**Accuracy**
Accuracy refers to the proportion of correctly classified predictions in relation to the total number. It is a standard measure for classification models, but it does not always make sense for unbalanced classes.
**Example:** 95 correct predictions out of 100 result in 95% accuracy.

**A/B Testing**

An A/B test compares two variants to see which one performs better. It requires a clean separation and random assignment.

**Example:** Two website versions with different button colors are tested.

**Active Learning**

An ML approach in which a model specifically decides which data to label for further improvement.

**Example:** The model deliberately selects unsafe images for annotation by humans.

**AdaBoost**

A boosting algorithm that combines weak classifiers into a strong model. Particularly effective for small data sets.

**Example:** Several small decision trees are combined to form an overall model.

**Aggregation**

Data is summarized, e.g. as a sum or average. Important in BI, SQL and Excel.

**Example:** Average revenue per month.

**Alias**

An alternate name, usually in SQL or code, for readability.

**Example:** `SELECT sales AS sales FROM table`.

**Algorithm**

A defined sequence of instructions to solve a problem. In data analysis, mostly learning methods.

**Example:** The k-means algorithm groups data into clusters.

**Alternative hypothesis**

Statistical term for the assumption that there is an effect or difference.

**Example:** The new drug active ingredient works better than the old one.

**Anomaly Detection**

Technique for detecting unusual patterns. Useful in fraud detection, log analysis, and quality control.

**For example,** a single user makes 100 purchases in one minute.

**API (Application Programming Interface)**

An interface for structured communication between programs. Indispensable for automation and data exchange.

**For example,** a weather API provides JSON data for a dashboard.

**ARIMA**

A time series model that combines autoregressive, integrated, and moving averages.

**Example:** Monthly sales of the last three years are estimated for the following year.

**Array**

A data structure for similar elements, efficient for numerical calculations.

**Example:** A NumPy array with 1,000 numbers.

**Artificial Intelligence (AI)**

Umbrella term for machines that solve tasks with cognitive abilities. Including ML and deep learning.

**Example:** An AI system diagnoses skin diseases based on images.

**AUC (Area Under the Curve)**

Classifier metric. Measures the area under the ROC curve – the closer to 1, the better.

**Example:** A model with AUC 0.95 separates classes very well.

**Autoencoder**

Neural network for data compression and reconstruction. Helpful for anomaly detection.

**Example:** Input image is reconstructed with minimal loss of information.

**Autocorrelation**

Relation of a value to itself via time difference. Important in time series analysis.

**Example:** Sales in December are high as in the previous year.

**Automation**

Replace manual processes with scripts or systems. Saves time and avoids errors.

**Example:** Daily import of sales data via Python script.

**Average**

The average value – sum of all values divided by number.

**Example:** Average of 5, 7, 8 is 6.67.

**Azure**

Microsoft's cloud platform with tools for data analysis, ML, databases. Competition with AWS and GCP.

**Example:** ETL process running in Azure Data Factory.

**Accuracy Paradox**

Phenomenon that high accuracy can still mean a bad model.

**Example:** 99% correct predictions in a dataset with 99% negatives.

**Augmented Analytics**

Form of analysis with AI support for insight generation and automation.

**Example:** A BI tool automatically explains anomalies.

**Attribute**

Characteristics of a data set, also known as features.

**Example:** Income, age, region of a customer.

**AutoML**

Automated selection, training, and tuning of ML models. For rapid prototypes.

**Example:** Google AutoML creates an image classification model without code.

**Atomic Operation**

Uninterruptible action, such as in database transactions. Guarantees consistency.

**Example:** `INSERT` into a table with a rollback option.

**Authentication**

Identity verification process. Often relevant for data access and APIs.

**Example:** Access to an SQL server only with a password.

**Autonomous System**

A fully self-running data or software system.

**Example:** A car creates its route ETAs based on live traffic data.

**Auto Regression (AR)**

Time series model that predicts current values from previous ones. Part of ARIMA.

**Example:** Today = 0.5×Yesterday + 0.3×The Day Before Yesterday

**ASCII**

Character encoding for letters, numbers, and symbols. Relevant for data imports and coding issues.

**Example:** Character "A" = ASCII 65

**Aliasing**

Phenomenon in which signals sampled too low are misinterpreted. Important in time series analyses.

**For example,** a weekly metric simulates a trend that disappears at daily resolution.

**Application Layer**

Layer in system architectures that is directly related to user interaction.

**Example:** A web app for visualizing analysis results.

**Analytic Function**

SQL functions that aggregate via Window Functions.

**Beispiel:** `ROW_NUMBER() OVER (PARTITION BY customer_id ORDER BY date)`

**Auto Scaling**

Automatically scale compute resources in the cloud based on workload.

**Example:** An ML model receives more RAM during load peaks.

**Association Rule Learning**

Technique for discovering rules and relationships in transactional data.

**Example:** Customers who buy beer also buy chips.

**Async**

Non-blocking execution of processes – important for parallel data processing or web requests.

**Example:** A web server processes several API requests at the same time.

**Backpropagation**
Traceability method for error correction in neural networks. It calculates how much each node contributed to the total deviation.
**Example:** In a CNN, the error is calculated back from the output layer to the input.

**Balanced Dataset**

A dataset with an even distribution of classes. Important for fair model valuation.

**Example:** 5,000 examples of spam and 5,000 examples of non-spam.

**Bar Chart**

Visualization of categorical data with bars of different heights.

**Example:** Number of sales per product category.

**Baseline Model**

A simple reference model to assess the performance of more complex models.

**Example:** Always predict the most common class.

**Batch Normalization**

Technique for accelerating and stabilizing the training of neural networks.

**For example,** values of a layer are scaled to mean 0 and variance 1.

**Batch Size**

Number of data points fed into the ML model at the same time. Affects training speed and model quality.

**Example:** Training with 128 examples per batch.

**Batch Processing**

Processing data in blocks instead of individually. Common in data warehousing and ETL.

**Example:** Night processing of all daily sales.

**Bayesian classifier**

A probabilistic model that calculates probabilities using Bayes' theorem.

**Example:** Naïve Bayes for spam classification.

**Bayesian Statistics**

Statistical approach that incorporates prior knowledge and iteratively updates probabilities.

**Example:** Probability of fraud increases after conspicuous behavior.

**Bias (distortion)**

Systematic error that leads to incorrect model results. Can be created by data or model structure.

**Example:** Unbalanced training data puts a group at a disadvantage.

**Big data**

Very large, fast-growing and diverse data volumes that overwhelm traditional processing.

**Example:** Billions of log data per day in an online shop.

**Binary Classification**

Classification with exactly two target classes.

**Example:** Fraud: Yes or No.

**Binning**

Classification of continuous variables into categories.

**Example:** Age groups such as 18–25, 26–35.

**BI (Business Intelligence)**

Entirety of technologies for data-driven decision support.

**Example:** Power BI or Tableau to visualize KPIs.

**Binary variable**

Variable with exactly two characteristics.

**Example:** Yes/No, 0/1, True/False.

**Blending**

Combination of several ML models, usually as an ensemble method.

**Example:** Combination of SVM and decision tree.

**Bloom Filter**

Probability-based data structure for fast quantity checking with memory savings.

**Example:** Checking whether an email has already been processed.

**Boolean Logic**

Logic system with truth values TRUE and FALSE.

**Example:** `WHERE active = TRUE AND country = 'DE'`

**Bootstrap**

Resampling technique for estimating distributions from samples.

**Example:** 1,000 draws with set-aside for confidence interval estimation.

**Box plot**

Diagram showing distributions including outliers.

**Example:** Distribution of income in five departments.

**Buffering**

Temporary storage of data for bridging or relief.

**Example:** Data from a stream is buffered in RAM.

**Bucket**

Single area in a scale divided into categories.

**Example:** Price range 0-10€, 10-50€, 50-100€.

**Business Analytics**

Form of analysis with a focus on business insights, strategic or operational.

**Example:** Why did revenue drop in Q2?

**Business Metric**

Key figure for the management of a company.

**Example:** Customer Lifetime Value (CLV).

**Byte**

Memory unit consisting of 8 bits. Most common unit of measurement for data volumes.

**For example,** a text file of 1,000 characters ≈ 1 KB.

**Bayesian Network**

Graph-based model for the representation of probabilistic dependencies.

**Example:** Model for disease symptoms and causes.

**Bias-Variance-Tradeoff**

Basic principle in ML: Models must balance between overfitting (variance) and underfitting (bias).

**Example:** Linear regression = high bias, low variance.

**Binary Tree**

Tree structure with a maximum of two child nodes per parent node.

**Example:** Decision tree for classification.

**Benchmarking**

Comparison of algorithms or systems based on defined metrics.

**Example:** Which classifier has the best AUC for the same data set?

**Business Rule**

Rule for controlling a process based on data.

**Example:** If users < 18 years of age → do not allow a purchase.

**Bias Mitigation**

Strategies for reducing bias in models or data.

**Example:** Fairness constraints in model training.

**Boolean Masking**

Technique for selecting certain elements with true/false arrays.

**Example:** `df[df['value'] > 100]`

**Broadcast Join**

SQL optimization, where a small table is distributed to all nodes.

**Example:** Small lookup table for country information broadcast in Spark.

**Caching**
: Caching of frequently used data to speed up access. Useful in web development, databases, and ML pipelines.
**For example,** a BI tool loads previously calculated aggregations from the cache.

**Categorical Variable**

Variables with discrete characteristics such as colors or countries. Mostly processed by one-hot-encoding.

**Example:** Color column with values: Red, Blue, Green.

**Centroid**

Center of a cluster used in k-means algorithms.

**Example:** The focus of a group of customers with similar buying behavior.

**Churn Rate**

Customer churn rate over a period of time. Important for subscription models.

**Example:** 15% monthly churn in a SaaS service.

**Classification**

ML task that divides data into discrete classes.

**Example:** Creditworthy vs. not creditworthy.

**Clean Data**

Error-free, clean data suitable for analysis or modeling.

**Example:** No duplicates, correct types, no spaces.

**Clustering**

Unsupervised learning to group similar data points.

**Example:** Customer segmentation based on buying behavior.

**Coefficient**

Weight in a model that indicates the strength and direction of a predictor.

**Example:** In a regression: Income has a positive influence on consumption.

**Collinearity**

High correlation between two or more independent variables. Makes interpretation more difficult.

**Example:** Height and weight are strongly correlated.

**Column Store**

Database system that stores data column-based – advantageous for analytical queries.

**Example:** BigQuery or Redshift.

**Confidence Interval**

Interval that contains the true parameter with a defined probability.

**Example:** 95% interval for mean: 4.1 to 4.8.

**Confusion Matrix**

Presentation of classification results with TP, FP, FN, TN.

**For example,** a model has 87% accuracy but many false positives.

**Correlation**

Measure of linear relationship between two variables. Values from -1 to +1.

**Example:** Advertising budget and revenue with correlation +0.84.

**CPU (Central Processing Unit)**

Central computing unit, performs logical operations and model training.

**Example:** Pandas usually calculates on the CPU.

**CSV (Comma-Separated Values)**

Text file with tabular data, fields separated by commas.

**Example:** Exporting a SQL table as a `data.csv`.

**Cross-Validation**

Method for stable model evaluation by repeated training/testing on different data splits.

**Example:** K-fold CV with k=5.

**Curse of Dimensionality**

High dimensionality issues – e.g., sparse data, overfitting.

**Example:** 1,000 features with only 100 observations.

**Cut-off-Point**

Threshold for classification decisions.

**Example:** Probability > 0.6 → loan approved.

**Custom Function**

Custom function in Python, SQL or Excel.

**Example:** `def berechne_rabatt(price): return price*0.85`

**Categorical Encoding**

Techniques for converting categorical variables to numerical formats.

**Example:** One-Hot, Label, Target Encoding.

**Control Chart**

Diagram for monitoring processes in quality control.

**Example:** Production line shows outliers in error frequency.

**Confidence Score**

Output value of a model that indicates how confident it is in its prediction.

**Example:** Image classification: 82% probability of "cat".

**Composite Key**

Primary key, which consists of multiple columns.

**Example:** Combination of "Order Number" + "Item ID".

**Contextual Bandit**

ML model for selecting actions in case of uncertainty, taking into account the context.

**Example:** Advertising based on user behavior.

**Confidence Level**

Specifies the degree of certainty with which a confidence interval contains the true value.

**Example:** 95% trust level → 5% probability of error.

**Click-Through-Rate (CTR)**

The percentage of clicks on an ad relative to the total number of impressions.

**Example:** 100 clicks with 1,000 views = 10% CTR.

**Canonical Correlation Analysis (CCA)**

Statistical method for studying the relationships between two sets of variables.

**Example:** Relationship between school performance and family background.

**Classification Report**

Standard output for evaluating a classification model: contains precision, recall, F1 score per class.

**Beispiel:** Scikit-learn `classification_report()`.

**Confidence Bound**

Upper or lower bound of a confidence interval.

**Example:** Cap = 7.9 at 95% interval.

**Cloud Computing**

Provision of IT resources on demand via the Internet.

**Example:** AWS, GCP or Azure offer computing power and storage space on demand.

**Cron Job**

Timed task on Linux to automate recurring processes.

**Example:** Daily update of a dashboard at 3:00 AM.

**Cost Function**

Function that measures the error of a model and is intended to be minimized.

**Example:** MSE in regression measures deviation between prediction and reality.

**Composite Index**

Database index that combines multiple columns to speed up queries.

**Beispiel:** Index auf `user_id` + `created_at`.

**Constraint**

Constraint in databases to enforce consistency rules.

**Example:** `NOT NULL`, `UNIQUE`, `FOREIGN KEY`.

**Confidence Ellipse**

Graphical representation of the confidence interval of two-dimensional data.

**Example:** Scatter plot with 95% ellipses for two features.

**Dashboard**
Visual interface for displaying key metrics, often interactively. Used in BI, monitoring and management.
**For example,** Power BI dashboard shows revenue trends and regional distributions.

**Data Analyst**

Role for analyzing, visualizing, and preparing data. Uses tools like SQL, Excel, Python.

**Example:** Analyzes the development of the conversion rate in the online shop.

**Data Cleaning**

Process to remove erroneous, missing, or duplicate data. A prerequisite for any reliable analysis.

**For example,** removing empty fields and incorrect data types from a CSV file.

**Data Engineer**

Specialist in building and maintaining data infrastructures such as pipelines, databases, cloud systems.

**Example:** Develops an ETL route for automated data integration.

**Data Governance**

Rules and processes for data quality, security, and access rights.

**Example:** Who is allowed to see personal data, who is not?

**Data Lake**

Unstructured data storage in raw form, often on Hadoop or S3.

