MATHÉMATIQUES ET INFORMATIQUE
Sciences
Université Paris Cité

Projet :

Introduction à la réconnaissance des formes

Laurent Wendling

Daniel Latorre
22313677

Contents

1 Introduction :

1.1 Contexte et Objectif du Projet

Dans le domaine en constante évolution de l'informatique, qui embrasse des défis complexes à l'intersection de l'analyse d'images, de l'apprentissage automatique et de l'intelligence artificielle, la reconnaissance des formes se présente comme un secteur clé nécessitant une compréhension approfondie des techniques et méthodologies avancées. Ce projet, issu du cursus d'Introduction à la reconnaissance des formes, vise à explorer de manière exhaustive les approches classiques de reconnaissance des formes. Il met ces approches à l'épreuve contre les défis réalistes présentés par la base de données BDshape, avec pour objectif d'évaluer et comparer ces méthodes en termes de précision, de robustesse et d'efficacité dans le traitement de données complexes.

1.2 Présentation de la Base de Données BDshape

La base de données BDshape, qui est l'élément central de cette étude, comprend une collection diversifiée de formes binaires réparties en neuf classes d'objets, telles que des animaux et des avions, chacune comportant onze échantillons. Ces formes, souvent partielles, occultées ou sujettes à des distorsions, offrent un cadre idéal pour tester et analyser la performance des différentes méthodologies de reconnaissance des formes.

1.3 Approfondissement des Méthodologies de Reconnaissance des Formes

Pour aborder ce projet, quatre méthodes de reconnaissance des formes ont été sélectionnées, chacune représentant une approche distincte et établie dans le domaine. La première, la Mesure du Degré d'Ellipticité (E34), repose sur l'analyse de la proximité d'une forme à une ellipse parfaite. Elle utilise la géométrie analytique et la théorie des ellipses pour évaluer le degré d'ellipticité de segments radiaux issus du centre de masse de la forme.

La seconde méthode, le Descripteur de Fourier Générique (GFD), utilise la transformation de Fourier pour convertir les formes spatiales en un profil de fréquence, permettant ainsi d'analyser leur structure globale. Cette méthode est particulièrement efficace dans les contextes où des formes complexes doivent être comparées malgré des variations de taille, de position et de petites distorsions.

La troisième méthode, la Signature Angulaire (SA), étudie les variations angulaires le long du contour d'une forme, offrant une analyse fine de la structure de contour en mesurant les angles formés par des lignes radiales issues du centre de la forme.

Enfin, la quatrième méthode, F0, se concentre sur l'évaluation des écarts d'une forme par rapport à un modèle géométrique de référence, souvent une forme géométrique simple. Elle mesure l'étendue et le type des distorsions par rapport à cette référence.

1.4 Introduction aux Méthodes de Classification

En complément des méthodes de reconnaissance des formes, ce projet envisage également l'utilisation de techniques de classification avancées, telles que les k-plus-proches voisins (knn) et les algorithmes de clustering comme les nuées dynamiques (k-means). Ces méthodes seront explorées pour leur potentiel à classer efficacement les formes analysées. Leur application, efficacité, et intégration avec les méthodes de reconnaissance des formes seront examinées en détail dans les sections ultérieures du rapport.

1.5 Objectif Principal

Ce projet a pour ambition de fournir une évaluation comparative des méthodes conventionnelles de reconnaissance des formes dans un contexte moderne. Notre objectif est d'enrichir notre compréhension de ces techniques et d'explorer des améliorations potentielles basées sur les résultats obtenus. Ce travail représente une opportunité précieuse d'appliquer nos connaissances théoriques à un problème concret, enrichissant ainsi notre expérience académique et professionnelle dans le domaine de l'informatique.

2 Méthodologie

2.1 Représentation des Données

La transformation des données brutes en un format interprétable par des algorithmes de classification nécessite une sélection méticuleuse des descripteurs qui peuvent capturer l'essence des formes en tenant compte de leurs propriétés topologiques et géométriques. Quatre descripteurs classiques ont été intégrés, détaillés ci-dessous, reconnus pour leur capacité à encapsuler des caractéristiques distinctes pour une variété de formes.

2.1.1 Mesure du Degré d'Ellipticité (E34):

Le degré d'ellipticité (E34) est une mesure quantitative dérivée des principes de l'analyse de forme et des moments géométriques. Il établit un cadre pour l'extraction de caractéristiques invariantes en considérant une forme comme une distribution de masse et en utilisant les moments jusqu'à un ordre défini pour caractériser sa similitude avec une ellipse canonique.

- **Application :** l'E34 est intrinsèquement liée à la caractérisation des objets dans des domaines où la symétrie et la régularité elliptique sont des indicateurs de phénomènes naturels ou de processus artificiels, comme dans l'analyse de formes biologiques à des fins de classification taxonomique ou dans la caractérisation des corps célestes pour des études astronomiques.
- **Calcul :** l'E34 est dérivé en calculant les moments centraux jusqu'au second ordre pour la forme en question. L'ellipse théorique de référence est déterminée par un ensemble de paramètres qui minimisent la somme des distances radiales pondérées au carré entre la périphérie de la forme et celle de l'ellipse. Les moments sont ensuite utilisés pour calculer un indice d'ellipticité global, qui reflète la similarité géométrique entre la forme étudiée et l'ellipse de référence, ajustée en fonction de la distribution de la densité de la forme.

2.1.2 Descripteur de Fourier Générique (GFD):

Le GFD est basé sur une analyse spectrale où la transformation de Fourier décompose une forme en une série de fréquences qui caractérisent son contenu spatial. Cette décomposition fournit une représentation holistique et multirésolution de la forme, capturant des informations allant des variations globales aux asymétries fines.

- **Application :** le GFD est utilisé pour identifier et comparer les formes au-delà des contraintes de positionnement, d'échelle et d'orientation, rendant cet outil particulièrement utile dans des domaines comme la classification des images de télédétection, l'analyse de textures et l'identification de motifs biologiques.
- **Calcul :** après la normalisation de la forme pour assurer l'invariance d'échelle, une transformation de Fourier bidimensionnelle est appliquée. Les fréquences sont ensuite classées et une sélection est faite pour extraire les coefficients de Fourier basse fréquence, qui sont les plus significatifs pour la forme globale, tandis que les hautes fréquences, qui représentent les détails fins et le bruit, sont souvent omises pour simplifier la représentation.

2.1.3 Signature Angulaire (SA):

La Signature Angulaire est une fonction qui trace les variations angulaires le long du contour d'une forme, produisant un ensemble de données qui résume les détails géométriques fins du contour. Elle est sensible aux subtilités locales qui sont souvent diluées dans les mesures globales.

- **Application :** la SA est une méthode de choix pour l'analyse détaillée de formes aux contours riches et informatifs, tels que les empreintes digitales, les cartes géographiques ou les caractéristiques morphologiques des feuilles des plantes.
- **Calcul :** le calcul de la SA implique l'établissement d'un système de coordonnées polaires centré sur la forme. Des rayons sont tracés depuis ce centre vers le contour extérieur, et les angles entre des vecteurs adjacents sont mesurés à intervalles réguliers. Ces mesures sont ensuite assemblées en un histogramme ou une fonction continue, formant la signature angulaire qui décrit les variations du contour de la forme.

2.1.4 Évaluation des Distorsions (F0):

F0 constitue une métrique de comparaison entre une forme donnée et une forme de référence préétablie, permettant d'évaluer la présence et l'intensité des distorsions structurelles. Elle est fondamentale pour déterminer l'intégrité géométrique ou la dégradation d'une forme.

- **Application :** cette méthode est prépondérante dans les industries de fabrication où la conformité aux spécifications est essentielle et dans les sciences de la santé pour le diagnostic de conditions pathologiques par l'analyse de déformations morphologiques.
- **Calcul :** l'évaluation de F0 nécessite la superposition de la forme à une référence géométrique standard, après quoi un calcul différentiel est effectué pour mesurer les écarts entre les deux formes. Ces différences sont exprimées par des mesures de distance point-à-point, telles que les distances euclidiennes ou les écarts de contour, ou par une analyse comparative des moments géométriques qui quantifie la distorsion en termes de distribution de la masse de la forme.

