Université de Poitiers

Projet Maison

Supervised Learning STDV

Rapport

Classification de tirages à développement chromogène

Étudiant :
Soufiane Lmezouari

Enseignante :
Mme Farida Enikeeva

15 avril 2025

Table des matières

Table des figures

Liste des tableaux

CHAPITRE 1

Ce projet s'inscrit dans le cadre du cours de Supervised Learning et a pour objectif l'application de méthodes de classification supervisée à un problème réel : identifier automatiquement le type de papier photographique utilisé dans des tirages à développement chromogène (Kodak, Agfa, Fuji), à partir de mesures spectroscopiques dans l'infrarouge proche (350–2500 nm). Ces données, fournies par le Musée national d'Histoire naturelle, permettent d'explorer la possibilité de caractériser les supports photographiques selon leur signature spectrale. Cela contribue à la fois à l'analyse artistique et technique des œuvres, ainsi qu'à la définition de stratégies de conservation adaptées. Le projet repose sur l'utilisation de plusieurs algorithmes de classification (LDA, QDA, K-NN, SVM, arbres de décision, etc.), appliqués sur des données brutes mais également réduites à l'aide de techniques telles que l'Analyse en Composantes Principales (ACP) et le critère de Fisher, dans un objectif d'optimisation des performances et d'interprétabilité.

1.1 Analyses Préliminaires

1.1.0.1 Préparation des données

Les données utilisées dans ce projet ont été chargées à partir de quatre fichiers texte distincts : `agfa.txt`, `fuji.txt`, `kodak.txt` et `inconnu.txt`. Chaque fichier correspond à un type de papier photographique (Agfa, Fuji, Kodak) ou à des observations non étiquetées. Après lecture des trois jeux de données étiquetés, une colonne `class` a été ajoutée à chacun afin d'identifier l'origine de chaque observation. Ces jeux ont ensuite été combinés à l'aide de la fonction `rbind`, pour former un ensemble de données complet contenant 312 observations (lignes) et 2152 variables (colonnes). Aucune donnée manquante n'a été détectée (`sum(is.na(data)) = 0`), ce qui permet d'aborder les étapes suivantes d'analyse sans traitement supplémentaire de nettoyage.

1.1.0.2 la distribution des variables et des classes

Une première analyse de la distribution des variables spectrales a été réalisée à l'aide d'histogrammes et de tests de normalité. À titre d'exemple, la distribution de la première variable montre une forme légèrement asymétrique, et donc c'est pas une loi normale

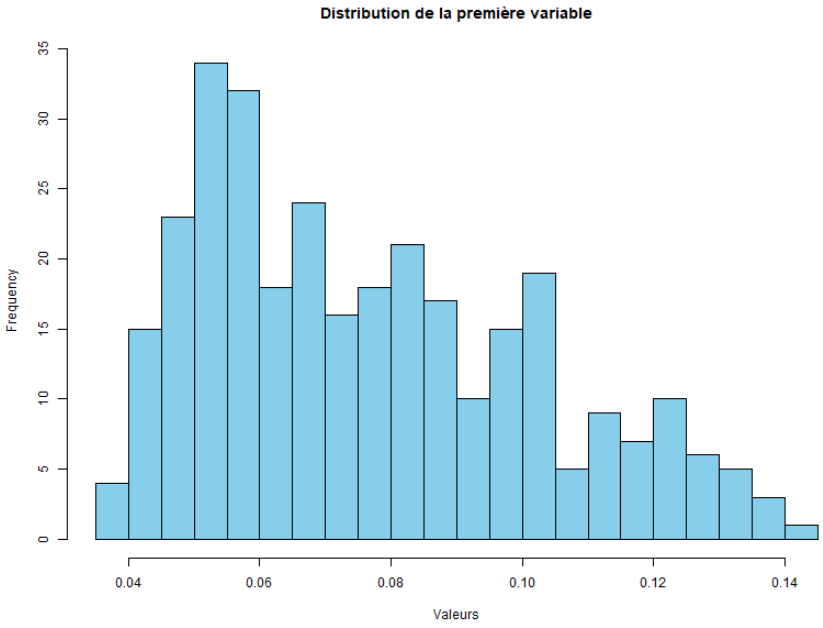

FIGURE 1.1 – Histogramme variable

Pour évaluer de manière plus systématique la distribution des variables, le test de Shapiro-Wilk a été appliqué à chacune des 2151 variables. Les résultats indiquent que seules 5 variables présentent une distribution conforme à la normalité (p-value > 0.05), ce qui confirme un comportement majoritairement non normal des données spectrales.

L'analyse de la répartition des classes met en évidence un déséquilibre notable dans le jeu de données. Le tableau suivant présente le nombre d'observations pour chaque type de papier :

Classe	**Nombre d'observations**
Agfa	118
Fuji	67
Kodak	127

TABLE 1.1 – Répartition des observations par classe de papier

Comme on peut le constater, la classe `fuji` est nettement sous-représentée, avec presque moitié moins d'exemples que `kodak`.

