Comprehensive Roadmap to Become a Data Scientist in 2025

What does Data Scientist do?



Analyze data to find useful insights



Build ML models for predictions



Leverage AI for complex challenges

Learn to Read Data

Start with **Python fundamentals**
- Data types
- Functions

Python Libraries

NumPy pandas

Learn to visualize data with

matplotlib seaborn

Move to **Data Cleaning** & **Preprocessing**

70% of your time will be spent cleaning messy data

Learn **SQL**
- SELECT and JOIN to retrieve data from databases.

Create **Dashboards** using BI Tools

Power BI tableau

Learn **Cloud Computing**
- SELECT and JOIN to retrieve data from databases.

EC2 amazon S3

Learn **Statistics**
- Probability Distributions & Hypothesis Testing

Get familiar with using **GenAI tools**

ChatGPT Claude

Month One

Month Two

Prediction and Forecasting

Learn **Machine Learning Algorithms**

Start with **Supervised ML Algorithms**
- Linear Regression
- Logistic Regression
- Decision Trees
- Random Forests

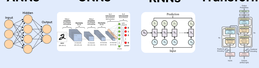
Master **Scikit Learn** for building **ML Models**

Learn **Unsupervised ML Algorithms**
- K-Means Clustering
- DBSCAN

Learn **Dimensionality Reduction**
- Principal Component Analysis (PCA)

Deep Dive into **Deep Learning Concepts**

ANNs **CNNs** **RNNs** **Transformers**



Learn **Frameworks**

PyTorch Hugging Face

to access some pre-trained models

Participate in Challenges/Create Projects

kaggle

Test Your Skills on the Ultimate AI Playground

DataHack Platform by Analytics Vidhya

experiment with datasets from platforms like

This practical exposure ensures you build a solid foundation for tackling real-world problems.

Month Three

Month Four

Model Deployment & Monitoring

It's time to learn how to make them useful in real-world applications

Begin with **Containerization** using

docker kubernetes

These tools allow you to deploy models efficiently in different environments.

Explore tools like **MLFlow** to **Track Experiments** and **Model Versions**

mlflow

Learn how to use **Monitoring tools** like

Prometheus Grafana

help detect performance issues, making you proficient in **maintaining production-ready systems**

To provide real-time or batch predictions Learn to create **REST APIs**

Flask FastAPI

Showcase deployment-ready projects on

GitHub

Month Five

Month Six

Get a Data Science Internship

Begin applying for "**Data Science Intern**" or "**Machine Learning Intern**" through

LinkedIn

indeed

Internships expose you to complexities:
- Missing Values
- Data Leakage
- Ambiguous Problem Definitions

These roles often involve working with messy, real-world data—a crucial skillset for any **Data Scientist**.

You'll collaborate with domain experts to prioritize tasks that matter most. This hands-on exposure solidifies your technical skills while building soft skills such as:

Teamwork Communication Time Management

Month Seven

Month Eight

Pick a Specialization

By now, you've gained foundational knowledge and practical experience.

It's time to narrow your focus and build expertise in a specialized area of

Data Science
- **Natural Language Processing(NLP)** Understanding of text-based models
- **Computer Vision(CV)** Focus on real-world applications like object detection

Natural Language Processing Path

Explore techniques like
- Named Entity Recognition (NER)
- Summarization
- Topic Modeling

Learn **Vector Representations**
- TF-IDF
- Word2Vec
- GloVe

Become **Expert** in

RNN Attention Mechanism Transformers

Computer Vision Path

Focus on real-world applications like object detection

YOLO Mask R-CNN

For Image Segmentation

Use Advanced Tools like

TensorFlow PyTorch for robust implementations

Month Nine

Month Ten

Large Language Models or Diffusion Models

AI technologies that define the forefront of

Modern Data Science
- **Large Langauge Models** if you have chosen NLP
- **Diffusion Models** if you have chosen CV

Large Language Models

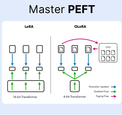
Dive into **Transformer-based architectures** such as

GPT-4 LLaMA by Meta T5

Learn **RAG Systems**



Master **PEFT**



Diffusion Models

Understand the core concepts behind Diffusion Models:
- Noise Scheduling
- Reverse Processes

Stable Diffusion DreamBooth

enable domain-specific fine-tuning for creating synthetic datasets or enhancing creative projects.

Month Eleven

Month Twelve

For More Amazing Infographics, Visit

www.analyticsvidhya.com

Analytics Vidhya