Assignment 1

Data set for predicting diabetes

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| S. No | Age (years) | Gender | BMI (kg/m²) | Blood Pressure (mm Hg) | Glucose Level (mg/dL) | Insulin Level (µU/mL) | Physical Activity (hours/week) | Diabetes (Yes/No) |
| 1 | 45 | Female | 28.0 | 130/85 | 160 | 15 | 2 | Yes |
| 2 | 50 | Male | 32.0 | 140/90 | 180 | 25 | 1 | Yes |
| 3 | 40 | Female | 24.0 | 120/80 | 130 | 10 | 4 | No |
| 4 | 35 | Male | 27.5 | 115/75 | 140 | 12 | 5 | No |
| 5 | 60 | Female | 31.0 | 135/85 | 170 | 20 | 1 | Yes |
| 6 | 55 | Male | 29.0 | 125/80 | 150 | 18 | 3 | No |
| 7 | 42 | Female | 26.0 | 130/80 | 145 | 14 | 2 | No |
| 8 | 48 | Male | 30.0 | 140/85 | 160 | 22 | 1 | Yes |

## Terminology Explanations

### Feature

Individual measurable properties used as inputs to the model. For example, age, gender, BMI, BP, Glucose level, Insulin level and physical activities.

### Label

Output variable that is predicted. For example, Diabetes column.

### Prediction

Model’s guess about the output label based on the input features. For example, the model predicts if the person has Diabetes.

### Outlier

Data points that are different from rest of the data. For example, a person with very high glucose level when compared to others.

### Test Data

Subset of data that is used to evaluate the model’s performance after training. For example, the last two rows can be used as test data.

### Training Data

Subset of data used to train the model. For example, the first four rows can be used as training data.

### Model

An algorithm that learns from the training data to make predictions. For example, a decision tree model.

### Validation Data

Subset of data used to tune the model’s parameters. For example, the middle two rows can be used as validation data.

### Hyperparameter

Parameters whose values are set before the training process. For example, learning rate or number of Epoch.

### Epoch

One complete pass through the entire training dataset during the training process, For example, if the model is trained over 5 epochs, it means the dataset is processed 5 times.

### Loss Function

A function that measures how well the model’s prediction matches the actual label. For example, Binary Cross-Entropy could be used to calculate the error.

### Learning Rate

A hyperparameter that controls how much the model’s weights are updated. For example, a learning rate of 0.01 could be used.

### Overfitting

Occurs when a model performs very well on training data but poorly on test data. For example, the model could predicts exact diabetes of the training data but fails to predict exact diabetes of the test data.

### Underfitting

When a model is too simple to capture the underlying patterns in the data, resulting in poor performance on both training and new data. For example, the model could predict ‘yes’ for all person irrespective of their features.

### Regularization

Techniques to prevent overfitting by adding a penalty to the loss function for more complex models. For example, L1 or L2 regularization could be applied to the loss function.

### Cross-Validation

A method to evaluate model’s performance by splitting the data into multiple subsets. For example, k-fold cross-validation.

### Feature Engineering

Process of creating new features or modifying existing features to improve model performance. For example, creating a ‘weight’ feature.

### Dimensionality Reduction

Techniques used to reduce the number of features while retaining important information. For example, using Principal Component Analysis (PCA).

### Bias

An error due to overly simplistic assumptions in the learning algorithm. High bias means underfitting.

### Variance

The model’s sensitivity to fluctuations in the training data. High variance means overfitting.