ISIS INGÉNIEURS CASTRES

Institut National Universitaire Champollion

INSA PARTENAIRE

Projet Tutoré :

Étude de la relation entre température, corporelle et taux de mortalité chez les chiots par l'analyse des données de l'ENVT via NeoCare

envt école nationale vétérinaire toulouse

GUENNOC Denoëla - AOUBIZA Imane - MAINTENANT Victor

PTUT 2021/2022 FIE4_Promo 2023

Remerciement

Nous tenions tout d'abord à remercier École Nationale Vétérinaire de Toulouse de nous avoir proposé ce projet tutoré.

Nous souhaitions remercier plus particulièrement le Dr. Amélie MUGNIER pour sa disponibilité, le temps qu'elle a nous a accordé tout au long du projet, ainsi que les conseils qu'elle a pu nous apporter pour la compréhension du sujet.

Nous remercions bien évidemment Mme MEGDICHE pour son accompagnement et son aide qu'elle nous a apporté pour nous guider au mieux dans l'avancement de notre projet.

Pour finir, nous remercions nos camarades Clémence DELESTRE, Lisa ESTEBE et Cécile ROEHRIG pour le travail qu'ils ont fourni avec Denoëla GUENNOC au premier semestre.

Tables des matières

Glossaire

Facture Selection : Processus utilisé en apprentissage automatique et en traitement de données. Il consiste à trouver un sous-ensemble de variables pertinentes

LSTM : Architecture de réseau neuronal récurrent artificiel utilisée dans le domaine du Deep Learning.

RNN : Réseau de neurones interconnectés interagissant non linéairement et pour lequel il existe au moins un cycle dans la structure.

Deep Learning : Ensemble de méthodes d'apprentissage automatique tentant de modéliser avec un haut niveau d'abstraction des données

Python : Langage de programmation interprété

Notebooks : Interface de programmation interactive permettant de combiner des sections en langage naturel et des sections en langage informatique.

Dataset : Ensemble de données où chaque valeur est associée à une variable et à une observation.

ENVT : École Nationale Vétérinaire de Toulouse.

NeoCare : Néonatalogie des Carnivores Reproduction et Élevage, centre dédié à l'élevage, la reproduction et la pédiatrie canine et féline.

1. Présentation du projet

1.1. Présentation de MOA (Maître d'Ouvrage)

ENVT, Ecole Nationale Vétérinaire de Toulouse, est une grande école formant environ un quart des vétérinaires de France. NéoCare, Néonatalogie des Carnivores Reproduction et Elevage, est quant à elle une unité de recherche de l'ENVT créée en 2016. Il s'agit d'un centre dédié à l'élevage, la reproduction et la pédiatrie canine et féline regroupant vétérinaires, étudiants, éleveurs, refuges, propriétaires et chercheurs dans un but clinique, de formation, de recherche et de service. NéoCare mène ainsi des travaux de recherche appliquée visant à améliorer la santé des chiots et des chatons sur les premiers mois de vie via des recherches sur le développement foetal, néonatal et pédiatrique et son impact sur leur vie adulte. Ce centre a pour but la production de données scientifiques robustes, la mise au point d'outils ou de solutions pratiques à l'usage des professionnels et la diffusion de connaissances vers les étudiants vétérinaires et éleveurs. Notre interlocutrice, Mme Amélie Mugnier, est une doctorante et résidente ECVPH (The European College of Veterinary Public Health) travaillant sur le poids de naissance des chiots et des chatons au sein de NéoCare.

1.2. Présentation Projet

En élevage canin, la mortalité avant la fin de la période de sevrage est relativement fréquente. Il est donc essentiel d'investiguer les déterminants de santé chez le chiot pour améliorer sa survie. Pour cela, certains paramètres pourraient faire office de marqueurs précoces utiles à l'anticipation de la survenue de problèmes de santé et donc de prodiguer des soins adaptés aux animaux le plus tôt possible. C'est le cas, par exemple, de la température, qui est un des éléments cruciaux du suivi clinique. En effet, la capacité de digestion du nouveau-né est directement affectée par sa température : en dessous de 35°C, le lait n'est plus digéré. Le biberonnage ne peut donc pas sauver le chiot s'il reste à cette température.
C'est au sein de ces travaux de recherche que s'inscrit notre projet. Nous réaliserons l'analyse de facteurs de risques, une approche expérimentale ayant déjà été réalisée en amont par les chercheurs.
L'objectif de ce projet est d'étudier les liens que pourraient avoir certaines caractéristiques d'un chiot et l'évolution de sa température sur son état de santé afin de pouvoir anticiper et prédire des épisodes de diarrhée, d'hospitalisation et ou de mort, et ce chez des chiots issus d'un même élevage. Pour cela, nous disposons d'un tableau de données, fourni par Mme Mugnier, regroupant les différentes données récoltées par l'équipe et les collaborateurs de NéoCare.
C'est grâce à l'exploitation de ces données lors de diverses méthodes d'analyse que nous essayerons de démontrer une corrélation entre les températures et caractéristiques et les états de santé chez le chiot et d'essayer de pouvoir les prédire.

1.3 Analyse des besoins et rendu

Ce projet a pour objectif d'étudier le lien qui existe entre les diverses caractéristiques des chiots et leurs températures et leur apparition de problèmes de santé (épisode de diarrhée, hospitalisation, mortalité) chez 168 chiots issus d'un même élevage.
Suite à notre rencontre au second semestre avec Mme Mugnier, nous avons décidé d'essayer de trouver tout ce qui pourrait être exploitable pour déterminer et anticiper des changements d'état de santé chez un chiot. Cela pouvait prendre la forme que nous souhaitions : graphique, température de seuil haut et bas à ne pas franchir, etc.

2. Déroulement du projet

2.1 Gestion de projet

En ce qui concerne la gestion de projet, nous avons établi un diagramme de Gantt (cf. Annexe 3) afin de visualiser dans le temps les différentes tâches que nous avions à faire tout au long du projet.

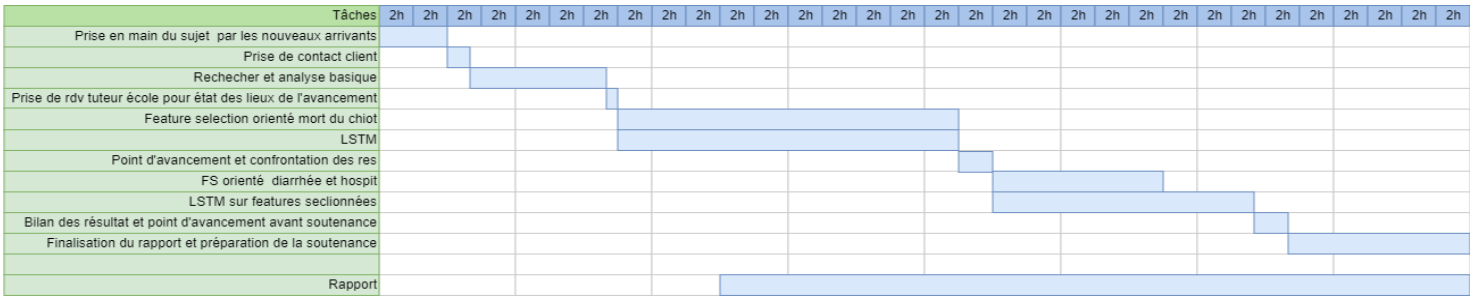


Figure 0. Diagramme de Gantt du projet

2.2 Problèmes rencontrés

2.2.1 Problèmes humains

Parmi les problèmes rencontrés, le principal fut le manque de temps. En effet, le nombre d'heures prévues dans l'emploi du temps ne suffisait pas à mener à bien ce projet. Il a donc fallu trouver du temps en dehors de ces créneaux, ce qui fut compliqué étant donné la charge de travail que nous avions dans les autres matières.

De plus, le changement dans le groupe dû au semestre d'étude à l'étranger entre le premier et deuxième semestre a été compliqué à gérer. L'un des membres n'étant rentré que début février à l'université, nous avons perdu du retard sur la prise de connaissance du projet et son commencement au second semestre.

2.2.2 Problèmes données et analyse

Lors de la récupération du dataset permettant de faire nos analyses, nous avons de suite remarquée que plusieurs valeurs concernant les températures de plusieurs chiots ainsi que certaines de leurs caractéristiques comme les taux d'infériorité de croissance, par exemple, n'était pas renseigné. Il a donc fallu revoir les valeurs et essayer de les remplir en essayant de suivre ce qui existait déjà dans le dataset. Pour les températures nous avons pris la température du jour-1 pour les jours n'ayant pas de valeurs.

