Analyse de données, classification et réduction de dimension – Projet Wine Dataset

Projet réalisé par : Alim Liadi

Objectifs du projet

- Analyser les relations entre variables (corrélations)
- Classifier des observations avec KNN
- Gérer des données manquantes
- Réduire la dimension (PCA, LDA)
- Vérifier la robustesse des modèles

Description du dataset

- Dataset Wine (178 observations, 13 variables)
- 3 classes : Types de vins
- Variables numériques (alcohol, flavanoids, proline, etc.)
- Tableau du dataset et des classes :

Partie A

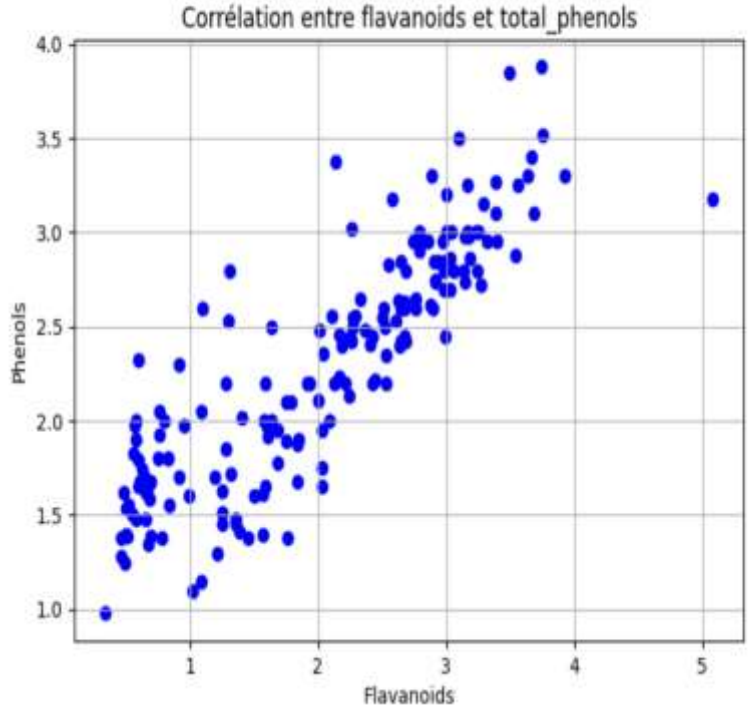
```
     alcohol  malic_acid   ash  alcalinity_of_ash  magnesium  total_phenols  \
0      14.23        1.71  2.43               15.6      127.0           2.80
1      13.20        1.78  2.14               11.2      100.0           2.65
2      13.16        2.36  2.67               18.6      101.0           2.80
3      14.37        1.95  2.50               16.8      113.0           3.85
4      13.24        2.59  2.87               21.0      118.0           2.80
..       ...         ...   ...                ...        ...            ...
173    13.71        5.65  2.45               20.5       95.0           1.68
174    13.40        3.91  2.48               23.0      102.0           1.80
175    13.27        4.28  2.26               20.0      120.0           1.59
176    13.17        2.59  2.37               20.0      120.0           1.65
177    14.13        4.10  2.74               24.5       96.0           2.05

     flavanoids  nonflavanoid_phenols  proanthocyanins  color_intensity   hue  \
0          3.06                  0.28             2.29             5.64  1.04
1          2.76                  0.26             1.28             4.38  1.05
2          3.24                  0.30             2.81             5.68  1.03
3          3.49                  0.24             2.18             7.80  0.86
4          2.69                  0.39             1.82             4.32  1.04
..          ...                   ...              ...              ...   ...
173        0.61                  0.52             1.06             7.70  0.64
174        0.75                  0.43             1.41             7.30  0.70
175        0.69                  0.43             1.35            10.20  0.59
176        0.68                  0.53             1.46             9.30  0.60
177        0.76                  0.56             1.35             9.20  0.61

     od280/od315_of_diluted_wines  proline
0                            3.92   1065.0
1                            3.40   1050.0
2                            3.17   1185.0
3                            3.45   1480.0
4                            2.93    735.0
..                            ...      ...
173                          1.74    740.0
174                          1.56    750.0
175                          1.56    835.0
176                          1.62    840.0
177                          1.60    560.0

[178 rows x 13 columns]
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2]
```

Partie A : Corrélations

- Utilisation du coefficient de Pearson
- Identification des variables fortement corrélées

Représentation graphique et observation de la corrélation

Représentatio graphique de la relation
Coefficient de corrélation de Pearson : r = 0.8645635000951151



Les 10 paires de variables les plus corrélées :

total_phenols	flavanoids	0.864564
flavanoids	total_phenols	0.864564
od280/od315_of_diluted_wines	flavanoids	0.787194
flavanoids	od280/od315_of_diluted_wines	0.787194
total_phenols	od280/od315_of_diluted_wines	0.699949
od280/od315_of_diluted_wines	total_phenols	0.699949
proanthocyanins	flavanoids	0.652692
flavanoids	proanthocyanins	0.652692
alcohol	proline	0.643720
proline	alcohol	0.643720

dtype: float64

Partie B1 : Classification KNN

- KNN (k=5)
- Normalisation obligatoire
- Prédiction sur une nouvelle observation :
  - new_d = [13, 2, 2, 20, 99, 2, 2, 0.4, 2, 5, 1, 2.5, 500]
  - Classe prédite : 1

