Studies in Big Data 96

G. P. Obi Reddy
Mehul S. Raval
J. Adinarayana
Sanjay Chaudhary *Editors*

Data Science in Agriculture and Natural Resource Management

Springer

Studies in Big Data

Volume 96

Series Editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland

The series “Studies in Big Data” (SBD) publishes new developments and advances in the various areas of Big Data- quickly and with a high quality. The intent is to cover the theory, research, development, and applications of Big Data, as embedded in the fields of engineering, computer science, physics, economics and life sciences. The books of the series refer to the analysis and understanding of large, complex, and/or distributed data sets generated from recent digital sources coming from sensors or other physical instruments as well as simulations, crowd sourcing, social networks or other internet transactions, such as emails or video click streams and other. The series contains monographs, lecture notes and edited volumes in Big Data spanning the areas of computational intelligence including neural networks, evolutionary computation, soft computing, fuzzy systems, as well as artificial intelligence, data mining, modern statistics and Operations research, as well as self-organizing systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

The books of this series are reviewed in a single blind peer review process.

Indexed by SCOPUS, EI Compendex, SCIMAGO and zbMATH.

All books published in the series are submitted for consideration in Web of Science.

More information about this series at http://www.springer.com/series/11970

G. P. Obi Reddy · Mehul S. Raval · J. Adinarayana ·
Sanjay Chaudhary
Editors

Data Science in Agriculture and Natural Resource Management

Springer

Editors
G. P. Obi Reddy
Division of Remote Sensing Application
ICAR-National Bureau of Soil Survey
and Land Use Planning
Nagpur, India

J. Adinarayana
Centre of Studies in Resources Engineering
Indian Institute of Technology Bombay
(IITB)
Mumbai, India

Mehul S. Raval
School of Engineering and Applied Science
Ahmedabad University
Ahmedabad, India

Sanjay Chaudhary
School of Engineering and Applied Science
Ahmedabad University
Ahmedabad, India

ISSN 2197-6503 ISSN 2197-6511 (electronic)
Studies in Big Data
ISBN 978-981-16-5846-4 ISBN 978-981-16-5847-1 (eBook)
https://doi.org/10.1007/978-981-16-5847-1

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

To my beloved parents and family members for their constant encouragement and support.

by G. P. Obi Reddy

In dedication to my wife—Hemal,
son—Hetav, and in memory of
mother—Charu and father—Shirish for their constant encouragement and support through thick and thin.

by Mehul S. Raval

To my parents and family members for their constant encouragement and support.

by J. Adinarayana

To my parents and family members for their strong support throughout.

by Sanjay Chaudhary

Foreword

The world of Data Science is rapidly changing due to fast-emerging technology changes in the fields of computers, data acquisition systems, and digital technologies. The last two decades witnessed tremendous developments in Information and Communication Technologies (ICT), including the Internet, cloud computing, sensors technology, computer processing, and storage and dissemination systems, which opens new avenues in digital data acquisition, processing, visualization, and disseminating valuable information of the planet earth to the general users, planners, and policy makers. This digital revolution, accompanied by the fast emerging of remote sensing platforms, provides an unprecedented amount of geospatial data on the status of planet earth resources, especially on natural resources, agriculture, and their dynamics, to develop various Earth Observation (EO) applications.

The edited volume on 'Data Science in Agriculture and Natural Resource Management' addressed the principles and applications of data science and emerging challenges in agriculture and natural resource management. The volume contains three important sections, namely, 'Data Science-Principles, Concepts and Applications, 'Data Science—Applications in Agriculture', and 'Data Science—Applications in Natural Resources Management.' The chapters in these sections dealt with specialized areas contributed by eminent experts from research institutes and universities of India and abroad with suitable illustrations and tables. The edited book is unique in the field of 'Data Science Applications in Agriculture and Natural Resource Management.' The book provides advanced knowledge on data science and added the latest knowledge in the field to benefit the global readers. The authors from renowned international organizations/Universities/Industries like International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), International Center for Agricultural Research in the Dry Areas (ICARDA), national organizations like ICAR-National Bureau of Soil Survey and Land Use Planning (ICAR-NBSS&LUP), Indian Institute of Technology Bombay (IITB), Ahmedabad University, Pandit Deendayal Energy University, University of Agricultural Science, Bangalore; and industrial entities such as NTT Data, International Business Machines (IBM) Corporation and Amnex Infotechnologies contributed the chapters to the edited volume with suitable case studies.

This book explored various facets of data science, the big data revolution, and EO applications in agriculture and natural resource management. It stimulates new ideas in data-driven research, new applications by integrating emerging data science techniques and robust prediction models in the emerging fields of cloud computing, artificial intelligence (AI), and deep learning, and ICT applications for geo-smart agriculture and sustainable natural resource management. I am sure the book realizes the aspirations and needs of geospatial, data science, agricultural, natural resources, and environmental scientists/faculty/students from traditional universities, agricultural universities, technological universities, research institutes, and academic colleges worldwide. It also helps planners, policymakers, and extension scientists plan and sustain natural resources for climate-resilient agriculture.

In summary, this book will be an added asset in the field of contemporary data science applications, which depicts the chapters ranging from fundamentals to the advanced mix of high-end information technology and agriculture and natural resource management. As these fields are dynamic, I wish the Editors will bring more editions of the influential book.

July 2021

Prof. Seishi Ninomiya
International Field Phenomics
Research Laboratory
The University of Tokyo
Tokyo, Japan

Preface

Agriculture and natural resources are the core and essential building blocks for the development of any society. It is important to efficiently use them for fulfilling growing food demand. In recent years, we are experiencing tremendous technology growth and the evolution of innovative applications using Data Science (DS). The technology is helping us to capture various types of data related to farming practices to harvesting to post-harvesting and supply chain.

