INTRODUCTION TO MACHINE LEARNING

What is Machine Learning?

Technical term:

A subset of AI that focuses on the development of algorithms and models that allow computer systems to learn and make predictions or decisions without being explicitly programmed.

Layman's term:

Is like teaching computers to learn from examples, kind of like how we learn from our experiences. Instead of giving computers specific rules to follow, we give them lots of examples and let them figure things out on their own.

Importance of Machine Learning

- Ability to solve complex problems (manpower)
- Make predictions
- Adapt to changing circumstances

Types of Machine Learning

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Supervised Learning

- Monitored
- Similar to teaching a child

Classification and Regression

Classification

- Sorting things using keywords
- Straight to the point

Regression

- Automation with a pattern
- Predicting numbers using data

Best Examples of Supervised Learning

- Social Media
- Microsoft Products
- E-Commerce Sites

Unsupervised Learning

The algorithm is trained on a dataset without explicit supervision or labeled output.

Primary Goals of Unsupervised Learning

- **Clustering**

  Grouping similar data points together into clusters or categories without prior knowledge of what those categories should be.

- **Dimensionality Reduction**

  Reducing the number of features or variables in the data while preserving the essential information and relationships.

- **Anomaly Detection**

  Identifying rare or unusual data points that do not conform to the expected patterns in the dataset.

Tools or Techniques used to achieve these goals

- **K-Means Clustering**

  is an **iterative algorithm** that partitions a dataset into K distinct, non-overlapping subsets (clusters), with each cluster represented by its centroid. It **divides the data into clusters.**

  *Just as K-Means organizes balloons into clusters based on their colors, **the algorithm groups data points into clusters based on their similarities**. It iteratively adjusts cluster positions to minimize the differences within each cluster, aiming for a neat and organized arrangement.*

- **Principal Component Analysis (PCA)**

  a **linear transformation technique** that reorients the axes of a high-dimensional dataset to create a new coordinate system where the first axis (principal component) corresponds to the direction of maximum variance in the data.

  *Just as PCA helps you adjust your view to concentrate on the main action at a sports event, it allows you to **reduce the complexity of data while retaining the most critical information**, making it easier to analyze and understand.*

- **t-Distributed Stochastic Neighbor Embedding (t-SNE)**

  a **nonlinear dimensionality reduction technique** that seeks to map high-dimensional data to a lower-dimensional space while preserving the relative similarities between data points.

  *Just as t-SNE helps arrange puzzle pieces on the board so that similar piece's cluster together, it **reorganizes complex data in a way that reveals patterns and groups,** making it simpler to understand and visualize the relationships between data points.*

Machine Learning: Workflow

Determine which phases are used during a machine learning project. Describes the steps of a machine learning implementation.

Goal of an ML Workflow

1. **Problem Solving**
   Use ML to solve real-world problems, make data-driven decisions, or gain valuable insights
2. **Interpretability and Ethical Considerations**
   Come up with novel solutions that meet the customer's demands/needs
3. **Model Development**
   Create accurate and effective ML models that capture patterns and relationships in the data
4. **Optimization and Development**
   Production
   def. Gathering real-time data after the system deployment

   Testing
   def. Deploying the system that is still on testing phase

6 Recommendations when Creating a ML Workflow

1. **Define Clear Objectives and Problem Statement**
   - Clearly define the problem or the question
   - Set specific and measurable objectives
2. **Data Quality and Preprocessing**
   - Gather quality and reliable data
   - Handle missing data and anomalies appropriately
   - Create the best model
3. **Split Data Properly**
   - Split dataset into training (70-80%), validation (10-15%), and testing (10-15%)
   - Ensure the data splits are representative of the overall dataset especially with imbalanced classes
4. **Select Suitable Algorithms and Models**
   - Chose ML algorithms and models that are well-suited to the nature of the problem
   - Experiment with multiple algorithms to find the best one
5. **Hyperparameter Tuning and Cross-Validation**
   - Fine-tune the hyperparameters of your models using techniques like grid search or random search.
   - Implement cross-validation to assess model performance more robustly and to avoid overfitting.
6. **Documentation and Version Control**
   - Keep thorough documentation of your workflow (data preprocessing steps, model selection criteria, and hyperparameter tuning results).

