BIG DATA
ANALYTICS
For Industry Practitioners

BIG DATA ANALYTICS

For Industry Practitioners

Jafreezal Jaafar
Mohd Hilmi Hasan
Izzatdin A Aziz
Amir Ali

Universiti Teknologi PETRONAS
SERI ISKANDAR • 2023

Big Data Analytics for Industry Practitioners

© Universiti Teknologi PETRONAS, 2023
Second Print, 2023

All rights reserved. No part of this publication may be reproduced or transmitted in any form or any means, including photocopying and recording, without the written permission of the copyright holder, application for which should be addressed to Universiti Teknologi PETRONAS. Such written permission must also be obtained before any part of this publication is stored in a retrieval system of any nature. The publisher, authors, contributors and endorsers of this publication each excludes liability for loss suffered by any person resulting in any way from the use of, or reliance on this publication.

Published by
UTP Press
Universiti Teknologi PETRONAS
Owned by Institute of Technology PETRONAS Sdn. Bhd.
32610 Seri Iskandar, Perak Darul Ridzuan, Malaysia

Typeset by:
IFZAHWORKZ SOLUTIONS SDN BHD
C-20-11, Zetapark, Kompleks Danau Kota
Jalan Taman Ibu Kota, Taman Ibu Kota
53300 Kuala Lumpur, MALAYSIA

Printed in Malaysia by:
Reka Cetak Sdn. Bhd.
No. 14, Jalan Jemuju Empat 16/13D,
Seksyen 16, 40200 Shah Alam,
Selangor Darul Ehsan
Tel: +603-5510 4254 / +603-5510 4758
www.rekacetak.com

Perpustakaan Negara Malaysia Cataloguing-in-Publication Data



Cataloguing-in-Publication Data
Perpustakaan Negara Malaysia
A catalogue record for this book is available from the National Library of Malaysia

ISBN 978-967-2880-52-3

Contents

Preface ix

1 INTRODUCTION **1**
Jafreezal Jaafar
1.1 Making Sense of Data 2
1.2 Characteristics of Big Data 2
1.3 Overview of Data Science 3
1.4 What is Big Data Analytics? 4
1.4.1 Data Analytics Process 4
1.4.2 Data Analytics Types 5
1.5 Who is Data Scientist? 6
1.6 Why Big Data Analytics 6
1.7 Issues and Challenges of Big Data analytics 7
1.8 Big Data Analytic Lifecycle 9
1.9 Summary 10

2 DATA QUALITY ASSESSMENT **11**
Emelia Akashah Patah Akhir
2.1 Data Quality Management 11
2.2 Definition of Data Quality 11
2.3 Importance of Data Quality 11
2.4 Data Quality Dimensions 12
2.5 Examples of Data Quality Dimensions Using Sales Dataset 14
2.5.1 Completeness 14
2.5.2 Uniqueness 15
2.5.3 Timeliness 16
2.5.4 Validity 17

2.5.5 Accuracy 18
2.5.6 Consistency 19
2.6 Case Study: Data Quality Assessment for Oil & Gas Organisation 20
2.6.1 Selected Dimensions 21
2.6.2 Completeness 21
2.6.3 Accuracy 21
2.6.4 Consistency 22
2.6.5 Timeliness 22
2.6.6 Data Quality Assessment Result 22
2.6.7 Case Study Conclusion 24
2.7 Summary 25

3 BIG DATA ANALYSIS 26
Emelia Akashah Patah Akhir
3.1 Descriptive Analytics 27
3.1.1 Descriptive Statistics 28
3.2 Predictive Analytics 34
3.2.1 Time Series Analysis 35
3.3 Prescriptive Analytics 38
3.3.1 Data-driven decision making 40
3.3.2 Covid-19 Producers, Storage, and Centers of Vaccination Model 41
3.4 Summary 46

4 SUPERVISED LEARNING 48
Said Jadid Abdulkadir
4.1 Introduction 48
4.2 Regression 49
4.2.1 Linear Regression 50
4.2.2 Logistic Regression 50
4.2.3 Polynomial Regression 51
4.2.4 Stepwise Regression 52
4.2.5 Ridge Regression 52
4.2.6 Lasso Regression 52
4.2.7 ElasticNet Regression 53
4.3 Classification 53
4.3.1 Logistic Regression 54
4.3.2 Naïve Bayes 54
4.3.3 K-Nearest Neighbors 55
4.3.4 Decision Tree 55
4.3.5 Random Forest 55
4.3.6 Support Vector Machine 55
4.4 Summary 56

5 UNSUPERVISED LEARNING **57**
Said Jadid Abdulkadir
5.1 Introduction 57
5.2 Clustering Models 58
5.2.1 K-Mean Clustering 58
5.2.2 Mean Shift Clustering 59
5.2.3 Density-Based Spatial Clustering of Applications with Noise (DBSCAN) 60
5.3 Dimensionality Reduction 60
5.3.1 Principal Component Analysis (PCA) 61
5.3.2 Factor Analysis (FA) 61
5.3.3 Linear Discriminant Analysis (LDA) 61
5.4 Anomaly Detection 62
5.4.1 Isolation Forest 62
5.5 Summary 63

6 PYTHON AND DATA PRE-PROCESSING **54**
Mohd Hilmi Hasan
6.1 Big Data 64
6.1.1 Data Dimensionality 65
6.1.2 Data Augmentation 65
6.2 Introduction of Python 66
6.2.1 Data Frame 66
6.2.2 Visualisation 70
6.2.3 Managing Empty Cells 74
6.2.4 Normalisation 77
6.2.5 Standardisation 78
6.2.6 Splitting Data 78
6.3 Summary 79

