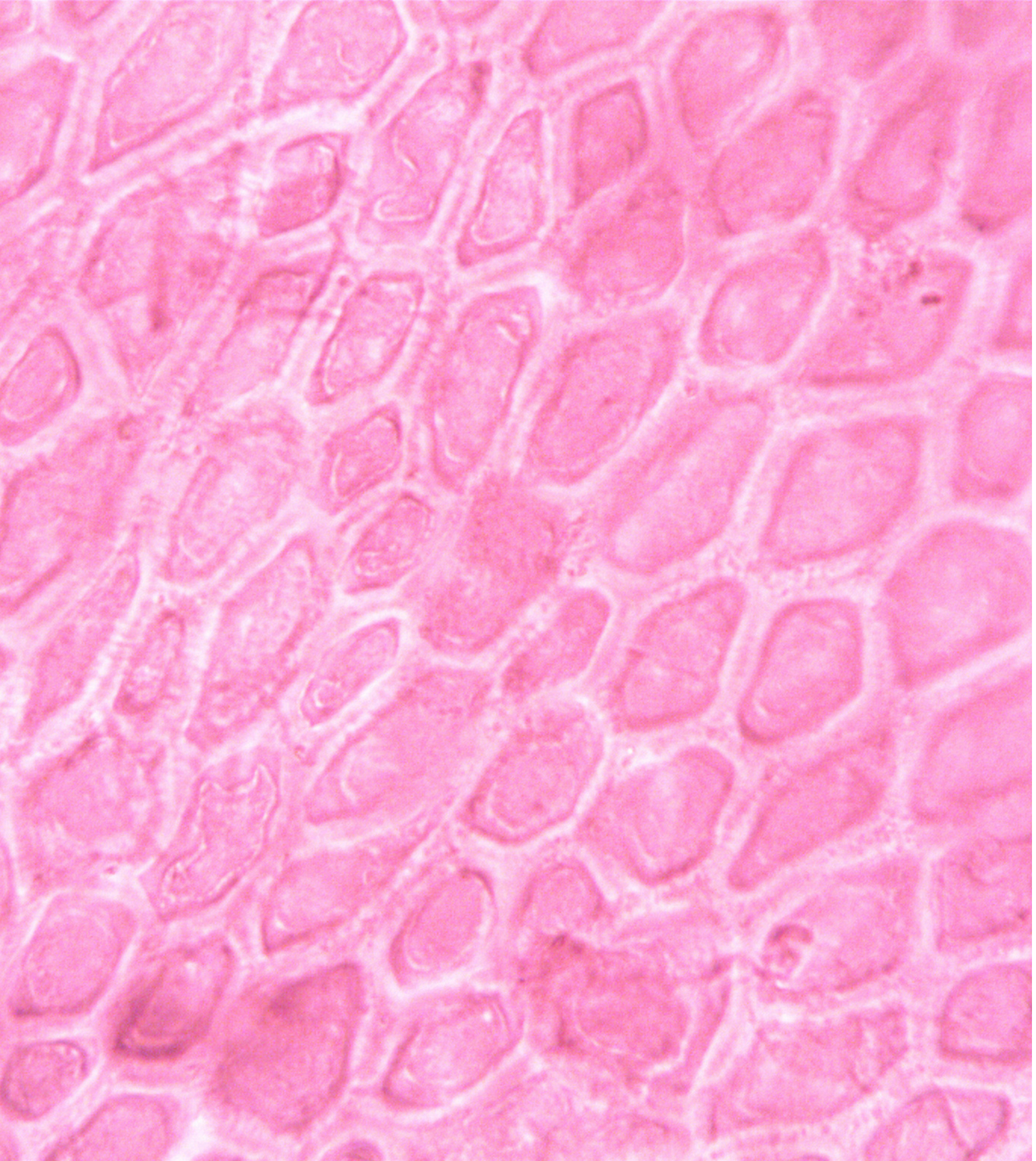


La détection d'anomalies de données déséquilibrées

Cas du cancer du sein

Présenté par : Paola Andrieu, Augustin Robert, Timéo Baudat

24 octobre 2024

L'INSTITUT agro Rennes Angers

Sommaire

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 déséquilibrés

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

(University of Wisconsin – Madison, n.d.)

