Projet Machine Learning

Boucherrab meziane, Bekda Lilia, Boukhemis Imène

Exercice 1

1. Étant donné des observations $d_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, on cherche à prédire les sorties $y_i \in \mathcal{Y}$ à partir des entrées $x_i \in \mathcal{X}$. On suppose que ces données d_n sont des réalisations d'un échantillon $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. L'objectif de l'apprentissage supervisé est de construire un "bon" prédicteur $f : \mathcal{X} \to \mathcal{Y}$ prédisant une variable Y à partir de données X, sans connaître la loi de probabilité $\mathbb{P}$. On suppose souvent que les couples (X_i, Y_i) sont indépendants et identiquement distribués (iid). On définit une fonction de perte $l(f(x), y)$ qui permet de mesurer la capacité de $f(x)$ à "prédire" y. On définit le risque comme étant la perte moyenne pour un nouveau couple : $R(f) = \mathbb{E}[l(Y, f(X))]$. L'objectif est de prédire le meilleur modèle, en minimisant ce risque.

Lorsque le problème est un problème de classification et que la variable à prédire est qualitative à deux modalités, cela signifie que $Y \in \{-1, +1\}$. Dans ce cas, on choisit la fonction de perte $l(f(x), y) = \mathbb{1}_{Y \neq f(X)}$ et le risque $R(f) = \mathbb{E}[l(Y, f(X))] = \mathbb{P}(Y \neq f(X))$.

2.Présentation des données : Histogramme : L'analyse des histogrammes des données sélectionnées révèle une division entre deux types de variables : continues et discrètes. Nous observons que la plupart des variables discrètes présentent une forte concentration autour de la valeur 0, suggérant une distribution asymétrique des données avec une tendance marquée vers cette valeur.

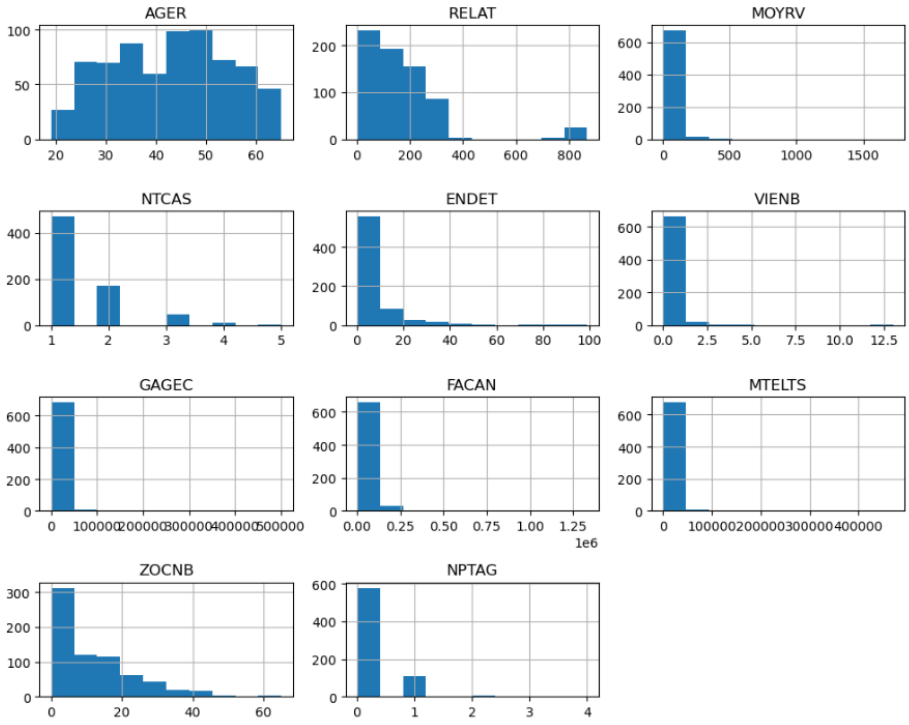

FIGURE 1 – Histogramme des variables

Nuages de points : Dans la figure ci-dessus, chaque nuage de points représente la distribution des données pour une variable spécifique. Les points dispersés dans chaque nuage illustrent la répartition des valeurs, qu'elles soient continues ou discrètes. Nous remarquons que certaines variables continues présentent une dispersion importante des points, ce qui indique une variabilité significative dans ces données. Pour les variables discrètes, nous observons des regroupements de points, suggérant une concentration autour de certaines valeurs. Ces observations soulignent l'importance de mettre à l'échelle et de standardiser les données afin d'assurer une meilleure performance lors de la modélisation, en permettant aux variables de contribuer de manière équilibrée et significative aux prédictions.

