

Préface

Introduction générale

Avant-propos

1. Pourquoi ce livre ?	23
2. À qui s'adresse ce livre ?	25
3. Comment est organisé ce livre ?	27
4. Comment lire ce livre ?	28
5. Quels sont les prérequis pour la lecture de ce livre ?	29
6. Qui est l'auteur ?	29
7. Remerciements	31

Partie 1 : La Data Science - Concepts généraux

Chapitre 1-1

La Data Science

1. Objectif du chapitre	33
2. L'objectif recherché en Machine Learning	34
3. Une expérimentation Machine Learning	38
3.1 Types de données	46
3.2 Préparation des données	48
4. Cycle de vie d'un projet Data Science	51
5. Les algorithmes du Machine Learning	54
6. Le problème de surapprentissage	56
7. Les paramètres et les hyperparamètres	57
8. Validation croisée	59

2 ————— Le Machine Learning

et l'IA générative avec Python

9. Données d'entraînement, de validation et de test	63
10. Métriques de performance	64
10.1 Métriques pour les problèmes de régression	66
10.2 Métriques pour la classification	69
10.2.1 Matrice de confusion binaire	69
10.2.2 Matrice de confusion générale	71
10.2.3 Exemple de matrice de confusion	72
10.2.4 La courbe ROC	74
10.3 Métriques pour le clustering	75
11. Conclusion	75

Partie 2 : Outils techniques de la Data Science - Python, Numpy, Pandas et Jupyter

Chapitre 2-1

Le langage Python

1. Objectif du chapitre	77
2. Python en deux mots	78
3. Installer l'interpréteur Python	78
4. Les bases de la programmation Python	81
4.1 Hello world avec Python	81
4.1.1 La fonction print	81
4.1.2 La fonction input	85
4.2 Les structures de données	85
4.2.1 Les variables numériques	86
4.2.2 Les chaînes de caractères	91
4.2.3 Le type booléen	93
4.2.4 Les listes	99
4.2.5 Les tuples	102
4.2.6 Les dictionnaires	103
4.2.7 Les ensembles	105

4.2.8	Liste vs tuple vs dictionnaire vs ensemble	109
4.3	Structurer un code Python.	110
4.3.1	L'indentation et les blocs de code	110
4.3.2	Écrire une instruction sur plusieurs lignes	111
4.3.3	Écrire plusieurs instructions sur une ligne	113
4.3.4	Les commentaires en Python.	113
4.4	Les instructions conditionnelles	113
4.4.1	Les conditions avec la structure if	114
4.4.2	Les conditions avec la structure if-else	115
4.4.3	Les conditions avec la structure if-elif-else	116
4.5	Les boucles.	119
4.5.1	La boucle for	119
4.5.2	La boucle for et la fonction zip	123
4.5.3	La boucle while	129
4.5.4	Contrôler les boucles avec break	131
4.5.5	Contrôler les boucles avec continue	132
4.6	Les fonctions.	133
4.6.1	Définir et utiliser une fonction sans paramètre	134
4.6.2	Les fonctions avec paramètres	136
4.6.3	Les valeurs par défaut des paramètres.	137
4.6.4	Renvoi de résultats	141
4.6.5	La portée des variables	142
4.6.6	Passage d'arguments à une fonction	145
4.6.7	Les fonctions récursives	148
4.7	Les listes en compréhension.	153
4.7.1	Les origines des listes en compréhension	153
4.7.2	Construire une liste avec les listes en compréhension . .	154
4.7.3	Application de fonction avec une liste en compréhension.	155
4.7.4	Utiliser if-else avec les listes en compréhension	156
4.7.5	Filtrer avec les listes en compréhension	157

4 — Le Machine Learning

et l'IA générative avec Python

4.8	Les expressions régulières	157
4.8.1	Regex sans caractères spéciaux	160
4.8.2	Regex avec caractères spéciaux	162
4.8.3	Regex avec les multiplicateurs	164
4.8.4	Regex avec un nombre d'occurrences limité	167
4.8.5	Regex avec groupage des résultats	168
4.8.6	Taille des motifs	169
4.8.7	Aller plus loin avec les expressions régulières	174
4.9	Gestion des exceptions	175
4.9.1	La levée des exceptions	175
4.9.2	Utiliser le bloc try-except	177
4.9.3	Gérer plusieurs exceptions	178
4.9.4	Utiliser la clause finally	180
4.9.5	Utiliser la structure try-except-finally-else	181
4.9.6	Lever une exception avec raise	183
5.	Conclusion	185

Chapitre 2-2

La bibliothèque NumPy

1.	Objectif du chapitre	187
2.	NumPy en deux mots	188
3.	Les tableaux NumPy	188
3.1	Création de tableaux NumPy	188
3.1.1	Créer un tableau à une dimension	189
3.1.2	Créer un tableau à plusieurs dimensions	189
3.2	Les dimensions d'un tableau NumPy	191
3.3	Le type et la taille d'un tableau NumPy	193
3.4	Fonction d'initialisation de tableaux NumPy	195
4.	Accéder aux données d'un tableau NumPy	197
4.1	Accès aux données d'un tableau à une dimension	197
4.2	Accès aux données d'un tableau à deux dimensions	199

