**Accuracy**
Accuracy bezeichnet den Anteil korrekt klassifizierter Vorhersagen im Verhältnis zur Gesamtanzahl. Sie ist ein Standardmaß für Klassifikationsmodelle, aber nicht immer sinnvoll bei unbalancierten Klassen.
**Beispiel:** 95 korrekte Vorhersagen von 100 ergeben 95 % Accuracy.

**A/B-Test**

Ein A/B-Test vergleicht zwei Varianten, um herauszufinden, welche besser abschneidet. Er erfordert eine saubere Trennung und zufällige Zuweisung.

**Beispiel:** Zwei Website-Versionen mit unterschiedlicher Buttonfarbe werden getestet.

**Active Learning**

Ein ML-Ansatz, bei dem ein Modell gezielt entscheidet, welche Daten es zur weiteren Verbesserung labeln lässt.

**Beispiel:** Das Modell wählt gezielt unsichere Bilder zur Annotation durch Menschen.

**AdaBoost**

Ein Boosting-Algorithmus, der schwache Klassifikatoren zu einem starken Modell kombiniert. Besonders effektiv bei kleinen Datensätzen.

**Beispiel:** Mehrere kleine Entscheidungsbäume werden zu einem Gesamtmodell verbunden.

**Aggregation**

Daten werden zusammengefasst, z. B. als Summe oder Durchschnitt. Wichtig in BI, SQL und Excel.

**Beispiel:** Durchschnittlicher Umsatz pro Monat.

**Alias**

Ein alternativer Name, meist in SQL oder Code, zur besseren Lesbarkeit.

**Beispiel:** `SELECT sales AS umsatz FROM table`.

**Algorithmus**

Eine definierte Abfolge von Anweisungen zur Lösung eines Problems. In der Datenanalyse meist Lernverfahren.

**Beispiel:** Der k-means-Algorithmus gruppiert Daten in Cluster.

**Alternative Hypothese**

Statistischer Begriff für die Annahme, dass ein Effekt oder Unterschied vorliegt.

**Beispiel:** Der neue Medikamentenwirkstoff wirkt besser als der alte.

**Anomalieerkennung (Anomaly Detection)**

Technik zur Erkennung ungewöhnlicher Muster. Nützlich in der Betrugserkennung, Loganalyse und Qualitätskontrolle.

**Beispiel:** Ein einzelner Nutzer tätigt 100 Käufe in einer Minute.

**API (Application Programming Interface)**

Eine Schnittstelle zur strukturierten Kommunikation zwischen Programmen. Unverzichtbar für Automatisierung und Datenaustausch.

**Beispiel:** Eine Wetter-API liefert JSON-Daten für ein Dashboard.

**ARIMA**

Ein Zeitreihenmodell, das autoregressive, integrierte und gleitende Mittelwerte kombiniert.

**Beispiel:** Monatsumsätze der letzten drei Jahre werden für das Folgejahr geschätzt.

**Array**

Eine Datenstruktur für gleichartige Elemente, effizient für numerische Berechnungen.

**Beispiel:** Ein NumPy-Array mit 1.000 Zahlen.

**Artificial Intelligence (AI)**

Überbegriff für Maschinen, die Aufgaben mit kognitiven Fähigkeiten lösen. Inklusive ML und Deep Learning.

**Beispiel:** Ein KI-System diagnostiziert Hautkrankheiten anhand von Bildern.

**AUC (Area Under the Curve)**

Kennzahl für Klassifikatoren. Misst die Fläche unter der ROC-Kurve – je näher an 1, desto besser.

**Beispiel:** Ein Modell mit AUC 0.95 trennt Klassen sehr gut.

**Autoencoder**

Neural Network zur Datenkomprimierung und -rekonstruktion. Hilfreich für Anomalieerkennung.

**Beispiel:** Eingabebild wird mit minimalem Informationsverlust rekonstruiert.

**Autokorrelation**

Beziehung eines Werts mit sich selbst über Zeitverschiebung. Wichtig in Zeitreihenanalyse.

**Beispiel:** Absatz im Dezember ist hoch wie im Vorjahr.

**Automatisierung**

Ersetzung manueller Prozesse durch Skripte oder Systeme. Spart Zeit und vermeidet Fehler.

**Beispiel:** Täglicher Import von Verkaufsdaten per Python-Skript.

**Average**

Der Durchschnittswert – summe aller Werte geteilt durch Anzahl.

**Beispiel:** Durchschnitt aus 5, 7, 8 ist 6.67.

**Azure**

Microsofts Cloudplattform mit Tools für Datenanalyse, ML, Datenbanken. Konkurrenz zu AWS und GCP.

**Beispiel:** ETL-Prozess läuft in Azure Data Factory.

**Accuracy Paradox**

Phänomen, dass hohe Accuracy trotzdem ein schlechtes Modell bedeuten kann.

**Beispiel:** 99 % korrekte Vorhersagen in einem Datensatz mit 99 % Negativen.

**Augmented Analytics**

Analyseform mit KI-Unterstützung für Insight-Generierung und Automatisierung.

**Beispiel:** Ein BI-Tool erklärt Anomalien automatisch.

**Attribute**

Merkmale eines Datensatzes, auch Features genannt.

**Beispiel:** Einkommen, Alter, Region eines Kunden.

**AutoML**

Automatisierte Auswahl, Training und Tuning von ML-Modellen. Für schnelle Prototypen.

**Beispiel:** Google AutoML erstellt ein Bildklassifikationsmodell ohne Code.

**Atomic Operation**

Nicht unterbrechbare Aktion, z. B. in Datenbanktransaktionen. Garantiert Konsistenz.

**Beispiel:** INSERT in eine Tabelle mit Rollback-Option.

**Authentication**

Prozess der Identitätsprüfung. Häufig bei Datenzugriff und APIs relevant.

**Beispiel:** Zugang zu einem SQL-Server nur mit Passwort.

**Autonomous System**

Ein vollständig selbstlaufendes Daten- oder Softwaresystem.

**Beispiel:** Ein Auto erstellt seine Routen-ETAs basierend auf Live-Verkehrsdaten.

**Auto Regression (AR)**

Zeitreihenmodell, das aktuelle Werte aus vorherigen vorhersagt. Bestandteil von ARIMA.

**Beispiel:** Heute = 0.5×Gestern + 0.3×Vorgestern

**ASCII**

Zeichencodierung für Buchstaben, Zahlen und Symbole. Relevant bei Datenimporten und Codierungsproblemen.

**Beispiel:** Zeichen „A“ = ASCII 65

**Aliasing**

Phänomen, bei dem zu niedrig beprobte Signale falsch interpretiert werden. Wichtig in Zeitreihenanalysen.

**Beispiel:** Ein wöchentlicher Messwert täuscht einen Trend vor, der bei täglicher Auflösung verschwindet.

**Application Layer**

Schicht in Systemarchitekturen, die direkt mit Nutzerinteraktion zu tun hat.

**Beispiel:** Eine Web-App zur Visualisierung von Analyseergebnissen.

**Analytische Funktion (Analytic Function)**

SQL-Funktionen, die über Fenster (Window Functions) aggregieren.

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

**Auto Scaling**

Automatische Skalierung von Rechenressourcen in der Cloud je nach Auslastung.

**Beispiel:** Ein ML-Modell erhält mehr RAM bei Lastspitzen.

**Assoziationsanalyse (Association Rule Learning)**

Technik zur Entdeckung von Regeln und Zusammenhängen in Transaktionsdaten.

**Beispiel:** Kunden, die Bier kaufen, kaufen auch Chips.

**Async (Asynchron)**

Nicht-blockierende Ausführung von Prozessen – wichtig bei paralleler Datenverarbeitung oder Webanfragen.

**Beispiel:** Ein Webserver verarbeitet mehrere API-Requests gleichzeitig.

**Backpropagation**
Rückführungsverfahren zur Fehlerkorrektur in neuronalen Netzen. Es berechnet, wie stark jeder Knoten zur Gesamtabweichung beigetragen hat.
**Beispiel:** In einem CNN wird der Fehler von der Ausgabeschicht bis zur Eingabe zurückgerechnet.

**Balanced Dataset**

Ein Datensatz mit gleichmäßiger Verteilung der Klassen. Wichtig für faire Modellbewertung.

**Beispiel:** Je 5.000 Beispiele für Spam und Nicht-Spam.

**Bar Chart**

Visualisierung kategorialer Daten mit Balken unterschiedlicher Höhe.

**Beispiel:** Anzahl der Verkäufe je Produktkategorie.

**Baseline Model**

Ein einfaches Referenzmodell, um die Leistung komplexerer Modelle zu beurteilen.

**Beispiel:** Immer die häufigste Klasse vorhersagen.

**Batch Normalization**

Technik zur Beschleunigung und Stabilisierung des Trainings neuronaler Netze.

**Beispiel:** Werte eines Layers werden auf Mittelwert 0 und Varianz 1 skaliert.

**Batch Size**

Anzahl der Datenpunkte, die gleichzeitig ins ML-Modell eingespeist werden. Beeinflusst Trainingstempo und Modellqualität.

**Beispiel:** Training mit 128 Beispielen pro Batch.

**Batch Processing**

Verarbeitung von Daten in Blöcken statt einzeln. Gängig in Data Warehousing und ETL.

**Beispiel:** Nachtverarbeitung aller Tagesverkäufe.

**Bayes'scher Klassifikator**

Ein probabilistisches Modell, das Wahrscheinlichkeiten per Bayes-Theorem berechnet.

**Beispiel:** Naive Bayes zur Spam-Klassifikation.

**Bayessche Statistik**

Statistikansatz, der Vorwissen einbezieht und Wahrscheinlichkeiten iterativ aktualisiert.

**Beispiel:** Wahrscheinlichkeit für Betrug steigt nach auffälligem Verhalten.

**Bias (Verzerrung)**

Systematischer Fehler, der zu falschen Modellergebnissen führt. Kann durch Daten oder Modellstruktur entstehen.

**Beispiel:** Unausgewogene Trainingsdaten benachteiligen eine Gruppe.

**Big Data**

Sehr große, schnell wachsende und vielfältige Datenmengen, die klassische Verarbeitung überfordern.

**Beispiel:** Milliarden Logdaten pro Tag in einem Onlineshop.

**Binary Classification**

Klassifikation mit genau zwei Zielklassen.

**Beispiel:** Betrug: Ja oder Nein.

**Binning**

Einteilung kontinuierlicher Variablen in Kategorien.

**Beispiel:** Altersgruppen wie 18–25, 26–35.

**BI (Business Intelligence)**

Gesamtheit von Technologien zur datengetriebenen Entscheidungsunterstützung.

**Beispiel:** Power BI oder Tableau zur Visualisierung von KPIs.

**Binäre Variable**

Variable mit genau zwei Ausprägungen.

**Beispiel:** Ja/Nein, 0/1, Wahr/Falsch.

**Blending**

Kombination mehrerer ML-Modelle, meist als Ensemble-Methode.

**Beispiel:** Kombination aus SVM und Entscheidungsbaum.

**Bloom Filter**

Wahrscheinlichkeitsbasierte Datenstruktur zur schnellen Mengenprüfung mit Speicherersparnis.

**Beispiel:** Prüfung, ob eine E-Mail schon verarbeitet wurde.

**Boolean Logic**

Logiksystem mit Wahrheitswerten TRUE und FALSE.

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

**Bootstrap**

Resampling-Technik zur Schätzung von Verteilungen aus Stichproben.

**Beispiel:** 1.000 Ziehungen mit Zurücklegen zur Konfidenzintervallschätzung.

**Boxplot**

Diagramm zur Darstellung von Verteilungen inklusive Ausreißern.

**Beispiel:** Verteilung der Einkommen in fünf Abteilungen.

**Buffering**

Temporäres Speichern von Daten zur Überbrückung oder Entlastung.

**Beispiel:** Daten aus einem Stream werden in RAM gepuffert.

**Bucket**

Einzelner Bereich in einer in Kategorien eingeteilten Skala.

**Beispiel:** Preisrange 0–10€, 10–50€, 50–100€.

**Business Analytics**

Analyseform mit Fokus auf Geschäftserkenntnisse, strategisch oder operativ.

**Beispiel:** Warum ist der Umsatz in Q2 gesunken?

**Business Metric**

Kennzahl zur Steuerung eines Unternehmens.

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

**Byte**

Speichereinheit bestehend aus 8 Bit. Häufigste Maßeinheit bei Datenmengen.

**Beispiel:** Eine Textdatei mit 1.000 Zeichen ≈ 1 KB.

**Bayesian Network**

Graphenbasiertes Modell zur Darstellung probabilistischer Abhängigkeiten.

**Beispiel:** Modell für Krankheitssymptome und Ursachen.

**Bias-Variance-Tradeoff**

Grundprinzip im ML: Modelle müssen zwischen Überanpassung (Varianz) und Unteranpassung (Bias) balancieren.

**Beispiel:** Lineare Regression = hoher Bias, niedrige Varianz.

**Binary Tree**

Baumstruktur mit maximal zwei Kindknoten pro Elternknoten.

**Beispiel:** Entscheidungsbaum zur Klassifikation.

**Benchmarking**

Vergleich von Algorithmen oder Systemen anhand definierter Metriken.

**Beispiel:** Welcher Klassifikator hat bei gleichem Datensatz die beste AUC?

**Business Rule**

Regel zur Steuerung eines Prozesses auf Basis von Daten.

**Beispiel:** Wenn Nutzer < 18 Jahre → kein Kauf erlaubt.

**Bias Mitigation**

Strategien zur Reduktion von Verzerrung in Modellen oder Daten.

**Beispiel:** Fairness Constraints beim Modelltraining.

**Boolean Masking**

Technik zur Selektion bestimmter Elemente mit True/False-Arrays.

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

**Broadcast Join**

SQL-Optimierung, bei der eine kleine Tabelle an alle Knoten verteilt wird.

**Beispiel:** Kleine Lookuptabelle für Länderinformationen in Spark ge-broadcastet.

**Caching**
Zwischenspeicherung häufig benötigter Daten zur Beschleunigung von Zugriffen. Nützlich in Webentwicklung, Datenbanken und ML-Pipelines.
**Beispiel:** Ein BI-Tool lädt zuvor berechnete Aggregationen aus dem Cache.

**Categorical Variable**

Variable mit diskreten Ausprägungen wie Farben oder Ländern. Meist durch One-Hot-Encoding aufbereitet.

**Beispiel:** Spalte „Farbe“ mit Werten: Rot, Blau, Grün.

**Centroid**

Mittelpunkt eines Clusters, genutzt in k-means-Algorithmen.

**Beispiel:** Der Schwerpunkt einer Gruppe von Kunden mit ähnlichem Kaufverhalten.

**Churn Rate**

Kundenabwanderungsrate über einen bestimmten Zeitraum. Wichtig für Subscription-Modelle.

**Beispiel:** 15 % monatlicher Churn in einem SaaS-Service.

**Classification**

ML-Aufgabe, bei der Daten in diskrete Klassen eingeteilt werden.

**Beispiel:** Kreditwürdig vs. nicht kreditwürdig.

**Clean Data**

Fehlerfreie, bereinigte Daten, die für Analyse oder Modellierung geeignet sind.

**Beispiel:** Keine Duplikate, korrekte Typen, keine Leerwerte.

**Clustering**

Unüberwachtes Lernen zur Gruppierung ähnlicher Datenpunkte.

**Beispiel:** Kundensegmentierung auf Basis von Kaufverhalten.

**Coefficient**

Gewicht in einem Modell, das die Stärke und Richtung eines Prädiktors angibt.

**Beispiel:** In einer Regression: Einkommen hat einen positiven Einfluss auf Konsum.

**Collinearity**

Hohe Korrelation zwischen zwei oder mehr unabhängigen Variablen. Erschwert Interpretation.

**Beispiel:** Größe und Gewicht korrelieren stark.

**Column Store**

Datenbanksystem, das Daten spaltenbasiert speichert – vorteilhaft für analytische Abfragen.

**Beispiel:** BigQuery oder Redshift.

**Confidence Interval**

Intervall, das mit definierter Wahrscheinlichkeit den wahren Parameter enthält.

**Beispiel:** 95 %-Intervall für Mittelwert: 4.1 bis 4.8.

**Confusion Matrix**

Darstellung von Klassifikationsergebnissen mit TP, FP, FN, TN.

**Beispiel:** Ein Modell hat 87 % Genauigkeit, aber viele False Positives.

**Correlation**

Maß für linearen Zusammenhang zwischen zwei Variablen. Werte von -1 bis +1.

**Beispiel:** Werbebudget und Umsatz mit Korrelation +0.84.

**CPU (Central Processing Unit)**

Zentrale Recheneinheit, führt logische Operationen und Modelltraining durch.

**Beispiel:** Pandas rechnet meist auf der CPU.

**CSV (Comma-Separated Values)**

Textdatei mit tabellarischen Daten, Felder durch Kommas getrennt.

**Beispiel:** Export einer SQL-Tabelle als `data.csv`.

**Cross-Validation**

Verfahren zur stabilen Modellbewertung durch wiederholtes Trainieren/Testen auf verschiedenen Datenaufteilungen.

**Beispiel:** K-fold CV mit k=5.

**Curse of Dimensionality**

Probleme bei hoher Dimensionalität – z. B. spärliche Daten, Overfitting.

**Beispiel:** 1.000 Features bei nur 100 Beobachtungen.

**Cut-off-Point**

Grenzwert für Klassifikationsentscheidungen.

**Beispiel:** Wahrscheinlichkeit > 0.6 → Kredit genehmigt.

**Custom Function**

Benutzerdefinierte Funktion in Python, SQL oder Excel.

**Beispiel:** `def berechne_rabatt(preis): return preis*0.85`

**Categorical Encoding**

Techniken zur Umwandlung kategorialer Variablen in numerische Formate.

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

**Control Chart**

Diagramm zur Überwachung von Prozessen in der Qualitätskontrolle.

**Beispiel:** Produktionslinie zeigt Ausreißer in Fehlerhäufigkeit.

**Confidence Score**

Ausgabewert eines Modells, der angibt, wie sicher es in seiner Vorhersage ist.

**Beispiel:** Bildklassifikation: 82 % Wahrscheinlichkeit für „Katze“.

**Composite Key**

Primärschlüssel, der aus mehreren Spalten besteht.

**Beispiel:** Kombination aus „Bestellnummer“ + „Artikel-ID“.

**Contextual Bandit**

ML-Modell zur Auswahl von Aktionen bei Unsicherheit, unter Berücksichtigung des Kontexts.

**Beispiel:** Werbeausspielung basierend auf Nutzerverhalten.

**Confidence Level**

Gibt an, mit welcher Sicherheit ein Konfidenzintervall den wahren Wert enthält.

**Beispiel:** 95 %-Vertrauensniveau → 5 % Irrtumswahrscheinlichkeit.

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

Anteil der Klicks auf eine Anzeige im Verhältnis zur Gesamtanzahl der Impressionen.

**Beispiel:** 100 Klicks bei 1.000 Views = 10 % CTR.

**Canonical Correlation Analysis (CCA)**

Statistisches Verfahren zur Untersuchung der Beziehungen zwischen zwei Variablensätzen.

**Beispiel:** Zusammenhang zwischen schulischen Leistungen und familiärem Hintergrund.

**Classification Report**

Standardausgabe zur Bewertung eines Klassifikationsmodells: enthält Precision, Recall, F1-Score pro Klasse.

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

**Confidence Bound**

Ober- oder Untergrenze eines Konfidenzintervalls.

**Beispiel:** Obergrenze = 7.9 bei 95 %-Intervall.

**Cloud Computing**

Bereitstellung von IT-Ressourcen über das Internet auf Abruf.

**Beispiel:** AWS, GCP oder Azure bieten Rechenleistung und Speicherplatz on demand.

**Cron Job**

Zeitgesteuerter Task unter Linux zur Automatisierung wiederkehrender Prozesse.

**Beispiel:** Tägliches Update eines Dashboards um 3:00 Uhr.

**Cost Function**

Funktion, die den Fehler eines Modells misst und minimiert werden soll.

**Beispiel:** MSE bei Regression misst Abweichung zwischen Vorhersage und Realität.

**Composite Index**

Datenbank-Index, der mehrere Spalten kombiniert, um Abfragen zu beschleunigen.

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

**Constraint**

Einschränkung in Datenbanken, um Konsistenzregeln zu erzwingen.

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

**Confidence Ellipse**

Grafische Darstellung des Konfidenzintervalls zweidimensionaler Daten.

**Beispiel:** Streudiagramm mit 95 %-Ellipsen für zwei Merkmale.

**Dashboard**
Visuelle Oberfläche zur Anzeige wichtiger Kennzahlen, oft interaktiv. Eingesetzt in BI, Monitoring und Management.
**Beispiel:** Power BI-Dashboard zeigt Umsatztrends und regionale Verteilungen.

**Data Analyst**

Rolle zur Analyse, Visualisierung und Aufbereitung von Daten. Verwendet Tools wie SQL, Excel, Python.

**Beispiel:** Analysiert die Entwicklung der Conversion Rate im Onlineshop.

**Data Cleaning**

Prozess zur Entfernung fehlerhafter, fehlender oder doppelter Daten. Voraussetzung für jede verlässliche Analyse.

**Beispiel:** Entfernen leerer Felder und falscher Datentypen aus einer CSV-Datei.

**Data Engineer**

Spezialist für Aufbau und Pflege von Dateninfrastrukturen wie Pipelines, Datenbanken, Cloud-Systemen.

**Beispiel:** Entwickelt eine ETL-Strecke zur automatisierten Datenintegration.

**Data Governance**

Regeln und Prozesse für Datenqualität, Sicherheit und Zugriffsrechte.

**Beispiel:** Wer darf personenbezogene Daten sehen, wer nicht?

**Data Lake**

Unstrukturierter Datenspeicher in Rohform, häufig auf Hadoop oder S3.

**Beispiel:** Speicherung aller Rohdaten aus Weblogs, APIs und externen Quellen.

**Data Mart**

Fokussierter Teil eines Data Warehouses für bestimmte Fachabteilungen.