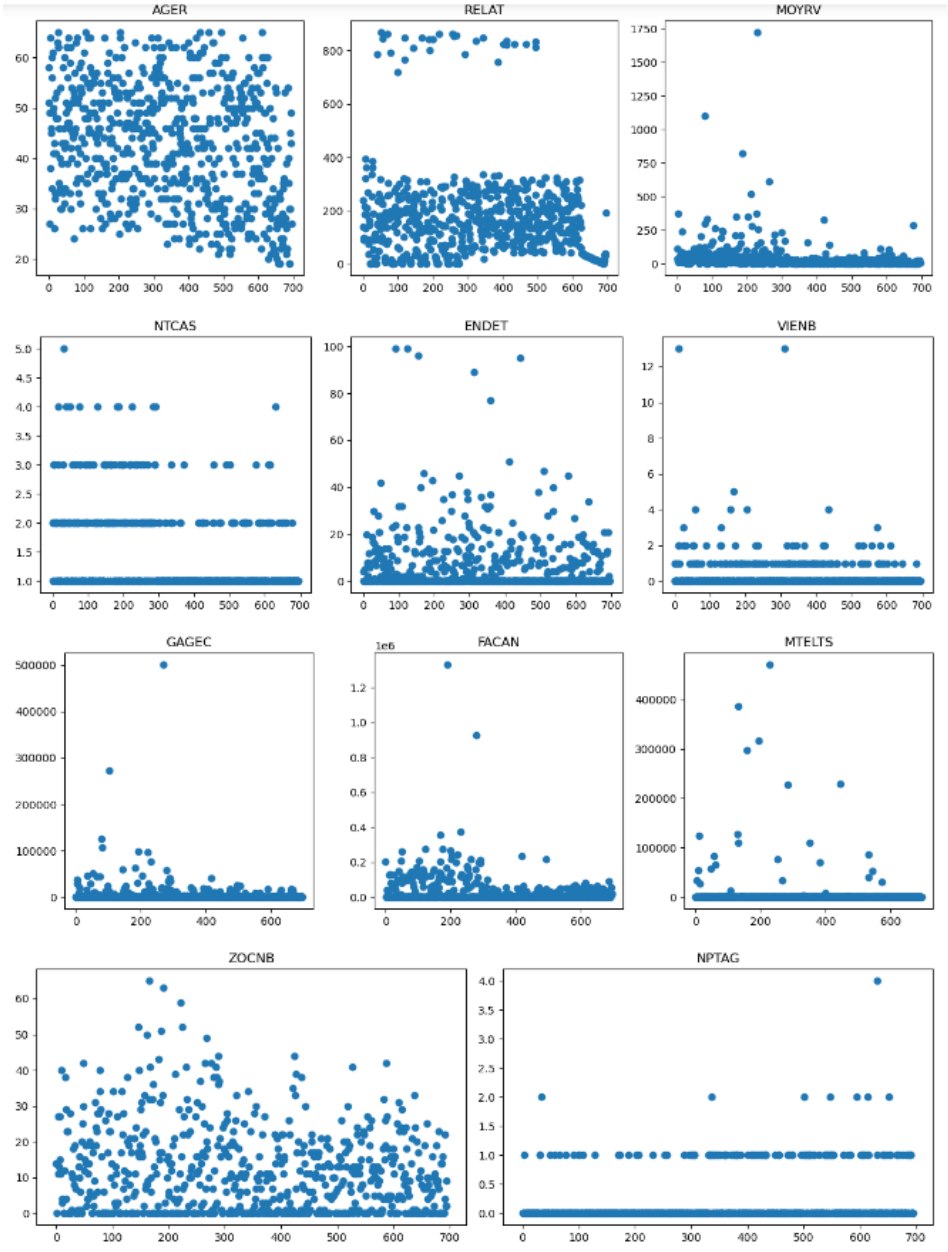

FIGURE 2 – Nuages de points

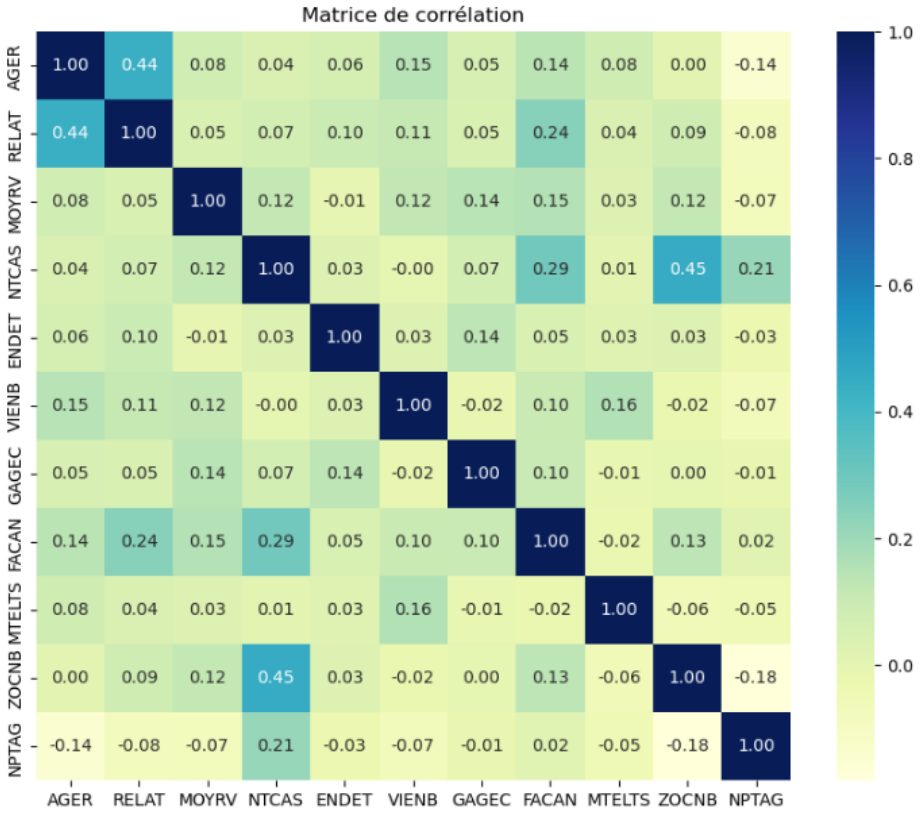

FIGURE 3 – Matrice de corrélation

La matrice de corrélation révèle plusieurs relations entre les variables de notre ensemble de données. Nous observons des corrélations positives significatives entre certaines paires de variables, telles que l'âge (AGER) et la relation (RELAT), ainsi que le nombre de transactions (NTCAS) et le facteur de canaux (FACAN). À l'inverse, des corrélations négatives significatives sont également présentes, comme entre l'âge (AGER) et le nombre de tags (NPTAG). Ces résultats soulignent l'importance des interactions entre les variables dans notre ensemble de données et mettent en lumière des pistes potentielles pour une analyse plus approfondie.

3.a.1 Régression logistique : Comme expliqué à la question 1., on cherche à prédire $Y \in \{-1, 1\}$. On a $Y|X \sim \text{Bernoulli}(p(X))$, avec $p(X) = \mathbb{P}(Y = 1|X)$ à modéliser. La régression logistique consiste à estimer β_0 et β_1 tel que :

$$log(\frac{\mathbb{P}(Y=1|X)}{\mathbb{P}(Y=0|X)} = \beta_0 + \beta_1 X \text{ ou de manière équivalente } \mathbb{P}(Y=1|X) = \frac{\exp \beta_0 + \beta_1 X}{1 + \exp \beta_0 + \beta_1 X}$$

. Les coefficients β sont estimés en maximisant la vraisemblance $L_n(\beta) = \prod_{i=1}^{n} P_\beta(Y = y_i|X = x_i)$.

La régression logistique pour les valriables qu'on a choisis nous donne la sortie suivante :

```
         Current function value: 0.433090
         Iterations 8
              precision    recall  f1-score   support

           0       0.84      0.88      0.86        86
           1       0.80      0.74      0.77        54

    accuracy                           0.83       140
   macro avg       0.82      0.81      0.82       140
weighted avg       0.83      0.83      0.83       140

                           Logit Regression Results
==============================================================================
Dep. Variable:                  CARVP   No. Observations:                  556
Model:                          Logit   Df Residuals:                      544
Method:                           MLE   Df Model:                           11
Date:                Thu, 18 Apr 2024   Pseudo R-squ.:                 0.3662
Time:                        20:26:49   Log-Likelihood:                -240.80
converged:                       True   LL-Null:                       -379.90
Covariance Type:            nonrobust   LLR p-value:                 3.380e-53
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0236      0.143     -0.166      0.868      -0.303       0.256
0              0.3007      0.127      2.359      0.018       0.051       0.551
1             -0.4447      0.133     -3.355      0.001      -0.705      -0.185
2              1.5599      0.399      3.907      0.000       0.777       2.342
3              1.7455      0.216      8.074      0.000       1.322       2.169
4             -0.0127      0.102     -0.125      0.901      -0.213       0.188
5             -0.1338      0.098     -1.369      0.171      -0.325       0.058
6              0.7481      0.370      2.021      0.043       0.022       1.474
7              0.9609      0.269      3.574      0.000       0.434       1.488
8              0.2162      0.147      1.468      0.142      -0.072       0.505
9             -0.6265      0.152     -4.132      0.000      -0.924      -0.329
10            -1.3113      0.219     -5.989      0.000      -1.740      -0.882
```

FIGURE 4 – Régression Logistique

Ce rapport de régression logistique met en évidence l'importance de certains coefficients dans la prédiction de la variable cible, comme confirmé par leurs p-values significatives. Les coefficients positifs ou négatifs dénotent l'impact des variables correspondantes sur la variable cible. Le pseudo R-carré, d'une valeur de 0,3662, fournit une mesure de l'ajustement global du modèle aux données. Les tests de Wald renforcent la validité des coefficients estimés.

