Rendu Final - SY09

Lange Mathilde - Trottet Silvia

17 juin 2023

1 Introduction

Le jeu de données sur lequel nous travaillons porte sur l'évolution des colonies d'abeilles aux Etats-Unis (nombre de colonies, rénovation, ajout, perte...) ainsi que l'identification de facteurs de stress pour leur santé susceptible d'influencer leur disparition. Il est trouvable à cette adresse : Cliquez ici pour ouvrir le Github.

La problématique étudiée est : Existe-t-il des paramètres qui influent sur les pertes de colonies et peut-on prédire ces paramètres, à partir des variations des colonies ?

Nous verrons dans une première partie le traitement réalisé sur les données, puis l'analyse exploratoire réalisée sur celles-ci. Enfin, nous détaillerons les différentes étapes de l'étude des paramètres influançant la variation des colonies grâce aux modèles de classification.

2 Traitement du jeu de données

- Fusion des 2 jeux de données : Les données portant sur les facteurs de stress étant séparées du jeu de données principal, nous avons procédé à une fusion des deux tables pour faciliter notre analyse.
- Suppression de colonne : La différence entre les colonnes "colony_max" et "colony_n" semble floue, les pourcentages de pertes et de rénovation sont calculés à partir de la colonne colony_max. Mais dans les données rassemblant tous les états (state = "United States") "colony_max" est à nulle (NaN), on décide donc de conserver les deux colonnes.
- Ajout de colonne : Il manque au jeu de données le pourcentage de colonies ajoutées, la colonne "colony_added_pct" a donc été ajoutée, (calculée à partir des colonnes "colony_added" et "colony_max").
- Suppression des lignes avec trop de valeurs nulles : pour la période avril-juin 2019, toutes les informations sur les colonies (celles ajoutées, perdues et rénovées) sont vides. Nous avons donc décidé de supprimer les données sur cette période.
- Les données ayant pour valeur "United States" à la colonne state, recensent les valeurs de tous les états, elles seront exploitées différement. Nous avons créé un dataset à part pour celles-ci et nous les avons supprimées du dataset principal.

3 Exploration

3.1 Evolution saisonière

Pour comprendre l'impact des facteurs de stress sur les colonies d'abeilles, nous devons étudier les variations des colonies en fonction des saisons.

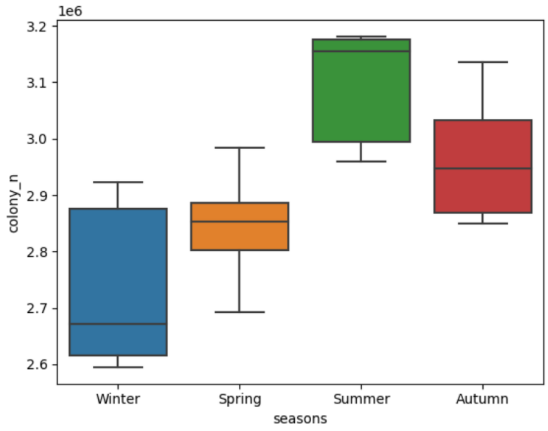

FIGURE 1 – Répartition du nombre de colonies en fonction des saisons

Globalement, le nombre de colonies d'abeilles est plus important en été et plus bas en hiver. En effet, au printemps, la ruche reprend son activité et la reine recommence à pondre (voir lien 1). La quantité maximale de colonies arrive ensuite en été. Les abeilles ont une petite espérance de vie (environ un mois), mais elles survivent plus longtemps en automne/hiver pour que la colonies puissent continuer d'exister à travers le temps.

