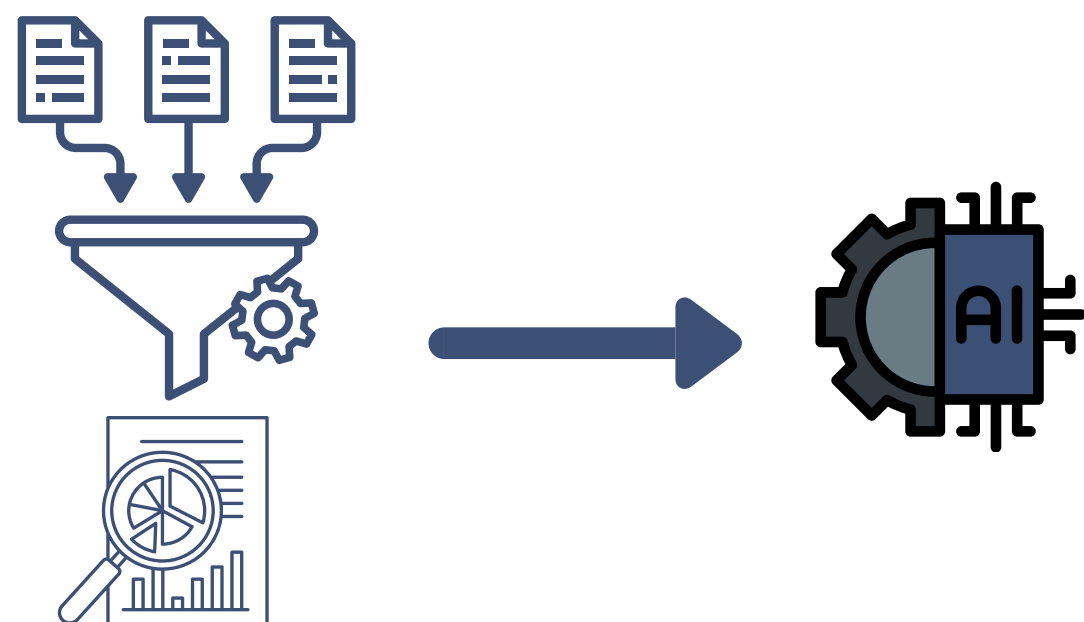


AUDIT INTELLIGENT : DéTECTION D'ANOMALIES DANS LES JOURNAUX COMPTABLES VIA MACHINE LEARNING

Réalisé par :
FALAQ Jad



Encadrant externe:
GHANAM Kamal
EL MOUADDIN Reda

PLAN

Introduction

I . Business Understanding

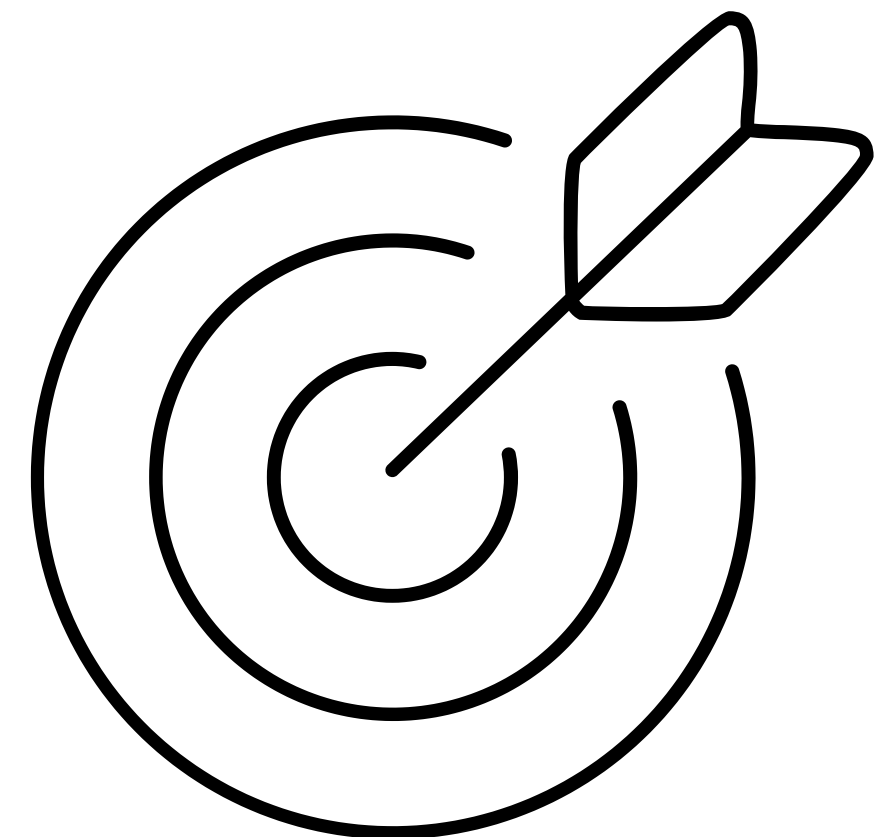
II . Data Understanding

III . Data Preprocessing

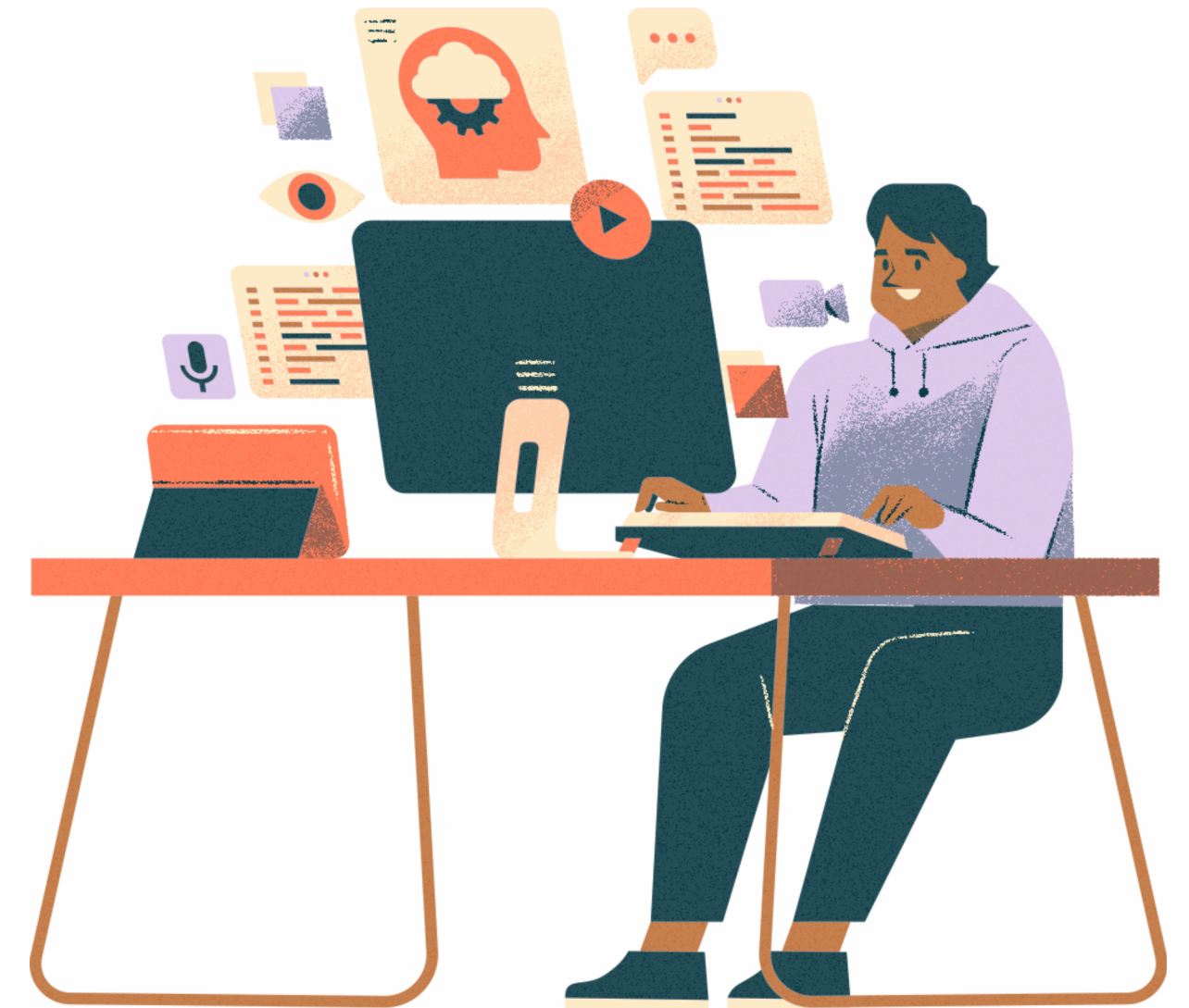
IV . Modeling

V . Evaluation

Conclusion



Introduction



Introduction

Introduction



Akwa Group est un grand conglomérat privé marocain fondé en 1932 par les familles Akhannouch et Wakrim. Il opère dans divers secteurs, notamment les hydrocarbures (pétrole et gaz) grâce à sa marque phare Afriquia, ainsi que le tourisme, les médias, l'immobilier et les énergies renouvelables. Le groupe bénéficie d'une forte présence nationale, sa filiale Afriquia gérant le plus grand réseau de stations-service du Maroc.



Introduction



La Direction des Systèmes d'Information (DSI) du groupe Akwa est structurée en trois pôles principaux :

Le pôle Réseaux, Le pôle Sécurité, et Le pôle Data

La DSI travaille en étroite coordination avec le service JDE, qui gère la comptabilité et les opérations financières via l'ERP Oracle JDE. Les données issues de ce système sont exploitées par le pôle Data pour améliorer le suivi et la performance des activités du groupe.



CRISP-DM. La méthodologie Cross-Industry Standard Process for Data Mining (CRISP-DM) comporte 6 étapes, commençant par la compréhension du métier et se terminant par le déploiement de la solution développée. Les tâches liées aux données sont traitées dans les étapes de compréhension des données, préparation des données et modélisation, puis évaluées dans l'étape d'évaluation. Des itérations peuvent être effectuées à tout moment, de la compréhension métier jusqu'à l'évaluation.

