**skolar**

-sneha latha reddy .p

**Week 3** : ( part 2B )

python libraries:

**Assignment-1 :**

**Explore numpy official documentation . summarize the essiential concepts and features covered in the beginner’s guide**

NumPy (****Numerical Python****) is an open source Python library that’s used in almost every field of science and engineering. It’s the universal standard for working with numerical data in Python, and it’s at the core of the scientific Python and PyData ecosystems.

INSTALLING NUMPY:

To install NumPy, we strongly recommend using a scientific Python distribution. If you’re looking for the full instructions for installing NumPy on your operating system, see [Installing NumPy](https://numpy.org/install/).

If you already have Python, you can install NumPy with:

conda install numpy

or

pip install numpy

If you don’t have Python yet, you might want to consider using [Anaconda](https://www.anaconda.com/). It’s the easiest way to get started.

## How to import NumPy

To access NumPy and its functions import it in your Python code like this:

**import** numpy **as** np

We shorten the imported name to np for better readability of code using NumPy. This is a widely adopted convention that makes your

code more readable for everyone working on it.

****Arrays****: The fundamental data structure in NumPy is the **ndarray**, which is a multi-dimensional array. These arrays can be created using various methods like conversion from other Python structures (lists, tuples), initializing with specific values, or creating special types of arrays like zeros, ones, or identity matrices.

****Array Indexing****: NumPy provides powerful indexing techniques for accessing elements and subarrays within arrays. This includes basic indexing, slicing, boolean indexing, and fancy indexing.

****Array Operations****: NumPy allows for efficient element-wise operations on arrays, including arithmetic operations (+, -, \*, /), element-wise functions (sin, cos, exp), and aggregation functions (sum, mean, max, min).

****Broadcasting****: Broadcasting is a powerful mechanism in NumPy that allows arrays with different shapes to be combined in arithmetic operations. This can greatly simplify code and make it more readable.

****Array Manipulation****: NumPy provides functions for reshaping, joining, splitting, and other manipulations of arrays. These operations are essential for data preprocessing and manipulation in scientific computing tasks.

****Universal Functions (ufuncs)****: Universal functions (ufuncs) are functions that operate element-wise on arrays, supporting broadcasting and type casting. NumPy provides a wide range of ufuncs for mathematical operations, including trigonometric functions, exponential functions, and more.

****Input and Output****: NumPy provides functions for reading from and writing to disk in various formats, including text files, binary files, and NumPy's own **.npy** format.

****Linear Algebra****: NumPy provides a rich set of functions for linear algebra operations, such as matrix multiplication, matrix decompositions (e.g., SVD, LU decomposition), solving linear equations, and computing eigenvalues and eigenvectors.

****Random Number Generation****: NumPy includes a random module for generating random numbers from various probability distributions. This module is useful for generating random data for simulations and statistical analysis.

**pandas: investigate pandas official documentation . summerize the key aspects of pandas for beginners**

Pandas is an open-source library built on top of NumPy, providing high-performance, easy-to-use data structures and data analysis tools for Python.

****Data Structures****: Pandas primarily deals with two main data structures: Series and DataFrame

****Series****: A one-dimensional array-like object that can hold any data type.

****DataFrame****: A two-dimensional labeled data structure with columns of potentially different types. It can be thought of as a spreadsheet or SQL table.

****Reading and Writing Data****: Pandas provides functions to read data from various file formats such as CSV, Excel, SQL databases, JSON, and HTML. Similarly, it provides functions to write data to these formats.

****Data Exploration****: Pandas offers numerous functions and methods to explore and understand the data. This includes functions to display the first few rows (**head()**), last few rows (**tail()**), basic statistics (**describe()**), data types (**dtypes**), and information about the DataFrame (**info()**).

****Indexing and Selection****: Pandas allows for various methods of indexing and selecting data from Series and DataFrame objects. This includes indexing by label (**loc[]**), indexing by integer position (**iloc[]**), boolean indexing, and using conditional expressions.

****Data Manipulation****: Pandas provides a rich set of tools for data manipulation, including:

****Data Alignment****: Automatic data alignment across different Series and DataFrame objects.

****Missing Data Handling****: Functions for detecting, removing, and filling missing values (**isnull()**, **dropna()**, **fillna()**).

****Data Filtering and Sorting****: Methods for filtering rows based on conditions (**query()**), and sorting rows and columns (**sort\_values()**, **sort\_index()**).

****Data Aggregation and Grouping****: Functions for aggregating data based on groups (**groupby()**) and applying functions to groups (**agg()**).

****Applying Functions****: Applying custom functions to Series or DataFrame objects (**apply()**, **applymap()**).

****Data Visualization****: Pandas integrates with Matplotlib, a popular plotting library, to provide basic plotting functionality directly from DataFrames and Series objects. This allows for quick and easy visualization of data.

****Time Series and Categorical Data****: Pandas includes specialized data structures and functions for working with time series data (**Timestamp**, **DatetimeIndex**, **resample()**) and categorical data (**Categorical**, **cut()**).

****Performance****: Pandas is optimized for performance, especially when dealing with large datasets. It leverages vectorized operations and integrates seamlessly with NumPy for efficient computation.

****Community and Resources****: The Pandas community is active and supportive, with a wealth of resources available including documentation, tutorials, forums, and online courses to help beginners get started and advance their skills.