1.1.0.3 Analyse des spectres moyens des trois classes

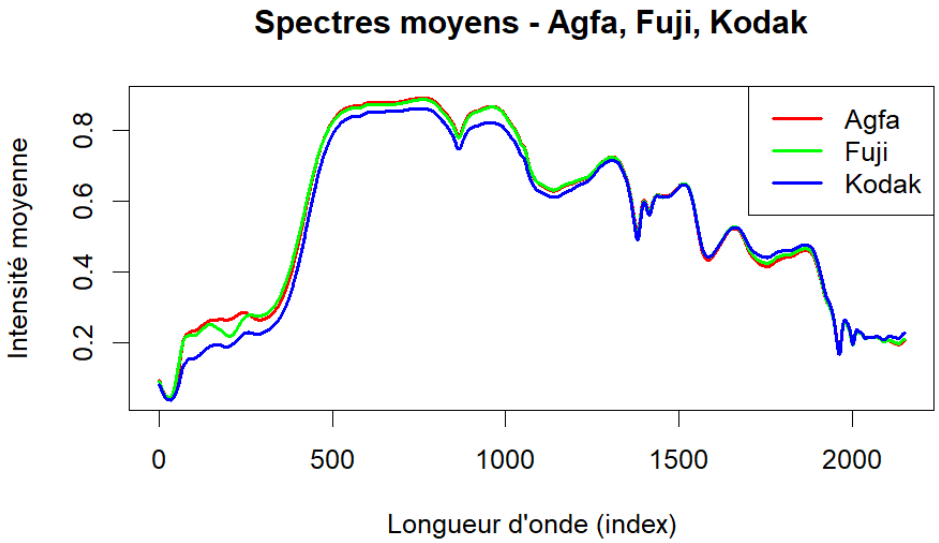

FIGURE 1.2 – Spectres moyens des papiers photographiques Agfa.

Le graphique présenté montre les intensités spectrales moyennes caractérisant les trois catégories de papier. Les courbes révèlent des tendances globalement comparables, bien que des écarts significatifs apparaissent dans les plages de 400 à 1000 nm et de 1400 à 2100 nm. Ces légères disparités expliquent la nécessité d'employer des approches statistiques sophistiquées pour déterminer les zones du spectre les plus discriminantes

1.1.0.4 Construction des ensembles d'apprentissage et de test

Afin de construire et d'évaluer les performances des modèles de classification, les données ont été divisées en deux sous-ensembles : un ensemble d'apprentissage (*training set*) et un ensemble de test (*test set*). Les 80 premières observations de la classe `agfa`, les 45 premières de `fuji` et les 85 premières de `kodak` ont été sélectionnées pour constituer l'ensemble d'apprentissage, soit un total de 210 observations. Les observations restantes (102 au total) ont été utilisées comme ensemble de test. Cette séparation a été réalisée de manière fixe à l'aide de l'instruction `set.seed(123)` afin d'assurer la reproductibilité des résultats. Les colonnes de classe ont ensuite été converties en facteur pour permettre une utilisation correcte avec les algorithmes de classification supervisée.

1.1.0.5 Centrage et réduction des variables

Avant d'appliquer les algorithmes de classification, les données ont été normalisées afin d'uniformiser l'échelle des variables spectrales. Étant donné que les mesures sont exprimées en intensité pour chaque longueur d'onde, il est essentiel de centrer et de réduire les données

pour éviter que certaines variables à grande variance n'influencent trop les modèles. Le prétraitement a été réalisé à l'aide de la fonction `preProcess` du package `caret`, en appliquant les méthodes `center` (centrage) et `scale` (réduction) sur l'ensemble d'apprentissage. Le modèle de transformation ainsi obtenu a ensuite été appliqué à l'ensemble de test pour garantir la cohérence du traitement. Les labels de classe ont ensuite été réintégrés pour permettre l'entraînement et l'évaluation des modèles sur les jeux standardisés.

1.1.0.6 Réduction de dimension par Analyse en Composantes Principales (ACP)

Le jeu de données initial est caractérisé par une forte corrélation entre les variables spectrales et un nombre de variables largement supérieur au nombre d'observations. Dans ce contexte, il est essentiel de réduire la dimension pour limiter le risque de surajustement, améliorer la stabilité des modèles, et réduire le temps de calcul.

Pour cela, nous avons appliqué l'**Analyse en Composantes Principales (ACP)** sur l'ensemble d'entraînement standardisé, afin de projeter les données dans un nouvel espace de dimension réduite, tout en conservant un maximum d'information.

Étapes réalisées :

1. **Application de l'ACP** sur les données d'apprentissage avec `ncp = 10` composantes principales retenues.
2. **Projection des données de test** sur ces mêmes composantes à l'aide de la fonction `predict()`.
3. **Construction de nouveaux jeux de données réduits** (train/test), intégrant uniquement les coordonnées sur les composantes retenues et la classe associée.
4. **Réorganisation des étiquettes de classes** et définition du vecteur `prior` correspondant aux proportions observées dans les données : 45 % agfa, 10 % fuji, 45 % kodak.

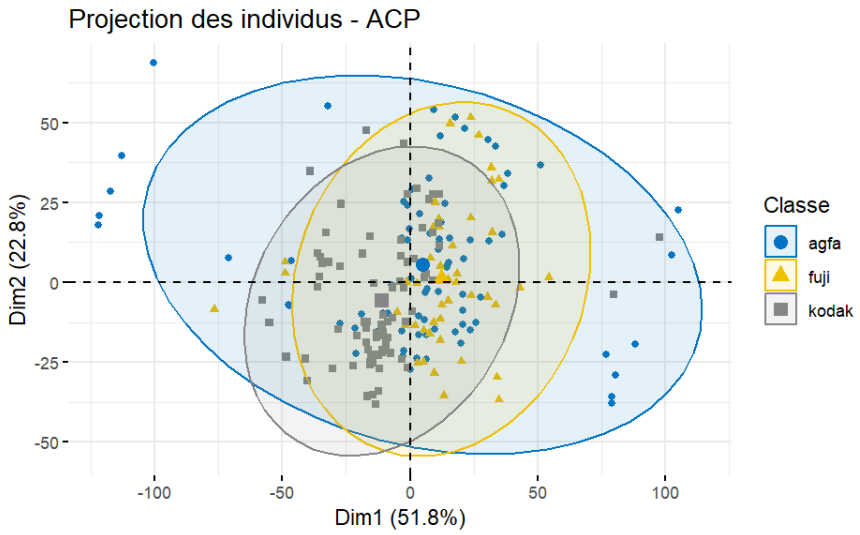

FIGURE 1.3 – Projection des individus

1.1.1 Critère de Fisher

Le critère de Fisher est un outil de réduction de dimension, permettant de sélectionner des variables informatives tout en allégeant la complexité des modèles.

