



La détection d'anomalies de données déséquilibrées

Cas du cancer du sein

Présenté par : Paola Andrieu, Augustin
Robert, Timéo Baudat

24 octobre 2024

Sommaire

■ Introduction

■ Méthodes et indicateurs

■ Local Outlier Factor

■ Isolation Forest

■ One-Class Support Vector Machine

■ Comparaison de méthodes



Introduction - La détection d'anomalies

- Identification d'individus qui présentent un écart par rapport à la normale
- Détection de fraudes bancaires, informatiques...
- En médecine, faible présence d'individus avec des cas graves (par comparaison à des personnes saines)
- -> Problèmes liées à des données déséquilibrées (classification)



Le jeu de données

Jeu de données sur le cancer du sein (domaine de la santé)

- 30 features numériques (caractères géométriques des différentes cellules)
- 1 variable factorielle : Bénigne (0) ou Maligne (1)

5.6% du jeu de données représente les cellules malignes (21 vs 357) -> on les considère comme des anomalies



Comment gérer les données déséquilibrées ?

- Ré-échantillonnage possible avec **oversampling** ou **undersampling**
 1. Oversampling (sur-échantillonnage): Générer des observations pour la classe minoritaire (par exemple : SMOTE, Adasyn)
 2. Undersampling (sous-échantillonnage): Supprimer les observations de la classe majoritaire (par exemple : Tomek)
- OU méthodes spécialisées pour la détection d'anomalies



Les méthodes de machine learning utilisées

- 1. Local Outlier Factor (LOF)**
- 2. Isolation Forest**
- 3. One-Class Support Vector Machine (OCSVM)**

Certaines méthodes ne sont pas adaptées à la situation de nos données (par exemple : DBscan qui est adapté aux données de petites dimensions)



Choix des indicateurs de performances

	Accuracy	Spécificité	Sensibilité	Balanced accuracy
Formule	$\frac{TP + TN}{FP + FN + TP + TN}$	$\frac{TN}{TN + FP}$	$\frac{TP}{TP + FN}$	$\frac{Spé + Sen}{2}$
Choix	<ul style="list-style-type: none">• Ne tient pas compte de la classe minoritaire	<ul style="list-style-type: none">• Ne s'intéresse qu'à la classe négative	<ul style="list-style-type: none">• Ne s'intéresse qu'à la classe positive	<ul style="list-style-type: none">• Version pondérée de l'accuracy• Tient compte du déséquilibre des classes

TP : Vrais positifs, TN : Vrais négatifs, FP : Faux positifs, FN : Faux négatifs, Spé : Spécificité, Sen : Sensibilité



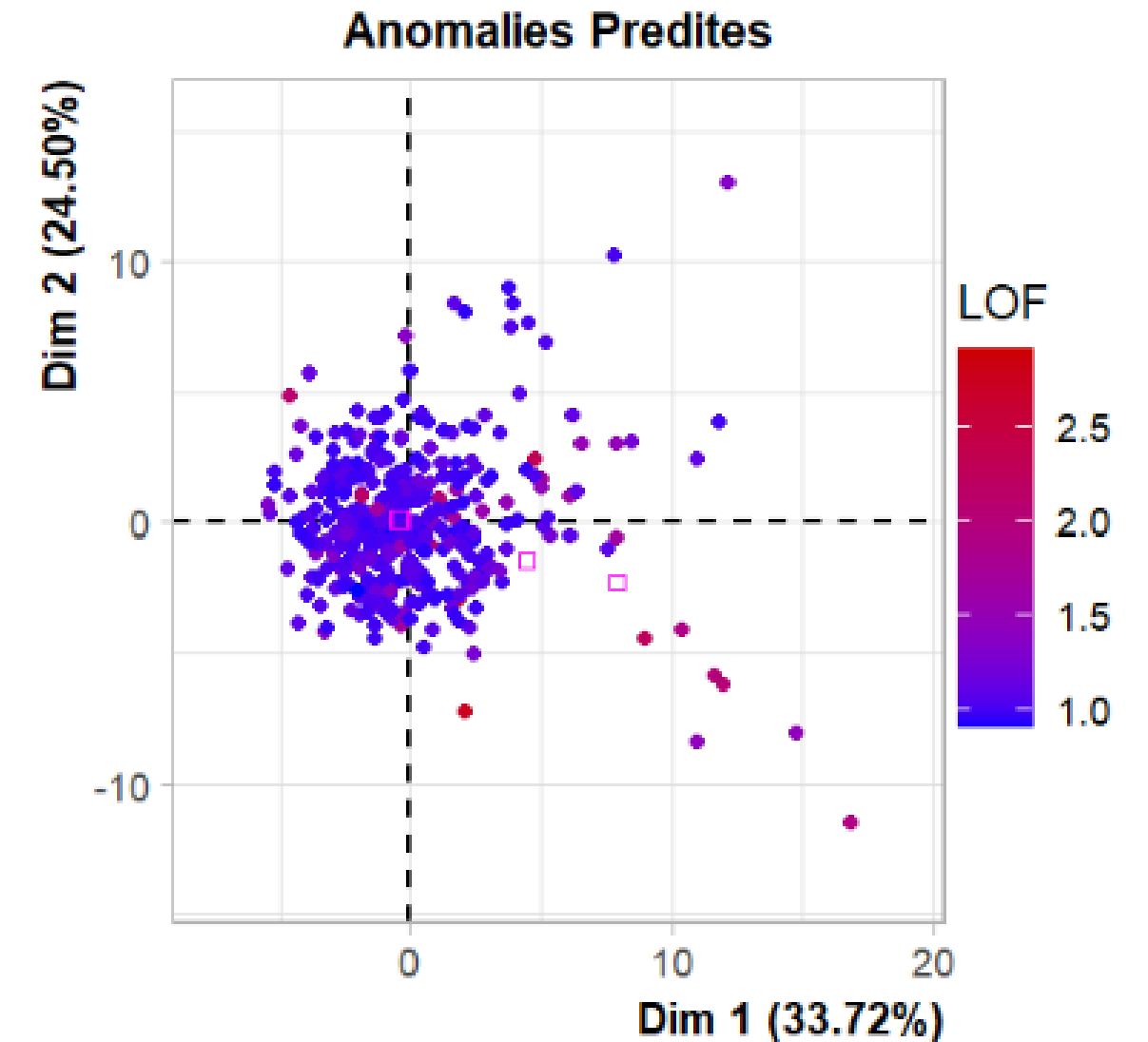
Local Outlier Factor – Principe

Assignation à chaque point d'une valeur de densité locale qui dépend du nombre de voisins proches
Comparaison de chaque densité locale à celle des voisins

Donne un score de “LOF” : plus il est élevé, plus l'observation a de chances d'être une anomalie

On fixe un seuil qui sépare les points en anomalie ou non selon le score de LOF

Deux hyperparamètres : le seuil du score de LOF et le nombre de voisins considérés