En dépit d'un R^2 relativement modeste, ce modèle de régression logistique semble être efficace pour la prédiction, compte tenu des variables incluses. Il est également important de noter l'évaluation visuelle de la performance du modèle à travers la courbe ROC et l'AUC, présentées dans la figure à la page X, permettant d'apprécier sa capacité à discriminer entre les classes positives et négatives.

3.a.2 Random Forest :

À partir de la taille des arbres, du nombre d'arbres à utiliser et le nombre de caractéristiques échantillonnées, le Random Forest crée dans un premier temps de nombreux sous-échantillons aléatoires de notre ensemble de données, tirage uniforme avec remise (bagging). Il construit ensuite des arbres de décision individuels pour chaque échantillon. Chaque arbre est entraîné sur une portion aléatoire afin de recréer une prédiction. La combinaison de tous ces modèles indépendants qui permettent de réduire la variance du modèle d'ensemble. Enfin, l'algorithme prédit un résultat pour chaque arbre. Le résultat le plus fréquent devient le résultat final de notre modèle, dans le cas de la classification.

L'un des avantages du Random Forest est sa précision parce qu'il combine les résultats de plusieurs sous-modèles non corrélés. Malheureusement, lorsque les données sont éparses ou lorsqu'il y a plusieurs valeurs manquantes dans le jeu de données, l'algorithme peut conduire à de mauvais résultats. Il ne permet pas de prédire des valeurs/situations qui n'existent pas dans le jeu de données. Il est donc très adapté pour les problèmes de classification.

Dans un random forest de classification, l'estimation finale consiste à choisir la catégorie de réponse la plus fréquente. Plutôt qu'utiliser tous les résultats obtenus, on procède à une sélection en recherchant la prévision qui revient le plus souvent.

```
Exactitude du modèle (Random Forest) : 0.85
Rapport de classification :
              precision    recall  f1-score   support

           0       0.89      0.86      0.88        86
           1       0.79      0.83      0.81        54

    accuracy                           0.85       140
   macro avg       0.84      0.85      0.84       140
weighted avg       0.85      0.85      0.85       140

Meilleurs hyperparamètres : {'n_estimators': 100, 'max_depth': None, 'min_samples_split': 10, 'min_samples_leaf': 4}
Importance des caractéristiques : [0.10357878 0.20714304 0.41564937 0.         0.17582978 0.
 0.         0.04562455 0.00939052 0.04278396]
```

Figure 5 – Random Forest

L'exactitude du modèle Random Forest est de 85%, avec des précisions de 89% pour la classe 0 et de 79% pour la classe 1. Le recall est de 86% pour la classe 0 et de 83% pour la classe 1. Les scores F1 sont respectivement de 88% et 81% pour ces classes. Globalement, la moyenne des scores précision, recall et F1 est de 84%. Les meilleurs hyperparamètres identifiés incluent 100 arbres, une profondeur maximale non spécifiée, un minimum de 10 échantillons pour diviser un nœud, et un minimum de 4 échantillons par feuille. En termes d'importance des caractéristiques, la troisième et la cinquième caractéristiques sont les plus importantes, les autres ayant une contribution relativement faible.

3.a.3 Boosting :

Le boosting est une méthode qui combine un ensemble d'apprenants faibles en un apprenant fort, afin de réduire les erreurs d'apprentissage, à l'aide d'arbres de décisions. Dans le boosting, un échantillon aléatoire de données est sélectionné, doté d'un modèle, puis entraîné séquentiellement (chaque modèle tente de compenser les faiblesses de son prédécesseur via des poids). À chaque itération, les règles faibles de chaque classificateur individuel sont combinées pour former une seule règle de prédiction forte.

Voici deux exemples de types de boosting :

— AdaBoost : Cette méthode fonctionne de manière itérative, en identifiant les points de données mal classifiés et en ajustant leurs poids pour réduire l'erreur d'entraînement. Le modèle continue à être optimisé de manière séquentielle jusqu'à ce qu'il donne le prédicteur le plus fort.

— Boosting de gradient : On ajoute séquentiellement des prédicteurs à un ensemble, chaque prédicteur corrigeant les erreurs de son prédécesseur. Il s'entraîne sur les erreurs résiduelles du prédicteur précédent, avec l'algorithme de descente de gradient.

Ainsi, le boosting permet de réduire les biais élevés et permet d'augmenter la précision prédictive lors de l'entraînement. Cependant, les valeurs aberrantes ou les données différentes du reste du jeu de données peuvent fausser considérablement les résultats, car chaque modèle tente de corriger les défauts de son prédécesseur.

Dans le cas de la classification, on construit l'arbre en partant d'une seule région contenant toutes les données et en divisant récursivement ces régions avec l'algorithme. On choisit d'attribuer par exemple la valeur -1 ou 1 à la région en fonction de la classe majoritaire.

3.a.4 Réseau de neurones :

Chaque neurone reçoit une entrée, effectue un calcul à l'aide d'une fonction d'activation, et transmet son résultat au neurone suivant. Les données dans un réseau de neurones passent à travers différentes couches. Ces couches transforment progressivement l'entrée brute en une représentation plus abstraite et utile, puis finalement un résultat est généré à la sortie. Dans le cas de l'apprentissage supervisé, l'algorithme s'entraîne avec des données étiquetées pour optimiser ses performances et atteindre les résultats attendus pour chaque entrée.

L'une des tâches dans lesquelles les réseaux de neurones sont les plus performants est la classification. Pour cela, ils utilisent un ensemble de données préalablement étiqueté afin de résoudre rapidement des problèmes complexes. Concrètement, un réseau de neurones peut par exemple identifier des motifs dans une série de photos et y appliquer des étiquettes qui lui ont été préalablement fournies. L'algorithme est en mesure de distinguer les caractéristiques les plus importantes de façon autonome avec les étiquettes dont il dispose. Par exemple, les réseaux de neurones sont souvent utilisés dans le cadre de la classification pour identifier des personnes sur des images ; gestes dans une vidéo ; des voix et identifier les intervenants ; transcrire des discours en texte ; classifier des mails en tant que spams...

