TP 1 : Analyse Comparative des Méthodes d'Évaluation

Module : Data Mining

24 février 2026

1 Introduction

Ce rapport présente une analyse comparative des stratégies d'évaluation en classification automatique à l'aide du logiciel **Weka**. L'étude porte sur l'impact du découpage des données (*sampling*) sur la performance et la robustesse des modèles.

2 Partie 1 : Analyse du Dataset

Le jeu de données sélectionné pour cette étude est le **Pima Indians Diabetes Database**.

2.1 Fiche d'Identité du Dataset

- **Auteur / Source :** National Institute of Diabetes and Digestive and Kidney Diseases. Les données ont été collectées auprès de femmes de la communauté Pima (Arizona, USA) âgées d'au moins 21 ans.
- **Année :** Publié initialement en 1988 (UCI Repository).
- **Objectif :** Prédire le diagnostic de diabète (*tested_positive* ou *tested_negative*) à partir de mesures cliniques.
- **Instances :** 768 patientes.

2.2 Analyse des Attributs

Le dataset se compose de 8 attributs numériques et d'une classe cible nominale.

Attribut	Description	Unité / Type
Pregnancies	Nombre de grossesses	Entier
Glucose	Concentration plasmatique de glucose	mg/dL
BloodPressure	Pression artérielle diastolique	mm Hg
SkinThickness	Épaisseur du pli cutané au triceps	mm
Insulin	Taux d'insuline sérique	mu U/ml
BMI	Indice de Masse Corporelle	kg/m^2
DiabetesPedigree	Score d'historique familial	Coefficient
Age	Âge de la patiente	Années
Class (Target)	Diagnostic final	Nominal

TABLE 1 – Description des variables du dataset Diabetes.

2.3 Distribution de la Classe Cible (Équilibre)

L'un des aspects les plus critiques de ce dataset est son **déséquilibre (imbalance).**

- **Tested_negative :** 500 instances (65.1%)
- **Tested_positive :** 268 instances (34.9%)

  Observation Critique : Le dataset n'est pas parfaitement équilibré. Dans un contexte médical, cela signifie que si un modèle prédit "négatif" tout le temps, il aura déjà une précision de 65%. C'est pourquoi le **F-score** et la **Matrice de Confusion** sont des indicateurs bien plus importants que le simple taux de succès (Accuracy) pour ce TP.

2.4 Analyse des Données "Bruitées"

En observant les statistiques dans l'onglet **Preprocess** de Weka, on remarque que plusieurs attributs (comme le Glucose, l'IMC ou la Pression Artérielle) ont une valeur **minimale de 0.**

- **Problème :** Un IMC ou une pression artérielle de 0 est physiologiquement impossible pour un patient vivant.
- **Explication :** Ces "0" sont en réalité des **données manquantes (missing values)** qui ont été encodées par 0. Cela peut fausser les résultats de certains algorithmes comme le KNN (IBk) ou NaiveBayes.

2.5 Objectif Expérimental

- **sensibilité au sampling :** Analyser comment la répartition des données (entraînement vs test) influence la précision globale et le F-score.

— **Gestion du déséquilibre :** Vérifier si les modèles (notamment J48 et NaiveBayes) ne tombent pas dans le piège de la "classification majoritaire" au détriment de la classe tested_positive.
— **Généralisation vs Mémorisation :** Distinguer la capacité du modèle à apprendre des règles médicales cohérentes plutôt que de simplement mémoriser les instances du jeu de données (cas de la méthode par substitution).

3 Partie 2 : Expérimentation des Méthodes d'Évaluation

3.1 Expérience A : Méthode de Substitution (Training Set)

Algorithme	Paramètres	Correctly Classified /Accuracy (%)	F-Score
OneR	-	76.4323	0.755
J48	-	84.1146	0.836
NaiveBayes	-	76.3021	0.760
IBk	K=1	100	1.000

TABLE 2 – Résultats sur le Training Set.