Local Outlier Factor - Résultats

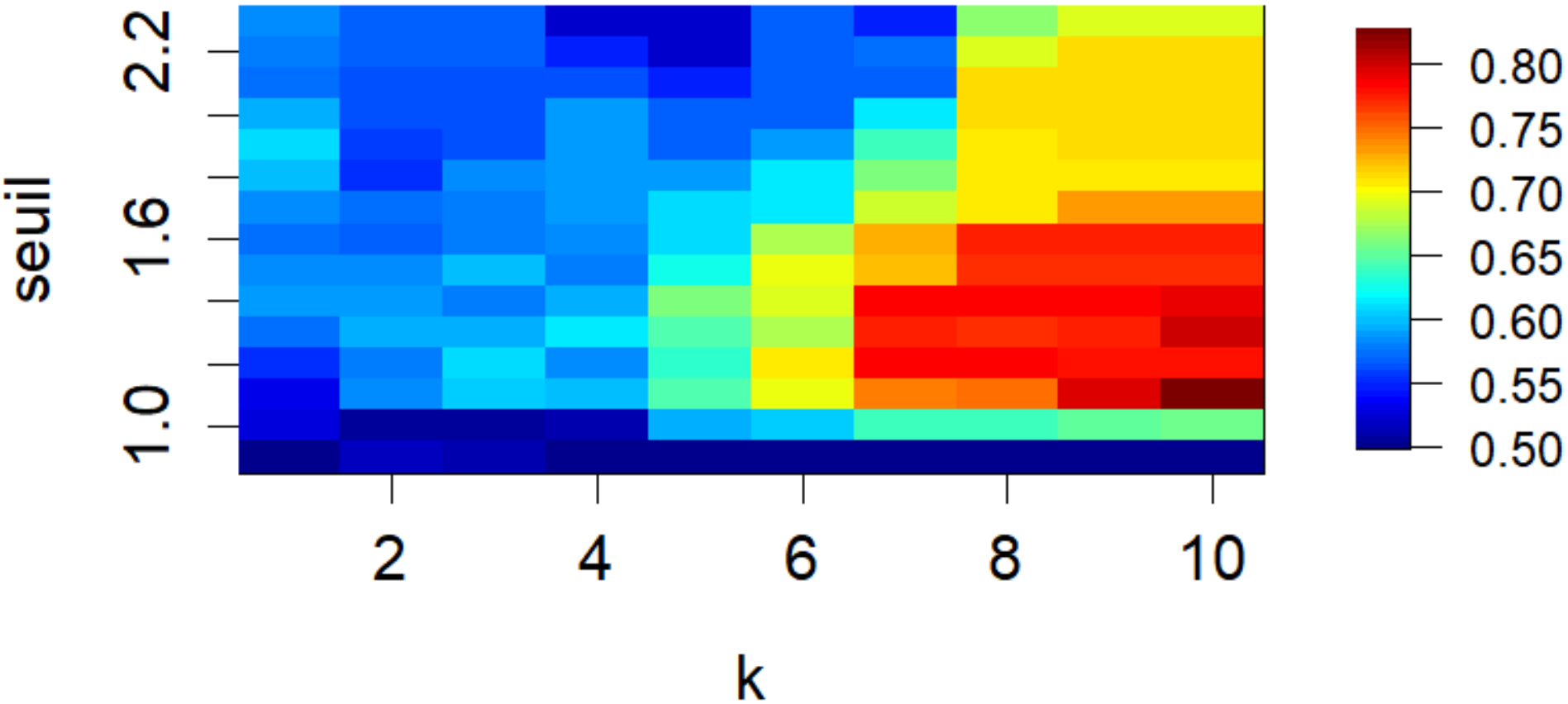
Balanced Accuracy : 0.82
Spécificité : 0.86

Reference		
Prediction	0	1
0	283	3
1	74	18

Accuracy : 0.7963
95% CI : (0.7521, 0.8358)
No Information Rate : 0.9444
P-Value [Acc > NIR] : 1

Kappa : 0.2508

Balanced Accuracy en fonction des paramètres



Mcnemar's Test P-Value : 1.496e-15

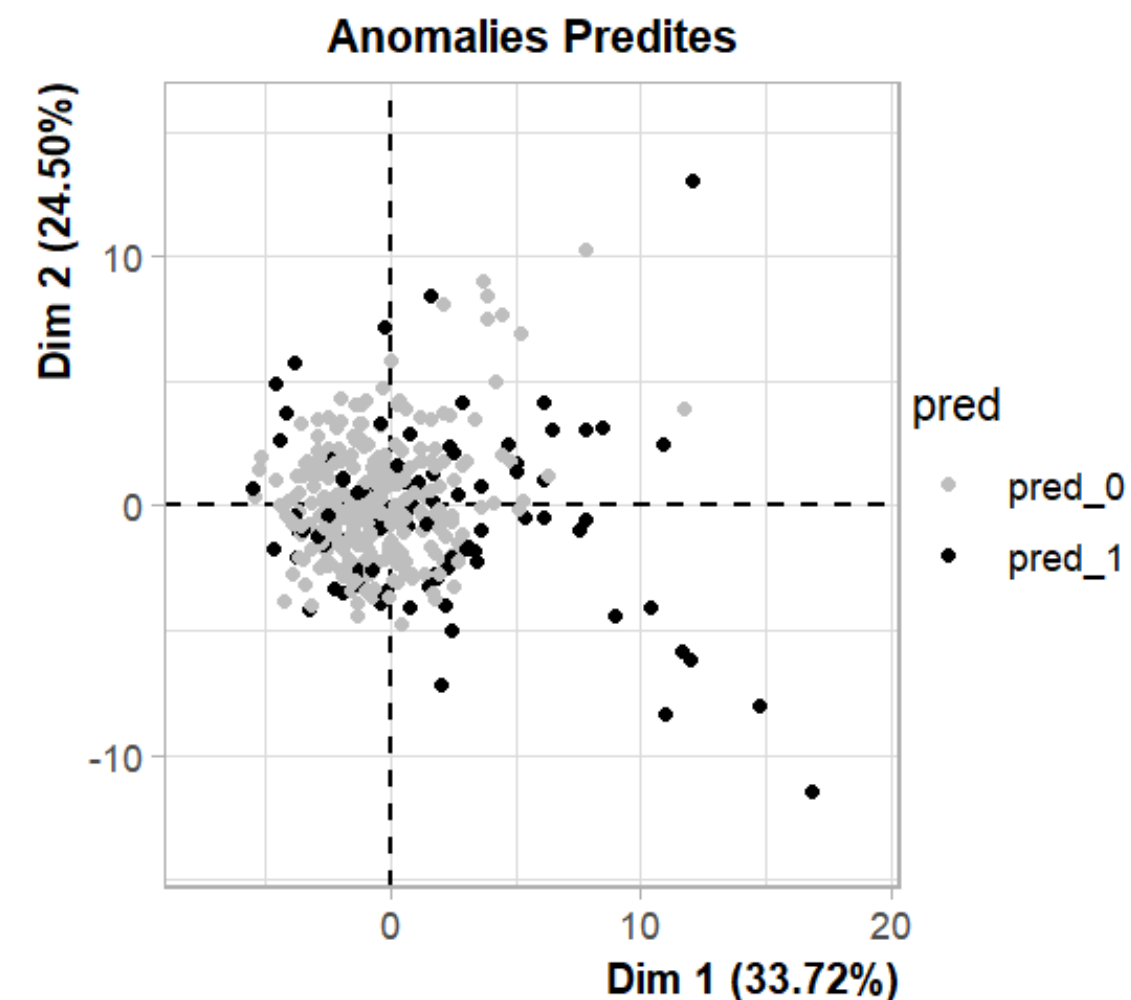
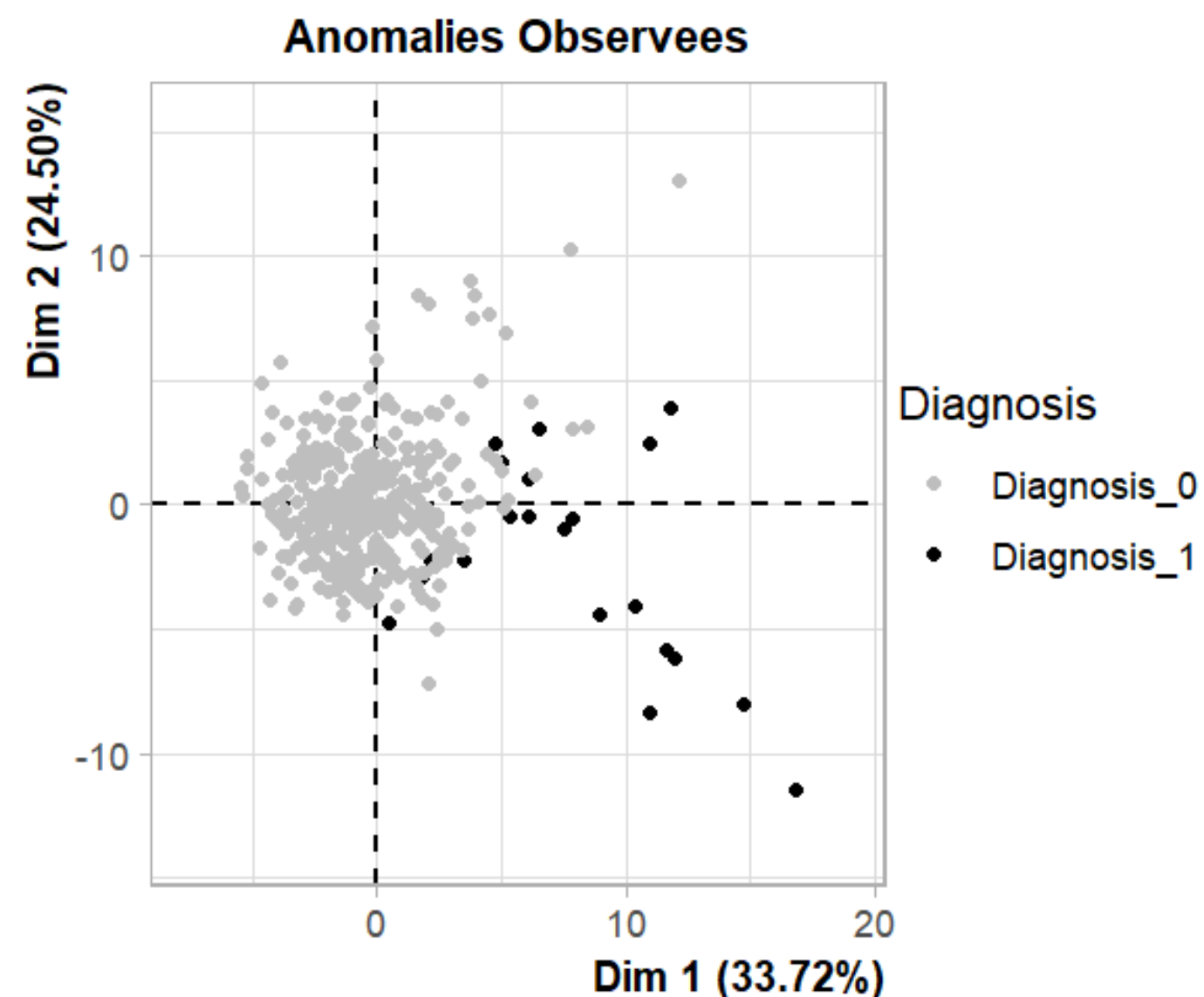
Sensitivity : 0.7927
Specificity : 0.8571
Pos Pred Value : 0.9895
Neg Pred Value : 0.1957
Prevalence : 0.9444
Detection Rate : 0.7487
Detection Prevalence : 0.7566
Balanced Accuracy : 0.8249

