Machine learning exam

This exam is timed and the duration of each question is displayed at the end of each question. A progress bar below each question also indicates the time remaining. **Questions are automatically skipped** as soon as the time runs out, and **there is no turning back.**

It is divided into 2 parts

1- **Oral exam:** you will read the questions on the screen and answer them orally to your examiner.

You will have a two-minute break in between to prepare.

2**- The written exam:** you will read the questions and answers on the screen, and you will have to answer via a google forms form by ticking the right answer(s).

ORAL EXAM 15 min

(2 min)

[redacted]

QUESTION 1

What is the difference between overfitting and underfitting? (1 min)

QUESTION 2

Explain the difference between linear regression and logistic regression. (1 min)

QUESTION 3

What is a selection bias, and what is its impact on a machine learning model? (1 min)

QUESTION 4

Compare boosting approaches (e.g., XGBoost) and bagging approaches (e.g., Random Forest). (1 min 30)

QUESTION 5

Ensemble methods, like Random Forest, reduce bias compared to single models.

True or False? Explain why if false. (1 min)

QUESTION 6

Explain how a Convolutional Neural Network (CNN) works and provide an example application. (1 min)

QUESTION 7

What is a "label" in supervised learning? (30 s)

QUESTION 8

How does cross-validation improve a model's performance? (1 min)

QUESTION 9

Explain how the K-Means algorithm works and give an example of its limitations. (1 min 30)

QUESTION 10

What is the problem of vanishing gradients? (1 min)

QUESTION 11

What is KNN, and in what context is it used? (1 min)

QUESTION 12

Name three types of supervised machine learning algorithms. (1 min)

QUESTION 13

How can overfitting be avoided? (1 min)

QUESTION 14

What is a confusion matrix, and what metrics can be derived from it? (1 min30)

WRITTEN EXAM 15 min

(2 min)

[redacted]

QUESTION 1

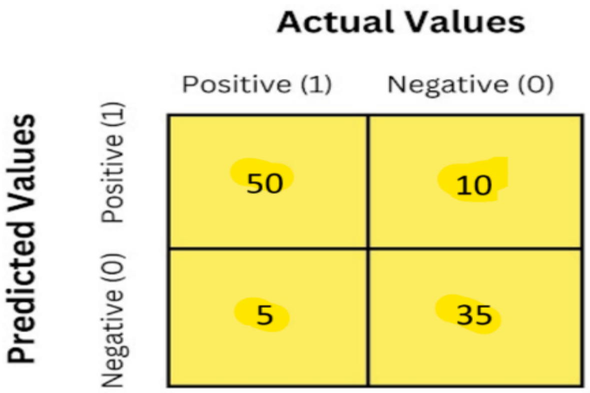
Name 3 metrics that can be used to evaluate the output of a classification model? *(1 min)*

QUESTION 2

What is the Recall value of the following confusion matrix. Provide the result in the form of a simplified fraction. (2 min)

The formula in the image defines **Recall (True Positive Rate)** as follows:

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

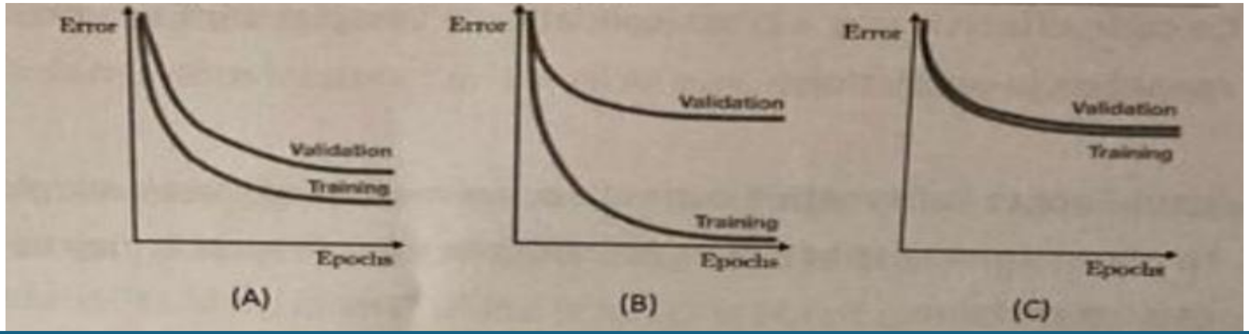


		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	50	10
	Negative (0)	5	35

QUESTION 3

Match each curve to the corresponding proposition : (1min30)