La figure suivante illustre l'évolution du critère de Fisher en fonction de l'indice des longueurs d'onde. Les pics (maximums locaux) signalent les zones spectrales les plus pertinentes pour différencier les types de papiers photographiques.

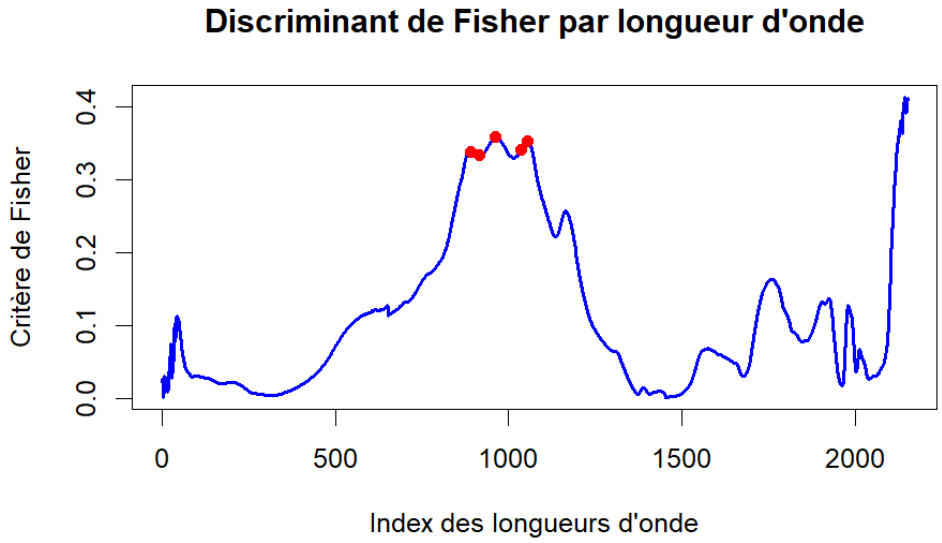

FIGURE 1.4 – Critère de Fisher selon la longueur d'onde

Sélection des longueurs d'onde (descripteurs)
Nous avons retenu **6 longueurs d'onde** correspondant aux maxima locaux, tout en excluant les extrémités du spectre (zones souvent instables ou peu fiables). Les longueurs sélectionnées sont les suivantes :

— **1238**, **1241**, **1264**, **1312**, **1386**, **1404** nm

Correspondance avec les bandes spectrales chimiques

— **1485 nm** (OH/NH) → *proche de 1404 nm*

Construction d'un jeu de données réduit
Un nouveau jeu de données a été construit à partir des **6 longueurs d'onde retenues**, en conservant uniquement les colonnes correspondantes ainsi que la classe de chaque échantillon. Ce jeu réduit sera utilisé pour comparer les performances des modèles avec ceux entraînés sur le jeu complet.

1.2 Alogorithmes

Dans le cadre de l'étape de classification, notre objectif est d'évaluer la capacité des algorithmes supervisés à prédire correctement la classe du papier photographique à partir des données spectrales. Pour chaque algorithme testé, nous appliquerons les trois configurations suivantes :

— **Modèle 1 – Données complètes :** classification réalisée sur l'ensemble des variables spectrales normalisées, sans réduction de dimension. Cette approche permet d'utiliser toute l'information disponible, mais au risque de surajuster les modèles en raison de la haute dimensionnalité.

— **Modèle 2 – Réduction par ACP :** classification réalisée après réduction de dimension par Analyse en Composantes Principales. Seules les composantes principales les plus informatives (10 composantes) sont conservées, afin de limiter la redondance et la corrélation entre variables.

— **Modèle 3 – Sélection par critère de Fisher :** classification basée sur un sous-ensemble réduit de longueurs d'onde, sélectionnées en fonction des maxima locaux du critère de Fisher. Cette méthode vise à conserver uniquement les variables les plus discriminantes entre les classes.

Cette approche comparative a pour but de mesurer l'impact de la réduction de dimension sur la performance des modèles de classification, en termes de précision, de robustesse et d'efficacité computationnelle.

1.3 Méthode LDA et QDA

1.3.1 LDA

Modèle 1

Préd.	Agfa	Fuji	Kodak
Agfa	38	0	0
Fuji	0	22	0
Kodak	0	0	42

Précision : 100%

Modèle 2

Préd.	Agfa	Fuji	Kodak
Agfa	29	10	0
Fuji	6	12	16
Kodak	3	0	26

Précision : 65,7%

Modèle 3

Préd.	Agfa	Fuji	Kodak
Agfa	36	14	0
Fuji	2	8	0
Kodak	0	0	42

Précision : 84,3%

TABLE 1.2 – Matrices de confusion obtenues pour les trois modèles de classification.

Le modèle 1 atteint une précision de **100 %**, mais cela suggère un possible *surapprentissage*, dû à la forte colinéarité entre les variables spectrales. À l'inverse, le modèle 3 (ACP) offre une bonne généralisation avec **84,3 %** de précision, tandis que le modèle 2 (Fisher) reste moins performant.

