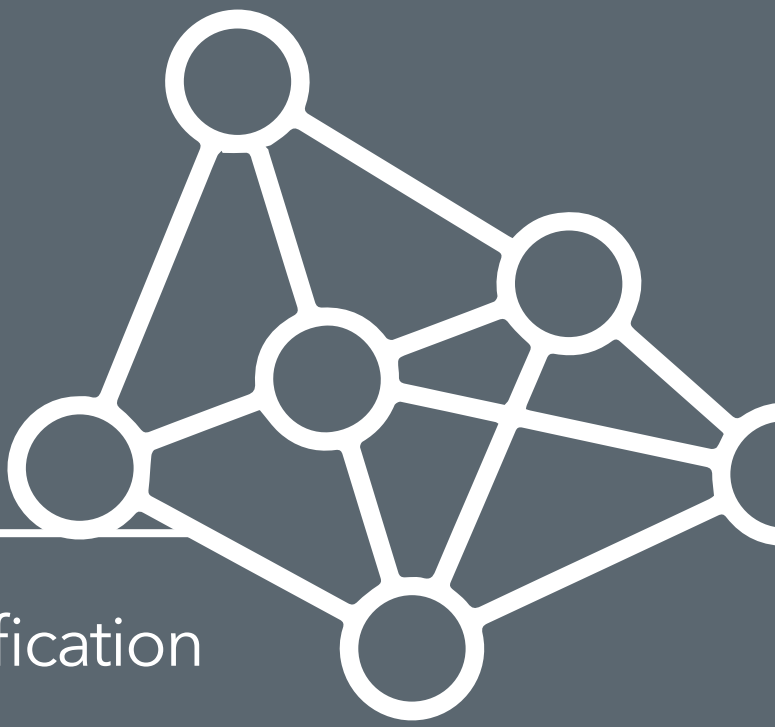
NCC education®
Awarding Great British Qualifications

SHORT COURSE IN DATA SCIENCE: Expert



Qualification Unit Specification
2020/21

Modification History

Version	Revision Description
V1.0	For release

Published by: NCC Education Limited, The Towers, Towers Business Park, Wilmslow Road, Didsbury, Manchester M20 2EZ UK

Tel: +44 (0) 161 438 6200 Fax: +44 (0) 161 438 6240 Email: info@nccedu.com
http://www.nccedu.com

1. About NCC Education

NCC Education is a UK-based awarding body, active in the UK and internationally. Originally part of the National Computing Centre, NCC Education started offering IT qualifications in 1976 and from 1997 developed its Higher Education portfolio to include Business qualifications, IT qualifications for school children and a range of Foundation qualifications.

With Centres in over forty countries, four international offices and academic managers worldwide, NCC Education strives to employ the latest technologies for learning, assessment and support. NCC Education is regulated and quality assured by Ofqual (the Office of Qualifications and Examinations Regulation, see www.ofqual.gov.uk) in England and Northern Ireland.

Overview and Objectives

This **Short Course in Data Science: Expert** focuses Data Science team leaders who are looking to step into senior management role. Participants will learn the structure of the data science pipeline, the goals for each stage, and how to keep the team on target.

This course provides knowledge on all the key techniques such as Predictive Analysis, Big Data, Web Scraping, Data Mining and many more.

Hardware and Software Requirements

Hardware: Learners need access to computers with Internet access.

Software: Learners must have access to Hadoop.

Short Course in Data Science: Expert

Title:	Short Course in Data Science: Expert

Guided Learning Hours	120 hours

Learning Outcomes; The Learner will:	**Objectives;** The Learner can:
1. Be able to explain ethical and privacy issues in data science conduct and apply etical practices	1.1 Explain ethical and privacy issues in data science conduct. 1.2 Apply ethical practices on data science project.
2. Undertand the life of a data scientist – identify problem, define question, define ideal dataset, obtain and analyse data, and distribute results.	2.1 Explain the whole data science process and how its different components interact. 2.2 Explain the skills needed to be a data scientist. 2.3 Identify the different fields interlinked with data science.
3. Understand the basics of optimization and how to formulate the problem.	3.1 Explain optimisation and operation research. 3.2 Identify the constraints and explain different classes in an optimisation problem.
4. Be able to formulate the linear programming and the integer programming	4.1 Explain and differentiate integer programming problem and linear programming problem. 4.2 Solve optimisation problem by simplex method of linear programming. 4.3 Explain the characteristics of linear programming problem.
5. Understand big data concepts and the importance of data mash-up	5.1 Discuss the generation and characteristics of big data. 5.2 Explain Hadoop distribution in detail and Hadoop architecture. 5.3 Discuss data mashups and the importance of data mashups. 5.4 Explain the implementation of web mashups and the architecture of wen mashups.
6. Be able to perform data collection using web scraping tool	6.1 Explain web scrapping and identify alternatives to web scrapping. 6.2 Explain document object model and regular expressions. 6.3 Use different libraries for web scrapping.
7. Be able to make predictions using multiple regression models	7.1 Explain predictive analytics. 7.2 Explain predictive modelling techniques. 7.3 Explain k-nearest neighbour's classification and k-means clustering.

8. Understand data mining process	8.1 Explain data mining foundations. 8.2 Explain data mining models and tasks. 8.3 Identify the purpose of data mining. 8.4 Explain cluster analysis. 8.5 Explain the partitioning methods, hierarchical methods, grid-based methods, and model-based clustering algorithms etc.
9. Be able to perform model evaluation and deployment	9.1 Explain modelling, evaluation and deployment. 9.2 Model selection, evaluation and deployment. 9.3 Review data science project.