Lorsque nous avions fait nos premières analyses succinctes sur les données nous avons dû demander de l'aide auprès de plusieurs professeurs de Deep Learning dans l'école afin d'avoir des pistes pour savoir ce qu'il était possible de faire pour essayer d'obtenir des résultats et répondre au besoin du client. En effet, nous ne nous y connaissions pas, pour la plupart, en Deep Learning et méthode d'analyse du fait que les cours de Big Data et Deep Learning n'ont été abordés qu'en fin de semestre.

Lorsque nous avons commencé à mettre en place des méthodes d'analyse des données et de prédictions nous nous sommes alors rendu compte que le dataset possédait un autre problème conséquent, Il n'y a pas assez de données. Nous n'avons des données que sur 168 chiots et seulement 9 sont morts ce qui n'est vraiment pas assez pour pouvoir essayer de faire des prédictions.

3. Méthode d'analyse

Après avoir récupéré les données finales et triées sur lesquelles les analyses porteront, nous avons appliqué plusieurs techniques d'analyse de données pour essayer de trouver la ou les meilleures pour notre cliente.
La première d'entre elles est une approche pas à pas pour essayer de voir si un type de données ou un pattern de données dans les températures ressortent. Nous avons aussi repris une des approches vues au premier semestre par l'ancien groupe, la corrélation de Pearson, afin de voir si sur ces données de nouvelles corrélations sortent.
Puis sur les conseils de madame MEGDICHE, nous avons mis en place deux nouvelles techniques pour mettre en évidence des colonnes et données qui pourraient répondre à nos attentes. Il s'agit des méthodes de Feature Selectiion et de LSTM.

3.1 Étude des données et Corrélation[1]

Dans un premier temps nous avons voulu voir s'il existe un pattern dans la température chez les chiots qui avait eu un problème de santé au cours de la période pédiatrique, que ce soit pour la diarrhée, l'hospitalisation ou la mort. Dans un premier temps nous avons dû reprendre les données de températures car certaines n'étaient pas renseignées, ce qui ne permettait pas d'avoir des données pertinentes. Pour remplacer les dates manquantes nous avons

[1] Cf. Notebook 'Tableau de Corrélation.ipynb'

décidé d'affecter les températures du jour précédent. Cette méthode nous permettait de suivre une température évolutive du chiot au jour par jour et d'éviter de faire de grands écarts de température, par rapport à une méthode de remplacement par la moyenne. Ce choix nous permettait de pouvoir mettre en avant une période avec des températures qui dépassent un certain pallier afin de voir si les chiots qui avaient eu un problème de santé ressortait.

Une fois le problème de température réglé nous avons affiché les températures de tous les chiots pour voir si celles des chiots qui avait eu un problème de santé ressortent et dépassent de manière unie celles des autres chiots. Cela nous aurait permis de mettre en avant une température maximale et/ou minimale à ne pas avoir pour un chiot.

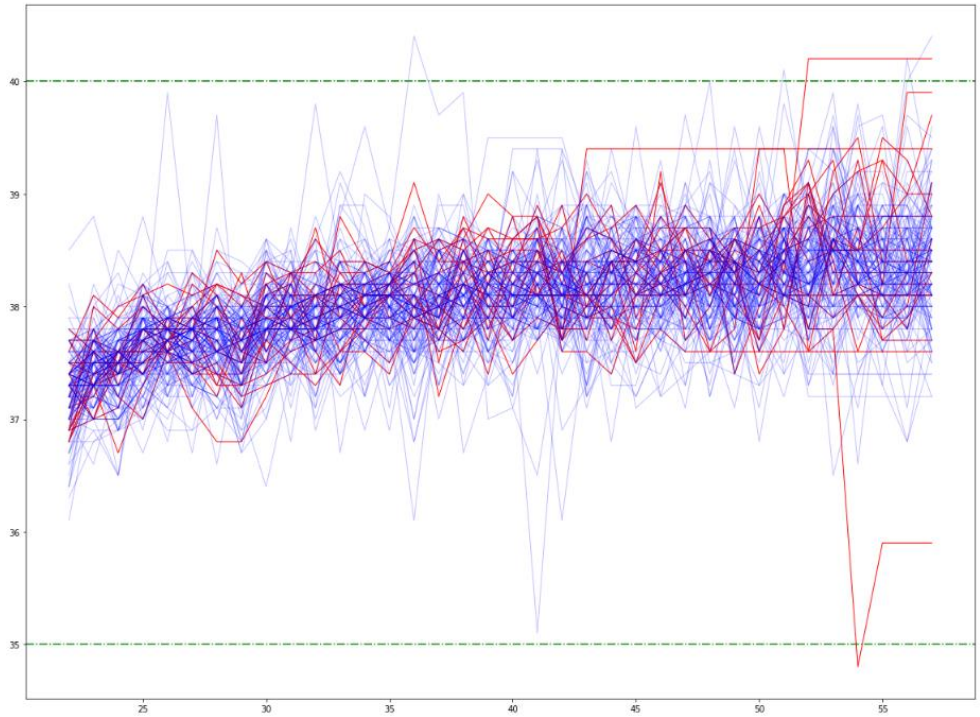


Figure 1. Affichage de toutes les températures pour la mort

Lors de l'affichage des températures de tous les chiots (en rouge ceux mort et en bleu les autres), on remarque qu'il est difficile de pouvoir isoler des patterns dans la température pour pouvoir trouver une piste de solution. En effet, il n'est pas possible de trouver une température max à ne pas dépasser ou à ne pas rester plus de x jours car plusieurs chiots ont eu des températures très hautes et sur une longue période. Cependant on peut poser une hypothèse sur une température basse à ne pas atteindre pour un chiot. En effet on voit qu'un seul des chiots a eu une température inférieure à 35°C et qu'il est mort.

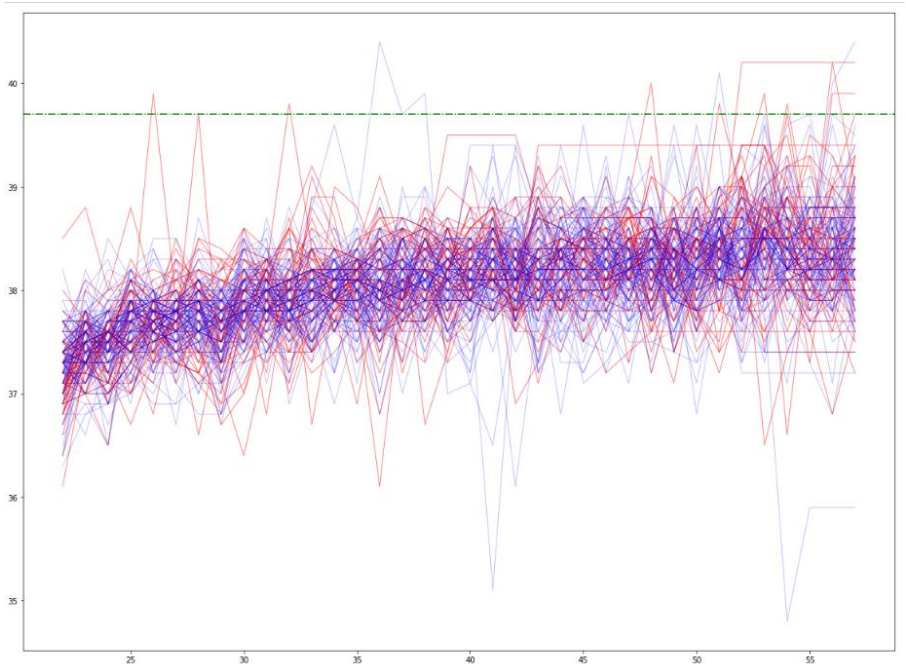


Figure 2. Affichage de toutes les températures pour la diarrhée

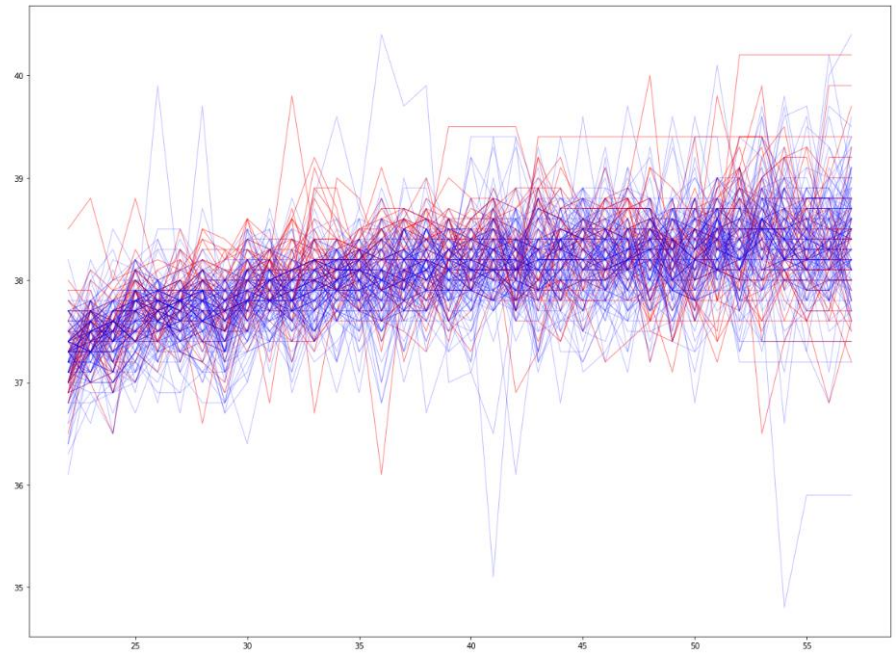


Figure 3. Affichage de toutes les températures pour l'hospitalisation

En ce qui concerne les températures pour la diarrhée et l'hospitalisation, rien ne ressort lors de l'affichage et de l'analyse. Il n'est donc pas possible de faire une première hypothèse quant à ces états et la température de manière certaine.