Le type de réseau de neurones vu en cours notamment pour la classification est le preceptron. Expliquons par exemple le perceptrion simple. Le preceptron simple est une fonction f des entrées $x = (x_1, ..., x_d)$ pondérées par un vecteur $w = (w_1, ..., w_d)$, complétées par un neurone de biais w_0, et une fonction d'activation $\sigma : \mathbb{R} \to \mathbb{R}$, $\hat{y} = f(x) = \sigma(w_0 + w_1x_1 + \ldots + w_dx_d)$. Le preceptron est entraîné par des mises à jour itératives de ses poids, qui sont estimés (à partir des données) en minimisant une fonction de perte. Le seuil du résultat permet de prendre une décision. Si on considère la variable de sortie binaire $\{-1, 1\}$ par exemple, pour le preceptron, on donne à une entrée x0 une valeur de sortie 1 si $\hat{p}$(x0) ≥ 0.5 et -1 sinon.

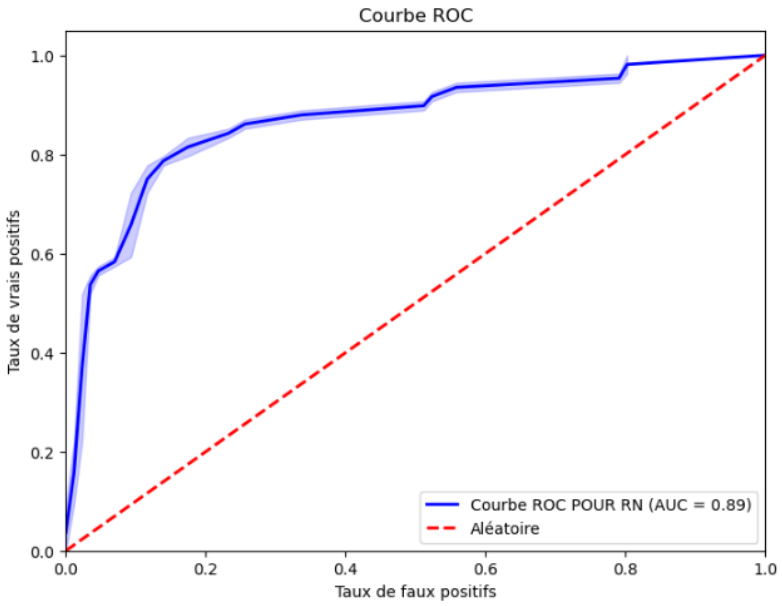

FIGURE 6 – Courbe Roc du rèseau de neuronne

Dans l'analyse de la courbe ROC, une observation cruciale émerge : l'aire sous la courbe (AUC) atteint une valeur remarquable de 0,89, une indication solide de l'efficacité de notre modèle dans la classification des données. Cette performance est soulignée par la proximité de cette mesure à la valeur maximale de 1, démontrant ainsi une capacité de classification robuste.

Comparaison des méthodes :

Exactitude du modèle de régression logistique : 0.8071428571428572
Exactitude du modèle de random forest : 0.85
Exactitude du modèle de gradient boosting : 0.8071428571428572
Exactitude du modèle de Réseau de neuronne : 0.8214285714285714

On remarque donc ici que le modèle le plus efficace pour la prédiction est le Random Forest.

3.b Validation croisée :

La validation croisée est une technique d'évaluation des performances des modèles d'apprentissage automatique. Elle consiste à diviser les données en ensembles d'entraînement et de test de manière itérative. À chaque itération, une partie des données est utilisée pour l'entraînement du modèle, tandis que le reste est utilisé pour le tester. Cette méthode permet d'estimer la capacité du modèle à généraliser sur de nouvelles données en moyennant les performances sur plusieurs itérations. En résumé, la validation croisée aide à évaluer la robustesse et la fiabilité d'un modèle en le testant sur différents sous-ensembles de données.

Pour les résultats de la validation croisée, on obtient le résultat suivant :

Scores de validation croisée : [0.88571429 0.83453237 0.81294964 0.78417266 0.45323741]
Précision moyenne : 0.7541212744090442

Ces scores de validation croisée représentent les performances du modèle Random Forest sur différents plis de données lors de la validation croisée. Les valeurs indiquent la précision du modèle pour chaque pli, avec des valeurs allant de 0.45 à 0.88. La précision moyenne du modèle sur l'ensemble des plis est de 0.75, ce qui suggère que le modèle a une performance généralement bonne mais peut varier selon les données utilisées.

Exercice 2

1. Apprentissage supervisé :

L'objectif de l'apprentissage supervisé est de prédire une variable cible (dans notre cas la concentration en ozone O3obs) à partir d'un ensemble de variables prédictives (les variables climatiques). Nous sommes confrontés à une tâche de régression. Le but de l'apprentissage est d'induire une fonction qui prédise les réponses associées à de nouvelles observations en commettant une erreur de prédiction la plus faible possible.

But de l'Apprentissage supervisé :

L'objectif de l'apprentissage supervisé est de prédire une sortie y_{true} en fonction des entrées x. En d'autres termes, il s'agit de créer un modèle qui peut estimer la valeur de y_{true} en se basant sur les informations fournies par x. Ce modèle est entraîné sur un ensemble de données où les valeurs de x et y_{true} sont connues, de sorte qu'il apprend à établir une relation entre les entrées et les sorties. Une fois que le modèle est entraîné, il peut être utilisé pour faire des prédictions sur de nouvelles données, où les valeurs de y ne sont pas initialement disponibles.

Fonction de perte et risque :

Dans le cas de régression, la fonction de perte couramment utilisée est la l'erreur quadratique moyenne (MSE). Cette fonction mesure l'écart entre la valeur prédite Y_{pred} et la valeur réelle y_true qui est en fonction des entrées x, elle est donnée par la formule :

$$l(Y, X) = (Y - f(X))^2 = (Y_{true} - Y_{pred})^2$$

Le risque théorique est l'espérance de la fonction de perte sur l'ensemble des données :

$$R(f) = \mathbb{E}[l(Y, f(X))] = \mathbb{E}[(Y - f(X))^2] = \mathbb{E}[(Y_{true} - Y_{pred})^2]$$

Le risque empirique est l'estimation du risque théorique sur un échantillon de données :

$$R_{emp}(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_{pred_i} - Y_{true_i})^2$$

Objectif :

L'objectif de l'apprentissage est de trouver un modèle (ou prédicteur) $\hat{f}$ qui minimise le risque empirique, c'est à dire :

$$\hat{f} = f_{\hat{\theta}} = \arg\min_{f_\theta, \theta \in \Theta} \frac{1}{n}\sum_{i=1}^{n} l\left(Y_i, f_\theta(X_i)\right)$$

2. Présentation des données :

Les données ont été extraites et mises en forme par le service concerné de MétéoFrance, de taille totale de 1041, décrite par les 10 variables suivantes :

JOUR : Le type de jour ; férié (1) ou pas (0).