Data Science is a genuinely multi-disciplinary area having an intersection of Computer Science, Mathematics, and Artificial Intelligence (AI). The problems in DS require analyzing discrete structured and unstructured high dimensional data sets, consisting of a collection of observables and their associated responses for predicting responses for previously unseen observations. Each farm and natural resource has a great source of data. DS has all the potential to make the best use of available voluminous, dynamic, and real-time data to generate actionable information. It enables us to make efficient decisions for the best utilization of natural resources and improve the yield of agricultural products.

We feel a significant need to present various aspects of 'Data Science in Agriculture and Natural Resource Management (DSANRM) in the form of a book. It will be helpful for students aiming to undertake projects in the DSANRM. At the same time, it will be a comprehensive quality resource for practitioners, researchers, decision-makers, and policymakers associated with this domain.

With the above aims, this book is organized into three sections:

1. Data Science—Principles, Concepts, and Applications
2. Data Science—Applications in Agriculture
3. Data Science —Applications in Natural Resource Management

The Part I comprises four chapters that focus on DS principles and concepts used in DSANRM. The Chapter "Data Science—Algorithms and Applications in Earth Observation" by Reddy and Kumar N. introduces DS and highlights its interdisciplinary nature. It uses Earth Observations technologies to discuss tools used in DS at various stages—collection, storing, processing, analysis, and visualization.

It shows how DS has been enriched with techniques derived from Machine Learning (ML) and Data Mining (DM). It introduces supervised algorithms for regression, classification, and various clustering algorithms. It discusses the use and challenges of ML algorithms in EO studies like drought monitoring. The Chapter "Emerging Technologies—Principles and Applications in Precision Agriculture" by Mishra, dives deeper and brings out all essential aspects of technology-driven Precision Agriculture (PA). It covers the complete gamut of on-field sensing using the Internet of Things (IoT), remote sensing using Unmanned Aerial Vehicles (UAVs), and satellites. It shows how edge computing and cloud computing can use Big Data Analytics (BDA) and ML to carry out the predictive analysis.

The Chapter "Data Science: Principles and Concepts in Modeling Decision Trees", by Divakaran presents a brief overview of statistical learning techniques like Linear and Non-linear Regression, Support Vector Machines (SVM), Decision Trees (DT), and Neural Networks (NN) for prediction problems in both regression and classification settings. Then, the study of Decision Trees is motivated by highlighting their computational efficiency and higher model interpretability. The chapter provides vital insights into the form of (i) the workings of standard decision tree-based statistical learning methods for making predictions; (ii) the understanding of these methods; and (iii) the use of Decision Trees in the context of Precision Agriculture (PA). Gandhi, in Chapter "Deep Reinforcement Learning for Agriculture: Principles and Use Cases" brings in another dimension of ML—Reinforcement Learning (RL). The chapter introduces RL concepts like Q learning and then details the connection between Deep Neural Networks (DNN) and RL. The chapter presents two fascinating case studies on—yield maximization and fruit detection. The yield case study shows how the RL agent can find the optimal strategy for setting the greenhouse parameters to maximize yield for cherry and tomatoes. In the second case study, RL is applied to identify products from the crop images. The chapter shows the architecture of the RL, implementation details, and results. The chapter is an essential milestone in linking theory and application, following Section II of the book.

Part II comprises a five-chapter dedicated to showing the use of DS in agriculture. Chapter "Computer Vision and Machine Learning in Agriculture" by Raval, Chaudhary, and Adinarayana link the sensing by computer vision (CV) and analysis by ML. Using PA as a domain, the chapter shows how CV and ML can be seamlessly integrated. It provides crucial discussions on public datasets and state-of-the-art algorithms for weed control, fruit detection, and plant phenotyping. The chapter ends with discussions on challenges and future paths. Chapter "An Architecture for Quality Centric Crop Production System by Kumar, Hiremath, and Chaudhary develops a recommendation system for farmers. The system helps farmers to map specific crop attributes with consumer requirements. The recommender uses endogenous and exogenous data obtained through IoT sensors and AI. Interestingly, the architecture is generic and can be extensible to a wide range of crops. Kumar J. et al. pens Chapter "Crop Classification for Precision Farming Using Machine Learning Algorithms and Sentinel-2 Data". The chapter focuses on remote sensing-based crop monitoring for small farm holding. It uses Sentinel—2 satellite images to delineate farm boundaries and perform crop classification using Random Forest (RF) classifier and Open—source Geographic Information system. The method is generic and applied to different sizes of farms and crops.

Chapter with the title "Machine Learning Approaches and Sentinel-2 Data in Crop Type Mapping," by Panjala, Gumma, and Teluguntla, uses Google Earth Engine (GEE) for crop monitoring. The chapter showcases evaluation of several ML supervised learning algorithms and spectral matching techniques using Sentinel—2 data to classify several crops. The chapter compares classifier results with agricultural statistics and shows that the RF classifier is the best crop classification. The last chapter of Part II, "Big Data Analytics for Climate-Resilient Food Supply Chains: Opportunities and Way Forward," written by Twarakavi et al., is about the reliable forecast of food production. The authors argue that current methods for forecasting food production levels are rudimentary and not scalable. The chapter develops and validates methodology by combining weather forecasts, remote sensing, scalable machine-learning methods, and cloud computing to estimate production risk at regional levels.

