionnel

La Figure 4.2 ci-dessous représente la modélisation de notre magasin de données :

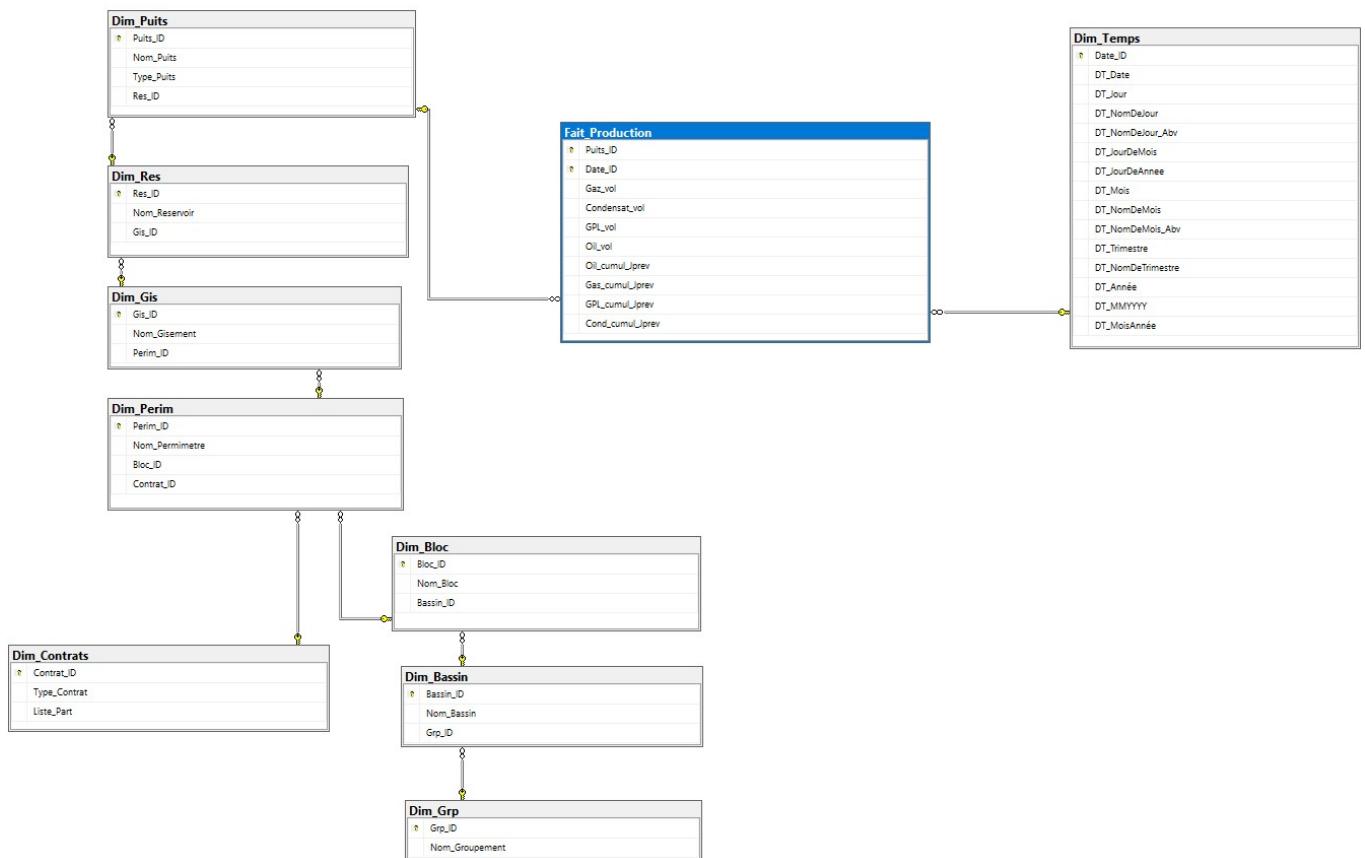


FIGURE 4.2 – Schéma du magasin de données

4.4 Conception de la zone d'alimentation

L'une des parties initiales de tout projet BI est l'identification des données décisionnelles les plus pertinentes à notre analyse afin de les importer depuis les sources et les stocker dans le DW.

4.4.1 Etude des sources de données

Comme expliqué précédemment, nous avons eu accès à des données sous différents formats. Etant donné le nombre de données perdues et non disponibles, un travail et une compréhension sur les données sources était nécessaire afin de répondre aux besoins.

- **Groupement Berkine :**

On retrouve pour ce groupement un fichier csv pour les données de production et un fichier Excel pour les données des contrats et type de contrats par périmètre. Les données de prévision pour ce champ n'ont pas été partagées.

Le fichier de production csv de ce groupement a été utilisé comme support pour adapter les autres fichiers à cette structure. Dans ce fichier, on retrouve tous les niveaux de granularité à compter du puits jusqu'au groupement, ainsi que les données de production huile, gaz, eau et d'autres paramètres d'injection, pression de tubing...

On remarque que le nom du puits est le nom du réservoir qu'il exploite suivi d'un numéro. Le nom du réservoir est le nom du gisement suivi de « - » et de son nom, idem pour les gisements qui suivent les noms de périmètres séparés toujours par « - ».

Exemple :

Nom périmètre : EMK
Nom gisement : EMK-EMN
Nom réservoir : EMN-Tagi

- **Groupement GSS :**

Pour ce groupement, nous avons les données de prévision et les données de production sous format Excel. Les données des contrats n'ont pas été partagées, il nous a alors été demandé d'affecter un contrat de notre choix.

Le fichier de production contient plusieurs feuilles de calcul, chacune d'elles comporte les données de production journalières d'huile, gaz, eau des puits d'un réservoir précis. Dans ce fichier, on retrouve uniquement le nom du puits et le nom de réservoir, le reste des hiérarchies doivent donc être générées.

- **Groupement GTIM :**

L'entreprise nous a fourni un dossier contenant des rapports journaliers au format Excel de 2018 à 2022. Chaque fichier contient le bilan de production global du groupement (production journalière, cumul mensuel, cumul annuel, cumul depuis l'origine), le bilan de production par périmètre, et enfin la production journalière par puits (chaque bilan est contenu dans une feuille de calcul Excel). Il faut donc extraire les informations né-

cessaires, les traiter et les concaténer afin d'obtenir, en fin de compte, un fichier global journalier de la production par puits. Nous avons également un autre fichier décrivant les gisements, les contrats et le bassin sédimentaire du groupement.

- **Groupement GTFT :**

Ce groupement a été partagé sous la forme d'une base de données Access contenant un grand nombre de tables. La table la plus pertinente pour notre cas est la table **Production Journalière** des 3 fluides par puits. La BDD contient également une table décrivant la hiérarchie entre les puits et les réservoirs. Nous avons également obtenu un fichier Excel regroupant des informations sur les contrats, le type de contrat et l'emplacement du groupement.

Cependant, il n'y a pas de tables bien structurées pour décrire chaque hiérarchie (comme une table pour les puits, les réservoirs, etc.) dans cette base Access, ce qui nous oblige à créer une nouvelle base de données contenant les tables nécessaires et les relations entre elles.

4.4.2 Processus ETL

Comme expliqué dans le chapitre 1, le processus d'ETL consiste en premier lieu à extraire les données depuis les sources, faire les transformations nécessaires selon les besoins et enfin les charger dans le DW.

Dans notre cas, la transformation de données a nécessité énormément de temps étant donné le nombre de données perdues et non disponibles. Le but ultime est de traiter deux sources de données hétérogènes :

- Charger la base de données ACCESS sur SQL Server
- Consolider toutes les données de production des fichiers (Excel, csv) des différents groupements dans un seul fichier de sortie final pour optimiser les traitements répétitifs et le remplissage de l'entrepôt.

On procède en premier lieu à la modification de chaque fichier source en effectuant une suite d'opérations tels que :

- Simplifier les noms des attributs.
- Supprimer les doublons
- Ajouter des attributs si nécessaire
- Nettoyer les valeurs manquantes, aberrantes...
- Réduire les valeurs d'attributs
- Convertir les types d'attributs pour avoir une structure homogène à celle de l'entrepôt de données.

On effectue par la suite, la fusion des fichiers de production modifiés dans un seul et même fichier de sortie final regroupant toutes les données de production de tous les groupements

réunis. A partir de ce fichier (Berkine, GSS, GTIM) et de la base de données sur SQL Server (GTFT), on charge les dimensions une à une à partir du niveau de hiérarchie le plus bas (Dimension Groupements jusqu'à la dimension Puits) et la table de faits par la suite.

Cette opération constitue la dernière étape du processus ETL (phase d'alimentation). Vu le volume de données, le transfert et l'insertion des données nettoyées et préparées dans l'entrepôt de données sur SQL Server prend beaucoup de temps. Les étapes de cette partie seront bien évidemment plus détaillées dans le prochain chapitre.

4.5 Conception de la zone de restitution

Une fois les dimensions et la table de fait alimentée, la dernière étape est la prospection des données dans un Dashboard afin de permettre aux décideurs de les visualiser quotidiennement et de contrôler l'évolution de production de chaque puits pour intervenir en cas de problème.

4.6 Conception de la solution prévisionnelle avec Time Series Analysis

Pour cette partie, nous allons effectuer CRISP DM pour la l'analyse de séries temporelle sur de production mensuelle (mars 1999 à avril 2022) de gaz sec du groupement GTFT récupérés depuis la base de données SQL SERVER déjà créée.

4.6.1 Compréhension des données

1. Nous avons effectué une analyse graphique de la série chronologique pour déterminer le type de tendance et de saisonnalité mais aussi le type de décompositions (additive ou multiplicative) à faire.
2. Pour extraire les composantes (Tendance, saisonnalité et résidus) de notre série et mieux comprendre sa nature et ses caractéristiques générales, nous avons effectué la décomposition des tendances saisonnières (Seasonal-trend decomposition).
3. Afin de traiter les valeurs aberrantes de production, la détection des anomalies à partir des résidus est primordiale.
4. Test de stationnarité : Vérifiant si la série chronologique étudiée a des propriétés statistiques constantes dans le temps.
 - Nous avons analysé le corrélogramme simple et partiel pour déterminer la stationnarité ou non de la série.
 - Nous avons calculé et tracé le graphique de la moyenne mobile et de l'écart type mobile Pour une analyse plus pertinente concernant la stationnarité.
 - Nous avons procédé au test de Dickey-Fuller sur la série.

4.6.2 Préparation des données

1. Au premier plan, nous devons traiter les outliers déjà détectés à partir des résidus (anomalies de production).
2. Nous allons rendre la série stationnaire si elle ne l'est pas, en traitant le changement de la moyenne par la différenciation pour enlever la tendance, ainsi que le changement de volatilité.
3. Nous avons divisé nos données en un ensemble d'entraînement et un ensemble de test. Cette division nous permet d'évaluer les performances de nos modèles sur des données futures non vues.

4.6.3 Modélisation

Après avoir analysé et préparé les données (Stationnarisation si nécessaire), cette étape consiste à modéliser notre série.

On doit choisir quel modèle convient le mieux à notre série temporelle entre ARMA, ARIMA et SARIMA mais aussi entre lissage exponentiel basique, lissage exponentiel double et lissage exponentiel triple (Holt-Winters).

En revanche, les modèles de lissage exponentiel (Simple, Double et Holt-winters) sont plus flexibles qui peuvent être appliqués à des séries non stationnaires, ils estiment les tendances et les saisons de manière adaptative donc pas besoin de réaliser une Stationnarisation de la série.

Les modèles utilisés sont :

- Saisonnalité Autoregressive Integrated Moving Average (SARIMA)
 - Holt-winters Exponential Smoothing
1. **SARIMA** : À l'aide des graphiques de l'autocorrélation (ACF) et de l'autocorrélation partielle (PACF), nous identifions les paramètres du modèle SARIMA, tels que les termes AR (autorégressifs), MA (moyenne mobile) et saisonniers. Ensuite, nous utilisons une technique d'estimation pour obtenir les meilleurs paramètres du modèle SARIMA pour notre série originale stationnaire , en comparant les valeurs AIC des différentes combinaisons de modèles possibles.
 2. **Holt-winters Exponential Smoothing** : d'après la décomposition (Seasonal-trend decomposition) déjà effectuer, on précise le type de tendance et de la saisonnalité de la série (additive ou multiplicative).

4.6.4 Évaluations

1. Évaluation des paramètres du modèle SARIMA :

- Nous interprétons le résumé du modèle SARIMA après application sur la série originale stationnaire (Métriques d'évaluation du modèle, évaluation des coefficients du modèle et tests de diagnostic sur les résidus). Nous traçons le corrélogramme de la

série résiduelle. Nous réalisons également un test de normalité des résidus.

2. Évaluation des modèles :

- Ajustement des modèles sur la série : pour les deux modèles (SARIMA et Holt-winters Exponential Smoothing), nous calculons la série estimée sur l'ensemble d'entraînement et la comparons avec la série originale (non stationnaire) graphiquement et avec du coefficient de détermination R^2 .
- Test Evaluation : Après avoir entraîné les deux modèles sur les données d'apprentissage, nous les avons évalués en utilisant les données de test sur une durée de 12 mois en calculant MAPE (Mean Absolute Percentage Error), MAE (Mean Absolute Error) et RMSE (Root Mean Square Error).

3. Comparaison des modèles :

- Nous avons effectué une comparaison entre les deux modèles (graphes et Mesures) pour déterminer le modèle optimal pour notre jeu de données afin de l'adopter comme solution.
4. Une comparaison finale doit être effectuer entre le modèle choisi et les résultats de prévision de la méthode Declin Curve Analysis.

4.6.5 Déploiement

Ce modèle implémenté nous permet de fournir des prévisions des valeurs futures de production de gaz du groupement GTFT pour un horizon défini de mois.

4.7 Conception de la solution prédictive en utilisant le Machine Learning

4.7.1 Compréhension des données

Pour un bon entraînement du modèle, il est impératif d'avoir un gros volume de données variées, cohérentes et complètes.

Pour notre cas, nous avons effectué l'analyse sur toutes les données citées précédemment, et nous avons constaté qu'aucune de ces sources n'est applicable au Machine Learning pour les raisons suivantes :

- Beaucoup de données manquantes.
- Pas de corrélation entre les variables et la production.
- Redondance des données.

Un autre dataset nous a alors été partagé. Ce fichier Excel contient les données de production journalières de Gaz et d'huile des puits du groupement **Berkine** avec les paramètres pression fond du puits, taille vanne de cuvelage, Gaz lift injecté, GOR...

Ce nouvel ensemble présente également une absence de corrélations entre les variables et la production d'huile, la prédiction de la production d'huile ne peut donc être effectuée.

Néanmoins, l'analyse concernant la prédiction de gaz est un peu plus satisfaisante que les précédentes. Malgré le manque de diversité et de variété des données, il existe une corrélation entre la production de gaz le volume de gaz lift injecté, la taille de la vanne, la température ambiante du puits, la pression, et le GOR.

4.7.2 Préparation des données

Une fois les données collectées et analysées, cette étape consiste à remplacer les valeurs manquantes, éliminer les valeurs aberrantes, sélectionner les variables et les standardiser afin de mettre les données sur une échelle commune avant de les préparer pour l'étape de modélisation. Afin de sélectionner les features les plus importantes, il faut étudier la relation entre les valeurs de ces attributs et la valeur de la production.

Avant d'effectuer la prédiction, le dataset est divisé en deux catégories :

- **Train** : pour l'entraînement du modèle.
- **Test** : pour évaluer si les valeurs prédites coïncident avec les données déjà disponibles.

4.7.3 Modélisation

En utilisant les données préparées, cette étape consiste à construire les modèles de prédiction choisis. Parmi les nombreux algorithmes de Machine Learning (classification supervisée, non-supervisée...), les plus appropriés à notre cas d'étude sont les algorithmes de régression.

Nous avons donc sélectionné les plus populaires et les plus utilisés dans le domaine pétrolier :

- Support Vector Machine SVR
- Decision Tree Regressor DTR
- Random Forest Regressor RFR
- Gradient Boosting Regressor GBR
- Extreme Gradient Boosting Regressor XGBOOST

Pour plus de précision, il est conseillé d'utiliser les méthodes pour rechercher les meilleurs paramètres du modèle.

4.7.4 Evaluation

Pour notre cas, l'évaluation de performance de nos modèles se fait à l'aide des indicateurs suivants :

- R^2
- RMSE
- MSE
- MAE

Afin de déterminer le modèle le plus performant et le valider comme modèle final, on procède à la comparaison des résultats obtenus.

4.7.5 Déploiement

En insérant les tests mesurés sur les paramètres d'un puits (pression du puits, GL injecté...), le modèle de prédiction permet aux utilisateurs finaux de déterminer les valeurs futures de production de gaz de ce puits.

Conclusion

Dans ce chapitre, nous avons présenté le schéma conceptuel de notre solution analytique ainsi que celle de la prédiction. Plus de détails sur l'application et la mise en œuvre de ces solutions seront apportées dans le chapitre suivant.

Chapitre 5

IMPLÉMENTATION ET MISE EN OEUVRE

Introduction

Arrivés à la fin de ce manuscrit, ce chapitre sera consacré à la présentation des outils utilisés, et éventuellement l'explication détaillée de la solution, et enfin, la comparaison des résultats obtenus en appliquant différentes méthodes de prédiction basées sur les modèles autorégressifs(séries chronologiques) et les modèles d'apprentissage automatique.

5.1 Présentation des outils utilisés

5.1.1 Outils de stockage : SQL Server

Produit par Microsoft, SQL Server 2019 est un système de gestion de bases de données relationnelles. Le stockage, la manipulation et l'analyse de ces données se font au sein de son moteur de bases de données. Ce service permet la réalisation de nombreuses applications, requêtes, et transactions, notamment grâce au langage T-SQL (Transact-SQL). [39]

SQL Server 2019 contient trois plateformes, indispensables pour réaliser un projet BI :

- Integration Services pour l'intégration des données dans l'entrepôt central.
- Analysis Services pour l'analyse des données et le stock dans des cubes multidimensionnels.
- Reporting Services pour la création et publication des rapports résultant des analyses.

5.1.2 Outils de collecte : SSIS

SQL Server Integration Services (SSIS) est une plate-forme permettant d'extraire, transformer et intégrer les données à partir de multiples sources (fichiers Excel, csv, fichiers plats, BDD relationnelles...) dans une ou plusieurs destinations. pour les ranger dans un entrepôt central.

5.1.3 Outils de restitution : Power BI

Microsoft Power BI est un ensemble d'outils d'analyse décisionnelle développée par Microsoft, il permet aux entreprises d'analyser et de visualiser des données résultant de différentes sources.[55]

Il permet aussi de partager de façon sécurisée les données en les transformant en informations exploitables sous forme de tableaux de bord afin d'aider à prendre des décisions stratégiques. Cette solution permet à tous les employés de comprendre les données et de les exploiter pour fournir des rapports et des analyses à leurs entreprises et d'augmenter leur productivité et leur créativité. L'outil Power BI Desktop permet notamment d'ordonner des données en provenance de nombreuses sources sur site ou sur le Cloud : bases de données, fichiers, services web.

5.1.4 Langage de programmation : Python

Python est un langage de programmation open source multiplateformes et orienté objet principalement utilisé pour le Scripting et l'automatisation de tâches simples mais fastidieuses, le Machine Learning et la Data Science. Python a fortement contribué à l'essor du Big Data.[50]

En effet, parmi ses qualités, Python permet notamment aux développeurs de se concentrer sur ce qu'ils font plutôt que sur la manière dont ils le font. Il a libéré les développeurs des contraintes de formes qui occupaient leur temps avec les langages plus anciens. Ainsi, développer du code avec Python est plus rapide qu'avec d'autres langages, grâce à ses nombreuses bibliothèques spécialisées. On cite :

- **Pandas** est une bibliothèque open source pour le traitement et l'analyse de données simples et intuitives en Python. Elle fournit une structure de données appelée Data Frame pour un traitement rapide et efficace des données avec indexation intégrée ; une flexibilité dans le traitement des données hétérogènes ou manquantes. Comme elle contient des outils pour lire et écrire dans des fichiers de différents formats (.csv, .txt, .xls, .sql, etc...).[51]
- **Scikit-learn (sklearn)** est une bibliothèque d'apprentissage automatique gratuite qui supporte l'apprentissage supervisé et non supervisé. Il fournit également une variété d'outils pour le réglage du modèle, le pré-traitement des données, le choix du modèle, l'évaluation du modèle et beaucoup d'autres utilitaires.[52]
- **Statsmodels** est un module Python qui propose des classes et des fonctions pour l'estimation de nombreux modèles statistiques différents, ainsi que pour la réalisation de tests statistiques et l'exploration de données statistiques. Une vaste gamme de statistiques de résultats est disponible pour chaque estimateur. Les résultats sont vérifiés par rapport aux packages statistiques existants afin de garantir leur exactitude.[54]
- **Sktim** offre un cadre ouvert, flexible et modulaire facile à utiliser pour une large gamme de tâches d'apprentissage automatique sur les séries temporelles. Il propose des interfaces compatibles avec scikit-learn et des outils de composition de modèles, dans le but de rendre l'écosystème plus utilisable et interopérable dans son ensemble. Nous construisons et soutenons une communauté ouverte, diversifiée et autonome, accueillant de nouveaux contributeurs issus du milieu universitaire et de l'industrie grâce à une documentation instructive, du mentorat et des ateliers.[53]

5.2 Volet BI

5.2.1 Mise en œuvre de l'ETL

Comme expliqué dans le chapitre précédent, plusieurs modifications ont été nécessaires. Pour ce faire, nous avons effectué les gros traitements de transformations sur Python. L'outil SSIS a été utilisé uniquement pour extraire les données sources depuis le fichier Excel et la BDD relationnelle SQL Server afin de les combiner et les stocker dans l'entrepôt central.

1. Prétraitement des données BI sur Python

- **Prévisions GTIM / GTFT**

La Figure 5.1 représente la feuille de calcul contient deux tableaux distincts pour chaque groupement.

FIGURE 5.1 – Préavisons des groupement GTFT / GTIM Brutes

Nous avons donc ajouté la colonne du nom du groupement et l'ID groupement afin de faire la jointure avec l'ID groupement de la dimension **Groupements** du DW.

Le groupement GTIM produit uniquement le Gaz Sec, on a donc ajouté les colonnes de prévisions GPL et Condensat initialisés à 0 ainsi que l'ID et le nom de groupement afin qu'il ait la même structure que le tableau GTFT pour concaténer les deux tableaux et optimiser le traitement décrit ci-dessous.

Comme les prévisions se font par groupement mensuellement, il a fallu diviser la donnée par le nombre de puits de chaque groupement et par 30 pour obtenir les prévisions journalières par puits. Comme ces groupements sont de type gaz, on a ajouté la colonne de prévisions d'huile à valeurs 0.