Une fois une première analyse succincte faite, nous avons décidé d'afficher un tableau de corrélation comme l'avait fait le groupe précédent lors du premier semestre afin de voir si de nouveaux liens de corrélations apparaissent.

Pour pouvoir calculer la corrélation entre les éléments du dataset nous avons supprimé les colonnes qui ne nous paraissaient pas pertinentes comme les numéros du chiot et de sa portée, son sexe et quelques autres. Nous avons aussi changé les valeurs de colonnes qui possédaient des données non renseignées. Nous avons décidé de valeurs qui nous semblaient pertinentes pour ne pas trop détériorer le dataset.

```
frame['IgGJ2'].fillna(0, inplace = True)
frame['TPI'].fillna(0, inplace = True)
frame['sex'].fillna('ND', inplace = True)
frame['InfCroissPed'].fillna(0, inplace = True)
frame['InfCroiss7'].fillna(0, inplace = True)
frame['InfCroiss8'].fillna(0, inplace = True)
frame['Min1inflcr'].fillna(0, inplace = True)
```

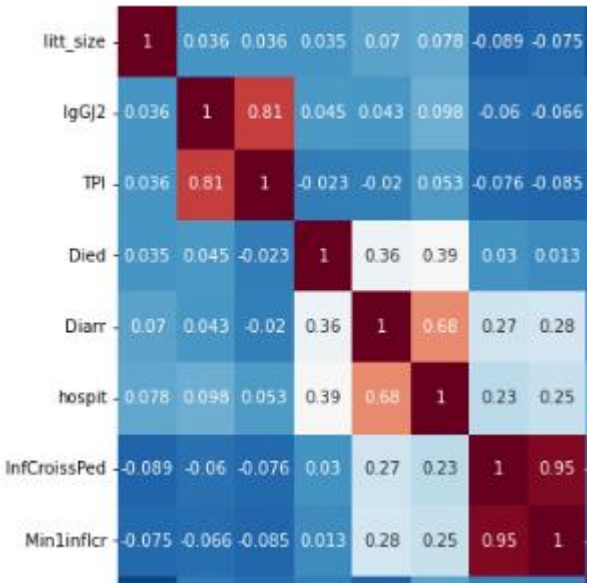
Figure 4. Ligne de commande pour changer les valeurs vides

Une fois les données toute renseignées nous avons lancé la création et l'affichage du tableau de corrélation (annexe 1, tableau complet). Le tableau complet est assez fourni et seules certaines parties semblent intéressantes comme la partie en haut à gauche avec les différents états et attributs du chiot et certaines températures. En ce qui concerne les gros pourcentages de corrélation en bas à droite et sur la diagonale basse du tableau, ils ne sont pas pertinents car ils nous indiquent des corrélations entre certaines températures et le plus souvent du jour précédent au jour suivant.
La partie en haut à gauche du tableau nous donne des pourcentages de corrélation élevés entre plusieurs colonnes. En effet, comme lors du premier semestre, on voit qu'il existe une forte corrélation entre :

- l'hospitalisation et la diarrhée
- l'hospitalisation et la mort
- la diarrhée et la mort

Mais on voit aussi apparaître une corrélation entre :

- la diarrhée et si le chiot a eu une infériorité de croissance
- l'hospitalisation et si le chiot a eu une infériorité de croissance



On voit aussi apparaître des corrélations très fortes entre le 'TPI' et 'IgGJ2' et 'InfCroissPed' et Min1inflcr, mais qui ne nous intéresse pas pour ce que nous recherchons.

Figure 5. Tableau de corrélation des attributs des chiots

En analysant le tableau obtenu, on remarque que la corrélation de certaines températures dans la colonne d'hospitalisation ressort plus que les autres. Ce sont les températures des jours 31, 32 et 38. Certes ces corrélations ne sont pas assez grandes pour pouvoir affirmer que la température de ces jours joue un rôle dans l'hospitalisation d'un chiot. Il faudra voir si ces données ressortent avec les autres modèles d'analyse que nous allons utiliser.

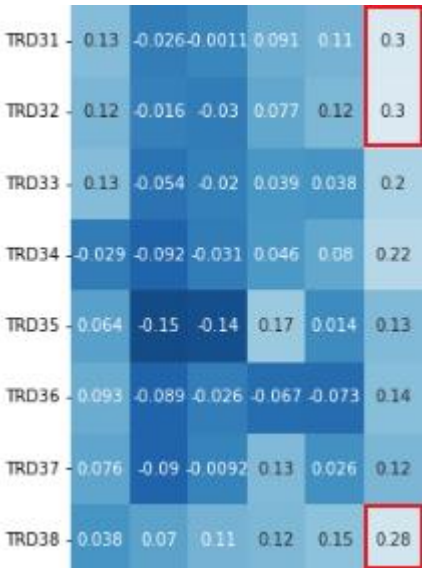


Figure 6. Corrélation température hospitalisation

3.2 Méthode d'analyse Feature Selection

Après avoir réalisé une première analyse des données succinctes et sur la corrélation entre les attributs, nous sommes ensuite partis sur l'utilisation d'une autre méthode d'analyse car nous n'avions pas encore eu de résultats satisfaisants. Nous nous sommes tournés premièrement vers la méthode Selection Feature.
Feature Selection est une méthode permettant de réduire la variable d'entrée de notre modèle en utilisant uniquement les données pertinentes et en éliminant le bruit dans les données. Il s'agit d'un processus consistant à choisir automatiquement des caractéristiques pertinentes pour notre modèle d'apprentissage automatique en fonction du type de problème que vous essayez de résoudre. Pour ce faire, nous incluons ou excluons des caractéristiques importantes sans les modifier. Cela permet de réduire le bruit dans nos données et de réduire la taille de nos données d'entrée.

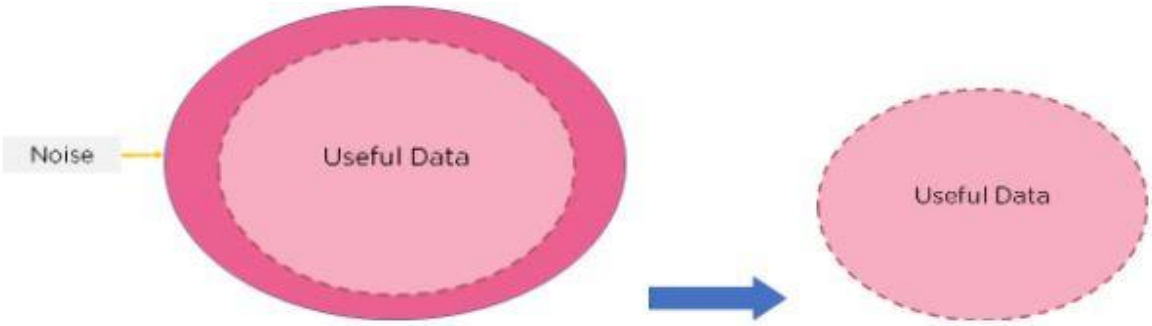


Figure 7. Schématisation Feature Selection [2]

[2]https://www.simplilearn.com/tutorials/machine-learning-tutorial/feature-selection-in-machine-learning#why_feature_selection

Il existe plusieurs types de Feature Selection, non supervisées et supervisées. Et pour la partie supervisée il y a trois autres types de modèle : Filter, Wrapper, Embedded.

