**Different Data types in Python**

* **Numeric Types**: int, float, complex
* **Sequence Types**: list, tuple, range
* **Text Type**: str
* **Mapping Type**: dict
* **Set Types**: set, frozenset
* **Boolean Type**: bool
* **Binary Types**: bytes, bytearray, memoryview
* **None Type**: NoneType

**Difference between List and Tuple**

* **List**:
  + Mutable: Can be changed after creation.
  + Syntax: [1, 2, 3]
  + Slower due to extra functionality.
  + Example: my\_list = [1, 2, 3]
* **Tuple**:
  + Immutable: Cannot be changed after creation.
  + Syntax: (1, 2, 3)
  + Faster due to immutability.
  + Example: my\_tuple = (1, 2, 3)

**What is a Linked List?**

A linked list is a linear data structure where elements are stored in nodes. Each node contains a data part and a reference (or link) to the next node in the sequence. Unlike arrays, linked lists do not have a fixed size.

**Explain List Comprehension**

List comprehension provides a concise way to create lists. It consists of brackets containing an expression followed by a for clause, then zero or more for or if clauses. The expressions can be anything, meaning you can put in all kinds of objects in lists.

Example:

squares = [x\*\*2 for x in range(10)]

**What is a Set?**

A set is an unordered collection of unique elements. Sets are mutable, but since they are unordered, they do not support indexing, slicing, or other sequence-like behavior.

Example:

my\_set = {1, 2, 3, 4}

**What is Shallow Copy?**

A shallow copy creates a new object, but inserts references into it to the objects found in the original. It copies the top-level structure, but not the nested objects.

Example:

import copy

original = [1, 2, [3, 4]]

shallow\_copy = copy.copy(original)

**What is a Deep Copy?**

A deep copy creates a new object and recursively adds copies of nested objects found in the original.

Example:

import copy

original = [1, 2, [3, 4]]

deep\_copy = copy.deepcopy(original)

**What is a Queue?**

A queue is a collection of elements that follows the First In First Out (FIFO) principle. Elements are added from the rear and removed from the front.

Example using collections.deque:

from collections import deque

queue = deque([1, 2, 3])

queue.append(4) # Add to the rear

queue.popleft() # Remove from the front

**What is Mutable and Immutable in Python?**

* **Mutable**: Objects that can be changed after creation. Examples: list, dict, set.
* **Immutable**: Objects that cannot be changed after creation. Examples: int, float, str, tuple.

**What are NumPy and Pandas?**

* **NumPy**: A library for numerical computing with powerful array objects.
* **Pandas**: A library for data manipulation and analysis, built on top of NumPy. It provides data structures like DataFrame.

**What is an Array?**

An array is a data structure that stores a collection of items, typically of the same type, at contiguous memory locations.

**Different types of slicing in Pandas**

* **Single-axis slicing**:

df.loc['row\_index'] # Row by label

df.iloc[0] # Row by index

* **Multi-axis slicing**:

df.loc['row\_index', 'column\_label'] # Row and column by label

df.iloc[0, 1] # Row and column by index

**What is AI?**

Artificial Intelligence (AI) is the simulation of human intelligence processes by machines, especially computer systems. It includes learning, reasoning, and self-correction.

**What is the Algorithm in AI?**

An algorithm in AI is a set of rules or calculations used to solve problems or perform tasks, typically involving data processing and decision making.

**How to Handle Missing Values?**

* **Remove**: Drop rows or columns with missing values.
* **Impute**: Fill missing values with mean, median, mode, or other methods like forward fill or backward fill.
* **Predict**: Use machine learning models to predict and fill missing values.

**What are Sampling and Outliers?**

* **Sampling**: Selecting a subset of data from a larger dataset for analysis.
* **Outliers**: Data points that differ significantly from other observations. They can be treated by removal, transformation, or capping.

**What is Imbalanced Data and How to Handle It?**

Imbalanced data occurs when classes in a dataset are not equally represented. It can be handled by:

* **Resampling**: Oversampling the minority class or undersampling the majority class.
* **Synthetic Data Generation**: Using techniques like SMOTE to create synthetic samples for the minority class.
* **Algorithm Adjustment**: Using algorithms that handle imbalance better or adjusting class weights.