On a également dû générer une colonne contenant l'ID de la date sous format « **MMYYYY** » pour effectuer la jointure avec l'attribut de la dimension temps du DW. Trois ans de prévisions mensuelles pour deux groupements, on obtient donc à la fin un fichier (Figure 5.64 en Annexe) de 72 lignes.

- **Prévisions GSS**

Le fichier comporte la colonne année et la prévision d'huile uniquement. On ajoute les colonnes de prévisions de **GPL** et de **Condensat** avec une valeur égale à 0, étant donné que c'est un groupement de type **Oil**.

Pour avoir la donnée journalière prévue, on effectue la même opération citée plus haut (i.e donnée/ 365 et sur le nombre de puits du groupement). De la même façon, on ajoute également l'ID et le nom du groupement pour effectuer la jointure lors du remplissage du DW.

1	Year	ID_Grp	Nom_Grp	Prev_Oil_J	Prev_Cond_M_J	Prev_GPL_M_J
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						

FIGURE 5.2 – Prévisions du regroupement GSS

- **Production Berkine** : Le fichier de production .csv (Figure 5.3) délimité par des virgules, contient toutes les colonnes importantes pour le remplissage des dimensions et des colonnes indésirables que nous avons supprimées.

FIGURE 5.3 – Production du Groupement Berkine

Quelques réservoirs ont été perdus, nous les avons donc remplis en effectuant une extraction du nom de gisement auquel il appartient (sur la même ligne) et en ajoutant l'extension qui lui correspond (comme expliqué dans le chapitre précédent).

Les types de certaines colonnes (éventuellement celles de la date et les débits de production) ont été modifiés car nous avons constaté que leurs types n'étaient pas conformes aux types des attributs des dimensions de notre DW.

Dans ce groupement, on retrouve 5 types de puits (**Gaz, Oil, Water, Water Injection et Gas Injection**). Afin de préserver le volume de données de ce fichier, nous avons calculé pour chaque type, la moyenne de production des fluides (Gaz Sec et Oil) et nous avons remplacé toutes les données de production manquantes par la moyenne correspondante.

Ci-dessous Figure 5.4, le fichier où se trouvent les contrats par périmètre :

FIGURE 5.4 – Contrats par périmètre

Pour ajouter ces informations à notre fichier final, nous avons relevé les périmètres et les contrats associés à chacun d'eux, ainsi que le type de contrats.

Figure 5.65 en Annexe représente le fichier final obtenu comporte 325 459 lignes.

- Production GSS

FIGURE 5.5 – Production du groupement GSS

Comme les données de production (Figure 5.5) sont réparties sur plusieurs feuilles de calcul, nous avons procédé à la fusion des feuilles afin d'avoir un fichier de sortie final regroupant tous les réservoirs réunis.

Afin de générer les colonnes manquantes (ID réservoirs, nom périmètre, nom gisement...), nous avons suivi la structure du grouement Berkine.

C'est-à-dire, à partir des noms des réservoirs, nous avons extrait la partie précédant le « - » qui représente le nom du gisement (il n'y a qu'un seul gisement, et donc un seul périmètre aussi).

L'ajout de la colonne ID réservoir a été générée en extrayant la partie qui suit le « - ».

Les autres ID (ID puits, ID périmètre, ID gisement...), ont été générés à l'aide d'une fonction « **.map** » et d'un dictionnaire qui attribue à chaque élément différent, un ID spécifique.

Les colonnes inutiles ont été supprimées et les colonnes type de puits, listes de partenaires, type de contrats ont été ajoutées.

On obtient en sortie (Figure 5.66 en Annexe), un fichier de 25 colonnes et 22 436 lignes.

• Production GTIM

Comme nous l'avons expliqué précédemment, il est nécessaire de consolider le contenu du dossier fourni afin d'obtenir un fichier unique de données de production pour ce groupement.

Figure 5.67 en Annexe représente les données sont partagées en 4 dossiers. Chacun d'entre eux contient les rapports journaliers de production d'une année (Figure 5.68 en Annexe).

Chacun de ces fichiers contient plusieurs feuilles de calcul (Figure 5.6), notamment la troisième page qui nous intéresse qui est le **Rapport journalier de production**.

The screenshot shows a Microsoft Excel spreadsheet titled "RAPPORT JOURNALIER DE PRODUCTION". At the top left, there are logos for SOLENTECH, CEPSA, and TOTAL. Below the title, it says "08) Puits producteurs de GAZ :". The main area is a large grid with various columns and rows of data. The columns include: Périmètre, Puits, Heure Marche (Jour, Cumul depuis origine), Dose (x/64), Pression tête (Barg), Temp. (deg.C), Pression Flow Line, Temp.FL on Line, Chemical injection, Pression Espace annulaire (Bar), Lj, SCSSV, SSV Header Pressure (bar), Vol. Gaz (10³ Sm3) (Jour, Cumul depuis origine), and Vol. Condensat (Tonne) (Jour, Cumul depuis origine). The rows represent different data points over time. At the bottom of the grid, there are tabs for "Bilan Gaz-Condensat", "Bilan par périmètre", "Etats des puits", and "COMMENTAIRE".

FIGURE 5.6 – Exemple d'un rapport journalier GTIM

Cette feuille est composée d'un tableau décrivant la production journalière de gaz par périmètre et par puits, avec des paramètres sur l'état de ces derniers tels que la Pression tête **WHP** (Barg), la Température **WHT** (deg.C), la Pression Flow Line **FLP** et la Température Flow Line **FLT**.

Il est clairement visible que la colonne de date est absente. Cette information est normalement incluse dans le nom du fichier, par exemple **canevas journalier 01-11-2018.xlsx**, ou alors elle peut être présente dans une cellule du même rapport Figure 5.7.

The screenshot shows the same "RAPPORT JOURNALIER DE PRODUCTION" spreadsheet as Figure 5.6. A red circle highlights the date "Du : 1 novembre 2018" located in the header area above the grid. The rest of the grid and interface are identical to Figure 5.6.

FIGURE 5.7 – L'emplacement de Date dans le fichier

Au premier plan et avec une analyse visuelle des fichiers Excel, nous avons remarqué que les dimensions des feuilles de calculs d'années différentes ne sont pas les mêmes, et dans leur globalité, chaque année possède une dimension unique (même emplacement du tableau et de la date dans les feuilles). Par conséquent, pour une première solution, nous avons décidé de gérer chaque année à part entière.

Cependant, après de nombreux tests, nous avons constaté que même pour une seule année, il existe des feuilles de différentes dimensions. Il est donc préférable de changer notre

approche pour la rendre plus générale, c'est-à-dire la gestion par dimensions (nombre de lignes et de colonnes) des feuilles.

Ainsi, nous avons continué à traiter dossier par dossier, mais le test se fait désormais par dimensions, et non par année. Au fur et à mesure, nous avons remarqué que même des feuilles d'années différentes pouvaient avoir des dimensions identiques.

L'idée est donc de générer un script pour un dossier d'une année donnée et de traiter tous les cas possibles, puis de le tester sur l'année suivante et de le modifier ou d'ajouter d'autres cas jusqu'à parvenir à une gestion totale de toutes les dimensions possibles des feuilles.

Nous avons identifié six cas de dimensions possibles pour ces feuilles : le nombre de colonnes peut prendre les valeurs **24, 27, 28, 29, 30 ou 31**, car il y a des colonnes dans le tableau avec une taille de 2 ou 3 colonnes Excel.

Pour chaque cas, nous devons également vérifier si le nombre de lignes est équivalent, c'est-à-dire si l'emplacement du tableau est le même pour tous les cas.

Cependant, ce n'est pas vrai même pour les feuilles ayant le même nombre de colonnes. Par conséquent, nous devons spécifier l'emplacement du tableau pour chaque cas afin de l'extraire correctement.

On propose un script sous Python Voici son PseudoCode (Algorithm 1) :

Algorithm 1 Traitement des fichiers XLSX GTIM

Require: Le chemin du dossier contenant les fichiers XLSX

```
1: Importer les bibliothèques nécessaires (pandas, openpyxl, re, os)
2: for all fichier In dossier do
3:   if fichier est de type XLSX then
4:     Lire le fichier
5:     Récupérer la feuille calcul du fichier
6:     Déterminer la dimension de la feuille
7:     Extraire la date en utilisant une expression régulière
8:     Convertir au format de date approprié
9:     Récupérer le tableau en DataFrame
10:    Modifier le schéma
11:    Supprimer les colonnes supplémentaires ou vides
12:    Remplir la colonne "périmètre"
13:    Supprimer les anomalies
14:    Ajouter une colonne "Date" au DataFrame
15:    Remplir chaque ligne avec la date récupérée précédemment
16:    if Le fichier CSV final existe déjà then
17:      Ajouter le DataFrame au fichier existant
18:    else
19:      Créer un nouveau fichier CSV et le remplir avec le DataFrame courant
20:  Le traitement est terminé
```

Plus de détails voir en Annexe.

Une fois le fichier consolidé et enregistré sous format csv, nous avons procédé à la transformation afin d'avoir un fichier de sortie final ayant une même structure que les fichiers précédents.

Comme expliqué plus haut, la génération de la colonne nom gisement se fait en suivant les noms de périmètre. Ensuite, pour avoir les noms des réservoirs, on a ajouté les extensions en suivant le nom du gisement correspondant.

Les noms de bassins, contrats et types contrats étant sur un autre fichier (Figure 5.69 en

Annexe), il a fallu récupérer ces colonnes depuis leurs sources et les ajouter au fichier de production GTIM. Cet ensemble de données (Figure 5.70 en Annexe) contient au total 35 732 lignes.

Une fois les modifications nécessaires à chaque fichier apportées, on les concatène afin d'avoir un fichier de sortie final sous format Excel qui regroupe donc les 3 groupements : **Berkine, GSS et GTIM**.

On vérifie que tous les types d'attributs sont bons et on ajoute les colonnes de production de **Condensat** et de **Gaz GPL** à valeurs 0 étant donné qu'aucun de ces groupements ne les produit.

On compte (Figure 5.8) au total 25 colonnes pour 383 628 lignes.

ID_Grp	Nom_Grp	D_Bassin	Nom_Bassin	ID_Bloc	Nom_Bloc	ID_Perimetre_Perf	ID_Gis	Nom_Gid	ID_Res	Nom_Res	ID_Puits	Puits	Type_Puit	Date	Baz	Sec	Densat	3Pt.	Torjò	Oli	mD.	Contr	Partenaire	Type_Cons	
1	GRP 1	1	Basin 1	1	Bloc 1	1	Gis 1	Gid 1	1	Res 1	1	Puits 1	1	1	2023-01-01	1	1	1	1	1	1	1	1	1	1
2	GRP 2	2	Basin 2	2	Bloc 2	2	Gis 2	Gid 2	2	Res 2	2	Puits 2	2	2	2023-01-01	2	2	2	2	2	2	2	2	2	2
3	GRP 3	3	Basin 3	3	Bloc 3	3	Gis 3	Gid 3	3	Res 3	3	Puits 3	3	3	2023-01-01	3	3	3	3	3	3	3	3	3	3
4	GRP 4	4	Basin 4	4	Bloc 4	4	Gis 4	Gid 4	4	Res 4	4	Puits 4	4	4	2023-01-01	4	4	4	4	4	4	4	4	4	4
5	GRP 5	5	Basin 5	5	Bloc 5	5	Gis 5	Gid 5	5	Res 5	5	Puits 5	5	5	2023-01-01	5	5	5	5	5	5	5	5	5	5
6	GRP 6	6	Basin 6	6	Bloc 6	6	Gis 6	Gid 6	6	Res 6	6	Puits 6	6	6	2023-01-01	6	6	6	6	6	6	6	6	6	6
7	GRP 7	7	Basin 7	7	Bloc 7	7	Gis 7	Gid 7	7	Res 7	7	Puits 7	7	7	2023-01-01	7	7	7	7	7	7	7	7	7	7
8	GRP 8	8	Basin 8	8	Bloc 8	8	Gis 8	Gid 8	8	Res 8	8	Puits 8	8	8	2023-01-01	8	8	8	8	8	8	8	8	8	8
9	GRP 9	9	Basin 9	9	Bloc 9	9	Gis 9	Gid 9	9	Res 9	9	Puits 9	9	9	2023-01-01	9	9	9	9	9	9	9	9	9	9
10	GRP 10	10	Basin 10	10	Bloc 10	10	Gis 10	Gid 10	10	Res 10	10	Puits 10	10	10	2023-01-01	10	10	10	10	10	10	10	10	10	10
11	GRP 11	11	Basin 11	11	Bloc 11	11	Gis 11	Gid 11	11	Res 11	11	Puits 11	11	11	2023-01-01	11	11	11	11	11	11	11	11	11	11
12	GRP 12	12	Basin 12	12	Bloc 12	12	Gis 12	Gid 12	12	Res 12	12	Puits 12	12	12	2023-01-01	12	12	12	12	12	12	12	12	12	12
13	GRP 13	13	Basin 13	13	Bloc 13	13	Gis 13	Gid 13	13	Res 13	13	Puits 13	13	13	2023-01-01	13	13	13	13	13	13	13	13	13	13
14	GRP 14	14	Basin 14	14	Bloc 14	14	Gis 14	Gid 14	14	Res 14	14	Puits 14	14	14	2023-01-01	14	14	14	14	14	14	14	14	14	14
15	GRP 15	15	Basin 15	15	Bloc 15	15	Gis 15	Gid 15	15	Res 15	15	Puits 15	15	15	2023-01-01	15	15	15	15	15	15	15	15	15	15
16	GRP 16	16	Basin 16	16	Bloc 16	16	Gis 16	Gid 16	16	Res 16	16	Puits 16	16	16	2023-01-01	16	16	16	16	16	16	16	16	16	16
17	GRP 17	17	Basin 17	17	Bloc 17	17	Gis 17	Gid 17	17	Res 17	17	Puits 17	17	17	2023-01-01	17	17	17	17	17	17	17	17	17	17
18	GRP 18	18	Basin 18	18	Bloc 18	18	Gis 18	Gid 18	18	Res 18	18	Puits 18	18	18	2023-01-01	18	18	18	18	18	18	18	18	18	18
19	GRP 19	19	Basin 19	19	Bloc 19	19	Gis 19	Gid 19	19	Res 19	19	Puits 19	19	19	2023-01-01	19	19	19	19	19	19	19	19	19	19
20	GRP 20	20	Basin 20	20	Bloc 20	20	Gis 20	Gid 20	20	Res 20	20	Puits 20	20	20	2023-01-01	20	20	20	20	20	20	20	20	20	20
21	GRP 21	21	Basin 21	21	Bloc 21	21	Gis 21	Gid 21	21	Res 21	21	Puits 21	21	21	2023-01-01	21	21	21	21	21	21	21	21	21	21
22	GRP 22	22	Basin 22	22	Bloc 22	22	Gis 22	Gid 22	22	Res 22	22	Puits 22	22	22	2023-01-01	22	22	22	22	22	22	22	22	22	22
23	GRP 23	23	Basin 23	23	Bloc 23	23	Gis 23	Gid 23	23	Res 23	23	Puits 23	23	23	2023-01-01	23	23	23	23	23	23	23	23	23	23
24	GRP 24	24	Basin 24	24	Bloc 24	24	Gis 24	Gid 24	24	Res 24	24	Puits 24	24	24	2023-01-01	24	24	24	24	24	24	24	24	24	24
25	GRP 25	25	Basin 25	25	Bloc 25	25	Gis 25	Gid 25	25	Res 25	25	Puits 25	25	25	2023-01-01	25	25	25	25	25	25	25	25	25	25
26	GRP 26	26	Basin 26	26	Bloc 26	26	Gis 26	Gid 26	26	Res 26	26	Puits 26	26	26	2023-01-01	26	26	26	26	26	26	26	26	26	26
27	GRP 27	27	Basin 27	27	Bloc 27	27	Gis 27	Gid 27	27	Res 27	27	Puits 27	27	27	2023-01-01	27	27	27	27	27	27	27	27	27	27
28	GRP 28	28	Basin 28	28	Bloc 28	28	Gis 28	Gid 28	28	Res 28	28	Puits 28	28	28	2023-01-01	28	28	28	28	28	28	28	28	28	28
29	GRP 29	29	Basin 29	29	Bloc 29	29	Gis 29	Gid 29	29	Res 29	29	Puits 29	29	29	2023-01-01	29	29	29	29	29	29	29	29	29	29
30	GRP 30	30	Basin 30	30	Bloc 30	30	Gis 30	Gid 30	30	Res 30	30	Puits 30	30	30	2023-01-01	30	30	30	30	30	30	30	30	30	30
31	GRP 31	31	Basin 31	31	Bloc 31	31	Gis 31	Gid 31	31	Res 31	31	Puits 31	31	31	2023-01-01	31	31	31	31	31	31	31	31	31	31
32	GRP 32	32	Basin 32	32	Bloc 32	32	Gis 32	Gid 32	32	Res 32	32	Puits 32	32	32	2023-01-01	32	32	32	32	32	32	32	32	32	32
33	GRP 33	33	Basin 33	33	Bloc 33	33	Gis 33	Gid 33	33	Res 33	33	Puits 33	33	33	2023-01-01	33	33	33	33	33	33	33	33	33	33
34	GRP 34	34	Basin 34	34	Bloc 34	34	Gis 34	Gid 34	34	Res 34	34	Puits 34	34	34	2023-01-01	34	34	34	34	34	34	34	34	34	34
35	GRP 35	35	Basin 35	35	Bloc 35	35	Gis 35	Gid 35	35	Res 35	35	Puits 35	35	35	2023-01-01	35	35	35	35	35	35	35	35	35	35
36	GRP 36	36	Basin 36	36	Bloc 36	36	Gis 36	Gid 36	36	Res 36	36	Puits 36	36	36	2023-01-01	36	36	36	36	36	36	36	36	36	36
37	GRP 37	37	Basin 37	37	Bloc 37	37	Gis 37	Gid 37	37	Res 37	37	Puits 37	37	37	2023-01-01	37	37	37	37	37	37	37	37	37	37
38	GRP 38	38	Basin 38	38	Bloc 38	38	Gis 38	Gid 38	38	Res 38	38	Puits 38	38	38	2023-01-01	38	38	38	38	38	38	38	38	38	38
39	GRP 39	39	Basin 39	39	Bloc 39	39	Gis 39	Gid 39	39	Res 39	39	Puits 39	39	39	2023-01-01	39	39	39	39	39	39	39	39	39	39
40	GRP 40	40	Basin 40	40	Bloc 40	40	Gis 40	Gid 40	40	Res 40	40	Puits 40	40	40	2023-01-01	40	40	40	40	40	40	40	40	40	40
41	GRP 41	41	Basin 41	41	Bloc 41	41	Gis 41	Gid 41	41	Res 41	41	Puits 41	41	41	2023-01-01	41	41	41	41	41	41	41	41	41	41
42	GRP 42	42	Basin 42	42	Bloc 42	42	Gis 42	Gid 42	42	Res 42	42	Puits 42	42	42	2023-01-01	42	42	42	42	42	42	42	42	42	42
43	GRP 43	43	Basin 43	43	Bloc 43	43	Gis 43	Gid 43	43	Res 43	43	Puits 43	43	43	2023-01-01	43	43	43	43	43	43	43	43	43	43
44	GRP 44	44	Basin 44	44	Bloc 44	44	Gis 44	Gid 44	44	Res 44	44	Puits 44	44	44	2023-01-01	44	44	44	44	44	44	44	44	44	44
45	GRP 45	45	Basin 45	45	Bloc 45	45	Gis 45	Gid 45	45	Res 45	45	Puits 45	45	45	2023-01-01	45	45	45	45	45	45	45	45	45	45
46	GRP 46	46	Basin 46	46	Bloc 46	46	Gis 46	Gid 46	46	Res 46	46	Puits 46	46	46	2023-01-01	46	46	46	46	46	46	46	46	46	46
47	GRP 47	47	Basin 47	47	Bloc 47	47	Gis 47	Gid 47	47	Res 47	47	Puits 47	47	47	2023-01-01	47	47	47	47	47	47	47	47	47	47
48	GRP 48	48	Basin 48	48	Bloc 48	48	Gis 48	Gid 48	48	Res 48	48	Puits 48	48	48	2023-01-01	48	48	48	48	48	48	48	48	48	48
49	GRP 49	49	Basin 49	49	Bloc 49	49	Gis 49	Gid 49	49	Res 49	49	Puits 49	49	49	2023-01-01	49	49	49	49	49	49	49	49	49	49
50	GRP 50	50	Basin 50	50	Bloc 50	50	Gis 50	Gid 50	50	Res 50	50	Puits 50	50	50	2023-01-01	50	50	50	50	50	50	50	50	50	50
51	GRP 51	51	Basin 51	51	Bloc 51	51	Gis 51	Gid 51	51	Res 51	51	Puits 51	51	51	2023-01-01	51	51	51	51	51	51	51	51	51	51
52	GRP 52	52	Basin 52	52	Bloc 52	52	Gis 52	Gid 52	52	Res 52	52	Puits 52	52	52	2023-01-01	52	52	52	52	52	52	52	52	52	52
53	GRP 53	53	Basin 53	53	Bloc 53	53	Gis 53	Gid 53	53	Res 53	53	Puits 53	53	53	2023-01-01	53	53	53	53	53	53	53	53	53	53
54	GRP 54	54	Basin 54	54	Bloc 54	54	Gis 54	Gid 54	54	Res 54	54	Puits 54	54	54	2023-01-01	54	54	54	54	54	54	54	54	54	54
55	GRP 55	55	Basin 55	55	Bloc 55	55	Gis 55	Gid 55	55	Res 55	55	Puits 55	55	55	2023-01-01	55	55	55	55	55	55	55	55	55	55
56	GRP 56	56	Basin 56	56	Bloc 56	56	Gis 56	Gid 56	56	Res 56	56	Puits 56	56	56	2023-01-01	56	56	56	56	56	56	56	56	56	56
57	GRP 57	57	Basin 57	57	Bloc 57	57	Gis 57	Gid 57	57	Res 57	57	Puits 57	57	57	2023-01-01	57	57	57	57	57	57	57	57	57	57
58	GRP 58	58	Basin 58	58	Bloc 58																				

FIGURE 5.8 – Fichier final comportant les 3 groupements

- Production GTFT :

Comme déjà expliqué dans la conception, nous sommes dans l'obligation de créer une base de données SQL Server en raison de l'absence totale de relations dans la base de données Microsoft Access.