**Beispiel:** Separater Bereich für Marketingdaten mit aggregierten KPIs.

**Data Mining**

Entdeckung von Mustern und Zusammenhängen in großen Datenmengen mittels statistischer Methoden.

**Beispiel:** Regel „Kunden, die X kaufen, kaufen auch Y“.

**Data Pipeline**

Automatisierter Datenfluss von Quelle zu Ziel mit Extraktion, Transformation, Speicherung.

**Beispiel:** Apache Airflow steuert tägliches Laden neuer Shopdaten ins Data Warehouse.

**Data Scientist**

Experte für Analyse, Modellierung und Prognose komplexer Daten mittels ML/AI.

**Beispiel:** Prognose von Retourenwahrscheinlichkeit mit Random Forest.

**Database (Datenbank)**

Strukturierte Ablage von Daten für effiziente Suche und Verarbeitung. Relationale (SQL) oder NoSQL-Varianten.

**Beispiel:** PostgreSQL speichert Kunden- und Transaktionsdaten.

**Datensatz (Record)**

Einzelne Zeile in einer Tabelle mit mehreren Attributen.

**Beispiel:** Kunde #123 mit Name, Geburtsdatum, Umsatz.

**Decision Tree**

ML-Modell mit if/else-Struktur zur Klassifikation oder Regression.

**Beispiel:** Baum entscheidet, ob Kredit gewährt wird.

**Deep Learning**

Teilgebiet des ML, das auf tiefen neuronalen Netzen basiert. Stark bei Bildern, Sprache, komplexen Mustern.

**Beispiel:** Spracherkennung auf Smartphones mit Deep Learning.

**Default Value**

Standardwert bei fehlender Eingabe.

**Beispiel:** Standardmäßig 0 bei leerem Feld.

**Denormalisierung**

Bewusste Redundanz zur Leistungssteigerung in Datenbanken.

**Beispiel:** Kundenname wird in jede Bestellzeile übernommen.

**Deployment**

Bereitstellung eines Modells oder Systems zur produktiven Nutzung.

**Beispiel:** ML-Modell wird via REST-API bereitgestellt.

**Derived Variable**

Abgeleitetes Merkmal, berechnet aus bestehenden Feldern.

**Beispiel:** Alter = Heute – Geburtsjahr.

**Descriptive Analytics**

Analyse historischer Daten zur Beschreibung von Entwicklungen.

**Beispiel:** Umsatzrückgang von 10 % im Vergleich zum Vorjahr.

**Dimensionality Reduction**

Verfahren zur Reduktion der Merkmalsanzahl bei hoher Dimensionalität.

**Beispiel:** PCA reduziert 500 Sensorwerte auf 10 Hauptkomponenten.

**DNN (Deep Neural Network)**

Mehrschichtiges neuronales Netz mit hohem Abstraktionsvermögen.

**Beispiel:** Klassifikation von Handschrift anhand von Pixelwerten.

**Docker**

Containertechnologie für portable, reproduzierbare Softwareumgebungen.

**Beispiel:** Ein Python-Analyse-Skript läuft unabhängig von Host-Systemen.

**Document Store**

NoSQL-Datenbank zur Speicherung von Dokumentenstrukturen wie JSON.

**Beispiel:** MongoDB speichert Userprofile als JSON-Objekte.

**Dropout**

Regulierungsmethode bei neuronalen Netzen zur Vermeidung von Overfitting.

**Beispiel:** 30 % der Neuronen werden pro Trainingsdurchlauf deaktiviert.

**Dummy Variable**

Künstlich erzeugte binäre Variable zur Kodierung kategorialer Merkmale.

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

**Data Wrangling**

Umfassender Begriff für das Bearbeiten und Umstrukturieren von Daten zur Analyse.

**Beispiel:** Spalten splitten, fehlende Werte ersetzen, Typen umwandeln.

**Drilldown**

Funktion in Dashboards zur Navigation von Aggregation in Detaildaten.

**Beispiel:** Klick auf „Region Bayern“ zeigt Städte an.

**Data Imputation**

Ersetzen fehlender Werte durch Schätzung oder Regel.

**Beispiel:** Durchschnittswert ersetzt fehlende Umsatzangaben.

**Decision Boundary**

Grenze, an der ein Klassifikationsmodell zwischen zwei Klassen unterscheidet.

**Beispiel:** Lineare Trennung zwischen Gut-/Schlecht-Kredit.

**Data Provenance**

Dokumentation der Herkunft und Transformation eines Datensatzes.

**Beispiel:** Ursprung, Änderungen und Zugriffe eines Datensatzes werden protokolliert.

**Data Drift**

Veränderung der Datenverteilung über Zeit, wodurch ML-Modelle an Genauigkeit verlieren können.

**Beispiel:** Kundentypen ändern sich durch Marktveränderung – Modell muss neu trainiert werden.

**EDA (Exploratory Data Analysis)**
Systematische Untersuchung von Daten vor der Modellierung. Dient zur Erkennung von Mustern, Ausreißern und Datenproblemen.
**Beispiel:** Boxplots, Korrelationen und Histogramme zur Bewertung eines Kundendatensatzes.

**Edge Case**

Ungewöhnlicher Eingabefall, der ein System an seine Grenzen bringt. Relevant für Tests und Fehlerresistenz.

**Beispiel:** Ein Kunde mit Alter 0 oder 120 Jahren.

**Elasticity (Elastizität)**

Maß für die Empfindlichkeit einer Zielgröße bei Änderung einer Einflussgröße.

**Beispiel:** -1,5 Preiselastizität = 1 % Preiserhöhung → 1,5 % weniger Nachfrage.

**Embedding**

Transformation von Objekten (z. B. Wörtern) in Vektoren mit fixer Länge.

**Beispiel:** Wort „Auto“ als Vektor [0.12, -0.44, ...] für neuronales Netz.

**Ensemble Learning**

Verfahren zur Kombination mehrerer Modelle zur Verbesserung der Genauigkeit.

**Beispiel:** Random Forest kombiniert viele Entscheidungsbäume.

**Entropy (Entropie)**

Maß für Unordnung in Daten, genutzt in Entscheidungsbäumen. Je höher, desto gemischter die Klassen.

**Beispiel:** Maximale Entropie bei 50 %/50 %-Verteilung zweier Klassen.

**Epoch**

Ein kompletter Durchlauf aller Trainingsdaten beim Modelltraining.

**Beispiel:** Das Modell wird über 100 Epochen trainiert.

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

Kernprozess zur Datenintegration: Extraktion aus Quelle, Transformation, Laden in Zielsystem.

**Beispiel:** Webshopdaten → Währungsumrechnung → Speichern in PostgreSQL.

**Evaluation Metric**

Kennzahl zur Bewertung von Modellen, z. B. Accuracy, RMSE, Precision.

**Beispiel:** Ein F1-Score von 0.81 bei Spam-Klassifikation.

**Excel**

Tabellenkalkulationstool mit hoher Relevanz in Reporting, Analyse und Visualisierung.

**Beispiel:** Pivot-Tabelle zur Analyse von Verkaufsdaten nach Region und Monat.

**Exponential Smoothing**

Zeitreihenmethode zur Glättung kurzfristiger Schwankungen.

**Beispiel:** Gewichtete Umsatzprognose mit stärkerem Fokus auf jüngste Werte.

**Extrapolation**

Vorhersage außerhalb des bekannten Datenbereichs – riskanter als Interpolation.

**Beispiel:** Prognose des Umsatzes für 2030 auf Basis von 2020–2024.

**Early Stopping**

Abbruch des Modelltrainings, sobald Validierungsfehler steigt. Verhindert Overfitting.

**Beispiel:** Modelltraining stoppt nach 34 Epochen.

**Entity**

Reales oder konzeptionelles Objekt, das in einer Datenbank erfasst wird.

**Beispiel:** Kunden, Produkte, Bestellungen.

**Event-based Data**

Zeitgestempelte Daten, die durch Aktionen erzeugt werden.

**Beispiel:** Klicks, Logins, Käufe in Websystemen.

**Explainability**

Verständlichkeit von Modellen und deren Entscheidungen für Menschen.

**Beispiel:** SHAP-Werte zeigen Einfluss einzelner Merkmale auf Modellentscheidungen.

**Exogenous Variable**

Einflussgröße von außen, nicht durch das Modell erklärt, aber berücksichtigt.

**Beispiel:** Wetterdaten bei Umsatzanalyse im Einzelhandel.

**Elastic Net**

Regularisierungstechnik, die Lasso (L1) und Ridge (L2) kombiniert.

**Beispiel:** Verwendung bei korrelierten Regressionsvariablen.

**Error Rate**

Anteil der falschen Vorhersagen. Ergänzung zur Accuracy.

**Beispiel:** 7 Fehlklassifikationen bei 100 Fällen → Error Rate: 7 %.

**Eins-gegen-Rest (One-vs-Rest)**

Klassifikationsstrategie für Multiklassen-Probleme.

**Beispiel:** 3 Modelle: Katze-vs-Rest, Hund-vs-Rest, Maus-vs-Rest.

**Euklidische Distanz**

Länge der kürzesten Verbindung zweier Punkte im Merkmalsraum.

**Beispiel:** Abstand zwischen zwei Kundenprofilen im 5D-Raum.

**Encoding**

Umwandlung von Variablen in eine numerische Repräsentation.

**Beispiel:** One-Hot-Encoding für Farbe: Blau = [0,1,0]

**Empirical Distribution**

Verteilung der beobachteten Datenwerte, ohne theoretisches Modell.

**Beispiel:** Histogramm der beobachteten Kundenalter.

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

Diagramm zur Strukturierung von Datenbanktabellen und ihren Beziehungen.

**Beispiel:** Beziehung: Kunde –> Bestellung (1:n).

**Execution Plan**

Beschreibung, wie eine Datenbankabfrage technisch ausgeführt wird.

**Beispiel:** PostgreSQL zeigt, wie ein JOIN ausgeführt wird (Index, Sortierung etc.).

**External Table**

Tabelle, die Daten außerhalb der Datenbank referenziert (z. B. in Data Lakes).

**Beispiel:** Hive-Tabelle, die auf Parquet-Dateien verweist.

**ETL Scheduler**

Werkzeug zur zeitlichen Steuerung von ETL-Prozessen.

**Beispiel:** Apache Airflow plant nächtliche ETL-Pipelines.

**Enrichment**

Ergänzung von Daten mit weiteren Attributen zur Verbesserung der Analyse.

**Beispiel:** Anreicherung von Transaktionsdaten mit Wetterdaten.

**Einschlusskriterium (Inclusion Criteria)**

Filterbedingung für Datenzugang oder Modelltraining.

**Beispiel:** Nur Kunden mit vollständigen Profildaten werden trainiert.

**Endpoint**

Adresse (z. B. URL), über die auf Daten oder Modelle per API zugegriffen wird.

**Beispiel:** `/predict/` nimmt Features entgegen und liefert Vorhersage.

**Embedding Layer**

Schicht in neuronalen Netzen, die diskrete Werte in kontinuierliche Vektoren abbildet.

**Beispiel:** User-ID → 16-dimensionale Darstellung für Empfehlungssystem.

**Error Analysis**

Gezielte Untersuchung von Modellfehlern zur Verbesserung der Performance.

**Beispiel:** Analyse, bei welchen Produkten ein Klassifikator regelmäßig versagt.

**Einschränkung (Constraint)**

Datenbankregel, die bestimmte Zustände erzwingt.

**Beispiel:** Spalte darf keine Nullwerte enthalten (`NOT NULL`).

**F1-Score**
Harmonisches Mittel von Precision und Recall. Gute Metrik bei unbalancierten Datensätzen.
**Beispiel:** F1 von 0.84 bedeutet solides Gleichgewicht zwischen Erkennung und Präzision.

**Factor Analysis**

Statistisches Verfahren zur Reduktion auf latente Variablen (Faktoren).

**Beispiel:** Mehrere Zufriedenheitsfragen ergeben einen „Service“-Faktor.

**Feature**

Eingabemerkmal eines Modells. Auch Attribut oder Prädiktor genannt.

**Beispiel:** Alter, Einkommen, Wohnort.

**Feature Engineering**

Erstellung und Transformation relevanter Merkmale für ML-Modelle.

**Beispiel:** Aus Datum das Quartal extrahieren.

**Feature Importance**

Kennzahl für den Einfluss eines Merkmals auf das Modell.

**Beispiel:** In einem Churn-Modell ist „letzter Login“ am wichtigsten.

**Feature Selection**

Auswahl der wichtigsten Merkmale zur Modellvereinfachung.

**Beispiel:** Elimination von redundanten oder irrelevanten Spalten.

**Federated Learning**

Dezentrale Modelltrainingsmethode ohne zentrale Datenspeicherung.

**Beispiel:** Mobilgeräte trainieren lokal ein gemeinsames Sprachmodell.

**Filter Function**

Funktion zur Datenselektion nach Bedingung.

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

**Float**

Datentyp für Fließkommazahlen.

**Beispiel:** 3.14159

**Forecasting**

Prognose zukünftiger Werte anhand historischer Daten.

**Beispiel:** Umsatzvorhersage mit Holt-Winters-Modell.

**Foreign Key**

Fremdschlüssel in relationalen Datenbanken. Verweist auf Primärschlüssel einer anderen Tabelle.

**Beispiel:** `customer_id` in Bestell-Tabelle verweist auf Kunden-Tabelle.

**Formel**

Rechenvorschrift zur automatisierten Berechnung.

**Beispiel:** Excel: =B2*C2

**Forward Selection**

Schrittweise Merkmalsauswahl für Regressionsmodelle.

**Beispiel:** Beginne mit leerem Modell und füge sukzessive Merkmale hinzu.

**Fourier Transformation**

Zerlegt Zeitreihen in Frequenzanteile.

**Beispiel:** Frequenzanalyse von Stromverbrauchsdaten.

**False Positive (FP)**

Fehlerhafte positive Vorhersage.

**Beispiel:** Spamfilter markiert legitime E-Mail als Spam.

**False Negative (FN)**

Fehlerhafte negative Vorhersage.

**Beispiel:** Ein Krebsfall bleibt unerkannt.

**FRAUD Detection**

System zur Erkennung von Betrugsmustern.

**Beispiel:** ML erkennt gefälschte Kreditkartentransaktionen.

**Frequency Table**

Tabelle mit Häufigkeiten von Ausprägungen.

**Beispiel:** 340 Nutzer aus Deutschland, 120 aus Österreich.

**Full Outer Join**

SQL-Verknüpfung, die alle Zeilen beider Tabellen zeigt, auch ohne Übereinstimmung.

**Beispiel:** Alle Kunden und alle Bestellungen – auch wenn keine Verbindung existiert.

**Function**

Wiederverwendbarer Codeblock mit Eingaben und Rückgabe.

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

**Fuzzy Matching**

Abgleich ähnlich geschriebener Texte mit Toleranz.

**Beispiel:** „Meier“ ≈ „Mayer“.

**Flat File**

Einfache Datei ohne relationale Struktur, meist CSV oder TXT.

**Beispiel:** Rohdatenexport aus einem alten CRM.

**Feature Map**

Zwischenausgabe von CNNs in Bildverarbeitung.

**Beispiel:** Aktivierungskarte nach Convolution-Operation.

**First Normal Form (1NF)**

Grundregel für relationale Datenbanken: keine Wiederholungsgruppen, atomare Werte.

**Beispiel:** Eine Spalte enthält nicht mehrere Telefonnummern.

**File System**

Strukturierte Ablage und Verwaltung von Dateien in Verzeichnissen.

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

**Fingerprinting**

Erkennung eines Benutzers/Objekts durch eindeutige Datenmuster.

**Beispiel:** Wiedererkennung von Geräten anhand Browserdaten.

**Finite State Machine**

Modell, das Zustände und Übergänge eines Systems beschreibt.

**Beispiel:** Klickfolge in einer App wird als Zustandsdiagramm modelliert.

**Fit (Modelltraining)**

Anpassung eines ML-Modells an Trainingsdaten.

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

**Feature Drift**

Änderung der Bedeutung oder Verteilung eines Merkmals über Zeit.

**Beispiel:** „Nutzeraktivität“ verliert Aussagekraft nach Produktupdate.

**Field (Feld)**

Einzelnes Datenattribut innerhalb eines Datensatzes.

**Beispiel:** „email“ in einer User-Tabelle.

**Flattening**

Umwandlung verschachtelter Datenstrukturen in flache Tabellenform.

**Beispiel:** JSON → DataFrame mit Spalten für jeden Schlüssel.

**Fact Table**

Kernbestandteil eines Data Warehouses, speichert messbare Ereignisse.

**Beispiel:** Verkaufstabelle mit Umsatz, Menge, Datum.

**Factless Fact Table**

Tabelle ohne numerische Kennzahlen, aber mit Beziehungen zur Analyse von Ereignissen.

**Beispiel:** Anwesenheitstabelle für Schüler – kein „Wert“, aber relational nutzbar.

**Gantt Chart**
Diagramm zur Visualisierung zeitlicher Abläufe von Projekten oder

Prozessen. Genutzt in Planung und Projektmanagement.
**Beispiel:** Darstellung von ETL-Jobs über eine Woche.

**Gaussian Distribution (Normalverteilung)**

Symmetrische, glockenförmige Verteilung vieler natürlicher Merkmale. Basis vieler statistischer Methoden.

**Beispiel:** Körpergröße in einer Bevölkerung.

**Gini Index**

Maß für Unreinheit einer Aufteilung in Entscheidungsbäumen. Je niedriger, desto homogener die Klassen.

**Beispiel:** Gini = 0 bei reinen Blättern.

**Git**

Versionskontrollsystem für Code. Erlaubt parallele Arbeit, Historie und Wiederherstellung.

**Beispiel:** `git commit -m "Datenbereinigung hinzugefügt"`

**GitHub**

Online-Plattform zur Verwaltung von Git-Repositories. Unterstützt Kollaboration, Reviews und Automatisierung.

**Beispiel:** Team teilt Notebooks über GitHub-Repo.

**Gradient Descent**

Optimierungsverfahren zur Minimierung von Fehlerfunktionen. Grundlage fast aller ML-Verfahren.

**Beispiel:** Training eines neuronalen Netzes.

**GPU (Graphics Processing Unit)**

Prozessor für parallele Berechnungen, besonders bei Deep Learning.

**Beispiel:** NVIDIA A100 beschleunigt CNN-Training.

**Granularity (Granularität)**

Detailtiefe von Daten oder Zeitintervallen. Fein = detailliert, grob = aggregiert.

**Beispiel:** Minutendaten vs. Monatsdurchschnitte.

**Graph Database**

NoSQL-Datenbank zur Speicherung vernetzter Daten. Nutzt Knoten und Kanten.

**Beispiel:** Neo4j speichert soziale Netzwerke.

**Grid Search**

Brute-Force-Methode zur Hyperparameteroptimierung durch Testen aller Kombinationen.

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

**Group By**

SQL-Kommando zur Gruppierung von Zeilen nach Spaltenwerten, oft kombiniert mit Aggregationen.

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

**Growth Rate**

Wachstumsrate eines Werts über die Zeit.

**Beispiel:** 8 % Umsatzwachstum pro Monat.

**Ground Truth**

Verifizierte Referenzdaten zur Modellvalidierung.

**Beispiel:** Manuell gelabelte Bilddaten für ein CNN.

**GUI (Graphical User Interface)**

Benutzeroberfläche zur Interaktion mit Software über visuelle Elemente.

**Beispiel:** Tableau Dashboard mit Drag-and-drop.

**GxP (Good x Practice)**

Regelwerke für Qualität und Sicherheit in regulierten Bereichen.

**Beispiel:** GMP in Pharma – gute Herstellpraxis.

**GMM (Gaussian Mixture Model)**

Clustering-Modell, das Daten als Mischung mehrerer Normalverteilungen modelliert.

**Beispiel:** Clustering von Kunden nach Verhalten.

**Gradient Boosting**

Boosting-Verfahren, das sequentiell Fehler reduziert.
Leistungsstark für strukturierte Daten.

**Beispiel:** XGBoost.

**Guesstimate**

Grobe, erfahrungsbasierte Schätzung.

**Beispiel:** Erwartete Rücklaufquote bei Umfrage = 30 %.

**Generalization**

Fähigkeit eines Modells, auf neue Daten korrekt zu reagieren.

**Beispiel:** Modell funktioniert auch auf unbekannten Kundendaten.

**Geoanalytics**

Analyse raumbezogener Daten mit geographischen Komponenten.

**Beispiel:** Heatmap der Verkaufszahlen pro Postleitzahl.

**Gaussian Naive Bayes**

Klassifikator, der Normalverteilungen pro Merkmal annimmt.

**Beispiel:** Schnell trainierter Textklassifikator.

**Greedy Algorithmus**

Algorithmus, der in jedem Schritt lokal beste Entscheidung trifft.
Nicht immer optimal.

**Beispiel:** Entscheidungsbaum-Split mit höchstem Informationsgewinn.

**Guided Analytics**

Interaktive Analyse mit vordefinierten Fragen oder Pfaden.

**Beispiel:** Nutzer klickt sich durch Dashboard zur Zielerkenntnis.

**Gamma Distribution**

Schiefe Wahrscheinlichkeitsverteilung für positive Werte.

**Beispiel:** Modellierung von Versicherungsforderungen.

**Gradient**

Vektor der partiellen Ableitungen, zeigt Richtung des stärksten Anstiegs einer Funktion.

**Beispiel:** Gradient in Backpropagation.

**Gaussian Kernel**

Funktion zur Gewichtung naher Datenpunkte in Kernel-Methoden.

**Beispiel:** SVM mit RBF-Kernel.

**Group Normalization**

Alternative zu Batch Normalization – robust bei kleinen Batchgrößen.

**Beispiel:** In CNNs für kleine Datenmengen.

**Graph Neural Network (GNN)**

Neural Network zur Verarbeitung von Graphstrukturen.

**Beispiel:** Vorhersage von Moleküleigenschaften anhand ihrer Struktur.

**Hash Function**
Funktion zur Umwandlung beliebiger Daten in einen festen Code. Häufig in Sicherheit, Indexierung oder Datenabgleich genutzt.
**Beispiel:** SHA-256 erzeugt aus einem Passwort einen einzigartigen Hash.

