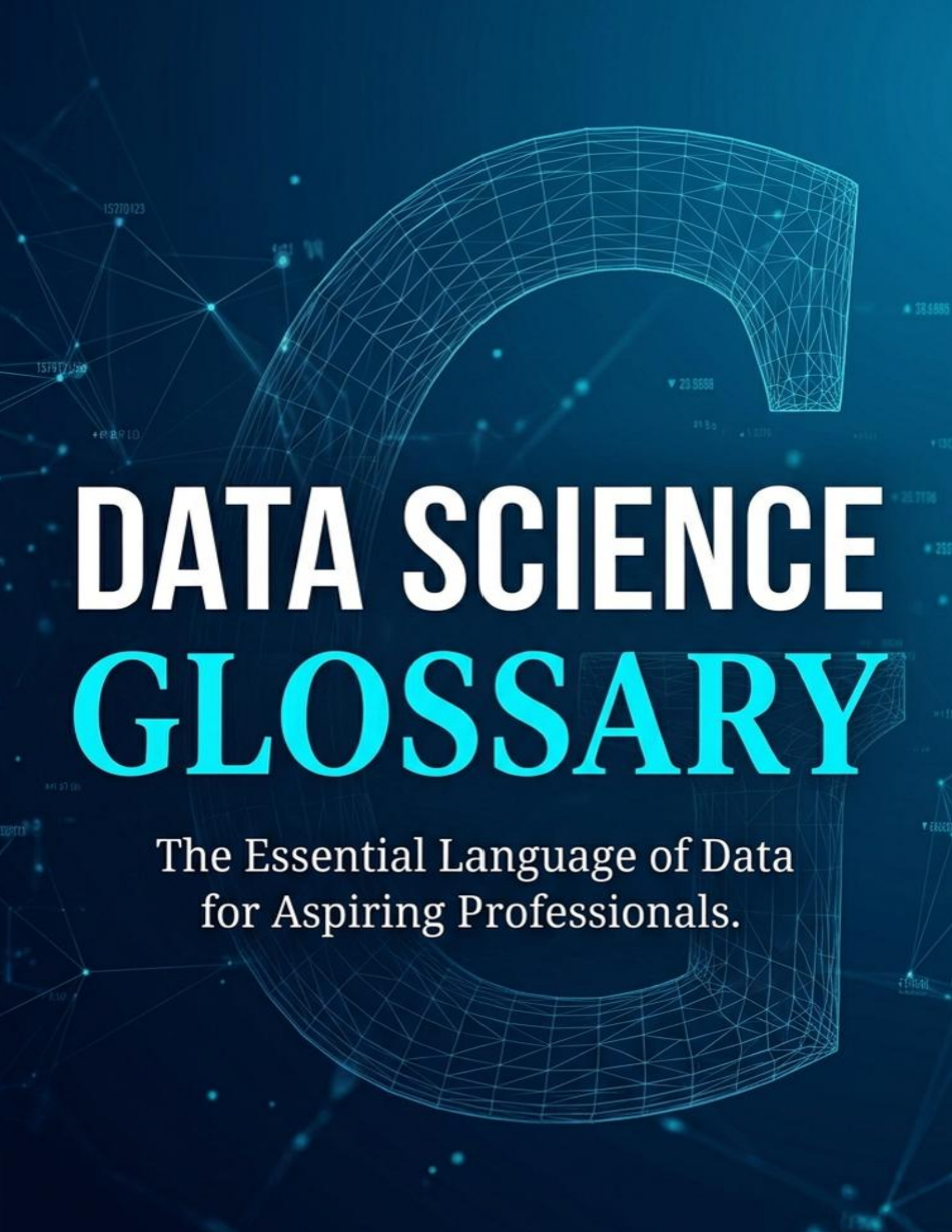