4.3	Accès aux données d'un tableau à trois dimensions	201
5.	Modifier les données d'un tableau NumPy	202
6.	Copier un tableau NumPy dans un autre tableau NumPy	203
7.	Algèbre linéaire avec NumPy	205
7.1	Opérations mathématiques de base avec NumPy	205
7.2	Opérations sur les matrices avec NumPy	207
8.	Réorganiser des tableaux NumPy	208
8.1	Restructurer un tableau NumPy	208
8.2	Superposer des tableaux NumPy	210
9.	Statistiques descriptives avec NumPy	211
10.	Lire des données NumPy à partir d'un fichier	213
11.	Les masques booléens avec NumPy	214
11.1	Créer et utiliser un masque booléen	214
11.2	Un masque avec plusieurs conditions	216
11.3	Les fonctions numpy.any et numpy.all	217
12.	Tableaux NumPy versus listes Python	219
12.1	Comparaison des tailles en mémoire	220
12.2	Comparaison des temps de calcul	221
12.2.1	Temps de calcul sur une liste Python	222
12.2.2	Temps de calcul sur un tableau NumPy	223
13.	Conclusion	224

Chapitre 2-3

La bibliothèque Pandas

1.	Objectif du chapitre	225
2.	C'est quoi, Pandas ?	226
3.	Installation de Pandas	227

6 — Le Machine Learning

et l'IA générative avec Python

4. DataFrame Pandas	228
4.1 Création d'un DataFrame à partir d'un dictionnaire	229
4.2 Création d'un DataFrame à partir d'un tableau NumPy	231
4.3 Chargement des données à partir de fichiers	232
4.3.1 Lecture des données d'un fichier CSV	233
4.3.2 Lecture d'un fichier texte	234
5. Accès aux données d'un DataFrame	235
5.1 Lire les lignes d'un DataFrame	236
5.1.1 Lire une ligne d'un DataFrame	236
5.1.2 Lire plusieurs lignes d'un DataFrame	236
5.1.3 Parcourir les lignes d'un DataFrame	237
5.1.4 Filtrer les lignes avec une condition	238
5.1.5 Filtrer les lignes avec plusieurs conditions	238
5.1.6 Filtrage avec des critères textuels	239
5.1.7 Réinitialiser les index	240
5.1.8 Filtrer avec les valeurs uniques	242
5.1.9 Filtrer avec une expression régulière	242
5.2 Accéder aux variables d'un DataFrame	243
5.2.1 Liste des variables d'un DataFrame	243
5.2.2 Accès aux valeurs d'une colonne	244
5.2.3 Accès à plusieurs colonnes	244
5.3 Lire une cellule spécifique avec les index	245
6. Modifier un DataFrame	245
6.1 Modifier les valeurs dans un DataFrame	245
6.2 Modifier la structure d'un DataFrame	246
6.2.1 Ajouter une variable à un DataFrame	246
6.2.2 Réordonner les variables d'un DataFrame	249
6.2.3 Supprimer une variable d'un DataFrame	250
6.2.4 Utiliser la méthode melt pour diminuer le nombre de variables	251
6.3 Appliquer une fonction sur une variable avec la méthode apply	253
6.4 Modification avec conditions	255

6.5	Ajouter des lignes dans un DataFrame	256
7.	Tri sur les données d'un DataFrame	257
7.1	Tri avec un seul critère	257
7.2	Tri avec plusieurs critères.	259
8.	Sauvegarder les données d'un DataFrame.	260
9.	Faire des statistiques sur un DataFrame	261
9.1	Faire un résumé direct	261
9.2	Faire un résumé par agrégation	262
9.3	Agrégation avec plusieurs paramètres.	264
10.	Lecture des fichiers de grande taille.	265
11.	Conclusion	267

Chapitre 2-4

Travailler avec Jupyter

1.	Objectif du chapitre	269
2.	Installation de l'environnement Anaconda et Jupyter.	270
3.	Travailler avec Jupyter	276
3.1	Les documents dans Jupyter	278
3.1.1	Créer un dossier	279
3.1.2	Renommer un dossier.	280
3.1.3	Déplacer un dossier.	281
3.1.4	Charger des documents	282
3.1.5	Supprimer des éléments	283
3.1.6	Navigation dans l'arborescence des dossiers	284
3.1.7	Créer un notebook	285
3.2	Utiliser un notebook Jupyter	286
3.2.1	Renommer un notebook	286
3.2.2	Les cellules Jupyter	287
3.2.3	Les fonctionnalités d'un notebook	291

8 — Le Machine Learning

et l'IA générative avec Python

3.3	Utiliser les widgets Jupyter	297
3.3.1	Le widget FloatSlider	297
3.3.2	Associer une fonction à un slider	298
3.3.3	Le widget interact	300
3.3.4	Le widget Image	302
3.3.5	Le widget DatePicker	302
4.	Conclusion	303