Au printemps, le nombre de colonies est bas, mais le % de pertes est également le plus faible. Il s'agit d'une période de développement pour la colonie, ce qui ex-

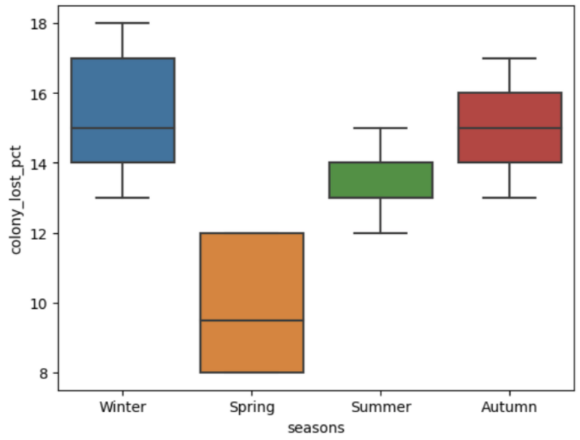

FIGURE 2 – Répartition des pertes de colonies en fonction des saisons

plique qu'elle ait moins de % de pertes. Par la suite, les pertes sont de plus en plus importantes. Dès l'été, la ponte diminue et à cause des fortes chaleurs, il est plus difficile de réguler la colonie. Nous pouvons donc penser que les facteurs de stress seront plus forts.

3.2 Les états

Nous avons souhaité comprendre où étaient situées les colonies d'abeilles et leur répartition aux États-Unis. Notons que la catégorie "Others States" représente cinq états (Alaska, Delaware, Nevada, New Hampshire et Rhode Island).

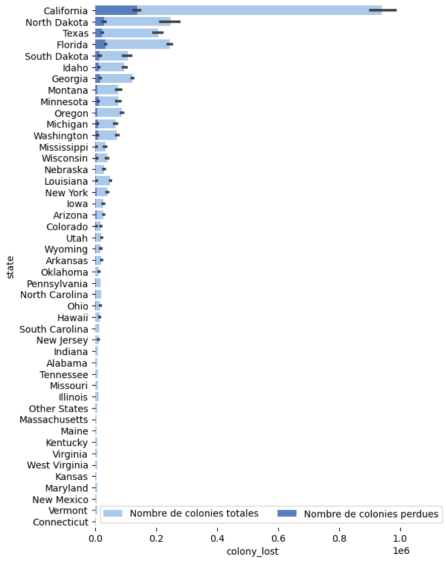

FIGURE 3 – Perte de colonies par rapport au nombre total de colonies par états

Les colonies d'abeilles sont regroupées principalement en Californie. De plus, les pertes ne représentent pas une grande partie du nombre total de colonies. Cela nous permet de mettre en perspective les résultats autour de l'étude des états.

3.3 Les facteurs de stress

Pour étudier les facteurs de stress, nous avons tout d'abord observé ses variations en fonction des saisons. Les facteurs de stress touchent en moyenne 12,2% des colonies en hiver, 16,3% au printemps, 16,8% en été et 16,1% en automne. Ces résultats semblent décalé d'une saison avec notre hypothèse précédente, indiquant des périodes de stress plus fortes dès le printemps et en baisse en hiver.

Nous avons ensuite tenté de représenter les pertes de colonies en fonction du pourcentage de colonies affectées par chaque facteur de stress, mais aucune linéarité ne semblait se dessiner. Nous avons donc décidé de sommer les pourcentages de chaque facteur pour représenter les pertes en fonction de tous les facteurs confondus :

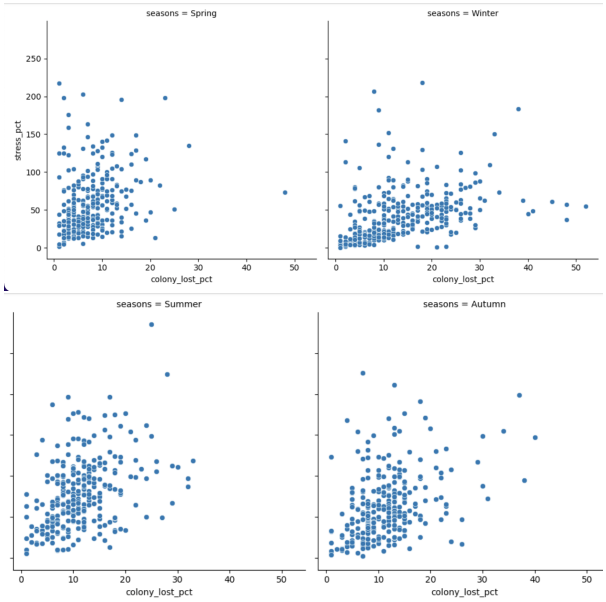

FIGURE 4 – Répartition des pertes de colonies en fonction de l'exposition à un ou plusieurs facteurs de stress, par saisons

