1. Record Nr. UNINA9910838326303321

Autore	Campesato Oswald
Titolo	Managing Datasets and Models
Pubbl/distr/stampa	Bloomfield : , : Mercury Learning & Information, , 2023 ©2023
ISBN	1-68392-950-0 1-68392-951-9
Edizione	[1st ed.]
Descrizione fisica	1 online resource (387 pages)
Disciplina	005.133
Soggetti	Python (Computer program language) COMPUTERS / Database Management / Data Mining
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Nota di contenuto	Front Cover -- Half-Title Page -- Title Page -- Copyright Page -- Dedication -- Contents -- Preface -- Chapter 1: Working with Data -- Import Statements for this Chapter -- Exploratory Data Analysis (EDA) -- Dealing with Data: What Can Go Wrong? -- Analyzing Missing Data -- Explanation of Data Types -- Data Preprocessing -- Working with Data Types -- What is Drift? -- What is Data Leakage? -- Model Selection and Preparing Datasets -- Types of Dependencies Among Features -- Data Cleaning and Imputation -- Summary -- Chapter 2: Outlier and Anomaly Detection -- Import Statements for this Chapter -- Working with Outliers -- Finding Outliers with NumPy -- Finding Outliers with Pandas -- Finding Outliers with Scikit-Learn (Optional) -- Fraud Detection -- Techniques for Anomaly Detection -- Working with Imbalanced Datasets -- Summary -- Reference -- Chapter 3: Cleaning Datasets -- Prerequisites for this Chapter -- Analyzing Missing Data -- Pandas, CSV Files, and Missing Data -- Missing Data and Imputation -- Skewed Datasets -- CSV Files with Multi-Row Records -- Column Subset and Row Subrange of Titanic CSV File -- Data Normalization -- Handling Categorical Data -- Working with Currency -- Working with Dates -- Working with Quoted Fields -- What is SMOTE? -- Data Wrangling -- Summary -- Chapter 4: Working with Models -- Import Statements for this Chapter -- Techniques for Scaling Data --

Examples of Splitting and Scaling Data -- The Confusion Matrix -- The ROC Curve and AUC Curve -- Exploring the Titanic Dataset -- Steps for Training Classifiers -- Diagram for Partitioned Datasets -- A KNN-Based Model with the wine.csv Dataset -- Other Models with the wine.csv Dataset -- A KNN-Based Model with the bmi.csv Dataset -- A KNN-Based Model with the Diabetes.csv Dataset -- SMOTE and the Titanic Dataset -- EDA and Data Visualization.
What about Regression and Clustering? -- Feature Importance -- What is Feature Engineering? -- What is Feature Selection? -- What is Feature Extraction? -- Data Cleaning and Machine Learning -- Summary -- Chapter 5: Matplotlib and Seaborn -- Import Statements for this Chapter -- What is Data Visualization? -- What is Matplotlib? -- Matplotlib Styles -- Display Attribute Values -- Color Values in Matplotlib -- Cubed Numbers in Matplotlib -- Horizontal Lines in Matplotlib -- Slanted Lines in Matplotlib -- Parallel Slanted Lines in Matplotlib -- Lines and Labeled Vertices in Matplotlib -- A Dotted Grid in Matplotlib -- Lines in a Grid in Matplotlib -- Two Lines and a Legend in Matplotlib -- Loading Images in Matplotlib -- A Checkerboard in Matplotlib -- Randomized Data Points in Matplotlib -- A Set of Line Segments in Matplotlib -- Plotting Multiple Lines in Matplotlib -- Trigonometric Functions in Matplotlib -- A Histogram in Matplotlib -- Histogram with Data from a Sqlite3 Table -- Plot a Best-Fitting Line with ggplot -- Plot Bar Charts -- Plot a Pie Chart -- Heat Maps -- Save Plot as a PNG File -- Working with SweetViz -- Working with Skimpy -- 3D Charts in Matplotlib -- Plotting Financial Data with Mplfinance -- Charts and Graphs with Data from Sqlite3 -- Working with Seaborn -- Seaborn Dataset Names -- Seaborn Built-In Datasets -- The Iris Dataset in Seaborn -- The Titanic Dataset in Seaborn -- Extracting Data from Titanic Dataset in Seaborn (1) -- Extracting Data from Titanic Dataset in Seaborn (2) -- Visualizing a Pandas Data Frame in Seaborn -- Seaborn Heat Maps -- Seaborn Pair Plots -- What is Bokeh? -- Introduction to Scikit-Learn -- The Digits Dataset in Scikit-Learn -- The Iris Dataset in Scikit-Learn (1) -- The Iris Dataset in Scikit-Learn (2) -- Advanced Topics in Seaborn -- Summary -- Appendix: Working with awk -- The awk Command.
Aligning Text with the printf() Statement -- Conditional Logic and Control Statements -- Deleting Alternate Lines in Datasets -- Merging Lines in Datasets -- Matching with Metacharacters and Character Sets -- Printing Lines Using Conditional Logic -- Splitting File Names with awk -- Working with Postfix Arithmetic Operators -- Numeric Functions in awk -- One-Line awk Commands -- Useful Short awk Scripts -- Printing the Words in a Text String in awk -- Count Occurrences of a String in Specific Rows -- Printing a String in a Fixed Number of Columns -- Printing a Dataset in a Fixed Number of Columns -- Aligning Columns in Datasets -- Aligning Columns and Multiple Rows in Datasets -- Removing a Column from a Text File -- Subsets of Column-Aligned Rows in Datasets -- Counting Word Frequency in Datasets -- Displaying Only "Pure" Words in a Dataset -- Working with Multi-Line Records in awk -- A Simple Use Case -- Another Use Case -- Summary -- Index.

Sommario/riassunto

This book contains a fast-paced introduction to data-related tasks in preparation for training models ondatasets. It presents a step-by-step, Python-based code sample that uses the kNN algorithm to manage a model on a dataset.Chapter One begins with an introduction to datasets and issues that can arise, followed by Chapter Two on outliers and anomaly detection. The next chapter explores ways for handling missing data and invalid data, and Chapter Four demonstrates how to train models with classification algorithms. Chapter 5 introduces

visualization toolkits, such as Sweetviz, Skimpy, Matplotlib, and Seaborn, along with some simple Python-based code samples that render charts and graphs. An appendix includes some basics on using awk. Companion files with code, datasets, and figures are available for downloading.FEATURES:Covers extensive topics related to cleaning datasets and working with modelsIncludes Python-based code samples and a separate chapter on Matplotlib and SeabornFeatures companion files with source code, datasets, and figures from the book