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 pout la détection d'anomalies

(Rouvière, 2023)

Les méthodes de machine learning utilisées

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

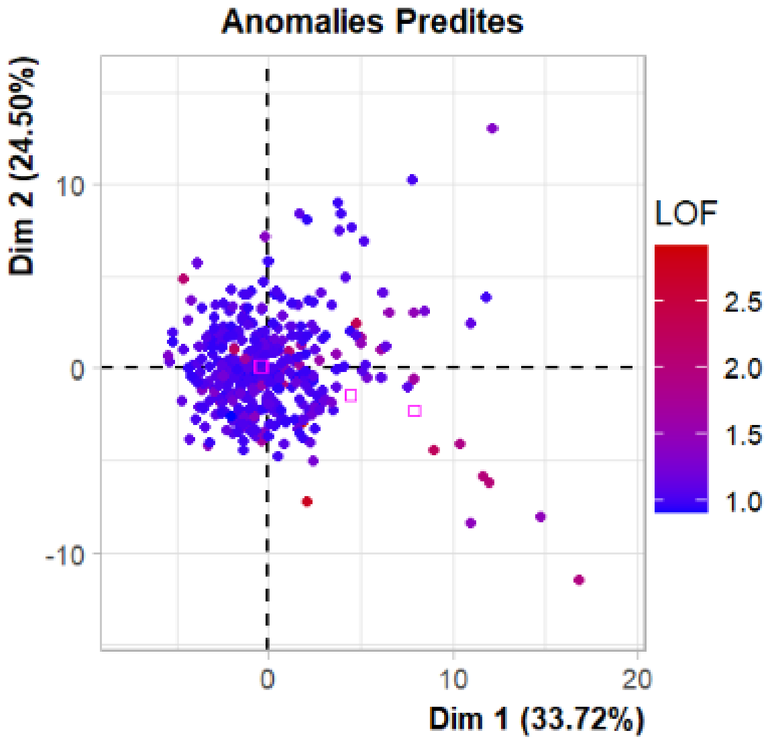
Certaines méthodes ne sont pas adaptées à la situation de nos données (par exemple : DBscan qui est adapté aux données de petites dimensions)

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	• Ne tient pas compte de la classe minoritaire	• Ne s'intéresse qu'à la classe négative	• Ne s'intéresse qu'à la classe positive	• Version pondérée de l'accuracy • Tient compte du déséquilibre des classes