'Positive' Class : 0



Local Outlier Factor - Résultats

	Reference	
Prediction	0	1
0	283	3
1	74	18

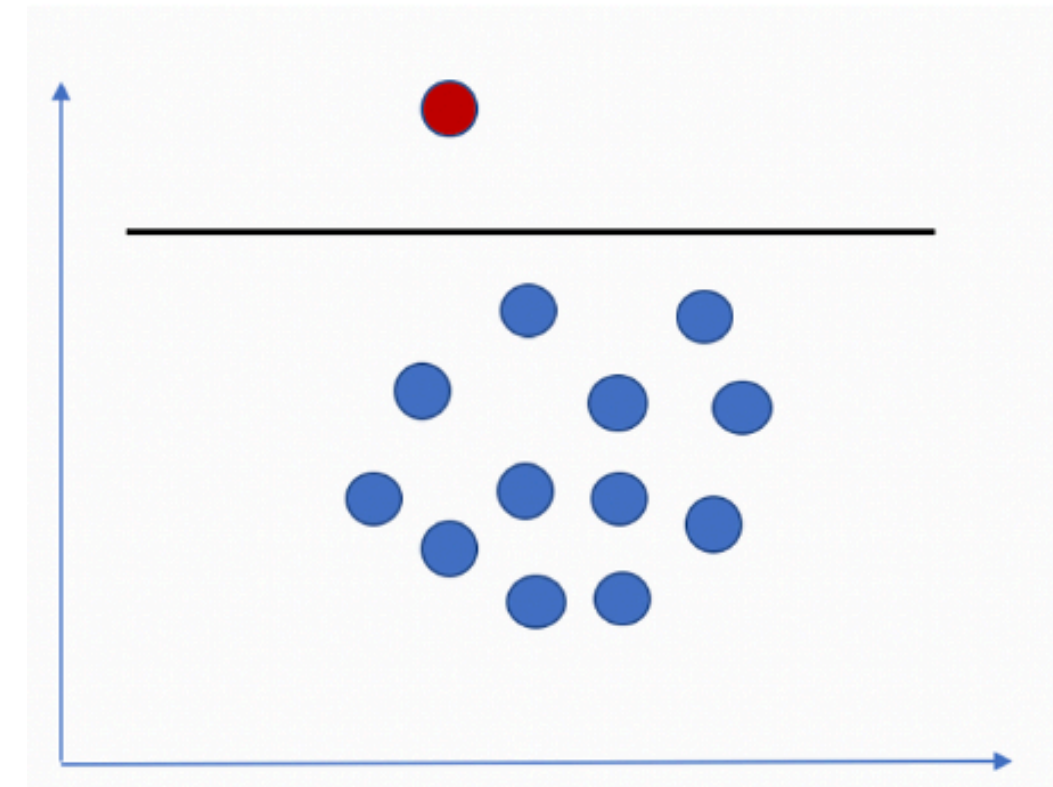




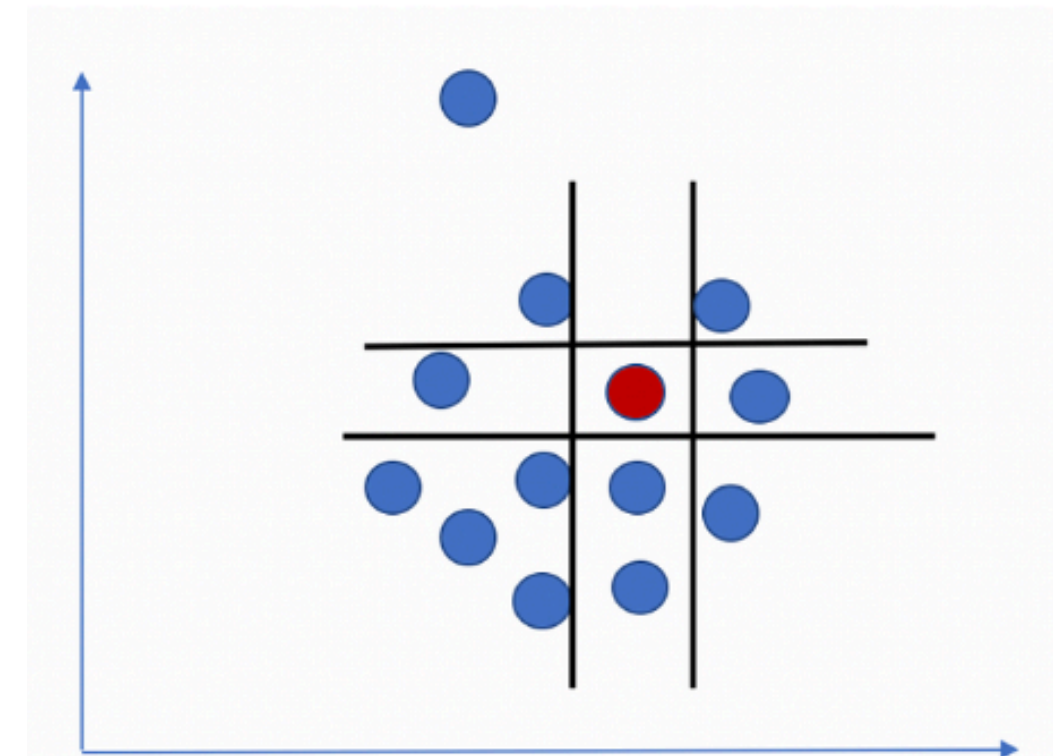
Isolation Forest - Principe

- Principe
 - Anomalies facilement isolables, car rares et différentes
- Concept clé
 - Isolation par partitionnement aléatoire (construction d'arbres de manière aléatoire)
 - Profondeur d'isolement
- Etapes de l'algorithme
 - Echantillonnage aléatoire
 - Construction de l'arbre d'isolation
 - Calcul du score d'anomalie
- Score d'anomalie
 - compris entre 0 et 1
 - d'autant plus grand que l'isolement est rapide
 - seuil à déterminer