O3obs : La concentration d'ozone effectivement observée le lendemain à 17h locales correspondant souvent au maximum de pollution observée.

MOCAGE : Prévision de cette pollution obtenue par un modèle déterministe de mécanique des fluides (équation de Navier et Stockes).

TEMPE : Température prévue par MétéoFrance pour le lendemain 17h.

RMH2O : Rapport d'humidité.

NO2 : Concentration en dioxyde d'azote.

NO : Concentration en monoxyde d'azote.

STATION : Lieu de l'observation : Aix-en-Provence, Rambouillet, Munchhausen, Cadarache et Plan de Cuques.

VentMO : D Force du vent.

VentANG : Orientation du vent.

NO2 et NO sont des variables liées à la concentration de dioxyde d'azote (NO2) et d'oxyde d'azote (NO). Elles sont souvent utilisées pour évaluer la qualité de l'air et les émissions polluantes.

Nous voulons prédire la variable cible est la concentration en ozone (**O3obs**) à partir des autres variables explicatives.

Voici quelques représentations graphiques de nos données :

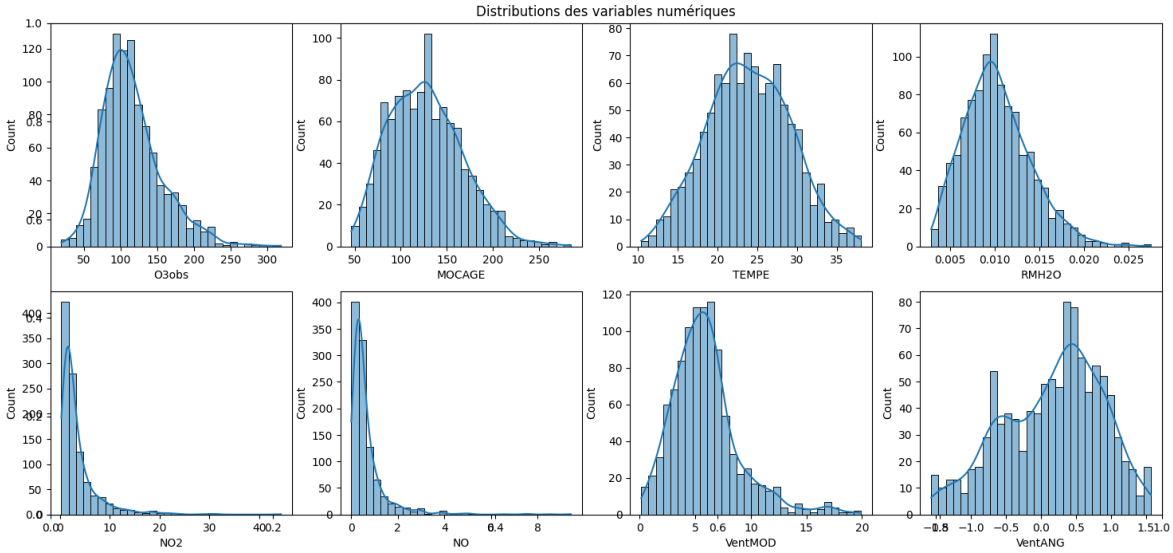

FIGURE 7 – Représentations graphiques des variables numériques brutes

On constate que les variables RMH2O, NO2, NO et ventMOD ne sont pas symétriques, et donc pas gaussiennes. On va alors les normaliser pour une meilleure étude comme suit :

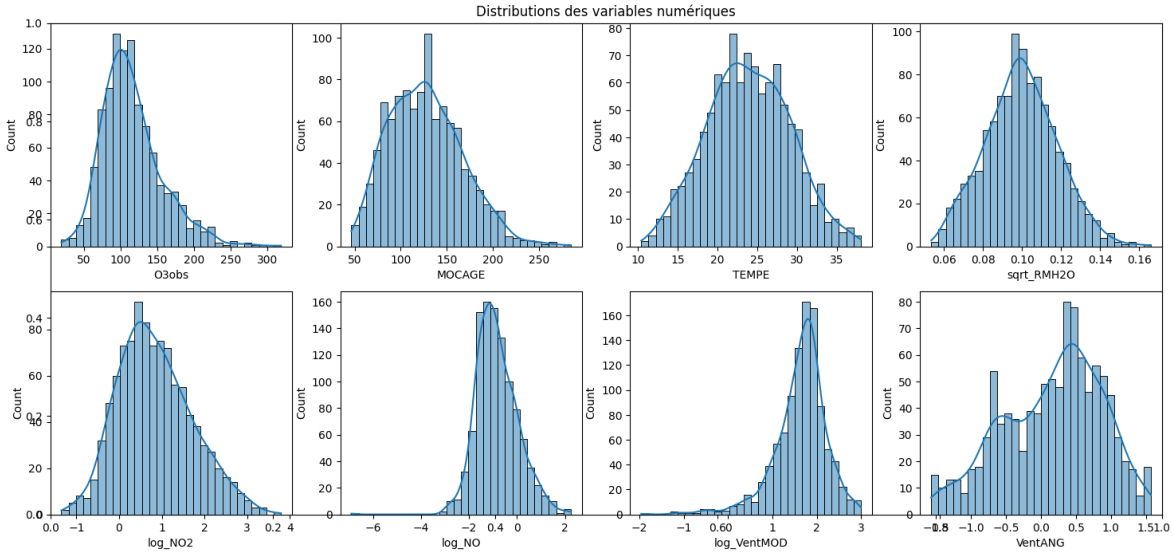

FIGURE 8 – Distribution des variables numériques

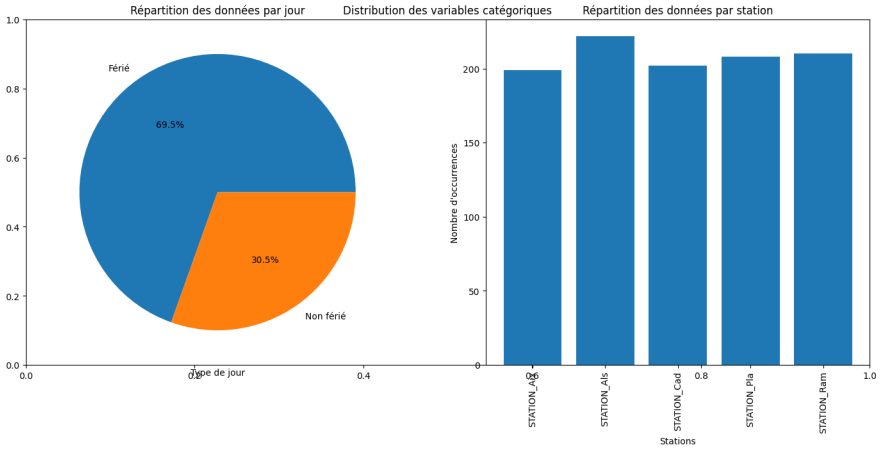

FIGURE 9 – Distribution des variables catégoriques

Voici la matrice de corrélation entre les variables de notre base de données, elle donnes les liens qui peuvent exister entre les variables.

