Détection d'Anomalies à l'aide d'Arbres de Décision Flous

18 JUIN 2021

Présenté par :
Mohamed Taieb SLAMA
Karem SFAR GANDOURA
Mohamed Iheb LANDOULSI

Encadré par :
Mme. Mariem CHATER

PLAN DE LA PRÉSENTATION

Introduction

Implémentation Isolation Forest

Adaptation aux Données Floues

Interprétation

Conclusion et Perspectives

INTRODUCTION

DÉTECTION D'ANOMALIES

ISOLATION FOREST

FUZZY LOGIC

PROBLÈMATIQUE

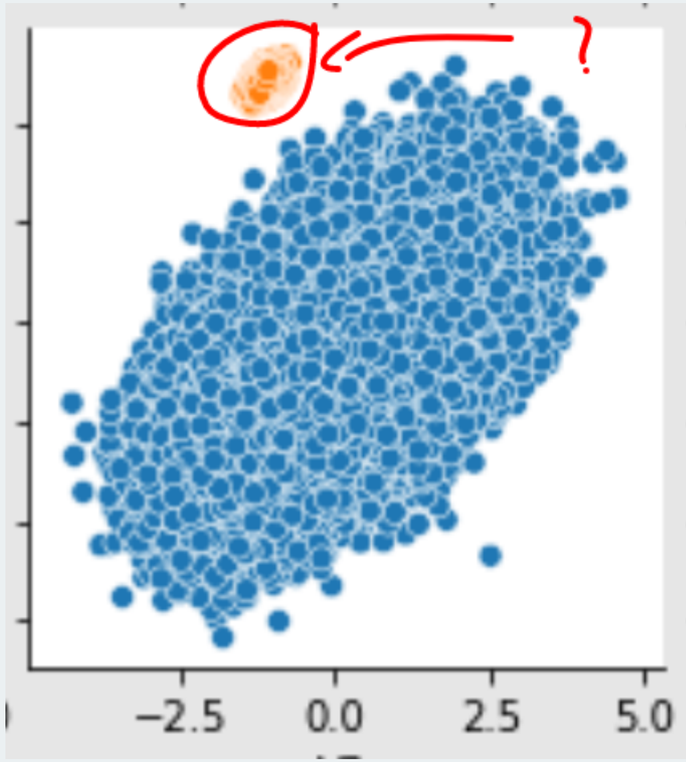
Détection d'anomalie

Identification les points qui dérivent du comportement normal de la base de données

Détection d'anomalies vs Classification

Supervisée vs Non supervisée

Isolation Forest ?



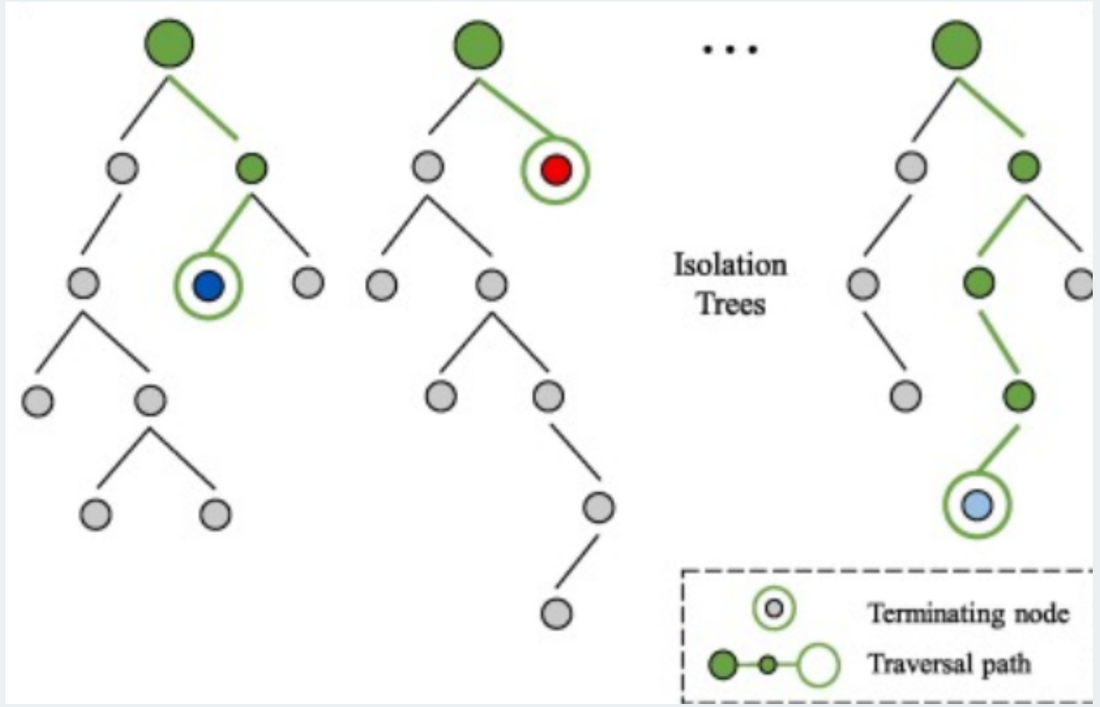
Isolation Forest

Fei Tony Lui & Zhi-Hua Zhou - 2008

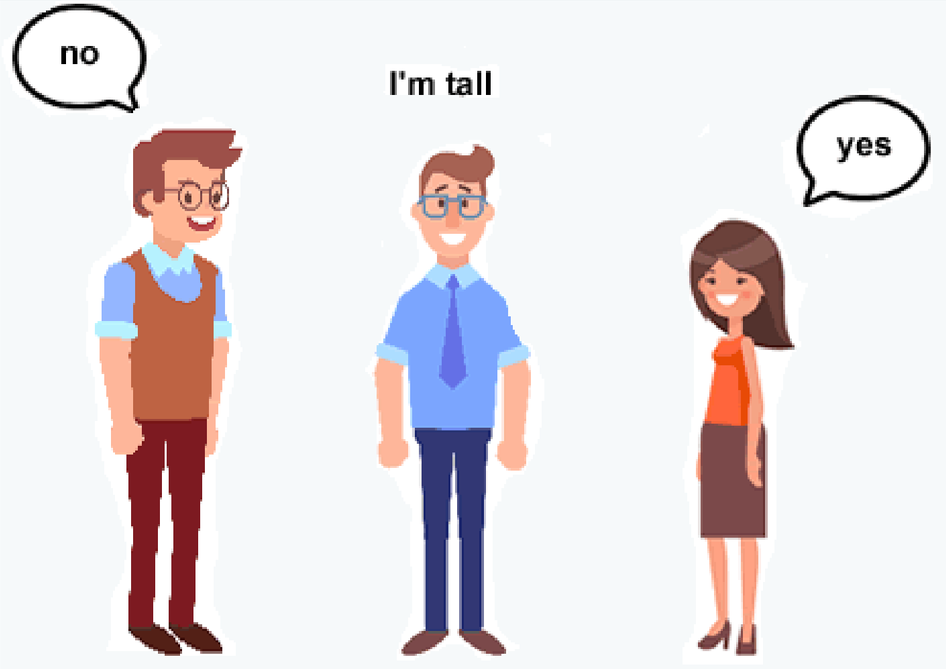
Principe Totalement Différent se basant sur l'isolation des éléments plutôt que le profiling des données normales