Tout d'abord il faut extraire les tables concernant la production :

- (a) **Table CATEGORY** contient la relation entre les puits et les réservoirs (Figure 5.71 en Annexe). On peut remarquer que plusieurs puits peuvent exploiter un même réservoir, ce qui implique une relation un-à-plusieurs **1-N**. Uniquement les colonnes **WELL** et **RESERVOIR** sont nécessaires pour le remplissage des nouvelles tables **Puits** et **Réservoirs** sous SQL SERVER.

(b) **Table DPRD** décrit la production quotidienne des trois fluides **Gaz Sec**, **Condensat**, **Gaz GPL** de chaque puits ainsi que d'autres informations supplémentaires telles que la composition des gaz, la température, la pression, la production d'eau, etc... (Figure 5.72 en Annexe)

Cette table sera directement intégrée dans la table **Production** de la base de données SQL SERVER avec quelques modifications à apporter.

Le schéma de la BDD pour ce groupement se présente dans la Figure 5.9 :

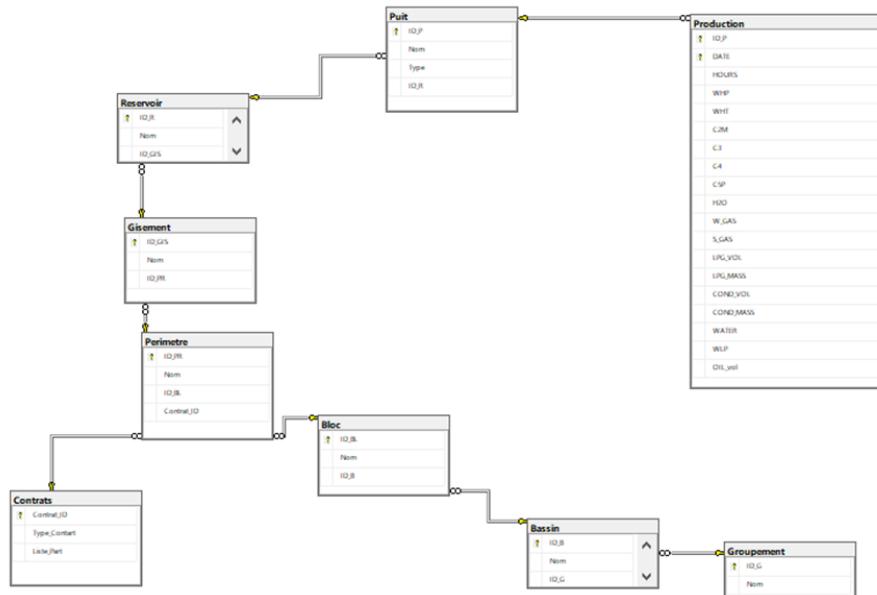


FIGURE 5.9 – Schéma de la Base de données GTFT

Pour procéder au remplissage, la base de données sous SQL Server est créée en suivant les hiérarchies comme suit :

— **Les tables Bloc, Bassin et Groupement :**

Ces tables sont générées manuellement dans un fichier Excel réparti en plusieurs feuilles spécifiques à chaque table étant donné qu'il n'y a qu'un seul bassin ,bloc et groupement.

— **Table Contrats :**

Les informations concernant les contrats et leurs types ainsi que les bassins sédimentaires sont extraites depuis le fichier Excel (Figure 5.73 en Annexe).

Les informations disponibles sur ACCESS sont extraites directement, et les autres informations sur les hiérarchies restantes sont générées de la même façon décrite précédemment.

— **Tables Réservoir, Gisement et Périmètre :**

Selon la logique expliquée pour le groupement **ELMERK** : le nom du réservoir n'est autre le nom du gisement suivi d'un code, cela est également observable dans la table **CATEGORY**.

Il suffit donc d'extraire le nom du gisement et par la même occasion le nom du périmètre à partir du nom réservoir car pour ce groupement l'entreprise nous a confirmé que le découpage des dimensions des périmètres et des gisements est le même, ce qui signifie qu'ils ont les mêmes noms.

Les clés primaires sont générées une à une.

La clé étrangère de la table **Réservoir** est associée à l'ID de son gisement à l'aide de la logique expliquée précédemment.

Exemple : **TFTW-R1** appartient au gisement **TFTW**.

La clé primaire du gisement est récupérée par **VLOOKUP** sur la table gisement. Le même processus est suivi pour la génération de la clé étrangère de la table **Gisement** associée à la clé primaire de la table **Périmètre** qui elle même est associée à la table **Bloc**.

— Table Puits :

Les noms de puits et réservoirs sont récupérés depuis la table **CATEGORY**, la clé primaire est générée, la clé étrangère est obtenue en faisant un VLOOKUP sur la table **Réservoir**, sans oublier d'ajouter une colonne pour le type de puits (les puits de GTFT sont tous de type **Gaz**).

— Table Production :

Importer directement la table **DPRD** en ajoutant la colonne **ID_puits** grâce à un VLOOKUP sur la table **Puits**, supprimer la colonne WELL et considérer (**ID_P**, **DATE**) comme clé primaire.

Le remplissage de la base de données SQL server déjà créée se fait en utilisant l'outil SSIS. Voici un exemple du remplissage de la table Puits (Figure 5.10)

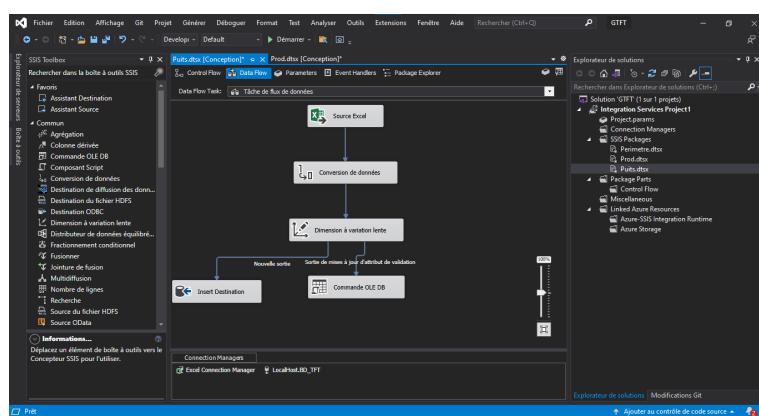


FIGURE 5.10 – Remplissage de la table Puit GTFT

- La lecture de la feuille de calculs « **Puits** » contenue dans le fichier Excel.
- La conversion du type de l'**ID_P**, **Nom**, **Type**, **ID_R** vers chaîne de caractères est obligatoire pour avoir le même type dans la BD SQL SERVER.
- La gestion des dimensions à variations lentes pour assurer la mise à jour et les nouveaux ajouts dans la base de données.

Remarque : pour la table Production une conversion de la date est aussi nécessaire.

2. Création des packages

Après conversion et nettoyage des données sur Python, nous avons procédé au remplissage du DW. Nous avons élaboré pour chaque remplissage des tables, un package différent sur SSIS (Figure 5.11)

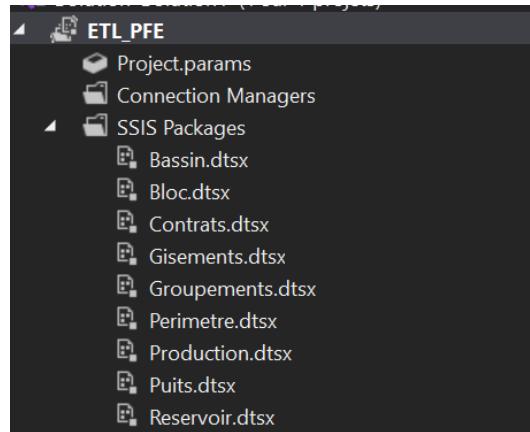


FIGURE 5.11 – Crédit à la création des Package

3. Chargement et mises à jour de toutes les dimensions

A l’exception de la dimension Temps générée et remplie par un script SQL (comme expliqué dans le chapitre précédent), le remplissage de toutes les dimensions s’effectue de la même façon en commençant de la dimension **Groupements** jusqu’à la dimension **Puits**. Les étapes sont décrites ci-dessous :

3.1 Extractions des données :

- BDD SQL Server du groupement TFT à l’aide du composant «Source OLE DB » où on sélectionne à chaque fois le nom de la table souhaitée.
- Fichier Excel des autres groupements GTIM, GSS, Berkine à l’aide du composant « Source Excel ».

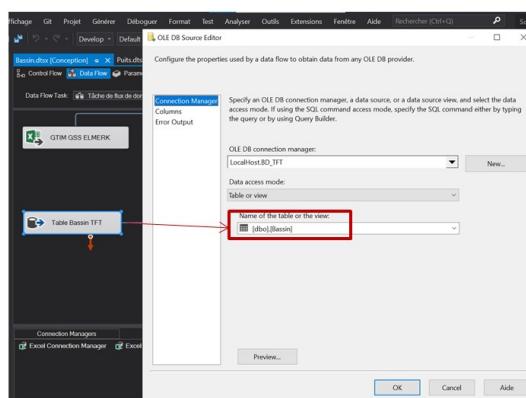


FIGURE 5.12 – Remplissage de la dimension Bassin

3.2 Conversion de données :

Uniquement pour les sources Excel afin de modifier les types des attributs (Figure 5.13).

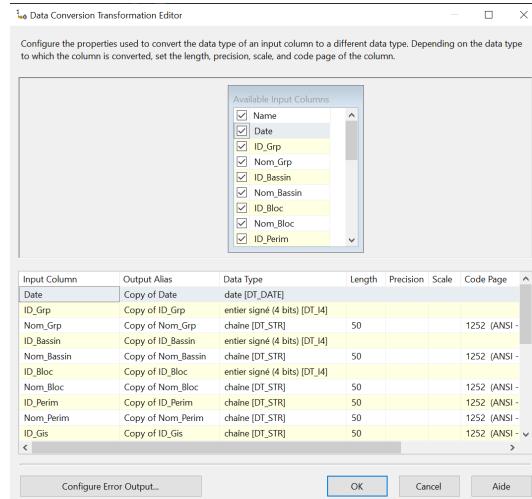


FIGURE 5.13 – Conversions du type dans le fichier Excel global

3.3 Tri selon l'ID de la dimension à remplir :

Uniquement pour les sources Excel pour extraire l'ID qu'une seule fois, et supprimer les doublons vu que le remplissage de la dimension nécessite uniquement un ID et le nom lui correspondant(Figure 5.14).

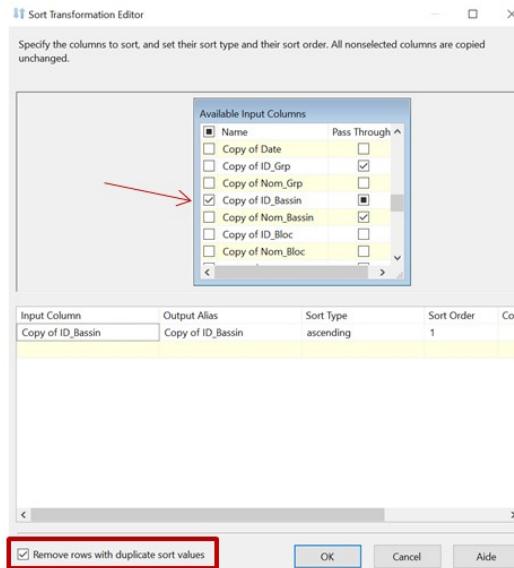


FIGURE 5.14 – Selection des ID Bassin Uniques

3.4 Union : Des deux sources de données hétérogènes.

3.5 Mise en œuvre de mécanisme de gestion des « dimensions à variations lentes »

qui détecte automatiquement lors de l'extraction, les nouveaux ajouts du fichier source et les lignes modifiées :

- Ajout des nouveaux enregistrements dans la « destination OLE DB » qui représente la table de dimension.
- Modification des enregistrements déjà présents dans la dimension et ayant subi une transformation dans leur fichier source.

Par exemple, pour la dimension **Contrats**, le type de contrats peut être modifié, ou encore l'ajout d'un nouveau partenaire dans un contrat déjà existant peut également être effectué.

Remplissage Dimension Contrats (Figure 5.15)

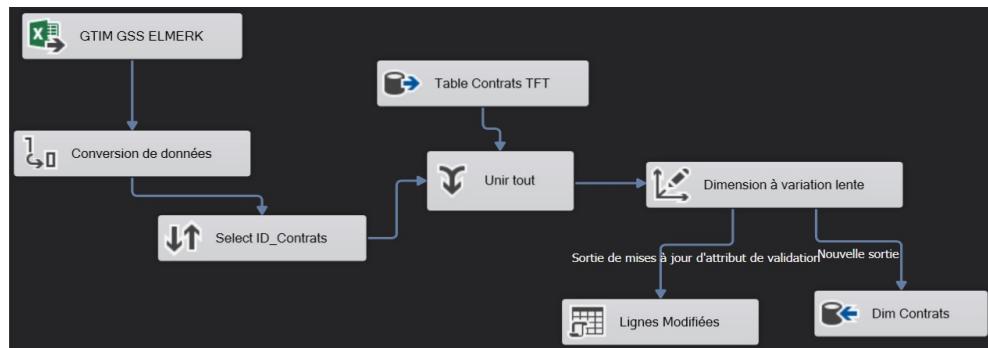


FIGURE 5.15 – Remplissage Dimension Contrats

4. Alimentation de la table Fait

Le remplissage de la table de fait se fait à partir des données des tables de dimensions et des données numériques extraites à partir des données sources :

4.1 Union :

Après avoir effectué les conversions nécessaires (types d'attributs), on procède à l'union des données de la table Production de la base de données SQL Server (Groupement TFT) et du fichier Excel (GTIM, GSS, Elmerk) afin de récupérer les données de production de chaque puits et les dates correspondantes.

4.2 Recherche des correspondances :

On effectue par la suite une recherche avec correspondance (Figure 5.16) avec la dimension Puits, afin de joindre les « **ID_Puits** » entre eux.

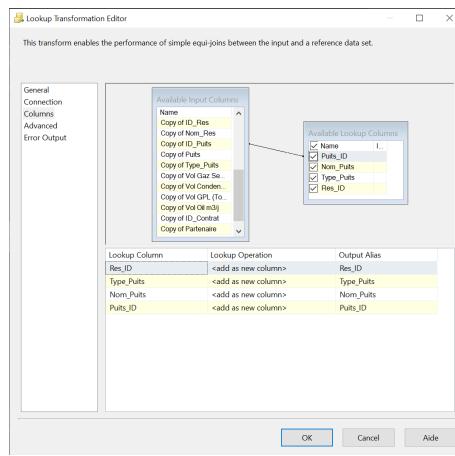


FIGURE 5.16 – Recherche par correspondance

A partir de la clé étrangère de la dimension Puits, on sélectionne « ID_Reservoir » et on effectue la jointure (Figure 5.17) avec la dimension Réservoirs.

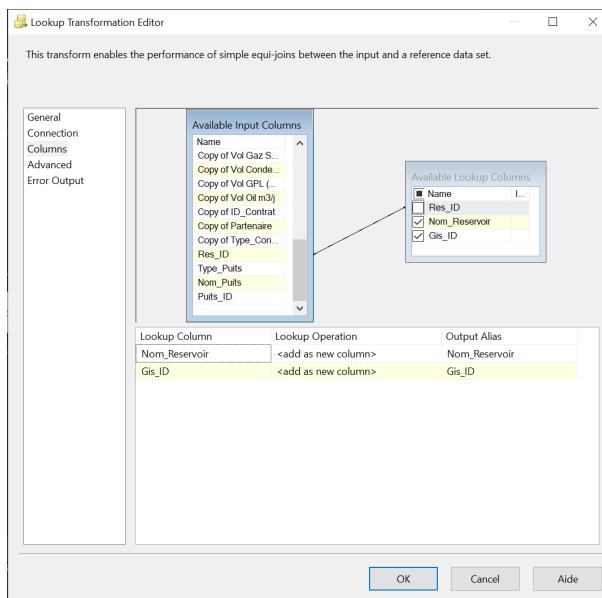


FIGURE 5.17 – Jointure avec la dimension Réservoirs

On effectue la même recherche pour toutes les dimensions en suivant les niveaux de hiérarchies. La jointure de la dimension Temps s'effectue entre la date récupérée après l'union des deux sources et l'ID de la dimension elle-même.

4.3 Jointure de fusion :

Afin de récupérer les données de prévisions annuelles du groupement GSS, on effectue une jointure de fusion (Figure 5.18) entre le fichier Excel « **Production GSS** » et les données résultant du processus antérieur. Cette opération nécessite un tri selon « **ID_Groupements** » et « **Year** ».

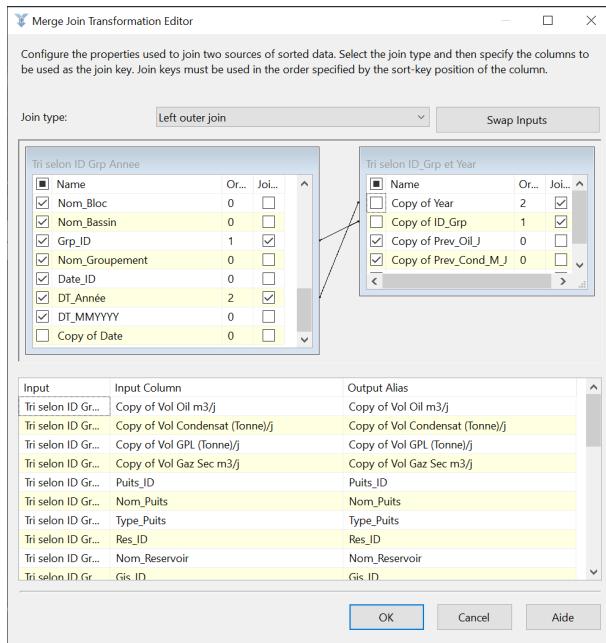


FIGURE 5.18 – Jointure de fusion GSS

De même pour les prévisions mensuelles des groupements GTFT et GTIM qui nécessitent des tris selon « **ID_Groupements** » et « **MMYYYY** » (Figure 5.19), qui comme expliqué précédemment, représente la clé du mois de l'année.

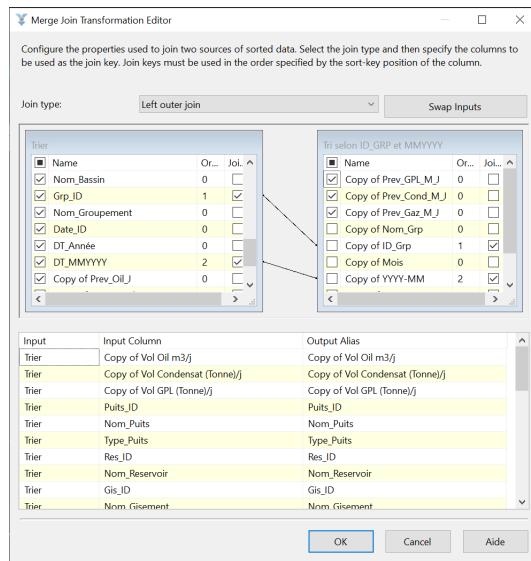


FIGURE 5.19 – Jointure de fusion GTFT et GTIM

Ci-dessous (Figure 5.20), les étapes de l'exécution du remplissage de la table de Fait :

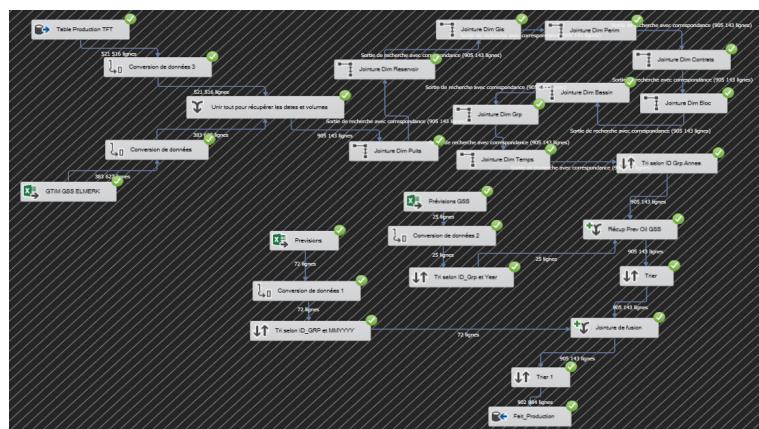


FIGURE 5.20 – Remplissage de la table Fait

5.2.2 Mise en œuvre du tableau de bord

Comme cité précédemment, Sonatrach possède plusieurs types de groupements :

- Groupements qui produisent uniquement du Gaz Sec
- Groupements qui produisent 3 fluides différents (GPL, Condensat, Gaz Sec)
- Groupements qui produisent de l'huile (et éventuellement du Gaz Sec)

Afin de mieux visualiser les données, nous avons mis à leur disposition un tableau de bord pour chaque type de groupements. Chaque feuille du tableau de bord prend en considération uniquement les groupements du type choisi.

Si un ingénieur souhaite voir le suivi de production d'un groupement de type Gaz, il se rend à la page correspondante. Nous n'avons malheureusement pas eu accès à toutes les données des groupements mais une fois ajoutées, il suffira juste d'appliquer le filtre en haut à gauche pour choisir le groupement en question.

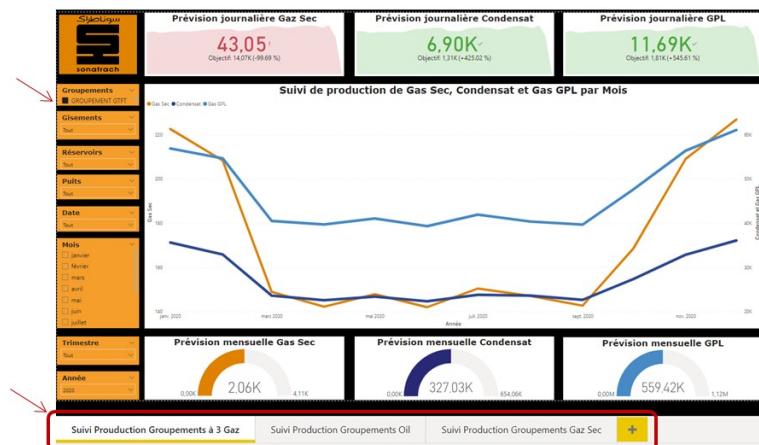


FIGURE 5.21 – Visualisation globale du groupement GTFT en 2020

D'autres filtres sont également mis à sa disposition, s'il souhaite analyser avec plus de précision la production par puits, gisements, réservoirs...