**Histogram**

Diagramm zur Darstellung der Häufigkeitsverteilung numerischer Werte in Intervallen (Bins).

**Beispiel:** Visualisierung der Altersverteilung in Kundenstammdaten.

**Hyperparameter**

Voreingestellte Modellparameter, die nicht durch Training gelernt werden, sondern manuell oder durch Optimierung bestimmt werden.

**Beispiel:** Lernrate, Anzahl Bäume in Random Forest.

**Hyperparameter Tuning**

Optimierung der Hyperparameter zur Verbesserung der Modellperformance.

**Beispiel:** Grid Search zur Auswahl der besten `max_depth` und `n_estimators`.

**Hypothesis Testing**

Statistisches Verfahren zum Testen einer Annahme über eine Population.

**Beispiel:** Test, ob die durchschnittliche Conversion Rate > 3 % ist.

**Heteroskedasticity**

Nicht-konstante Varianz der Fehler in einem Regressionsmodell. Kann zu verzerrten Ergebnissen führen.

**Beispiel:** Residuen steigen mit dem Einkommen.

**Heuristic**

Vereinfachte Regel zur schnellen Problemlösung, nicht garantiert optimal.

**Beispiel:** „Wenn Nutzer >10x klickt, ist er interessiert."

**Holdout Set**

Teil der Daten, der nicht zum Training, sondern ausschließlich zur abschließenden Bewertung eines Modells verwendet wird.

**Beispiel:** 80/10/10-Split: Training/Validation/Holdout.

**HDFS (Hadoop Distributed File System)**

Verteiltes Dateisystem zur Speicherung großer Datenmengen auf Clustern.

**Beispiel:** Rohdaten werden in Blöcken über mehrere Server verteilt gespeichert.

**Head (Tabellenfunktion)**

Zeigt die ersten n Zeilen eines Datensatzes an.

**Beispiel:** `df.head(5)` zeigt die ersten fünf Zeilen eines DataFrames.

**Hierarchical Clustering**

Clusteranalyse, bei der Daten schrittweise zu immer größeren Gruppen zusammengefügt werden.

**Beispiel:** Dendrogramm zeigt die Hierarchie von Kundenclustern.

**Homogeneity**

Maß für die Ähnlichkeit von Gruppen oder Clustern. Höher = ähnlicher.

**Beispiel:** Cluster mit reinem Alter 20–25 ist hoch homogen.

**Host (Server)**

Rechner oder Dienst, auf dem Anwendungen oder Datenbanken laufen.

**Beispiel:** PostgreSQL läuft auf `analytics.company.com`

**HTML (HyperText Markup Language)**

Standardauszeichnungssprache für Webseiten. Relevant für Webscraping.

**Beispiel:** Extraktion von Daten aus `<table>`-Elementen.

**HTTP (Hypertext Transfer Protocol)**

Protokoll für Datenübertragung im Web. Wichtig bei API-Calls und Web-Scraping.

**Beispiel:** REST-API liefert JSON über HTTP GET.

**Heuristic Algorithm**

Algorithmus, der mit Faustregeln arbeitet, um Lösungen effizient zu finden.

**Beispiel:** K-nearest-neighbor mit einfachem Abstandsmaß.

**Histogram Equalization**

Bildverarbeitungstechnik zur Kontrastanpassung durch Neuskalierung der Helligkeitsverteilung.

**Beispiel:** Verbesserung der Lesbarkeit von Röntgenbildern.

**Hinge Loss**

Verlustfunktion für lineare Klassifikation, v. a. bei SVMs.

**Beispiel:** Bestraft falsch klassifizierte Punkte mit Abstand zur Entscheidungsgrenze.

**Hamming Distance**

Anzahl unterschiedlicher Zeichen in zwei gleich langen Strings.

**Beispiel:** „10101“ vs. „11100“ → Hamming-Distanz = 3.

**Hash Join**

Join-Strategie in Datenbanksystemen, bei der Hash-Tabellen verwendet werden, um schnelle Übereinstimmungen zu finden.

**Beispiel:** Join zwischen großen Tabellen in PostgreSQL.

**Hierarchical Indexing**

Mehrstufiger Index in Pandas oder SQL, oft zur Gruppierung und Abfrage von Multi-Level-Daten.

**Beispiel:** MultiIndex aus „Region“ und „Jahr“.

**Heatmap**

Farbige Visualisierung von Korrelations- oder Häufigkeitsdaten in Matrixform.

**Beispiel:** Korrelation zwischen Features im Datensatz.

**Heuristic Threshold**

Schwellenwert, der empirisch oder erfahrungsbasiert gewählt wurde.

**Beispiel:** Kreditvergabe bei Score > 0.6.

**Hybrid Model**

Kombination verschiedener Modellarten oder Algorithmen, oft ML + regelbasiert.

**Beispiel:** Empfehlungssystem mit kollaborativem Filter + Content-based Matching.

**Hyperplane**

Trennfläche in höherdimensionalen Räumen, genutzt in SVMs zur Klassenabgrenzung.

**Beispiel:** Zwei Klassen im 3D-Raum werden durch eine Ebene getrennt.

**Hyperparameter Optimization**

Systematische Suche nach den besten Hyperparametern.

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

**Human-in-the-Loop**

Systeme, bei denen Menschen gezielt in den Entscheidungsprozess eingebunden werden.

**Beispiel:** Mensch überprüft Anomalien, die vom Modell markiert wurden.

**Hazard Function**

Wahrscheinlichkeit des Eintretens eines Ereignisses zu einem bestimmten Zeitpunkt, gegeben dass es noch nicht eingetreten ist.

**Beispiel:** Ausfallwahrscheinlichkeit einer Maschine pro Stunde.

**High Cardinality**

Spalte mit sehr vielen unterschiedlichen Werten – problematisch für One-Hot-Encoding.

**Beispiel:** E-Mail-Adressen, IDs, URLs.

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

Darstellung kategorialer Variablen als binäre Spalten.

**Beispiel:** „Rot“ → [1,0,0], „Blau“ → [0,1,0], „Grün“ → [0,0,1].

**ID (Identifier)**
Eindeutiger Schlüssel zur Unterscheidung von Datensätzen. Wird meist als Primärschlüssel verwendet.
**Beispiel:** `user_id = 1023` identifiziert einen bestimmten Kunden.

**Imbalanced Dataset**

Datensatz mit ungleich verteilten Klassen. Kann Klassifikationsmodelle stark beeinflussen.

**Beispiel:** 95 % „Nicht-Betrug“, 5 % „Betrug“.

**Imputation**

Verfahren zum Ersetzen fehlender Werte.

**Beispiel:** Fehlende Temperaturwerte durch Mittelwert auffüllen.

**Index (SQL/Pandas)**

Struktur zur schnellen Datenzugriffsoptimierung. In Pandas zusätzlich zur Zeilenidentifikation.

**Beispiel:** Index auf `customer_id` beschleunigt Abfragen.

**Independent Variable**

Unabhängige Variable in einer Analyse, Prädiktor.

**Beispiel:** Werbebudget als Einflussfaktor auf Umsatz.

**Inferential Statistics**

Verfahren zur Verallgemeinerung von Stichprobenergebnissen auf Populationen.

**Beispiel:** Konfidenzintervalle, Hypothesentests.

**Information Gain**

Maß für Reduktion der Unreinheit durch ein Attribut (v. a. Entscheidungsbäume).

**Beispiel:** Alter reduziert Entropie der Kaufentscheidung stark → hoher Gain.

**Inner Join**

SQL-Verknüpfung, die nur passende Zeilen beider Tabellen zurückgibt.

**Beispiel:** Nur Kunden mit mindestens einer Bestellung werden angezeigt.

**Instance**

Einzelnes Beispiel oder Datenpunkt in einem Datensatz.

**Beispiel:** Ein Kunde mit Attributen: Alter, Geschlecht, Umsatz.

**Interquartile Range (IQR)**

Spannweite zwischen 25. und 75. Perzentil. Robust gegen Ausreißer.

**Beispiel:** IQR für Alter liegt zwischen 30 und 50 → IQR = 20.

**Interpolation**

Schätzung fehlender Werte zwischen bekannten Punkten.

**Beispiel:** Temperatur am 15. durch Mittelwert von 14. und 16. geschätzt.

**Interpretability**

Grad, in dem ein Modell für Menschen verständlich ist.

**Beispiel:** Entscheidungsbaum ist gut interpretierbar, ein neuronales Netz eher nicht.

**Interval Data**

Numerische Daten mit gleichen Abständen, aber ohne echten Nullpunkt.

**Beispiel:** Temperatur in °C – 0 °C bedeutet nicht „keine Temperatur“.

**Intersection**

Schnittmenge zweier Datenmengen oder Mengenoperation in SQL.

**Beispiel:** Nutzer, die sowohl gekauft als auch bewertet haben.

**Iterative Process**

Wiederholender Prozess zur Verfeinerung von Modellen oder Workflows.

**Beispiel:** Feature-Engineering → Modelltraining → Evaluation → zurück.

**Isolation Forest**

ML-Verfahren zur Anomalieerkennung durch zufällige Partitionierung.

**Beispiel:** Auffällige Kredittransaktionen werden isoliert.

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

Annahme in der Statistik, dass Datenpunkte unabhängig und aus derselben Verteilung stammen.

**Beispiel:** Münzwürfe sind i.i.d., Einkommen nicht unbedingt.

**Identity Matrix**

Quadratische Matrix mit Einsen auf der Diagonalen, sonst Nullen.

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

**Indicator Variable**

Binäre Variable zur Kennzeichnung kategorialer Ausprägungen.

**Beispiel:** Geschlecht: „weiblich = 1“, sonst 0.

**Interaction Effect**

Wechselwirkung zwischen zwei oder mehr unabhängigen Variablen.

**Beispiel:** Werbeeffekt hängt vom Geschlecht UND Alter ab.

**Incremental Learning**

Modelltraining in kleinen Schritten ohne komplettes Neulernen.

**Beispiel:** Modell aktualisiert sich stündlich mit neuen Nutzerdaten.

**Inertia (K-Means)**

Summe der Abstände aller Punkte zu ihren Clusterzentren.

**Beispiel:** Ziel ist minimale Inertia → enge Cluster.

**Input Layer**

Erste Schicht eines neuronalen Netzes, nimmt Rohdaten auf.

**Beispiel:** 10 Neuronen für 10 Eingabefeatures.

**Image Recognition**

Erkennung von Objekten oder Mustern in Bildern mittels ML.

**Beispiel:** Modell identifiziert Katzen auf Fotos.

**Indexing (Pandas)**

Zugriff auf Datenzeilen oder -spalten durch Labels oder Positionen.

**Beispiel:** `df.loc['row1']` oder `df.iloc[0]`

**Inter-Rater Reliability**

Maß für Übereinstimmung zwischen mehreren Beurteilenden.

**Beispiel:** Zwei Ärzte stellen dieselbe Diagnose → hohe Reliabilität.

**IQR-Based Outlier Detection**

Ausreißererkennung auf Basis von IQR.

**Beispiel:** Werte außerhalb [Q1 – 1.5×IQR, Q3 + 1.5×IQR] gelten als Ausreißer.

**Imbalanced Learning**

ML-Techniken zur besseren Handhabung ungleicher Klassenverteilungen.

**Beispiel:** Verwendung von SMOTE zur Erzeugung synthetischer Minoritätsbeispiele.

**Integer**

Ganzzahliger Datentyp ohne Dezimalstellen.

**Beispiel:** 1, 42, -7 – aber nicht 3.14.

**Inferencing**

Anwendung eines trainierten Modells auf neue Daten.

**Beispiel:** Vorhersage des Kaufverhaltens für neue Nutzer.

**Indicator Matrix**

Matrixform für One-Hot-Encoding kategorialer Daten.

**Beispiel:** 3 Kategorien → 3 Spalten, je 0 oder 1.

**Identity Column (SQL)**

Automatisch hochzählende Spalte zur eindeutigen ID-Vergabe.

**Beispiel:** `id INT AUTO_INCREMENT`

**In-Memory Computing**

Verarbeitung großer Datenmengen direkt im RAM zur Beschleunigung.

**Beispiel:** Apache Spark verarbeitet Daten im Speicher statt auf Festplatte.

**Information Retrieval**

Prozess des Suchens und Findens relevanter Informationen in großen Datenmengen.

**Beispiel:** Suche nach Produktrezensionen mit bestimmtem Keyword.

**Instruction Set**

Befehlssatz eines Prozessors oder Systems, relevant in Low-Level-Datenverarbeitung.

**Beispiel:** SIMD-Instruktionen zur Parallelverarbeitung von Matrizen.

**Jaccard Distance**
Maß für die Unähnlichkeit zwischen zwei Mengen, definiert als 1 minus der Jaccard-Ähnlichkeit.
**Beispiel:** Zwei Listen mit 60 % Überschneidung → Distance = 0.4

**Jaccard Index**

Maß zur Berechnung der Ähnlichkeit zwischen zwei Mengen, definiert als Größe der Schnittmenge geteilt durch Größe der Vereinigungsmenge.

**Beispiel:** Zwei Mengen mit 4 gleichen und 6 unterschiedlichen Elementen ergeben 0.4.

**Jaccard Loss**

Verlustfunktion, die auf der Jaccard-Ähnlichkeit basiert, verwendet in Bildsegmentierung.

**Beispiel:** Semantic-Segmentation-Netzwerke.

**Jaccard Similarity Coefficient**

Alternativer Begriff für den Jaccard Index; wird häufig bei Clustering oder Recommender-Systemen verwendet.

**Beispiel:** Vergleich von Nutzerinteressen durch binäre Vektoren.

**Jaccard Similarity Matrix**

Matrix mit paarweisen Jaccard-Scores zwischen Mengen oder Dokumenten.

**Beispiel:** Ähnlichkeitsvergleich von Texten in einem Recommender-System.

**Jaccard Thresholding**

Verfahren zur Auswahl ähnlicher Paare basierend auf Mindestwert für den Jaccard Index.

**Beispiel:** Nur Paare mit Jaccard > 0.5 werden verknüpft.

**JAR File (Java Archive)**

Komprimiertes Archivformat für Java-Klassen, Konfigurationen und Bibliotheken.

**Beispiel:** Ein Apache-Spark-Job wird als ausführbares JAR übergeben.

**Java**

Plattformunabhängige, objektorientierte Programmiersprache, häufig in Enterprise- und Big-Data-Anwendungen.

**Beispiel:** Hadoop-MapReduce-Programme werden meist in Java geschrieben.

**Java EE (Enterprise Edition)**

Erweiterung von Java für Web- und Unternehmensanwendungen, mit Fokus auf Skalierbarkeit und Modularität.

**Beispiel:** Webservice mit Authentifizierung via Java EE.

**Java Native Interface (JNI)**

Schnittstelle zur Einbindung von C/C++-Code in Java-Anwendungen.

**Beispiel:** Java ruft eine Bibliothek zur Bildverarbeitung in C auf.

**Java Server Pages (JSP)**

Technologie zur serverseitigen Generierung dynamischer HTML-Inhalte in Java.

**Beispiel:** JSP-Seite zeigt Analyseergebnisse auf Knopfdruck an.

**Java Virtual Machine (JVM)**

Virtuelle Maschine, die Java-Bytecode in Maschinencode übersetzt und ausführt.

**Beispiel:** Apache Spark läuft auf der JVM.

**JavaBeans**

Java-Komponenten mit definierten Getter- und Setter-Methoden für die strukturierte Datenmodellierung.

**Beispiel:** `getName()` und `setName()` als Datenzugriff.

**JavaScript**

Skriptsprache zur dynamischen Webentwicklung, auch genutzt für Visualisierungstools.

**Beispiel:** D3.js-Visualisierungen im Browser.

**Jena (Apache Jena)**

Framework für semantische Webanwendungen und Verarbeitung von RDF-Daten.

**Beispiel:** SPARQL-Abfragen auf Wissensgraphen.

**Jenkins**

Open-Source-Automatisierungstool für Continuous Integration/Delivery.

**Beispiel:** Pipeline zur täglichen Ausführung von ETL-Jobs.

**Jensen-Shannon-Divergenz**

Maß zur Bewertung der Ähnlichkeit zwischen Wahrscheinlichkeitsverteilungen.

**Beispiel:** Vergleich von Sprachmodellen zweier Nachrichtentexte.

**Jitter (Visualisierung)**

Künstliches Streuen überlappender Punkte in einem Plot zur besseren Lesbarkeit.

**Beispiel:** Punktwolke mit Jitter bei identischen X-Werten.

**Job Queue**

System zur Verwaltung und Abarbeitung asynchroner Prozesse oder Aufgaben.

**Beispiel:** Warteschlange zur Bildverarbeitung auf einem Server.

**Joblib**

Python-Bibliothek zur Parallelisierung und Serialisierung von Aufgaben und Modellen.

**Beispiel:** Modell als `.pkl` speichern mit `joblib.dump()`.

**Join (SQL)**

Operation zur Kombination von Zeilen aus zwei Tabellen basierend auf einem gemeinsamen Attribut.

**Beispiel:** `JOIN kunden ON kunden.id = bestellungen.kundennr`.

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

Varianten des SQL-Joins mit unterschiedlichen Ergebnismengen.

**Beispiel:** LEFT JOIN zeigt alle Kunden, auch ohne Bestellung.

**JSON (JavaScript Object Notation)**

Textbasiertes, hierarchisches Format zum Speichern und Übertragen strukturierter Daten.

**Beispiel:** `{ "name": "Anna", "alter": 30 }`.

**JSDOM**

JavaScript-Implementierung des DOM in Node.js-Umgebungen.

**Beispiel:** Testen von Webseiten ohne echten Browser.

**JupyterHub**

Multi-User-Plattform zur Bereitstellung von Jupyter Notebooks in Teams und Bildungseinrichtungen.

**Beispiel:** Data-Science-Kurs mit zentralem Notebook-Server.

**JupyterLab**

Moderne, erweiterte Benutzeroberfläche für Jupyter Notebooks mit Tabs und Terminals.

**Beispiel:** Gleichzeitiges Öffnen von CSV, Code und Plot.

**Jupyter Notebook**

Webbasierte Entwicklungsumgebung für Python, die Code, Text und Visualisierungen kombiniert.

**Beispiel:** Explorative Datenanalyse mit Pandas und Seaborn.

**Jupyter Themes**

Anpassbare Designpakete zur Modifikation des Erscheinungsbilds von Jupyter.

**Beispiel:** Dunkler Hintergrund für bessere Lesbarkeit.

**Jupyter Widgets**

Interaktive Bedienelemente in Jupyter, etwa Slider oder Dropdowns.

**Beispiel:** Regler für Parameter in einer ML-Demo.

**JWT (JSON Web Token)**

Standardformat zur sicheren Übertragung von Informationen zwischen Parteien.

**Beispiel:** Zugriffstoken für geschützte APIs.

**Jaro-Winkler Distance**

Ähnlichkeitsmaß für Strings mit Fokus auf kleine Vertauschungen.

**Beispiel:** Vergleich von „Data“ und „Dtaa“ ergibt hohe Ähnlichkeit.

**k-Anonymity**
Datenschutzprinzip, das sicherstellt, dass Daten nicht eindeutig auf Einzelpersonen zurückgeführt werden können, wenn sie in Gruppen von mindestens *k* nicht unterscheidbaren Personen vorliegen.
**Beispiel:** Eine Tabelle ist 3-anonym, wenn jede Kombination von Quasi-Identifikatoren mindestens dreimal vorkommt.

**Kaggle**

Online-Plattform für Datenanalysewettbewerbe, Tutorials und Community-Projekte. Sie bietet offene Datasets und eine interaktive Jupyter-Umgebung.

**Beispiel:** Teilnahme an einem Wettbewerb zur Vorhersage von Wohnpreisen.

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

Statistisches Maß zur Bewertung der Übereinstimmung zwischen zwei Klassifikatoren unter Berücksichtigung zufälliger Übereinstimmungen.

**Beispiel:** Vergleich von menschlicher Klassifikation und Modellklassifikation.

**KDE (Kernel Density Estimation)**

Nichtparametrisches Verfahren zur Schätzung der Wahrscheinlichkeitsdichte einer Zufallsvariablen.

**Beispiel:** Glättung eines Histogramms zur Analyse der Datenverteilung.

**Kendall's Tau**

Korrelationskoeffizient für Rangdaten, der die Übereinstimmung zweier Rangordnungen bewertet.

**Beispiel:** Vergleich der Rangfolge von Produkten durch zwei Algorithmen.

**Kernel Trick**

Methode in der SVM, um nichtlineare Daten durch Transformation in einen höherdimensionalen Raum trennbar zu machen.

**Beispiel:** Verwendung eines RBF-Kernels für komplexe Klassifikationsprobleme.

**Key-Value Store**

Einfaches NoSQL-Datenbanksystem, bei dem Daten als Schlüssel-Wert-Paare gespeichert werden.

**Beispiel:** Redis oder Amazon DynamoDB.

**K-Fold Cross Validation**

Technik zur Modellvalidierung, bei der Daten in $k$ Teile geteilt und das Modell mehrfach trainiert und getestet wird.

**Beispiel:** 10-Fold Cross Validation für robustes Modell-Scoring.

**KMeans**

Beliebter Clustering-Algorithmus, der Daten in *k* Gruppen einteilt, basierend auf deren Ähnlichkeit.

**Beispiel:** Kundensegmentierung nach Kaufverhalten.

**KMedoids**

Clustering-Verfahren ähnlich KMeans, aber robuster gegen Ausreißer, da reale Datenpunkte als Clusterzentren gewählt werden.

**Beispiel:** Clustering von Usern basierend auf ihren Surfmustern.

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

Einfacher Klassifikationsalgorithmus, der die Klasse eines Punktes basierend auf der Mehrheitsklasse seiner *k* nächsten Nachbarn bestimmt.

**Beispiel:** Handschriftenerkennung basierend auf Pixeln.

**Knowledge Graph**

Netzwerk aus Entitäten und deren Beziehungen, das Wissen strukturiert repräsentiert.