Partie 3 : Les statistiques

Chapitre 3-1 Statistiques

1.	Objectif du chapitre	305
2.	Les statistiques descriptives	306
2.1	Paramètres de position	306
2.1.1	La moyenne	306
2.1.2	Le mode	307
2.1.3	La médiane	307
2.1.4	Les quartiles	310
2.2	Paramètres de dispersion	310
2.2.1	La variance	311
2.2.2	Calcul de la variance avec la formule de Koenig	311
2.2.3	L'écart-type	311
2.2.4	L'écart interquartile	312
3.	Les lois de probabilité	312
4.	La loi normale	314
5.	L'échantillonnage	318
5.1	Principe de l'échantillonnage	318
5.2	Résultats sur la distribution des moyennes	319
5.3	Résultats sur la distribution des proportions	328
5.4	Théorème central limite	332

6. Les statistiques inférentielles	333
6.1 Estimation ponctuelle	334
6.2 Estimation de la moyenne par intervalle de confiance.	338
6.3 Estimation d'une proportion par intervalle de confiance.	342
6.4 Test d'hypothèse.	346
6.4.1 Tests paramétriques	347
6.4.2 Tests non paramétriques	347
6.4.3 Construire un test d'hypothèse	348
6.5 Types de tests d'hypothèse	351
6.5.1 Test de conformité	351
6.5.2 Test d'adéquation	352
6.5.3 Tests d'homogénéité.	353
6.5.4 Test d'indépendance de variables	354
6.6 Exemple numérique de test de conformité d'une moyenne.	355
6.7 Le paradoxe de Simpson.	358
7. Les séries temporelles	361
7.1 Techniques d'analyse des séries temporelles.	362
7.1.1 La décomposition des séries temporelles.	362
7.1.2 Lissage des données	362
7.1.3 Modèles de prévision	362
7.2 Stationnarité des séries temporelles	363
7.2.1 Tests de stationnarité.	363
7.2.2 Transformation pour rendre une série stationnaire	363
8. Conclusion	364

Partie 4 : Les grands algorithmes du Machine Learning

Chapitre 4-1

La régression linéaire et polynomiale

1. Objectif du chapitre	365
2. La régression linéaire simple	366
2.1 La régression linéaire simple de point de vue géométrique . . .	367
2.2 La régression linéaire simple de point de vue analytique . . .	368
2.2.1 La méthode des moindres carrés	368
2.2.2 Quelques considérations statistiques sur les données . .	370
3. La régression linéaire multiple	371
3.1 La méthode des moindres carrés pour la régression multiple .	372
3.2 La méthode de la descente de gradient	373
3.3 Exemple de régression linéaire multiple	374
3.3.1 Définition du jeu de données utilisées	374
3.3.2 Régression linéaire multiple avec Scikit-learn	375
3.3.3 Importer les modules Scikit-learn	376
3.3.4 Lecture des données dans un DataFrame	377
3.3.5 Normalisation des données	378
3.3.6 Construction d'un modèle linéaire	381
3.3.7 Évaluation d'un modèle linéaire.	383
3.3.8 Évaluer le futur comportement d'un modèle	387
3.3.9 Cross-validation avec KFold.	390
4. La régression polynomiale	398
4.1 Exemple de régression polynomiale.	399
4.1.1 Construction d'un modèle polynomial.	399
4.1.2 Le coefficient de détermination R^2	406
4.1.3 R^2 et les valeurs extrêmes	408
4.1.4 Modèle polynomial et surapprentissage	408
5. Aller plus loin avec les modèles de régression.	414
5.1 La régularisation Lasso	414
5.2 La régularisation Ridge.	415

6. Conclusion	415
---------------------	-----

Chapitre 4-2

La régression logistique

1. Objectif du chapitre	417
2. La régression logistique	418
3. Prédire les survivants du Titanic	422
3.1 Définition du jeu de données Titanic	422
3.2 Réalisation du modèle de régression logistique	423
3.2.1 Chargement des modules Scikit-learn	423
3.2.2 Lecture des données	424
3.2.3 Traitement des valeurs manquantes	425
3.2.4 Transformation de variables	426
3.2.5 Sélection des variables	428
3.2.6 Traitement des variables catégorielles	430
3.2.7 Entraînement du modèle logistique	431
3.2.8 Le seuil de décision	432
4. L'algorithme One-vs-All	436
5. Conclusion	436

Chapitre 4-3

Arbres de décision et Random Forest

1. Objectif du chapitre	437
1.1 Construction d'un arbre de décision	438
1.2 Prédire la classe d'appartenance avec un arbre de décision ...	441
1.3 Considérations théoriques sur les arbres de décision	442
1.3.1 Choix de la variable de segmentation	443
1.3.2 Profondeur d'un arbre de décision	445
2. Problème de surapprentissage avec un arbre de décision	449
3. Random Forest	449

