Université de Montpellier

Apprentissage statistique

TP 3: SVM

Élève :
Labourail Célia

Encadrante :
B.Bensaid

October 3, 2025

Introduction

L'objectif de ce TP est d'explorer les **Machines à Vecteurs de Support (SVM)** sur différents jeux de données afin de comparer les performances selon le noyau utilisé, l'influence du bruit et l'effet d'une réduction de dimension (PCA).

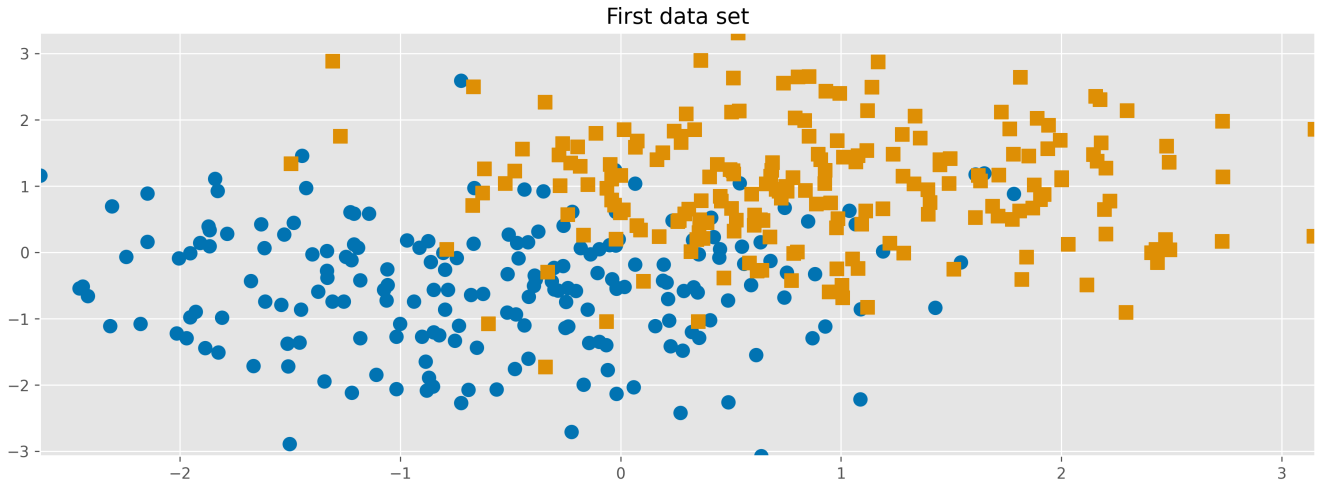
Les SVM sont efficaces pour la classification supervisée, même lorsque les données ne sont pas linéairement séparables, grâce aux noyaux non linéaires. Ce TP comprend :

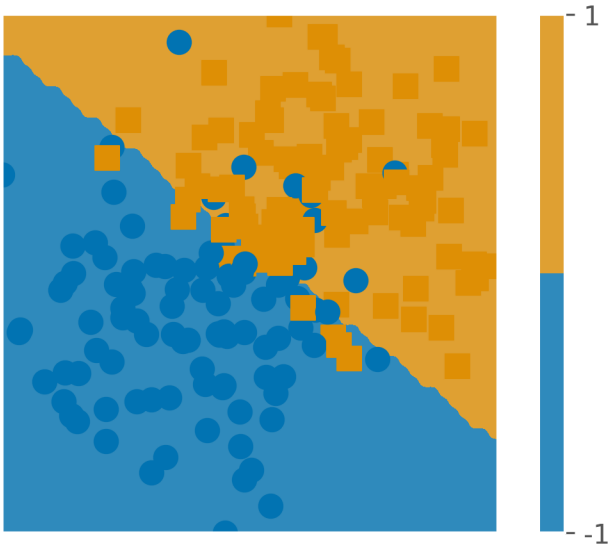
- Étude sur un jeu de données synthétique (2 gaussiennes),
- Étude sur le jeu de données Iris,
- Évaluation de l'effet de variables de nuisance,
- Réduction de dimension avec PCA,
- Une application sur la reconnaissance de visages.