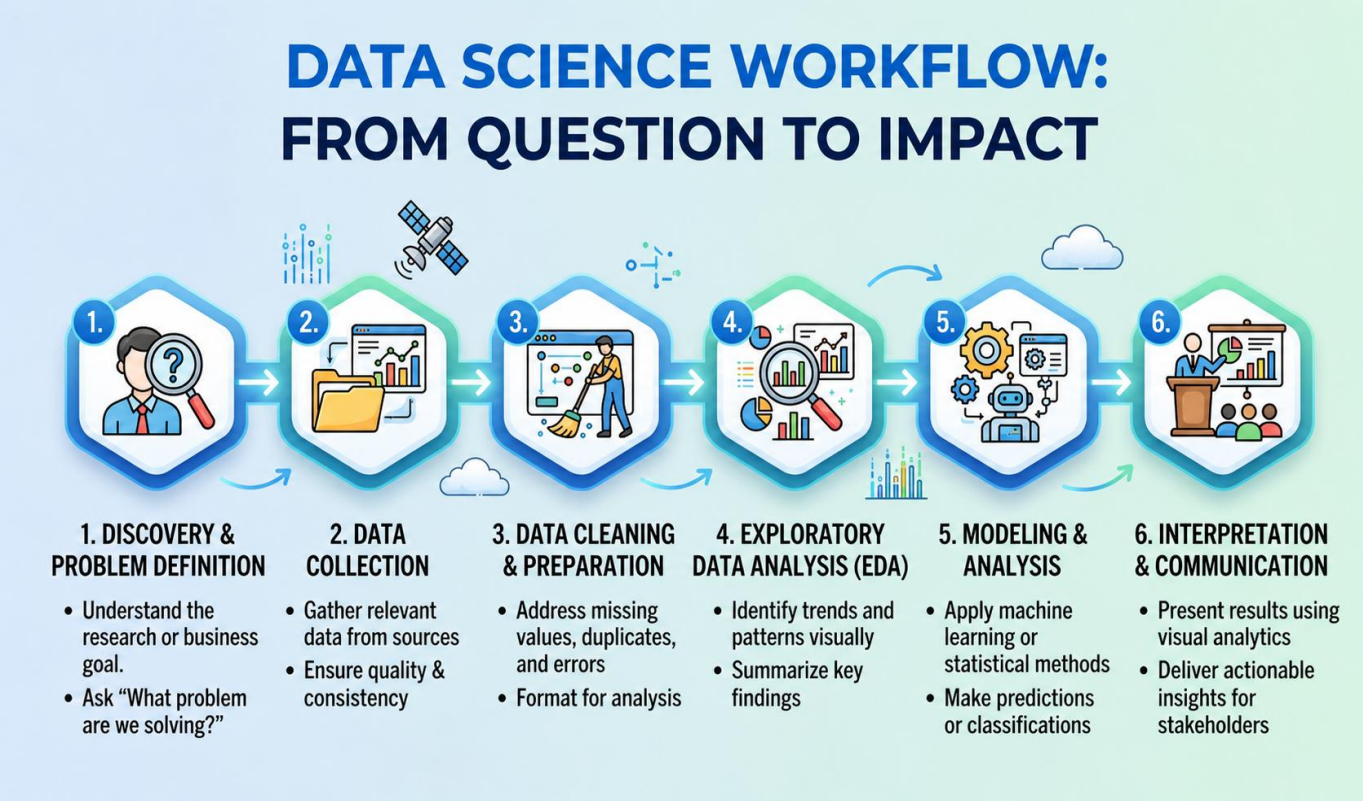

