GTI-770

Alessandro Koerich

ÉTS

# Projet de session

Olivier Arteau

Simon Grondin

27 juillet 2015

# Description du problème

Le problème consiste à classifier correctement le type de musique en fonction de différentes primitives fournies. Un ensemble d'entraînement est donné avec des étiquettes et un ensemble de test est donné sans étiquette. Il faut donc correctement classifier les éléments de l'ensemble de test.

Les classes de sortie sont les suivantes : Big Band, Blues Contemporary, Country Traditional, Dance, Electronica, Experimental, Folk International, Gospel, Grunge Emo, Hip Hop Rap, Jazz Classic, Metal Alternative, Metal Death, Metal Heavy, Pop  Contemporary, Pop Indie, Pop Latin, Punk, Reggae, Rnb Soul, Rock Alternative, Rock College, Rock Contemporary, Rock Hard et Rock Neo Psychedelia. Ce qui va rendre la tâche très ardu est que plusieurs de ces classes sont très près une de l'autre. Par exemple, les styles Rock, Metal et Pop ont plusieurs déclinaisons qui ont peu de différences entre elles. Le nombre élévé de classes de sortie risque de faire en sorte que les taux de succès seront relativement très faibles même avec un bon modèle.

# Conception de la méthode

## Choix d'ensembles de données et d'algorithmes d'apprentissage

Afin de déterminer quelles données d'entrées allaient être choisies, nous avons roulé des tests minimaux sur les ensembles de données afin d'avoir une vague idée de quels ensembles ont le plus de potentiel. Les algorithmes d'apprentissage sélectionnés dans cette étape ont été choisis en fonction de leur vitesse et de leur diversité. Pour ce faire nous avons effectué les étapes suivantes pour chaque ensemble de données :

1. 10% des données ont été selectionnées.
2. La méthode de validation des données a été définie à “Cross-Validation” “10 folds”
3. Nous avons observé les résultats de l'apprentissage avec l'algorithme Bayes Naive (Supervised Descritization)
4. Nous avons observé les résultats de l'apprentissage avec l'algorithme J48.
5. Nous avons observé les résultats de l'apprentissage avec l'algorithme KNN (k=3) avec les données filtrés avec le filtre “Standardize”.

Les résultats obtenus sont les suivants :

| **Ensemble de données** | **Nombre d'attributs** | **Bayes** | **J48** | **KNN** |
|---|---|---|---|---|
| msd-jmirderivatives_dev | 99 | 15% | 11% | 12% |
| msd-jmirlpc_dev | 23 | 14% | 10% | 12% |
| msd-jmirmfccs_dev | 29 | 16% | 11% | 15% |
| msd-jmirmoments_dev | 13 | 13.5% | 10% | 11% |
| msd-jmirspectral_dev | 19 | 14.5% | 11% | 12% |
| msd-marsyas_dev_new | 127 | 16% | 13.5% | 17% |
| msd-mvd_dev | 423 | 12% | 9% | 10.5% |
| msd-rh_dev_new | 63 | 11% | 9% | 9% |
| msd-ssd_dev | 169 | 16% | 12.5% | 16% |
| msd-trh_dev | 421 | 11% | 9% | 10% |

À partir de ces données, nous avons déterminer nous avons sélectionné l'ensemble de données “msd-marsyas_dev_new”, “msd-ssd_dev” et “msd-jmirmfccs_dev”.

Nous nous sommes aussi basé sur la littérature pour sélection de l'ensemble de données “msd-rh_dev_new”, puisqu'une publication listée mentionnait être capable d'obtenir de très bon taux de succès avec un ensemble de données similaire lorsqu'il était utilisée avec SVM. On a effectivement été capable d'obtenir les taux de succès suivant avec un 10% des données et la même méthode de validation.

Voici les résultats obtenus avec différent noyau de SVM :

| **Noyau** | **% d'exemple bien classifié** |
|---|---|
| linéaire | 16% |
| polynomiale | 14.5% |
| radiale | 17.5% |
| sigmoïdale | 6.5% |

À la lumière des résultats obtenus précédemment, les combinaisons d'ensemble de données et d'algorithme retenues sont les suivantes :

| **Algorithme** | **Ensemble de données** |
|---|---|
| KNN (k=3) | msd-jmirmfccs_dev, msd-marsyas_dev_new, msd-ssd_dev |
| Bayes (Supervised Discretization) | msd-jmirmfccs_dev, msd-marsyas_dev_new, msd-ssd_dev |
| SVM (radial) | msd-rh_dev_new |

Étant donnée que pour l'algorithme KNN, l'efficacité du paramètre “k” varie en fonction de la quantité de données, ce paramètre a été ajusté sur le modèle final lors de la validation du modèle. Il a été determiné que pour KNN, le “k” optimal était “1”. Ce qui est très peu étonnant lorsque l'on a autant de données d'entraînement.