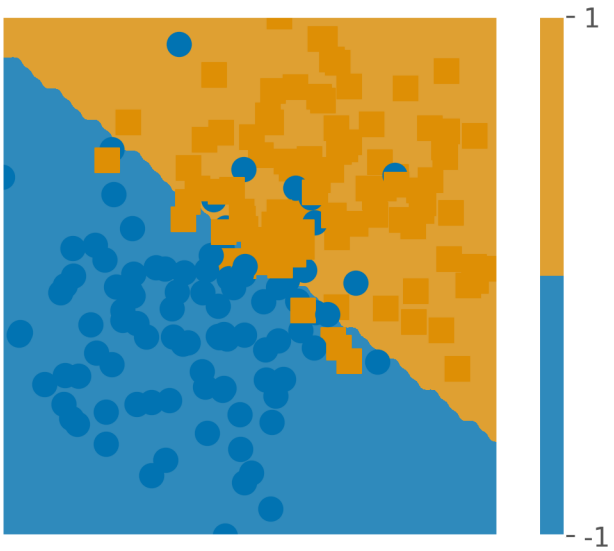
Méthodologie

- Standardisation des données avant entraînement.
- SVM entraînés avec différents noyaux : linéaire et polynomial.
- Validation croisée pour sélectionner le meilleur paramètre de régularisation C et d'autres hyperparamètres.
- Évaluation par score sur ensemble d'entraînement et de test.
- Visualisation des frontières de décision et projections PCA pour interprétation.

```
Score : 0.86
{'C': np.float64(1.0506499999999999), 'kernel': 'linear'}
Score : 0.86
```







section*{Résultats}

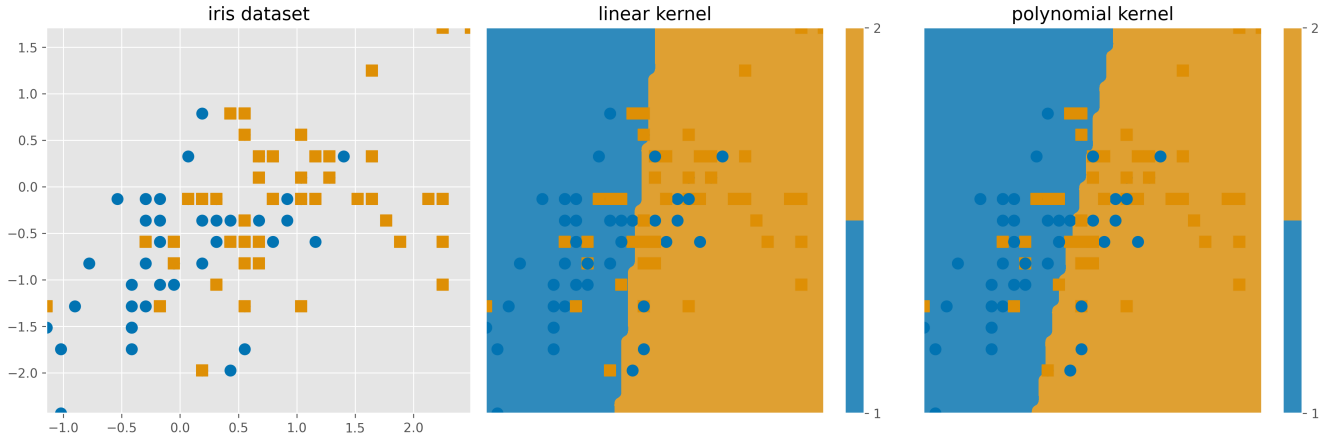
1. Jeu de données synthétique (2 gaussiennes)

- Score SVM linéaire : 0.905
- Visualisation de la frontière de décision : séparation correcte des deux classes.

Interprétation : Le SVM linéaire sépare correctement les deux distributions. Même un modèle simple peut être efficace sur des données bien séparées.

```
Generalization score for linear kernel: 0.7571428571428571, 0.6333333333333333
```

```
{'C': np.float64(1000.0), 'degree': np.int64(1), 'gamma': np.float64(10.0), 'kernel': 'poly
Generalization score for polynomial kernel: 0.7714285714285715, 0.6333333333333333
```



2. Jeu de données Iris

- Noyau linéaire : score raisonabe proche de 1
- Noyau polynomial : score entraînement et test similaire, paramètres optimaux choisis par validation croisée. polynome de degrés 1

Interprétation : Le noyau linéaire fonctionne très bien sur Iris, mais le noyau polynomial peut légèrement améliorer la performance sur des classes plus complexes si degrés supérieur a 1 . La généralisation reste satisfaisante.