Rapide

Faible coût en mémoire



LA LOGIQUE FLOUE

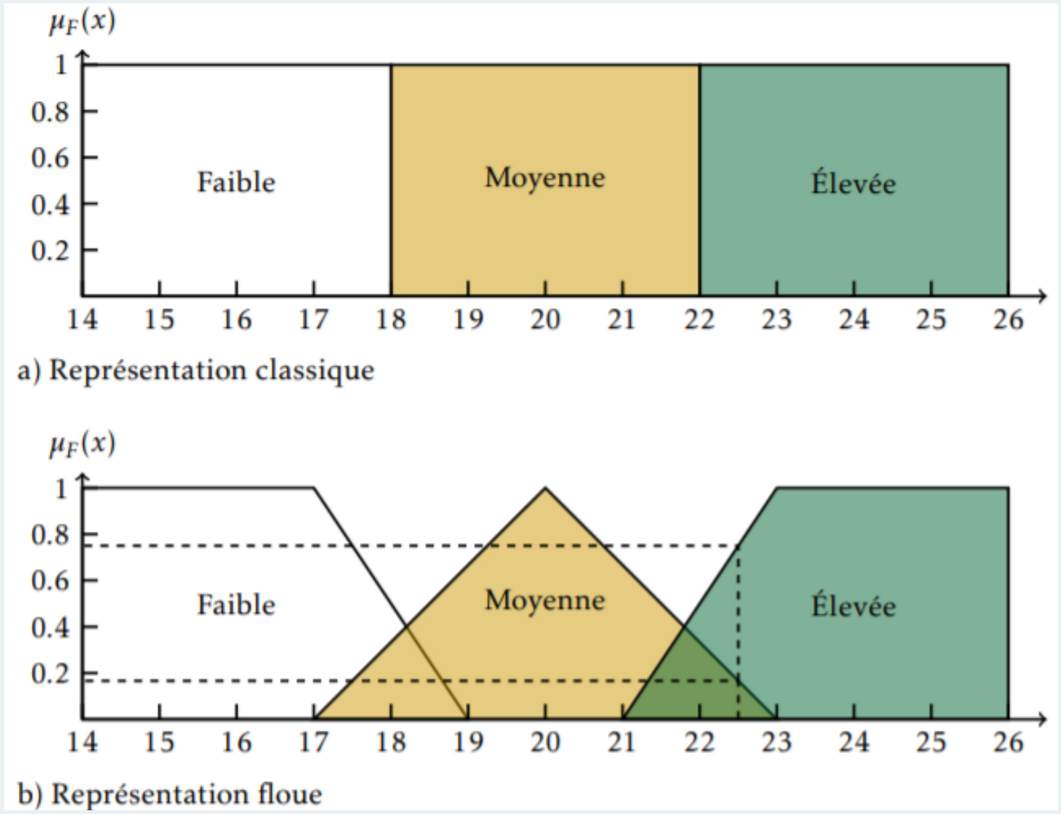


LOGIQUE CLASSIQUE **VS** LOGIQUE FLOUE

Logique Booléenne (0 ou 1)	Raisonnement proche du cerveau humain
Chaque variable (affirmation) est soit vraie soit fausse	Valeur de vérité de la variable entre 0 (faux) et 1 (vrai) => Vérité Partielle

FONCTION D'APPARTENANCE

Permet le passage vers la logique floue



Exemple: Pour une valeur de 22.5, la variable appartient à 75% à la classe Elevée, 16.7% à la classe Moyenne et 0% à la classe Faible.

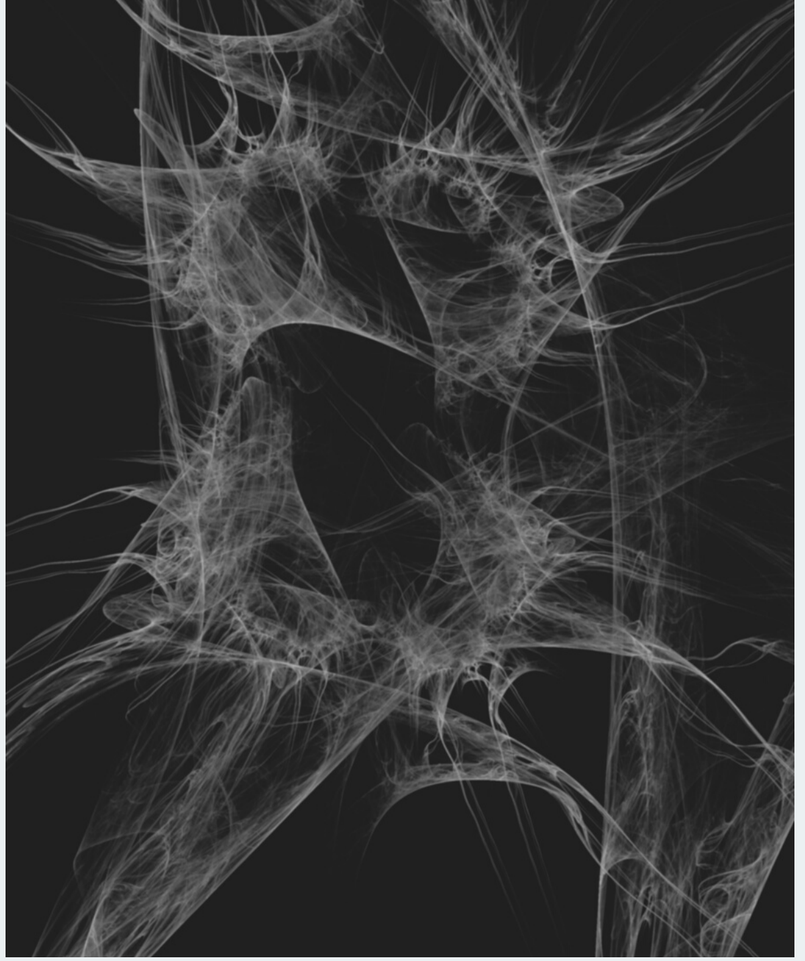
La Logique Floue & Le Machine Learning

RANDOM FORESTS & DECISION TREES

L'incertitude et l'utilisation d'un langage proche de l'humain ont contribué à la réalisation de plusieurs algorithme de Machine Learning avec des arbres de décisions flous.

DEEP LEARNING

Plusieurs algorithmes de réseaux de neurones se basant sur la logique floue ont été explorés.
Exemple: Pour le Load Forecasting (Prévision de charge énergétique)



Serait-il intéressant d'essayer d'adapter
l'algorithme d'isolation forest à la logique floue ?