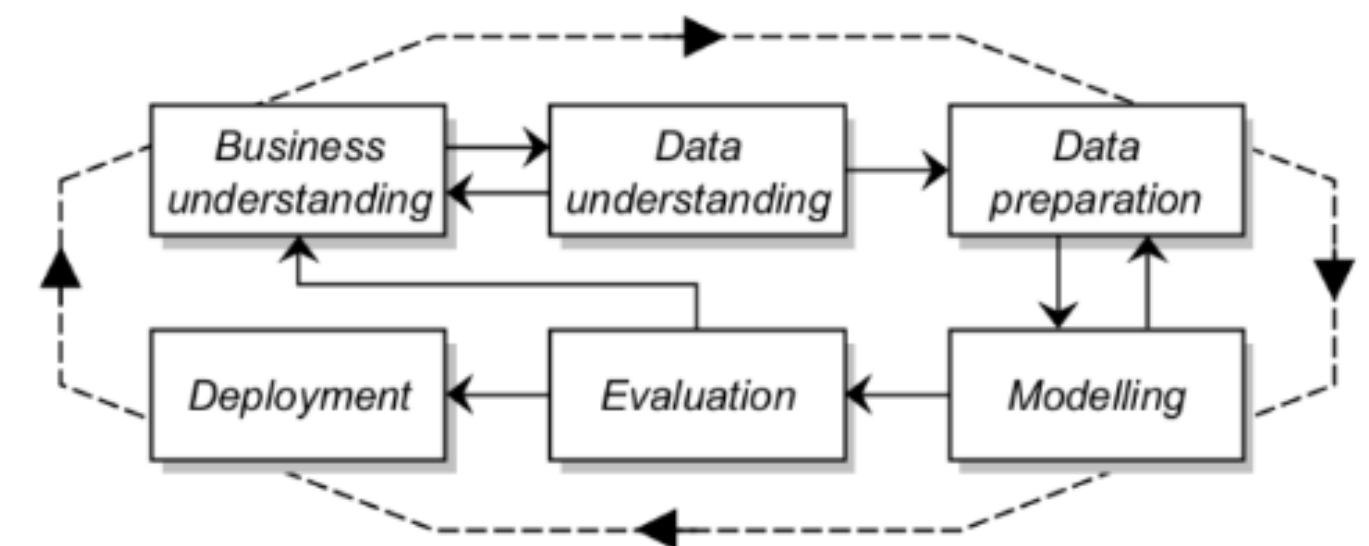


FIGURE 2.1 – Illustration des étapes CRISP-DM



Data Understanding



Notre data est constituée de trois fichiers : Afriware, jde_f03b11 et jde_f0911.

Il s'agit de transactions d'une société appartenant au groupe Akwa, enregistrées sur une période allant du début de l'année 2025 jusqu'au 14 juillet 2025.



Notre tâche consiste à identifier les transactions présentes dans le fichier Afriware mais absentes des autres fichiers via Machine Learning.





Data Understanding



Le fichier **afriware** regroupe les factures issues du système Afriware.

- **TypeFacture** : nature de la facture (par exemple, facture client, avoir).
- **NumeroFacture** : identifiant unique attribué à chaque facture dans Afriware.
- **NumeroLigne** : numéro de ligne associé à une facture dans Afriware. Contrairement au fichier `jde_f0911`, ce numéro correspond à l'entrée (ou ligne) d'une facture donnée, et non pas au nombre de lignes rattachées à un client.
- **CodeClient** : code identifiant le client destinataire de la facture.
- **MontantHT** : montant hors taxes facturé.
- **MontantTTC** : montant toutes taxes comprises.
- **Taxes** : montant des taxes appliquées à la facture.
- **DateCreation** : date initiale de création de la facture dans le système.
- **DateModification** : date de la dernière mise à jour de la facture.
- **DateEDI** : date de transfert de la facture vers le système JDE via EDI.
- **ReferenceEDI** : identifiant de référence du transfert EDI. Si ce champ est renseigné, la facture doit exister dans JDE ; s'il est vide ou nul, la facture n'a pas encore été transférée.
- **CentreAnalyse** : centre analytique rattaché à la facture.
- **CompteProduit** : compte produit associé à la ligne de facture.
- **DateFacture** : date d'encaissement ou de facturation effective de la facture.



Data Understanding



Le fichier **jde_f0911** correspond aux écritures du grand livre issues du système JDE.

- **GLKCO** : numéro de la société émettrice.
- **GLDCT** : type de facture (même rôle que TypeFacture dans Afriware).
- **GLDOC** : numéro de facture (équivalent à NumeroFacture).
- **GLDGJ** : jour de la facture exprimé en format julien, nécessitant une conversion.
- **GLJELN** : numéro de ligne associé à un client dans JDE. Contrairement à NumeroLigne d'Afriware (qui est le rang de la ligne au sein d'une facture), ce numéro indique le nombre de lignes rattachées au client et peut donc différer.
- **GLMCU** : centre analytique.
- **GLOBJ** : compte produit.
- **GLSUB** : sous-compte produit.
- **GLAA** : montant hors taxes.
- **GLU** : montant hors taxes restant à payer.
- **GLAN8** : code client.
- **GLCTRY** : siècle correspondant à la facture.
- **GLFY** : année fiscale de la facture.
- **GLPN** : mois de la facture.

Le fichier **jde_f03b11** contient des informations sur les factures et leur état de règlement

- **RPKCO** : numéro de la société.
- **RPDCT** : type de facture.
- **RPDOC** : numéro de facture.
- **RPSFX** : jour de la facture (en format julien, à convertir).
- **RPAG** : montant TTC.
- **RPAAP** : montant TTC restant à payer.
- **RPAN8** : code client.
- **RPCTY** : siècle associé à la facture.
- **RPFY** : année de la facture.
- **RPPN** : mois de la facture.



Data Understanding



TABLE 3.4 – Comptage des transactions présente dans le fichier Afriware et manquante dans le fichier JDE_311.

Label	Nombre
Données normales	432,826
Anomalies	49,846

TABLE 3.5 – Comptage des transactions présente dans le fichier Afriware et manquante dans le fichier JDE_911.

Label	Nombre
Données normales	432,826
Anomalies	49,846





Data Cleaning



Nous avons vérifié le jeu de données pour détecter d'éventuelles valeurs manquantes dans les enregistrements, et il n'y en avait aucune.