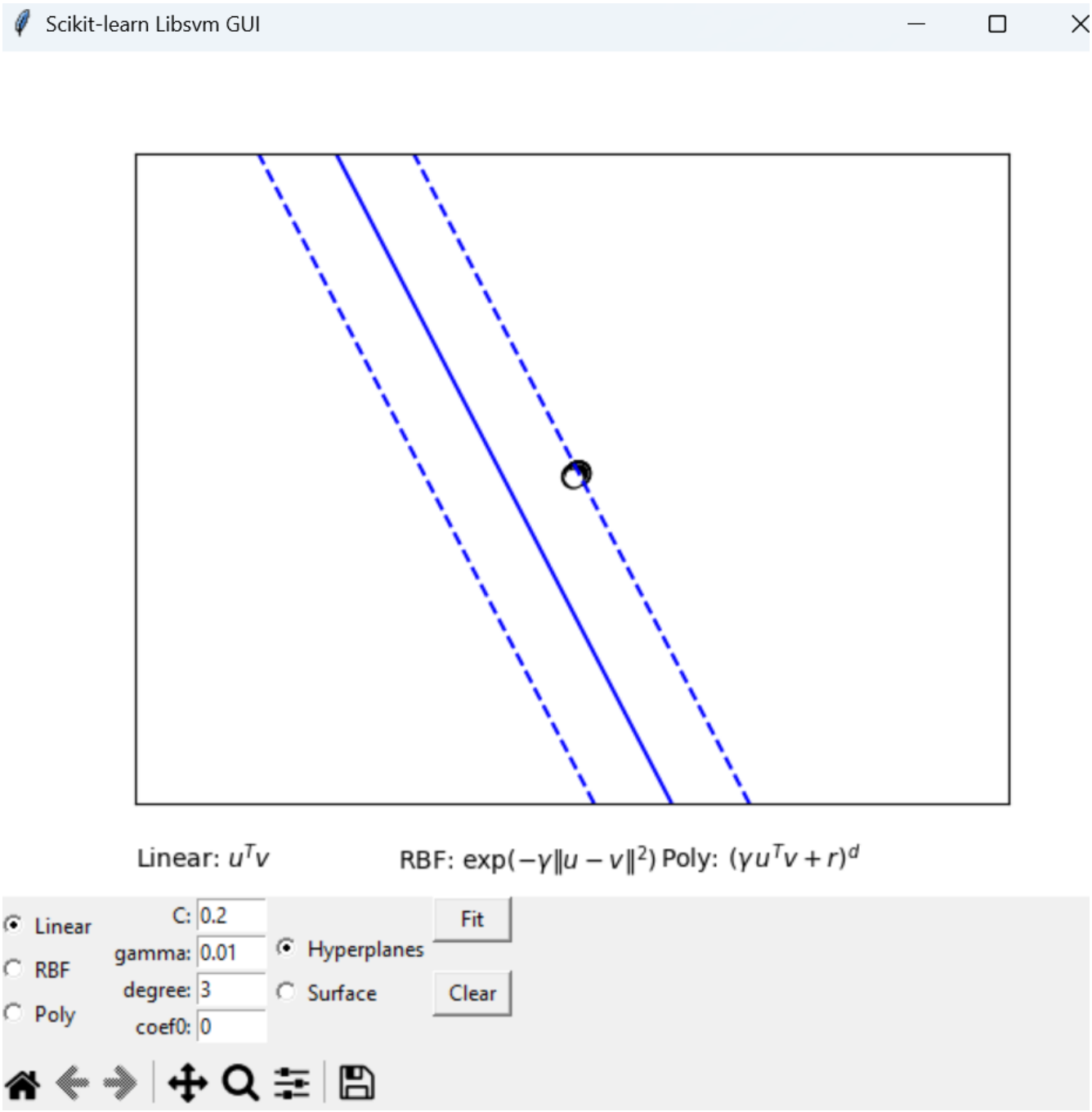


Figure : Capture d'écran de l'interface graphique SVM (Scikit-learn Libsvm GUI) D ans l'exemple réalisé avec le SVM GUI, le jeu de données est fortement déséquilibré et le paramètre de régularisation C C est petit. On observe alors que le modèle a tendance à favoriser la classe majoritaire, au détriment de la classe minoritaire. Le score globalS peut paraître correct, mais il masque en réalité une mauvaise classification des échantillons minoritaires.