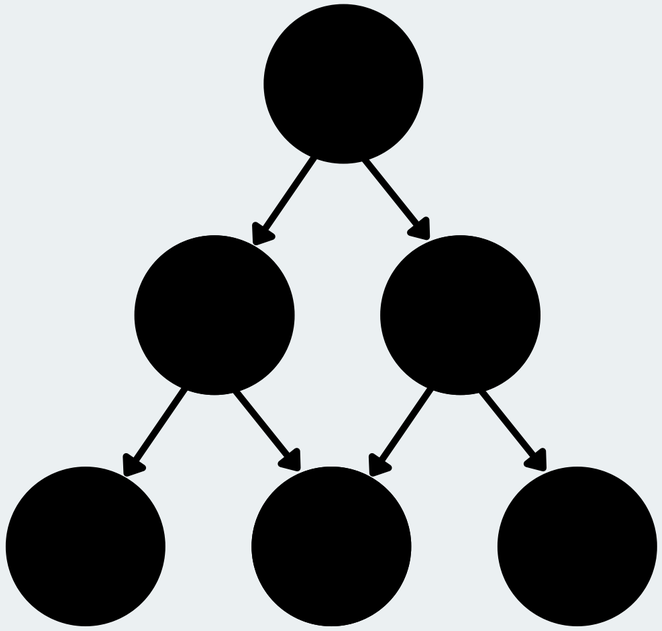
PROBLÈMATIQUE

ISOLATION FOREST

ALGORITHME ET FONCTIONNEMENT

IMPLEMENTATION / REALISATION

ALGORITHME ET FONCTIONNEMENT: APPRENTISSAGE



La phase d'apprentissage est similaire à l'apprentissage des arbres de décision et se présente comme suit:

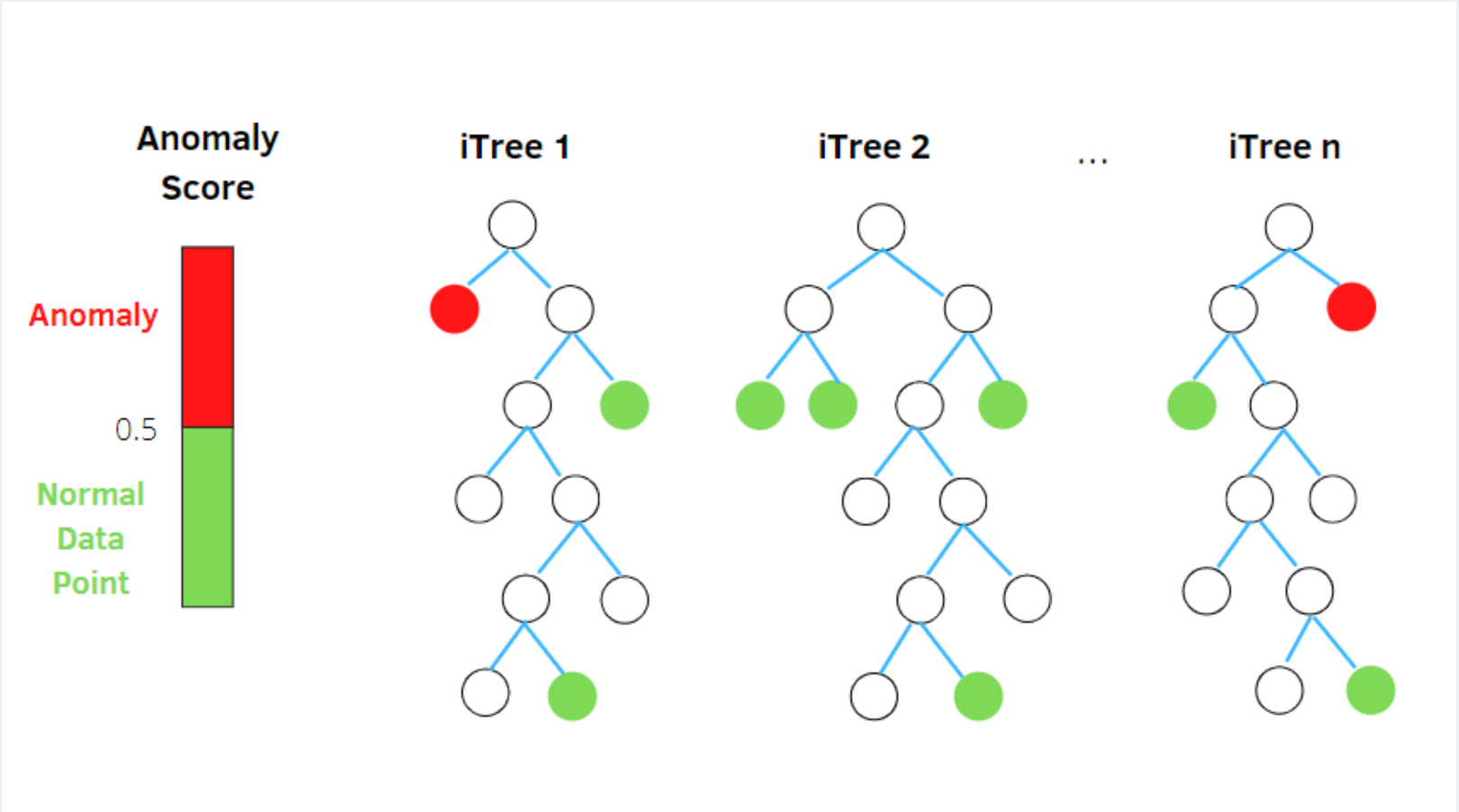
- Le noeud **racine contient tous les éléments**
- Pour chaque noeud nous choisissons **aléatoirement** un attribut **a** et une valeur de split **s**
- x[a] > s => fils 1 sinon fils 2

L'Isolation Forest est un algorithme ensembliste qui apprend en créant un nombre t donné d'arbres.

ALGORITHME ET FONCTIONNEMENT: ÉVALUATION

Anomaly score :
calculé en fonction de la profondeur d'un élément dans l'arbre !

Le score final étant la moyenne des scores de tous les arbres



UNE IMPLEMENTATION EN LOGIQUE CLASSIQUE ?

+ Création d'un algorithme modulaire où la migration vers la logique floue engendrera un minimum de points de doutes

RÉSULTATS: COMPARATIF AVEC L'ETAT DE L'ART

IMPLEMENTATION

MULCROSS DATASET:
-> 0.9471
HEART DISEASE:
-> 0.649
CREDIT CARDS:
-> 0.9475

SCI-KIT LEARN

MULCROSS DATASET:
-> 0.9493
HEART DISEASE:
-> 0.67655
CREDIT CARDS:
-> 0.9507

H2O

MULCROSS DATASET:
-> 0.9292
HEART DISEASE:
-> 0.5339
CREDIT CARDS:
-> 0.9343

Comparaison du Score AUC ROC de différentes implémentation d' Isolation Forest

ADAPTATION AUX DONNÉES FLOUES

MODULE DE FUZZIFICATION

PARTITIONNEMENT FLOU: ALPHA-COUPES

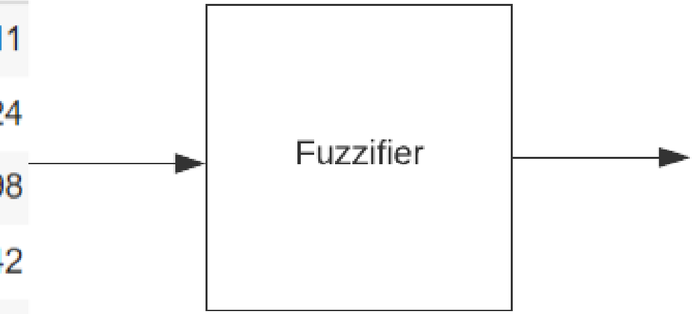
PARTITIONNEMENT FLOU: JANIKOW

DIFFERENCES AVEC LES DECISIONS TREES

MODULE DE FUZZIFICATION

Faute de présence de bases de données floues sur internet, nous avons dû implémenter notre propre module de fuzzification

