Compétition Kaggle “Spaceship Titanic”

Jonathan Duckes

Sommaire

1. Contexte du Projet
2. Gestion de Projet
3. Données Utilisées
4. Présentation du dataset
5. Exploration des Données
6. Processus de Préparation des Données
7. Analyse des Données
8. Modélisation
9. Résultats et Discussion
10. Soumission Kaggle
11. Difficultés Rencontrées
12. Conclusion
13. Perspectives Futures
14. Questions et Réponses
15. Références

Contexte du Projet

01

DESCRIPTION DE L'UNIVERS FICTIF DU SPACESHIP TITANIC:

- Voyage vers TRAPPIST-1e
- Transport dimensionnel des passagers suite à un accident

02

OBJECTIF DE LA COMPÉTITION KAGGLE:

- Prédire le transport des passagers
- Analyse des données passagers

03

IMPORTANCE ET PERTINENCE DU PROJET:

- Application techniques Data Science
- Compétences en :
  - Exploration de données
  - Préparation des données
  - Modélisation

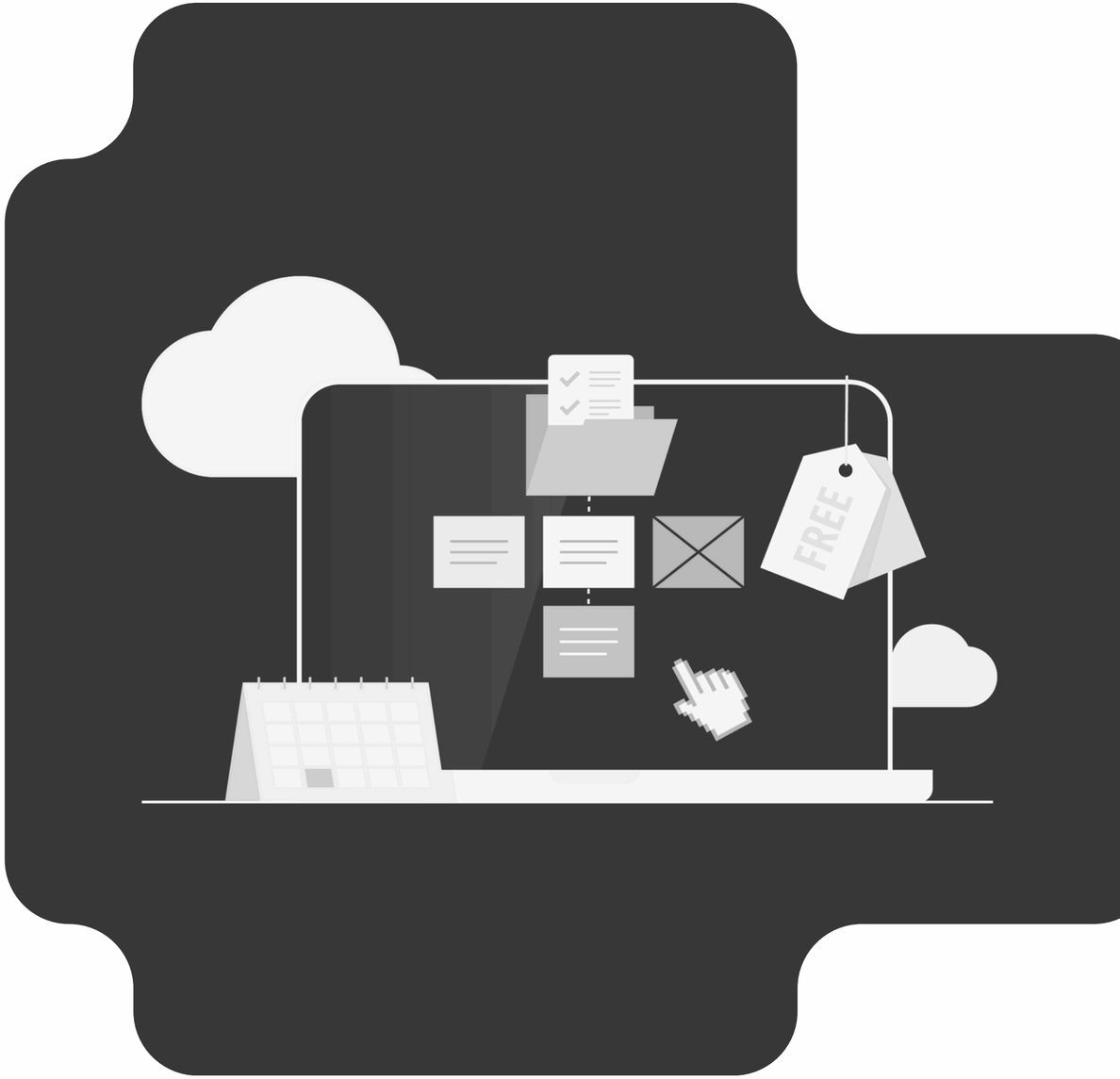
Gestion de projet

Organisation du Travail :

Répartition des tâches: Microsoft To Do

Documentation :

Lecture de cours, documentation d’outils utilisés



Données utilisées

PRÉSENTATION DU DATASET:

