Women in Data Science Worldwide | Trinidad & Tobago

Data Science Essentials

Here is a quick, comprehensive reference of terms for beginners in data science.

Term	Meaning
Data Science	A multidisciplinary field that uses scientific methods, algorithms, processes, and systems to extract knowledge and insights from structured and unstructured data.
R	A programming language and software environment primarily used for statistical computing and graphics.
Python	A high-level programming language known for its simplicity and readability, widely used in data analysis, machine learning, and web development.
Google Sheets	A web-based spreadsheet application offered by Google that allows users to create and edit spreadsheets online while collaborating with others in real-time.
Google Colab	An online platform provided by Google that allows users to write and execute Python code in a collaborative environment using Jupyter notebooks.
MATLAB	A programming language and environment primarily used for numerical computing, data analysis, and visualization.
IBM SPSS	Statistical software used for analyzing data and generating statistics. It offers advanced features for data management, analysis, and reporting.

Visualization	The graphical representation of data and information to facilitate understanding and communication.

Term	Meaning
Data	Raw facts or observations typically stored in a structured format (e.g., databases, spreadsheets) or unstructured format (e.g., text documents, images).
Dataset	A collection of data typically organized in tabular form with rows representing observations and columns representing variables or attributes.
Data Science	A multidisciplinary field that uses scientific methods, algorithms, processes, and systems to extract knowledge and insights from structured and unstructured data.
Machine Learning	A subset of artificial intelligence (AI) that enables systems to learn from data and improve performance on specific tasks without being explicitly programmed.
Algorithms	Step-by-step procedures or instructions for solving a problem or performing a task, often used in data analysis and machine learning to extract insights from data.
Regression	A statistical method used to model and analyze the relationship between a dependent variable and one or more independent variables.

Classification	A machine learning task that involves categorizing input data into predefined classes or categories based on their features.

Term	Meaning
Clustering	A machine learning technique used to group similar data points together in order to discover underlying patterns or structures in the data.
Feature Extraction	The process of selecting, transforming, or encoding raw data into a format that is suitable for machine learning algorithms to process and analyze.
Overfitting	A phenomenon in machine learning where a model learns to capture noise or random fluctuations in the training data, leading to poor performance on unseen data.
Cross-Validation	A technique used to assess the performance and generalization ability of a machine learning model by splitting the data into multiple subsets for training and testing.
Bias-Variance Tradeoff	The balance between the error introduced by bias (underfitting) and the error introduced by variance (overfitting) when building and evaluating predictive models.

Data Cleaning	The process of identifying and correcting errors, inconsistencies, and missing values in a dataset to improve its quality and reliability for analysis.
Exploratory Data Analysis (EDA)	The process of examining and visualizing data to understand its characteristics, patterns, and relationships before applying formal statistical techniques.
Hypothesis Testing	A statistical method used to make inferences or conclusions about a population based on sample data, typically involving testing the validity of a hypothesis.

Term	**Meaning**
Big Data	Extremely large and complex datasets that require advanced techniques and technologies to store, manage, and analyze effectively.
Data Mining	The process of discovering patterns, trends, and insights from large datasets using statistical and machine learning techniques.
Supervised Learning	A type of machine learning where the model is trained on labeled data, meaning each input is associated with a corresponding output, to learn the mapping between inputs and outputs.
Unsupervised Learning	A type of machine learning where the model is trained on unlabeled data and learns patterns or structures without explicit guidance, typically used for clustering and dimensionality reduction tasks.

Reinforcement Learning	A type of machine learning where an agent learns to make decisions by interacting with an environment, receiving feedback in the form of rewards or penalties, and adjusting its actions accordingly to maximize cumulative rewards over time.
Neural Networks	A type of machine learning model inspired by the structure and function of the human brain, composed of interconnected nodes (neurons) organized into layers, used for various tasks such as image recognition, natural language processing, and predictive modeling.
Deep Learning	A subset of machine learning that uses deep neural networks with multiple layers (deep architectures) to learn complex representations of data, often achieving state-of-the-art performance in tasks such as image and speech recognition, and natural language understanding.
Term	**Meaning**
Natural Language Processing (NLP)	A field of artificial intelligence that focuses on the interaction between computers and human language, enabling machines to understand, interpret, and generate human language data.
Feature Engineering	The process of creating new features or transforming existing features in a dataset to improve the performance of machine learning models by making them more representative of the underlying patterns in the data.

Dimensionality Reduction	The process of reducing the number of features (dimensions) in a dataset while preserving its important structure or relationships, commonly used to address the curse of dimensionality and improve model efficiency and performance.
Model Evaluation	The process of assessing the performance and effectiveness of a machine learning model on unseen data using various metrics and techniques such as accuracy, precision, recall, F1-score, and ROC-AUC curve.
Ensemble Learning	A machine learning technique that combines multiple models (learners) to improve predictive performance, robustness, and generalization by aggregating their predictions through methods such as averaging, voting, or stacking.
Time Series Analysis	A statistical method used to analyze and forecast time-dependent data (time series) by identifying patterns, trends, seasonality, and irregularities to make predictions or infer insights for decision-making purposes.
Term	**Meaning**
Anomaly Detection	The process of identifying unusual or abnormal patterns or events in data that deviate significantly from the expected behavior, often used for fraud detection, system monitoring, and quality control.

Bias	Systematic errors or inaccuracies introduced into a model or analysis due to flawed assumptions, limitations in data collection, or inherent prejudice, which can lead to unfair or misleading results.
Variance	The variability or spread of model predictions or estimates across different datasets or samples, reflecting the sensitivity of the model to fluctuations in the training data and its ability to generalize to unseen data.