DATA SCIENCE with KNIME

Data Exploration, Machine Learning and Visualization using CODELESS Visual Programming

Dr. NICKHOLAS ANTING, Ph.D.

NASPSOFT

DATA SCIENCE with KNIME: Data Exploration, Machine Learning and Visualization using CODELESS Visual Programming

by Dr. Nickholas Anting Anak Guntor, Ph.D.

Published by,
NASPSOFT Intelligence Academy
83300 Batu Pahat, Johor, Malaysia.
www.naspsoft.com

ISBN: 978-629-96941-0-6

First published: August 2022

Perpustakaan Negara Malaysia Cataloguing-in-Publication Data

Nickholas Anting, 1988-
Data Science with KNIME: Data Exploration, Machine Learning and Visualization using CODELESS Visual Programming / Nickholas Anting.
ISBN 978-629-96941-0-6
1. Data sets.
2. Visual Programming (Computer science).
3. Machine Learning.
4. Programming by example (Computer science).
I. Title.
005.7

Printed by Swan Printing Sdn. Bhd.

The book is dedicated to my family, mentors, collaborators, friends, and students for continuous support and encouragement for success.

About the Author

Dr. Nickholas Anting Anak Guntor, Ph.D., is the founder, director and consultant of the NASPSOFT Intelligence Academy, a consulting and training company specializing in data science, machine learning and artificial intelligence. Dr. Nickholas is a Ph.D. holder in Civil Engineering from Universiti Teknologi Malaysia. With actual data science and machine learning experience in research, he is able to deliver valuable knowledge to the student with his effective method and high-quality educational materials. He is a certified KNIME trainer for basic and advanced levels at NASPSOFT and trains professionals from various industries, including finance, engineering, medical, sport, marketing and education.

Dr. Nickholas Anting is also a full-time senior lecturer and researcher at the Faculty of Civil Engineering and Built Environment, Universiti Tun Hussein Onn Malaysia. The author specialises in advanced computational engineering for structural performance analysis. He teaches courses related to Data Science, Civil Engineering and Computer Programming for the faculty.

From 2019 to 2021, Dr. Nickholas was appointed as a data science consulting committee for the Ministry of Higher Education, Malaysia, on a big data analytics project to build a predictive intelligence model on student placement in higher learning institutions. Dr. Nickholas and his team are actively working with local and international organizations in Malaysia to establish data department, produce data talents, and create impacts with insights and data intelligence services.

Dr. Nickholas is also the author of "*Machine Learning with KNIME Analytics Platform*" and "*Programming for Beginners with Python*".

The author's website is *www.nickholasanting.com*. It is a platform the author uses to share knowledge, information and tips. Dr. Nickholas Anting is reachable for consultation by email at *nickholasanting@gmail.com*.

About the Technical Editor

Dr. Alvin Lim Meng Siang, Ph.D., is a full-time associate professor at the Faculty of Civil Engineering and Built Environment, Universiti Tun Hussein Onn Malaysia. The editor's background is dynamic computational in geotechnical modelling. He teaches subjects related to Civil Engineering, Data Science and Computer Programming to the faculty. Dr. Alvin is also an active writer and published two books, entitled "*Visual Analytics with Power BI*" and "*Programming for Beginners with Python*". He is also a certified KNIME trainer for basic and advanced levels.

About the Reviewers

Victor Palacios has a master's degree in data science and teaches courses on data science and machine learning at KNIME. He is currently a data scientist at KNIME. His initial projects and interests were in the deception detection space and healthcare, but his love of data science originally blossomed from natural language processing in the translation field.

Dr. Rosaria Silipo, PhD, now head of data science evangelism at KNIME, has spent 25+ years in applied AI, predictive analytics and machine learning at Siemens, Viseca, Nuance Communications, and private consulting. Sharing her practical experience in a broad range of industries and deployments, including IoT, customer intelligence, financial services, social media, and cybersecurity, Rosaria has authored 50+ technical publications, including her recent books: "Guide to Intelligent Data Science" (Springer) and "Codeless Deep Learning with KNIME" (Packt).