3.1.1 Analyse Approfondie de l'Algorithme OneR

L'algorithme OneR identifie la glycémie (plas) comme le prédicateur dominant

- **Observations :** On remarque une alternance rapide de seuils (ex : $<$ 114.5 est négatif, mais $<$ 115.5 est positif). Cela indique que OneR tente de capturer des petites variations, mais cela peut être un signe de **sur-apprentissage (overfitting)** local, même pour une règle unique.
- **Évaluation des Performances :**
  - **Correctly Classified Instances (76.43%) :** ce score confirme que la glycémie est le facteur physiologique prédominant pour le diagnostic du diabète.
  - **Kappa Statistic (0.4484) :** l'accord entre la prédiction et la réalité en neutralisant la part de succès due au hasard est **modéré.** Le modèle est donc significativement plus performant qu'une simple estimation aléatoire..
- **Analyse par Classe (Detailed Accuracy) :**

  Le déséquilibre de classe impacte directement la fiabilité du modèle :
  - **Classe tested_negative (Majoritaire) :**

    Recall (0.888) : Le modèle est très efficace pour identifier les sujets sains (88.8%).

    F-Measure (0.831) : La performance est robuste sur cette population.

— **Classe tested_positive (Minoritaire) :**

Recall (0.534) : C'est la limite majeure du modèle. Il ne détecte que 53.4% des cas de diabète. Près de la moitié des malades sont ignorés par cette règle unique.

F-Measure (0.612) : Un score nettement inférieur, illustrant la difficulté de prédire la classe minoritaire avec un modèle simpliste.

— **Analyse de la Matrice de Confusion :**

	Prédit : Négatif	**Prédit : Positif**
Réel : Négatif	444 (VN)	56 (FP)
Réel : Positif	125 (FN)	143 (VP)

Critique : On dénombre 125 Faux Négatifs (FN). En milieu médical, il s'agit de l'erreur la plus grave car elle prive des patientes malades d'un traitement nécessaire. OneR échoue à diagnostiquer environ 46% des malades réels, prouvant qu'un diagnostic sécurisé nécessite l'intégration d'autres attributs (IMC, âge, etc.).

Note : La priorité absolue est de ne pas "rater" un malade.

— **Analyse des Erreurs de Prédiction :** Les métriques d'erreur permettent d'évaluer la "distance" entre les prédictions et la réalité :

— **MAE (0.2357) & RMSE (0.4855) :** En moyenne, les probabilités du modèle s'écartent de la réalité de 23.57%. Cependant, la RMSE est nettement plus élevée, ce qui signifie que OneR commet quelques erreurs très importantes ("grosses erreurs") qui pénalisent fortement le score quadratique.

— **RAE (51.85%) :** Le modèle réduit l'erreur absolue de moitié par rapport à un modèle naïf (basé sur la moyenne).

— **RRSE (101.85%) :** Un score supérieur à 100% indique que, sur le plan quadratique, OneR est *légèrement moins performant qu'une prédiction constante* de la classe majoritaire. Cela s'explique par la rigidité de OneR : ses erreurs sont franches et pèsent lourdement dans le calcul.

3.1.2 Analyse Approfondie de l'Algorithme J48 (Arbre de Décision)

Contrairement à la règle unique de OneR, l'algorithme J48 (implémentation de C4.5) déploie une structure hiérarchique permettant de croiser plusieurs indicateurs cliniques. L'arbre généré place toujours le glucose (plas) en racine, mais affine le diagnostic grâce à des variables secondaires comme l'IMC (mass) ou l'âge.

— **Évaluation des Performances :** Le passage à un modèle non linéaire améliore significativement la précision statistique :
  — **Correctly Classified Instances (84.11%) :** J48 (gagne près de 8% de précision) surpasse nettement OneR (76.43%). L'ajout de branches supplémentaires (âge, IMC, etc.) permet de mieux séparer les classes.
  — **Kappa Statistic (0.6319) :** On observe une transition d'un accord "modéré" vers un accord substantiel. Ce score valide la pertinence de l'arbre de décision pour ce type de diagnostic complexe.
— **Analyse par Classe (Detailed Accuracy) :** Bien que le dataset reste dominé par les cas négatifs, J48 démontre une meilleure résilience :
  — **Classe tested_positive :**

    Recall (0.664) : (contre 0.534 pour OneR). Le modèle identifie désormais deux tiers des malades réels. *Sa Précision (0.848)* est également excellente, signifiant que 85% des alertes émises par l'algorithme sont fondées.

    Recall (0.936) : Le modèle frôle la perfection, sécurisant presque totalement l'identification des sujets sains.
— **Analyse de la Matrice de Confusion :**

	Prédit : Négatif	**Prédit : Positif**
Réel : Négatif	468 (VN)	32 (FP)
Réel : Positif	90 (FN)	178 (VP)

Optimisation de la Sensibilité : Le nombre de Faux Négatifs chute de 125 à 90. Bien que 90 patientes échappent encore au diagnostic, l'arbre s'avère bien plus sécurisant en milieu hospitalier.