## Modèle d'apprentissage

3 stratégies d'apprentissage seront combinées pour faire l'apprentissage.

### Stratégie #1

3 modèles KNN seront construits pour les ensembles de données msd-jmirmfccs_dev, msd-marsyas_dev_new et msd-ssd_dev. Ces 3 modèles seront combinés avec un vote.

### Stratégie #2

3 modèles Naive Bayes seront construit pour les ensembles de données msd-jmirmfccs_dev, msd-marsyas_dev_new et msd-ssd_dev. Ces 3 modèles seront combinés avec un vote.

## Stratégie #3

Un module SVM sera construit avec l'ensemble de données “msd-rh_dev_new”.

# Description de l'implémentation

L'implémentation a été faite en Java avec l'API de Weka et la libraire LibSVM. Elle permet à la fois de faire l'apprentissage et l'évaluation du modèle.

## Utilisation

Pour exécuter le projet, chargez le dossier Code comme un projet Eclipse. Ensuite, ajouter une configuration de démarrage. Les paramètres suggérés pour les arguments de la VM sont “-Xms4g -Xmx4g”, sans ces paramètres le programme risque d'échouer avec des exceptions reliés à la mémoire. Ensuite les arguments du programme sont les suivants et ils sont tous obligatoires :

[dossier des données d'entraînement] [dossier des données de test] [dossier où les modèles sont sauvegardés] [fichier de résultats] [faire apprentissage] [faire évaluation]

### Paramètres

**dossier des données d'entraînement** : Dossier où se trouve tous les fichiers “.arff” d'entraînement. Les fichiers doivent être nommé “msd-[nom de l'ensemble]_[suffixe].arff” (les fichiers fournit sur Moodle ont déjà ce format).

**dossier des données de test** : Dossier où se trouve tous les fichiers “.arff” de test. Les fichiers doivent être nommé “msd-[nom de l'ensemble]_[suffixe].arff” (les fichiers fournit sur Moodle ont déjà ce format).

**dossier où les modèles sont sauvegardés** : Dossier où les modèles sérializés sont sauvegardés. Lors de l'apprentissage, les modèles seront écrit dans ce dossier. Lors de l'évaluation, les modèles seront lu dans ce dossier.

**fichier de résultats** : Fichier où sera écrit la liste des étiquettes de l'ensemble de test.

**faire apprentissage** : “true” ou “false” selon que l'apprentissage doit être fait ou non.

**faire évaluation** : “true” ou “false” selon que l'évaluation doit être fait ou non.

### Exemple

../Data/ ../Test/ ../Model/ ../resultat.txt true true

## Apprentissage

L'apprentissage construit les trois modèles qui correspondent aux trois stratégies mentionnées plus tôt et construit aussi un modèle qui correspond à un vote de ces trois modèles. Ces modèles sont

sauvegardés de façon serialisée sur le disque. Ceci permet de reprendre le processus d'évaluation, s'il y a eu des problèmes et ce sans avoir à reprendre l'apprentissage.

L'apprentissage utilise directement les classes de Weka pour faire l'apprentissage. Les paramètres utilisés sont les mêmes que les paramètres par défaut de Weka à l'exception de ceux qui ont été mentionnés dans le rapport.

# Évaluation

L'évaluation utilise le modèle qui combine tous les modèles et lui fournit l'ensemble de données qui combine toutes les colonnes des ensembles de données utilisé. Une méthode utilitaire a été faite pour combiner ensemble plusieurs ensemble de données. Cette méthode s'assure aussi de conserver une seule colonne de classe.

L'évaluation utilise directement les classes de Weka pour faire l'évaluation du modèle d'entraînement. Ils sont utilisés pour évaluer l'ensemble des données de tests et afficher le taux de succès.

# Difficultés rencontrées

## Normalisation des données

Étant donnée que l'on utilisait l'algorithme KNN, il était intéressant de normaliser les données avec le filtre “Standardize” pour éviter que certaines colonnes aient trop d'importance. Par contre, il fallait que les données de tests et d'entraînement soient normalisés de la même façon. Pour ce faire, une méthode utilitaire a été faite pour la normalisation. Cette méthode utilise l'ensemble à normaliser et l'ensemble de test. Ceci permet de normaliser l'ensemble de test et d'entraînement de la même façon.

## Colonnes utilisées

Pour l'évaluation du modèle unifié, toutes les colonnes doivent être passées en paramètre. Par contre les algorithmes d'apprentissage ont besoin seulement d'un sous-ensemble de ces données et ils assument que les données se trouvent aux mêmes index que lors de l'apprentissage. Ceci fait en sorte que les résultats sont totalement erronés lors de l'évaluation. Pour règler ce problème, un filtre (AttributeClassifier) a été créé. Lors de l'apprentissage, ce filtre garde en mémoire le nom des colonnes utilisées et lors de l'évaluation, il fournit seulement les colonnes nécessaires dans le même ordre au modèle.