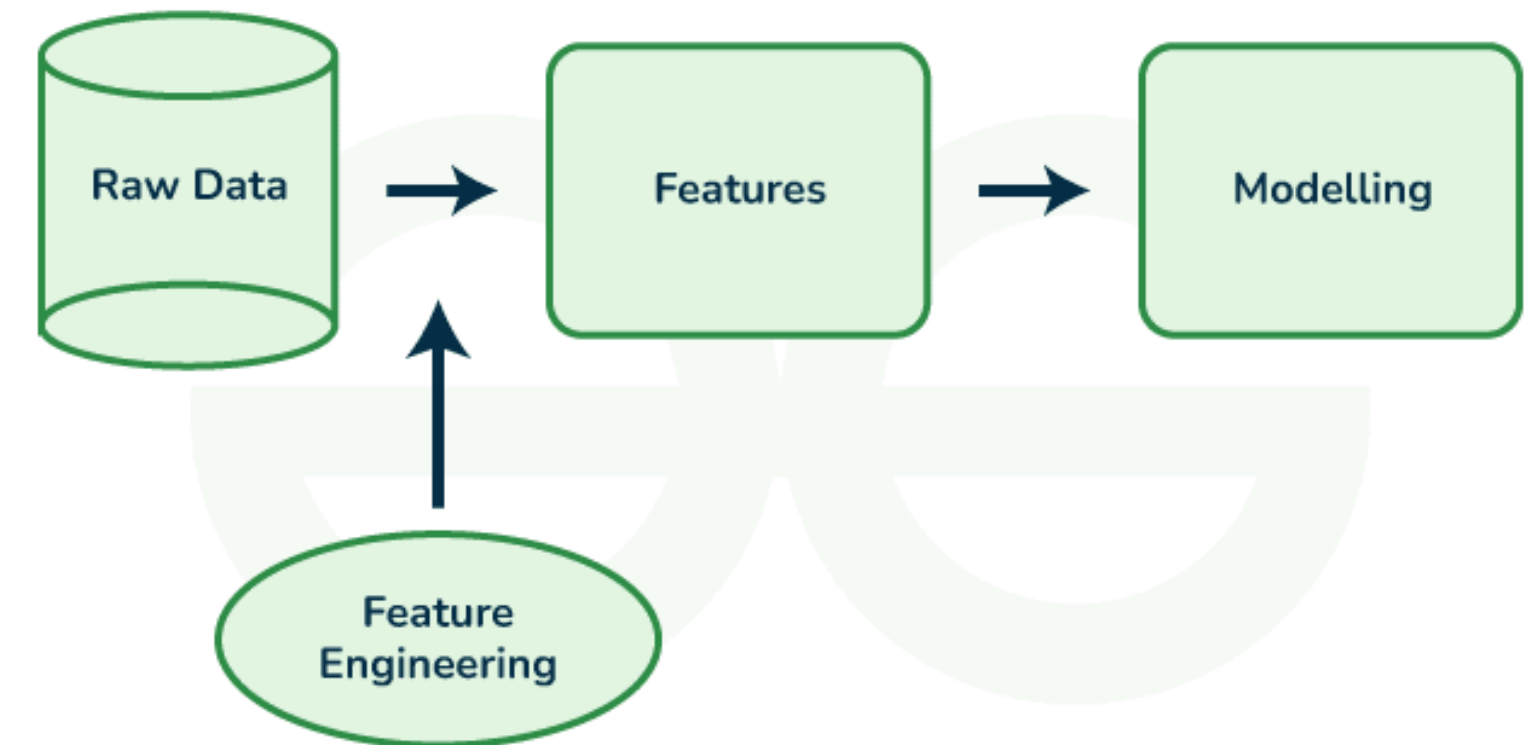


La suppression des doublons a été effectuée afin de garantir l'absence de lignes répétées dans les données

Feature Engineering

Dans le cadre de l'ingénierie des caractéristiques (feature engineering), certaines colonnes du jeu de données ont été supprimées afin de ne conserver que les variables pertinentes pour l'analyse.

Cette étape permet ainsi de simplifier le modèle et de se concentrer sur les caractéristiques réellement informatives, telles que les montants ou les codes analytiques.



