CREDO SYSTEMZ
Simplifying IT

Data Science Course Syllabus

Phase 1: Python Programming

Week 1: Python Basics & Core Programming Concepts

Python Fundamentals

- Setting up Python Environment (Jupyter Notebook, VS Code, Anaconda)
- Python Data Types: Integers, Floats, Strings, Booleans
- Variables and Memory Management
- Operators: Arithmetic, Comparison, Logical, Bitwise
- Type Casting & String Manipulation

Data Structures & Control Flow

- Lists, Tuples, Sets, Dictionaries (Usage in Data Processing)
- Loops: for, while, and Iterators
- Conditional Statements: if-elif-else
- List & Dictionary Comprehensions (Efficient Data Processing)

Functions

- Defining & Calling Functions
- *args and **kwargs
- Lambda Functions

Working with Files & Exception Handling

- Reading & Writing Files (CSV, JSON, TXT)
- Handling Exceptions (try-except-finally)
- Logging in Python

  - **Hands-on Exercise**

Week 2: NumPy, Pandas, and Data Handling

NumPy for Numerical Computations

- Creating NumPy Arrays (ndarray, dtype, shape)
- Indexing, Slicing, and Broadcasting
- Mathematical Operations (Dot Product, Matrix Multiplication)
- Handling Missing Data in Arrays
- Statistical Operations (mean, std, var, percentile)

Pandas for Data Manipulation

- Loading Datasets (read_csv, read_json, read_excel)
- Pandas Series vs. DataFrame (Data Structures)

DataFrame Operations

- Indexing, Filtering, Sorting
- GroupBy & Aggregations (sum, mean, count, apply)
- Handling Missing Data (dropna, fillna, interpolate)
- Merging, Concatenation & Joins in Pandas

- **Hands-on Exercise**

Week 3: Data Visualization, APIs, and Automation

Data Visualization with Matplotlib & Seaborn

- **Matplotlib Basics**: Line, Bar, Scatter, and Pie Charts
- **Seaborn for Statistical Visualization**:
  - Histograms, KDE plots
  - Boxplots & Violin Plots (Detecting Outliers)
  - Heatmaps (Correlation Matrices)
- **Customization**: Legends, Colors, Titles, Annotations

Working with APIs & Web Scraping

- Making API Calls using requests
- Handling JSON Responses
- Introduction to Web Scraping (BeautifulSoup, Scrapy)
- Automating Data Collection

Regular Expressions (Regex) for Text Processing

- Pattern Matching & String Manipulation (re.search, re.findall)
- Cleaning Text Data (Removing Special Characters, Tokenization)
- Applying Regex in Pandas

- **Hands-on Exercise**

Phase 2: Statistics & Probability

Week 4: Descriptive Statistics & Data Understanding

Understanding Data Types & Distributions

- Types of Data: Categorical vs. Numerical (Discrete & Continuous)
- Measures of Central Tendency: Mean, Median, Mode
- Measures of Dispersion: Variance, Standard Deviation, Range, IQR
- Skewness & Kurtosis: Identifying Asymmetry and Shape of Distributions

Data Visualization

- Histograms & Boxplots (for detecting outliers)
- KDE (Kernel Density Estimation) plots
- Scatterplots & Pairplots (Exploring relationships)

Outliers & Data Cleaning

- Z-score & Standardization
- IQR (Interquartile Range) Method for Outlier Detection
- Winsorization & Transformation Techniques (Log, Square Root, etc.)

- **Hands-on Exercise**

Weeks 5 and 6: Probability Theory & Inferential Statistics

Probability Foundations

- Basic Probability Rules (Addition, Multiplication, Conditional Probability)
- Bayes' Theorem & Applications (Spam detection, Medical diagnosis, etc.)