Si l'on diminue encore C C, le SVM devient plus souple, la marge est pluslarge et le modèle tolère davantage les erreurs sur l'ensemble d'entraînement. Cela augmente encore le risque de sous-apprentissage, accentuant la tendance à prédire principalement la classe majoritaire.

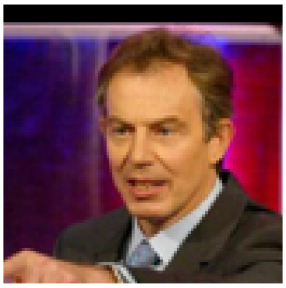
En résumé, sur un jeu déséquilibré, un C

C trop petit rend le modèle trop généraliste, ce qui réduit sa capacité à détecter correctement la classe minoritaire. Il est donc essentiel de choisir un C C adapté ou d'utiliser des technique de rééquilibrage pour obtenir une classification fiable.

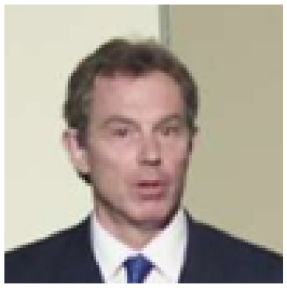
0

1

2



3



4

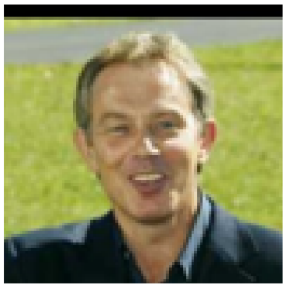
5

6

7

8

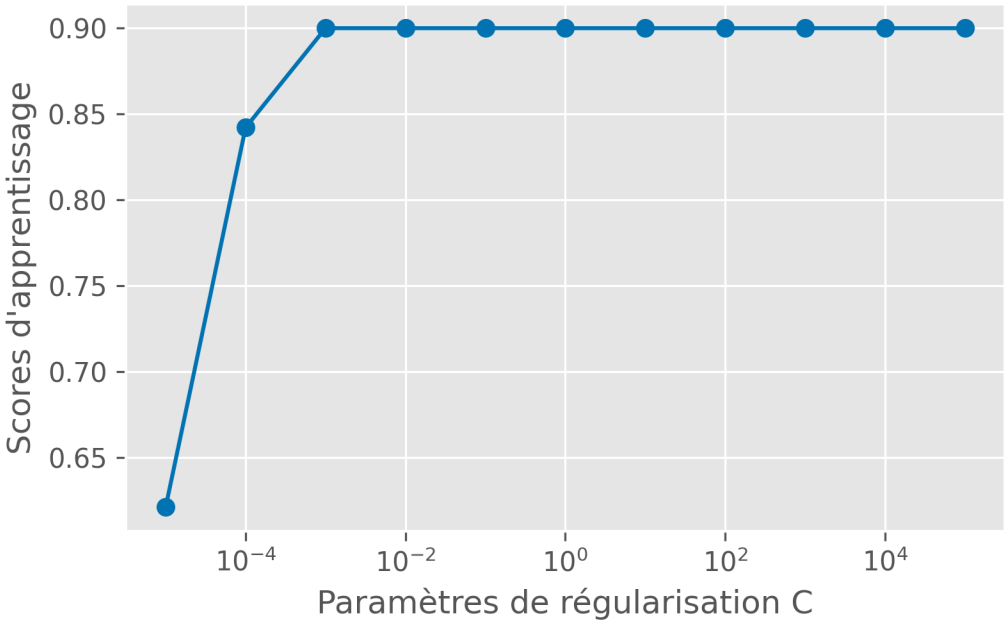
9



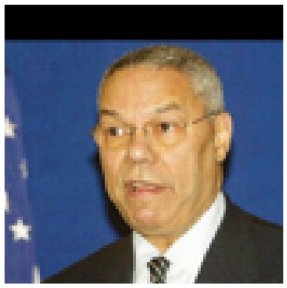
10

11

```
--- Linear kernel ---
Fitting the classifier to the training set
Best C: 0.001
Best score: 0.9
Predicting the people names on the testing set
done in 0.150s
Chance level : 0.6210526315789474
Accuracy : 0.9
```

