**Module 1: From Problem to Approach and From Requirements to Collection**

* Video: Course Introduction
* Reading: Helpful Tips for Course Completion
* Reading: Syllabus

**Lesson 1: From Problem to Approach**

* Video: Data Science Methodology Overview
* Video: Business Understanding
* Video: Analytic Approach
* Hands-on Lab: From Problem to Approach
* Reading: Lesson 1 Summary: From Problem to Approach
* Practice Quiz: From Problem to Approach
* Glossary: From Problem to Approach
* Graded Quiz: From Problem to Approach

**Lesson 2: From Requirements to Collection**

* Video: Data Requirements
* Video: Data Collection
* Hands-on Lab: From Requirements to Collection
* Reading: Lesson 2 Summary: From Requirements to Collection
* Practice Quiz: From Requirements to Collection
* Glossary: From Requirements to Collection
* Graded Quiz: From Requirements to Collection

**Module 2: From Understanding to Preparation and from Modeling to Evaluation**

**Lesson 1: From Understanding to Preparation**

* Video: Data Understanding
* Data Preparation - Concepts
* Data Preparation - Case Study
* Hands-on Lab: From Understanding to Preparation
* Reading: Lesson 1 Summary: From Understanding to Preparation
* Practice Quiz: From Understanding to Preparation
* Glossary: From Understanding to Preparation
* Graded Quiz: From Understanding to Preparation

**Lesson 2: From Modeling to Evaluation**

* Video: Modeling - Concepts
* Video: Modeling - Case Study
* Video: Evaluation
* Hands-on Lab: From Modeling to Evaluation
* Reading: Lesson 2 Summary:From Modeling to Evaluation
* Practice Quiz: From Modeling to Evaluation
* Glossary: From Modeling to Evaluation
* Graded Quiz: From Modeling to Evaluation

**Module 3: From Deployment to Feedback**

* Video: Deployment
* Video: Feedback
* Video: Storytelling
* Video: Course Summary
* Reading: Module 3 Summary:From Deployment to Feedback
* Practice Quiz: From Deployment to Feedback
* Glossary: From Deployment to Feedback
* Graded Quiz: From Deployment to Feedback

**Module 4: Final Project and Assessment**

**Final Project**

* Video: Introduction to CRISP-DM
* Reading: Final Assignment Overview
* Peer Review: Final Assignment

**Course Summary and Final Quiz**

* Reading: Review What You Learned
* Graded Quiz: Final Quiz

**Course Wrap Up**

* Reading: Congratulations and Next Steps
* Reading: Thanks from the Course Team
* Reading: IBM Digital Badge