- Nombre d'entrées: Train : 8693 Test : 4277
- Nombre de colonnes: 14

	PassengerId	HomePlanet	CryoSleep	Cabin	Destination	Age	VIP	RoomService	FoodCourt	ShoppingMall	Spa	VRDeck	Name	Transported
0	0001_01	Europa	False	B/0/P	TRAPPIST-1e	39.0	False	0.0	0.0	0.0	0.0	0.0	Maham Ofracculy	False
1	0002_01	Earth	False	F/0/S	TRAPPIST-1e	24.0	False	109.0	9.0	25.0	549.0	44.0	Juanna Vines	True
2	0003_01	Europa	False	A/0/S	TRAPPIST-1e	58.0	True	43.0	3576.0	0.0	6715.0	49.0	Altark Susent	False
3	0003_02	Europa	False	A/0/S	TRAPPIST-1e	33.0	False	0.0	1283.0	371.0	3329.0	193.0	Solam Susent	False
4	0004_01	Earth	False	F/1/S	TRAPPIST-1e	16.0	False	303.0	70.0	151.0	565.0	2.0	Willy Santantines	True

Exploration des données

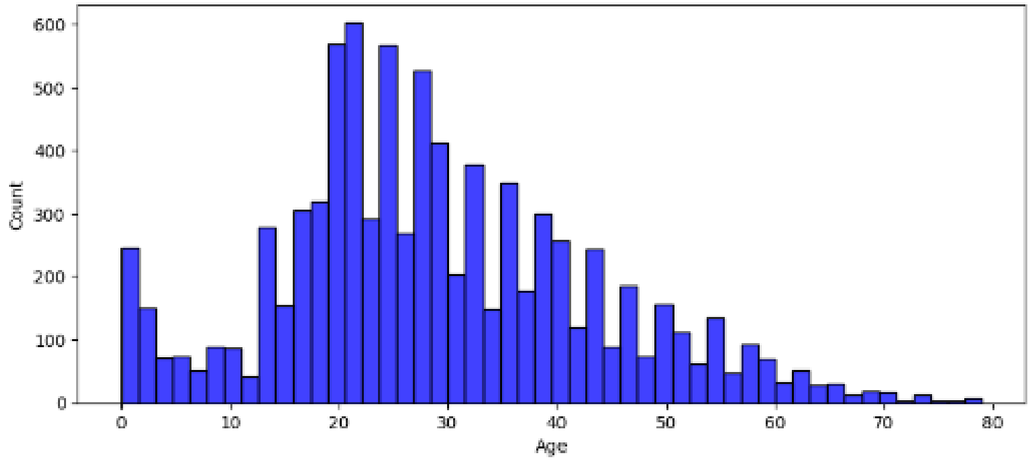
STATISTIQUES ET INFORMATIONS

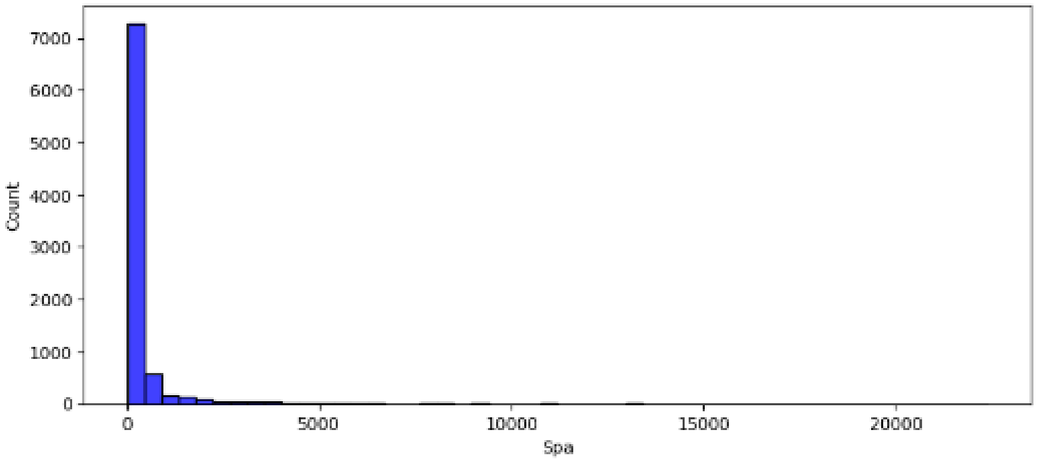
	Age	RoomService	FoodCourt
count	8514.000000	8512.000000	8510.000000
mean	28.827930	224.687617	458.077203
std	14.489021	666.717663	1611.489240
min	0.000000	0.000000	0.000000
25%	19.000000	0.000000	0.000000
50%	27.000000	0.000000	0.000000
75%	38.000000	47.000000	76.000000
max	79.000000	14327.000000	29813.000000

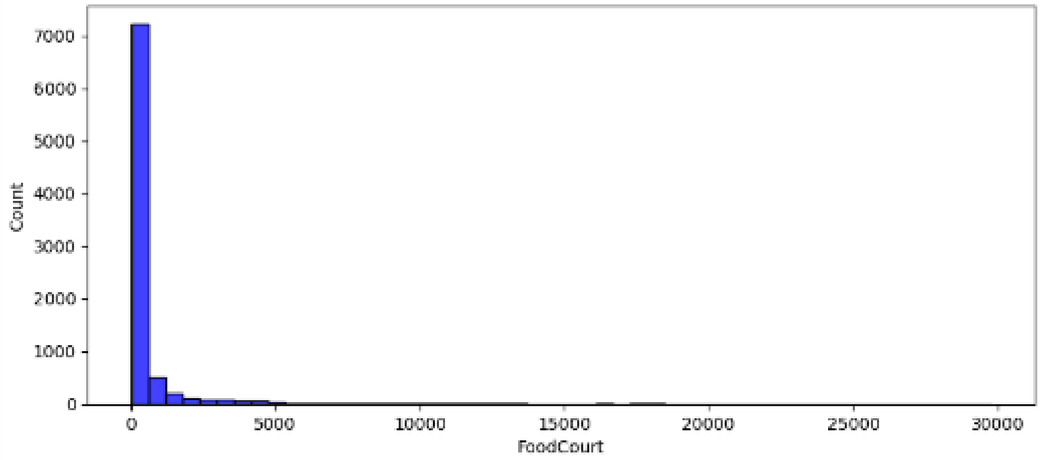
	ShoppingMall	Spa	VRDeck
count	8485.000000	8510.000000	8505.000000
mean	173.729169	311.138778	304.854791
std	604.696458	1136.705535	1145.717189
min	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000
75%	27.000000	59.000000	46.000000
max	23492.000000	22408.000000	24133.000000