- Use version control systems to track changes in your code and experiment configurations. This helps in reproducibility.

PYTHON

Definition:

A widely-used programming language in the field of machine learning and data science due to its simplicity and a rich ecosystem of libraries.

Rich Ecosystem of Libraries:

- **NumPy** - Essential for efficient numerical operations and array handling.
- **Pandas** - Simplifies data manipulation with versatile DataFrames.
- **Scikit-Learn** - Offers a broad set of tools for machine learning tasks.
- **Matplotlib** - Comprehensive data visualization library for creating various plots and charts.
- **Seaborn** - Enhances data visualization, particularly for statistical data.
- **TensorFlow** - A powerful deep learning framework for neural network development.
- **PyTorch** - A flexible deep learning framework known for its ease of use and adaptability.
- **Jupyter Notebooks** - Enables interactive code and visualization within shareable documents.

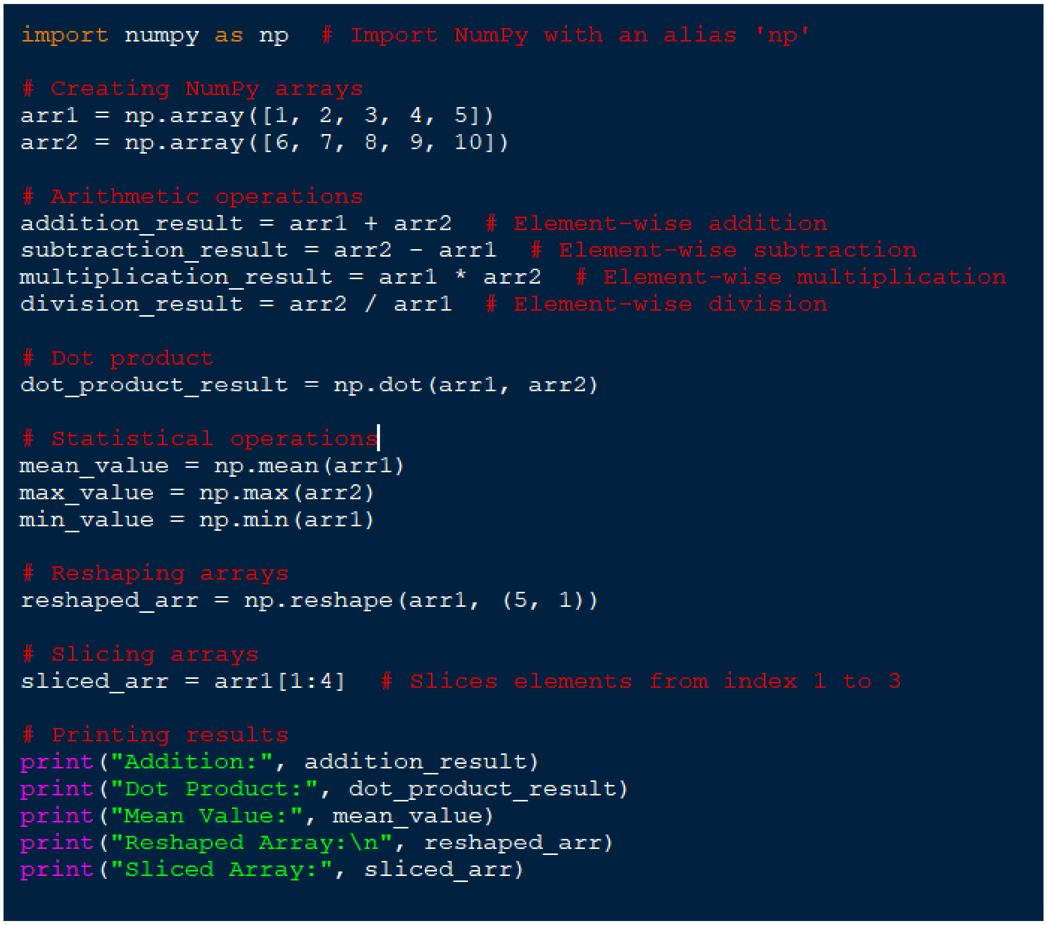
NumPy

- Short for "Numerical Python,"
- A fundamental Python library for numerical operations
- Provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays efficiently.
- A Swiss Army knife for working with numbers in computer programming.

