Model answer

Exercise 1: (6 Points)

Q1: Misconception about AI (1 P):
AI systems can think and understand exactly like humans.

Q2:Difference between traditional AI and modern AI (1 P):
Traditional AI relies on hand-coded rules and logic, while modern AI uses data-driven learning techniques like machine learning.

Q3:Two main uses of AI in finance (1 P):

Fraud detection, Algorithmic trading, Risk Management,Customer Service,Personalized Banking,Regulatory Compliance

Q4:Goal of supervised learning (1 P):
To learn a function that maps inputs to correct outputs using labeled data.

Q5:Type of data used in unsupervised learning (1 P):
Unlabeled data (no explicit output labels).

Q6:Two metrics to evaluate regression models (1 P):

Mean Squared Error (MSE) and R-squared (Coefficient of determination)

Exercise 2: (8 Points)

1. **Problem type:**
Classification (binary classification: Risk = Yes/No). **The Target variable:** Risk (Yes or No). **(1 P)**
2. Build a decision tree using the ID3 algorithm to classify the target variable.

1.**Initial Entropy of Dataset (S):**

Total Patients: 10, "Risk = Yes": 6, "Risk = No": 4

1. Entropy(S)=−(0.6log2(0.6))−(0.4log2(0.4))=0.971 **(0,5 P)**
2. **Information Gain for each attribute (first level):**

Young Adult (3 patients: 1, 5, 8) **(1 P)**

Risk = No: 1 (Patient 1), Risk = Yes: 2 (Patients 5, 8)

E(SYoungAdult)=−(1/3log2(1/3))−(2/3log2(3/2))

E(SYoungAdult)=−(0.33333×−1.58496)−(0.66667×−0.58496)

E(SYoungAdult)=−(−0.52832)−(−0.38997)=0.91829

Middle-aged (3 patients: 2, 4, 10), Risk = No: 1 (Patient 4), Risk = Yes: 2 (Patients 2, 10)

E(SMiddle−aged)=−(1/3 log2 (1/3))−(2/3log2(2/3))=0.91829

Senior (4 patients: 3, 6, 7, 9), Risk = No: 2 (Patients 3, 7), Risk = Yes: 2 (Patients 6, 9)

E(SSenior)=−(2/4log2(2/4))−(2/4log2(2/4))

E(SSenior)=−(0.5×−1)−(0.5×−1)=0.5+0.5=1.0

IG(S,Age Group)=E(S)−[3/10E(SYoungAdult)+3/10E(SMiddle−aged)+4/10E(SSenior)]
IG(S,Age Group)=0.97095−[3/10(0.91829)+3/10(0.91829)+4/10(1.0)]
IG(S,Age Group)=0.97095−[0.275487+0.275487+0.4] IG(S,Age Group)=0.97095−0.950974=0.01998

b) Information Gain for 'Exercise Frequency'**(1 P)**

Daily (3 patients: 1, 4, 10), Risk = No: 2 (Patients 1, 4), Risk = Yes: 1 (Patient 10)

E(S,Daily)=−(2/3log2(2/3))−(1/3log2(1/3))=0.91829

Sedentary (4 patients: 2, 6, 8, 9) Risk = No: 0, Risk = Yes: 4 (Patients 2, 6, 8, 9)

E(S,Sedentary)=−(0log2(0))−(1log2(1))=0−0=0 (Pure node)

Light/Occasional (3 patients: 3, 5, 7), Risk = No: 2 (Patients 3, 7), Risk = Yes: 1 (Patient 5)

E(S,Light/Occasional)=−(2/3log2(2/3))−(1/3log2(1/3))=0.91829

IG(S,Exercise Frequency)=E(S)−[3/10E(SDaily)+4/10E(SSedentary)+3/10E(SLight/Occasional)]
IG(S,Exercise Frequency)=0.97095−[3/10 (0.91829)+4/10 (0)+ 3/10 (0.91829)]
IG(S,Exercise Frequency)=0.97095−[0.275487+0+0.275487]
IG(S,Exercise Frequency)=0.97095−0.550974=0.41998

c) Information Gain for 'Smoking Status'**(1 P)**

Non-smoker (5 patients: 1, 4, 5, 7, 10) , Risk = No: 3 (Patients 1, 4, 7), Risk = Yes: 2 (Patients 5, 10)