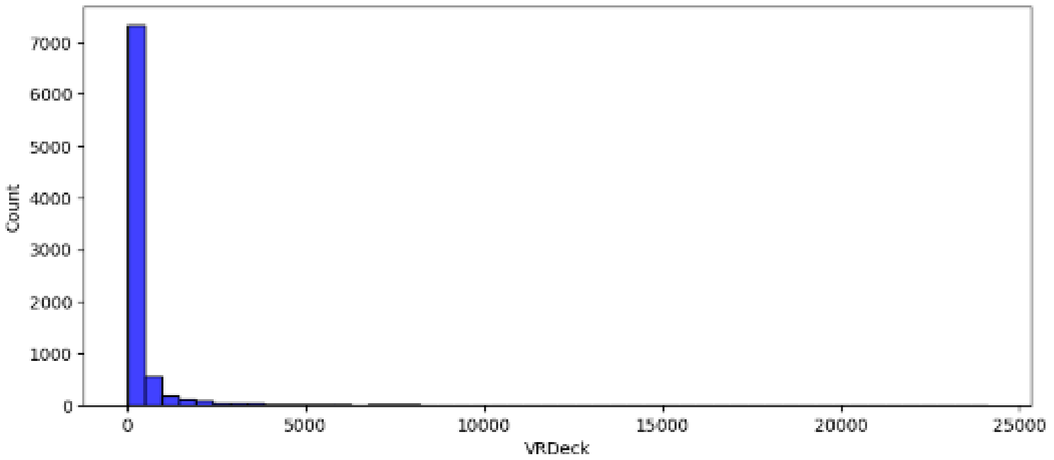
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8693 entries, 0 to 8692
Data columns (total 14 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   PassengerId   8693 non-null   object
 1   HomePlanet    8492 non-null   object
 2   CryoSleep     8476 non-null   object
 3   Cabin         8494 non-null   object
 4   Destination   8511 non-null   object
 5   Age           8514 non-null   float64
 6   VIP           8490 non-null   object
 7   RoomService   8512 non-null   float64
 8   FoodCourt     8510 non-null   float64
 9   ShoppingMall  8485 non-null   float64
 10  Spa           8510 non-null   float64
 11  VRDeck        8505 non-null   float64
 12  Name          8493 non-null   object
 13  Transported   8693 non-null   bool
dtypes: bool(1), float64(6), object(7)
memory usage: 891.5+ KB