12 ————— Le Machine Learning

et l'IA générative avec Python

4. Exemple de Random Forest avec Scikit-learn	450
5. Conclusion	456

Chapitre 4-4

L'algorithme k-means

1. Objectif du chapitre	457
2. k-means du point de vue géométrique	458
3. k-means du point de vue algorithmique	465
4. Application de k-means avec Scikit-learn	467
5. L'algorithme k-means et les valeurs extrêmes.	474
6. Choisir le k de k-means	480
6.1 Déterminer k avec la méthode Elbow	483
6.2 Déterminer k avec le coefficient de silhouette	489
7. Les limites de k-means	495
8. Avantages et inconvénients de l'algorithme k-means	500
9. Quelques versions de l'algorithme k-means	501
10. Conclusion	501

Chapitre 4-5

Support Vector Machine

1. Objectif du chapitre	503
2. Le SVM du point de vue géométrique.	504
3. Le SVM du point de vue analytique	509
4. Données non linéairement séparables.	513
4.1 Le Kernel Trick	515
4.2 La condition de Mercer.	517
4.3 Exemple de fonction noyau	517

5.	Détecter les fraudes de cartes de crédit	518
5.1	Les données des transactions de cartes de crédit	518
5.2	Application de l'algorithme SVM pour la détection des transactions bancaires frauduleuses	519
5.2.1	Application de l'algorithme SVM sur les données creditcard.csv	519
5.2.2	Application du SVM sur un sous-ensemble de creditcard.csv	526
5.2.3	Application du SVM sur des données normalisées.	528
5.3	Les paramètres de l'algorithme SVM.	533
5.3.1	Le paramètre Kernel pour la variation de la fonction noyau.	534
5.3.2	Le paramètre C	535
5.3.3	Le paramètre Gamma.	539
5.3.4	Le paramètre C versus le paramètre Gamma	541
5.3.5	Tuning des hyperparamètres d'un SVM avec GridSearchCV.	541
6.	Conclusion	545

Chapitre 4-6

Analyse en composantes principales

1.	Objectif du chapitre	547
2.	Pourquoi l'ACP ?	548
3.	L'ACP du point de vue géométrique	550
4.	L'ACP du point de vue analytique.	552
5.	Indicateurs de la qualité de la représentation des données	555
5.1	Indicateurs liés aux individus	555
5.1.1	Score des individus	556
5.1.2	Qualité de la représentation des individus	556
5.1.3	Contribution des individus	556

14 _____ Le Machine Learning

et l'IA générative avec Python

5.2	Indicateurs liés aux variables	557
5.2.1	Le cercle des corrélations	557
5.2.2	Qualité de la représentation des variables.	558
5.2.3	Contribution des variables.	559
6.	Exemple d'ACP avec Python	559
6.1	Déterminer le nombre de facteurs pertinents.	564
6.2	Interprétation des résultats sur les individus	569
6.2.1	Représentation des individus.	570
6.2.2	Calcul de la qualité de la représentation des individus .	574
6.2.3	Calcul de la contribution des individus.	575
6.3	Interprétation des résultats sur les variables	576
6.3.1	Tracer un cercle des corrélations	576
6.3.2	Calcul de la qualité de la représentation des variables .	579
6.3.3	Calcul des contributions des variables	580
6.4	Représentation de nouveaux individus.	581
7.	Conclusion	583

Chapitre 4-7

Les réseaux de neurones

1.	Objectif du chapitre	585
2.	Modélisation d'un neurone artificiel	586
2.1	Le neurone biologique	587
2.2	Le neurone artificiel	588
3.	Architecture d'un réseau de neurones	590
4.	L'algorithme de rétropropagation	593
5.	Exemple d'un réseau de neurones avec Scikit-learn	604
6.	Conclusion	611

Partie 5 : Le Deep Learning et le traitement automatique du langage

Chapitre 5-1

Le Deep Learning avec TensorFlow

1. Objectif du chapitre	613
2. Le Deep Learning : notions générales	614
2.1 Réseau de neurones avec plusieurs couches d'entrée	617
2.2 Réseau de neurones avec plusieurs couches de sortie	618
2.3 Réseau de neurones avec des branchements conditionnels	619
2.4 Réseau de neurones avec de la récurrence RNN	620
2.5 Réseau de neurones avec couches de convolution CNN	621
2.6 Éviter le surapprentissage avec les couches Dropout	623
2.7 Le Transfer Learning	625
3. Introduction à TensorFlow	629
3.1 Installer TensorFlow	629
3.1.1 Créer un environnement virtuel	630
3.1.2 Installer des bibliothèques dans un environnement virtuel avec Anaconda	634
3.1.3 Installer la bibliothèque TensorFlow	636
3.1.4 Tester TensorFlow	637
3.2 Opérations élémentaires avec les tensors	638
3.2.1 Travailler avec les tensors	639
3.2.2 Les tensors variables	643
3.2.3 Initialiser les tensors	643
3.2.4 Opérations algébriques avec les tensors	644
4. Les réseaux de neurones avec Sequential API	645
4.1 Charger les données	646
4.2 Définir un MLP avec Sequential API	651
4.3 Accéder aux éléments d'un réseau de neurones	653
4.4 Initialisation des poids et des biais d'un réseau de neurones ..	655
4.5 Compiler un réseau de neurones	657