## Nom de colonne pas unique

Pour fournir un seul ensemble de données au modèle unifié, les colonnes des ensembles de données utilisés doivent être combinées. Par contre, l'ensemble de données “ssd” et “rh” ont les mêmes nom de colonne ce qui empêche ces ensembles de données d'être combinés. Pour règler ce problème, les noms de colonnes sont renommés avec un préfix unique au modèle lors de leur chargement. Ceci évite donc le problème des noms de colonne qui sont présents en double.

# Validation et analyse des résultats

## Validation préliminaire

Pour faire la validation préliminaire de l'ensemble du modèle, on a utilisé 5% des données pour l'entraînement et 5% des données pour les tests. Chaque stratégie a été validée individuellement.

### Stratégie Naive Bayes

Taux de succès : 16%

Matrice de confusion :

| | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 26 | 2 | 5 | 2 | 0 | 1 | 1 | 0 | 0 | 3 | 52 | 2 | 0 | 0 | 1 | 3 | 9 | 0 | 1 | 3 | 0 | 0 | 1 | 0 | 0 |
| 14 | 8 | 13 | 3 | 0 | 7 | 3 | 2 | 1 | 25 | 90 | 17 | 4 | 0 | 7 | 7 | 19 | 1 | 3 | 12 | 5 | 0 | 1 | 7 | 0 |
| 22 | 9 | 36 | 5 | 0 | 5 | 2 | 1 | 2 | 42 | 94 | 29 | 4 | 0 | 31 | 10 | 14 | 0 | 3 | 24 | 4 | 0 | 8 | 7 | 0 |
| 6 | 3 | 3 | 69 | 14 | 21 | 1 | 1 | 0 | 103 | 111 | 35 | 8 | 5 | 24 | 5 | 12 | 1 | 39 | 10 | 3 | 0 | 3 | 11 | 0 |
| 2 | 3 | 1 | 32 | 8 | 23 | 0 | 0 | 0 | 52 | 144 | 26 | 10 | 5 | 9 | 8 | 5 | 2 | 13 | 2 | 5 | 0 | 3 | 7 | 0 |
| 11 | 2 | 4 | 17 | 5 | 48 | 2 | 1 | 1 | 17 | 147 | 23 | 35 | 6 | 6 | 9 | 12 | 4 | 9 | 5 | 5 | 1 | 1 | 12 | 0 |
| 16 | 5 | 12 | 11 | 0 | 11 | 8 | 0 | 0 | 37 | 140 | 27 | 8 | 2 | 7 | 11 | 15 | 0 | 17 | 13 | 3 | 0 | 4 | 10 | 0 |
| 7 | 3 | 6 | 6 | 0 | 5 | 3 | 4 | 0 | 33 | 59 | 52 | 5 | 2 | 19 | 8 | 6 | 0 | 3 | 13 | 4 | 0 | 8 | 3 | 0 |
| 2 | 0 | 0 | 2 | 0 | 11 | 0 | 0 | 2 | 11 | 17 | 113 | 16 | 4 | 3 | 4 | 2 | 0 | 2 | 1 | 2 | 2 | 2 | 9 | 0 |