D'après le graphique, il semblerait y avoir une corrélation entre les facteurs de stress et la perte de colonies. Cependant, nous n'avons pas pu mettre en évidence un facteur expliquant davantage la perte qu'un autre.

4 Visualisation des données et apprentissage supervisé

4.1 Analyse en Composante Principale

Afin d'améliorer notre compréhension du jeu de données, nous avons appliqué une ACP. Cette technique nous a permis de réduire la dimensionnalité des données et d'explorer les relations entre les variables, facilitant ainsi l'identification de classes ou de tendances significatives.

Nous l'avons réalisé sur toutes les variables représentant des pourcentages (ajout, perte, renovation, exposition aux 6 types de stress). Nous avons exclu les autres données quantitatives (relatives au nombre de colonies par État) afin d'éviter de regrouper les données en fonction de la taille de l'État au lieu de leurs variations (certains États, comme la Californie, possèdent un nombre de colonies bien supérieur à la moyenne dû à leur taille).

Nous avons finalement 9 variables (contre 17 au départ), que nous avons centrées-réduites. Voici un apperçu des résultats de notre ACP :

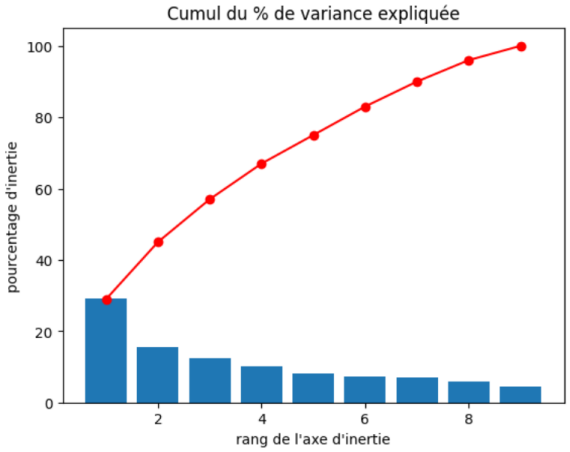

FIGURE 5 – Répartition de l'inertie le long des différents axes

Les 2 premiers axes de notre ACP expliquent 45% de l'inertie totale, et on atteint 90% avec 7 axes. Cela ne fait pas une très grande différence avec nos 9 variables d'origine. Toutefois l'ACP nous permet de visualiser nos données (voir la représentation dans le plan factoriel plus bas).

4.2 Différentes classifications

Les données concernant les variations des colonies (perte, ajout et rénovation) étant quantitatives, il nous a semblé plus pertinent de réaliser l'apprentissage sur des variables qualitatives suceptible d'influencer ces variations. À partir de notre analyse exploratoire, nous avons déterminé trois variables qualitatives qui pouvaient avoir un potentiel impact sur les variations des colonies : les Etats, l'exposition aux divers facteurs de stress et les saisons.

4.2.1 États

Nous avons effectué une première étude sur les États.

Puisque nos données comportent 46 États différents, nous souhaitions créer des classes pour les regrouper et avoir de meilleurs résultats de classification. Pour cela, nous avons appliqué la méthode des K-means. Après l'analyse exploratoire, nous savons que la taille de la Californie est beaucoup plus grande que les autres, nous avons donc enlevé les données relatives aux nombres de colonies. Nous avons cependant conservé la variable des saisons en utilisant la méthode de l'"encodage onehot".

