2nd Lab Homework

Jayesh Kirtane
Student ID: M24DS008

Q1)

Imagine a machine learning model designed to diagnose a rare disease that affects only 1% of the population. In a test set of 10,000 patients, only 100 actually have the disease. The confusion matrix from the model's predictions might look like this:

True Positives (TP): 80 patients correctly diagnosed with the disease.

False Negatives (FN): 20 patients who have the disease but were incorrectly diagnosed as not having it.

False Positives (FP): 100 patients who do not have the disease but were incorrectly diagnosed as having it.

True Negatives (TN): 9,800 patients correctly diagnosed as not having the disease.

So which performance parameter should you prefer for model performance?

Answer 1

For diagnosing a rare disease, the cost of a false negative is high. The cost of missing a diagnosis in these cases, can be extremely high, potentially leading to a lack of necessary treatment and worsening of the patient's condition.

Under such curcumstances, the **Recall** metric is crucial. The Recall focuses on the model's ability to correctly identify all positive instances. In medical diagnosis, it is vital to catch as many true cases as possible.

The formula for Recall is given as:

Recall = $\frac{TP}{TP+FN}$

Based on the given data:

TP = 80

FN = 20

FP = 100

TN = 9800

Substituting the values in formula we get:

Recall = $\frac{80}{80+20}$

Recall = 0.8

A recall of 0.8 indicates the model is successfully identifying 80% of all actual cases, which indicates a strong performance.