| **Term** | **Definition** |
| --- | --- |
| **Analytic Approach** | The process of selecting the appropriate method or path to address a specific data science question or problem. |
| **Analytics** | The systematic analysis of data using statistical, mathematical, and computational techniques to uncover insights, patterns, and trends. |
| **Business Understanding** | The initial phase of data science methodology involves seeking clarification and understanding the goals, objectives, and requirements of a given task or problem. |
| **Clustering Association** | An approach used to learn about human behavior and identify patterns and associations in data. |
| **Cohort** | A group of individuals who share a common characteristic or experience is studied or analyzed as a unit. |
| **Cohort study** | An observational study where a group of individuals with a specific characteristic or exposure is followed over time to determine the incidence of outcomes or the relationship between exposures and outcomes. |
| **Congestive Heart Failure (CHF)** | A chronic condition in which the heart cannot pump enough blood to meet the body's needs, resulting in fluid buildup and symptoms such as shortness of breath and fatigue. |
| **CRISP-DM** | Cross-Industry Standard Process for Data Mining is a widely used methodology for data mining and analytics projects encompassing six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. |
| **Data analysis** | The process of inspecting, cleaning, transforming, and modeling data to discover useful information, draw conclusions, and support decision-making. |
| **Data cleansing** | The process of identifying and correcting or removing errors, inconsistencies, or inaccuracies in a dataset to improve its quality and reliability |
| **Data science** | An interdisciplinary field that combines scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. |
| **Data science methodology** | A structured approach to solving business problems using data analysis and data-driven insights. |
| **Data scientist** | A professional using scientific methods, algorithms, and tools to analyze data, extract insights, and develop models or solutions to complex business problems. |
| **Data scientists** | Professionals with data science and analytics expertise who apply their skills to solve business problems. |
| **Data-Driven Insights** | Insights derived from analyzing and interpreting data to inform decision-making |
| **Decision tree** | A supervised machine learning algorithm that uses a tree-like structure of decisions and their possible consequences to make predictions or classify instances. |
| **Decision Tree Classification Model** | A model that uses a tree-like structure to classify data based on conditions and thresholds provides predicted outcomes and associated probabilities. |
| **Decision Tree Classifier** | A classification model that uses a decision tree to determine outcomes based on specific conditions and thresholds. |
| **Decision-Tree Model** | A model used to review scenarios and identify relationships in data, such as the reasons for patient readmissions |
| **Descriptive approach** | An approach used to show relationships and identify clusters of similar activities based on events and preferences |
| **Descriptive modeling** | Modeling technique that focuses on describing and summarizing data, often through statistical analysis and visualization, without making predictions or inferences |
| **Domain knowledge** | Expertise and understanding of a specific subject area or field, including its concepts, principles, and relevant data |
| **Goals and objectives** | The sought-after outcomes and specific objectives that support the overall goal of the task or problem. |
| **Iteration** | A single cycle or repetition of a process often involves refining or modifying a solution based on feedback or new information. |
| **Iterative process** | A process that involves repeating a series of steps or actions to refine and improve a solution or analysis. Each iteration builds upon the previous one. |
| **Leaf** | The final nodes of a decision tree where data is categorized into specific outcomes. |
| **Machine Learning** | A field of study that enables computers to learn from data without being explicitly programmed, identifying hidden relationships and trends. |
| **Mean** | The average value of a set of numbers is calculated by summing all the values and dividing by the total number of values. |
| **Median** | When arranged in ascending or descending order, the middle value in a set of numbers divides the data into two equal halves. |
| **Model (Conceptual model)** | A simplified representation or abstraction of a real-world system or phenomenon used to understand, analyze, or predict its behavior. |
| **Model building** | The process of developing predictive models to gain insights and make informed decisions based on data analysis. |
| **Pairwise comparison (correlation)** | A statistical technique that measures the strength and direction of the linear relationship between two variables by calculating a correlation coefficient. |
| **Pattern** | A recurring or noticeable arrangement or sequence in data can provide insights or be used for prediction or classification. |
| **Predictive model** | A model used to determine probabilities of an action or outcome based on historical data. |
| **Predictors** | Variables or features in a model that are used to predict or explain the outcome variable or target variable. |
| **Prioritization** | The process of organizing objectives and tasks based on their importance and impact on the overall goal. |
| **Problem solving** | The process of addressing challenges and finding solutions to achieve desired outcomes. |
| **Stakeholders** | Individuals or groups with a vested interest in the data science model's outcome and its practical application, such as solution owners, marketing, application developers, and IT administration. |
| **Standard deviation** | A measure of the dispersion or variability of a set of values from their mean; It provides information about the spread or distribution of the data. |
| **Statistical analysis** | Stand deviations are applied to problems that require counts, such as yes/no answers or classification tasks. |
| **Statistics** | The collection, analysis, interpretation, presentation, and organization of data to understand patterns, relationships, and variability in the data. |
| **Structured data (data model)** | Data organized and formatted according to a predefined schema or model and is typically stored in databases or spreadsheets. |
| **Text analysis data mining** | The process of extracting useful information or knowledge from unstructured textual data through techniques such as natural language processing, text mining, and sentiment analysis. |
| **Threshold value** | The specific value used to split data into groups or categories in a decision tree. |

Data Science Methodology

| **Term** | **Definition** |
| --- | --- |
| **Analytics team** | A group of professionals, including data scientists and analysts, responsible for performing data analysis and modeling. |
| **Data collection** | The process of gathering data from various sources, including demographic, clinical, coverage, and pharmaceutical information. |
| **Data integration** | The merging of data from multiple sources to remove redundancy and prepare it for further analysis. |
| **Data Preparation** | The process of organizing and formatting data to meet the requirements of the modeling technique. |
| **Data Requirements** | The identification and definition of the necessary data elements, formats, and sources required for analysis. |
| **Data Understanding** | A stage where data scientists discuss various ways to manage data effectively, including automating certain processes in the database. |
| **DBAs (Database Administrators)** | The professionals who are responsible for managing and extracting data from databases. |
| **Decision tree classification** | A modeling technique that uses a tree-like structure to classify data based on specific conditions and variables. |
| **Demographic information** | Information about patient characteristics, such as age, gender, and location. |
| **Descriptive statistics** | Techniques used to analyze and summarize data, providing initial insights and identifying gaps in data. |
| **Intermediate results** | Partial results obtained from predictive modeling can influence decisions on acquiring additional data. |
| **Patient cohort** | A group of patients with specific criteria selected for analysis in a study or model. |
| **Predictive modeling** | The building of models to predict future outcomes based on historical data. |
| **Training set** | A subset of data used to train or fit a machine learning model; consists of input data and corresponding known or labeled output values. |
| **Unavailable data** | Data elements are not currently accessible or integrated into the data sources. |
| **Univariate** | Modeling analysis focused on a single variable or feature at a time, considering its characteristics and relationship to other variables independently. |
| **Unstructured data** | Data that does not have a predefined structure or format, typically text images, audio, or video, requires special techniques to extract meaning or insights. |
| **Visualization** | The process of representing data visually to gain insights into its content and quality. |