```

DISTRIBUTION: VARIABLES NUMÉRIQUES

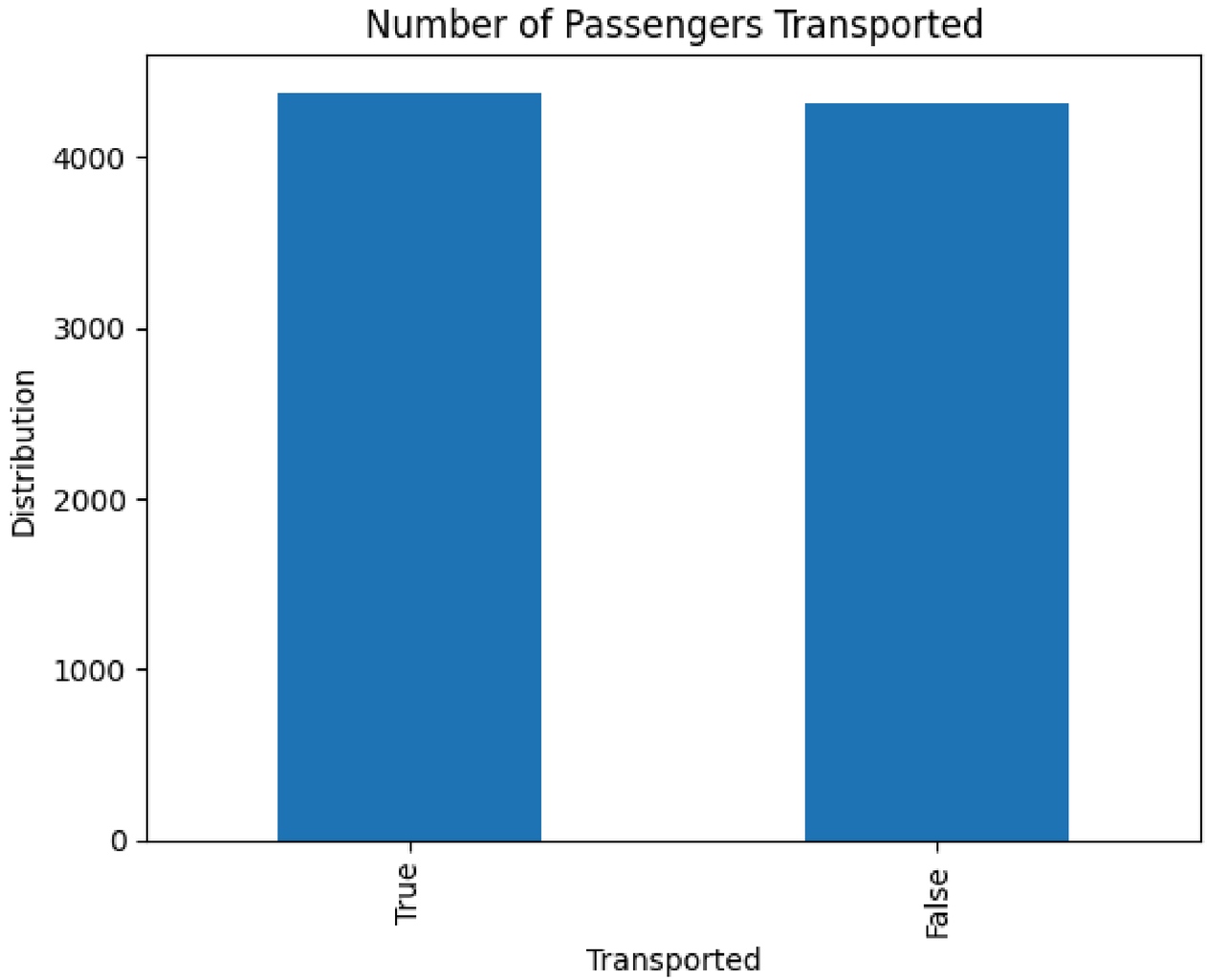




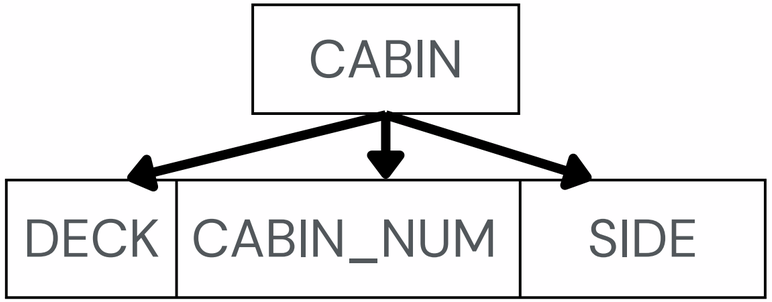




Distribution: Transported



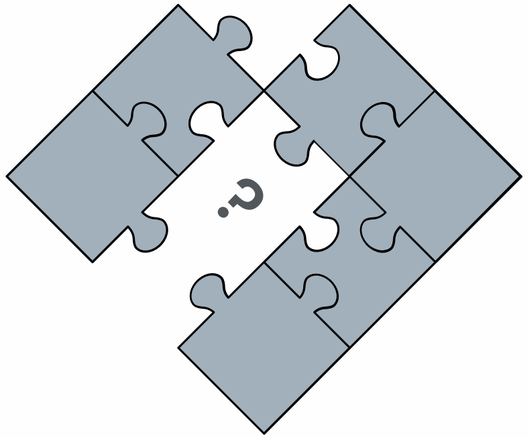
PROCESSUS PRÉPARATION DES DONNÉES

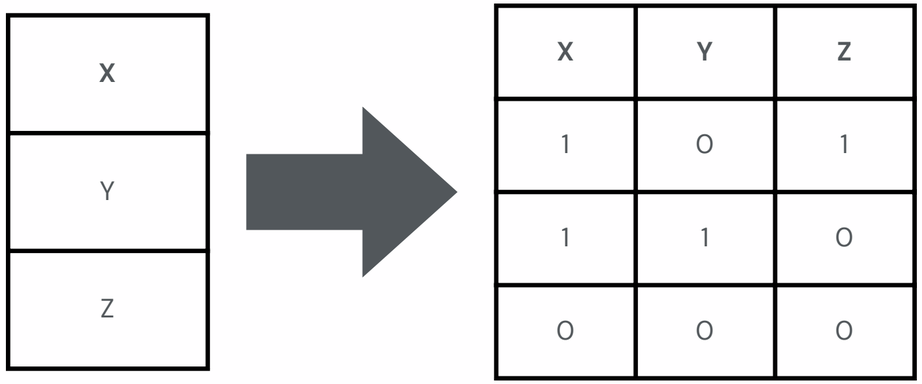


Séparation de Cabin en Deck, Cabin_num, Side.

Remplacement des valeurs manquantes

- Cabin_num : la valeur la plus courante
- Age : la médiane
- Variables de dépenses : zéro
- Colonnes catégorielles : le mode



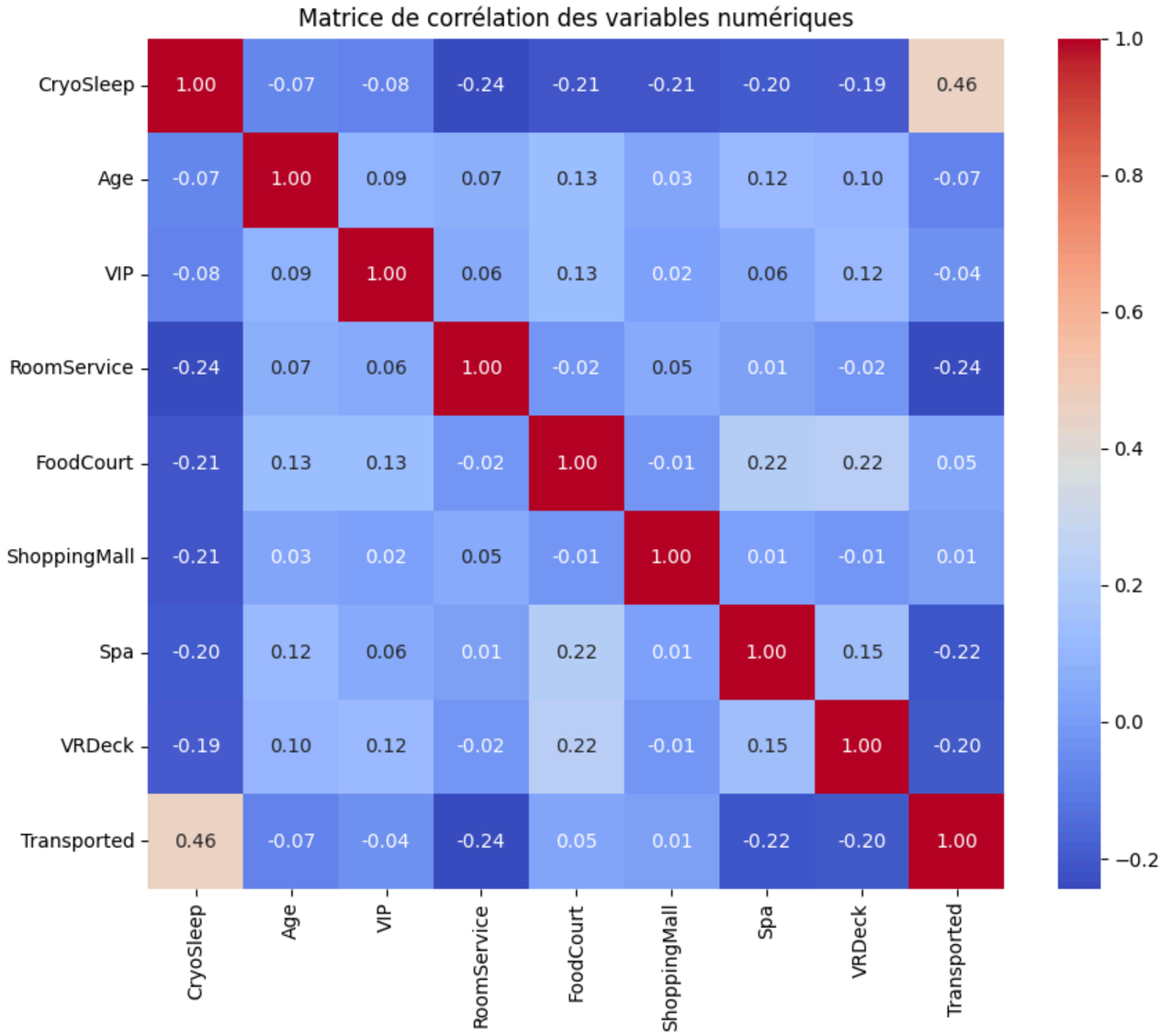


Encodage des variables catégorielles

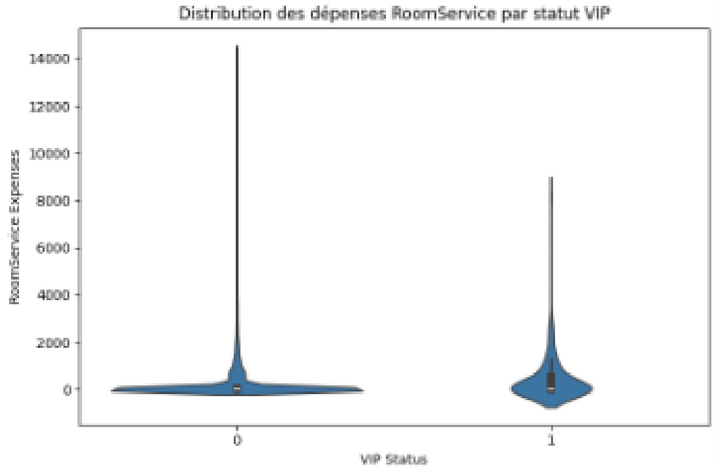
ANALYSE DES DONNÉES

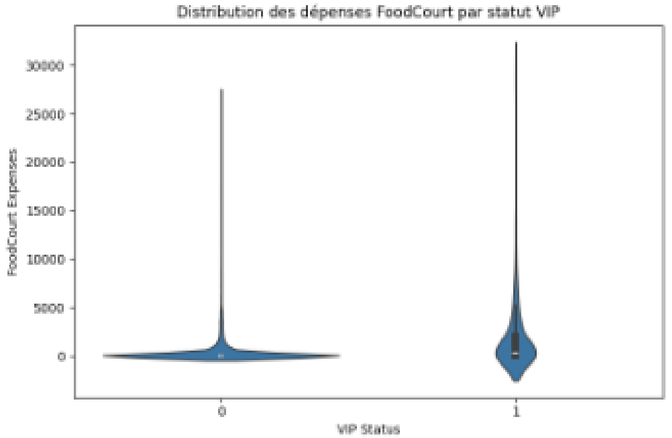
MATRICE DE CORRÉLATION

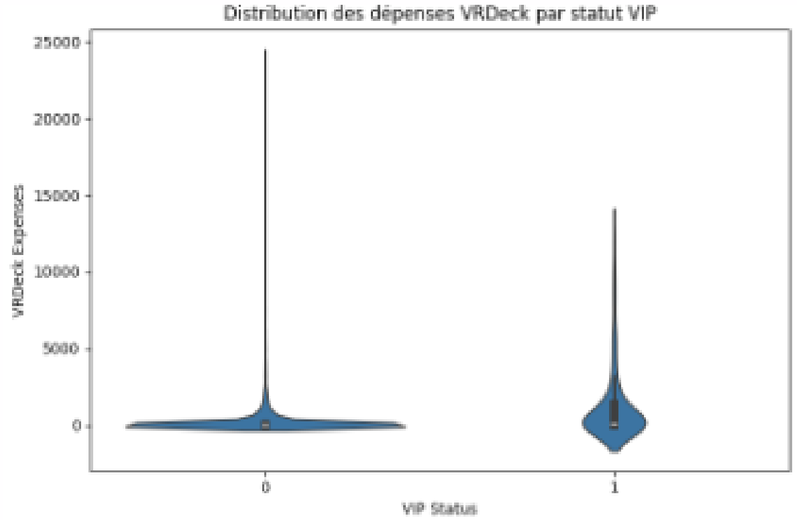
- CRYOSLEEP ET TRANSPORTED : CORRÉLATION MODÉRÉE POSITIVE (0.46)