2.2 Méthodes de Classification

Cette partie du rapport se concentre sur l'examen approfondi des méthodes des k-moyennes et des k-plus Proches Voisins (knn), deux techniques fondamentales en apprentissage automatique. Nous explorerons leurs principes théoriques, leur formulation mathématique, et leurs applications pratiques. Notre objectif est de fournir une compréhension claire et détaillée de ces méthodes, soulignant leur fonctionnement, leurs avantages, et leurs limites. Cette analyse vise à éclairer les praticiens sur l'utilisation efficace et judicieuse de ces outils dans divers scénarios d'analyse de données.

2.2.1 Méthode des k-plus-proches voisins (knn)

2.2.1.1 Principes Mathématiques

La méthode des k-plus-proches voisins est une technique de classification et de régression supervisée. La classe d'un nouveau point de données est déterminée par la majorité des classes de ses k voisins les plus proches. La distance entre les points de données est souvent calculée à l'aide de la distance euclidienne, définie comme $d(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2}$, où x_i et x_j sont deux points de données et n est le nombre de caractéristiques.

2.2.1.2 Mécanismes et efficacité

knn commence par calculer la distance entre le point de test et tous les points de données existants. Il identifie ensuite les k points les plus proches et détermine la classe ou la valeur de sortie en fonction de ces voisins. Pour la classification, c'est la classe majoritaire parmi les k voisins qui est attribuée, et pour la régression, c'est la moyenne ou la médiane de leurs valeurs.

L'algorithme est basé sur le principe que des points similaires se trouvent généralement à proximité les uns des autres dans l'espace des caractéristiques. Cette méthode est donc particulièrement efficace pour les données où cette proximité traduit une similarité réelle en termes de classe ou de valeur. Sa simplicité et sa nature non paramétrique la rendent robuste et adaptable à divers types de données.

2.2.1.3 Avantages, limites et applications

La méthode des k-plus Proches Voisins (knn) est largement utilisée pour sa simplicité et sa flexibilité, ce qui la rend adaptée à une variété d'applications allant de la reconnaissance de formes à la recherche médicale. Sa force réside dans sa capacité à effectuer des classifications précises en se basant sur les caractéristiques des exemples les plus proches. Par exemple, en reconnaissance de formes, knn peut identifier la catégorie d'une nouvelle image en la comparant à des images étiquetées. Dans les systèmes de recommandation, elle peut suggérer des produits ou des films en trouvant des utilisateurs aux goûts similaires. En recherche médicale, knn aide à classifier des tumeurs comme bénignes ou malignes en analysant les cas les plus proches.

Cependant, la knn a ses limites. Sa sensibilité aux données de grande dimension peut réduire son efficacité, car dans des espaces de haute dimension, la distance entre les points de données devient moins significative, un phénomène connu sous le nom de "malédiction de la dimensionnalité". De plus, knn nécessite une importante quantité de mémoire pour stocker toutes les données de formation, ce qui peut être problématique avec de très

grands ensembles de données. La performance de la méthode dépend également du choix du nombre de voisins k, qui doit être soigneusement sélectionné pour éviter les problèmes de surajustement ou de sous-ajustement.

2.2.2 Méthode des k-moyennes (kmeans)

2.2.2.1 Principes Mathématiques

La méthode des k-moyennes est une technique de clustering non supervisée qui vise à partitionner un ensemble de données en k clusters distincts. L'objectif est de minimiser la variance intra-cluster, ce qui est mathématiquement exprimé par la somme des carrés des distances entre les points de données et le centroïde de leur cluster respectif. La fonction de coût est définie comme $J = \sum_{i=1}^{K} \sum_{x \in S_i} \|x - \mu_i\|^2$, où S_i est le i-ème cluster, x est un point de données dans S_i, et μ_i est le centroïde du i-ème cluster.

2.2.2.2 Mécanismes et efficacité

L'algorithme des k-moyennes commence par choisir aléatoirement k points comme centroïdes initiaux. Chaque point de données est ensuite attribué au cluster dont le centroïde est le plus proche, en fonction de la distance euclidienne. Après cette affectation, les centroïdes de chaque cluster sont recalculés comme étant la moyenne de tous les points assignés à ce cluster. Cette étape d'itération continue jusqu'à ce que les positions des centroïdes ne changent plus de manière significative, indiquant la convergence de l'algorithme.

L'algorithme est efficace car il optimise une fonction de coût claire, réduisant la variance au sein de chaque cluster. En minimisant les distances intra-cluster, l'algorithme regroupe les points de données de manière à ce que les membres de chaque cluster soient aussi proches que possible les uns des autres, tout en étant aussi éloignés que possible des points dans d'autres clusters. Cette approche itérative garantit que les clusters formés sont compacts et bien séparés les uns des autres, à condition que les données soient bien adaptées à ce type de regroupement.

2.2.2.3 Avantages, limites et applications

La méthode des k-moyennes est reconnue pour sa simplicité conceptuelle et son efficacité, en particulier lorsqu'il s'agit de traiter de grands ensembles de données. Cette méthode est efficace pour regrouper les données en clusters distincts, ce qui est particulièrement utile dans des domaines tels que la segmentation de marché, où elle peut aider à identifier différents groupes de consommateurs, ou en bioinformatique, où elle est utilisée pour regrouper des données génétiques ou protéiques présentant des caractéristiques similaires. Elle est également utilisée en analyse de documents pour regrouper des textes en fonction de leur contenu ou de leur thème.

Cependant, l'algorithme des k-moyennes présente certaines limitations. Il est notamment sensible à la manière dont les centroïdes initiaux sont choisis, ce qui peut influencer la qualité finale des clusters. De plus, l'algorithme tend à être moins performant avec des clusters de formes non sphériques ou de tailles très différentes, car il se base sur la minimisation de la distance euclidienne, ce qui favorise les clusters de forme sphérique. Ces limites doivent être prises en compte lors de l'utilisation de la méthode des k-moyennes pour garantir des résultats pertinents et fiables.

2.2.3 Comparaison et choix entre knn et kmeans

Dans le domaine de l'analyse de données et de l'apprentissage automatique, la sélection de la méthode appropriée entre les k-moyennes et les k-plus Proches Voisins revêt une importance capitale. Cette décision doit être guidée par une compréhension approfondie des caractéristiques et des applications de chaque méthode.

2.2.3.1 Nature et objectifs des données

La méthode des k-moyennes est une technique de clustering non supervisée, idéale pour explorer des ensembles de données où les étiquettes ou les catégories ne sont pas prédéfinies. Cette méthode est particulièrement efficace pour révéler des groupements naturels ou des structures cachées au sein des données. Son application est fréquente dans des domaines tels que la segmentation de marché, où elle aide à identifier des groupes de consommateurs avec des comportements ou des préférences similaires, ou en analyse textuelle pour regrouper des documents par thèmes.

À l'opposé, la knn est une méthode supervisée, nécessitant des données étiquetées pour la classification ou la régression. Elle est particulièrement adaptée aux situations où la relation entre les caractéristiques des données et

les catégories ou les valeurs à prédire est moins explicite. Dans des domaines tels que la reconnaissance d'images ou la prédiction médicale, la knn excelle en utilisant les étiquettes des exemples les plus proches pour prédire la classe ou la valeur d'un nouvel échantillon.

2.2.3.2 Complexité computationnelle et performance

La complexité computationnelle est un autre critère important dans le choix entre ces deux méthodes. La k-moyennes est généralement plus adaptée pour traiter de grands volumes de données en raison de sa structure algorithmique qui converge vers une solution après un nombre défini d'itérations. En revanche, la knn peut être plus exigeante en termes de calcul et de stockage, surtout avec de grands ensembles de données, car elle nécessite de comparer chaque nouvel échantillon à tous les échantillons de l'ensemble de formation.