**Beispiel:** Google Knowledge Graph zur Verbesserung von Suchergebnissen.

**Kolmogorov-Smirnov-Test**

Nichtparametrischer Test zur Bewertung, ob eine Stichprobe einer Referenzverteilung folgt.

**Beispiel:** Prüfung, ob Daten normalverteilt sind.

**Kolmogorov-Komplexität**

Maß für die Informationsmenge eines Objekts, definiert als Länge des kürzesten Programms, das es erzeugt.

**Beispiel:** Zufallszahlen haben hohe Kolmogorov-Komplexität.

**Komplexitätsklasse**

Klassifikation von Problemen nach ihrem rechnerischen Aufwand.

**Beispiel:** P, NP, NP-schwer im Kontext von Algorithmus-Analyse.

**Konfidenzintervall**

Bereich, der den wahren Wert eines Parameters mit einer bestimmten Wahrscheinlichkeit enthält.

**Beispiel:** "Der Mittelwert liegt mit 95% Sicherheit zwischen 10 und 12."

**Konfusionsmatrix**

Tabelle zur Bewertung von Klassifikationsmodellen, die wahre/nicht wahre Positiv/Negativ-Werte darstellt.

**Beispiel:** Analyse der Treffergenauigkeit eines Spam-Filters.

**Kontingenztabelle**

Kreuztabelle zur Darstellung von Häufigkeiten zweier kategorialer Merkmale.

**Beispiel:** Verteilung von Geschlecht und Zustimmung zu einer Aussage.

**Kontinuierliche Variable**

Variable mit unendlich vielen möglichen Ausprägungen in einem Intervall.

**Beispiel:** Temperatur, Gewicht, Einkommen.

**Konvergenz (Numerik)**

Eigenschaft eines Algorithmus, sich einem stabilen Wert oder Lösung näherungsweise anzunähern.

**Beispiel:** Gradientenverfahren in der linearen Regression.

**Korrelation**

Statistischer Zusammenhang zwischen zwei Variablen.

**Beispiel:** Positiver Zusammenhang zwischen Werbung und Umsatz.

**Korrelationsmatrix**

Matrix mit paarweisen Korrelationskoeffizienten mehrerer Variablen.

**Beispiel:** Vergleich von Aktienrenditen.

**Kovarianz**

Maß für die gemeinsame Variabilität zweier Zufallsvariablen.

**Beispiel:** Wenn x steigt und y dazu, ist Kovarianz positiv.

**Kritischer Wert**

Grenzwert, ab dem ein statistisches Testergebnis als signifikant gilt.

**Beispiel:** t-Kritisch = 2.01 für df=20 bei α = 0.05.

**Kruskal-Wallis-Test**

Nichtparametrischer Test zur Analyse von Unterschieden zwischen mehr als zwei Gruppen.

**Beispiel:** Vergleich von Benutzerbewertungen mehrerer Produkte.

**k-d Tree**

Datenstruktur zur schnellen Suche in mehrdimensionalen Räumen.

**Beispiel:** Effiziente Nachbarsuche in k-NN-Algorithmen.

**Kollinearität**

Problem in der Regression, wenn unabhängige Variablen hoch korreliert sind.

**Beispiel:** Gewicht und BMI gleichzeitig in einer Regressionsanalyse.

**Kreuztabelle**

Synonym für Kontingenztabelle, oft in Excel und Statistiksoftware verwendet.

**Beispiel:** Darstellung der Kundenanzahl pro Region und Geschlecht.

**Kumulative Verteilung**

Funktion, die angibt, mit welcher Wahrscheinlichkeit eine Zufallsvariable kleiner oder gleich einem Wert ist.

**Beispiel:** 80% der Werte liegen unter x = 15.

**Kurtosis (Wölbung)**

Statistisches Maß für die "Spitzigkeit" einer Verteilung.

**Beispiel:** Hohe Kurtosis bei stark konzentrierten Daten um den Mittelwert.

**K-Anonymisierung**

Praktische Umsetzung der k-Anonymity, oft durch Generalisierung oder Suppression.

**Beispiel:** Alter 31 wird zu Altersgruppe 30–39.

**KPI (Key Performance Indicator)**

Kennzahl zur Bewertung von Prozessen, Leistung oder Zielerreichung.

**Beispiel:** Conversion Rate, Churn Rate.

**Kernkomponente**

Zentrales Element oder Einflussfaktor in einem komplexen System.

**Beispiel:** Feature mit starker Gewichtung in einem Modell.

**Knowledge Discovery in Databases (KDD)**

Gesamter Prozess der Mustererkennung in Daten, inkl. Vorverarbeitung, Modellierung und Interpretation.

**Beispiel:** Data-Mining-Projekt zur Betrugserkennung.

**Kombinatorik**

Teilgebiet der Mathematik zur Zählung möglicher Kombinationen und Anordnungen.

**Beispiel:** Anzahl der möglichen Passwortvarianten mit 3 Zeichen.

**Label Encoding**
Verfahren zur Umwandlung kategorialer Variablen in numerische Werte, indem jeder Kategorie eine Ganzzahl zugewiesen wird.
**Beispiel:** "rot" = 0, "grün" = 1, "blau" = 2

**Lag Feature**

Zeitversetzte Variable in Zeitreihenanalysen, um vergangene Werte zur Prognose künftiger Zustände zu nutzen.

**Beispiel:** Temperatur von gestern als Feature für heute.

**Lagrange-Multiplikator**

Mathematische Methode zur Berücksichtigung von Nebenbedingungen bei Optimierungsproblemen.

**Beispiel:** Optimierung eines Modells unter Ressourcenbeschränkung.

**Lambda-Funktion (Python)**

Anonyme Kurzfunktion, die mit dem `lambda`-Schlüsselwort definiert wird.

**Beispiel:** `lambda x: x**2` ergibt das Quadrat von x.

**Laplacian (Graphentheorie)**

Matrix zur Beschreibung der Struktur eines Graphen, oft verwendet bei Clustering oder Graph-basierten ML-Algorithmen.

**Beispiel:** Laplace-Matrix bei Spectral Clustering.

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

Regressionsmethode mit L1-Regularisierung, die Koeffizienten auf null setzen kann.

**Beispiel:** Feature-Selektion durch Lasso-Regression.

**Latente Variable**

Nicht direkt beobachtbare Variable, die Einfluss auf beobachtete Daten hat.

**Beispiel:** Kundenloyalität als latente Einflussgröße auf Kaufverhalten.

**Latent Dirichlet Allocation (LDA)**

Themenmodellierungsverfahren zur Entdeckung latenter Themen in Textdaten.

**Beispiel:** Extraktion von Themen aus Nutzerbewertungen.

**Layer (NN)**

Ebene in einem neuronalen Netz, die aus Knoten besteht und Transformationen durchführt.

**Beispiel:** Eingabeschicht, versteckte Schicht, Ausgabeschicht.

**Leaky ReLU**

Aktivierungsfunktion in neuronalen Netzen, die auch für negative Werte einen kleinen Gradienten liefert.

**Beispiel:** `f(x) = x` für $x > 0$, `f(x) = 0.01x` sonst.

**Leistung (Statistik)**

Wahrscheinlichkeit, dass ein Test eine falsche Nullhypothese korrekt ablehnt (Power).

**Beispiel:** Ein Test mit 80% Leistung erkennt einen echten Effekt mit 80% Wahrscheinlichkeit.

**Likelihood**

Wahrscheinlichkeit, dass ein Modell gegebene Daten erzeugt, wichtig für Maximum-Likelihood-Schätzungen.

**Beispiel:** Likelihood einer Normalverteilung bei gegebenen Messwerten.

**Likelihood Ratio Test**

Vergleich zweier verschachtelter Modelle über das Verhältnis ihrer Likelihoods.

**Beispiel:** Test, ob ein zusätzliches Feature die Modellgüte verbessert.

**Lineare Regression**

Statistisches Modell, das den Zusammenhang zwischen einer abhängigen und unabhängigen Variable durch eine lineare Gleichung beschreibt.

**Beispiel:** Umsatz = a + b * Werbekosten

**Linear Discriminant Analysis (LDA)**

Klassifikationsverfahren, das Merkmalsräume so transformiert, dass Klassen gut trennbar sind.

**Beispiel:** Trennung von Spam- und Nicht-Spam-Mails.

**Lineare Unabhängigkeit**

Eigenschaft einer Variablenmenge, dass keine Variable als Linearkombination der anderen darstellbar ist.

**Beispiel:** Features mit hoher Korrelation sind nicht linear unabhängig.

**Linkage (Clustering)**

Strategie zur Berechnung der Distanz zwischen Clustern in hierarchischen Verfahren.

**Beispiel:** Single-Linkage verbindet die nächsten Punkte zweier Cluster.

**Little's MCAR Test**

Statistischer Test, um zu prüfen, ob fehlende Werte zufällig (MCAR) sind.

**Beispiel:** Diagnose von Ausfällen in Umfragedaten.

**Local Outlier Factor (LOF)**

Verfahren zur Identifikation lokaler Ausreißer durch Dichtevergleich mit Nachbarn.

**Beispiel:** Detektion seltener Ereignisse in Sensordaten.

**Log-Loss (Logarithmic Loss)**

Verlustfunktion für probabilistische Klassifikatoren, die falsche, sichere Vorhersagen stark bestraft.

**Beispiel:** `-log(p)` bei p = 0.01 ergibt hohen Verlust.

**Logarithmische Transformation**

Transformation zur Reduktion von Schiefe in rechts-schiefen Verteilungen.

**Beispiel:** Anwendung von `log(x+1)` auf Einkommensdaten.

**Logistische Regression**

Klassifikationsmodell, das Wahrscheinlichkeiten für binäre Klassen vorhersagt.

**Beispiel:** Vorhersage, ob ein Kunde kündigt.

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

Spezielle Form eines rekurrenten neuronalen Netzes, das Langzeitabhängigkeiten lernen kann.

**Beispiel:** Textgenerierung aus Sequenzen.

**Look-Up Table**

Tabelle zur schnellen Zuordnung von Eingabewerten zu Ausgabewerten.

**Beispiel:** Mapping von Codes zu Kategorien.

**Loss Function**

Funktion zur Quantifizierung des Fehlers eines Modells.

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

**Low Cardinality**

Kategoriales Merkmal mit wenigen verschiedenen Ausprägungen.

**Beispiel:** "Geschlecht" oder "Wochentag".

**Lurking Variable**

Verdeckter Einflussfaktor, der eine scheinbare Beziehung zwischen zwei beobachteten Variablen erklärt.

**Beispiel:** Eiskonsum und Ertrinken werden beide durch das Wetter beeinflusst.

**Lückenanalyse (Gap Analysis)**

Vergleich des Ist-Zustands mit dem Soll-Zustand zur Identifikation von Optimierungspotenzialen.

**Beispiel:** Umsatzsoll = 10M€, Ist = 8M€, Gap = 2M€.

**LZ77**

Algorithmus zur verlustfreien Datenkompression mittels Erkennung von Wiederholungsmustern.

**Beispiel:** Grundlage des ZIP-Dateiformats.

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

Effizienter Kompressionsalgorithmus mit hoher Kompressionsrate.

**Beispiel:** 7z-Archivformat.

**L1-Regularisierung**

Regulierungsmethode zur Bestrafung großer Koeffizienten in Modellen, fördert Sparsamkeit.

**Beispiel:** Lasso-Regression.

**L2-Regularisierung**

Bestraft große Koeffizienten quadratisch, stabilisiert das Modell.

**Beispiel:** Ridge-Regression.

**Latent Semantic Analysis (LSA)**

Textanalyseverfahren, das latente Bedeutungsbeziehungen zwischen Wörtern ermittelt.

**Beispiel:** Dokumentenclustering nach Inhalt.

**Latent Space**

Abstrakter Merkmalsraum, in den Daten durch ein Modell projiziert werden.

**Beispiel:** Repräsentation von Bildern in einem Autoencoder.

**Lemmatization**

Textvorverarbeitungsschritt zur Rückführung von Wörtern auf ihre Grundform.

**Beispiel:** "went" wird zu "go".

**Machine Learning (ML)**
Oberbegriff für Methoden, bei denen Modelle aus Daten lernen, ohne explizit programmiert zu sein.
**Beispiel:** Ein Modell lernt, Spam-Mails zu erkennen.

**Manifold Learning**

Nichtlineare Dimensionenreduktion zur Entdeckung von niedrigdimensionalen Strukturen in hochdimensionalen Daten.

**Beispiel:** t-SNE oder Isomap zur Datenvisualisierung.

**MapReduce**

Verteiltes Programmiermodell zur Verarbeitung großer Datenmengen.

**Beispiel:** Google nutzt MapReduce zur Indexierung des Webs.

**Marginalisierung**

Integration über unwichtige Variablen, um Verteilungen zu vereinfachen.

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

**Markov-Kette**

Modell, bei dem die Wahrscheinlichkeit des nächsten Zustands nur vom aktuellen Zustand abhängt.

**Beispiel:** Wettermodell: Sonne → Regen mit definierter Übergangswahrscheinlichkeit.

**Markov Decision Process (MDP)**

Mathematisches Modell für Entscheidungsfindung unter Unsicherheit.

**Beispiel:** Optimale Strategie im Reinforcement Learning.

**MAE (Mean Absolute Error)**

Durchschnittlicher absoluter Fehler zwischen Prognose und Beobachtung.

**Beispiel:** MAE von 2 bedeutet, dass Vorhersagen im Schnitt um 2 Einheiten abweichen.

**Mean**

Arithmetisches Mittel einer Zahlenreihe.

**Beispiel:** Mittelwert von [2, 4, 6] ist 4.

**Mean Imputation**

Ersatz fehlender Werte durch den Mittelwert der Spalte.

**Beispiel:** Fehlende Alterseinträge werden durch den Durchschnitt ersetzt.

**Mean Shift**

Clustering-Verfahren, das Dichte-Maxima identifiziert und Cluster daran ausrichtet.

**Beispiel:** Gruppierung von Kunden in Dichteregionen.

**Mean Squared Error (MSE)**

Durchschnitt des quadratischen Fehlers zwischen Vorhersage und Realwert.

**Beispiel:** Großer MSE zeigt starke Abweichung.

**Median**

Zentraler Wert einer sortierten Datenreihe.

**Beispiel:** Median von [1, 3, 9] ist 3.

**Median Imputation**

Ersatz fehlender Werte durch den Median.

**Beispiel:** Robuste Methode bei Ausreißern.

**Membership Inference Attack**

Angriff, bei dem ermittelt wird, ob bestimmte Daten beim Training eines Modells verwendet wurden.

**Beispiel:** Angriff auf ein ML-Modell, um Trainingsdaten zu extrahieren.

**Memory-Based Learning**

Lernmethode, bei der alle Beispiele gespeichert und zur Vorhersage herangezogen werden.

**Beispiel:** k-NN speichert alle Datenpunkte.

**Meta-Learning**

"Lernen zu lernen": Modelle lernen, wie sie neue Aufgaben schnell und effizient lösen.

**Beispiel:** Few-Shot Learning bei Bildklassifikation.

**Metric Learning**

Lernen einer Distanzfunktion, die relevante Ähnlichkeiten korrekt abbildet.

**Beispiel:** Gesichtsvergleich auf Basis gelernter Ähnlichkeitsmetriken.

**Min-Max-Normalisierung**

Skalierung von Werten auf einen definierten Bereich, meist [0, 1].

**Beispiel:** Werte zwischen 5 und 10 werden auf 0–1 gestreckt.

**Minimum Description Length (MDL)**

Prinzip der Modellwahl basierend auf der Kürze der Beschreibung von Daten plus Modell.

**Beispiel:** Präferenz für einfache Modelle mit guter Erklärungskraft.

**Minimum Spanning Tree**

Teilgraph mit minimaler Gesamtkantengewichtung, der alle Knoten verbindet.

**Beispiel:** Netzwerkoptimierung bei Kabelverbindungen.

**Missing Completely at Random (MCAR)**

Fehlende Daten sind völlig zufällig und unabhängig von beobachteten oder unbeobachteten Werten.

**Beispiel:** Sensorfehler ohne systematische Ursache.

**Missing Not at Random (MNAR)**

Fehlende Daten hängen mit den fehlenden Werten selbst zusammen.

**Beispiel:** Hohe Einkommen werden überdurchschnittlich oft nicht angegeben.

**Missing at Random (MAR)**

Fehlende Daten hängen nur mit beobachteten Werten zusammen.

**Beispiel:** Alter beeinflusst Wahrscheinlichkeit für fehlende Einkommensangabe.

**Mode**

Der am häufigsten vorkommende Wert in einem Datensatz.

**Beispiel:** Modus von [1, 2, 2, 3] ist 2.

**Model Drift**

Verlust der Modellgenauigkeit über Zeit durch Änderungen in den Daten.

**Beispiel:** Ein Empfehlungsmodell altert, wenn sich Nutzerverhalten ändert.

**Model Interpretability**

Verständlichkeit der Entscheidungslogik eines Modells für Menschen.

**Beispiel:** Entscheidungsbaum ist besser interpretierbar als ein neuronales Netz.

**Model Selection**

Auswahl des besten Modells anhand von Validierungskriterien.

**Beispiel:** Vergleich mehrerer Regressionsmodelle mit Cross-Validation.

**Model Zoo**

Sammlung vortrainierter Modelle, oft mit offenen Gewichten und Dokumentation.

**Beispiel:** TensorFlow Hub oder HuggingFace Transformers.

**Model-Based Clustering**

Clustering basierend auf der Annahme, dass Daten von einer Mischung statistischer Modelle stammen.

**Beispiel:** Gaussian Mixture Models.

**Modelkomplexität**

Grad an Freiheitsgraden und Parametern eines Modells.

**Beispiel:** Neuronale Netze mit vielen Schichten sind komplexer als lineare Modelle.

**Monte Carlo Simulation**

Zufallsbasierte Simulation zur Näherung von Wahrscheinlichkeitsverteilungen.

**Beispiel:** Prognose von Projektrisiken durch viele Durchläufe.

**Multikollinearität**

Problem in der Regression, wenn unabhängige Variablen stark korrelieren.

**Beispiel:** Gewicht und BMI als Regressoren.

**Multilabel-Klassifikation**

Klassifikationsproblem mit mehreren zutreffenden Labels pro Instanz.

**Beispiel:** Ein Film kann gleichzeitig als Komödie und Action gelten.

**Multivariate Analyse**

Analyse mehrerer abhängiger Variablen gleichzeitig.

**Beispiel:** Gleichzeitige Vorhersage von Gewicht und Blutdruck.

**Mutual Information**

Maß für die Abhängigkeit zwischen zwei Variablen.

**Beispiel:** MI = 0 bei unabhängigen Variablen.

**MVP (Minimum Viable Product)**

Einfachste funktionsfähige Version eines Produkts zur Überprüfung am Markt.

**Beispiel:** Prototyp einer App mit Kernfunktionalität.

**MXNet**

Deep-Learning-Framework mit Fokus auf Performance und Skalierbarkeit.

**Beispiel:** Verwendung von MXNet für GPU-Training in der Cloud.

**MySQL**

Beliebtes relationales Datenbankmanagementsystem (RDBMS).

**Beispiel:** Speicherung strukturierter Transaktionsdaten.

**Naive Bayes**
Ein einfaches, probabilistisches Klassifikationsverfahren basierend auf Bayes' Theorem mit der Annahme bedingter Unabhängigkeit der Merkmale.
**Beispiel:** Spam-Filter klassifizieren E-Mails als Spam/Nicht-Spam basierend auf Wortwahrscheinlichkeiten.

**Named Entity Recognition (NER)**

Verfahren aus der natürlichen Sprachverarbeitung zur Identifikation benannter Entitäten wie Namen, Orte, Organisationen in Texten.

**Beispiel:** Erkennung von "Berlin" als Stadt in einem Text.

**NAND-Gatter**

Logikgatter in der Digitaltechnik, das ein Ausgangssignal liefert, wenn nicht beide Eingänge 1 sind. Es ist universell einsetzbar.

**Beispiel:** Grundlage für Speicherlogik in CPUs.

**Natural Language Processing (NLP)**

Teilgebiet der KI zur Verarbeitung, Analyse und Generierung natürlicher Sprache.

**Beispiel:** Chatbots, maschinelle Übersetzung, Textklassifikation.

**Natural Logarithm (ln)**

Logarithmus zur Basis e (Euler-Zahl, ca. 2,718), verwendet in exponentiellem Wachstum und Zerfall sowie in vielen ML-Algorithmen.

**Beispiel:** ln(x), oft genutzt in log-linearen Modellen.

**Negative Binomial Distribution**

Wahrscheinlichkeitsverteilung für die Anzahl der Fehlversuche bis zum r-ten Erfolg.

**Beispiel:** Modellierung der Anzahl von Kundenanrufen bis zur dritten Beschwerde.

**Negative Sampling**

Technik zum effizienten Training neuronaler Netzwerke bei sehr großen Ausgabemengen durch gezielte Auswahl negativer Beispiele.

**Beispiel:** Training von Word2Vec-Modellen.

**Nested Queries (SQL)**

Abfragen innerhalb anderer Abfragen, oft als Subqueries bezeichnet.

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

**Neural Network**

Maschinelles Lernmodell, das aus Schichten vernetzter künstlicher Neuronen besteht und komplexe Muster in Daten erkennt.

**Beispiel:** Bilderkennung oder Spracherkennung durch Deep Learning.

**NLP Pipeline**

Verarbeitungskette für Textdaten in NLP, oft bestehend aus Tokenisierung, Stoppwortentfernung, Lemmatisierung etc.

**Beispiel:** Analyse von Kundenfeedback durch strukturierte Schritte.

**Noise (Statistik)**

Unsystematische, zufällige Störungen oder Fehler in Daten, die nicht durch das Modell erklärt werden können.

**Beispiel:** Messfehler in Sensorwerten.

**Noise Reduction**

Verfahren zur Entfernung oder Minimierung von Rauschen in Daten.

**Beispiel:** Glättung von Zeitreihen durch Moving Average.

**Nominal Variable**

Kategoriale Variable ohne natürliche Reihenfolge.

**Beispiel:** Farben: Rot, Blau, Grün.

**Normalization**

Skalierung von numerischen Werten auf einen einheitlichen Bereich, oft zwischen 0 und 1.

