Assignment 1

Data set of Hemoglobin Level

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| S.No | Age (years) | Gender | Weight (kg) | Height (cm) | Blood Pressure (mm Hg) | Iron Intake (mg/day) | Exercise (hours/week) | Hemoglobin Level (g/dL) |
| 1 | 25 | Male | 70 | 175 | 120/80 | 15 | 5 | 14.5 |
| 2 | 30 | Female | 60 | 160 | 110/70 | 12 | 3 | 13.2 |
| 3 | 35 | Male | 80 | 180 | 130/85 | 20 | 4 | 15.0 |
| 4 | 40 | Female | 65 | 165 | 115/75 | 18 | 2 | 14.0 |
| 5 | 45 | Male | 85 | 170 | 140/90 | 25 | 6 | 15.5 |
| 6 | 50 | Female | 70 | 155 | 125/80 | 10 | 1 | 12.5 |
| 7 | 55 | Male | 90 | 175 | 135/85 | 22 | 3 | 16.0 |
| 8 | 60 | Female | 75 | 160 | 120/75 | 17 | 4 | 13.8 |

In the above data set,

* **Feature**: Individual measurable properties (e.g., Age, Gender, Weight, Height, Blood Pressure, Iron Intake, Exercise) used as inputs to the model.
* **Label:** The output variable that the model aims to predict (e.g., Hemoglobin Level).
* **Prediction:** Predictions about possible outcomes of hemoglobin level based on given input data. For example, given that a person is 50 years old female, who weighs 70kg and 155cm tall, has BP of 125/80 mm Hg, takes 10 mg of iron per day, exercises 1 hour/week, the model predict that she has hemoglobin level of 12.5 g/dL.
* **Outlier:** A data point that deviates significantly from the rest of the data (e.g., if there was a person with hemoglobin level of 10 g/dL or 20 g/dL in this dataset).
* **Test Data:** A subset of data used to evaluate the performance of a model. Here, the last two datasets can be used as test data.
* **Training Data:** The subset of data used to train the model. Here, the first four datasets can be used as training data.
* **Model:** The model is built using training data to predict hemoglobin levels based on the inputs. Here, Linear regression model may be used.
* **Validation Data:** A subset of data used to tune the parameters of a model. Here, the middle datasets can be used as validation data.
* **Hyperparameter:** A parameter whose value is set before the learning process begins. Here, learning rate may be set to 0.01.
* **Epoch:** The one entire passing of training data through the algorithm. If the model is trained over 10 epochs, it means the dataset is processed 10 times.
* **Loss Function:** Process that quantifies the error margin between a model's prediction and the actual target value*.* For example, Mean Squared Error (MSE) could be used to calculate the error.
* **Learning Rate:** A parameter that provides the model a scale of how much model weights should be updated. Here, a learning rate of 0.01 may be used.
* **Overfitting:** Overfitting occurs when the model cannot generalize and fits too closely to the training dataset instead. For example, the model may predict exact hemoglobin levels for the training data set but fails to find exact level for new person.
* **Underfitting:** Underfitting occurs when the model has not learned the patterns in the training data well and is unable to generalize on the new data. For example, a model that predicts a constant hemoglobin level of 14.5 g/dL for all individuals regardless of different inputs or features.
* **Regularization:** Techniques used to prevent overfitting by adding a penalty to the loss function for large weights. Common regularization techniques include L1 and L2 regularization.
* **Cross-Validation:** A technique for evaluating models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data. For example, k-fold cross-validation where the data is split into 5 parts, and each part is used as validation once.
* **Feature Engineering:** The process of reworking a data set to improve the training of a machine learning model. For example, a new feature BMI can be created.
* **Dimensionality Reduction:** A method for representing a given dataset using a lower number of features while retaining important information. For example, Principal Component Analysis (PCA) might reduce the features from 7 to 3.
* **Bias:** A sort of mistake in which some aspects of a dataset are given more weight than others. High bias can lead to underfitting.
* **Variance:** The changes in the model when using different portions of the training data set. High variance can lead to overfitting.