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

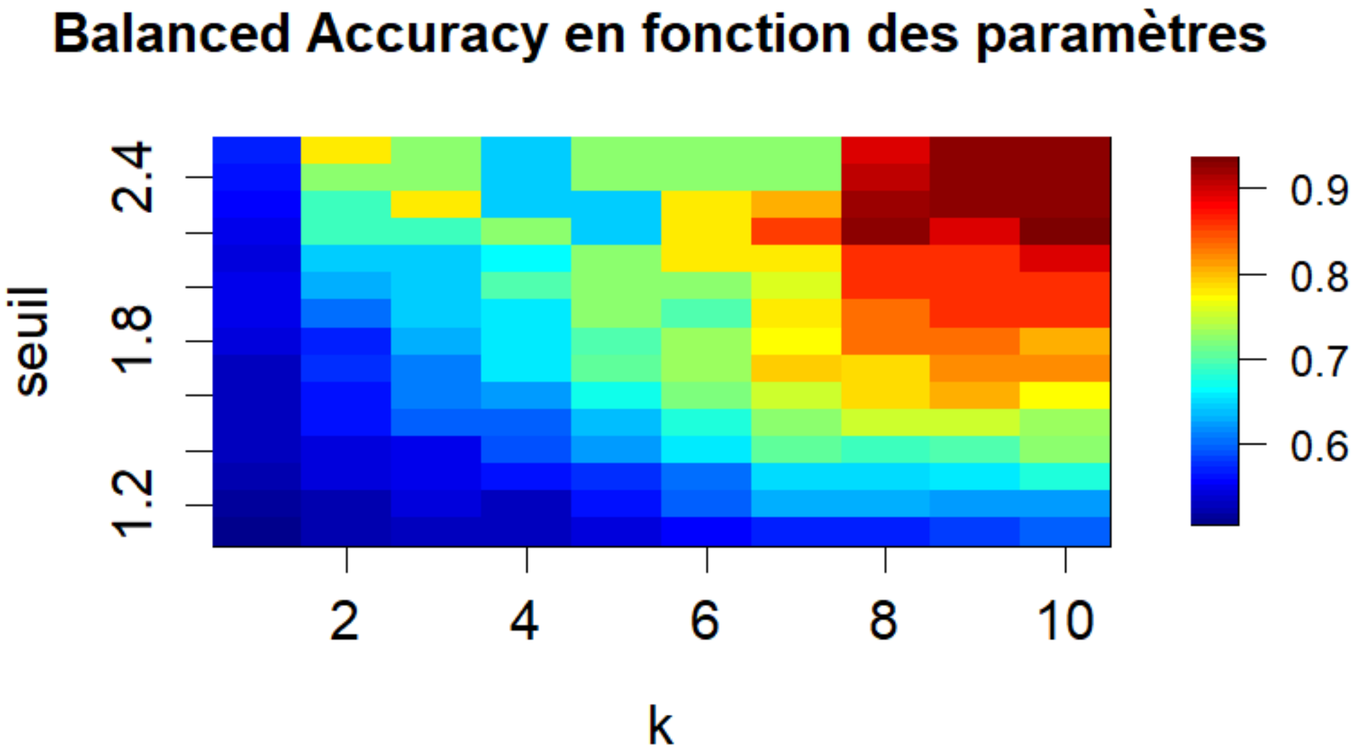


(Jayaswal, 2020)