Remarque : la correlation de deux variables ne signifie pas forcement qu'il y a un lien explicatif entre elle.

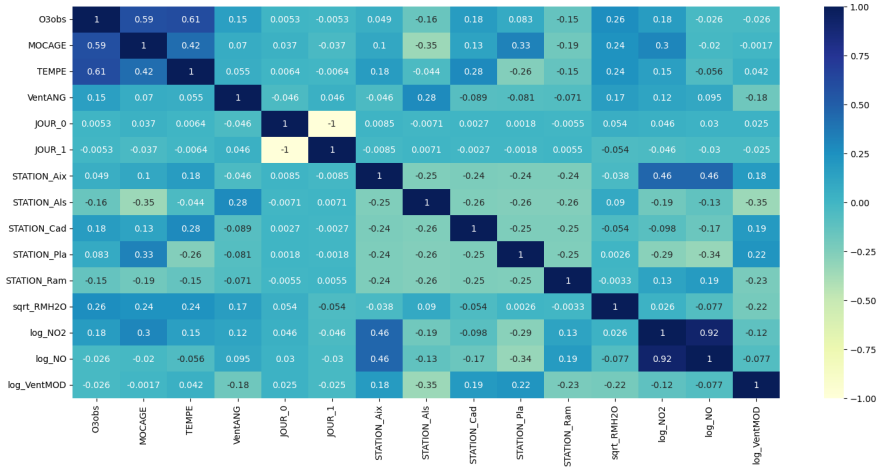

FIGURE 10 – Matrice de corrélation

On remarque par ailleurs que la variable O3obs est :

1.**Moyennement** corrélée avec les variables MOCAGE, TEMPE avce des coefficients de 0.59 e 0.61 respectivement.

2. **Légèrement** corrélée avec les variables VentANG, log_NO2, STATION_ALS, STATION_Cad, STATION_Ram et sqrt_RMH2O.

3. **Faiblement** corrélée avec JOUR80, JOUR_1, STATION_Aix, STATION_Pla, log_NO et lo_h_VentMOD.

Ces corrélation peuvent s'expliquer logiquement, en effet, le type du jour(ferié ou pas), n'a pas d'impacte sur la concentration en Ozone, mais le lien avec la station est également logique, la concentration peut différer selon l'altitude ou l'endroit où les études ont été faites.

Ceci peut être utile pour l'application des modèles de régression dans la suite de l'exercice

3.a Modèle théorique : Soit Y la variable à prédire (ici O3obs) et $X_1, X_2, ..., X_p$ (ici les autres variables telles que MOCAGE, NO2, TEMPE...) les variables explicatives. La régression linéaire multiple modélise la relation entre Y et les X_i comme suit :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$$

Avec :

— β_0 est l'ordonnée à l'origine.
— β_i est le coefficient de régression de la variable X_i pour $i = 1, ..., p$, qui mesurent l'effet de chaque variable indépendante sur la variable dépendante, toutes les autres variables restant constantes. Un coefficient de régression positif indique une relation positive entre X_i et Y, tandis qu'un coefficient négatif indique une relation négative.
— ϵ est le terme d'erreur, non observés, indépendants et identiquement distribués ; $\mathbf{E}[\epsilon] = 0, \mathbf{V}[\epsilon] = \sigma$.

Hypothèses du modèle de régression linéaire multiple :
Pour que le modèle de régression linéaire multiple soit valide, plusieurs hypothèses doivent être satisfaites :

— Linéarité : La relation entre Y et les X_i est additive linéaire.
— Les β sont supposés constants.
— Homoscédasticité : La variance des résidus (erreurs) est constante pour toutes les valeurs des variables indépendantes.
— Indépendance des erreurs : Les erreurs ϵ sont indépendantes les unes des autres, si l'on calcule Y pour différents X_i.
— Normalité des erreurs : Les erreurs sont distribuées selon une loi normale.
— Absence de colinéarité : Les variables explicatives ne sont pas colinéaires.

3.b. On a considéré trois modèles de régressions linéaires multiples. Le premier modèle est une régression linéaire multiple avec toutes les variables.
On a considéré pour le deuxième modèle une régression en sélectionnant les 4 variables (MOCAGE, TEMPE, sqrt_RMH2O et log_NO2) qui semblaient influencer le plus la concentration en ozone. Ce deuxième modèle offre un ajustement légèrement inférieur, mais avec un nombre réduit de variables, ce qui peut être bénéfique pour la simplicité et la robustesse du modèle. Cependant, comme pour le modèle 1, la non-normalité potentielle des erreurs reste un point à adresser pour s'assurer de la validité des résultats.

On a fait de même pour un troisième modèle en considérant cette fois 2 variables (MOCAGE, TEMPE). Ce modèle simple à deux variables offre un compromis intéressant entre ajustement et complexité. Il est plus facile à interpréter et moins susceptible au surajustement que les modèles précédents. Cependant, la non-normalité potentielle des erreurs reste un point à considérer.
Comme la non-normalité des erreurs est confirmée, des techniques de transformation des données ou des modèles alternatifs (par exemple, des régressions plus robustes) peuvent être nécessaires pour obtenir des résultats plus fiables. En affichant la distributions des résidus ainsi que les QQ-plots en constate que les résidus ne sont pas gaussiens, les modèles de régressions choisis peuvent être améliorés dans ce cas. Reste alors à choisir le meilleurs modèle suivant certains critères. On propose ensuite une piste pour améliorer le modèle.
Conclusion : Le gain d'ajustement en passant du modèle 3 au modèle 2 est très faible (0.004), et ne justifie pas l'ajout de deux variables supplémentaires. Le gain d'ajustement en passant du modèle 2 au modèle 1 est également faible (0.034), et les variables supplémentaires du modèle 1 ne sont pas toutes significatives. Le modèle 3 est le plus simple et le plus facile à interpréter. En tenant compte de tous ces éléments, le modèle 3 semble être un bon choix. Il est le plus simple, tout en conservant un ajustement correct et en ne négligeant aucune variable significative.

4.
Random Forest : Les forêts aléatoires introduites par Leo BREIMAN au début des années 2000 sont une méthode de classification et de régression par apprentissage supervisé.

L'algorithme est basé sur l'assemblage d'arbres de décisions indépendantes. Chaque arbre dispose d'une vision parcellaire du problème du fait d'un tirage aléatoire :
— **Tree bagging :** Consiste à faire un tirage aléatoire avec remplacement sur observation.
— **Feature sampling :** consiste à faire un tirage aléatoire sur les variables.
Tous les arbres de décisions sont assemblés, la prédiction faite par RF pour les données inconnues est donc la moyenne.

