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)

QUESTION 1

Explain the difference between supervised and unsupervised machine learning. (1 min)

QUESTION 2

What is the F1-score metric, and when is it preferred?

(1 min 30)

QUESTION 3

What are the hyperparameters of a decision tree? (4 at least) (1 min)

QUESTION 4

Compare cross-entropy and mean squared error (MSE) as cost functions. (1 min)

QUESTION 5

What is the model selection problem? (1 min)

QUESTION 6

Provide examples of approaches to deal with highly imbalanced datasets. (1 min 30)

QUESTION 7

Why is it important to split a dataset into training and test sets in a supervised ML task? (1 min)

QUESTION 8

What is feature engineering, and why is it important?
(1 min)

QUESTION 9

How does the Random Forest algorithm improve over a single decision tree? (1 min)

QUESTION 10

Increasing the number of features in a dataset always improves a model's performance.
True or False? Explain why if false. *(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

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

WRITTEN EXAM 15 min

(2 min)

QUESTION 1

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

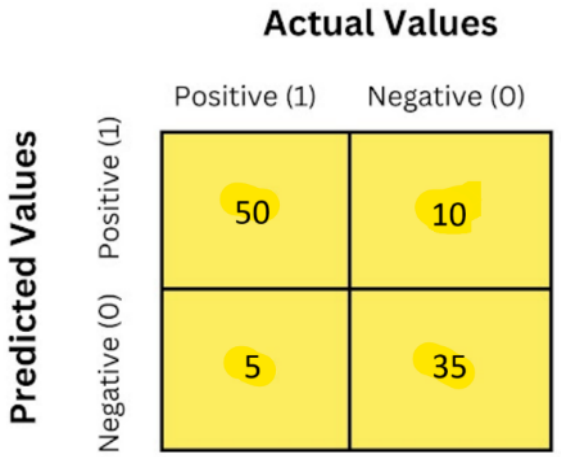
QUESTION 2

What is the formula for accuracy? *(1 min)*

QUESTION 3

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:

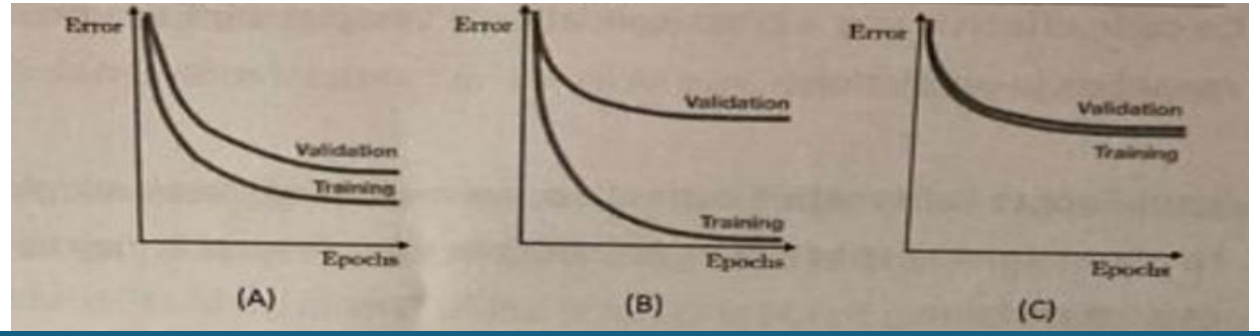


$$Recall = \frac{TP}{TP + FN}$$

QUESTION 4

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 5

To measure the effectiveness of a split while constructing a Decision Tree, we can use: (1 min)

a) The Classification error metric

b) The Gini index metric

c) The entropy metric

d) None of the above

QUESTION 6

Which technique can be used to handle missing data in a dataset? (1 min)

a) Removing columns with missing values.
b) Imputation using the mean or median.
c) Using predictive models to estimate the missing values.
d) All of the above.

QUESTION 7

Which of the following scenarios can be considered an unsupervised machine learning problem? (1 min)

a) Predicting the price of a house.
b) Identifying groups of customers with similar behaviors.
c) Determining whether an email is spam or not.
d) Predicting future sales of a product.

QUESTION 8

What does bagging stand for? (1 min)

a) Boosted Aggregation
b) Bootstrap Aggregation
c) Bayesian Aggregation
d) Balanced Aggregation

QUESTION 9

Choose the correct answer(s): (1 min)

a) The more over-fitted a model is, the larger its bias-error is.
b) The more over-fitted a model is, the larger its variance-error is
c) Regularization is a technique to reduce the bias-error
d) Regularization is a technique to reduce the variance-error.

QUESTION 10

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 11

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 12

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)