While Part II is focused on agriculture, Part III brings attention to Natural Resource Management with five chapters. Chapter "Machine Learning Algorithms for Optical Remote Sensing Data Classification and Analysis" by Reddy and Kumar A., introduces advanced ML algorithms for remote sensing analysis and classification. The RF, SVM, and Classification and Regression Tree (CART) were applied on GEE platform to derive land-use-land-cover (LULC) classes. The chapter demonstrates the immense potential for advanced ML algorithms in remote sensing data classification. Chapter "Geo-Big Data in Digital Augmentation and Accelerating Sustainable Agroecosystems"—presents ongoing digital augmentation efforts for rice fallows in India. The authors, Krishna and Biradar, argue about the need for augmentation and digitization to accelerate an agroecological transition. The chapter shows the capabilities and limitations of mapping rice-based farming systems using Sentinel series data during the *Kharif* season in a monsoon climate. The chapter shows a novel approach for multi-scale mapping (at state or region level) to micro-scale mapping (at district/village level) to assist informed decision-making.

Chapter "Transforming Soil Paradigms with Machine Learning" by Kumari S. et al.—brings attention to the most important natural resource—soil. The chapter shows that ML and BDA can be effectively used to predict, identify, and classify soil resources. Primarily, it shows the use of Digital Soil Mapping (DSM) and ML algorithms to predict soil properties. The soil-related studies continue in Chapter "Remote Sensing and Machine Learning for Identification of Salt-affected Soils". The authors Kumar N. et al. showcase ML techniques in identifying salt-affected soils (SAS) in Indo-Gangetic plains in India. The chapter studies salinity indices and principal components on images of the Landsat—8 satellite. The results were validated with field observations and GEE data. The final chapter, 14, "Geoportal Platforms for Sustainable Management of Natural Resources" by Reddy, uncovers Geo-portals' role in disseminating geospatial information. The chapter weaves the exciting development of the Bhoomi geoportal and its services. It provides Web Map Services (WMS) to access, query, visualize, and disseminate information. The chapter also highlights fundamental challenges in the development of such portals, such as cross-domain thematic applications.

Nagpur, India	G. P. Obi Reddy
Surat, India	Mehul S. Raval
Mumbai, India	J. Adinarayana
Ahmedabad, India	Sanjay Chaudhary
July 2021	

Acknowledgements

Writing or editing a book with excellence is an asymptotic process towards perfection—there is always a scope to improvise. The editors are eternally grateful to all the authors of the book—Nirmal Kumar, Shakti Mishra, Srikrishnan Divakaran, Ratnik Gandhi, Vikaskumar, Deepak Hiremath, Jay Prakash Kumar, Dimple Singhania, Sanket Nayankumar Patel, Mihir Dakwala, Pranay Panjala, Murali Krishna Gumma, Parthasaradhi Teluguntla, Navin Kumar C. Twarakavi, Kamal Das, Mohamed Akram Zaytar, Fred Otieno, Jitendra Singh, Bruno Silva, Komminist Weldemariam, Shantanu Godbole, K. C. Arun Kumar, Gopal Krishna and Chandrashekhar Biradar, Kumari Sweta, S. Dharumarajan, Amar Suputhra, M. Lalitha, R. Vasundhara, B. Kalaiselvi, B. N. Shashi Kumar, A. Sathish, Rajendra Hegde, M. S. S. Nagaraju, and R. K. Naitam. The book would not have seen a timely outcome without their expert contribution and real time responses to reviewers' comments.

Big thanks to Professor Seishi Ninomiya, International Field Phenomics Research Laboratory, Graduate School of Agriculture and Life Sciences, the University of Tokyo for penning Foreword and words of advice.

The editors express their gratitude to all associated with 8th International Conference on Big Data Analytics conference (BDA2020) conference and co-located Virtual Workshop on Data Science for Agriculture and Natural Resource Management (DSANRM2020), where the seeds of the book were sown.

The editors would like to thank leadership at their parent organizations—ICAR-National Bureau of Soil Survey and Land Use Planning (ICAR—NBSS&LUP), Nagpur, India; Ahmedabad University, Ahmedabad, India; and Indian Institute of Technology Bombay (IITB), Mumbai for providing time and freedom to work on this book project.

The editors also acknowledge the support from Prof. Dr. Janusz Kacprzyk Series Editor “Big Data” of Springer, and Jennifer Sweety Johnson and Suresh Dharmalingam from Springer Nature for project coordination of the book.

Nagpur, India
Surat, India
Mumbai, India
Ahmedabad, India
July 2021

G. P. Obi Reddy
Mehul S. Raval
J. Adinarayana
Sanjay Chaudhary

Contents

Part I Data Science—Principles, Concepts and Applications

Part II Data Science Applications in Agriculture

Part III Data Science Applications in Natural Resource Management

About the Editors

G. P. Obi Reddy holds Ph.D. and is working as Principal Scientist in the field of remote sensing and GIS applications at Division of Remote Sensing Application, ICAR-National Bureau of Soil Survey and Land Use Planning, Nagpur, India. He has significantly contributed in the field of remote sensing along with GIS applications in digital terrain modelling, landforms mapping, soil-landscape modelling, land degradation assessment, agroecology and development of soil information systems. He is instrumental in the design and development of the Land Resource Information System (LRIS) of India and Bhoomi geoportal. On ICAR deputation, he visited Sri Lanka, The Netherlands, Nepal and South Africa. He has published 102 research articles in reputed national/international journals, 7 books, 53 book chapters and 15 technical bulletins. Currently he is acting as a Chairman, Data Content Standards of DST-NSDI and Member, ISO/TC 211/WG 04 on "Geospatial Services". He is Recipient of Indian National Geospatial Award-2007 and National Geospatial Award for Excellence-2013 from Indian Society of Remote Sensing, Dehradun, Outstanding Scientist Award-2016–17 from ICAR-NBSS&LUP, Nagpur and Fellow of Indian Society of Soil Survey and Land Use Planning, Nagpur.