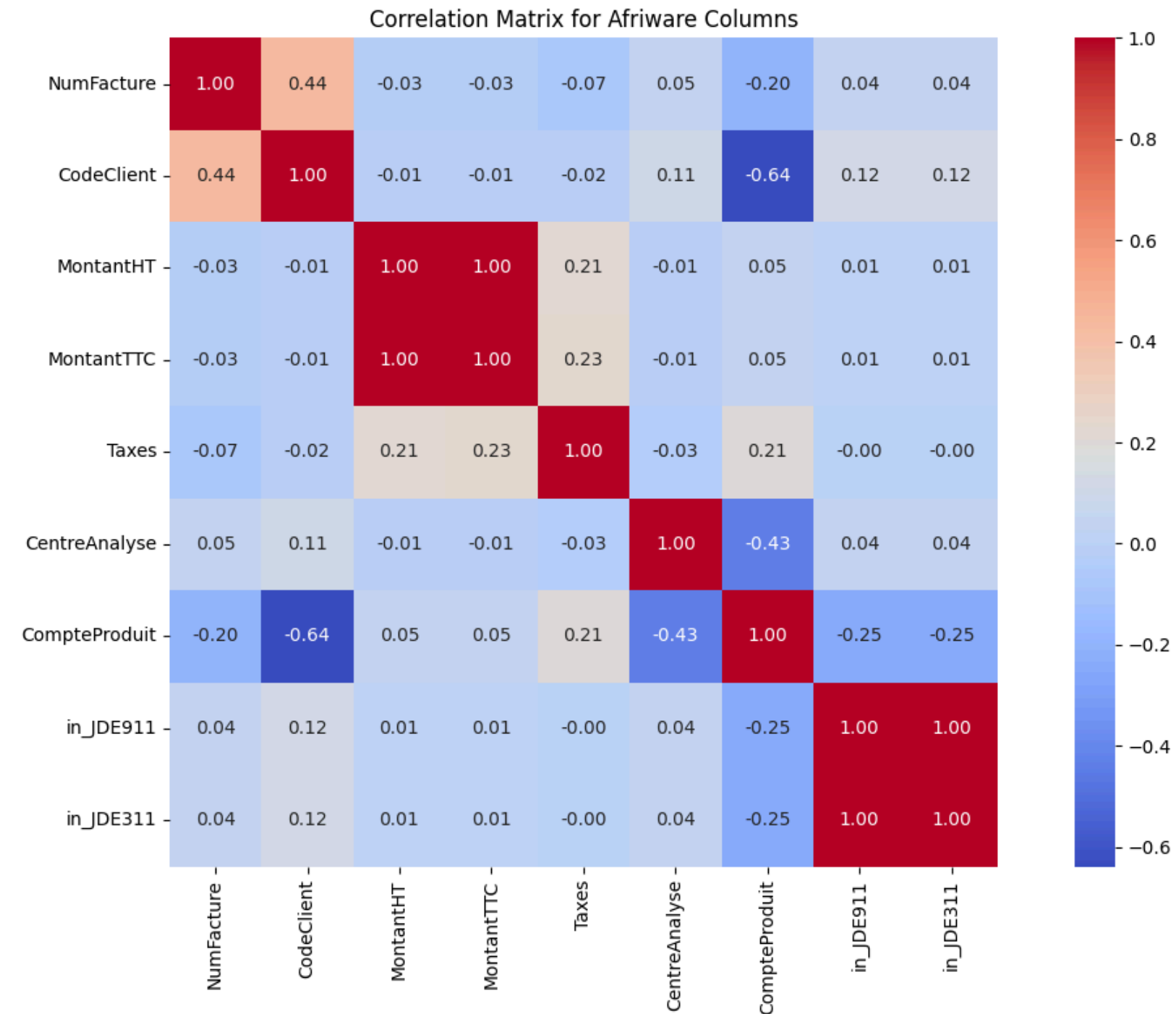
- **TypeFacture** : nature de la facture (par exemple, facture client, avoir).
- **NumeroFacture** : identifiant unique attribué à chaque facture dans Afriware.
- **NumeroLigne** : numéro de ligne associé à une facture dans Afriware. Contrairement au fichier `jde_f0911`, ce numéro correspond à l'entrée (ou ligne) d'une facture donnée, et non pas au nombre de lignes rattachées à un client.
- **CodeClient** : code identifiant le client destinataire de la facture.
- **MontantHT** : montant hors taxes facturé.
- **MontantTTC** : montant toutes taxes comprises.
- **Taxes** : montant des taxes appliquées à la facture.
- **DateCreation** : date initiale de création de la facture dans le système.
- **DateModification** : date de la dernière mise à jour de la facture.
- **DateEDI** : date de transfert de la facture vers le système JDE via EDI.
- **ReferenceEDI** : identifiant de référence du transfert EDI. Si ce champ est renseigné, la facture doit exister dans JDE ; s'il est vide ou nul, la facture n'a pas encore été transférée.
- **CentreAnalyse** : centre analytique rattaché à la facture.
- **CompteProduit** : compte produit associé à la ligne de facture.
- **DateFacture** : date d'encaissement ou de facturation effective de la facture.

Feature Selection

- **NumeroFacture** : identifiant unique attribué à chaque facture dans Afriware.
- **CodeClient** : code identifiant le client destinataire de la facture.
- **CentreAnalyse** : centre analytique rattaché à la facture.
- **CompteProduit** : compte produit associé à la ligne de facture.



Data Preprocessing





Data Preprocessing



```

1 GQ,22393510,2,1522,7514110.8,8265521.88,751411.08,12/31/2024 13:15,1/3/2025 7:18,1/15/2025 14:03,EDI,1198.0,711103.1105,1/15/2025 0:00
2 GQ,22393510,1,1522,1560537.9,1716591.69,156053.79,12/31/2024 13:15,1/3/2025 7:18,1/15/2025 14:03,EDI,1198.0,711100.1101,1/15/2025 0:00
3 GQ,22393511,1,1522,1560537.9,1716591.69,156053.79,1/2/2025 8:58,1/3/2025 6:12,1/15/2025 14:03,EDI,1198.0,711100.1101,1/15/2025 0:00
4 GQ,22393511,2,1522,7514110.8,8265521.88,751411.08,1/2/2025 8:58,1/3/2025 6:12,1/15/2025 14:03,EDI,1198.0,711103.1105,1/15/2025 0:00
5 GQ,22393512,2,1522,6977388.6,7675127.46,697738.86,1/2/2025 14:17,1/4/2025 3:26,1/15/2025 14:03,EDI,1198.0,711103.1105,1/15/2025 0:00
6 GQ,22393512,1,1522,2184753.06,2403228.42,218475.36,1/2/2025 14:17,1/4/2025 3:26,1/15/2025 14:03,EDI,1198.0,711100.1101,1/15/2025 0:00
7 GQ,22393513,1,1522,2223041.76,2445345.99,222304.23,1/3/2025 13:32,1/6/2025 6:48,1/15/2025 14:03,EDI,1198.0,711104.1105,1/15/2025 0:00
8 GQ,22393514,1,1522,1560537.9,1716591.69,156053.79,1/3/2025 13:32,1/6/2025 6:48,1/15/2025 14:03,EDI,1198.0,711100.1101,1/15/2025 0:00
9 GQ,22393514,2,1522,5367222.0,5903944.2,536722.2,1/3/2025 13:32,1/6/2025 6:48,1/15/2025 14:03,EDI,1198.0,711103.1105,1/15/2025 0:00
10 GQ,22393515,1,1522,1560537.9,1716591.69,156053.79,1/4/2025 17:37,1/7/2025 10:52,1/15/2025 14:03,EDI,1198.0,711100.1101,1/15/2025 0:00
11 GQ,22393515,2,1522,7514110.8,8265521.88,751411.08,1/4/2025 17:37,1/7/2025 10:52,1/15/2025 14:03,EDI,1198.0,711103.1105,1/15/2025 0:00
12 GQ,22393516,2,1523,3729563.46,4102519.86,372956.4,12/31/2024 13:23,1/1/2025 18:11,1/15/2025 14:03,EDI,1124.0,711103.1105,1/15/2025 0:00
13 GQ,22393516,1,1523,620287.74,682316.46,62028.72,12/31/2024 13:23,1/1/2025 18:11,1/15/2025 14:03,EDI,1124.0,711100.1101,1/15/2025 0:00
14 GQ,22393517,2,1523,3995960.85,4395556.8,399595.95,1/1/2025 12:39,1/2/2025 13:12,1/15/2025 14:03,EDI,1124.0,711103.1105,1/15/2025 0:00
15 GQ,22393517,1,1523,930431.61,1023474.69,93043.08,1/1/2025 12:39,1/2/2025 13:12,1/15/2025 14:03,EDI,1124.0,711100.1101,1/15/2025 0:00