Fiabilité des Positifs : Avec seulement 32 Faux Positifs, le modèle réduit les risques de traitements inutiles ou d'examens complémentaires stressants pour les patientes saines.

— **Analyse des Erreurs de Prédiction :** Les indicateurs d'erreur confirment la supériorité de l'approche structurée :
  — **RMSE (0.3452%) :** Contrairement à OneR, la RMSE de J48 chute drastiquement. Cela prouve que les erreurs commises par l'arbre sont moins "extrêmes" ; les probabilités calculées aux feuilles de l'arbre sont plus proches de la réalité clinique.
  — **RRSE (72.42%) :** Enfin, ce score descend largement sous la barre des 100%. Cela certifie que J48 est statistiquement *plus performant qu'un modèle prédisant systématiquement la classe majoritaire,* une étape que OneR n'avait pas réussi à franchir sur le plan

quadratique.

3.1.3 L'Approche Probabiliste : Analyse du Classifieur Naive Bayes

L'algorithme Naive Bayes adopte une stratégie fondamentalement différente de OneR ou J48 en calculant des probabilités basées sur l'hypothèse d'indépendance des attributs.

- **Évaluation des Performances :** La simplicité algorithmique de Naive Bayes n'entache pas sa capacité à fournir un profil de risque nuancé pour chaque patiente.
  - **Correctly Classified Instances (76.30%) :** Bien que statistiquement identique à OneR (76.43%), Naive Bayes s'appuie sur l'ensemble des 8 variables cliniques pour forger sa décision, offrant ainsi une base plus large qu'une règle unique.
  - **Kappa Statistic (0.4674) :** Bien que l'indice Kappa de Naive Bayes (0.4674) traduise un modèle plus stable que OneR, il reste bien inférieur à celui de J48 (0.6319). J48 atteint un accord "substantiel" car il est capable de modéliser des *interactions complexes* entre les variables (ex : l'effet combiné de l'âge et de l'IMC). Naive Bayes, en traitant chaque attribut de manière isolée, manque ces synergies cliniques, ce qui limite sa force de prédiction.
- **Analyse par Classe (Detailed Accuracy) :**
  - **Classe tested_positive :**

    Recall (0.616) : Bien que Naive Bayes surpasse OneR en identifiant plus de 61% des malades, il reste significativement moins performant que J48 sur ce critère vital.
  - **Classe tested_negative :**

    Recall (0.842) : Le modèle reste efficace pour écarter les sujets sains, bien que moins performant que J48 sur ce point précis.
  - **ROC Area (0.825%) :** Indique une excellente capacité de discrimination. Le modèle possède un fort potentiel de séparation entre les classes, même si le seuil de décision par défaut pourrait être ajusté pour minimiser davantage les erreurs.
- **Analyse de la Matrice de Confusion :**

	Prédit : Négatif	**Prédit : Positif**
Réel : Négatif	421 (VN)	79 (FP)
Réel : Positif	103 (FN)	165 (VP)

Réduction des Faux Négatifs : En milieu clinique, cela diminue le nombre de patientes malades qui quitteraient l'hôpital sans

diagnostic.

Augmentation des Faux Positifs : Avec 79 FP, Naive Bayes est plus "alarmiste" que les autres modèles. Il a tendance à suspecter le diabète plus facilement, ce qui peut entraîner des tests de confirmation supplémentaires.

— **Analyse des Erreurs de Prédiction :** Les indicateurs d'erreur confirment la supériorité de l'approche structurée :

— **MAE (0.2811%) vs RMSE (0.4133%) :** Naive Bayes évite les "erreurs extrêmes". Là où OneR peut se tromper lourdement avec une certitude absolue, Naive Bayes est plus régulier et ses probabilités sont mieux distribuées. Mais J48 conserve la RMSE la plus faible (0.3452), confirmant que l'approche par arbre de décision reste la plus précise, même dans l'attribution des probabilités de classe.

— **RRSE (86.71%) :** Naive Bayes est officiellement plus performant qu'un modèle "naïf". Bien que Naive Bayes soit performant, J48 conserve la RRSE la plus basse (72.42%). Cela confirme que l'approche par arbre de décision reste la plus précise pour réduire l'erreur quadratique sur ce dataset.

