HAX907X Support Vector Machine

SCAIA Matteo GILLET Louison

2025–2026

Table des matières

Table des figures

1 Classification de la classe 1 contre la classe 2

Dans un premier temps, on évite que les données soient triées par classe en permutant aléatoirement les lignes de `X` et les valeurs correspondantes de `y`, grâce à l'option `shuffle=True` [1] de la fonction `train_test_split`. L'ensemble est ensuite scindé en deux parties : un jeu d'apprentissage représentant 75% des données et un jeu de test représentant les 25% restants.

1.1 Noyau linéaire

En se basant sur les définitions et notations de l'énoncé, nous utilisons un noyau linéaire de la forme :

$$K(x, x') = \langle x, x' \rangle$$

Nous obtenons la sortie suivante :
Avec un SVM à noyau linéaire, le meilleur paramètre trouvé est $C \approx 0.074$. Cela montre que le modèle accepte plus facilement des erreurs sur les points mal classés afin de privilégier une frontière de décision plus souple. Le modèle obtient un score d'apprentissage d'environ 0.75. Ce résultat indique qu'il se trompe déjà sur près de 25 % des données d'entraînement. Sur l'ensemble de test, il obtient un score de 0.68. Ces résultats traduisent une faible capacité de généralisation, le modèle ne parvient pas à bien dissocier les deux classes.

1.2 Noyau polynomial

On cherche à comparer ces résultats avec un SVM basé sur un noyau polynomial. Ce noyau est de la forme :

$$K(x, x') = (\alpha + \beta\langle x, x' \rangle)^{\delta} \quad \text{pour un } \delta > 0$$

Cette fois, nous obtenons les résultats suivants :
La validation croisée a retenu un polynôme de degré 1, ce qui correspond en réalité à un modèle linéaire. Cela explique le fait que les scores d'apprentissage et de test soient identiques au SVM à noyau linéaire.

Nous proposons finalement de visualiser graphiquement nos données afin de mieux comprendre leur répartition et les phénomènes observés.

1. Pour la reproductibilité, on fixe la graine à 42 : `random_state=42`.

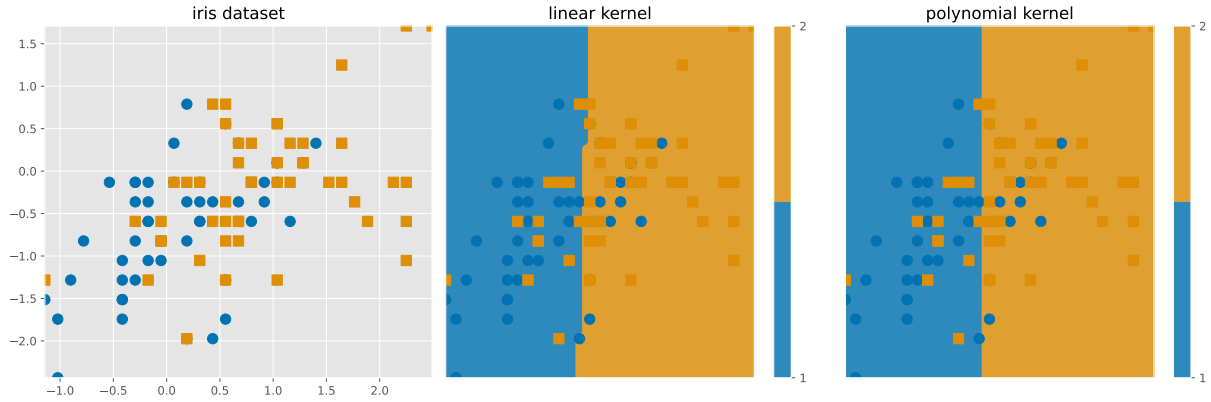

FIGURE 1 – Représentation du jeu de données Iris et classification par SVM à noyau linéaire et polynomial

On observe sur la figure 1 que de nombreux points des deux classes se chevauchent ce qui rend impossible une séparation nette par une frontière, qu'elle soit linéaire ou polynomiale. On retrouve ce problème dans les deux représentations où l'on voit que de nombreux points se retrouvent mal classifiés.
En conclusion, ces résultats montrent que dans l'espace réduit aux deux premières variables, la classification des deux classes d'iris est difficile car elles ne sont pas linéairement séparables.

