

# Data Science and Machine Learning Course Syllabus

## 1. Data Science and Machine Learning Theory



- Statistics as the Foundation
- Distributions
- What is Data Science?
- Types of Data
- What is Machine Learning?
- How Machine Learning is just fancy statistics?
- Show an example using stats (mean, std, histogram, mode)
- Explain that ML models are probability distributions

## 2. Types of Machine Learning



- Unsupervised Learning
- Supervised Learning
- Semi-Supervised Learning
- Reinforcement Learning

## 3. Python Basics



- Data Types
- For Loops
- If Conditions

- Imports and Packages
- Directory Navigation (os Module)
- Data Structures
  - Lists and List Comprehension
  - Dictionaries and Dictionary Comprehension
  - Sets
  - NumPy Arrays and Operations
  - Indexing
- Functions and Return Statement
- Utility Functions
  - Lambda Function
  - Map()
  - Filter()
  - Len()
  - Sorted() and list.sort()
  - Reversed()
  - Enumerate()
  - Zip()
  - Datetime Module
  - Random Module
  - Math Module
  - Pickle

## 4. Pandas



- What is a DataFrame?
- Data Readers
- Pandas Functions
  - Df.head()
  - Df.tail()

- Df.info()
- Df.describe()
- Df.shape
- Df.columns
- Df.loc[]
- Df.iloc[]
- Df.groupby()
- Df.sort\_values()
- Df.drop()
- Df.dropna()
- Df.fillna()
- Export to Files
- Plotting DataFrames

## 5. Visualization with Python



- Types of Plots
  - Bar Plot
  - Line Plot
  - Scatter Plot
  - Histograms
  - Distributions
  - Box Plot
- Libraries
  - Matplotlib
  - Plotly
  - Seaborn
  - Altair
- Finding Patterns in Plots Visualization is a straightforward tool to spot patterns in data

## 6. Machine Learning



- Why We Need It? How It's Just Fancy Statistics?
- Building on the Material Delivered Earlier, Start with ML Models
  - Unsupervised
    - K-means Clustering
    - Hierarchical Clustering
    - DBSCAN
  - Supervised
    - Regression
      - MSE (Mean Squared Error)
      - MAE (Mean Absolute Error)
    - Classification
      - Logistic Regression
      - Decision Trees
      - Random Forest
      - K-nearest Neighbors
      - Support Vector Machines
  - Loss Functions
    - Regression
    - Classification
      - Binary Cross-Entropy Loss
  - Metrics
    - Regression
    - Classification
  - Clustering

- Silhouette Score
- Calinski-Harabasz Index

## 7. Feature Engineering



- Feature Crossing: Describe feature engineering techniques, starting with feature crossing to create new features by combining existing ones.
- Dimensionality Reduction: Explain dimensionality reduction techniques like PCA (Principal Component Analysis) and when to use them.
- Feature Importance and Selection: Discuss methods for assessing feature importance and selecting the most relevant features for modeling.

## 8. ML and Data Processing Pipelines



- Collection: Explain data collection methods, both manual and automated, and the importance of clean data.
- Cleaning: Discuss data cleaning techniques to handle missing values, outliers, and errors.
- Preprocessing: Cover preprocessing steps such as normalization and scaling.
- Train-Test Split: Explain the importance of splitting data into training and testing sets for model evaluation.
- Model Training and Evaluation: Describe the process of model training, hyperparameter tuning, and model evaluation using the metrics mentioned earlier.
- Deployment: Highlight the deployment phase, where models are put into production for real-world use.

## 9. Deep Learning



- Feed Forward Networks: Introduce neural networks, gradient descent, backpropagation, activation functions, and hyperparameters.
- RNNs and CNNs: Explain recurrent neural networks (RNNs) for text sequences, tokenization, embeddings, and convolutional neural networks

(CNNs) for image analysis.

## 10. Applications



- Tabular Example: Provide an example of solving a tabular data problem using machine learning.
- Sentiment Analysis: Explain sentiment analysis as an application of NLP (Natural Language Processing).
- Image Classification: Discuss image classification using deep learning models like CNNs.

## 11. Tools



- Keras and TensorFlow: Introduce Keras and TensorFlow for building and training deep learning models.

## 12. Transfer Learning



- Keras Applications and Pretrained Models: Explain how to leverage pretrained deep learning models for various tasks using Keras applications.

## 13. Power BI, Tableau, PySpark, SQL



- Power BI and Tableau: Describe these tools as powerful data visualization and business intelligence platforms. Explain how to create interactive dashboards and reports.
- PySpark: Discuss PySpark for big data processing, including data manipulation and machine learning on large datasets.
- SQL: Emphasize SQL skills for data retrieval and manipulation from relational databases.