Syllabus

Topic No	Title	Content
1	Data Science Process	• What is Data Science? • What do Data Scientists do? • How to be a data scientist. • What should you learn? • Data Science Process • Data Science Ethics • Data Scientific Method ***Learning Outcome: 1, 2***
2	Optimisation and How to Formulate the Problem	• Optimisation & Operation Research • Mathematical Optimisation Problem • Classification of optimisation problems • Integer Programming Problem • Stochastic Programming Problem • Linear Programming I: Simplex method • Standard Form of a Linear Programming Problem • Characteristics of a Linear Programming Problem • Transformation of LP Problems into Standard Form ***Learning Outcome: 3 & 4***
3	Big Data Concepts	• What is big data? • Big data characteristics • Big Data: 6V's • Cloud Computing • Why Study Big Data Technologies? • Big Data Open Source Tools • Philosophy to Scale for Big Data Processing • What is Hadoop? • Why use Hadoop? • What are the core parts of a Hadoop distribution? • Common Hadoop Distributions ***Learning Outcome: 5***

4	Data Mashups and Big Data Infrastructure	• Data Analytics and Mash-up • Data/Web mashup • Architecture of web mashup • Implementation Architecture • Why Cloud Computing? • Advantages and Disadvantages of Cloud Computing • How Cloud Computing Works • Challenges of Cloud Computing • Layers of Cloud Computing • Components of Cloud Computing • Big Data • Hadoop Architecture • Hadoop With Big Data • Map Reduce • Why Data Analytics? • Types of Data Analytics • Big Data Analytics ***Learning Outcome: 5***
5	Data Collection with Web Scrapping Tools	• What's wrong with scraping? • Before writing a scraper • Alternatives to scraping • Why scrape web? • Harvesting options • Watch out for spider traps! • Web scraping, the easy way • Document Object Model (DOM) • Regular Expressions ***Learning Outcome: 6***

6	Introduction to Predictive Analysis and Prediction to Multiple Regression Model	• Data Mining Foundations • Problem Types and Model Paradigms • Model Performance Considerations • Recognising a Strong Model • K-Nearest Neighbours Classification • Decision Tree Models • K-Means Clustering • Multiple Regression • Formal Definition of the model • Estimating the parameters of the model • Analysis of the variance table • F-test for the overall fit of the model • Interval estimation • Selecting the best Regression Equation • Example: Sales Forecasting • Interpreting the final model ***Learning Outcome: 7***
7	Introduction to Logistic Regression	• What is Logistic Regression • Questions • Assumptions • Background • Plain old regression • An alternative – the ogive function • The Logistics Function • The Logit • Conversion ***Learning Outcome: 7***

8	Time Series & Modelling Predictive Modelling Techniques	• Objective of Time Series Analysis • Classical Decomposition: An Example • Trend • Residuals • Trend and Residual Variation • Time Series Models • Gaussian White Noise • Random Walk • Time Series Modelling • Nonlinear transformations • Differencing and Trend • Differencing and Seasonal Variation • What are predictive analytics? • Why predictive analytics? • What do predictive analytics do? • Predictive Modelling Techniques ***Learning Outcome: 7***

9	Data Mining, Data Exploration, Cleaning and Visualisation	<u>First Lecture:</u> • What is Data Mining? • Data Mining Definitions • Why Mine Data? Commercial Viewpoint • Why Mine Data? Scientific Viewpoint • Database Processing vs. Data Mining Processing • Data Mining: Classification Schemes • Decisions in Data Mining • Data Mining Tasks • Data Mining Models and Tasks • Classification: Definition • Classification: Applications • Classification Techniques <u>Second Lecture:</u> • Why do we need data mining? • The data analysis pipeline • Data Quality • Sampling • Types of Sampling • A Data Mining Challenge • Data Collection • Mining Task • Exploratory analysis of data • Post Processing ***Learning Outcome: 8***

10	Data Transformation and Reduction	• Data Transformation • Data Transformation Challenges • Data Transformation Strategies • Normalisation • Normalisation Methods • Data Reduction Strategies • The Curse of dimensionality • The Curse of Dimensionality: Principal Components Analysis • The Curse of Dimensionality: Factor Analysis • Data Sampling • Binning and Reduction of Cardinality ***Learning Outcome: 9***
11	Cluster Analysis	• What is Cluster Analysis? • Types of Data in Cluster Analysis • A Categorisation of Major Clustering Methods • Partitioning Methods • Hierarchical Methods • Grid-Based Methods • Model-Based Clustering Methods • Outlier Analysis ***Learning Outcome: 8***

12	Decision Tree and Model Evaluation & Deployment	• Decision Tree and Classification Task
		• Building Decision Tree
		• Built Decision Tree Algorithm
		• Node Splitting in BuildDT Algorithm
		• Entropy and its Meaning
		• Decision Tree Induction Techniques
		• CRISP-DM Phases
		• Modelling
		  • Select Modelling Technique
		  • General Test Design
		  • Build Models
		  • Access Models
		• Evaluation
		  • Evaluate Results
		  • Review Process
		  • Determine next steps
		• Deployment
		  • Plan deployment
		  • Plan Monitoring and Maintenance
		  • Produce Final Report
		  • Review Project
		Learning Outcome: 9