2.2.3.3 Conclusion

En somme, la décision d'utiliser la k-moyennes ou la knn doit être prise en considération de la nature des données, de l'objectif de l'analyse et des contraintes computationnelles. Comprendre les points forts et les limites de chaque méthode permet une utilisation plus éclairée et efficace, contribuant ainsi à la réussite des projets d'analyse de données. La maîtrise de ces méthodes est donc essentielle pour les praticiens souhaitant exploiter pleinement le potentiel de l'apprentissage automatique dans divers contextes d'application.

3 Procédure de Test

3.1 Évaluation de la Performance du Modèle k-Plus-Proches Voisins (knn)

L'analyse de la performance du modèle k-plus Proches Voisins (knn) est essentielle pour déterminer son efficacité et sa fiabilité dans diverses applications. Plusieurs méthodes et métriques clés sont employées pour une évaluation complète.

3.1.1 Taux d'Erreur de Classification

3.1.1.1 Fondements Mathématiques

Le taux d'erreur de classification, calculé par la formule Taux d'Erreur $= \frac{\text{Nombre de prédictions incorrectes}}{\text{Nombre total de prédictions}}$, est une mesure directe de l'efficacité du modèle knn. Ce taux, exprimé en pourcentage, indique clairement la proportion d'erreurs commises par le modèle. Un faible taux d'erreur est synonyme d'une haute précision, reflétant l'aptitude du modèle à classer correctement les données. Inversement, un taux élevé suggère des inexactitudes dans les prédictions du modèle, signalant ainsi un besoin de révision des paramètres ou de la qualité des données utilisées pour l'entraînement.

3.1.1.2 Interprétation

Le taux d'erreur de classification, une mesure directe de la précision du modèle knn, peut varier de 0% à 100%. Un taux proche de 0% indique une précision quasi parfaite, signifiant que le modèle réussit à classifier correctement presque toutes les instances. À l'opposé, un taux approchant 100% révèle un modèle largement inexact, souvent incapable de faire des prédictions fiables. Cette mesure est cruciale pour évaluer la pertinence du modèle pour une tâche donnée, et un taux élevé peut indiquer le besoin de réexaminer les paramètres du modèle, la qualité des données ou l'adéquation du modèle lui-même.

3.1.1.3 Cas d'Utilisation et Exemples

Le taux d'erreur de classification est un indicateur statistique essentiel dans l'évaluation des modèles prédictifs en contextes professionnels. Par exemple, dans le secteur financier, l'utilisation d'un modèle knn avec un taux d'erreur minimisé est impératif pour la prévision de la solvabilité des emprunteurs. Cela permet aux institutions financières de réduire les risques de crédit. En médecine, l'application d'un modèle knn avec un faible taux d'erreur est crucial pour le diagnostic précis des maladies, contribuant ainsi à des interventions médicales plus ciblées et efficaces.

3.1.2 Matrice de Confusion

3.1.2.1 Fondements Mathématiques

La matrice de confusion est un outil statistique essentiel qui catégorise les prédictions du modèle knn en vrais positifs, faux positifs, vrais négatifs, et faux négatifs. Cette répartition permet une analyse détaillée des performances du modèle, révélant non seulement la fréquence des erreurs de classification mais aussi leur nature. Une forte occurrence de vrais positifs et de vrais négatifs indique une bonne capacité de classification du modèle, tandis que des nombres élevés de faux positifs ou de faux négatifs peuvent révéler des faiblesses spécifiques dans la capacité du modèle à distinguer correctement entre les classes.

3.1.2.2 Interprétation

Dans la recherche et l'analyse de données, la matrice de confusion est un outil vital pour comprendre la capacité de classification d'un modèle knn. Dans le domaine de la sécurité publique, par exemple, l'analyse détaillée fournie par la matrice de confusion permet d'affiner la performance d'un système de reconnaissance faciale, en équilibrant avec précision la détection des menaces réelles et la minimisation des fausses alarmes.

3.1.2.3 Cas d'Utilisation et Exemples

Le taux d'erreur de classification est un indicateur statistique essentiel dans l'évaluation des modèles prédictifs en contextes professionnels. Par exemple, dans le secteur financier, l'utilisation d'un modèle knn avec un taux d'erreur minimisé est impératif pour la prévision de la solvabilité des emprunteurs. Cela permet aux institutions financières de réduire les risques de crédit. En médecine, l'application d'un modèle knn avec un faible taux d'erreur est crucial pour le diagnostic précis des maladies, contribuant ainsi à des interventions médicales plus ciblées et efficaces.

3.1.3 Rappel, Précision et Score F1 : Indicateurs Combinés de Performance

3.1.3.1 Fondements Mathématiques

Le Rappel $= \frac{\text{VP}}{\text{VP+FN}}$ et la Précision $= \frac{\text{VP}}{\text{VP+FP}}$, complétés par le Score F1 $= 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision+Rappel}}$, offrent une perspective équilibrée sur la performance du modèle knn. Le rappel, mesurant la capacité du modèle à détecter correctement les instances positives, est crucial dans les contextes où les instances manquées peuvent avoir des conséquences graves. La précision, quant à elle, évalue la justesse des prédictions positives du modèle. Le score F1, étant la moyenne harmonique de rappel et de précision, fournit une mesure composite qui équilibre ces deux aspects. Un score F1 élevé indique que le modèle maintient un bon équilibre entre la sensibilité et la spécificité, ce qui est particulièrement important dans des applications où il est essentiel de minimiser à la fois les faux positifs et les faux négatifs.

3.1.3.2 Interprétation

Le rappel et la précision, ainsi que le score F1, offrent une perspective équilibrée de la performance du modèle knn. Un rappel élevé indique que le modèle est performant pour détecter les instances positives, mais peut inclure un nombre important de faux positifs. Inversement, une précision élevée suggère que les prédictions positives sont fiables, mais certains cas positifs peuvent être manqués. Le score F1, en tant que moyenne harmonique de ces deux mesures, fournit un équilibre entre la sensibilité et la spécificité, reflétant une performance globale solide du modèle. Un score F1 élevé est particulièrement souhaitable, car il indique que le modèle maintient un bon équilibre entre la détection de toutes les instances positives et la minimisation des fausses alertes.

3.1.3.3 Cas d'Utilisation et Exemples

Ces métriques sont fondamentales dans les études académiques où l'équilibre entre la détection des instances positives et la réduction des fausses alertes est essentiel. Dans le domaine biomédical, par exemple, un modèle knn ayant un score F1 élevé est indispensable pour l'identification précise de maladies rares, où le coût d'un faux négatif ou d'un faux positif peut être extrêmement élevé.

3.1.4 Spécificité et Taux de Faux Positifs : Mesures de Performance Négative

3.1.4.1 Fondements Mathématiques

La Spécificité = $\frac{VN}{VN+FP}$ et le taux de faux positifs FPR = $\frac{FP}{VN+FP}$ sont deux indicateurs clés de la capacité du modèle knn à classer correctement les instances négatives. Une spécificité élevée indique que le modèle est efficace pour détecter les cas négatifs, tandis qu'un faible taux de faux positifs signifie que le modèle ne classe pas fréquemment les instances négatives comme positives. Ces mesures sont particulièrement pertinentes dans les situations où les erreurs de classification négative peuvent avoir des répercussions importantes.

3.1.4.2 Interprétation

La spécificité et le taux de faux positifs fournissent des informations complémentaires sur la capacité du modèle à classifier correctement les instances négatives. Une spécificité élevée, associée à un faible taux de faux positifs, est idéale, indiquant que le modèle est précis dans l'identification des cas négatifs et minimise les erreurs de classification positive. Ces mesures sont particulièrement importantes dans des contextes où les erreurs de faux positifs ont des conséquences graves.

3.1.4.3 Cas d'Utilisation et Exemples

La spécificité et le taux de faux positifs sont des mesures critiques dans les applications où les erreurs peuvent avoir des conséquences graves. Dans les systèmes de surveillance, un modèle knn nécessite une spécificité élevée pour éviter des alarmes non justifiées tout en maintenant une vigilance accrue pour les vraies menaces, un équilibre délicat mais essentiel pour la sécurité et la fiabilité du système.

