Assignment - 1

Q.1 Air-traffic data problem

→ Dataset:

Days	Season	Fog	Rain	Class
Weekday	Spring	None	None	on time
Weekday	Winter	None	Slight	on time
Weekday	Winter	None	None	on time
Holiday	Winter	High	Slight	late
Saturday	Summer	Normal	None	on time
Weekday	Autumn	Normal	None	very late
Holiday	Summer	High	slight	on time
Sunday	summer	Normal	None	on time
Weekday	Winter	High	Heavy	Very late
Weekday	Summer	None	Slight	on time
Saturday	Spring	High	Heavy	cancelled
Weekday	Summer	High	Slight	on time
Weekday	winter	Normal	None	late
Weekday	summer	High	None	on time
Weekday	winter	Normal	Heavy	Very late
Saturday	Autumn	High	Slight	on time
Weekday	Autumn	None	Heavy	on time
Holiday	Spring	Normal	Slight	on time
Weekday	Spring	Normal	None	on time
Weekday	Spring	Normal	Heavy	on time

A = [day, season, fog, rain]

C = [on time, late, very late, cancelled]

Attributes:

① Day:

	On-time	late	Very late	Cancelled
Weekday	9/14 = 0.64	1/2 = 0.5	3/3 = 1	0/1 = 0
Saturday	2/14 = 0.14	0/2 = 0.0	0/3 = 0	1/1 = 1
Sunday	1/14 = 0.07	0/2 = 0	0/3 = 0	0/1 = 0
Holiday	2/14 = 0.14	1.0/2 = 1.0	0/3 = 0	0/1 = 0

② Season:

	On-time	late	Very late	Cancelled
Spring	4/14 = 0.29	0/2 = 0	0/3 = 0	1.0/1 = 1.0
Summer	6/14 = 0.43	0/2 = 0	0/3 = 0	0/1 = 0
Autumn	2/14 = 0.14	0/2 = 0	1/3 = 0.33	0/1 = 0
Winter	2/14 = 0.14	2/2 = 1	2/3 = 0.67	0/1 = 0

③ Fog:

	On-time	late	Very-late	Cancelled
None	5/14 = 0.36	0/2 = 0	0/3 = 0	0/1 = 0
High	4/14 = 0.29	1/2 = 0.5	1/3 = 0.33	1/1 = 1
Normal	5/14 = 0.36	1/2 = 0.5	2/3 = 0.67	0/1 = 0

④ Rain:

	On-time	late	Very late	Cancelled
None	6/14 = 0.43	1/2 = 0.5	1/3 = 0.33	0/1 = 0
Slight	6/14 = 0.43	1/2 = 0.5	0/3 = 0	0/1 = 0
Heavy	2/14 = 0.14	0/2 = 0	2/3 = 0.67	1/1 = 1

Prior probability:

On-time	late	very late	Cancelled
14/20 = 0.70	2/20 = 0.10	3/20 = 0.15	1/20 = 0.05

given instance :

Weekday, winter, High, Heavy, ?

① Case 1 :

on time = 0.70 × 0.64 × 0.14 × 0.29 × 0.14 = 0.0013

② Case 2 :
late = 0.10 × 0.50 × 1 × 0.50 × 0.00 = 0.0

③ Case 3 :
Very late = 0.15 × 1 × 0.67 × 0.33 × 0.67 = 0.0222

④ Case 4 :
Cancelled = 0.05 × 0.0 × 0.0 × 1.0 × 1.0 = 0.000

∵ case 3 is strongest
∴ The correct classification for give instance is "very late".

∴ Weekday, winter, High, Heavy, ? = very late.

Q.2

	Male	Female	Total
fiction	250 (90)	200 (360)	450
non-fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

⟶

$$e_{ij} = \frac{\text{count}(A = a_i) \times \text{count}(B = b_j)}{n}$$

$$\therefore e_{11} = \frac{\text{count(male)} \times \text{count(fiction)}}{n} = \frac{300 \times 450}{1500} = \underline{\underline{90}}$$

$$e_{12} = \frac{\text{count (female)} \times \text{count (fiction)}}{n} = \frac{1200 \times 450}{1500} = \underline{\underline{360}}$$

$$e_{13} = \frac{\text{count(male)} \times \text{count (non-fiction)}}{n} = \frac{300 \times 1050}{1500} = \underline{\underline{210}}$$

$$e_{14} = \frac{\text{count (female)} \times \text{count (non-fiction)}}{n} = \frac{1200 \times 1050}{1500} = \underline{\underline{840}}$$

in any row, sum of expected frequencies must equal total observed frequency for that row and sum of expected frequencies in any column must also equal, total observed frequency for that column

$$\because \chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

Q.2

→ $\therefore x^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360}$

$+ \frac{(1000-840)^2}{840}$

$= 284.44 + 121.90 + 71.11 + 30.48$

$= \underline{\underline{507.93}}$

degree of freedom,

$(2-1)(2-1) = 1$

for 1 degree of freedom, the x^2 value needed to reject hypothesis at 0.001 significance level is 10.828

$\because$ our value is above this, we can reject the hypothesis that gender and preferred reading are independent and conclude that two attributes are strongly correlated for given group of people.