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

**Normalverteilung**

Glockenförmige Wahrscheinlichkeitsverteilung mit Mittelwert und Standardabweichung. Häufige Annahme in der Statistik.

**Beispiel:** Körpergrößenverteilung in einer Population.

**Null Hypothesis (H0)**

Annahme, dass kein Effekt oder Unterschied vorliegt. Grundlage für viele statistische Tests.

**Beispiel:** "Die Werbemaßnahme hatte keinen Einfluss auf den Umsatz."

**Null Value**

Spezielle Kennzeichnung fehlender oder undefinierter Werte in Datenbanken oder Programmen.

**Beispiel:** NULL in SQL bedeutet kein Wert vorhanden.

**Numerical Feature**

Merkmal mit kontinuierlichen oder diskreten numerischen Werten.

**Beispiel:** Alter, Preis, Temperatur.

**Numerical Integration**

Berechnung von Näherungswerten für bestimmte Integrale, wenn keine analytische Lösung möglich ist.

**Beispiel:** Trapezregel oder Monte-Carlo-Verfahren zur Flächenberechnung.

**NumPy**

Python-Bibliothek für numerische Berechnungen und effiziente Array-Operationen. Grundlage vieler Datenanalyse-Tools.

**Beispiel:** Vektoroperationen mit `numpy.array()`.

**N-gram**

Sequenz von N aufeinanderfolgenden Elementen (z. B. Wörtern oder Zeichen) in Textdaten, genutzt zur Sprachmodellierung.

**Beispiel:** Trigramm von "Ich liebe dich" = ["Ich liebe", "liebe dich"]

**NaN (Not a Number)**

Spezielle Darstellung ungültiger oder fehlender numerischer Werte in Programmen wie Python oder R.

**Beispiel:** Division durch null ergibt NaN in `pandas`.

**Nearest Neighbor Search**

Algorithmus zur Suche der nächsten Punkte im Merkmalsraum, Grundlage für k-NN und Clustering.

**Beispiel:** Empfehlung ähnlicher Produkte.

**Nested Cross Validation**

Kombination aus zwei verschachtelten Cross-Validation-Schleifen für faire Modell- und Hyperparameterbewertung.

**Beispiel:** äußere CV zur Performance-Messung, innere CV zur Hyperparameter-Optimierung.

**NetworkX**

Python-Bibliothek zur Analyse und Visualisierung komplexer Netzwerke.

**Beispiel:** Soziale Netzwerkanalyse oder Transportnetzwerke.

**Newton-Raphson-Verfahren**

Iteratives Verfahren zur Lösung nichtlinearer Gleichungen.

**Beispiel:** Wurzelbestimmung oder Maximum-Likelihood-Schätzungen.

**Node (Graphentheorie)**

Einzelnes Element in einem Netzwerk oder Baumstruktur, z. B. ein Nutzer in einem sozialen Netzwerk.

**Beispiel:** Jeder Knoten in einem Entscheidungsbaum ist ein Node.

**NoSQL**

Datenbanktechnologien, die nicht auf relationalen Tabellen basieren, oft dokumenten- oder graphenbasiert.

**Beispiel:** MongoDB oder Cassandra.

**Nullmodell**

Einfaches Basismodell ohne erklärende Variablen, dient als Referenz zur Bewertung komplexerer Modelle.

**Beispiel:** Mittelwertmodell als Vergleich für lineare Regression.

**Numerische Stabilität**

Maß für die Robustheit numerischer Algorithmen gegenüber Rundungsfehlern.

**Beispiel:** Verwendung stabiler Matrizenoperationen in ML.

**Nyquist-Theorem**

Theorem aus der Signalverarbeitung, das die minimale Abtastfrequenz zur exakten Rekonstruktion eines Signals beschreibt.

**Beispiel:** Audio-Sampling muss mit mindestens doppelter Frequenz erfolgen.

**NamedTuple (Python)**

Datentyp in Python zur Definition von Tupeln mit benannten Feldern, ähnlich wie Klassen.

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

**Nesterov Momentum**

Optimierungsverfahren mit vorausschauendem Gradienten, verbessert die Konvergenz bei neuronalen Netzen.

**Beispiel:** Trainingsbeschleunigung im Vergleich zu klassischem Momentum.

**Noise Injection**

Technik zur Erhöhung der Modellrobustheit durch absichtliches Einfügen von Rauschen in Trainingsdaten.

**Beispiel:** Bildrauschen in der Bildklassifikation.

**Normalized Mutual Information (NMI)**

Metrik zur Bewertung der Übereinstimmung zweier Clusterings, skaliert auf [0,1].

**Beispiel:** Vergleich von Clustering-Ergebnissen mit Ground Truth.

**Numerisches Differenzieren**

Approximation der Ableitung durch Differenzenquotienten.

**Beispiel:** Finite-Differenzen-Methode in Optimierungsverfahren.

**Newton-Verfahren (Multivariabel)**

Erweiterung des Newton-Raphson-Verfahrens auf mehrere Variablen zur Optimierung.

**Beispiel:** Einsatz bei nichtkonvexen Zielfunktionen in ML.

**Nützlichkeit (Utility)**

Maß für den Wert oder Nutzen einer Handlung oder Vorhersage, häufig in Entscheidungsbäumen oder Recommender-Systemen.

**Beispiel:** Empfehlung mit maximalem erwarteten Nutzen.

**Object Detection**
Verfahren des maschinellen Sehens zur Lokalisierung und Klassifizierung mehrerer Objekte in einem Bild.

**Beispiel:** Erkennung von Fahrzeugen und Fußgängern in Echtzeit für autonome Fahrzeuge.

**Object-Oriented Programming (OOP)**

Programmierparadigma, das Daten und Verhalten in Objekten kapselt. Erleichtert Wiederverwendbarkeit, Modularität und Wartung.

**Beispiel:** In Python definierte Klassen zur Modellierung von Datentransformationen.

**Observability**

Fähigkeit, den internen Zustand eines Systems durch externe Outputs zu bestimmen.

**Beispiel:** Logfiles und Metriken zur Analyse von Datenpipelines.

**Observation**

Einzelner Datenpunkt in einem Datensatz, meist eine Zeile.

**Beispiel:** Ein Kunde mit allen Attributen in einer CRM-Tabelle.

**Occam's Razor**

Prinzip, nach dem bei gleicher Güte das einfachere Modell bevorzugt wird.

**Beispiel:** Wahl eines linearen Modells statt eines tiefen neuronalen Netzes bei gleicher Leistung.

**OCR (Optical Character Recognition)**

Technik zur automatischen Texterkennung in Bildern oder gescannten Dokumenten.

**Beispiel:** Digitalisierung von Rechnungen im PDF-Format.

**Octile Distance**

Metrik zur Berechnung von Distanzen in Rastern mit diagonalen Bewegungen.

**Beispiel:** Pfadfindung in Gitterkarten.

**ODS (Operational Data Store)**

Zentraler Speicher operativer Daten für Reporting und Analyse in nahezu Echtzeit.

**Beispiel:** Daten aus mehreren Systemen für ein Dashboard zusammengeführt.

**Offline Learning**

Modelltraining auf einem statischen, vorher bekannten Datensatz.

**Beispiel:** Klassifikatortraining auf historischem Nutzerverhalten.

**OGNL (Object Graph Navigation Language)**

Ausdruckssprache zur Navigation und Manipulation von Objektgraphen, z. B. in Java-Frameworks.

**Beispiel:** Zugriff auf verschachtelte Werte in JavaBeans.

**OLS (Ordinary Least Squares)**

Standardverfahren zur Schätzung linearer Regressionsmodelle.

**Beispiel:** Minimierung der quadrierten Fehler zur Anpassung einer Regressionslinie.

**One-Hot-Encoding**

Kategoriale Kodierung, bei der jede Kategorie als binärer Vektor dargestellt wird.

**Beispiel:** „rot“, „blau“, „grün“ → [1,0,0], [0,1,0], [0,0,1]

**One-Class SVM**

Support-Vector-Machine für Anomalieerkennung in einer einzigen Klasse.

**Beispiel:** Erkennung von Betrug basierend auf „normalem“ Verhalten.

**Online Learning**

Lernverfahren, bei dem das Modell schrittweise mit neuen Daten aktualisiert wird.

**Beispiel:** Anpassung eines Empfehlungssystems in Echtzeit.

**Ontology**

Formale Beschreibung von Begriffen und deren Beziehungen innerhalb eines Wissensgebiets.

**Beispiel:** Datenmodell für medizinische Diagnosen mit ICD-Begriffen.

**Open Data**

Daten, die frei nutzbar, weiterverwendbar und weiterverbreitbar sind.

**Beispiel:** Verkehrsdaten einer Stadtregierung für Entwickler freigegeben.

**Open Source**

Software, deren Quellcode öffentlich zugänglich ist und modifiziert werden darf.

**Beispiel:** Python-Bibliotheken wie Pandas oder Scikit-learn.

**Operationalization**

Übersetzung abstrakter Konzepte in messbare Variablen.

**Beispiel:** „Kundenzufriedenheit“ wird durch eine Umfrage mit 5-Punkte-Skala operationalisiert.

**Optimization**

Prozess der Verbesserung eines Modells, Algorithmus oder Systems durch Feinjustierung.

**Beispiel:** Hyperparameter-Tuning mit Grid Search.

**Optimizer**

Algorithmus zur Anpassung der Modellparameter während des Lernprozesses.

**Beispiel:** Adam-Optimizer in neuronalen Netzen.

**Ordinal Data**

Daten mit natürlicher Reihenfolge, aber ohne festen Abstand zwischen Werten.

**Beispiel:** Zufriedenheitsumfrage: „sehr schlecht“ bis „sehr gut“.

**Outlier**

Datenpunkt, der signifikant von anderen abweicht.

**Beispiel:** Einkommen von 1.000.000 € in einem Datensatz mit Mittelwert 50.000 €.

**Outlier Detection**

Verfahren zur Identifikation von Ausreißern.

**Beispiel:** Isolation Forest oder Z-Score.

**Output Layer**

Letzte Schicht in einem neuronalen Netz, die die finale Vorhersage liefert.

**Beispiel:** Softmax-Schicht für Klassifikation mit mehreren Klassen.

**Overfitting**

Modellanpassung, die sich zu stark an Trainingsdaten orientiert und Generalisierungsfähigkeit verliert.

**Beispiel:** Komplexes Modell mit 100% Trainingsgenauigkeit, aber schlechter Testleistung.

**Oversampling**

Technik zur Erhöhung der Anzahl seltener Klassen in unbalancierten Datensätzen.

**Beispiel:** SMOTE zur künstlichen Erzeugung von Minoritätsklassen.

**Own Join**

Join-Operation, bei der eine Tabelle mit sich selbst verknüpft wird.

**Beispiel:** Hierarchien in Mitarbeiterdaten analysieren.

**Ox Metrics**

Softwarepaket zur ökonometrischen Modellierung und Zeitreihenanalyse.

**Beispiel:** Durchführung von ARIMA-Modellen.

**Out-of-Bag Error**

Fehlerschätzung bei Bagging-Verfahren, basierend auf Daten, die beim Bootstrapping nicht verwendet wurden.

**Beispiel:** Random Forest verwendet OOB-Daten zur internen Validierung.

**Out-of-Sample Performance**

Modellleistung auf unbekannten, nicht im Training verwendeten Daten.

**Beispiel:** Validierungsergebnisse im Holdout-Set.

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

Wörter, die im Trainingsvokabular eines NLP-Modells nicht enthalten sind.

**Beispiel:** Umgang mit neuen Slang-Begriffen in Chatbots.

**Outlier Score**

Numerischer Wert, der angibt, wie stark ein Datenpunkt ein Ausreißer ist.

**Beispiel:** LOF-Score > 1.5 gilt oft als Ausreißer.

**Ordinal Encoding**

Zuordnung von Ganzzahlen zu geordneten kategorialen Variablen.

**Beispiel:** „niedrig“ = 0, „mittel“ = 1, „hoch“ = 2.

**Oracle**

System oder Komponente, das als allwissend angenommen wird und für Vergleichszwecke dient.

**Beispiel:** Theorie-Modell mit perfektem Wissen als Benchmark.

**Operational Metric**

Kennzahl zur Überwachung betrieblicher Abläufe und Datenverarbeitungssysteme.

**Beispiel:** Latenzzeit oder Fehlerrate einer Pipeline.

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

Strategie zur Erweiterung binärer Klassifikatoren auf Mehrklassenprobleme.

**Beispiel:** Drei binäre Modelle für Klassen A vs B+C, B vs A+C, C vs A+B.

**Ordinal Logistic Regression**

Regressionsmodell für ordinal skalierte Zielvariablen.

**Beispiel:** Analyse von Kundenzufriedenheitsskalen.

**Online Analytical Processing (OLAP)**

Technologie zur schnellen multidimensionalen Analyse großer Datenmengen.

**Beispiel:** Drilldown von Quartalsumsätzen nach Region und Produkt.

**OpenAI API**

Programmierschnittstelle zur Nutzung von Sprachmodellen und KI-Diensten von OpenAI.

**Beispiel:** Textgenerierung durch einen API-Call aus einer Python-Anwendung.

**P-Value (p-Wert)**
Statistisches Maß zur Bewertung der Signifikanz eines Ergebnisses. Ein niedriger p-Wert deutet darauf hin, dass ein beobachtetes Ergebnis nicht durch Zufall erklärbar ist.
**Beispiel:** Ein p-Wert von 0.01 bedeutet, dass die

Wahrscheinlichkeit für das Ergebnis unter der Nullhypothese bei 1 % liegt.

**Pandas**

Python-Bibliothek zur Datenmanipulation und -analyse. Sie stellt leistungsfähige Datenstrukturen wie DataFrames bereit.

**Beispiel:** `df = pd.read_csv("daten.csv")` lädt eine CSV-Datei in ein DataFrame.

**Parameter**

Feste Werte in einem statistischen Modell, die geschätzt werden müssen. Sie definieren das Verhalten des Modells.

**Beispiel:** In einer linearen Regression ist die Steigung ein Parameter.

**Parquet**

Spaltenbasiertes Speicherformat, optimiert für große Datenmengen. Unterstützt effiziente Abfragen und Kompression.

**Beispiel:** Speicherung eines DataFrames in `data.parquet` für schnelle Analysen.

**Partial Dependence Plot (PDP)**

Visualisierung des Einflusses eines Merkmals auf das Modell-Ergebnis, unter Kontrolle aller anderen Merkmale.

**Beispiel:** PDP zeigt, wie sich der Hauspreis bei zunehmender Wohnfläche entwickelt.

**Partitioning**

Aufteilung von Daten in logische oder physische Teile, z. B. bei Datenbanken oder Data Lakes.

**Beispiel:** Monatsweise Partitionierung einer Tabelle zur Performance-Steigerung.

**Pearson-Korrelation**

Maß für linearen Zusammenhang zwischen zwei Variablen. Werte reichen von –1 (negativ) bis +1 (positiv).

**Beispiel:** Korrelation zwischen Lernzeit und Prüfungsergebnis.

**Percentile**

Schwellenwerte, die eine Verteilung in 100 gleiche Teile teilen.

**Beispiel:** Das 90. Perzentil ist der Wert, unter dem 90 % der Daten liegen.

**Permutationstest**

Nichtparametrischer Test zur Bestimmung der Signifikanz durch zufälliges Neuanordnen der Daten.

**Beispiel:** Vergleich von Mittelwerten zweier Gruppen durch Permutation.

**Pipelines**

Reihenfolge von Verarbeitungsschritten, z. B. bei Datenvorverarbeitung und ML-Training.

**Beispiel:** Skalieren, Feature-Engineering, Modelltraining – alles in einer Pipeline.

**Pivot-Tabelle**

Excel-Funktion zur schnellen Aggregation und Analyse großer Datenmengen.

**Beispiel:** Summierung von Umsatz nach Produkt und Region.

**Plotly**

Interaktive Visualisierungsbibliothek in Python. Unterstützt dynamische Grafiken für Web und Dashboarding.

**Beispiel:** `plotly.express.scatter()` für interaktive Scatterplots.

**Poisson-Verteilung**

Wahrscheinlichkeitsverteilung für Ereignisse mit konstanter durchschnittlicher Rate.

**Beispiel:** Anzahl der Anrufe pro Stunde im Callcenter.

**Polynomial Regression**

Regressionsmodell mit nichtlinearer Beziehung, durch Polynomterme höherer Ordnung.

**Beispiel:** Vorhersage der Verkaufszahlen bei wachsender Werbung mit abnehmender Grenzwirkung.

**Population**

Gesamtheit aller Elemente, über die statistische Aussagen gemacht werden.

**Beispiel:** Alle Bürger eines Landes in einer Umfrage.

**Portierung (Porting)**

Übertragung von Code oder Daten von einer Plattform auf eine andere.

**Beispiel:** Portierung eines Skripts von R nach Python.

**Precision (Genauigkeit)**

Anteil der korrekt als positiv klassifizierten Elemente an allen als positiv klassifizierten.

**Beispiel:** 80 % Precision bedeutet: Von 100 als „krank“ vorhergesagten Personen sind 80 wirklich krank.

**Precision-Recall-Kurve**

Visualisierung von Precision und Recall bei verschiedenen Schwellenwerten.

**Beispiel:** Entscheidungsgrundlage bei unbalancierten Datensätzen.

**Predictive Modeling**

Erstellung von Modellen zur Vorhersage zukünftiger Ereignisse basierend auf historischen Daten.

**Beispiel:** Prognose des Kundenabsprungs mit ML-Modell.

**Prescriptive Analytics**

Analytischer Ansatz, der nicht nur vorhersagt, was passieren wird, sondern auch Handlungsempfehlungen gibt.

**Beispiel:** Empfehlung von Preisänderungen basierend auf Nachfrageprognose.

**Principal Component Analysis (PCA)**

Dimensionsreduktionsverfahren, das Daten auf Basis der Hauptvarianzrichtungen projiziert.

**Beispiel:** Reduktion von 100 Features auf 3 Hauptkomponenten.

**Priorwahrscheinlichkeit**

Subjektive Anfangswahrscheinlichkeit vor Beobachtung von Daten, z. B. in Bayes-Theorie.

**Beispiel:** Vorerwartung, dass 5 % der Kunden kündigen.

**Probability Density Function (PDF)**

Funktion, die die Wahrscheinlichkeitsverteilung einer stetigen Zufallsvariablen beschreibt.

**Beispiel:** Glockenkurve bei Normalverteilung.

**Probability Mass Function (PMF)**

Entsprechung der PDF für diskrete Zufallsvariablen.

**Beispiel:** Anzahl der gewürfelten Sechsen in 10 Würfen.

**Process Mining**

Technik zur Analyse realer Geschäftsprozesse anhand von Logdaten.

**Beispiel:** Entdeckung ineffizienter Abläufe in Supportprozessen.

**Profiling (Datenprofiling)**

Analyse der Struktur, Qualität und Eigenschaften von Datenbeständen.

**Beispiel:** Erkennung von Dubletten, Nullwerten, Inkonsistenzen.

**Prophet (Facebook)**

Open-Source-Tool für Zeitreihenprognosen mit einfacher API.

**Beispiel:** Prognose saisonaler Verkaufszahlen.

**Protokoll (Logging)**

Systematische Aufzeichnung von Prozessen, Fehlern oder Transaktionen.

**Beispiel:** Speichern von Anfragen an eine API zur Fehlersuche.

**Python**

Weit verbreitete Programmiersprache im Data-Science-Bereich, bekannt für Lesbarkeit und umfangreiche Bibliotheken.

**Beispiel:** Verwendung von `numpy`, `pandas`, `scikit-learn` zur Datenanalyse.

**PyTorch**

Python-Framework für Deep Learning mit dynamischer Berechnungsgrafik.

**Beispiel:** Aufbau und Training neuronaler Netze mit GPU-Unterstützung.

**PySpark**

Python-Schnittstelle für Apache Spark zur verteilten Datenverarbeitung.

**Beispiel:** Verarbeitung großer CSV-Dateien im Cluster.

**Pseudocode**

Programmiernahe Beschreibung von Algorithmen in Klartextform, unabhängig von Programmiersprache.

**Beispiel:** Beschreibung eines Sortierverfahrens in strukturiertem Text.

**P-Wert-Korrektur**

Anpassung von p-Werten bei multiplen Tests zur Kontrolle des Fehlertyps.

**Beispiel:** Bonferroni-Korrektur bei mehreren Hypothesentests.

**pandas_profiling**

Python-Bibliothek zur schnellen automatisierten Datenanalyse und Erstellung eines EDA-Reports.

**Beispiel:** `df.profile_report()` generiert PDF mit Statistiken und Plots.

**Point-Biseriale Korrelation**

Korrelation zwischen einer binären und einer metrischen Variable.

**Beispiel:** Zusammenhang zwischen Geschlecht und Einkommen.

**Poisson-Prozess**

Stochastischer Prozess für zählbare Ereignisse über kontinuierliche Zeit.

**Beispiel:** Modellierung von Anrufen im Callcenter.

**Power BI**

BI-Tool von Microsoft für Dashboards, Reports und Datenvisualisierung.

**Beispiel:** Verbindung zu Excel und Visualisierung von Verkaufszahlen.

**PostgreSQL**

Leistungsfähiges, objektrelationales Open-Source-Datenbanksystem.

**Beispiel:** Nutzung von SQL und JSON-Funktionen zur Datenanalyse.

**Probability Calibration**

Anpassung von Vorhersagewahrscheinlichkeiten zur besseren Interpretation.

**Beispiel:** Platt Scaling bei unkalibrierten Modellen.

**Prediction Interval**

Intervall, das zukünftige Einzelbeobachtungen mit definierter Wahrscheinlichkeit einschließt.

**Beispiel:** Prognose von Temperatur morgen: 17–21 °C mit 95 % Sicherheit.

**Precision Medicine**

Ansatz in der Medizin, bei dem Entscheidungen auf individuellen Patientendaten basieren.