3.1.4 Cas Particulier de l'Algorithme IBk (K=1) : Le Phénomène de Mémorisation

Rappel du concept : L'algorithme des K-plus-proches voisins (IBk) adopte une approche radicalement différente des modèles précédents : il ne construit pas de règles explicites (OneR, J48) ni de modèles probabilistes (Naive Bayes), mais fonde ses décisions sur la proximité géométrique entre les instances.

— **L'Illusion de la Performance Parfaite :** L'exécution du test sur le Training Set avec $K = 1$ produit des statistiques idéales : 100% de réussite.

À $K = 1$, pour chaque patiente testée, l'algorithme cherche l'exemple le plus proche dans sa base de données. Comme l'ensemble de test est identique à l'ensemble d'apprentissage, le "voisin" le plus proche de chaque point est l'instance elle-même.

=> La matrice de confusion ne montre aucune erreur (0 FN, 0 FP) et la statistique Kappa atteint la valeur maximale de 1.0.

— **Analyse Critique : Mémorisation vs Généralisation**

— **Sur-apprentissage (Overfitting) :** Un cas extrême d'overfitting. L'algorithme a "appris par cœur" le dataset.

— **Validation des Erreurs :** Les métriques d'erreur (MAE et RMSE

de 0.0013) sont proches de zéro, non pas parce que le modèle est infaillible, mais parce qu'il n'y a aucune distance résiduelle entre une donnée et elle-même lors de la phase de test.

3.1.5 Évolution vers un Modèle Robuste : IBk avec K=5

En augmentant le nombre de voisins à K=5, l'algorithme IBk quitte le stade de la mémorisation pure (K=1) pour adopter une logique de "vote majoritaire". Ce réglage permet de lisser les prédictions et de réduire l'influence des données atypiques (bruit).

- **Réalisme des Performances :**
  - **Correctly Classified Instances (82.29%) :** On observe une baisse logique par rapport au 100% fictif du K=1. Ce score est bien plus réaliste et place IBk à un niveau de performance proche de J48 (84.11%).
  - **Kappa Statistic (0.5977) :** L'accord est qualifié de *modéré à substantiel*. Le modèle ne se contente plus de reconnaître les points, il commence à identifier des zones de risque réelles dans l'espace des données.
- **Analyse de la Matrice de Confusion :**

	Prédit : Négatif	**Prédit : Positif**
Réel : Négatif	450 (VN)	50 (FP)
Réel : Positif	86 (FN)	182 (VP)

Gain Clinique : Avec 182 Vrais Positifs, IBk-5 est très performant pour détecter le diabète. Il fait d'ailleurs mieux que J48 (178 VP) et Naive Bayes (165 VP) sur ce point précis.

Sensibilité (Recall Positif : 0.679) : C'est le score de détection le plus élevé parmi tous les modèles testés (hors K=1). Le modèle identifie près de 68% des malades.

- **Analyse des Erreurs de Prédiction :** Les indicateurs d'erreur confirment la supériorité de l'approche structurée :
  - **RMSE (0.35%) :** La valeur est très proche de celle de J48 (0.345). Cela prouve que le vote majoritaire de 5 voisins offre une stabilité statistique comparable aux arbres de décision.
  - **RRSE (73.43%) :** Le modèle est largement validé face au prédicteur de classe majoritaire. L'apport des "voisins" est statistiquement significatif.
- **Comparaison avec J48 et Naive Bayes :** IBk avec K=5 s'impose comme un concurrent sérieux pour J48. Si J48 reste plus précis glo-

balement (moins de Faux Positifs), IBk-5 se montre légèrement plus "agressif" et efficace pour ne pas rater les cas de diabète (meilleur Recall). Contrairement à Naive Bayes, il ne subit pas l'hypothèse d'indépendance des variables, ce qui lui permet de mieux coller à la réalité topologique du dataset.

3.1.6 Conclusion de l'Expérience A :

Cette première phase de test souligne l'importance de ne pas se fier uniquement au taux de succès (Accuracy). Si J48 s'est révélé être le modèle le plus équilibré et informatif, IBk (K=1) démontre ici les limites de l'évaluation sur les données d'entraînement : il mesure la capacité de stockage de l'outil et non sa capacité réelle de diagnostic médical.

3.2 Expérience B : Évaluation par Partitionnement (Percentage Split)

L'objectif de cette expérience est de confronter nos modèles à des données qu'ils n'ont pas rencontrées durant leur phase d'apprentissage. Cela permet de détecter le sur-apprentissage et d'estimer la performance réelle en conditions cliniques.