Filter methods	Wrapper methods	Embedded methods
Generic set of methods which do not incorporate a **specific machine learning algorithm.**	Evaluates on a **specific machine learning algorithm** to find optimal features.	Embeds (fix) features during **model building process.** Feature selection is done by observing each iteration of model training phase.
Much **faster** compared to Wrapper methods in terms of time complexity	**High computation time** for a dataset with many features	Sits **between Filter methods and Wrapper methods** in terms of time complexity
Less prone to **over-fitting**	High chances of **over-fitting** because it involves training of machine learning models with different combination of features	Generally used to reduce **over-fitting** by **penalizing** the coefficients of a model being too large.
Examples – **Correlation, Chi-Square test, ANOVA, Information gain** etc.	Examples - **Forward Selection, Backward elimination, Stepwise selection** etc.	Examples - **LASSO, Elastic Net, Ridge Regression** etc.

Figure 8. Différence entre les trois type de méthode[3]

Pour notre analyse nous nous sommes tournés vers ces trois modèles que nous avons appliqué à chaque type d'état de santé du chiot pour voir quelles caractéristiques pouvaient ressortir et si certaines venaient à se répéter.

3.2.1 Feature Selection Wrapper[4]

Avec la méthode Wrapper, le processus de sélection des caractéristiques est basé sur un algorithme d'apprentissage automatique spécifique que l'on essaie d'adapter à un ensemble de données donné. Il suit une approche de recherche dite gloutonne en évaluant toutes les combinaisons possibles de caractéristiques par rapport au critère d'évaluation. Le critère d'évaluation est simplement la mesure de performance qui dépend du type de problème. Enfin, il sélectionne la combinaison de caractéristiques qui donne les résultats optimaux pour l'algorithme d'apprentissage automatique spécifié.

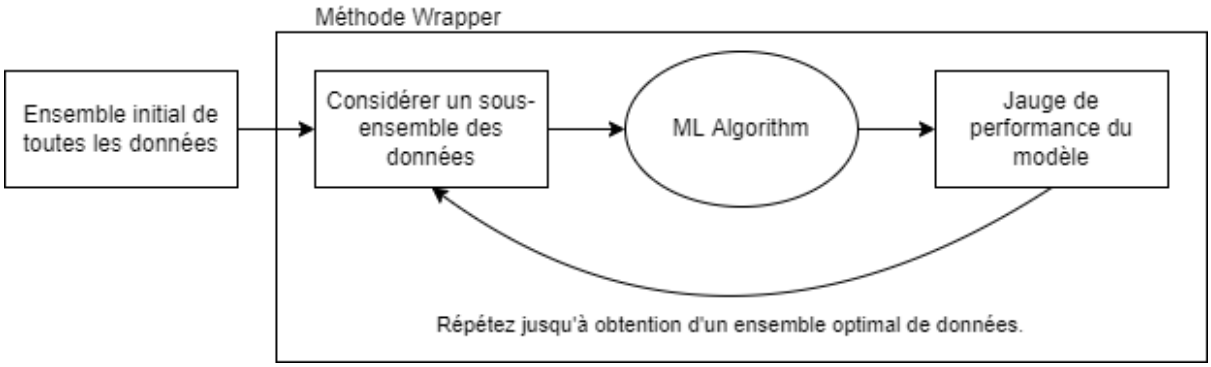


Figure 9. Explication méthode Wrapper

[3]https://www.analyticsvidhya.com/blog/2020/10/a-comprehensive-guide-to-feature-selection-using-wrapper-methods-in-python/

[4] Cf. Notebook 'Feature Selection Wrapper.ipynb'

3.2.1.1 État ‘Died’

Dans un premier temps il a fallu reprendre les données du dataset et remplir les données qui n'étaient pas renseignées par des données qui nous semblaient les plus pertinentes pour ne pas abîmer le dataset. Une fois fait, il a fallu créer le modèle puis le lancer et afficher les 10 colonnes qui ressortent le plus.

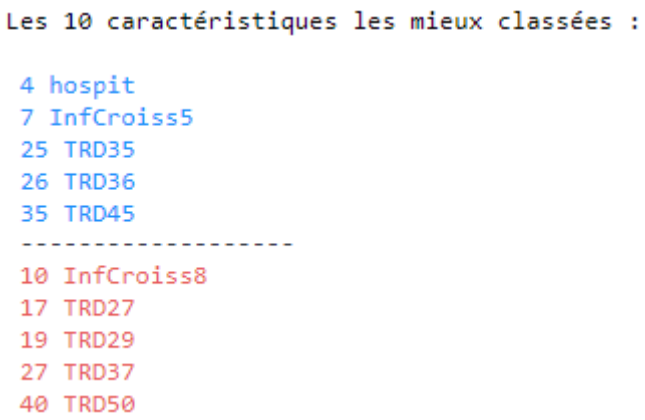


Figure 10. Affichage des caractéristiques pour la mort avec la méthode Wrapper

Le modèle wrapper nous donne les résultats précédents. On retrouve en bleu les cinq premiers caractères les mieux classés sur les 47 utilisés. Et en rouge les cinq suivants. On voit apparaître, en plus de l’hospitalisation que l’on avait déjà retrouvé plus tôt, le caractère d’infériorité de croissance de la semaine 5 et les températures des jours 35, 36 et 45. Puis dans la deuxième partie le caractère d’infériorité de croissance de la 8ème semaine et les températures des jours 27, 29, 37 et 50.
En se penchant sur les autres types de modèle il faudra voir si ces caractéristiques reviennent pour en tirer une hypothèse.

3.2.1.2 État ‘Diarr’

On réitère le procédé mais en définissant la cible sur le caractère de la diarrhée.

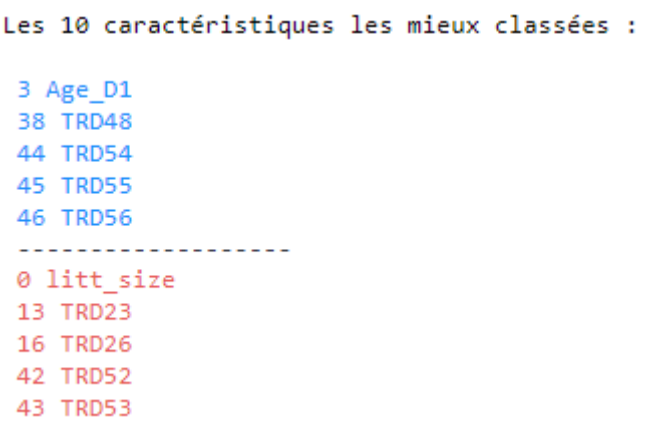


Figure 11. Affichage des caractéristiques pour la diarrhée avec la méthode Wrapper

Avec le caractère diarrhée comme cible on obtient les résultats ci-dessus. On peut voir que le premier des caractères qui ressort est l'âge de la première diarrhée. Cela semble cohérent et normal puisque tout chiot qui a eu de la diarrhée a un âge de première diarrhée renseigné mais cela ne peut pas trop nous aider pour répondre à nos attentes. Puis, vient une série de températures des jours 48, 54, 55 et 56. Dans la deuxième partie des résultats on retrouve des températures encore une fois, des jours 23, 26, 52 et 53, qu'il faudra venir retrancher avec les autres résultats. Et on retrouve aussi la données 'litt_size' qui correspond à la taille de la portée du chiot. Si ce résultat revient il faudra regarder de plus près si la taille d'une portée peut avoir un impact sur la santé d'un chiot.

3.2.1.3 État 'hospit'

On recommence, mais avec comme cible l'hospitalisation des chiots.

```
Les 10 caractéristiques les mieux classées :

 2 Died
 4 Diarr
 5 InfCroissPed
 14 TRD24
 21 TRD31
 -------------------
 6 InfCroiss4
 8 InfCroiss6
 15 TRD25
 34 TRD44
 38 TRD48
```

Figure 12. Affichage des caractéristiques pour l'hospitalisation avec la méthode Wrapper

En prenant la caractéristique d'hospitalisation comme cible, on obtient les résultats précédents. Les états de mort et de diarrhée ressortent en premier comme on a déjà pu le voir auparavant. On retrouve ensuite la caractéristique d'infériorité de croissance sur la période de la 4ème à la 8ème semaine. Puis, on a la température de la semaine 24 et 31. Dans les cinq caractéristiques qui ressortent ensuite, on retrouve les caractéristiques d'infériorité de croissance pour les semaines 4 et 6 et la température des jours 25, 44 et 48.

3.2.2 Feature Selection Filter[5]

Dans cette méthode, les caractéristiques sont éliminées en fonction de leur relation avec la sortie, ou de leur corrélation avec la sortie. Nous utilisons la corrélation pour vérifier si les caractéristiques sont positivement ou négativement corrélées aux étiquettes de sortie et éliminons les caractéristiques en conséquence.