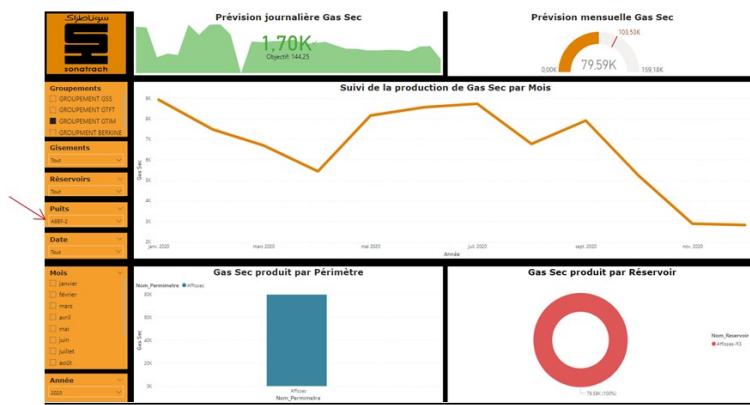


FIGURE 5.22 – Visualisation du suivi de production du puits « ABBF-2 » en 2020

Lorsque l'ingénieur souhaite voir l'état d'avancement de la production pour s'assurer que le taux de production de la veille a atteint le taux journalier et mensuel prévus, il filtre à la date J-1.



FIGURE 5.23 – 7 janvier 2020 : Prévision journalière du Groupement GTIM atteinte, cumul mensuel prévu non atteint

Autres détails En Annexe

5.3 Volet Prédition

5.3.1 Prévision avec Time Series Analysis :

Comme déjà expliquer, nous avons effectué l'analyse des séries temporelles sur la production mensuelle (Mars 1999 à Avril 2022) de gaz sec du groupement GTFT.

1. Compréhension des données (Analyse brute de la série temporelle)

Apres avoir récupéré les données depuis la BD SQL SERVER (en effectuant un group by Date monthly pour obtenir la production mensuelle de gaz Sec). Notre DataSet (Table 5.1) comporte 277 lignes de production gaz Sec ($\times 10^3 m^3$) mensuelle.

Date Mensuelle	Production ($10^3 m^3$)
2000-01	522.2
2000-02	518.71
2000-03	573.67
2000-04	549.03
2000-05	566.85
2000-06	543.51
2000-07	559.53
2000-08	551.91
2000-09	478.81
2000-10	432.98
2000-11	538.58
2000-12	566.46

TABLE 5.1 – Production mensuelle de gaz de l'année 2000

a. Analyse graphique de la série :

Nous procérons au tracage du graphe (Figure 5.24) de la série et effectuer une première analyse visuelle.

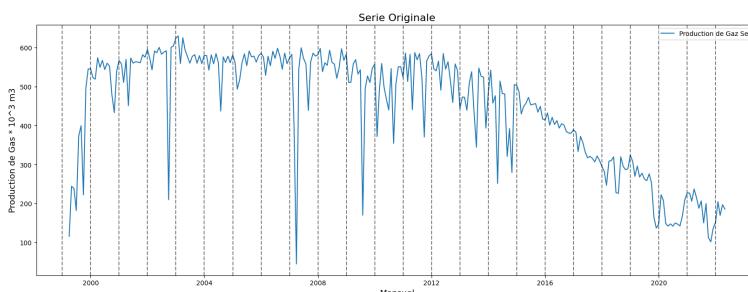


FIGURE 5.24 – Graphe de la Production de Gaz mensuelle de 2000 à 2022

- D'après ce graphe (Figure 5.24) on remarque que notre série temporelle comporte une tendance qui diminue manière exponentielle au fil du temps (non linéaire) : c'est une tendance multiplicative.
- Un effet de saisonnalité peut être visible aussi et on remarque que les variations saisonnières diminuent en fonction de la tendance donc c'est le cas d'une saisonnalité multiplicative.

- Nous pouvons dire aussi que notre série suit un schéma multiplicatif car les variations saisonnières diminuent en fonction de tendance (proportionnelles).

b. Décomposition de la série (Seasonal-trend decomposition) :

Nous effectuons une Seasonal-trend décomposition (Figure 5.25) de notre série pour diviser ses composantes et mieux les étudier.

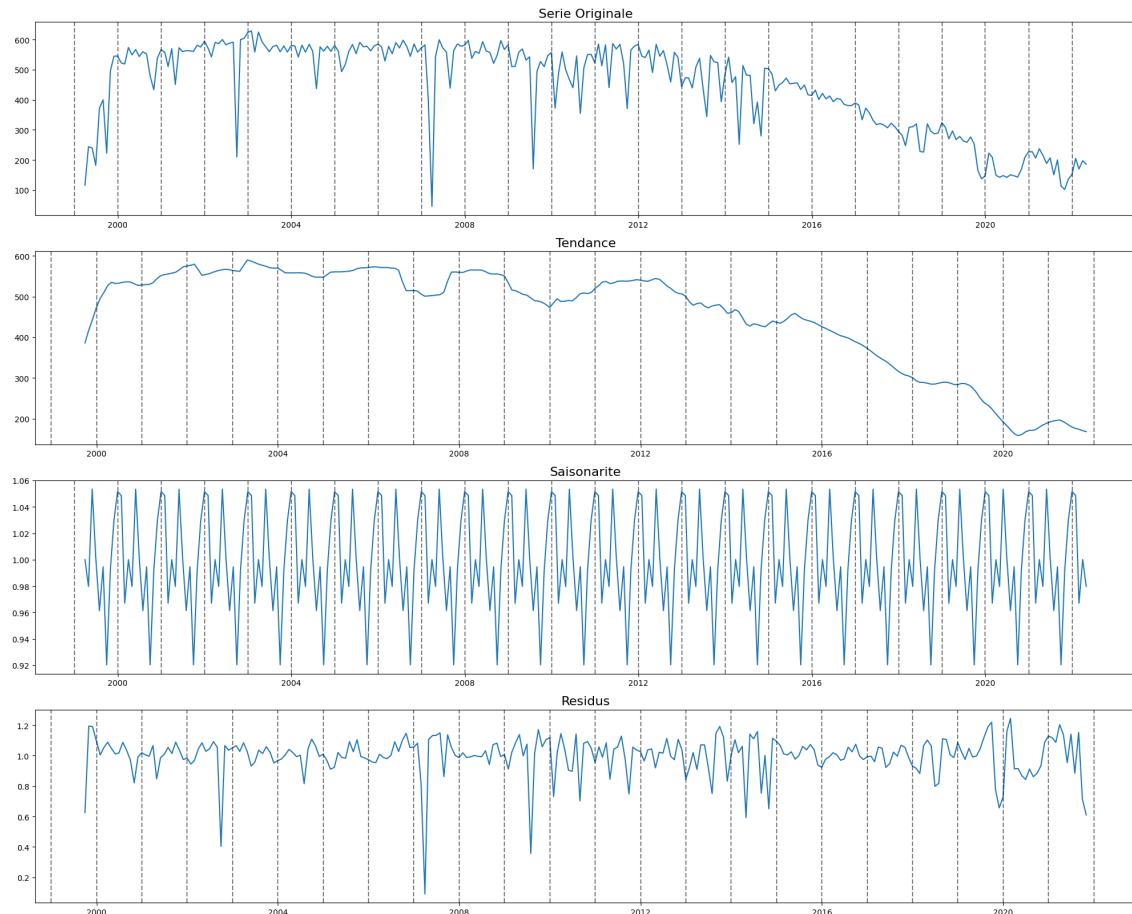


FIGURE 5.25 – Seasonal-trend décomposition

- D'après la Figure 5.25 On confirme la présence de l'effet de tendance multiplicative ainsi la saisonnalité, cette dernière a été bien capturée vu la précision que notre série suit un schéma multiplicatif ((Trend * Seasonality) + residuals).
- Cette décomposition nous a aussi fourni les résidus (erreurs) de la série.

Nous pouvons visualiser (Figure 5.26) la série sans cet effet en calculant Estimation (Trend * Seasonality).

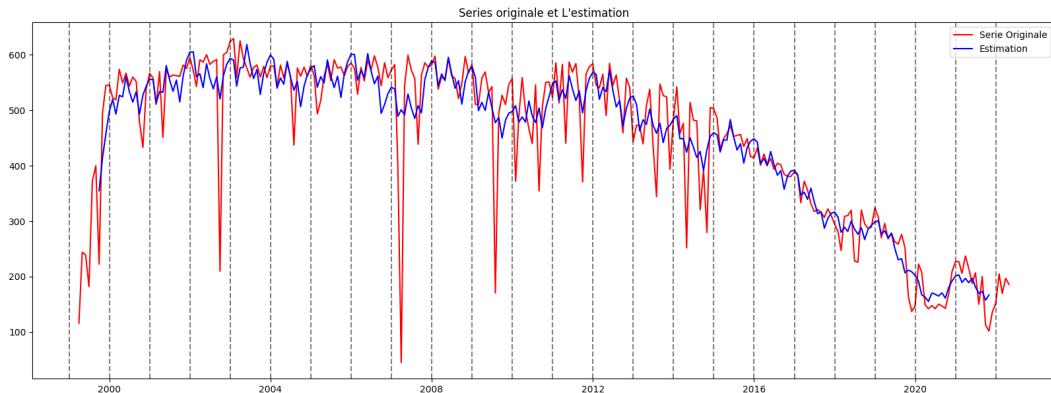


FIGURE 5.26 – Comparaison entre la série originale et son estimation (Trend * Saisonnalité)

c. Détection des anomalies (Outliers de Production) :

Ces résidus vont nous aider à détecter les potentielles anomalies de production présentes dans notre série (Valeurs aberrantes).

Premièrement, nous traçons (Figure 5.27) la Box Plot des résidus :

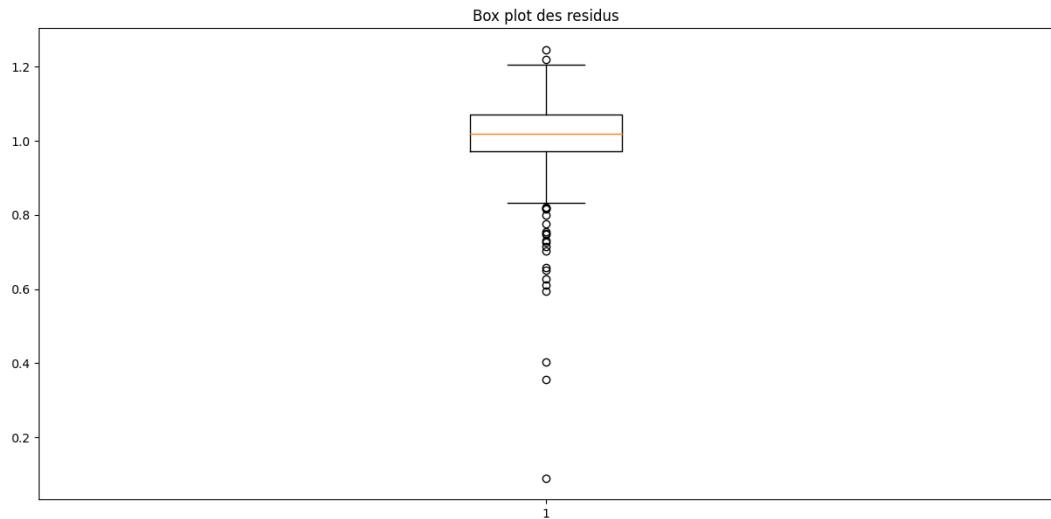


FIGURE 5.27 – Box plot des résidus

Nous pouvons remarquer qu'il y ait des anomalies, nous avons utilisé la méthode de détection des valeurs aberrantes par intervalle interquartile sur résidus.

Cette méthode nous a permis d'identifier les observations qui se situent en dehors de l'intervalle interquartile (Figure 5.28), suggérant ainsi des valeurs potentiellement aberrantes.

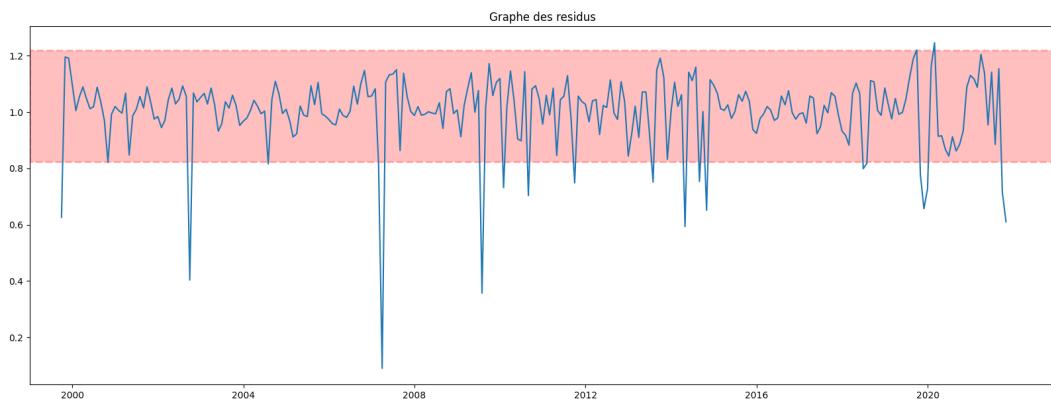


FIGURE 5.28 – Détection par intervalle inter-quartile

Nous récupérons les dates des anomalies et on les affiche sur le graphique de la série originale.

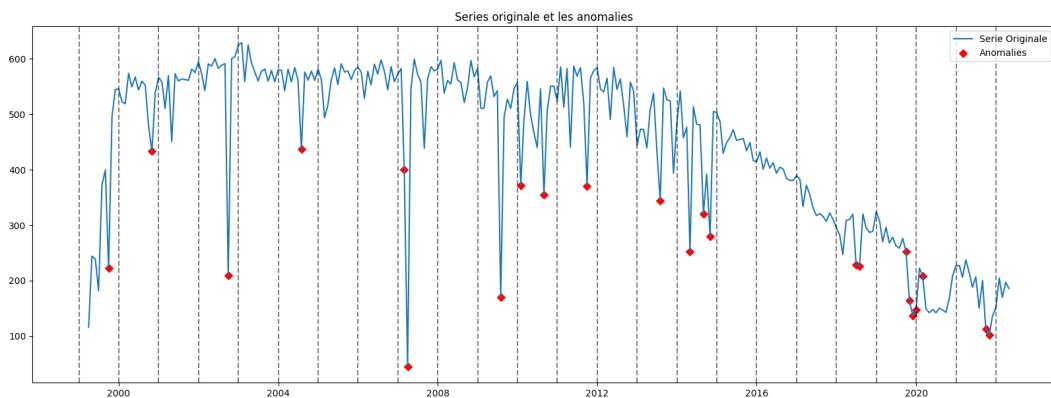


FIGURE 5.29 – présentation des anomalies

d. Test de stationnarité :

Le test de stationnarité est une étape importante, permettant de vérifier si la série chronologique étudiée possède des propriétés statistiques constantes au fil du temps. Cela garantit l'adéquation des modèles et leur capacité à fournir des prévisions précises.

- **ACF and PACF :**

Nous traçons tout d'abord les graphes d'autocorrélation simple et partiel (de $277/4 = 46$ lags) de la série.

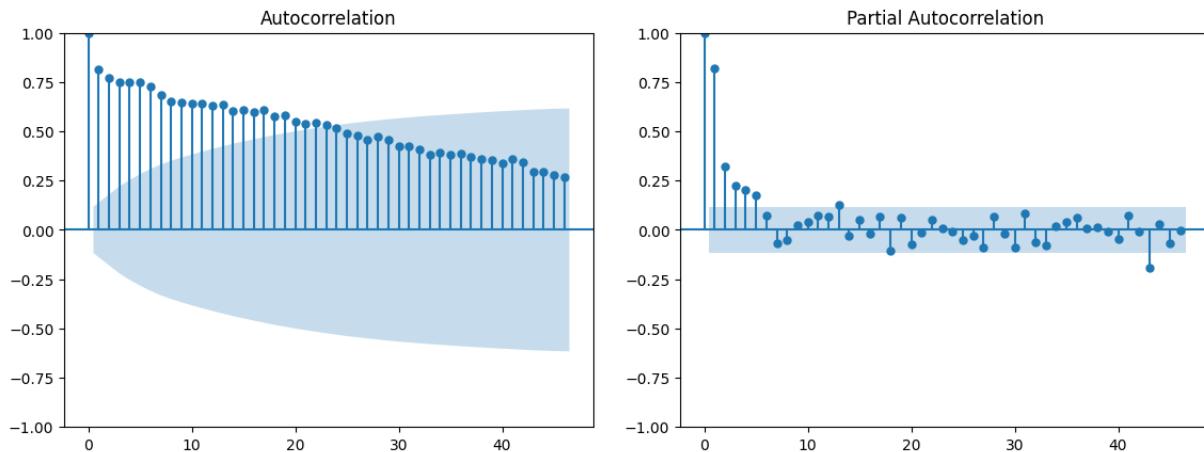


FIGURE 5.30 – Graphe ACF et PACF

L'analyse du corrélogramme simple et partiel (Figure 5.30), nous indique une non stationnarité de la série. En effet, la fonction d'autocorrélation simple ne décroît pas de manière rapide vers zéro et le premier terme du corrélogramme partiel est très important.

- Calcul de moyenne mobile et écart type mobile et application du Dickey Fuller Test :**

Nous avons conçu une fonction qui calcule la moyenne mobile ainsi que l'écart type mobile avec une fenêtre de 12 mois en effectuant aussi le Dickey Fuller Test.

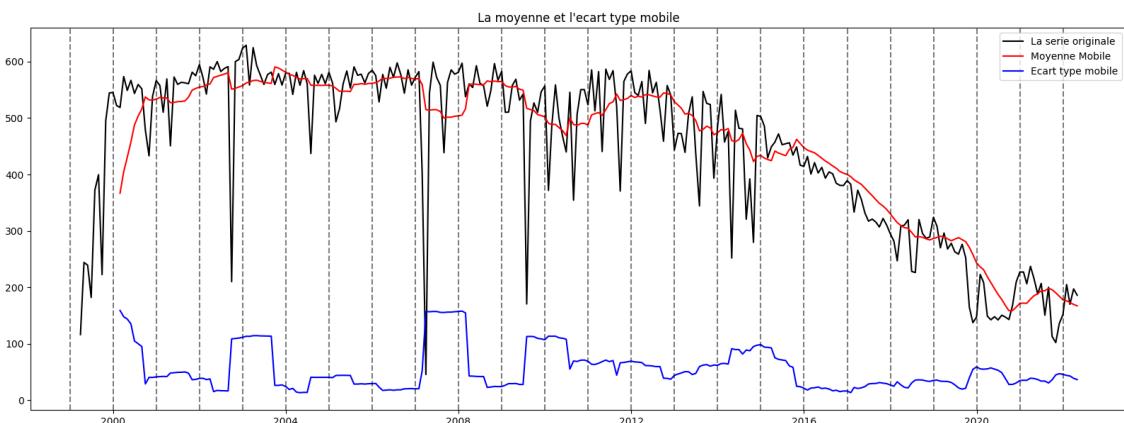


FIGURE 5.31 – Moyenne mobile et écart type mobile

Nous remarquons (Figure 5.31) que la moyenne mobile est pas constante dans le temps de même pour l'écart type (volatilité pas stable) donc on confirme que la série n'est pas stationnaire.

Résultats du test de Dickey-Fuller	
Valeur-p	0.784273
Nombre de lags utilisés	5
Nombre d'observations utilisées	272

TABLE 5.2 – Résultats du test de Dickey-Fuller de la série

La valeur p associée au test de Dickey-Fuller (Table 5.2) est très élevée (0.784273), ce qui suggère qu'il n'y a pas suffisamment de preuves pour rejeter l'hypothèse de non-stationnarité. Cela suggère que la série présente des caractéristiques de non-stationnarité.

2. Préparation des données (Transformation de la série temporelle)

(a) Traitement des anomalies :

Les anomalies déjà détectées précédemment (Voir Figure 5.29) en été remplacées par la moyenne mobile avec une fenêtre de 12 mois, on couvre une année complète, ce qui permet de garder les variations saisonnières.

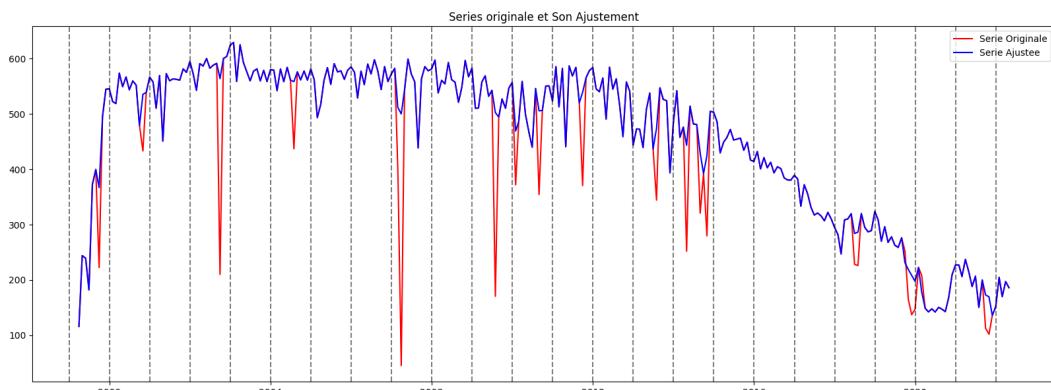


FIGURE 5.32 – Série ajustée

(b) Stationarisation :

Ce traitement permet de déterminer le nombre de différenciations nécessaires pour rendre une série temporelle stationnaire. Nous avons aussi éliminé les changements de volatilité pour obtenir des prédictions plus fiables.

- **La première différenciation pour enlever la tendance :** la première différenciation de la série temporelle consiste à prendre la différence entre les observations successives. Cela permet de réduire la tendance présente dans la série en éliminant les variations entre les valeurs.

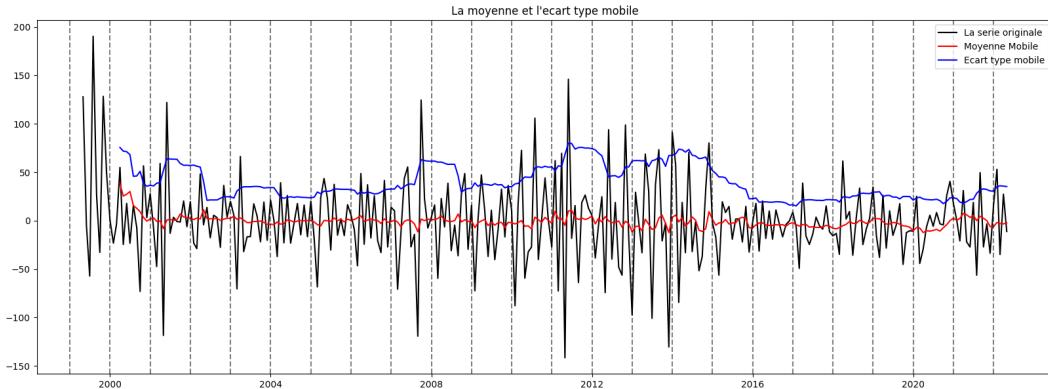


FIGURE 5.33 – la série après diff (1)

Déjà Nous remarquons (Figure 5.33) que la série est plus stable et stationnaire dans le temps. Mais on remarque un effet de volatilité qui doit être gérée.