Probability Distributions

- **Discrete Distributions**: Bernoulli, Binomial, Poisson
- **Continuous Distributions**: Uniform, Normal, Exponential
- Central Limit Theorem (CLT) and its importance in large-scale data analysis

Hypothesis Testing & Confidence Intervals

- Statistical Hypothesis Testing
- Null Hypothesis (H0) & Alternative Hypothesis (H1)
- p-value & Significance Level (α)

Types of Hypothesis Tests

- **Z-test** (for large samples, known variance)
- **T-test** (for small samples, unknown variance)
- **Chi-square test** (for categorical data independence)
- **ANOVA** (Comparing multiple groups)

Confidence Intervals

- How to interpret confidence intervals?
- Bootstrapping for interval estimation

- **Hands-on Exercise**

Phase 3: Classical Machine Learning

Week 7: Introduction to Machine Learning & Supervised Learning

Machine Learning Basics

- What is Machine Learning? Types: Supervised, Unsupervised, Reinforcement Learning
- ML Workflow: Data Collection, Preprocessing, Feature Engineering, Model Training, Evaluation

Linear Regression (Predicting Continuous Variables)

- Simple Linear Regression (One Predictor)
- Multiple Linear Regression (Multiple Predictors)
- Assumptions of Linear Regression (Linearity, No Multicollinearity,
- Homoscedasticity)
- Feature Selection Techniques:
- Recursive Feature Elimination (RFE)
- LASSO & Ridge Regularization
- Evaluating Regression Models:
- RMSE, MAE, R^2 Score

- **Hands-on Exercise**

Week 8: Classification Models & Model Evaluation

Logistic Regression (Binary Classification)

- Sigmoid Function & Decision Boundary
- Cost Function for Classification
- ROC Curve, AUC, Precision-Recall Tradeoff
- Handling Class Imbalance (Oversampling & Undersampling, SMOTE)

Decision Trees & Random Forests

- How Decision Trees Work (Gini Index, Entropy, Information Gain)
- Overfitting & Pruning in Decision Trees

- Random Forests for Ensemble Learning
- Feature Importance in Tree-based Models

Support Vector Machines (SVMs)

- How SVM Works: Finding the Optimal Hyperplane
- Kernel Trick (Linear, Polynomial, RBF Kernels)
- Tuning C and Gamma for Model Performance

- Hands-on Exercise

Week 9: Unsupervised Learning & Dimensionality Reduction

Clustering Techniques (Grouping Similar Data)

- K-Means Clustering:
- Choosing the Optimal K (Elbow Method, Silhouette Score)
- Hierarchical Clustering
- DBSCAN (Density-Based Clustering)

Principal Component Analysis (PCA)

- Reducing Dimensions to Improve Model Efficiency
- Eigenvectors & Eigenvalues Intuition
- Explained Variance & Selecting the Right Number of Components

Feature Engineering & Data Preprocessing

- Handling Missing Values (Mean/Median Imputation, KNN Imputation)
- Encoding Categorical Variables (One-Hot Encoding, Label Encoding, Target
- Encoding)
- Feature Scaling (Standardization, Min-Max Scaling, Robust Scaling)
- Feature Extraction Techniques

- Hands-on Exercise

Week 10: Advanced ML Concepts & Hyperparameter Tuning

Gradient Boosting Algorithms (XGBoost, LightGBM, CatBoost)

- Why Boosting Works?
- XGBoost & LightGBM Optimization Techniques
- Regularization in Boosting Models

Model Selection & Hyperparameter Tuning

- Grid Search vs. Random Search vs. Bayesian Optimization
- Automated ML Tuning using Optuna
- Cross-Validation Techniques (K-Fold, Stratified K-Fold, Leave-One-Out)

- Hands-on Exercise

Week 11: Time Series Forecasting & Real-World Applications

Time Series Forecasting

- Introduction to Time Series Data
- Stationarity & Differencing Techniques
- ARIMA, SARIMA, and Prophet Models

Model Deployment & Interpretability

- Deploying ML Models using Flask/FastAPI
- Explainable AI (SHAP, LIME for Feature Interpretability)

- Hands-on Exercise

Phase 4: Deep Learning (20 hours)

Week 12: Neural Networks & Deep Learning Foundations

Introduction to Deep Learning & Neural Networks

- Why Deep Learning?

