PEA
PETROLEUM ENGINEERS ASSOCIATION

ACCREDITED

2 Weeks Online Practical Training On

Data Mining for Oil & Gas Professionals

python™

Nashat Jumaah Omar

ZOOM

SCAN
TO JOIN
PEA
COMMUNITY

REGISTRATION OPEN!

REGISTER NOW!

Reach out to us at

+91 6205464268/
+91 70194 95792

www.peassociations.com

ABOUT THIS TRAINING

- The Data Mining Training equips professionals in the oil and gas industry with practical skills to extract valuable insights from large datasets. Participants will learn the full data mining process, including data preparation, exploration, association rule mining, and predictive modeling using Python.

- The training includes hands-on exercises and industry-relevant case studies to apply techniques for predictive maintenance, process optimization, and safety improvements. By the end, attendees will be able to turn raw data into strategic knowledge, driving informed decisions and operational efficiency.

SKILLS GAINED

- Introduction to Python For Data Science
- Data preparation and cleaning
- Exploratory data analysis (EDA)
- Association rule mining
- Predictive modeling techniques
- Application of Python for data mining
- Practical case study analysis
- Enhancing predictive maintenance strategies

AUDIENCE

- Reservoir Engineers.
- Production engineers.
- Chemical engineers.
- Drilling engineers.
- Geologists and petrophysics
- AL and workover engineers.
- Any one with exposure to Large Volumes of Data

PREREQUISITE

- No knowledge is required
- A working laptop with Windows 10 OS, MacOS or GNU Linux Distro

WHAT YOU WILL GET FROM JOINING

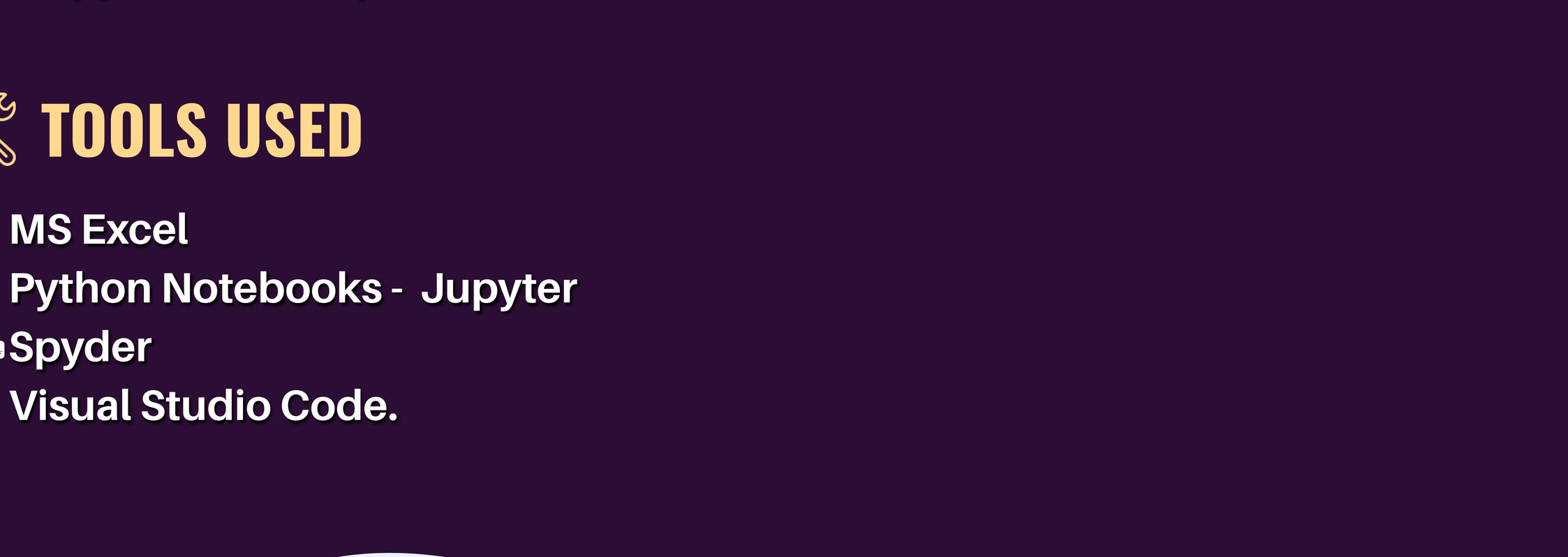
- Access to Video Recordings on daily basis.
- Study materials ppt, pdf
- Many Oil and Gas Datasets.
- Learn Python through anaconda package