**Example:** Storage of all raw data from weblogs, APIs and external sources.

**Data Mart**

Focused part of a data warehouse for specific business departments.

**Example:** Separate area for marketing data with aggregated KPIs.

**Data Mining**

Discovery of patterns and relationships in large amounts of data using statistical methods.

**Example:** Rule "Customers who buy X also buy Y".

**Data Pipeline**

Automated data flow from source to destination with extraction, transformation, storage.

**Example:** Apache Airflow controls the daily loading of new shop data into the data warehouse.

**Data Scientist**

Expert in analysis, modeling and forecasting of complex data using ML/AI.

**Example:** Predicting the probability of returns with Random Forest.

**Database**

Structured storage of data for efficient search and processing. Relational (SQL) or NoSQL variants.

**Example:** PostgreSQL stores customer and transaction data.

**Record**

Single row in a table with multiple attributes.

**Example:** Customer #123 with name, date of birth, sales.

**Decision Tree**

ML model with if/else structure for classification or regression.

**Example:** Baum decides whether to grant credit.

**Deep Learning**

A subfield of ML based on deep neural networks. Strong in images, language, complex patterns.

**Example:** Speech recognition on smartphones with deep learning.

**Default Value**

Default value for missing input.

**Example:** By default, 0 if the field is empty.

**Denormalization**

Intentional redundancy to improve performance in databases.

**Example:** Customer name is copied to each order line.

**Deployment**

Providing a model or system for productive use.

**Example:** ML model is provided via REST API.

**Derived Variable**

Derived feature calculated from existing fields.

**Example:** Age = Today – year of birth.

**Descriptive Analytics**

Analysis of historical data to describe developments.

**Example:** Revenue decline of 10% compared to the previous year.

**Dimensionality Reduction**

Method for reducing the number of features at high dimensionality.

**Example:** PCA reduces 500 sensor values to 10 main components.

**DNN (Deep Neural Network)**

Multi-layered neural network with high abstraction capacity.

**Example:** Classification of handwriting based on pixel values.

**Docker**

Container technology for portable, reproducible software environments.

**Example:** A Python analysis script runs independently of host systems.

**Document Store**

NoSQL database for storing document structures such as JSON.

**Example:** MongoDB stores user profiles as JSON objects.

**Dropout**

Regulation method in neural networks to avoid overfitting.

**Example:** 30% of neurons are deactivated per training run.

**Dummy Variable**

Artificially generated binary variable for encoding categorical features.

**Example:** Geschlecht_männlich = 1, Geschlecht_weiblich = 0.

**Date Wrangling**

Broad term for editing and restructuring data for analysis.

**Example:** Split columns, replace missing values, convert types.

**Drilldown**

Function in dashboards to navigate aggregation in detail data.

**Example:** Click on "Region Bavaria" to display cities.

**Data Imputation**

Replace missing values with estimation or rule.

**Example:** Average value replaces missing sales information.

**Decision Boundary**

Boundary at which a classification model distinguishes between two classes.

**Example:** Linear separation between good/bad credit.

**Data Provenance**

Documentation of the origin and transformation of a data set.

**Example:** The origin, changes and accesses of a data record are logged.

**Data Drift**

Changes in the distribution of data over time, which can cause ML models to lose accuracy.

**Example:** Customer types change due to market change – model must be retrained.

**EDA (Exploratory Data Analysis)**
Systematic examination of data before modeling. Used to detect patterns, outliers, and data issues.
**Example:** Box plots, correlations, and histograms to evaluate a customer record.

**Edge Case**

Unusual input case that pushes a system to its limits. Relevant for testing and bug resistance.

**Example:** A customer who is 0 or 120 years old.

**Elasticity**

Measure of the sensitivity of a target variable when an influencing variable changes.

**Example:** -1.5 price elasticity = 1% price increase → 1.5% less demand.

**Embedding**

Transformation of objects (such as words) into fixed-length vectors.

**Example:** Word "car" as a vector [0.12, -0.44, ...] for neural network.

**Ensemble Learning**

Method of combining multiple models to improve accuracy.

**Example:** Random Forest combines many decision trees.

**Entropy (Entropie)**

Measure of disorder in data, used in decision trees. The higher, the more mixed the classes.

**Example:** Maximum entropy at 50%/50% distribution of two classes.

**Epoch**

A complete run-through of all training data during model training.

**Example:** The model is trained over 100 epochs.

**ETL (Extract, Transform, Load)**

Core process for data integration: extraction from source, transformation, loading into target system.

**Example:** Webshop data → currency conversion → storage in PostgreSQL.

**Evaluation Metric**

Key figure used to evaluate models, e.g. Accuracy, RMSE, Precision.

**Example:** An F1 score of 0.81 in spam classification.

**Excel**

Spreadsheet tool with high relevance in reporting, analysis and visualization.

**Example:** Pivot table to analyze sales data by region and month.

**Exponential Smoothing**

Time series method to smooth out short-term fluctuations.

**Example:** Weighted sales forecast with a stronger focus on recent stocks.

**Extrapolation**

Prediction outside the known data range – riskier than interpolation.

**Example:** Forecast revenue for 2030 based on 2020-2024.

**Early Stopping**

Discontinuation of model training as soon as validation errors increase. Prevents overfitting.

**Example:** Model training stops after 34 epochs.

**Entity**

Real or conceptual object that is recorded in a database.

**Example:** Customers, Products, Orders.

**Event-based Data**

Time-stamped data generated by actions.

**Example:** clicks, logins, purchases in web systems.

**Explainability**

Comprehensibility of models and their decisions for humans.

**Example:** SHAP values show the influence of individual features on model decisions.

**Exogenous Variable**

Influencing variable from outside, not explained by the model, but taken into account.

**Example:** Weather data in retail sales analysis.

**Elastic Net**

Regularization technique that combines Lasso (L1) and Ridge (L2).

**Example:** Use for correlated regression variables.

**Error Rate**

Proportion of incorrect predictions. Addition to Accuracy.

**Example:** 7 misclassifications in 100 cases → error rate: 7%.

**One-vs-Rest**

Classification strategy for multi-class problems.

**Example:** 3 models: Cat-vs-Rest, Dog-vs-Rest, Mouse-vs-Rest.

**Euclidean distance**

Length of the shortest connection of two points in the feature space.

**Example:** Distance between two customer profiles in 5D space.

**Encoding**

Conversion of variables into a numerical representation.

**Example:** One-hot-encoding for color: blue = [0,1,0]

**Empirical Distribution**

Distribution of the observed data values, without a theoretical model.

**Example:** Histogram of observed customer ages.

**Entity-Relationship-Model (ER-Model)**

Diagram for structuring database tables and their relationships.

**Example:** Relationship: Customer – > order (1:n).

**Execution Plan**

Description of how a database query is technically executed.

**Example:** PostgreSQL shows how to perform a JOIN (index, collation, etc.).

**External Table**

Table that references data outside the database (for example, in data lakes).

**Example:** Hive table referencing Parquet files.

**ETL Scheduler**

Tool for timing ETL processes.

**Example:** Apache Airflow plans nightly ETL pipelines.

**Enrichment**

Enrichment of data with additional attributes to improve the analysis.

**Example:** Enrichment of transaction data with weather data.

**Inclusion Criteria**

Filter condition for data access or model training.

**For example,** only customers with full profile data will be trained.

**Endpoint**

Address (e.g., URL) used to access data or models via API.

**Example:** `/predict/` takes features and provides prediction.

**Embedding Layer**

Layer in neural networks that maps discrete values into continuous vectors.

**Example:** User ID → 16-dimensional representation for recommendation system.

**Error Analysis**

Targeted investigation of model errors to improve performance.

**Example:** Analysis of which products a classifier regularly fails for.

**Constraint**

Database rule that enforces certain states.

**Example:** Column must not contain null values (`NOT NULL`).

**F1 Score**
Harmonic mean of Precision and Recall. Good metric for unbalanced datasets.
**Example:** F1 of 0.84 means solid balance between detection and precision.

**Factor Analysis**

Statistical method for reducing to latent variables (factors).

**Example:** Several satisfaction questions result in a "service" factor.

**Feature**

Input characteristic of a model. Also called attribute or predictor.

**Example:** Age, income, place of residence.

**Feature Engineering**

Create and transform relevant features for ML models.

**Example:** Extract the quarter from date.

**Feature Importance**

Measure of the influence of a feature on the model.

**Example:** In a churn model, "last login" is the most important.

**Feature Selection**

Selection of the most important features for model simplification.

**Example:** Elimination of redundant or irrelevant columns.

**Federated Learning**

Decentralized model training method without central data storage.

**Example:** Mobile devices train a common language model locally.

**Filter Function**

Function for data selection by condition.

**Example:** Pandas: `df[df['alter'] > 30]`

**Float**

Data type for floating-point numbers.

**Example:** 3.14159

**Forecasting**

Forecast future values based on historical data.

**Example:** Revenue forecasting with the Holt-Winters model.

**Foreign Key**

Foreign keys in relational databases. Refers to primary keys of another table.

**Example:** `customer_id` in Order Table refers to Customer Table.

**Formula**

Calculation rule for automated calculation.

**Example:** Excel: `=B2*C2`

**Forward Selection**

Step-by-step feature selection for regression models.

**Example:** Start with an empty model and add features successively.

**Fourier Transformation**

Breaks down time series into frequency components.

**Example:** Frequency analysis of electricity consumption data.

**False Positive (FP)**

Faulty positive prediction.

**Example:** Spamfilter marks legitimate email as spam.

**False Negative (FN)**

Incorrect negative prediction.

**Example:** A case of cancer remains undetected.

**FRAUD Detection**

System for detecting fraud patterns.

**Example:** ML detects fake credit card transactions.

**Frequency Table**

Table with frequencies of characteristics.

**Example:** 340 users from Germany, 120 from Austria.

**Full Outer Join**

SQL join, which shows all rows of both tables, even without a match.

**Example:** All customers and all orders – even if there is no connection.

**Function**

Reusable code block with inputs and returns.

**Example:** `def quadrat(x): return x*x`

**Fuzzy Matching**

Comparison of similarly written texts with tolerance.

**Example:** "Meier" ≈ "Mayer".

**Flat File**

Simple file with no relational structure, mostly CSV or TXT.

**Example:** Raw data export from an old CRM.

**Feature Map**

Intermediate output of CNNs in image processing.

**Example:** Activation card after convolution operation.

**First Normal Form (1NF)**

Basic rule for relational databases: no repetition groups, atomic values.

**For example,** a column does not contain multiple phone numbers.

**File System**

Structured storage and management of files in directories.

**Example:** Hadoop Distributed File System (HDFS).

**Fingerprinting**

Recognition of a user/object through unique data patterns.

**Example:** Recognition of devices based on browser data.

**Finite State Machine**

Model that describes states and transitions of a system.

**Example:** Click sequence in an app is modeled as a state diagram.

**Fit (model training)**

Adaptation of an ML model to training data.

**Example:** `model.fit(X_train, y_train)`

**Feature Drift**

Change in the meaning or distribution of a characteristic over time.

**Example:** "User activity" loses significance after product update.

**Field**

Single data attribute within a record.

**Example:** "email" in a user table.

**Flattening**

Conversion of nested data structures into flat table form.

**Example:** JSON → DataFrame with columns for each key.

**Fact Table**

Core component of a data warehouse, stores measurable events.

**Example:** Sales table with sales, quantity, date.

**Factless Fact Table**

Table without numerical measures, but with relationships to analyze events.

**Example:** Attendance table for students - not a "value", but relationally usable.

**Gantt Chart**
diagram for visualizing the timeline of projects or processes. Used in planning and project management.
**Example:** Representation of ETL jobs over a week.

**Gaussian Distribution (Normal Distribution)**

Symmetrical, bell-shaped distribution of many natural features. Basis of many statistical methods.

**Example:** Height in a population.

**Gini Index**

Measure of impurity of a division in decision trees. The lower, the more homogeneous the classes.

**Example:** Gini = 0 for pure leaves.

**Git**

Code version control system. Allows parallel work, history and recovery.

**Example:** `git commit -m "Data cleansing added"`

**GitHub**

Online platform for managing Git repositories. Supports collaboration, reviews, and automation.

**Example:** Team shares notebooks via GitHub repo.

**Gradient Descent**

Optimization procedures to minimize error functions. Basis of almost all ML procedures.

**Example:** Training a neural network.

**GPU (Graphics Processing Unit)**

Processor for parallel computation, especially in deep learning.

**Example:** NVIDIA A100 accelerates CNN training.

**Granularity**

Depth of detail of data or time intervals. Fine = detailed, coarse = aggregated.

**Example:** Minute data vs. monthly averages.

**Graph Database**

NoSQL database for storing networked data. Uses knots and edges.

**Example:** Neo4j stores social networks.

**Grid Search**

Brute force method for hyperparameter optimization by testing all combinations.

**Beispiel:** `max_depth` + `n_estimators` für Random Forest.

**Group By**

SQL command for grouping rows by column values, often combined with aggregations.

**Beispiel:** `SELECT region, SUM(sales) FROM data GROUP BY region`

**Growth Rate**

Growth rate of a value over time.

**Example:** 8% revenue growth per month.

**Ground Truth**

Verified reference data for model validation.

**Example:** Manually labeled image data for a CNN.

**GUI (Graphical User Interface)**

User interface for interacting with software via visual elements.

**Example:** Tableau Dashboard with drag-and-drop.

**GxP (Good x Practice)**

Regulations for quality and safety in regulated areas.

**Example:** GMP in Pharma – Good Manufacturing Practice.

**GMM (Gaussian Mixture Model)**

Clustering model that models data as a mix of multiple normal distributions.

**Example:** Clustering of customers by behavior.

**Gradient Boosting**

Boosting process that sequentially reduces errors. Powerful for structured data.

**Example:** XGBoost.

**Guesstimate**

Rough, experience-based estimate.

**Example:** Expected survey response rate = 30%.

**Generalization**

Ability of a model to respond correctly to new data.

**Example:** Model also works on unknown customer data.

**Geoanalytics**

Analysis of spatial data with geographical components.

**Example:** Heatmap of sales figures per zip code.

**Gaussian Naive Bayes**

Classifier that assumes normal distributions per feature.

**Example:** Quickly trained text classifier.

**Greedy Algorithm**

Algorithm that makes the best local decision at every step. Not always optimal.

**Example:** Decision tree split with the highest information gain.

**Guided Analytics**

Interactive analysis with predefined questions or paths.

**Example:** User clicks through dashboard to goal insight.

**Gamma Distribution**

Skewed probability distribution for positive values.

**Example:** Insurance claims modeling.

**Gradient**

Vector of partial derivatives, points in the direction of the strongest increase of a function.

**Example:** Gradient in backpropagation.

**Gaussian Kernel**

Function for weighting close data points in kernel methods.