**Beispiel:** Personalisierte Krebsbehandlung auf Basis von Genomdaten.

**Q-Q Plot (Quantile-Quantile Plot)**
Grafische Methode zum Vergleich zweier Verteilungen, indem deren Quantile gegeneinander aufgetragen werden.
**Beispiel:** Normalverteilungs-Q-Q-Plot zeigt, ob Daten normalverteilt sind (Punkte liegen auf der Diagonalen).

**Q-Learning**

Bestärkendes Lernverfahren, bei dem ein Agent aus Belohnungen lernt, welche Aktionen in welchem Zustand am lohnendsten sind.

**Beispiel:** Ein autonomer Agent lernt, Hindernissen auszuweichen und Belohnungen zu sammeln.

**Quadratic Loss**

Verlustfunktion, bei der die Differenz zwischen Vorhersage und wahrem Wert quadriert wird.

**Beispiel:** Mean Squared Error (MSE) ist eine Form von Quadratic Loss.

**Quadratic Programming (QP)**

Optimierungsproblem mit quadratischer Zielfunktion und linearen Nebenbedingungen.

**Beispiel:** Portfoliomanagement mit Risiko-Minimierung.

**Qualitative Data**

Nicht-numerische, kategoriale Daten, die Zustände oder Merkmale beschreiben.

**Beispiel:** Farben, Produktkategorien, Kundenmeinungen.

**Quantile**

Werte, die eine Verteilung in gleich große Intervalle aufteilen.

**Beispiel:** Das 25%-Quantil (Q1) ist der Wert, unter dem 25% der Daten liegen.

**Quantile Regression**

Regressionsverfahren, das nicht den Mittelwert, sondern ein bestimmtes Quantil der Zielvariablen modelliert.

**Beispiel:** Prognose des 90%-Quantils der Lieferzeit.

**Quantitative Data**

Zahlenbasierte Daten, die gemessen oder gezählt werden können.

**Beispiel:** Alter, Umsatz, Temperatur.

**Quantization**

Reduktion der Genauigkeit von Werten auf eine diskrete Menge, oft in ML zur Modellkomprimierung genutzt.

**Beispiel:** Komprimierung neuronaler Netze durch 8-Bit-Quantisierung.

**Query**

Anfrage an eine Datenbank, um bestimmte Informationen zu extrahieren.

**Beispiel:** `SELECT * FROM kunden WHERE land = 'DE'` ist eine SQL-Query.

**Query Optimization**

Prozess der Verbesserung der Ausführungsgeschwindigkeit von Datenbankanfragen.

**Beispiel:** Nutzung von Indexen und Join-Strategien zur Query-Beschleunigung.

**Queue**

Datenstruktur, bei der Elemente in der Reihenfolge ihres Eintreffens verarbeitet werden (FIFO).

**Beispiel:** Warteschlange bei Ereignisverarbeitung.

**QuickSort**

Effizienter, rekursiver Sortieralgorithmus mit Divide-and-Conquer-Ansatz.

**Beispiel:** Sortieren eines Arrays mit Pivot-Element.

**Quota Sampling**

Nicht-zufällige Stichprobenmethode, bei der bestimmte Gruppenanteile gezielt erhoben werden.

**Beispiel:** 50% Frauen, 50% Männer in einer Umfrage.

**Quasi-Experiment**

Studie mit experimentellem Design ohne zufällige Zuweisung zu Gruppen.

**Beispiel:** Untersuchung der Wirkung einer Preisänderung ohne Zufallsstichprobe.

**Quasi-Newton Method**

Näherungsverfahren zur numerischen Optimierung, das auf einer Annäherung der Hesse-Matrix basiert.

**Beispiel:** BFGS-Algorithmus zur Minimierung einer Kostenfunktion.

**Quadrant Analysis**

Analyseverfahren zur Klassifikation von Datenpunkten anhand von zwei Achsen, meist in vier Quadranten unterteilt.

**Beispiel:** Priorisierung von Aufgaben nach Wichtigkeit und Dringlichkeit.

**Quantum Computing**

Rechenparadigma, das quantenmechanische Zustände für parallele Informationsverarbeitung nutzt.

**Beispiel:** Qubits statt Bits für exponentielle Rechenleistung.

**Query Plan**

Interner Ausführungsplan einer Datenbank zur Bearbeitung einer Anfrage.

**Beispiel:** Darstellung der Schritte eines SQL-Joins zur Analyse der Performance.

**Quality Assurance (QA)**

Systematische Prozesse zur Sicherstellung der Qualität von Daten, Modellen und Software.

**Beispiel:** Validierung von Datenpipelines und Unit Tests für ML-Modelle.

**Quantitative Trait**

Merkmal, das durch kontinuierlich messbare Werte beschrieben wird und oft durch mehrere Gene beeinflusst ist.

**Beispiel:** Körpergröße oder Blutzuckerwert.

**Quadratic Mean (RMS)**

Wurzel aus dem Durchschnitt der quadrierten Werte; robust gegen Ausreißer.

**Beispiel:** Berechnung der Effektivspannung.

**Query Language**

Programmiersprache zur Formulierung von Datenbankanfragen.

**Beispiel:** SQL, GraphQL.

**Quality Score**

Bewertung der Qualität eines Datenpunkts, einer Vorhersage oder eines Modells.

**Beispiel:** Bewertung von Anzeigeneffektivität im Marketing.

**Quadratwurzeltransformation**

Transformation zur Reduktion der Schiefe in zählbaren Daten.

**Beispiel:** `sqrt(x)` bei zählbaren Ereignissen wie Unfallzahlen.

**Quantile Normalization**

Methode zur Normalisierung mehrerer Verteilungen auf dieselbe Quantilverteilung.

**Beispiel:** Vergleich von Genexpressionen über verschiedene Experimente hinweg.

**Quicksight (AWS)**

BI-Tool von Amazon zur Datenvisualisierung und Dashboard-Erstellung.

**Beispiel:** Erstellung interaktiver Verkaufs-Dashboards für E-Commerce.

**Quorum**

Minimale Anzahl an Teilnehmern, die für eine Entscheidung oder ein Systemverhalten erforderlich ist.

**Beispiel:** Replikationssysteme in verteilten Datenbanken.

**Qubit**

Elementare Informationseinheit im Quantencomputing mit Überlagerungszuständen.

**Beispiel:** Ein Qubit kann gleichzeitig 0 und 1 sein.

**Query Federation**

Technik, um Anfragen über mehrere Datenquellen hinweg gleichzeitig auszuführen.

**Beispiel:** Kombinieren von Daten aus S3, Redshift und MySQL in einer Abfrage.

**Queueing Theory**

Mathematische Theorie zur Modellierung von Warteschlangenprozessen.

**Beispiel:** Optimierung von Callcenter-Kapazitäten.

**Query Caching**

Zwischenspeicherung von Abfrageergebnissen zur Beschleunigung wiederholter Zugriffe.

**Beispiel:** Redis als Cache für komplexe SQL-Reports.

**QGIS (Quantum GIS)**

Open-Source-Software zur Bearbeitung, Analyse und Visualisierung geographischer Daten.

**Beispiel:** Darstellung von Kundenstandorten auf einer Karte.

**Quasi-Poisson Regression**

Variante der Poisson-Regression, die Überdispersion in zählbaren Daten berücksichtigt.

**Beispiel:** Modellierung von Anrufzahlen bei variierender Tageslast.

**Quality Control Chart**

Diagramm zur Überwachung der Qualität in Prozessen durch statistische Grenzen.

**Beispiel:** SPC-Kontrollkarte zur Produktionsüberwachung.

**Quadratmatrix**

Matrix mit gleicher Anzahl von Zeilen und Spalten, wichtig für lineare Algebra und Eigenwertanalysen.

**Beispiel:** Kovarianzmatrix.

**Quantitative PCR (qPCR)**

Laborverfahren zur quantitativen Bestimmung von DNA/RNA-Mengen.

**Beispiel:** Nachweis viraler Lasten in medizinischen Tests.

**Quadraturregel**

Numerisches Verfahren zur Näherung von Integralen.

**Beispiel:** Trapezregel zur Flächenberechnung unter Kurven.

**Query Result Cache**

Speicherbereich, in dem Datenbankantworten für schnelleren Zugriff zwischengelagert werden.

**Beispiel:** Oracle Query Result Cache.

**Quasi-Binomial Model**

Generalisiertes lineares Modell, das Überdispersion bei binären Ergebnissen erlaubt.

**Beispiel:** Modellierung von Conversion-Raten bei Online-Werbung.

**Quantitative Forecasting**

Prognosemethode, die numerische Zeitreihendaten verwendet.

**Beispiel:** Absatzprognose auf Basis historischer Verkaufszahlen.

**Quotientenkorrelation**

Verhältnisbasierte Korrelation zweier Variablen, genutzt bei Dimensionsreduktion.

**Beispiel:** Anteil von Kategorie-A-Käufen an Gesamtkäufen.

**Quadratfehler (Squared Error)**

Fehlerwert, der durch Quadrieren der Differenz zwischen Ist- und Sollwert berechnet wird.

**Beispiel:** (y - ŷ)^2 bei Vorhersageabweichungen.

**Quick Ratio**

Kennzahl für kurzfristige Liquidität eines Unternehmens.

**Beispiel:** (Umlaufvermögen - Vorräte) / kurzfristige Verbindlichkeiten.

**R (Programmiersprache)**
Statistisch orientierte Programmiersprache mit starker Unterstützung für Datenanalyse, Visualisierung und wissenschaftliches Rechnen.
**Beispiel:** R wird häufig in der akademischen Forschung verwendet, z. B. für lineare Modelle oder ggplot2-Visualisierungen.

**R-Squared ($R^2$, Bestimmtheitsmaß)**

Statistisches Maß zur Bewertung der Erklärkraft eines Regressionsmodells. Werte nahe 1 bedeuten hohe Erklärungsgüte.

**Beispiel:** Ein $R^2$ von 0{,}85 bedeutet, dass 85 % der Varianz durch das Modell erklärt werden.

**Random Forest**

Ensemble-Lernverfahren, das viele Entscheidungsbäume kombiniert, um Klassifikationen oder Regressionen robust durchzuführen.

**Beispiel:** Ein Random Forest kann zur Vorhersage von Kreditrisiken verwendet werden.

**Random Sampling**

Zufällige Auswahl von Beobachtungen aus einer Population zur Schätzung oder Analyse.

**Beispiel:** Zufälliges Ziehen von 1.000 Kunden für eine Umfrage.

**Random Variable (Zufallsvariable)**

Variable, deren Wert vom Ergebnis eines Zufallsprozesses abhängt.

**Beispiel:** Augenzahl eines geworfenen Würfels.

**Range (Spannweite)**

Differenz zwischen dem größten und kleinsten Wert in einem Datensatz.

**Beispiel:** Bei den Werten 2, 4, 6 ist die Range 6 - 2 = 4.

**Rank Transformation**

Transformation numerischer Werte in ihre Rangfolge.

**Beispiel:** Werte [100, 50, 75] werden zu Rängen [3, 1, 2].

**Rasch-Modell**

Statistisches Modell zur Skalierung latenter Merkmale, oft in der Psychometrie.

**Beispiel:** Analyse von Schülerantworten in standardisierten Tests.

**Rate Limiting**

Technik zur Begrenzung der Anzahl von API-Anfragen innerhalb eines Zeitfensters.

**Beispiel:** Max. 1000 Anfragen pro Stunde bei einem Webservice.

**Rationalisierung**

Datenbereinigung durch Entfernen redundanter oder irrelevanter Informationen.

**Beispiel:** Zusammenfassung doppelter Einträge in einer Kundenliste.

**Raw Data**

Unverarbeitete Rohdaten, wie sie direkt aus Quellen generiert werden.

**Beispiel:** CSV-Export aus einer Sensor-API.

**Recall**

Kennzahl für Klassifikationsmodelle, die misst, wie viele der tatsächlich positiven Fälle korrekt erkannt wurden.

**Beispiel:** 80 % Recall bedeutet: 80 % aller Positiven wurden erkannt.

**Receiver Operating Characteristic (ROC)**

Kurve zur Visualisierung der Trade-Offs zwischen Sensitivität und Spezifität.

**Beispiel:** Fläche unter der ROC-Kurve (AUC) als Maß für Modellgüte.

**Recoding**

Umschüsselung oder Umcodierung von Variablenwerten in eine andere Struktur.

**Beispiel:** "m" und "f" werden zu 0 und 1.

**Recursive Feature Elimination (RFE)**

Feature-Selektionstechnik, bei der schrittweise weniger relevante Merkmale entfernt werden.

**Beispiel:** RFE mit Random Forest zur Reduktion von 100 auf 10 wichtige Features.

**Regression**

Statistisches Verfahren zur Modellierung von Zusammenhängen zwischen abhängigen und unabhängigen Variablen.

**Beispiel:** Vorhersage von Hauspreisen auf Basis von Größe, Lage, Zustand.

**Regression Tree**

Entscheidungsbaum-Modell zur Vorhersage numerischer Zielvariablen.

**Beispiel:** Entscheidungspfade zur Vorhersage von Gehaltswerten.

**Regularization**

Technik zur Vermeidung von Overfitting durch Bestrafung großer Koeffizienten.

**Beispiel:** L1- und L2-Regularisierung in Regressionsmodellen.

**Reinforcement Learning**

Lernparadigma, bei dem ein Agent durch Belohnung und Bestrafung optimale Strategien erlernt.

**Beispiel:** Trainingsprozess eines Schachprogramms.

**Relationale Datenbank**

Datenbanksystem mit tabellarischer Struktur und Beziehungen über Schlüssel.

**Beispiel:** MySQL-Datenbank für Kundendaten.

**Relationship**

Verknüpfung zwischen Tabellen in relationalen Datenbanken.

**Beispiel:** Fremdschlüssel verbindet Kunden und Bestellungen.

**Relative Häufigkeit**

Anteil eines Werts an der Gesamtheit.

**Beispiel:** 40 von 200 Kunden haben gekauft → 20 % relative Häufigkeit.

**Replikation**

Wiederholung einer Analyse, um Ergebnisse zu überprüfen.

**Beispiel:** Reproduktion eines ML-Modells mit neuen Daten.

**Residual (Residuum)**

Differenz zwischen beobachtetem und vorhergesagtem Wert.

**Beispiel:** Beobachtung = 10, Prognose = 8 → Residuum = 2.

**Residual Sum of Squares (RSS)**

Summe der quadrierten Residuen, Maß für Modellgüte in der Regression.

**Beispiel:** Niedriger RSS weist auf gute Modellanpassung hin.

**Resolution (Rasterdaten)**

Maß für die Detailgenauigkeit von Rasterbildern oder -daten.

**Beispiel:** 10x10 Pixelauflösung pro km in einem Höhenmodell.

**Resampling**

Techniken wie Bootstrapping oder Cross-Validation zur Erhöhung der Robustheit von Schätzungen.

**Beispiel:** Bootstrapping zur Konfidenzintervallschätzung.

**REST API**

Webservice-Architektur auf Basis von HTTP-Methoden.

**Beispiel:** GET, POST, PUT, DELETE auf Ressourcen wie /kunden.

**Resultatmatrix**

Matrix mit Ausgabeergebnissen eines Analyseverfahrens oder eines Modells.

**Beispiel:** Konfusionsmatrix bei Klassifikation.

**Retail Analytics**

Datenanalyse im Einzelhandel zur Optimierung von Sortiment, Preisgestaltung, Lagerhaltung.

**Beispiel:** Analyse von Kassendaten zur Prognose von Abverkäufen.

**Ridge Regression**

Regressionsform mit L2-Regularisierung zur Vermeidung von Overfitting.

**Beispiel:** Stabilisierung bei multikollinearen Daten.

**Right Join**

SQL-Operation, bei der alle Zeilen der rechten Tabelle und passende der linken ausgegeben werden.

**Beispiel:** Kunden ohne Bestellung tauchen nicht auf, aber alle Bestellungen werden angezeigt.

**ROC AUC**

Metrik zur Bewertung von Klassifikationsmodellen, misst Fläche unter der ROC-Kurve.

**Beispiel:** AUC-Wert von 0.95 zeigt hohe Klassifikationsqualität.

**Root Mean Square Error (RMSE)**

Wurzel des mittleren quadratischen Fehlers, gebräuchliches Maß für Prognosegüte.

**Beispiel:** RMSE = 2 bedeutet durchschnittlich 2 Einheiten Abweichung.

**Round()**

Funktion zur Rundung numerischer Werte.

**Beispiel:** round(3.14159, 2) ergibt 3.14.

**Row-Level Security**

Technik zur Einschränkung des Datenzugriffs auf Zeilenebene, z. B. in BI-Tools.

**Beispiel:** Nutzer sieht nur Verkaufsdaten seiner Region.

**R Script**

Datei mit R-Befehlen zur wiederholbaren Analyse und Visualisierung.

**Beispiel:** Automatisiertes Reporting in RMarkdown.

**Runtime**

Ausführungszeit eines Programms oder Modells, oft kritisch bei Big Data.

**Beispiel:** Python-Skript läuft in 12 Sekunden durch.

**Run-Length Encoding (RLE)**

Kompressionsmethode, bei der Wiederholungen durch Zählen codiert werden.

**Beispiel:** "AAAABBB" → "4A3B".

**Rückpropagation (Backpropagation)**

Lernverfahren für neuronale Netze, das Fehler zurück durch das Netz propagiert.

**Beispiel:** Training eines CNN durch Minimierung des Fehlers.

**Rückschlussstatistik**

Teilgebiet der Statistik zur Generalisierung von Stichproben auf Grundgesamtheiten.

**Beispiel:** Konfidenzintervall, Hypothesentest.

**Sample**
Teilmenge aus einer größeren Population, die für Analysen verwendet wird.

**Beispiel:** 500 Kundenbefragungen aus einer Datenbank mit 10.000 Einträgen.

**Sampling Bias**

Verzerrung durch eine nicht-repräsentative Stichprobe.

**Beispiel:** Online-Umfrage nur unter jungen Nutzern.

**Sampling Rate**

Frequenz, mit der Daten erfasst werden.

**Beispiel:** Ein Sensor, der alle 10 Sekunden misst, hat eine Sampling Rate von 0{,}1 Hz.

**Sankey Diagram**

Visualisierung von Flüssen und Verzweigungen, z. B. bei Energieflüssen oder Nutzerpfaden.

**Beispiel:** Darstellung der Conversion-Ströme in einem Online-Shop.

**Scalability**

Fähigkeit eines Systems, bei wachsendem Datenvolumen effizient zu bleiben.

**Beispiel:** Cloud-Datenbanken skalieren horizontal bei großen Datenmengen.

**Scaling**

Transformation von Features auf einen vergleichbaren Wertebereich.

**Beispiel:** Min-Max- oder Z-Score-Normalisierung.

**Scenario Analysis**

Modellierung alternativer Zukunftsszenarien zur Risikoabschätzung.

**Beispiel:** Umsatz bei +10 % bzw. -10 % Nachfrage.

**Scatterplot**

Diagramm zur Darstellung von Beziehungen zwischen zwei numerischen Variablen.

**Beispiel:** Alter vs. Einkommen.

**Schema**

Struktur und Definition von Tabellen, Feldern und Beziehungen in einer Datenbank.

**Beispiel:** Datenbankschema mit "Kunden", "Bestellungen", "Produkten".

**Scientific Method**

Strukturierter Prozess zur Erkenntnisgewinnung durch Hypothesenbildung und -prüfung.

**Beispiel:** Hypothese testen: "Mehr Licht erhöht Produktivität".

**Scikit-learn**

Populäre ML-Bibliothek in Python für Klassifikation, Regression, Clustering.

**Beispiel:** `RandomForestClassifier()` aus scikit-learn.

**Score**

Numerische Bewertung einer Vorhersage oder eines Modells.

**Beispiel:** Kreditrisiko-Score = 0{,}82

**Scoring Function**

Funktion zur Berechnung eines Scores für Klassifikation oder Ranking.

**Beispiel:** Log-Loss oder ROC-AUC als Scoring-Funktion.

**Script**

Programmdatei zur automatisierten Ausführung von Befehlen.

**Beispiel:** Python-Script zur Datenbereinigung.

**Seasonality**

Regelmäßige, wiederkehrende Muster in Zeitreihen.

**Beispiel:** Umsatzrückgang im Januar.

**Second Normal Form (2NF)**

Datenbanknormalform, die Teilschlusselabhängigkeiten vermeidet.

**Beispiel:** Trennung von Artikel- und Bestellinformationen.

**Segmentation**

Aufteilung von Daten in Gruppen mit ähnlichen Merkmalen.

**Beispiel:** Clusteranalyse zur Kundensegmentierung.

**Selectivity**

Anteil der Zeilen in einer Tabelle, die durch eine Abfrage getroffen werden.

**Beispiel:** SELECT * WHERE status = 'aktiv' mit 5 % Selectivity.

**Self-Join**

Verknüpfung einer Tabelle mit sich selbst.

**Beispiel:** Hierarchische Struktur in einer Mitarbeiterliste.

**Semi-Structured Data**

Daten mit teilweise definierter Struktur.

**Beispiel:** JSON- oder XML-Dokumente.

**Sensitivity (Recall)**

Maß für die korrekte Erkennung positiver Fälle.

**Beispiel:** 90 % Sensitivity = 90 % der Kranken erkannt.

**Sentiment Analysis**

Analyse von Meinungen und Gefühlen in Textdaten.

**Beispiel:** Bewertung von Tweets als positiv, neutral oder negativ.

**Sequence Model**

ML-Modelle, die auf sequentielle Daten spezialisiert sind.

**Beispiel:** LSTM für Text- oder Audiosequenzen.

**Shannon Entropy**

Maß für die Unbestimmtheit oder Informationsmenge einer Verteilung.

**Beispiel:** Maximale Entropie bei Gleichverteilung.

**Shapiro-Wilk-Test**

Statistischer Test auf Normalverteilung.

**Beispiel:** p-Wert < 0{,}05 deutet auf Abweichung von der Normalverteilung hin.

**Sharding**

Verteilung von Daten auf mehrere Server zur Lastverteilung.

**Beispiel:** Horizontales Sharding in verteilten NoSQL-Systemen.

**Sharpe Ratio**

Verhältnis von Überrendite zur Volatilität eines Investments.