Mehul S. Raval holds Ph.D. and is Associate Dean—Experiential Learning and Professor at Ahmedabad University, India. His research interest includes computer vision, and he has contributed to problems in surveillance, medical imaging, biometrics and agriculture. His academic pursuits involve visits to under Sakura Science Fellowship (2015) to Okayama University, Japan, Argosy visiting Associate Professor at Olin College of Engineering, MA, the USA (2016), and Sacred Heart University, the USA (2019). He serves as Member, technical program committee, for leading national and international conferences, workshops and symposiums. Currently, he chairs IEEE Computational Intelligence Society—Gujarat Chapter and served in IEEE Gujarat section under various capacities. He has received research grants from Board of Research in Nuclear Science and DST—Govt. of India. He has published 79 scholarly works in journals, magazines, conferences, workshops at the national and international stage. He reviews IEEE, ACM, Springer, Elsevier, IET and other leading publishers and has supervised three Ph.D. students.

J. Adinarayana holds Ph.D. and is working as Teaching/Research Faculty Member from 1986 and currently Institute Chair Professor at Centre of Studies in Resources Engineering (CSRE), IIT Bombay, India. His areas of expertise include agro-informatics in contemporary agriculture. As a Team Leader, he led various interdisciplinary national and international R&D projects on digital agriculture. He is the Immediate Past President of Asia- Pacific Federation for Information Technology in Agriculture (APFITA) Board and the Current President of INSAIT. He served as an Editorial Board Member of 'Regional Geoderma' (Elsevier) & CIGR journals. Currently, he is a Member, CIGR Board on Technical Section VII (IT Systems); Vice-Chair of Agriculture and Open & Sharing Data Working Groups of the Asia-Pacific Advanced Network (APAN), Expert Committee Member in various national organizations, including ICAR, ISRO, MoA&FW and DST. He is Recipient of JSPS Invitation Fellowship, DST Young Scientist Project, INSA Visiting Scientist/Faculty awards, Visiting and Adjunct Faculty/Scientist at Univ. of Tokyo (2013), Univ. of Bonn (2007), Univ. of Aston/UK (1991); Edith Cowan Univ./Perth (2010-17). He guided about 12 Ph.D.s and more than 30 M.Techs and presented research results in more than 100 research papers/book chapters in International Journals/publishers.

Sanjay Chaudhary holds Ph.D. and is Professor at the School of Engineering and Applied Science and Dean of Students of Ahmedabad University. During 2001 to 2013, he was Professor as well as Dean (Academic Programs) at Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, India. His research areas are cloud computing, blockchain technology, big data analytics and ICT applications in agriculture and rural development. He has authored eight books and nine book chapters. He has published more than 125 research papers in international conferences, workshops and journals. He has received research grants from leading organizations including IBM, Microsoft and Department of Science and Technology, Govt. of India. He is Vice President of INSAIT. He has guided seven Ph.D. students and more than 32 M.Tech. students. He holds a doctorate degree in computer science from Gujarat Vidyapith.

Part I
Data Science—Principles, Concepts and Applications

Data Science—Algorithms and Applications in Earth Observation

G. P. Obi Reddy and Nirmal Kumar

Abstract Data Science (DS) is a fast-evolving interdisciplinary field and it encompasses various scientific approaches, procedures, and systems to abstract information, and insights from large datasets. The vital components of DS are identifying the reliable data sources, data curation, model building, planning and evaluation, results communication, and visualization. DS tools are being widely used for data collection, storage, extraction, cleaning, analysis, and visualization. In recent times, DS is enriched by various techniques like Artificial Intelligence (AI), Machine Learning (ML), Data Mining (DM), predictive analysis, pattern recognition, and visualization. Decision tree algorithms, Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Logistic Regression (LR), Neural networks are some of the important ML algorithms. In recent years, Earth Observation (EO) satellites generated voluminous spatial data, and integration of these datasets with ground-based observational data significantly enhanced the capabilities in mapping, monitoring, and forecasting various earth system processes. ML algorithms have vast potential in monitoring earth resources to provide timely information for site-specific interventions, and implementation. This chapter aims to illustrate various DS tools, algorithms, and the potential applications of ML algorithms in important EO applications like drought monitoring, vegetation monitoring and assessment, digital soil mapping, soil moisture monitoring, Soil Organic Carbon (SOC) assessment, crop type mapping, and precision agriculture. The challenges associated with DS applications in EO are also discussed in the chapter.