- a) (A) Good Fit, (B) Underfit, (C) Overfit
- b) (A) Good Fit, (B) Overfit, (C) Underfit
- c) (A) Underfit, (B) Overfit, (C) Good Fit
- d) (A) Underfit, (B) Good Fit, (C) Overfit



QUESTION 4

What does the term "cross-validation" mean? (1 min)

a) A technique to normalize data before training.
b) A method to evaluate the performance of a model by dividing the data into multiple subsets.
c) A technique to detect outliers in a dataset.
d) A method to balance classes in an imbalanced dataset.

QUESTION 5

Among the following scenarios, choose the one(s) that might be considered as a regression problem: (3min)

a) We collect a set of data on the top 500 firms in France. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary

b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charger for the product, marketing budget, competition price, and ten other variables.

c) We are interested in predicting the % change in the US dollar in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week record the % change in the dollar, the % change in the US market, the % change in the British market, and the % change in the German market

d) An emergency room in a hospital measures 10 variables (e.g. blood pressure, age, etc.) of newly admitted patients. Based on these variables, the hospital can determine the life-risk (high/low) of each patient. A decision has to be taken whether to put the patient in an intensive-care unit based on his life-risk level.

QUESTION 6

What does bagging stand for? (1 min)

a) Boosted Aggregation

b) Bootstrap Aggregation

c) Bayesian Aggregation

d) Balanced Aggregation

QUESTION 7

Which of the following algorithms uses bagging? (1 min)

a) Random Forest

b) Gradient Boosting

c) XGBoost

d) K-Nearest Neighbors

QUESTION 8

- **What is a limitation of Grid Search?** (1 min 30)

a) It cannot tune hyperparameters for regression models.

b) It requires large computational resources for high-dimensional grids.

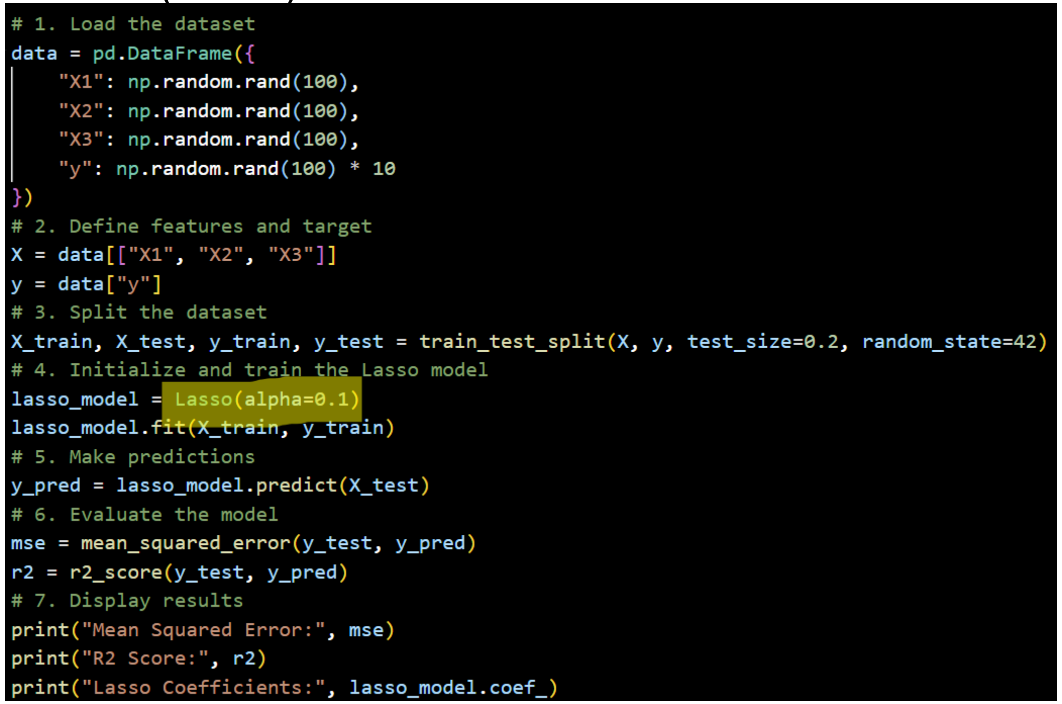
c) It only works for tree-based models.

d) It cannot handle categorical hyperparameters.