**Example:** SVM with RBF kernel.

**Group Normalization**

Alternative to batch normalization – robust for small batch sizes.

**Example:** In CNNs for small amounts of data.

**Graph Neural Network (GNN)**

Neural network for processing graph structures.

**Example:** Prediction of molecular properties based on their structure.

**Hash Function**
function to convert arbitrary data into solid code. Commonly used in security, indexing, or data reconciliation.
**Example:** SHA-256 generates a unique hash from a password.

**Histogram**

Graph showing the frequency distribution of numerical values in intervals (bins).

**Example:** Visualization of the age distribution in customer master data.

**Hyperparameter**

Preset model parameters that are not learned through training, but are determined manually or by optimization.

**Example:** Learning rate, number of trees in Random Forest.

**Hyperparameter Tuning**

Hyperparameter optimization to improve model performance.

**Example:** Grid Search to select the best `max_depth` and `n_estimators`.

**Hypothesis Testing**

Statistical method for testing an assumption about a population.

**Example** : Test if the average conversion rate is > 3%.

**Heteroskedasticity**

Non-constant variance of errors in a regression model. May lead to distorted results.

**Example:** Residuals increase with income.

**Heuristic**

Simplified rule for quick problem solving, not guaranteed optimal.

**Example:** "If user clicks >10x, he is interested."

**Holdout Set**

Part of the data that is not used for training, but only for the final evaluation of a model.

**Example:** 80/10/10 split: Training/Validation/Holdout.

**HDFS (Hadoop Distributed File System)**

Distributed file system for storing large amounts of data on clusters.

**Example:** Raw data is stored in blocks spread over several servers.

**Head (Table Function)**

Displays the first n rows of a record.

**For example,** `df.head(5)` shows the first five rows of a DataFrame.

**Hierarchical Clustering**

Cluster analysis, in which data is gradually merged into larger and larger groups.

**Example:** Dendrogram shows the hierarchy of customer clusters.

**Homogeneity**

Measure of similarity of groups or clusters. Higher = more similar.

**Example:** Cluster with pure age 20–25 is highly homogeneous.

**Host (Server)**

Computer or service running applications or databases.

**Example:** PostgreSQL runs on `analytics.company.com`

**HTML (HyperText Markup Language)**

Standard markup language for web pages. Relevant for web scraping.

**Example:** Extraction of data from `<table>` elements.

**HTTP (Hypertext Transfer Protocol)**

Protocol for data transmission on the web. Important for API calls and web scraping.

**Example:** REST API delivers JSON over HTTP GET.

**Heuristic Algorithm**

Algorithm that works with rules of thumb to find solutions efficiently.

**Example:** K-nearest-neighbor with simple distance measure.

**Histogram Equalization**

Image processing technology for contrast adjustment by rescaling the brightness distribution.

**Example:** Improving the readability of X-ray images.

**Hinge Loss**

Loss function for linear classification, especially for SVMs.

**Example:** Penalizes misclassified points at a distance from the decision boundary.

**Hamming Distance**

Number of different characters in two strings of equal length.

**Example:** "10101" vs. "11100" → Hamming distance = 3.

**Hash Join**

Join strategy in database systems where hash tables are used to find quick matches.

**Example:** Join between large tables in PostgreSQL.

**Hierarchical Indexing**

Multi-level index in pandas or SQL, often used to group and query multi-level data.

**Example:** MultiIndex of "Region" and "Year".

**Heatmap**

Color visualization of correlation or frequency data in matrix form.

**Example:** Correlation between features in the dataset.

**Heuristic Threshold**

Threshold value chosen empirically or based on experience.

**Example:** Lending at score > 0.6.

**Hybrid Model**

Combination of different types of models or algorithms, often ML+ rule-based.

**Example:** Recommendation system with collaborative filter + content-based matching.

**Hyperplane**

Separation surface in higher-dimensional spaces, used in SVMs for class demarcation.

**Example:** Two classes in 3D space are separated by a layer.

**Hyperparameter Optimization**

Systematic search for the best hyperparameters.

**Example:** Random Search, Grid Search, Bayesian Optimization.

**Human-in-the-Loop**

Systems in which people are specifically involved in the decision-making process.

**Example:** Human checks anomalies that have been marked by the model.

**Hazard Function**

Probability of an event occurring at a certain point in time, given that it has not yet occurred.

**Example:** probability of failure of one machine per hour.

**High Cardinality**

Column with a lot of different values – problematic for one-hot encoding.

**Example:** Email addresses, IDs, URLs.

**Hot Encoding (One-Hot-Encoding)**

Representation of categorical variables as binary columns.

**Example:** "Red" → [1,0,0], "Blue" → [0,1,0], "Green" → [0,0,1].

**ID (Identifier)**
Unique key used to distinguish records. Mostly used as a primary key.
**For example,** `user_id = 1023` identifies a specific customer.

**Imbalanced Dataset**

Dataset with unevenly distributed classes. Can strongly influence classification models.

**Example:** 95% "Non-fraud", 5% "Fraud".

**Imputation**

Procedure for replacing missing values.

**Example:** Fill in missing temperature values with average value.

**Index (SQL/Pandas)**

Structure for fast data access optimization. In pandas, in addition to line identification.

**Example:** Index on `customer_id` speeds up queries.

**Independent Variable**

Independent variable in an analysis, predictor.

**Example:** Advertising budget as an influencing factor on sales.

**Inferential Statistics**

Method for generalizing sample results to populations.

**Example:** confidence intervals, hypothesis tests.

**Information Gain**

Measure of reduction of impurity by an attribute (especially decision trees).

**Example:** Age greatly reduces the entropy of the purchase decision → high gain.

**Inner Join**

SQL join, which returns only matching rows of both tables.

**For example,** only customers with at least one order will be displayed.

**Instance**

Single example or data point in a dataset.

**Example:** A customer with attributes: age, gender, revenue.

**Interquartile Range (IQR)**

Range between 25th and 75th percentile. Robust against outliers.

**Example:** IQR for age is between 30 and 50 → IQR = 20.

**Interpolation**

Estimation of missing values between known points.

**Example:** Temperature on the 15th estimated by mean value of 14th and 16th.

**Interpretability**

Degree in which a model is understandable to humans.

**Example:** Decision tree is easy to interpret, a neural network is not.

**Interval Data**

Numerical data with equal distances but no true zero.

**Example:** Temperature in °C – 0 °C does not mean "no temperature".

**Intersection**

Intersection of two data sets or set operation in SQL.

**Example:** Users who have both purchased and rated.

**Iterative Process**

Repetitive process to refine models or workflows.

**Example:** Feature engineering → model training → evaluation → back.

**Isolation Forest**

ML method for anomaly detection through random partitioning.

**Example:** Conspicuous credit transactions are isolated.

**i.i.d. (independent and identically distributed)**

Assumption in statistics that data points come independently and from the same distribution.

**Example:** Coin tosses are usual, income is not necessarily.

**Identity Matrix**

Square matrix with ones on the diagonal, otherwise zeros.

**Example:** `I = [[1,0,0],[0,1,0],[0,0,1]]`

**Indicator Variable**

Binary variable for marking categorical characteristics.

**Example:** Gender: "female = 1", otherwise 0.

**Interaction Effect**

Interaction between two or more independent variables.

**Example:** Advertising effect depends on gender AND age.

**Incremental Learning**

Model training in small steps without complete relearning.

**Example:** Model updates with new user data every hour.

**Inertia (K-Means)**

Sum of the distances of all points to their cluster centers.

**Example:** The goal is minimal inertia → tight clusters.

**Input Layer**

First layer of a neural network, records raw data.

**Example:** 10 neurons for 10 input features.

**Image Recognition**

Recognition of objects or patterns in images using ML.

**Example:** Model identifies cats in photos.

**Indexing (Pandas)**

Access data rows or columns by labels or positions.

**Example:** `df.loc['row1']` or `df.iloc[0]`

**Inter-Rater Reliability**

Measure of agreement between multiple assessors.

**Example:** Two doctors make the same diagnosis → high reliability.

**IQR-Based Outlier Detection**

Outlier detection based on IQR.

**Example:** Values outside [Q1 – 1.5×IQR, Q3 + 1.5×IQR] are considered outliers.

**Imbalanced Learning**

ML techniques to better handle unequal class distributions.

**Example:** Using SMOTE to generate synthetic minority examples.

**Integer**

Integer data type with no decimal places.

**Example:** 1, 42, -7 – but not 3.14.

**Inferencing**

Applying a trained model to new data.

**Example:** Predicting buying behavior for new users.

**Indicator Matrix**

Matrix form for one-hot encoding of categorical data.

**Example:** 3 categories → 3 columns, 0 or 1 each.

**Identity Column (SQL)**

Automatically incrementing column for unique ID assignment.

**Example:** `id INT AUTO_INCREMENT`

**In-Memory Computing**

Process large amounts of data directly in RAM for acceleration.

**Example:** Apache Spark processes data in memory instead of disk.

**Information Retrieval**

Process of searching and finding relevant information in large amounts of data.

**Example:** Search for product reviews with a specific keyword.

**Instruction Set**

Instruction set of a processor or system, relevant in low-level computing.

**Example:** SIMD instructions for parallel processing of matrices.

**Jaccard Distance measure**
of the dissimilarity between two sets, defined as 1 minus the Jaccard similarity.
**Example:** Two lists with 60% overlap → Distance = 0.4

**Jaccard Index**

Measure used to calculate the similarity between two sets, defined as the magnitude of the intersection divided by the magnitude of the union set.

**Example:** Two sets with 4 of the same and 6 different elements result in 0.4.

**Jaccard Loss**

Loss function based on the Jaccard similarity used in image segmentation.

**Example:** Semantic segmentation networks.

**Jaccard Similarity Coefficient**

Alternative term for the Jaccard Index; is often used in clustering or recommender systems.

**Example:** Comparison of user interests using binary vectors.

**Jaccard Similarity Matrix**

Matrix with pairwise Jaccard scores between sets or documents.

**Example:** Similarity comparison of texts in a recommender system.

**Jaccard Thresholding**

Procedure for selecting similar pairs based on minimum value for the Jaccard Index.

**Example:** Only pairs with Jaccard > 0.5 are linked.

**JAR File (Java Archive)**

Compressed archive format for Java classes, configurations, and libraries.

**Example:** An Apache Spark job is passed as an executable JAR.

**Java**

Platform-independent, object-oriented programming language, commonly used in enterprise and big data applications.

**Example:** Hadoop MapReduce programs are usually written in Java.

**Java EE (Enterprise Edition)**

Extension of Java for web and enterprise applications, with a focus on scalability and modularity.

**Example:** Web service with authentication via Java EE.

**Java Native Interface (JNI)**

Interface for integrating C/C++ code into Java applications.

**Example:** Java calls a library for image processing in C.

**Java Server Pages (JSP)**

Technology for server-side generation of dynamic HTML content in Java.

**Example:** JSP page displays analysis results at the touch of a button.

**Java Virtual Machine (JVM)**

Virtual machine that translates and executes Java bytecode into machine code.

**Example:** Apache Spark runs on the JVM.

**JavaBeans**

Java components with defined getter and setter methods for structured data modeling.

**Example:** `getName()` and `setName()` as data access.

**JavaScript**

Scripting language for dynamic web development, also used for visualization tools.

**Example:** D3.js visualizations in the browser.

**Jena (Apache Jena)**

Framework for semantic web applications and processing of RDF data.

**Example:** SPARQL queries on knowledge graphs.

**Jenkins**

Open-source automation tool for continuous integration/delivery.

**Example:** Pipeline to execute ETL jobs on a daily basis.

**Jensen-Shannon divergence**

Measure for evaluating the similarity between probability distributions.

**Example:** Comparison of language models of two news texts.

**Jitter (visualization)**

Artificially scattering overlapping points in a plot for better readability.

**Example:** Point cloud with jitter at identical X values.

**Job Queue**

System for managing and processing asynchronous processes or tasks.

**Example:** Queue for image processing on a server.

**Joblib**

Python library for parallelization and serialization of tasks and models.

**Example:** Save model as `.pkl` with `joblib.dump()`.

**Join (SQL)**

Operation to combine rows from two tables based on a common attribute.

**Example:** `JOIN customers ON kunden.id = orders.customer no`.

**Joins (inner, outer, left, right)**

Variants of the SQL join with different result sets.

**Example:** LEFT JOIN shows all customers, even without an order.

**JSON (JavaScript Object Notation)**

Text-based, hierarchical format for storing and transmitting structured data.

**Example:** `{ "name": "Anna", "older": 30 }`.

**JSDOM**

JavaScript implementation of the DOM in Node.js environments.

**Example:** Testing websites without a real browser.

**JupyterHub**

Multi-user platform for deploying Jupyter notebooks to teams and educational institutions.

**Example:** Data science course with a central notebook server.

**JupyterLab**

Modern, advanced user interface for Jupyter notebooks with tabs and terminals.

**Example:** Opening CSV, code, and plot at the same time.

**Jupyter Notebook**

Web-based development environment for Python that combines code, text, and visualizations.

**Example:** Exploratory data analysis with Pandas and Seaborn.

**Jupyter Themes**

Customizable design packs to modify the look and feel of Jupyter.

**Example:** Dark background for better readability.

**Jupyter Widgets**

Interactive controls in Jupyter, such as sliders or dropdowns.

**Example:** Slider for parameters in an ML demo.

**JWT (JSON Web Token)**

Standard format for the secure transfer of information between parties.

**Example:** Access tokens for protected APIs.

**Jaro-Winkler Distance**

Similarity measure for strings with a focus on small swaps.

**Example:** Comparison of "Data" and "Dtaa" results in high similarity.

**k-anonymity data**
protection principle, which ensures that data cannot be uniquely traced back to individuals if it exists in groups of at least *k* indistinguishable individuals.
**Example:** A table is 3-anonymous if each combination of quasi-identifiers occurs at least three times.

**Kaggle**

Online platform for data analysis competitions, tutorials and community projects. It offers open datasets and an interactive Jupyter environment.

**Example:** Participation in a competition to predict housing prices.

**Kappa Score (Cohen's Kappa)**

Statistical measure used to evaluate the agreement between two classifiers, taking into account random matches.

**Example:** Comparison of human classification and model classification.

**KDE (Kernel Density Estimation)**

Nonparametric method for estimating the probability density of a random variable.

**Example:** Smoothing a histogram to analyze data distribution.

**Kendall's Tau**

Correlation coefficient for rank data, which evaluates the agreement of two rankings.

**Example:** Comparison of the ranking of products by two algorithms.

**Kernel Trick**

method in SVM to make nonlinear data separable by transformation into a higher-dimensional space.

**Example:** Using an RBF kernel for complex classification problems.

**Key-Value Store**

Simple NoSQL database system where data is stored as key-value pairs.