**Beispiel:** Sharpe Ratio = (Rendite - risikofreier Zins) / Standardabweichung.

**Shotgun Stochastic Search**

Suchverfahren zur Modellauswahl in hochdimensionalen Datenräumen.

**Beispiel:** Auswahl relevanter Gene in Bioinformatik.

**Shrinkage**

Regularisierungsprinzip zur Verringerung von Modellkomplexität.

**Beispiel:** Ridge-Regression.

**Signal-to-Noise Ratio**

Verhältnis von nützlicher Information zu Rauschen.

**Beispiel:** 10:1 ist ein hoher SNR.

**Silhouette Score**

Maß für die Qualität eines Clusterings.

**Beispiel:** Score nahe 1 = klare Clustertrennung.

**Simulation**

Nachbildung von Prozessen zur Analyse von Szenarien.

**Beispiel:** Monte-Carlo-Simulation zur Risikobewertung.

**Singular Value Decomposition (SVD)**

Matrixfaktoriserung für Dimensionsreduktion und Latent Space Learning.

**Beispiel:** Empfehlungssysteme.

**Skewness (Schiefe)**

Maß für die Asymmetrie einer Verteilung.

**Beispiel:** Positive Schiefe bei Einkommen.

**Slack Space**

Unbenutzter Speicherbereich in Blöcken auf Datenträgern.

**Beispiel:** In forensischer Datenanalyse relevant.

**Sliding Window**

Technik zur Analyse sequentieller Daten durch Verschieben eines festen Fensters.

**Beispiel:** Moving Average mit Fenstergröße 5.

**Softmax**

Aktivierungsfunktion, die Wahrscheinlichkeitsverteilung über Klassen erzeugt.

**Beispiel:** Softmax-Ausgabe bei Klassifikation.

**Spearman-Korrelation**

Nichtparametrisches Maß für Rangkorrelation.

**Beispiel:** Bewertung von Korrelation bei nicht-normalverteilten Daten.

**SQL**

Sprache für die Abfrage und Manipulation relationaler Datenbanken.

**Beispiel:** `SELECT name FROM kunden WHERE ort = 'Berlin'`

**Stacked Generalization (Stacking)**

Ensemble-Methode zur Kombination mehrerer Modelle.

**Beispiel:** Meta-Classifier aggregiert Vorhersagen verschiedener Modelle.

**Standardabweichung**

Maß für die Streuung von Daten um den Mittelwert.

**Beispiel:** Std = 5 bedeutet, dass Daten im Schnitt 5 Einheiten abweichen.

**Standardfehler**

Schätzwert für die Streuung einer Statistik (z. B. Mittelwert).

**Beispiel:** StdError = 0{,}2 für Mittelwert einer Stichprobe.

**Standardisierung**

Skalierung von Variablen auf Mittelwert 0 und Std 1.

**Beispiel:** Z-Transformation für Regressionsanalyse.

**Stationarität**

Eigenschaft von Zeitreihen, deren statistische Eigenschaften sich nicht ändern.

**Beispiel:** Differenzieren zur Erreichung von Stationarität.

**Statistische Signifikanz**

Wahrscheinlichkeit, dass ein beobachteter Effekt kein Zufall ist.

**Beispiel:** p < 0{,}05 bedeutet statistisch signifikant.

**Stetige Variable (kontinuierlich)**

Variable mit unendlich vielen Ausprägungen im Wertebereich.

**Beispiel:** Körpergröße in cm.

**Stichprobe**

Auswahl von Datenpunkten zur Analyse einer Population.

**Beispiel:** 1.000 Kundenbefragungen.

**Streuung (Dispersion)**

Maß für die Verteilung der Werte um das Zentrum.

**Beispiel:** Hohe Streuung bei stark variierenden Einkommen.

**Stochastik**

Teilgebiet der Mathematik, das sich mit Zufall und Wahrscheinlichkeiten befasst.

**Beispiel:** Anwendung in Prognosemodellen.

**Stratifiziertes Sampling**

Aufteilung der Population in Schichten zur gezielten Stichprobenziehung.

**Beispiel:** Altersgruppen bei Wahlumfragen.

**Streaming Data**

Kontinuierlich eingehende Daten in Echtzeit.

**Beispiel:** Sensordaten oder Weblogs.

**Strukturierte Daten**

Daten mit klar definierter Struktur in Tabellenform.

**Beispiel:** Excel-Tabelle mit Kundendaten.

**Subsampling**

Teilweise Auswahl aus großem Datensatz.

**Beispiel:** 10 % der Logdaten für erste Analyse.

**Supervised Learning**

ML-Ansatz mit gelabelten Trainingsdaten.

**Beispiel:** Klassifikation von E-Mails als Spam oder Nicht-Spam.

**Support Vector Machine (SVM)**

ML-Algorithmus zur Klassifikation durch Trennung mit maximalem Abstand.

**Beispiel:** Erkennung von Handschrift in Bildern.

**Synthetische Daten**

Künstlich erzeugte Daten zur Modellierung oder zum Training.

**Beispiel:** Generierte Kundendaten zum Testen eines Dashboards.

**Syntax**

Grammatikregeln einer Programmiersprache.

**Beispiel:** Python: `if x > 0:`

**Systematische Verzerrung**

Verzerrung durch methodische Fehler oder Voreingenommenheit.

**Beispiel:** Schiefe Stichprobe bei nicht-randomisierter Auswahl.

**Tabellenstruktur**
Definiert die Organisation von Spalten und Datentypen in einer Datenbanktabelle.
**Beispiel:** Eine Kundentabelle mit Name (Text), Alter (Integer) und Stadt (Text).

**Tabellenverknüpfung**

Verbindung von zwei oder mehr Tabellen über gemeinsame Schlüssel.

**Beispiel:** Kunden-ID verbindet "Kunden"- und "Bestellungen"-Tabelle.

**Target Variable**

Die zu prognostizierende oder zu klassifizierende Zielvariable in einem ML-Modell.

**Beispiel:** "Kaufentscheidung" in einem Modell zur Vorhersage von Konversionen.

**Tidy Data**

Strukturierte Daten, bei denen jede Variable eine Spalte, jede Beobachtung eine Zeile ist.

**Beispiel:** Zeitreihendaten in langem Format.

**t-SNE (t-distributed Stochastic Neighbor Embedding)**

Nichtlineare Dimensionsreduktion zur Visualisierung hochdimensionaler Daten.

**Beispiel:** Darstellung von Wortvektoren in 2D.

**Target Encoding**

Kodierung kategorialer Variablen basierend auf dem Mittelwert der Zielvariablen.

**Beispiel:** Durchschnittliche Konversionsrate pro Werbekanal.

**TensorFlow**

Open-Source-Bibliothek von Google für Machine Learning und Deep Learning.

**Beispiel:** `tf.keras.Sequential()` zur Modellerstellung.

**Testdaten**

Datensatz zur Prüfung der Modellgüte, getrennt vom Trainingsdatensatz.

**Beispiel:** 20 % der Daten für die finale Validierung.

**Teststatistik**

Wert, der aus einer Stichprobe berechnet wird, um eine Hypothese zu testen.

**Beispiel:** t-Wert beim t-Test.

**Text Mining**

Extraktion strukturierter Informationen aus unstrukturierten Texten.

**Beispiel:** Themenextraktion aus Kundenbewertungen.

**Textklassifikation**

Zuordnung von Texten zu vordefinierten Kategorien.

**Beispiel:** E-Mail als Spam oder Nicht-Spam.

**TF-IDF (Term Frequency-Inverse Document Frequency)**

Gewichtungsmaß für Worte in Texten, um wichtige Begriffe hervorzuheben.

**Beispiel:** Häufiges Wort in einem Dokument, selten in anderen.

**Threshold**

Grenzwert zur Entscheidung bei Klassifikationen.

**Beispiel:** Wahrscheinlichkeit > 0{,}5 = Klasse 1.

**Time Series Analysis**

Analyse zeitlich geordneter Daten zur Prognose oder Mustererkennung.

**Beispiel:** Absatzentwicklung pro Monat.

**Time to Event**

Zeit bis zum Eintreten eines bestimmten Ereignisses.

**Beispiel:** Zeit bis zum ersten Kauf nach Newsletter-Anmeldung.

**Tokenisierung**

Aufteilung von Text in kleinere Einheiten wie Wörter oder Sätze.

**Beispiel:** "Data Science ist cool" → ["Data", "Science", "ist", "cool"].

**Top-K Accuracy**

Metrik, bei der das richtige Label unter den Top-K-Vorhersagen sein muss.

**Beispiel:** In Top-3-Vorhersagen ist die richtige Klasse enthalten.

**Tracking Code**

Skript oder Tag zur Erfassung von Nutzeraktionen.

**Beispiel:** Google Analytics Tracking Pixel.

**Trainingsdaten**

Daten, auf denen ein ML-Modell lernt.

**Beispiel:** Historische Verkaufszahlen zum Modelltraining.

**Transformationsfunktion**

Funktion zur Umwandlung von Daten für bessere Modellleistung.

**Beispiel:** Log-Transformation bei stark rechts-schiefen Daten.

**Transponieren**

Vertauschen von Zeilen und Spalten in einer Matrix oder Tabelle.

**Beispiel:** `df.T` in Pandas.

**True Negative**

Fall, in dem ein negatives Beispiel korrekt als negativ klassifiziert wird.

**Beispiel:** Gesunder Patient wird korrekt als gesund erkannt.

**True Positive**

Fall, in dem ein positives Beispiel korrekt erkannt wird.

**Beispiel:** Kranker Patient korrekt als krank erkannt.

**T-Test**

Statistischer Test zum Vergleich von Mittelwerten zweier Gruppen.

**Beispiel:** Vergleich der Durchschnittsausgaben von Männern und Frauen.

**Type I Error (Alpha-Fehler)**

Falsch-positiver Fehler: Ablehnung der Nullhypothese, obwohl sie wahr ist.

**Beispiel:** Gesunder wird als krank diagnostiziert.

**Type II Error (Beta-Fehler)**

Falsch-negativer Fehler: Nullhypothese wird beibehalten, obwohl sie falsch ist.

**Beispiel:** Kranker wird als gesund eingestuft.

**Typkonvertierung (Type Casting)**

Umwandlung von Datentypen innerhalb eines Programms oder einer Analyse.

**Beispiel:** `int("42")` in Python ergibt eine Ganzzahl.

**UAT (User Acceptance Testing)**
Letzte Phase des Softwaretests, in der reale Nutzer prüfen, ob das System ihren Anforderungen entspricht.
**Beispiel:** Ein BI-Tool wird von Endanwendern getestet, bevor es live geht.

**UDAF (User-Defined Aggregate Function)**

Benutzerdefinierte Funktion zur Aggregation mehrerer Werte in SQL-ähnlichen Sprachen.

**Beispiel:** Eigene Median-Funktion in Apache Hive.

**UDF (User-Defined Function)**

Benutzerdefinierte Funktion, die in SQL-, Python- oder Spark-Umgebungen verwendet wird.

**Beispiel:** Eigene Berechnungslogik mit `@udf` in PySpark.

**UI (User Interface)**

Schnittstelle zwischen Mensch und System zur Bedienung von Software.

**Beispiel:** Dashboard-Oberfläche mit Filterelementen.

**ULID (Universally Unique Lexicographically Sortable Identifier)**

Alternative zu UUIDs, die sortierbar ist.

**Beispiel:** ULID = 01F8MECHZX3TBDSZ7XRADM79XV

**UMAP (Uniform Manifold Approximation and Projection)**

Algorithmus zur Dimensionsreduktion und Datenvisualisierung.

**Beispiel:** Visualisierung von Kundensegmenten im 2D-Raum.

**Unbalanced Data**

Datensätze, in denen Klassenverteilungen stark ungleich sind.

**Beispiel:** 95 % Nicht-Spam, 5 % Spam.

**Uncertainty**

Unsicherheit über den wahren Wert oder das Modell.

**Beispiel:** Prognose: Umsatz = 10 Mio ± 0,5 Mio.

**Underfitting**

Modell ist zu einfach und bildet die Datenstruktur nicht ausreichend ab.

**Beispiel:** Lineares Modell bei nichtlinearer Beziehung.

**Undersampling**

Technik zur Reduktion der Mehrheitklasse in unbalancierten Daten.

**Beispiel:** Reduktion der Nicht-Spam-Mails.

**Univariate Analysis**

Analyse einer einzelnen Variable.

**Beispiel:** Histogramm der Altersverteilung.

**Unit Test**

Automatisierter Test einzelner Funktionen oder Module.

**Beispiel:** `test_mean_function()` in Python.

**Unnormalized Data**

Daten ohne Skalierung oder Normalisierung.

**Beispiel:** Einkommen in Euro, Alter in Jahren, Gewicht in kg.

**Unstructured Data**

Daten ohne feste Struktur, oft Text, Bilder oder Audio.

**Beispiel:** E-Mails, PDFs, Chatverläufe.

**Unsupervised Learning**

ML-Methode ohne gelabelte Trainingsdaten.

**Beispiel:** Clustering-Algorithmus (z. B. K-Means).

**Update Anomaly**

Dateninkonsistenz durch redundante Speicherung ohne Normalisierung.

**Beispiel:** Änderung einer Adresse muss in mehreren Tabellen erfolgen.

**Upsampling**

Künstliches Vergrößern der Minderheitsklasse.

**Beispiel:** Kopieren von Spam-Mails zur Klassenbalance.

**Upper Bound**

Obere Grenze eines Konfidenzintervalls oder Parameters.

**Beispiel:** Vertrauensbereich 95 %: [10, 15] → Upper Bound = 15

**URI (Uniform Resource Identifier)**

Eindeutige Adresse zur Identifikation von Ressourcen im Web.

**Beispiel:** `https://api.server.com/data/123`

**URL Encoding**

Kodierung spezieller Zeichen in URLs.

**Beispiel:** Leerzeichen wird zu `%20`

**UUID (Universally Unique Identifier)**

128-bit-Wert zur eindeutigen Identifikation.

**Beispiel:** `550e8400-e29b-41d4-a716-446655440000`

**UX (User Experience)**

Gesamterfahrung eines Nutzers mit einem System.

**Beispiel:** Ladezeiten, Navigation und visuelle Gestaltung eines Dashboards.

**Utility Function**

Funktion zur Bewertung von Entscheidungen oder Ergebnissen.

**Beispiel:** Auswahl zwischen Modellen basierend auf Kosten/Nutzen.

**Validation Set**
Datensatz, der zur Bewertung eines Modells während der Trainingsphase verwendet wird.
**Beispiel:** Aufteilung eines Datensatzes in 70 % Training, 15 % Validierung, 15 % Test.

**Value at Risk (VaR)**

Statistische Kennzahl zur Quantifizierung des potenziellen Verlusts in einem bestimmten Zeitraum bei gegebenem Konfidenzniveau.

**Beispiel:** Ein Tages-VaR von 5 % bei 1 Mio. EUR = 50.000 EUR Verlust mit 95 % Wahrscheinlichkeit.

**Variance**

Maß für die Streuung der Daten um den Mittelwert.

**Beispiel:** Hohe Varianz bedeutet große Unterschiede zwischen den Werten.

**Variance Inflation Factor (VIF)**

Maß für Multikollinearität in Regressionsmodellen.

**Beispiel:** VIF > 10 weist auf starke Korrelation mit anderen Variablen hin.

**Variable**

Eine messbare Eigenschaft oder ein Merkmal, das analysiert wird.

**Beispiel:** Alter, Einkommen oder Klickrate.

**Variable Importance**

Maß für die Relevanz einer Variable in einem Modell.

**Beispiel:** Feature-Importances bei Random Forests.

**Variance Threshold**

Feature-Selection-Methode, die Merkmale mit geringer Varianz entfernt.

**Beispiel:** Filtert Spalten, deren Werte fast immer gleich sind.

**Vector**

Mathematische Struktur zur Darstellung von Datenpunkten in n-dimensionalem Raum.

**Beispiel:** [1.2, 3.4, 0.5] als Eingabe für ein neuronales Netz.

**Vectorization**

Umwandlung von Daten in numerische Vektoren, oft für maschinelles Lernen.

**Beispiel:** Text zu TF-IDF-Vektoren.

**Version Control**

System zur Verwaltung von Änderungen am Code oder an Daten.

**Beispiel:** Git mit Commit-Historie.

**Vertical Scaling**

Erhöhung der Ressourcen eines einzelnen Servers.

**Beispiel:** Mehr RAM oder CPU für eine Datenbankinstanz.

**Visualization**

Darstellung von Daten in visueller Form zur Erkennung von Mustern.

**Beispiel:** Balkendiagramm, Heatmap oder Boxplot.

**VLOOKUP**

Excel-Funktion zum vertikalen Suchen in Tabellen.

**Beispiel:** Suchen eines Produktnamens anhand der ID.

**Volatility**

Maß für die Schwankungsbreite von Zeitreihen oder Märkten.

**Beispiel:** Aktien mit hoher Volatilität haben stark schwankende Kurse.

**Voting Classifier**

Ensemble-Lernverfahren, bei dem mehrere Modelle abstimmen.

**Beispiel:** Mehrheit entscheidet über Klassifikation.

**VAE (Variational Autoencoder)**

Neuronales Netz zur Dimensionsreduktion und Generierung von Daten.

**Beispiel:** Bildkompression oder synthetische Datengenerierung.

**Variance Explained**

Anteil der Gesamtvarianz, der durch ein Modell oder eine Komponente erklärt wird.

**Beispiel:** 80 % erklärte Varianz in PCA.

**Virtual Join**

Verknüpfung zweier Tabellen in der Abfrage ohne physische Zusammenführung.

**Beispiel:** View in SQL.

**Volumetrische Daten**

3D-Daten, häufig in Medizin oder Geowissenschaften.

**Beispiel:** CT-Scans oder seismische Datenwürfel.

**Voice Recognition**

Technologie zur Umwandlung von gesprochener Sprache in Text.

**Beispiel:** Google Assistant erkennt Befehle per Sprache.

**Variance-Bias Tradeoff**

Grundlegendes Konzept im maschinellen Lernen: Balance zwischen Unter- und Überanpassung.

**Beispiel:** Komplexe Modelle riskieren Overfitting, einfache Underfitting.

**Vector Database**

Spezialisierte Datenbank für semantische Suchen mit Embeddings.

**Beispiel:** Verwenden von FAISS oder Pinecone zur Ähnlichkeitssuche.

**Video Analytics**

Automatisierte Analyse von Videodaten.

**Beispiel:** Erkennung von Objekten oder Bewegungen in Überwachungsvideos.

**Violin Plot**

Visualisierung, die Boxplot mit Dichteverteilung kombiniert.

**Beispiel:** Darstellung von Notenverteilungen nach Fach.

**Virtual Machine**

Virtuelles System mit eigener Betriebssysteminstanz.

**Beispiel:** Ubuntu-VM auf Windows zur Datenanalyse.

**View (SQL)**

Virtuelle Tabelle, basierend auf gespeicherten Abfragen.

**Beispiel:** `CREATE VIEW aktive_kunden AS SELECT * FROM kunden WHERE status='aktiv'`

**Von Neumann Architecture**

Rechenarchitektur mit gemeinsamem Speicher für Daten und Programme.

**Beispiel:** Grundlage moderner Computerdesigns.

**Vector Space Model**

Modell zur Darstellung von Texten oder Dokumenten als Vektoren.

**Beispiel:** TF-IDF-Vektoren in NLP.

**Visual Regression Testing**

Testmethode zum Erkennen von UI-Änderungen durch Bildvergleiche.

**Beispiel:** Differenzvergleich von Screenshots bei Webänderungen.

**Vulnerability Assessment**

Bewertung von Schwachstellen in IT-Systemen.

**Beispiel:** Scans auf unsichere Ports oder veraltete Software.

**WAAS (Workspace as a Service)**
Cloud-Dienstleistung, die eine komplette virtuelle Arbeitsumgebung bereitstellt.
**Beispiel:** Remote-Teams nutzen WAAS für sicheren Zugriff auf Daten und Software.

**Wahrscheinlichkeit**

Maß für die Erwartung, dass ein bestimmtes Ereignis eintritt.

**Beispiel:** Die Wahrscheinlichkeit, dass bei einem Münzwurf "Kopf" erscheint, ist 0{,}5.

**Wald-Test (Wald Statistic)**

Statistischer Test zur Signifikanzprüfung von Regressionskoeffizienten.

**Beispiel:** Einsatz bei Logit-Modellen zur Prüfung einzelner Einflussgrößen.

**Warehouse (Data Warehouse)**

Zentrale Datenbank zur Analyse und Berichterstattung über große Datenmengen.

**Beispiel:** Speicherung historischer Verkaufszahlen zur Analyse von Trends.

**Wasserfalldiagramm (Waterfall Chart)**

Grafik zur Darstellung kumulativer Veränderungen.

**Beispiel:** Gewinnentwicklung eines Unternehmens über mehrere Quartale.

**Wavelet Transformation**

Technik zur Analyse von Signalen oder Zeitreihen in verschiedenen Frequenzbereichen.

**Beispiel:** Kompression von Audio- oder Bilddaten.

**Web Scraping**

Automatisierte Extraktion von Daten aus Webseiten.

**Beispiel:** Preissammlung von Produkten aus einem Online-Shop.

**Weight Initialization**

Ausgangswertzuweisung für neuronale Netze, beeinflusst Trainingsverlauf.

**Beispiel:** He-Initialization für ReLU-Aktivierungen.

**Weighted Average**

Durchschnitt, bei dem unterschiedliche Werte unterschiedlich stark gewichtet werden.

**Beispiel:** Durchschnittsnote unter Berücksichtigung von Kreditpunkten.

**White Noise**

Zufallsrauschen mit konstanter spektraler Leistungsdichte.

**Beispiel:** Residuenmodellierung in Zeitreihenanalyse.

**Whitening**

Vorverarbeitung, bei der Korrelationen zwischen Features entfernt werden.

**Beispiel:** PCA Whitening vor Clustering.

**Whisker Plot (Boxplot)**

Grafische Darstellung von Median, Quartilen und Ausreißern.

