DATA SCIENCE AND MACHINE LEARNING

CAPSTONE PROJECTS

DATA SCIENCE

- Data science is an interdisciplinary field that utilizes scientific methods, algorithms, processes, and systems to extract insights and knowledge from structured and unstructured data.
- It involves the application of statistical analysis, machine learning, data mining, and various other techniques to interpret complex data sets and solve analytical problems.
- Data scientists often work with large volumes of data to uncover patterns, trends, and correlations that can inform decision-making and drive business strategies across various industries.

MACHINE LEARNING

- Machine learning is a branch of artificial intelligence (AI) that focuses on creating systems and algorithms that can learn from and make predictions or decisions based on data, without being explicitly programmed to do so.
- In essence, it's about enabling computers to learn from experience and improve their performance over time. Machine learning algorithms can recognize patterns in data, make predictions, classify information, and optimize processes, among other tasks, by iteratively learning from data.

DATA COLLECTION METHODOLOGY

- **Objective**: Collect weather data for a specific location using a weather API.
- **Steps**:

1. **Sign Up and Get API Key**: Visit the website of a weather data provider, such as OpenWeatherMap or WeatherAPI, and sign up for an account to obtain an API key.
2. **Install Necessary Libraries**: Install any necessary libraries for making HTTP requests and handling JSON data in your preferred programming language (e.g., Python).
3. **Write Code to Make API Requests**: Write code to make HTTP requests to the weather API endpoint, passing your API key and the location you want to get weather data for.
4. **Handle API Responses**: Handle the API responses, parsing the JSON data returned by the API to extract the weather information you need (e.g., temperature, humidity, wind speed).

- **Handle API Responses**: Handle the API responses, parsing the JSON data returned by the API to extract the weather information you need (e.g., temperature, humidity, wind speed).
- **Store Data**: Store the collected weather data in a suitable data structure or format (e.g., CSV file, database) for further analysis or use.
- **Repeat as Needed**: Set up your code to run periodically or as needed to continuously collect updated weather data for the desired location.
- **Handle Errors and Edge Cases**: Implement error handling and consider edge cases, such as API rate limits, network errors, and missing data, to ensure the reliability and robustness of your data collection process.
- **Test and Debug**: Test your code thoroughly to ensure it collects the correct data and handles various scenarios correctly. Debug any issues that arise during testing.
- **Document Your Code**: Document your code, including comments and explanations of key steps and functions, to make it understandable and maintainable for yourself and others.
- **Deploy and Monitor**: Deploy your code in a suitable environment and set up monitoring to track its performance and ensure that it continues to collect data reliably over time.
- By following these steps, you can complete the data collection task using an API in a lab environment effectively.

PERFORM DATA WRANGLING

- Data wrangling, also known as data munging, is the process of cleaning, structuring, and enriching raw data into a format suitable for analysis. Here's a basic outline of how you might perform data wrangling:
- **Data Collection**
- **Data Inspection**
- **Data Cleaning**
- **Data Transformation**
- **Data Enrichment**
- **Iterative Process**

PERFORM EXPLORATARY DATA ANALYSIS USING VISUALIZATION AND SQL

- Certainly! Exploratory Data Analysis (EDA) is a crucial step in understanding your data before diving into more advanced analysis or modeling. Using both visualization and SQL can provide valuable insights into your dataset. Here's how you can perform EDA using both approaches:
- **Exploratory Data Analysis with Visualization**
- **Exploratory Data Analysis with SQL**
- By combining visualization techniques with SQL queries, you can effectively perform exploratory data analysis to gain insights into your dataset from multiple perspectives, facilitating better decision-making and hypothesis generation for subsequent analysis or modeling tasks.

Combine Visualization and SQL

DATA COLLECTION –SpaceX API

Github url : srinithinkothalanka (github.com)

DATA COLLECTION -SCRAPING

- **Web scraping (or data scraping) is a technique used to collect content and data from the internet.**
- This data is usually saved in a local file so that it can be manipulated and analyzed as needed.
- If you've ever copied and pasted content from a website into an Excel spreadsheet, this is essentially what web scraping is, but on a very small scale.

