Data Science (Mids) - Jaish Khan

Introduction to Data Science

Data Science → Data science involves problem-solving through the utilization of data. The primary objective of data science is to extract knowledge from data. It combines techniques from various disciplines and employs scientific principles for data analysis.

- **Process of Data Science**
    1. *Data Collection and Understanding*
    2. *Data Cleaning and Formatting*
    3. *Data Analysis*
    4. *Problem Identification*
    5. *Solution Development Using Data*

Data Analysis → Data analysis aims to uncover useful information from data using techniques like statistics and algorithms.

Data Scientists → They are not strictly computer scientists, mathematicians, statisticians, or domain experts but possess a blend of skills from these areas. A good data scientist has proficiency in data analysis and machine learning.

- **Skills**
    - *Quantitative Skills*: Math, algorithms, statistics.
    - *Technical Skills*: Programming, infrastructures.
    - *Skeptical Mindset*: Formulate hypotheses but maintain a critical perspective.
    - *Collaborative Abilities*: Teamwork, communication skills.
- **Types**
    - *Polymath*: All-rounders with expertise in all areas.
    - *Data Evangelist*: Focus on data analysis, insights dissemination, and action.
    - *Data Preparer*: Specialize in querying and preparing data for analysis.
    - *Data Shaper*: Skilled in both analyzing and preparing data.
    - *Data Analyzer*: Primarily focused on data analysis.
    - *Platform Builder*: Responsible for data collection and infrastructure development.
    - *Moonlighter*: Part-time data scientists who contribute in their spare time.
    - *Insight Actor*: Utilize the outcomes of data science and take action based on insights.

Business Intelligence: According to Gartner's IT Glossary, business intelligence involves best practices that facilitate access to and analysis of information to enhance decision-making and performance.

Feature	Business Intelligence	Data Science
Depth of Insights	Provides a summary of past data.	Offers deeper insights and predicts future trends.
Time Focus	Concentrates on past events.	Focuses on present and future trends
Techniques	Uses dashboards, alerts, and queries.	Employs optimization, predictive modeling, and forecasting.
Data Types	Primarily uses structured data from data warehouses.	Works with diverse data types, often unstructured.
Common Questions	Addresses questions like "What happened?" and "When did...?"	Explores questions like "What if...?" and "How can we...?"

> **Big Data** → Big data consists of high-volume, high-velocity, and/or high-variety information assets that necessitate cost-effective and innovative information processing methods.

The Three Vs of Big Data

- **Volume**: The scale of data needs to be "big," although there's no strict definition. Gartner suggests it should demand innovative forms of information processing.
- **Velocity**: Represents the speed at which new data is generated and the speed at which data must be processed and analyzed. Real-time processing is often required.
- **Variety**: Encompasses the diversity of data types and sources, including:
    1. *Structured Data*: Data with defined types and structure, such as comma-separated values.
    2. *Semi-Structured Data*: Textual data with a parseable pattern, like XML files with a schema.
    3. *Quasi-Structured Data*: Textual data with irregular formats that can be formatted with effort, like clickstream data.
    4. *Unstructured Data*: Data without inherent structure and often in multiple formats, such as websites and videos.

? What to do with Data?

1. Aggregation and Statistics: Data warehousing and OLAP (Online Analytical Processing).
2. Indexing, Searching, and Querying: Keyword-based search and pattern matching (XML/RDF).
3. Knowledge Discovery: Data mining and statistical modeling.

- **Real-Life Examples of Data Science** → Understanding Consumer Behavior and Political Campaigns.

Machine Learning

Machine learning involves the creation and utilization of models that learn from data. Key components include:
* *Data*: Any measurable or recordable information.
* *Model*: A representation of the mathematical relationship between variables.
* *Evaluation*: Assessing the effectiveness of the model.

- **Types of Machine Learning**
  - **Regression**: Predicting continuous values (e.g., predicting house prices).
  - **Classification**: Categorizing data into predefined classes (e.g., classifying emails as spam or not spam).
  - **Clustering**: Grouping data points based on similarities without predefined labels (e.g., customer segmentation based on purchasing behavior).
- **Machine Learning Workflow**
  - The workflow consists of three phases:
    - **Training Phase**: The model learns patterns from the training data.
    - **Test Phase**: The trained model is applied to unseen test data to make predictions.
    - **Evaluation Phase**: The model's performance is assessed by comparing predictions to actual values using metrics like accuracy, precision, or error measures.

> **Traditional CS vs. Machine Learning**
>
> In traditional computer science, programs are written to process data and produce output. In machine learning, the program learns from the data to generate output, essentially "programming itself.