16 — Le Machine Learning

et l'IA générative avec Python

4.6	Entraîner un réseau de neurones	658
4.7	Analyser les résultats de l'entraînement d'un réseau de neurones	660
4.8	Évaluer un réseau de neurones.	662
4.9	Prédire avec un réseau de neurones pour la classification	662
5.	Utiliser Functional API.	664
5.1	Un modèle Functional API avec plusieurs couches d'entrée . .	665
5.2	Un modèle Functional API avec plusieurs couches de sortie . .	668
6.	Opérations avancées sur les réseaux de neurones.	671
6.1	Monitorer un réseau de neurones	671
6.1.1	Contrôler les critères d'arrêt avec les callbacks	671
6.1.2	TensorBoard	674
6.1.3	Sauvegarder un réseau de neurones.	678
6.1.4	Charger et utiliser un réseau de neurones.	678
6.2	Réseaux de neurones de convolution	680
6.3	Réutiliser un réseau de neurones	684
6.4	Le Transfer Learning.	686
6.4.1	Chargement des données locales	688
6.4.2	Chargement du modèle VGG16	689
6.4.3	Extraction des features.	689
6.4.4	Étendre un modèle	690
6.4.5	Chargement des données de test pour le Transfer Learning.	691
7.	Aller plus loin avec le Deep Learning et TensorFlow	694
8.	Conclusion	695

Chapitre 5-2

Le Deep Learning avec OpenCV

1. Objectif du chapitre	697
2. Introduction à OpenCV	698
2.1 Qui utilise OpenCV ?	698
2.2 Exemples de cas d'utilisation d'OpenCV	699
3. L'architecture d'OpenCV	700
4. Installer et tester OpenCV	703
5. Manipuler les images avec OpenCV	704
5.1 Charger une image	704
5.2 Capturer le flux d'une vidéo	706
5.3 Ajouter des objets dans une image	708
5.4 Ajouter des objets dans une vidéo	710
5.5 Gérer les clics de la souris sur une vidéo	712
5.6 Comprendre la structure d'une image	714
5.7 Modifier les pixels d'une image	716
5.8 Flouter une image	717
6. La détection et la reconnaissance d'objets	719
6.1 La détection faciale sur une image	719
6.2 La détection faciale sur une vidéo	721
6.3 Traquer les mouvements	723
6.4 Détecter des objets avec YOLO	725
7. Conclusion	729

Chapitre 5-3

Les réseaux de neurones antagonistes génératifs

1. Objectif du chapitre	731
2. Introduction au GAN	732
2.1 Comprendre les différents types de modèles d'IA générative	733
2.2 Définition et origine des GAN	733

18 — Le Machine Learning

et l'IA générative avec Python

2.3	Importance des GAN dans l'apprentissage profond	734
2.4	Les différents types de GAN	735
2.5	Quelques exemples de GAN	735
2.6	Avantages et inconvénients des GAN	736
2.6.1	Avantages	736
2.6.2	Inconvénients	736
3.	Fonctionnement des GAN	737
4.	Mes premiers pas avec PyTorch	744
4.1	Tester PyTorch dans Google Colab	745
4.2	Transformer une image en tensors	748
4.3	Appliquer des filtres sur les images	749
4.3.1	Accentuer les bords dans une image	749
4.3.2	Détecter des bords verticaux	751
4.3.3	Détecter des bords horizontaux	753
4.3.4	Appliquer un noyau gaussien	755
4.3.5	Donner un effet de gravure à une image	755
5.	Développer des réseaux de neurones avec PyTorch	756
5.1	Entraîner un modèle à rejeter les mauvaises images	758
5.2	Entraîner un modèle à accepter les images réalistes	768
6.	Générer des images réalistes avec un GAN	769
6.1	Chargement des vraies images de référence	771
6.2	L'entrée du Generator ou le vecteur latent	775
6.2.1	Pourquoi le vecteur latent est-il important ?	775
6.2.2	Quel est l'effet de modifier la taille du vecteur latent ?	776
6.3	Définition du réseau de neurones Generator	777
6.4	Définition d'un réseau de neurones Discriminator	779
6.5	Tester le Generator et le Discriminator avant la boucle d'apprentissage	781
6.6	Implémentation de la boucle d'apprentissage du GAN	784
6.7	Tester le Generator et le Discriminator	791
7.	Conclusion	794

Chapitre 5-4

Le traitement automatique du langage

1. Objectif du chapitre	795
2. NLP : concepts généraux	796
2.1 Le nettoyage des données textuelles	798
2.1.1 Suppression des stopwords	798
2.1.2 Appliquer le Stemming sur un texte	800
2.1.3 Appliquer la Lemmatization sur un texte	800
2.1.4 Stemming versus Lemmatization	800
2.2 Vectorisation des données textuelles	801
2.2.1 La vectorisation par comptage d'occurrences des mots	802
2.2.2 La vectorisation avec TF-IDF	804
2.2.3 La vectorisation avec N-Gram	806
2.2.4 Feature Engineering sur des documents	807
3. Exemple complet pour la détection des spams	808
3.1 Installation de la NLTK	809
3.2 Modèle de détection de spams	810
4. Conclusion	817