Isolating an anomalous point

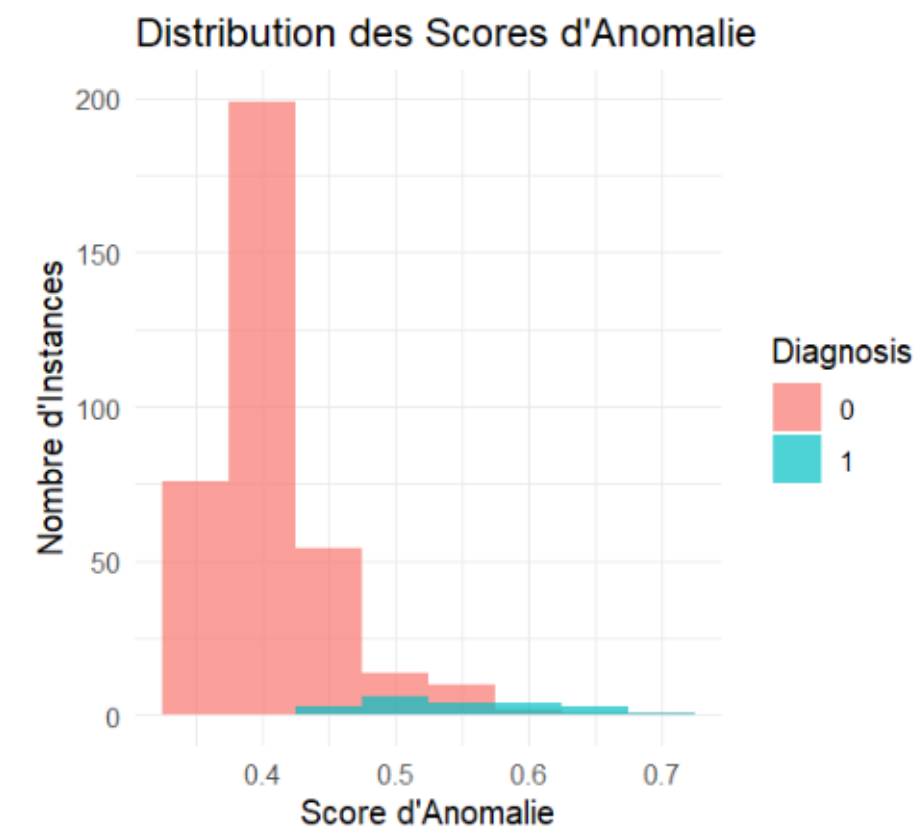
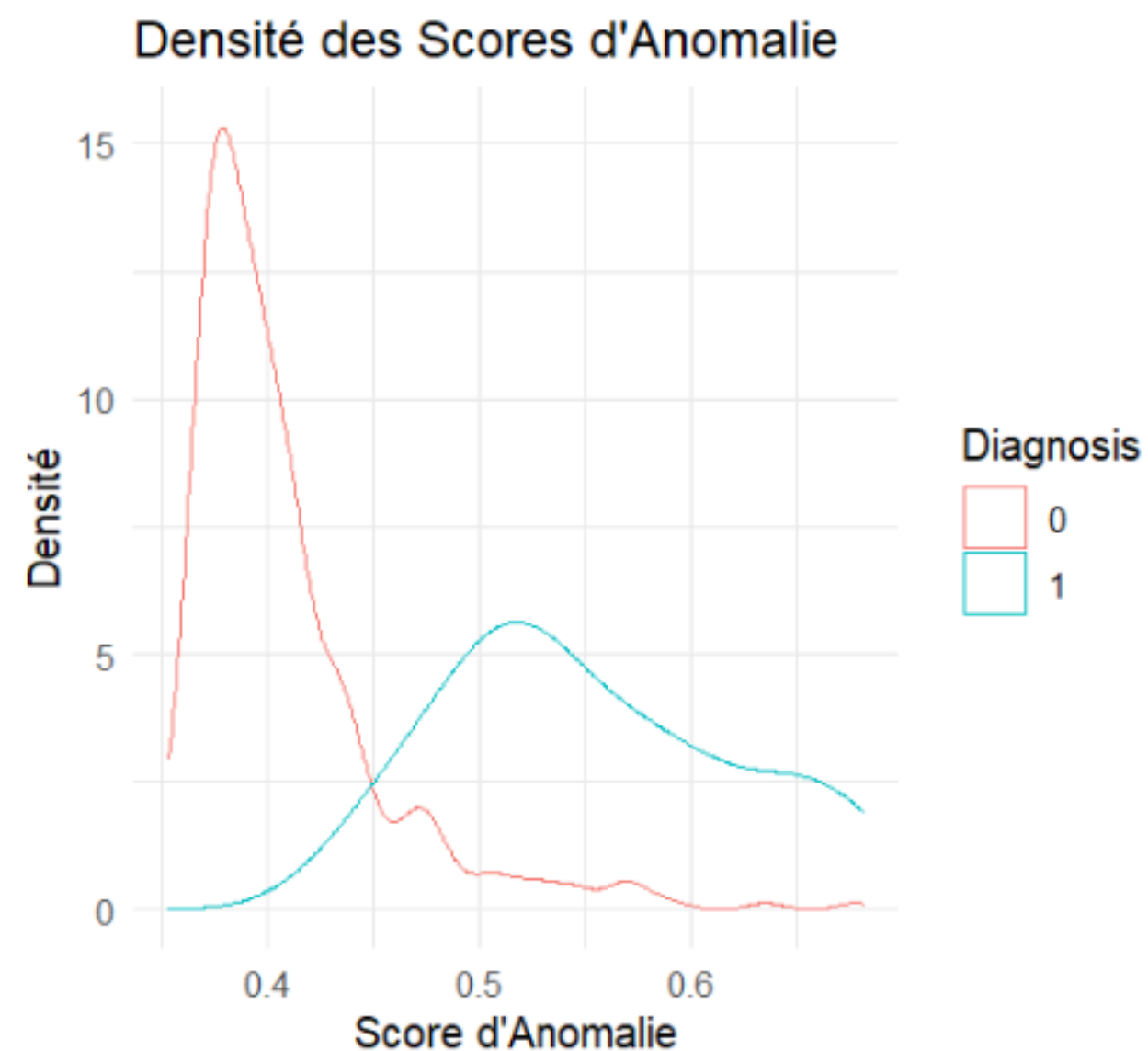