- VIP ET AUTRES DÉPENSES : FAIBLES CORRÉLATIONS NÉGATIVES

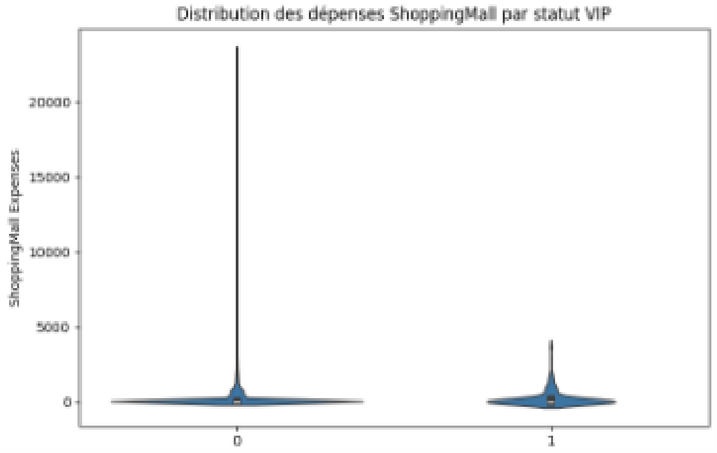


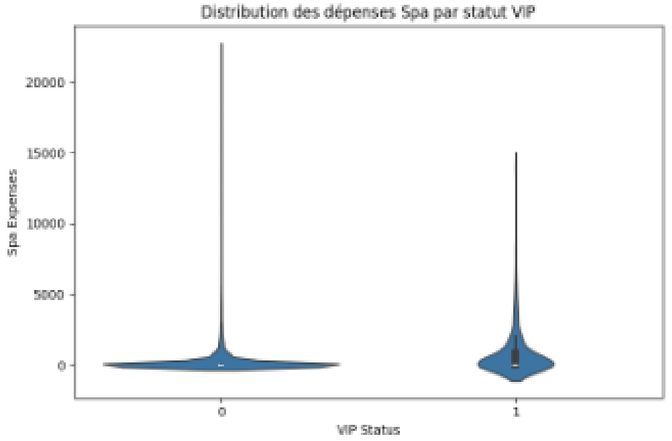
DISTRIBUTION DES DÉPENSES EN FONCTION DU STATUT VIP





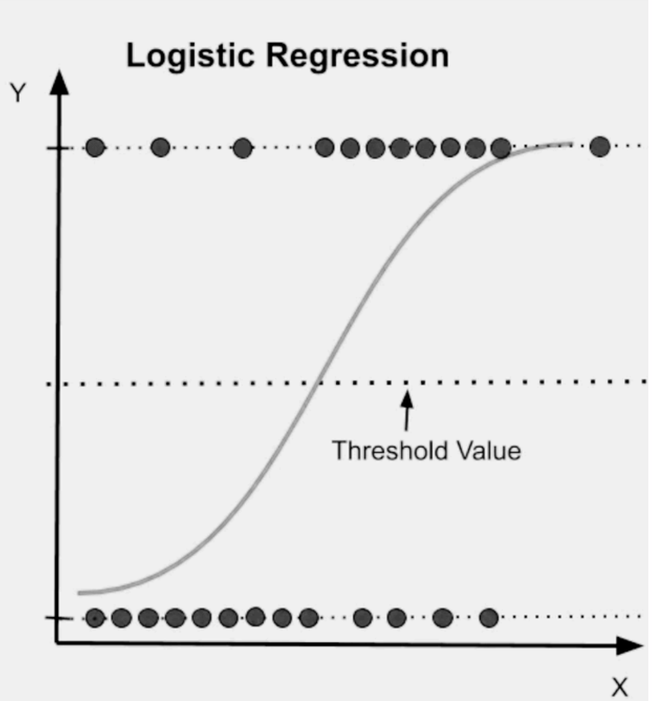






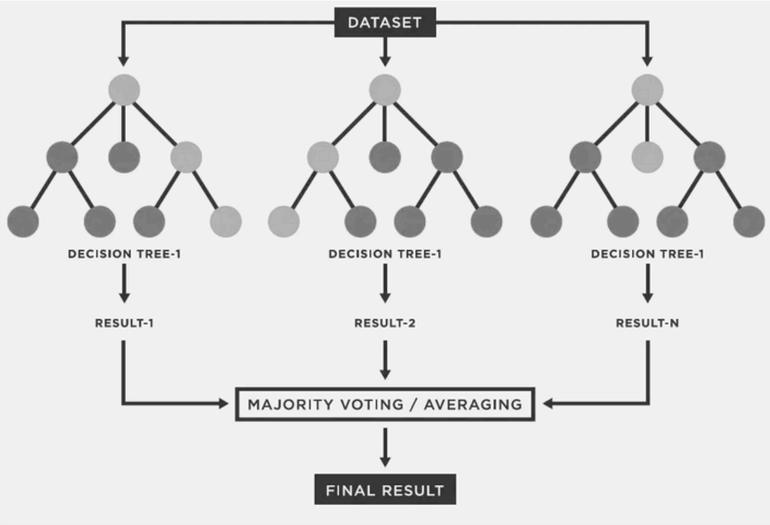
MODÉLISATION

CHOIX DES MODÈLES:



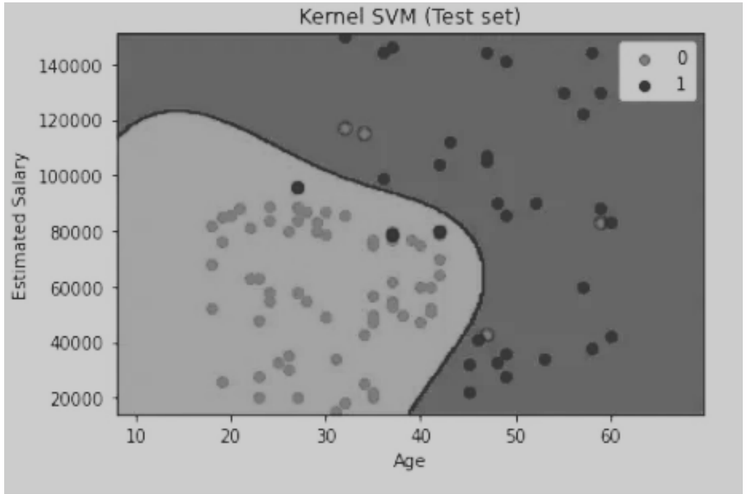
Régression Logistique:

- Modèle de classification binaire
- Prédit des probabilités
- Interprétabilité



Forêt Aléatoire (Random Forest):
Corrige le surapprentissage

Robustes aux valeurs aberrantes



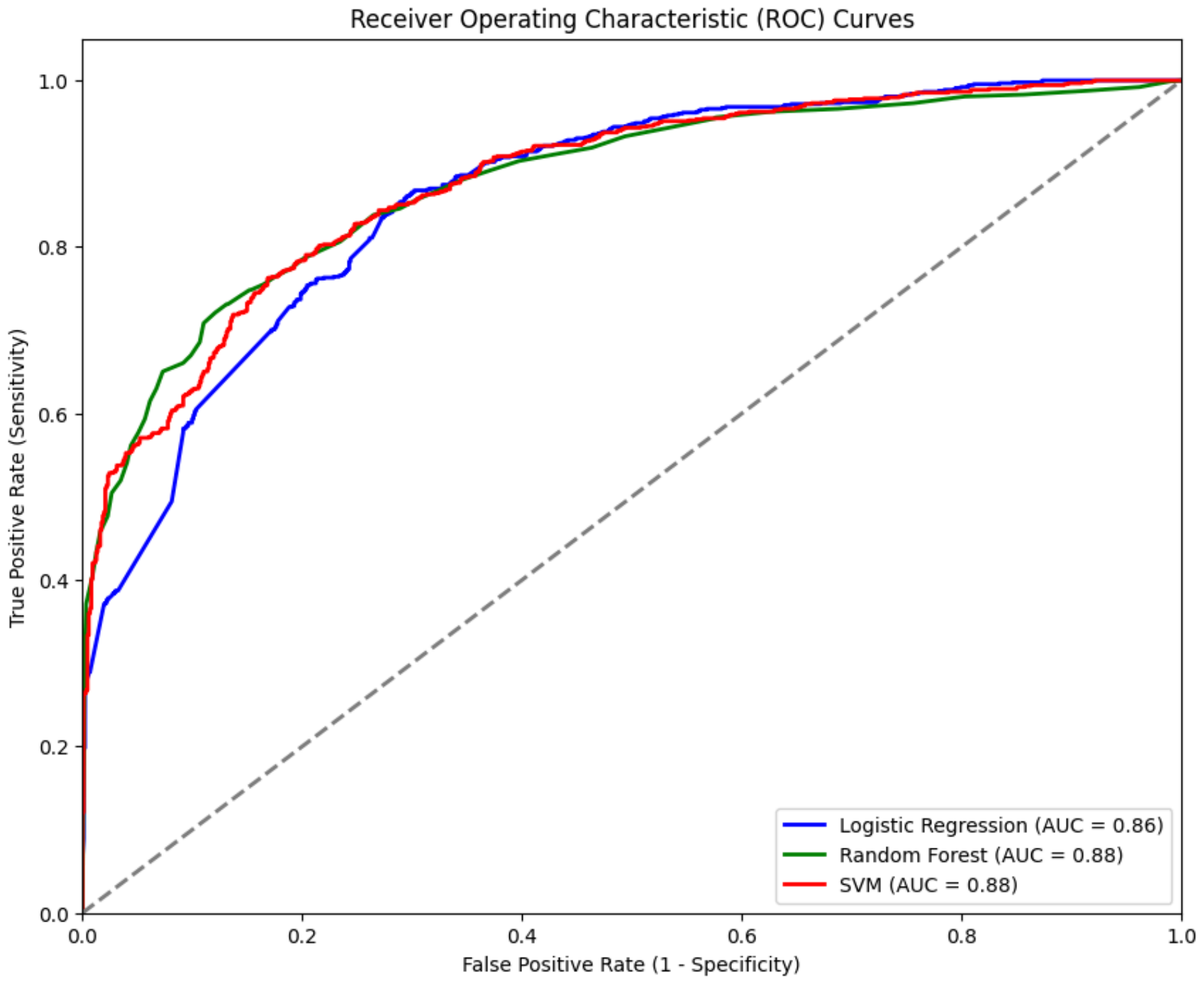
Support Vector Machine (SVM) RBF:
Maximisation de la marge

Relations non-linéaires

RÉSULTATS ET DISCUSSION

MÉTRIQUES

		Regression logistiques		Random Forest		SVM	
		Valeurs observées		Valeurs observées		Valeurs observées	