V1	V2
-0.203950	0.363011
-0.761118	2.436424
-0.209979	1.131098
0.836812	0.650342
0.454204	1.560128



	V1;poor	V1;average	V1;good	V2;poor	V2;average	V2;good
X0	0.397257	0.602743	0.000000	0.000000	0.769698	0.230302
X1	0.000000	0.758485	0.241515	0.574197	0.425803	0.000000
X2	0.000000	0.963961	0.036039	0.000000	0.552493	0.447507
X3	0.000000	0.690766	0.309234	0.000000	0.655924	0.344076

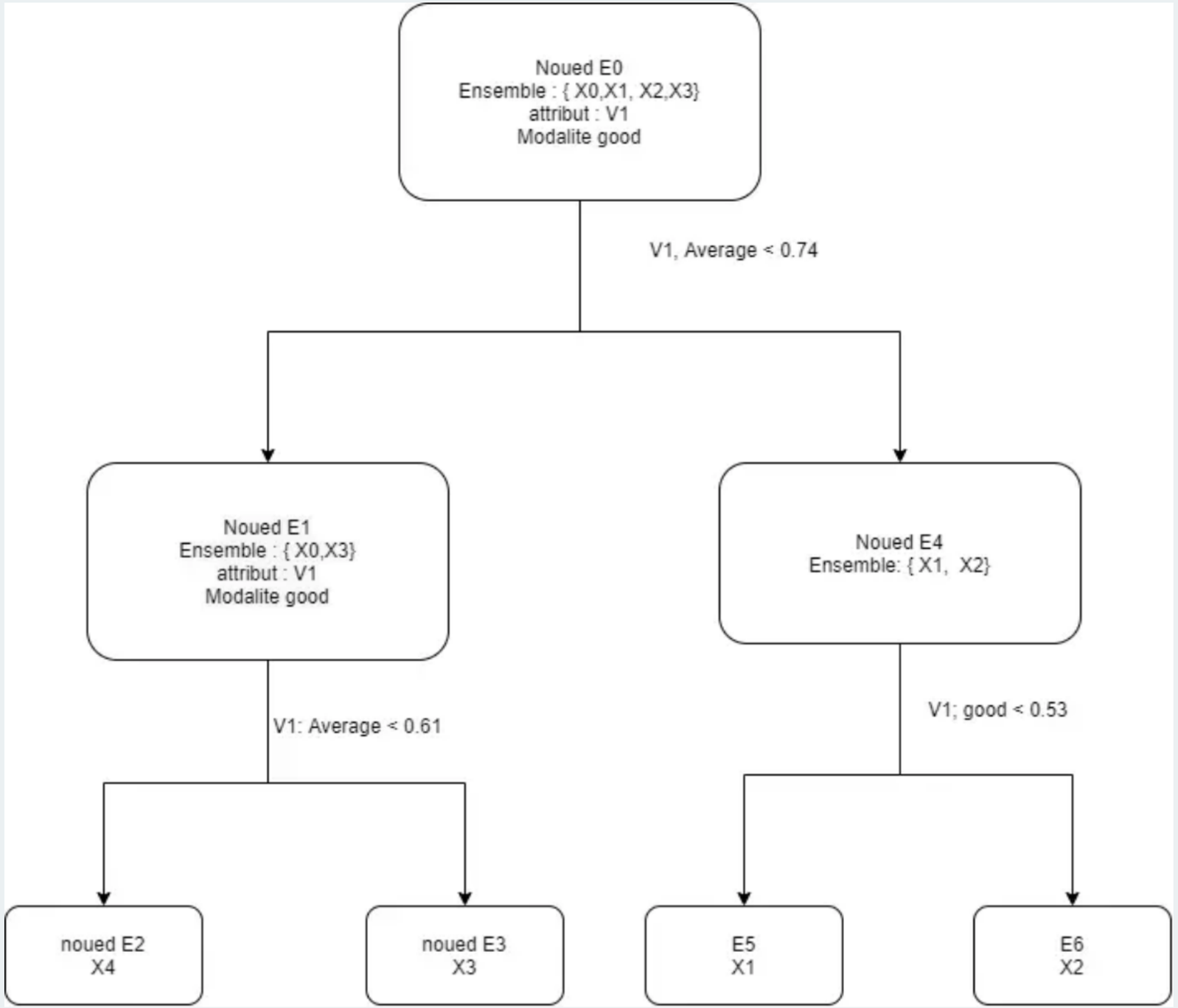
UN PROBLÈME DE PARTITIONNEMENT ?

+ Le changement fondamentale apporté par la logique floue est qu'un élément e appartient avec degré à un ensemble

+ Cela sucite le problème de partitionnement d'un esemble flou, qui va où ?

LES PARTITIONNEMENT ALPHA-COUPE

SOLUTION INTUITIVE, MAIS À QUEL PRIX?


Noued E0
Ensemble : { X0,X1, X2,X3}
attribut : V1
Modalite good
V1, Average < 0.74
Noued E1
Ensemble : { X0,X3}
attribut : V1
Modalite good
Noued E4
Ensemble: { X1, X2}
V1: Average < 0.61
V1; good < 0.53
noued E2
X4
noued E3
X3
E5
X1
E6
X2