DATA SCIENCE
GLOSSARY
The Essential Language of Data
for Aspiring Professionals.



Key Terms & Definitions

This glossary is designed as a companion for your Data Science learning journey. It provides concise, accessible definitions of the core terms and concepts you will encounter throughout the book and in your wider study of data science. Entries are arranged alphabetically for ease of reference. Whether you are encountering a term for the first time or need a quick reminder, this glossary will support your learning journey.

A

Term	Definition
Algorithm	A step-by-step procedure or set of rules for solving a problem or performing a computation. In data science, algorithms are used to process data, identify patterns, and make predictions.
Anomaly Detection	The process of identifying unusual patterns or outliers in data that deviate significantly from expected behaviour. Used in fraud detection, network security, and quality control.
API (Application Programming Interface)	A set of rules and protocols that allows different software applications to communicate with each other. Data scientists use APIs to access external data sources and services.

Term	Definition
Artificial Intelligence (AI)	The simulation of human intelligence processes by machines, including learning, reasoning, and self-correction. Data science is a key discipline within the broader field of AI.
Autocorrelation	The correlation of a time series with a delayed copy of itself. It measures the degree to which past values of a variable predict its future values.

B

Term	Definition
Bagging (Bootstrap Aggregating)	An ensemble machine learning technique that trains multiple models on different random subsets of the training data and combines their predictions to reduce variance and overfitting.
Bayesian Inference	A statistical method that updates the probability of a hypothesis as new evidence becomes available, combining prior beliefs with observed data.
Bias	A systematic error in a model or dataset that causes predictions to consistently deviate from true values. Bias can arise from flawed data collection, model assumptions, or algorithmic design.
Big Data	Datasets that are too large or complex to be processed by traditional data management tools. Characterised by the 3 Vs: Volume (scale), Velocity (speed of generation), and Variety (diversity of types).
Binary Classification	A supervised learning task where the goal is to categorise input data into one of exactly two classes, such as spam vs. not spam or positive vs. negative.
Boosting	An ensemble method that sequentially trains weak learners, each correcting the errors of its predecessor, to build a strong predictive model. Examples include AdaBoost and Gradient Boosting.

C

Term	Definition
Categorical Variable	A variable that takes on a limited number of distinct values representing categories or groups (e.g., gender, colour, country). Often encoded numerically before use in machine learning models.
Central Limit Theorem (CLT)	A fundamental statistical theorem stating that the distribution of sample means approaches a normal distribution as the sample size grows, regardless of the population's original distribution.

Term	Definition
Classification	A supervised learning task in which a model learns to assign input data to predefined categories or labels. Common algorithms include logistic regression, decision trees, and support vector machines.
Clustering	An unsupervised learning technique that groups similar data points together based on their characteristics, without using predefined labels. Common algorithms include K-Means and DBSCAN.
Confusion Matrix	A table used to evaluate the performance of a classification model by showing the counts of true positives, true negatives, false positives, and false negatives.
Correlation	A statistical measure that expresses the extent to which two variables are linearly related. Values range from -1 (perfect negative) to +1 (perfect positive), with 0 indicating no linear relationship.
Cross-Validation	A model evaluation technique that splits data into multiple subsets (folds), training the model on some folds and validating on others, to give a more robust estimate of model performance.

D

Term	Definition
Data Cleaning	The process of detecting and correcting errors, inconsistencies, and missing values in a dataset to improve data quality before analysis.
Data Lake	A centralised repository that stores large volumes of raw, structured, semi-structured, and unstructured data in its native format until it is needed for analysis.
Data Mining	The process of discovering patterns, correlations, and insights from large datasets using statistical, mathematical, and computational techniques.
Data Pipeline	A series of automated data processing steps that move and transform data from one system to another, typically from raw sources to analysis-ready formats.
Data Wrangling	The process of transforming and mapping raw data into a more usable format for analysis. Also called data munging, it includes cleaning, restructuring, and enriching data.
Decision Tree	A supervised learning model that makes predictions by learning a hierarchy of if-then rules from training data. The model resembles a tree structure with branches representing decision points.
Dimensionality Reduction	The process of reducing the number of features in a dataset while preserving important information. Common techniques include Principal Component Analysis (PCA) and t-SNE.

E

Term	Definition
Ensemble Method	A machine learning approach that combines multiple models to produce better predictive performance than any individual model. Examples include Random Forests, Bagging, and Boosting.
Epoch	One complete pass through the entire training dataset during the training of a neural network. Models are typically trained over many epochs to converge on optimal weights.
Exploratory Data Analysis (EDA)	An approach to analysing datasets that uses visual and statistical methods to summarise their main characteristics, discover patterns, and identify anomalies before formal modelling.