- **Enlever le changement dans la volatilité :** Pour enlever le changement dans la volatilité d'une série temporelle, on peut diviser la production de chaque mois par l'écart-type de la production de l'année correspondante pour ce mois. Cette méthode permet de normaliser la volatilité en prenant en compte les variations saisonnières.

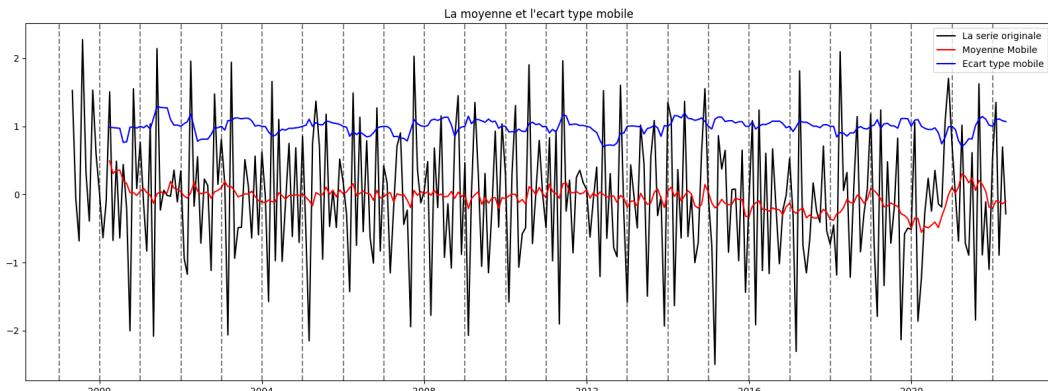


FIGURE 5.34 – La série temporelle stable

Nous pouvons effectuer un dernier Dickey Fuller Test (Figure 5.3) pour s'assurer de la stationnarité de la série.

Résultats du test de Dickey-Fuller

Valeur-p	0.000025
Nombre de lags utilisés	12
Nombre d'observations utilisées	264

TABLE 5.3 – Dickey Fuller Test de la série stationnaire

Dans ce cas (Figure 5.3), la valeur-p associée au test est de 0.000025, ce qui est extrêmement faible. Cela signifie que nous pouvons rejeter l'hypothèse nulle de non-stationnarité et conclure que la série temporelle est stationnaire.

(c) **Séparation :**

Dans notre analyse de la série temporelle, nous avons divisé nos données en un ensemble d'entraînement et un ensemble de test. Cette division nous permet d'évaluer les performances de nos modèles sur des données futures non vues.

L'ensemble d'entraînement est utilisé pour ajuster les modèles et est généralement constitué des observations les plus anciennes. Dans notre cas, nous avons utilisé une division avec un horizon de prévision de 12 mois, ce qui signifie que nous avons conservé les 12 dernières périodes de la série comme ensemble de test.

3. Modélisation :

(a) **Choix du modèle :**

- Nous avons exclu le modèle ARMA en raison de la non-stationnarité des données, qui nécessitent une transformation pour devenir stationnaires. L'ARMA ne prend pas en compte directement l'intégration des données, rendant son application inappropriée dans notre cas. En revanche, nous avons opté pour le modèle SARIMA plutôt que l'ARIMA en raison de la présence évidente d'une composante saisonnière dans les données temporelles.
- Le modèle de lissage exponentiel basique n'est pas applicable en raison de la tendance significative présente dans les données. De même, le modèle de lissage exponentiel double n'est pas approprié en raison de la présence de variations saisonnières. Par conséquent, nous avons choisi d'utiliser le modèle de lissage exponentiel triple. Ce modèle nous permet de prendre en compte les tendances ainsi que les variations saisonnières, offrant ainsi une représentation plus complète et précise de notre série temporelle.

(b) **Identification et Estimation du modèle SARIMA**

Au premier plan, Nous allons analyser les graphes ACF et PACF (Figure 5.35) de 36 lags de la série stationnaire pour identifier les paramètres du modèle.

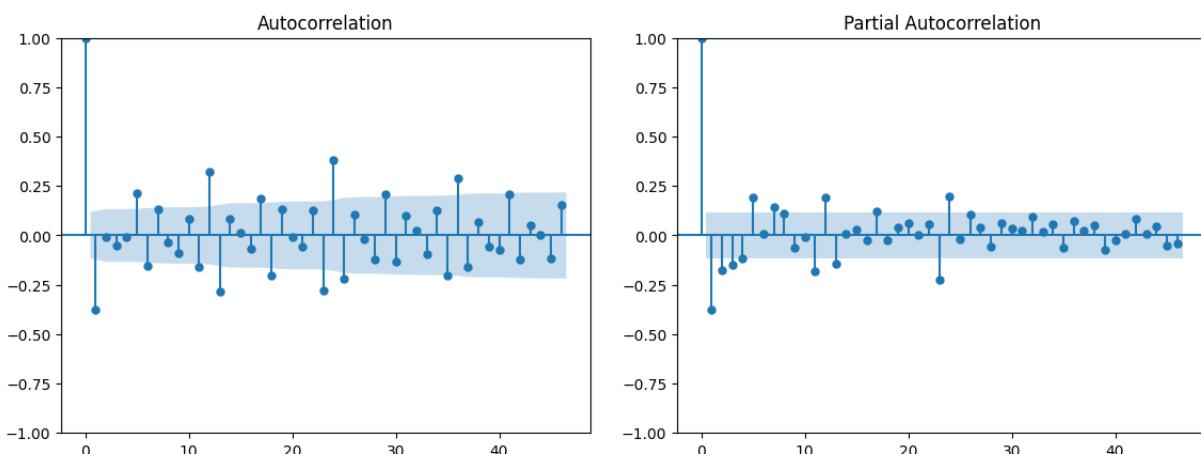


FIGURE 5.35 – ACF et PACF de la série stationnaire

En analysant les fonctions d'autocorrélation (ACF) et d'autocorrélation partielle (PACF) (Figure 5.35) , nous avons observé que l'ordre du modèle est de 1 pour la composante MA (Moyenne Mobile) et de 3 pour la composante AR (Auto Régressive). De plus, nous avons identifié un effet saisonnier avec 3 retards pour la composante MA et 2 retards pour la composante AR.

Pour déterminer le meilleur modèle, nous avons testé toutes les combinaisons possibles en variant les ordres de l'AR, de l'AR saisonnier, du MA et du MA saisonnier, afin de trouver celui qui présente le plus faible critère d'information d'Akaike (AIC) (Figure 5.36).

```
Meilleur Modele :  
SARIMA (3, 0, 0, 1, 0, 3, 12) , AIC : 672.5112378312507
```

FIGURE 5.36 – Modèle SARIMA Estimé

(c) Identification du modèle Holt-winters Exponential Smoothing

D'après la décomposition saisonnière de notre série temporelle, nous avons identifié que la tendance et la saisonnalité sont de nature "multiplicative". Cela signifie que ces composantes ont un effet proportionnel et multiplicatif sur nos données observées. Dans le cas d'une tendance multiplicative, chaque période subit une variation proportionnelle par rapport à la période précédente.

Pour la saisonnalité multiplicative, les variations relatives diminuent de manière proportionnelle à la moyenne de notre série. En prenant en compte cette nature multiplicative, nous spécifions cette caractéristique dans notre modèle, afin de capturer au mieux les variations de tendance et de saisonnalité qui ont un impact multiplicatif sur nos données.

La fonction du modèle sous Sktime utilise une heuristique pour l'estimation des facteurs de lissage initiaux α , β , γ et δ puis les estimer dans le cadre du processus d'ajustement .

4. Evaluation :

(a) Evaluation des paramètres du modèle SARIMA :

Après avoir appliquer le modèle SARIMA estimé (sur la série originale stationnaire Figure 5.34), nous avons évalué ces paramètres ainsi ces résidus, en étudiant le résumé fourni après l'estimation contenant les tests statistiques.

Résultats SARIMA		
Dep. Variable :	S_GAS	
No. Observations :	277	
Model :	SARIMAX(3, 0, 0)(1, 0, [1, 2, 3], 12)	
Date :	Fri, 16 Jun 2023	
AIC	672.511	
Time :	14 :24 :31	
Sample :	04-30-1999 04-30-2022	
Paramètres du modèle		
Variable	Coeff.	P> z
ar.L1	-0.3333	0.000
ar.L2	-0.2807	0.000
ar.L3	-0.1376	0.017
ar.S.L12	0.9788	0.000
ma.S.L12	-0.9154	0.000
ma.S.L24	0.1737	0.047
ma.S.L36	-0.1253	0.097
sigma2	0.6047	0.000
Tests de diagnostic		
Test	Valeurs S	Valeur p
Ljung-Box (L1) (Q)	0.10	0.75
Jarque-Bera (JB)	2.59	0.27
Hétéroscédasticité (H)	1.37	0.13

TABLE 5.4 – Résultats SARIMA

- Métriques d'évaluation du modèle :

Critère d'information d'Akaike (AIC) : La valeur de l'AIC est de 672.511. L'AIC est une mesure de la qualité relative du modèle par rapport à d'autres modèles potentiels.

- Coefficients du modèle :

- Les coefficients (coef) estimés pour chaque terme du modèle SARIMA sont présentés dans le tableau. Les valeurs des coefficients indiquent la force et la direction de l'effet de chaque terme sur la série temporelle.
- Les colonnes "P>|z|" indiquent les valeurs p associées à chaque coefficient. Des valeurs p faibles (généralement inférieures à 0,05) indiquent une signification statistique.

- Tests de diagnostic sur les résidus :

- Test Ljung-Box (Q) : Le test Ljung-Box évalue l'autocorrélation des résidus. Dans ce cas, la valeur p est de 0.75, ce qui suggère que les résidus ne présentent pas d'autocorrélation significative.

- Test de Jarque-Bera (JB) : Le test de Jarque-Bera évalue la normalité des résidus. Ici, la valeur p est de 0.27, indiquant que les résidus semblent suivre approximativement une distribution normale.
- Hétéroscédasticité (H) : La colonne "Heteroskedasticity (H)" affiche une valeur de 1.37. Une valeur proche de 1 indique une homoscédasticité, ce qui signifie que la variance des résidus est relativement constante.

Nous allons aussi tracer les résidus ainsi que leurs graphes ACF et PACF (Figure 5.37) (de 66 lags) et leurs distribution : On peut remarquer l'absence de corrélation simple

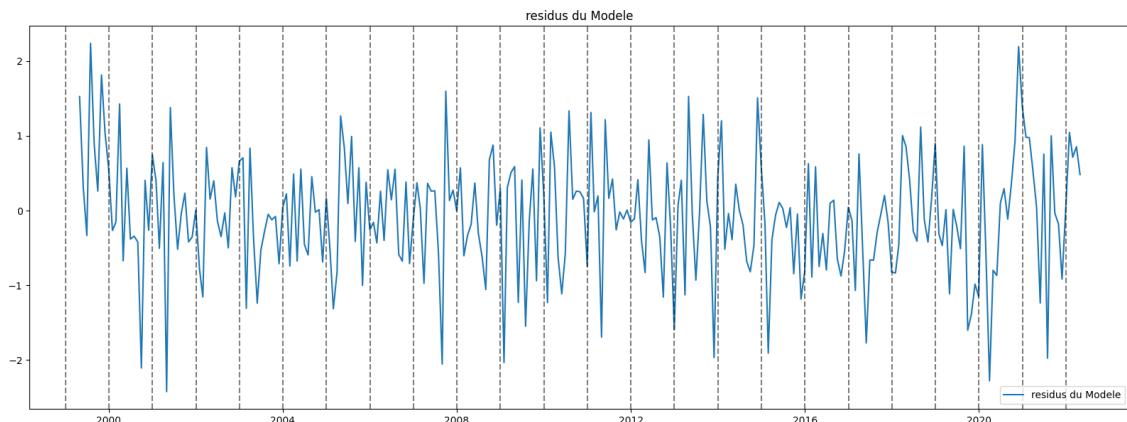


FIGURE 5.37 – Graphe des résidus de l'estimation

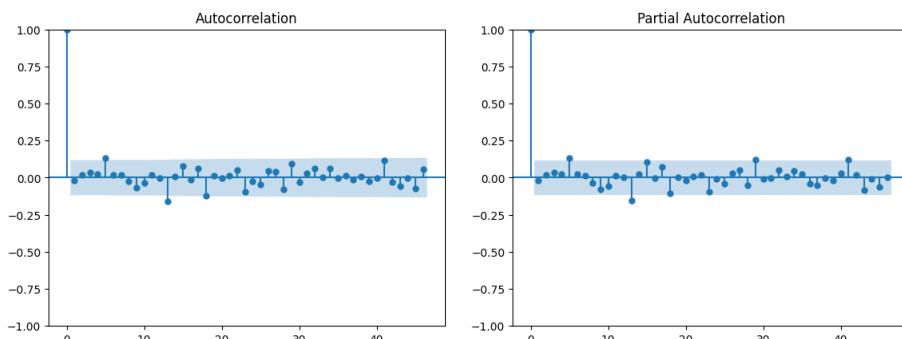


FIGURE 5.38 – Graphe ACF and PACF des résidus

et partiel (Figure 5.38) entre les lags des résidus.

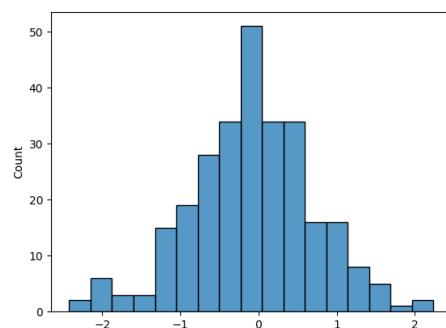


FIGURE 5.39 – Distribution des résidus

Une distribution normale (Figure 5.39) peut être remarquée. Ces résultats suggèrent que ce modèle SARIMA présente un ajustement raisonnable aux données, avec des coefficients significatifs et des résidus satisfaisants en termes d'autocorrélation et de normalité.

(b) Evaluation des modèles

- Evaluation de l'ajustement des modèles sur la série :**

Nous avons effectué l'entraînement des modèles SARIMA et Holt-Winters Exponential Smoothing sur l'ensemble d'entraînement , puis nous avons calculé le coefficient de détermination (R^2) pour évaluer leur ajustement.

— **Modèle SARIMA :** En utilisant le modèle estimé précédemment (Figure 5.36) SARIMA (3,0,0) (1,0,3,12) avec $d = 1$ car on a effectué la première différenciation pour rendre la série stationnaire. Donc on estime la série avec le modèle SARIMA (3,1,0) (1,0,3,12) et on obtient le résultat suivant :

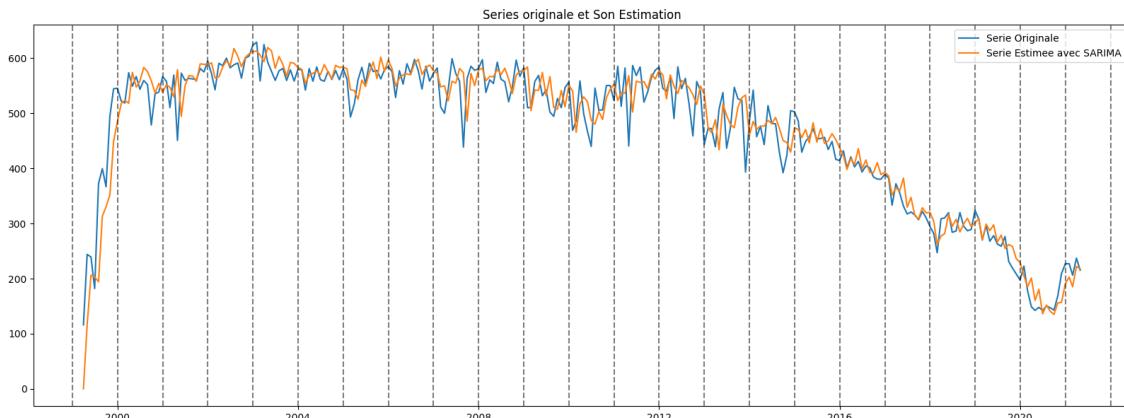


FIGURE 5.40 – Estimation avec SARIMA

Modèle	R^2
SARIMA	91.078637

TABLE 5.5 – Coefficient de détermination SARIMA

Le modèle SARIMA ajusté a obtenu un coefficient de détermination (R^2) de 91.078637. Ce résultat indique que le modèle SARIMA a réussi à expliquer environ 91.1% de la variabilité des données observées (Table 5.57). Cela suggère un ajustement relativement bon du modèle aux données de la série temporelle.

— Modèle Holt-Winters Exponential Smoothing :

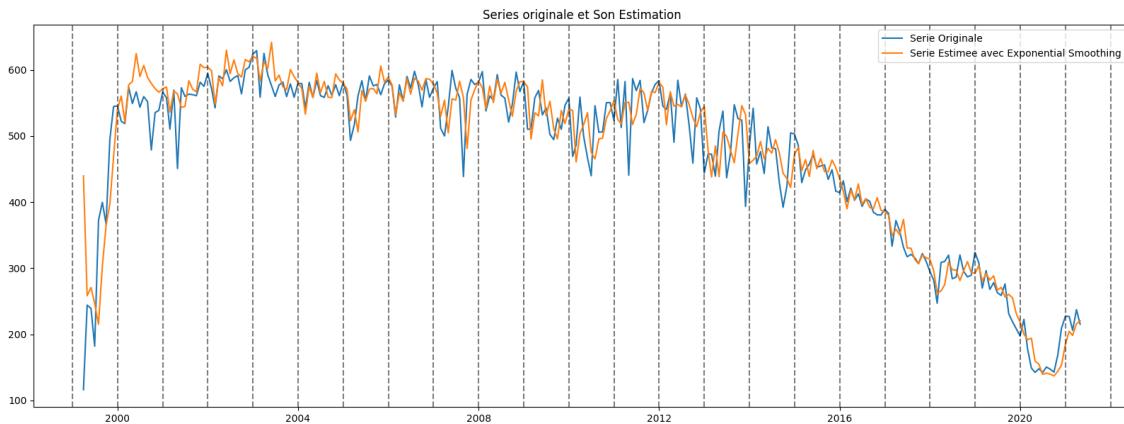


FIGURE 5.41 – Estimation avec Holt-Winters Exponential Smoothing

Modèle	R ²
Holt-winters Exponential Smoothing	89.909746

TABLE 5.6 – Coefficient de détermination Holt-Winters Exponential Smoothing

Le modèle Holt-Winters Exponential Smoothing ajusté a obtenu un coefficient de détermination (R^2) de 89.909746. Ce résultat indique que le modèle Holt-Winters a réussi à expliquer (Figure 5.41) environ 90% de la variabilité des données observées .

- **Test Evaluation**

Après avoir entraîné les modèles SARIMA et Holt-Winters Exponential Smoothing sur les données d'apprentissage, nous les avons évalués en utilisant les données de test sur une durée de 12 mois. Afin d'évaluer leur performance, nous avons utilisé les mesures suivantes : MAPE (Mean Absolute Percentage Error), MAE (Mean Absolute Error) et RMSE (Root Mean Square Error).

— Modèle SARIMA :

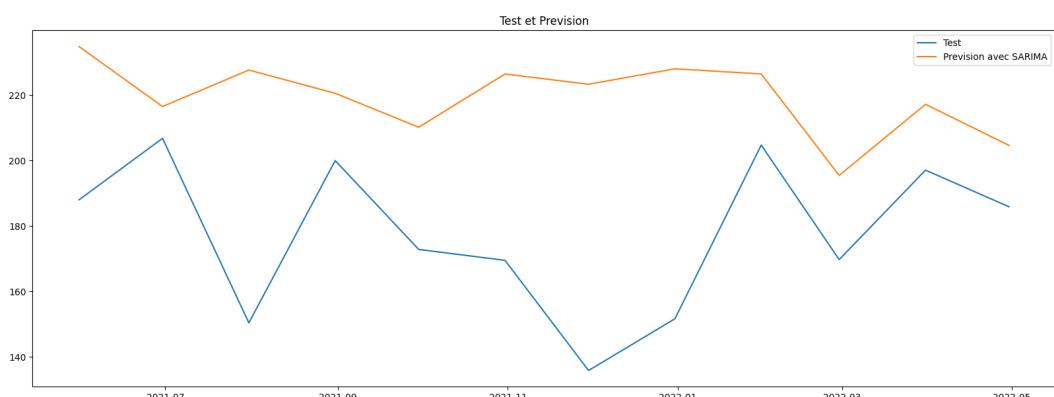


FIGURE 5.42 – Prévision du test avec SARIMA

Modèle	MAPE	MAE	RMSE
SARIMA	0.25597	41.565882	48.891918

TABLE 5.7 – Calcul des mesures de SARIMA

Erreur absolue moyenne (MAE) : La valeur de MAE est de 41.565882. Cela signifie que, en moyenne, les prédictions du modèle ont une différence absolue de 41.565882 par rapport aux valeurs réelles de l'ensemble de test.

Erreur quadratique moyenne (RMSE) : La valeur de RMSE est de 48.891918. Le RMSE mesure l'écart moyen entre les prédictions du modèle et les valeurs réelles, en tenant compte des différences quadratiques. Dans ce cas, le RMSE indique que, en moyenne, les prédictions du modèle ont une erreur d'environ 48.891918. **Erreur relative moyenne (MAPE)** : La valeur de MAPE est de 0.25597, exprimée en pourcentage. Le MAPE est une mesure de l'erreur relative moyenne du modèle par rapport aux valeurs réelles. Une valeur de 0.25597 indique que, en moyenne, les prédictions du modèle ont une erreur relative de 25.597%.