Ing. Barbora Stetinova, MBA, is an expert in data science, artificial intelligence and business intelligence. Currently, she works as AI and RPA process analyst at Fortuna Entertainment Group. At the same time, she is also a data science consultant, trainer and conference speaker at Leadership Synergy Community. Barbora obtained her MBA from the University of New York in Prague. In 2019, she crowned the winner of the Data Cup 2019 Czech Republic.

Acknowledgements

Writing a book and getting it published is always an exciting journey. Many individuals contribute to making this book available for readers. I want to take this opportunity to thank a number of important people and organizations for their contributions, directly or indirectly, who made this book possible.

First, I want to thank Dr. Alvin Lim, my technical editor, who spent a lot of time reading, checking and advising the book's contents. Dr. Alvin is always a person who has done his job professionally.

Special thanks to the team of reviewers, Victor, Rosaria and Barbora, who are willing to give feedback and suggestions to ensure the book is suitable for publication. Their recommendations are significant for the book since they are actual data scientists with expertise in the KNIME Analytics Platform.

Also, I would like to thank Universiti Tun Hussein Onn Malaysia for the opportunity to serve as an educator and researcher, which taught me a lot about writing and publication.

Finally, I want to thank my parents and my wife, Evyna Ernestia, for all the continuous support they have given me. I love you all.

Preface

Rapid and continuous innovation in digital technologies has accelerated vast amounts of data produced daily. Regardless of how people define it, the impact of Big Data is a reality. The term big data is not only limited to the size of the data but also describes the complexity of raw information generated by digital devices. Giant technology corporations are at the forefront of leading in utilizing technology to harness the enormous value and opportunities in Big Data. At the same time, others are taking a leap of faith to adopt a data-driven culture in their organization to remain competitive.

Data Science is an analytics process capturing value and insights from Big Data. It is a body of knowledge that has merged and popularised in recent years due to the increased demand for data talents. Moreover, many courses and self-learning materials have been developed, and even academic programs have been offered by higher learning institutions to produce more data science professionals.

With teaching in mind, this book is written for professionals and students who want to learn data science with practical hands-on and case studies. Realising that writing a programming code is a challenge for newcomers, this book teaches how to do end-to-end data science projects with visual programming tools, the KNIME Analytics Platform. Visual programming is a CODELESS technique that applies drag-and-drop to the functional nodes to execute specific tasks. It is a total replacement of programming tools, such as Python and R.

I hope this book helps to accelerate your efforts to be a data professional. All the best to all readers.

Dr. Nickholas Anting, Ph.D.
Director & Author, NASPSOFT Intelligence Academy
August 2022

Contents

Introduction

This book provides practical approaches to learning data science with visual programming tools, KNIME Analytics Platform. Visual programming is a CODELESS technique that applies drag, drop and execute mechanisms into the functional nodes. In other words, NO CODING requires. This book is written for professionals and students who want to learn practical data science without worrying about coding. The book is packed with theories and practical hands-on that guide readers through step-by-step procedures to execute actual data science tasks with KNIME software.

The contents of the book are structured into nine chapters. Chapter 1 introduces the reader to the fundamental of data science and big data, real applications of data science in industries and the data analytics lifecycle. Chapter 2 teaches readers technical skills to use the KNIME Analytics Platform as a programming tool for data science tasks. It is a primary tool to execute all examples and hands-on in this book.

Chapter 3 concentrates on exploratory data analysis, teaching the readers to perform data exploration using univariate and multivariate analysis. This chapter also covers the topic of inferential statistics to conduct hypothesis testing.

Chapters 4 through 8 focus on predictive analytics with machine learning. The chapters cover a range of advanced analytical machine learning models, including regression, classification and clustering.

Chapter 9 is a chapter dedicated to demonstrating the actual analytics project with the KNIME Analytics Platform. Real-business datasets are used for the practical hands-on in this chapter.

All materials used for this book, including datasets and saved files, can be downloaded from `https://bit.ly/naspsoft_knimebook`.

Let’s get started.