**Example:** Redis or Amazon DynamoDB.

**K-Fold Cross Validation**

Model validation technique, which involves splitting data into *k* parts and training and testing the model multiple times.

**Example:** 10-Fold Cross Validation for robust model scoring.

**KMeans**

Popular clustering algorithm that divides data into *k* groups based on their similarity.

**Example:** Customer segmentation by buying behavior.

**KMedoids**

Clustering methods similar to KMeans, but more robust against outliers because real data points are chosen as cluster centers.

**Example:** Clustering of users based on their browsing patterns.

**K-NN (K-Nearest Neighbors)**

Simple classification algorithm that determines the class of a point based on the majority class of its *k* nearest neighbors.

**Example:** Handwriting recognition based on pixels.

**Knowledge Graph**

network of entities and their relationships, which represents knowledge in a structured way.

**Example:** Google Knowledge Graph to improve search results.

**Kolmogorov-Smirnov-Test**

Nonparametric test to evaluate whether a sample follows a reference distribution.

**Example:** Checking whether data is normally distributed.

**Kolmogorov Complexity**

Measure of the amount of information of an object, defined as the length of the shortest program it generates.

**Example:** Random numbers have high Kolmogorov complexity.

**Complexity class**

Classification of problems according to their computational effort.

**Example:** P, NP, NP-difficult in the context of algorithm analysis.

**Confidence interval**

Range that contains the true value of a parameter with a certain probability.

**Example:** "The mean value is 95% certain to be between 10 and 12."

**Confusion matrix**

Table for evaluating classification models that represents true/untrue positive/negative values.

**Example:** Analysis of the accuracy of a spam filter.

**Contingency table**

Cross-table for the representation of frequencies of two categorical characteristics.

**Example:** Gender distribution and agreement with a statement.

**Continuous Variable**

Variable with an infinite number of possible characteristics in an interval.

**Example:** temperature, weight, income.

**Convergence (numerics)**

Property of an algorithm to approximate a stable value or solution.

**Example:** Gradient method in linear regression.

**Korrelation**

Statistical relationship between two variables.

**Example:** Positive correlation between advertising and sales.

**Correlation matrix**

Matrix with pairwise correlation coefficients of several variables.

**Example:** Comparison of stock returns.

**Covariance**

Measure of the joint variability of two random variables.

**Example:** If x rises and y increases, covariance is positive.

**Critical Value**

Threshold above which a statistical test result is considered significant.

**Example:** t-Critical = 2.01 for df=20 at α = 0.05.

**Kruskal-Wallis-Test**

Nonparametric test to analyze differences between more than two groups.

**Example:** Comparison of user reviews of multiple products.

**k-d Tree**

Data structure for quick search in multidimensional spaces.

**Example:** Efficient neighbor search in k-NN algorithms.

**Collinearity**

Problem in regression when independent variables are highly correlated.

**Example:** Weight and BMI at the same time in a regression analysis.

**Cross Table**

Synonym for contingency table, often used in Excel and statistical software.

**Example:** Display of the number of customers per region and gender.

**Cumulative distribution**

Function that specifies the probability that a random variable is less than or equal to a value.

**Example:** 80% of the values are below x = 15.

**Kurtosis (Wölbung)**

Statistical measure of the "sharpness" of a distribution.

**Example:** High kurtosis with highly concentrated data around the mean.

**K-anonymization**

Practical implementation of k-anonymity, often through generalization or suppression.

**Example:** Age 31 becomes age group 30–39.

**KPI (Key Performance Indicator)**

Key figure for evaluating processes, performance or goal achievement.

**Example:** Conversion rate, churn rate.

**Core**

Central element or influencing factor in a complex system.

**Example:** Feature with heavy weighting in a model.

**Knowledge Discovery in Databases (KDD)**

Entire process of pattern recognition in data, including pre-processing, modeling and interpretation.

**Example:** Data mining project for fraud detection.

**Combinatorics**

Subfield of mathematics for the counting of possible combinations and arrangements.

**Example:** Number of possible password variants with 3 characters.

**Label Encoding**
Method for converting categorical variables into numeric values by assigning an integer to each category.
**Example:** "red" = 0, "green" = 1, "blue" = 2

**Lag Feature**

Time-shifted variable in time-series analyses to use past values to predict future states.

**Example:** Yesterday's temperature as a feature for today.

**Lagrange Multiplier**

Mathematical method for considering constraints in optimization problems.

**Example:** Optimizing a model under resource constraint.

**Lambda Function (Python)**

Anonymous short function defined with the `lambda` keyword.

**Example:** `lambda x: x**2` gives the square of x.

**Laplacian (graph theory)**

Matrix to describe the structure of a graph, often used in clustering or graph-based ML algorithms.

**Example:** Laplace matrix in Spectral Clustering.

**Lasso (Least Absolute Shrinkage and Selection Operator)**

Regression method with L1 regularization that can set coefficients to zero.

**Example:** Feature selection by lasso regression.

**Latent Variable**

Non-directly observable variable that affects observed data.

**Example:** Customer loyalty as a latent influencing factor on purchasing behavior.

**Latent Dirichlet Allocation (LDA)**

Topic modeling techniques for discovering latent issues in textual data.

**Example:** Extraction of topics from user reviews.

**Layer (NN)**

Layer in a neural network that consists of nodes and performs transformations.

**Example:** Input Layer, Hidden Layer, Output Layer.

**Leaky ReLU**

Activation function in neural networks that also provides a small gradient for negative values.

**Example:** $f(x) = x$ for $x > 0$, $f(x) = 0.01x$ else.

**Performance (statistics)**

Probability that a test will correctly reject a false null hypothesis (power).

**Example:** A test with 80% power detects a real effect with 80% probability.

**Likelihood**

Probability that a model will produce given data, important for maximum likelihood estimates.

**Example:** Likelihood of a normal distribution given measured values.

**Likelihood Ratio Test**

Comparison of two nested models via the ratio of their likelihoods.

**Example:** Test whether an additional feature improves the model quality.

**Linear regression**

Statistical model that describes the relationship between a dependent and independent variable by means of a linear equation.

**Example:** Revenue = a + b * Advertising costs

**Linear Discriminant Analysis (LDA)**

Classification procedure that transforms feature spaces in such a way that classes can be easily separated.

**Example:** Separation of spam and non-spam mails.

**Linear independence**

Property of a variable set that no variable can be represented as a linear combination of the other.

**For example,** features with high correlation are not linearly independent.

**Linkage (Clustering)**

Strategy for calculating the distance between clusters in hierarchical methods.

**Example:** Single-linkage connects the nearest points of two clusters.

**Little's MCAR Test**

Statistical test to check if missing values are random (MCAR).

**Example:** Diagnosing failures in survey data.

**Local Outlier Factor (LOF)**

Method for identifying local outliers by density comparison with neighbors.

**Example:** Detection of rare events in sensor data.

**Log-Loss (Logarithmic Loss)**

Loss function for probabilistic classifiers, which severely penalises false, safe predictions.

**Example:** `-log(p)` at p = 0.01 results in high loss.

**Logarithmic transformation**

Transformation for the reduction of skewness in right-skewed distributions.

**Example:** Applying `log(x+1)` to income data.

**Logistic regression**

Classification model that predicts probabilities for binary classes.

**Example** : Predicting whether a customer will cancel.

**Long Short-Term Memory (LSTM)**

Special form of a recurrent neural network that can learn long-term dependencies.

**Example:** Text generation from sequences.

**Look-Up Table**

Table for quickly mapping input values to output values.

**Example:** Mapping codes to categories.

**Loss Function**

Function for quantifying the error of a model.

**Example:** Mean Squared Error in Regression.

**Low Cardinality**

Categorical characteristic with few different characteristics.

**Example:** "Gender" or "Day of the Week".

**Lurking Variable**

Hidden influencing factor that explains an apparent relationship between two observed variables.

**Example:** Ice consumption and drowning are both affected by the weather.

**Gap Analysis**

Comparison of the actual state with the target state to identify optimization potential.

**Example:** Sales target = 10M€, Ist = 8M€, Gap = 2M€.

**LZ77**

Algorithm for lossless data compression by detecting repetition patterns.

**Example:** Basis of the ZIP file format.

**LZMA (Lempel-Ziv-Markov chain algorithm)**

Efficient compression algorithm with high compression rate.

**Example:** 7z archive format.

**L1 regularization**

Regulatory method of penalizing large coefficients in models, promotes thriftiness.

**Example:** Lasso regression.

**L2 regularization**

Penalizes large coefficients square, stabilizes the model.

**Example:** Ridge regression.

**Latent Semantic Analysis (LSA)**

Text analysis method that determines latent meaning relationships between words.

**Example:** Document clustering by content.

**Latent Space**

Abstract feature space into which data is projected by a model.

**Example:** Representation of images in an autoencoder.

**Lemmatization**

Text pre-processing step to return words to their basic form.

**Example:** "went" becomes "go".

**Machine learning (ML)**
Umbrella term for methods in which models learn from data without being explicitly programmed.
**Example:** A model learns to recognize spam emails.

**Manifold Learning**

Nonlinear dimension reduction for the discovery of low-dimensional structures in high-dimensional data.

**Example:** t-SNE or Isomap for data visualization.

**MapReduce**

Distributed programming model for processing large amounts of data.

**Example:** Google uses MapReduce to index the web.

**Marginalization**

Integration via unimportant variables to simplify distributions.

**Example:** $P(X) = \int P(X,Y)\, dY$.

**Markov Necklace**

Model in which the probability of the next state depends only on the current state.

**Example:** Weather model: Sun → rain with a defined transition probability.

**Markov Decision Process (MDP)**

Mathematical model for decision-making under uncertainty.

**Example:** Optimal strategy in reinforcement learning.

**MAE (Mean Absolute Error)**

Average absolute error between forecast and observation.

**Example:** MAE of 2 means that predictions differ by an average of 2 units.

**Mean**

Arithmetic mean of a series of numbers.

**Example:** Mean of [2, 4, 6] is 4.

**Mean Imputation**

Replace missing values with the mean value of the column.

**Example:** Missing age entries are replaced by the average.

**Mean Shift**

Clustering method that identifies density maxima and aligns clusters with them.

**Example:** Grouping customers into density regions.

**Mean Squared Error (MSE)**

Average of the quadratic error between the prediction and the real value.

**Example:** Large MSE shows strong deviation.

**Median**

Central value of a sorted data series.

**Example:** Median of [1, 3, 9] is 3.

**Median Imputation**

Replacement of missing values by the median.

**Example:** Robust method for outliers.

**Membership Inference Attack**

Attack that determines whether certain data was used in training a model.

**Example:** Attack on an ML model to extract training data.

**Memory-Based Learning**

Learning method in which all examples are stored and used to predict.

**Example:** k-NN stores all data points.

**Meta-Learning**

"Learning to learn": Models learn how to solve new tasks quickly and efficiently.

**Example:** Few-shot learning in image classification.

**Metric Learning**

Learning a distance function that correctly maps relevant similarities.

**Example:** Face comparison based on learned similarity metrics.

**Min-Max Normalization**

Scaling values to a defined range, usually [0, 1].

**Example:** Values between 5 and 10 are stretched to 0–1.

**Minimum Description Length (MDL)**

Principle of model selection based on the brevity of the description of data plus model.

**Example:** Preference for simple models with good explanatory power.

**Minimum Spanning Tree**

Subgraph with minimal total edge weighting that connects all nodes.

**Example:** Network optimization for cable connections.

**Missing Completely at Random (MCAR)**

Missing data is completely random and independent of observed or unobserved values.

**Example:** Sensor error without a systematic cause.

**Missing Not at Random (MNAR)**

Missing data is related to the missing values themselves.

**Example:** High incomes are disclosed more often than average.

**Missing at Random (MAR)**

Missing data is only related to observed values.

**Example:** Age influences the probability of missing income information.

**Fashion**

The most common value in a data set.

**Example:** Mode of [1, 2, 2, 3] is 2.

**Model Drift**

Loss of model accuracy over time due to changes in the data.

**Example:** A recommendation model ages as user behavior changes.

**Model Interpretability**

Comprehensibility of the decision logic of a model for humans.

**Example:** Decision tree is easier to interpret than a neural network.

**Model Selection**

Selection of the best model based on validation criteria.

**Example:** Comparing multiple regression models with cross-validation.

**Model Zoo**

Collection of pre-trained models, often with open weights and documentation.

**Example:** TensorFlow Hub or HuggingFace Transformers.

**Model-Based Clustering**

Clustering based on the assumption that data comes from a mix of statistical models.

**Beispiel:** Gaussian Mixture Models.

**Model Complexity**

Degree of freedom and parameters of a model.

**Example:** Neural networks with many layers are more complex than linear models.

**Monte Carlo Simulation**

Random-based simulation to approximate probability distributions.

**Example:** Forecasting project risks through many runs.

**Multicollinearity**

Problem in regression when independent variables are strongly correlated.

**Example:** Weight and BMI as regressors.

**Multi-label classification**

Classification issue with multiple applicable labels per instance.

**Example:** A film can be considered a comedy and action at the same time.

**Multivariate analysis**

Analysis of several dependent variables at the same time.

**Example:** Simultaneous prediction of weight and blood pressure.

**Mutual Information**

Measure of the dependence between two variables.

**Example:** MI = 0 for independent variables.

**MVP (Minimum Viable Product)**

Simplest functional version of a product for testing on the market.

**Example:** Prototype an app with core functionality.

**MXNet**

Deep learning framework with a focus on performance and scalability.

**Example:** Using MXNet for GPU training in the cloud.

**MySQL**

Popular relational database management system (RDBMS).

**Example:** Storage of structured transaction data.

**Naïve Bayes**
A simple, probabilistic classification method based on Bayes' theorem with the assumption of conditional independence of traits.
**For example,** spam filters classify emails as spam/non-spam based on word probabilities.

**Named Entity Recognition (NER)**

Methods from natural language processing for the identification of named entities such as names, places, organizations in texts.

**Example:** Recognition of "Berlin" as a city in a text.

**NAND Gates**

Logic gate in digital technology that provides an output signal if both inputs are not 1. It can be used universally.

**Example:** Basis for memory logic in CPUs.

**Natural Language Processing (NLP)**

A subfield of AI for processing, analyzing, and generating natural language.

**Example:** chatbots, machine translation, text classification.

**Natural Logarithm (ln)**

Logarithm to base e (Euler number, about 2.718), used in exponential growth and decay, as well as in many ML algorithms.

**Example:** ln(x), often used in log-linear models.

**Negative Binomial Distribution**

Probability distribution for the number of failed attempts until the rth success.

**Example:** Modeling the number of customer calls up to the third complaint.

**Negative Sampling**

Technique for efficient training of neural networks with very large output quantities by targeted selection of negative examples.