| 2 | 1 | 1 | 9 | 2 | 3 | 0 | 0 | 0 | 326 | 68 | 15 | 3 | 2 | 17 | 1 | 3 | 0 | 32 | 9 | 0 | 0 | 0 | 2 | 0 |
| 10 | 5 | 5 | 6 | 0 | 9 | 3 | 1 | 0 | 9 | 225 | 5 | 0 | 0 | 0 | 1 | 4 | 0 | 6 | 13 | 0 | 0 | 1 | 7 | 0 |
| 3 | 1 | 2 | 5 | 0 | 19 | 0 | 1 | 0 | 16 | 30 | 224 | 85 | 7 | 15 | 6 | 3 | 6 | 1 | 0 | 5 | 1 | 3 | 9 | 0 |
| 0 | 0 | 0 | 2 | 0 | 16 | 0 | 0 | 0 | 1 | 22 | 78 | 187 | 7 | 2 | 1 | 0 | 1 | 0 | 0 | 2 | 2 | 1 | 4 | 0 |
| 1 | 0 | 1 | 10 | 0 | 16 | 0 | 0 | 0 | 14 | 29 | 121 | 123 | 17 | 1 | 8 | 1 | 2 | 1 | 1 | 1 | 1 | 0 | 13 | 0 |
| 19 | 1 | 8 | 13 | 3 | 6 | 2 | 0 | 1 | 107 | 74 | 62 | 5 | 2 | 37 | 11 | 13 | 1 | 19 | 22 | 7 | 0 | 10 | 7 | 0 |
| 14 | 6 | 22 | 21 | 4 | 58 | 3 | 1 | 4 | 32 | 141 | 107 | 48 | 13 | 16 | 43 | 10 | 11 | 7 | 13 | 14 | 2 | 12 | 16 | 0 |
| 4 | 3 | 5 | 9 | 1 | 2 | 0 | 2 | 0 | 84 | 40 | 27 | 1 | 2 | 20 | 3 | 15 | 1 | 11 | 13 | 0 | 0 | 2 | 4 | 0 |
| 3 | 2 | 1 | 3 | 0 | 21 | 0 | 2 | 1 | 14 | 20 | 127 | 55 | 13 | 5 | 7 | 4 | 17 | 2 | 3 | 3 | 1 | 1 | 14 | 0 |
| 0 | 3 | 2 | 5 | 0 | 4 | 0 | 0 | 0 | 61 | 38 | 11 | 1 | 0 | 3 | 1 | 10 | 0 | 28 | 2 | 1 | 0 | 0 | 3 | 0 |
| 13 | 1 | 5 | 8 | 1 | 3 | 1 | 0 | 0 | 62 | 62 | 9 | 1 | 0 | 11 | 1 | 12 | 0 | 12 | 23 | 0 | 0 | 3 | 0 | 0 |
| 7 | 2 | 10 | 13 | 0 | 31 | 1 | 0 | 1 | 24 | 92 | 123 | 48 | 4 | 12 | 11 | 6 | 9 | 3 | 7 | 13 | 3 | 9 | 14 | 0 |
| 5 | 1 | 6 | 15 | 3 | 40 | 3 | 1 | 5 | 38 | 100 | 112 | 56 | 4 | 18 | 29 | 7 | 11 | 4 | 6 | 10 | 7 | 8 | 23 | 0 |
| 29 | 7 | 9 | 10 | 1 | 17 | 9 | 0 | 0 | 66 | 125 | 77 | 11 | 3 | 58 | 14 | 17 | 1 | 16 | 16 | 14 | 1 | 21 | 12 | 0 |
| 3 | 2 | 1 | 8 | 0 | 25 | 4 | 2 | 1 | 19 | 74 | 142 | 37 | 8 | 12 | 5 | 13 | 6 | 4 | 11 | 4 | 1 | 2 | 35 | 0 |
| 0 | 6 | 1 | 9 | 0 | 30 | 1 | 2 | 0 | 26 | 40 | 119 | 71 | 7 | 5 | 6 | 5 | 4 | 4 | 2 | 9 | 2 | 0 | 14 | 0 |