— Modèle Holt-Winters Exponential Smoothing

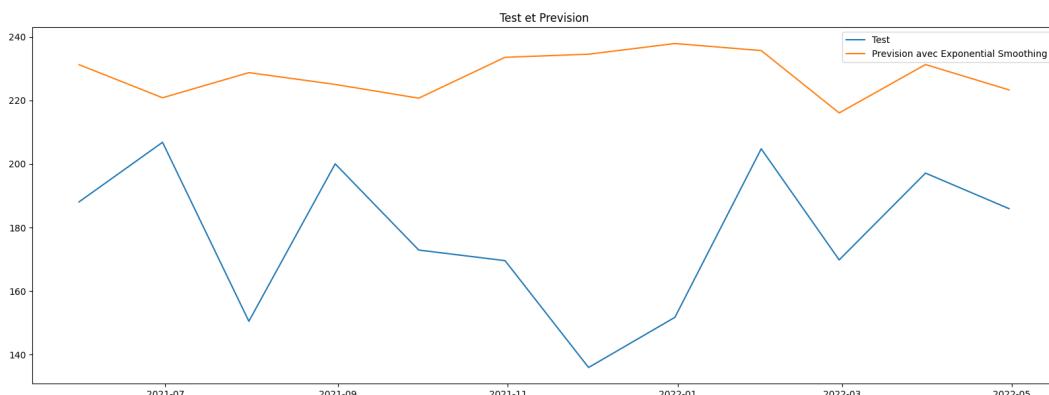


FIGURE 5.43 – Prévision du test avec Holt-Winters Exponential Smoothing

Modèle	MAPE	MAE	RMSE
Holt-Winters	0.307497	50.50635	56.303785

TABLE 5.8 – Calcul des mesures de Holt-Winters Exponential Smoothing

Erreur absolue moyenne (MAE) : La valeur de MAE est de 50.50635. Cela signifie que, en moyenne, les prédictions du modèle ont une différence absolue de 50.50635 par rapport aux valeurs réelles de la série temporelle.

Erreur quadratique moyenne (RMSE) : La valeur de RMSE est de 56.303785.

Erreur relative moyenne (MAPE) : La valeur de MAPE est de 0.307497 signifie que, en moyenne, les prédictions du modèle ont une erreur relative de 30.7497%.

(c) Comparaison des modèles

Nous avons comparé l'ajustement des deux modèles, ainsi que leurs prévisions sur les données de test. Ensuite, nous avons effectué une comparaison des mesures d'évaluation entre les deux modèles.

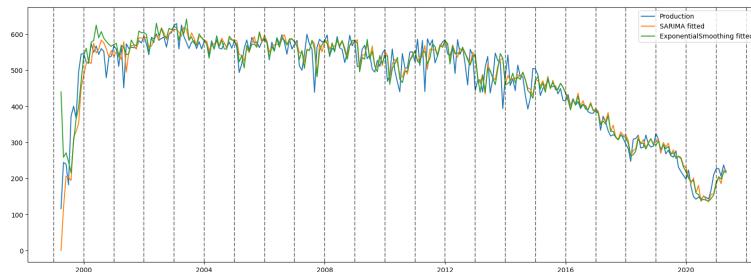


FIGURE 5.44 – Ajustement des 2 modèles.

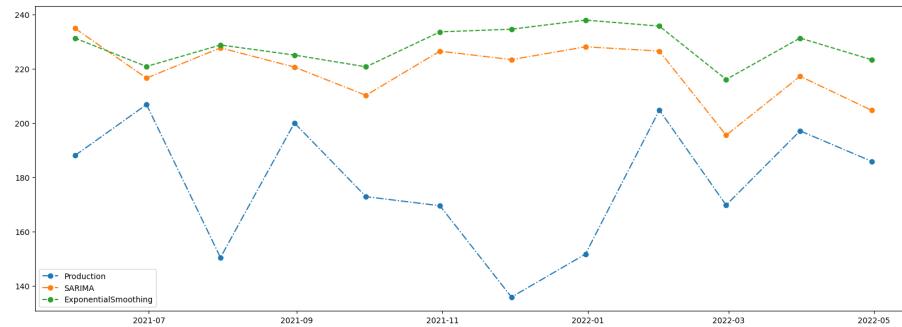


FIGURE 5.45 – Prévisions du test

Modèle	MAPE	MAE	RMSE	R^2
SARIMA	0.255970	41.565882	48.891918	91.078637
Holt-Winters	0.307497	50.506350	56.303785	89.909746

TABLE 5.9 – Résultats des modèles SARIMA et Holt-Winters

En analysant ces mesures (Figure 5.9), nous constatons que le modèle SARIMA présente de meilleures performances globales en termes de précision de prévision. Il affiche un MAPE plus bas (25.5970%) par rapport au modèle Holt-Winters Exponential Smoothing (30.7497%).

De plus, le MAE et le RMSE du modèle SARIMA sont également inférieurs à ceux du modèle Holt-Winters Exponential Smoothing, ce qui indique une meilleure adéquation aux données de test.

Cependant, il est important de noter que le modèle SARIMA obtient un coefficient de détermination (R^2) légèrement supérieur (91.078637%) par rapport au modèle Holt-Winters Exponential Smoothing (89.909746%), ce qui suggère une meilleure capacité à expliquer la variabilité des données observées.

En conclusion, le modèle SARIMA offre de meilleures performances en termes de précision de prévision.

(d) Comparaison avec DCA

L’entreprise nous a transmis les prévisions réalisées par Declin Curve Analysis afin de les comparer à notre modèle SARIMA. Leur modèle a été élaboré à partir des données de production de Gaz Sec du groupement GTFT à partir de 2011, car comme

nous pouvons le constater sur le graphe de production, le déclin de la production débute en 2011.

Nous avons leurs prévisions sur les données de test (Horizon = 12 mois). Ensuite, nous avons effectué une comparaison des mesures d'évaluation entre les deux modèles.

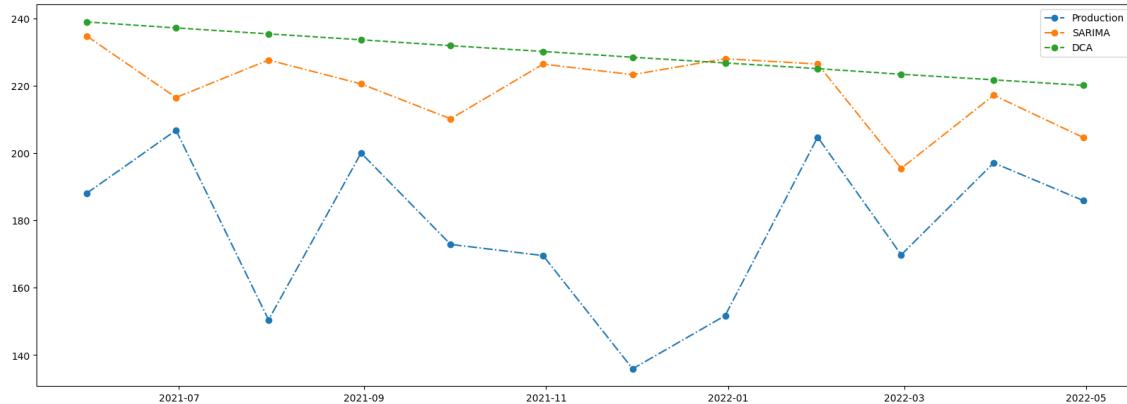


FIGURE 5.46 – Prévisions du Test des deux modèles

Modèle	MAPE	MAE	RMSE
SARIMA	0.255970	41.565882	48.891918
DCA	0.312772	51.716653	56.541119

TABLE 5.10 – Résultats des modèles SARIMA et DCA

Nous constatons (Table 5.10) que le modèle SARIMA offre de meilleures performances de prévision et d'ajustement que l'analyse de courbe de déclin (DCA). Il présente un MAPE plus bas (25.5970%) par rapport à l'analyse de courbe de déclin (DCA) (31.2772%), ainsi qu'un MAE et un RMSE inférieurs. Cela indique une meilleure adéquation aux données de test et une précision accrue dans la prédiction des valeurs réelles de la série temporelle.

SARIMA capture efficacement (Figure 5.46) les schémas temporels complexes, offrant ainsi une modélisation plus précise des séries temporelles. L'analyse de la courbe de déclin convient davantage lorsque le déclin suit une trajectoire spécifique et prévisible de manière régulière. Donc le modèle SARIMA offre une adaptabilité supérieure, une plus grande flexibilité dans les ajustements et des performances de prévision améliorées par rapport à l'analyse de la courbe de déclin.

5. Déploiement

Les prévisions sont calculées pour un horizon $h = 12$, c'est-à-dire, pour la période allant de mai 2022 à avril 2023.

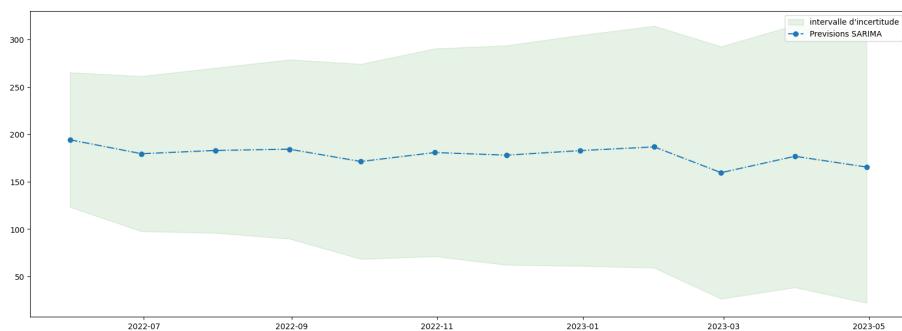


FIGURE 5.47 – Prévisions moyenne de production de gaz (Mai 2022 à Avril 2023) et intervalle d’incertitude.

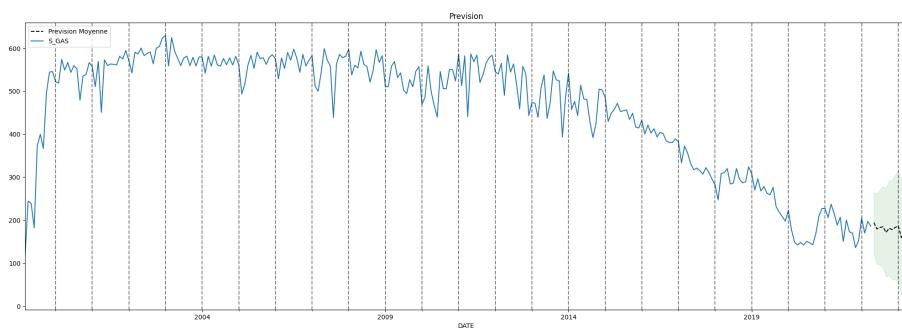


FIGURE 5.48 – Prévisions moyenne de production de gaz (Mai 2022 à Avril 2023)

Intervalles permettent de quantifier l’incertitude associée à chaque valeur prédictive et d’estimer la plage probable dans laquelle la véritable valeur future pourrait se situer.

Date Mensuelle	Production (10^3 m^3)
2022-05	194,20
2022-06	179,50
2022-07	182,92
2022-08	184,29
2022-09	171,28
2022-10	180,76
2022-11	177,96
2022-12	182,79
2023-01	186,70
2023-02	159,52
2023-03	176,79
2023-04	165,45

TABLE 5.11 – Tableau des Prévisions de production de gaz (Mai 2022 à Avril 2023)

5.3.2 Solution Prédition avec le Machine Learning

1. Compréhension des données

Après importation du dataset à l'aide du package **Pandas** sur Python, on procède à son exploration. Notre Dataset (Figure 5.49) se compose de 21 colonnes et 82702 lignes.

	RECORD_DATE	WELL_NAME	H_ON	OIL_VOL_STB	GAS_VOL_MMSCF	GL_VOL_MMSCF	GL_RATE_MMSCF	WH_P	WH_T	CHOKE
82698	2023-05-21	BKNE-B-002B	24.0	1188.1790	15.397	0.000	0.000	1240.03	66.0	160.0
82699	2023-05-21	BKNE-B-004	24.0	759.6245	11.322	0.000	0.000	1359.97	59.0	70.0
82700	2023-05-21	BKNE-B-009	24.0	909.4830	0.100	6.694	6.694	1260.04	40.0	160.0
82701	2023-05-21	BKNE-B-010B	24.0	1483.9300	7.921	0.000	0.000	1280.06	63.0	160.0
82702	2023-05-21	BKNE-B-011	24.0	768.5818	0.032	3.967	3.967	1389.99	45.0	160.0

FIGURE 5.49 – Dataset de production journalière par puits

En explorant le Dataset, on remarque que le pourcentage des valeurs nulles pour les données de production de gaz et d'huile est égal à 0, et donc aucun traitement n'est nécessaire.

Figure 5.50 ci-dessous relève les pourcentages des données manquantes du dataset :

```

RECORD_DATE      0.000000
WELL_NAME       0.000000
H_ON            1.438884
OIL_VOL_STB     0.000000
GAS_VOL_MMSCF   0.000000
GL_VOL_MMSCF    0.000000
GL_RATE_MMSCF   44.561866
WH_P             17.549545
WH_T             18.173464
CHOKE           0.000000
FL_P             0.000000
B_ANNULUS_P     49.553221
A_ANNULUS_P     49.693481
GOR              44.778303
WCUT            62.007424
SALINITY         29.198457
dtype: float64

```

FIGURE 5.50 – Pourcentage de valeurs nulles

Par contre pour les autres colonnes dont le pourcentage est très élevé, il est nécessaire d'éliminer les colonnes complètement car elles n'ont aucune valeur ajoutée et remplacer ces valeurs risquerait d'induire en erreur notre modèle.

Par ailleurs, les valeurs des colonnes **GOR**, **SALINITY**, **WHP**, **WHT** dont le pourcentage est moins élevé peuvent être remplacées.

Comme expliqué dans les chapitres précédents, les principales conditions pour juger que les données sont de bonne qualité et sont appropriées au Machine Learning sont la variété, le volume et l'absence de données manquantes.

Ci-dessous, on peut voir dans Figure 5.51 la distribution des paramètres par rapport à la production de gaz des puits :

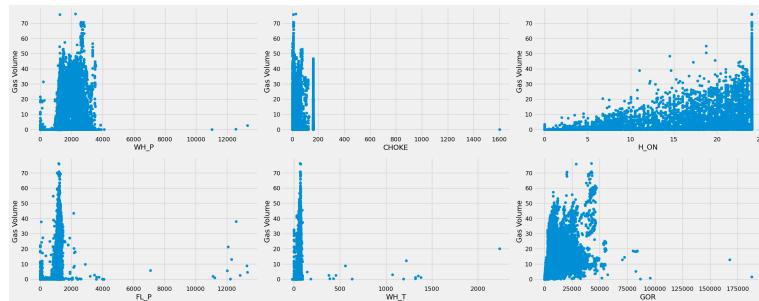


FIGURE 5.51 – Distribution des paramètres

On remarque que les données ne sont pas très variées et risquent donc de diminuer la performance de nos modèles.

Pour une analyse plus approfondie, on procède à l'exploration des données de chaque puits indépendamment. A l'aide des fonctions de Pandas, Figure 5.79 en Annexe affiche le nombre de lignes pour chaque puits.

On retrouve 20 puits et la production de chacun d'entre eux est affichée dans Figure 5.80. (Voir annexe).

Exemple : La Figure 5.52 suivante représente le suivi de production du puits **024** :

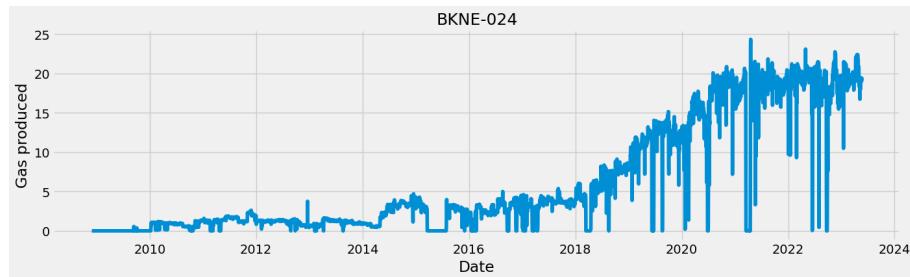


FIGURE 5.52 – Production du Puits 024 après suppression des valeurs aberrantes

L'analyse effectuée nous a permis de voir que la production des puits est instable : ceci est dû au fait que l'on provoque la fermeture du puits et en le rallumant, on injecte le Gaz Lift afin de redynamiser sa production.

Les valeurs de production égales à 0 sont donc insignifiantes et doivent être supprimées pour un bon entraînement du modèle.

On remarque également que chaque puits a un comportement différent :

- La majorité des puits ont une production très réduite voir presque nulle
- Seulement 5 ont une production plus ou moins élevée.

Afin d'analyser le comportement de ces puits et vérifier leur similarité, les boxplots affichés dans Figure 5.53 suivante nous permettent d'avoir la distribution exacte de la

production de chaque puits en précisant les valeurs extrêmes ainsi que la médiane.

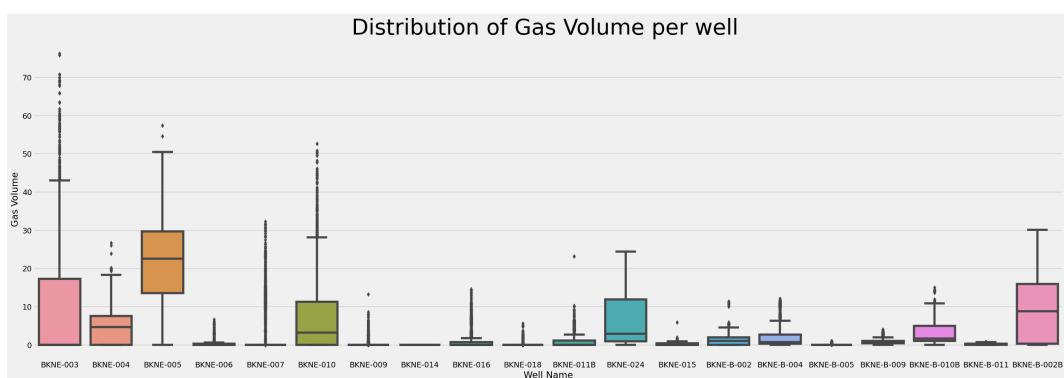


FIGURE 5.53 – Distribution des valeurs de production de gaz par puits

On remarque que La majorité des puits ont une production très réduite voir presque nulle, tandis que les puits **003, 004, 005, 010, 024, 002B** ont une production plus distribuée. Cette analyse nous permet de constater que les résultats ne risquent pas d'être concluants.

2. Préparation des données

Après l'analyse des données, nous procérons au nettoyage et à la transformation.

2.1 Substitution des valeurs manquantes

En général, les valeurs manquantes sont remplacées par la moyenne ou la médiane. Mais pour notre cas les lignes de notre ensemble de données représentent le production journalière par puits, et donc les valeurs des paramètres entre deux jours consécutifs ne varient pas beaucoup, voir pas du tout.

En effet, comme explique dans le chapitre 2, les mesures sont calculées de façon hebdomadaire ou mensuelle en raison du coût des machines.

Les valeurs manquantes des paramètres sont donc remplacées par la valeur précédente non nulle.

2.2 Suppression des valeurs aberrantes

Les valeurs de production égales à 0 sont supprimées.

Figure 5.81 en Annexe affiche le nombre de lignes obtenu de chaque puits après suppression.

Figure 5.54 ci-dessous affiche la production de gaz du puits **024** avant et après suppression des valeurs aberrantes :

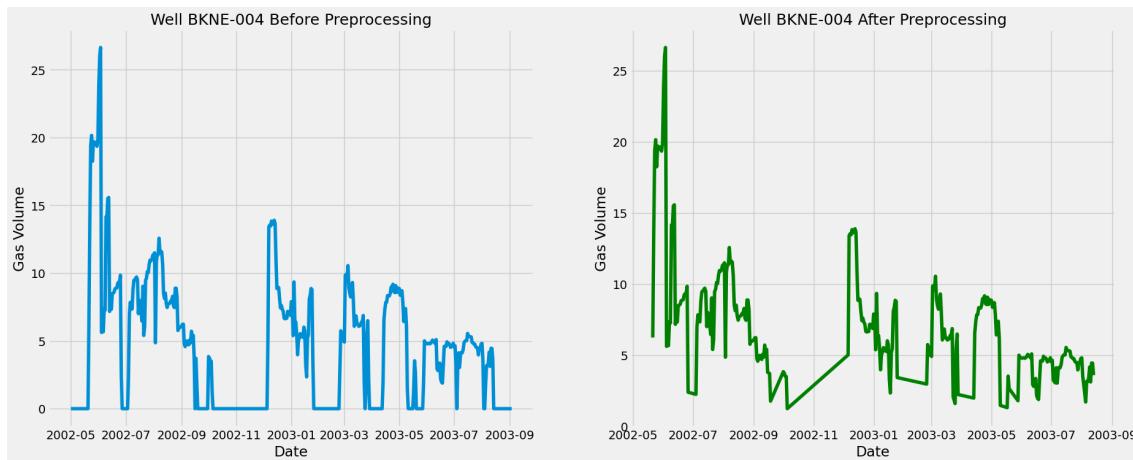


FIGURE 5.54 – Production de gaz par puits

Les valeurs extrêmes des attributs sont traitées à l'aide d'une fonction qui calcule le premier quartile, le troisième quartile et la médiane.

Les instances supérieures au troisième quartile (limite haute) sont remplacées par cette valeur. De même pour les instances en dessous du premier quartile (limite basse) qui sont remplacées par cette dernière.