Le défaut des arbres de décisions est que sa performence est sensible aux données de départ, l'ajout de nouvelles données peut modifier les résultats du modèles. Pour contrer ce problème, l'utilisation de

plusieurs arbres est idéale. l'idée est que plutôt d'avoir un estimateur global complexe qui fait tout, on utilise plusieurs estimateurs simples avec une vision parcellaire. L'assemblage de ces estimateurs rend la méthode Random Forest très performante.

Le principe du random forest a été présenté dans l'exercice 1. Une forêt aléatoire permet d'appliquer une régression. Reposant sur un système de bagging (voir exercice 1), le random forest de régression consiste schématiquement à calculer la moyenne des prévisions obtenues par l'ensemble des estimations des arbres décisionnels de la forêt aléatoires.

Réseau de neurones : Le principe général a été également présenté dans l'exercice 1. Dans le cas de la régression logistique par exemple, en reprenant l'exemple du preception, la fonction d'activation peut être par exemple la fonction logistique $\sigma = \frac{e^t}{1+e^t}$ et utiliser l'erreur quadratique de la régression (question 1).

Boosting : Dans le cas de la régression, on construit l'arbre en divisant la région grâce des critéres (choisis en fonction du type de la régression...) (cf exercice 1).

On a appliqué dans un premier temps random forest. On a eu $R^2 = 0.48$ et une erreur $MSE = 0.50$. En analysant le graphe de valeurs réelles par rapport aux valeurs prédites, puis en regardant le score des features (variables explicatives), on a pris en compte celles qui contribuaient le plus aux prédictions du modèle. Les variables les plus significatives étaient MOCAGE et TEMPE, comme à la question 3 et on a donc choisi d'éliminer les autres. La moyenne des scores obtenu après validation croisée est : 0,44.

Pour la méthode RN, on a une erreur $MSE = 0.5$. Après validation croisée, on a une moyenne des scores de -0,24.

On a choisi de faire un gradient boosting. On a obtenu une erreur $MSE = 0.47$. On a eu un score moyen de 0.37 à la validation croisée.

Pour les méthodes Random Forest, RN et Boosting, on a obtenu les (mean squared error) MSE respectifs suivants : 0.44, 0.5, 0.47 et une moyenne des scores suivants après validation croisée : 0.44, -0.24, 0.37.

Le modèle de Random Forest a la meilleure performance en termes de MSE et de scores de validation croisée. Il a une moyenne de scores de validation croisée plus élevée par rapport aux autres modèles, indiquant une meilleure généralisation sur de nouvelles données. Le modèle de Boosting a également de bonnes performances, mais légèrement inférieures à Random Forest. Le modèle de RN semble être le moins performant avec une MSE plus élevée et des scores de validation croisée négatifs, indiquant une mauvaise généralisation.

Conclusion :

Sur la base de ces résultats, le modèle de Random Forest semble être le choix optimal en raison de ses meilleures performances globales en termes de MSE et de scores de validation croisée. Il offre un bon équilibre entre précision et généralisation sur de nouvelles données.

5. Estimation de l'erreur de prédiction par validation croisée :

— Régression linéaire : 0.46
— Random Forest : 0.38
— Réseaux de Neurones : 0.36
— Boosting : 0.42

La meilleure méthode est Réseaux de Neurones avec une erreur de prédiction de 0.36.

5 a. L'erreur empirique avec des données de test pour la méthode Réseaux de Neurones : 0.46. Cette erreur a été calculée en faisant une $MeanSquarred$ des erreurs entre les sorties et les prédicions, sur le "meilleur modèle d'entrainement", c'est-à-dire le modèle qui a la petite erreur moyenne lors de la validation croisée.

5.b On a choisi de calculer la prévision pour la méthode que l'on a choisi réseau de neurones, mais aussi pour la méthode random forest et boosting. On trouve ces prévisions de concentration :

Individu 1 :

— Valeur réelle en ozone : -1.1322886704865944
— Prédiction Random Forest : -0.8279898702148195
— Prédiction Réseau de Neurones : -1.1324355923281249
— Prédiction Boosting : -1.0407200071753848

Individu 2
— Valeur réelle en ozone : -0.5222269217460684
— Prédiction Random Forest : -0.8123722894470631
— Prédiction Réseau de Neurones : -0.5988223188958702
— Prédiction Boosting : -0.714583385294772

Individu 3
— Valeur réelle en ozone : -1.0834837305873524
— Prédiction Random Forest : -0.5432130459027423
— Prédiction Réseau de Neurones : -0.7134953033038128
— Prédiction Boosting : -0.5032291868138989

Pour chaque individu, la prédiction qui a la valeur la plus proche de la valeur réelle a été obtenue par le réseau de neurones.

Globalement, les trois modèles semblent être capables de faire des prédictions assez proches des valeurs réelles en ozone pour les individus sélectionnés, bien que chacun puisse avoir ses forces et ses faiblesses, comme on a pu le voir dans d'autres questions, en fonction des données et des paramètres utilisés pour l'entraînement. Il faudrait effectuer davantage de tests sur plus d'individus pour pouvoir confirmer ces résultats.

Exercice 3

1. Il y avait tout d'abord environ 500 passagers en $3^{\text{ème}}$ classe, plus de 200 passagers en $1^{\text{ère}}$ classe et moins de 200 passagers en 2^{nd} classe. On remarque tout d'abord qu'il y a presque quatre fois plus de décès chez les passagers de $3^{\text{ème}}$ classe par rapport aux deux autres classes. En effet, il y a eu 372 décès chez les passagers de $3^{\text{ème}}$ classe alors qu'on observe 80 décès chez les passagers de $1^{\text{ère}}$ classe et 97 décès chez les passagers de 2^{nd} classe. On a presque un nombre égal de décès/survivants chez les passagers de 2^{nd} classe, avec 87 survivants. Presque le quart des passagers de $3^{\text{ème}}$ classe ont survécu (119 personnes) et enfin on observe le plus de survivants chez les passagers de $3^{\text{ème}}$ classe 136 survivants sur plus de 200 passagers.

On choisit pour les deux prochaines questions le modèle de régression logistique, car la variable dépendante Survivor est binaire et on cherche à modéliser la probabilité de succès en fonction des autres variables, qui sont explicatives.

2. a) On considère la variable binaire Survived (1 pour survécu, 0 pour non survécu) notée Y et la variable binaire Sexmale (0 pour féminin, 1 pour masculin) notée X. On peut s'attendre à ce qu'il y ait plus de femmes chez les survivants, car lors de ces situations d'urgences, la procédure consiste à sauver en priorité les femmes et les enfants.

