Case Studies in Data Science

Additional Learning Resources – Chris Santosh John

RMIT UNIVERSITY

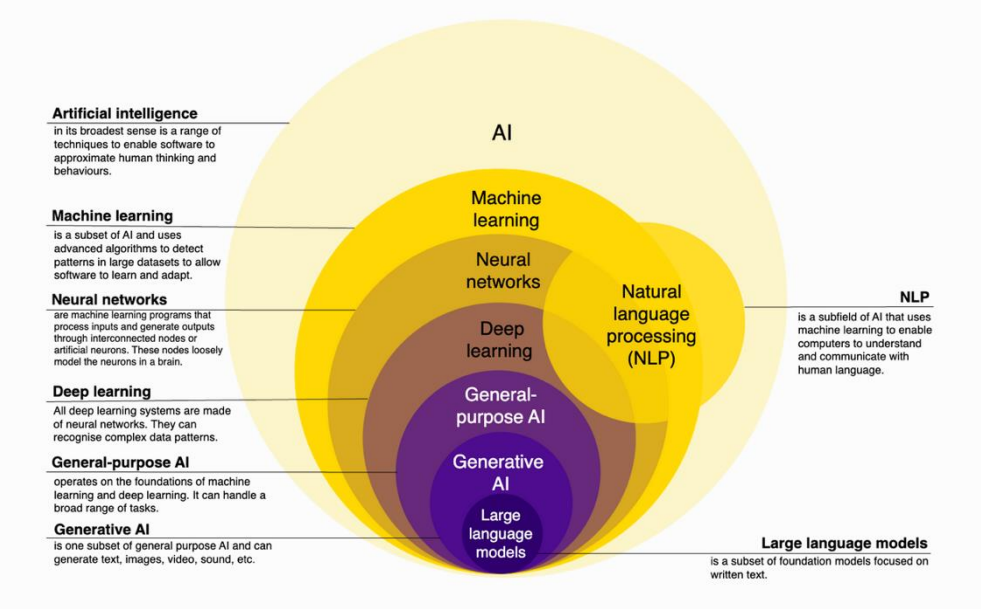
Purpose

Resourcefulness is a vital soft skill that you should focus on developing alongside your projects. It involves the ability to solve problems creatively and overcome challenges by **effectively using resources** at your disposal.

While the resource links provided in this document **offer a good starting point, they are not exhaustive**. You are encouraged to proactively seek out additional resources, tools, and information tailored to the specific needs of your project.

RMIT UNIVERSITY

ML vs AI vs AGI



AI subsets (diagram by Yang et al. [1])

- **Machine Learning (ML)** is a subset of artificial intelligence that involves the use of algorithms and statistical models to enable computers to learn from and make predictions or decisions based on data. **Common Types include** Supervised and Unsupervised Learning.
- **Artificial Intelligence (AI)** is a **broad field** of computer science focused on creating systems capable of performing tasks that typically require human intelligence.
- **Artificial General Intelligence (AGI)** refers to a level of AI where machines possess the ability to understand, learn, and apply intelligence across a **wide array of tasks at a level comparable to human beings.**

RMIT UNIVERSITY

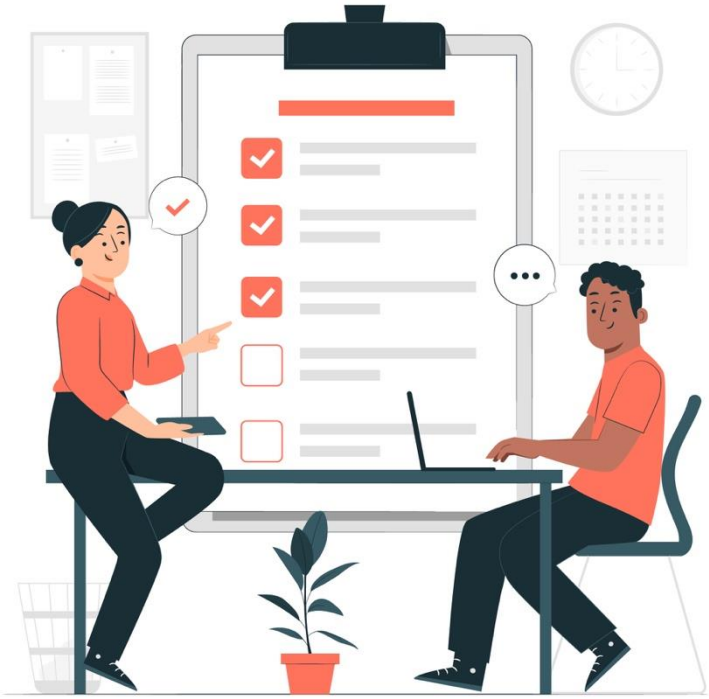
ML Workflow

Every ML project revolves around the data the machine intends to analyze and model. The success of the project depends on a structured workflow, where each step seamlessly integrates into the next. Here are the common steps in a typical ML workflow:

Machine Learning Guide for Beginners - Link [2]

1. **Data Analysis**
2. **Data Preparation/Pre-processing**
3. **Model Implementation**
4. **Model Evaluation**

***Note*:** *Each step of the process funnels into the next, making it essential to complete each stage with the highest quality to ensure the success of the ML project.*



RMIT UNIVERSITY

Data Analysis

Exploratory Data Analysis (**EDA**) is a critical step in the machine learning workflow that involves examining and visualizing data to uncover patterns, anomalies, and insights.