1.3 Remarque

Par ailleurs, pour mettre en évidence l'impact crucial du mélange initial des données, nous proposons d'analyser les performances du modèle en fixant l'option `shuffle=False` dans la fonction `train_test_split`. On obtient les scores suivants :
On observe que les scores d'apprentissage sont proches de ceux obtenus précédemment. En revanche, les scores de test chutent de manière drastique ce qui révéle une très mauvaise capacité de généralisation. Cela montre que lorsque les données ne sont pas mélangées, l'ensemble d'entraînement peut ne pas être représentatif de l'ensemble de test, et le modèle apprend alors des caractéristiques spécifiques aux données d'entraînement plutôt que des motifs généraux.

2 SVM GUI

On cherche à évaluer l'impact du choix du noyau et du paramètre C dans un SVM. Pour cela, on génère un jeu de données très déséquilibré de deux classes, qu'on représente sur la figure 2.

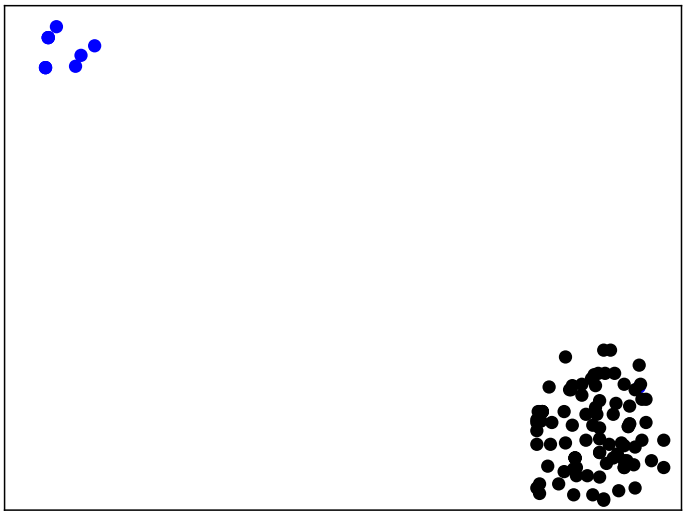

FIGURE 2 – Représensation d'un jeu de données très déséquilibré

Ensuite, on ajuste le SVM pour différentes valeurs de C. On observe qu'en réduisant ce paramètre, le modèle accepte plus facilement les erreurs de classification.

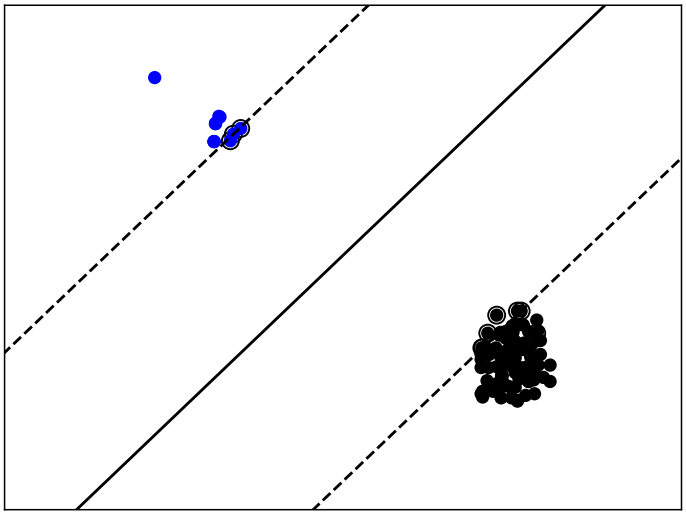

FIGURE 3 – Classification par SVM à noyau linéaire pour C=1

Sur la figure 3 la frontière de décision sépare correctement les deux classes et se situe globalement à égale distance des groupes de points. La séparation reste équilibrée et la classe minoritaire est encore bien prise en compte.