## Stratégie SVM

Taux de succès : 16.5%

Matrice de confusion :

| | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 7 | 1 | 0 | 14 | 1 | 0 | 0 | 4 | 21 | 1 | 0 | 1 | 5 | 10 | 0 | 0 | 0 | 0 | 3 | 3 | 38 | 3 | 0 |
| 0 | 1 | 17 | 8 | 1 | 10 | 0 | 0 | 0 | 23 | 34 | 8 | 1 | 0 | 10 | 33 | 1 | 2 | 0 | 1 | 3 | 19 | 63 | 9 | 5 |
| 0 | 2 | 31 | 5 | 1 | 13 | 1 | 0 | 0 | 23 | 29 | 7 | 5 | 3 | 28 | 42 | 0 | 1 | 1 | 0 | 10 | 21 | 123 | 5 | 1 |
| 0 | 0 | 1 | 228 | 16 | 30 | 1 | 1 | 0 | 74 | 5 | 9 | 5 | 6 | 21 | 38 | 1 | 1 | 3 | 0 | 4 | 14 | 22 | 6 | 2 |
| 0 | 0 | 3 | 92 | 27 | 69 | 0 | 0 | 0 | 62 | 14 | 3 | 4 | 3 | 15 | 32 | 1 | 2 | 2 | 0 | 4 | 10 | 10 | 5 | 2 |
| 0 | 0 | 4 | 28 | 20 | 79 | 0 | 0 | 0 | 26 | 30 | 16 | 23 | 10 | 8 | 48 | 1 | 8 | 1 | 0 | 11 | 26 | 35 | 8 | 1 |
| 0 | 0 | 11 | 36 | 8 | 40 | 1 | 0 | 0 | 31 | 38 | 12 | 7 | 6 | 20 | 43 | 2 | 8 | 1 | 1 | 13 | 13 | 58 | 6 | 2 |
| 0 | 1 | 7 | 9 | 1 | 23 | 1 | 0 | 0 | 12 | 15 | 17 | 2 | 0 | 21 | 30 | 0 | 2 | 0 | 0 | 21 | 16 | 59 | 10 | 2 |
| 0 | 0 | 3 | 8 | 1 | 15 | 0 | 0 | 0 | 15 | 1 | 36 | 10 | 5 | 8 | 27 | 1 | 6 | 1 | 0 | 12 | 20 | 26 | 7 | 3 |
| 0 | 0 | 8 | 58 | 3 | 12 | 1 | 0 | 0 | 304 | 10 | 7 | 4 | 1 | 20 | 18 | 1 | 6 | 3 | 0 | 2 | 3 | 22 | 10 | 3 |
| 0 | 0 | 12 | 8 | 9 | 46 | 0 | 0 | 0 | 13 | 99 | 2 | 1 | 1 | 5 | 37 | 1 | 0 | 0 | 0 | 5 | 13 | 54 | 3 | 1 |
| 0 | 0 | 2 | 19 | 3 | 30 | 0 | 0 | 1 | 31 | 9 | 98 | 47 | 21 | 17 | 29 | 1 | 20 | 0 | 0 | 13 | 31 | 37 | 20 | 13 |
| 0 | 0 | 1 | 4 | 2 | 29 | 0 | 0 | 0 | 11 | 0 | 42 | 127 | 19 | 5 | 18 | 0 | 21 | 1 | 0 | 7 | 11 | 6 | 14 | 8 |
| 0 | 0 | 0 | 14 | 7 | 18 | 0 | 0 | 0 | 23 | 8 | 49 | 87 | 24 | 6 | 29 | 0 | 20 | 0 | 0 | 9 | 17 | 20 | 20 | 10 |
| 0 | 0 | 12 | 83 | 2 | 20 | 0 | 0 | 1 | 50 | 18 | 17 | 2 | 4 | 48 | 33 | 1 | 2 | 0 | 0 | 10 | 19 | 93 | 11 | 4 |
| 0 | 0 | 16 | 48 | 9 | 65 | 0 | 1 | 0 | 34 | 43 | 36 | 24 | 7 | 29 | 110 | 2 | 9 | 4 | 2 | 16 | 42 | 88 | 24 | 9 |
| 0 | 0 | 6 | 29 | 2 | 11 | 1 | 0 | 0 | 48 | 7 | 11 | 2 | 1 | 20 | 28 | 2 | 0 | 4 | 0 | 10 | 11 | 39 | 13 | 4 |
| 0 | 0 | 4 | 14 | 4 | 15 | 1 | 0 | 0 | 26 | 1 | 48 | 40 | 21 | 9 | 32 | 0 | 27 | 1 | 0 | 7 | 16 | 20 | 15 | 18 |
| 0 | 0 | 2 | 41 | 6 | 3 | 0 | 0 | 0 | 59 | 2 | 2 | 0 | 3 | 13 | 19 | 0 | 2 | 4 | 0 | 2 | 5 | 3 | 4 | 3 |
| 0 | 0 | 5 | 23 | 1 | 10 | 0 | 0 | 0 | 42 | 15 | 9 | 0 | 0 | 17 | 33 | 1 | 0 | 0 | 1 | 4 | 8 | 53 | 5 | 1 |
| 0 | 1 | 7 | 15 | 7 | 37 | 0 | 1 | 0 | 23 | 28 | 36 | 26 | 10 | 23 | 65 | 1 | 10 | 2 | 0 | 23 | 25 | 79 | 23 | 1 |
| 0 | 0 | 7 | 30 | 7 | 40 | 0 | 0 | 0 | 39 | 32 | 40 | 26 | 11 | 16 | 99 | 2 | 5 | 3 | 0 | 20 | 37 | 68 | 27 | 3 |
| 0 | 0 | 21 | 31 | 3 | 31 | 0 | 3 | 0 | 42 | 39 | 16 | 3 | 3 | 45 | 65 | 2 | 3 | 2 | 1 | 15 | 25 | 167 | 14 | 3 |
| 0 | 0 | 2 | 14 | 2 | 35 | 0 | 0 | 0 | 21 | 20 | 50 | 21 | 13 | 26 | 75 | 1 | 10 | 1 | 0 | 14 | 30 | 46 | 34 | 4 |
| 0 | 0 | 2 | 27 | 1 | 29 | 0 | 1 | 0 | 29 | 8 | 47 | 33 | 19 | 19 | 41 | 0 | 24 | 3 | 0 | 9 | 21 | 22 | 15 | 13 |