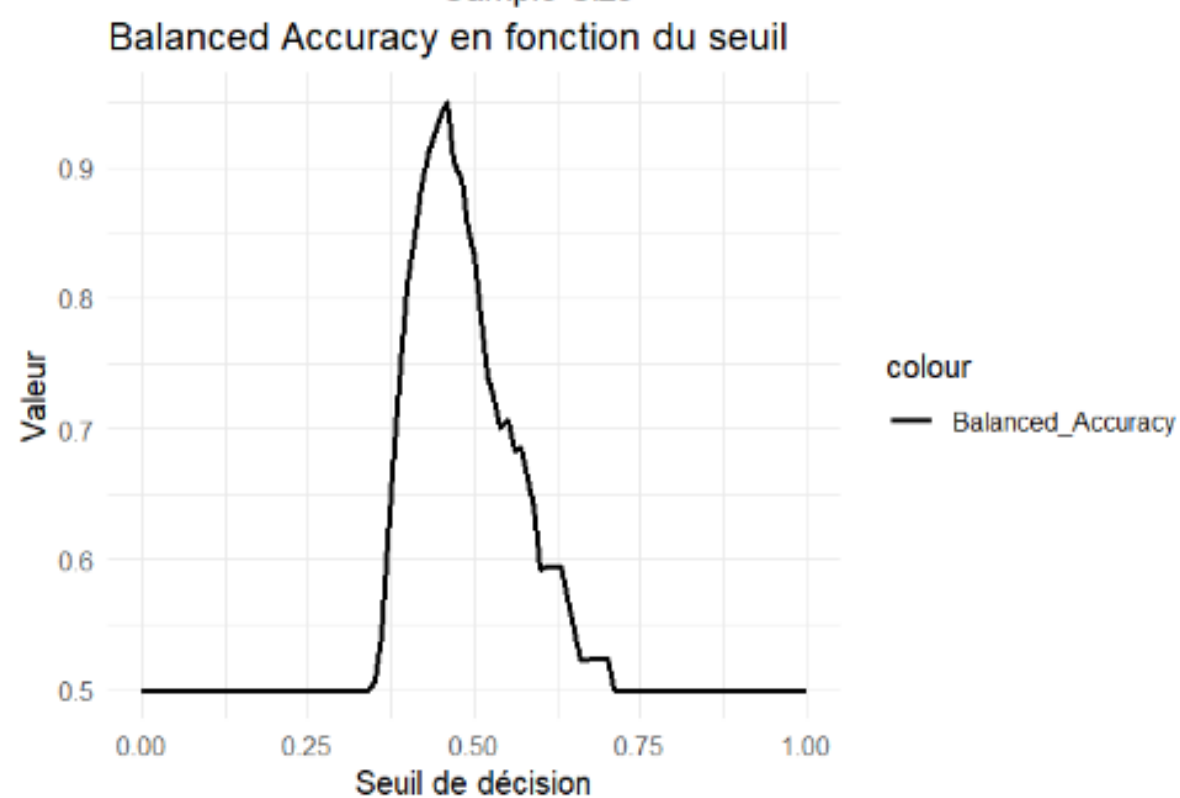
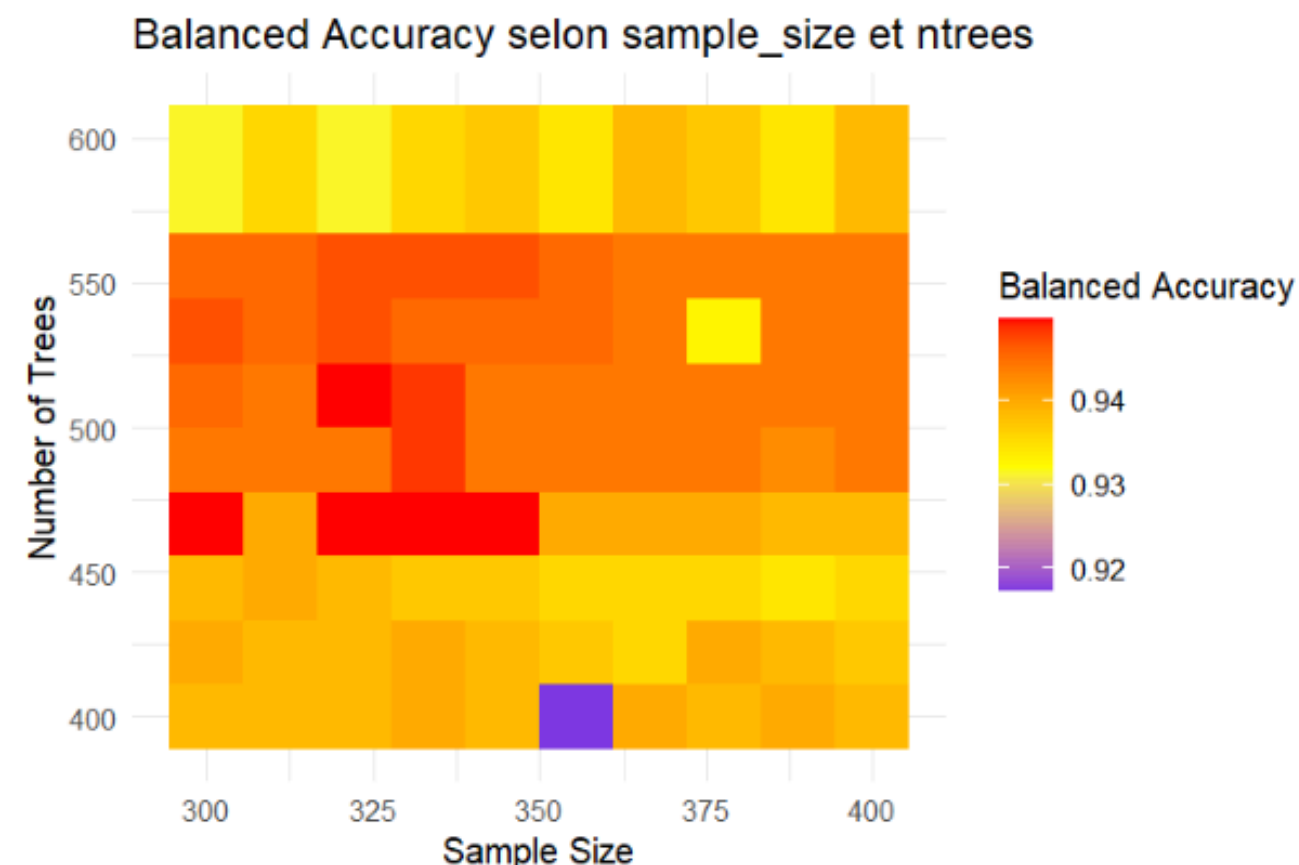