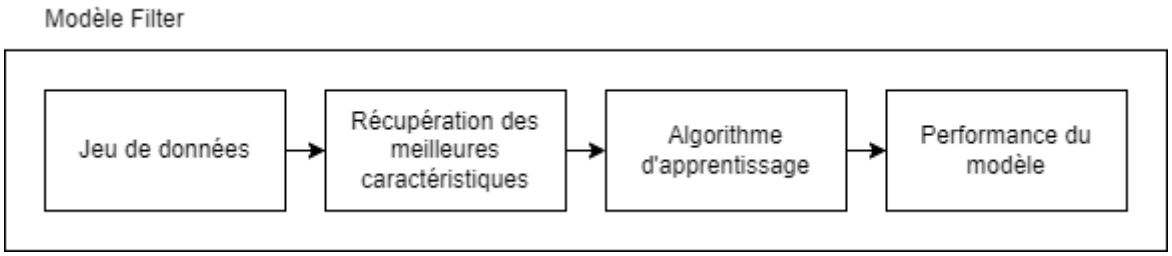


Figure 13. Explication méthode Filter

Pour mettre en place cette méthode il a fallu récupérer les données de notre dataset puis enlever les données qui n'étaient pas du tout pertinentes comme le nom, le sexe, la taille, … Puis nous l'avons appliqué sur tous les états de santé du chiot pour voir si une caractéristique ressortait.

3.2.2.1 État 'Died'

Dans un premier temps nous avons nettoyé et retiré toutes les caractéristiques que nous ne voulions pas et qui n'étaient pas pertinentes pour faire notre sélection.
Les caractéristiques constantes, qui sont le type de caractéristiques qui ne contiennent qu'une seule valeur pour toutes les sorties de l'ensemble de données. Les caractéristiques constantes ne fournissent aucune information qui puisse aider à la classification de l'enregistrement en question. Les caractéristiques quasi constantes, comme leur nom l'indique, sont les caractéristiques qui sont presque constantes. Ces caractéristiques ne sont pas très utiles pour faire des prédictions. Les caractéristiques dupliquées sont les caractéristiques qui ont des valeurs similaires. Puis, les caractéristiques dupliquées, qui n'apportent aucune valeur ajoutée à l'apprentissage de l'algorithme. Elles ajoutent plutôt des frais généraux et un retard inutile au temps d'apprentissage.
Tous ces changements n'ont pas trop abîmé notre dataset puisque qu'une seule colonne a été retirée lors de ces transformations.

Une fois le nettoyage terminé, nous avons créé des caractéristiques d'entrée, nos prédicteurs, et la variable cible. Pour ce test la variable cible est l'état mort du chiot et les caractéristiques d'entrée toutes les autres colonnes de notre dataset.

```
X = data.iloc[:,0:50]
Y = data['Died']
```

Figure 14. Définition de nos variables d'entrée et la cible

[5] Cf. Notebook 'Feature Selection Filter.ipynb'

Une fois la cible et les entrées définies, il ne reste plus qu'à trouver les corrélations entre la variable cible et les prédicteurs indépendants, puis de l'afficher.

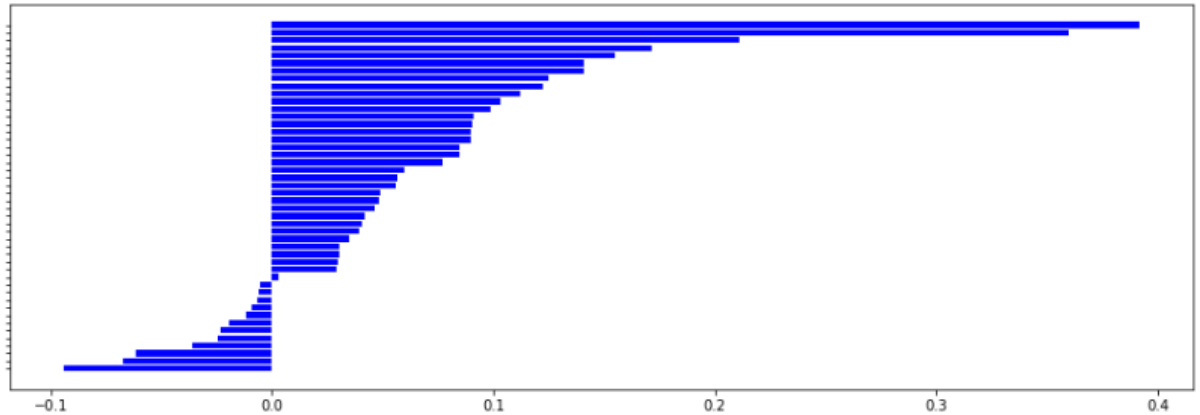


Figure 15. Affichage entier de la méthode Filter sur la mort

Figure 16. Affichage des 6 plus gros coefficients

Les plus gros coefficients de corrélation que nous retrouvons pour la mort avec cette méthode sont les caractéristiques de l'hospitalisation et de la diarrhée mais que nous connaissions déjà. On voit ensuite sortir trois températures, celle du jour 44, 35 et 50. Lors de la première corrélation, la température à ces jours n'était pas autant ressortie, surtout pour celle du jour 44 (0.22 de corrélation). Il faudra donc voir si avec d'autres méthodes elle sort de nouveau.

3.2.2.2 État 'Diarr'

Nous avons par la suite fait le même test avec l'état de diarrhée des chiots. Nous avons suivi la même procédure pour le nettoyage des données et lors du nommage des données d'entrées et la cible mais en définissant pour cette fois l'état de diarrhée des chiots comme cible. Puis, on affiche les résultats.

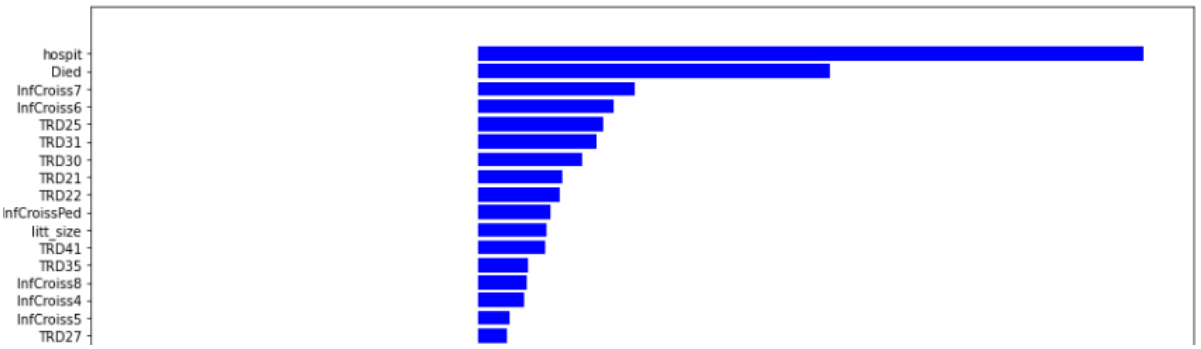


Figure 17. Affichage de la méthode Filter sur la diarrhée

Cette fois-ci lors de l'analyse des résultats on retrouve bien nos attributs mort et hospitalisation, mais on voit apparaître aussi ceux qui définissent si un chiot a eu une infériorité de croissance sur une semaine définie, ici c'est pour les semaines 7 et 6. Données qui ne ressortaient pas auparavant, mais qui n'ont pas un énorme coefficient ici non plus.

3.2.2.3 État 'hospit'

Puis nous avons réalisé la même chose pour l'état de santé hospitalisé.

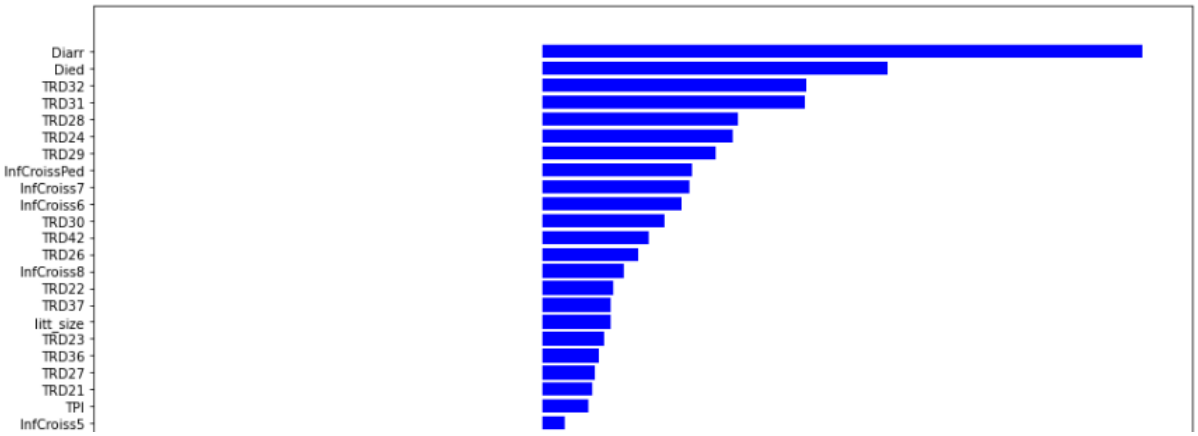


Figure 18. Affichage de la méthode Filter sur l'hospitalisation

Pour l'état d'hospitalisation du chiot on retrouve bien l'état de diarrhée et de mort et comme vu plus tôt on voit bien ressortir les températures du jour 32, 31 et 28, mais on voit aussi que la température du jour 24 ressort plus cette fois-ci.

