**VIJAYA GOPIKHA S**

**22CSR237**

**ML intern Assignment - 1**

**Dataset Table:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Patient ID** | **Age** | **Cholesterol Level (mg/dL)** | **Blood Pressure (mmHg)** | **Heart Disease** |
| 001 | 55 | 200 | 140/90 | Yes |
| 002 | 45 | 190 | 130/85 | No |
| 003 | 65 | 230 | 150/95 | Yes |
| 004 | 50 | 180 | 125/85 | No |
| 005 | 60 | 210 | 145/85 | Yes |
| 006 | 40 | 160 | 110/75 | no |

**Terminologies:**

**Features**:

* Features are the input variables used by a machine learning model to make predictions (output).

**Eg**: In the dataset, the features are ‘Age’, ‘Cholesterol Level(mg/dL)’, ‘Blood pressure (mmHg)’.

**Label:**

* The label is the output variable that the model is trying to predict.

**Eg** : In this dataset, the label is Heart Disease, which indicates whether a patient has heart disease (Yes) or not (No).

**Prediction:**

* A prediction is the output of the model after it has been trained. Using the features, the model predicts the label.

**Eg** : Predicting whether a patient has heart disease based on their age, cholesterol level, and blood pressure.

**Outlier:**

* An outlier is a data point that differs significantly from other observations..It is like as the odd one out. It’s a data that doesn’t fit the pattern of the rest of the data.

**Eg** : If a patient had an exceptionally high cholesterol level of 400 mg/dL, it might be considered an outlier because it’s so different from the other data points.

**Test Data:**

* Test data is a subset of the dataset used to evaluate the performance of a trained model. It is not used in training
* This is the portion of the dataset we set aside to see how well our model is doing. It’s like a practice test for our model. We might use the last two records(005,006) to test our model in this case.

**Eg**:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 005 | 60 | 210 | 145/85 | Yes |
| 006 | 40 | 160 | 110/75 | No |

**Training Data:**

* Training data is the subset of the dataset used to train the model.

**Eg**: Here 001 and 004 could be used as training data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 10 | 80 | 9 | Yes | A |
| 7 | 70 | 7 | No | B |
| 12 | 85 | 10 | Yes | A |

**Model:**

* A model is a mathematical representation of the relationships between features and the label.

**Eg** : A decision tree model predicting heart disease based on age, cholesterol level, and blood pressure.

**Validation Data:**

* Validation data is a subset of the training data used to tune the hyperparameters of a model. It helps in validating the model's performance during training.

**Eg**: using patient id 003 and 004 data to validate the model.

**Hyperparameter:**

* Hyperparameters are the parameters set before the learning process begins, like the learning rate or the number of epochs.

**Eg**: The depth of the decision tree.

**Epoch:**

* An epoch is one complete pass through the entire training dataset.

**Eg**: If you train the model for 10 epochs, it means the model has seen the entire training dataset 10 times.

**Loss Function:**

* A loss function measures how well the model’s predictions match the actual labels.

**Eg**: Using binary cross-entropy loss to measure the difference between predicted and actual heart disease status.

**Learning Rate:**

* The learning rate controls how much the model's weights are adjusted with respect to the loss gradient.

**Eg** : A learning rate of 0.01 means the model is updates weights slowly.

**Overfitting:**

* Overfitting occurs when the model learns the training data too well, including noise and details that don’t matter.

**Eg**: An overfitted model performs well on training data but poorly on new data.

**Underfitting:**

* Underfitting occurs when a model is too simple to capture the underlying pattern in the data.

**Eg** : A model that performs poorly on both the training and test data

**Regularization:**

* Regularization techniques prevent overfitting by adding a penalty to the loss function.

**Eg**: Adding a penalty to the loss function for large coefficients in a linear regression model.

**Cross – Validation:**

* Cross-validation is a method to ensure the model performs well by splitting the data into multiple parts and training/testing several times

.

**Eg**: 5 fold cross validation involves splitting the data into 5 parts, training the model on 4 parts and testing on the remaining part repeating this process 5 times

**Feature Engineering:**

* This involves creating new features from existing ones to improve model performance

**Eg**: Combining Age and Cholesterol Level to create a new feature.

**Dimensionality Reduction :**

* Dimensionality reduction techniques reduce the number of features while retaining important information.

**Eg**: Using Principal Component Analysis (PCA) to reduce Age, Cholesterol Level, and Blood Pressure to two principal components.

**Bias:**

* This is an error introduced by simplifying assumptions made by the model. A high-bias model might make overly simple predictions

**Eg** :Assuming heart disease only depends on age and ignoring cholesterol level and blood pressure.

**Variance:**

* Variance is the model's sensitivity to small fluctuations in the training data.

**Eg**: A high variance model may predict heart disease accurately for training data but poorly for new, unseen test data.