- **Data Structures**
  - **Set**: Used for storing unique, unordered data without duplicates.
  - **Array**: Suitable for storing large, homogeneous, numerical data efficiently.
  - **List**: A general-purpose structure for ordered data that can contain duplicates.

Regression

A supervised learning technique for analyzing and predicting relationships between variables. It aims to understand how changes in independent variables affect the dependent variable.

- *Dependent Variable*: The outcome variable being predicted. (Target)
- *Independent Variables*: The predictor variables that influence the dependent variable. (Features)
- *Coefficients*: Quantify the relationship between independent and dependent variables.
- *Error Term* (ϵ): Represents the unexplained variation in the model.

Types of Regression

1. **Linear Regression** → Models the relationship between variables using a straight line.
2. *Simple Linear Regression*: Involves one independent variable.
3. *Multiple Linear Regression*: Involves multiple independent variables.
4. **Logistic Regression** → Used for predicting categorical outcomes, often binary (yes/no).
5. **Polynomial Regression** → Extends linear regression to model non-linear relationships using polynomial terms.
6. **Ridge and Lasso Regression** → Regularized linear regression methods that prevent overfitting.
7. *Ridge Regression*: Penalizes large coefficients.
8. *Lasso Regression*: Can shrink coefficients to zero, helping with feature selection.
9. **Support Vector Regression (SVR)** → Adapts Support Vector Machines for regression tasks, finding an optimal hyperplane to fit data.
10. **Decision Tree and Random Forest Regression** → Tree-based methods for regression.
11. *Decision Trees*: Creates a tree-like model for predictions.
12. *Random Forest*: An ensemble of decision trees that improves prediction robustness.

Key Metrics for Regression Evaluation

1. **R-squared (R^2)**: Measures the proportion of variance in the dependent variable explained by the model. Higher values (closer to 1) indicate a better fit.
2. **Mean Squared Error (MSE)**: The average of squared differences between actual and predicted values. Lower MSE indicates better performance.
3. **Root Mean Squared Error (RMSE)**: The square root of MSE, providing error in the same units as the dependent variable. Sensitive to large errors due to squaring.
4. **Mean Absolute Error (MAE)**: The average of absolute differences between actual and predicted values, offering an interpretable error measure. Less sensitive to outliers, providing a more robust error measure.

> Applications of Regression → Predicting Prices, Risk Assessment, Medical Research, Marketing etc.

- **Training**: Finding the best-fitting line that minimizes the error between predicted and actual house prices.
- **Prediction**: Using the trained model (equation of the line) to predict prices for new houses based on their features.
- **Model Complexity and Evaluation**: More complex models (e.g., higher-order polynomials) may fit the training data perfectly but can lead to overfitting, where the model performs poorly on unseen data. Evaluation metrics like RMSE and MAE are used to assess model performance on test data, helping to choose the best model that generalizes well to new data.
- **Training Error vs. Test Error**
    - Training error → the error on the data used to train the model.
    - Test error → the error on a separate dataset not used for training, which reflects the model's ability to generalize.

Classification

- **Definition**: Classification is a supervised learning technique used to categorize data into predefined classes.
- **Objective**: Predict the class or category of new data based on patterns learned from labeled data.

Types of Classification

- **Binary Classification**: Involves two classes (e.g., YES/NO, MALE/FEMALE, SPAM/NOT SPAM).
- **Multi-Class Classification**: Involves more than two classes (e.g., handwritten digit recognition (0-9)).
- **Multi-Label Classification**: Assigns multiple labels to each instance (e.g., tagging a social media post with multiple categories).

Classification Algorithms

1. **Logistic Regression** → A linear model for binary classification.
    - **Example**: Predicting whether a patient has a disease based on features like age and weight.
2. **Decision Trees** → A tree-like structure where Internal nodes represent decision rules and Leaf nodes represent outcomes.
    - **Example**: Classifying if a customer will churn based on their service usage.
    - **Concepts**:
        - *Predictors/Attributes*: Features used to make decisions.