```

Data Aggregation

```

GQ,22393510,9074648.7,9982113.57,907464.87,1/15/2025 0:00,12/31/2024 13:15,1/3/2025 7:18,1/15/2025 14:03,EDI,1522,1198.0,711100.1101
GQ,22393511,9074648.7,9982113.57,907464.87,1/15/2025 0:00,1/2/2025 8:58,1/3/2025 6:12,1/15/2025 14:03,EDI,1522,1198.0,711100.1101
GQ,22393512,9162141.66,10078355.88,916214.22,1/15/2025 0:00,1/2/2025 14:17,1/4/2025 3:26,1/15/2025 14:03,EDI,1522,1198.0,711100.1101
GQ,22393513,2223041.76,2445345.99,222304.23,1/15/2025 0:00,1/3/2025 13:32,1/6/2025 6:48,1/15/2025 14:03,EDI,1522,1198.0,711104.1105
GQ,22393514,6927759.9,7620535.89,692775.99,1/15/2025 0:00,1/3/2025 13:32,1/6/2025 6:48,1/15/2025 14:03,EDI,1522,1198.0,711100.1101
GQ,22393515,9074648.7,9982113.57,907464.87,1/15/2025 0:00,1/4/2025 17:37,1/7/2025 10:52,1/15/2025 14:03,EDI,1522,1198.0,711100.1101
GQ,22393516,4349851.2,4784836.32,434985.12,1/15/2025 0:00,12/31/2024 13:23,1/1/2025 18:11,1/15/2025 14:03,EDI,1523,1124.0,711100.1101
GQ,22393517,4926392.46,5419031.49,492639.03,1/15/2025 0:00,1/1/2025 12:39,1/2/2025 13:12,1/15/2025 14:03,EDI,1523,1124.0,711100.1101

```



Train-Test Data Split

Nous avons divisé notre data en deux parties :

- 80 % pour l'ensemble d'entraînement,
- 20 % pour l'ensemble de test.

Les variables d'entrée (X) correspondent au fichier Afriware, tandis que la variable cible (Y) correspond à la colonne in_JDE311 (resp. JDE911).





Models Performance Measure

		Predicted class	
		Non-anomaly (0)	Anomaly (1)
Actual class	Non-anomaly (0)	<i>True Negative (TN)</i>	<i>False Positive (FP)</i>
	Anomaly (1)	<i>False Negative (FN)</i>	<i>True Positive (TP)</i>



Models Performance Measure

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Recall_1 / Specificity = \frac{TP}{TP + FN}$$

$$Recall_0 / Sensitivity = \frac{TN}{TN + FP}$$

$$Precision_1 = \frac{TP}{TP + FP}$$

$$Precision_0 = \frac{TN}{TN + FN}$$

$$F1 - score_1 = \frac{2 * Recall_1 * Precision_1}{Recall_1 + Precision_1}$$

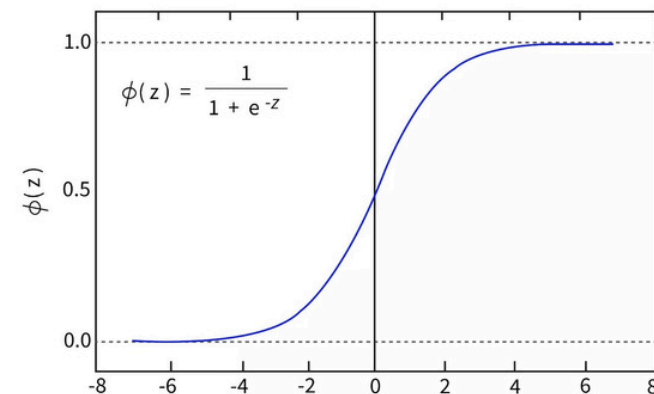
$$F1 - score_0 = \frac{2 * Recall_0 * Precision_0}{Recall_0 + Precision_0}$$

$$Metric_avg_{macro} = \frac{Metric_0 + Metric_1}{2}$$

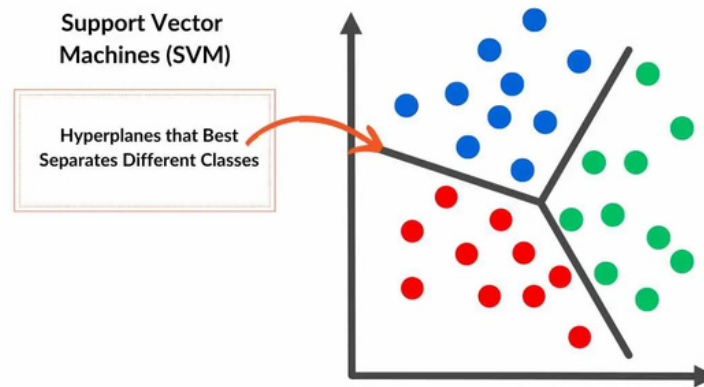
$$Metric_avg_{weighted} = \frac{Metric_0 * Support_0 + Metric_1 * Support_1}{Support_0 + Support_1}$$

Nous avons entraîné cinq modèles d'apprentissage automatique supervisés ,
en utilisant un algorithme différent pour chaque modèle, afin de détecter les
valeurs aberrantes dans l'ensemble de données prétraité.