1.3.2 Taux d'erreur et du temps de calcul

Modèle	Taux d'erreur	Temps d'entraînement (s)	Temps de test (s)
LDA Modèle 1 (complet)	0 %	0.13	0.02
LDA Modèle 2 (Fisher)	34.31 %	0.02	0.00
LDA Modèle 3 (ACP)	15.69 %	0.03	0.00

TABLE 1.3 – Performances de la méthode LDA selon les trois modèles (données complètes, Fisher, ACP).

La méthode LDA s'avère particulièrement performante sur les données complètes, avec un taux d'erreur nul. La réduction par ACP permet de maintenir de bons résultats, tandis que la sélection de variables par le critère de Fisher dégrade nettement la précision. Le modèle 1 apparaît donc comme le plus fiable pour cette méthode.

1.3.3 QDA

L'algorithme QDA a également été testé dans le cadre des trois modèles définis précédemment. Toutefois, l'application du modèle 1 (sans réduction de dimension) a échoué en raison de la contrainte mathématique imposée par QDA : il nécessite que chaque classe dispose d'un nombre suffisant d'observations pour estimer les matrices de covariance. Or, dans notre cas, certaines variables sont trop nombreuses par rapport aux observations dans certaines classes, ce qui a entraîné l'erreur suivante : `some group is too small for 'qda'`. Par conséquent, seuls les modèles réduits par ACP et par critère de Fisher ont pu être évalués pour cette méthode.

Modèle 2 (Fisher)

Préd.	Agfa	Fuji	Kodak
Agfa	30	10	1
Fuji	8	12	11
Kodak	0	0	30

Précision : 70,6%

Modèle 3 (ACP)

Préd.	Agfa	Fuji	Kodak
Agfa	29	7	0
Fuji	9	15	0
Kodak	0	0	42

Précision : 84,3%

TABLE 1.4 – Matrices de confusion pour QDA

1.3.4 Taux d'erreur et temps de calcul

Modèle	Taux d'erreur	Temps d'entraînement (s)	Temps de test (s)
QDA Modèle 1 (complet)	— (erreur technique)	—	—
QDA Modèle 2 (Fisher)	29.41 %	0.02	0.02
QDA Modèle 3 (ACP)	15.69 %	0.02	0.02

TABLE 1.5 – Performances de la méthode QDA selon les trois modèles.

le modèle avec ACP obtient la meilleure performance (15,69 % d'erreur), tandis que le modèle Fisher reste moins performant (29,41 %). Le modèle 3 est donc à privilégier pour cette méthode.

1.4 CART

Dans cette section, nous explorons la méthode CART en comparant différentes variantes : l'arbre par défaut, l'arbre maximal, l'élagage basé sur le cp optimal (xerror) et l'élagage selon la règle du 1-SE.

Modèle	Taux d'erreur
Arbre par défaut	0.245
Arbre maximal	0.206
Arbre élagué (cp optimal)	0.206
Arbre élagué (règle du 1-SE)	0.196

TABLE 1.6 – Comparaison des modèles CART selon le taux d'erreur

L'arbre élagué selon la règle du 1-SE, qui présente le taux d'erreur le plus faible parmi les modèles testés, est illustré ci-dessous.

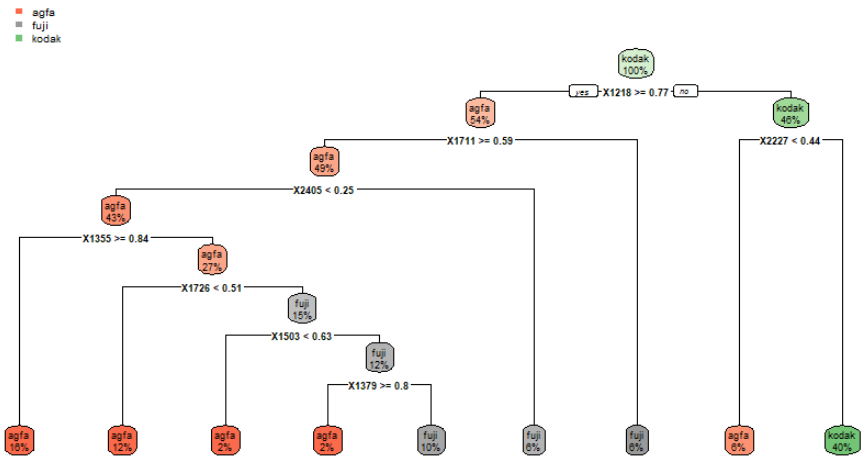

FIGURE 1.5 – Arbre élagué selon la règle 1-SE

Le modèle CART élagué selon la règle du 1-SE a ensuite été appliqué aux jeux de données réduits obtenus par les méthodes de Fisher et ACP.

1.4.1 Matrices de confusion (CART : 1-SE)

Données complètes

Préd.	Agfa	Fuji	Kodak
Agfa	35	12	0
Fuji	3	5	0
Kodak	0	5	42

Précision : 80,39%

Fisher

Préd.	Agfa	Fuji	Kodak
Agfa	30	14	17
Fuji	8	8	0
Kodak	0	0	25

Précision : 61,76%

ACP

Préd.	Agfa	Fuji	Kodak
Agfa	36	7	0
Fuji	1	15	19
Kodak	1	0	23

Précision : 72,55%

TABLE 1.7 – Matrices de confusion obtenues pour CART (élagage 1-SE)

1.4.2 Tableau des erreurs et des temps d'exécution (CART : 1-SE)

Méthode	**Taux d'erreur (%)**	**Temps entraînement (s)**	**Temps prédiction (s)**
Complète	19.61	1.59	0.70
Fisher	38.24	0.49	1.34
ACP	27.45	1.13	0.59

TABLE 1.8 – Performances de la méthode CART (élagage 1-SE) selon les trois données.