Data Preparation

| **Term** | **Definition** |
| --- | --- |
| **Automation** | Using tools and techniques to streamline data collection and preparation processes. |
| **Data Collection** | The phase of gathering and assembling data from various sources. |
| **Data Compilation** | The process of organizing and structuring data to create a comprehensive data set. |
| **Data Formatting** | The process of standardizing the data to ensure uniformity and ease of analysis. |
| **Data Manipulation** | The process of transforming data into a usable format. |
| **Data Preparation** | The phase where data is cleaned, transformed, and formatted for further analysis, including feature engineering and text analysis. |
| **Data Preparation** | The stage where data is transformed and organized to facilitate effective analysis and modeling. |
| **Data Quality** | Assessment of data integrity and completeness, addressing missing, invalid, or misleading values. |
| **Data Quality Assessment** | The evaluation of data integrity, accuracy, and completeness. |
| **Data Set** | A collection of data used for analysis and modeling. |
| **Data Understanding** | The stage in the data science methodology focused on exploring and analyzing the collected data to ensure that the data is representative of the problem to be solved. |
| **Descriptive Statistics** | Summary statistics that data scientists use to describe and understand the distribution of variables, such as mean, median, minimum, maximum, and standard deviation. |
| **Feature** | A characteristic or attribute within the data that helps in solving the problem. |
| **Feature Engineering** | The process of creating new features or variables based on domain knowledge to improve machine learning algorithms' performance. |
| **Feature Extraction** | Identifying and selecting relevant features or attributes from the data set. |
| **Interactive Processes** | Iterative and continuous refinement of the methodology based on insights and feedback from data analysis. |
| **Missing Values** | Values that are absent or unknown in the dataset, requiring careful handling during data preparation. |
| **Model Calibration** | Adjusting model parameters to improve accuracy and alignment with the initial design. |
| **Pairwise Correlations** | An analysis to determine the relationships and correlations between different variables. |
| **Text Analysis** | Steps to analyze and manipulate textual data, extracting meaningful information and patterns. |
| **Text Analysis Groupings** | Creating meaningful groupings and categories from textual data for analysis. |
| **Visualization techniques** | Methods and tools that data scientists use to create visual representations or graphics that enhance the accessibility and understanding of data patterns, relationships, and insights. |

Modelling to evaluation

| **Term** | **Definition** |
| --- | --- |
| **Binary classification model** | A model that classifies data into two categories, such as yes/no or stop/go outcomes. |
| **Data compilation** | The process of gathering and organizing data required for modeling. |
| **Data modeling** | The stage in the data science methodology where data scientists develop models, either descriptive or predictive, to answer specific questions. |
| **Descriptive model** | A type of model that examines relationships between variables and makes inferences based on observed patterns. |
| **Diagnostic measure based tuning** | The process of fine-tuning the model by adjusting parameters based on diagnostic measures and performance indicators. |
| **Diagnostic measures** | The evaluation of a model's performance of a model to ensure that the model functions as intended. |
| **Discrimination criterion** | A measure used to evaluate the performance of the model in classifying different outcomes. |
| **False-positive rate** | The rate at which the model incorrectly identifies negative outcomes as positive. |
| **Histogram** | A graphical representation of the distribution of a dataset, where the data is divided into intervals or bins, and the height of each bar represents the frequency or count of data points falling within that interval. |
| **Maximum separation** | The point where the ROC curve provides the best discrimination between true-positive and false-positive rates, indicating the most effective model. |
| **Model evaluation** | The process of assessing the quality and relevance of the model before deployment. |
| **Optimal model** | The model that provides the maximum separation between the ROC curve and the baseline, indicating higher accuracy and effectiveness. |
| **Receiver Operating Characteristic (ROC)** | Originally developed for military radar, the military used this statistical curve to assess the performance of binary classification models. |
| **Relative misclassification cost** | This measurement is a parameter in model building used to tune the trade-off between true-positive and false-positive rates. |
| **ROC curve (Receiver Operating Characteristic curve)** | A diagnostic tool used to determine the optimal classification model's performance. |
| **Separation** | Separation is the degree of discrimination achieved by the model in correctly classifying outcomes. |
| **Statistical significance testing** | Evaluation technique to verify that data is appropriately handled and interpreted within the model. |
| **True-positive rate** | The rate at which the model correctly identifies positive outcomes. |