F

Term	Definition
Feature	An individual measurable property or characteristic of the data used as input to a machine learning model. Also known as an independent variable or predictor.
Feature Engineering	The process of using domain knowledge to create, transform, or select features from raw data to improve model performance.
Feature Scaling	The process of normalising or standardising the range of features in a dataset so that no single feature dominates others due to its scale. Common methods include min-max scaling and standardisation.
F1 Score	A classification metric that balances precision and recall by computing their harmonic mean. Useful when class imbalance is present and both false positives and false negatives are costly.

G

Term	Definition
Gradient Descent	An optimisation algorithm that iteratively adjusts model parameters by moving in the direction of the steepest decrease in the loss function, aiming to find the minimum loss.
Gradient Boosting	An ensemble technique that builds models sequentially, with each model correcting the residual errors of the previous one using gradient descent in function space.

H

Term	Definition
Hyperparameter	A configuration setting external to a model that is set before training begins and controls the learning process (e.g., learning rate, number of trees). Distinct from parameters, which are learned from data.
Hypothesis Testing	A statistical method used to determine whether there is enough evidence in a sample of data to infer that a certain condition holds for the entire population.

I

Term	Definition
Imputation	The process of replacing missing values in a dataset with estimated values, such as the mean, median, mode, or values predicted by a model.
Imbalanced Dataset	A dataset where the classes are not represented equally. For example, a fraud detection dataset where fraudulent transactions are far rarer than legitimate ones.

K

Term	Definition
K-Fold Cross-Validation	A cross-validation technique that partitions data into K equal subsets (folds), trains the model K times — each time using a different fold as the validation set — and averages the results.
K-Means Clustering	An unsupervised algorithm that partitions data into K clusters by iteratively assigning points to the nearest centroid and updating centroids until convergence.
K-Nearest Neighbours (KNN)	A simple supervised learning algorithm that classifies a data point based on the majority class of its K nearest neighbours in the feature space.

L

Term	Definition
Label	The target output variable in a supervised learning task. Also called the dependent variable or response variable. For example, in email classification, 'spam' or 'not spam' are labels.

Term	Definition
Lasso Regression (L1 Regularisation)	A regression technique that adds a penalty equal to the absolute value of coefficients to the loss function, encouraging sparsity by shrinking some coefficients to exactly zero.
Latent Variable	A variable that is not directly observed or measured but is inferred from other observed variables. Commonly used in dimensionality reduction and topic modelling.
Learning Rate	A hyperparameter that controls how much the model's weights are adjusted with respect to the loss gradient at each update step during training.
Linear Regression	A supervised learning algorithm that models the relationship between a continuous dependent variable and one or more independent variables by fitting a linear equation to the data.
Logistic Regression	A classification algorithm that models the probability of a binary outcome using a logistic (sigmoid) function. Despite its name, it is used for classification, not regression.
Loss Function	A function that measures the difference between a model's predictions and the actual target values. The goal of training is to minimise the loss function.

M

Term	Definition
Machine Learning (ML)	A subset of AI that enables systems to learn and improve from experience without being explicitly programmed. Models are trained on data to identify patterns and make decisions.
Mean Absolute Error (MAE)	A regression performance metric that measures the average absolute difference between predicted and actual values. It is robust to outliers compared to MSE.
Mean Squared Error (MSE)	A regression metric that measures the average squared difference between predicted and actual values. Squaring penalises larger errors more heavily.
Model Deployment	The process of integrating a trained machine learning model into a production environment so that it can make predictions on new, real-world data.
Multicollinearity	A situation in regression analysis where two or more predictor variables are highly correlated, making it difficult to determine their individual effects on the response variable.