Keywords Data Science · Data engineering · Earth observation · Machine learning

G. P. O. Reddy (✉) · N. Kumar
ICAR-National Bureau of Soil Survey and Land Use Planning, Amravati Road, Nagpur 440033, India
e-mail: GPO.Reddy@icar.gov.in

G. P. O. Reddy et al. (eds.), *Data Science in Agriculture and Natural Resource Management*, Studies in Big Data 96, https://doi.org/10.1007/978-981-16-5847-1_1

1 Introduction

Earth Observation (EO) technologies and applications play a pivotal role in achieving global challenges like no poverty, zero hunger, and climate change (Andries et al. 2019). EO technologies aimed to gather and record information about a point of interest with the help of sensors usually mounted on a satellite without contacting the object physically (Campbell and Wynne 2011). In recent years, digital data produced through EO technologies are playing a key role in monitoring, and analyzing various physical, chemical and biological processes of planet Earth (Yang et al. 2013). In general, EO data acquired in the form of images collected by various sensors in different time scales are cost-effective and time-efficient to provide solutions at global, regional, and local scales (Roy et al. 2017; Reddy 2018a). The United Nations recognized the potential of EO data to supplement the official statistics generation processes, and complement traditional sources of various environmental data (UN 2017). Time-series remote sensing technologies have demonstrated their potential in EO especially in mapping, and monitoring of earth resources to realize sustainable development (Reddy et al. 2017; Reddy and Singh 2018). To address various emerging issues, there is a need to apply new approaches, and techniques to effectively process, and analyze voluminous EO data received from the various ground, air, and space-borne sensors. In recent years, a host of data science-based Machine Learning (ML) algorithms like supervised, and unsupervised learning, K-means, neural net processing, and pattern recognition have enabled us to transform the field of EO research and applications.

Data Science (DS) covers various scientific approaches, procedures, and systems, which are used to abstract information, and insights from large datasets (Dhar 2013). DS applies mathematical, and statistical concepts and computer tools for processing, extract insights and information from big data. The distributed processing, monitoring of workflows, visualization, and performance monitoring are core parts of the DS domain. DS techniques help in finding the useful patterns, and relationships within the data. In recent years, many organizations adopted DS tools for collection, storing, and processing of data as a part of their regular operations, and services. Artificial Intelligence (AI) and ML are related to each other and are often used interchangeably in DS. AI is the superset; it has ML and Deep Learning (DL) as subsets. AI describes machines that can execute tasks resembling those of humans through artificially modeling human intelligence. ML is a subsection of AI that device means by which systems can automatically learn, and improve from experience. ML algorithms play a vital role in effectively handling EO data as these datasets are huge in quantity, and formats. The focus of ML is the depiction of the input data, and simplification of patterns for use from big data.

In the past three decades, the advanced capabilities in computing and sensor technology have tremendously transformed the field of remote sensing, data collection systems, processing, and producing valuable products, and services. The important portals like USGS Earth Explorer and Google Earth Engine (GEE) (Gorelick et al. 2017), Copernicus Sentinel Hub (https://scihub.copernicus.eu/), and ISRO-Bhuvan (https://bhuvan.nrsc.gov.in/home/index.php) provide public access to obtain enormous EO resources data to the scientific communities to build various EO applications. Integration of data obtained from space-based EO satellites and field-based observations is a formidable combination to develop complex, and comprehensive data-driven EO applications. In addition, multi-resolution, and multi-temporal EO data divulge various planet earth processes, and features, which are not able to observe through conventional methods. Thus, the frequency and coverage of large quantities of EO data provide voluminous data in the form of images, and maps, such large datasets could not be easily generated through ground-based technologies alone. Advances made in spectral, spatial, and temporal resolutions of remote sensing sensors technology, and ever-growing number, and diversity of EO satellites, implementation of open data access policies, and ever-increasing demand for innovative applications prompted to adopt innovative approaches, and platforms to store, manage, process, and analyze voluminous EO data (Guo 2017; Dhu et al. 2017; Baumann et al. 2019). Data generated through EO technologies has immense potential in developing the applications and services in managing the planet's resources and environmental monitoring. However, processing large quantities of EO data, and converting it into useful information necessitates robust computational resources and it remains a major challenge to build tailor-made applications to meet the user requirements (Giuliani et al. 2017). This chapter aimed to discuss various components, tools, techniques, and potential applications of DS in emerging EO like drought monitoring, vegetation monitoring and assessment, digital soil mapping, soil moisture monitoring, SOC assessment, crop mapping, and precision agriculture. The performance of the RF algorithm in SOC assessment, and challenges associated with DS applications in EO were also discussed in the chapter.

2 Data Science and Its Components

DS principles and concepts help to mine both structured and unstructured large datasets in order to recognize the patterns, trends, and extract insights. This is an inter-disciplinary field and it includes statistics, inference, computer science, predictive analytics, ML algorithms, and new technologies, which immensely helps to infer insights from big data. Big data in DS undergo numerous processes like data discovery, development, warehousing, cleaning, classification, analysis, and validation for visualization of patterns, and insights. (Ngiam and Khor 2019). Big data implies voluminous datasets that have been generated to develop possible applications and provide solutions through the extraction of value, and knowledge (Gandomi and Haider 2015). The main concerns in DS are efficient capture, storage, extraction, process, analysis, and visualization of the information from the enormous datasets. The important components of DS can be grouped into data discovery, data preparation, model planning, model building, model operationalization, results communication, and discovery/visualization (Fig. 1 and Table 1).