Local Outlier Factor - Résultats

Balanced Accuracy : 0.93

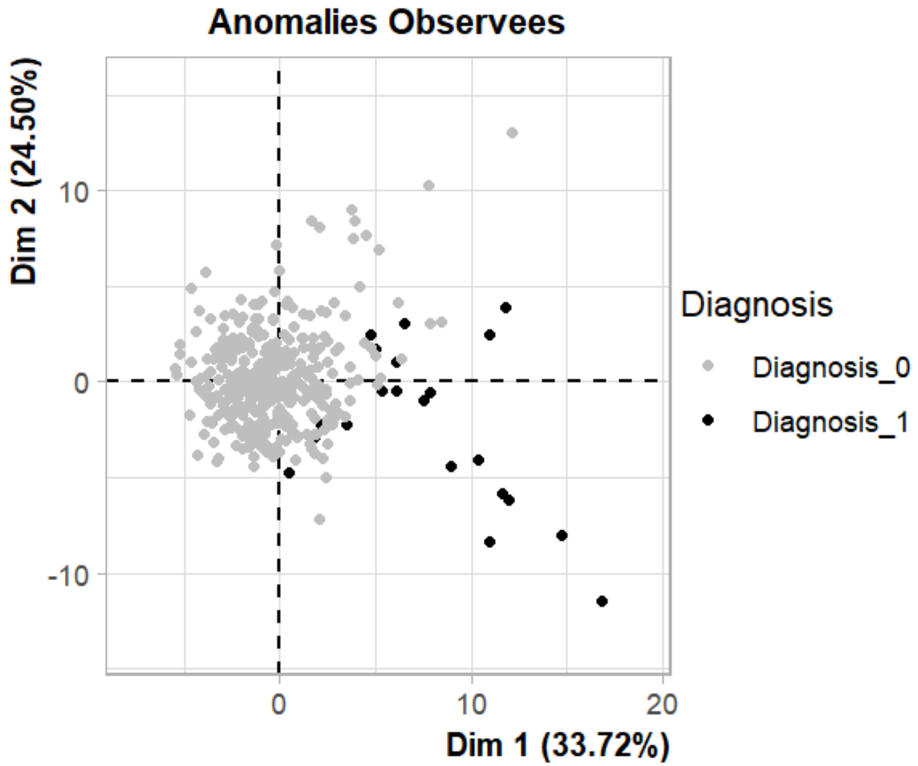
Kappa : 0.57

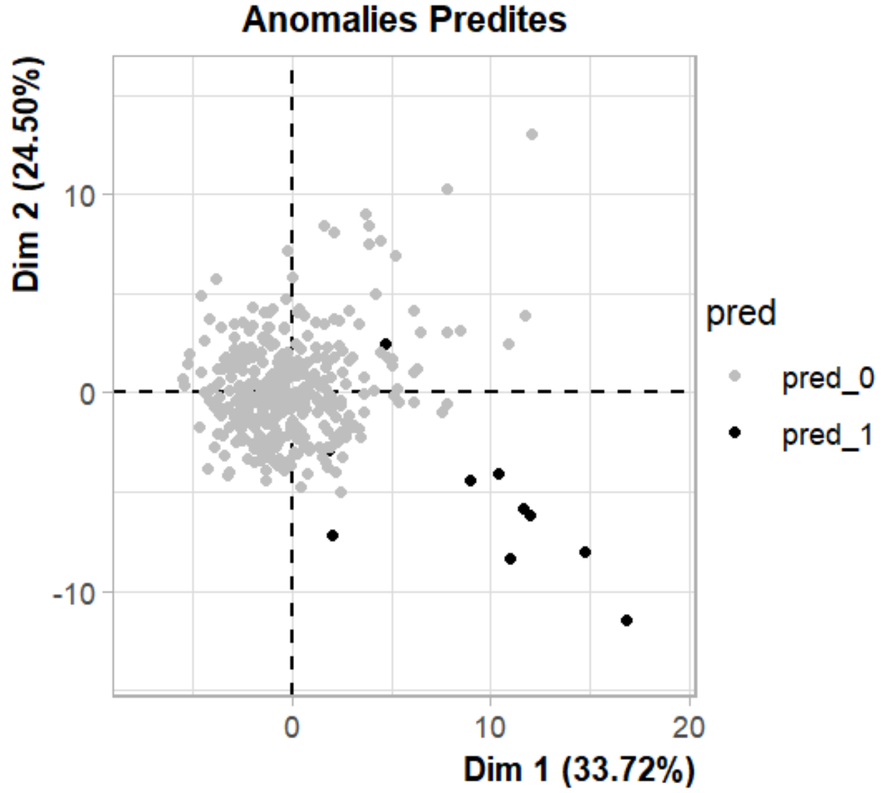



```
          0   1
0 356   1
1  12   9

               Accuracy : 0.9656
                 95% CI : (0.9419, 0.9816)
    No Information Rate : 0.9735
    P-Value [Acc > NIR] : 0.867406

                  Kappa : 0.5651

 Mcnemar's Test P-Value : 0.005546

            Sensitivity : 0.9674
            Specificity : 0.9000
         Pos Pred Value : 0.9972
         Neg Pred Value : 0.4286
             Prevalence : 0.9735
         Detection Rate : 0.9418
   Detection Prevalence : 0.9444
      Balanced Accuracy : 0.9337

       'Positive' Class : 0
```

