الجامعة اللبنانية
كلية الإعلام
الفرع الأول

Third Year

Advanced Data Science

Semester VI Date:

Final Exam

Duration: **2024 / 2025**

Instructor : Dr. K. Danach

SECTION 1: 20 Multiple Choice Questions (MCQs). Each question has only one answer. 1 point/question, 20 points in total.

1. What does the .loc[] method in pandas primarily use for indexing?
   (A) Labels
   B) Integer positions
   C) Column names only
   D) Boolean values only

2. What is the result of df.iloc[2] in pandas?
   (A) The third row by position
   B) The row labeled "2"
   C) All rows with index > 2
   D) The last column in the DataFrame

3. In matplotlib, what function is typically used to display a line plot?
   (A) plt.plot()
   B) plt.hist()
   C) plt.scatter()
   D) plt.boxplot()

4. Which of the following is used to detect outliers using the IQR method?
   (A) Q3 + 1.5 * IQR

   B) Q2 + IQR

   C) Mean ± IQR

   D) Q1 + IQR / 2

5. What does the .isnull() function in pandas return?
   (A) A DataFrame of Boolean values

   B) The number of missing values

   C) A list of NaN rows

   D) Only columns with NaN

6. What method is used to fill missing values with the column mean?
   (A) df.fillna(df.mean())

   B) df.dropna()

   C) df.replace(0)

   D) df.interpolate()

7. Which matplotlib function sets the label of the X-axis?
   (A) plt.xlabel()

   B) plt.title()

   C) plt.xaxis()

   D) plt.set_xlabel()

8. What is **standardization** in feature scaling?
   (A) Transforming data to have a mean of 0 and std deviation of 1

   B) Rescaling data to [0, 1]

   C) Encoding strings into numbers

   D) Filling missing values

9. Which pandas function is used to calculate basic descriptive statistics?
   (A) df.describe()

   B) df.stats()

   C) df.summary()

   D) df.info()

10. Which evaluation metric is defined as **TP / (TP + FP)**?
   (A) Precision

   B) Recall

   C) Accuracy

   D) F1 Score

11. In pandas, what does `df.iloc[:, 1]` return?

   A) First row

   (B) Second column

   C) All rows after the first

   D) Last column

12. Which of the following is used to normalize data to a range [0, 1]?

   A) Z-score normalization

   (B) Min-max scaling

   C) Label encoding

   D) Mean substitution

13. What is the default behavior of `dropna()` in pandas?

   A) Drops all columns

(B) Drops any row with at least one NaN

C) Fills missing values with 0

D) Replaces NaNs with the column mode

14. Which visualization is best suited for detecting outliers in a dataset?

A) Line chart

(B) Boxplot

C) Histogram

D) Pie chart

15. What does `df[df['col'] > 10]` return?

A) All rows where 'col' > 10

(B) A copy of the DataFrame with filtered rows

C) An error

D) Only the 'col' column values

16. What does `accuracy_score(y_true, y_pred)` measure?

(A) Correct predictions over total predictions

B) TP / (TP + FP)

C) TN / (TN + FP)

D) FN / (FN + TP)

17. What is the shape of a DataFrame with 5 rows and 3 columns?

(A) (5, 3)

B) (3, 5)

C) (5)

D) (3)

18. Which function displays the first few rows of a DataFrame?

(A) df.head()

B) df.tail()

C) df.describe()

D) df.first()

19. Which scaling technique is not affected by outliers?

(A) RobustScaler

B) MinMaxScaler

C) StandardScaler

D) Log Transform

20. Which of the following plots is most appropriate for visualizing relationships between two numerical variables?

(A) Scatter plot

B) Bar chart

C) Pie chart

D) Histogram

SECTION 2: 10 True and False questions. 1.5 point/question, 15 points in total

1. `df.loc[2]` retrieves the row labeled with index 2 in a pandas DataFrame. T
2. The `.iloc[]` method in pandas accesses rows or columns by their integer position. T
3. In matplotlib, `plt.plot()` is commonly used to create line charts. T
4. The IQR (Interquartile Range) is calculated as Q3 - Q1. T
5. The expression `Q3 + 1.5 * IQR` is used as a threshold to detect upper outliers. T
6. The function `df.fillna(method='ffill')` fills missing values using forward fill. T

7. Standardization transforms data to have a mean of 0 and a standard deviation of 1. T
8. Normalization typically scales values to fit in a range between 0 and 1. T
9. Precision is calculated as TP / (TP + FP), where TP is True Positive and FP is False Positive. T
10. Accuracy is always a better evaluation metric than precision for imbalanced datasets. F

SECTION 3: 10 matching questions. 1.5 point /question, 15 points in total

Match cells from the left column with cells on the right column, by writing the corresponding number in the empty column.

Choice	Column A	Answer	Column B
1.	df.loc[]	d	Scaling values to [0, 1] range ✓ 9
2.	df.iloc[]	f	Used to create line plots in matplotlib ✓ b
3.	dropna()	c	Removes rows with missing values ✓ c
4.	plt.plot()	b	Selects rows/columns by label in pandas ✓ d
5.	IQR	e	Measures data spread between Q1 and Q3 ✓ e
6.	Normalization	a	Selects rows/columns by integer position in pandas ✓ f
7.	Standardization	h	(TP + TN) / Total predictions ✓ g
8.	Accuracy	g	Converts data to have mean = 0 and std = 1 ✓ h
9.	Precision	i	TP / (TP + FP) ✓ i
10.	Outlier	j	A value significantly different from the rest ✓ j

SECTION 4: Case Studies. Diverse 10 points /question, 50 points in total

Python Case Study Exercise: Health Insurance Dataset Analysis

🎯 Objective:

You are given a dataset containing records of customers from a health insurance provider. Your task is to explore the dataset, clean it, preprocess it, and train a **Decision Tree Classifier** to predict whether a customer will buy insurance.

📁 Dataset: insurance_data.csv

Columns:

- Age (numeric)
- Gender (categorical: Male/Female)
- Salary (numeric)
- Education (categorical: High School, Bachelor's, Master's, PhD)
- City (categorical: e.g., Beirut, Tripoli, Saida)
- Buy_Insurance (target: Yes/No)

✅ Part 1 – Reading and Exploring the Data (10 pts)

Instructions:

1. Read the dataset using pandas.
2. Display the first 5 and last 5 rows.
3. Show summary statistics using .describe().
4. Check for missing values in all columns.

✅ Part 2 – Handling Missing Values and Outliers (10 pts)

Instructions:

1. Fill missing values in Age and Salary with the **median**.
2. Drop rows with missing values in the City column.
3. Remove outliers in the Salary column using the **IQR method**.

✓ Part 3 – Encoding and Indexing (10 pts)

Instructions:

1. Encode Gender and City using one-hot encoding.
2. Encode Education as ordinal:
   High School < Bachelor's < Master's < PhD
3. Convert Buy_Insurance to binary: Yes → 1. No → 0
4. Use .loc to select rows where Buy_Insurance is 1.
5. Use .iloc to display the first 3 rows and first 4 columns.

✓ Part 4 – Scaling and Splitting (10 pts)

Instructions:

1. Apply **standardization** to numeric columns: Age, Salary, and Education.
2. Apply **normalization** to the same columns (in parallel).
3. Use **standardized data** to split the dataset into 70% training and 30% testing sets.

✓ Part 5 – Classification and Evaluation (10 pts)

Instructions:

1. Train a **Decision Tree Classifier**.
2. Make predictions on the test set.
3. Evaluate the model using **accuracy** and **F1-score**.