2. b) On peut considérer un modèle de régression logistique qui prédit la probabilité de survie des passagers selon leur sexe. Le modèle théorique pourrait donc être formulé comme suit :

$$P(Y = 1|X) = \frac{\exp(\beta_0 + \beta_1 \cdot X)}{1 + \exp(\beta_0 + \beta_1 \cdot X)}$$

où β_0 et β_1 sont les coefficients à estimer. Grâce aux codes du sujet, on lit $\beta_0 = 1.06$ et $\beta_1 = -2.51$. Ce qui donne donc

$$P(Y = 1|X) = \frac{\exp(1.06 - 2.51 \cdot X)}{1 + \exp(1.06 - 2.51 \cdot X)}$$

En classification binaire avec perte 0-1, le prédicteur est

$$f^*(X) = \begin{cases} 1 & \text{si } P(Y = 1|X) \geq \frac{1}{2} \\ -1 & \text{sinon} \end{cases}$$

Si X = 1, c'est-à-dire si l'on considère un homme, on calcule $P(Y = 1|X = 1) = \frac{\exp(1.06-2.51)}{1+\exp(1.06-2.51)} = 0.19 \geq \frac{1}{2}$ donc $f^*(X) = 0$, autrement dit, on prédit la mort de l'homme.

Si X = 0, c'est-à-dire si l'on considère une femme, on calcule $P(Y = 1|X = 0) = \frac{\exp(1.06)}{1+\exp(1.06)} = 0.74 \geq \frac{1}{2}$ donc $f^*(X) = 1$, autrement dit, on prédit la survie de la femme.

2. c) Le coefficient associé à Sexmale est négatif (-2.513710), ce qui indique que les passagers masculins ont une probabilité de survie plus faible que les passagers féminins. Cela suggère donc que le sexe a un effet significatif sur les chances de survie, les passagers masculins étant moins susceptibles de survivre que les passagers féminins, ce qui est cohérent avec la procédure à suivre lors de telles catastrophes.

3. a) En retirant la variable "Fare" du modèle, on suppose que son effet sur la survie des passagers est négligeable, ce qui peut être une hypothèse réaliste, car la survie des passagers ne dépend pas à priori du prix dépensé pour le billet. Néanmoins, comme on l'a vu à la question 1, la classe des passagers est quant à elle à prendre en compte. Pour les mêmes arguments qu'à la question 2, le sexe ainsi que l'âge sont à prendre en compte
On reprend la même méthode qu'à la question 2. On note cette fois la variable aléatoire binaire X_S qui vaut 0 si le passager n'est pas un homme, 1 sinon, on note X_A la variable aléatoire décrivant l'âge et X_{C2} et X_{C3} la variable aléatoire binaire qui décrit la classe (2 ou 3) du passager. Le modèle théorique est donc

$$P(Y = y|X = x) = \frac{\exp(\beta_0 + \beta_1 \cdot x_S + \beta_2 \cdot x_{C2} + \beta_3 \cdot x_{C3} + \beta_4 \cdot x_A)}{1 + \exp(\beta_0 + \beta_1 \cdot x_S + \beta_2 \cdot x_{C2} + \beta_3 \cdot x_{C3} + \beta_4 \cdot x_A))}$$

où $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ sont les coefficients à estimer.
Grâce aux codes fournis, on lit $\beta_0 = 3.54, \beta_1 = -2.61, \beta_2 = -1.12, \beta_3 = -2.33, \beta_4 = -0.03$.
Finalement,

$$P(Y = y|X = x) = \frac{\exp(3.54 - 2.61 \cdot x_S - 1.12 \cdot x_{C2} - 2.33 \cdot x_{C3} - 0.03 \cdot x_A)}{1 + \exp(3.54 - 2.61 \cdot x_S - 1.12 \cdot x_{C2} - 2.33 \cdot x_{C3} - 0.03 \cdot x_A)}$$

Comme précédemment, le prédicteur est :

$$f^*(X) = \begin{cases} 1 & \text{si } P(Y = 1|X) \geq \frac{1}{2} \\ -1 & \text{sinon} \end{cases}$$

3. b) Le coefficient associé à la variable Age est -0.03330. Ce coefficient indique que les passagers plus âgés ont tendance à avoir des chances de survie moins élevées que les passagers plus jeunes. De plus, la p-value associé à la variable Age est très faible (< 0.05), ce qui signifie que l'âge a un effet significatif sur les chances de survie des passagers, même après avoir pris en compte les autres variables du modèle. Ceci confirme que dans ce cas de situation, la survie des enfants est privilégiée.

3. c) Pour le passager 1, le modèle prédit qu'il va survivre. Il s'agit d'une fille de 5 ans, de 3$^{\text{ème}}$ classe, sa probabilité de survie est de 0.741. Le modèle ne s'est pas trompé, car, cet enfant a survécu, ce qui confirme les résultats de la question 3 b).

Pour le passager 2, il s'agit d'une femme de 20 ans, de 3$^{\text{ème}}$ classe, sa probabilité de survie est de 0.634. Le modèle a prédit sa mort, sûrement à cause de sa classe, elle a cependant survécu.

Enfin le passager 3 est un homme de 50 ans, de 2$^{\text{nd}}$ classe, sa probabilité de survie est de 0.135. Le modèle a prédit sa mort, ce qui a en effet été le cas, sûrement à cause de son genre et de son âge.

Pour le passager 1, on calcule :

$$P(Y = y|X = x) = \frac{\exp(3.54 - 2.33 - 0.03 \cdot 5)}{1 + \exp(3.54 - 2.33 - 0.03 \cdot 5)} = 0.742$$

ce qui est cohérent avec le résultat de l'énoncé qui est de 0.741 (au centième près, car on a pris des valeurs approchées des β).

Bibliographie

- Cours de Madame Fermin

- Exercice 1 :

— https ://aws.amazon.com/fr/what-is/boosting/
— https ://blent.ai/blog/a/random-forest-comment-ca-marche
— https ://ryax.tech/fr/deep-learning-comprendre-les-reseaux-de-neurones-artificiels-artificial-neural-networks/
— https ://www.journaldunet.fr/intelligence-artificielle/guide-de-l-intelligence-artificielle/1501905-random-forest-ou-foret-aleatoire/

- Exercice 2 :

— https ://hal.science/hal-03049016/file/MLTP.pdf
— https ://www.iro.umontreal.ca/ slacoste/teaching/apprentissage-fall2014/notes/cours1.pdf
— https ://www.math.univ-toulouse.fr/ besse/Wikistat/pdf/st-l-inf-intRegmult.pdf
— https ://songrcs.medium.com/versioning-your-dataset-and-models-using-modeldb-10b0ee3873ed
— https ://perso.lpsm.paris/ liautaud/projects/KeRF.pdf
— https ://www.journaldunet.fr/intelligence-artificielle/guide-de-l-intelligence-artificielle/1501905-random-forest-ou-foret-aleatoire