Data Science Notes

# 1. What is Data Science?

* Definition: Data Science is an interdisciplinary field that uses scientific methods, algorithms, and systems to extract insights and knowledge from structured and unstructured data.
* Key Elements: Statistics & Mathematics, Programming (Python, R, SQL, etc.), Data Engineering (ETL, pipelines, storage), Machine Learning & AI, Domain Knowledge.
* Applications: Predictive analytics, Recommendation systems, Fraud detection, Healthcare diagnostics, Social media sentiment analysis.

# 2. Data Science Activities

* Problem Definition → Understand business problem & define objectives.
* Data Collection → Gather data from multiple sources (databases, APIs, logs, IoT devices, social media).
* Data Exploration → Initial analysis to understand patterns, anomalies, relationships.
* Data Cleaning & Preparation → Handle missing values, outliers, duplicates, and format inconsistencies.
* Data Analysis & Feature Engineering → Apply statistical methods, derive new features.
* Model Building → Train machine learning models (regression, classification, clustering, etc.).
* Model Evaluation → Test accuracy, precision, recall, F1 score, etc.
* Deployment → Integrate models into production systems (APIs, dashboards).
* Monitoring & Maintenance → Track performance, retrain as data evolves.

# 3. Data Exploration

* Definition: The initial step of analyzing data sets visually and statistically to summarize their main characteristics.
* Techniques: Descriptive statistics, Visualizations, Correlation analysis, Identifying outliers and missing values.
* Tools: Pandas, NumPy, Matplotlib, Seaborn, Tableau, Power BI.

# 4. Data Statistics & Visualization

* Data Statistics:
* - Descriptive Statistics: Mean, median, mode, standard deviation, skewness, kurtosis.
* - Inferential Statistics: Hypothesis testing, confidence intervals, p-values.
* - Correlation & Regression: Relationships between variables.
* Visualization:
* - Helps to communicate insights effectively.
* - Types: Distribution plots, Relationship plots, Comparison plots, Trend analysis.
* - Best Practices: Use appropriate chart type, avoid clutter, use color coding wisely.

# 5. Data Hygiene

* Definition: Practices to keep data clean, consistent, and reliable.
* Common Issues: Missing values, Duplicate records, Outliers, Inconsistent formats.
* Techniques: Data validation at entry, Regular audits, Handling missing values, Normalization & standardization, Proper metadata documentation.

# 6. Unstructured Data Storage & Analysis

* Unstructured Data: Data that doesn’t follow a fixed schema (e.g., text, images, videos, social media posts, emails).
* Storage Options: NoSQL Databases (MongoDB, Cassandra), Data Lakes (Hadoop HDFS, AWS S3, Azure Data Lake), Search & Indexing Systems (Elasticsearch, Solr).
* Analysis Techniques: Text (NLP, sentiment analysis, topic modeling), Image/Video (Computer Vision, deep learning), Audio (Speech recognition, sound classification), Big Data Tools (Spark, Hadoop, Kafka).
* Challenges: Scalability, Data quality & preprocessing, Storage costs & retrieval speed.

**Step 1: Sample Unstructured Data**

import pandas as pd

# Sample customer reviews (unstructured text data)

data = {

"ReviewID": [1, 2, 3, 4, 5],

"ReviewText": [

"The product quality is excellent and delivery was fast!",

"Terrible customer service. I will not buy again.",

"Average experience, product was okay but packaging was bad.",

"Absolutely love it! Great value for money.",

"The item arrived broken and support was unhelpful."

]

}

df = pd.DataFrame(data)

print(df)

**Step 2: Store Data (like NoSQL )**

# Save unstructured data (text reviews) into a CSV file

df.to\_csv("customer\_reviews.csv", index=False)

**Step 3: Analyze Text Data (NLP Basics)**

We’ll do:

* **Tokenization**
* **Sentiment Analysis** (positive/negative)

from textblob import TextBlob

# Sentiment analysis

df["Sentiment"] = df["ReviewText"].apply(lambda x: TextBlob(x).sentiment.polarity)

# Categorize sentiment

df["SentimentLabel"] = df["Sentiment"].apply(lambda x: "Positive" if x > 0 else ("Negative" if x < 0 else "Neutral"))

print(df[["ReviewText", "Sentiment", "SentimentLabel"]])

**Step 4: Visualization of Results**

import seaborn as sns

import matplotlib.pyplot as plt

sns.countplot(x="SentimentLabel", data=df, palette="viridis")

plt.title("Sentiment Distribution of Customer Reviews")

plt.show()