Module Quiz and Answer

1. What is the main difference between Linear Regression and Logistic Regression?

a) Linear Regression uses trees, Logistic Regression uses lines
b) Linear Regression predicts continuous values, Logistic Regression predicts categories
c) Logistic Regression can only work with numbers
d) Linear Regression is faster than Logistic Regression

Ans: b) Linear Regression predicts continuous values, Logistic Regression predicts categories

ব্যাখ্যা: Linear Regression একটি continuous value predict করে যেমন বাড়ির দাম।
কিন্তু Logistic Regression একটি category predict করে — যেমন survived/not survived।

2. In the Titanic dataset used in this module, what does the target column Survived contain?

a) 0 and 1
b) True and False
c) "yes" and "no"
d) 1 and 2

Ans: c) "yes" and "no"

ব্যাখ্যা: এই module-এর dataset-এ Survived column-এ "yes" এবং "no" string value আছে।
এটা standard 0/1 নয়, তাই পরে encoding দরকার হয়।

3. What does the "Accuracy" metric measure in a classification model, and in which scenario is it most likely to provide a misleading evaluation?

a) It measures only the correct positive predictions; misleading when false negatives are high.
b) It measures the overall proportion of correct predictions; misleading when the dataset is highly imbalanced.
c) It measures the ratio of true positives to false positives; misleading when classes are equally distributed.
d) It measures the model's ability to find all actual positives; misleading when false positives are costly.

Ans: b) It measures the overall proportion of correct predictions; misleading when the dataset is highly imbalanced.

ব্যাখ্যা**:** Accuracy হলো মোট প্রেডিকশনের মধ্যে মডেলটি কতগুলো সঠিক প্রেডিকশন (True Positives + True Negatives) করেছে তার অনুপাত। এর সূত্র: `(TP + TN) / Total Predictions`। তবে ডেটাসেট যদি imbalanced হয় (যেমন ১০০ জন রোগীর মধ্যে ৯৫ জন সুস্থ এবং ৫ জন অসুস্থ), তখন মডেলটি না বুঝেই সবাইকে "সুস্থ" প্রেডিক্ট করলেও ৯৫% Accuracy দেখাতে পারে। মডেলটি আসলে অসুস্থ রোগীদের শনাক্ত করতে ব্যর্থ হলেও Accuracy অনেক হাই দেখায়, যা অত্যন্ত misleading।

4. What does the stratify=y parameter do in train_test_split?

a) It sorts the data before splitting
b) It ensures both train and test sets have a similar ratio of survived vs not survived
c) It removes duplicate rows
d) It balances the number of rows in train and test

Ans: b) It ensures both train and test sets have a similar ratio of survived vs not survived

ব্যাখ্যা: stratify=y দিলে train এবং test উভয় set-এ target class-এর proportion একই থাকে। যদি 40% passenger survive করে থাকে, তাহলে দুটোতেই প্রায় 40% survive থাকবে।
Learning: Stratified splitting prevents class imbalance between train and test sets.

5. In a classification model, what does the "Recall" metric primarily evaluate, and in which scenario is it most critical to maximize it?

a) The proportion of correct positive predictions;
b) The ability to find all actual positive cases; important when false negatives are highly dangerous.
c) The overall correctness of the model; important for perfectly balanced datasets.
d) Important for Regression Performance .

Ans: b) The ability to find all actual positive cases; important when false negatives are highly dangerous.

ব্যাখ্যা**:** Recall (বা Sensitivity) মূলত পরিমাপ করে যে, বাস্তবে যতগুলো "Positive" কেস ছিল, মডেল তার মধ্যে কতগুলোকে সঠিকভাবে শনাক্ত করতে পেরেছে। এর সূত্র হলো: `True Positives / (True Positives + False Negatives)`। যখন False Negative-এর ঝুঁকি অনেক বেশি থাকে (যেমন: রোগ নির্ণয় বা ব্যাংক ফ্রড ডিটেকশন, যেখানে একজন অসুস্থ মানুষকে ভুল করে সুস্থ বললে মারাত্মক ক্ষতি হতে পারে), তখন Recall-এর উপর সবচেয়ে বেশি জোর দেওয়া হয়।

6. What does the "Precision" metric primarily evaluate in a classification model?

a) The proportion of actual negative cases that were correctly identified
b) The overall percentage of correct predictions across all classes
c) The proportion of positive predictions that were actually correct
d) The harmonic mean of True Positives and False Negatives

Ans**:** c) The proportion of positive predictions that were actually correct

ব্যাখ্যা**:** Precision মূলত পরিমাপ করে মডেল যতগুলোকে "Positive" হিসেবে প্রেডিক্ট করেছে, তার মধ্যে আসলেই কতগুলো Positive ছিল। এর সূত্র হলো: True Positives / (True Positives + False Positives)। যখন False Positive কমানো আমাদের মূল লক্ষ্য হয় (যেমন: দরকারি ইমেইল যেন ভুল করে Spam ফোল্ডারে না যায়), তখন Precision-এর উপর বেশি জোর দেওয়া হয়।

7. Which encoder is used for the Pclass column, and why?

a) OneHotEncoder, because Pclass has no order
b) LabelEncoder, because it is the simplest

c) OrdinalEncoder, because first class is better than third class
d) StandardScaler, because Pclass is numerical

Ans: c) OrdinalEncoder, because first class is better than third class

ব্যাখ্যা: Pclass-এ একটি natural order আছে — third < second < first।
OrdinalEncoder এই order maintain করে encode করে।
Learning: Use OrdinalEncoder when categories have a meaningful order.

8. What does remainder='drop' mean in ColumnTransformer?

a) It drops rows with missing values
b) It removes columns not mentioned in the transformers list
c) It deletes duplicate columns
d) It drops the target column automatically

Ans: b) It removes columns not mentioned in the transformers list

ব্যাখ্যা: ColumnTransformer-এ শুধু নির্দিষ্ট column গুলোর জন্য pipeline define করা হয়।
remainder='drop' মানে বাকি সব column (Name, Ticket, PassengerId) automatically বাদ দেওয়া হবে।
Learning: remainder='drop' ensures only selected and transformed features reach the model.

9. What does class_weight='balanced' do in LogisticRegression?

a) It makes the model train faster
b) It gives more importance to the minority class during training
c) It balances the number of rows in train and test
d) It sets all feature weights to equal values

Ans: b) It gives more importance to the minority class during training

ব্যাখ্যা: যদি dataset-এ একটি class অনেক কম থাকে, তাহলে model সেই class ignore করতে পারে।
class_weight='balanced' দিলে model automatically minority class-কে বেশি গুরুত্ব দেয়।
Learning: Use class_weight='balanced' when your dataset has unequal class distribution.

10. What does the predict() method return in Logistic Regression?

a) Probability scores between 0 and 1
b) The predicted class label for each input
c) The model accuracy score
d) The loss value during training

Ans: b) The predicted class label for each input

ব্যাখ্যা: predict() method প্রতিটি input row-এর জন্য একটি class label return করে — যেমন "yes" অথবা "no"। এটি probability নয়, সরাসরি prediction। probability দেখতে চাইলে predict_proba() ব্যবহার করতে হয়।
Learning: predict() returns class labels, predict_proba() returns probability scores.