Widely used in scientific computing, data analysis, and machine learning for the following **purposes**:

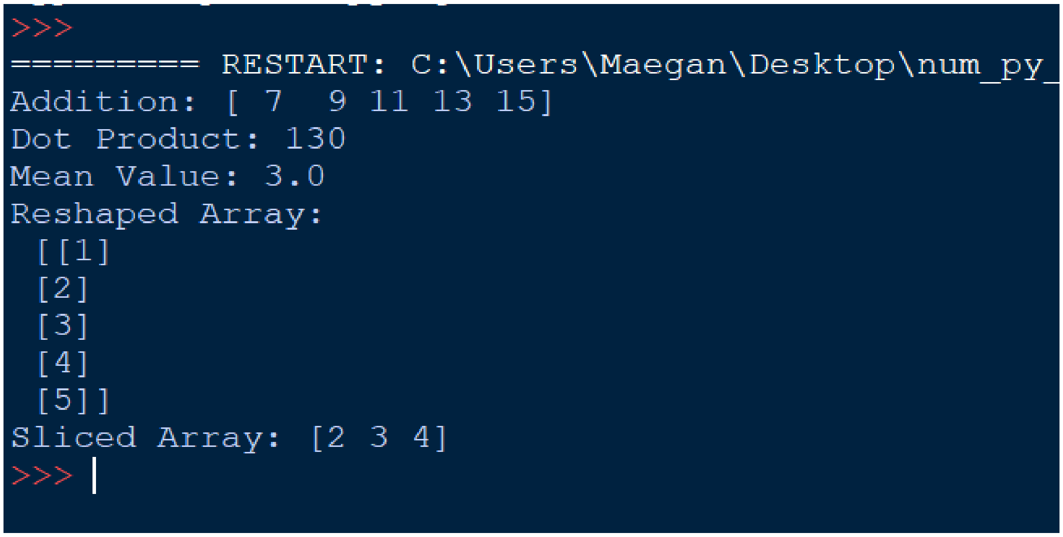
- Efficient storage and manipulation of large datasets.
- Performing mathematical and logical operations on arrays.
- Integration with other data science libraries for data preprocessing and analysis.

Example:



```
import numpy as np  # Import NumPy with an alias 'np'

# Creating NumPy arrays
arr1 = np.array([1, 2, 3, 4, 5])
arr2 = np.array([6, 7, 8, 9, 10])

# Arithmetic operations
addition_result = arr1 + arr2  # Element-wise addition
subtraction_result = arr2 - arr1  # Element-wise subtraction
multiplication_result = arr1 * arr2  # Element-wise multiplication
division_result = arr2 / arr1  # Element-wise division

# Dot product
dot_product_result = np.dot(arr1, arr2)

# Statistical operations
mean_value = np.mean(arr1)
max_value = np.max(arr2)
min_value = np.min(arr1)

# Reshaping arrays
reshaped_arr = np.reshape(arr1, (5, 1))

# Slicing arrays
sliced_arr = arr1[1:4]  # Slices elements from index 1 to 3

# Printing results
print("Addition:", addition_result)
print("Dot Product:", dot_product_result)
print("Mean Value:", mean_value)
print("Reshaped Array:\n", reshaped_arr)
print("Sliced Array:", sliced_arr)
```

Output:



```
>>>
========= RESTART: C:\Users\Maegan\Desktop\num_py_
Addition: [ 7  9 11 13 15]
Dot Product: 130
Mean Value: 3.0
Reshaped Array:
 [[1]
 [2]
 [3]
 [4]
 [5]]
Sliced Array: [2 3 4]
>>>
```