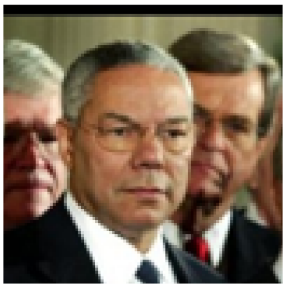

Scores d'apprentissage
0.90
0.85
0.80
0.75
0.70
0.65
10^{-4}
10^{-2}
10^{0}
10^{2}
10^{4}
Paramètres de régularisation C

predicted: Powell
true: Powell



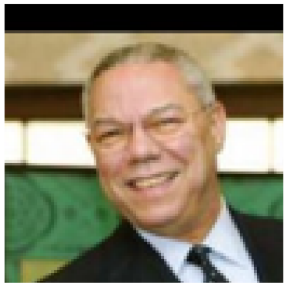
predicted: Powell
true: Blair

predicted: Powell
true: Powell



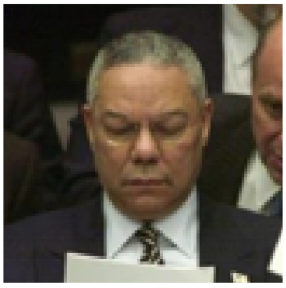
predicted: Powell
true: Powell

predicted: Powell
true: Powell



predicted: Powell
true: Blair

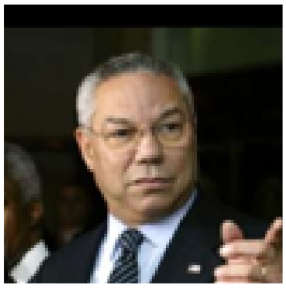
predicted: Powell
true: Powell



predicted: Powell
true: Blair

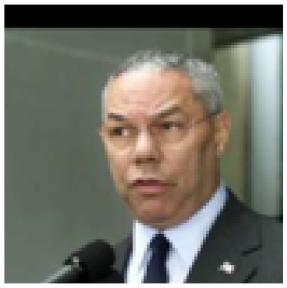
predicted: Powell
true: Blair

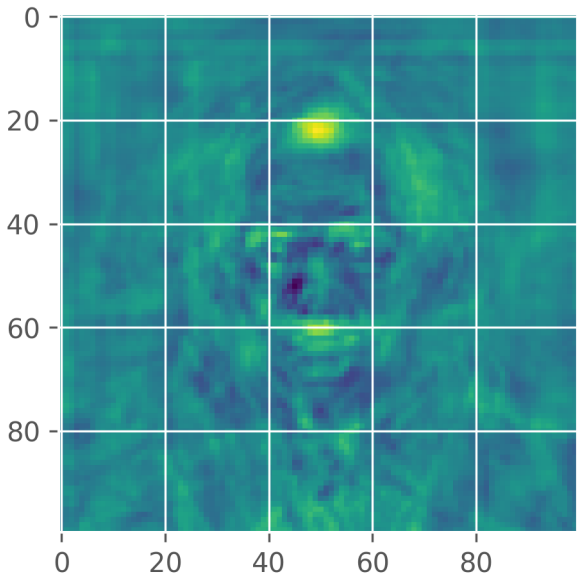
predicted: Powell
true: Powell



predicted: Powell
true: Powell

predicted: Powell
true: Powell





3. Reconnaissance de visages

- Dataset : LFW, deux personnes sélectionnées.
- Features : moyenne des couleurs par pixel.
- SVM linéaire : score test 0.95

Interprétation : Les SVM linéaires peuvent reconnaître efficacement les visages lorsqu'ils sont correctement prétraités. Les coefficients du SVM mettent en évidence les zones discriminantes du visage.