Pour déterminer le nombre de classes, nous avons utilisé la méthode du coude. Ici, la variation de la courbe montre qu'il faudrait utiliser 5 classes.

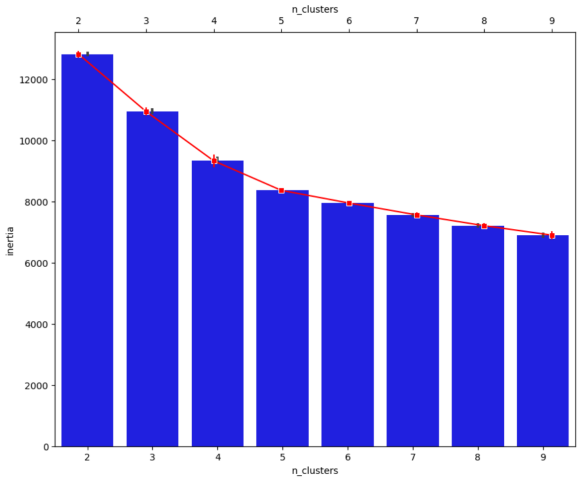

FIGURE 6 – Inertie intra-classes en fonction du nombre de clusters - méthode du coude

Après avoir appliqué la méthode des K-means, nous avons observé la répartition des états dans le 5 classes. Le dendrogramme permet de visualiser le regroupement par CAH.

Malheureusement, ces dernières ne regroupaient pas bien les États. Chaque État était répartis plus ou moins équitablement entre toutes les classes. Il n'aurait donc pas été possible d'associer un État à une seule classe ce qui rend donc l'interprétation de nos résultats de classification sans intérêt.

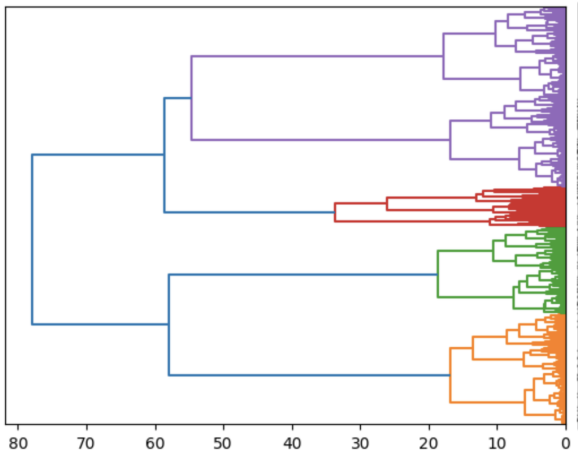

FIGURE 7 – CAH sur les classes du K-means

Les variations et les stress d'une colonie ne permettent donc pas de déterminer son appartenance à un état ou à un groupe d'états. Ceci indique une homogénéité des profils des États. Les problèmes de pertes des colonies sont donc répartis similairement sur tous les États-Unis.

4.2.2 Facteurs de stress

Lors de l'analyse exploratoire, nous avions émis l'hypothèse d'une corrélation entre l'exposition à un ou plusieurs facteurs et la perte de colonies.

Cette hypothèse vient se confirmer avec le cercle des corrélations :

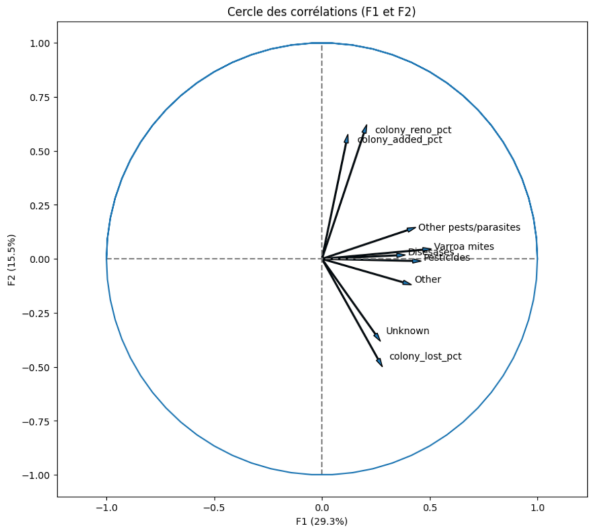

FIGURE 8 – Cercle des corrélations pour l'ACP selon les 2 premiers axes

D'après le graphique, on observe un lien plus ou moins fort entre la perte des colonies ("colony_lost_pct") et les différents facteurs de stress ("Varroa_mites", "Pesticides"...) sur le premier axe factoriel.

