1.

Record Nr.	UNINA9910768172903321
Autore	Qi Zhixin
Titolo	Dirty Data Processing for Machine Learning / / by Zhixin Qi, Hongzhi Wang, Zejiao Dong
Pubbl/distr/stampa	Singapore : , : Springer Nature Singapore : , : Imprint : Springer, , 2024
ISBN	981-9976-57-X
Edizione	[1st ed. 2024.]
Descrizione fisica	1 online resource (141 pages)
Altri autori (Persone)	WangHongzhi DongZejiao
Disciplina	005.7
Soggetti	Artificial intelligence - Data processing Data mining Big data Data Science Data Mining and Knowledge Discovery Big Data
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Nota di contenuto	Chapter 1. Introduction -- Chapter 2. Impacts of Dirty Data on Classification and Clustering Models -- Chapter 3. Dirty-Data Impacts on Regression Models -- Chapter 4. Incomplete Data Classification with View-Based Decision Tree -- Chapter 5. Density-Based Clustering for Incomplete Data -- Chapter 6. Feature Selection on Inconsistent Data -- Chapter 7. Cost-Sensitive Decision Tree Induction on Dirty Data.
Sommario/riassunto	In both the database and machine learning communities, data quality has become a serious issue which cannot be ignored. In this context, we refer to data with quality problems as “dirty data.” Clearly, for a given data mining or machine learning task, dirty data in both training and test datasets can affect the accuracy of results. Accordingly, this book analyzes the impacts of dirty data and explores effective methods for dirty data processing. Although existing data cleaning methods improve data quality dramatically, the cleaning costs are still high. If we knew how dirty data affected the accuracy of machine learning models, we could clean data selectively according to the accuracy requirements instead of cleaning all dirty data, which entails substantial costs.

However, no book to date has studied the impacts of dirty data on machine learning models in terms of data quality. Filling precisely this gap, the book is intended for a broad audience ranging from researchers inthe database and machine learning communities to industry practitioners. Readers will find valuable takeaway suggestions on: model selection and data cleaning; incomplete data classification with view-based decision trees; density-based clustering for incomplete data; the feature selection method, which reduces the time costs and guarantees the accuracy of machine learning models; and cost-sensitive decision tree induction approaches under different scenarios. Further, the book opens many promising avenues for the further study of dirty data processing, such as data cleaning on demand, constructing a model to predict dirty-data impacts, and integrating data quality issues into other machine learning models. Readers will be introduced to state-of-the-art dirty data processing techniques, and the latest research advances, while also finding new inspirations in this field.