DR. GANESH ACADEMY

Email: drganeshacademy@gmail.com

End-to-End Data Science Syllabus – Dr. Ganesh Academy

1. Python for Data Science (Foundation)

A. Core Python Programming
- Data types, Variables, I/O
- Functions, Scoping, Lambda, Map/Filter
- Control Structures, Loops
- Exception Handling

B. Data Structures
- Lists, Tuples, Dictionaries, Sets
- Comprehensions and Generators

C. Working with Files & APIs
- Reading and writing text, CSV, JSON
- Consuming REST APIs using `requests`

D. Essential Libraries
- NumPy: arrays, slicing, broadcasting, vectorized operations
- Pandas: Series, DataFrames, indexing, grouping, pivoting, reshaping

2. Data Preprocessing & Cleaning

A. Data Cleaning
- Handling missing values
- Removing duplicates
- Dealing with outliers
- String manipulation

B. Data Transformation
- Feature scaling (StandardScaler, MinMaxScaler)
- Encoding categorical variables (OneHot, Label Encoding)
- Discretization, Binning
- Log transformations
- Feature extraction (date parts, text features)

C. Data Integration
- Merging, joining datasets
- Concatenation and aggregation

3. Exploratory Data Analysis (EDA)

- Summary statistics
- Correlation matrix
- Histograms, Box plots, Pair plots
- Group-wise analysis
- Visual EDA with Seaborn and Plotly

4. Statistics & Probability

A. Descriptive Statistics
- Mean, Median, Mode
- Variance, Standard Deviation

B. Inferential Statistics
- Population vs Sample
- Hypothesis Testing
- t-test, ANOVA, Chi-Square
- Confidence Intervals

C. Probability
- Bayes Theorem
- Distributions: Normal, Binomial, Poisson
- Central Limit Theorem

5. Machine Learning (Detailed)

A. Supervised Learning

- Regression: Linear, Polynomial, Ridge, Lasso
- Classification: Logistic, kNN, Decision Trees, Random Forest, SVM

B. Model Evaluation
- Accuracy, Precision, Recall, F1-score
- Confusion Matrix, ROC-AUC
- Cross-validation
- Bias-Variance Trade-off

C. Unsupervised Learning
- Clustering: k-Means, DBSCAN, Hierarchical
- Dimensionality Reduction: PCA, t-SNE

D. Model Deployment
- Saving models using Pickle/Joblib
- Loading and using in production

6. SQL for Data Science

A. Basics
- SELECT, WHERE, ORDER BY
- DISTINCT, LIMIT, ALIAS

B. Intermediate SQL
- JOINs (INNER, LEFT, RIGHT, FULL)
- GROUP BY, HAVING
- Subqueries, CTEs

C. Advanced SQL
- Window functions (ROW_NUMBER, RANK, LAG, LEAD)
- Aggregations
- Nested queries
- Date/Time functions

D. Practice
- Real-world datasets (Sales, HR, E-commerce)

7. PySpark for Big Data

A. PySpark Basics

- Setting up SparkSession
- RDDs vs DataFrames

B. DataFrame API
- Read/Write CSV, JSON, Parquet
- Filter, Select, GroupBy, Join

C. PySpark SQL
- Creating temporary views
- Writing SQL queries on Spark DataFrames

D. MLlib (Spark ML)
- VectorAssembler
- Feature Transformers
- Building classification/regression models

8. Visualization & Dashboards

A. Basic Visualization
- Matplotlib: line, bar, pie charts
- Seaborn: distribution plots, heatmaps
- Plotly: interactive plots (scatter, 3D, maps)

B. Streamlit (Rapid Prototyping)
- st.write, st.dataframe, st.plotly_chart
- Forms, sliders, buttons, inputs
- Upload CSV, run predictions
- Use with trained ML models
- Deploy to Streamlit Cloud

C. Dash by Plotly
- Layout and callbacks
- Input/Output linking
- Advanced components: DataTables, Graphs
- Real-time updates
- Deployment with Flask/Gunicorn

9. Capstone Projects

Beginner Level

- EDA on Titanic Dataset
- MovieLens recommender system
- Netflix Revenue Dashboard (SQL + Plotly)

Intermediate Level
- Customer churn prediction with Streamlit app
- Product recommendation engine using PySpark
- Sales forecasting using Time Series

Advanced Level
- MLOps with FastAPI + ML + Streamlit + Docker
- Real-time dashboards using Kafka + Spark Streaming + Dash
- End-to-end NLP pipeline (TF-IDF, classification, API)

10. Optional Advanced Tracks

- Deep Learning (TensorFlow/PyTorch)
- NLP (Spacy, Transformers)
- Time Series Forecasting (ARIMA, Prophet)
- MLOps (Docker, GitHub Actions, MLflow)
- Cloud (GCP BigQuery, AWS SageMaker)
- CI/CD for Data Science Projects

Tools & Platforms
- IDEs: Jupyter, VSCode, Colab
- Version Control: Git, GitHub
- Data Platforms: Kaggle, UCI ML Repo
- Task Automation: Makefile, pip-tools, Conda
- Deployment: Streamlit Cloud, Render, Hugging Face Spaces