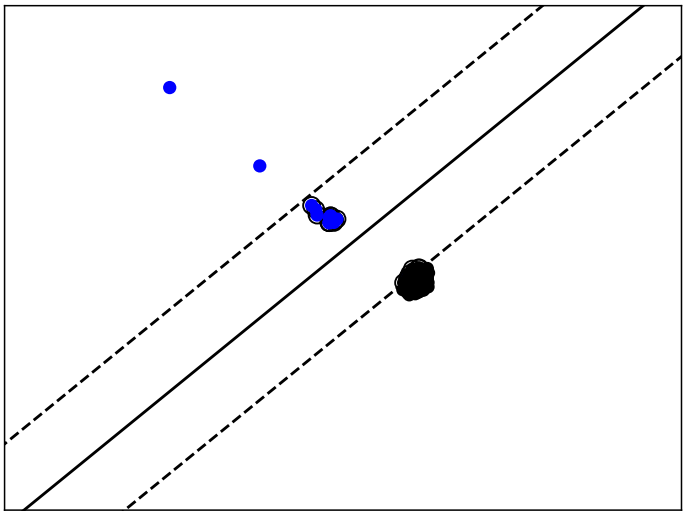

FIGURE 4 – Classification par SVM à noyau linéaire pour C=0.1

Lorsque l'on diminue C à 0.1, comme illustré sur la figure 4, la frontière se déplace progressivement vers la classe minoritaire. En effet, un nombre important de points se retrouve dans la marge. Ces points ne sont pas suffisamment éloignés de l'hyperplan pour permettre une bonne classification et entraînent une pénalité.

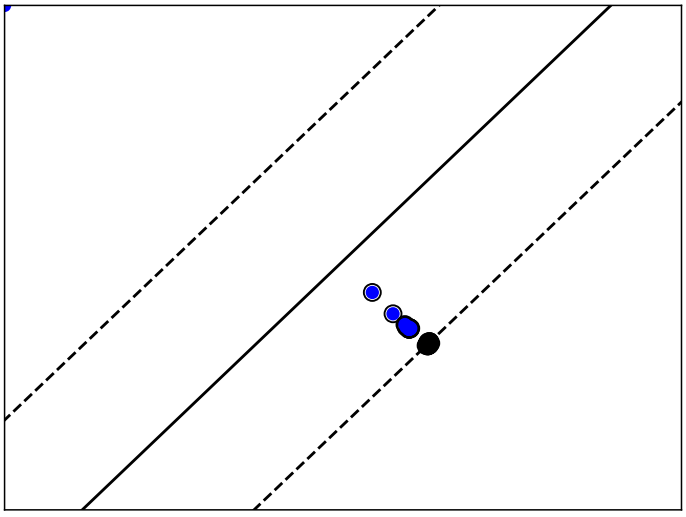

FIGURE 5 – Classification par SVM à noyau linéaire pour C=0.01

Enfin, en réduisant encore C à 0.01, la figure 5 montre que la classe minoritaire est quasiment ignorée. La frontière de décision s'aligne presque exclusivement sur la bonne classification de la classe majoritaire, au détriment complet de la minorité.

Ainsi, dans le cas d'un dataset très déséquilibré, la diminution de C entraîne systématiquement un déplacement de la frontière de décision vers la classe minoritaire, renforçant le biais en faveur de la classe majoritaire.

En pratique, il peut être corrigé en pondérant davantage les erreurs sur la classe minoritaire à l'aide du paramètre class_weight="balanced" de la classe SVC. Une autre solution consiste à recalibrer les probabilités (probability=True) ou encore à recourir à des techniques de rééchantillonnage (oversampling de la minorité ou undersampling de la majorité).

3 Classification de visages

Nous nous intéressons à présent à la problématique de la classification des visages. Nous utilisons ici un jeu de données qui est un extrait prétraité de "Labeled Faces in the Wild" (LFW). Dans ce rapport, nous avons entraîné notre modèle sur deux visages, ceux de Tony Blair et de Colin Powell. Une version équivalente, réalisée sur deux autres visages (Donald Rumsfeld et Colin Powell) est présentée en annexe.

3.1 Impact du paramètre de régularisation

Afin d'analyser l'impact du paramètre de régularisation, nous étudions la variation de l'erreur de prédiction en fonction de la valeur de C, représentée sur une échelle logarithmique comprise entre 10^{-5} et 10^{5}.
Nous obtenons une valeur optimale $C = 0.001$, associée à un score de 1 sur l'échantillon d'apprentissage. Ce résultat suggère que le modèle est parvenu à séparer parfaitement les données d'entraînement.