Plus particulièrement, on observe une corrélation bien plus prononcée avec le facteur "Unkown" que les autres. Cette observation suggère l'existence d'un impact différent sur la perte selon le type de stress. Certains types de stress ne présenteraient donc pas la même "toxicité" auprès des colonies (particulièrement le facteur "Unkown").

Nous avons également utiliser les cinq types de stress comme labels pour faire de la classification. Les résultats n'ont pas été concluants. Ainsi, on ne peut pas discriminer les types de stress des uns des autres. Ceci s'explique surement par une variance très similaire pour certains stress, notamment "Varroa mites", "Diseases" et "Pesticides" observé sur le cercle des corrélations.

4.2.3 Saisons

Finalement, nous avons classifié les colonies selon les saisons pour confirmer l'hypothèse de leur lien avec les variations et les différents types de stress.

Visualisation

Les saisons représentant déjà des classes, la clusterisation n'a pas été nécessaire. Nous avons pu, à partir de notre ACP, visualiser directement la répartition des saisons sur le plan factoriel.

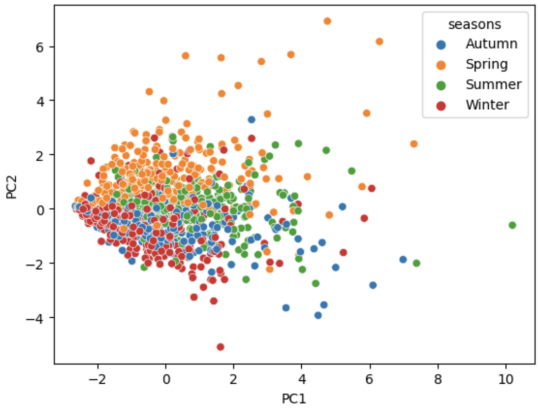

FIGURE 9 – Visualisation des saisons sur le plan factoriel

Analyse de la performance des différents classifieurs des saisons

Nous avons comparé les résultats de six modèles différents :

— LinearDiscriminantAnalysis

— QuadraticDiscriminantAnalysis
— Naive Bayes
— Decision Tree
— Random Forest
— K plus proches voisins

Pour chaque classifieur, nous avons utilisé une validation croisée à 10 itérations pour évaluer les classifieurs obtenus. Concernant la méthode des K plus proches voisins, nous avons utilisé un algorithme de recherche exhaustive afin de déterminer les meilleurs paramètres, avec 80% de nos données réservées à l'entraînement et 20% au test. Ainsi, le K optimal (nombre de voisins) est 10.

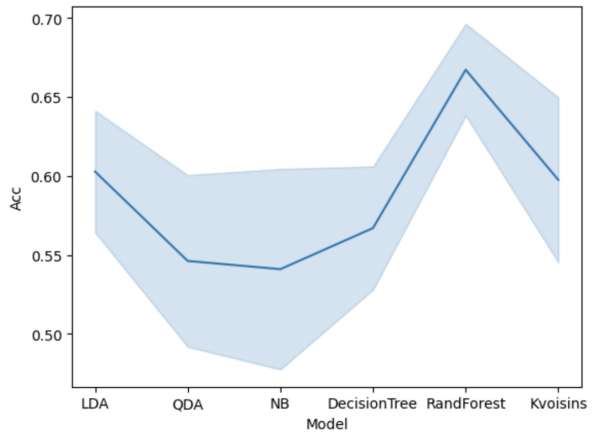

FIGURE 10 – Précision des différents modèles de classifieurs

Les résultats ne sont pas parfaits, ce qui permet de dire que les données de variations et de stress n'expliquent pas entièrement les saisons. En revanche, nous pouvons tout de même noter que le meilleur résultat est obtenu par la forêt aléatoire avec une précision de 67% environ.