Chapitre 5-5

Le prompt engineering

1. Objectif du chapitre	819
2. Le prompt engineering	820
2.1 Concepts généraux	820
2.2 Les tokens	820
2.3 Comment écrire un prompt efficace et précis ?	822
3. Exemples de prompts	822
3.1 Les prompts simples et naïfs	823
3.2 Les prompts zero-shot	823
3.3 Les prompts few-shot	827
3.4 Les prompts Chain-of-Thought	830

20 ————— Le Machine Learning

et l'IA générative avec Python

3.5	Les prompts de type Generated-Knowledge-Prompting	835
3.6	Les prompts Directional-stimulus-prompting	836
3.7	Les prompts OPRO.	837
3.8	Résoudre les problèmes de logique avec des prompts	840
3.9	Faire des résumés avec Chain-of-Density	842
3.10	Générer du code avec les LLM	844
4.	Conclusion	853

Annexe

La programmation orientée objet avec Python

1.	Programmation orientée objet avec Python	855
1.1	Pourquoi la programmation orientée objet?	855
1.2	Classes et objets	857
1.2.1	Définir une classe	857
1.2.2	La fonction <code>__init__</code>	858
1.2.3	Instanciation d'un objet	860
1.2.4	Les attributs d'un objet	861
1.2.5	Les méthodes d'objet	863
1.2.6	Les attributs de classe	865
1.2.7	Les méthodes de classe	866
1.2.8	Les méthodes statiques	868
1.2.9	Sécuriser les attributs	869
1.3	L'héritage	873
1.3.1	L'héritage simple	873
1.3.2	L'héritage multiple	877
1.4	Les classes abstraites	879
1.5	Les interfaces	881

1.6	Les méthodes spéciales	885
1.6.1	Afficher un objet avec la fonction <code>print()</code>	885
1.6.2	Personnaliser les accès aux attributs d'une classe	888
1.6.3	Vérifier la validité d'un attribut	889
1.6.4	Comparer deux objets	891
1.6.5	Rendre les objets callable	892
2.	Les modules	893
2.1	Importer des modules	894
2.2	Le module principal	897
3.	Pour aller plus loin avec Python.	900
	Index	901

Les éléments à télécharger sont disponibles à l'adresse suivante :

<http://www.editions-eni.fr>

Saisissez la référence de l'ouvrage **EI2MLPYTSL** dans la zone de recherche et validez. Cliquez sur le titre du livre puis sur le bouton de téléchargement.

Chapitre 1

Introduction

1. Pourquoi un livre sur le Machine Learning ?	13
2. Python et Scikit-learn : les raisons du choix	14
3. À qui s'adresse ce livre ?	15
4. Organisation du livre et éléments en téléchargement	16
5. Datasets utilisés dans ce livre	17
5.1 Iris de Fisher, 1936	18
5.2 Titanic, 1994	18
5.3 Boston, 1978	19

Chapitre 2

Le Machine Learning : vue d'ensemble

1. Un peu de vocabulaire	21
2. Les métiers de la data	24
3. La croissance du Machine Learning	28
4. Formes d'apprentissage et tâches de ML	29
4.1 Apprentissage supervisé	30
4.1.1 Classification	30
4.1.2 Classification : le cas des images	31
4.1.3 Régression	32
4.1.4 Prévision	33
4.2 Apprentissage non supervisé	33
4.2.1 Clustering	33
4.2.2 Réduction de dimensions	34

2 **Machine Learning**

Implémentation en Python avec Scikit-learn

4.2.3	Système de recommandations	34
4.2.4	Associations	34
4.3	Apprentissage par renforcement	35
4.3.1	Comportements	36
4.3.2	Jeux et stratégies	37
4.4	Apprentissage semi-supervisé	37
4.5	Cas des IA génératives	38
4.6	Synthèse des différentes formes d'apprentissage et tâches	39
5.	Méthodologie CRISP-DM	39
5.1	Vue d'ensemble	40
5.2	Business Understanding	41
5.3	Data Understanding	43
5.3.1	Carte d'identité du dataset	43
5.3.2	Description des champs	44
5.3.3	Statistiques descriptives	44
5.4	Data Preparation	46
5.5	Modeling	47
5.6	Évaluation	48
5.7	Deployment	49

Chapitre 3

La pile technologique en Python

1.	Les outils de la Data Science	51
1.1	Les outils intégrés	52
1.2	L'auto ML	52
1.3	Les outils de développement	53
2.	Langage Python	53
2.1	Présentation	53
2.2	Brève présentation de R	54
2.3	Python ou R ?	55
2.4	Python 2 vs Python 3	56

3. Jupyter	57
3.1 Caractéristiques de Jupyter	57
3.2 Avantages de Jupyter pour la Data Science	60
3.3 Installation et utilisation de Jupyter	61
4. Bibliothèques de Machine Learning	63
4.1 NumPy	64
4.2 Pandas	67
4.3 Matplotlib	68
4.4 Scikit-learn	71
5. Bibliothèques de Deep Learning	72