- Comparison: ML vs. Deep Learning vs. AI
- Biological Neurons vs. Artificial Neurons
- Activation Functions: ReLU, Sigmoid, Tanh, Softmax
- Forward & Backpropagation in Neural Networks
- Loss Functions (MSE, Cross-Entropy)

Building Feedforward Neural Networks (FNNs)

- Understanding Weights & Biases
- Gradient Descent & Optimizers (SGD, Adam, RMSprop)
- Overfitting & Regularization: Dropout, L1/L2 Regularization
- Batch Normalization for Stable Training

- Hands-on Exercise

Week 13: Convolutional Neural Networks (CNNs) for Computer Vision

CNN Architecture & Applications

- Why CNNs for Image Data?
- Convolution Operation & Filters
- Stride, Padding, and Pooling Layers
- Flattening & Fully Connected Layers

Building & Training CNN Models

- Popular Architectures: LeNet, AlexNet, VGG, ResNet
- Transfer Learning with Pretrained Models
- Data Augmentation & Handling Imbalanced Datasets

- Hands-on Exercise

Week 14: Recurrent Neural Networks (RNNs)

Sequence Modeling & Recurrent Networks

- Why RNNs for Sequential Data?
- Vanishing Gradient Problem & LSTMs/GRUs

- Hands-on Exercise

Week 15: NLP

Introduction to NLP

- What is NLP?
  - Real-world applications (Search engines, chatbots, sentiment analysis)
- Challenges in NLP
  - Ambiguity, Context Understanding, Sarcasm
- Hands-on:
  - Implementing basic text preprocessing

Text Preprocessing Techniques

- Tokenization
  - Word & Sentence Tokenization
- Text Cleaning
  - Stopword Removal, Lemmatization vs. Stemming
- Hands-on:
  - Preprocessing text using NLTK & spaCy

Text Representation Techniques

- Traditional Approaches
  - Bag of Words (BoW) & TF-IDF
- Word Embeddings
  - Word2Vec, GloVe, FastText
- Contextual Embeddings
  - BERT, GPT
- Hands-on:
  - Implementing embeddings using Scikit-learn & Gensim

Named Entity Recognition (NER) & POS Tagging

- POS Tagging & Dependency Parsing
  - Understanding syntactic structures

- Named Entity Recognition (NER)
  - Using spaCy for entity extraction
- Hands-on:
  - Extracting key entities from resumes

Sequence Modeling & Recurrent Networks

- Why RNNs for Sequential Data?
  - Time-series & language modeling use cases
- Challenges with RNNs
  - Vanishing Gradient Problem
- LSTMs & GRUs
  - How they solve RNN limitations
- Hands-on:
  - Training an LSTM for sentiment analysis

Sentiment Analysis & Text Classification

- Traditional Machine Learning Approaches
  - Naive Bayes, SVM for text classification
- Deep Learning Approaches
  - Transformer-based Sentiment Analysis (BERT, DistilBERT)
- Hands-on:
  - Fine-tuning BERT for text classification

Text Similarity & Semantic Search

- Text Similarity Metrics
  - Cosine Similarity, Jaccard Similarity
- Semantic Search
  - Using BERT & SBERT
- Hands-on:
  - Implementing SBERT for FAQ retrieval

Machine Translation & Text Generation

- Sequence-to-Sequence Models
  - LSTMs, Transformers for translation

- Neural Machine Translation (NMT)
  - Using MarianMT for multilingual translation
- Hands-on:
  - English-to-French translation with Hugging Face

Topic Modeling & Text Summarization

- Topic Modeling Techniques
  - LDA (Latent Dirichlet Allocation)
- Summarization Approaches
  - Extractive (TextRank), Abstractive (BART)
- Hands-on:
  - Summarizing long articles with BART

Phase 5: Generative AI and Prompting techniques (20 hours)

PART 1: FUNDAMENTALS OF GENERATIVE AI

Introduction to Generative AI

- What is Generative AI?
  - Difference between Discriminative & Generative Models
- Real-world Applications of Generative AI
  - Image generation, text generation, code generation, music synthesis
- Hands-on:
  - Exploring OpenAI's GPT and DALL·E models

Foundation Models in Generative AI

- Overview of Large Language Models (LLMs)
  - GPT, BERT, T5, LLaMA, Mistral
- Pre-training & Fine-tuning of LLMs
  - Supervised Fine-tuning vs. Reinforcement Learning from Human Feedback (RLHF)
- Hands-on:
  - Fine-tuning GPT-3.5/4 with domain-specific datasets

