ASSIGNMENT-1

SHRI SASHMITHA S

AIDS

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| CITY | COUNTRY | CONTINENT | POPULATION  (2024) | POPULATION  (2023) | GROWTH RATE |
| Tokyo | Japan | Asia | 37115035 | 37194105 | -0.0021 |
| Delhi | India | Asia | 33807403 | 32941309 | 0.0263 |
| Shanghai | China | Asia | 29867918 | 29210808 | 0.0225 |
| Dhaka | Bangladesh | Asia | 10000 | 23209616 | 0.0313 |
| Sao Paulo | Brazil | South America | 22806704 | 22619736 | 0.0083 |
| Cairo | Egypt | Africa | 22623874 | 22183201 | 0.0199 |
| Mexico | Mexico | North America | 22505315 | 22281442 | 25 |

**1.FEATURE –** Feature represent the input data or independent variables used to describe various aspects of the object which is under study. (For example city,country,continent etc..are the features,attributes or dimensions.)

2**.LABEL** – A label is the target variable or dependent variable. They are known outcomes associated with the input data used for training the model. (For example growth rate in the above dataset.)

**3.PREDICTION -** Refers to the process of using a trained model to estimate the output or target variable for new, unseen data based on the patterns it has learned from the training data.( For example predicting the growth rate of a city from the population growth.)

**4.OUTLIER –** The outlier is the data point that varies significantly from the rest of the observations.(for example the valu 10000 in population (2024) attribute and 25 in growth rate attribute is examples of outliers.)

**5.TEST DATA –** The test data is the new unseen data used for testing the trained model which is similar to training data in the distribution of features. It is used to check whether the model predicts the label correctly or not.

**6.TRAINING DATA –** It is the data used to train the machine learning model to predict the outcome. The data that is given as input to make the ML model to learn.

7.**MODEL –** A model refers to mathematical representation or algorithm that is trained to make predictions based on the data.

8.**VALIDATION DATA –** A validation data refers to the part of data taken from training data for model evaluation and hyper parameters fine-tuning.

9.**HYPERPARAMETER -** Hyperparameters are the parameters set before the learning process begins and govern the training of machine learning models. They are not learned from the data but are crucial for controlling the behavior and performance of the model.

10.**EPOCH -** An epoch in the context of training neural networks is a single pass through the entire training dataset.

11.**LOSS FUNCTION –** The loss function quantifies the difference between the predicted output values and the actual target values.

12.**LEARNING RATE -** The learning rate is a crucial hyperparameter in machine learning, especially in training neural networks.It determines the size of the steps the model takes towards the minimum of the loss function.

13.**OVERFITTING -** Overfitting occurs when a machine learning model learns the training data too well, capturing noise and outliers in addition to the underlying patterns. This results in a model that performs well on the training data but poorly on test data.

14.**UNDERFITTING –** Underfitting occurswhen the machine learning model doesn’t learn the underlying pattern in the training data which results in the poor performance on the new unseen data or test data.

15.**REGULARIZATION -** Regularization is a technique used to prevent overfitting by adding a penalty term to the loss function. The penalty term discourages the model from fitting the noise in the training data.

16.**CROSS VALIDATION -** Cross-validation is a statistical method used to evaluate and compare the performance of machine learning models. It helps ensure that the model generalizes well to unseen data by using different subsets of the data for training and validation.

17.**FEATURE ENGINEERING** - It is the process of transforming raw data into useful features such that it is suitable for training a machine learning model.

18.**DIMENSIONALITY REDUCTION** – Dimensionality reduction is the process of reducing the number of features in the dataset without affecting the integrity of the data.

19**.BIAS**- A bias is the difference between the average prediction and the actual output values.

20.**VARIANCE** – Variance is the amount that the prediction will change if different training datasets were used.It measures how scattered are the predicted values from the correct value due to different training datasets.