En combinant taux d'erreur et temps de calcul , on observe que le modèle CART élagué selon la règle du 1-SE appliquée aux données complètes est un peu meilleur que les autres.

1.5 KNN

Dans notre cas, nous avons évalué différents modèles de K-NN à l'aide d'une validation croisée afin de déterminer la valeur optimale de k, et nous avons utilisé le taux d'erreur pour mesurer la performance de classification sur l'échantillon de test. Comme pour les autres méthodes, l'analyse a été effectuée selon les trois stratégies définies : sans réduction de dimension, avec ACP, et avec sélection par critère de Fisher.

1.5.1 Le k optimal

Le meilleur k qu'on a trouvé c'est 1 , mais on va éviter les extrémités donc on prend le deuxième meilleur k qui vaut 3.

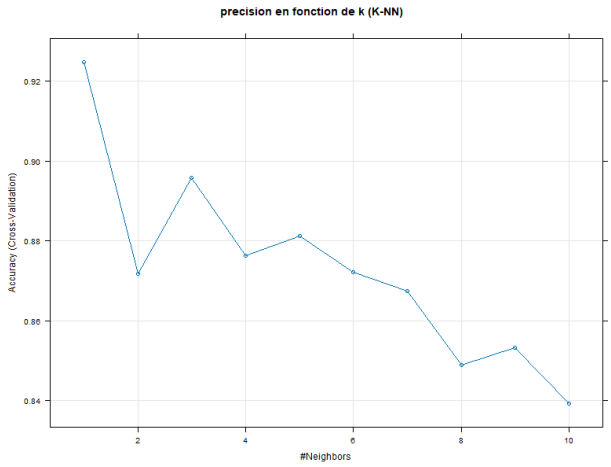

FIGURE 1.6 – précision en fonction du nombre de voisins k

1.5.2 Matrices de confusion

Modèle 1

Préd.	Agfa	Fuji	Kodak
Agfa	37	4	0
Fuji	1	18	6
Kodak	0	0	36

Précision : 89,2%

Modèle 2

Préd.	Agfa	Fuji	Kodak
Agfa	34	16	13
Fuji	4	6	0
Kodak	0	0	29

Précision : 67,65%

Modèle 3

Préd.	Agfa	Fuji	Kodak
Agfa	36	11	1
Fuji	2	11	3
Kodak	0	0	38

Précision : 83,3%

TABLE 1.9 – Matrices de confusion obtenues pour K-NN

La méthode K-NN donne ses meilleurs résultats sans réduction de dimension, avec une précision de 89,2%. L'utilisation de l'ACP permet de conserver de bonnes performances (83,3%), tandis que le critère de Fisher mène à une précision plus faible (63,7%).

1.5.3 Taux d'erreur et du temps de calcul

Modèle	**Taux d'erreur**	**Temps d'entraînement (s)**	**Temps de test (s)**
K-NN Modèle 1 (complet)	10.78 %	12.28	0.09
K-NN Modèle 2 (Fisher)	36.27 %	0.09	0.02
K-NN Modèle 3 (ACP)	16.67 %	0.02	0.02

TABLE 1.10 – Comparaison des performances de la méthode K-NN

La méthode K-NN obtient de très bons résultats dans sa version sans réduction, avec un taux d'erreur faible (10,78 %) malgré un temps de calcul plus élevé. La réduction par ACP

permet de maintenir de bonnes performances (16,67 %) avec un gain de temps notable. En revanche, la sélection par le critère de Fisher baisse fortement la qualité de la classification (36,27 %), bien que rapide à exécuter.

1.6 SVM

Nous avons appliqué l'algorithme de classification SVM. Plusieurs noyaux ont été testés, notamment les noyaux linéaire, radial et sigmoid. Après validation croisée, c'est le **noyau linéaire** qui a donné les meilleurs résultats pour les trois modèles. Les prédictions ont ensuite été réalisées sur l'échantillon de test selon les trois approches de données : complètes, réduites par Fisher, et réduites par ACP.

1.6.1 Matrices Confusion

Modèle 1

Préd.	Agfa	Fuji	Kodak
Agfa	38	15	8
Fuji	0	7	0
Kodak	0	0	34

Précision : 77,45%

Modèle 2

Préd.	Agfa	Fuji	Kodak
Agfa	38	22	10
Fuji	0	0	0
Kodak	0	0	32

Précision : 68,63%

Modèle 3

Préd.	Agfa	Fuji	Kodak
Agfa	35	6	0
Fuji	3	16	0
Kodak	0	0	42

Précision : 91,18%

TABLE 1.11 – Matrices de confusion obtenues pour SVM

Le modèle 3 (réduction par ACP) offre les meilleures performances avec la méthode SVM, atteignant une précision de 91,18 %.

1.6.2 Taux d'erreur et du temps de calcul

Méthode	**Taux d'erreur (%)**	**Temps entraînement (s)**	**Temps prédiction (s)**
Complète	22.55	1.11	0.20
Fisher	31.37	0.09	0.08
ACP	8.82	0.03	0.02

TABLE 1.12 – Performances de la méthode SVM.

Parmi les trois modèles testés, le SVM avec réduction par ACP se démarque nettement, combinant un taux d'erreur très faible (8,82 %) et un temps de calcul minimal. À l'inverse, la réduction par critère de Fisher entraîne une perte significative de performance.

