ML intern Assignment - 1

By THIVISHA S

22CSR225

Dataset Table:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Patient**  **ID** | **Admission**  **Date** | **Name** | **Age** | **Diagnosis** | **Treatment** | **DischargeDate** |
| 201 | 2024-07-20 | John Doe | 34 | Flu | Medication | 2024-07-25 |
| 202 | 2024-07-21 | Jane Smith | 45 | Broken Arm | Surgery | 2024-07-29 |
| 203 | 2024-07-22 | Alice Brown | 29 | Migraine | Therapy | 2024-07-24 |
| 204 | 2024-07-23 | Bob White | 56 | Diabetes | Diet | 2024-08-01 |
| 205 | 2024-07-24 | Charlie Black | 40 | Hypertension | Medication | 2024-07-30 |

Terminologies:

**Feature:**

**Features** are individual measurable properties or characteristics used as input for a model to make predictions of output.

Eg : In this dataset, the features are ADMISSION DATE,TREATMENT,AGE, DIAGNOSIS.

**Label:**

**Label**  is the target variable that the model is trained to predict, representing the outcome of interest.

Eg : In this dataset, the label is DISCHARGE DATE.

**Prediction :**

**Prediction** is the output generated by a model based on input features, estimating the label or outcome.

Eg : Given a patient with Age = 50, Diagnosis = Diabetes, and Treatment = Medication, predict their DISCHARGE DATE.

**Outlier:**

**Outlier**  is a data point that significantly deviates from the other values in a dataset.

**Eg:** If the Length Of Stay of 20 days could be considered an outlier compared to the rest of the data where the range is from 2 to 9 days for MEDICATION.

**Test data:**

**Test data** refers to a subset of data used to evaluate the performance of a machine learning model. It helps assess how well the model generalizes to new, unseen data.

**Eg:** You would use AGE,DIAGNOSIS,TREATMENT test data to see how accurately the model predicts the DISCHARGE DATE these patients.

**Training Data:**

**Training data**  is the subset of a dataset used to train a machine learning model. It contains input features and corresponding labels, allowing the model to learn patterns and relationships.

**Eg:** This data is used to train the model to predict the Discharge Date based on features like AGE, DIAGNOSIS, TREATMENT and ADMISSION DATE.

**Model:**

A **model**  is an algorithm that learns from training data to predict outcomes for new data based on identified patterns and input features.

**Eg:** A Linear Regression model could be trained to predict DISCHARGE DATE based on features like AGE, DIAGNOSIS, TREATMENT and ADMISSION DATE. For instance, if the model is trained on this data, it might predict that a new patient with Age = 40, Diagnosis = Hypertension, and Treatment = Medication will have a DISCHARGE DATE of 2024-07-30.

**Validation data:**

**Validation data**  is a subset of data used to assess a model's performance and fine-tune its parameters during training. It helps ensure the model generalizes well to new, unseen data.

**Eg:** If you train a model on 80% of patient data, you can use the remaining 20% as validation data to check how accurately the model predicts ` DISCHARGE DATE `. This helps ensure the model's predictions are reliable on unseen data.

**Hyper parameters:**

**Hyper parameters** are pre-set values that are not learned from the data but are specified before training the model to guide its learning process.

**Eg:** In predicting `Discharge Date` using a decision tree model with the hospital data, the maximum depth of the tree is a hyper parameter set before training that controls how complex the tree can become.

**Epoch:**

A**n epoch**  is when the model looks at every piece of training data once to learn from it.

**Eg:** If you set 50 epochs, the model will review the entire hospital dataset 50 times to better predict `Discharge Date`.

**Loss function:**

A loss function measures prediction errors.

**Eg:** The Mean Squared Error (MSE) loss function would compute the average squared difference between the model's predicted `Discharge Date` and the actual `Discharge Date` for each patient.

**Learning Rate:**

The learning rate controls how much the model’s weights are adjusted with each iteration during training.

**Eg:** If the learning rate is set too high, the model might overshoot the optimal `Discharge Date` predictions ; if set too low, training might be too slow.

**Over fitting:**

**Over fitting** is when a model performs well on training data but poorly on new data because it has learned too many details from the training set.

**Eg:** In predicting `Discharge Date`, an over fitted model might give accurate predictions for known patients but fail to generalize to new patients.

**Under fitting:**

**Under fitting** occurs when a model is too simple to capture the underlying patterns in the data, resulting in poor performance on both training and new data.

**Eg:** In predicting `Discharge Date`, an under fitted model might produce in accurate predictions for both the training set and new patients because it doesn’t capture enough complexity.

**Regularization:**

**Regularization** is a technique used to prevent overfitting by adding a penalty to the model for having overly complex parameters.

**Eg:** In predicting `DischargeDate`, regularization might limit the size of coefficients in a regression model to avoid overly complex solutions that fit the training data too closely.

**Cross-validation:**

**Cross-validation** is a technique for assessing a model's performance by dividing the data into multiple subsets, training on some, and validating on others to ensure the model generalizes well.

**Eg** :In predicting `DischargeDate`, cross-validation might involve splitting the hospital data into 5 subsets, training the model on 4 subsets, and testing it on the remaining 1 subset, repeating this process to ensure robust performance.

**Feature engineering:**

**Feature engineering** involves creating or transforming input features to improve a model's performance by making the data more informative and relevant.

**Eg** :For predicting `DischargeDate`, feature engineering might involve creating new features like the number of days between `AdmissionDate` and `DischargeDate` or encoding categorical variables such as `Diagnosis` and `Treatment`.

**Dimensionality reduction:**

**Dimensionality reduction**  is the process of reducing the number of input features in a dataset while preserving as much relevant information as possible.

**Eg** :In predicting `DischargeDate`, dimensionality reduction might involve using techniques like Principal Component Analysis (PCA) to reduce a large set of features (e.g., various medical metrics) to a smaller set of principal components that capture most of the variance in the data.

**Bias:**

**Bias** refers to the error introduced by approximating a real-world problem with a simplified model, which can lead to systematic errors in predictions.

**Eg** :In predicting `DischargeDate`, high bias might occur if the model assumes all patients with similar features have the same discharge date, leading to inaccurate predictions for individual cases.

**Variance:**

**Variance** refers to the error introduced by a model's sensitivity to small fluctuations in the training data, leading to inconsistent predictions on new data.

**Eg** :In predicting `DischargeDate`, high variance might occur if the model learns detailed patterns specific to the training data, resulting in highly variable predictions for new patients.