        - *Target/Class*: The outcome to be predicted.
        - *Tree Split*: Dividing data based on attribute values.
    - **Impurity Measures**:
        - *Entropy*: Measures the disorder or randomness in a set.
        - *Gini Impurity*: Another measure of impurity.
        - *Misclassification Error*: The rate of incorrect classifications.
        - Splits aim to reduce impurity in child nodes, making them more homogeneous.
    - **Rule Induction** → Extracting rules from decision trees to represent classification logic.
3. **k-Nearest Neighbors (kNN)** → Classifies data points based on the majority class of their k-nearest neighbors.
    - **Example**: Predicting tumor type (malignant/benign) based on similar cases.
    - **Measures of Proximity**:
        - *Distance*: Commonly used metric (e.g., Euclidean distance).
        - *Correlation Similarity*: Measures the linear relationship between data points.
        - *Simple Matching Coefficient*: Counts the number of matching attributes.
        - *Jaccard Similarity*: Measures similarity based on the ratio of shared attributes to the total number of attributes.
        - *Cosine Similarity*: Measures similarity based on the angle between data point vectors.
4. **Naive Bayes** → A probabilistic classifier based on Bayes' theorem, assuming feature independence.
    - **Example**: Classifying document categories (text classification).
    - **Concepts**:
        - *Class Conditional Probability*: Probability of observing features given a class.
        - *Posterior Probability*: Probability of a class given the observed features.
        - *Prior Probability*: Initial probability of a class before observing features.
    - **Issues**:
        - *Incomplete Training Set*: Addressed using Laplace correction.
        - *Continuous Numeric Attributes*: Handle using probability density functions.
        - *Attribute Independence Assumption*: May need to remove correlated attributes.
5. **Neural Networks** → Inspired by the human brain, neural networks consist of interconnected nodes (neurons) that process and transmit information. They can learn complex patterns from data.
6. **Support Vector Machines (SVM)** → Finds the optimal hyperplane to separate classes in a high-dimensional space.

    - **Example**: Classifying emails as work, personal, or spam.
    - **Concepts**:
        - *Boundary*: The decision boundary separating classes.
        - *Margin*: The distance between the boundary and the closest data points.
        - *Kernel Trick*: Transforms data to higher dimensions to improve separability.

Key Metrics for Classification Evaluation:

1. **Precision**: Measures the accuracy of positive predictions.
2. **Recall**: Measures the ability to identify all positive instances.
3. **F1-Score**: A harmonic mean of precision and recall, providing a balanced performance measure.
4. **Confusion Matrix**: A table that summarizes classification results, showing true positives, true negatives, false positives, and false negatives.

> Applications of Classification → Medical Diagnosis, Customer Segmentation, Fraud Detection, Natural Language Processing (NLP), Image Recognition.

Clustering

An unsupervised learning technique that organizes data into groups (clusters) based on similarity. The goal is to achieve high similarity within clusters and low similarity between clusters.

- **Key Difference from Classification**: Clustering discovers labels directly from data without predefined classes or labeled examples.
- **Concept of Similarity**
    - Similarity measures the strength of the relationship between data items, indicating how alike they are.
    - Clustering relies on similarity measures to group similar data objects together.
- **Distance Measures**
    - Distance measures are often used to quantify similarity.
    - **Euclidean Distance** → Straight-line distance between points.
    - **Manhattan Distance** → Distance calculated as the sum of absolute differences along each dimension.

Types of Clustering

- **Partitional Algorithms**: Create partitions and evaluate them based on a criterion.
- **Hierarchical Algorithms**: Build a hierarchical decomposition of objects based on a criterion.

Clustering Algorithms

K-Means Algorithm, A popular partitional clustering algorithm.

1. Decide on the number of clusters, *k*.
2. Initialize *k* cluster centers (randomly or using a more informed approach).
3. Assign objects to their nearest cluster center.
4. Recalculate cluster centers based on the current object assignments.
5. Repeat steps 3 and 4 until convergence (no objects change clusters).

- **Strengths**:
  - Relatively efficient.
  - Easy to implement.
- **Weaknesses**:
  - The concept of "mean" needs to be defined, making it unsuitable for categorical data.
  - It's a heuristic (approximation) that can get stuck in local optima.
  - The number of clusters (*k*) needs to be specified in advance.
  - Sensitive to noisy data and outliers.
  - Tends to find spherical clusters and may not perform well with other cluster shapes.
- **Elbow Method** → Use an objective function (e.g., within-cluster sum of squares) to evaluate clustering for different *k* values. Plot the objective function against *k* and look for an "elbow" point, which suggests a suitable number of clusters.

> Applications of Clustering → EDA (Exploratory Data Analysis), Color Compression, Finding Distribution Centers, Making Recommendations.

Clustering Evaluation → A good partitioning algorithm aims to minimize an objective function that quantifies the quality of the clustering.

Data Mining Process

CRISP-DM Process (Cross-Industry Standard Process for Data Mining) is a widely used framework for data mining projects.

- **Stages of CRISP-DM**:
  - *Business Understanding*: Defining the project objectives and requirements from a business perspective.
  - *Data Understanding*: Collecting, exploring, and assessing the quality of data.
  - *Data Preparation*: Cleaning, transforming, and preparing data for modeling.

