

## **Complete Solutions for All Questions (All 3 Images)**

### **1. Pearson Correlation, Interpretation, Imputation, Encoding, Normalization, SMC & Jaccard**

Computed correlation  $r \approx 0.978$  (strong positive correlation).

Two imputation methods: mean/median and KNN/regression imputation.

Ordinal encoding: HighSchool=0, Bachelor=1, Master=2, PhD=3.

One-hot encoding for Department categories {Sales, Ops, R&D;}.

Label encoding misleads linear models because it produces false ordering.

Min-max normalized age(35) = 0.523. Z-score = 0.

SMC = 0.6667, Jaccard = 0.6 (better for sparse baskets).

## Numeric Tables (Correlation/Regression/Decision Tree)

### Decision Tree: Entropy Table

Class	Count	Probability	$-p \log_2 p$
Yes	4	0.667	0.38997
No	2	0.333	0.52832

### Regression Deviations Table

x	y	$x-\bar{x}$	$y-\bar{y}$	Product	$(x-\bar{x})^2$
1	3	-2	-2.8	5.6	4
2	5	-1	-0.8	0.8	1
3	6	0	0.2	0	0
4	7	1	1.2	1.2	1
5	8	2	2.2	4.4	4

### Regression Errors Table

x	y	$\hat{y}$	Error	Error	Error^2
1	3	3.4	-0.4	0.4	0.16
2	5	4.6	0.4	0.4	0.16
3	6	5.8	0.2	0.2	0.04
4	7	7.0	0.0	0.0	0.00
5	8	8.2	-0.2	0.2	0.04

### **Decision Tree (Entropy, IG, Tree Construction)**

Overall entropy = 0.9183.

$IG(\text{Studied}) = 0.459$ ,  $IG(\text{GPA})=0.252 \rightarrow$  first split = Studied.

Studied=T → Passed=Yes.

Studied=F → Split on GPA → H=Yes; L,M=No.

## **Machine Learning Type Identification**

- i) Fraud detection → Supervised (because labels 'fraud/legit' exist).
- ii) Grouping songs → Unsupervised (clustering by patterns).
- iii) Drone flight optimization → Reinforcement learning (rewards).

## **Supervised vs Unsupervised vs Semi-supervised vs RL**

Supervised: trained with labeled data.

Unsupervised: no labels; find patterns.

Semi-supervised: some labeled + many unlabeled.

Reinforcement: agent learns via rewards.

## **Linear Regression Full Solution**

Regression line:  $y = 1.2x + 2.2$ .

Prediction at  $x=3.2 \rightarrow y=6.04$ .

$R^2=0.973$ ,  $MAE=0.24$ ,  $MSE=0.08$ .