Examen	**Session Principale/ Semestre 2**
Matière: AI programming Fundamentals	**Classe : Licence 2 H**
Enseignant(e): Khouloud Chalbi	**Date:** 28/05/2024
Documents non autorisés	**Duration: 1H30**

Exercice n°1: Choose the correct answer. (15 pts)
Ps: 1 option is valid per question

1. Training a model using labeled data and using this model to predict the labels for new data is known as _____.
   a) Unsupervised Learning
   b) Clustering
   c) Density Estimation
   d) Supervised Learning

2. Training a model using categorically labelled data to predict labels for new data is known as _____.
   a) Regression
   b) Classification
   c) Feature Extraction
   d) Clustering

3. Training a model using labelled data where the labels are continuous quantities to predict labels for new data is known as _____.
   a) Regression
   b) Feature Extraction
   c) Classification
   d) Clustering

4. What is the key difference between supervised and unsupervised learning?
   a) The presence of labeled data in supervised learning
   b) The use of deep neural networks in unsupervised learning
   c) The requirement for large datasets in supervised learning
   d) The absence of algorithms in unsupervised learning

5. What is the primary purpose of a validation set in machine learning?
   a) To train the model
   b) To test the model
   c) To evaluate the model during training
   d) To compare multiple models

6. Which of the following is an example of a classification metric?
   a) Mean Absolute Error (MAE)
   b) Root Mean Squared Error (RMSE)
   c) Area Under the Receiver Operating Characteristic (ROC-AUC)
   d) R-squared (R2)

7. What is overfitting in the context of machine learning models?
   a) Fitting a model with insufficient data
   b) Fitting a model too closely to the training data
   c) Fitting a model with too few features
   d) Fitting a model to the validation set

8. What is feature importance in the context of tree-based models like Random Forests?
   a) The importance of including a feature in the dataset
   b) The correlation between features
   c) The contribution of each feature to the model's predictions
   d) The number of times a feature appears in the dataset

9. What is the difference between precision and recall?
   a) Both measure the same thing
   b) Precision focuses on false positives, recall focuses on false negatives
   c) Precision focuses on false negatives, recall focuses on false positives
   d) Precision and recall are unrelated metrics

10. The key purpose of splitting the dataset into training and test sets is:
    a) To speed up the training process
    b) To reduce the number of features we need to consider as input to the learning algorithm
    c) To estimate how well the learned model will generalize to new data
    d) To reduce the amount of labelled data needed for evaluating classifier accuracy

11. Given a dataset with 10,000 observations and 50 features plus one label, what would be the dimensions of X_train, y_train, X_test, and y_test? Assume a train/test split of 75%/25%.

a)
- X_train: (7500, 50)
- y_train: (7500,)
- X_test: (2500, 50)
- y_test: (2500,)

b)
- X_train: (10000, 50)
- y_train: (10000,)
- X_test: (10000, 50)
- y_test: (10000,)

c)
- X_train: (2500, 50)
- y_train: (2500,)
- X_test: (7500, 50)
- y_test: (7500,)

d)
- X_train: (2500,)
- y_train: (2500, 50)
- X_test: (7500,)
- y_test: (7500, 50)

e)
- X_train: (7500, 51)
- y_train: (7500,1)
- X_test: (2500, 51)
- y_test: (2500,1)

12. Given the following confusion matrix:

	Predicted Positive	Predicted Negative
Condition Positive	96	4
Condition Negative	8	19

What would be the accuracy value?

a) 0.906
b) 0.896
c) 1.256
d) 1.096
e) None of the above

13. Given the following confusion matrix:

	Predicted Positive	Predicted Negative
Condition Positive	102	56
Condition Negative	17	78

What would be the precision value?

a) 0.898
b) 0.857
c) 0.906
d) 1.096
e) None of the above

14. Given the following confusion matrix:

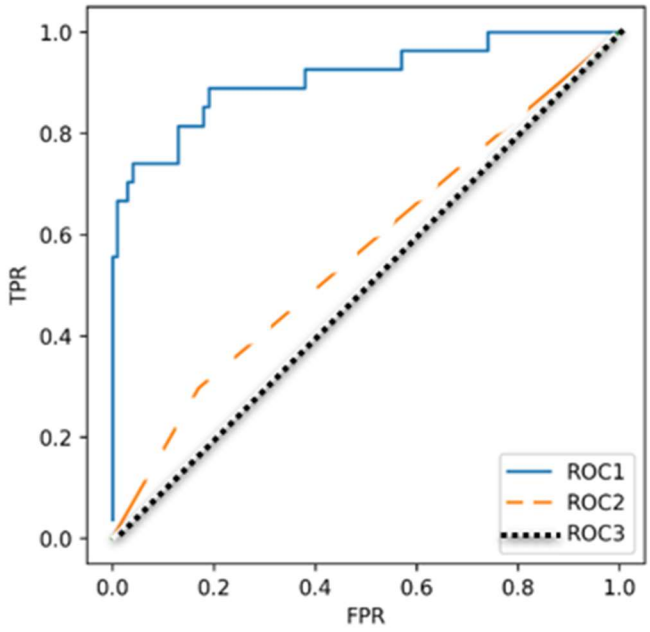
	Predicted Positive	Predicted Negative
Condition Positive	102	56
Condition Negative	17	78

