A logo of a company

Description automatically generatedA close-up of a logo

Description automatically generated

**Joint Tech Internship Community Program**

**ASSIGNMENT 1**

**Name:** Indhuja T

**Roll No:** 22CSR072

**Email:**indhujathamarai@gmail.com

**Example:** Predicting Student Grades

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STUDENT ID** | **HOURS STUDIED** | **ATTENDANCE**  **(%)** | **HOME WORK SCORE(%)** | **GRADE** |
| 1 | 5 | 90 | 80 | A |
| 2 | 3 | 80 | 70 | B |
| 3 | 7 | 95 | 90 | A |
| 4 | 2 | 75 | 60 | C |
| 5 | 6 | 85 | 85 | A |

**Terminologies Explanation:**

**Feature:**

An attribute or characteristic of the data used by the model to make predictions. Features are the input variables used to make predictions.

**Example:**Student id, Hours studied, Attendance, Home work scoreare features here.

**Label:**

The target or response variable that the model is trying to predict.

**Example:** **Grade** is the Label.

**Prediction:**

The output of a model that is generated based on the input features.

**Example:**A model predicts that a student who studied for 4 hours, had an attendance of 85%, and a homework score of 80 will get a grade of B.

**Outlier:**

An outlier is a data point that differs significantly from other observations. Outliers can occur due to variability in the data or errors in the data collection process.

**Example:** A student who studied for 10 hours and got a grade of C is an outlier, as most students who study for 10 hours get higher grades.

**Test Data:**

Test data is a subset of the data that is used to evaluate the performance of the trained model.

**Example:** 20% of the student dataset is set aside as test data to evaluate the model's performance.

**Training data:**

Training data is the subset of the data used to train the model.

**Example:** 80% of the student dataset is used to train the model.

**Model:**

A model in machine learning is a mathematical representation of a real-world process. It is created by training an algorithm on a dataset and is used to make predictions or decisions based on new input data.

**Validation Data:**

Validation data is a subset of the data used to tune the model's hyperparameters and to prevent overfitting. It is separate from both the training and test data.

**Example:** 10% of the student dataset is set aside as validation data to tune the model's hyperparameters.

**Hyperparameter:**

A parameter that is set before training a model, such as learning rate or number of epochs.

**Example:** The learning rate of the model is set to 0.01 and the number of epochs is set to 100.

**Epoch:**

An epoch is one complete pass through the entire training dataset.

**Example:** The model is trained for 100 epochs, with each epoch processing the entire training dataset.

**Loss Function:**

A mathematical function that measures the difference between the model's predictions and the actual labels.

**Learning rate:**

A hyperparameter that controls how quickly the model learns from the data.

**Example:** A learning rate of 0.01 means that the model will update its parameters by 1% of the gradient at each step.

**Overfitting:**

When a model is too complex and performs well on the training data but poorly on new, unseen data.

**Example:** A model that is trained on a small dataset of students and performs well on the training data but poorly on new data is overfitting.

**Underfitting:**

When a model is too simple and performs poorly on both the training data and new, unseen data.

**Example:** A model that is trained on a large dataset of students but performs poorly on both the training data and new data is underfitting.

**Regularization:**

A technique used to prevent overfitting by adding a penalty term to the loss function. **Example:** L1 regularization

**Cross validation:**

A technique used to evaluate the performance of a model by training and testing it on multiple subsets of the data.

**Feature Engineering:**

The process of selecting and transforming raw data into features that are suitable for modelling **.**

**Example:** Feature engineering is used to transform the raw data of student grades into features such as Hours Studied, Attendance, and Homework Score.

**Dimensionality reduction:**

The process of reducing the number of features in a dataset to prevent the curse of dimensionality.

**Example:** Principal component analysis (PCA) is used to reduce the number of features in the dataset.