LIBRARIES COVERED

- Python Core Library.
- Matplotlib
- Plolty
- Pandas
- Dask
- Streamlit
- Pypdf (Python PDF Reader).
- PathLib
- Numpy
- Sci-kit Learn
- Scipy for Curve Optimization

TOOLS USED

- MS Excel
- Python Notebooks - Jupyter
- Spyder
- Visual Studio Code.



MODULE 1

TOPICS

- A gentle introduction to Python Programming Language
- Data types and Structures in Python
- Introduction to Data Visualization
- Working with Tabulated Data using Pandas
- Basics of Data Cleaning and Transformation using Pandas.
- Creating Calculations and Data Exports.

EXERCISES

- Oil and Gas Data Reading and excel connection to python
- Simple Reservoir Data Visualization.
- Filtering Reservoir Data based on Wells (single or Multiple)
- Cleaning and organizing historical data, with proper datetime conversion.

WHAT YOU WILL LEARN

- Understand the Basics of Python Programming: Gain a foundational understanding of the Python programming language, including its syntax, structure, and common uses.
- Navigate the Python Environment and Ecosystem: Familiarize yourself with the Python environment, including IDEs, libraries, and tools that form the Python ecosystem.
- Identify and Utilize Python Data Types and Structures: Learn to work with various data types and structures in Python, such as lists, dictionaries, tuples, and sets, for efficient data management.

MODULE 2

TOPICS

- Introduction to Exploratory Data Analysis
- Importance of EDA in the data science workflow
- Overview of the EDA process: understanding the dataset, data cleaning, and initial insights
- Tools and libraries for EDA: Pandas, Matplotlib, Seaborn, and Plotly
- Understanding the Dataset
- Exploring dataset structure: rows, columns, data types
- Inspecting the first and last few rows, and summary statistics
- Identifying missing values, duplicates, and outliers
- Descriptive statistics: mean, median, mode, variance, standard deviation
- Visualizing Data Distributions
- Visualizing the distribution of numeric variables: histograms, boxplots, and density plots
- Exploring categorical data: bar charts, pie charts, and count plots
- Using pair plots and scatter matrix plots for feature relationships
- Identifying skewness, kurtosis, and data transformation needs

EXERCISES

- Exercise 1: Dataset Overview
- Load a dataset and explore its basic structure using head (), tail (), and info ().
- Calculate and interpret summary statistics such as mean, median, and standard
- deviation.
- Identify missing values and duplicates in the dataset and apply strategies to handle them.
- Exercise 2: Visualizing Data Distributions
- Create histograms and boxplots to visualize the distribution of numerical features.
- Generate bar charts and count plots for categorical variables to understand their frequency distribution.
- Identify skewness in the data and suggest appropriate transformations.

WHAT YOU WILL LEARN

- Understanding and Preparing the Dataset
- Learn how to explore and understand the structure of a dataset using Pandas methods like head (), info (), and summary statistics.
- Gain skills in identifying and handling missing data, duplicates, and outliers to prepare data for further analysis.
- Visualizing Data Distributions
- Master the creation of various plots to visualize the distribution of numerical and categorical features.
- Learn to interpret histograms, boxplots, count plots, and bar charts to understand the spread and frequency of data.
- Identify skewness and understand how to transform data for better analysis.