**Assignment - 02**

**scikit learning : review the skikit-learn tutorial for beginner summarize the basic concepts and functionalities provided by scikit learning**

scikit-learn is a powerful machine learning library in Python that provides a wide range of tools and algorithms for data mining and analysis. Here's a summary of the basic concepts and functionalities covered in the scikit-learn tutorial for beginners:

****Introduction to scikit-learn****: The tutorial starts with an overview of scikit-learn, highlighting its key features and advantages. It emphasizes scikit-learn's user-friendly interface, extensive documentation, and active community support.

****Data Representation in scikit-learn****: scikit-learn represents data in the form of arrays or matrices. Features are typically stored as columns in a 2D array (often referred to as X), while target variables are stored in a separate 1D array (often referred to as y).

****Loading Datasets****: scikit-learn provides built-in datasets for practice and experimentation, including toy datasets like Iris and digits, as well as larger real-world datasets. These datasets can be loaded using dedicated functions like ‘**load\_iris()’** or ‘**load\_digits()’**.

****Data Preprocessing****: Before applying machine learning algorithms, it's essential to preprocess the data. scikit-learn provides various preprocessing techniques such as scaling, normalization, encoding categorical variables, handling missing values, and feature selection.

****Supervised Learning Algorithms****: scikit-learn offers a wide range of supervised learning algorithms for classification and regression tasks. Some popular algorithms include:

**Linear Models:** Linear Regression, Logistic Regression

Nearest Neighbors: K-Nearest Neighbors

Tree-Based Models: Decision Trees, Random Forests

Support Vector Machines (SVM)

Ensemble Methods: Gradient Boosting, AdaBoost

****Unsupervised Learning Algorithms****: scikit-learn also provides algorithms for unsupervised learning tasks such as clustering and dimensionality reduction. Some common algorithms include:

Clustering: K-Means, DBSCAN, Hierarchical clustering

Dimensionality Reduction: Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE)

****Model Training and Evaluation****: scikit-learn follows a consistent API for model training and evaluation. Models are trained using the ‘**fit()’** method, and predictions are made using the ‘**predict()’** method. Additionally, scikit-learn provides tools for evaluating model performance, including metrics like accuracy, precision, recall, F1-score, and ROC-AUC.

****Cross-Validation****: scikit-learn offers utilities for cross-validation, a technique used to assess a model's performance. Cross-validation helps to estimate how well a model will generalize to unseen data by splitting the dataset into multiple subsets and evaluating the model on each subset.

****Hyperparameter Tuning****: scikit-learn provides tools for hyperparameter tuning, including techniques like Grid Search and Random Search, to find the optimal hyperparameters for machine learning models.

****Integration with Other Libraries****: scikit-learn integrates well with other Python libraries such as NumPy, Pandas, and Matplotlib, allowing for seamless data manipulation, analysis, and visualization in machine learning workflows.

**Matplotlib and Seaborn** : compare matplot lib and seaborn for data visualization ,summarize the strength and weeknesses of each library and when to use them

Matplotlib and Seaborn are both popular Python libraries used for data visualization, but they have different strengths and weaknesses. Here's a comparison of the two:

****Matplotlib:****

****Strengths:****

\* Matplotlib is a versatile library that provides a wide range of plotting functions, allowing for highly customizable plots.

\* It offers fine-grained control over every aspect of the plot, including line styles, markers, colors, and annotations.

\* Matplotlib is well-integrated with other Python libraries, making it easy to incorporate plots into larger projects.

\* It has been around for a long time and has a large user base, resulting in extensive documentation and community support.

****Weaknesses:****

\* Matplotlib can have a steep learning curve for beginners due to its low-level syntax and extensive customization options.

\* Creating complex plots from scratch can be time-consuming, especially for beginners.

\* Matplotlib's default styles are considered somewhat basic, requiring additional customization for aesthetically pleasing plots.

****When to Use:****

\* Matplotlib is suitable for users who need full control over their plots and are willing to invest time in customization.

\* It is a good choice for creating publication-quality plots and embedding them in reports or presentations.

\* Matplotlib is also useful for creating interactive plots in Jupyter notebooks or web applications.

****Seaborn:****

****Strengths:****

\* Seaborn is built on top of Matplotlib and provides a high-level interface for creating attractive statistical visualizations.

\* It offers simplified functions for common plot types, such as scatter plots, histograms, box plots, and heatmaps, with sensible default settings.

\* Seaborn has built-in support for visualizing statistical relationships in datasets, such as linear regression plots, pair plots, and categorical plots.

\* It comes with several stylish themes and color palettes that enhance the aesthetics of plots without requiring manual customization.

****\* Weaknesses:****

\* Seaborn's focus on statistical visualization may limit its flexibility for creating non-standard plots or custom visualizations.

\* While Seaborn provides high-level abstractions, users may sometimes need to revert to Matplotlib for fine tuning or customizations beyond Seaborn's capabilities.

\* Seaborn's integration with other libraries may not be as seamless as Matplotlib, particularly for more complex plotting requirements.

****\*When to Use:****

\* Seaborn is ideal for users who want to quickly create visually appealing plots with minimal code and effort.

\* It is well-suited for exploratory data analysis and visualizing relationships between variables in datasets.

\* Seaborn is particularly useful for users working with statistical data and seeking to visualize distributions, correlations, and trends effectively.