3.2.3 Feature Selection Embedded[6]

Les méthodes intégrées combinent les aspects avantageux des méthodes Filter et Wrapper et ressemblent fortement à la méthode Wrapper en termes de fonctionnement. À première vue, les deux méthodes sélectionnent les caractéristiques en fonction de la procédure d'apprentissage du modèle d'apprentissage automatique. Cependant, les méthodes Wrapper considèrent les caractéristiques non importantes de manière itérative en fonction de la métrique d'évaluation, tandis que les méthodes Embedded effectuent la sélection des caractéristiques et l'apprentissage de l'algorithme en parallèle. Le processus de sélection des caractéristiques fait partie intégrante de ce modèle de classification / régression. Les méthodes Wrapper et Filter sont des processus discrets, dans le sens où les caractéristiques sont soit conservées, soit écartées. Cependant, cela peut souvent entraîner une variance élevée. À l'inverse, les méthodes intégrées sont plus continues et ne souffrent donc pas tant que cela d'une variabilité élevée.

[6] Cf. Notebook 'Feature Selection Embedded.ipynb'

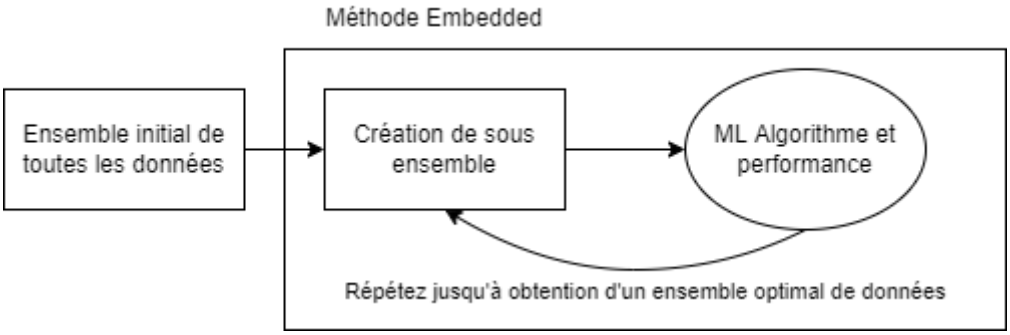


Figure 19. Fonctionnement méthode Embedded

Pour notre projet nous ne traiterons et ne parlerons que de la méthode de régression Ridge car c'est la seule qui nous a donné des résultats probants et exploitables.

3.2.3.1 État 'Died'

Comme pour pouvoir utiliser les précédentes méthodes, nous avons dans un premier temps, récupéré les valeurs de notre dataset, éliminé les colonnes qui ne nous importaient pas et défini les colonnes que nous utiliserons comme variables d'entrée et notre cible, pour notre premier cas la cible est l'état de mort du chiot.

```
TRD45          0.269147
hospit         0.185809
Diarr          0.163864
TRD50          0.153662
TRD35          0.135219
TRD37          0.133265
TRD25          0.120417
TRD54          0.113093
TRD23          0.086972
TRD40          0.078591
InfCroiss7     0.038270
TRD56          0.037458
TRD33          0.032804
```

Figure 20. Meilleur coefficient pour la cible 'Died'

Après exécution de notre modèle on obtient les valeurs ci-dessus. Nous avons décidé de manière arbitraire de n'afficher que les valeurs supérieures à un coefficient de 0.03 car dans une régression de Ridge les coefficients des paramètres non importants à se rapprocher de 0 et les caractéristiques ayant des coefficients négatifs ne contribuent que très peu au modèle et sont donc moins pertinentes.
Dans les résultats obtenus, on voit apparaître la température du jour 45 en premier, puis comme souvent on retrouve les états de diarrhée et d'hospitalisation des chiots et enfin les températures des jours 50 et 35. Les coefficients de ces caractéristiques ne sont pas très élevés mais ils pourraient nous permettre de prédire l'état de mort d'un chiot.

3.2.3.2 État ‘Diarr’

On réitère la mise en place de la méthode mais en prenant comme cible l’état de diarrhée des chiots.

hospit	0.714695
Died	0.224922
TRD39	0.162603
TRD44	0.156917
TRD53	0.149231
TRD48	0.122986
InfCroiss8	0.121642
InfCroiss5	0.118985
InfCroiss4	0.095640
TRD26	0.082417
TRD49	0.074685
TRD43	0.067677
TRD55	0.063568
InfCroissPed	0.059824
InfCroiss6	0.036969
TRD28	0.033200
TRD47	0.031558

Figure 21. Meilleur coefficient pour la cible ‘Diarr’

Cette fois, le premier résultat que nous donne la méthode a un coefficient très important (supérieur à 0.7) est c’est l’hospitalisation du chiot. Vient ensuite l’état de mort et les températures des jours 39, 44 et 53, mais leurs coefficients montrent qu’ils sont moins pertinents pour le modèle. Néanmoins, il faudra voir si des températures sont apparues dans d’autres modèles.

3.2.3.3 État ‘hospit’

On applique la même chose pour l’état d’hospitalisation d’un chiot.

Diarr	0.484349
TRD38	0.214908
TRD24	0.210057
TRD31	0.175193
Died	0.172844
TRD32	0.135998
TRD34	0.114144
TRD21	0.108962
TRD51	0.103572

Figure 22. Meilleur coefficient pour la cible ‘hospit’

Comme lors du premier set de valeur avec l’état ‘Died’, les valeurs sont assez homogènes. Nous retrouvons l’état de diarrhée, puis les températures des jours 38, 24, 31 et enfin l’état de mort.

Il ne reste plus qu’à corréler les résultats obtenus avec tous les modèles de Features Selection pour voir quelles caractéristiques semblent être les plus importantes et pertinentes à utiliser pour pouvoir prédire les différents états de santé d’un chiot.

3.2.4 Résultat pour les trois méthodes de Feature Selection

État ‘Died’ :

Wrapper : ‘hospit’, ‘InfCroiss5’, ‘TRD35’, ‘TRD36’, ‘TRD45’
Filter : ‘hospit’, ‘Diarr’, ‘TRD44’, ‘TRD35’, ‘TRD50’
Embedded : ‘TRD45’, ‘hospit’, ‘Diarr’, ‘TRD50’, ‘TRD35', ‘TRD37’

État ‘Diarr’ :

Wrapper : ‘Age_D1’, ‘TRD48’, ‘TRD54’, ‘TRD55’, ‘TRD56’
Filter : ‘hospit’, ‘Died’, ‘InfCroiss7’, ‘InfCroiss8’, ‘TRD25’
Embedded : ‘hospit’, ‘Died’, ‘TRD39’, ‘TRD44’, ‘TRD53’

État ‘hospit’ :

Wrapper : ‘Died’, ‘Diarr’, ‘InfCroissPed’, ‘TRD24’, ‘TRD31’
Filter : ‘Died’, ‘Diarr’, ‘TRD31’, ‘TRD32’, ‘TRD28’
Embedded : ‘Diarr’, ‘TRD38’, ‘TRD24’, ‘TRD31’, ‘Died’

En analysant le tableau récapitulatif des résultats des trois modèles on observe que selon l’état de santé du chiot plusieurs caractéristiques ressortent fortement et semblent donc être de bons paramètres à utiliser par des algorithmes de prédiction.