Transformer Architecture & Mechanisms

- Understanding Transformer-based Models
  - Self-attention, positional encoding, multi-head attention
- Differences between Auto-regressive & Auto-encoding Models
  - GPT vs. BERT vs. T5
- Hands-on:
  - Implementing a Transformer from scratch in PyTorch

PART 2: PROMPT ENGINEERING & TECHNIQUES

Introduction to Prompting Techniques

- What is Prompt Engineering?
- Importance of Effective Prompts in AI Models
- Hands-on:
  - Experimenting with basic prompts using OpenAI's API

Types of Prompting Techniques

- Zero-shot Prompting
  - Generating responses
- Few-shot Prompting
  - Providing contextual examples to guide AI behavior
- Chain-of-Thought (CoT) Prompting
  - Step-by-step reasoning for complex problem-solving
- Hands-on:
  - Testing different prompting techniques for text generation

Advanced Prompt Engineering Strategies

- Self-consistency Prompting
  - Improving reliability in AI-generated responses
- Role-based & Persona-based Prompting
  - Using AI to mimic specific styles & behaviors
- Hands-on:
  - Designing prompts for creative writing & chatbots

Prompt Optimization & Debugging

- Prompt Refinement Strategies
- Handling Model Hallucinations & Biases
- Hands-on:
  - Debugging AI-generated outputs with iterative prompt tuning

PART 3: GENERATIVE AI APPLICATIONS

Text Generation & AI Writing Assistants

- Natural Language Generation (NLG)
- AI for Content Creation & Summarization
- Hands-on:
  - Generating blog articles & summaries using GPT

Image Generation with Diffusion Models

- Understanding Diffusion Models (Stable Diffusion, DALL·E)
- Customizing AI-generated images with control prompts
- Hands-on:
  - Generating and editing AI images using prompt-based techniques

Multimodal AI & Interactive Applications

- Combining Text, Images, and Audio in AI Models
- Understanding CLIP (Contrastive Language-Image Pretraining)
- Hands-on:
  - Implementing multimodal AI using CLIP

AI in Code Generation & Productivity Tools

- AI-powered Code Assistants (GitHub Copilot, Code Llama)
- AI-driven Workflow Automation
- Hands-on:
- Using AI for code generation and debugging

PART 4: ETHICS, DEPLOYMENT & PROJECTS

Ethical Considerations in Generative AI

- Bias, Misinformation & Hallucinations in AI
- Copyright & Legal Issues in AI-generated Content
- Hands-on:
  - Evaluating AI model bias with real-world examples

Deploying Generative AI Models

- Model Deployment Strategies
- APIs, Cloud Deployment, Containers
  - Optimizing AI Model Performance & Cost
- Hands-on:
  - Deploying a GPT-based chatbot using FastAPI

Phase 6: Ancillary Skills for AI & ML Practitioners (10 hours)

Week 1: Big Data Processing & Distributed Computing

Introduction to Big Data for AI

- Why Big Data is Important for AI/ML?
- Challenges in Handling Large-Scale AI Workloads
- Big Data Ecosystem: Spark, and Dask

Scalable Data Processing for AI Pipelines

- Apache Spark for Data Processing
- Pandas vs. Dask vs. Spark for ML Workflows
- ETL Pipelines for AI: Batch vs. Streaming Data Processing

- Hands-on Exercise

Week 2: Cloud Computing & AI Workflows

Cloud AI Services & Model Deployment

- Overview of Cloud AI Platforms: AWS, GCP, Azure

s

- AutoML & Managed AI Services (AWS SageMaker, GCP Vertex AI)
- Scaling AI Models with Cloud GPUs & TPUs

Serverless & Containerized AI Deployments

- Dockerizing AI Applications for Portability
- Kubernetes for AI Model Scaling
- Deploying AI Models as Serverless Functions (AWS Lambda, GCP Cloud
- Functions)
  - Hands-on Exercise

Week 3: ML Engineering & Software Development Best Practices

ML Engineering & Best Coding Practices

- Structuring ML Code for Maintainability
- Modularization, OOP & Functional Programming in AI
- Unit Testing for AI Codebases (pytest, unittest)

CI/CD Pipelines for AI Models

- Introduction to CI/CD for ML (Continuous Integration & Deployment)
- Using GitHub Actions for ML Pipelines
- Automating ML Workflows with MLflow & Kubeflow
  - Hands-on Exercise