Data Science & Machine Learning Program

### Week-by-Week Training Schedule

| Week | Focus Area | Topics Covered | Projects |
| --- | --- | --- | --- |
| **1** | **Foundations & EDA** | Python for DS, Pandas, NumPy, Matplotlib/Seaborn, GitHub setup | Iris Dataset Classification |
| **2** | **Data Cleaning & Regression** | Data wrangling, missing values, encoding, scaling, feature engineering | House Price Regression, BigMart Sales Prediction |
| **3** | **Classification Problems** | Logistic regression, decision trees, random forests, and evaluation metrics | Loan Prediction, Titanic Survival Prediction |
| **4** | **NLP Applications** | Text preprocessing, TF-IDF, sentiment analysis, Naive Bayes | Twitter Sentiment Analysis, Fake News Detection |
| **5** | **Deep Learning Intro** | Keras/PyTorch, CNNs, regularization, validation techniques | MNIST Digit Recognition, Wine Quality Prediction |
| **6** | **Computer Vision** | Advanced CNNs, transfer learning, OpenCV applications | Traffic Sign Recognition, Driver Drowsiness Detection |
| **7** | **ML Systems** | Unsupervised learning, clustering, recommendation systems | Movie Recommender, Customer Segmentation, Fraud Detection |
| **8** | **Capstone & Career** | End-to-end ML project, portfolio optimization, interview prep | Covid-19 X-ray Classification OR Custom capstone |

## 

Detailed Weekly Curriculum

### **Week 1: Foundations & Exploratory Data Analysis (EDA)**

**Focus:** Building strong Python foundations for data science

#### Session 1: Python Fundamentals & Environment Setup

**Topics:**

- Python refresher: data types, control structures, functions

- Jupyter notebook setup and best practices

- Introduction to data science libraries ecosystem

- Git and GitHub setup for portfolio development

**Practical Activities:**

- Set up development environment (Anaconda, Jupyter, VS Code)

- Create GitHub repository structure

- Basic Python exercises with data structures

#### Session 2: Data Manipulation with Pandas & NumPy

**Topics:**

- Pandas DataFrames: creation, indexing, filtering, grouping

- NumPy arrays: mathematical operations, broadcasting

- Data loading from various sources (CSV, JSON, APIs)

- Basic data cleaning techniques

**Practical Activities:**

- Load and explore multiple datasets

- Practice data manipulation exercises

- Introduction to the **Iris Dataset**

#### Session 3: Data Visualisation & EDA Techniques

**Topics:**

- Matplotlib fundamentals: plots, subplots, customisation

- Seaborn for statistical visualisation

- EDA methodology: univariate, bivariate, multivariate analysis

- Statistical summaries and distributions

**Project:** **Iris Dataset Classification**

- Complete EDA with visualisations

- Build first ML model (basic classification)

- Document findings and push to GitHub

**Week 1 Deliverable:** Comprehensive Jupyter notebook with EDA + ML model on GitHub

### **Week 2: Data Preprocessing & Regression Analysis**

**Focus:** Advanced data cleaning and regression modelling

#### Session 4: Advanced Data Cleaning & Feature Engineering

**Topics:** - Handling missing values: imputation strategies - Outlier detection and treatment - Data type conversions and categorical encoding - Feature scaling and normalisation techniques

**Practical Activities:** - Work with messy datasets - Implement various imputation methods - Create feature engineering pipeline

#### Session 5: Regression Fundamentals & Implementation

**Topics:** - Linear regression theory and assumptions - Multiple regression and polynomial features - Regularisation techniques (Ridge, Lasso, Elastic Net) - Model evaluation metrics for regression

**Project Start:** **House Price Regression** - Load and explore housing dataset - Feature engineering for price prediction - Implement multiple regression models

#### Session 6: Advanced Regression & Business Applications

**Topics:** - Cross-validation techniques - Hyperparameter tuning with GridSearch - Feature selection methods - Business interpretation of model results

**Projects:** - Complete **House Price Regression** with model tuning - Start **BigMart Sales Prediction** project - Compare different regression approaches

**Week 2 Deliverable:** Two complete regression projects with documentation and model comparison

### **Week 3: Classification Algorithms & Evaluation**

**Focus:** Binary and multi-class classification problems

#### Session 7: Classification Fundamentals

**Topics:** - Logistic regression: theory and implementation - Decision trees: splitting criteria, pruning - Classification vs regression: key differences - Probability interpretation in classification

**Project Start:** **Loan Prediction** - Explore loan approval dataset - Implement logistic regression - Feature importance analysis

#### Session 8: Ensemble Methods & Advanced Classification

**Topics:** - Random Forest: bagging and feature randomness - Gradient Boosting and XGBoost - Ensemble voting methods - Handling imbalanced datasets

**Practical Activities:** - Compare individual vs ensemble models - Implement SMOTE for imbalanced data - Cross-validation for model selection

#### Session 9: Model Evaluation & Business Metrics

**Topics:** - Confusion matrix and classification report - ROC curves and AUC interpretation - Precision, recall, F1-score trade-offs - Business-specific evaluation metrics

**Project:** **Titanic Survival Prediction** - Complete end-to-end classification pipeline - Comprehensive model evaluation - Feature engineering for survival prediction

**Week 3 Deliverable:** Two classification projects with comprehensive evaluation metrics and business insights

### **Week 4: Natural Language Processing (NLP)**

**Focus:** Text data processing and sentiment analysis

#### Session 10: Text Preprocessing & NLP Fundamentals

**Topics:** - Text cleaning: tokenisation, stemming, lemmatisation - Stop words removal and text normalisation - Regular expressions for text processing - Introduction to NLTK and spaCy libraries