3.1.5 Validation Croisée

3.1.5.1 Fondements Mathématiques

La validation croisée, en divisant l'ensemble des données en plusieurs sous-ensembles et en évaluant le modèle sur ces partitions, sert à tester la robustesse et la généralisabilité du modèle knn. Cette méthode permet d'assurer que la performance du modèle n'est pas seulement le résultat d'un ajustement spécifique aux données d'entraînement mais qu'il est capable de maintenir une performance constante sur différents échantillons de données. Une performance stable à travers les différentes itérations de validation croisée est un indicateur de la fiabilité et de l'applicabilité du modèle dans des conditions variées.

3.1.5.2 Interprétation

La validation croisée évalue la capacité du modèle knn à généraliser au-delà des données d'entraînement spécifiques. Une performance stable et élevée à travers différentes itérations de validation croisée indique que le modèle est robuste et fiable, capable de bien fonctionner sur diverses données. Cette méthode est essentielle pour garantir que la performance du modèle n'est pas uniquement le résultat d'un ajustement trop spécifique à un ensemble de données particulier.

3.1.5.3 Cas d'Utilisation et Exemples

Dans le milieu académique et les industries de recherche, la validation croisée est utilisée pour tester la robustesse et la généralisabilité des modèles knn. Par exemple, en marketing, la capacité d'un modèle knn à prédire les tendances de consommation à travers différents segments de marché peut être évaluée en utilisant la validation croisée. Cela garantit que le modèle est adaptable et fiable, fournissant ainsi des informations précieuses pour des stratégies de marketing data-driven.

3.2 Évaluation de la performance du Modèle kmeans

Les modèles de clustering k-means jouent un rôle crucial dans l'analyse de données non supervisée. Evaluer la qualité de ces modèles est essentiel pour garantir que les clusters identifiés sont significatifs et utiles. Les méthodes d'évaluation standard incluent le Silhouette Score, le Within-Cluster Sum of Squares (WCSS), et le Between-Cluster Sum of Squares (BCSS), chacune offrant une perspective unique sur la performance des clusters formés.

3.2.1 Silhouette Score

3.2.1.1 Fondements Mathématiques

Le Silhouette Score mesure à quel point chaque point dans un cluster est proche des points du même cluster par rapport aux points des clusters les plus proches. La formule $S(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}}$ repose sur deux composantes : $a(i)$ est la distance moyenne de i à tous les autres points de son propre cluster, et $b(i)$ est la distance moyenne la plus faible de i à tous les points d'un autre cluster. Cette formule évalue donc l'adéquation d'un point à son cluster par rapport à son prochain meilleur fit dans un autre cluster.

La logique ici est de quantifier à quel point chaque point est "bien placé" : un score élevé (proche de +1) indique qu'un point est bien plus proche des points de son propre cluster que de ceux d'un autre cluster, signifiant un bon clustering.

3.2.1.2 Interprétation

L'interprétation du Silhouette Score repose sur sa plage de valeurs allant de -1 à +1. Un score proche de +1 révèle une excellente séparation des clusters, indiquant que les points sont non seulement bien regroupés au sein de leur propre cluster, mais également nettement séparés des autres clusters. Cela traduit une clarté et une précision élevées dans l'assignation des points aux clusters. À l'inverse, un score proche de -1 suggère que les points ont été probablement mal assignés, car ils se trouvent plus proches des points d'autres clusters que de ceux de leur propre cluster. Les scores autour de 0 sont indicatifs de chevauchements entre les clusters, révélant une séparation insuffisante et une possible ambiguïté dans la formation des clusters. Par conséquent, un score moyen élevé sur l'ensemble des données signale un clustering réussi, tandis qu'un score moyen faible ou négatif appelle à une réévaluation des paramètres de clustering.

3.2.1.3 Cas d'Utilisation et Exemples

Le Silhouette Score est un outil de diagnostic robuste dans l'optimisation des modèles de clustering, particulièrement dans la détermination du nombre adéquat de clusters. En contexte professionnel, par exemple dans une étude de segmentation de clientèle, l'application de ce score permet d'identifier la configuration de clustering qui maximise la cohérence interne des groupes tout en assurant leur distinction mutuelle. Supposons une entreprise effectuant une analyse de segmentation; un Silhouette Score élevé pour un nombre spécifique de clusters révélerait une congruence optimale entre les membres de chaque segment. Ce niveau de précision dans le clustering est crucial pour le ciblage marketing, la personnalisation de l'offre, et l'amélioration de l'expérience client.

3.2.2 Within-Cluster Sum of Squares (WCSS)

3.2.2.1 Fondements Mathématiques

Le WCSS est une mesure de la variance interne des clusters. La somme $\sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - \mu_k||^2$ calcule la distance au carré entre chaque point x_i et le centroid μ_k de son cluster k. L'idée est de mesurer à quel point les points d'un même cluster sont proches les uns des autres, en supposant que des clusters bien définis auront des points proches de leur centroid.

Ce calcul est basé sur le principe de minimisation de la variance interne des clusters, une idée fondamentale dans la théorie statistique pour évaluer l'homogénéité au sein d'un groupe.

3.2.2.2 Interprétation

Le WCSS est une mesure qui évalue la compacité des clusters. Des valeurs faibles de WCSS indiquent que les points au sein de chaque cluster sont proches de leur centroid, reflétant une forte cohésion interne. Cette proximité suggère que les membres du cluster partagent des caractéristiques similaires, ce qui est un indicateur d'un bon clustering. D'un autre côté, des valeurs élevées de WCSS signifient que les points sont dispersés et éloignés de leur centroid, indiquant une variabilité importante au sein des clusters. Cela peut être le signe que le cluster contient des points qui ne lui appartiennent pas naturellement ou que le nombre de clusters choisi est insuffisant. Cependant, il est crucial de se rappeler que le WCSS tend à diminuer avec l'augmentation du nombre de clusters, d'où l'importance d'éviter le sur-ajustement en choisissant un nombre de clusters qui offre un équilibre entre la minimisation du WCSS et la prévention de la fragmentation excessive des données.

3.2.2.3 Cas d'Utilisation et Exemples

Dans le domaine académique et professionnel, le WCSS est fréquemment employé pour identifier le point de saturation où l'augmentation du nombre de clusters n'entraîne plus une amélioration significative de la compacité des clusters. Cette méthode, souvent désignée sous le nom de 'méthode du coude', est essentielle pour éviter à la fois le sur-ajustement et le sous-ajustement des modèles de clustering. Dans une application pratique, comme l'analyse de données transactionnelles, le point d'inflexion sur la courbe du WCSS indique le nombre optimal de catégories pour regrouper les transactions. La détermination précise de ce nombre permet une segmentation efficace et informée, essentielle pour l'analyse des comportements de consommation et la prise de décisions stratégiques.

3.2.3 Between-Cluster Sum of Squares (BCSS)

3.2.3.1 Fondements Mathématiques

Le BCSS évalue la séparation entre les différents clusters. La formule $\sum_{k=1}^{K} n_k||\mu_k - \mu||^2$ calcule la distance au carré entre le centroid de chaque cluster μ_k et le centroid global μ, pondérée par le nombre de points dans chaque cluster n_k. Cette mesure est conçue pour refléter à quel point les clusters sont distincts les uns des autres en termes de leur emplacement moyen.

Ce raisonnement s'aligne sur la notion que des clusters bien séparés seront centrés autour de points éloignés les uns des autres, indiquant une bonne diversité entre les groupes.

3.2.3.2 Interprétation

L'interprétation du BCSS se concentre sur l'évaluation de la séparation entre les différents clusters. Un BCSS élevé est indicatif de clusters bien séparés, où chaque groupe est nettement distinct des autres. Cela est particulièrement important dans les situations où une distinction claire entre les groupes est recherchée, par exemple, lorsqu'on veut identifier des segments de marché distincts ou des catégories clairement différenciées au sein des données. Un faible BCSS, en revanche, implique que les clusters ne sont pas suffisamment distincts, se chevauchent ou sont trop proches les uns des autres, ce qui peut être problématique pour des applications nécessitant une segmentation claire. La comparaison du BCSS avec le WCSS offre une perspective plus complète, où un ratio élevé de BCSS par rapport au WCSS est souvent interprété comme un signe de clusters bien définis et distincts, offrant une séparation claire entre les différents groupes au sein de l'ensemble de données.