MODULE 3

TOPICS

- Introduction to Relations and Correlation
- Understanding relationships between variables in data
- Difference between correlation and causation
- Types of relationships: linear, non-linear, monotonic, and non-monotonic
- Introduction to Pearson’s correlation coefficient and other correlation measures
- Exploring Correlation in Data
- Interpreting correlation coefficients (positive, negative, zero correlation)
- Visualizing correlations: heatmaps, scatter plots, and pair plots
- Understanding and handling multicollinearity
- Exploring other correlation methods: Spearman's rank, Kendall’s Tau
- Data Fitting
- Introduction to data fitting and its importance in data science and machine learning
- Overview of different data fitting techniques: linear regression, polynomial fitting, and spline fitting
- Fitting a line to data: simple linear regression model
- Exploring advanced data fitting techniques: multiple regression, non-linear regression, and curve fitting
- Model evaluation: R-squared, Mean Squared Error (MSE), and residual analysis

EXERCISES

- Calculate and interpret the Pearson correlation coefficient between numerical features in a dataset.
- Visualize correlations using a heatmap and scatter plots.
- Identify highly correlated variables and discuss the implications of multicollinearity.
- Use Spearman's rank correlation to analyze the relationship between ordinal variables.
- Compare Pearson and Spearman correlation methods using different types of datasets.
- Create scatter plots to visualize linear and non-linear relationships between features.

WHAT YOU WILL LEARN

- Understanding Relationships and Correlation
- Gain a deep understanding of different types of relationships between variables, including linear, non-linear, and monotonic relationships.
- Exploring and Visualizing Correlation
- Learn how to visualize correlations using scatter plots, heatmaps, and pair plots.
- Understand the concept of multicollinearity and how it can impact model performance, as well as strategies to mitigate it.
- Data Fitting Techniques
- Understand the basics of data fitting, and learn how to apply linear regression to fit a line to data.
- Gain expertise in fitting more complex models, such as polynomial regression and non-linear regression, to better capture relationships in data.
- Learn how to evaluate model performance using metrics like R-squared, MSE, and residual analysis.
- Advanced Data

MODULE 4

TOPICS

- Introduction to Document Data Mining
- Overview of document data mining and its significance
- Types of documents: structured vs. unstructured data
- Key techniques used in document data mining: text preprocessing, feature extraction, and model building
- Applications of document data mining in various industries: sentiment analysis, topic modeling, and text classification
- Text Preprocessing
- Text cleaning: removing stop words, punctuation, and irrelevant characters
- Tokenization: breaking down text into words or phrases
- Stemming and Lemmatization: reducing words to their root form
- Handling case sensitivity and text normalization

EXERCISES

- Reading Full Folders Containing PDFs.
- Extract PDF data from Drilling Reports.
- Extract Patterns from Text Data.
- Search in Documents

WHAT YOU WILL LEARN

- Learn the principles of document data mining and the significance of extracting insights from unstructured text data.
- Gain an understanding of various document types and how they impact text mining approaches.
- Master essential preprocessing steps such as text cleaning, tokenization, stemming, and lemmatization to prepare text data for analysis.
- Learn how to handle case sensitivity and normalize text for better analysis and model performance.

MODULE 5

TOPICS

- Introduction to the Concept of classification
- Voting and Decision Trees
- Introduction to KNN method
- Introduction to the Decision Tree and Random Forest Methods
- Python Plotting techniques

EXERCISES

- Classifying ESP Operational Problems.
- Predicting Flow Regime Type

WHAT YOU WILL LEARN

- Understand the role of labels and events in the oil and gas industry, and how they influence data-driven decision-making.
- Learn the fundamentals of labeled data and how classification models use these labels to predict outcomes.
- Explore decision trees and their related algorithms, and how they are used in machine learning to make decisions based on data.

MODULE 6

TOPICS

- Introduction to Continuous Data and Corresponding Relationships
- Relationship Visualization and Correlation Matrix
- Introduction to Regression Analysis
- Linear Regression Fundamentals
- Support Vector Regression (SVR)
- Xtreme Gradient Regression (XGBoost Library)

EXERCISES

- Training ML to Behave like PROSPER software
- Predicting Hydrocarbon Properties using ML

WHAT YOU WILL LEARN

- Understand continuous data and how it differs from categorical data, as well as the relationships that can be formed between continuous variables.
- Learn how to visualize relationships between variables using various plots and generate correlation matrices to understand variable interdependencies.
- Explore regression analysis as a key technique for modeling relationships between variables and predicting continuous outcomes.
- Gain a deep understanding of linear regression, its assumptions, and how it is used to fit straight-line relationships between variables.
- Learn how to implement Support Vector Regression (SVR), a powerful algorithm for dealing with both linear and non-linear relationships in data.
- Discover the XGBoost library for performing extreme gradient boosting, a highly efficient and accurate machine learning algorithm for regression and classification tasks.