Le résultat obtenu pour chacune des colonnes est affichée dans Figure 5.55 ci-après :

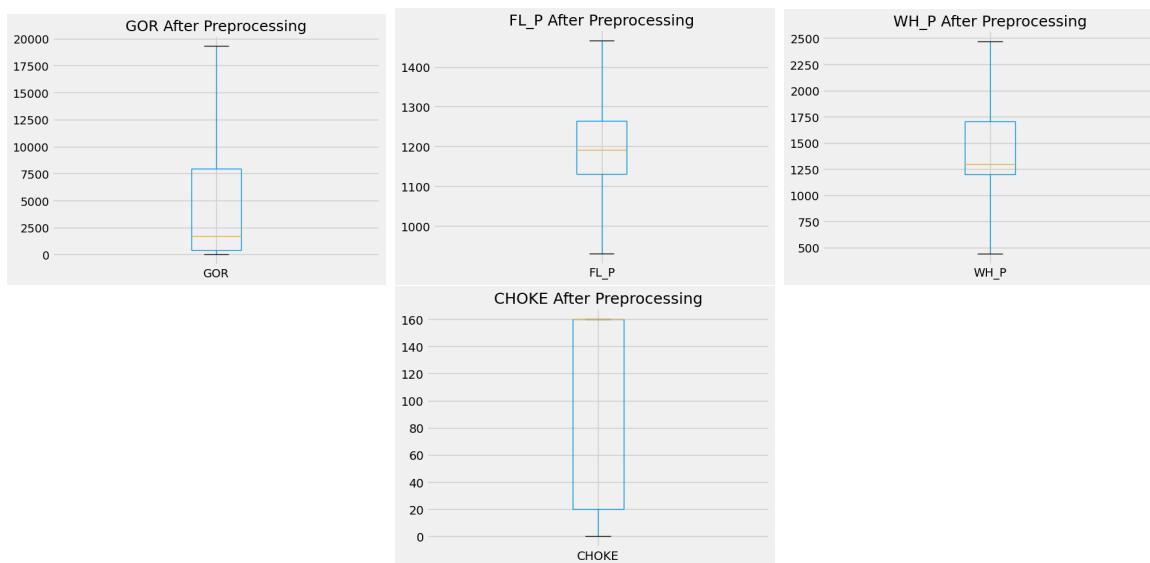


FIGURE 5.55 – Distribution des paramètres après ajustement des valeurs aberrantes

2.3 Sélection des features

Après avoir effectué les transformations nécessaires, nous pouvons à présent sélectionner les caractéristiques les plus importantes à insérer dans nos modèles en analysant la matrice de corrélation dans Figure 5.56 ci-dessous :

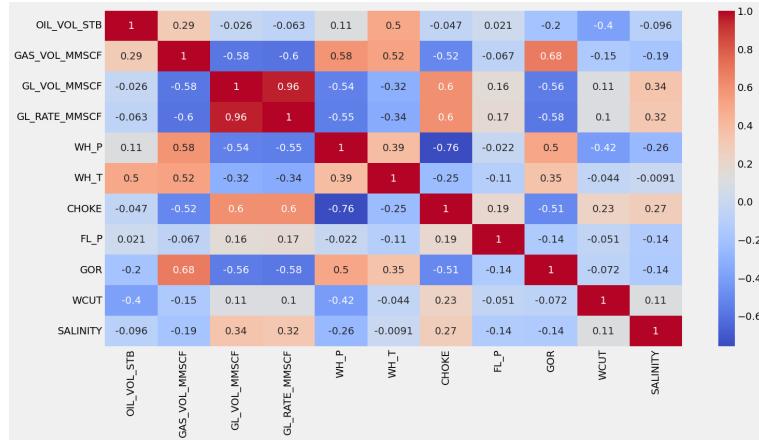


FIGURE 5.56 – Matrice de corrélation globale

Pour notre cas, on remarque que :

- Il existe une forte corrélation négative entre la production de gaz, le volume de **Gaz Lift** injecté, ainsi que la taille de Vanne de cuvelage (**CHOKE**)
- La production de gaz est corrélée positivement à la Pression tête de puits **WHP**, température ambiante autour du puits (**WHT**), ainsi qu'une forte coorelation avec Gaz Oil Ratio (**GOR**).

On peut ainsi affirmer que les attributs qui ont le plus d'impact sur la production de gaz sont :

- GL_VOL
- GL_RATE
- WHP
- WHT
- GOR
- CHOKE

Le reste des colonnes dont la corrélation est réduite peuvent donc être supprimées.

2.4 Standardisation des données

Une fois les features sélectionnées, nous procédons à la standardisation afin que toutes les données soient sur une même échelle et aient un même degré d'importance.

Pour cela, **Scikit-learn** propose plusieurs outils de prétraitement comme le **StandardScaler** qui centre les données autour de la moyenne et divise le tout par la variance.

Voir Figure 5.82 en Annexe pour avoir un aperçu des données standardisées.

2.5 Division du dataset en deux parties Train et Test

On effectue par la suite la séparation de notre dataset en deux sous-ensembles :

- 85% pour l'entraînement du modèle.
- 15% pour vérifier si les résultats obtenus sont exacts .

3. Modélisation

3.1 Choix des meilleurs paramètres pour les algorithmes de prédition

Avant d'appliquer les modèles de prédition et pour plus d'efficacité, on effectue un Grid Search qui teste plusieurs combinaisons (Cross Validation) afin de déterminer les meilleurs paramètres à insérer dans notre algorithme.

Figure 5.57 ci-dessous illustre un exemple de recherche des meilleurs paramètres par Grid Search pour l'algorithme Random Forest Regressor :

```
param_grid = {  
    'max_depth': [5,9,15,110],  
    'max_features': [2,3,5],  
    'min_samples_leaf': [2, 3, 5],  
    'min_samples_split': [8, 10, 12],  
    'n_estimators': [100, 500,1000]  
}  
  
rf = RandomForestRegressor()  
grid_search = GridSearchCV(estimator = rf, param_grid = param_grid, cv = 3, n_jobs = -1, verbose = 2)  
rf_grid_result=grid_search.fit(Xtrain,ytrain)
```

FIGURE 5.57 – Exemple recherche des meilleurs paramètres par Grid Search

Pour notre cas, l'ensemble de données est divisé en sous-ensembles. Pour chaque sous-ensemble, le modèle est formé à l'aide des données des deux autres sous-ensembles et le rendement du modèle est évalué à l'aide des données intégrées. L'erreur de test moyenne sur tous les sous-ensembles détermine par la suite, le meilleur score du **Grid Search**. Cette méthode appelée **Cross Validation** fonctionne assez bien pour les petits ensembles de données.

Table 5.12 ci-dessous résume les paramètres utilisés pour chacun des modèles appliqués :

Model	Parameters	Values
Support Vector Regressor SVR	Kernel C Gamma Epsilon Degree	Rbf 10 Scale 0.01 3
Decision Tree Regressor DTR	Criterion Max_depth Max features Min_samples_leaf Splitter Min_samples_split	Friedman_mse 12 3 3 Best 10
Random Forest Regressor RFR	Bootstrap max_depth Max features Min_samples_leaf Min_samples_split n_estimators	True 15 5 2 8 1000
Gradient Boosting Regressor GBR	learning_rate max_depth n_estimators Min_samples_split subsample	0.1 8 500 2 0.6
Extreme Gradient Boosting Regressor	learning_rate max_depth Gamma reg_alpha n_estimators subsample	0.1 12 10 0.1 500 0.9

TABLE 5.12 – Paramètres des modèles et leurs and Valeurs

3.2 Exécution des algorithmes de prédiction

Après avoir sélectionné les meilleurs paramètres, nous avons importé la bibliothèque **Scikit Learn** pour faire appel aux algorithmes de prédiction. **Nous procédon**s en premier lieu à l'entraînement du modèle : cette étape consiste à actionner notre modèle afin de le familiariser et l'habituer à nos données.

En effet, en insérant les features et la valeur de production à la fois, l'algorithme apprend automatiquement qu'à chaque valeur d'un certain paramètre x , lui correspond une valeur de gaz produite y .

Figure 5.58 suivante affiche un exemple d'entraînement de modèle avec les paramètres Grid Search obtenu :

```
rf=RandomForestRegressor(bootstrap = True,
                         max_depth =15,
                         max_features =5,
                         min_samples_leaf = 2,
                         min_samples_split = 8,
                         n_estimators = 1000)

rf.fit(Xtrain,ytrain)
rf.score(Xtrain,ytrain)
```

FIGURE 5.58 – Exemple Entrainement du modèle Random Forest Regressor

4. Evaluation

L'étape de test nous permet de confirmer l'exactitude de notre entraînement. C'est-à-dire qu'en insérant uniquement les valeurs paramétrées, notre modèle devrait maintenant être capable de déterminer la valeur de production adéquate à chacune de ces valeurs insérées. Pour procéder à l'évaluation, nous avons testé sur les données du Test set préalablement séparés du Train set (15% l'ensemble de données complet).

Figures 5.59, 5.60, 5.61, 5.62 et 5.63 ci-dessous représentent les résultats des algorithmes **SVR**, **DTR**, **RF**, **GBR**, **XGBOOST** (resp.) implémentés :

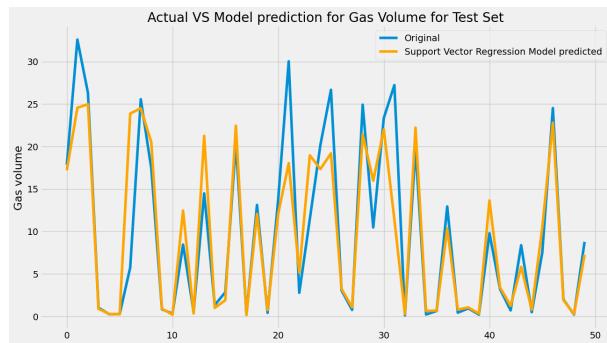


FIGURE 5.59 – Support Vector Regression Model Predictions

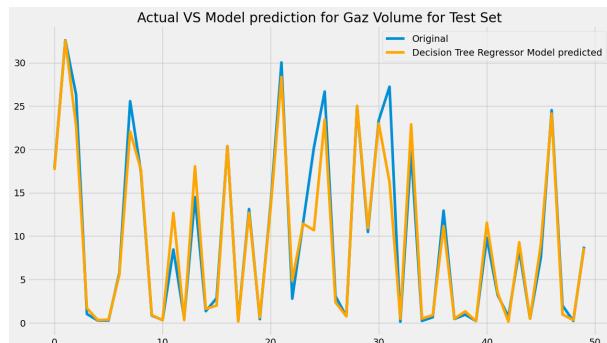


FIGURE 5.60 – Decision Trees Regression Model Predictions

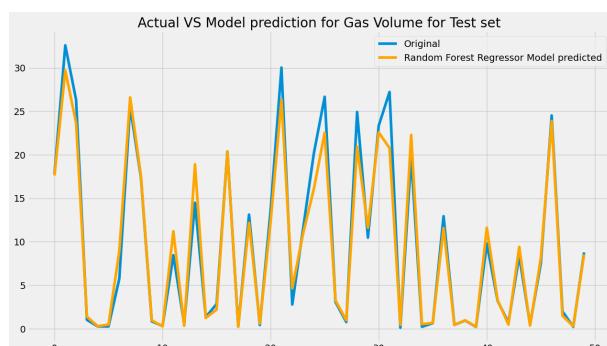


FIGURE 5.61 – Random Forest Regression Model Predictions

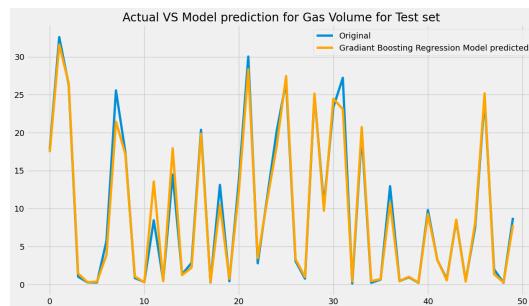


FIGURE 5.62 – Gradient Boosting Regression Model Predictions

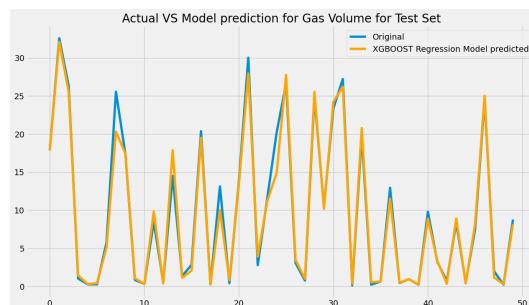


FIGURE 5.63 – Extreme Gradient Boosting Regression Model Predictions XGBOOST

Table 5.13 ci-dessous présente un tableau comparatif entre les différents algorithmes : On constate pour notre cas, que le modèle **Random Forest Regressor** est le plus performant.

Model	R^2	MAE	MSE	RMSE
Support Vector Regressor SVR	0.8436	1.9277	13.1628	6.5814
Decision Tree Regressor DTR	0.9077	1.3343	7.7671	3.8835
Random Forest Regressor RFR	0.9390	1.0227	5.1336	2.5668
Gradient Boosting Regressor GBR	0.9349	1.1097	5.4750	2.7375
Extreme Gradient Boosting Regressor	0.9377	1.0418	5.2490	2.6214

TABLE 5.13 – Model Evaluation Metrics

5. Déploiement

En insérant les valeurs des paramètres mesurés des puits, on peut avoir la production journalière de gaz prédite.

Pour notre cas, le modèle ne peut prédire d'autres données inconnues, car, comme démontré ci-dessus, les données de puits sont peu variées et avec une faible production quotidienne. Cela est dû à des limitations dans la collecte des données et des contraintes de ressources pour effectuer les mesures de production et des paramètres sur chaque puits individuellement.

Conclusion

Il a été question dans ce dernier chapitre, de décrire les étapes suivies pour l'implémentation de notre solution proposée. En premier lieu nous avons présenté les étapes de construction des différentes zones de la solution BI à savoir la zone d'entreposage, la zone ETL et la zone de restitution. Ensuite, les différentes phases de mise en œuvre de notre approche de prédiction et les comparaisons entre les résultats obtenus avec celles de l'entreprise.

CONCLUSION GÉNÉRALE

SONATRACH, plus précisément la division Associations ne dispose d'aucun système d'information décisionnel pour le suivi de ses activités. L'élaboration des rapports analytiques est assurée de façon manuelle et classique et connaît des retards, ce qui n'aide pas les décideurs. Ces derniers souhaitent avoir un système qui fournit les rapports de visualisation de l'évolution de la production sous différents axes d'analyse dans des délais raisonnables afin de prendre les décisions au moment opportun et aussi une solution pour la prédition de production pétrolière afin de mieux anticiper les risques.

Dans le cadre de notre stage, nous avons accompagné les décideurs avec des outils analytiques qui permettent une accessibilité rapide aux informations dont ils ont besoin pour prendre des décisions. Notre projet consiste à mettre en place une solution d'analyse décisionnelle et prévisionnelle pour le suivi de la production dans le domaine pétrolier.

Après avoir rappelé les différents concepts théoriques liés aux systèmes d'information décisionnels, l'entrepôt de données, la modélisation multidimensionnelle et les notions du Data Mining ; nous avons évoqué les approches existantes en matière de prédition et de prévision, à savoir les séries chronologiques et les algorithmes d'apprentissage automatique en mettant l'accent sur leur application dans divers domaines et en particulier dans le domaine pétrolier.

S'ensuit alors l'étape d'étude de l'existant dans laquelle nous avons récolté les besoins exprimés par les décideurs de l'entreprise. A l'issue de cette partie, nous avons proposé notre solution, qui consistait à la mise en place d'une solution BI pour l'analyse et une solution Data Mining qui, à son tour se scinde en solution prévisionnelle et solution prédictive.

Afin de concrétiser notre solution BI, nous tenons à rappeler ci-dessous les principales étapes par lesquelles nous sommes passés :

- Extraction, Préparation et alimentation de données (ETL) sur Python et SQL Server Integration Services.
- Conception du magasin de données sur SQL Server.
- Conception de la zone de restitution en utilisant Microsoft Powerbi.

En ce qui concerne le volet DataMining , nous avons suivi la méthode CRISP-DM composée de 6 étapes, en commençant par l'étape compréhension du contexte métier et menant à l'étape déploiement en mettant particulièrement l'accent sur l'utilisation de modèles d'apprentissage automatique et séries chronologiques. Nous avons exploré comment ces modèles peuvent être adaptés pour répondre aux défis spécifiques de ce domaine complexe, tels que la prédition de la production.

Nos résultats ont démontré que l'utilisation de l'apprentissage automatique dans le domaine pétrolier offre des avantages significatifs en termes de précision, d'efficacité et de prise de dé-

cision éclairée. Les modèles développés ont montré une capacité à prédire avec précision la production future.

L'utilisation de l'analyse des séries chronologiques nous a fourni des informations précieuses sur les tendances, les schémas saisonniers et les variations de la production. Les résultats obtenus nous permettent d'améliorer la rentabilité, d'anticiper les fluctuations futures et de prendre des décisions stratégiques.

Cette étude a été très bénéfique et très instructive dans la mesure où elle nous a permis de nous familiariser avec le domaine professionnel et nous a ouvert de nouvelles perspectives pour l'application de l'analyse décisionnelle et prévisionnelle dans ce secteur vital de l'économie.

Nous avons mis en pratique les connaissances acquises tout au long de notre cursus universitaire dans le but de développer nos performances techniques et les enrichir. Cependant, des recherches supplémentaires sont nécessaires pour surmonter les défis spécifiques et pour explorer d'autres domaines d'application de l'analyse décisionnelle et prévisionnelle.

Cependant, nous avons également souligné certaines limites et défis associés à l'application de l'apprentissage automatique dans le domaine pétrolier, tels que la disponibilité limitée des données.

Il est important de noter que la prédiction de production de gaz est une tâche complexe et qu'elle comporte une certaine marge d'erreur. Les résultats peuvent varier en fonction de nombreux facteurs, notamment les incertitudes géologiques, les changements dans les conditions de marché... Par conséquent, nous pouvons proposer les perspectives d'évolution suivantes :

- Enrichir les analyses, en ajoutant un système décisionnel pour chaque sujet d'analyse : à savoir, le forage, l'exploration pour une meilleure maîtrise sur les différents aspects liés à la production.
- Améliorer la collecte de données : Nous devons aborder la question de la non disponibilité des données en travaillant sur l'amélioration des processus de collecte de données.
- Développer une plateforme analytique intégrée : Nous pouvons envisager de regrouper nos solutions dans une plateforme intégrée conviviale. Une telle plateforme permettrait aux décideurs d'accéder facilement aux informations nécessaires à partir d'un seul endroit, en facilitant la prise de décision rapide.
- Exploiter de nouvelles sources de données : Pour enrichir nos analyses et nos prédictions, nous devons considérer l'exploration de nouvelles sources de données. Par exemple, l'intégration de données géographiques ou de capteurs IoT peut apporter des informations supplémentaires précieuses les facteurs géologiques et l'état d'équipements de l'entreprise, ce qui peut contribuer à des décisions plus éclairées.

ANNEXES

	Mois	ID_Grp	Nom_Grp	Prev_Oil_J	Year	month	YYYY-MM	Prev_Gaz_M_J	Prev_Cond_M_J	Prev_GPL_M_J
2		1	GROUPEMENT				1999-01	1,00000000	1,00000000	1,00000000
3		1	GROUPEMENT				1999-02	1,00000000	1,00000000	1,00000000
4		1	GROUPEMENT				1999-03	1,00000000	1,00000000	1,00000000
5		1	GROUPEMENT				1999-04	1,00000000	1,00000000	1,00000000
6		1	GROUPEMENT				1999-05	1,00000000	1,00000000	1,00000000
7		1	GROUPEMENT				1999-06	1,00000000	1,00000000	1,00000000
8		1	GROUPEMENT				1999-07	1,00000000	1,00000000	1,00000000
9		1	GROUPEMENT				1999-08	1,00000000	1,00000000	1,00000000
10		1	GROUPEMENT				1999-09	1,00000000	1,00000000	1,00000000
11		1	GROUPEMENT				1999-10	1,00000000	1,00000000	1,00000000
12		1	GROUPEMENT				1999-11	1,00000000	1,00000000	1,00000000
13		1	GROUPEMENT				1999-12	1,00000000	1,00000000	1,00000000
14		1	GROUPEMENT				2000-01	1,00000000	1,00000000	1,00000000
15		1	GROUPEMENT				2000-02	1,00000000	1,00000000	1,00000000
16		1	GROUPEMENT				2000-03	1,00000000	1,00000000	1,00000000
17		1	GROUPEMENT				2000-04	1,00000000	1,00000000	1,00000000
18		1	GROUPEMENT				2000-05	1,00000000	1,00000000	1,00000000
19		1	GROUPEMENT				2000-06	1,00000000	1,00000000	1,00000000
20		1	GROUPEMENT				2000-07	1,00000000	1,00000000	1,00000000
21		1	GROUPEMENT				2000-08	1,00000000	1,00000000	1,00000000
22		1	GROUPEMENT				2000-09	1,00000000	1,00000000	1,00000000
23		1	GROUPEMENT				2000-10	1,00000000	1,00000000	1,00000000
24		1	GROUPEMENT				2000-11	1,00000000	1,00000000	1,00000000
25		1	GROUPEMENT				2000-12	1,00000000	1,00000000	1,00000000
26		1	GROUPEMENT				2001-01	1,00000000	1,00000000	1,00000000
27		1	GROUPEMENT				2001-02	1,00000000	1,00000000	1,00000000
28		1	GROUPEMENT				2001-03	1,00000000	1,00000000	1,00000000

FIGURE 5.64 – Prévisions des groupements GTFT / GTIM finaux

FIGURE 5.65 – Fichier final de Production Berkine

FIGURE 5.66 – Ficher final de production par le groupement GSS

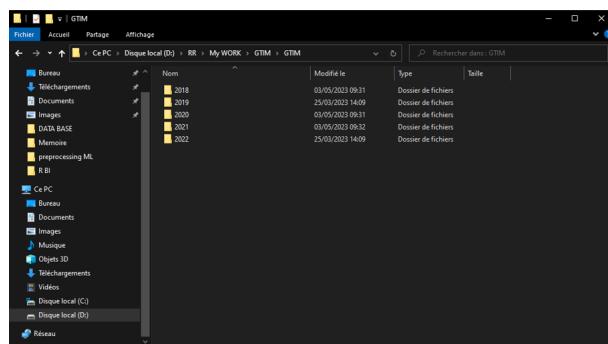


FIGURE 5.67 – Données partagées par l’entreprise

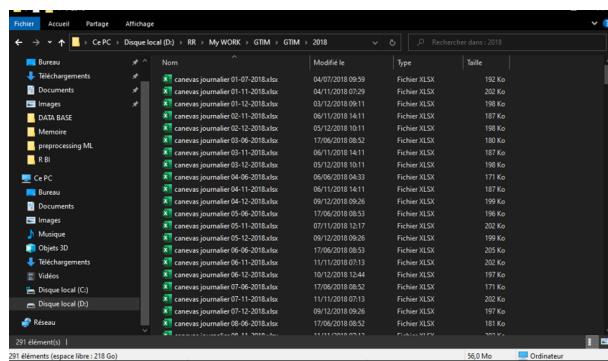


FIGURE 5.68 – Répertoire des rapports journaliers 2018 du groupement GTIM

Details du PseudoCode GTIM

1. Lire un fichier du dossier et vérifier qu'il est bien de type **XLSX**.
2. Récupérer la feuille du fichier.
3. Effectuer un test pour déterminer la dimension de cette feuille.
4. Extraire la date de la cellule en utilisant une expression régulière (en tenant compte de la présence de plusieurs formats de dates) et la convertir au format de date approprié.
5. Récupérer le tableau et effectuer quelques modifications (modifier le schéma, supprimer les colonnes supplémentaires ou vides, remplir la colonne "périmètre" avec les valeurs manquantes, supprimer certaines anomalies).
6. Ajouter une colonne "Date" au dataframe (le tableau modifié) et remplir chaque ligne avec la date récupérée précédemment.
7. Créer un nouveau fichier **CSV** qui servira de fichier final et ajouter le dataframe (si le fichier existe déjà, le nouveau dataframe sera directement ajouté au fichier existant).

Remarque : la condition de ne pas ajouter un dataframe ou bien une ligne déjà existante est gérée automatiquement via **Pandas**. On refait ce traitement pour chaque fichier du dossier jusqu'à obtenir le fichier global sous format csv.