3.2.3.3 Cas d'Utilisation et Exemples

Le BCSS joue un rôle crucial dans les études où la démarcation nette entre les différents groupes est primordiale. Dans un cadre d'étude de marché, par exemple, une entreprise cherchant à distinguer des niches de consommateurs distinctes bénéficiera grandement d'un BCSS élevé. Ce dernier indique que les clusters identifiés présentent des caractéristiques significativement différentes, facilitant ainsi l'élaboration de stratégies marketing différenciées. Un ratio élevé de BCSS par rapport au WCSS confirme l'efficacité du clustering en démontrant que non seulement les membres de chaque cluster sont homogènes, mais aussi que chaque cluster est distinct et unique par rapport aux autres. Cette distinction est fondamentale pour des approches marketing ciblées et pour l'identification précise des besoins et préférences des consommateurs.

4 Résultats et Analyse

4.1 Impact de la Distance de Minkowski sur knn

4.1.1 Exemple d'analyse des résultats pour la mesure d'ellipticite (E34)

4.1.1.1 Analyse Métrique à $p = 1$

La précision globale, variant entre 0.40% et 0.60%, souligne une efficacité limitée du modèle dans l'identification correcte des instances positives. Ce niveau faible de précision peut indiquer des problèmes tels que la sur-généralisation ou une difficulté à saisir les nuances fines des données. Ceci est particulièrement préoccupant dans les scénarios où la détection précise des cas positifs est essentielle.

Parallèlement, le rappel global, oscillant lui aussi entre 0.40 et 0.60, mesure la capacité du modèle à détecter les vrais positifs. Des valeurs modérées dans cette métrique suggèrent une performance restreinte en matière de

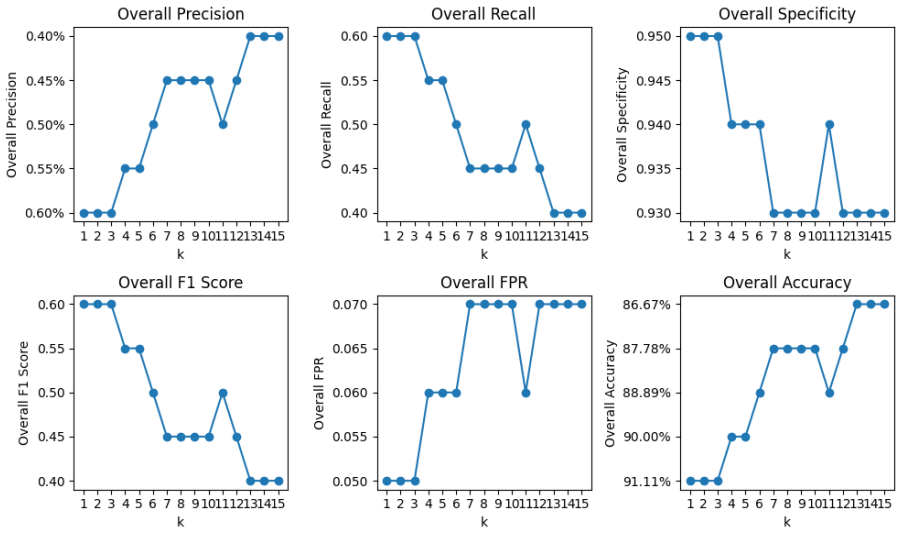

Figure 1: Résultat knn sur E34 pour k allant de 1 à 15 avec $p = 1$

détection des instances positives, ce qui peut être alarmant dans des contextes où capturer tous les cas positifs est crucial, comme dans les diagnostics médicaux ou la détection de fraude.

En ce qui concerne la spécificité globale, se situant entre 0.93 et 0.95, nous observons une capacité relativement élevée du modèle à reconnaître correctement les instances négatives. Cela indique une certaine efficacité dans l'identification des vrais négatifs, bien que cette force ne doive pas masquer les faiblesses du modèle dans d'autres domaines, notamment en termes de précision et de rappel.

Le score F1, avec des valeurs allant de 0.40 à 0.60, reflète un équilibre imparfait entre précision et rappel. Ce score modéré indique que le modèle présente des compromis et des limitations dans la gestion des instances positives et négatives.

Concernant le taux de faux positifs, qui se situe entre 0.05 et 0.07, il révèle la proportion d'instances négatives incorrectement classées comme positives. Bien que ce taux soit relativement bas, il souligne une tendance du modèle à classer incorrectement une minorité d'instances négatives, ce qui peut avoir des implications importantes dans des contextes où les faux positifs sont particulièrement coûteux ou dangereux.

L'exactitude globale du modèle, variant de 86.67% à 91.11%, fournit une vue d'ensemble de sa performance. Ces valeurs, bien que semblant indiquer une performance raisonnable, doivent être interprétées avec prudence, en raison des faibles performances du modèle en termes de précision et de rappel. L'exactitude peut être trompeuse, surtout dans des cas où l'ensemble de données présente un déséquilibre des classes.

Enfin, l'analyse des différentes valeurs de k révèle une diminution de la plupart des métriques avec l'augmentation de k, suggérant une tendance à la sur-généralisation pour des valeurs plus élevées. Cela correspond à la nature du modèle knn, où un k plus élevé peut diluer la sensibilité du modèle aux variations locales dans l'espace des caractéristiques. De plus, l'utilisation d'une métrique de distance fixe ($p = 1$, correspondant à la distance de Manhattan) peut limiter l'optimisation du modèle pour toutes les distributions de données, ce qui pourrait expliquer certaines de ses limitations de performance.

En conclusion, cette analyse met en lumière que, bien que le modèle knn démontre une capacité raisonnable à classifier correctement les instances négatives, il présente des limitations significatives en termes de précision, de rappel, et d'équilibre entre ces deux métriques. Ces résultats soulignent la nécessité d'ajustements dans le choix de k, la métrique de distance, ou d'autres paramètres du modèle, ainsi que l'importance d'une évaluation plus détaillée des besoins spécifiques du domaine d'application pour optimiser la performance globale du modèle.

4.1.1.2 Analyse Métrique à $p = 2$

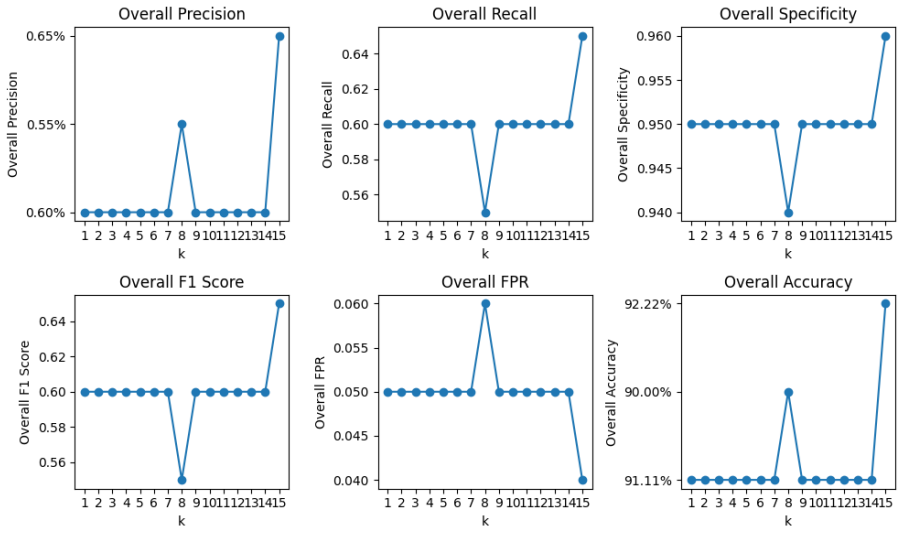

Figure 2: Résultat knn sur E34 pour k allant de 1 à 15 avec $p = 2$

La précision globale, oscillant principalement autour de 0.60%, montre que le modèle possède une capacité limitée à identifier correctement les instances positives. Cette tendance reste constante malgré des variations dans la valeur de k, à l'exception d'une légère amélioration à 0.65% pour $k = 15$. Cette augmentation marginale pour $k = 15$ pourrait indiquer une amélioration de la performance du modèle dans la classification des cas positifs, bien que cette amélioration soit faible.