**Practical Activities:** - Process raw text data from various sources - Build text preprocessing pipeline - Explore Twitter and news datasets

#### Session 11: Feature Extraction from Text

**Topics:** - Bag of Words (BoW) model - TF-IDF (Term Frequency-Inverse Document Frequency) - N-grams and character-level features - Word embeddings introduction

**Project Start:** **Sentiment Analysis** - Amazon reviews dataset exploration - Implement BoW and TF-IDF approaches - Compare feature extraction methods

#### Session 12: NLP Classification & Applications

**Topics:** - Naive Bayes for text classification - Support Vector Machines for text - Model evaluation for NLP tasks - Handling multi-class text classification

**Projects:** - Complete **Twitter Sentiment Analysis** - **Fake News Detection** implementation - **Email Classification** (spam vs ham)

**Week 4 Deliverable:** Three NLP projects demonstrating different text classification approaches

### **Week 5: Introduction to Deep Learning**

**Focus:** Neural networks and computer vision basics

#### Session 13: Neural Network Fundamentals

**Topics:** - Perceptron and multi-layer perceptrons - Backpropagation algorithm - Activation functions and their properties - Introduction to TensorFlow/Keras

**Practical Activities:** - Build neural network from scratch - Implement simple MLP with Keras - Understand gradient descent visualisation

#### Session 14: Convolutional Neural Networks (CNNs)

**Topics:** - CNN architecture: convolution, pooling, fully connected layers - Feature maps and filters - CNN for image classification - Transfer learning concepts

**Project:** **Handwritten Digit Recognition (MNIST)** - Load and preprocess MNIST dataset - Build CNN from scratch - Compare with traditional ML approaches

#### Session 15: Deep Learning Best Practices

**Topics:** - Overfitting prevention: dropout, early stopping - Batch normalisation and optimisation - Model architecture design principles - Hyperparameter tuning for deep learning

**Project:** **Wine Quality Prediction** - Apply deep learning to tabular data - Compare neural networks with ensemble methods - Document model architecture decisions

**Week 5 Deliverable:** Two deep learning projects with architecture documentation and performance analysis

### **Week 6: Computer Vision Applications**

**Focus:** Advanced image processing and real-world CV applications

#### Session 16: OpenCV Fundamentals & Image Processing

**Topics:** - Image loading, manipulation, and basic operations - Colour spaces and transformations - Edge detection and contour analysis - Basic image filtering and morphological operations

**Project:** **Color Detection** - Implement colour detection system - Real-time colour tracking with webcam - HSV colour space applications

#### Session 17: Advanced CNN Architectures & Transfer Learning

**Topics:** - Pre-trained models: VGG, ResNet, MobileNet - Transfer learning strategies - Fine-tuning vs feature extraction - Model optimisation for deployment

**Project:** **Traffic Signs Recognition** - Use German Traffic Sign Recognition Benchmark - Implement transfer learning with pre-trained CNN - Real-time traffic sign detection

#### Session 18: Real-time Computer Vision Applications

**Topics:** - Facial landmark detection - Real-time video processing - Integration of deep learning with OpenCV - Model deployment considerations

**Project:** **Driver Drowsiness Detection** - Eye blink detection using facial landmarks - Real-time drowsiness alert system - Combine OpenCV with deep learning models

**Week 6 Deliverable:** Three computer vision applications with real-time capabilities and deployment documentation

### **Week 7: Machine Learning Systems & Recommendation Engines**

**Focus:** Unsupervised learning and production ML systems

#### Session 19: Unsupervised Learning & Clustering

**Topics:** - K-means clustering algorithm and implementation - Hierarchical clustering methods - DBSCAN for density-based clustering - Dimensionality reduction with PCA

**Project:** **Customer Segmentation** - RFM analysis for customer segmentation - K-means clustering implementation - Business insights from clustering results

#### Session 20: Recommendation Systems

**Topics:** - Collaborative filtering vs content-based filtering - Matrix factorisation techniques - Similarity metrics and nearest neighbours - Evaluation metrics for recommendation systems

**Project Start:** **Movie Recommendation System** - Implement collaborative filtering - Content-based filtering with movie features - Hybrid recommendation approach

#### Session 21: Anomaly Detection & Production Systems

**Topics:** - Statistical methods for anomaly detection - Isolation Forest and One-Class SVM - Model deployment with Streamlit/Gradio - API development for ML models

**Projects:** - Complete **Movie Recommendation System** with web interface - **Credit Card Fraud Detection** using anomaly detection - Deploy recommendation system as web application

**Week 7 Deliverable:** Deployed web applications for recommendation system and fraud detection

### **Week 8: Capstone Project & Career Preparation**

**Focus:** End-to-end ML project and job readiness

#### Session 22: Capstone Project Planning & Advanced Applications

**Topics:** - Project planning and requirement analysis - Advanced deep learning applications - Medical image analysis with CNNs - Transfer learning for specialised domains

**Capstone Project Options:** - **Covid-19 X-ray Classification** (medical imaging) - **Human Activity Recognition** (time series + sensors) - **Custom project** based on student interest

#### Session 23: Model Development & Evaluation

**Topics:** - End-to-end ML pipeline development - Model versioning and experiment tracking - Advanced evaluation techniques - A/B testing concepts for ML

**Activities:** - Develop complete capstone project - Implement comprehensive evaluation framework - Document methodology and results

#### Session 24: Portfolio Optimization & Career Preparation

**Topics:** - GitHub portfolio structuring and documentation - Technical resume optimisation - LinkedIn profile enhancement - Interview preparation and mock interviews

**Career Activities:** - Portfolio review and optimisation - README documentation for all projects - Resume and LinkedIn profile updates - Mock technical interviews with feedback

**Week 8 Deliverable:** Complete portfolio with capstone project + career preparation materials