PARTITIONNEMENT ALPHA COUTE

INCONVÉNIENT

+ Retour à la logique classique; nous avons des éléments d'origine floues qui

- **soit appartiennent à la partition**
- **soit n'y appartiennent pas.**

LES PARTITIONNEMENTS JANIKOW

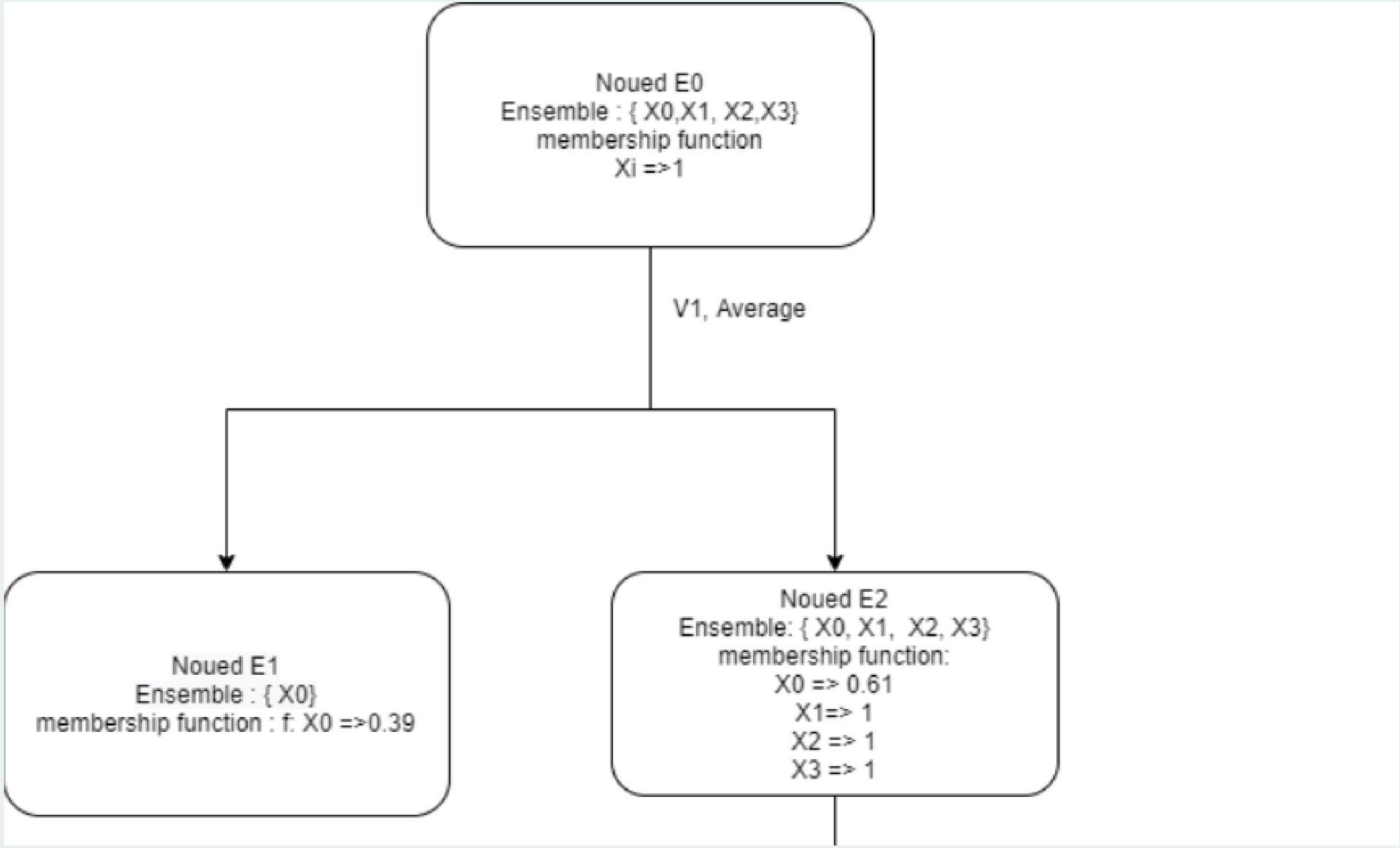
Chaque élement appartient à toutes les partitions, mais à des degrés d'appartenances différents

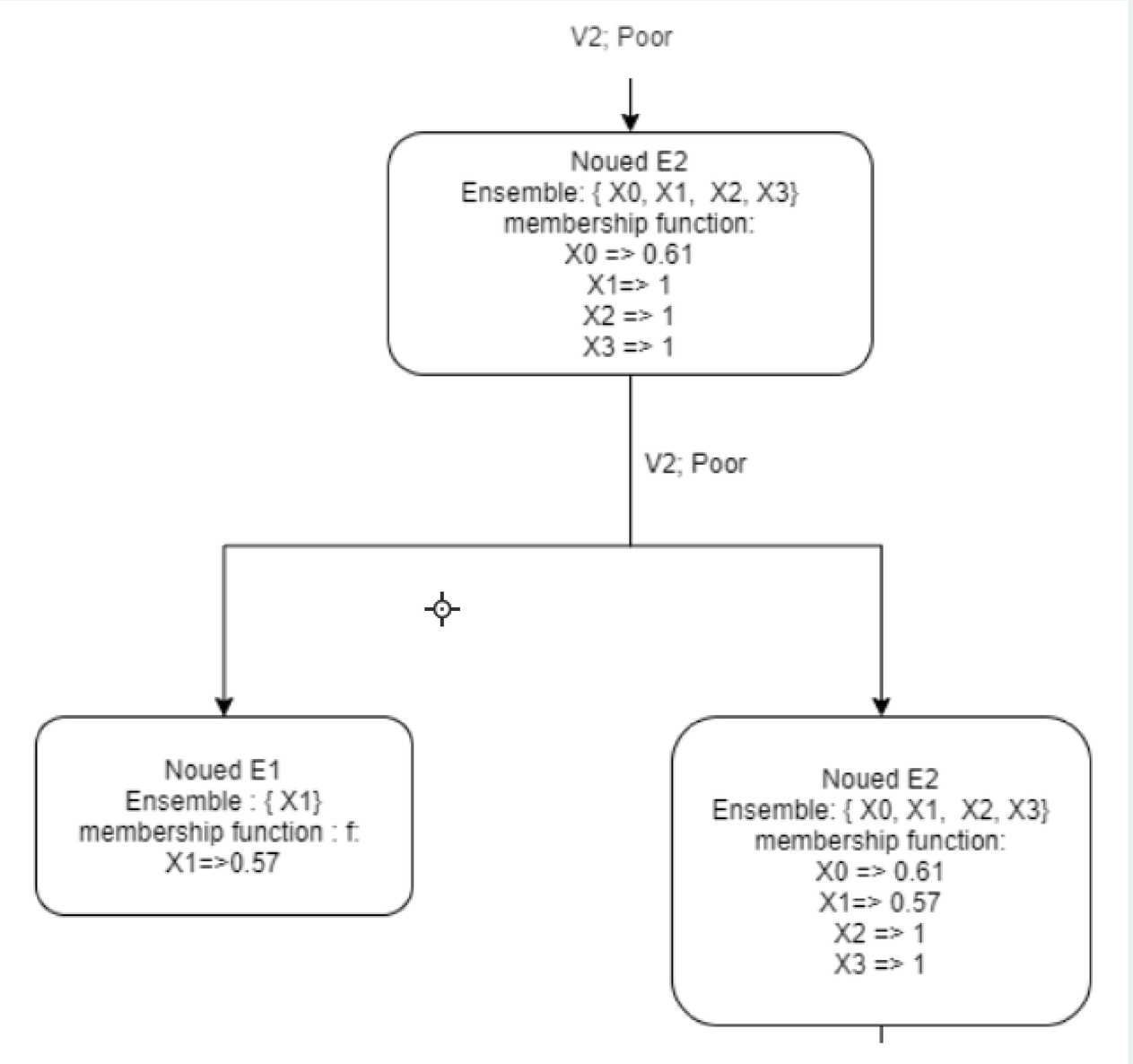
-> **Calcul d'appartenance au sous ensemble**

En choisissant un attribut et une modalité selon la quelle faire le split, on calcul l'apparatenance d'un element e
au sous-ensemble grace à :