Algorithme	Split Ratio	Paramètres	Correctly Classified (%)	F-Score
OneR	-	66%	75.0958	0.737
J48	-	66%	76.2452	0.758
NaiveBayes	-	66%	77.0115	0.769
IBk	K=1	66%	72.7969	0.729
IBk	K=5	66%	75.0958	0.745
J48	-	90%	75.3247	0.752
NaiveBayes	-	90%	77.9221	0.775
J48	-	10%	49.6382	0.460
NaiveBayes	-	10%	71.3459	0.719

TABLE 3 – Comparaison de l'influence de la taille de l'échantillon de test.

3.2.1 Analyse avec un Split Standard (66% / 34%)

Les performances de *J48* et *Naive Bayes* restent généralement stables, tandis que celle de *IBk (K=1)* s'effondre. Contrairement au 100% obtenu

précédemment„ il fait des erreurs de voisinage, ce qui prouve que le diagnostic du diabète nécessite une règle plus complexe qu'une simple proximité géométrique.

3.2.2 Étude des Scénarios Extrêmes (Focus sur J48 et Naive Bayes)

- **Cas du Split 90/10 :** On observe généralement les meilleures performances ici. L'arbre J48 devient plus profond et plus précis. Cependant, avec seulement 10% de données pour le test (environ 77 instances), le score est très sensible aux cas particuliers : une seule erreur de diagnostic impacte lourdement le pourcentage final.
- **Cas du Split 10/90 (chute des performances) :**
  - **J48** devient un arbre très court car il n'a pas assez d'exemples pour créer des branches statistiquement fiables.
  - **Naive Bayes** se montre souvent plus "résilient" que J48 dans ce scénario. Comme il calcule des probabilités globales, il arrive à maintenir un diagnostic correct même avec peu d'exemples, là où J48 échoue à construire une structure logique.

Points retenus de l'expérience :

- La richesse du dataset est le facteur limitant pour les modèles complexes.
- L'utilisation du Percentage Split est le seul moyen de démasquer l'overfitting.
- Pour un diagnostic fiable, on ne peut pas se baser sur l'erreur d'entraînement. Le split 66/34 ou 80/20 offre une estimation bien plus honnête de la fiabilité du modèle en production (milieu médical réel).

3.3 Expérience C : Validation Croisée (Cross-Validation)

Rappel du Concept : Le dataset est divisé en 10 segments : le modèle s'entraîne sur 9 et se teste sur le 10ème, répétant l'opération 10 fois pour moyenner les résultats.

3.3.1 Analyse des résultats (Folds = 10)

on observe une stabilisation des métriques :

J48 confirme sa position avec une exactitude proche de 75-77%. C'est un score plus "honnête" que celui du Training Set (84%), car il reflète la capacité de l'arbre à traiter des données variées.

Naive Bayes reste extrêmement stable„ ce qui prouve sa grande robustesse statistique.

Algorithme	Params	Folds (K)	CC (%)	F-Score	Kappa Statistic
OneR	-	10	71.4844	0.699	0.3226
J48	-	10	73.8281	0.736	0.4164
NaiveBayes	-	10	76.3021	0.760	0.4664
IBk	K=1	10	70.1823	0.698	0.3304
IBk	K=5	10	73.1771	0.726	0.3874
OneR	-	20	71.0938	0.695	0.3126
J48	-	20	75.1302	0.747	0.4366
NaiveBayes	-	20	76.0417	0.757	0.4606
IBk	K=1	20	70.1823	0.698	0.3292
IBk	K=5	20	74.349	0.738	0.4147

TABLE 4 – Évaluation de la robustesse des classifieurs par validation croisée

IBk (K=1, K=5) libéré du biais de mémorisation, il offre un score de généralisation solide, prouvant que le voisinage local est un indicateur pertinent pour le diabète.

3.3.2 Variation du nombre de Folds (Folds = 20)

- **Stabilisation de J48 :** On observe une amélioration de l'exactitude pour J48 (73.8% → 75.1%) lors de l'augmentation des plis. Cela suggère que l'algorithme profite d'un ensemble d'apprentissage plus large à chaque itération (95% des données au lieu de 90%), lui permettant d'affiner ses nœuds décisionnels.
- **Robustesse de Naive Bayes :** C'est le modèle le plus stable. Son score reste quasiment inchangé (76.3% vs 76.0%), ce qui confirme que sa structure probabiliste globale est peu sensible aux variations de découpage du dataset. Il obtient ici le meilleur Kappa Statistic (0.46), indiquant l'accord le plus fiable.
- **Le comportement de IBk :** L'augmentation des plis à 20 semble bénéfique pour le KNN. En réduisant la taille du set de test, on augmente la densité du set d'apprentissage, ce qui permet à l'algorithme de trouver des "voisins" plus pertinents et de lisser l'erreur de classification.