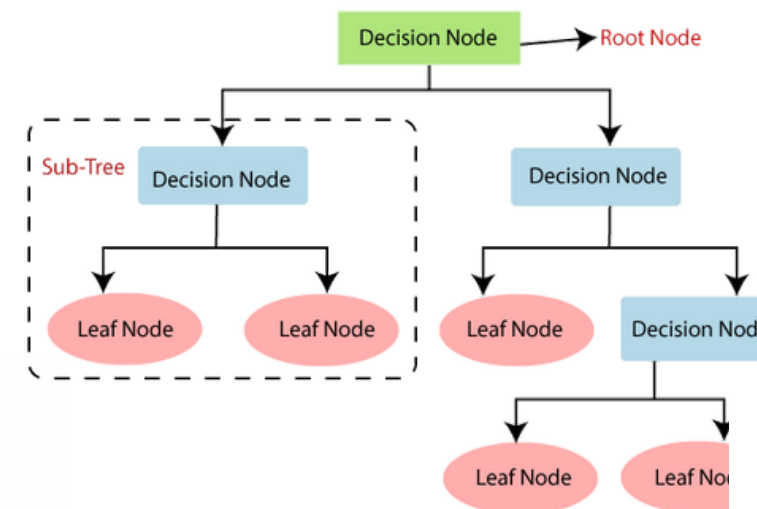
Logistic Regression



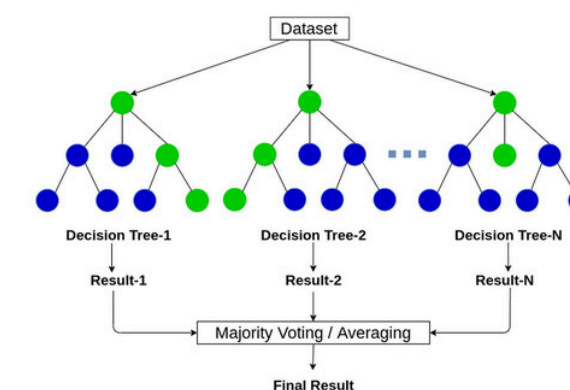
SVM



Decision Tree



Random Forest



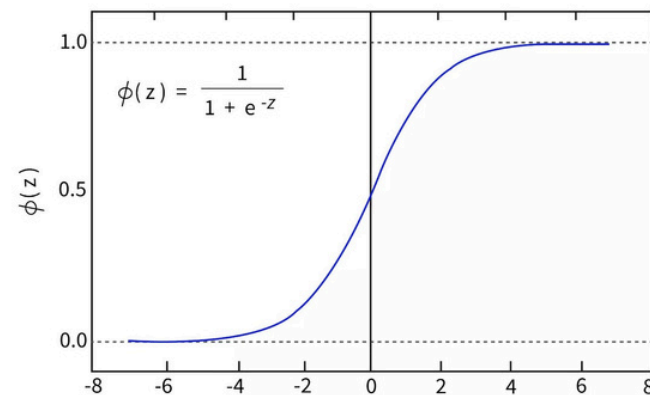
KNN





Supervised Machine Learning Modeling

Logistic Regression



$$P(y = 1 | x) = \frac{e^{a+bx}}{1 + e^{a+bx}} = \frac{1}{1 + e^{-(a+bx)}}$$

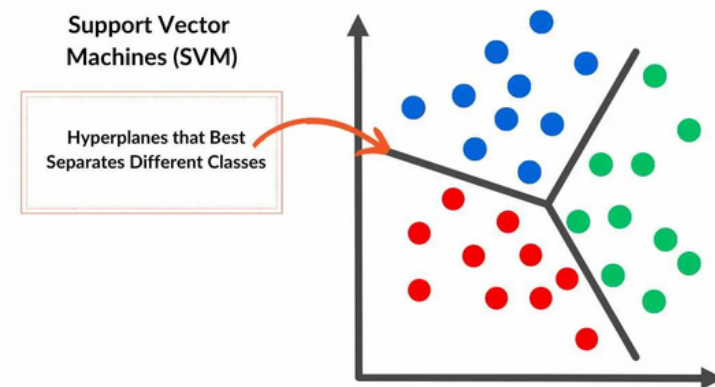
TABLE 3.8 – Rapport de classification du modèle de régression logistique.

Classe	Précision	Rappel	F1-Score	Support
Présente (0)	0,97	0,67	0,79	29 034
Manquante (1)	0,19	0,80	0,31	2 815
Exactitude (Accuracy)			0,68	31 849
Moyenne macro	0,58	0,73	0,55	31 849
Moyenne pondérée	0,90	0,68	0,75	31 849



Supervised Machine Learning Modeling

SVM



$$F(x) = \sum_{i=1}^N \alpha_i y_i k(x_i, x) + b$$

TABLE 3.9 – Rapport de classification du modèle SVM.

