|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Name** | **Inmates Gender** | **Bed Room Per Room** | **Wi-fi** | **Type** | **Sports equipments** | **Veg** | **Non-Veg** | **Price/Year** |
| 1 | Ganga | Male | 4 | Yes | Private | Yes | Yes | Yes | 2,00,000 |
| 2 | Erode Stays | Female | 3 | Yes | Govt | No | Yes | No | 75,000 |
| 3 | Girls Nest | Female | 4 | No | Govt | Yes | Yes | No | 70,000 |
| 4 | Amirtha  Hostels | Male | 3 | No | Private | No | Yes | Yes | 75,000 |
| 5 | Traditional  Hostel | Male | 2 | No | Private | Yes | Yes | Yes | 1,00,000 |
| 6 | 1980 hostel | Male | 1 | No | Private | No | Yes | Yes | 1,05,000 |

Terminologies and examples

1. **Feature:** These are the individual measurable properties and like parameters (Eg. Id, Name, Inmates Gender …., Non-veg).
2. **Label :** The value the model aims to predict in this case price/year is the label the model aims to predict with the given values.
3. **Outlier:** The data point whose value is significantly different from other records (Eg. Price/Year in the id 1).
4. **Prediction:** The trained model with above mentioned features to derive to a solution which is label with new values, i.e. the model gives us price/year for the new value of features like (7, Sam Hostel, Female, 2, No, Private, Yes, Yes , Yes ).
5. **Outliers:** The certain value of price/year is unusually high or low without following the trend of all other data is called outliers (Eg. Data at id 1 is an outlier which has unusual high data when the model learns the relationship between each feature to come up with the price/year.
6. **Test Data:** Test Data is a set of data which are not used when training model but which will be used after model gets trained to test how accurate the model produces the required data which is label with some actual data.
7. **Model:** Model is an algorithm trained using set of features and labels to find a set of rules, patterns to come up with the required value when new values are given to the model.
8. **Validation Data:** These are a subset of data set aside from training data and testing data to check for the correctness of the model among various models before evaluating the model with the test data.
9. **Hyperparameters:** These are the external parameters used in the model training which is different from the training data but this parameter is considered as configurations before starting to train the model which is crucial for model accuracy.
10. **Epoch:** This refers to the one complete pass through the training data when the model goes through each data to learn the pattern among the given data(Eg. In this case let’s say data from id 2 to 5 may be considered training data).
11. **Loss Function:** Loss function is a mathematical expression used to find how well or bad a model’s performance by following certain methods like Mean Squared Error (MSE) which is sum of the square of the difference between actual value of the given features and the predicted value by total number of records.
12. **Learning Rate:** It is the crucial hyperparameter used to assess how the model’s loss function value changes when we change specific value of the input parameters, it is called gradient of loss function per each parameter in the dataset.
13. **Overfitting:** Overfitting is a problem in machine learning where the model learns all the given data including outliers to the extent such that it is unable to produce or generalize the predictions for the new upcoming data, so model should be trained considering this problem.
14. **Underfitting:** It is also a problem in machine learning when the model is too simple to learn the pattern between the features becoming unable to predict the label. This can be solved by introducing more parameters that can make the difference clear to the model and by increasing complexity of the model.
15. **Regularization:** It is a technique in machine learning to reduce the complexity of the model if the model exhibits behaviours which are like the overfitting of data.
16. **Cross Validation:** It is one of the powerful methods where data is partitioned into subsets of test and training data set to boost the performance of the model and improve the generalization of the result by the model.
17. **Feature Engineering:** It involves contextual understanding of the problem and deciding the features for a particular model which is expected to understand the differences and patterns between the given data to improve the generalization of the solution predicted by the model for the new values given to the model.
18. **Dimensionality Reduction:** It is another practice in machine learning where the number of features in a dataset is reduced to only some important features which leads to the correct interpretation of the relation between the data by the model to improve the accuracy of the prediction of the solution by the model.
19. **Bias:** It is the systematic error occurs in a model when it deviates from the patterns in the data due to the failure of the model to understand the underlying relation between the features making an inaccurate and inappropriate error in the prediction.
20. **Variance:** Variance refers to the change of the model’s behaviour in predicting results when it is trained on various different subset of training data.