Le rappel global suit une tendance similaire à celle de la précision, avec des valeurs stables autour de 0.60% et une légère hausse à 0.65% pour $k = 15$. Ces valeurs, bien qu'étant constantes, ne sont pas particulièrement élevées, ce qui pourrait soulever des inquiétudes dans les situations où il est crucial de détecter tous les cas positifs.

En ce qui concerne la spécificité globale, avec des valeurs fluctuant entre 0.94 et 0.96, le modèle montre une bonne capacité à identifier correctement les instances négatives. Cela suggère que le modèle est fiable pour reconnaître les négatifs, réduisant ainsi le risque de faux positifs.

Le score F1, qui mesure l'équilibre entre la précision et le rappel, reste autour de 0.60%, avec une légère augmentation pour $k = 15$. Cela indique un équilibre constant dans la performance du modèle, mais ces valeurs ne sont pas exceptionnellement élevées. Le score F1 est un indicateur de l'efficacité globale du modèle dans la gestion des instances positives et négatives.

Le taux de faux positifs, restant faible entre 0.04 et 0.05, montre que le modèle a tendance à classer correctement les instances négatives, un point positif dans les contextes où les faux positifs peuvent être problématiques.

L'exactitude globale, avec des valeurs entre 90.00% et 92.22%, suggère une capacité générale du modèle à classer correctement une grande majorité des instances. Néanmoins, cette performance doit être interprétée avec prudence, en raison des limitations observées en précision et en rappel.

En utilisant $p = 2$, le modèle s'appuie sur la distance euclidienne, une méthode fréquemment utilisée pour sa simplicité et son efficacité. Cependant, l'efficacité de cette métrique dépend fortement de la nature des données traitées et de l'espace des caractéristiques.

En conclusion, bien que le modèle knn avec ces paramètres démontre une précision globale raisonnable, il présente des faiblesses en termes de précision et de rappel, surtout dans la classification des cas positifs. Ces faiblesses signalent la nécessité d'ajustements potentiels dans le choix des paramètres, des caractéristiques, ou des méthodes de traitement des données. Il est également crucial de considérer le contexte spécifique d'application pour évaluer l'adéquation des performances du modèle, en particulier dans des domaines où une détection précise des instances positives est essentielle.

4.1.1.3 Analyse Métrique à $p = 3$

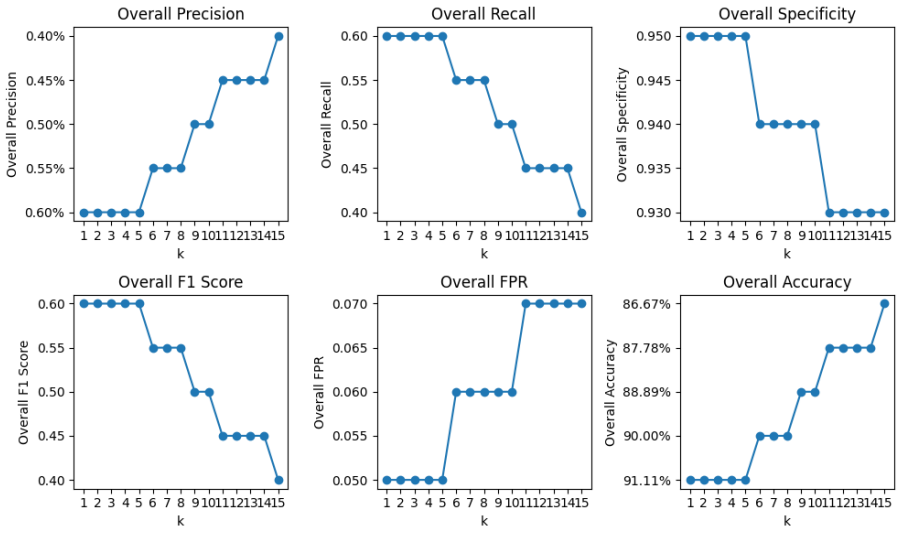

Figure 3: Résultat knn sur E34 pour k allant de 1 à 15 avec $p = 3$

la précision globale, nous observons une diminution progressive, passant de 0.60% à 0.40% à mesure que k augmente. Cette tendance décroissante indique que le modèle devient moins apte à classifier correctement les instances positives avec un k plus élevé, suggérant une possible sur-généralisation et une perte de sensibilité dans la détection des vrais positifs.

Le rappel global suit une trajectoire semblable, diminuant de 0.60 à 0.40. Cette baisse progressive montre que le modèle devient moins efficace pour identifier les instances positives, un point critique dans des scénarios où capturer tous les cas positifs est essentiel.

En ce qui concerne la spécificité globale, les variations sont légères, oscillant entre 0.93 et 0.95, et reflètent une performance stable et relativement élevée du modèle dans l'identification des instances négatives. Cela suggère que, bien que le modèle soit généralement fiable pour détecter les négatifs, cette fiabilité peut légèrement diminuer avec des valeurs de k plus élevées.

Le score F1 global, qui équilibre la précision et le rappel, montre également une diminution progressive, passant de 0.60% à 0.40%. Cette baisse signale un équilibre de moins en moins optimal entre précision et rappel, traduisant une efficacité globale décroissante du modèle dans la gestion des instances positives et négatives.

Concernant le taux de faux positifs, on note une légère augmentation, de 0.05 à 0.07. Bien que ces valeurs demeurent relativement faibles, cette hausse indique une tendance croissante du modèle à mal classifier des instances négatives comme positives, ce qui peut être problématique dans des contextes où les conséquences des faux positifs sont significatives.

L'exactitude globale, quant à elle, diminue de 91.11% à 86.67%. Malgré des valeurs relativement élevées, elles doivent être interprétées avec prudence en tenant compte des performances en précision, en rappel et en score F1. L'exactitude, prise isolément, peut masquer des déficiences spécifiques dans la classification des instances positives et négatives, surtout en considérant la baisse observée dans les autres métriques.

Enfin, avec p fixé à 3, le modèle utilise une métrique de distance moins courante que la distance euclidienne ($p = 2$) ou la distance de Manhattan ($p = 1$). Cette spécificité de mesure peut influencer la manière dont les distances sont calculées dans l'espace des caractéristiques, et par conséquent, affecter la performance du modèle, particulièrement pour des valeurs de k plus élevées.

En résumé, cette analyse met en lumière que, bien que le modèle knn soit globalement compétent, il présente des limitations notables en termes de précision, de rappel et d'équilibre général entre ces deux métriques, surtout à des valeurs plus élevées de k et avec la métrique de distance $p = 3$. Ces constats soulignent l'importance d'ajuster soigneusement les paramètres du modèle pour optimiser sa performance dans des contextes spécifiques.

4.1.1.4 Conclusions sur les Configurations de k et p

La comparaison des performances d'un modèle k-nearest neighbors (knn) en utilisant trois différentes métriques de distance ($p = 1, 2$ et 3) révèle des aspects importants sur la manière dont la métrique de distance influence la performance globale du modèle. Chaque métrique apporte ses caractéristiques distinctes, affectant différemment la précision, le rappel, la spécificité, le score F1, le taux de faux positifs et l'exactitude du modèle.

Avec la métrique $p = 1$, correspondant à la distance de Manhattan, on observe une performance initiale modeste en termes de précision et de rappel, qui tend à diminuer avec l'augmentation de la valeur de k. Cette tendance suggère une difficulté croissante du modèle à classifier correctement les instances positives à mesure que le nombre de voisins augmente. En revanche, la spécificité reste élevée à travers toutes les valeurs de k, indiquant une capacité constante du modèle à identifier correctement les instances négatives. Cependant, le score F1 et l'exactitude montrent une baisse progressive, soulignant une diminution globale de la performance du modèle, en particulier pour des valeurs élevées de k.