1.7 Régression logistique multinomiale par approche un-contre-tous

Nous avons appliqué la méthode de classification "un-contre-tous" en combinant plusieurs régressions logistiques binaires, une par classe. Le choix des variables pour chaque modèle a été effectué automatiquement grâce à la fonction `stepAIC`.

un-contre-tous : un modèle par classe

```
models <- list()
probs_test <- matrix(0, nrow = nrow(X_test), ncol = length(classes))
colnames(probs_test) <- classes
for (k in classes) {

  y_bin <- ifelse(y_train == k, 1, 0)
  df_k <- data.frame(X_train, y_bin = factor(y_bin))
```

Sélection par stepAIC et les prédictions finales sont obtenues en prenant, pour chaque observation, la classe ayant la probabilité la plus élevée.

```
  model_null <- glm(y_bin ~ 1, data = df_k, family = binomial)
  model_full <- glm(y_bin ~ ., data = df_k, family = binomial)
  model_k <- stepAIC(model_null, scope = list(lower = model_null, upper = model
      _full),
  direction = "both", trace = FALSE)
  models[[k]] <- model_k
  probs_test[, k] <- predict(model_k, newdata = X_test, type = "response")
}

pred_class <- apply(probs_test, 1, function(p) classes[which.max(p)])
```

1.7.1 Ajustement des modèles et comparaison

L'ajustement du modèle complet et du modèle ACP avec la régression logistique a généré les avertissements `glm.fit: algorithm did not converge` et `fitted probabilities numerically 0 or 1 occurred`, liés à des problèmes de colinéarité ou de séparation quasi-parfaite. En revanche, le modèle Fisher n'a pas rencontré ces problèmes.

Modèle 1			
Préd.	Agfa	Fuji	Kodak
Agfa	38	6	1
Fuji	0	16	7
Kodak	0	0	34
Précision : 86,27%			

Modèle 2			
Préd.	Agfa	Fuji	Kodak
Agfa	26	7	0
Fuji	7	15	16
Kodak	5	0	26
Précision : 65,69%			

Modèle 3			
Préd.	Agfa	Fuji	Kodak
Agfa	32	10	0
Fuji	4	12	0
Kodak	2	0	42
Précision : 84,31%			

TABLE 1.13 – Matrices de confusion pour la régression logistique .

La régression logistique donne les meilleurs résultats avec le modèle complet (86,27 %) et le modèle ACP (84,31 %). Le modèle Fisher est moins performant, avec un taux d'erreur plus élevé.

1.7.2 Comparaison des performances de la régression logistique

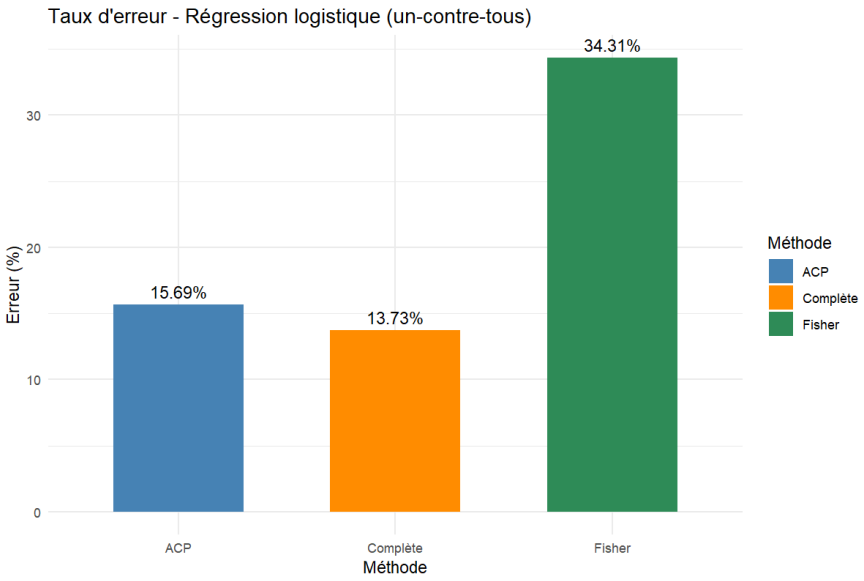

FIGURE 1.7 – Taux d'erreur - Régression logistique (un-contre-tous)

Méthode	Taux d'erreur (%)	Temps entraînement (s)
Complète	13.73	1.11
Fisher	34.31	0.09
ACP	15.69	0.03

TABLE 1.14 – Performances de la régression logistique selon les trois modèles.

Le modèle complet donne les meilleures performances (13,73 % d'erreur), suivi de près par le modèle ACP. Le modèle Fisher, bien que rapide, reste moins précis.

1.8 Régression logistique multinomiale

La régression logistique multinomiale est une extension de la régression logistique binaire, utilisée pour des tâches de classification à plus de deux classes. Elle permet d'estimer, pour chaque observation, la probabilité d'appartenance à chacune des classes, et d'assigner la classe avec la probabilité maximale. Dans notre cas, la méthode `multinom()` du package `nnet` a été utilisée pour ajuster les modèles sur les trois versions des données : complètes, réduites par Fisher et réduites par ACP.