Local Outlier Factor - Résultats

	0	1
0	356	1
1	12	9

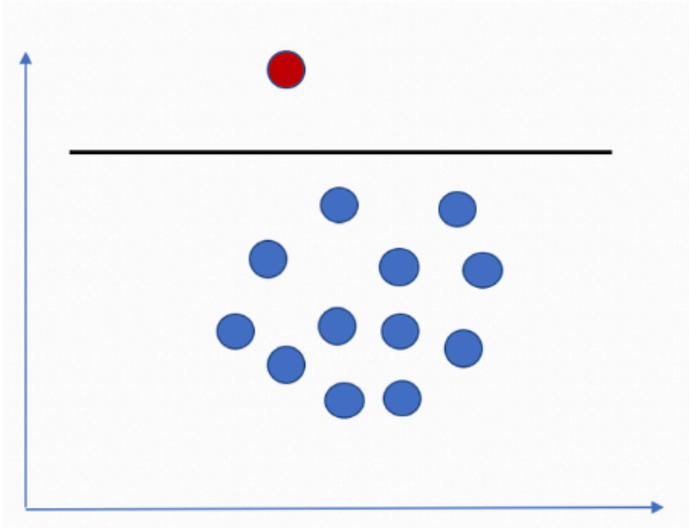




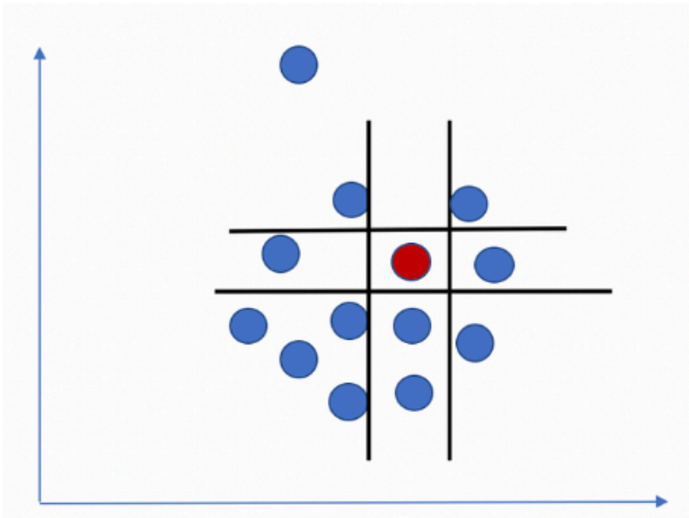
Isolation Forest - Principe

- 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

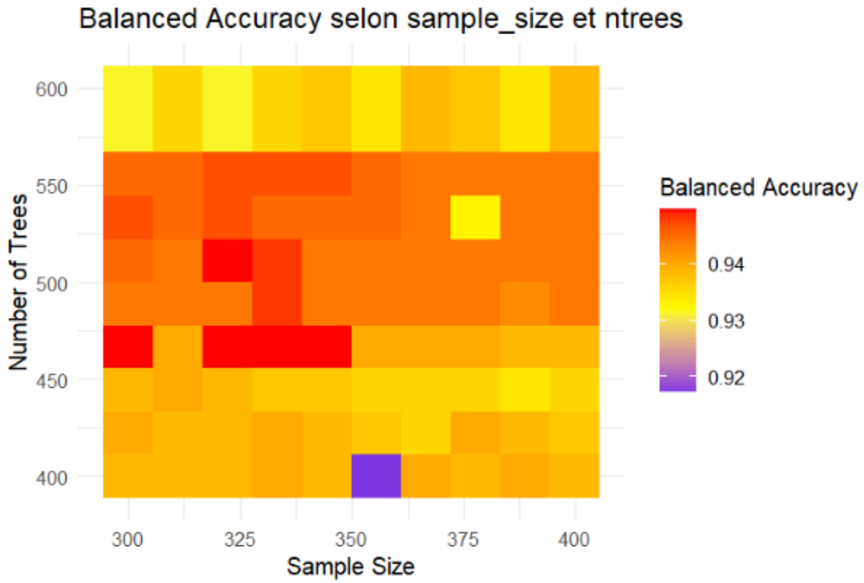
Isolating an anomalous point

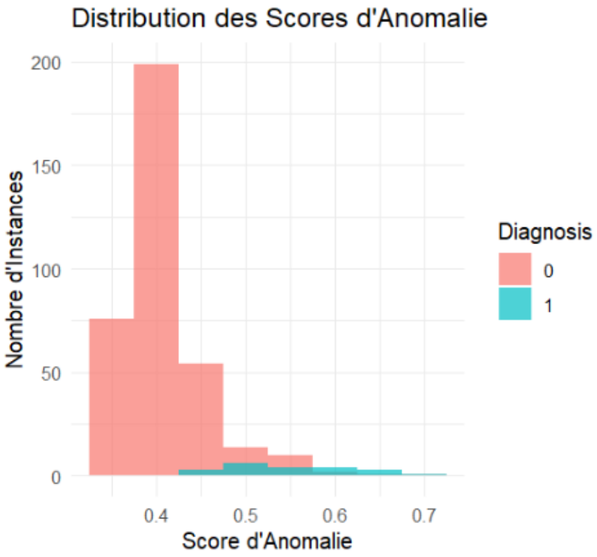


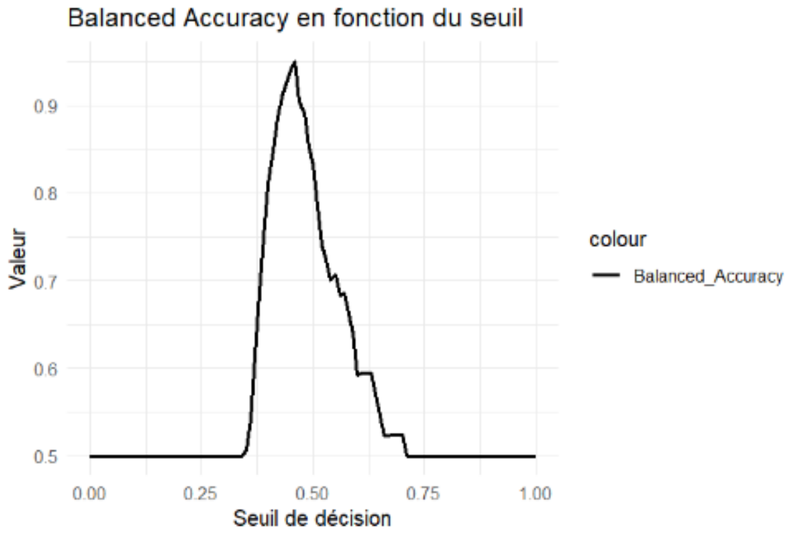