-Son **appartenance** à l'ensemble à partitionner

-Sa function **d'appartenance** à la **modalité** considérée


Noued E0
Ensemble : { X0,X1, X2,X3}
membership function
Xi =>1
V1, Average
Noued E1
Ensemble : { X0}
membership function : f: X0 =>0.39
Noued E2
Ensemble: { X0, X1, X2, X3}
membership function:
X0 => 0.61
X1=> 1
X2 => 1
X3 => 1


V2; Poor
Noued E2
Ensemble: { X0, X1, X2, X3}
membership function:
X0 => 0.61
X1=> 1
X2 => 1
X3 => 1
V2; Poor
Noued E1
Ensemble : { X1}
membership function : f:
X1=>0.57
Noued E2
Ensemble: { X0, X1, X2, X3}
membership function:
X0 => 0.61
X1=> 0.57
X2 => 1
X3 => 1

INTÉRPRÉTATION

DATASETS UTILISÉS

MESURES UTILISÉES

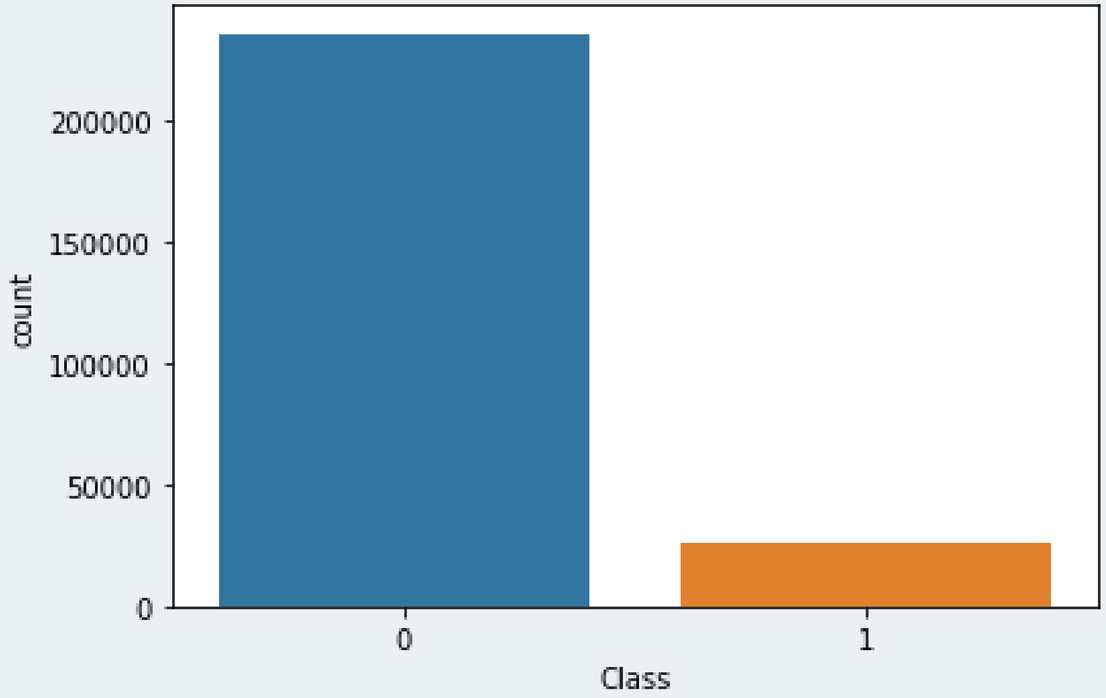
RÉSULTATS

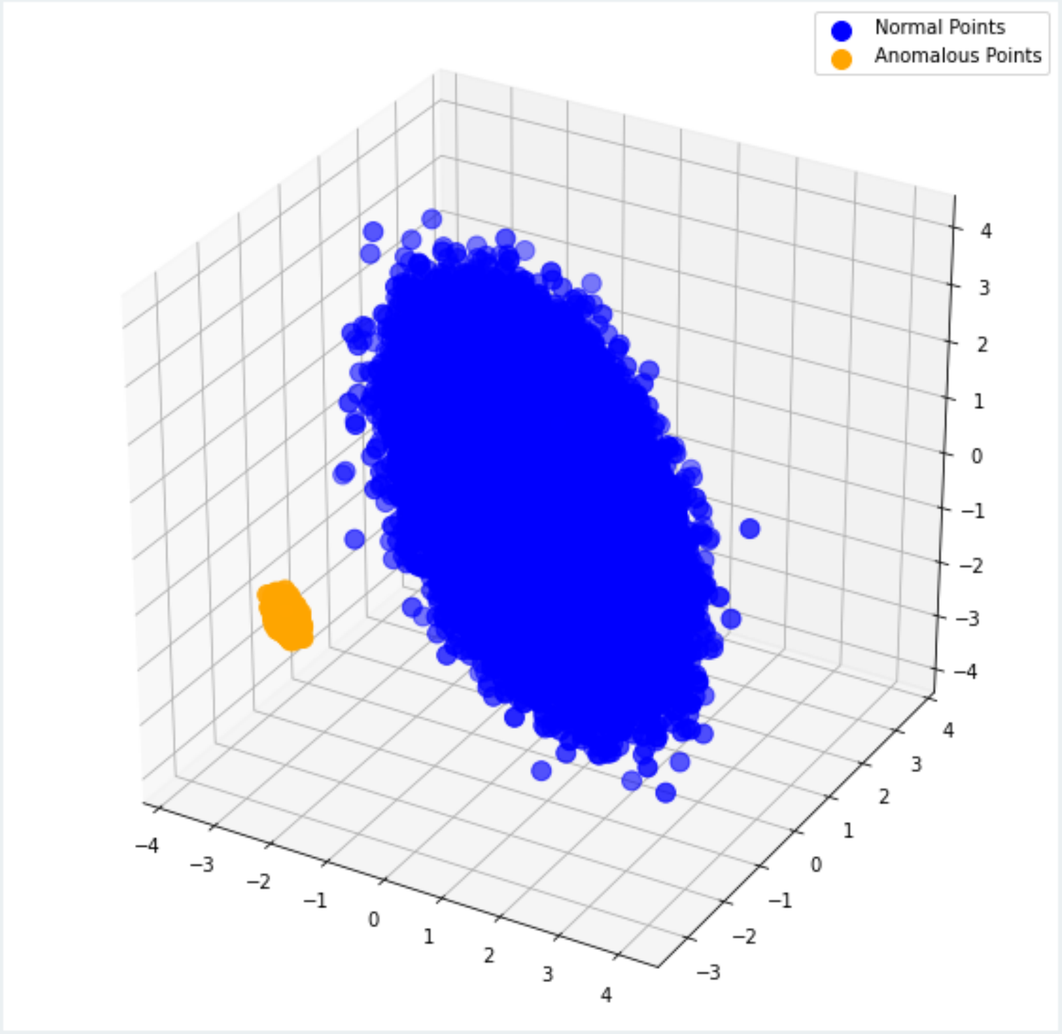
LES DATASETS UTILISÉS (1/3)