**Beispiel:** Vergleich der Einkommensverteilung verschiedener Regionen.

**Wide Format**

Datenstruktur mit einer Spalte pro Variable und einer Zeile pro Beobachtungseinheit.

**Beispiel:** Pivotierte Tabelle mit Monatsumsätzen als Spalten.

**Wilcoxon-Test**

Nichtparametrischer Test für gepaarte Stichproben.

**Beispiel:** Vorher-Nachher-Vergleich von Trainingsdaten.

**Window Function**

SQL-Funktion zur Berechnung über Datenzeilen mit Bezug zur aktuellen Zeile.

**Beispiel:** Laufender Durchschnitt in einer Zeitreihe mit `OVER(PARTITION BY...)`.

**Winsorizing**

Technik zur Behandlung von Ausreißern durch Begrenzung extremer Werte.

**Beispiel:** Setzen aller Werte über dem 95. Perzentil auf genau dieses Perzentil.

**Wirtschaftskennzahl (Key Performance Indicator)**

Quantitative Messgröße zur Bewertung wirtschaftlicher Leistung.

**Beispiel:** Umsatzwachstum, EBITDA, Return on Investment.

**Word Embedding**

Darstellung von Wörtern als Vektoren im kontinuierlichen Raum.

**Beispiel:** Word2Vec, GloVe.

**Word Cloud**

Visualisierung, bei der Wörter je nach Häufigkeit unterschiedlich groß dargestellt werden.

**Beispiel:** Analyse dominanter Begriffe in Kundenbewertungen.

**Working Directory**

Aktuelles Verzeichnis, in dem ein Skript arbeitet oder Dateien speichert.

**Beispiel:** Pfad in Python über `os.getcwd()` auslesen.

**Workload**

Menge an Aufgaben oder Daten, die ein System oder Benutzer in einer Zeit verarbeiten muss.

**Beispiel:** Hohe CPU-Auslastung bei parallelem Datenimport.

**Wrapper Method**

Feature Selection durch wiederholte Modellierung mit unterschiedlichen Variablensätzen.

**Beispiel:** Rekursive Eliminierung unwichtiger Features bei Regression.

**Wurzel-MSE (Root Mean Squared Error)**

Fehlermetrik für Regressionsmodelle – Quadratwurzel des mittleren quadratischen Fehlers.

**Beispiel:** RMSE von 5{,}2 für ein Vorhersagemodell des Umsatzes.

**WYSIWYG (What You See Is What You Get)**

Oberflächenkonzept, bei dem das Ergebnis der Darstellung direkt der Ansicht entspricht.

**Beispiel:** Dashboard-Editoren mit Live-Vorschau.

**Wöchentliche Saisonalität**

Regelmäßiges Muster im Wochenrhythmus innerhalb einer Zeitreihe.

**Beispiel:** Höherer Traffic auf Webseiten am Montag und Freitag.

**X-Achse**
Horizontale Achse in einem Koordinatensystem oder Diagramm.
**Beispiel:** In einem Liniendiagramm stellt die X-Achse meist die Zeit dar.

**X-bar Chart**

Qualitätskontroll-Diagramm zur Überwachung von Mittelwerten in Prozessen.

**Beispiel:** Tägliche Mittelwertkontrolle der Produktmaße in einer Fertigung.

**XGBoost**

Hochleistungs-Boosting-Framework für maschinelles Lernen.

**Beispiel:** Wird zur Teilnahme an Kaggle-Wettbewerben verwendet.

**XML (eXtensible Markup Language)**

Textbasiertes Datenformat zur Darstellung hierarchisch strukturierter Daten.

**Beispiel:** Produktdaten als XML-Datei mit verschachtelten Tags.

**XPath**

Abfragesprache für die Navigation in XML-Dokumenten.

**Beispiel:** `//Produkt/Preis` extrahiert den Preis aus jedem Produkt-Tag.

**XOR (exclusive OR)**

Logische Operation mit wahr, wenn genau ein Operand wahr ist.

**Beispiel:** XOR(1, 0) = 1, XOR(1, 1) = 0.

**XSS (Cross-Site Scripting)**

Sicherheitslücke, bei der Angreifer Skripte in Webanwendungen einschleusen.

**Beispiel:** Ein manipuliertes Eingabefeld führt JavaScript aus.

**X-Intercept**

Der Punkt, an dem eine Linie die X-Achse schneidet.

**Beispiel:** Bei $f(x) = 2x - 4$ liegt der X-Intercept bei $x = 2$.

**X-Wert**

Unabhängige Variable oder Feature in einer Analyse.

**Beispiel:** In einem Modell zur Gehaltsvorhersage ist "Berufserfahrung" ein X-Wert.

**XOML**

XML-basiertes Format für Workflows in Microsoft-Technologien.

**Beispiel:** Workflows in alten .NET-Automatisierungsprojekten.

**X-Querverweis**

Verweis auf andere Inhalte oder Daten innerhalb einer Datenbank oder eines Berichts.

**Beispiel:** KPI-Bericht verlinkt auf zugehörige Rohdaten.

**X-Test**

Teilmenge der Daten, die für die Prüfung eines Modells reserviert wird.

**Beispiel:** X_train und X_test zur Trennung von Trainings- und Testdaten.

**X-Space**

Merkmalsraum für Eingabedaten in einem Modell.

**Beispiel:** Alle Features zusammen bilden den X-Space.

**XOR-Gate**

Elektronisches Gatter, das eine XOR-Funktion realisiert.

**Beispiel:** Wird in digitalen Schaltungen verwendet.

**XAI (Explainable AI)**

Techniken zur Erklärung von Entscheidungen von ML-Modellen.

**Beispiel:** SHAP-Werte für Entscheidungsbäume.

**Xref (Cross-reference)**

Bezugssystem zur Verknüpfung von Datenpunkten oder Dokumenten.

**Beispiel:** Tabellenblatt mit Zellverweisen auf andere Sheets.

**X-Modellierung**

Strukturierung von Prozessen mit parallelen und sequentiellen Abläufen.

**Beispiel:** X-förmige Prozessverzweigung in BPMN.

**XPL (eXtensible Processing Language)**

Programmiersprache für datengetriebene Workflows.

**Beispiel:** Verwendung zur Datenumwandlung und -verarbeitung in XML.

**X-Means Clustering**

Erweiterung von K-Means mit automatischer Bestimmung der Clusteranzahl.

**Beispiel:** Identifikation optimaler Clusterzahl bei Kundenklassifikation.

**X-Y-Diagramm**

Zweidimensionale Visualisierung numerischer Beziehungen.

**Beispiel:** Streudiagramm mit Gewicht (X) und Blutdruck (Y).

**X-R Chart**

Diagramm zur Überwachung von Mittelwert und Spannweite in der Qualitätssicherung.

**Beispiel:** Tägliche Kontrolle von Prozessabweichungen.

**X.509-Zertifikat**

Standard für die Struktur von digitalen Zertifikaten.

**Beispiel:** TLS/SSL-Zertifikate zur Website-Verschlüsselung basieren auf X.509.

**X-Pipeline**

Datentransformations- oder Modellpipeline mit mehreren Schritten.

**Beispiel:** Daten-Cleaning → Feature Engineering → Modellierung.

**X-Feature**

Einzelnes Eingabemerkmal im ML-Kontext.

**Beispiel:** Alter ist ein X-Feature bei einer Vorhersage.

**X-Variablen**

Sammelbegriff für unabhängige Variablen in statistischen Modellen.

**Beispiel:** In der linearen Regression y = ax + b ist x die X-Variable.

**X-Strategie**

Abstrakter Begriff für explorative Vorgehensweise in Datenanalysen.

**Beispiel:** Unstrukturierte Datenanalyse mit offenen Hypothesen.

**X-Header**

Zusätzliche HTTP-Header zur Informationsübertragung im Web.

**Beispiel:** `X-Requested-With: XMLHttpRequest` zur Identifikation von AJAX-Aufrufen.

**X-Formular**

Strukturierte Datenmaske zur Eingabe oder Anzeige von Daten.

**Beispiel:** Formular mit Drop-downs, Checkboxes und Eingabefeldern.

**X.25**

Früher Standard für paketvermittelte Netzwerke.

**Beispiel:** Wurde in Banken und Kreditkartennetzen genutzt.

**Y-axis**
Die vertikale Achse in einem Diagramm oder Koordinatensystem. Sie stellt meist abhängige Variablen wie gemessene Werte oder

Ergebnisse dar.
**Beispiel:** In einem Liniendiagramm zeigt die Y-Achse den Umsatz pro Monat.

**YAML (YAML Ain't Markup Language)**

Ein menschenlesbares Datenformat zur Konfiguration und zum Datenaustausch, häufig in DevOps und ML-Projekten genutzt. YAML ist einfacher strukturiert als JSON, unterstützt aber komplexe Datenhierarchien.

**Beispiel:** Definition von Trainingsparametern für ein ML-Modell.

**Yarn (Hadoop YARN)**

„Yet Another Resource Negotiator" – ein Framework zur Ressourcenverwaltung im Hadoop-Ökosystem. Es ermöglicht das Ausführen verteilter Datenverarbeitungsanwendungen.

**Beispiel:** YARN verwaltet Ressourcen für MapReduce-Jobs in einem Hadoop-Cluster.

**Yeo-Johnson Transformation**

Eine Transformation zur Normalisierung von Daten, ähnlich der Box-Cox-Transformation, aber auch für negative Werte geeignet. Verbessert die Linearisierung und die Modellierbarkeit von Variablen.

**Beispiel:** Anwendung auf stark schiefe Daten wie Nettovermögen.

**Yield Curve**

Grafische Darstellung von Zinssätzen über verschiedene Laufzeiten hinweg. Eine inverse Zinsstrukturkurve kann auf wirtschaftliche Rezessionen hinweisen.

**Beispiel:** Analyse der Renditekurve zur Konjunkturprognose.

**YOLO (You Only Look Once)**

Ein Echtzeit-Objekterkennungsalgorithmus im Bereich Deep Learning. YOLO erkennt Objekte in Bildern oder Videos mit hoher Geschwindigkeit.

**Beispiel:** Einsatz in Überwachungssystemen zur Erkennung von Personen.

**YTD (Year-to-Date)**

Kennzahl, die den Zeitraum vom Jahresbeginn bis zum aktuellen Datum beschreibt. Häufig verwendet in Finanzanalysen zur Performancebewertung.

**Beispiel:** YTD-Umsatz = Gesamtumsatz vom 1. Januar bis heute.

**Yule-Simon-Verteilung**

Eine Wahrscheinlichkeitsverteilung, die bei der Analyse von „Long Tail"-Phänomenen wie der Häufigkeit seltener Ereignisse nützlich ist.

**Beispiel:** Modellierung der Wortverteilung in einem Textkorpus.

**Yule's Q-Koeffizient**

Ein Maß für die Assoziation zweier binärer Variablen. Q = (ad - bc)/(ad + bc), basierend auf einer 2×2-Kontingenztafel.

**Beispiel:** Untersuchung der Beziehung zwischen zwei Ja/Nein-Antworten.

**Y-Funktion**

Oberbegriff für eine mathematische Funktion, die von einer unabhängigen X-Variable abhängt. In der Statistik meist das zu erklärende Merkmal.

**Beispiel:** Y = 3X + 2 ist eine lineare Funktion.

**Yield**

Allgemein der Ertrag oder das Ergebnis einer Operation oder eines Invests. In Programmierung auch ein Schlüsselwort in Python zur Erzeugung von Generatoren.

**Beispiel:** Python-Funktion mit `yield` erzeugt Lazy-Sequenzen.

**Y-Achsenbeschriftung**

Die Beschreibungseinheit oder Beschriftung auf der Y-Achse eines Diagramms. Sie vermittelt die Bedeutung der angezeigten Werte.

**Beispiel:** „Einnahmen in EUR“ als Achsenbeschriftung.

**Y-Splitter**

In der Datenpipeline ein Mechanismus zur Verzweigung von Datenströmen basierend auf Bedingungen.

**Beispiel:** Weiterleitung von Daten mit „Status = Fehler“ an eine separate Pipeline.

**Yellowbrick**

Python-Toolkit zur Visualisierung und Analyse von ML-Modellen. Ergänzt scikit-learn mit visuellen Diagnosewerkzeugen.

**Beispiel:** Visualisierung der Modellgüte mittels ROC-Kurve mit Yellowbrick.

**Yield Spread**

Differenz zwischen den Renditen zweier Anleihen. Ein Maß für Risiko und Investorenvertrauen.

**Beispiel:** Höherer Spread = höheres wahrgenommenes Risiko.

**Y-Codierung (One-Hot Encoding Zielvariable)**

Verfahren zur Kodierung von Zielvariablen mit mehreren Klassen für ML-Modelle.

**Beispiel:** Zielvariable „Farbe“ mit Klassen „rot“, „blau“, „grün“ wird zu [1,0,0], [0,1,0], [0,0,1].

**Youden-Index**

Kennzahl zur Optimierung des Schwellenwerts in binären Klassifikationsmodellen.

**Beispiel:** Maximiere Sensitivität + Spezifität - 1 zur Auswahl des besten Cutoffs.

**Y-Randomization**

Validierungstechnik zur Überprüfung von Overfitting in ML-Modellen. Zielvariable wird zufällig permutiert und Modell erneut trainiert.

**Beispiel:** Deutlich schlechtere Leistung nach Y-Randomization spricht gegen Overfitting.

**Yield Forecasting**

Vorhersage von Erträgen oder Produktionsausbringung, z. B. in der Landwirtschaft oder Fertigung.

**Beispiel:** ML-Modell zur Prognose der Maisernte auf Basis von Wetterdaten.

**Yield Management**

Dynamische Preisgestaltung zur Optimierung von Auslastung und Ertrag, z. B. im Flugverkehr oder Hotelwesen.

**Beispiel:** Preisanpassung je nach Buchungszeitpunkt und Nachfrage.

**YTD-Analyse**

Analyse der kumulierten Entwicklungen seit Jahresbeginn. Nützlich zur Bewertung saisonaler Trends.

**Beispiel:** Vergleich der YTD-Performance verschiedener Geschäftsbereiche.

**Z-Score**
Standardisierte Kennzahl, die angibt, wie viele Standardabweichungen ein Wert vom Mittelwert entfernt ist.
**Beispiel:** Z = (Wert - Mittelwert) / Standardabweichung.

**Z-Test**

Statistischer Test zur Überprüfung von Hypothesen bei bekanntem Populationsstandardabweichung.

**Beispiel:** Test auf Mittelwertunterschied mit bekanntem σ.

**Zero-Inflated Model**

Modelltyp für Daten mit überproportional vielen Nullen, z. B. in Zählvariablen.

**Beispiel:** Unfallstatistik mit vielen Nullmeldungen.

**Zero-Shot Learning**

ML-Ansatz, bei dem ein Modell Aufgaben lösen kann, ohne für diese Beispiele gesehen zu haben.

**Beispiel:** Textklassifikation mit rein beschreibenden Klassenlabels.

**Zero Trust Architecture**

Sicherheitskonzept in IT, bei dem kein Nutzer oder Gerät automatisch vertraut wird.

**Beispiel:** Zugriffskontrollen bei jeder API-Anfrage.

**Zero-Based Budgeting**

Planungsmethode, bei der alle Ausgaben von Grund auf neu begründet werden müssen.

**Beispiel:** Jeder Kostenposten wird jährlich neu validiert.

**Zentrale Tendenz**

Lageparameter wie Mittelwert, Median oder Modus, die die Verteilung beschreiben.

**Beispiel:** Median-Einkommen in einer Region.

**Zeitreihe (Time Series)**

Daten, die zeitlich geordnet sind, oft mit gleichmäßigen Abständen.

**Beispiel:** Tägliche Aktienkurse.

**Zeitreihenanalyse**

Analyseverfahren zur Modellierung zeitabhängiger Daten.

**Beispiel:** ARIMA-Modell zur Prognose von Verkaufszahlen.

**Zeitverzögerung (Lag)**

Verzögerter Einfluss einer Variable in einer Zeitreihe.

**Beispiel:** Umsatz heute hängt vom Wetter gestern ab (Lag-1).

**Zeitfensteranalyse (Rolling Window)**

Analyse innerhalb eines sich bewegenden Zeitraums.

**Beispiel:** Gleitender Durchschnitt der letzten 30 Tage.

**Zielgröße (Target Variable)**

Die vorherzusagende oder zu erklärende Variable in einem Modell.

**Beispiel:** Preis eines Hauses in einer Regression.

**Zielwert (Label)**

Klassifikation oder Regressionswert, mit dem das Modell trainiert wird.

**Beispiel:** "Spam" oder "Nicht-Spam" als Label.

**Zufallsfehler (Random Error)**

Nicht systematische Abweichung vom wahren Wert.

**Beispiel:** Messfehler durch Rauschen.

**Zufallsforest (Random Forest)**

Ensemble-ML-Verfahren aus Entscheidungsbäumen zur Klassifikation oder Regression.

**Beispiel:** Klassifikation von Kundenabwanderung.

**Zufallsvariable**

Variable, deren Wert von einem Zufallsprozess abhängt.

**Beispiel:** Augenzahl beim Würfeln.

**Zufallszahlengenerator**

Algorithmus zur Erzeugung scheinbar zufälliger Werte.

**Beispiel:** `random()` in Python.

**Zufallsstichprobe**

Stichprobe, bei der jede Einheit der Grundgesamtheit die gleiche Auswahlwahrscheinlichkeit hat.

**Beispiel:** Zufällig gezogene Teilnehmer für eine Umfrage.

**Zugrundeliegende Verteilung (Underlying Distribution)**

Die angenommene oder beobachtete Verteilung, auf der eine Analyse basiert.

**Beispiel:** Normalverteilung bei IQ-Werten.

**Zugriffsrechte (Access Rights)**

Regeln, welche Nutzer welche Daten lesen, schreiben oder verändern dürfen.

**Beispiel:** Nur Admins dürfen Benutzerdaten löschen.

**Zugriffszeit (Access Time)**

Zeitspanne zwischen Anfrage und Erhalt von Daten.

**Beispiel:** Zugriff auf SQL-Datenbank dauert 15 ms.

**Zugriffslog**

Protokoll über Datenbank- oder Systemzugriffe.

**Beispiel:** Logdateien bei Webservern.

**Zuverlässigkeit (Reliability)**

Maß für die Konsistenz oder Reproduzierbarkeit von Messungen oder Modellen.

**Beispiel:** Ein Test liefert bei Wiederholung ähnliche Ergebnisse.

**Zuvor bekannte Labels (Known Labels)**

In Supervised Learning bekannte Zielvariablen im Trainingsdatensatz.

**Beispiel:** E-Mail-Daten mit bekannter Spam-Klassifikation.

**Z-Transformation**

Standardisierung einer Variable durch Subtraktion des Mittelwerts und Division durch die Standardabweichung.

**Beispiel:** Anwendung vor dem Trainieren eines linearen Modells.

**Zoomable Chart**

Interaktive Visualisierung mit Zoom-Funktion.

**Beispiel:** Zoombares Liniendiagramm mit D3.js.

**Zensierte Daten (Censored Data)**

Beobachtungen, bei denen nur ein Teilwert bekannt ist. Häufig in Überlebenszeitanalysen.

**Beispiel:** Patientenstudie endet vor dem Todeszeitpunkt.

**Zielgruppenanalyse**

Identifikation und Beschreibung von Nutzergruppen für gezielte Ansprache.

**Beispiel:** Analyse von Website-Besuchern nach Alter und Herkunft.

**Zonierung (Zoning)**

Segmentierung eines geografischen oder logischen Bereichs zur Analyse oder Steuerung.

**Beispiel:** Unterteilung einer Stadt in Cluster für Verkehrsanalysen.

**Zuverlässigkeitsschätzung**

Statistische Abschätzung, wie konsistent ein Verfahren bei Wiederholung ist.

**Beispiel:** Cronbachs Alpha in der Psychometrie.

**Zweistichproben-Test**

Statistischer Vergleich zweier Gruppen hinsichtlich Mittelwert oder Verteilung.

**Beispiel:** T-Test zwischen Kontroll- und Behandlungsgruppe.

**Zweidimensionale Normalverteilung**

Verteilung zweier korrelierter metrischer Variablen.

**Beispiel:** Größe und Gewicht.

**Zyklische Komponente**

Langfristige, wiederkehrende Schwankung in Zeitreihen.

**Beispiel:** Konjunkturzyklen in Wirtschaftsdaten.

**Zweistufiges Modell (Two-Stage Model)**

Modellierungsansatz in zwei Schritten, oft bei Endogenität.

**Beispiel:** 2SLS in ökonometrischen Modellen.

**Zentralwert (Median)**

Der mittlere Wert einer geordneten Verteilung.

**Beispiel:** Bei [3, 5, 7] ist 5 der Median.

**Zielkonflikt**

Widersprüchliche Anforderungen in einem Optimierungskontext.

**Beispiel:** Kosten senken vs. Qualität sichern.

**Zufallsprozesse**

Modelle zur Beschreibung stochastischer Abläufe.

**Beispiel:** Markov-Kette.

**Zellreferenz (Excel)**

Bezug auf eine Zelle oder Zellbereich in einer Tabelle.

**Beispiel:** `=A1 + B2`.

**Zugriffspfad (Access Path)**

Pfad, über den ein Datenbanksystem Daten aus Tabellen liest.

**Beispiel:** Indexzugriff vs. Full Table Scan.

**Zielerreichungsgrad**

Messgröße zur Bewertung, wie nahe eine Maßnahme dem gesetzten Ziel kommt.

**Beispiel:** Erfüllung von KPIs zu 90 %.

**Zahlendarstellung (Number Format)**

Darstellung von Zahlenwerten in Computern oder Tabellen.

**Beispiel:** Dezimal-, Prozent- oder Währungsformat.

**Zeitstempel (Timestamp)**

Zeitmarkierung für Ereignisse oder Datenpunkte.

**Beispiel:** 2025-05-28 13:32:00 als Logeintrag.

**Zugangskontrolle (Access Control)**

Mechanismus zur Einschränkung des Zugriffs auf Daten oder Systeme.

**Beispiel:** Rollenbasierte Zugriffskontrolle (RBAC).