MODULE 7

TOPICS

- Introduction to Time-Bounded Data in the Oil and Gas Industry
- Understanding Typical Decline Curve Analysis (DCA) and Its Limitations
- Introduction to Time Series Analysis (TSA)
- Short-Term Production Prediction Using Time Series Analysis
- Simple Moving Average (SMA) and Exponential Moving Average (EMA)
- Introduction to Auto Regressive (AR) Models

EXERCISES

- Predicting Shale Production Decline using Auto Regression Models.
- Predicting Water Cut Based on WHP, Qo, Qg

WHAT YOU WILL LEARN

- Understand the concept of time-bounded data in the oil and gas industry and its importance for production and reservoir analysis.
- Learn the fundamentals of Decline Curve Analysis (DCA), a traditional tool in oil and gas, and explore its limitations when predicting long-term production.
- Dive into Time Series Analysis (TSA), a modern approach to handling temporal data, and learn how it differs from DCA.
- Recognize patterns and components in time series data, such as trends, seasonality, and cycles, and how these components influence predictions.
- Predicting future values in a time series based on its past behavior.

MODULE 8

TOPICS

- Introduction to Unsupervised Learning
- Overview of unsupervised learning and its applications in data analysis
- Key concepts: no labeled data, finding hidden structures, and pattern recognition
- Unsupervised learning tasks: anomaly detection, similarity discovery, and clustering
- Anomaly Detection
- Definition and importance of anomaly detection in various industries (e.g., fraud detection, network security, etc.)
- Types of anomalies: point anomalies, contextual anomalies, and collective anomalies
- Techniques for anomaly detection: statistical methods, distance-based methods, and density-based methods
- Common algorithms: K-Nearest Neighbors (KNN), One-Class SVM, Isolation Forest,
- Local Outlier Factor (LOF)
- Understanding similarity measures: Euclidean distance, cosine similarity, and Jaccard similarity
- Applications of similarity discovery: document similarity, recommendation systems, image recognition, etc.
- Techniques for measuring and visualizing similarities between data points

EXERCISES

- Exercise 1: Anomaly Detection with KNN and LOF
- Apply K-Nearest Neighbors (KNN) to detect point anomalies in a given dataset.
- Use Local Outlier Factor (LOF) to identify dense regions of anomalies in a dataset.
- Evaluate the performance of both techniques using precision, recall, and F1-score.
- Exercise 2: Similarity Discovery with Cosine Similarity
- Calculate the cosine similarity between pairs of documents in a text corpus.
- Use similarity measures to cluster similar documents and evaluate the clusters.
- Apply dimensionality reduction techniques like PCA or t-SNE to visualize high-dimensional similarity relationships.

WHAT YOU WILL LEARN

- Unsupervised Learning Techniques
- Understand the principles and applications of unsupervised learning tasks, including anomaly detection, similarity discovery, and clustering.
- Learn how to identify and analyze patterns in unlabeled data and discover hidden structures within datasets.
- Anomaly Detection
- Master the techniques used for anomaly detection, including statistical, distance-based, and density-based methods.
- Learn how to apply K-Nearest Neighbors (KNN), One-Class SVM, Isolation Forest, and Local Outlier Factor (LOF) for detecting anomalies in datasets.
- Evaluate the performance of anomaly detection algorithms using precision, recall, and F1-score.
- Similarity Discovery
- Gain a deep understanding of similarity measures like Euclidean distance, cosine similarity, and Jaccard similarity.
- Learn how to apply similarity measures in practical applications such as document similarity and recommendation systems.
- Use dimensionality reduction techniques (PCA, t-SNE) to visualize and discover similarities in high-dimensional data.