Dans le but de garder un certain niveau de simplicité dans cette présentation, nous allons parler seulement de 3 datasets :

- Mulcross (base de données synthétisée):

Nombre d'éléments = 262 144
% d'anomalies = 0.09999847412109375 %



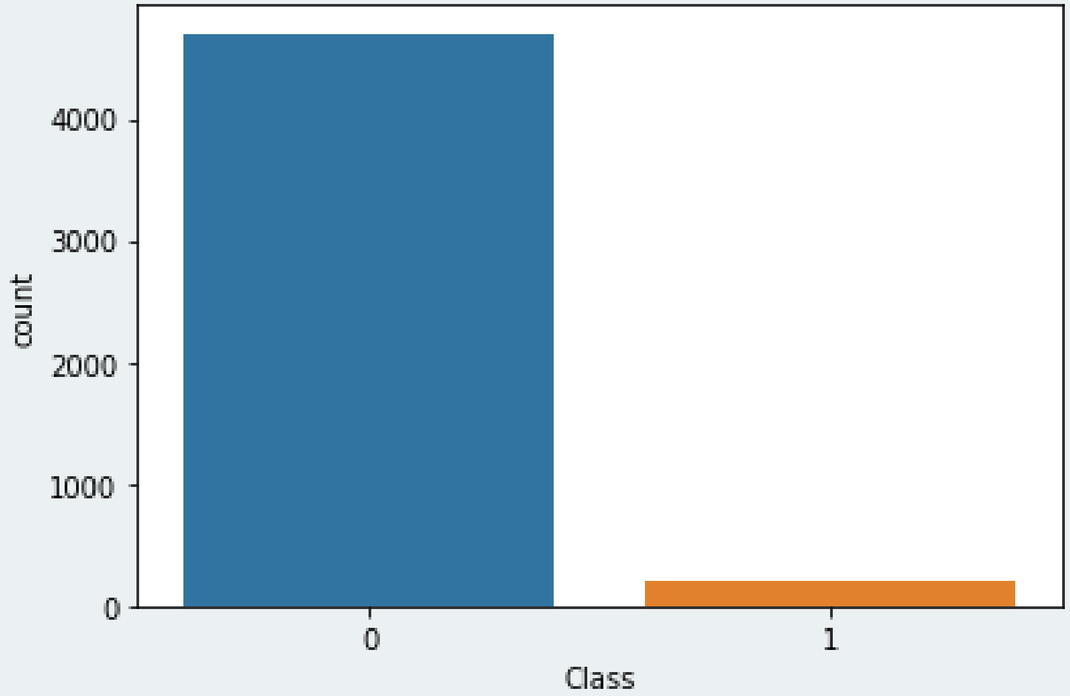


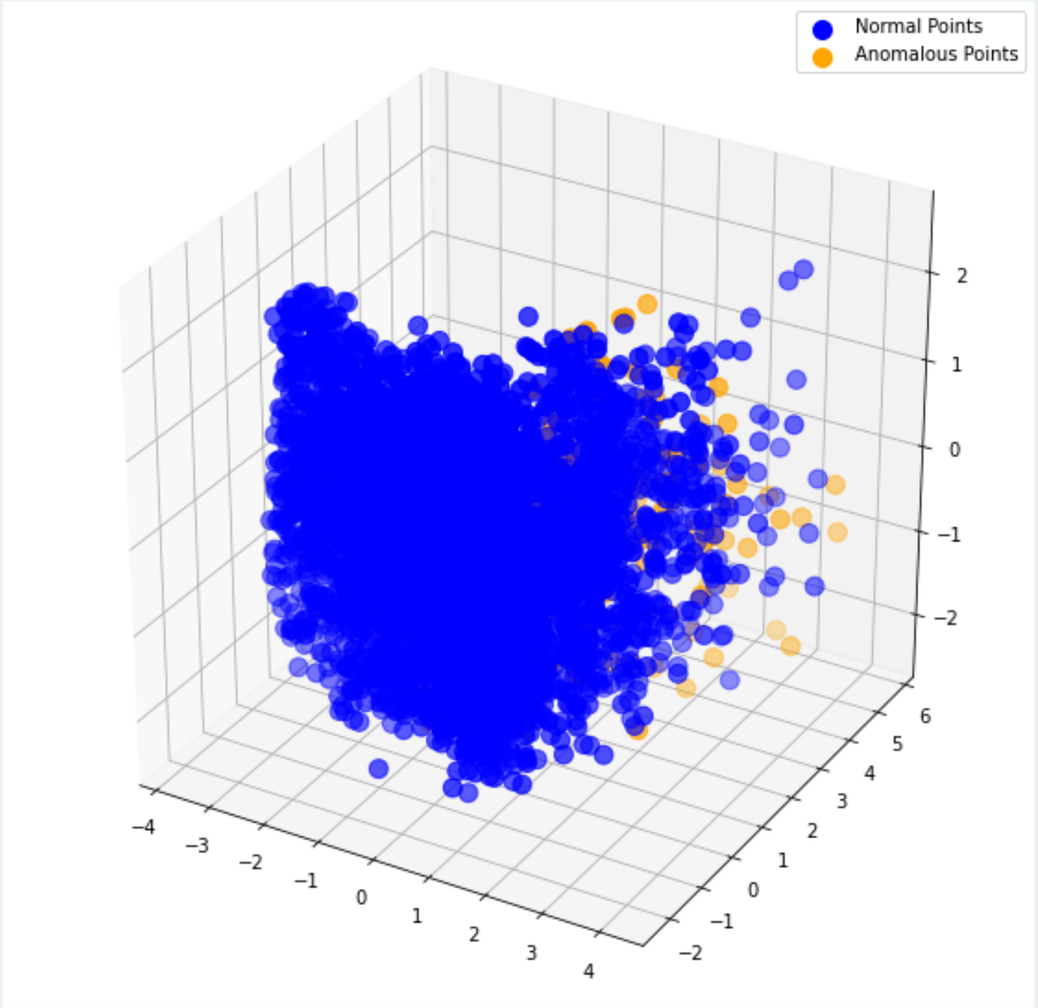
LES DATASETS UTILISÉS (2/3)

- Heart Disease :