3.3 Méthode d'analyse LSTM

3.3.1 Architecture d'un LSTM

En parallèle de nos recherches décrites plus haut, nous avons de même essayé de créer un réseau neuronal récurrent (RNN) particulier, appelé LTSM (Long Short Term Memory). Ce modèle de réseau neuronal a été introduit par Hochreiter & Schmidhuber en 1997.
La caractéristique de ce RNN est qu'il est capable d'apprendre des dépendances à long terme, autrement dit, de pouvoir utiliser l'information précédente. En effet, pour faire son apprentissage, le RNN classique utilise la méthode de la descente du gradient afin de mettre à jour les poids entre ses neurones selon la formule suivante :

$$w := w - \alpha \cdot Fw$$

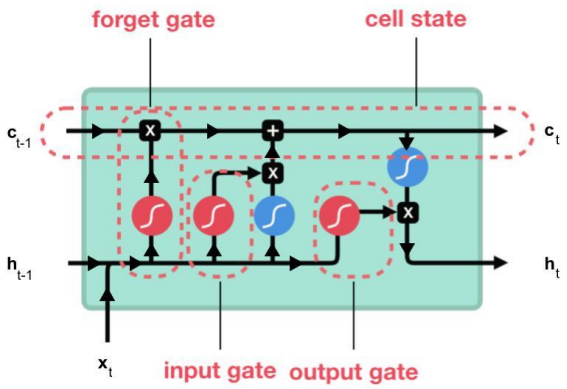
avec :
w un poids du réseau
α la vitesse d'apprentissage du réseau
Fw le gradient du réseau par rapport au poids w

Le problème de cette mise à jour des poids est qu'étant donné que les données d'entrée vont de droite à gauche, le produit de la vitesse d'apprentissage du réseau par le gradient du réseau par rapport au poids w tend vers 0 rapidement. Les poids des premières couches de neurones ne sont quasiment pas modifiés, d'où un apprentissage médiocre.

C'est là qu'intervient le LSTM, avec ses cellules composées de trois "portes". Ces trois portes sont des zones de calculs régulant le flot d'informations.
Les trois portes sont les suivantes :

- La porte d'oubli (*forget gate*)
- La porte d'entrée (*porte d'entrée*)
- La porte de sortie (*output gate*)

On peut modéliser la cellule comme ci-suit :

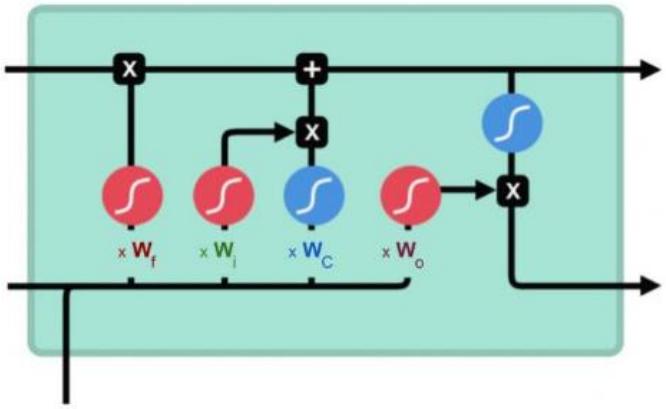


A l'intérieur de la cellule, on peut réaliser de nombreuses opérations, dont la représentation est donnée en annexe.

En mettant en place les trois portes de la cellule, on obtient 4 matrices de poids :

- *Wf* : pondère l'entrée de la porte d'oubli (forget gate)
- *Wi* : pondère l'entrée de la porte d'entrée (input gate)
- *WC* : pondère les données qui vont se combiner à la porte d'entrée pour mettre à jour l'état de la cellule (cell state)
- *Wo* : pondère l'entrée de la porte de sortie (output gate)

La dynamique de ces poids contrôlant les portes peut se voir ci-contre :



Décrivons le rôle de chaque porte afin de mieux comprendre le fonctionnement du LSTM.

Porte	Rôle
Porte d'oubli	❖ Choix de l'information (information de l'état précédent concaténé à la donnée en entrée) à conserver ❖ Application de la fonction sigmoïde : si sortie proche de 0, oubli de l'information, sinon, mémorisation pour la suite
Porte d'entrée	❖ Extraction de l'information ❖ Application de la sigmoide aux données concaténées et d'une tanh
Etat de la cellule (*cell state*)	❖ Obtention de l'état de la cellule avec ➢ Multiplication de la sortie de l'oubli avec l'ancien état de la cellule ➢ Addition du résultat à la sortie de la porte d'entrée
Porte de sortie	❖ Choix du nouvel état de la cellule normalisé

Nous avons donc implémenté un modèle de LSTM comprenant deux couches LSTM avec 30 neurones à activation tanh, puis une couche entièrement connectée de 30 neurones avec une activation de type softmax. L'optimizer choisi est l'optimizer ADAM, avec un taux d'apprentissage de 0,01. Cette configuration nous permet d'obtenir 12 210 paramètres à calculer, ce qui est raisonnable.
L'entraînement s'est effectué avec 20 epochs, soit 20 tours de l'ensemble de batches de 50 éléments chacun.
Afin d'avoir un modèle ayant une justesse la plus haute possible, nous avons modulé le nombre de couches, de neurones, d'epochs et de neurones.

L'idée est de prendre en variables les deux températures ayant été mises en valeur lors du feature selection, afin de prédire la mort des chiots. Malheureusement, le modèle n'a pas pu être entraîné jusqu'au bout, faute de morts de chiots dans le dataset.
Nous avons alors tenté de prédire les épisodes de diarrhée, ayant plus de cas positifs, et cela s'est conclu par l'entraînement du modèle avec une justesse (*accuracy*) de 1.
Cette accuracy peut sembler être bonne mais la justesse du modèle concernant les données de validation sont très faibles (inférieures à 20%). Par ailleurs, la val_loss est NaN (Not a Number), ce qui indique un mauvais entraînement du modèle. Cela n'est pas étonnant, mais on semble être sur la bonne piste avec plus de données.

4. Conclusion

Suite aux différentes analyses effectuées des données à notre disposition, il est difficile de faire apparaître de façon claire une corrélation entre la température des chiots et leur mortalité ou plus largement entre leur température et leur état de santé. En effet, même si le bon sens, couplé à des décennies d'études du vivant, de l'animal et de sa santé, nous poussent à affirmer qu'un chiot dépassant une limite haute ou basse de température a un très fort risque de mortalité, les résultats de nos recherches ne sont pas probants.

Après avoir utilisé différentes méthodes pour tenter d'extraire des données fournies une corrélation, plusieurs critères sont apparus comme potentiellement déterminants. Néanmoins, comme susmentionné dans ce rapport, l'échantillon sur lequel nous avons été amenés à travailler s'est avéré trop restreint pour confirmer ces corrélations.

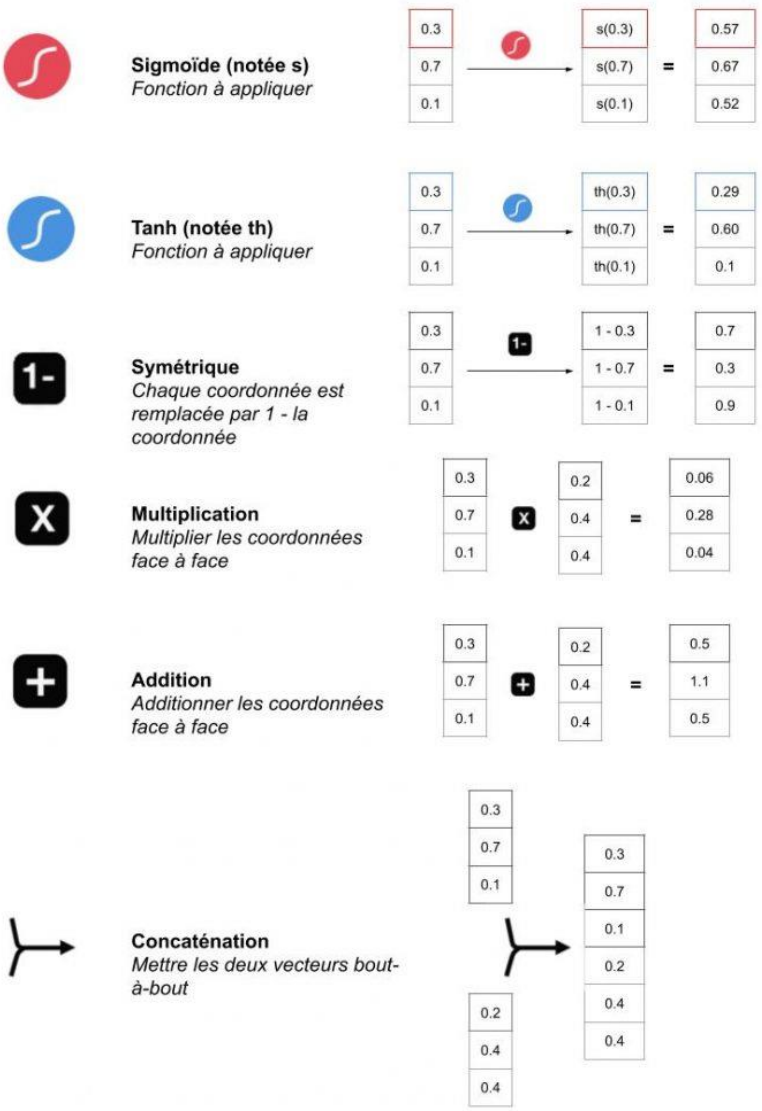
Ne pouvant disposer d'un panel plus important de chiots, et donc de données, au moment de notre analyse, notre étude s'est positionnée comme un travail préparatoire à un futur projet. En effet, nous avons cherché à pré-sélectionner des critères semblant être corrélés avec l'état de santé des chiots. Ces caractéristiques seront à confirmer lors d'une prochaine étude, sur un set de données plus étendu. Cela permettra de gagner du temps au début du projet en s'intéressant immédiatement aux critères sélectionnés par nos méthodes.

