

ASSIGNMENT 1:

SUBMITTED BY,

SAFRIN BARGANA A

22CSR166

III-BE-CSE-C

DATA SET ON STUDENT DETAILS

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ROLL NO | NAME | GRADE | AGE | MAJOR | GPA | CREDITS EARNED |
| 22CSR165 | KHAN | B | 19 | IT | 8.11 | 15 |
| 22CSR166 | SAFRIN | O | 20 | CSE | 9.19 | 24 |
| 22CSR167 | JAAN | C | 20 | EIE | 5.00 | 7 |
| 22CSR168 | INAYAH | A | 21 | ECE | 9.10 | 23 |
| 22CSR169 | ABSIN | A+ | 22 | CIVIL | 8.77 | 20 |

TERMINOLOGIES

1.FEATURES

A feature is a measurable property of some data-sample that is used as input for a model for training and serving. A feature should have predictive power for the model it is being used in.

**Example**: Roll no, Name, Grade, Age, Major, GPA, Credits earned.

2.LABEL

1.The label is the target variable that the model is trying to predict.

2. For instance, if predicting the GPA based on other features, the GPA would be the label.

3.PREDICTION

1.Prediction is the outcome of model from the data give.

2.In detail, Student’s GPA can be calculated using credits earned and grade points.

4.OUTLIER

An outlier is a data point significantly different from other data points in a dataset.

**Example:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ROLL NO | NAME | Grade | AGE | MAJOR | GPA | CREDITS EARNED |
| 22CSR167 | JAAN | C | 20 | EIE | 5.00 | 7 |

5.TEST DATA

A test dataset is a collection of data points (SUBSET OF DATASET) that the model hasn’t seen during its training process.

**Example**:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ROLL NO | NAME | GRADE | AGE | MAJOR | GPA | CREDITS EARNED |
| 22CSR169 | ABSIN | A+ | 22 | CIVIL | 8.77 | 20 |

6.TRAINING DATA

Training data is large dataset that is used to teach a machine learning model. Training data is used to teach prediction models on how to extract features that are relevant to required outcome.

**Example:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ROLL NO | NAME | GRADE | AGE | MAJOR | GPA | CREDITS EARNED |
| 22CSR166 | SAFRIN | O | 20 | CSE | 9.19 | 24 |
| 22CSR167 | JAAN | C | 20 | EIE | 5.00 | 10 |
| 22CSR168 | INAYAH | A | 21 | ECE | 9.10 | 23 |

7.MODEL

A machine learning model to recognize patterns or behaviours based on previous experience or data works by analysing a dataset and learning the relationships between the features and the target variable.

8.VALIDATION DATA

Validation data acts like a checkpoint between training and testing. It helps you fine-tune your model's settings to find the best configuration before you test it on new data. Validation data is distinguish from test data and training data.

**Example**:

Validation data (Roll No 22CSR169) is used to fine-tune the model's parameters, ensuring it performs optimally before being tested on new, unseen data (Roll No 22CSR167).

9.HYPERPARAMETER

Hyperparameters in Machine learning are those parameters that are explicitly defined by the user to control the learning process**.** These hyperparameters are used to improve the learning of the model, and their values are set before starting the learning process of the model.

10.EPOCH

An epoch is a complete **iteration** through the entire training dataset in one cycle for training the machine learning model.

**Example**:

The entire student management dataset trained at on time that is one epoch.

11.Loss function

The loss function also referred to as the error function, is a crucial component in machine learning that quantifies the difference between the predicted outputs of a machine learning algorithm and the actual target values.

12.Learning rate

The learning rate is a hyper-parameter used to govern the pace at which an algorithm updates or learns the values of a parameter estimate.

13.Overfitting

Overfitting is an undesirable machine learning behaviour that occurs when the machine learning model gives accurate predictions for training data but not for new data(overfit model can give inaccurate predictions and cannot perform well for all types of new data.).

14.Underfitting

When a model has not learned the patterns in the training data well and is unable to generalize well on the new data, it is known as underfitting. An underfit model has poor performance on the training data and will result in unreliable predictions.

15.Regularization

Regularization is a technique to prevent the model from overfitting by adding extra information to it.

16.Cross-validation

Cross validation is a technique used in machine learning to evaluate the performance of a model on unseen data. It involves dividing the available data into multiple folds or subsets, using one of these folds as a validation set, and training the model on the remaining folds. This process is repeated multiple times, each time using a different fold as the validation set. Finally, the results from each validation step are averaged to produce a more robust estimate of model’s performances.

Steps involved:

* **Split the Data**
* Train and test
* Average results

17.Feature engineering

**Feature engineering is the pre-processing step of machine learning, which extracts features from raw data**. It helps to represent an underlying problem to predictive models in a better way, which as a result, improve the accuracy of the model for unseen data. The predictive model contains predictor variables and an outcome variable, and while the feature engineering process selects the most useful predictor variables for the model.

18.Dimensionality reduction

Dimensionality reduction is a process and technique to reduce the number of dimensions or features in a dataset. The goal of dimensionality reduction is to decrease the data set's complexity by reducing the number of features while keeping the most important properties of the original data.

19.Bias

Bias is simply defined as the inability of the model because of that there is some difference or error occurring between the model’s predicted value and the actual value. These differences between actual or expected values and the predicted values are known as error or bias error or error due to bias. Bias is a systematic error that occurs due to wrong assumptions in the machine learning process.

20.Variance

Variance in machine learning refers to the variability of model predictions for a given input. It quantifies how much the predictions of a model fluctuate when trained on different subsets of the training data.