Partie B2 : Données manquantes

Méthode 1 : moyenne

Méthode 2 : imputation KNN

Impact sur la classe prédite

Données manquantes

- Tableau avec valeurs manquantes : new_d = [X, X, 3, 15, 80, 3, 1, 0.3, 2, 5, 1, 2.5, 500]

- Imputation par la moyen :
  - Resultats : : [12.97, 2.34, 3, 15, 80, 3, 1, 0.3, 2, 5, 1, 2.5, 500]
  - Classe predicte : 0
  - Impacte sur la classe predicte : Les valeurs remplies sont des valeurs générales, pas spécifiques à un type de vin.

- Imputation knn :
  - Resultats : [12.258, 1.502, 3, 15, 80, 3, 1, 0.3, 2, 5, 1, 2.5, 500]
  - Classe predicte : 1
  - Impacte sur la classe predicte : Les valeurs manquantes sont remplies en fonction des vins similaires (voisins proches).

Partie C1 : Comparaison des techniques de réduction

Nous avons comparé deux méthodes principales :

- PCA (Analyse en Composantes Principales) est une technique non supervisée qui projette les données sur les axes expliquant le plus de variance. Elle conserve l'information globale mais ne tient pas compte des classes.
- LDA (Linear Discriminant Analysis) est une technique supervisée qui cherche à maximiser la séparation entre les classes. Étant donné que le dataset Wine comporte 3 classes, LDA permet une réduction maximum à 2 dimensions

PCA :

- PCA testé avec 2, 3, 5, 7, 10 composantes
- Perte d'information si composantes trop faibles (les composantes supplémentaires apportent du bruit et réduisent la qualité de précision)
- Affichage des résultats :
  - PCA avec 2 composantes :1.0000
  - PCA avec 3 composantes : 0.9722
  - PCA avec 5 composantes : 0.9444
  - PCA avec 7 composantes : 0.9722
  - PCA avec 10 composantes : 0.9444

LDA :

- LDA testé avec 2 composantes car il y a 3 classes
- LDA testé avec 2 composantes : 1.0

Partie C2 : Impact de la réduction sur la performance de classification

Nous avons appliqué la classification KNN avant et après réduction de dimensionnalité. Sans réduction, la précision du modèle est d'environ 95–99%.

- Avec PCA, la précision dépend du nombre de composantes : 2 composantes : baisse de précision (≈ 85–93%) 3 à 5 composantes : précision proche du modèle original 7 composantes : quasi aucune perte
- Avec LDA, la réduction à 2 dimensions offre les meilleures performances (≈ 96–100%), supérieures à PCA lorsque le nombre de composantes est faible.

Conclusion: LDA est la méthode la plus efficace pour maintenir la performance de classification, car elle préserve l'information discriminante. PCA devient compétitif lorsque le nombre de composantes est supérieur ou égal à 5.

Partie D : Validation et robustesse

- **Utilisation de K-flod, k=5**
  - Pour vérifier que nos résultats ne dépendent pas d'un seul découpage du dataset, nous avons utilisé la validation croisée K-fold, qui est la méthode la plus répandue pour évaluer la robustesse d'un modèle.

- **Principe de fonctionnement**
  - Le principe de cette méthode est de diviser les données en K blocs, d'entraîner le modèle sur K–1 blocs, puis de tester sur le bloc restant. Cette opération est répétée K fois avec un bloc différent, puis les précisions obtenues sont moyennées.

Synthese des resultats

Modèles	Moyenne	Variance
Knn	0.955	0.00084
PCA 2D	0.966	0.00012
PCA 3D	0.961	0.00111
PCA 5D	0.966	0.00074
PCA 7D	0.955	0.00081
LDA	0.994	0.00012

A savoir :

- Si le modèle a une très forte variance, il est instable donc il risque de surapprendre
- Si le modèle a une faible variance, il est stable, donc il généralise bien

Conclusion : Les faibles variances obtenues, surtout avec PCA 2D et LDA 2D, montrent que nos modèles généralisent bien et ne surapprennent pas.

Améliorations futures

- Tester des modèles plus puissants comme SVM ou RandomForest
- Utiliser GridSearchCV pour optimiser les paramètres
- Faire une sélection automatique des meilleures variables
- Tester des méthodes d’imputation plus avancées
- Explorer d’autres techniques de réduction de dimension comme UMAP
- Améliorer la normalisation selon les modèles

Conclusion

Ce projet a permis d’appliquer et de comparer plusieurs approches essentielles en apprentissage automatique. Les résultats montrent que les choix de prétraitement, d’imputation et de réduction de dimension influencent fortement la performance de la classification. La normalisation s’est révélée indispensable, tandis que LDA s’est démarquée comme la technique la plus efficace pour maintenir de bonnes performances même en dimension très réduite. Enfin, la validation croisée a confirmé la robustesse des modèles présentés. Ce travail offre une vision complète de l’impact des différentes étapes de préparation et de transformation des données sur un processus de classification supervisée.