Data Analysis for AI/ML

Jonathan Atiene

Techcrity V5 Mentorship

OUTLINE

1. Intro to AI/ML
2. Setting up your python environments
3. Introduction to libraries
4. Exploratory data analysis
5. Understanding data distribution
6. Code session
7. Questions
8. Assignments

Learning Outcomes

By the end of this module, you will be able to:

- Understand the fundamentals of AI/ML.
- Work with Python for data analysis.
- Perform exploratory data analysis (EDA).
- Clean and preprocess datasets effectively.

What is Artificial Intelligence (AI)?

- Definition of AI
- Applications of AI in real life
- Subfields of AI (ML, NLP, Computer Vision, Robotics)

What is Machine Learning (ML)?

- ML is a subset of AI that enables machines to learn from data.
- Supervised, Unsupervised, and Reinforcement Learning.
- Spam detection, image recognition, recommendation systems.

SETTING UP PYTHON ENV.

python -m venv ./.venv

source .venv/bin/activate

Intro to jupyter notebook

AI/ML libraries

- Pandas
- Numpy
- Matplotlib
- Seaborn
- Scikit learn

Data Exploration

Steps:

- Load dataset
- Handle missing values
- Generate descriptive statistics
- Visualize data

DATA VISUALIZATION

Visualizing Data

- **Pie Chart:** Survived vs Not Survived to show proportions.
- **Histogram:** Age distribution to observe skewness.
- **Boxplot:** Fare distribution to detect outliers.
- **Bar Plot:** Pclass vs Survival Rate to compare survival rates across passenger classes.
- **Count Plot:** Embarked vs Survived to identify variations across embarkation points.
- **Heatmap:** Correlation between Pclass, Fare, and Survived to analyze feature importance.

DATA PROCESSING

Steps:

- Encode Ordinal variables
  - Label Encoding: Assigning numbers to categories .
- Encode categorical variables
  - One-Hot Encoding: Creating binary columns.
- Normalize numerical data
- Save the cleaned dataset

Assignment

1. Remove the names of the titanic column
2. Implement Label Encoding
3. Implement some visualization with reasons
4. Save an encoded dataset.