Nombre d'éléments = 4 909

% d'anomalies = 0.04257486249745366 %





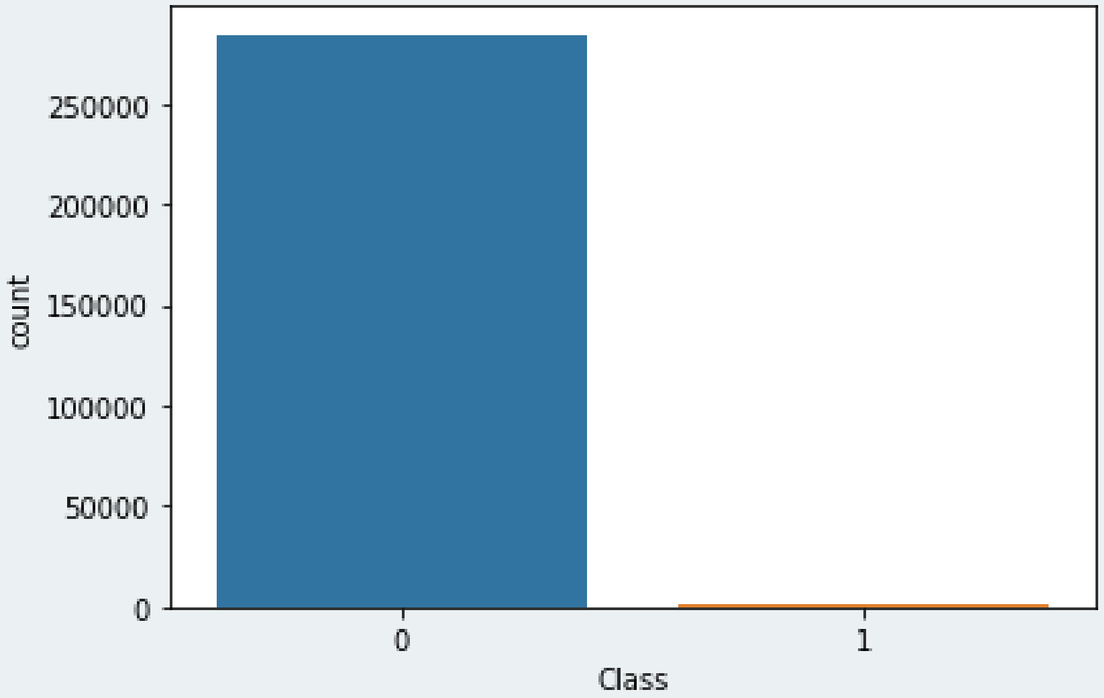
INTÉRPRÉTATION

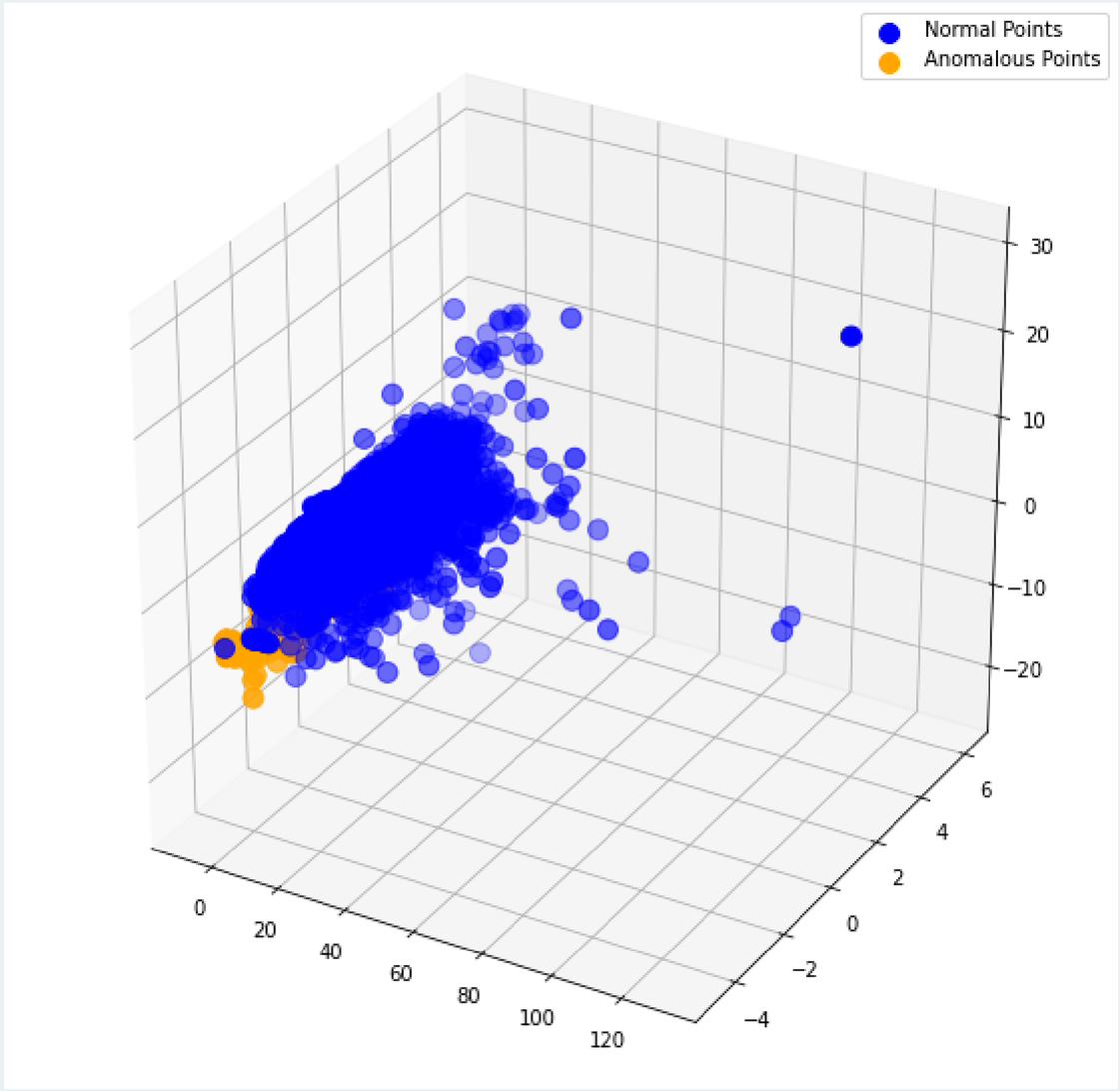
LES DATASETS UTILISÉS (3/3)

- Credit Cards :

Nombre d'éléments = 284 807

% d'anomalies = 0.04257486249745366 %





INTÉRPRÉTATION

LES MESURES UTILISÉES

Les métriques utilisées pour l'évaluation sont principalement 2 :

F-score (ou F1)

Une mesure qui combine la précision et le rappel est leur moyenne harmonique

AUC ROC

Une mesure qui calcule la probabilité qu'une anomalie posséde un score inférieur à celui d'une instance normale

RÉSULTATS

- Mulcross

	F-score	AUC ROC
Valeurs	0.242371	0.947159

- Heart Disease

	F-score	AUC ROC
Valeurs	0.225392	0.649005

- Credit Card

	F-score	AUC ROC
Valeurs	0.157326	0.947561

CONCLUSION

L'UTILISATION DE L'ALGORITHME ISOLATION FOREST AVEC DES DONNÉES FLOUES EST AUSSI PERFORMANT QUE L'UTILISATION DE DONNÉES CLASSIQUE, TOUT EN AJOUTANT LE FACTEUR COMPRÉHENSION VU LA SIMPLICITÉ DE L'INFORMATION QUANT AU CERVEAU HUMAIN.

CE RÉSULTAT PEUT DONC S'AVÉRER ENCORE PLUS UTILE SI ON CHANGE D'AUTRES FACTEURS OU PEUT ÊTRE MÊME SI ON CHANGE D'ALGORITHME.

Merci pour votre Attention