Pour représenter les frontières de décision, nous avons utilisé les données issues de l'ACP. En se rappelant que les deux premiers axes n'expliquent que 45% des données, nous pouvons tout de même visualiser le résulat du LDA. Sur ces données, les scores d'accuracy du modèle par cross-validation sont d'environ 52%.

Lorsque l'on ajoute des données de taille de colonies, c'est-à-dire le nombre de colonies, le nombre de colonies perdues, ajoutées, rénovées, nous trouvons des résultats légèrements meilleurs.

— LDA : 61%
— QDA : 59%
— Naive Bayes : 50%
— Decision Tree : 60%

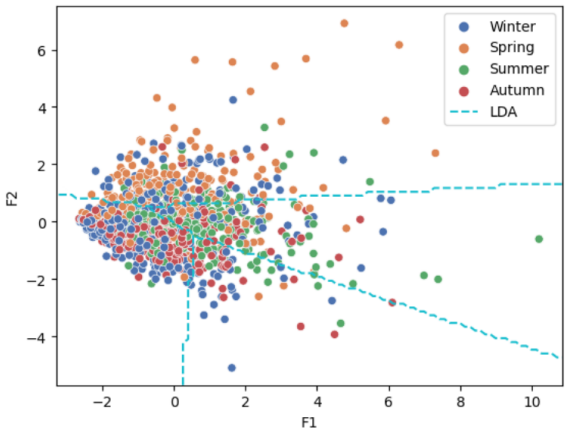

FIGURE 11 – Visualisation des frontières de décision du LDA sur les données de l'ACP

— Random Forest : 72%
— K plus proches voisins : 62%

Ainsi, à partir de données quantitatives sur une colonie, nous pourrions prédire la saison à laquelle le recueillement des données a été faite.

5 Conclusion

L'analyse des données a permis de montrer le fonctionnement des colonies d'abeilles en fonction des saisons. Il existe bien des tendances dans les pertes qu'elles peuvent subir en fonction de différents paramètres. L'apprentissage non-supervisé a été une autre manière de montrer le lien entre saison et variations et stress. L'utilisation de différents modèles permet de trouver le plus efficient pour nos données.

Nous avons également pu observé la similarité des variations des colonies entre les Etats. S'il avait fallu étudier plus précisemment les spécificités des Etats, des données supplémentaires auraient été nécessaires.

Une autre piste intéressante à explorer est le lien entre les facteurs de stress et les pertes des colonies. Nous avons montré qu'il existait une corrélation entre les deux. Pour prédire le % de perte, nous avons fait des essais de regression linéaire, en utilisant les stress, l'état et la saison, mais ils n'ont pas été très concluants. Pour cette raison, nous ne les avons pas inclus dans ce rapport. Il serait intéressant d'utiliser d'autres méthodes pour comprendre plus en détails la toxicité de chaque stress sur les colonies.

6 Bibliographie

1. Fonctionnement d'une colonie d'abeille : http://www.mes-abeilles.com/les-abeilles/la-colonie-dabeilles/
2. Le cycle de vie des abeilles en fonction des saisons : https://mesabeilles.fr/les-abeilles/le-cycle-de-vie-des-abeilles-au-fil-des-saisons
3. Sur les Varroa Mites : https://beeaware.org.au/archive-pest/varroa-mites
4. Cercle de corrélation : https://colab.research.google.com/github/OpenClassrooms-Student-Center/4525281-realisez-une-analyse-exploratoire-de-donnees/blob/main/notebooks/P2C6_TP_cours_effectuees.ipynb
5. StandardScaler : https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html