**Types of ML Algorithms Used**

* **Supervised Learning**: Linear regression, logistic regression, decision trees, SVM, etc.
* **Unsupervised Learning**: K-means clustering, hierarchical clustering, PCA, etc.
* **Reinforcement Learning**: Q-learning, Deep Q Networks, etc.

**Difference between Analytics & Analysis**

* **Analytics**: The discovery, interpretation, and communication of meaningful patterns in data.
* **Analysis**: The process of examining data to draw conclusions.

**Types of Outlier Detection Methods**

* **Statistical Methods**: Z-score, IQR method.
* **Machine Learning Methods**: Isolation Forest, DBSCAN.
* **Visualization Methods**: Box plots, scatter plots.

**What are Outliers, Should They Be Treated on Healthcare Domain Data?**

Outliers are data points significantly different from others. In healthcare data, outliers should be carefully analyzed before treatment, as they might indicate important anomalies.

**If the Dataset is Skewed What Kind of Treatment Will You Do?**

* **Transformation**: Log, square root, or Box-Cox transformations to normalize data.
* **Resampling**: Use stratified sampling to balance the dataset.

**How Does Linear Regression Work?**

Linear regression models the relationship between a dependent variable and one or more independent variables by fitting a linear equation.

**How Does Logistic Regression Work?**

Logistic regression models the probability of a binary outcome based on one or more predictor variables using a logistic function.

**Why is Linear Regression Best?**

Linear regression is simple, interpretable, and works well when the relationship between variables is linear. It also has low variance and high bias, which can prevent overfitting.

**What is Underfitting & Overfitting?**

* **Underfitting**: A model too simple to capture the underlying patterns in data.
* **Overfitting**: A model too complex, capturing noise along with the underlying patterns.

**Differentiation between Accuracy, F1 Score, Precision, Recall**

* **Accuracy**: The ratio of correctly predicted observations to the total observations.
* **Precision**: The ratio of true positive observations to the predicted positive observations.
* **Recall**: The ratio of true positive observations to all the actual positive observations.
* **F1 Score**: The harmonic mean of precision and recall.

**What are Multiclass and Multilabel Problems?**

* **Multiclass**: Classification problems with more than two classes.
* **Multilabel**: Each sample can belong to multiple classes.

**What is a Positive Class and a Negative Class in Machine Learning?**

In binary classification, the positive class is the class of interest (usually labeled as 1), and the negative class is the other class (usually labeled as 0).

**What is Ridge and Lasso Regression?**

* **Ridge Regression**: Adds L2 regularization term to the loss function to penalize large coefficients.
* **Lasso Regression**: Adds L1 regularization term to the loss function, which can shrink coefficients to zero, effectively performing feature selection.

**What is a Correlation Heatmap?**

A graphical representation of the correlation matrix, where colors represent the strength and direction of correlations between variables.

**Explain Distribution Plots**

Distribution plots (e.g., histograms, KDE plots) visualize the distribution of data points within a dataset, showing frequencies or probabilities.

**Difference between Seaborn and Matplotlib**

* **Matplotlib**: A low-level plotting library, highly customizable.
* **Seaborn**: Built on top of Matplotlib, provides high-level interface for drawing attractive and informative statistical graphics.

**What is Correlation and How it Ranges?**

Correlation measures the strength and direction of the linear relationship between two variables. It ranges from -1 to 1, where -1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 indicates no correlation.

**Why Do We Have to Balance Data When It Has a 70/30 Ratio?**

Balancing data helps improve the performance of machine learning models, as imbalanced data can lead to biased predictions towards the majority class.

**What is the F1 Score?**

The F1 Score is the harmonic mean of precision and recall, providing a single metric to balance the two, especially useful in imbalanced datasets.

**If the Dataset is Skewed What Kind of Treatment Will You Do?**

For a skewed dataset, you can use log transformation, square root transformation, or other methods to normalize the distribution.