E(SNon−smoker)=−(3/5log2(3/5))−(2/5log2(2/5))

E(SNon−smoker)=−(0.6×−0.736965594)−(0.4×−1.321928095)=0.97095

Smoker (3 patients: 2, 6, 8), Risk = No: 0, Risk = Yes: 3 (Patients 2, 6, 8)

E(SSmoker)=0 (Pure node)

Ex-smoker (2 patients: 3, 9), Risk = No: 1 (Patient 3), Risk = Yes: 1 (Patient 9)

E(S,Ex−smoker)=−(1/2log2(1/2))−(1/2log2(1/2))

E(S,Ex−smoker)=−(0.5×−1)−(0.5×−1)=0.5+0.5=1.0

IG(S,Smoking Status)=E(S)−[5/10E(S,Non−smoker)+3/10E(S,Smoker)+2/10 E(S,Ex−smoker)]
IG(S,Smoking Status)=0.97095−[5/10(0.97095)+3/10(0)+2/10(1.0)]
IG(S,Smoking Status)=0.97095−[0.485475+0+0.2] IG(S,Smoking Status)=0.97095−0.685475=0.28548

d) Information Gain for 'BMI Category'**(1 P)**

Normal (3 patients: 1, 4, 10), Risk = No: 2 (Patients 1, 4), Risk = Yes: 1 (Patient 10)

E(S,Normal)=−(2/3log2(2/3))−(1/3log2(1/3))=0.91829

Obese (4 patients: 2, 6, 8, 9), Risk = No: 0, Risk = Yes: 4 (Patients 2, 6, 8, 9)

E(S,Obese)=0 (Pure node)

Overweight (3 patients: 3, 5, 7), Risk = No: 2 (Patients 3, 7), Risk = Yes: 1 (Patient 5)

E(S,Overweight)=−(2/3log2(2/3))−(1/3log2(1/3))=0.91829

IG(S,BMI Category)=E(S)−[3/10E(S,Normal)+4/10E(S,Obese)+3/10E(S,Overweight)]
IG(S,BMI Category)=0.97095−[3/10(0.91829)+4/10(0)+3/10(0.91829)]
IG(S,BMI Category)=0.97095−[0.275487+0+0.275487]
IG(S,BMI Category)=0.97095−0.550974=0.41998

Age Group: =0.0202

Exercise Frequency: =0.4202

Smoking Status: =0.2855

BMI Category: =0.4202

Root Node Selection: Both Exercise Frequency and BMI Category have the highest Information Gain (=0.4202). In the case of a tie, ID3 can arbitrarily choose one. For this solution, **'BMI Category'** is selected as the root node.**(0,25 P)**

Tree Building Process:

Split by 'BMI Category':

'Obese' Branch (Patients 2, 6, 8, 9): All 4 patients have 'Risk = Yes'. This is a **pure leaf node (Risk = Yes)**.

'Normal' Branch (Patients 1, 4, 10): This subset contains (1 'Yes', 2 'No'). This is an **impure node** (Entropy=0.918). Further splitting is required.

'Overweight' Branch (Patients 3, 5, 7): This subset contains (1 'Yes', 2 'No'). This is an **impure node** (Entropy=0.918). Further splitting is required.

Split 'Normal' BMI Branch (Patients 1, 4, 10) by 'Age Group':

Age Group is the best attribute for this subset (Information Gain =0.251).