**Example:** Training Word2Vec models.

**Nested Queries (SQL)**

Queries within other queries, often referred to as subqueries.

**Beispiel:** `SELECT * FROM users WHERE id IN (SELECT user_id FROM orders)`

**Neural Network**

Machine learning model consisting of layers of networked artificial neurons that recognizes complex patterns in data.

**Example:** Image recognition or speech recognition through deep learning.

**NLP Pipeline**

Processing chain for text data in NLP, often consisting of tokenization, stop word removal, lemmatization, etc.

**Example:** Analysis of customer feedback through structured steps.

**Noise (statistics)**

Unsystematic, random glitches or errors in data that cannot be explained by the model.

**Example:** Measurement errors in sensor values.

**Noise Reduction**

Method of removing or minimizing noise in data.

**Example:** Smoothing of time series by moving average.

**Nominal Variable**

Categorical variable with no natural order.

**Example:** Colors: Red, Blue, Green.

**Normalization**

Scaling numeric values to a uniform range, often between 0 and 1.

**Example:** `x' = (x - min) / (max - min)`

**Normal distribution**

Bell-shaped probability distribution with mean and standard deviation. Frequent assumption in statistics.

**Example:** Body size distribution in a population.

**Null Hypothesis (H0)**

Assumption that there is no effect or difference. Basis for many statistical tests.

**Example:** "The advertising measure had no influence on sales."

**Null Value**

Special marking of missing or undefined values in databases or programs.

**Example:** NULL in SQL means no value exists.

**Numerical Feature**

Feature with continuous or discrete numeric values.

**Example:** age, price, temperature.

**Numerical Integration**

Calculation of approximations for certain integrals when no analytical solution is possible.

**Example:** Trapezoidal rule or Monte Carlo method for area calculation.

**NumPy**

Python library for numerical calculations and efficient array operations. The foundation of many data analysis tools.

**Example:** Vector operations with `numpy.array()`.

**N-gram**

Sequence of N consecutive elements (e.g., words or characters) in text data, used for language modeling.

**Example:** Trigram of "I love you" = ["I love", "love you"]

**Yy**

Special representation of invalid or missing numeric values in programs such as Python or R.

**Example:** Dividing by zero equals NaN in `pandas`.

**Nearest Neighbor Search**

Algorithm for searching for the nearest points in the feature space, basis for k-NN and clustering.

**Example:** Recommendation of similar products.

**Nested Cross Validation**

Combination of two nested cross-validation loops for fair model and hyperparameter evaluation.

**Example:** outer CV for performance measurement, inner CV for hyperparameter optimization.

**NetworkX**

Python library for analyzing and visualizing complex networks.

**Example:** Social network analysis or transport networks.

**Newton-Raphson method**

Iterative method for solving nonlinear equations.

**Example:** Root determination or maximum likelihood estimates.

**Node (graph theory)**

A single element in a network or tree, such as a user on a social network.

**For example,** each node in a decision tree is a node.

**NoSQL**

Database technologies that are not based on relational tables, often document- or graph-based.

**Example:** MongoDB or Cassandra.

**Null model**

Simple basic model without explanatory variables, serves as a reference for evaluating more complex models.

**Example:** Mean model as a comparison for linear regression.

**Numerical stability**

Measure of the robustness of numerical algorithms against rounding errors.

**Example:** Using stable matrix operations in ML.

**Nyquist-Theorem**

Signal processing theorem that describes the minimum sampling frequency for the exact reconstruction of a signal.

**Example:** Audio sampling must be done at at least twice the frequency.

**NamedTuple (Python)**

Data type in Python for defining tuples with named fields, similar to classes.

**Beispiel:** `Point = namedtuple('Point', ['x', 'y'])`

**Nesterov Momentum**

Predictive gradient optimization method, improves neural network convergence.

**Example:** Training acceleration compared to classic momentum.

**Noise Injection**

Technique to increase model robustness by intentionally inserting noise into training data.

**Example:** Image noise in image classification.

**Normalized Mutual Information (NMI)**

Metric used to evaluate the correspondence of two clusterings, scaled to [0,1].

**Example:** Comparing clustering results to ground truth.

**Numerical differentiation**

Approximation of the derivative by difference quotients.

**Example:** Finite difference method in optimization methods.

**Newtonian method (multivariable)**

Extension of the Newton-Raphson method to several variables for optimization.

**Example:** Use in non-convex objective functions in ML.

**Utility**

Measure of the value or benefit of an action or prediction, often found in decision trees or recommender systems.

**Example:** Recommendation with maximum expected benefit.

**Object Detection**

Machine vision method for locating and classifying several objects in an image.

**Example:** Real-time detection of vehicles and pedestrians for autonomous vehicles.

**Object-Oriented Programming (OOP)**

Programming paradigm that encapsulates data and behavior in objects. Facilitates reusability, modularity, and maintenance.

**Example:** Classes defined in Python for modeling data transformations.

**Observability**

Ability to determine the internal state of a system through external outputs.

**Example:** Log files and metrics for analyzing data pipelines.

**Observation**

Single data point in a data set, usually one row.

**Example:** A customer with all attributes in a CRM table.

**Occam's Razor**

Principle according to which the simpler model is preferred for the same quality.

**Example:** Choosing a linear model instead of a deep neural network with the same performance.

**OCR (Optical Character Recognition)**

Technology for automatic text recognition in images or scanned documents.

**Example:** Digitization of invoices in PDF format.

**Octile Distance**

Metric used to calculate distances in grids with diagonal movements.

**Example:** Pathfinding in grid maps.

**ODS (Operational Data Store)**

Central repository of operational data for near real-time reporting and analysis.

**Example:** Data from multiple systems merged into one dashboard.

**Offline Learning**

Model training on a static, previously known dataset.

**Example:** Classifier training on historical user behavior.

**OGNL (Object Graph Navigation Language)**

Expression language for navigating and manipulating object graphs, e.g. in Java frameworks.

**Example:** Accessing nested values in JavaBeans.

**OLS (Ordinary Least Squares)**

Standard Methods for Estimating Linear Regression Models.

**Example:** Minimizing squared errors to fit a regression line.

**One-Hot-Encoding**

Categorical encoding, where each category is represented as a binary vector.

**Example:** "red", "blue", "green" → [1,0,0], [0,1,0], [0,0,1]

**One-Class SVM**

Support vector machine for anomaly detection in a single class.

**Example:** Detection of fraud based on "normal" behavior.

**Online Learning**

Learning method in which the model is gradually updated with new data.

**Example:** Customization of a recommendation system in real time.

**Ontology**

Formal description of terms and their relationships within a field of knowledge.

**Example:** Data model for medical diagnoses with ICD terms.

**Open Data**

Data that can be freely used, reused and redistributed.

**Example:** Traffic data of a city government released to developers.

**Open Source**

Software whose source code is publicly available and may be modified.

**Example:** Python libraries such as Pandas or Scikit-learn.

**Operationalization**

Translation of abstract concepts into measurable variables.

**Example:** "Customer satisfaction" is operationalized by a survey with a 5-point scale.

**Optimization**

The process of improving a model, algorithm or system by fine-tuning it.

**Example:** Hyperparameter tuning with Grid Search.

**Optimizer**

Algorithm for adjusting the model parameters during the learning process.

**Example:** Adam optimizer in neural networks.

**Ordinal Data**

Data with natural order, but no fixed distance between values.

**Example:** Satisfaction survey: "very bad" to "very good".

**Outlier**

Data point that differs significantly from others.

**Example:** Income of €1,000,000 in a data set with an average value of €50,000.

**Outlier Detection**

Procedure for identifying outliers.

**Example:** Isolation Forest or Z-Score.

**Output Layer**

Last layer in a neural network that provides the final prediction.

**Example:** Softmax layer for classification with multiple classes.

**Overfitting**

Model adaptation that is too strongly oriented towards training data and loses generalization capability.

**Example:** Complex model with 100% training accuracy but poor test performance.

**Oversampling**

Technique for increasing the number of rare classes in unbalanced datasets.

**Example:** SMOTE for the artificial creation of minority classes.

**Own Join**

Join operation in which a table is joined to itself.

**Example:** Analyze hierarchies in employee data.

**Ox Metrics**

Software package for econometric modelling and time series analysis.

**Example:** Execution of ARIMA models.

**Out-of-Bag Error**

Error estimation in bagging procedures based on data that was not used in bootstrapping.

**Example:** Random Forest uses OOB data for internal validation.

**Out-of-Sample Performance**

Model performance on unknown data not used in training.

**Example:** Validation results in the holdout set.

**Out-of-Vocabulary (OOV)**

Words that are not included in the training vocabulary of an NLP model.

**Example:** Dealing with new slang terms in chatbots.

**Outlier Score**

Numeric value that indicates how much of an outlier a data point is.

**Example:** LOF score > 1.5 is often considered an outlier.

**Ordinal Encoding**

Mapping integers to ordered categorical variables.

**Example:** "low" = 0, "medium" = 1, "high" = 2.

**Oracle**

System or component that is assumed to be omniscient and is used for comparison purposes.

**Example:** Theory model with perfect knowledge as a benchmark.

**Operational Metric**

Key figure for monitoring operational processes and data processing systems.

**Example:** Latency or error rate of a pipeline.

**One-vs-Rest (OvR)**

Strategy for extending binary classifiers to multiclass problems.

**Example:** Three binary models for classes A vs B+C, B vs A+C, C vs A+B.

**Ordinal Logistic Regression**

Regression model for ordinally scaled target variables.

**Example:** Analysis of customer satisfaction scales.

**Online Analytical Processing (OLAP)**

Technology for fast multidimensional analysis of large amounts of data.

**Example:** Drill down quarterly sales by region and product.

**OpenAI API**

Programming interface for the use of language models and AI services from OpenAI.

**Example:** Text generation by an API call from a Python application.

**P-Value Statistical**
measure used to evaluate the significance of an outcome. A low p-value indicates that an observed result cannot be explained by chance.
**For example,** a p-value of 0.01 means that the probability of the outcome below the null hypothesis is 1%.

**Pandas**

Python library for data manipulation and analysis. It provides powerful data structures such as DataFrames.

**For example,** `df = pd.read_csv("daten.csv")` loads a CSV file into a DataFrame.

**Parameter**

Fixed values in a statistical model that need to be estimated. You define the behavior of the model.

**Example:** In a linear regression, the slope is a parameter.

**Parquet**

Column-based storage format, optimized for large amounts of data. Supports efficient queries and compression.

**Example:** Storing a DataFrame in `data.parquet` for quick analysis.

**Partial Dependence Plot (PDP)**

Visualization of the influence of a feature on the model result, controlling for all other features.

**Example:** PDP shows how the house price develops as living space increases.

**Partitioning**

Breaking down data into logical or physical parts, such as databases or data lakes.

**Example:** Partitioning a table on a monthly basis to increase performance.

**Pearson correlation**

Measure of linear relationship between two variables. Values range from -1 (negative) to +1 (positive).

**Example:** Correlation between learning time and exam result.

**Percentile**

Thresholds that divide a distribution into 100 equal parts.

**For example,** the 90th percentile is the value below which 90% of the data falls.

**Permutation test**

Nonparametric test to determine significance by randomly rearranging the data.

**Example:** Comparison of mean values of two groups by permutation.

**Pipelines**

Order of processing steps, e.g. in data preprocessing and ML training.

**Example:** scaling, feature engineering, model training – all in one pipeline.

**Pivot Table**

Excel function for fast aggregation and analysis of large amounts of data.

**Example:** Summation of sales by product and region.

**Plotly**

Interactive visualization library in Python. Supports dynamic graphics for web and dashboarding.

**Example:** `plotly.express.scatter()` for interactive scatter plots.

**Poisson Distribution**

Probability distribution for events with a constant average rate.

**Example:** Number of calls per hour in the call center.

**Polynomial Regression**

Regression model with nonlinear relation, using higher-order polynomial terms.

**Example:** Predicting sales figures as advertising grows with decreasing marginal effect.

**Population**

Totality of all elements about which statistical statements are made.

**Example:** All citizens of a country in a survey.

**Porting**

Transfer code or data from one platform to another.

**Example:** Porting a script from R to Python.

**Precision**

Proportion of elements correctly classified as positive out of all elements classified as positive.

**Example:** 80% Precision means: Out of 100 people predicted to be "sick", 80 are really sick.

**Precision Recall Curve**

Visualization of precision and recall at different thresholds.

**Example:** Basis for decision-making in unbalanced data sets.

**Predictive Modeling**

Create models to predict future events based on historical data.

**Example:** Forecasting customer abandonment with ML model.

**Prescriptive Analytics**

Analytical approach that not only predicts what will happen, but also provides recommendations for action.

**Example:** Recommending price changes based on demand forecasting.

**Principal Component Analysis (PCA)**

Dimensional reduction method that projects data based on the main directions of variance.

**Example:** Reduction of 100 features to 3 main components.

**Priority probability**

Subjective initial probability before observing data, e.g. in Bayesian theory.

**Example:** Expectation that 5% of customers will quit.

**Probability Density Function (PDF)**

Function that describes the probability distribution of a continuous random variable.

**Example:** Bell curve with normal distribution.

**Probability Mass Function (PMF)**

Equivalent of PDF for discrete random variables.

**Example:** Number of sixes rolled in 10 rolls.

**Process Mining**

Technology for analyzing real business processes based on log data.

**Example:** Discovery of inefficiencies in support processes.

**Profiling (data profiling)**

Analysis of the structure, quality and properties of data sets.

**Example:** Detection of duplicates, null values, inconsistencies.

**Prophet (Facebook)**

Open-source tool for time series forecasting with a simple API.

**Example:** Forecasting seasonal sales figures.

**Logging**

Systematic recording of processes, errors or transactions.

**Example:** Saving requests to a debugging API.

**Python**

Widely used programming language in the data science field, known for readability and extensive libraries.

**Example:** Using `numpy`, `pandas`, `scikit-learn` for data analysis.

**PyTorch**

Python framework for deep learning with dynamic computational graphing.

**Example:** Building and training neural networks with GPU support.

**PySpark**

Python interface for Apache Spark for distributed computing.

**For example,** processing large CSV files in the cluster.

**Pseudocode**

Programming-related description of algorithms in plain text form, independent of programming language.

**Example:** Description of a sorting procedure in structured text.

**P-value correction**

Adjustment of p-values in multiple tests to control the type of error.

**Example:** Bonferroni correction for multiple hypothesis tests.

**pandas_profiling**

Python library for fast automated data analysis and creation of an EDA report.

**Example:** `df.profile_report()` generates PDF with statistics and plots.

**Point-Biserial Correlation**

Correlation between a binary and a metric variable.