Annexe

Annexe 1 : Glossaire du dataset fourni

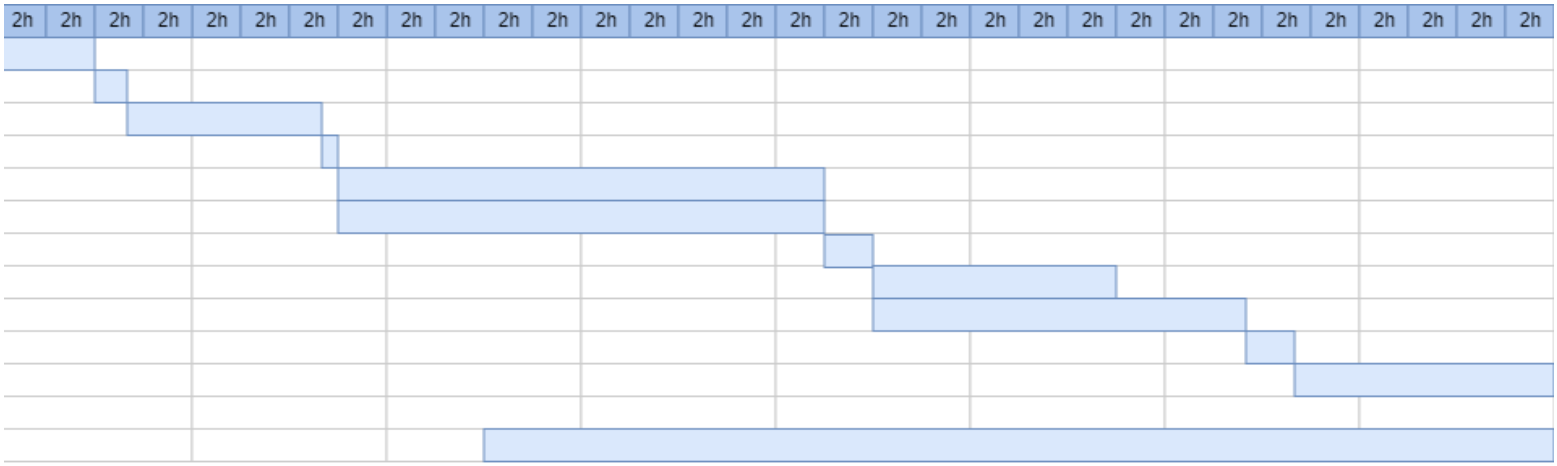
Caractérisation	ID	Numéro d'Identification du chiot
	litter_ID_late	Numéro d'identification de la portée
	Size	Format racial (S, M, L)
	litt_size	Taille de la portée
	sex	Sexe du chiot
	group	Group (Late = supplémenté IgY late ou Control)
	IgGJ2	Valeur Immunoglobuline G (IgG) au jour 2
	TPI	Défaut de transfert de l'immunité passive ou pas (oui si IgG J2 < 2.3)
Santé	Died	État de mort du chiot
	Diarr	État de diarrhée du chiot
	Age_D1	Jour de la première diarrhée du chiot (peut être nul)
	AgeD2	Jour de la deuxième diarrhée du chiot (peut être nul)
	hospit	État d'hospitalisation du chiot
	InfCroissPed	Présence ou non d'un taux de croissance sur la période pédiatrique inférieur au seuil
	InfCorissX	Présence ou non d'un taux de croissance sur la sur la semaine X (entre 4 et 8) inférieur au seuil
	Min1inflcr	Présence ou non d'au moins une semaine avec une inflection de croissance (GR < seuil) au cours de la période pédiatrique
Température	TRDXX	Température rectale du chiot au jour XX (entre 21 et 56)

Annexe 2 : Opérations réalisées par les cellules



Annexe 3 : Diagramme de Gantt

Tâches
Prise en main du sujet par les nouveaux arrivants
Prise de contact client
Rechecher et analyse basique
Prise de rdv tuteur école pour état des lieux de l'avancement
Feature selection orienté mort du chiot
LSTM
Point d'avancement et confrontation des res
FS orienté diarrhée et hospit
LSTM sur features seclionnées
Bilan des résultat et point d'avancement avant soutenance
Finalisation du rapport et préparation de la soutenance
Rapport



Bibliographie

Feature Selection :

Brownlee, J. (2021, 29 juin). *An Introduction to Feature Selection*. Machine Learning Mastery.
https://machinelearningmastery.com/an-introduction-to-feature-selection/

Menon, K. (2021, 16 septembre). *Everything You Need to Know About Feature Selection In Machine Learning*. Simplilearn.Com.
https://www.simplilearn.com/tutorials/machine-learning-tutorial/feature-selection-in-machine-learning

Garg, S. (2022, 6 janvier). *Feature Selection Using Filter Method : Python Implementation from Scratch*. Medium.
https://medium.com/mlearning-ai/feature-selection-using-filter-method-python-implementation-from-scratch-375d86389003

Verma, V. (2020, 29 décembre). *Feature Selection using Wrapper Method - Python Implementation*. Analytics Vidhya.
https://www.analyticsvidhya.com/blog/2020/10/a-comprehensive-guide-to-feature-selection-using-wrapper-methods-in-python/

Tzini, E. (2022, 30 mars). *Feature Selection : Embedded Methods - Analytics Vidhya*. Medium.
https://medium.com/analytics-vidhya/feature-selection-embedded-methods-a7940036973f

Ephraim, E. O. (2022, 5 janvier). *Embedded Feature Selection in Machine Learning.* Myrtle's Blog.
https://myrtle.hashnode.dev/embedded-feature-selection-in-machine-learning

Bhunia, A. (2020, 11 mars). *T101 : Embedded method-Feature selection techniques in machine learning*. UpSkillPoint.
https://www.upskillpoint.com/machine%20learning/2020/03/11/feature-selection-using-embedded-method/

LSTM :

Keras documentation : LSTM layer. (s. d.). Keras.
https://keras.io/api/layers/recurrent_layers/lstm/

Brownlee, J. (2020, 20 octobre). *Multivariate Time Series Forecasting with LSTMs in Keras*. Machine Learning Mastery.
https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/

Towards AI Editorial Team. (2022, 18 mars). *Main Types of Neural Networks and its Applications*. Towards AI.
https://towardsai.net/p/machine-learning/main-types-of-neural-networks-and-its-applications-tutorial-734480d7ec8e

K.H., P. (2021, 22 décembre). *Multi-Step Multivariate Time-Series Forecasting using LSTM*. Medium.
https://pangkh98.medium.com/multi-step-multivariate-time-series-forecasting-using-lstm-92c6d22cd9c2

Pang, M. (s. d.). *GitHub - mikepang98/TimeSeriesLSTM: Fully coded with Google Colab.* GitHub.
https://github.com/mikepang98/TimeSeriesLSTM

J. (2020, 14 décembre). *Multivariate Time Series Forecasting with LSTMs in Keras*. Analytics Vidhya.
https://www.analyticsvidhya.com/blog/2020/10/multivariate-multi-step-time-series-forecasting-using-stacked-lstm-sequence-to-sequence-autoencoder-in-tensorflow-2-0-keras/

Brownlee, J. (2020, 20 octobre). *Multivariate Time Series Forecasting with LSTMs in Keras*. Machine Learning Mastery.
https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/

Lendave, V. (2021, 9 juillet). *How To Do Multivariate Time Series Forecasting Using LSTM*. Analytics India Magazine.
https://analyticsindiamag.com/how-to-do-multivariate-time-series-forecasting-using-lstm/

Srinidhi, S. (2020, 9 janvier). *Label Encoder vs. One Hot Encoder in Machine Learning*. Medium.
https://contactsunny.medium.com/label-encoder-vs-one-hot-encoder-in-machine-learning-3fc273365621

C.O.L.A.H. (s. d.). *Understanding LSTM Networks*. GitHub. Consulté le août 2015, à l'adresse
https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Messaoud, Y. (2018, 27 novembre). *LSTM, Intelligence artificielle sur des données chronologiques*. Medium.
https://medium.com/smileinnovation/lstm-intelligence-artificielle-9d302c723eda

R., L. (2020, 12 février). *Comprendre le fonctionnement d'un LSTM et d'un GRU en schémas*. Pensée Artificielle.
https://penseeartificielle.fr/comprendre-lstm-gru-fonctionnement-schema/