## Stratégie KNN

Taux de succès : 18.75%

Matrice de confusion :

| | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 13 | 8 | 27 | 4 | 1 | 4 | 7 | 2 | 0 | 4 | 9 | 2 | 0 | 2 | 2 | 3 | 3 | 0 | 1 | 1 | 2 | 4 | 7 | 5 | 1 |
| 10 | 31 | 32 | 4 | 3 | 7 | 16 | 5 | 3 | 5 | 16 | 7 | 0 | 0 | 11 | 16 | 8 | 2 | 1 | 5 | 11 | 10 | 25 | 12 | 9 |
| 10 | 13 | 131 | 6 | 6 | 9 | 12 | 10 | 3 | 5 | 13 | 5 | 0 | 2 | 14 | 19 | 9 | 1 | 1 | 6 | 13 | 7 | 46 | 6 | 5 |
| 6 | 5 | 22 | 153 | 33 | 15 | 14 | 11 | 2 | 73 | 10 | 9 | 5 | 5 | 32 | 17 | 8 | 1 | 10 | 6 | 6 | 18 | 20 | 5 | 2 |
| 7 | 13 | 16 | 50 | 44 | 20 | 19 | 3 | 2 | 44 | 18 | 9 | 2 | 3 | 10 | 25 | 8 | 2 | 7 | 5 | 10 | 14 | 19 | 5 | 5 |
| 13 | 10 | 27 | 33 | 25 | 25 | 11 | 7 | 5 | 11 | 18 | 17 | 11 | 13 | 12 | 25 | 3 | 7 | 4 | 3 | 15 | 36 | 25 | 21 | 6 |
| 11 | 20 | 43 | 24 | 6 | 21 | 34 | 14 | 5 | 25 | 22 | 6 | 0 | 1 | 26 | 13 | 13 | 3 | 1 | 4 | 5 | 13 | 30 | 12 | 5 |
| 4 | 4 | 26 | 10 | 3 | 4 | 9 | 17 | 6 | 9 | 7 | 16 | 2 | 6 | 31 | 18 | 7 | 0 | 1 | 9 | 15 | 13 | 24 | 7 | 1 |
| 1 | 2 | 10 | 4 | 1 | 5 | 2 | 3 | 26 | 1 | 4 | 53 | 12 | 15 | 2 | 8 | 2 | 5 | 1 | 0 | 13 | 13 | 6 | 10 | 6 |
| 6 | 6 | 12 | 45 | 14 | 7 | 5 | 3 | 2 | 295 | 1 | 12 | 0 | 5 | 32 | 10 | 7 | 0 | 6 | 7 | 2 | 6 | 10 | 2 | 1 |
| 22 | 14 | 30 | 8 | 8 | 15 | 22 | 7 | 1 | 6 | 97 | 1 | 0 | 1 | 8 | 8 | 9 | 1 | 0 | 6 | 4 | 13 | 22 | 6 | 1 |
| 1 | 4 | 14 | 5 | 8 | 13 | 3 | 6 | 33 | 7 | 2 | 121 | 38 | 24 | 12 | 19 | 3 | 14 | 0 | 2 | 30 | 23 | 13 | 29 | 18 |
| 1 | 1 | 2 | 1 | 3 | 7 | 1 | 4 | 10 | 0 | 2 | 70 | 126 | 37 | 3 | 10 | 0 | 10 | 0 | 1 | 9 | 14 | 2 | 4 | 8 |
| 2 | 4 | 4 | 8 | 3 | 9 | 2 | 9 | 11 | 5 | 3 | 82 | 63 | 60 | 5 | 8 | 2 | 19 | 1 | 1 | 15 | 15 | 1 | 21 | 8 |
| 7 | 9 | 41 | 36 | 3 | 5 | 6 | 22 | 7 | 44 | 17 | 20 | 2 | 3 | 71 | 22 | 18 | 3 | 1 | 4 | 15 | 11 | 49 | 9 | 5 |
| 9 | 18 | 54 | 34 | 20 | 23 | 23 | 14 | 6 | 17 | 12 | 34 | 15 | 29 | 27 | 77 | 10 | 9 | 2 | 6 | 38 | 65 | 42 | 23 | 11 |
| 3 | 7 | 24 | 7 | 5 | 7 | 4 | 10 | 1 | 20 | 3 | 2 | 1 | 3 | 25 | 9 | 37 | 4 | 6 | 8 | 9 | 9 | 34 | 8 | 3 |
| 1 | 1 | 6 | 9 | 3 | 8 | 2 | 7 | 14 | 5 | 1 | 39 | 27 | 38 | 4 | 17 | 3 | 35 | 2 | 3 | 15 | 25 | 10 | 27 | 17 |
| 1 | 2 | 13 | 27 | 4 | 3 | 3 | 4 | 1 | 48 | 5 | 6 | 1 | 2 | 8 | 6 | 5 | 1 | 14 | 0 | 2 | 7 | 4 | 5 | 1 |
| 6 | 10 | 23 | 10 | 7 | 1 | 9 | 9 | 1 | 35 | 10 | 1 | 1 | 0 | 22 | 9 | 10 | 1 | 3 | 14 | 2 | 5 | 31 | 3 | 5 |
| 7 | 15 | 39 | 17 | 10 | 17 | 10 | 10 | 14 | 11 | 7 | 37 | 19 | 18 | 23 | 32 | 4 | 10 | 1 | 1 | 39 | 41 | 25 | 24 | 12 |
| 4 | 14 | 35 | 17 | 7 | 26 | 15 | 8 | 6 | 15 | 14 | 31 | 18 | 28 | 30 | 55 | 6 | 9 | 1 | 2 | 35 | 55 | 32 | 32 | 17 |
| 9 | 16 | 75 | 15 | 8 | 12 | 15 | 19 | 8 | 21 | 15 | 18 | 3 | 3 | 56 | 43 | 14 | 5 | 1 | 9 | 34 | 23 | 91 | 17 | 4 |
| 10 | 10 | 28 | 11 | 6 | 15 | 7 | 9 | 22 | 3 | 8 | 53 | 13 | 26 | 10 | 23 | 3 | 11 | 2 | 3 | 32 | 28 | 20 | 50 | 16 |
| 1 | 9 | 13 | 15 | 7 | 12 | 7 | 8 | 8 | 12 | 3 | 41 | 22 | 37 | 12 | 27 | 4 | 16 | 2 | 1 | 16 | 22 | 17 | 22 | 29 |