**Example:** Relationship between gender and income.

**Poisson Process**

Stochastic process for countable events over continuous time.

**Example:** Call center call modeling.

**Power BI**

Microsoft BI tool for dashboards, reports, and data visualization.

**Example:** Connection to Excel and visualization of sales figures.

**PostgreSQL**

Powerful, object-relational open source database system.

**Example:** Using SQL and JSON functions for data analysis.

**Probability Calibration**

Adjustment of prediction probabilities for better interpretation.

**Example:** Platt scaling on uncalibrated models.

**Prediction Interval**

Interval that includes future individual observations with a defined probability.

**Example:** Forecast of tomorrow's temperature: 17–21 °C with 95% certainty.

**Precision Medicine**

Approach in medicine where decisions are based on individual patient data.

**Example:** Personalized cancer treatment based on genome data.

**Q-Q Plot (Quantile-Quantile Plot)**
Graphical method for comparing two distributions by plotting their quantiles against each other.
**Example:** Normal distribution Q-Q plot shows whether data is normally distributed (points are on the diagonal).

**Q-Learning**

Reinforcement learning process in which an agent learns from rewards which actions are most rewarding in which state.

**Example:** An autonomous agent learns to avoid obstacles and collect rewards.

**Quadratic Loss**

Loss function, where the difference between prediction and true value is squared.

**Example:** Mean Squared Error (MSE) is a form of Quadratic Loss.

**Quadratic Programming (QP)**

Optimization problem with quadratic objective function and linear constraints.

**Example:** Portfolio management with risk minimization.

**Qualitative Data**

Non-numeric, categorical data that describes states or characteristics.

**Example:** colors, product categories, customer opinions.

**Quantile**

Values that divide a distribution into equal intervals.

**Example:** The 25% quantile (Q1) is the value below which 25% of the data is located.

**Quantile Regression**

Regression method that does not model the mean, but a certain quantile of the target variable.

**Example:** Forecasting the 90% quantile of the delivery time.

**Quantitative Data**

Number-based data that can be measured or counted.

**Example:** age, turnover, temperature.

**Quantization**

Reduction of the precision of values to a discrete set, often used in ML for model compression.

**Example:** Compression of neural networks by 8-bit quantization.

**Query**

Request to a database to extract certain information.

**Example:** `SELECT * FROM customers WHERE land = 'DE'` is an SQL query.

**Query Optimization**

Process of improving the execution speed of database queries.

**Example:** Using indexes and join strategies to accelerate queries.

**Cue**

Data structure in which elements are processed in the order in which they arrive (FIFO).

**Example:** Queue for event processing.

**QuickSort**

Efficient, recursive sorting algorithm with a divide-and-conquer approach.

**Example:** Sorting an array with a pivot element.

**Quota Sampling**

Non-random sampling method in which certain group proportions are specifically collected.

**Example:** 50% women, 50% men in a survey.

**Quasi-Experiment**

Study with experimental design without random assignment to groups.

**Example:** Investigation of the effect of a price change without a random sample.

**Quasi-Newton Method**

Approximation method for numerical optimization based on an approximation of the Hesse matrix.

**Example:** BFGS algorithm to minimize a cost function.

**Quadrant Analysis**

Analysis method for classifying data points on the basis of two axes, usually divided into four quadrants.

**Example:** Prioritizing tasks according to importance and urgency.

**Quantum Computing**

Computational paradigm that uses quantum mechanical states for parallel information processing.

**Example:** qubits instead of bits for exponential computing power.

**Query Plan**

Internal execution plan of a database for processing a request.

**Example:** Presentation of the steps of an SQL join to analyze performance.

**Quality Assurance (QA)**

Systematic processes to ensure the quality of data, models and software.

**Example:** Validation of data pipelines and unit tests for ML models.

**Quantitative Trait**

Trait that is described by continuously measurable values and is often influenced by several genes.

**Example:** height or blood sugar level.

**Quadratic Mean (RMS)**

root from the average of squared values; robust against outliers.

**Example:** Calculation of the RMS voltage.

**Query Language**

Programming language for formulating database queries.

**Example:** SQL, GraphQL.

**Quality Score**

Evaluate the quality of a data point, prediction, or model.

**Example:** Evaluation of ad effectiveness in marketing.

**Square Root Transformation**

Transformation to reduce skewness in countable data.

**Example:** `sqrt(x)` for countable events such as accident numbers.

**Quantile Normalization**

Method for normalizing multiple distributions to the same quantile distribution.

**Example:** Comparison of gene expressions across different experiments.

**Quicksight (AWS)**

Amazon's BI tool for data visualization and dashboard creation.

**Example:** Creating interactive sales dashboards for e-commerce.

**Quorum**

Minimum number of participants required for a decision or system behavior.

**Example:** replication systems in distributed databases.

**Qubit**

Elementary unit of information in quantum computing with superposition states.

**Example:** A qubit can be 0 and 1 at the same time.

**Query Federation**

technique to execute queries across multiple data sources at the same time.

**Example:** Combining data from S3, Redshift, and MySQL into one query.

**Queueing Theory**

Mathematical theory for modeling queue processes.

**Example:** Optimization of call center capacities.

**Query Caching**

Caching of query results to speed up repeated accesses.

**Example:** Redis as a cache for complex SQL reports.

**QGIS (Quantum GIS)**

Open source software for editing, analysing and visualising geographical data.

**Example:** Display of customer locations on a map.

**Quasi-Poisson Regression**

Variant of Poisson regression that takes into account overdispersion in countable data.

**Example:** Modeling call counts with varying daily loads.

**Quality Control Chart**

Diagram to monitor quality in processes by statistical limits.

**Example:** SPC control card for production monitoring.

**Square Matrix**

Matrix with equal number of rows and columns, important for linear algebra and eigenvalue analysis.

**Example:** Covariance matrix.

**Quantitative PCR (qPCR)**

Laboratory method for the quantitative determination of DNA/RNA quantities.

**Example:** Detection of viral loads in medical tests.

**Quadrature rule**

Numerical method for approximating integrals.

**Example:** Trapezoid rule for area calculation under curves.

**Query Result Cache**

Storage area where database responses are cached for faster access.

**Example:** Oracle Query Result Cache.

**Quasi-Binomial Model**

Generalized linear model that allows overdispersion in binary results.

**Example:** Modeling conversion rates in online advertising.

**Quantitative Forecasting**

Forecasting method that uses numeric time series data.

**Example:** Sales forecast based on historical sales figures.

**Quotient correlation**

Ratio-based correlation of two variables, used in dimension reduction.

**Example:** Category A purchases as a percentage of total purchases.

**Squared Error**

Error value, which is calculated by squaring the difference between the actual and target value.

**Example:** (y - ŷ)^2 for forecast deviations.

**Quick Ratio**

Measure of a company's short-term liquidity.

**Example:** (Current Assets - Inventories) / Current Liabilities.

**R (Programming Language)**
Statistically oriented programming language with strong support for data analysis, visualization, and scientific computing.
**Example:** R is often used in academic research, such as for linear models or ggplot2 visualizations.

**R-squared ($R^2$, coefficient of determination)**

Statistical measure used to evaluate the explanatory power of a regression model. Values close to 1 mean high explanatory quality.

**For example,** an $R^2$ of 0{,}85 means that 85% of the variance is explained by the model.

**Random Forest**

Ensemble learning method that combines many decision trees to robustly perform classifications or regressions.

**Example:** A random forest can be used to predict credit risk.

**Random Sampling**

Random selection of observations from a population for estimation or analysis.

**Example:** Randomly drawing 1,000 customers for a survey.

**Random Variable**

Variable whose value depends on the result of a random process.

**Example:** Number of a dice rolled.

**Range**

Difference between the largest and smallest value in a data set.

**Example:** For the values 2, 4, 6, the range 6 - 2 = 4.

**Rank Transformation**

Transformation of numerical values into their ranking.

**Example:** Values [100, 50, 75] become ranks [3, 1, 2].

**Rasch model**

Statistical model for scaling latent features, often in psychometrics.

**Example:** Analysis of student answers in standardized tests.

**Rate Limiting**

Technique to limit the number of API requests within a time window.

**Example:** Max. 1000 requests per hour for a web service.

**Rationalization**

Data cleansing by removing redundant or irrelevant information.

**Example:** Consolidation of duplicate entries in a customer list.

**Raw Data**

Unprocessed raw data as generated directly from sources.

**Example:** CSV export from a sensor API.

**Recall**

Classification model metric that measures how many of the actual positive cases were correctly detected.

**Example:** 80% recall means: 80% of all positives were recognized.

**Receiver Operating Characteristic (ROC)**

Curve to visualize the trade-offs between sensitivity and specificity.

**Example:** Area under the ROC curve (AUC) as a measure of model quality.

**Recoding**

Encoding or recoding variable values into another structure.

**Example:** "m" and "f" become 0 and 1.

**Recursive Feature Elimination (RFE)**

Feature selection technique, in which less relevant features are gradually removed.

**Example:** RFE with random forest to reduce from 100 to 10 important features.

**Regression**

Statistical method for modelling relationships between dependent and independent variables.

**Example:** Predicting house prices based on size, location, condition.

**Regression Tree**

Decision tree model for predicting numerical target variables.

**Example:** Decision paths for predicting salary values.

**Regularization**

Technique to avoid overfitting by penalizing large coefficients.

**Example:** L1 and L2 regularization in regression models.

**Reinforcement Learning**

Learning paradigm in which an agent learns optimal strategies through reward and punishment.

**Example:** Training process of a chess program.

**Relational database**

Database system with tabular structure and relationships via keys.

**Example:** MySQL database for customer data.

**Relationship**

Link between tables in relational databases.

**Example:** Foreign key connects customers and orders.

**Relative frequency**

Share of a value in the total.

**Example:** 40 out of 200 customers have purchased → 20% relative frequency.

**Replication**

Repetition of an analysis to verify results.

**Example:** Reproducing an ML model with new data.

**Residual (Residuum)**

Difference between observed and predicted value.

**Example:** Observation = 10, Forecast = 8 → Residual = 2.

**Residual Sum of Squares (RSS)**

Sum of the squared residuals, measure of model quality in the regression.

**Example:** Low RSS indicates good model fit.

**Resolution (raster data)**

Measure of the level of detail of raster images or data.

**Example:** 10x10 pixel resolution per km in an elevation model.

**Resampling**

Techniques such as bootstrapping or cross-validation to increase the robustness of estimates.

**Example:** Bootstrapping for confidence interval estimation.

**REST API**

Web service architecture based on HTTP methods.

**Example:** GET, POST, PUT, DELETE on resources such as /customer.

**Results Matrix**

Matrix with output results of an analysis method or model.

**Example:** Confusion matrix in classification.

**Retail Analytics**

Data analysis in retail to optimize assortment, pricing, warehousing.

**Example:** Analysis of cash register data to forecast sales.

**Ridge Regression**

Regression form with L2 regularization to avoid overfitting.

**Example:** Stabilization for multicollinear data.

**Right Join**

SQL operation that outputs all rows of the right table and matching the left table.

**Example:** Customers without an order do not show up, but all orders are displayed.

**ROC AUC**

Metric used to evaluate classification models, measures area under the ROC curve.

**Example:** AUC value of 0.95 shows high classification quality.

**Root Mean Square Error (RMSE)**

Root of mean square error, common measure of forecast quality.

**Example:** RMSE = 2 means an average of 2 units of deviation.

**Round()**

Function for rounding numeric values.

**Example:** round(3.14159, 2) equals 3.14.

**Row-Level Security**

Technique for restricting data access at the row level, e.g. in BI tools.

**Example:** User only sees sales data from their region.

**R Script**

File with R commands for repeatable analysis and visualization.

**Example:** Automated reporting in RMarkdown.

**Runtime**

Execution time of a program or model, often critical for big data.

**Example:** Python script runs through in 12 seconds.

**Run-Length Encoding (RLE)**

Compression method in which repetitions are encoded by counting.

**Example:** "AAAABBB" → "4A3B".

**Backpropagation**

Learning method for neural networks that propagates errors back through the network.

**Example:** Training a CNN by minimizing the error.

**Inference statistics**

Subfield of statistics for the generalization of samples on populations.

**Example:** confidence interval, hypothesis test.

**Sample**
subset from a larger population that is used for analysis.
**Example:** 500 customer surveys from a database with 10,000 entries.

**Sampling Bias**

Biased by a non-representative sample.

**Example:** Online survey only among young users.

**Sampling Rate**

Frequency at which data is collected.

**Example:** A sensor that measures every 10 seconds has a sampling rate of 0{,}1 Hz.

**Sankey Diagram**

Visualization of flows and branches, e.g. in energy flows or user paths.

**Example:** Representation of conversion streams in an online shop.

**Scalability**

Ability of a system to remain efficient as the volume of data grows.

**Example:** Cloud databases scale horizontally with large amounts of data.

**Scaling**

Transformation of features to a comparable range of values.

**Example:** Min-Max or Z-Score normalization.

**Scenario Analysis**

Modelling of alternative future scenarios for risk assessment.

**Example:** Revenue at +10% or -10% demand.

**Scatterplot**

Diagram showing relationships between two numeric variables.

**Example:** Age vs. income.

**Schema**

Structure and definition of tables, fields, and relationships in a database.

**Example:** Database schema with "Customers", "Orders", "Products".

**Scientific Method**

Structured process for gaining knowledge through hypothesis formation and testing.

**Example:** Test hypothesis: "More light increases productivity".

**Scikit-learn**

Popular ML library in Python for classification, regression, clustering.

**Example:** `RandomForestClassifier()` from scikit-learn.

**Score**

Numerical evaluation of a prediction or model.

**Example:** Credit risk score = 0{,}82

**Scoring Function**

Function for calculating a score for classification or ranking.

**Example:** Log loss or ROC-AUC as a scoring function.

**Script**

Program file for automated execution of commands.

**Example:** Python script for data cleansing.

**Seasonality**

Regular, recurring patterns in time series.

**Example:** Decline in sales in January.

**Second Normal Form (2NF)**

Database normal form that avoids subkey dependencies.

**Example:** Separation of item and order information.

**Segmentation**

Splitting data into groups with similar characteristics.

**Example:** Cluster analysis for customer segmentation.

**Selectivity**

Percentage of rows in a table that are hit by a query.

**Example:** SELECT * WHERE status = 'active' with 5% selectivity.

**Self-Join**

Link a table to itself.

**Example:** Hierarchical structure in an employee list.

**Semi-Structured Data**

Data with partially defined structure.

**Example:** JSON or XML documents.

**Sensitivity (Recall)**

Measure for the correct detection of positive cases.

**Example:** 90% sensitivity = 90% of patients detected.

**Sentiment Analysis**

Analysis of opinions and feelings in text data.