Lorsque la métrique de distance est fixée à $p = 2$, qui correspond à la distance euclidienne, les résultats suggèrent une meilleure stabilité. La précision et le rappel se maintiennent à un niveau relativement constant, indiquant une cohérence dans la performance du modèle, indépendamment de la valeur de k. La spécificité, bien qu'un peu inférieure à celle obtenue avec $p = 1$, reste élevée, affirmant la capacité du modèle à distinguer les négatifs. De plus, le score F1 et l'exactitude sont légèrement supérieurs par rapport à ceux obtenus avec $p = 1$, reflétant une performance globale plus équilibrée et efficace.

Enfin, avec $p = 3$, une métrique Minkowski avec un paramètre supérieur à 2, on constate une baisse plus marquée des performances. La précision et le rappel diminuent notablement, en particulier pour des valeurs élevées de k, ce qui indique une perte d'efficacité du modèle à identifier correctement les cas positifs. Cette tendance est également reflétée dans une baisse de la spécificité et une augmentation du taux de faux positifs, suggérant que le modèle devient moins fiable dans la classification des négatifs. En conséquence, le score F1 et l'exactitude subissent également une diminution significative, indiquant une performance globale plus faible.

En conclusion, le choix de la métrique de distance est un facteur clé dans la détermination de la performance d'un modèle knn. Les métriques $p = 1$ et $p = 2$ semblent offrir un meilleur équilibre global en termes de précision, de rappel, de spécificité et d'exactitude, tandis que la métrique $p = 3$, bien qu'intéressante dans sa conceptualisation, peut conduire à une baisse de performance, particulièrement pour des valeurs plus élevées de k. Ces observations soulignent l'importance de sélectionner une métrique de distance appropriée en fonction des spécificités des données et des objectifs de classification.

4.2 Impact de la Distance de Minkowski sur kmeans

4.2.1 Exemple d'analyse des résultats pour le Descripteur de Fourier Générique (GFD)

4.2.1.1 Analyse Métrique à $p = 2$

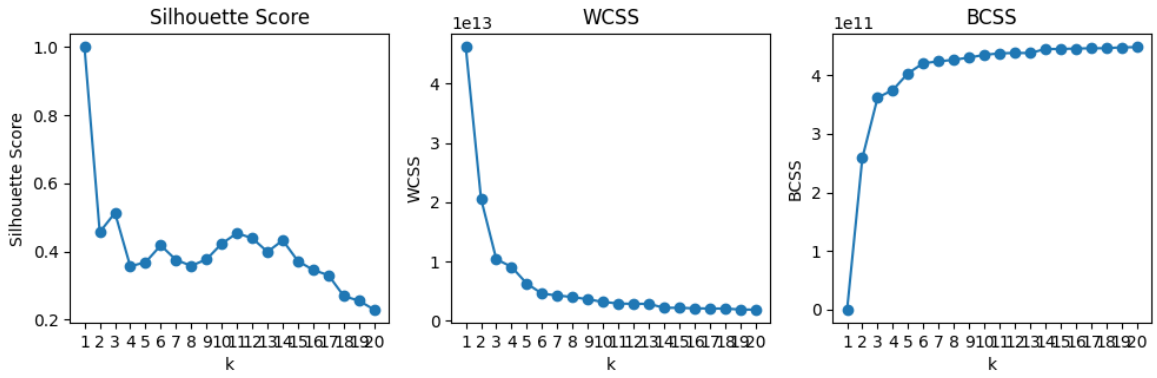

Figure 4: Résultat kmeans sur GFD pour k allant de 1 à 20 avec $p = 3$

Dans ces résultats, on observe que le score atteint la valeur maximale de 1.0 pour $k = 1$, mais cette valeur est peu informative car elle résulte d'un seul cluster. En analysant les scores pour les autres valeurs de k, on remarque un pic à $k = 5$ avec un score de 0.417505, indiquant une bonne séparation et une forte adéquation des points

dans leurs clusters à ce stade. Pour des valeurs de k supérieures, le score diminue progressivement, suggérant une diminution de la qualité du clustering.

Les résultats montrent une diminution continue du WCSS à mesure que k augmente, ce qui est attendu car l'ajout de clusters rapproche les points de leurs centres. Une forte baisse est observable entre $k = 1$ et $k = 4$, suivie d'une stabilisation relative après $k = 5$. Cette tendance suggère que des valeurs de k autour de 5 pourraient être optimales en termes de cohésion interne.

Dans ces données, le BCSS augmente avec k, ce qui est conforme aux attentes. Toutefois, tout comme pour le WCSS, la croissance du BCSS ralentit après $k = 5$, indiquant que les bénéfices en termes de séparation entre les clusters deviennent moins marqués pour des valeurs plus élevées de k.

4.2.1.2 Conclusion sur la Configuration

En intégrant ces analyses, $k = 5$ émerge comme le nombre optimal de clusters pour un équilibre entre la cohésion interne et la séparation externe. Cette conclusion est particulièrement pertinente dans des contextes où la précision de la segmentation est cruciale, comme dans le marketing personnalisé, la gestion des ressources humaines, ou la santé publique, où une bonne segmentation peut conduire à des interventions plus ciblées et efficaces.

Il est également important de noter que ces résultats sont spécifiques à la configuration du modèle et aux données utilisées. D'autres ensembles de données ou des modifications dans les paramètres du modèle pourraient mener à des conclusions différentes. Par conséquent, il est essentiel d'effectuer une telle analyse de manière itérative, en ajustant les paramètres et en interprétant les résultats dans le contexte spécifique de leur application.

En somme, cette analyse approfondie souligne l'importance d'une évaluation rigoureuse des résultats de clustering pour obtenir des informations précieuses et applicables, en tenant compte de l'équilibre entre la séparation et la cohésion des clusters pour une segmentation efficace et significative.

5 Discussion des résultats

5.1 Analyse des résultats pour le modèle knn ($p = 2$)

La méthode E34, malgré une spécificité et une exactitude relativement élevées, montre des limites en termes de précision et de rappel. Ces résultats indiquent que bien que la méthode soit compétente pour identifier correctement les cas négatifs, elle peine à détecter de manière fiable les cas positifs. Cette tendance est maintenue à travers diverses instances de k, suggérant une certaine cohérence dans sa capacité à éviter les faux positifs, mais avec une propension notable à manquer des cas positifs réels. Cette caractéristique pourrait la rendre moins adaptée aux applications où la détection précise des positifs est cruciale.

La méthode F0 affiche des performances globalement supérieures à E34, particulièrement en termes de précision et de rappel. Ce profil suggère une meilleure aptitude à identifier correctement à la fois les cas positifs et négatifs. La stabilité des valeurs de précision, de rappel, de spécificité, et d'exactitude à travers différentes valeurs de k indique une robustesse significative, rendant F0 appropriée pour des applications nécessitant un équilibre raisonnable entre la détection des positifs et des négatifs.

GFD se distingue par ses performances exceptionnelles dans presque tous les paramètres. La haute précision et le rappel élevé, combinés à une spécificité et une exactitude remarquables, révèlent une méthode extrêmement compétente pour distinguer précisément les classes positives et négatives. La constance de ces résultats à travers diverses instances de k renforce la confiance dans sa fiabilité et son applicabilité dans un large éventail de scénarios, en particulier ceux où la précision des prédictions est primordiale.

La méthode SA, quant à elle, montre une tendance à des performances décroissantes, en particulier dans les dernières instances de k. Bien que sa spécificité et son exactitude restent relativement élevées, la baisse de la précision et du rappel soulève des questions sur sa fiabilité et sa constance. Cette méthode pourrait être moins fiable pour des applications où une identification constante et précise des positifs et des négatifs est nécessaire.

