**ML TERMINOLOGIES**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Patient ID** | **Glucose** | **BP** | **BMI** | **Age** | **Outcome** |
| 1 | 148 | 72 | 33.6 | 50 | 1 |
| 2 | 82 | 66 | 26.6 | 31 | 0 |
| 3 | 183 | 64 | 23.3 | 32 | 1 |
| 4 | 89 | 66 | 28.1 | 21 | 0 |
| 5 | 137 | 40 | 43.1 | 33 | 1 |
| 6 | 170 | 90 | 33 | 35 | 1 |
| 7 | 78 | 50 | 31 | 26 | 1 |
| 8 | 115 | 69 | 35.3 | 29 | 0 |
| 9 | 197 | 70 | 30.5 | 53 | 1 |
| 10 | 198 | 68 | 30.3 | 34 | 0 |

**FEATURES:** Features are the independent variables or attributes describing the various aspects of the data. It is given as input to train the model. (Example: Glucose, BP, BMI, and Age are the features of the dataset)

**LABEL:** The label is the dependent variable or target variable that the model tries to predict. It is the output of the data. The label may be numerical or categorical based on the features. (Example: Outcome is the label of the dataset).

**PREDICTION:** The output generated by the trained model for new unseen data(test data) based on the training data patterns it learned. (Example: The model predicts the Outcome ‘0’ or ‘1’ for new unseen data).

**OUTLIERS:** Outliers are known as the extreme points, which are the data objects whose values are too different than the rest of the objects. (Example: If the Glucose value is 290 then it would be an outlier).

**TEST DATA:** Test data is the new unseen data used to evaluate the trained model’s performance. (Example: Glucose(110), BP(78), BMI(26.8), and Age(32) are given as input to the trained model and it should predict the outcome based on the training data).

**TRAIN DATA:** Train data are the data that is used to train the model to predict the outcome. (Example: The given dataset is training data).

**MODEL:** An algorithm or mathematical representation that predicts the outcome based on the input data. (Example: Logistic regression model predicting the ‘Outcome’).

**VALIDATION DATA:** A subset of the data taken from the train data for model evaluation and tuning model parameters to prevent overfitting. (Example: Patients 7 and 8 can be taken as validation data).

**HYPERPARAMETER:** Hyperparameters are the parameters set before training the model that governs the training process. It is crucial for controlling the behavior and performance of the model. (Example: Learning rate, Number of Epochs)

**EPOCH:** One complete pass through the entire training dataset by the learning algorithm. (Example: Training the model on the dataset for 10 epochs).

**LOSS FUNCTION:** It is a mathematical function that measures the difference between the predicted output model and the actual output. (Example: Binary cross-entropy used for classification).

**LEARNING RATE:** It is a hyperparameter that controls the size of the steps taken by the algorithm to update the model parameters during training. (Example: A learning rate of 0.01).

**OVERFITTING:** Overfitting occurs when the model learns the details and noise in the training data to an extent and might not be able to classify new unseen samples and performs poorly on test data. It results in high variance

**UNDERFITTING:** Underfitting occurs when the model is too simple to capture the underlying patterns in the data, resulting in poor performance on both training and test datasets. (Example: A model with too few features performs poorly on both train and test data).

**REGULARIZATION:** It is a technique to avoid overfitting by adding a penalty to the loss function to prevent the model from fitting noise. (Example: L1 or L2 regularization can be used).

**CROSS-VALIDATION:** It is a statistical technique to evaluate the performance of the model by partitioning the data into subsets, training the model on some of the subsets, and validating it on the remaining subsets. It reduces the risk of overfitting. (Example: k-fold cross-validation).

**FEATURE ENGINEERING:** It is a process of transforming existing features, and creating new features from the raw data to improve the performance and accuracy of the model. (Example: Age\_Group feature can be created by using the Age feature. ‘Young’ if age<30 else ‘Old’).

**DIMENSIONALITY REDUCTION:** It is a process of reducing the number of features or variables in the dataset without affecting the integrity of the data. (Example: Using Principal Component Analysis to reduce features).

**BIAS:** Bias is the difference between the average prediction of a model and the actual output values.

**VARIANCE**: It measures how much the predictions of a model fluctuate for different training datasets.