Chapitre 4

Chargement et analyse des données

1. La phase de Data Understanding	75
2. Chargement des données	76
3. Création de la carte d'identité du dataset	78
4. Description des champs	79
4.1 Gestion des types	80
4.2 Détection des données manquantes	83
5. Statistiques descriptives sur les champs	85
5.1 Types de données	85
5.2 Analyse des données numériques	86
5.3 Graphiques sur les données numériques	89
5.3.1 Histogramme	89
5.3.2 Nuage de points	91
5.3.3 Boîtes à moustaches	94
5.4 Analyse sur les données catégorielles	96
5.5 Graphiques sur les données catégorielles	98
5.6 Autres données	102

4 **Machine Learning**

Implémentation en Python avec Scikit-learn

5.7	Analyse croisée des données.	103
5.7.1	Entre des variables numériques	103
5.7.2	Entre des variables numériques et une variable catégorielle.	104
5.7.3	Entre des variables catégorielles.	105
6.	Préparer la phase suivante	106

Chapitre 5

Préparation des données

1.	La phase de Data Preparation	109
2.	Limiter les données.	110
2.1	Supprimer des colonnes	111
2.2	Supprimer des enregistrements	112
3.	Séparer les datasets.	113
3.1	Proportion Entraînement/Test	114
3.2	Séparation aléatoire	114
3.3	Séparation stratifiée	116
4.	Traiter les données manquantes	118
5.	Préparer les attributs numériques	119
5.1	Validation des données	120
5.1.1	Validation sémantique des données	120
5.1.2	Validation statistique des données	121
5.2	Feature engineering	123
5.3	Discrétisation	124
5.3.1	Intervalles égaux.	125
5.3.2	Répartition par quantile.	125
5.3.3	Répartition manuelle	126

5.4	Normalisation	127
5.4.1	Normalisation min-max	128
5.4.2	Normalisation standard	128
5.4.3	Normalisation robuste	129
5.4.4	Comparaison	129
6.	Préparer les variables catégorielles	133
6.1	Validation des données	133
6.2	Modification des catégories	134
6.2.1	Ordonner ou réordonner des catégories	134
6.2.2	Modifier la liste des catégories	135
6.3	Quantification	138
7.	Les données particulières	140
7.1	Préparer les dates	141
7.1.1	Le format datetime64	141
7.1.2	Extraire des composantes	143
7.1.3	Gérer les écarts	144
7.2	Préparer les chaînes de caractères	145
7.2.1	Préparer les chaînes	146
7.2.2	Effectuer une recherche dans les chaînes	147
7.2.3	Extraire des sous-chaînes	148
7.2.4	Autres méthodes	149
8.	Automatiser la préparation	150
8.1	Création de pipelines de traitement	150
8.2	Paramètres des opérations et code Pandas	151
8.3	Pipelines avec Scikit-learn	152
8.3.1	Création d'un Transformer	152
8.3.2	Utilisation des Transformer	153
8.3.3	Inconvénients de Scikit-learn	155
8.4	Autres possibilités	155

6 **Machine Learning**

Implémentation en Python avec Scikit-learn

Chapitre 6

Modélisation et évaluation

1. Phase de modélisation	157
2. Création d'un ensemble de validation	158
3. Préparation des datasets	160
3.1 Dataset Iris	160
3.2 Dataset Titanic	161
3.3 Dataset Boston	163
4. Création des modèles	164
4.1 Processus itératif	164
4.2 Création d'un modèle en Scikit-learn	165
4.3 Évaluation d'un modèle	166
4.4 Validation croisée	166
4.5 Sauvegarde et chargement d'un modèle	167
5. Amélioration des modèles (fine-tuning)	169
5.1 Optimisation des hyperparamètres	169
5.2 Application en Scikit-learn	170
5.3 Sur- et sous-apprentissage	172
6. Méthodes ensemblistes	174
6.1 Bagging	174
6.2 Boosting	176
6.3 Stacking	177

Chapitre 7

Algorithmes de classification

1. La tâche de classification	179
1.1 Définition	179
1.2 Exemples de cas pratiques	180
1.3 Préparation spécifique des données	181

2. Évaluation des modèles	181
2.1 Matrices de confusion	183
2.1.1 Cas de la classification binaire	183
2.1.2 Cas de la classification multiclasse	185
2.2 Indicateurs dérivés de la matrice de confusion	186
2.2.1 Accuracy	187
2.2.2 Rappel et précision	188
2.2.3 F1-score	191
2.2.4 Sensibilité et spécificité	192
2.3 La courbe ROC et l'AUC	194
2.3.1 Prédiction et probabilité	194
2.3.2 Taux de vrais et faux positifs	195
2.3.3 Courbe ROC	195
2.3.4 Aire sous la courbe (AUC)	198
2.4 Choix des indicateurs d'évaluation	199
3. Les arbres de décision et algorithmes dérivés	199
3.1 Arbres de décision	200
3.1.1 Sortie de l'arbre	201
3.1.2 Choix du point de coupure	201
3.1.3 Critères d'arrêt	203
3.1.4 Exploitation de l'arbre	204
3.2 Random Forests	207
3.3 XGBoost (eXtreme Gradient Boosting)	209
4. K-Nearest Neighbors	211
5. Logistic Regression	213
5.1 Régression logistique binaire	213
5.2 Régression logistique polytomique	215
5.3 Application avec Scikit-learn	215
6. Naive Bayes	218
6.1 Principe général	218
6.2 Calcul des différentes probabilités	219
6.3 Application avec Scikit-learn	220