1.8.1 Ajustement des modèles et comparaison

Modèle 1			
Préd.	Agfa	Fuji	Kodak
Agfa	36	1	0
Fuji	0	21	1
Kodak	2	0	41
Précision : 96,08%			

Modèle 2			
Préd.	Agfa	Fuji	Kodak
Agfa	29	7	0
Fuji	7	15	16
Kodak	2	0	26
Précision : 68,63%			

Modèle 3			
Préd.	Agfa	Fuji	Kodak
Agfa	33	11	0
Fuji	5	11	0
Kodak	0	0	42
Précision : 84,31%			

TABLE 1.15 – Matrices de confusion pour la régression logistique multinomiale

Le modèle complet obtient d'excellents résultats avec une précision de 96,08 %, tandis que les performances chutent avec le modèle Fisher. Le modèle ACP conserve une bonne précision (84,31 %) tout en réduisant la dimension.

1.8.2 Performances de la régression logistique multinomiale

Méthode	**Taux d'erreur (%)**	**Temps entraînement (s)**
Complète	3.92	7.18
Fisher	31.37	0.03
ACP	15.69	0.03

TABLE 1.16 – Performances de la régression logistique multinomiale

Le modèle complet donne d'excellents résultats, avec un taux d'erreur très faible (3,92 %) mais un temps de calcul plus important. Le modèle ACP offre un bon compromis entre performance et rapidité, tandis que le modèle Fisher reste moins fiable.

1.9 Classification par forêts aléatoires

L'algorithme des forêts aléatoires repose sur la construction d'un grand nombre d'arbres de décision, dont les prédictions sont agrégées pour améliorer la robustesse du modèle. Dans notre cas, l'algorithme `randomForest()` a été appliqué avec l'option `importance=TRUE` afin d'identifier les variables les plus pertinentes pour la classification des papiers photographiques. L'importance des variables a été évaluée selon deux critères : Mean Decrease Accuracy et Mean Decrease Gini. Les figures ci-dessous présentent les variables ou composantes les plus discriminantes selon les trois configurations de données : modèle complet, réduction par Fisher, et réduction par ACP.

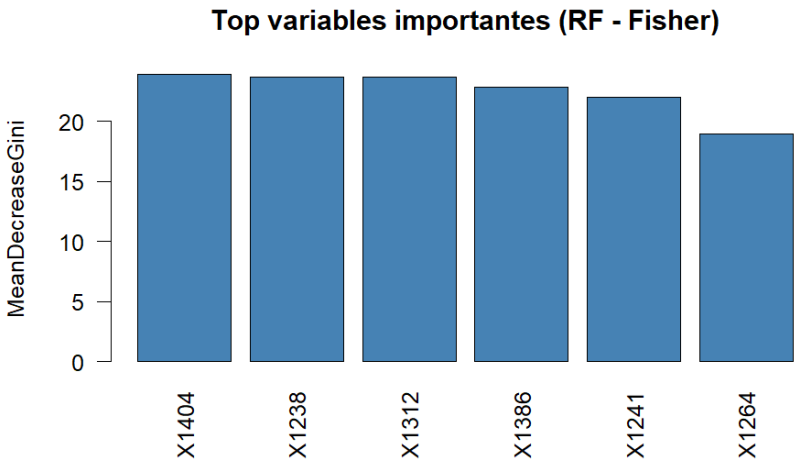

FIGURE 1.8 – Variables importantes — Random Forest (modèle complet)

L'algorithme Random Forest appliqué aux données complètes a mis en évidence les variables les plus discriminantes selon deux critères : la baisse de la précision (*MeanDecreaseAccuracy*) et la baisse de l'indice de Gini (*MeanDecreaseGini*). Ces variables spectrales

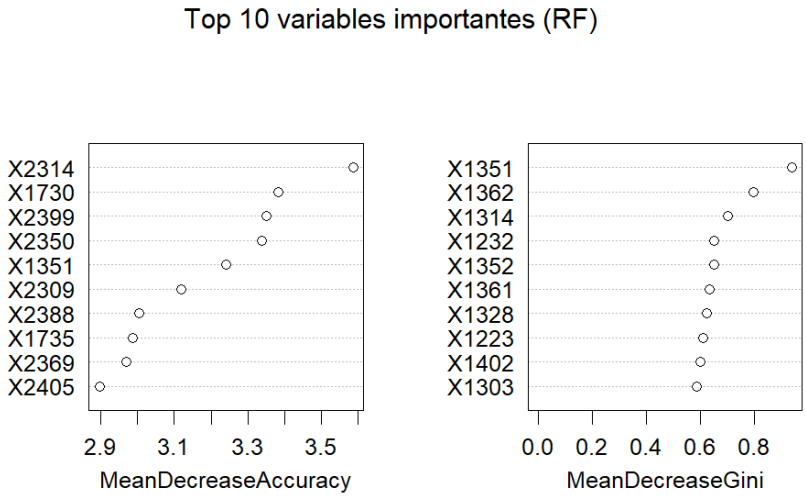

FIGURE 1.9 – Variables importantes — Random Forest (modèle réduit par Fisher)

Dans le cas du modèle réduit par sélection de variables via le critère de Fisher, Random Forest identifie les longueurs d'onde les plus pertinentes à partir du critère de Gini. Ces variables représentent les zones du spectre les plus utiles pour discriminer les trois classes de papier photo.