7 DATA VISUALISATION **80**
Norshakirah Aziz
7.1 Introduction 80
7.2 How to choose right visualisation for your data? 82
7.3 BI solution for Oil and Gas Industry 85
7.3.1 Case Study 1: Colorado Oil & Gas dashboard 85
7.3.2 Case study 2: Bold BI Analytics 86
7.4 Summary 87

8 APPLICATION OF BIG DATA ANALYTICS **89**
Izzatdin Abdul Aziz, Amir Ali
8.1 Background 89
8.2 Methodology 91
8.3 Error Measures and Scores 92
8.4 Data Preparation 93
8.5 Data Cleansing 93
8.6 Data Aggregation 96
8.7 Modeling 98
8.8 Time to Fail Range 98
8.9 Tag Counts 99
8.10 Tag Selection 100
8.11 Model Selection Based on Error Measure 101
8.12 Model Execution and Presentation 103
8.13 Validation 112
8.14 Deliberation 113
8.15 Summary 114

Authors' Biodata 115
Bibliography 118
Index 123

Preface

Data is anything that may be used by a computer e.g., number, letter, symbol, voice, or image. Meanwhile, Big Data is generally defined as extraordinarily huge data collections that have outgrown standard data processing technologies. Big Data can be structured, unstructured, or semi-structured. The data (or Big Data) possess valuable insights whereby human can make sense out of them for enhancing business decision making, getting public opinion/sentiment, and many other purposes. Therefore, there comes the Big Data Analytics technology that has the capability to carry out this generation of sense out of the Big Data. The technology, in general, is based on various Artificial Intelligence-based techniques.

This book covers the topics on Big Data Analytics which include theory and applications. Chapter 1 introduces what Big Data is and what are its characters. It further discusses about how to analyse Big Data using Big Data Analytics and the role of Data Science. It also briefly describes the responsibilities of Data Analyst, followed by the issues and challenges of Big Data. It guides how to make sense out of the Big Data. Finally, it describes the life cycle of Big Data Analytics.

Chapter 2 discusses about data quality that becomes one of the critical phases in any decision making or when the data is going to be used to produce another result or analysis e.g.: predictive analytics. The main idea of implementing data quality assessment is to avoid severe consequences, e.g., decision making based on unhealthy data or no quality data will lead to inaccurate decision being made that will harm the organisation. Thus, organisations' attention in assessing their data quality is very crucial to improve their informed decision making.

Chapter 3 presents Big Data analysis that covers mathematical tasks, statistical evaluation, and algorithms. This chapter also covers the discussions on descriptive, predictive, and prescriptive analytics. These three types of analysis are the implemented/utilised widely in the area of Big Data Analytics.

Chapter 4 covers supervised machine learning algorithms whereby a clear description about them is presented. The chapter also discusses about the most common tasks in Big Data analytics i.e., regression and classification. Seven types of regression are explained as well as six types of most widely used classification algorithms. This chapter serves as a guideline for researchers working on the domain of supervised learning where the inputs

are provided as a labelled dataset, a model can learn from it, the desired output for each dataset is provided as well. The main purpose of supervised learning is to learn the function of mapping the input to the desired output for prediction and classification problems.

Chapter 5 covers unsupervised machine learning algorithms whereby a clear description of them is explained. The chapter also presents the most common tasks used in Big Data analytics which are clustering, dimensionality reduction and anomaly detection. Three popular clustering algorithms are explained as well as three dimensionality reduction techniques. Additionally, anomaly detection is described. This chapter serves as a guideline for researchers working on the domain of unsupervised learning where the inputs are provided as a labelled dataset, a model can learn from it, the desired output for each dataset is not provided.

Chapter 6 presents Python and data pre-processing whereby it gives the basic concepts of Python programming, reading data and storing it into a data frame, removing the Null values and visualising the data. It further briefs about data normalisation, data scaling or standardisation and data splitting using python codes.

Chapter 7 describes the history of data visualisation. Furthermore, the chapter describes the selection of right visualisation according to the data in hand. Hence, various types of visualisation and their purposes and properties are presented in this chapter. Lastly, the chapter presents the visualisation application in the Oil and Gas industry.

Chapter 8 covers the procedures used to apply data analytics to a real-world oil and gas issue, predicting machine failure based on known past failure instances. This chapter's case study is based on an offshore gas compressor. Field managers and offshore and onshore operation engineers with extensive expertise in managing vital equipment in oil and gas platforms, particularly gas compressors, assist in this task. This chapter assumes the dataset has been thoroughly preprocessed. Preprocessing data was discussed previously in this book's chapter. In this case, the data preparation stages are given briefly.

Overall, the Big Data analytics techniques and algorithms are important to generate valuable senses out of our data. A lot of challenges may occur in conducting the works related to Big Data analytics, particularly in dealing with huge size of data and uncertainty/incompleteness of data in their raw format. The contents of this book are hoped to assist the readers in conducting works and research with enthusiasm, tenacity, and dedication to develop new data analytics techniques to provide new solutions to keep up with the ever-changing challenges and issues.

Jafreezal Jaafar
Mohd Hilmi Hasan
Izzatdin A Aziz
Amir Ali
Seri Iskandar, 2022