DATA WRANGLING

- Data Wrangling is the process of gathering, collecting, and transforming Raw data into another format for better understanding, decision-making, accessing, and analysis in less time. Data Wrangling is also known as Data Munging.

EDA AND INTERACTIVE VISUAL ANALYTIC METHODLOGY

- Exploratory Data Analysis (EDA) and Interactive Visual Analytics are both essential methodologies in the field of data science and analytics.

1. **Exploratory Data Analysis (EDA):** EDA is the process of analyzing data sets to summarize their main characteristics, often employing visual methods. The primary goal of EDA is to understand what the data can tell us beyond the formal modeling or hypothesis testing task. Techniques in EDA include summary statistics, visualization (histograms, box plots, scatter plots, etc.), and data cleaning.
2. **Interactive Visual Analytics:** This methodology combines visual representation of data with interactive capabilities to facilitate exploration and understanding. Interactive visual analytics tools allow users to manipulate and interact with data visualizations in real-time. Users can zoom, filter, drill-down, and perform other operations to gain deeper insights into the data. Interactive visual analytics often involves the use of tools like Tableau, Power BI, or custom-built applications.

PREDICTIVE ANAYLSIS METHODOLOGY

- Predictive analysis methodology involves using data, statistical algorithms, and machine learning techniques to identify the likelihood of future outcomes based on historical data. Here's a structured approach to predictive analysis:
- **Data Collection and Preparation**
- **Exploratory Data Analysis (EDA)**
- **Feature Engineering**
- **Model Selection**
- **Model Training**

EDA WITH VISUALIZATION RESULTS

- Certainly! Exploratory Data Analysis (EDA) with visualization results can provide valuable insights into the underlying patterns, relationships, and distributions within the data. Here's an example of an EDA process along with visualization results:

1. **Data Understanding:**
    1. Obtain a dataset containing information about customer transactions in a retail store.
2. **Initial Data Inspection:**
    1. Check the first few rows of the dataset to understand its structure and the type of information it contains.
    2. Look for missing values, outliers, and inconsistencies.
3. **Summary Statistics:**
    1. Calculate summary statistics for numerical variables such as mean, median, standard deviation, minimum, maximum, etc.
    2. For categorical variables, count the frequency of each category.
4. **Univariate Analysis:**
    1. Visualize the distribution of numerical variables using histograms, box plots, or density plots.
    2. Examine the frequency distribution of categorical variables using bar charts or pie charts.

EDA WITH SQL RESULTS

- Let's assume we have a database containing information about sales transactions in a retail store. We'll perform EDA using SQL queries:

1. **Data Understanding:**
    1. Understand the structure of the database tables and the type of information they contain.
2. **Initial Data Inspection:**
    1. Use SQL queries to retrieve the first few rows of each table to understand the data's structure and content.
3. **Summary Statistics:**
    3. Calculate summary statistics for numerical columns using SQL aggregate functions.

INTERACTIVE MAP WITH FOLIUM

- Folium is a Python library that allows you to create interactive maps using Leaflet.js, which is a leading open-source JavaScript library for interactive maps. Here's a basic example of how to create an interactive map using Folium:
- EXAMPLE PYTHON PROGRAM:

```
import folium
 m = folium.Map(location=[latitude, longitude],
zoom_start=10)
 folium.Marker([latitude1, longitude1],
popup='Marker 1').add_to(m)
folium.Marker([latitude2, longitude2],
popup='Marker 2').add_to(m)
folium.CircleMarker(
 location=[latitude3, longitude3],
 radius=50,
 popup='Circle Marker’,
 color='red’,
 fill=True,
fill_color='red’
).add_to(m)
 m.save('interactive_map.html')
```

THANK YOU

Edx