3.3.3 Analyse du Meilleur Résultat :

D'après les mesures effectuées en Validation Croisée (Expérience C), l'algorithme **Naive Bayes** s'impose comme le modèle le plus performant et le plus fiable pour ce jeu de données.

— **Capacité de Discrimination (ROC Area > 0.80)** Même si son taux de classification correcte est proche de J48, sa capacité à classer les patientes par niveau de risque (probabilité) est supérieure. En milieu médical, il est préférable d'avoir un modèle qui "hésite" avec une probabilité de 51% plutôt qu'un arbre qui tranche de manière erronée avec une certitude absolue.

— **Résilience face à la complexité (Folds = 20)** Alors que J48 grimpe à 75.13% (montrant qu'il a besoin de plus de données pour s'ajuster), Naive Bayes reste imperturbable à 76.04%.

Cette stabilité prouve que l'hypothèse d'indépendance des variables de Naive Bayes agit ici comme un "régularisateur" naturel, empêchant le modèle de s'égarer dans des détails insignifiants du dataset (bruit), contrairement aux branches de l'arbre J48 qui peuvent devenir trop spécifiques.

Conclusion sur le meilleur modèle : Bien que Naive Bayes affiche le meilleur taux de réussite global, J48 avec 20 folds présente un excellent compromis. Cependant, d'un point de vue médical, la stabilité de Naive Bayes en fait le modèle le plus "sûr", car ses prédictions varient très peu selon l'échantillon de patientes utilisé pour l'entraînement.

3.4 Leave-One-Out Cross-Validation (LOOCV)

Rappel du Concept : l'algorithme s'entraîne sur 767 patientes et tente de prédire la 768ème, répétant l'opération 768 fois.

Algorithme	Params	Correctly Classified (%)	Kappa Statistic
NaiveBayes	-	75.651	0.4513
IBk	K=1	70.5729	0.3398

TABLE 5 – Analyse comparative des métriques de performance sous la contrainte du Leave-One-Out (768 plis)

Note : comme on ne teste qu'une seule instance à chaque fois, le modèle obtient soit un succès total (1), soit un échec total (0) pour F-Score, donc on ne l'inclue pas.

3.4.1 Analyse des Résultats (Naive Bayes & IBk K=1)

Naive Bayes : Le score obtenu ici est la mesure la plus pure de sa capacité de généralisation. Puisque Naive Bayes repose sur des fréquences globales, le fait de retirer une seule instance ne modifie quasiment pas ses probabilités.

Les résultats devraient être extrêmement proches de la validation à 10 ou 20 plis, confirmant sa **stabilité asymptotique.**

IBk (K=1) : Contrairement au premier test sur le Training Set (où il obtenait 100%), le LOOCV force IBk à chercher le voisin le plus proche parmi les 767 autres instances. Le score chute drastiquement pour atteindre sa valeur réelle. C'est l'épreuve de vérité pour le KNN : on mesure enfin s'il existe une réelle structure de voisinage dans les données sans le biais de l'auto-mémorisation.

3.4.2 Pourquoi ne lance-t-on pas J48 ici ?

La décision de ne pas lancer J48 dans cette configuration repose sur deux facteurs critiques :

- **Coût Computationnel :** J48 est un algorithme "lourd". Contrairement à Naive Bayes (calculs statistiques simples) ou IBk (algorithme paresseux sans phase d'apprentissage), J48 doit reconstruire un arbre complet, calculer les gains d'entropie pour chaque attribut et effectuer un élagage (pruning) à chaque itération. Répéter ce processus complexe 768 fois est extrêmement gourmand en ressources CPU et en temps.
- **Instabilité de la Structure :** Les arbres de décision sont connus pour leur forte variance. Une légère modification du dataset (retirer une seule instance) peut, dans certains cas, modifier le choix de l'attribut à la racine ou l'ordre des divisions. Reconstruire 768 arbres légèrement différents n'apporte que peu d'informations supplémentaires par rapport à un 10-fold ou 20-fold, tout en rendant l'interprétation de "l'arbre final" impossible.