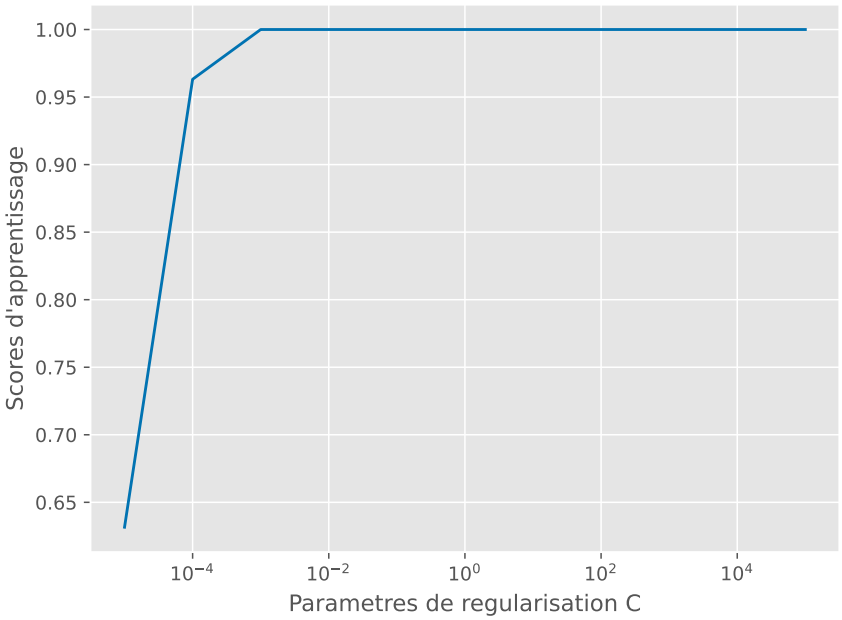

FIGURE 6 – Score d'apprentissage en fonction du paramètre de régularisation C

La figure 6 montre que l'augmentation de C améliore le score d'apprentissage. En revanche, un choix trop élevé de C pourrait amener le modèle à trop bien s'ajuster aux données

d'entraînement, au détriment de sa capacité de généralisation, on risque un phénomène de sur-apprentissage.

Nous utilisons désormais un SVM à noyau linéaire avec le paramètre C optimal sur notre jeu de données.
Nous obtenons le résultat suivant :
Pour évaluer la performance de notre SVM à noyau linéaire sur le jeu de visages, nous la comparons à un niveau de hasard. La proportion d'images correctement classées au hasard est de 62,1%, ce qui correspond au niveau de chance. En utilisant le paramètre de régularisation optimal C, le SVM linéaire atteint une précision de 90,5% sur le jeu de test, représentant une amélioration significative par rapport au hasard.

FIGURE 7 – Comparaison des résultats d'un SVM linéaire avec C optimal

La figure 7 présente un extrait des prédictions de notre modèle pour la reconnaissance faciale. On constate que sur ce petit échantillon, le modèle a parfaitement classé les images.

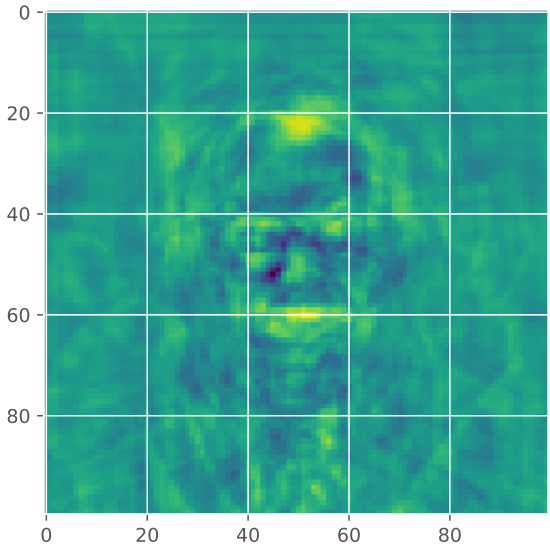

FIGURE 8 – Visualisation des coefficients

De plus, la figure 8 illustre que le modèle parvient à repérer les traits distinctifs des visages, en accordant un poids particulièrement élevé aux coefficients associés aux yeux, à la bouche et aux cheveux.

Ainsi, le classificateur est capable de distinguer efficacement les deux individus à partir des caractéristiques extraites des images, et qu'il généralise correctement à de nouvelles données. Le choix de C optimal a permis un compromis efficace entre sous-apprentissage et sur-apprentissage.

3.2 Introduction de variables de nuisances

Afin d'étudier la robustesse de notre modèle, nous ajoutons maintenant du bruit aux données d'entrée. Plus précisément, nous générons des variables de bruit suivant une distribution gaussienne de moyenne nulle et de variance égale à 1, que nous concaténons aux caractéristiques initiales.
Nous trouvons les résultats suivants.
On observe que malgré un nombre d'échantillons d'apprentissage identique, l'ajout de variables bruitées entraîne une chute du score de généralisation de $0,91$ à $0,49$. Cela illustre clairement que l'augmentation du nombre de variables, lorsqu'elles ne sont pas pertinentes, diminue la capacité du modèle à généraliser correctement.