Isolating a normal point





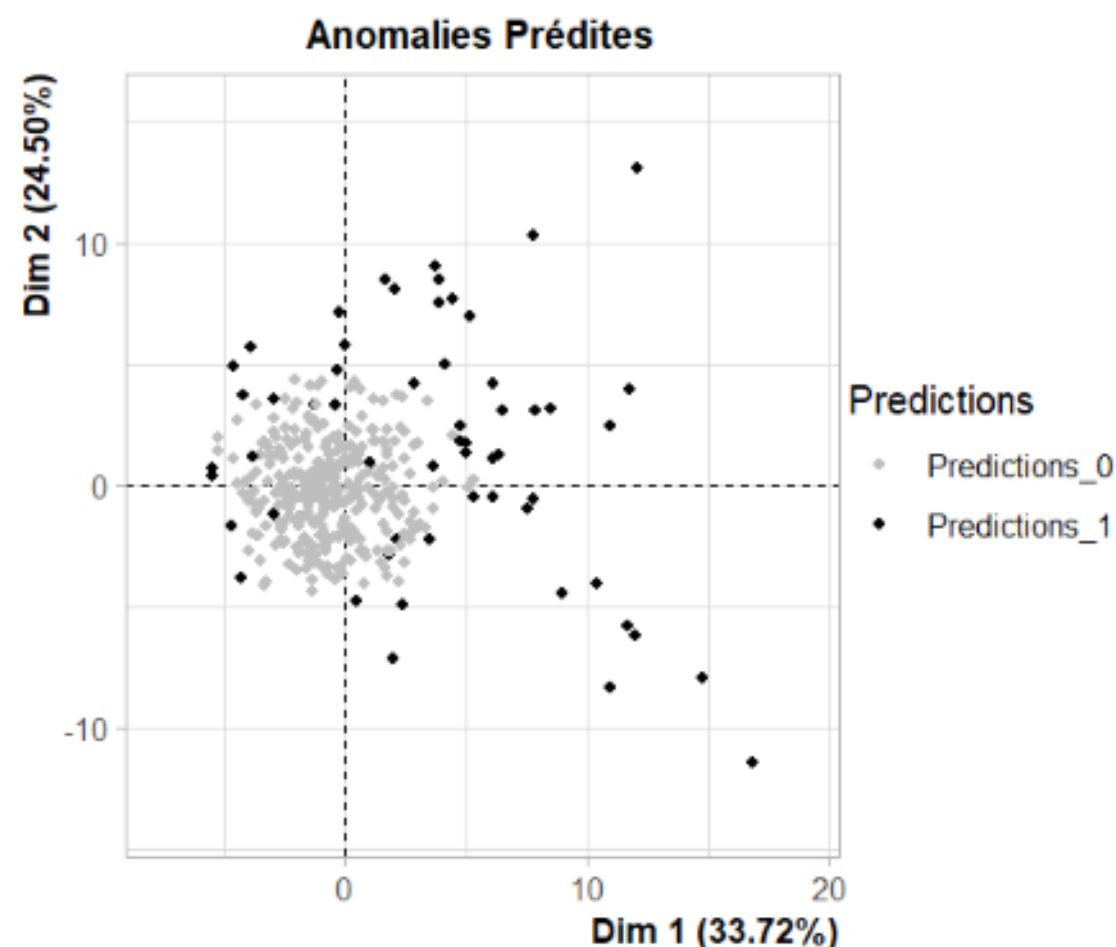
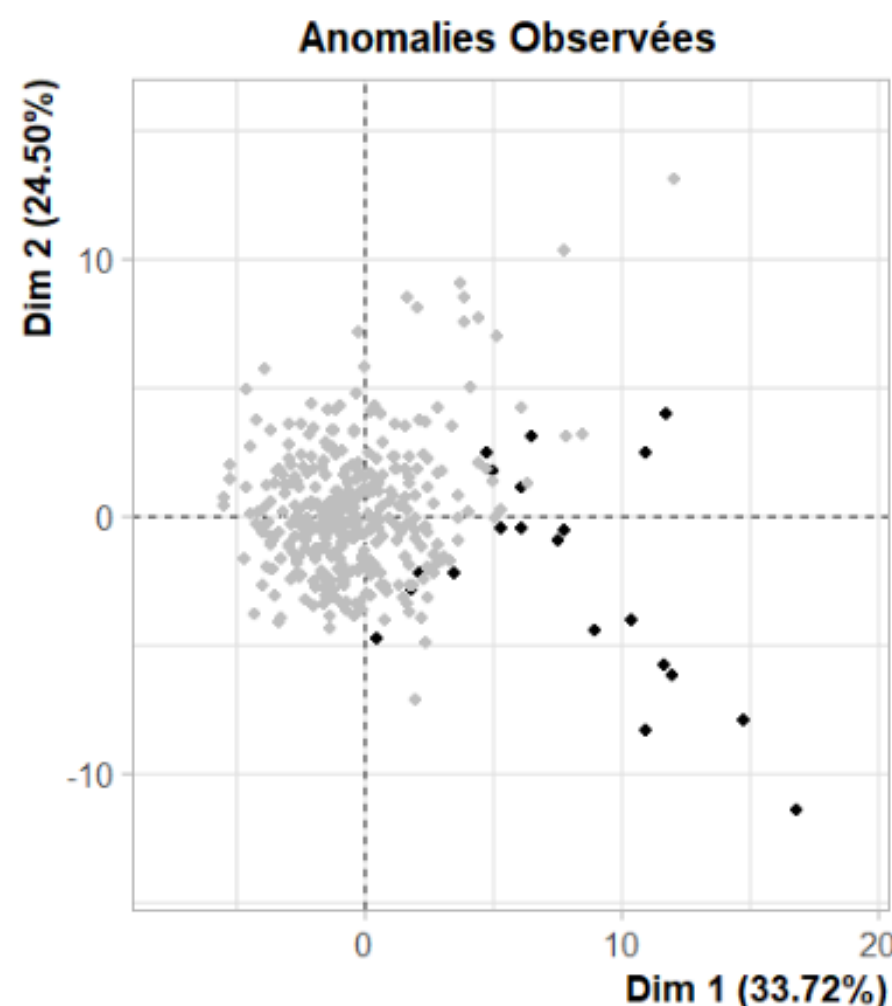
Isolation Forest - Résultats





Isolation Forest - Résultats

Confusion Matrix and Statistics



Reference

Prediction	0	1
0	321	0
1	36	21

Accuracy : 0.9048
95% CI : (0.8706, 0.9324)
No Information Rate : 0.9444
P-Value [Acc > NIR] : 0.9993

