GUIDED PROJECTS

Module 1: Python for Data Science

Case Study 1: Numpy, Pandas, Matplotlib

Problem Statement:

A small retail store is looking to manage and optimize their inventory. They want to identify which products are in high demand, which products are not selling. The dataset has daily sales data for each product.

Dataset: Sales.csv

Dataset description:

- **Order ID** - Each order receives its own Order ID that will not be duplicated.
- **Product** – The name of the product that has been sold.
- **Quantity Ordered** – The total item quantity ordered in the initial order.
- **Price Each** - The price of each product.
- **Order Date** - This is the date the customer is requesting the order be shipped.
- **Purchase Address** – Billing address.

Approach:

- Load the dataset into a Pandas DataFrame.
- Data exploration, Data cleaning.
- Use pandas to calculate key metrics such as
    - Total sales for each product.
    - Average sales for each product per day.
- Use Matplotlib to create visualizations and identify
    - Highest selling products. Which products are sold the most?
    - Which city had highest number of sales?

Case Study 2: Pandas, Seaborn

Problem Statement:

Analyzing Social Media Data - Dataset contains social media data such as the number of likes, comments, shares, etc. You need to analyze the data and create visualizations to identify popular trends and patterns. You can use Pandas to clean and manipulate the data and Seaborn to represent the data using visualizations.

Dataset: social media influencers – youtube.csv

Case Study 3: Pandas, Seaborn

Problem Statement:

The dataset contains Happiness Score for 153 countries along with the factors used to explain the score. Create visualizations for the following using Seaborn.

- Scatter plot to show the relationship between GDP per capita and happiness score.
- Heat map to show the correlations between different variables.
- Bar plot to show the Happiness score for the top 10 countries.
- Box plot to show the distribution of each variable in the dataset.
- Pair Plot to show the relationship between each variable in the dataset.

Dataset: happiness_report.csv

Module 2: Statistics

Case Study 1: Hypothesis Testing

Use Chi Square Test to investigate whether there is a relationship between gender and smoking status in a population. (Take the significance level to be 0.05).

- **Null hypothesis (H0):** There is no association between gender and smoking status
- **Alternative hypothesis (Ha):** There is an association between gender and smoking status.

Dataset: tips.csv

Case Study 2: Hypothesis Testing

Effect of Website Speed on Conversion Rate

- Use descriptive statistics to explore the data and identify patterns or trends.
- Perform a hypothesis test to analyse whether there is a significant relationship between website speed and user behavior. Use a t-test to determine whether there is a significant difference in conversion rate between two groups of users with different page load times.

Dataset: website_speed.csv

- Page load time
- Bounce rate
- Conversion rate.

Module 3: Machine Learning 1

Case Study 1:

Heart Disease Classification using Logistic Regression

Train a Logistic Regression model to classify which patients are most likely to suffer from heart disease in the near future.

Dataset: Use the features given in the **heart-disease.csv** dataset.

Approach:

- Import the dataset and do the necessary pre-processing.
- Build a Logistic Regression model to classify a person as having heart disease or not.
- Evaluate the model using
    - Confusion matrix
    - Classification Report
    - Accuracy

Case Study 2:

Application of Linear Regression for Housing Price Prediction

Develop a Linear Regression model for predicting the housing price based on certain factors like house area, bedrooms, furnished, nearness to mainroad, etc.

Dataset: house_price.csv

Approach:

- Import the dataset and do the necessary pre-processing.
- Develop a Linear Regression model to predict prices.
- Evaluate the model using
    - R2 score
    - RMSE

Module 4: Machine Learning 2

Case Study 1: Random Forest

Insurance Premium prediction

Develop a Random Forest Regressor model based on several features of individual such as age, physical/family condition and location against their existing medical expense to be used for predicting future medical expenses of individuals that help medical insurance to make decision on charging the premium.

Dataset: insurance.csv

Case Study 2: SVM

Iris species - Classification

We are given a dataset containing information about different types of flowers (setosa, versicolor, and virginica). Our task is to build a machine learning classification model using SVC that can predict the type of the flower based on the given features.

Dataset: Iris.csv

Module 5: Time Series

Case Study 1:

Stock price prediction

Build an ARIMA model and perform Time Series analysis to predict the future stock prices using historical data.

Dataset: time_series_Microsoft_Stock.csv

The dataset contains the stock information of Microsoft from 04/01/2015 to 04/01/2021.

Case Study 2:

Time Series Forecasting of sales data

Develop a Time Series model using SARIMA that can forecast future sales for a given business based on historical data. The model should be able to identify patterns and trends in the data, consider any seasonality or trends, and make predictions for future 2 years.

Dataset: time_series_revenue.csv