    - *Modeling*: Selecting and applying data mining algorithms to build models.
    - *Evaluation*: Assessing the performance and validity of the models.
    - *Deployment*: Implementing the models and integrating them into business processes.
- **Five Stages of Data Mining**
    - **Prior Knowledge**: Acquiring background information about the problem's objective, subject area, and available data.
    - **Data Preparation**: This stage involves:
        - Data exploration to understand its characteristics.
        - Data quality assessment and handling of missing values.
        - Data type conversion and transformation as needed.
        - Identification and handling of outliers.
        - Feature selection to choose relevant variables.
        - Data sampling if dealing with very large datasets.
    - **Modeling**:
        - Building a model using selected algorithms and training data.
        - Splitting the data into training and test sets.
        - Training the model on the training data and evaluating it on the test data.
        - Selecting the best-performing model based on evaluation results.
    - **Application**:
        - Ensuring the model is ready for deployment in a production environment.
        - Integrating the model into technical systems.
        - Monitoring the model's response time and performance.
        - Considering the need for remodeling as new data becomes available.
        - Facilitating the assimilation of insights from the model into business processes.
    - **Knowledge**:
        - Gaining posterior knowledge and insights derived from the data mining process.

Exploratory Data Analysis (EDA)

Sources of Data

- Internal Sources (Business-centric data or Scientific experimental data)
- External Sources (Public government databases or Stock market data)

Collected Data → Data gathered specifically for the project.
Online Data → Data obtained from APIs (e.g., Google Maps, Facebook, Twitter) or Web

scraping is used to extract data from websites.

Variables and Data Types → Different types of variables and data types need to be considered during EDA.

Common Data Issues

1. *Missing values*: Need to be handled appropriately, either by imputation or other methods.
2. *Wrong values*: Require detection and correction to ensure data accuracy.
3. *Messy format/representation*: Data may need restructuring or cleaning for analysis.

> Causes of messiness → Variables stored in both rows and columns, multiple features in a single column, etc.

Data Pre-processing

- The goal is to prepare data for analysis and machine learning, including:
  1. Data parsing and formatting
  2. Data profiling to assess data quality and quantity
  3. Data cleaning
  4. Data engineering tasks like outlier detection, feature engineering, and data augmentation.
- **Population vs. Sample Data**
  - *Population*: The entire set of objects or events under study.
  - *Sample*: A representative subset of the population.
  - Samples are often used due to the impracticality of collecting or analyzing the entire population data.
- **Techniques**:
  - **Exploring Individual Variables**:
    - Summary statistics (mean, median, mode)
    - Measures of spread (range, variance, standard deviation)
    - Distribution analysis (histograms)
  - **Assessing Interactions Between Variables**:
    - Correlation analysis
    - Analysis of Variance (ANOVA)
  - **Multidimensional Data Exploration**:
    - Clustering
    - Dimensionality reduction techniques (e.g., Principal Component Analysis - PCA)

Statistical Terms

Summary Statistics → The choice between mean and median depends on data distribution and the presence of outliers. The mean is sensitive to outliers.

- *Mean*: The average value.
- *Median*: The middle value when data is sorted.
- *Mode*: The most frequent value.

Measures of Spread → Understanding data spread is crucial for assessing data variability and potential outliers.

- *Range*: The difference between the maximum and minimum values.
- *Variance*: Measures how data points are spread out from the mean.
- *Standard Deviation*: The square root of the variance.

Data Visualization

Data visualization complements summary statistics and provides a visual representation of data patterns and relationships.

Types of Data Visualization

1. **Distribution Visualization**: Shows how a variable is distributed over a range of values (e.g., histograms).
2. **Relationship Visualization**: Depicts the relationship between two or more variables (e.g., scatter plots).
3. **Comparison Visualization**: Compares trends in different variables or datasets (e.g., multiple histograms, box plots).
4. **Composition Visualization**: Illustrates the breakdown of a dataset into subgroups (e.g., pie charts, stacked area graphs).

Visualization Techniques

1. **Histogram**: Used for visualizing the distribution of a single variable.
2. **Scatter Plot**: Shows the relationship between two variables.
3. **Box Plot**: Useful for comparing distributions, showing quartiles, median, outliers, and comparing a variable across groups.
4. **Pie Chart**: Depicts the composition of a dataset, showing the proportion of each category.
5. **Stacked Area Graph**: Visualizes trends in composition over time.

> Color-coding can be used in visualizations to represent categorical variables.

- **Multidimensional Visualization**
  - 3D visualizations can be used to explore relationships between three variables, but they are not always effective.