ANNEXES

Puits	Maifeld	gisement	groupement	Partenaire	Type de contrat	Loi de hydrocarbures	Bassin sédimentaire
01-01							
01-02							
01-03							
01-04							
01-05							
01-06							
01-07							
01-08							
01-09							
01-10							
01-11							
01-12							
01-13							
01-14							
01-15							
01-16							
01-17							
01-18							
01-19							
01-20							
01-21							
01-22							
01-23							
01-24							
01-25							
01-26							
01-27							
01-28							
01-29							
01-30							
01-31							
01-32							
01-33							
01-34							
01-35							
01-36							
01-37							
01-38							
01-39							
01-40							
01-41							
01-42							
01-43							
01-44							
01-45							
01-46							
01-47							
01-48							
01-49							
01-50							
01-51							
01-52							
01-53							
01-54							
01-55							
01-56							
01-57							
01-58							
01-59							
01-60							
01-61							
01-62							
01-63							
01-64							
01-65							
01-66							
01-67							
01-68							
01-69							
01-70							
01-71							
01-72							
01-73							
01-74							
01-75							
01-76							
01-77							
01-78							
01-79							
01-80							
01-81							
01-82							
01-83							
01-84							
01-85							
01-86							
01-87							
01-88							
01-89							
01-90							
01-91							
01-92							
01-93							
01-94							
01-95							
01-96							
01-97							
01-98							
01-99							
01-100							
01-101							
01-102							
01-103							
01-104							
01-105							
01-106							
01-107							
01-108							
01-109							
01-110							
01-111							
01-112							
01-113							
01-114							
01-115							
01-116							
01-117							
01-118							
01-119							
01-120							
01-121							
01-122							
01-123							
01-124							
01-125							
01-126							
01-127							
01-128							
01-129							
01-130							
01-131							
01-132							
01-133							
01-134							
01-135							
01-136							
01-137							
01-138							
01-139							
01-140							
01-141							
01-142							
01-143							
01-144							
01-145							
01-146							
01-147							
01-148							
01-149							
01-150							
01-151							
01-152							
01-153							
01-154							
01-155							
01-156							
01-157							
01-158							
01-159							
01-160							
01-161							
01-162							
01-163							
01-164							
01-165							
01-166							
01-167							
01-168							
01-169							
01-170							
01-171							
01-172							
01-173							
01-174							
01-175							
01-176							
01-177							
01-178							
01-179							
01-180							
01-181							
01-182							
01-183							
01-184							
01-185							
01-186							
01-187							
01-188							
01-189							
01-190							
01-191							
01-192							
01-193							
01-194							
01-195							
01-196							
01-197							
01-198							
01-199							
01-200							
01-201							
01-202							
01-203							
01-204							
01-205							
01-206							
01-207							
01-208							
01-209							
01-210							
01-211							
01-212							
01-213							
01-214							
01-215							
01-216							
01-217							
01-218							
01-219							
01-220							
01-221							
01-222							
01-223							
01-224							
01-225							
01-226							
01-227							
01-228							
01-229							
01-230							
01-231							
01-232							
01-233							
01-234							
01-235							
01-236							
01-237							
01-238							
01-239							
01-240							
01-241							
01-242							
01-243							
01-244							
01-245							
01-246							
01-247							
01-248							
01-249							
01-250							
01-251							
01-252							
01-253							
01-254							

FIGURE 5.72 – La table DPRD

<input checked="" type="checkbox"/> Maifold	<input checked="" type="checkbox"/> gisement	<input checked="" type="checkbox"/> groupement	<input checked="" type="checkbox"/> Partenaire	<input checked="" type="checkbox"/> Type de contrat	<input checked="" type="checkbox"/> Loi de hydrocarbures	<input checked="" type="checkbox"/> Bassin sédimentaire
958	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
959	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
960	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
961	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
962	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
963	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
964	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
965	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
966	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
967	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
968	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
969	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
970	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
971	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
972	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
973	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
974	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
975	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
976	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
977	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
978	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
979	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
980	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
981	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
982	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
983	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
984	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
985	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
986	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
987	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
988	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
989	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
990	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
991	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
992	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
993	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
994	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
995	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
996	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
997	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
998	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	
999	1917-0000	00000	Sonatrach Totalenergies	Concession	'95-07'	

FIGURE 5.73 – Les Contrats du groupement GTFT

Details Tableau de bord

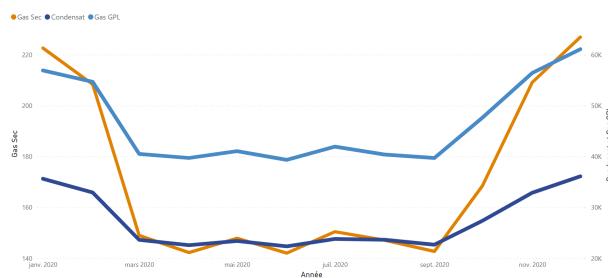


FIGURE 5.74 – Suivi de production de Gaz Sec, GPL, Condensat du groupement GTFT pour l'année 2020

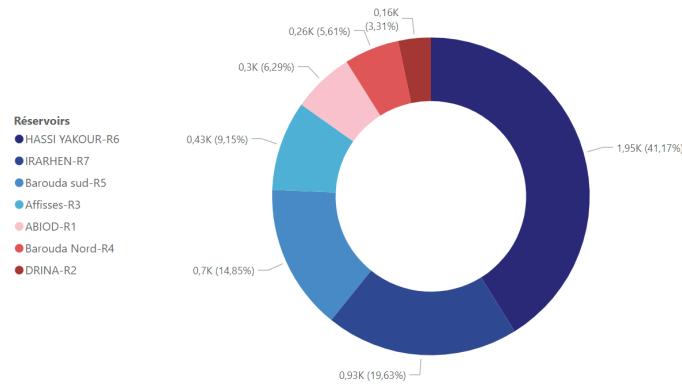


FIGURE 5.75 – Taux de production de Gaz Sec par réservoir du groupement GTIM pour l'année 2020

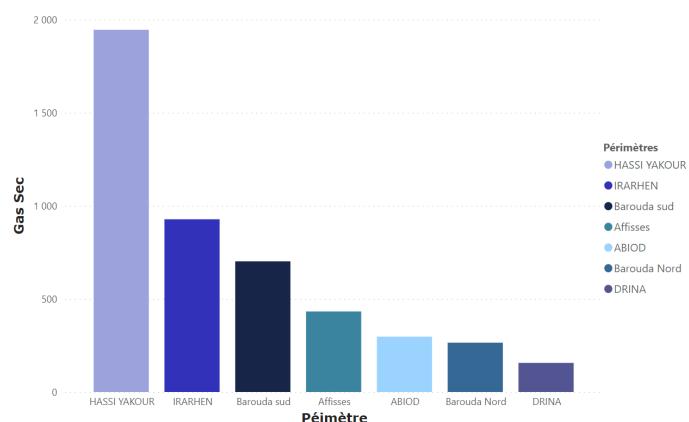


FIGURE 5.76 – Taux de production de Gaz Sec par périmètre du groupement GTIM pour l'année 2020

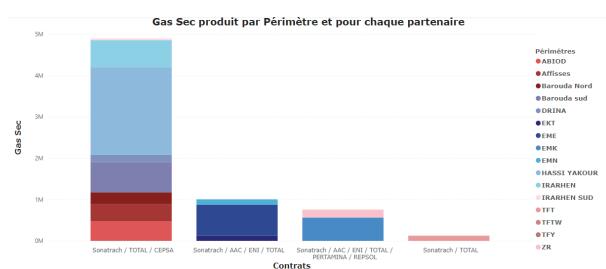


FIGURE 5.77 – Liste des partenaires et taux de production de Gaz Sec par périmètre en 2020

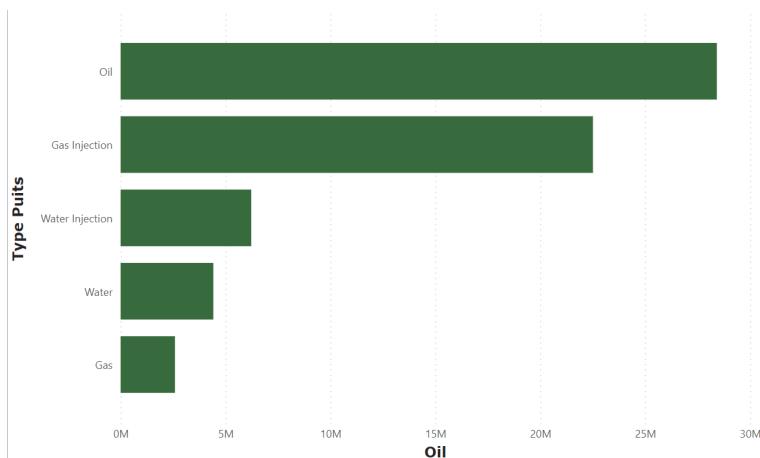


FIGURE 5.78 – Taux de production de Oil par type de puits en 2020

Details du ML

BKNE-003	7687
BKNE-005	7687
BKNE-010	7131
BKNE-009	6965
BKNE-016	6875
BKNE-024	5283
BKNE-B-004	4698
BKNE-B-009	4318
BKNE-B-011	4184
BKNE-B-010B	4184
BKNE-007	4171
BKNE-018	3667
BKNE-B-002B	2841
BKNE-011B	2699
BKNE-B-002	2517
BKNE-015	2511
BKNE-006	2441
BKNE-B-005	1743
BKNE-014	610
BKNE-004	491

Name: WELL_NAME, dtype: int64

FIGURE 5.79 – Nombre de lignes par puits

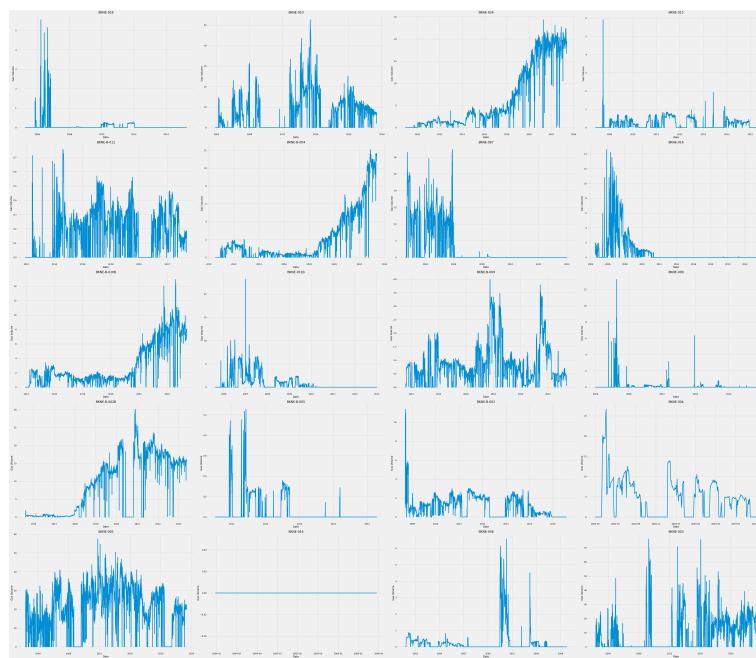


FIGURE 5.80 – Production de gaz par puits

```
BKNE-005      6494
BKNE-024      4542
BKNE-B-004    3831
BKNE-010      3782
BKNE-B-010B   3762
BKNE-B-009    3703
BKNE-003      3575
BKNE-B-011    2767
BKNE-B-002B   2476
BKNE-016      1897
BKNE-B-002    1840
BKNE-009      1494
BKNE-015      1244
BKNE-011B    1016
BKNE-007      996
BKNE-006      963
BKNE-018      731
BKNE-004      307
BKNE-B-005    279
Name: WELL_NAME, dtype: int64
```

FIGURE 5.81 – Nombre de lignes par puits après suppression des valeurs aberrantes

```
array([[-0.95210201, -0.96607161,  1.96874749, -0.02426407, -1.41009971,
       -0.1505992 ],
      [-0.95210201, -0.96607161,  1.96874749,  0.97206819, -1.42456031,
       -0.16148435],
      [-0.95210201, -0.96607161,  1.96874749,  1.23775679, -1.36671791,
       -0.16144551],
      ...,
      [ 2.63419563,  2.42225076, -0.52641487, -1.02059634,  0.83129336,
       -0.79450456],
      [-0.95210201, -0.96607161, -0.48507033,  0.50711313,  0.83129336,
       0.10403789],
      [ 1.17321062,  1.04191686, -0.25804706, -0.68848558,  0.83129336,
       -0.80624658]])
```

FIGURE 5.82 – Aperçu des données standardisées

Bibliographie

- [1] GHEBACHE Bouchrar ABADINE Meriem. La prédition de la perte (churn) de clients pour une entreprise commerciale cas des opérateurs téléphoniques. Master's thesis, US-THB, 2020.
- [2] Igor N. Aizenberg, Leonid Sheremetov, Luis Villa-Vargas, and Jorge Martínez Muñoz. Multilayer neural network with multi-valued neurons in time series forecasting of oil production. *Neurocomputing*, 175 :980–989, 2016.
- [3] Jan J Arps. Analysis of decline curves. *Transactions of the AIME*, 160(01) :228–247, 1945.
- [4] Abdrahmane AW. Mise en place d'un entrepôt de données pour l'aide à la décision médicale, 2014. Consulté avril 2023.
- [5] Ana Azevedo and Manuel Filipe Santos. Kdd, semma and crisp-dm : a parallel overview. *IADS-DM*, 2008.
- [6] Régis Bourbonnais and Virginie Terraza. Analyse des séries temporelles. Technical report, 2022.
- [7] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis : forecasting and control*. John Wiley & Sons, 2015.
- [8] Leo Breiman. Random forests. *Mach. Learn.*, 45(1) :5–32, 2001.
- [9] Chris Chatfield. The holt-winters forecasting procedure. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 27(3) :264–279, 1978.
- [10] Tianqi Chen and Carlos Guestrin. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, aug 2016.
- [11] N. Chithra Chakra, Ki-Young Song, Madan M. Gupta, and Deoki N. Saraf. An innovative neural forecast of cumulative oil production from a petroleum reservoir employing higher-order neural networks (honns). *Journal of Petroleum Science and Engineering*, 106 :18–33, 2013.
- [12] N. Chithra Chakra, Ki-Young Song, Madan M. Gupta, and Deoki N. Saraf. An innovative neural forecast of cumulative oil production from a petroleum reservoir employing higher-order neural networks (honns). *Journal of Petroleum Science and Engineering*, 106 :18–33, 2013.
- [13] Jaesung Choi, David Roberts, and Eunsu Lee. Forecasting oil production in north dakota using the seasonal autoregressive integrated moving average (s-arima). *Natural Resources*, 6 :16–26, 01 2015.
- [14] DataValue Consulting. Schéma sur le traitement des données, date inconnue. Consulté avril 2023.

- [15] Rainer Dietrich, Manfred Opper, and Haim Sompolinsky. Statistical mechanics of support vector networks. *Physical review letters*, 82(14) :273–297, 1999.
- [16] SF Ding, BJ Qi, and HY Tan. An overview on theory and algorithm of support vector machines. *Journal of University of Electronic Science and Technology of China*, 40(1) :2–10, 2011.
- [17] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.
- [18] Jean-Michel Franco and Sandrine de Lignerolles. *Piloter l’entreprise grâce au data warehouse*. Eyrolles, 2000.
- [19] Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780) :1612, 1999.
- [20] Ridha Gharbi and G Ali Mansoori. An introduction to artificial intelligence applications in petroleum exploration and production. *Journal of Petroleum Science and Engineering - J PET SCI ENGINEERING*, 49 :93–96, 12 2005.
- [21] J.-F. GHOZZI. *Informatique : conception et manipulation de bases de données dimensionnelles à contraintes*. Thèse doctorale en informatique, Université Paul Sabatier - Toulouse III, 2004.
- [22] Jiawei Han, Micheline Kamber, and Jian Pei. Data mining concepts and techniques third edition. *University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University*, 2012.
- [23] Trevor Hastie, Jerome H. Friedman, and Robert Tibshirani. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2001.
- [24] ZK He, GB Liu, XJ Zhao, and MH Wang. Overview of gaussian process regression. *Control and Decision*, 28(8) :1121–1129, 2013.
- [25] John Hirschmiller, Anton Biryukov, Bertrand Groulx, Brian Emmerson, and Scott Quinell. The importance of integrating subsurface disciplines with machine learning when predicting and optimizing well performance—case study from the spirit river formation. In *SPE Annual Technical Conference and Exhibition*. OnePetro, 2019.
- [26] S.L. Ho and M. Xie. The use of arima models for reliability forecasting and analysis. *Computers Industrial Engineering*, 35(1) :213–216, 1998.
- [27] W. H. Inmon. *Building the Data Warehouse,3rd Edition*. John Wiley Sons, Inc., USA, 3rd edition, 2002.
- [28] Martin Längkvist, Lars Karlsson, and Amy Loutfi. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognit. Lett.*, 42 :11–24, 2014.
- [29] Lulu Liao, Gensheng Li, Hongbao Zhang, Jiangpeng Feng, Yijin Zeng, Ke Ke, and Zhifa Wang. Well completion optimization in canada tight gas fields using ensemble machine learning. In *Abu Dhabi International Petroleum Exhibition & Conference*. OnePetro, 2020.
- [30] Luis Martí, Nayat Sánchez Pi, José Manuel Molina, and Ana Cristina Bicharra Garcia. Anomaly detection based on sensor data in petroleum industry applications. *Sensors*, 15(2) :2774–2797, 2015.

- [31] Dr. Rajesh Mehrotra and Regilal Gopalan. FACTORS INFLUENCING STRATEGIC DECISION-MAKING PROCESS FOR THE OIL/GAS INDUSTRIES OF UAE- A STUDY. *International Journal of Marketing & Financial Management*, ISSN : 2348 –3954 (Online) ISSN : 2349 –2546 (Print),, Volume 5(Issue 1, Jan-2017) :pp 62–69, January 2017.
- [32] Sid Ahmed Djallal Midouni, Jérôme Darmont, and Fadila Bentayeb. Approche de modélisation multidimensionnelle des données complexes : application aux données médicales,. In *Journées Francophones sur les Entrepôts de Données et l'Analyse en ligne*, 2009.
- [33] Douglas C Montgomery, Cheryl L Jennings, and Murat Kulahci. *Introduction to time series analysis and forecasting*. John Wiley & Sons, 2015.
- [34] Roar Nyboe. Fault detection and other time series opportunities in the petroleum industry. *Neurocomputing*, 73 :1987–1992, 06 2010.
- [35] Wilfredo Palma. *Time series analysis*. John Wiley & Sons, 2016.
- [36] Yogendra Narayan Pandey, Ayush Rastogi, Sribharath Kainkaryam, Srimoyee Bhattacharya, and Luigi Saputelli. Machine learning in the oil and gas industry. *Mach Learning in Oil Gas Industry*, 2020.
- [37] David Parmenter. *Key performance indicators : developing, implementing, and using winning KPIs*. John Wiley & Sons, 2015.
- [38] J. Ross Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1) :81–106, 1986.
- [39] GUENNOUN Nada Raihan and BAHRI Somia. Mise en place d'une solution business intelligence au profit de easy relay. Mémoire de projet de fin d'études, Ecole Nationale Supérieur D'Informatique ESI.
- [40] Margy Ross and Ralph Kimball. *The data warehouse toolkit : the definitive guide to dimensional modeling*. John Wiley & Sons, 2013.
- [41] Hemlata Sahu, Shalini Shrma, and Seema Gondhalakar. A brief overview on data mining survey. *International Journal of Computer Technology and Electronics Engineering*, 1(3) :114–21, 2011.
- [42] Gilbert Saporta and Gilles Stoltz. Gilbert saporta : un parcours éclectique, statistique et société, volume 8, n° 1. *Statistique et Société*, 2020.
- [43] Yousef Sheikhi Garjan and Mehdi Ghaneezabadi. Machine learning interpretability application to optimize well completion in montney. In *SPE Canada Unconventional Resources Conference*. OnePetro, 2020.
- [44] A. (MC2i Groupe) SIBAI. La méthode crisp : une solution pour réussir vos projets big data. Consulté juin 2023.
- [45] Hedong Sun. *Advanced production decline analysis and application*. Gulf professional publishing, 2015.
- [46] Carlo Vercellis. *Business intelligence : data mining and optimization for decision making*. John Wiley & Sons, 2011.
- [47] Shuhua Wang and Shengnan Chen. Insights to fracture stimulation design in unconventional reservoirs based on machine learning modeling. *Journal of Petroleum Science and Engineering*, 174 :682–695, 2019.

- [48] Rüdiger Wirth and Jochen Hipp. Crisp-dm : Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1. Manchester, 2000.
- [49] Liang Xue, Yuetian Liu, Yifei Xiong, Yanli Liu, Xuehui Cui, and Gang Lei. A data-driven shale gas production forecasting method based on the multi-objective random forest regression. *Journal of Petroleum Science and Engineering*, 196 :107801, 2021.

Webographie

- [50] *Python Software Foundation*. Documentation python 3. <http://docs.python.org/3/>.
- [51] *PyData*. Documentation de pandas. <https://pandas.pydata.org/docs/>.
- [52] *Scikit-learn*. Guide de l'utilisateur scikit-learn. https://scikit-learn.org/stable/user_guide.html.
- [53] *Sktime Team*. Documentation de sktime. <https://www.sktime.net/en/latest/>.
- [54] *Statsmodels*. Documentation de statsmodels. <https://www.statsmodels.org/stable/index.html>.
- [55] *Microsoft*. Power BI Documentation. <https://learn.microsoft.com/fr-fr/power-bi/>

Résumé

SONATRACH, souhaite analyser ses données de production et automatiser le processus de prédiction pour mieux contrôler le fonctionnement de chaque puits individuellement et anticiper les risques.

Notre projet intitulé « **Conception d'une solution décisionnelle pour la production de gaz et prédictions** » vise à combiner la BI, les séries chronologiques Time Series et les techniques de ML d'une part pour élaborer des rapports analytiques visant à avoir un suivi journalier de production et d'autre part pour anticiper les valeurs futures de production afin de prendre les bonnes décisions au moment opportun. Il s'agit en premier lieu de concevoir un entrepôt de données afin de stocker les données de production des années précédentes et les présenter ensuite sous forme de tableau de bord pour répondre à leurs besoins d'analyse de suivi. Par ailleurs, Time Series offre des résultats plus précis que **Decline Curve Analysis DCA**. Toutefois, cette méthode basée sur l'historique de production de tout le groupement présente des inconvénients dans le cas du forage d'un nouveau puits ou dans le cas où l'on souhaite avoir la production exacte par puits. Afin de permettre à l'entreprise une meilleure maîtrise sur les puits individuellement, les techniques de ML fournissent une prédiction en se basant sur les paramètres de production.

Mots clés : Business Intelligence, Data Mining, Prédiction, Prévision, Machine Learning, Time Series Analysis...