**For Categorical Data Which Type of Imputation?**

* **Mode Imputation**: Replace missing values with the most frequent value (mode).
* **Category Imputation**: Use 'missing' as a new category.

**What is the Random State in a Model?**

Random state is a seed value used to initialize the random number generator, ensuring reproducibility of results.

**Why is the Standard Scaler Used and What's the Formula and What is the Scaled Value Range?**

The Standard Scaler standardizes features by removing the mean and scaling to unit variance.

* **Formula**: z=(x−μ)σz = \frac{(x - \mu)}{\sigma}z=σ(x−μ)​
* **Scaled Value Range**: Typically ranges from -3 to 3.

**Difference between SVM and Logistic Regression**

* **SVM**: Finds the optimal hyperplane that maximizes the margin between classes.
* **Logistic Regression**: Estimates probabilities using a logistic function to classify data points.

**What is the CAP Theorem?**

The CAP theorem states that in a distributed data store, it is impossible to simultaneously provide all three of the following guarantees:

* **Consistency**: Every read receives the most recent write.
* **Availability**: Every request receives a response, without guarantee that it contains the most recent write.
* **Partition Tolerance**: The system continues to operate despite network partitions.

**How Will You Predict the Price of iPhone in E-commerce Sites - Explain with Block Diagram**

1. **Data Collection**: Collect historical price data, features like specifications, demand, reviews.
2. **Data Preprocessing**: Clean, handle missing values, normalize data.
3. **Feature Selection**: Select relevant features affecting price.
4. **Model Building**: Use algorithms like linear regression, decision trees.
5. **Model Evaluation**: Evaluate model using metrics like RMSE, MAE.
6. **Prediction**: Deploy the model to predict future prices.

**Which Distance Measure Do We Use in Case of Categorical Variables in K-NN?**

* **Hamming Distance**: Measures the distance between two categorical vectors.

**Explain About Exception Handling**

Exception handling manages errors gracefully using try, except, else, and finally blocks.

try:

# Code that might raise an exception

except SomeException:

# Code to handle the exception

else:

# Code to run if no exceptions were raised

finally:

# Code to run no matter what

**Explain About the SQL Constraints**

* **PRIMARY KEY**: Uniquely identifies each record in a table.
* **FOREIGN KEY**: A field in a table that is a primary key in another table, creating a relationship.
* **UNIQUE**: Ensures all values in a column are different.
* **NOT NULL**: Ensures a column cannot have NULL value.
* **CHECK**: Ensures all values in a column satisfy a specific condition.
* **DEFAULT**: Sets a default value for a column if no value is specified.

**Explain About Data Architecture**

Data architecture involves the design, management, and optimization of data systems and workflows, ensuring data is properly stored, accessed, and utilized. Components include databases, data warehouses, data lakes, and data pipelines.

**Different Types of Databases**

* **Relational Databases**: SQL databases like MySQL, PostgreSQL.
* **NoSQL Databases**: MongoDB, Cassandra.
* **In-Memory Databases**: Redis, Memcached.
* **Graph Databases**: Neo4j, Amazon Neptune.
* **Time-Series Databases**: InfluxDB, TimescaleDB.

**Explain About ETL**

ETL (Extract, Transform, Load) is a data integration process:

* **Extract**: Gather data from various sources.
* **Transform**: Convert data to the desired format.
* **Load**: Load transformed data into a target database or data warehouse.

**Explain Views and Keys in SQL**

* **Views**: Virtual tables representing a subset of data from one or more tables. They simplify complex queries and enhance security.
* **Keys**:
  + **Primary Key**: Uniquely identifies each record.
  + **Foreign Key**: Creates relationships between tables.
  + **Unique Key**: Ensures all values in a column are unique.
  + **Composite Key**: Combines two or more columns to create a unique identifier.

**What is Dropout in Computer Vision?**

Dropout is a regularization technique where randomly selected neurons are ignored during training, preventing overfitting by making the network less sensitive to specific neurons.

**What are Time and Space Complexity?**

* **Time Complexity**: The computational complexity that describes the amount of time it takes to run an algorithm.
* **Space Complexity**: The computational complexity that describes the amount of memory space an algorithm uses during execution.