3.3 Réduction de dimension

Afin d'améliorer la capacité de généralisation du classifieur, nous appliquons une réduction de dimension par Analyse en Composantes Principales, ce qui permet de projeter les données dans un sous-espace de plus petite dimension tout en conservant de l'information.
Nous avons choisi de mettre en place une boucle faisant varier le nombre de composantes principales retenues lors de la réduction de dimension par ACP. L'objectif est d'observer

l'influence de cette réduction progressive sur les performances du classifieur SVM, à la fois sur l'échantillon d'apprentissage et sur l'échantillon de test.

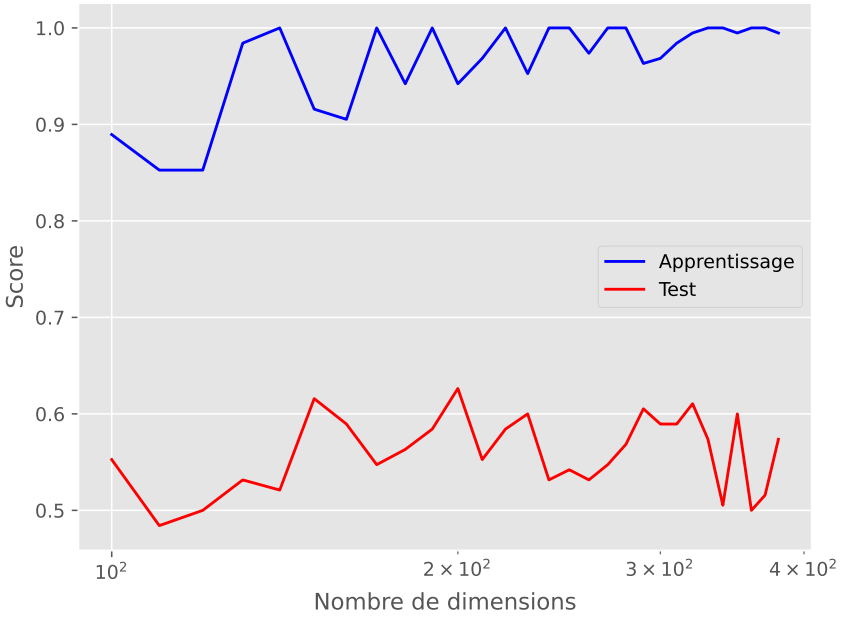

FIGURE 9 – Influence de la réduction de dimensions sur les scores

Les résultats présentés par la figure 9 montrent que l'augmentation du nombre de dimensions retenues par l'ACP n'améliore pas considérablement le score de test. On observe donc un compromis, conserver trop de dimensions favorise le sur-apprentissage, tandis qu'une réduction trop forte risque de dégrader la représentation des données.

3.4 Biais dans le pré-traitement

Dans notre prétraitement des données, un biais se situe au niveau de la conversion des images couleur en niveaux de gris.
En moyennant les trois canaux de couleur, nous perdons toute information liée à la couleur. Or, certaines personnes peuvent être mieux identifiables grâce à des caractéristiques dûes aux couleurs comme la teinte de peau, les cheveux, les vêtements et l'éclairage. Cette transformation peut donc introduire un biais d'information : le modèle ne peut plus exploiter certains indices potentiellement discriminants et pourrait être moins performant pour certaines images ou individus. Ainsi, la conversion en niveaux de gris simplifie le modèle mais modifie la représentation originale des données, ce qui constitue un biais dans le prétraitement.

4 Annexe

4.1 Classification de visages (Donald Rumsfeld, Colin Powell)

Afin d'analyser l'impact du paramètre de régularisation, nous étudions la variation de l'erreur de prédiction en fonction de la valeur de C, représentée sur une échelle logarithmique comprise entre 10^{-5} et 10^5.

Nous obtenons une valeur optimale $C = 0.001$, associée à un score de 1 sur l'échantillon d'apprentissage. Ce résultat suggère que le modèle est parvenu à séparer parfaitement les données d'entraînement.