What would be the recall value?

a) 0.887
b) 0.857
c) 0.906
d) 1.025
e) None of the above

15. Given the following models and AUC scores, match each model to its corresponding ROC curve.

- Model 1 test set AUC score: 0.91
- Model 2 test set AUC score: 0.50
- Model 3 test set AUC score: 0.56



a)
- Model 1: Roc 1
- Model 2: Roc 2
- Model 3: Roc 3

b)
-
- Model 1: Roc 1
- Model 2: Roc 3
- Model 3: Roc 2

c)
- Model 1: Roc 2
- Model 2: Roc 3
- Model 3: Roc 1

d)
- Model 1: Roc 3
- Model 2: Roc 2
- Model 3: Roc 1

e) Not enough information is given.

16. Which of the following is true of the R-Squared regression score metric?

a) A model that always predicts the mean of y would get a score of 0.5
b) The score can sometimes be negative.
c) The highest possible score is 1.0

17. Which of the following indicates a good fit of a linear regression Model?
a) High R-squared and low mean squared error (RMSE)
b) Low R-squared and high RMSE
c) High R-squared and high RMSE
d) Low R-squared and low RMSE

18. Which of the following is a key assumption of linear regression related to independent variables/features?
    a) They must be normally distributed
    b) They must be categorical
    c) There should be a high correlation between them
    d) They should be skewed

19. When is the use of Logistic regression preferred over Linear regression?
    a) When the dependant variable (target) is continuous
    b) When there is a linear relathionship
    c) When the variance of the error terms is constant
    d) When the dependant variable (target) is categorical

20. What is the margin in the context of SVMs (Support Vector Machines)?
    a) The distance between the closest data points of different classes
    b) The average distance between all data points
    c) The total number of support vectors
    d) The difference in means of two classes

21. What is the difference between a hard margin and a soft margin in SVMs?
    a) Hard margin is more flexible than soft margin
    b) Soft margin allows for some misclassification, while hard margin does not
    c) Hard margin is suitable for linearly inseparable data

22. What is the difference between bagging and boosting in ensemble learning?
    a) Bagging increases model diversity, boosting decreases it
    b) Bagging trains models sequentially, boosting trains them in parallel
    c) Bagging combines predictions using voting, boosting combines predictions using weighted averaging
    d) Bagging trains each model independently, boosting focuses on examples misclassified by previous models

23. Which of the following is an example of a bagging algorithm?
    a) Linear regression
    b) Decision Trees
    c) Random forest
    d) SVMs

24. What is the primary idea behind the “ensemble” nature of bagging?
    a) Combining models to increase interpretability
    b) Sequentially training models to achieve a cumulative effect
    c) Aggregating models to form a hierarchical structure
    d) Training multiple models independently and combining their predictions

25. What is the pandas method that outputs statistical analysis of the numerical values?
    a) Dataframe.info()
    b) Dataframe.describe()
    c) Dataframe.unique()
    d) Dataframe.head()

26. What does feature scaling help with in machine learning models?
    a) It ensures features contribute equally to model prediction
    b) It converts categorical data into numbers so that it can be used by the model
    c) It increases model training speed by reducing data size
    d) It creates new features from existing ones to improve model accuracy

27. What does feature encoding help with in machine learning models?
    a) It ensures features contribute equally to model prediction
    b) It converts categorical data into numbers so that it can be used by the model
    c) It increases model training speed by reducing data size
    d) It creates new features from existing ones to improve model accuracy

28. What does feature engineering help with in machine learning models?
    a) It ensures features contribute equally to model prediction
    b) It converts categorical data into numbers so that it can be used by the model
    c) It increases model training speed by reducing data size
    d) It creates new features from existing ones to improve model accuracy

29. In the context of missing data, what does "imputation" mean?
    a) The process of deleting records with missing values from the dataset
    b) Estimating missing values based on other observations (mean, median,…)
    c) The technique of visualizing data to identify patterns
    d) The method of preventing missing data during data collection

30. Fill in the blanks with the correct terminologies:
    "To encode categorical data, we use "one-hot encoder" for _____ features, and "label encoder" for _____ features"
    a) Ordinal, nominal
    b) Nominal, ordinal
    c) Categorical, numerical
    d) Numerical, categorical

Exercice n°2 : (5 pts)

1. Name the steps of the Crisp'DM method in machine learning projects
2. This is a sample dataset :

X1	X2	X3	Y
25.6	Red	Low	10000
10.4	Green	Low	25052
15.8	Yellow	High	15125
22.6	Yellow	Medium	8965

a) What are the features that need to be transformed / preprocessed and why?
b) What is(are) the preprocessing technique(s) to be performed
c) For feature "X2" and "X3" , write a line of code that would help extract all possible values (categories/classes)
d) Draw the sample dataset after transformation of only feature "X2" and "X3" knowing that possible classes for feature "X2" are Red/Green/Yellow/Orange , and for feature "X3" the possible classes are Low/Medium/High.