DATA SCIENCE PROJECTS

SQL • Python • Pandas • Machine Learning • NLP • AWS • Projects & Implementation Tasks

Project briefs are aligned to the supplied course source. Tasks are written as implementation activities so learners can build, test, document and present each project.

PROJECT 1: Email Extraction Project

SOURCE COVERAGE

Python functions, file I/O and Regex basics. The source explicitly includes an email extraction project.

OBJECTIVE

Create a Python utility that reads text/file input and extracts email addresses.

PROJECT TASKS

1. Prepare sample text/file input.
2. Read the input using Python file I/O.
3. Create and test the Regex pattern for email extraction.
4. Handle multiple matches and repeated values.
5. Validate the extracted output.
6. Save the results and document the approach.

DELIVERABLES

- Python source
- Sample input/output
- Regex explanation
- Extracted result file
- README

PROJECT 2: Machine Learning Model Evaluation & Pipeline Project

SOURCE COVERAGE

Supervised learning, Linear Regression, Metrics & Evaluation, Decision Trees, Random Forests, K-Means, DBSCAN, PCA, Hyperparameter Tuning and Pipelines.

OBJECTIVE

Build a structured ML workflow from model preparation through evaluation and pipeline construction.

PROJECT TASKS

1. Prepare the dataset and define features/target.
2. Build a Linear Regression model.
3. Calculate and interpret evaluation metrics.
4. Train a Decision Tree and Random Forest.
5. Compare model behavior and evaluation results.
6. Apply K-Means and DBSCAN for an unsupervised learning exercise.
7. Apply PCA and dimensionality reduction.
8. Perform hyperparameter tuning.
9. Create a repeatable ML Pipeline.
10. Document model selection and final findings.

DELIVERABLES

- Prepared dataset
- Regression model
- Tree/ensemble comparison
- Clustering output
- PCA output
- Tuned pipeline
- Final report

Greens Technologies

PROJECT 3: NLP Text Processing Project

SOURCE COVERAGE

Text preprocessing, POS tagging, Bag of Words, TF-IDF, Embeddings, SpaCy NER and Transformer models (BERT) overview.

OBJECTIVE

Build a progressive NLP workflow from raw text to feature extraction and entity recognition.

PROJECT TASKS

1. Collect and prepare sample text.
2. Apply text preprocessing.
3. Perform POS tagging.
4. Create Bag of Words representation.
5. Create TF-IDF features.
6. Experiment with embeddings.
7. Apply SpaCy NER.
8. Review where transformer models and BERT fit into the workflow.
9. Compare the outputs from the different representations.

DELIVERABLES

- Preprocessing pipeline
- POS output
- BoW/TF-IDF features
- Embedding output
- NER output
- NLP comparison report

PROJECT 4: AWS Data Workflow Project

SOURCE COVERAGE

IAM, S3, Lambda, CloudWatch, EventBridge, Glue, Redshift, Step Functions and SageMaker ML Pipelines.

OBJECTIVE

Apply the AWS section as an integrated cloud workflow.

PROJECT TASKS

1. Design the AWS architecture.
2. Set up IAM access requirements.
3. Create the S3 storage flow.
4. Implement a Lambda step.
5. Add CloudWatch logging.
6. Configure EventBridge automation.
7. Build Glue crawler/job/pipeline components.
8. Use Redshift for warehousing.
9. Add Step Functions orchestration.
10. Document SageMaker ML Pipeline integration.

DELIVERABLES

- Architecture diagram
- AWS workflow
- ETL configuration
- Monitoring plan
- Orchestration flow
- Presentation

Greens Technologies