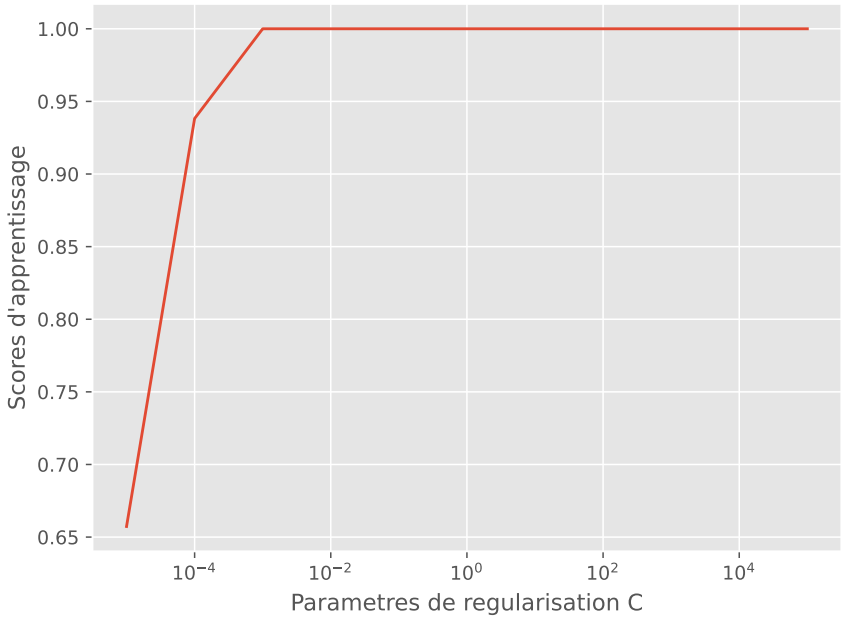

FIGURE 10 – Score d'apprentissage en fonction du paramètre de régularisation C

La figure 10 montre que l'augmentation de C améliore le score d'apprentissage. En revanche, un choix trop élevé de C pourrait amener le modèle à trop bien s'ajuster aux données d'entraînement, au détriment de sa capacité de généralisation, on risque un phénomène de sur-apprentissage.

Nous utilisons désormais un SVM à noyau linéaire avec le paramètre C optimal sur notre jeu de données.
Nous obtenons le résultat suivant :
Pour évaluer la performance de notre SVM à noyau linéaire sur le jeu de visages, nous la comparons à un niveau de hasard. La proportion d'images correctement classées au hasard est de 66,1%, ce qui correspond au niveau de chance. En utilisant le paramètre de régularisation optimal C, le SVM linéaire atteint une précision de 89,3% sur le jeu de test, représentant une amélioration significative par rapport au hasard.

FIGURE 11 – Comparaison des résultats d'un SVM linéaire avec C optimal

La figure 11 présente un extrait des prédictions de notre modèle pour la reconnaissance faciale. On constate que sur ce petit échantillon, le modèle a parfaitement classé les images.

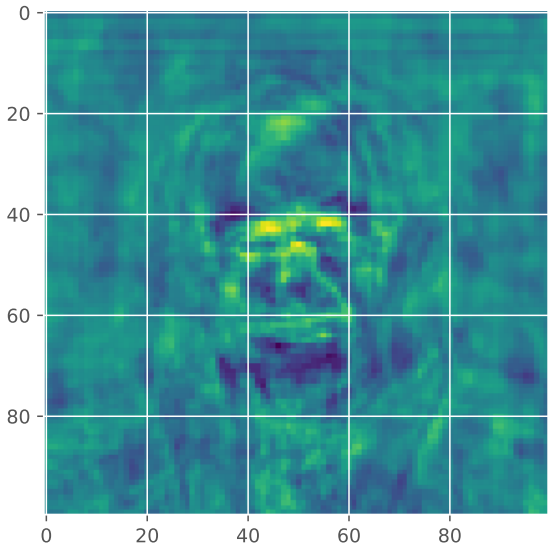

FIGURE 12 – Visualisation des coefficients

De plus, la figure 12 illustre que le modèle parvient à repérer les traits distinctifs des visages,

en accordant un poids particulièrement élevé aux coefficients associés aux yeux, à la bouche et aux cheveux. Dans cette nouvelle étude, le modèle semble également distinguer les lunettes.

Ainsi, le classificateur est capable de distinguer efficacement les deux individus à partir des caractéristiques extraites des images, et qu'il généralise correctement à de nouvelles données. Le choix de C optimal a permis un compromis efficace entre sous-apprentissage et sur-apprentissage.