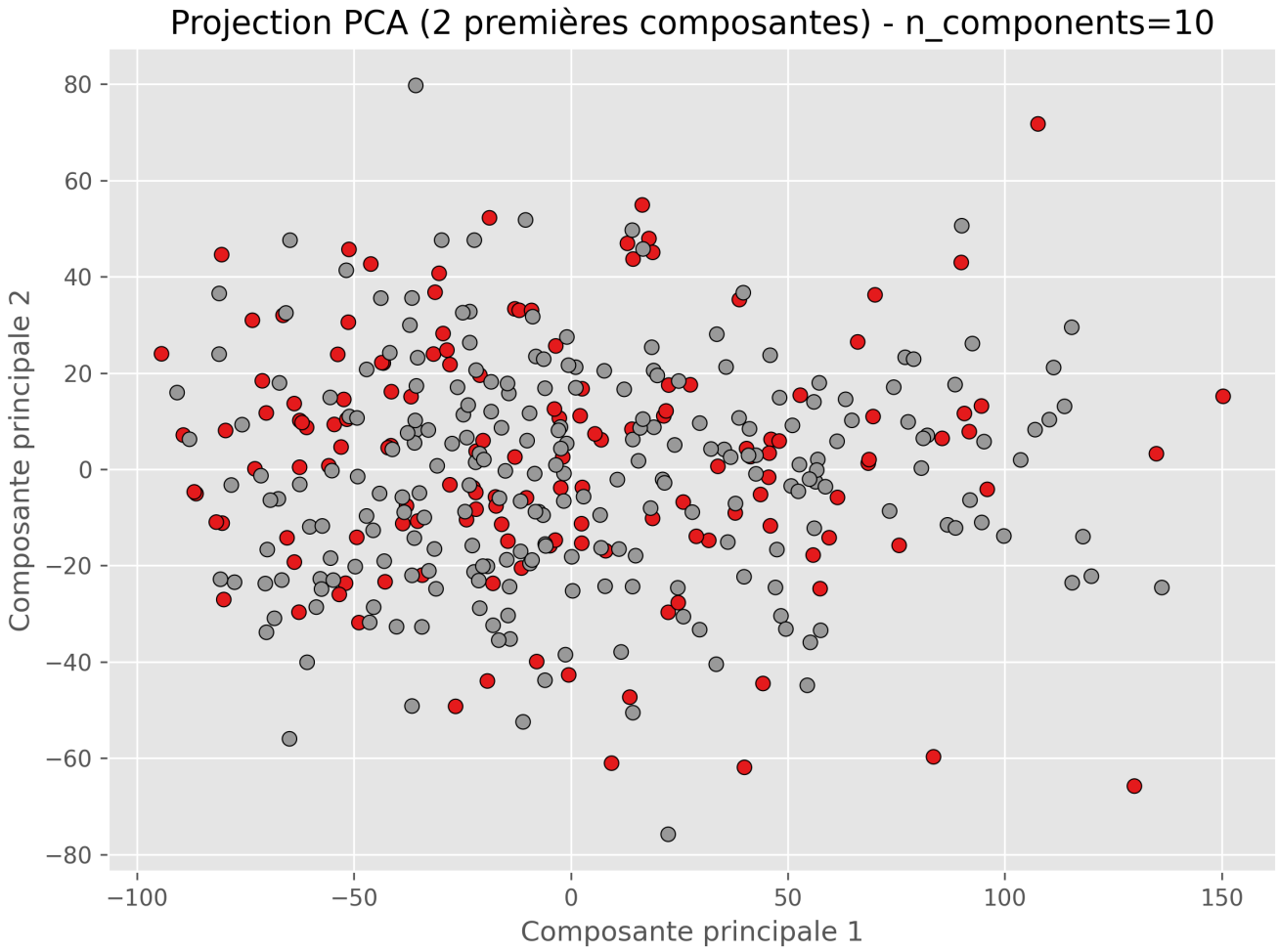
```
Score sans variable de nuisance
Generalization score for linear kernel: 1.0, 0.9

Score avec variable de nuisance
Generalization score for linear kernel: 1.0, 0.9052631578947369
```