Classe	Précision	Rappel	F1-Score	Support
Présente (0)	0.94	0.77	0.85	29034
Manquante (1)	0.17	0.49	0.26	2815
Exactitude (Accuracy)			0.75	31849
Moyenne macro	0.56	0.63	0.55	31849
Moyenne pondérée	0.87	0.75	0.80	31849



Supervised Machine Learning Modeling

Decision Tree

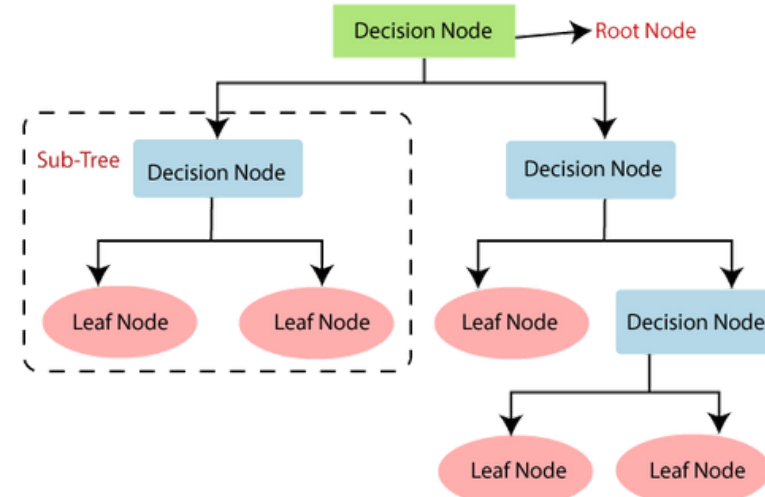


TABLE 3.10 – Rapport de classification du modèle Decision Tree.

Classe	Précision	Rappel	F1-score	Support
Présente (0)	0.98	0.98	0.98	29020
Manquante (1)	0.82	0.82	0.82	2829
Accuracy			0.97	31849
Macro avg	0.90	0.90	0.90	31849
Weighted avg	0.97	0.97	0.97	31849



Supervised Machine Learning Modeling

Random Forest

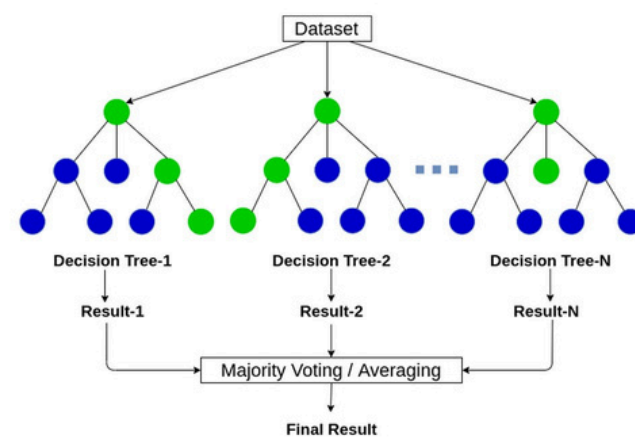


TABLE 3.11 – Rapport de classification du modèle de Forêt Aléatoire.

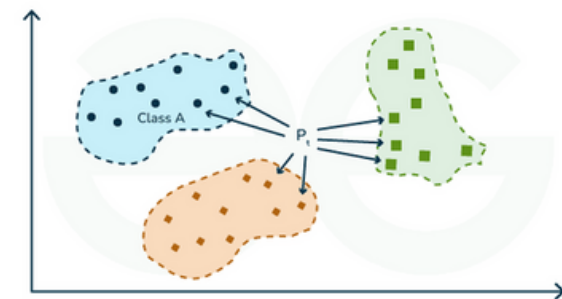
Classe	Précision	Rappel	F1-Score	Support
Présente (0)	0.98	0.99	0.99	29 020
Manquante (1)	0.93	0.78	0.85	2 829
Exactitude (Accuracy)			0.98	31 849
Moyenne macro	0.95	0.89	0.92	31 849
Moyenne pondérée	0.97	0.98	0.97	31 849



Supervised Machine Learning Modeling

KNN

K Nearest Neighbors



K Nearest Neighbors



$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

TABLE 3.12 – Rapport de classification du modèle KNN.

Classe	Précision	Rappel	F1-Score	Support
Présente (0)	0.97	0.97	0.97	29034
Manquante (1)	0.72	0.71	0.71	2815
Exactitude (Accuracy)			0.95	31849
Moyenne macro	0.84	0.84	0.84	31849
Moyenne pondérée	0.95	0.95	0.95	31849



Supervised Machine Learning Evaluation

Le tableau montre que les modèles basés sur des arbres, en particulier le Decision Tree et le Random Forest, obtiennent les meilleures performances en termes de rappel moyen macro, respectivement 0,90 et 0,885, ce qui indique leur efficacité pour détecter les transactions manquantes.

TABLE 3.13 – Comparaison des performances des modèles supervisés appliqués à la détection des transactions manquantes.

No.	Algorithme	TN	FN	FP	TP	Recall Avg Macro
1	Logistic Regression	19453	9581	563	2252	0.735
2	Support Vector Machines	22364	6670	1463	1379	0.63
3	Decision Tree	28440	580	509	2320	0.90
4	Random Forest	28730	290	622	2207	0.885
5	K-Nearest Neighbour	28163	871	817	1998	0.84



Perspectives

Intégration des modèles non supervisés : ajustement des seuils et optimisation des paramètres pour une détection plus efficace..

Intégration du Deep Learning : apprentissage de patterns complexes pour repérer des anomalies plus subtiles.

Extension du cadre de détection : inclusion d'autres types de fraudes (TVA, doublons, écritures fictives, etc.).

Combinaison de modèles : approche hybride supervisée/non supervisée pour plus de robustesse.

Amélioration de la qualité des données : standardisation et normalisation des fichiers pour renforcer la précision.



Conclusion

De la compréhension du besoin métier jusqu'à l'évaluation du modèle, ce projet m'a permis d'explorer l'ensemble du cycle de vie des données dans le contexte de l'audit intelligent. J'ai développé des compétences solides dans les domaines du data engineering (collecte et préparation des données), du data analysis (analyse et visualisation), et du data science (modélisation et apprentissage automatique).



MERCI

Pour Votre Attention