Isolating a normal point

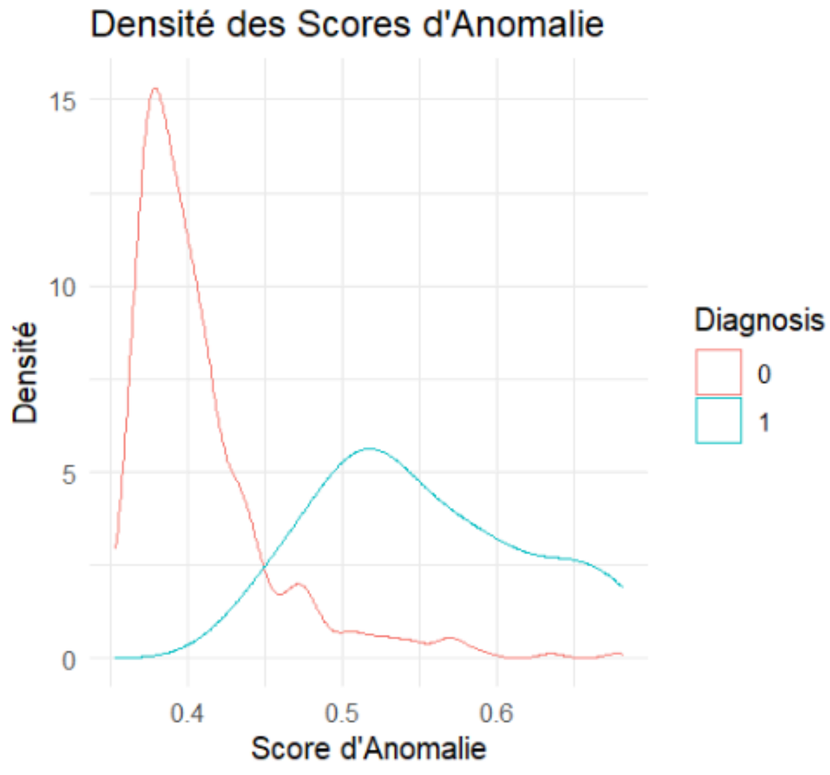


Isolation Forest - Résultats

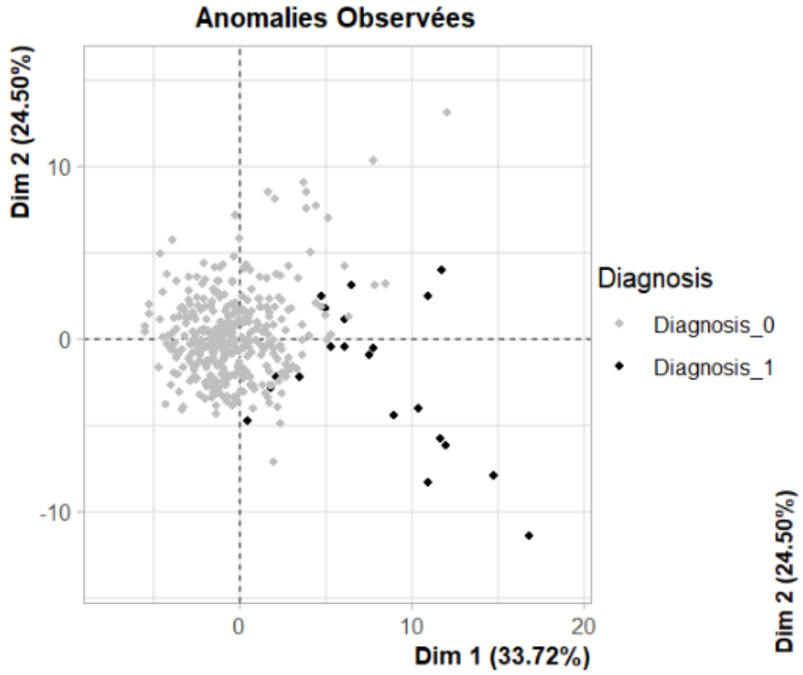


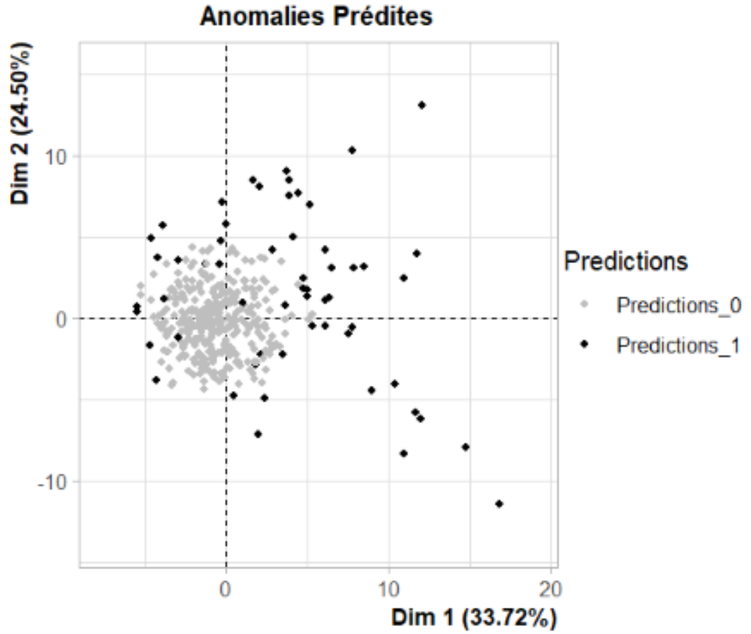






Isolation Forest - Résultats

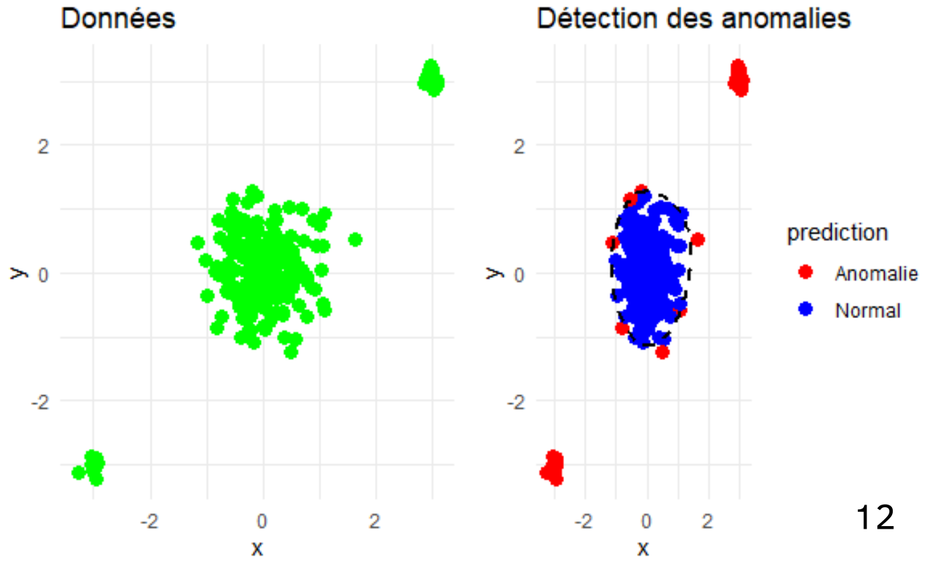





```
Confusion Matrix and Statistics

          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 ; un γ élevé génère une frontière plus complexe.

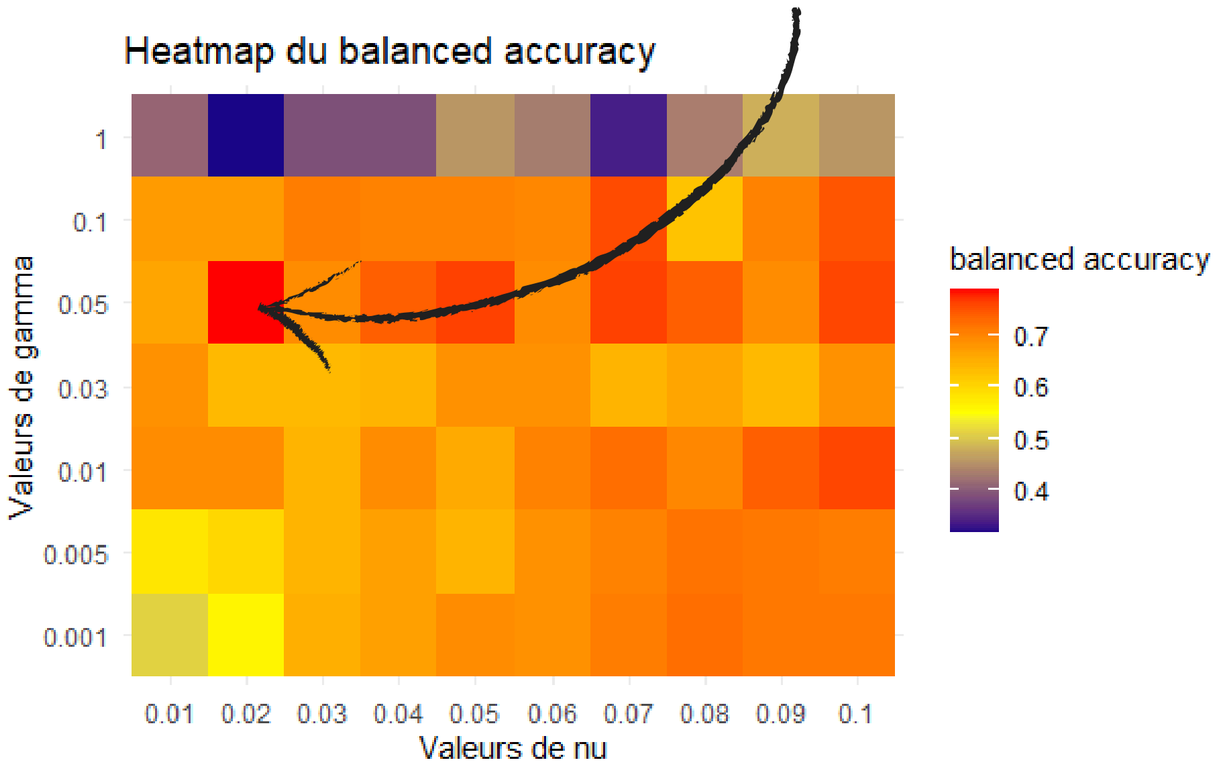


One-Class Support Vector Machine - Résultats

Sélection de la plage de 2 paramètres :

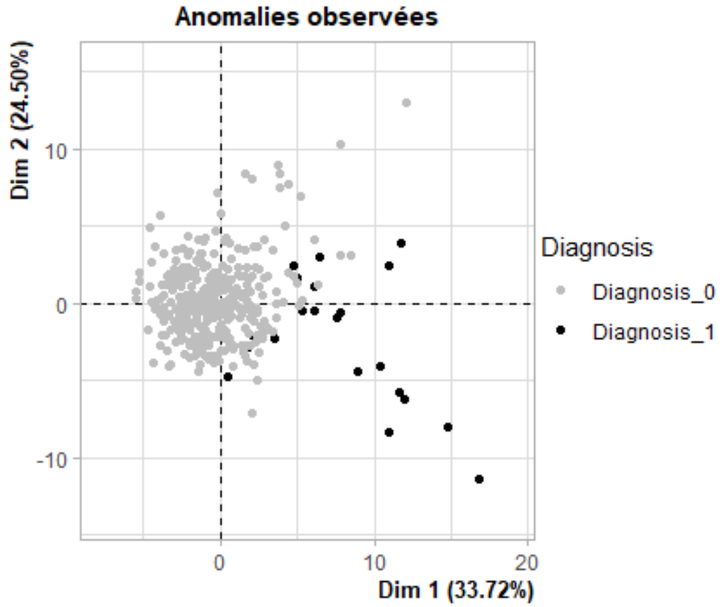
- <u>Paramètre ν :</u> en lien avec la proportion d'anomalies -> inférieur à 10%
- <u>Paramètre γ :</u> inversement proportionnel au nombre de features (ici 30 donc valeur de base = 0.03)