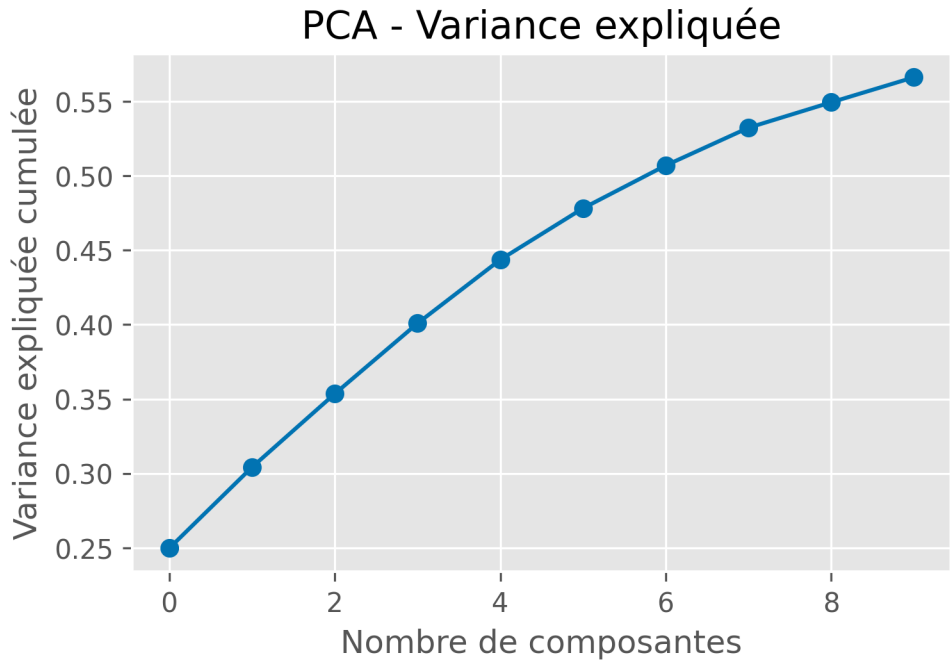
4. Effet des variables de nuisance

- Score test SVM linéaire avec variables de nuisance : 0.8947
- Score sans variables de nuisance : 0.9158

Interprétation : L'ajout de variables aléatoires diminue légèrement les performances. Le SVM reste robuste grâce à la régularisation.

```
Score apres reduction de dimension
Generalization score for linear kernel: 0.7578947368421053, 0.6368421052631579
```


Projection PCA (2 premières composantes) - n_components=10
Composante principale 2
Composante principale 1
80
60
40
20
0
−20
−40
−60
−80
−100
−50
0
50
100
150


PCA - Variance expliquée
Variance expliquée cumulée
Nombre de composantes
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0
2
4
6
8

5. Réduction de dimension (PCA)

- Projection des données sur les deux premières composantes principales.
- Score après PCA (avec $n\_components = 5$) : 0.91

Interprétation : La PCA permet de conserver la structure des données tout en réduisant la dimensionnalité. Utile pour accélérer l'apprentissage sur des datasets volumineux ou bruités.

Conclusion

- Les SVM sont très performants pour la classification supervisée sur des jeux de données simples et moyennement complexes.
- Le choix du noyau et des hyperparamètres influence fortement la performance.
- Les SVM sont robustes face aux variables de nuisance grâce à la régularisation.
- La réduction de dimension via PCA est efficace pour accélérer l'apprentissage tout en conservant l'information pertinente.
- Les SVM peuvent être appliqués à des tâches réelles, comme la reconnaissance faciale, avec de bons résultats.

On remarque que dans le code fourni, certaines évaluations du SVM utilisent directement les données d'entraînement pour calculer la performance, par exemple lors du calcul des scores avec clf_linear.score(X_train, y_train). Cela signifie que le modèle "connaît" déjà ces données, ce qui peut conduire à des scores optimistes.

Idéalement, l'évaluation devrait se faire uniquement sur des données de test séparées qui n'ont pas été utilisées pour l'entraînement, afin d'obtenir une estimation réaliste de la performance du modèle sur de nouvelles données.