Matrice de confusion	Valeurs prédites	676	185	730	131	684	177
		209	669	221	657	184	694
Taux de bonne classification		0,773		0,798		0,792	
Rappel(Sensibilité)		0,762		0,748		0,790	
Specifité		0,785		0,848		0,794	



COURBE ROC ET AUC

- Régression logistique: 0.862
- Random Forest: 0.877
- SVM: 0.878

SOUMISSION KAGGLE

- Prétraitement des Données de Test
- Utilisation du modèle SVM
- Création du fichier de soumission au format requis par Kaggle.
- Résultats :

881	**JGelo**		0.79939	1	16d

DIFFICULTÉS RENCONTRÉES

- Évaluation des modèles
  - Implémentation des formules
  - Compréhension des résultats obtenues
- Planification et organisation
  - Surestimation du temps mise à disposition
  - Manque d’intensité de travail
- Adaptation et flexibilité
  - Ajustement des méthodes de cours

CONCLUSION

Exploration et Préparation des Données

Implémentation et évaluation de plusieurs modèles

Évaluation des Performances

Soumission Kaggle

PERSPECTIVES FUTURES

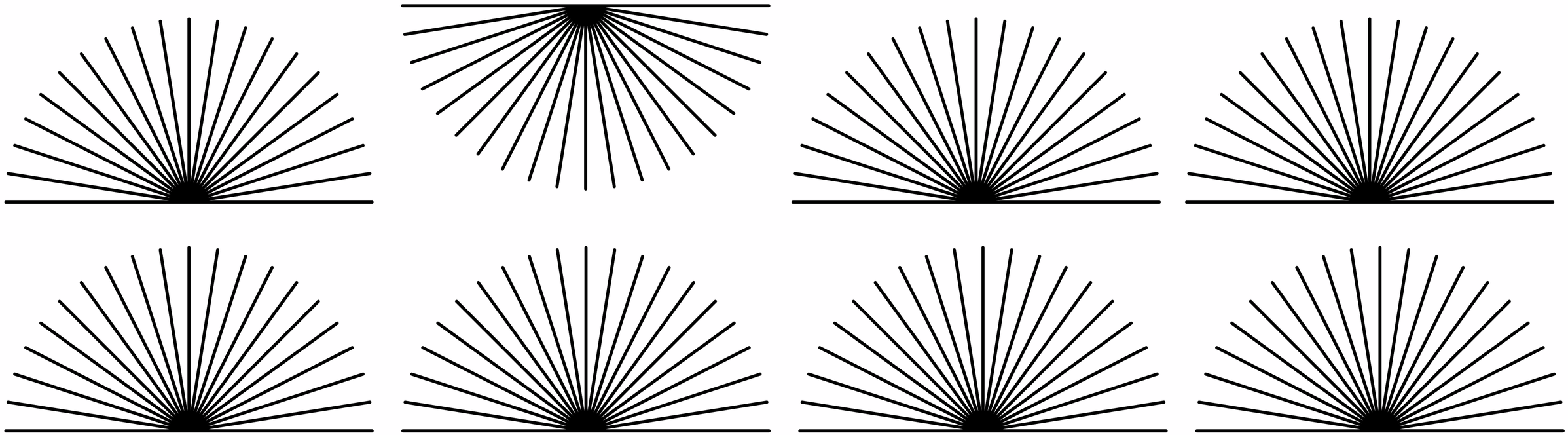
Amélioration des Modèles

Analyse des Erreurs

Création de Nouvelles Caractéristiques

Développement Professionnel

MERCI POUR VOTRE ATTENTION !



QUESTIONS ET RÉPONSES

RÉFÉRENCES

- Kaggle. Kaggle : Your machine learning and data science community. https://www.kaggle.com, 2010.
- M. Servajean S. Lèbre, A. Sallaberry. Rentrée m1 miashs septembre 2023, 2023.
- Ryan Holbrook Addison Howard, Ashley Chow. Spaceship titanic, 2022.
- C. Trottier and M. Amico. Régression logistique et modèles log-linéaires, 2024. Cours de Master MIASHS.
- Sophie Lèbre Marine Demangeot. Classification supervisée et non supervisée. Application d'éléments du cours.
- Project Jupyter. Jupyter notebook, 2014.
- Microsoft Corporation. Microsoft to do, 2017.
- Gusthema. Spaceship titanic with tfdf, 2023.
- Wikipedia contributors. Pearson correlation coefficient.
- Maximilien Servajean. Random forest tp, 2023.
- Wes McKinney. Pandas.
- Travis E. Oliphant. Numpy.
- David Cournapeau. Scikit-learn.
- Michael Droettboom John Hunter. Matplotlib.
- Michael Waskom. Seaborn.
- eliot robot Gusthema. spaceship-titanic-with-tfdf.ipynb, 2022.
- Gwenaël Richomme Catherine Trottier Sandra Bringay, Sophie Lèbre. Introduction à la science des données. Application d'éléments du cours.
- Pierre Lafaye De Micheaux. Analyse de données multidimensionnelles. Application d'éléments du cours.
- Forum : Kaggle competition spaceship titanic. Forum de discussion.
- Stackoverflow. Pour l'explication d'erreurs, debugging, et propositions de solutions de code .