8 **Machine Learning**

Implémentation en Python avec Scikit-learn

7. Support Vector Machine	223
7.1 Présentation générale	223
7.1.1 Marge et support vector	223
7.1.2 Kernels	225
7.1.3 Avantages	226
7.2 Application avec Scikit-learn	226

Chapitre 8

Algorithmes de régression

1. La tâche de régression	229
1.1 Définition	229
1.2 Exemples de cas pratiques	230
1.3 Préparation spécifique des données	230
2. Entraînement et évaluation des modèles	231
2.1 Notion d'erreurs	233
2.2 Indicateurs dérivés de la mesure d'erreurs	234
2.2.1 Erreur absolue moyenne	234
2.2.2 Erreur quadratique moyenne	235
2.2.3 Racine de l'erreur quadratique moyenne	236
2.2.4 Coefficient de détermination et variance expliquée	236
2.2.5 Autres indicateurs	238
2.3 Choix des indicateurs d'évaluation	239
3. Utilisation des algorithmes de classification	239
3.1 Principe général	239
3.2 Arbres de décision et algorithmes dérivés	240
3.2.1 Arbres de décision	240
3.2.2 Random Forest	243
3.2.3 XGBoost	243
3.3 K-plus proches voisins (KNN)	244
3.4 Support Vector Machine (SVM)	245

4. Régression linéaire et variantes	246
4.1 Régression linéaire	246
4.2 Application dans Scikit-learn	247
4.3 Problème de la colinéarité	249
4.4 Ridge Regression	250
4.5 Régression Lasso	252
5. Régression polynomiale	254
5.1 Principe	254
5.2 Régression polynomiale et Scikit-learn	255
6. Cas particulier de la prédiction	257
6.1 Prédiction et séries temporelles	257
6.2 Préparation des données	260
6.3 Application en Scikit-learn	262
6.4 Utilisation de modèles spécifiques	267
6.4.1 Limites de l'approche en régression linéaire	267
6.4.2 Algorithmes dédiés aux séries temporelles	269

Chapitre 9

Algorithmes d'apprentissage non supervisés

1. Les tâches en apprentissage non supervisé	275
2. Clustering	276
2.1 Définition	276
2.2 Exemples de cas pratiques	277
2.3 Algorithmes basés sur les distances	278
2.3.1 Principe de l'algorithme K-Means	278
2.3.2 Implémentation avec Scikit-learn	279
2.3.3 Variantes de l'algorithme K-Means	283
2.4 Algorithmes basés sur la densité	284
2.4.1 Principe général	284
2.4.2 Implémentation de DBSCAN en Scikit-learn	285
2.4.3 Variante de DBSCAN : OPTICS	286

10 **Machine Learning**

Implémentation en Python avec Scikit-learn

3. Réduction des dimensions	286
3.1 Définition	286
3.2 Exemples de cas pratiques	287
3.3 Détection des axes principaux	288
3.4 Création de nouveaux axes	289
3.4.1 Principal Component Analysis (PCA)	289
3.4.2 Linear Discriminant Analysis (LDA)	292
4. Systèmes de recommandation	294
4.1 Définition	294
4.2 Principales approches	295
4.2.1 Modèles basés sur la popularité (popularity-based filtering)	295
4.2.2 Modèles basés sur le contenu (content-based filtering)	295
4.2.3 Modèles basés sur les autres utilisateurs (collaborative filtering)	296
4.2.4 Méthodes hybrides	297
5. Association	297
5.1 Définition	297
5.2 Évaluation des algorithmes	298
5.2.1 Le support	299
5.2.2 L'indice de confiance	300
5.2.3 Le lift	300
5.3 Algorithme « APriori »	301
5.3.1 Étape 1 : réalisation des comptages des ensembles	301
5.3.2 Étape 2 : création et test des règles	303

Chapitre 10

Évaluation et déploiement

1. Phase d'évaluation.	305
1.1 Principe global.	305
1.2 Évaluation métier des résultats	306
1.3 Revue du processus.	307
1.4 Étapes suivantes	307
2. Phase de déploiement	308
2.1 Planification du déploiement.	309
2.2 Monitoring et maintenance.	309
2.3 Rapport final et documentation	310
3. Déploiement et MLOps	311
3.1 Retours sur le DevOps	311
3.2 Apparition du MLOps	312
3.3 Tâches couvertes par le MLOps.	313
3.4 Critères de choix.	314

Conclusion

1. Le Machine Learning, une compétence clé	319
2. Mener un projet jusqu'au bout	320
3. Au-delà de la méthodologie	322
4. Expérimentation et expérience	322
5. Pour aller plus loin	323

Index	325
------------	-----