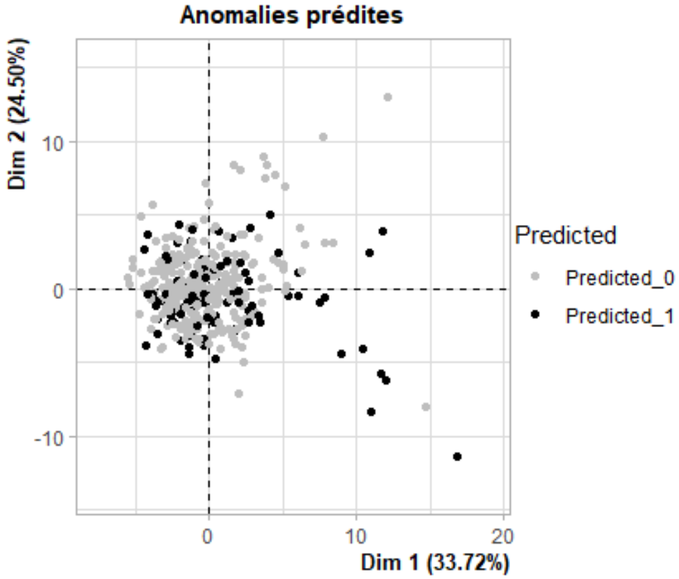
Couple de valeurs optimale



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	0.93	0.95	0.79

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)

(Togbe et al., 2020)

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 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