Kappa : 0.4977

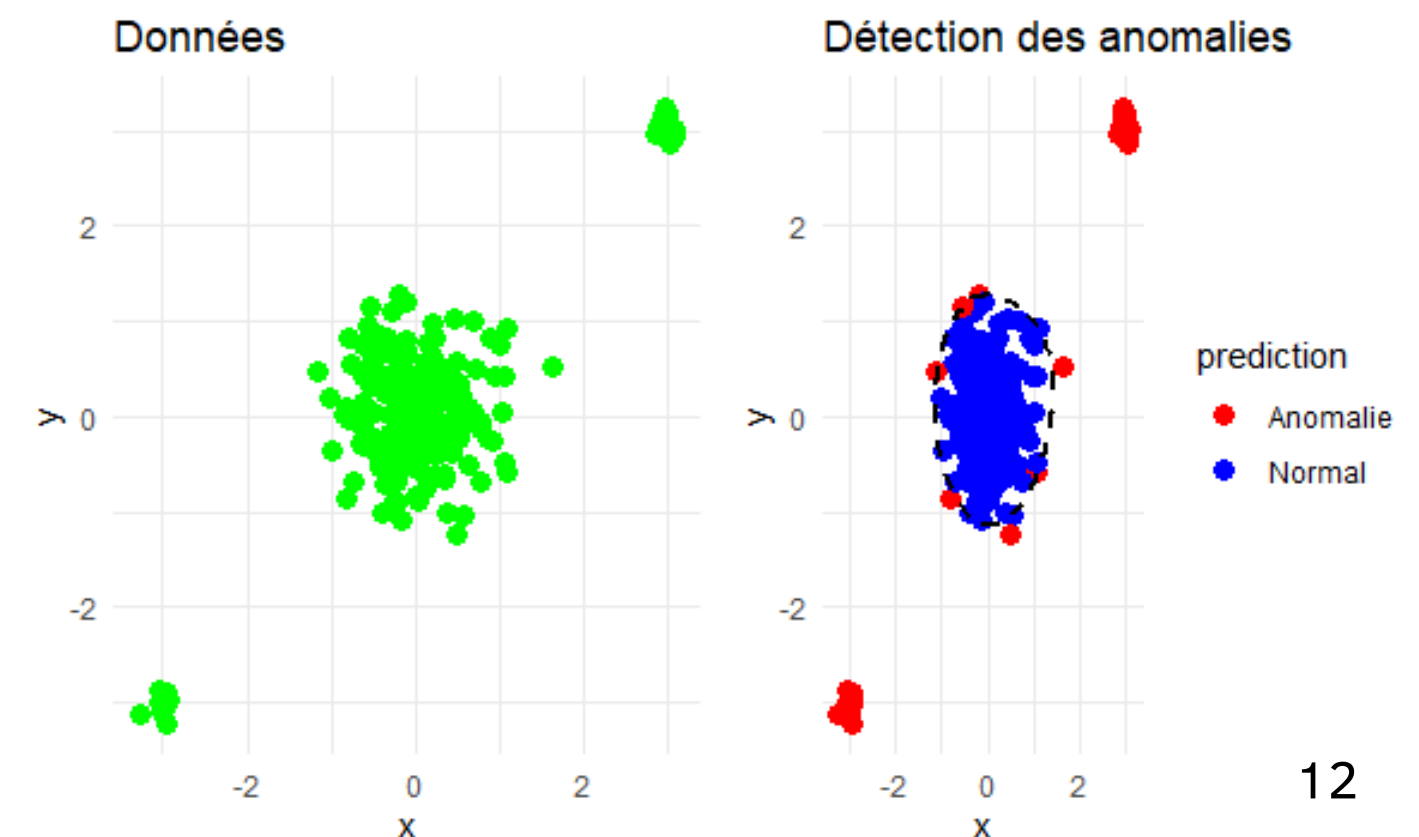
Mcnemar's Test P-Value : 5.433e-09

Sensitivity : 0.8992
Specificity : 1.0000
Pos Pred Value : 1.0000
Neg Pred Value : 0.3684
Prevalence : 0.9444
Detection Rate : 0.8492
Detection Prevalence : 0.8492
Balanced Accuracy : 0.9496



One-Class Support Vector Machine - Principe

- Apprentissage sur une **seule classe** (points normaux) pour déterminer une frontière qui enveloppe la majorité des points normaux
- Transformation des données via un noyau (comme les SVM)
- Tous les points qui ne sont pas dans la zone normale dans l'hyperplan sont considérés comme des **anomalies**
- Paramètre ν : contrôle la proportion d'erreurs acceptées (points hors de la frontière) et la fraction de support vectors.
- Paramètre γ : détermine l'influence d'un point de données individuel sur la forme de la frontière ; un γ élevé génère une frontière plus complexe.



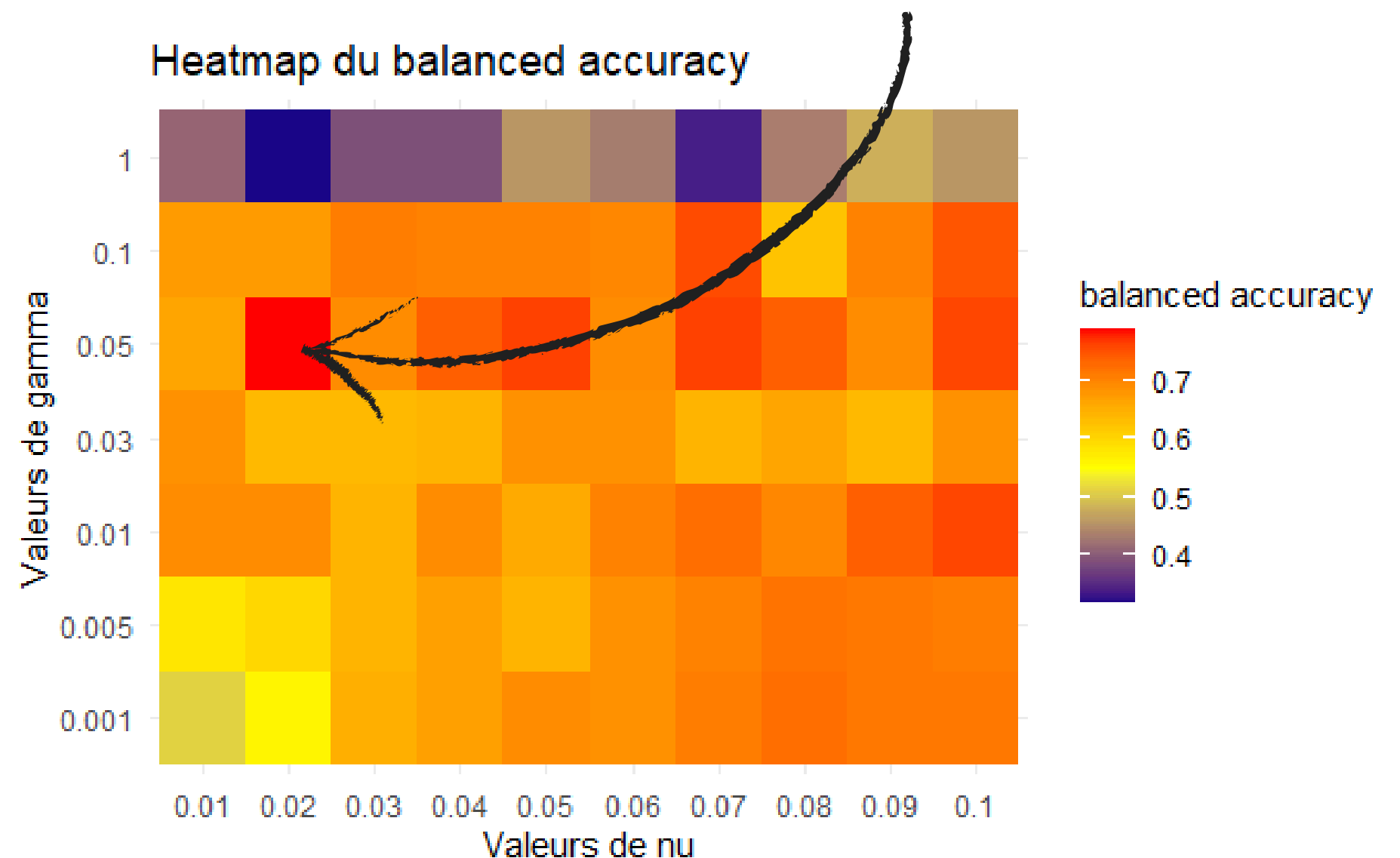


One-Class Support Vector Machine - Résultats

Sélection de la plage de 2 paramètres :

- Paramètre ν : en lien avec la proportion d'anomalies -> inférieur à 10%
- Paramètre γ : inversement proportionnel au nombre de features (ici 30 donc valeur de base = 0.03)

Couple de valeurs optimale





One-Class Support Vector Machine - Résultats

Confusion Matrix and Statistics

Reference		
Prediction	0	1
0	273	4
1	84	17

Accuracy : 0.7672

95% CI : (0.7213, 0.8089)