**Example:** Rating tweets as positive, neutral or negative.

**Sequence Model**

ML models that specialize in sequential data.

**Example:** LSTM for text or audio sequences.

**Shannon Entropy**

Measure of the indeterminacy or amount of information of a distribution.

**Example:** Maximum entropy with equal distribution.

**Shapiro-Wilk-Test**

Statistical test for normal distribution.

**Example:** p-value < 0{,}05 indicates deviation from the normal distribution.

**Sharding**

Distribute data across multiple servers for load balancing.

**Example:** Horizontal sharding in distributed NoSQL systems.

**Sharpe Ratio**

Ratio of excess return to volatility of an investment.

**Example:** Sharpe ratio = (return - risk-free interest rate) / standard deviation.

**Shotgun Stochastic Search**

Search methods for model selection in high-dimensional data spaces.

**Example:** Selection of relevant genes in bioinformatics.

**Shrinkage**

Regularization principle to reduce model complexity.

**Example:** Ridge regression.

**Signal-to-Noise Ratio**

Ratio of useful information to noise.

**Example:** 10:1 is a high SNR.

**Silhouette Score**

Measure of the quality of a clustering.

**Example:** Score close to 1 = clear cluster separation.

**Simulation**

Replicate processes to analyze scenarios.

**Example:** Monte Carlo simulation for risk assessment.

**Singular Value Decomposition (SVD)**

Matrix factorization for dimension reduction and latent space learning.

**Example:** Recommendation systems.

**Skewness**

Measure of the asymmetry of a distribution.

**Example:** Positive skewness in income.

**Slack Space**

Unused space in blocks on disks.

**Example:** Relevant in forensic data analysis.

**Sliding Window**

Technique for analyzing sequential data by moving a fixed window.

**Example:** Moving Average with window size 5.

**Softmax**

Activation function that generates probability distribution across classes.

**Example:** Softmax output on classification.

**Spearman Correlation**

Nonparametric measure of rank correlation.

**Example:** Evaluation of correlation for non-normally distributed data.

**SQL**

Language for querying and manipulating relational databases.

**Example:** `SELECT name FROM kunden WHERE ort = 'Berlin'`

**Stacked Generalization (Stacking)**

Ensemble method for combining several models.

**Example:** Meta-Classifier aggregates predictions from different models.

**Standard deviation**

Measure of the dispersion of data around the mean.

**Example:** Std = 5 means that data deviates by an average of 5 units.

**Standard error**

Estimate for the dispersion of a statistic (e.g., mean).

**Example:** StdError = 0{,}2 for mean of a sample.

**Standardization**

Scaling variables to mean 0 and hrs 1.

**Example:** Z-transform for regression analysis.

**Stationarity**

Property of time series whose statistical properties do not change.

**Example:** Differentiation to achieve stationarity.

**Statistical significance**

Probability that an observed effect is not a coincidence.

**Example:** $p < 0{,}05$ means statistically significant.

**Continuous Variable (Continuous)**

Variable with an infinite number of characteristics in the value range.

**Example:** Height in cm.

**Sample**

Selection of data points to analyze a population.

**Example:** 1,000 customer surveys.

**Dispersion**

Measure of the distribution of values around the center.

**Example:** High dispersion with widely varying incomes.

**Stochastics**

A branch of mathematics that deals with chance and probabilities.

**Example:** Application in forecasting models.

**Stratified Sampling**

Dividing the population into layers for targeted sampling.

**Example:** Age groups in election polls.

**Streaming Data**

Continuously incoming data in real time.

**Example:** sensor data or weblogs.

**Structured data**

Data with a clearly defined structure in tabular form.

**Example:** Excel spreadsheet with customer data.

**Subsampling**

Partial selection from large data set.

**Example:** 10% of the log data for initial analysis.

**Supervised Learning**

ML approach with labeled training data.

**Example:** Classification of emails as spam or non-spam.

**Support Vector Machine (SVM)**

ML algorithm for classification by separation with maximum distance.

**Example:** Handwriting recognition in images.

**Synthetic data**

Artificially generated data for modeling or training.

**Example:** Customer data generated to test a dashboard.

**Syntax**

Grammar rules of a programming language.

**Example:** Python: `if x > 0:`

**Systematic bias**

Biased by methodological errors or bias.

**Example:** Skewed sample in non-randomized selection.

**Table Structure**
Defines the organization of columns and data types in a database

table.
**Example:** A customer table with name (text), age (integer), and city (text).

**Table Linking**

Joining two or more tables via shared keys.

**Example:** Customer ID joins "Customers" and "Orders" tables.

**Target Variable**

The target variable to be predicted or classified in an ML model.

**Example:** "Purchase Decision" in a conversion prediction model.

**Tidy Data**

Structured data where every variable is a column, every observation is a row.

**Example:** Long format time series data.

**t-SNE (t-distributed Stochastic Neighbor Embedding)**

Nonlinear dimension reduction for the visualization of high-dimensional data.

**Example:** Representation of word vectors in 2D.

**Target Encoding**

Encoding categorical variables based on the mean of the target variable.

**Example:** Average conversion rate per advertising channel.

**TensorFlow**

Google's open-source library for machine learning and deep learning.

**Example:** `tf.keras.Sequential()` for model creation.

**Test**

Dataset to check the model quality, separate from the training dataset.

**Example:** 20% of the data for final validation.

**Test statistics**

Value calculated from a sample to test a hypothesis.

**Example:** t-value in the t-test.

**Text Mining**

Extraction of structured information from unstructured texts.

**Example:** Topic extraction from customer reviews.

**Text Classification**

Assignment of texts to predefined categories.

**Example:** Email as spam or non-spam.

**TF-IDF (Term Frequency-Inverse Document Frequency)**

Weighting measure for words in texts to highlight important terms.

**Example:** Common word in one document, rare in others.

**Threshold**

Threshold for decision on classifications.

**Example:** Probability > 0{,}5 = class 1.

**Time Series Analysis**

Analysis of time-ordered data for forecasting or pattern recognition.

**Example:** Sales development per month.

**Time to Event**

Time until a specific event occurs.

**Example:** Time until first purchase after newsletter registration.

**Tokenization**

Splitting text into smaller units such as words or sentences.

**Example:** "Data Science is cool" → ["Data", "Science", "is", "cool"].

**Top-K Accuracy**

Metric where the right label must be among the top K predictions.

**Example:** Top 3 predictions contain the correct class.

**Tracking Code**

Script or tag to capture user actions.

**Example:** Google Analytics Tracking Pixel.

**Workout**

Data on which an ML model learns.

**Example:** Historical sales figures for model training.

**Transformation function**

Ability to transform data for better model performance.

**Example:** Log transformation for data that is heavily skewed to the right.

**Transpose**

Swapping rows and columns in a matrix or table.

**Example:** `df.T` in pandas.

**True Negative**

Case where a negative example is correctly classified as negative.

**Example:** Healthy patient is correctly recognized as healthy.

**True Positive**

Case in which a positive example is correctly recognized.

**Example:** Sick patient correctly recognized as sick.

**T-Test**

Statistical test for comparing mean values of two groups.

**Example:** Comparison of the average expenditure of men and women.

**Type I Error**

False-positive error: Rejection of the null hypothesis even though it is true.

**Example:** Healthier is diagnosed as sick.

**Type II Error (Beta Error)**

False-negative error: Null hypothesis is retained even though it is false.

**Example:** Sick person is classified as healthy.

**Type Casting**

Conversion of data types within a program or analysis.

**Example:** `int("42")` in Python yields an integer.

**UAT (User Acceptance Testing)**
Final stage of software testing, in which real users check whether the system meets their requirements.
**Example:** A BI tool is tested by end users before it goes live.

**UDAF (User-Defined Aggregate Function)**

Custom function to aggregate multiple values in SQL-like languages.

**Example:** Own median function in Apache Hive.

**UDF (User-Defined Function)**

Custom function used in SQL, Python, or Spark environments.

**Example:** Own calculation logic with `@udf` in PySpark.

**UI (User Interface)**

Interface between humans and systems for operating software.

**Example:** Dashboard interface with filter elements.

**ULID (Universally Unique Lexicographically Sortable Identifier)**

Alternative to UUIDs that is sortable.

**Example:** ULID = 01F8MECHZX3TBDSZ7XRADM79XV

**UMAP (Uniform Manifold Approximation and Projection)**

Algorithm for dimension reduction and data visualization.

**Example:** Visualization of customer segments in 2D space.

**Unbalanced Data**

Datasets in which class distributions are highly unequal.

**Example:** 95% non-spam, 5% spam.

**Uncertainty**

Uncertainty about the true value or model.

**Example:** Forecast: Revenue = 10 million ± 0.5 million

**Underfitting**

The model is too simple and does not adequately reflect the data structure.

**Example:** Linear model in a nonlinear relationship.

**Undersampling**

Technique for reducing the majority class in unbalanced data.

**Example:** Reduction of non-spam mails.

**Univariate Analysis**

Analysis of a single variable.

**Example:** Histogram of the age distribution.

**Unit Test**

Automated testing of individual functions or modules.

**Example:** `test_mean_function()` in Python.

**Unnormalized Data**

Data without scaling or normalizing.

**Example:** Income in euros, age in years, weight in kg.

**Unstructured Data**

Data without a fixed structure, often text, images or audio.

**Example:** e-mails, PDFs, chat histories.

**Unsupervised Learning**

ML method without labeled training data.

**Example:** Clustering algorithm (e.g. K-Means).

**Update Anomaly**

Data inconsistency due to redundant storage without normalization.

**Example:** Changing an address must be done in several tables.

**Upsampling**

Artificially enlarging the minority class.

**Example:** Copying spam emails for class balance.

**Upper Bound**

Upper bound of a confidence interval or parameter.

**Example:** 95% Trust Range: [10, 15] → Upper Bound = 15

**URI (Uniform Resource Identifier)**

Unique address for identifying resources on the web.

**Example:** `https://api.server.com/data/123`

**URL Encoding**

Encoding special characters in URLs.

**Example:** Space becomes `%20`

**UUID (Universally Unique Identifier)**

128-bit value for unique identification.

**Example:** `550e8400-e29b-41d4-a716-446655440000`

**UX (User Experience)**

A user's overall experience with a system.

**Example:** loading times, navigation and visual design of a dashboard.

**Utility Function**

Function for evaluating decisions or results.

**Example:** Choosing between models based on cost/benefit.

**Validation Set**
Dataset used to evaluate a model during the training phase.
**Example:** Splitting a dataset into 70% training, 15% validation, 15% testing.

**Value at Risk (VaR)**

Statistical measure used to quantify the potential loss in a given period of time at a given confidence level.

**Example:** A daily VaR of 5% at EUR 1 million = EUR 50,000 loss with 95% probability.

**Variance**

Measure of the dispersion of data around the mean.

**Example:** High variance means large differences between the values.

**Variance Inflation Factor (VIF)**

Measure of multicollinearity in regression models.

**For example,** VIF > 10 indicates strong correlation with other variables.

**Variable**

A measurable property or feature that is being analyzed.

**Example:** age, income or click-through rate.

**Variable Importance**

Measure of the relevance of a variable in a model.

**Beispiel:** Feature-Importances bei Random Forests.

**Variance Threshold**

Feature selection method that removes features with low variance.

**Example:** Filters columns whose values are almost always the same.

**Vector**

Mathematical structure for representing data points in n-dimensional space.

**Example:** [1.2, 3.4, 0.5] as input for a neural network.

**Vectorization**

Converting data into numerical vectors, often for machine learning.

**Example:** Text to TF-IDF vectors.

**Version Control**

System for managing changes to code or data.

**Example:** Git with commit history.

**Vertical Scaling**

Increase the resources of a single server.

**For example,** more RAM or CPU for a database instance.

**Visualization**

Presenting data in visual form to recognize patterns.

**Example:** Bar chart, heat map, or box plot.

**VLOOKUP**

Excel function for searching vertically in tables.

**For example,** search for a product name by ID.

**Volatility**

Measure of the fluctuation range of time series or markets.

**Example:** Stocks with high volatility have strongly fluctuating prices.

**Voting Classifier**

Ensemble learning procedure in which several models coordinate.

**Example:** Majority decides on classification.

**VAE (Variational Autoencoder)**

Neural network for dimension reduction and generation of data.

**Example:** Image compression or synthetic data generation.

**Variance Explained**

Proportion of total variance explained by a model or component.

**Example:** 80% declared variance in PCA.

**Virtual Join**

Join two tables in the query without physically merging.

**Example:** View in SQL.

**Volumetric data**

3D data, often used in medicine or geosciences.

**Example:** CT scans or seismic data cubes.

**Voice Recognition**

Technology to convert spoken speech to text.

**Example:** Google Assistant recognizes commands by voice.

**Variance-Bias Tradeoff**

Basic concept in machine learning: balance between under- and over-adaptation.

**Example:** Complex models risk overfitting, simple underfitting.

**Vector Database**

Specialized database for semantic searches with embeddings.

**Example:** Using FAISS or Pinecone to search for similarities.

**Video Analytics**

Automated analysis of video data.

**Example:** Detection of objects or movements in surveillance videos.

**Violin Plot**

Visualization that combines boxplot with density distribution.

**Example:** Presentation of grade distributions by subject.

**Virtual Machine**

Virtual system with its own operating system instance.

**Example:** Ubuntu VM on Windows for data analysis.

**View (SQL)**

Virtual table, based on saved queries.

**Example:** `CREATE VIEW aktive_kunden AS SELECT * FROM kunden WHERE status='aktiv'`

**Von Neumann Architecture**

Compute architecture with shared memory for data and programs.

**Example:** Foundation of modern computer design.

**Vector Space Model**

Model for representing text or documents as vectors.

**Example:** TF-IDF vectors in NLP.

**Visual Regression Testing**

Test method for detecting UI changes through image comparisons.

**Example:** Difference comparison of screenshots for web changes.

**Vulnerability Assessment**

Assessment of vulnerabilities in IT systems.

**Example:** Scans for insecure ports or outdated software.

**WAAS (Workspace as a Service)**
Cloud service that provides a complete virtual work environment.
**Example:** Remote teams use WAAS for secure access to data and software.

**Probability**

Measure of the expectation that a certain event will occur.

**Example:** The probability that "heads" will appear in a coin toss is 0{,}5.

**Wald-Test (Wald Statistic)**

Statistical test for the significance test of regression coefficients.

**Example:** Use in Logit models to test individual influencing variables.

**Warehouse (Data Warehouse)**

Central database for analysis and reporting on large amounts of data.

**Example:** Storing historical sales figures to analyze trends.

**Waterfall Chart**

Graph showing cumulative changes.

**Example:** Profit development of a company over several quarters.

**Wavelet Transformation**

Technology for analyzing signals or time series in different frequency ranges.