En conclusion, l'analyse des performances de ces méthodes révèle que GFD est la plus performante, offrant une grande fiabilité et précision dans la reconnaissance des formes. Cette méthode se présente comme la plus appropriée pour des applications où la précision des prédictions est de la plus haute importance. F0 se positionne comme un choix solide pour des contextes nécessitant un équilibre entre la détection des positifs et des négatifs. E34 et SA, bien que possédant certaines forces, présentent des limites qui doivent être prises en compte

dans leur application pratique. Cette analyse souligne l'importance d'une sélection rigoureuse de la méthode de reconnaissance des formes en fonction des exigences spécifiques de l'application et du contexte.

5.2 Analyse des résultats pour le modèle kmeans ($p = 2$)

Le score de silhouette, dans ce contexte, révèle des tendances intéressantes. Pour la méthode E34, nous observons une diminution progressive du score, suggérant une baisse de la qualité des clusters en termes de cohésion et de séparation avec l'augmentation du nombre de k. Cela pourrait indiquer que, pour cette méthode, augmenter le nombre de clusters ne favorise pas nécessairement une meilleure structure des données. Pour F0, la variabilité des scores de silhouette reflète une incohérence dans la formation des clusters. GFD présente des scores généralement modérés, indiquant une qualité de clustering acceptable, mais suggérant aussi une marge d'amélioration dans la séparation et la cohésion des clusters. SA, en revanche, montre une qualité de clustering acceptable pour un nombre réduit de clusters, mais cette qualité se détériore avec l'augmentation de k, indiquant une perte de définition et de séparation des clusters dans des configurations plus complexes.

En ce qui concerne WCSS et BCSS, E34 montre une réduction du WCSS, indiquant la formation de clusters plus compacts, mais un BCSS très bas, suggérant que la séparation entre ces clusters est insuffisante. F0, avec une forte réduction du WCSS et un BCSS élevé, offre un profil plus équilibré, indiquant des clusters bien définis et distincts. GFD, avec une réduction du WCSS et une augmentation modeste du BCSS, semble tendre vers une meilleure définition et séparation des clusters, bien que ces indicateurs suggèrent que les clusters ne sont pas parfaitement optimisés. SA, avec une baisse constante du WCSS mais un BCSS relativement bas, indique des clusters compacts mais avec une distinction limitée entre les groupes pour des valeurs de k plus élevées.

Ces observations mettent en évidence les forces et les faiblesses de chaque méthode. E34 et SA peuvent être limitées dans des scénarios nécessitant une distinction claire entre différentes catégories. F0 semble offrir un meilleur équilibre entre la formation de clusters compacts et leur séparation claire, ce qui la rend appropriée pour des applications nécessitant une délimitation nette entre les groupes. GFD offre également un compromis intéressant, bien que ses performances puissent encore être optimisées.

En conclusion, cette analyse approfondie met en relief l'importance de choisir avec soin la méthode de clustering, en prenant en compte les caractéristiques spécifiques du dataset et les exigences de l'application. L'équilibre entre la densité interne des clusters et leur distinction mutuelle est crucial pour un clustering efficace et pertinent. Cette étude fournit des insights précieux pour orienter le choix des méthodes de clustering dans des applications futures, en soulignant la nécessité d'une approche globale qui considère à la fois la proximité interne des clusters et leur séparation claire.

5.3 Améliorations pour knn et kmeans

5.3.1 Normalisation et Standardisation des Données

La normalisation et la standardisation sont cruciales dans le traitement des données pour knn. La normalisation ajuste les données pour qu'elles se situent dans une plage de 0 à 1, ce qui est particulièrement utile pour les données dont les variables varient dans des gammes différentes. La standardisation, quant à elle, rééchelonne les données pour avoir une moyenne de 0 et un écart-type de 1, ce qui est bénéfique pour les modèles sensibles aux variations d'échelle. Ces méthodes garantissent une équité dans le traitement de chaque variable, permettant au modèle de mieux interpréter et de classer les données.

5.3.2 kmeans++ pour l'Initialisation des Centroids

L'utilisation de kmeans++ pour l'initialisation des centroids dans l'algorithme KMeans a montré une amélioration significative par rapport à l'initialisation aléatoire standard. Kmeans++ sélectionne de manière itérative les centroids initiaux en maximisant leur distance les uns des autres, ce qui conduit à une meilleure répartition initiale des clusters et à une convergence plus rapide et plus stable de l'algorithme.

5.4 Exploration de Nouvelles Méthodes

5.4.1 Réseaux de Neurones et Deep Learning

L'application des réseaux de neurones, en particulier des architectures de deep learning, a révolutionné le domaine de la reconnaissance de formes. Ces modèles sont capables de capturer des caractéristiques complexes et de haut

niveau à travers des couches cachées, rendant possible la classification précise dans des contextes où les méthodes traditionnelles échouent. Leur capacité à apprendre des représentations de données non linéaires les rend idéaux pour des tâches complexes et variées.

5.4.2 Support Vector Machines (SVM)

Les SVM sont une autre technique puissante, particulièrement efficace dans les tâches de classification binaire. Grâce à leur capacité à maximiser la marge entre différentes classes et à gérer efficacement des espaces de grande dimension, les SVM offrent une approche robuste et précise pour la reconnaissance des formes.

5.4.3 Méthodes de Réduction de Dimensionnalité

Techniques comme l'Analyse en Composantes Principales (PCA) et l'Analyse Factorielle Linéaire (LDA) sont cruciales pour réduire la dimensionnalité des ensembles de données tout en préservant les informations essentielles. Ces méthodes facilitent le traitement des données, en réduisant le bruit et en simplifiant la structure des données sans sacrifier des informations clés.

5.4.4 Clustering Hiérarchique

Contrairement aux méthodes de clustering traditionnelles comme kmeans, le clustering hiérarchique offre une vue plus nuancée de la structuration des données. Cette technique construit une hiérarchie de clusters et est particulièrement utile pour visualiser et comprendre la structure des données à différents niveaux d'agrégation.

6 Conclusion

Cette étude approfondie des performances des méthodes E34, F0, GFD et SA, appliquées à des modèles knn et kmeans, met en lumière des différences significatives dans leur efficacité et applicabilité dans des contextes de reconnaissance de formes et de clustering. L'analyse a mis en évidence que chaque méthode présente des avantages et des inconvénients distincts, qui doivent être soigneusement considérés en fonction des objectifs spécifiques de chaque application.

Pour le modèle knn, GFD a démontré une supériorité notable, affichant d'excellentes performances dans presque tous les paramètres évalués. Sa haute précision et son rappel élevé, couplés à une spécificité et une exactitude remarquables, la rendent idéale pour des applications où la précision des prédictions est essentielle. F0, bien qu'inférieure à GFD, a montré un profil de performances équilibré, le rendant adapté à des situations nécessitant un compromis entre la détection des cas positifs et négatifs. E34 et SA, en revanche, bien qu'ayant certaines qualités, présentent des limites significatives en termes de précision et de rappel, ce qui pourrait limiter leur utilisation dans des applications exigeant une grande fiabilité.

Dans le contexte du modèle kmeans, l'analyse des scores de silhouette, du WCSS et du BCSS a révélé que F0 se distingue par sa capacité à former des clusters bien définis et distincts, avec une forte réduction du WCSS et un BCSS élevé. Cette caractéristique la rend particulièrement adaptée aux applications où une séparation claire entre les groupes est nécessaire. GFD offre un équilibre raisonnable entre la compacité des clusters et leur séparation, bien que ses performances puissent être sujettes à des optimisations futures. E34 et SA, en dépit d'avoir des clusters compacts, montrent des difficultés à maintenir une séparation claire entre les clusters, en particulier à des valeurs plus élevées de k.

En somme, cette analyse souligne l'importance d'une sélection méthodique et contextualisée des techniques de reconnaissance de formes et de clustering. Les résultats de cette étude offrent des orientations précieuses pour le choix de méthodes adaptées en fonction des caractéristiques spécifiques des données et des exigences des applications pratiques. Ils mettent également en évidence la nécessité d'une évaluation multidimensionnelle des performances, prenant en compte non seulement la précision et la spécificité, mais aussi la capacité à former des clusters distincts et cohérents. Ces conclusions ouvrent la voie à des recherches futures pour affiner davantage ces méthodes et explorer leur applicabilité dans des scénarios plus diversifiés.