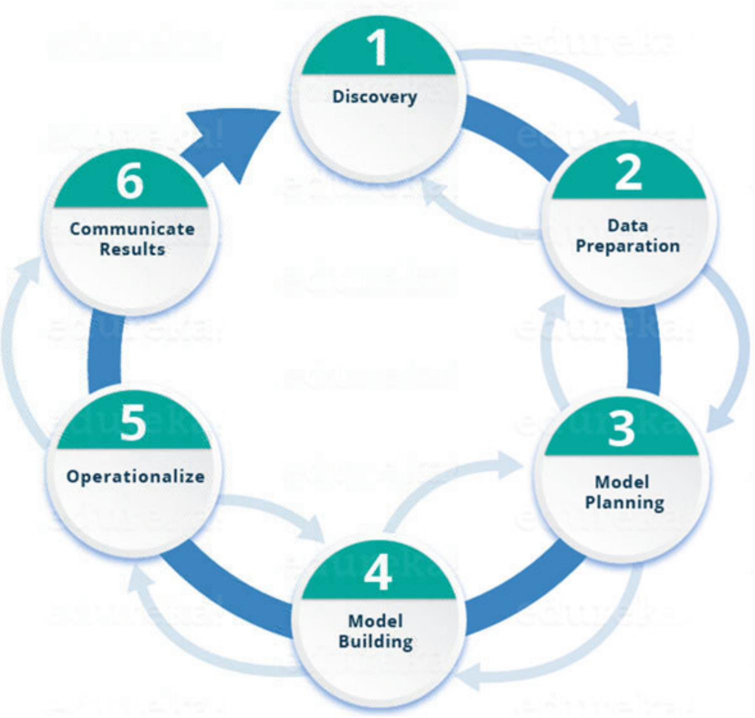

Fig. 1 Data science and its components (*Source* https://www.edureka.co/blog/what-is-data-science/)

3 Data Science: Tools

In DS, tools can be defined as an equipment or device that can be used to carry out a specific task. However, technique follows detailed scientific procedures to accomplish a particular task. Data scientists often follow operational methods/techniques on the data by using various devices as tools. The combination of tools and techniques often are used in data acquisition, refinement, manipulation, analysis, and visualization. DS tools are used at various stages of data processes such as data collection, storing, processing, analysis, and visualization. DS tools can be categorized according to important steps involved in the processing data like data acquisition, storage, extraction, cleaning, analysis, and visualization (Table 2).

Table 1 Various components of Data Science (DS) and their brief description

DS components	Description
Discovery	In DS, discovery includes problem identification, which is a fundamental and at the same time one of the important components. It includes the finding of relevant data, and their formats from numerous sources to accomplish basic necessities, priorities, and objectives of the task. While framing the research problem, the project objectives, manpower requirement, technology, timelines, budget, data, and the end goal are to be clearly identified
Data preparation	It includes data cleaning, reduction, integration, and transformation. It refers to identifying the data gaps, incorrect, unrelated parts of the data to replace, modify, update, or delete the unwanted data. After performing these tasks, the cleaned data can be used for further process, and analysis. In this phase, data engineering plays a key role in data acquiring, storing, retrieving, and transforming the data, and developing metadata of the data
Model building	Model building is a critical stage and its selection depends on the project objectives, nature, and volume of data. In this stage, depending on the model, method, or technique selected, model variables are to be differentiated. In model building, datasets are to be developed for training as well as validation purposes. Association, classification as well as clustering methods can be adopted in building the robust models by using popular model building tools
Model planning	In this phase, the relationship between the input variables needs to be clearly determined depending on the methods, and techniques adopted. The tools like exploratory data analytics, and visualization can be used to comprehend the relations between variables. The potential tools like SQL Analysis Services, R, SAS, and Python can be used in the model planning process
Model evaluation	Methods for evaluating the performance of the adopted model can be divided into two categories, namely holdout, and cross-validation, which uses a test dataset to evaluate the model performance, its accuracy, and other characteristics. At this stage, if needed the required changes can be made in the model to get the best results. In case the desired accuracy is not obtained, the model building process is to be revisited, select the appropriate model, and then perform the model evaluation to obtain the best result as per the objectives
Visualization	In DS, data visualization is an important phase and assumes greater importance, where representing data in a visual context to easily understand the importance and limitations of data. Various data visualization tools enhance the capabilities in the representation of data in a graphical form. Visualization is also an important component in data discovery, and decision-making process. Data visualization helps to translate information into easily understandable formats such as a map or graph to abstract the insights, identify the patterns, trends, and outliers in large datasets

(continued)

Table 1 (continued)

DS components	Description
Results communication	It is an important component of DS to communicate the results of the study to the intended audience to reach the goal set in the initial phase. The core objective of this phase is to bring out final technical reports by focusing on the highlights of the study. In results communication, the technical details of the task completed need to be furnished for its full deployment, implementation, and findings

4 Data Science: Techniques

The important DS techniques are AI, ML, natural learning process, data modeling, probability and statistics, predictive analysis, contextual analysis, DM, pattern recognition, and visualization (Fig. 2). Suitable DS techniques to be used to extract important information from heterogeneous structured or unstructured data.

4.1 Artificial Intelligence (AI)

AI applies the principles of science and engineering in making smart machines specifically with intelligent computer programs (McCarthy et al. 2020). It enables the machines with the ability to mimic human behavior, particularly with cognitive functions. In recent times, the scope of AI has increased many folds as the intelligence of smart machines embedded with ML capabilities created huge impacts in various fields (Davenport et al. 2020). AI gives machine thinking abilities to achieve higher efficiency in smart ways in a time-efficient and cost-effective manner. AI can make systems self-dependent and has the ability of decision making by using different algorithms. AI has the ability to augment the automation process, business outcomes, cost reduction, and revenue generation.