**Example:** Compression of audio or image data.

**Web Scraping**

Automated extraction of data from websites.

**Example:** Price collection of products from an online shop.

**Weight Initialization**

Initial value assignment for neural networks, affects training progression.

**Example:** He initialization for ReLU activations.

**Weighted Average**

Average, in which different values are weighted differently.

**Example:** Average grade taking credit points into account.

**White Noise**

Random noise with constant spectral power density.

**Example:** Residual modeling in time series analysis.

**Whitening**

Pre-processing, which removes correlations between features.

**Example:** PCA whitening before clustering.

**Whisker Plot (Boxplot)**

Graphical representation of medians, quartiles, and outliers.

**Example:** Comparison of the income distribution of different regions.

**Wide Format**

Data structure with one column per variable and one row per observation unit.

**Example:** Pivoted table with monthly sales as columns.

**Wilcoxon Test**

Nonparametric test for paired samples.

**Example:** Before-and-after comparison of training data.

**Window Function**

SQL function for calculation over data rows with reference to the current row.

**Example:** Running average in a time series with `OVER(PARTITION BY...)`.

**Winsorizing**

Technique for treating outliers by limiting extreme values.

**Example:** Set all values above the 95th percentile to exactly that percentile.

**Key Performance Indicator**

Quantitative metric used to evaluate economic performance.

**Example:** Sales growth, EBITDA, return on investment.

**Word Embedding**

Representation of words as vectors in continuous space.

**Example:** Word2Vec, GloVe.

**Word Cloud**

Visualization in which words are displayed in different sizes depending on the frequency.

**Example:** Analysis of dominant terms in customer reviews.

**Working Directory**

Current directory where a script works or stores files.

**Example:** Read `path in Python via` os.getcwd().

**Workload**

Amount of tasks or data that a system or user needs to process in one time.

**Example:** High CPU usage during parallel data import.

**Wrapper Method**

Feature selection through repeated modeling with different sets of variables.

**Example:** Recursive elimination of unimportant features on regression.

**Wurzel-MSE (Root Mean Squared Error)**

Regression Model Error Metric - Square root of the mean square error.

**Example:** RMSE of 5{,}2 for a predictive model of sales.

**WYSIWYG (What You See Is What You Get)**

Surface concept in which the result of the display corresponds directly to the view.

**Example:** Dashboard editors with live preview.

**Weekly seasonality**

Regular pattern in a weekly rhythm within a time series.

**Example:** Higher traffic to websites on Monday and Friday.

**X-Axis**
Horizontal axis in a coordinate system or diagram.
**Example:** In a line chart, the X-axis usually represents time.

**X-bar Chart**

Quality control diagram for monitoring averages in processes.

**Example:** Daily average control of product dimensions in a production facility.

**XGBoost**

High-performance machine learning boosting framework.

**Example:** Used to participate in Kaggle competitions.

**XML (eXtensible Markup Language)**

Text-based data format for the presentation of hierarchically structured data.

**Example:** Product data as an XML file with nested tags.

**XPath**

Query language for navigating XML documents.

**Example:** `Product/Price` extracts the price from each product tag.

**XOR (exclusive OR)**

Logical operation with true if exactly one operand is true.

**Example:** XOR(1, 0) = 1, XOR(1, 1) = 0.

**XSS (Cross-Site Scripting)**

Vulnerability in which attackers inject scripts into web applications.

**Example:** A manipulated input field executes JavaScript.

**X-Intercept**

The point at which a line intersects the X-axis.

**Example:** If f(x) = 2x − 4, the X intercept is x = 2.

**X-value**

Independent variable or feature in an analysis.

**For example,** in a salary prediction model, "work experience" is an X value.

**XOML**

XML-based format for workflows in Microsoft technologies.

**Example:** Workflows in old . NET automation projects.

**X Cross Reference**

Reference to other content or data within a database or report.

**Example:** KPI report links to related raw data.

**X-Test**

Subset of data that is reserved for testing a model.

**Example:** X_train and X_test to separate training and test data.

**X-Space**

Feature space for input data in a model.

**Example:** All features together form the X-Space.

**XOR-Gate**

Electronic gate that realizes an XOR function.

**Example:** Used in digital circuits.

**XAI (Explainable AI)**

Techniques for explaining decisions of ML models.

**Example:** SHAP values for decision trees.

**Xref (Cross-reference)**

Reference system for linking data points or documents.

**Example:** Spreadsheet with cell references to other sheets.

**X-Modeling**

Structuring processes with parallel and sequential flows.

**Example:** X-shaped process branching in BPMN.

**XPL (eXtensible Processing Language)**

Programming language for data-driven workflows.

**Example:** Use for data conversion and processing in XML.

**X-Means Clustering**

Extension of K-Means with automatic determination of the number of clusters.

**Example:** Identification of optimal cluster number in customer classification.

**X-Y diagram**

Two-dimensional visualization of numerical relationships.

**Example:** Scatter plot with weight (X) and blood pressure (Y).

**X-R Chart**

Diagram for monitoring the mean value and range in quality assurance.

**Example:** Daily control of process deviations.

**X.509 Certificate**

Standard for the structure of digital certificates.

**Example:** TLS/SSL certificates for website encryption are based on X.509.

**X-Pipeline**

Multi-step data transformation or model pipeline.

**Example:** Data cleaning → feature engineering → modeling.

**X-Feature**

Single input characteristic in the ML context.

**Example:** Age is an X-feature in a prediction.

**X-Variables**

Collective term for independent variables in statistical models.

**Example:** In linear regression y = ax + b, x is the X variable.

**X-Strategy**

Abstract term for exploratory approach in data analysis.

**Example:** Unstructured data analysis with open hypotheses.

**X-Header**

Additional HTTP headers for transmitting information on the web.

**Example:** `X-Requested-With: XMLHttpRequest` to identify AJAX calls.

**X-Form**

Structured data mask for entering or displaying data.

**Example:** Form with drop-downs, checkboxes and input fields.

**X.25**

Formerly standard for packet-switched networks.

**Example:** Used in banks and credit card networks.

**Y-axis**
The vertical axis in a diagram or coordinate system. It usually

represents dependent variables such as measured values or results.
**Example:** In a line chart, the Y-axis shows the sales per month.

**YAML (YAML Ain't Markup Language)**

A human-readable data format for configuration and data exchange, often used in DevOps and ML projects. YAML is structured more simply than JSON, but supports complex data hierarchies.

**Example:** Definition of training parameters for an ML model.

**Yarn (Hadoop YARN)**

"Yet Another Resource Negotiator" – a framework for resource management in the Hadoop ecosystem. It allows you to run distributed computing applications.

**For example,** YARN manages resources for MapReduce jobs in a Hadoop cluster.

**Yeo-Johnson Transformation**

A transformation to normalize data, similar to the Box-Cox transformation, but also suitable for negative values. Improves linearization and variable modelability.

**Example:** Application to heavily skewed data such as net assets.

**Yield Curve**

Graphical representation of interest rates across different maturities. An inverted yield curve can indicate economic recessions.

**Example:** Analysis of the yield curve for the economic forecast.

**YOLO (You Only Look Once)**

A real-time object recognition algorithm in the field of deep learning. YOLO detects objects in images or videos at high speed.

**Example:** Use in surveillance systems to detect people.

**YTD (Year-to-Date)**

Key figure that describes the period from the beginning of the year to the current date. Commonly used in financial analysis to evaluate performance.

**Example:** YTD revenue = total revenue from January 1 to today.

**Yule Simon Distribution**

A probability distribution useful in analyzing "long tail" phenomena such as the frequency of rare events.

**Example:** Modelling the word distribution in a text corpus.

**Yule's Q coefficient**

A measure of the association of two binary variables. Q = (ad - bc)/(ad + bc), based on a 2×2 contingency table.

**Example:** Examining the relationship between two yes/no answers.

**Y function**

Generic term for a mathematical function that depends on an independent X variable. In statistics, this is usually the characteristic to be explained.

**Example:** Y = 3X + 2 is a linear function.

**Yield**

In general, the return or result of an operation or investment. In programming, also a keyword in Python for generating generators.

**Example:** Python function with `yield` creates lazy sequences.

**Y-axis labeling**

The unit of description or label on the Y-axis of a chart. It conveys the meaning of the displayed values.

**Example:** "Revenue in EUR" as an axis label.

**Y-Splitter**

In the data pipeline, a mechanism for branching data streams based on conditions.

**Example:** Routing data with "Status = Error" to a separate pipeline.

**Yellowbrick**

Python toolkit for visualizing and analyzing ML models. Complements scikit-learn with visual diagnostic tools.

**Example:** Visualization of model quality using ROC curve with Yellowbrick.

**Yield Spread**

Difference between the yields of two bonds. A measure of risk and investor confidence.

**Example:** Higher spread = higher perceived risk.

**Y-Coding (One-Hot Encoding Target Variable)**

Method for coding multi-class target variables for ML models.

**Example:** Target variable "color" with classes "red", "blue", "green" becomes [1,0,0], [0,1,0], [0,0,1].

**Youden-Index**

Measure for optimizing the threshold in binary classification models.

**Example:** Maximize Sensitivity + Specificity - 1 to select the best cutoff.

**Y-Randomization**

Validation technique for verifying overfitting in ML models. Target variable is randomly permuted and model is retrained.

**Example:** Significantly worse performance after Y-randomization speaks against overfitting.

**Yield Forecasting**

Predict yields or production output, e.g. in agriculture or manufacturing.

**Example:** ML model for forecasting the corn harvest based on weather data.

**Yield Management**

Dynamic pricing to optimize occupancy and revenue, e.g. in air travel or hospitality.

**Example:** Price adjustment depending on booking time and demand.

**YTD-Analysis**

Analysis of cumulative developments since the beginning of the year. Useful for evaluating seasonal trends.

**Example:** Comparing the YTD performance of different business units.

**Z-Score**
Standardized metric that indicates how many standard deviations a value is away from the mean.
**Example:** Z = (Value - Mean) / Standard Deviation.

**Z-Test**

Statistical test to test hypotheses in the presence of known population standard deviation.

**Example:** Test for mean difference with known σ.

**Zero-Inflated Model**

Model type for data with a disproportionate number of zeros, e.g. in counting variables.

**Example:** Accident statistics with many zero reports.

**Zero-Shot Learning**

ML approach, where a model can solve tasks without having seen examples for them.

**Example:** Text classification with purely descriptive class labels.

**Zero Trust Architecture**

Security concept in IT where no user or device is automatically trusted.

**Example:** Access controls on every API request.

**Zero-Based Budgeting**

Planning method in which all expenses have to be justified from scratch.

**Example:** Each cost item is revalidated annually.

**Central tendency**

Location parameters such as mean, median, or mode that describe the distribution.

**Example:** Median income in a region.

**Time Series**

Data that is ordered in time, often with equal spacing.

**Example:** Daily stock prices.

**Time series analysis**

Analysis methods for modelling time-dependent data.

**Example:** ARIMA model for forecasting sales figures.

**Time Delay (Lag)**

Delayed influence of a variable in a time series.

**Example:** Turnover today depends on the weather yesterday (Lag-1).

**Time Window Analysis (Rolling Window)**

Analysis within a moving time period.

**Example:** Moving average of the last 30 days.

**Target variable**

The variable to be predicted or explained in a model.

**Example:** Price of a house in a regression.

**Target Value (Label)**

Classification or regression score used to train the model.

**Example:** "Spam" or "Non-Spam" as a label.

**Random Error**

Non-systematic deviation from true value.

**Example:** Measurement error due to noise.

**Random Forest**

Ensemble ML method from decision trees for classification or regression.

**Example:** Classification of customer churn.

**Random variable**

Variable whose value depends on a random process.

**Example:** Number of points when rolling the dice.

**Random Number Generator**

Algorithm for generating seemingly random values.

**Example:** `random()` in Python.

**Random sampling**

Sample where each unit of the population has the same probability of selection.

**Example:** Randomly drawn participants for a survey.

**Underlying Distribution**

The assumed or observed distribution on which an analysis is based.

**Example:** Normal distribution for IQ values.

**Access Rights**

Rules about which users are allowed to read, write or change which data.

**Example:** Only admins are allowed to delete user data.

**Access Time**

Time span between request and receipt of data.

**For example,** access to SQL database takes 15 ms.

**Access log**

Log of database or system access.

**Example:** Log files for web servers.

**Reliability**

Measure of consistency or reproducibility of measurements or models.

**Example:** A test gives similar results when repeated.

**Known Labels**

Target variables in the training dataset known in supervised learning.

**Example:** Email data with a known spam classification.

**Z-Transformation**

Standardization of a variable by subtracting the mean and dividing by the standard deviation.

**Example:** Application before training a linear model.

**Zoomable Chart**

Interactive visualization with zoom function.

**Example:** Zoomable line chart with D3.js.

**Censored Data**

Observations in which only a partial value is known. Often in survival time analyses.

**Example:** Patient study ends before the time of death.

**Target group analysis**

Identification and description of user groups for targeted targeting.

**Example:** Analysis of website visitors by age and origin.

**Zoning**

Segmentation of a geographic or logical area for analysis or control.

**Example:** Dividing a city into clusters for traffic analysis.

**Reliability Estimation**

Statistical estimation of how consistent a procedure is when repeated.

**Example:** Cronbach's alpha in psychometrics.

**Two-Sample Test**

Statistical comparison of two groups with regard to mean value or distribution.

**Example:** T-test between control and treatment group.

**Two-dimensional normal distribution**

Distribution of two correlated metric variables.

**Example:** Height and weight.

**Cyclic component**

Long-term, recurring fluctuation in time series.

**Example:** Business cycles in economic data.

**Two-Stage Model**

Two-step modelling approach, often at endogeneity.

**Example:** 2SLS in econometric models.

**Central value (median)**

The mean value of an ordered distribution.

**Example:** For [3, 5, 7], 5 is the median.

**Zielkonflikt**

Conflicting requirements in an optimization context.

**Example:** Reduce costs vs. ensure quality.

**Random processes**

Models for the description of stochastic processes.

**Example:** Markov chain.

**Cell Reference (Excel)**

Reference to a cell or range of cells in a table.

**Example:** `=A1 + B2`.

**Access Path**

Path through which a database system reads data from tables.

**Example:** Index access vs. Full Table Scan.

**Degree of target achievement**

Metric to evaluate how close a measure comes to the set goal.

**Example:** 90% of KPIs are met.

**Number Format**

Representation of numerical values in computers or tables.

**For example** , decimal, percentage, or currency format.

**Timestamp**

Time marker for events or data points.

**Example:** 2025-05-28 13:32:00 as log entry.

**Access Control**

Mechanism for restricting access to data or systems.

**For example,** role-based access control (RBAC).