Some common elements in the data analysis phase include:

- **Understanding Data Distribution**
- **Identifying Data Quality Issues**
- **Feature Selection and Engineering**
- **Detecting Patterns through visualizations**

Resources:

1. Performing basic EDA - Link [3]
2. Extensive Data Analysis Playlist – Link [10]

RMIT UNIVERSITY

Data Preparation/Pre-processing

Data preparation/preprocessing are essential stages in the machine learning workflow. They involve **transforming raw data into a clean and usable format**, which is crucial for building effective machine learning models. This includes:

- **Data Cleaning -** Link [4]
- **Data Transformation**
  - **Scaling -** Link [5]
  - **Encoding -** Link [6]
  - **Feature Engineering -** Link [7]
- **Dimensionality Reduction –** Link [8]

Note: *The order and necessity of these steps depend on the specific problem being addressed and the characteristics of the data.* ***This indicates the need for thorough data analysis to determine the appropriate pre-processing techniques.***

RMIT UNIVERSITY

Model Implementation & Evaluation

Model implementation involves selecting and training a machine learning algorithm to solve a specific problem.

Model evaluation is the process of assessing how well the trained model performs on unseen data.

Choosing the right machine learning algorithm depends on the problem type and data characteristics.

Choosing ML Models - Link [9]

References

[1] Yang, F., Goldenfein, J., & Nickels, K. (2024). GenAI Concepts. Melbourne: ARC Centre of Excellence for Automated Decision-Making and Society RMIT University, and OVIC. DOI: 10.60836/psmc-rv23

[2] S. Ray, "Top 10 Machine Learning Algorithms to Use in 2024," *Analytics Vidhya*, Sep. 08, 2017. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/?utm_source=reading_list&utm_medium=https://www.analyticsvidhya.com/blog/2021/08/how-to-perform-exploratory-data-analysis-a-guide-for-beginners/ (accessed Aug. 08, 2024).

[3] N. Vanawat, "How To Perform Exploratory Data Analysis -A Guide for Beginners," *Analytics Vidhya*, Aug. 12, 2021. https://www.analyticsvidhya.com/blog/2021/08/how-to-perform-exploratory-data-analysis-a-guide-for-beginners/ (accessed Aug. 08, 2024).

[4] Upwork, "6 Data Cleaning Steps for Preparing Your Data ," *Upwork.com*, Nov. 14, 2022. https://www.upwork.com/resources/data-cleaning-steps (accessed Aug. 08, 2024).

RMIT UNIVERSITY

References

[5] B. McShane, "An Extensive Guide to Scaling as a Method of Data-Preprocessing," *Medium*, Jan. 02, 2024. https://medium.com/@broganmcshane.code/an-extensive-guide-to-scaling-as-a-method-of-data-preprocessing-be88b8f861f0 (accessed Aug. 08, 2024).

[6] GeeksforGeeks, "One Hot Encoding in Machine Learning," *GeeksforGeeks*, Jun. 12, 2019. https://www.geeksforgeeks.org/ml-one-hot-encoding/ (accessed Aug. 08, 2024).

[7] G. Sharma, "Complete Guide to Feature Engineering: Zero to Hero," *Analytics Vidhya*, Sep. 21, 2021. https://www.analyticsvidhya.com/blog/2021/09/complete-guide-to-feature-engineering-zero-to-hero/ (accessed Aug. 08, 2024).

[8] J. Frost, "Principal Component Analysis Guide & Example," *https://statisticsbyjim.com/*. https://statisticsbyjim.com/basics/principal-component-analysis/ (accessed Aug. 08, 2024).

[9] Y. Kinha, "An easy guide to choose the right Machine Learning algorithm," *KDnuggets*, Feb. 17, 2022. https://www.kdnuggets.com/2020/05/guide-choose-right-machine-learning-algorithm.html (accessed Aug. 08, 2024).

RMIT UNIVERSITY

References

[10] M. Keith, “Learn Exploratory Data Analysis (EDA) in Python”, YouTube Playlist, Jul 25, 2023, https://www.youtube.com/playlist?list=PLe9UEU4oeAuV7RtCbL76hca5ELO_IELk4 (accessed Aug. 08, 2024).