4.2 Internet of Things (IoT)

IoT provides adaption properties to the machine and it helps the machine to communicate with the other machine to achieve stabilization among different machines connected in the network. IoT plays an important role in application development by using data analytics tools. IoT architecture consists of perception, network, middleware, application, and business layers (Atzori et al. 2010). IoT enables the remote handling of the equipment and it increases the operating range of the network. IoT has the ability to transform the present world to realize more efficient industries, automated machinery control, smart farming, and smarter cities. IoT applications

Table 2 Data science tools and their utilities

Data science tools	Utilities
Data acquisition tools	Data can be acquired through various equipment/devices such as satellites, in situ measurements, online portals and surveys, personnel interviews, etc. Data generated through different sources need to be cleaned and converted into a usable format to perform various data analysis tasks. "CREODIAS" as a cloud platform (https://creodias.eu/) facilitates to perform operations like search, view, and process satellite products, and develop applications in EO. "WEkEO" (https://www.wekeo.eu/) provides Copernicus data covering Sentinel satellites data in various fields EO. It also offers cloud computing packages with integrated service delivery tools to develop business solutions
Data storage tools	Data storage tools are used widely for storing large amounts of data through shared computers. These tools offer a platform to integrate servers to access data seamlessly. "Microsoft Excel" is extensively used to handle small to medium size of datasets. "Apache Cassandra" tool uses SQL (Structured Query Language) and CSL (Cassandra Structure Language) to connect with the database and retrieve required data from the stored data in different servers. **"**Hadoop" has the capabilities to store a large volume of data and perform various data processing tasks in a distributed architecture. "Apache Hadoop" facilitates the storage of high volume of data among clusters of systems and performs various computations with ease. "Hive" as a data warehouse architecture facilitates to quickly query the data stored in several databases, and file systems
Data extraction tools	Data extraction tools or web scraping tools allow the users to extract information, and data automatically from websites. "OctoParse" (https://www.octoparse.com/) provides output in structured spreadsheets for easy to use and perform further operations. "Content Grabber" can be used to extract data from web sources and provide structured data as output. Web scraping tools can also be made into functional Application Programming Interfaces (APIs) for use by researchers as APIs create interfaces for programmers to easily access, and process published data
Data cleaning tools	These tools are often integrated with databases and are time-saving in searching, sorting, and filtering data (Blei and Smyth 2017). "Data Cleaner" tool works to improve the quality of data through the removal of data redundancy and transforming it into a single record. "OpenRefine" tool cleans data before performing the data transformation task to convert it into the desired form
Data analysis tools	These tools perform various data analysis tasks including data modeling to derive meaningful information from the input dataset and it helps in the decision-making process to determine the problem. "R programming language" is widely used in performing statistical computations, and graphics. "Python" programming language is also used in application development and the outputs in CSV formats. Similar data analysis tools like SAS, Jupyter, R Studio, MATLAB, and Excel were also widely used in data analysis

(continued)

Table 2 (continued)

Data science tools	Utilities
Data visualization tools	The important data visualization tools are R, Jupyter, and Tableau. "Python" provides data visualization facilities with a wide range of graphical representations of data. "Tableau" software provides various user-friendly data visualization capabilities to visualize the results in various forms. "Orange" is another data visualization tool and it supports various visualization tasks like data extraction and analysis. DataWrapper, Qlik, and Gephi are some of the popular data visualization tools that support data visualization with the excellent support of graphical representation

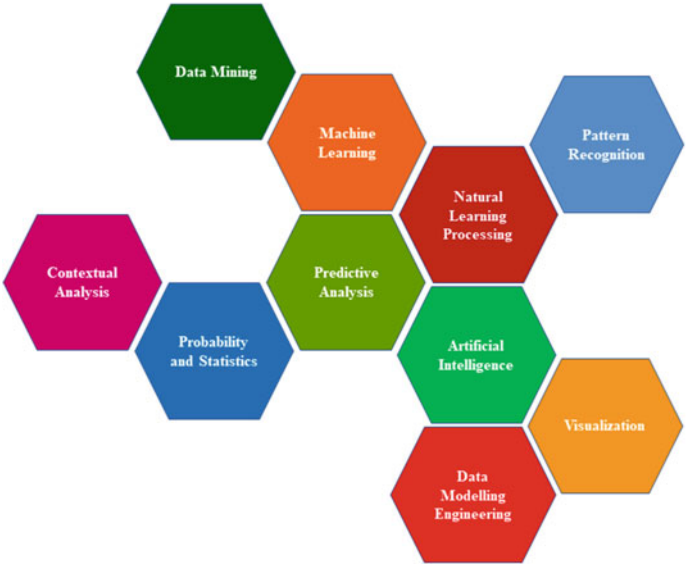

Fig. 2 Data science techniques (*Source* Sood and Rinehart 2016)

are being widely used in big data analytics, and cloud computing (Baseca et al. 2019). The capabilities of IoTs were widely applied in various fields like management, monitoring, control, and unmanned machinery systems (Ojha et al. 2015). Security services such as privacy, and authentication are some of the crucial factors for modern technologies to realize the full potential of IoT services (Atzori et al. 2017).

4.3 Data Mining (DM)

DM is a subdomain of AI and it denotes the extraction of hidden analytical information from big data (Weiss and Indurkhya 1998). DM techniques originated from the classical methods of statistics, databases, parallel computing, pattern recognition, AI, and visualization. It extracts the hitherto unknown, predictive information, hidden patterns from the available big data. It also helps in sorting the particular information from a huge dataset, and provides proper information to the particular machines. The important DM techniques like clustering, patterns recognition and classification, association, regression, outlier detection, and prediction are being widely used in various applications. Knowledge Discovery in Databases (KDD) approach of DM is useful to identify the underlying relationships, and characterize the knowledge in a given dataset (Hira and Deshpande 2015).

4.4 Data Visualization

Data visualization elucidates the significance of data by representing it in a visual context for its easy understanding. In general, the complex patterns, trends, and correlations in big data are difficult to detect and comprehend, however, by applying data visualization techniques, which can be transformed in the form of graphs, charts, scattered plots, gauges, etc. Innovative data visualization tools offer advanced functionalities and allow data analytics at several levels with the support of human intelligence, and reasoning capabilities. In agriculture, with the use of appropriate visualization techniques, it is possible to find out the patterns, connections, or similarities in the observed datasets. GPS-enabled tractors and field sensors generate voluminous spatial data that are not so easy to analyze and understand. However, the deployment of visualization techniques helps to transform them in a graphical format to easily understand and optimize the resources.