QUESTION 9

Refer to the following code and answer the questions

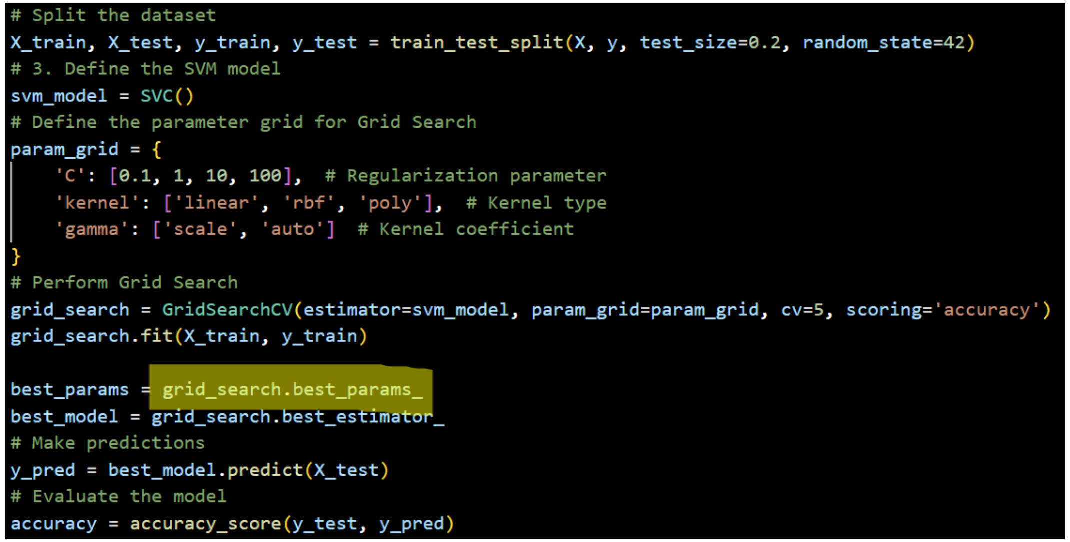
Why do we calculate both MSE and R2 score and what is the difference between these metrics? (1 min)



```
# 1. Load the dataset
data = pd.DataFrame({
    "X1": np.random.rand(100),
    "X2": np.random.rand(100),
    "X3": np.random.rand(100),
    "y": np.random.rand(100) * 10
})
# 2. Define features and target
X = data[["X1", "X2", "X3"]]
y = data["y"]
# 3. Split the dataset
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# 4. Initialize and train the Lasso model
lasso_model = Lasso(alpha=0.1)
lasso_model.fit(X_train, y_train)
# 5. Make predictions
y_pred = lasso_model.predict(X_test)
# 6. Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
# 7. Display results
print("Mean Squared Error:", mse)
print("R2 Score:", r2)
print("Lasso Coefficients:", lasso_model.coef_)
```

QUESTION 10

What does grid_search.best_params_ return? (2 min)



```
# Split the dataset
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# 3. Define the SVM model
svm_model = SVC()
# Define the parameter grid for Grid Search
param_grid = {
    'C': [0.1, 1, 10, 100],  # Regularization parameter
    'kernel': ['linear', 'rbf', 'poly'],  # Kernel type
    'gamma': ['scale', 'auto']  # Kernel coefficient
}
# Perform Grid Search
grid_search = GridSearchCV(estimator=svm_model, param_grid=param_grid, cv=5, scoring='accuracy')
grid_search.fit(X_train, y_train)

best_params = grid_search.best_params_
best_model = grid_search.best_estimator_
# Make predictions
y_pred = best_model.predict(X_test)
# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
```

END OF EXAM

(2 min)

[redacted]