N

Term	Definition
Natural Language Processing (NLP)	A branch of AI focused on enabling computers to understand, interpret, and generate human language. Applications include sentiment analysis, translation, and chatbots.
Neural Network	A machine learning model inspired by the human brain, composed of interconnected layers of nodes (neurons) that transform inputs into outputs through learned weights.
Normalisation	The process of scaling features to a standard range to ensure that no single feature disproportionately influences the model due to its magnitude.
Null Hypothesis	The default assumption in hypothesis testing that there is no effect or no difference between groups. A sufficiently low p-value leads to rejection of the null hypothesis.

O

Term	Definition
One-Hot Encoding	A technique for representing categorical variables as binary vectors, creating a new binary column for each category to make them suitable for use in machine learning algorithms.
Outlier	A data point that differs significantly from other observations. Outliers can distort statistical analyses and model training, and may indicate data errors or rare events of interest.
Overfitting	A modelling error where a model learns the noise and specific patterns of the training data too well, resulting in poor generalisation to new, unseen data.

P

Term	Definition
P-Value	The probability of observing results at least as extreme as the current data, given that the null hypothesis is true. A low p-value (typically < 0.05) suggests the results are statistically significant.
Precision	The proportion of true positive predictions among all positive predictions made by the model. A high precision means the model has few false positives.
Principal Component Analysis (PCA)	An unsupervised dimensionality reduction technique that transforms features into a smaller set of uncorrelated components that capture the maximum variance in the data.

R

Term	Definition
Random Forest	An ensemble learning method that builds multiple decision trees on random subsets of the data and features, then aggregates their predictions by voting (classification) or averaging (regression).
Recall (Sensitivity)	The proportion of actual positives that were correctly identified by the model. High recall means few false negatives. Also referred to as the True Positive Rate.
Regularisation	A technique used to prevent overfitting by adding a penalty to the model's loss function for large or complex coefficients. Common forms include L1 (Lasso) and L2 (Ridge) regularisation.
Regression	A supervised learning task where the goal is to predict a continuous numerical output variable. Examples include predicting house prices or temperature.
ROC Curve (Receiver Operating Characteristic)	A graph showing the trade-off between the True Positive Rate (Recall) and the False Positive Rate at various classification thresholds. The AUC (Area Under the Curve) summarises overall performance.

S

Term	Definition
Sampling	The process of selecting a subset of data from a larger population for analysis. Proper sampling ensures that the subset is representative of the whole.
Standardisation (Z-Score Normalisation)	Rescaling features so they have a mean of zero and a standard deviation of one. Calculated as (value - mean) / standard deviation.
Supervised Learning	A type of machine learning where the model is trained on labelled data, learning to map inputs to known output labels. Examples include classification and regression.
Support Vector Machine (SVM)	A supervised learning algorithm that finds the optimal hyperplane separating classes in the feature space, maximising the margin between the nearest data points of each class.

T

Term	Definition
Test Set	A portion of the dataset held back and not used during model training or validation, used exclusively to evaluate the final model's performance on unseen data.

Term	Definition
Time Series	A sequence of data points indexed in time order, typically collected at successive, equally spaced points in time. Analysed to identify trends, seasonal patterns, and forecasts.
Training Set	The portion of the dataset used to train a machine learning model. The model learns patterns from this data by adjusting its parameters to minimise prediction error.
Transfer Learning	A machine learning technique where a model trained on one task is adapted and fine-tuned for a different but related task, reducing the need for large amounts of labelled data.

U

Term	Definition
Underfitting	A modelling problem where the model is too simple to capture the underlying patterns in the data, resulting in poor performance on both training and test datasets.
Unsupervised Learning	A machine learning approach where the model learns patterns and structure from unlabelled data. Common tasks include clustering, dimensionality reduction, and anomaly detection.

V

Term	Definition
Validation Set	A portion of the dataset used to tune model hyperparameters and evaluate model performance during training, separate from both the training set and the final test set.
Variance	In statistics, variance measures the spread of data around the mean. In machine learning, high model variance means the model is sensitive to small fluctuations in training data (overfitting).

W

Term	Definition
Weight	A learnable parameter in a machine learning model that is adjusted during training to minimise the loss function. Weights determine the strength and direction of influence of each feature.