Feedback to Deployment

| **Term** | **Definition** |
| --- | --- |
| **Browser-based application** | An application that users access through a web browser, typically on a tablet or other mobile device, to provide easy access to the model's insights. |
| **Cyclical methodology** | An iterative approach to the data science process, where each stage informs and refines the subsequent stages. |
| **Data collection refinement** | The process of obtaining additional data elements or information to improve the model's performance. |
| **Data science model** | The result of data analysis and modeling that provides answers to specific questions or problems. |
| **Feedback** | The process of obtaining input and comments from users and stakeholders to refine and improve the data science model. |
| **Model refinement** | The process of adjusting and improving the data science model based on user feedback and real-world performance. |
| **Redeployment** | The process of implementing a refined model and intervention actions after incorporating feedback and improvements. |
| **Review process** | The systematic assessment and evaluation of the data science model's performance and impact. |
| **Solution deployment** | The process of implementing and integrating the data science model into the business or organizational workflow. |
| **Solution owner** | The individual or team responsible for overseeing the deployment and management of the data science solution. |
| **Stakeholders** | Individuals or groups with a vested interest in the data science model's outcome and its practical application, such as solution owners, marketing, application developers, and IT administration. |
| **Storytelling** | Storytelling is the art of conveying your message, or ideas through a narrative structure that engages, entertains, and resonates with the audience. |
| **Test environment** | A controlled setting where the data science model is evaluated and refined before full-scale implementation. |

* Foundational methodology, a cyclical, iterative data science methodology developed by John Rollins, consists of 10 stages, starting with Business Understanding and ending with Feedback.
* CRISP-DM, an open source data methodology, combines several data-related methodology stages into one stage and omits the Feedback stage resulting in a six-stage data methodology.
* The primary goal of the Business Understanding stage is to understand the business problem and determine the data needed to answer the core business question.
* During the Analytic Approach stage, you can choose from descriptive diagnostic, predictive, and prescriptive analytic approaches and whether to use machine learning techniques.
* During the Data Requirements stage, scientists identify the correct and necessary data content, formats, and sources needed for the specific analytical approach.
* During the Data Collection stage, expert data scientists revise data requirements and make critical decisions regarding the quantity and quality of data. Data scientists apply descriptive statistics and visualization techniques to thoroughly assess the content, quality, and initial insights gained from the collected data, identify gaps, and determine if new data is needed, or if they should substitute existing data.
* The Data Understanding stage encompasses all activities related to constructing the data set. This stage answers the question of whether the collected data represents the data needed to solve the business problem. Data scientists might use descriptive statistics, predictive statistics, or both.
* Data scientists commonly apply Hurst, univariates, and statistics such as mean, median, minimum, maximum, standard deviation, pairwise correlation, and histograms.
* During the Data Preparation stage, data scientists must address missing or invalid values, remove duplicates, and validate that the data is properly formatted. Feature engineering and text analysis are key techniques data scientists apply to validate and analyze data during the Data Preparation stage.
* The end goal of the Modeling stage is that the data model answers the business question. During the Modeling stage, data scientists use a training data set. Data scientists test multiple algorithms on the training set data to determine whether the variables are required and whether the data supports answering the business question. The outcome of those models is either descriptive or predictive.
* The Evaluation stage consists of two phases, the diagnostic measures phase, and the statistical significance phase. Data scientists and others assess the quality of the model and determine if the model answers the initial Business Understanding question or if the data model needs adjustment.
* During the Deployment stage, data scientists release the data model to a targeted group of stakeholders, including solution owners, marketing staff, application developers, and IT administration.,
* During the Feedback stage, stakeholders and users evaluate the model and contribute feedback to assess the model’s performance.
* The data model’s value depends on its ability to iterate; that is, how successfully the data model incorporates user feedback.