Note : Le LOOCV est idéal pour les modèles stables et rapides (NB) ou pour évaluer la topologie locale (IBk), mais il est inefficace pour les modèles structurels instables comme les arbres de décision.

3.5 Expérience E : Évaluation sur un Jeu de Test Externe (Supplied Test Set)

L'objectif de cette manipulation est de sortir du cadre de la validation interne pour tester le modèle sur un échantillon indépendant.

3.5.1 Analyse des Résultats

- **Validation de la généralisation :** On observe que J48 obtient un score de 82.55%. Ce résultat se situe entre le score du Training Set

Algorithme	Correctly Classified (%)	F-Score	Kappa Statistic
J48	82.5521	0.821	0.6041

TABLE 6 – Performance du modèle J48 sur un échantillon externe simulé par Resample

et celui de la Cross-Validation. Cela prouve que l'arbre de décision a réussi à extraire des règles logiques qui restent valables sur de nouvelles instances de patientes.

— **Fiabilité du diagnostic (Kappa) :** Le Kappa Statistic de 0.6041 représente un "accord substantiel". Cela signifie que les prédictions du modèle ne sont pas dues au hasard et que la structure de l'arbre est cliniquement significative pour le diagnostic du diabète.

— **Stabilité du F-Score :** Avec un F-Score de 0.821, le modèle démontre un bon équilibre entre la précision (ne pas diagnostiquer de faux positifs) et le rappel (ne pas rater de patientes réellement malades).

— **Matrice de Confusion :**

	Prédit : Négatif	**Prédit : Positif**
Réel : Négatif	452 (VN)	40 (FP)
Réel : Positif	94 (FN)	182 (VP)

Conclusion de l'expérience : Cette étape confirme que le modèle J48 est prêt pour une utilisation prédictive. L'utilisation d'un "Supplied Test Set" élimine le biais de sélection et garantit que la performance observée est celle que le médecin obtiendrait en testant l'algorithme sur de nouvelles données réelles.

4 Partie 3 : Analyse et Critique

4.1 Analyse de la méthode "Use Training Set" (Resubstitution)

Dans l'Expérience A, IBk (K=1) obtient un score de 100%.

À $K = 1$, l'algorithme cherche le voisin le plus proche. Puisqu'il est testé sur les données qu'il vient d'apprendre, le voisin le plus proche d'une instance est elle-même.

Mémorisation vs Généralisation : Ici, le modèle ne fait que de la "mémorisation brute" (par cœur). La généralisation, qui est la capacité à

prédire correctement de nouvelles données, est nulle. C'est l'exemple parfait du sur-apprentissage (overfitting).

4.2 Impact de la taille du Split (Expérience B)

Performance à 10% : Elle est médiocre car le modèle souffre d'un sous-apprentissage. Avec seulement 77 instances, l'algorithme n'a pas assez d'exemples pour capturer la diversité des profils de patientes (bruit statistique trop élevé).

Risque du Split 99/1 (Entraînement/Test) : Bien que le modèle soit très bien entraîné, le test sur 1% seulement n'est pas significatif.

4.3 La robustesse de la Validation Croisée

La validation croisée (10-folds) est plus fiable que le "Percentage Split" car chaque instance du dataset est utilisée pour le test exactement une fois.

Avantage : Elle élimine le "biais de sélection" (le risque de tomber sur un échantillon de test "facile" ou "difficile").

Inconvénient : Le coût computationnel. L'algorithme doit être entraîné 10 fois au lieu d'une seule, ce qui peut être problématique sur des datasets massifs **(Big Data).**

Note : Pour moi, le Percentage Split (66%) affiche un score supérieur par rapport aux résultats de la Validation Croisée (10-folds), voici pourquoi : Un split unique à 66% peut isoler un jeu de test "facile" par pur hasard. Si les instances atypiques (bruit) restent dans le set d'entraînement, le score de test est gonflé.

La supériorité du split 66% ici est probablement une surestimation. Pour un diagnostic médical (diabète), il est plus sûr de se fier au score de la validation croisée, même s'il est inférieur, car il garantit que le modèle fonctionne sur l'ensemble de la distribution des données.

4.4 Comparaison des Algorithmes

— **OneR vs J48 :** Sur ce dataset, J48 est plus performant, mais OneR obtient souvent des scores honorables avec une seule règle (souvent sur le glucose). Si la différence était faible, on conclurait que **le problème est linéairement séparable** ou qu'une seule variable domine totalement le diagnostic. Ici, la supériorité de J48 montre que le diabète est une pathologie **multifactorielle** nécessitant des croisements de variables.