### Stratégie combiné

Taux de succès : 17%

Matrice de confusion :

| | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 7 | 1 | 0 | 13 | 1 | 0 | 0 | 4 | 21 | 1 | 0 | 1 | 5 | 10 | 0 | 0 | 0 | 0 | 3 | 3 | 38 | 3 | 0 |
| 0 | 1 | 17 | 8 | 1 | 10 | 0 | 0 | 0 | 23 | 35 | 8 | 1 | 0 | 10 | 33 | 1 | 2 | 0 | 1 | 3 | 19 | 62 | 9 | 5 |
| 0 | 2 | 31 | 5 | 1 | 13 | 1 | 0 | 0 | 23 | 29 | 7 | 5 | 3 | 28 | 42 | 0 | 1 | 1 | 0 | 10 | 21 | 123 | 5 | 1 |
| 0 | 0 | 1 | 226 | 16 | 29 | 1 | 1 | 0 | 77 | 6 | 9 | 5 | 6 | 20 | 38 | 1 | 1 | 3 | 0 | 4 | 14 | 22 | 6 | 2 |
| 0 | 0 | 3 | 90 | 27 | 66 | 0 | 0 | 0 | 64 | 17 | 3 | 4 | 3 | 15 | 32 | 1 | 2 | 2 | 0 | 4 | 10 | 10 | 5 | 2 |
| 0 | 0 | 4 | 26 | 20 | 78 | 0 | 0 | 0 | 27 | 32 | 16 | 24 | 10 | 8 | 48 | 1 | 7 | 1 | 0 | 11 | 26 | 35 | 8 | 1 |
| 0 | 0 | 11 | 35 | 7 | 40 | 1 | 0 | 0 | 32 | 39 | 12 | 7 | 6 | 20 | 43 | 2 | 8 | 1 | 1 | 13 | 13 | 58 | 6 | 2 |
| 0 | 1 | 6 | 9 | 1 | 23 | 1 | 0 | 0 | 13 | 15 | 17 | 2 | 0 | 21 | 30 | 0 | 2 | 0 | 0 | 21 | 16 | 59 | 10 | 2 |
| 0 | 0 | 3 | 8 | 1 | 15 | 0 | 0 | 0 | 14 | 1 | 39 | 11 | 5 | 8 | 27 | 1 | 5 | 1 | 0 | 11 | 20 | 26 | 6 | 3 |
| 0 | 0 | 8 | 38 | 3 | 12 | 1 | 0 | 0 | 329 | 10 | 7 | 4 | 1 | 19 | 18 | 1 | 6 | 3 | 0 | 2 | 3 | 20 | 9 | 2 |
| 0 | 0 | 12 | 9 | 7 | 39 | 0 | 0 | 0 | 13 | 113 | 2 | 1 | 1 | 5 | 33 | 1 | 0 | 0 | 0 | 5 | 13 | 52 | 3 | 1 |
| 0 | 0 | 2 | 17 | 3 | 30 | 0 | 0 | 1 | 31 | 9 | 104 | 53 | 18 | 17 | 28 | 1 | 20 | 0 | 0 | 11 | 31 | 35 | 18 | 13 |
| 0 | 0 | 1 | 3 | 2 | 29 | 0 | 0 | 0 | 10 | 0 | 44 | 139 | 15 | 5 | 18 | 0 | 17 | 1 | 0 | 7 | 10 | 6 | 12 | 7 |
| 0 | 0 | 0 | 13 | 6 | 18 | 0 | 0 | 0 | 23 | 8 | 51 | 95 | 21 | 6 | 29 | 0 | 18 | 0 | 0 | 9 | 14 | 20 | 20 | 10 |
| 0 | 0 | 12 | 77 | 2 | 18 | 0 | 0 | 1 | 58 | 21 | 16 | 2 | 4 | 46 | 33 | 1 | 2 | 0 | 0 | 10 | 19 | 93 | 11 | 4 |
| 0 | 0 | 16 | 48 | 9 | 65 | 0 | 1 | 0 | 36 | 43 | 36 | 24 | 7 | 27 | 110 | 2 | 9 | 4 | 2 | 16 | 42 | 88 | 24 | 9 |
| 0 | 0 | 6 | 29 | 2 | 11 | 1 | 0 | 0 | 48 | 7 | 11 | 2 | 1 | 20 | 28 | 2 | 0 | 4 | 0 | 10 | 11 | 39 | 13 | 4 |
| 0 | 0 | 4 | 14 | 4 | 15 | 1 | 0 | 0 | 26 | 1 | 48 | 41 | 21 | 9 | 32 | 0 | 26 | 1 | 0 | 7 | 16 | 20 | 15 | 18 |
| 0 | 0 | 2 | 37 | 6 | 3 | 0 | 0 | 0 | 64 | 2 | 2 | 0 | 3 | 13 | 19 | 0 | 2 | 4 | 0 | 2 | 5 | 3 | 4 | 2 |
| 0 | 0 | 5 | 21 | 1 | 10 | 0 | 0 | 0 | 46 | 15 | 8 | 0 | 0 | 17 | 33 | 1 | 0 | 0 | 1 | 4 | 8 | 52 | 5 | 1 |
| 0 | 1 | 7 | 15 | 7 | 36 | 0 | 1 | 0 | 23 | 29 | 36 | 26 | 10 | 23 | 65 | 1 | 10 | 2 | 0 | 23 | 25 | 79 | 23 | 1 |
| 0 | 0 | 7 | 28 | 7 | 38 | 0 | 0 | 0 | 40 | 35 | 40 | 28 | 11 | 16 | 97 | 2 | 5 | 3 | 0 | 20 | 37 | 68 | 27 | 3 |
| 0 | 0 | 21 | 31 | 3 | 29 | 0 | 3 | 0 | 42 | 41 | 16 | 3 | 3 | 45 | 65 | 2 | 3 | 2 | 1 | 15 | 25 | 167 | 14 | 3 |
| 0 | 0 | 2 | 14 | 2 | 34 | 0 | 0 | 0 | 21 | 21 | 50 | 21 | 13 | 26 | 75 | 1 | 10 | 1 | 0 | 14 | 30 | 46 | 34 | 4 |
| 0 | 0 | 2 | 26 | 1 | 29 | 0 | 1 | 0 | 30 | 8 | 48 | 35 | 19 | 19 | 39 | 0 | 23 | 3 | 0 | 9 | 21 | 22 | 15 | 13 |