If Age Group = 'Young Adult' (Patient 1): 'Risk = No'. This is a **pure leaf node.**

If Age Group = 'Middle-aged' (Patients 4, 10): This subset is (1 'Yes', 1 'No'). All remaining attributes provide 0 Information Gain. This becomes a leaf node with **'Risk = No' (by majority vote).**

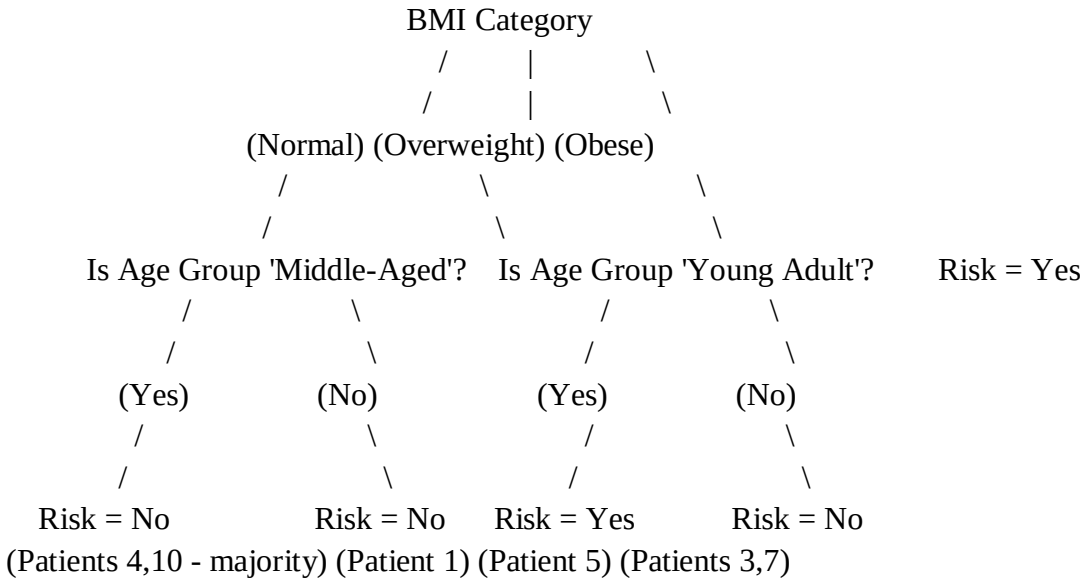
Split 'Overweight' BMI Branch (Patients 3, 5, 7) by 'Age Group':

Age Group is the best attribute for this subset (Information Gain ≈0.918).

If Age Group = 'Young Adult' (Patient 5): 'Risk = Yes'. This is a **pure leaf node.**

If Age Group = 'Senior' (Patients 3, 7): 'Risk = No'. This is a **pure leaf node.** **(0,75 P)**

3. Draw the final decision tree. (1 P)



4. Use the decision tree to predict the result for the following case:

Case: Age Group = Young Adult, Exercise Frequency = Sedentary, Smoking Status = Smoker, BMI Category = Overweight.

Prediction:

Start at the Root Node: **'BMI Category'**. The case has 'BMI Category = Overweight'. Follow the 'Overweight' branch.**(0,5 P)**

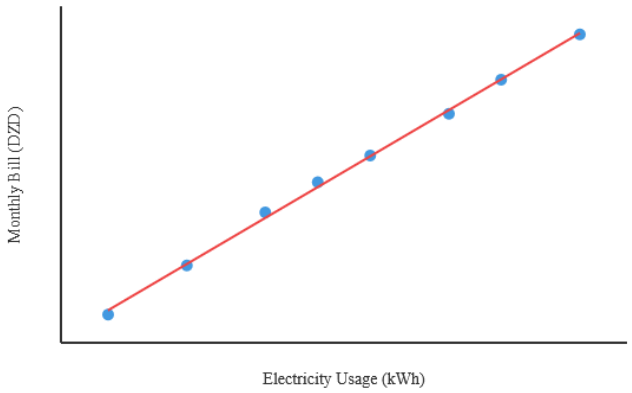
Next Node: **'Age Group'**. The case has 'Age Group = Young Adult'. Follow the 'Young Adult' branch

Leaf Node: This path leads to a leaf node. **Predicted Outcome for the given case: Risk = Yes.(2 P)**

Exercise 3: (6 Points): Linear Regression for Electricity Usage

1. Plot the data on a graph.

The data points are plotted with 'Electricity usage (kWh)' on the x-axis and 'Monthly bill (DZD)' on the y-axis. A linear regression line is also drawn to show the relationship.