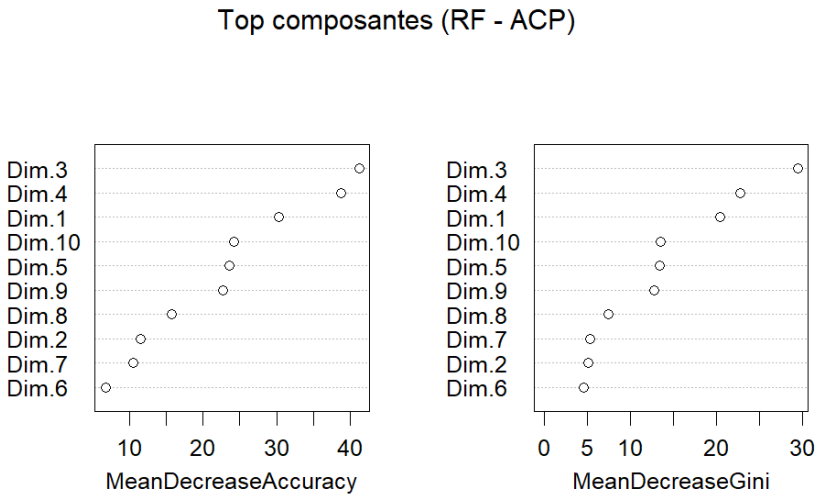

FIGURE 1.10 – Composantes importantes — Random Forest (modèle réduit par ACP)

Lorsque les données sont transformées par l'ACP, l'importance est évaluée non plus par variable spectrale, mais par composante principale. Ici, les composantes `Dim.3`, `Dim.4` et `Dim.1` apparaissent comme les plus déterminantes pour la classification selon les critères de précision et de pureté des nœuds.

1.9.0.1 Ajustement des modèles et comparaison

Modèle 1

Préd.	Agfa	Fuji	Kodak
Agfa	32	15	14
Fuji	4	7	0
Kodak	2	0	28

Précision : 65,69%

Modèle 2

Préd.	Agfa	Fuji	Kodak
Agfa	33	17	13
Fuji	5	5	0
Kodak	0	0	29

Précision : 65,69%

Modèle 3

Préd.	Agfa	Fuji	Kodak
Agfa	37	4	0
Fuji	1	18	12
Kodak	0	0	30

Précision : 83,33%

TABLE 1.17 – Matrices de confusion pour Random Forest s

Les performances du modèle Random Forest varient selon la configuration utilisée. Le modèle ACP obtient les meilleurs résultats (83,33 %), tandis que les modèles complet et Fisher présentent des taux d'erreur plus élevés, proches de 34 %.

1.9.1 Performances de l'algorithme Random Forest

Modèle	Taux d'erreur (%)	Temps entraînement (s)
Top variables (complet)	34.31	0.14
Fisher	34.31	0.06
ACP	16.67	0.16

TABLE 1.18 – Performances de l'algorithme Random Forest selon les trois configurations de données.

Le modèle ACP se démarque avec un taux d'erreur réduit (16,67 %) et un temps d'entraînement raisonnable (0,16 s), ce qui en fait le plus efficace parmi les trois. Les modèles complet et Fisher obtiennent des performances identiques (34,31 % d'erreur), mais le modèle Fisher est plus rapide à entraîner (0,06 s contre 0,14 s). Cela montre que la réduction par ACP permet d'optimiser à la fois la précision et l'efficacité.

1.10 Application des meilleurs classifieurs sur les données inconnues

Méthode	Modèle utilisé	Taux d'erreur (%)
K-NN	ACP (k = 5)	16.67
LDA	Complet	0.00
QDA	ACP	15.69
SVM	ACP	8.82
Régression logistique	Complet	13.73
Logistique Multinomiale	Complet	3.92
Random Forest	ACP	16.67

TABLE 1.19 – Taux d'erreur des meilleurs modèles par méthode de classification.

Après avoir comparé les performances des différents algorithmes, nous avons retenu deux modèles particulièrement efficaces pour la classification des papiers photographiques : le modèle SVM avec données réduites par ACP et la régression logistique multinomiale appliquée au jeu de données complet.

Ces deux modèles ont ensuite été utilisés pour prédire les classes des observations contenues dans le fichier `inconnu.txt`, dont les étiquettes ne sont pas connues à l'avance.

Prédictions SVM (ACP) :

1	2	3	4	5	6	7	8	9
agfa	agfa	agfa	fuji	fuji	fuji	agfa	kodak	kodak

Prédictions Régression Logistique Multinomiale (complète) :

1	2	3	4	5	6	7	8	9
agfa	agfa	agfa	fuji	fuji	fuji	kodak	kodak	kodak

On observe que les deux modèles donnent des résultats cohérents sur l'ensemble des neuf observations, avec seulement un léger décalage dans la classe prédite pour l'observation 7. Globalement, ces prédictions confirment la capacité des modèles choisis à généraliser sur de nouvelles données.

CONCLUSION

Dans ce projet, nous avons testé plusieurs méthodes de classification pour reconnaître automatiquement le type de papier photographique (*Agfa*, *Fuji* ou *Kodak*) à partir de leurs spectres.

Nous avons utilisé des méthodes classiques comme **LDA**, **QDA** ou la **régression logistique**, ainsi que des modèles plus avancés comme **SVM**, **Random Forest** et **K-NN**.

Pour améliorer les résultats, nous avons réduit le nombre de variables grâce à deux techniques : **l'Analyse en Composantes Principales (ACP)** et la **sélection par le critère de Fisher**. Cela permet de conserver les informations les plus importantes tout en évitant les problèmes liés à la grande dimension des données.

Parmi tous les modèles testés, les meilleurs résultats ont été obtenus avec :

- **SVM avec ACP**, qui donne une précision élevée (91,18 %) avec un temps de calcul court,
- **La régression logistique multinomiale** sur les données complètes, avec une précision de 96,08 %.

Ces deux modèles ont ensuite été utilisés pour prédire les classes d'un nouveau jeu de données sans étiquettes. Les résultats étaient cohérents, ce qui montre que les modèles sont capables de bien généraliser.

Ce projet met en avant l'importance de bien préparer les données et de choisir un bon algorithme pour réussir une tâche de classification.