## Validation finale

La validation des résultats a été faite en séparant l'ensemble d'entraînement en 2 groupes de données. Le premier groupe qui contient le 2/3 des données a été utilisé pour l'entraînement et l'autre 1/3 a été utilisé pour faire la validation des résultats.

Cette méthode de validation nous a permis d'obtenir un taux de succès de 20%.  Les stratégies individuellement atteingnaient des taux de succès autour de 17% à 18%. On peut donc conclure que leur combinaison a effectivement amélioré la performance du modèle.

On estime que l'entraînement final avec toutes les données devrait être capable d'obtenir autour de 20.5%-21%. Ce qui est tout à fait respectable lorsque l'on tient compte du fait qu'un choix aléatoire obtient au taux de succès de 4%.

## Analyse

Ce que l'on peut voir lors de l'analyse préliminaire est que la stratégie de Bayes et la stratégie de SVM ont chaqu'un des biais d'apprentissage différents qui font en sorte qu'il favorise le choix de certaines classes de façon disproportionnée. Il fait aussi en sorte que certaines classes sont pratiquement jamais choisie. Par contre la stratégie KNN n'a pas ces biais et obtenient un bon taux de succès individuellement. Chaque stratégie a une matrice de confusion différente ce qui est intéressant lorsque l'on veut combiner des stratégies ensemble. Par contre étant donné que deux des stratégies ont des biais d'apprentissage important, cela fait en sorte que la stratégie globale est moins bonne que la stratégie KNN à elle seule. Pour pallier à ce problème, il aurait fallu choisir des algorithmes qui gèrent mieux les biais d'apprentissage et/ou faire l'apprentissage avec un sous-ensemble de données plus uniforme.