No Information Rate : 0.9444

P-Value [Acc > NIR] : 1

Kappa : 0.2056

Mcnemar's Test P-Value : <2e-16

Sensitivity : 0.7647

Specificity : 0.8095

Pos Pred Value : 0.9856

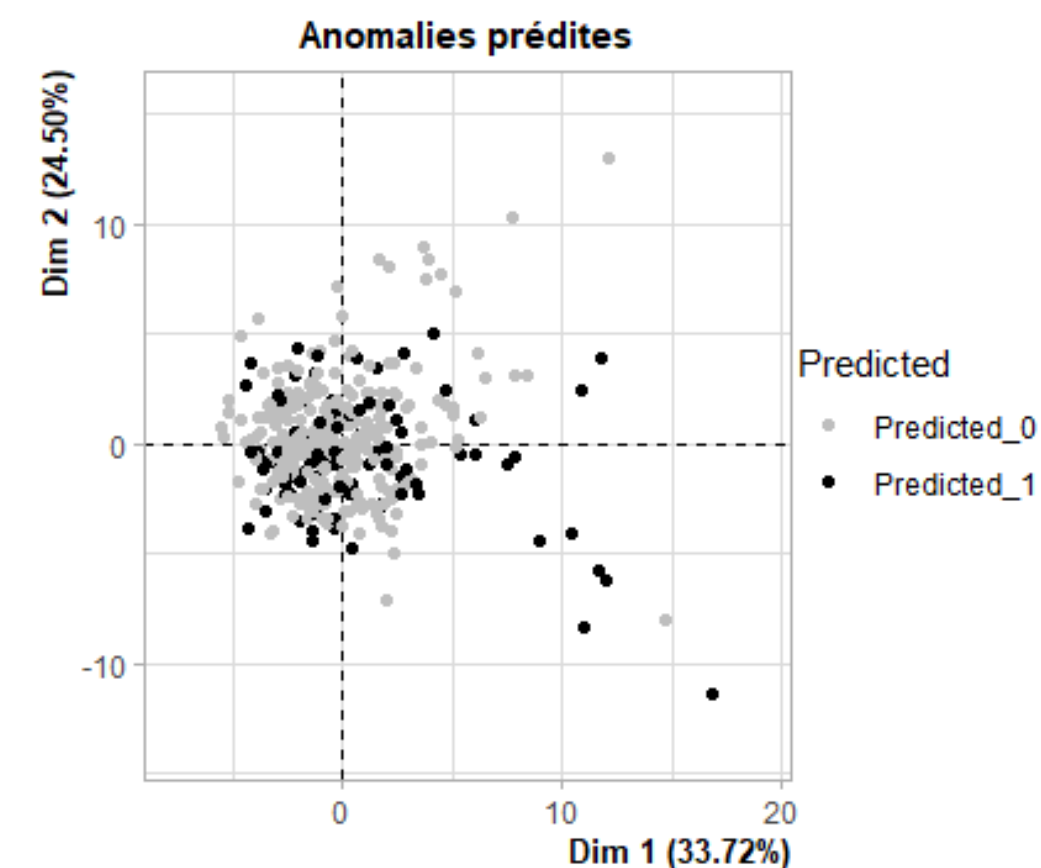
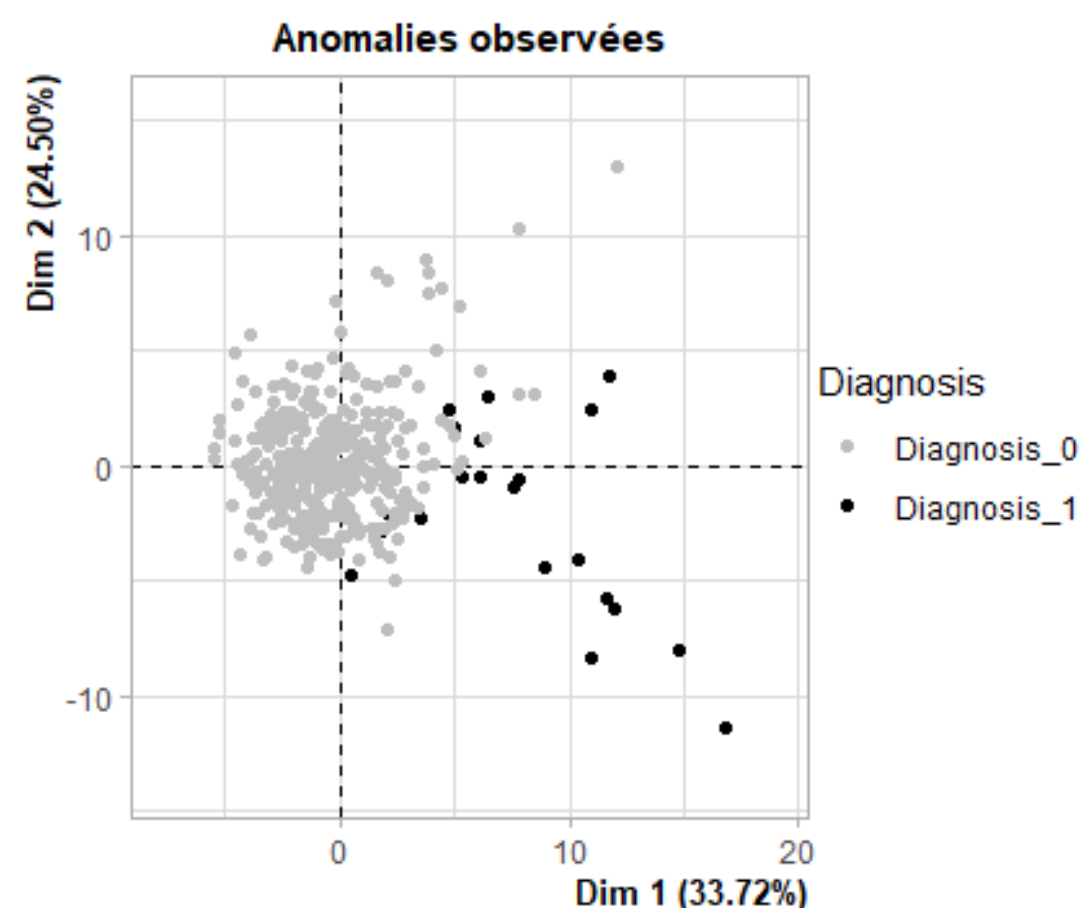
Neg Pred Value : 0.1683

Prevalence : 0.9444

Detection Rate : 0.7222

Detection Prevalence : 0.7328

Balanced Accuracy : 0.7871





Comparaison des méthodes

	LOF	IF	OCSVM
Balanced accuracy	0.82	0.95	0.79
Spécificité	0.86	1	0.81

En plus d'obtenir de meilleurs résultats, Isolation Forest reste la méthode la plus rapide (puis OC-SVM et enfin LOF qui reste la plus longue)



Conclusion

- La détection d'anomalies est adaptée à de **nombreux domaines** (sécurité, économie, santé, capteurs automatiques, etc)
- **Isolation Forest** reste la méthode la plus performante par rapport à OCSVM et LOF
- Limite : problèmes si trop d'anomalies regroupées : peuvent ne pas être détectées (exemple d'un capteur qui peut générer des données anormales sur une durée t)



Bibliographie

Hu Y, Shan WM and Y, Australia (2022) Rlof: R Parallel Implementation of Local Outlier Factor(LOF).

Liu FT, Ting KM, Zhou Z-H (2008) Isolation Forest. 2008 Eighth IEEE Int. Conf. Data Min. IEEE, Pisa, Italy, pp 413–422

Rouvière L (2023) Données déséquilibrées.

Togbe MU, Chabchoub Y, Boly A, Chiky R (2020) Etude comparative des méthodes de détection d'anomalies.

University of Wisconsin–Madison (n.d.) Breast Cancer Wisconsin (Diagnostic) Data Set.

<https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>

Vaibhav Jayaswal (2020) Local Outlier Factor (LOF) — Algorithm for outlier identification, dans Towards Data Science

Des questions ?

Merci pour votre écoute

Paola Andrieu

paola.andrieu@agrocampus-ouest.fr

Augustin Robert

augustin.robert@agrocampus-ouest.fr

Timéo Baudat

timeo.baudat@agrocampus-ouest.fr