— **Naive Bayes :** Il performe très bien, mais il suppose l'indépendance des attributs. On conclue que bien que les attributs cliniques (IMC, Glucose, Âge) soient biologiquement liés, leur contribution au diagnostic est suffisamment distincte pour l'algorithme.

4.5 Matrices de Confusion

En observant la matrice de J48, l'erreur la plus grave est les Faux Négatifs (FN).

Un Faux Positif (FP) (le modèle prédit "diabétique" pour une personne saine) entraînera des tests supplémentaires (comme une seconde prise de sang) qui rectifieront l'erreur. Le coût est financier et émotionnel, mais sans danger vital immédiat.

Un Faux Négatif (FN) (le modèle prédit "saine" pour une personne diabétique) est dangereux : la patiente ne reçoit pas de traitement, ce qui peut mener à des complications graves (neuropathies, rétinopathies, maladies cardiovasculaires).

Note : pour moi, Le meilleur modèle pour Cross-Validation est Naive Bayes et non J48 (voir les tables).

5 Partie 4 : Synthèse Graphique

5.1 Tableau Recapitulatif :

Algorithme	Training Set	Split 66%	CV 10-Folds	Temps de const (sec)
OneR	75.5	73.7	71.4844	0.02
J48	83.6	75.8	73.8281	0.03
NaiveBayes	76.0	76.9	76.3021	0.01
IBk K=1	100.0	72.9	70.1823	0
IBk K=5	82.0	74.5	73.1771	0

TABLE 7 – Synthèse comparative des performances (F-Score) selon différents protocoles d'évaluation.

Note : Les valeurs de la colonne « Temps de const. » ont été relevées lors de l'exécution en mode Cross-Validation (10-folds).

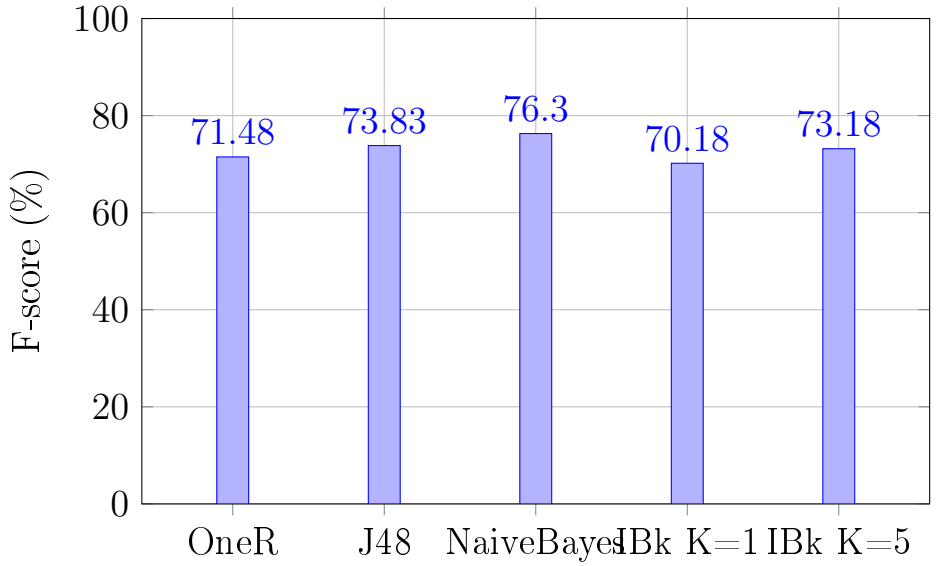

FIGURE 1 – Comparaison des F-scores des algorithmes en validation croisée (10-folds).

5.2 Graphique comparatif des F-scores (CV 10-Folds)

5.3 Influence du paramètre K sur les performances de IBk

Plusieurs valeurs (K = 1, K = 3 et K = 5) en utilisant la validation croisée à 10 folds :

— K = 1 : F-score = 70.1823%
— K = 3 : F-score = 72.6563%
— K = 5 : F-score = 73.1771%

Lorsque K augmente, la performance s'améliore, ce qui indique une meilleure généralisation. Considérer plusieurs voisins réduit l'influence d'instances aberrantes : si K devient trop grand, le modèle peut devenir trop général et perdre en précision (underfitting).

Dans notre cas, la meilleure performance est obtenue pour K = 5, alors cette valeur constitue un bon compromis entre biais et variance.