4.5 Machine Learning (ML)

ML algorithms were grouped into three categories, which include supervised, unsupervised, and reinforcement learning (RL) (Fig. 3). The supervised algorithms need output values in the training data (Berry et al. 2019), on the other hand, unsupervised algorithms need the input values in the training dataset to find out the hidden patterns in the dataset through clustering or dimension reduction algorithms (Baştanlar and Özuysal 2014). Whereas, while building the systems, RL algorithms learn from the interface with the environment by using rewards, and punishment rulesets, where every event is random (Osband et al. 2020; Busoniu et al. 2008). ML algorithms were widely applied in various fields of EO like smart agriculture, crop monitoring, yield

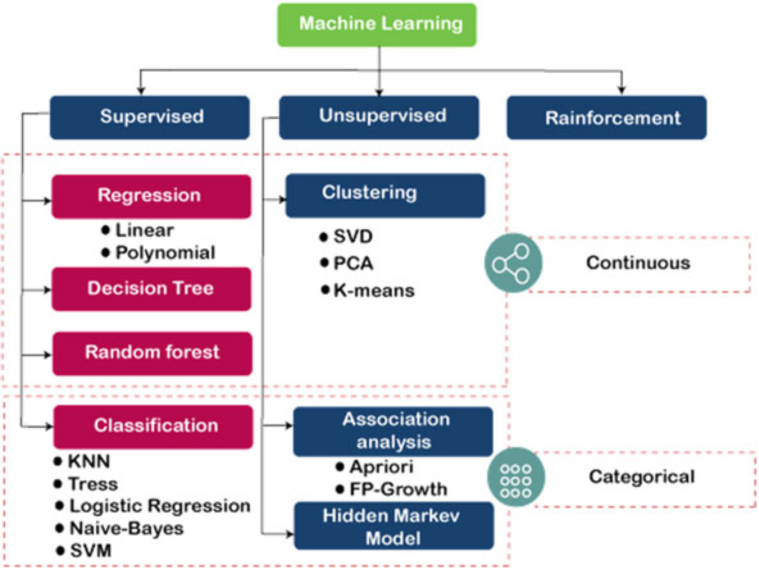

Fig. 3 Important categories of machine learning algorithms (*Source* https://www.edureka.co/blog/machine-learning-algorithms/)

prediction, disease detection, prediction of soil properties, soil and water management, etc. ML algorithms have the ability to overcome human limitations in terms of speed, accuracy, reliability, consistency, and transparency in performing the assigned tasks. The absence of quality and consistency, and any inherent biases within the dataset always pose hindrances in realizing the full potential of ML algorithms.

4.5.1 Supervised Machine Learning

Supervised learning involves training a system as per the labeled data to predict the output from every input as it trains the system according to its expectations. In this process, trained systems give predictions in classification and detect fault (Cui et al. 2019). This method helps to approximate a function between the input, and output data. Here, the system learns the training dataset's classifiers thereafter automatically applies this classification to an unknown dataset. Regression analysis (Linear and polynomial regression), Decision Tree (DT) algorithms, Random Forest (RF), and classification algorithms like K-Nearest Neighbor (K-NN), Logistic regression (LR), Naive Bayes Classifier (NBC), and Support Vector Machine (SVM) are the examples of supervised ML algorithms.

Regression Analysis

It is used to estimate the relationship among the variables, which have a reason, and result relationships (dependent and independent variable). However, multilinear regression models work with one dependent variable, and more than one independent variable. Linear regression works with straight regression lines and predicts the dependent variable outcomes based on the independent variables. Polynomial regression works with curved regression lines to predict the best fitting data points. LR works with success/failure probabilities, and binary variables to predict possible outcomes. The applications of these techniques are directly associated with discrete, and continuous distribution of variables. Regression analysis is a widely used technique to determine the correlations between several variables based on experimental or observed data (Lei et al. 2016). Selecting the right regression technique to perform the task depends on the type of dataset.

Decision Tree

DT learning is a popular classification technique used in ML applications. The DT algorithm offers support in decision-making, and it defines the logical structure, associated uncertainties, and constructive results of the decisions (Khader et al. 2013). DT algorithms contain a hierarchy of both internal as well as external nodes, which are linked by branches. In the model, a serial tree represents a logical model as a binary tree constructed using a training dataset to foresee the value of a target variable by making use of predictor variables. In the DT model, each node generates its child nodes until either the subgroups are too small to undertake the similar meaningful division or further splitting is not possible to produce statistically significant subgroups. Some sections in the sample may have outcomes in a big tree, and some of the links may generate outliers. Such branches are required to be removed. DT models are efficient in multi-feature extraction, and removing outliers without overfitting. DT algorithms like Chi-square–Automatic–Interaction–Detection (CHID), and Classification and Regression Tree (CART) are widely used to build decision trees.

Random Forest (RF)

As a “tree-based” classifier and extension to the CART algorithm, RF improves the overall model performance (Breiman 2001) and is applied to the prediction of discrete as well as categorical variables. In the RF algorithm, many trees are created, however, while growing the trees no pruning takes place. In RF for every individual tree, a subset of the prediction variables is only used. In the RF algorithm, the input dataset determines the number of predictors to be used in building each tree as well as the number of trees to be built in the forest (Liaw and Wiener 2018). In the RF