A logo of a company

Description automatically generated A close-up of a logo

Description automatically generated

**Joint Tech Internship Community Program**

**TASK - 1**

**Candidate ID:2024060014**

**Candidate Name: Janarthikan V**

**Submitted by Janarthikan.v**

**(janarthikanvivek23@gmail.com)**

**List of Terminologies:**

**Feature**

* A feature is an individual measurable property or characteristic of a phenomenon being observed. In machine learning, features are the input variables used to make predictions.
* Features are crucial because the quality and relevance of the features used can significantly impact the performance of a machine learning model.
* Good features enable a model to learn better patterns, while irrelevant or poor-quality features can lead to poor model performance.

Types of Features:

* Numerical Features
* Categorical Features
* Ordinal Features
* Boolean Features

**Labels**

The label is the output variable that the machine learning model is trained to predict. It is also known as the dependent variable, target, or outcome variable.

**Characteristics of Labels**:

1. **Known Outcomes**: During training, the labels are known values that correspond to the input features.
2. **Ground Truth**: Labels represent the ground truth against which the model’s predictions are compared.

**Prediction**

* The output of a machine learning model. It is the estimated label produced by the model for given input features.

**Outliers**

* An outlier is an observation that is substantially different from the other observations.

Types of Outliers:

* Univariate Outliers
* Multivariate Outliers

**Test Data**

* A subset of the dataset used to evaluate the performance of a trained machine learning model.
* This data is not used during the training process.

**Training Data**

The subset of the dataset used to train the machine learning model. It includes both the input features and their corresponding labels.

**Model**

* ML model is a specific instance of machine learning algorithm that has been trained on a set of data. The specific parameter and relationship between variables learned by the model during training are stored in the model. The model are used to make predications of new model.

**Validation Data**

* A subset of the dataset used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters.

**Hyperparameter**

**Model Hyperparameters**: These define the model architecture and complexity, such as the number of layers in a neural network or the number of trees in a random forest.

**Training Hyperparameters**: These control the training process, such as the learning rate, batch size, and the number of epochs.

**Epoch**

* One complete pass through the entire training dataset. In neural networks, multiple epochs are typically used to train the model.

**Loss Function**

* A function that measures how well the model's predictions match the true labels. The goal of training a model is to minimize the loss function.

**Learning Rate**

* The learning rate is a hyperparameter that controls how much to change the model in response to the estimated error each time the model weights are updated. It determines the step size at each iteration while moving toward a minimum of the loss function.

**Overfitting**

Overfitting happens when a model learns the training data too well, including the noise, causing it to perform poorly on new data.

**Underfitting**

Underfitting occurs when a model is too simple to capture the underlying structure of the data. It fails to learn the training data well enough and also performs poorly on new data.

**Regularization**

* Regularization prevents overfitting by adding a penalty for large coefficients. For example Ridge regression adds a penalty proportional to the square of the coefficients.

**Cross Validation**

* The cross validation is a very important technique that we use in machine learning in order to get a more reliable evaluation metric such as an accuracy score .

**Feature Engineering**

* Feature engineering creates new features from raw data to improve model performance.
* For example: extracting the hour of the day from a timestamp.

**Dimensionality Reduction**

* Dimensionality reduction decreases the number of features while keeping important information.
* For example:- using Principal Component Analysis (PCA) to simplify a dataset.

**Bias**

* Bias is an error from making overly simplistic assumptions in a model.
* For example: using a linear model for a non-linear relationship.

**Variance**

* Variance measures a model's sensitivity to small fluctuations in the training dataset.
* For example:- a highly complex model like an overfitted decision tree might perform poorly on new data.