(1 P)

2. Compute the Linear Regression Line using the slope (m) and the intercept (b).

The linear regression line is represented by the equation: y=mx+b.

Calculations for Slope (m) and Intercept (b) using Least Squares Method:

Number of data points (n) = 8

X (kWH)	Y (DZD)	$X \cdot Y$	X^2
120	540	$120 \times 540 = 64800$	$120^2 = 14400$
150	670	$150 \times 670 = 100500$	$150^2 = 22500$
180	810	$180 \times 810 = 145800$	$180^2 = 32400$

X (KWH)	Y (DZD)	X • Y	X^2
200	890	$200 \times 890 = 178000$	$200^2 = 40000$
220	960	$220 \times 960 = 211200$	$220^2 = 48400$
250	1070	$250 \times 1070 = 267500$	$250^2 = 62500$
270	1160	$270 \times 1160 = 313200$	$270^2 = 72900$
300	1280	$300 \times 1280 = 384000$	$300^2 = 90000$

$\Sigma X = 1690$ $\Sigma Y = 7380$ $\Sigma XY = 1,665,000$ $\Sigma X^2 = 383,100$

$$n \cdot \sum(xy) = 8 \cdot 1{,}665{,}000 = 13{,}320{,}000$$

$$\sum x \cdot \sum y = 1690 \cdot 7380 = 12{,}472{,}200$$

$$\text{Numérateur} = 13{,}320{,}000 - 12{,}472{,}200 = \boxed{847{,}800}$$

$$n \cdot \sum(x^2) = 8 \cdot 383{,}100 = 3{,}064{,}800$$

$$(\sum x)^2 = 1690^2 = 2{,}856{,}100$$

$$\text{Dénominateur} = 3{,}064{,}800 - 2{,}856{,}100 = \boxed{208{,}700}$$

$$m = \frac{847{,}800}{208{,}700}$$

m= 4,062

$b = \bar{y} - m.\bar{x}$

$\bar{x}=\frac{\sum x}{n}$

$\bar{y}=\frac{\sum y}{n}$

m=4,062, $\sum x$=1690, $\sum y = 7380$

$$4.062 \times 211.25 = 858.4875$$

$$b = 922.5 - 858.4875 = \boxed{64.0125}$$

$$b = 922.5 - (4.062 \times 211.25)$$

$$\boxed{y = 4.062x + 64.0125}$$

3. Use the regression model to predict the monthly bill for: 160 kWh and 280 kWh.

Prediction for 160 kWh:**(0,5 P)**

$$y = 4.062 \times 160 + 64.0125$$

$$y = 649.92 + 64.0125 = \boxed{713.93 \text{ DZD}}$$

Prediction for 280 kWh:**(0,5 P)**

$$y = 4.062 \times 280 + 64.0125$$

$$y = 1137.36 + 64.0125 = \boxed{1201.37 \text{ DZD}}$$

3. Using Python and scikit-learn, write the code to split the data into training and test sets (with 20% test size), train a linear regression model on the training data, and predict the target values for the test data.(2 P bonus)

```
import numpy as np

from sklearn.model_selection import train_test_split

from sklearn.linear_model import LinearRegression

# Electricity usage (kWh) - Features (X)

X = np.array([120, 150, 180, 200, 220, 250, 270, 300]).reshape(-1, 1)
```

```
# Monthly bill (DZD) - Target (y)

y = np.array([540, 670, 810, 890, 960, 1070, 1160, 1280])

# 4. Split the data into training and test sets (20% test size)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train a linear regression model on the training data

model = LinearRegression()

model.fit(X_train, y_train)

# Predict the target values for the test data

y_predicted = model.predict(X_test)

5. Print the predicted values for the test.

# 5. Print the predicted values for the test set along with the actual target value

print(y_predicted)
```