SRUTHI S
2019110059

DA Assignment-1

(1) The classification technique that can be used to Map this tuple into an accurate class is 'Naive Bayes' Classifier'. This is because we need to find the most likely classification.
We assume each feature of the tuple makes an independent & equal contribution to the outcome
With Naive Bayes we try to find the Maximum Likelihood.
Using Bayes' theorem,

$$P(y|x_1,\dots,x_n) = \frac{P(x_1|y)\cdot P(x_2|y)\dots P(x_n|y)\cdot P(y)}{P(x_1)\cdot P(x_2)\dots P(x_n)}$$

For the given Air traffic data,
There 20 tuples. ~~with~~

~~P(on time) =~~ $P(\text{on time}) = \frac{14}{20} = 0.7$ [14 instances with class as on time]

$P(\text{late}) = \frac{2}{20} = 0.1$ [2 instances with class as late]

$P(\text{Very late}) = \frac{3}{20} = 0.15$ [3 instances with class Very late]

$P(\text{cancelled}) = \frac{1}{20} = 0.05$ [1 instance with class cancelled]

Probability with respect to each attribute

Finding conditional probabilites for each attribute

~~F Days.~~ <u>For Days</u>:

Days	OnTime	Late	Very Late	Cancelled
Weekday	9/14	1/2	3/3	0/1
Holiday	2/14	1/2	0/3	0/1
Saturday	2/14	0/2	0/3	1/1
Sunday	1~~2~~/14	0/2	0/3	0/1

<u>For Season</u>:

Season	OnTime	Late	Very Late	Cancelled
Spring	4/14	0/2	0/3	1/1
Winter	2/14	2/2	2/3	0/1
Summer	6~~5~~/14	0/2	0/3	0/1
Autumn	2/14	0/2	1/3	0/1

<u>For Rain</u>:

~~Fog~~ Rain	OnTime	Late	Very Late	Cancelled
None	6/14	1/~~3~~2	1/3	0/1
Slight	6/14	1/2	~~2~~ 0/3	0/1
Heavy	2/14	0/2	2/3	1/1

<u>For Fog</u>:

SRUTHI S
2019110059

Page No
Date

Fog	OnTime	Late	Very Late	Cancelled
None	5/14	0/2	0/~~2~~3	0/1
High	4/14	1/2	1/3	1/1
Normal	5/14	1/2	2/3	0/1

Let's find the probability for each case,

~~Case 1 : (Day = Weekday, Season = Winter, Fog = High, Rain = None) = Instance 1~~
~~∴ OnTime~~
~~P(Weekday / Instance 1) = P(WeekDay) . P(Day = WeekDay) . P~~

~~Case 1:~~ Let Instance 1 = (Day = WeekDay, Season = Winter, Fog = High, Rain = None)

Case 1: OnTime

P(OnTime / Instance 1) = P(OnTime) . P(Day = WeekDay / Ontime) . P(Season = Winter / Ontime) . P(Fog = High / OnTime) . P(Rain = None / Ontime)

$$= 0.7 \times \frac{9}{14} \times \frac{2}{14} \times \frac{4}{14} \times \frac{6}{14} = 0.0078717$$

Case 2: ~~Let~~ Late

P(Late / Instance 1) = P(Late) . P(Day = WeekDay / Late) . P(Season = Winter / Late) . P (Fog = High / Late) . P(Rain = None / Late)

$$= 0.1 \times \frac{1}{2} \times \frac{2}{2} \times \frac{1}{2} \times \frac{1}{2} = \boxed{0.0125}$$

SRUTHI S
2019110059

Case 3: Very Late

P(Very Late / Instance 1) = P(Very Late) · P(Day = WeekDay / Very Late) · P(Season = Winter / Very Late) · P(Rain = None / Very Late) · P(Fog = High / Very Late)

$= 0.15 \times \frac{3}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{1}{3}$

$= 0.0111$

Case 4: Cancelled

P(Cancelled / Instance 1) = P (Cancelled) · P(Day = Weekday / Cancelled) · P(Season = Winter / Cancelled) · P(~~Rain~~ Fog = High / Cancelled) · P(Rain = None / Cancelled)

$= 0.05 \times \frac{0}{1} \times \frac{0}{1} \times \frac{1}{1} \times \frac{0}{1} = 0$

The Highest probability occurs for the case ~~when~~ of ~~Very~~ Late. Hence, when the Day is a Weekday, Season is Winter, Fog is High & Rain is None, the class is most likely to be '<u>Late</u>'

(2) Given: Sample size (n) = 1500

Contingency table

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

The values in '( )' are expected values.

Hypothesis:

H0: Prefferred Reading & gender are

SRUTHI S
2019110069

independent of each other
Ha: Prefferred Reading & gender are not independent of each other

We will ~~perr~~ perform a chi square test on the given data to test the Hypothesis.

$$X^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

where O_{ij} = Observed frequency
e_{ij} = Expected frequency

m & n are no of rows & columns respectively.

$$\therefore X^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$

$$= \frac{(160)^2}{90} + \frac{(-160)^2}{210} + \frac{(-160)^2}{360} + \frac{(160)^2}{840}$$

$$= (160)^2 \left[\frac{1}{90} + \frac{1}{210} + \frac{1}{360} + \frac{1}{840} \right]$$

$$= 507.93650$$

~~Comparing~~ Degree of freedom is $(m-1) \times (n-1)$

Here $m = 2$, $n = 2$

$\therefore$ Degree of freedom $= (2-1)\times(2-1) = 1$

~~comp~~
Value X^2 with degree of freedom 1 and

SRUTHI S
2019110059

0.01 significance level from the standard statistical table is 6.635
As value ~~of~~ obtained is > 6.635 we reject the null hypothesis that ~~G~~ Preferred reading & gender are independent of each other. Hence, concluding that Gender and Preferred reading are 'strongly correlated' with each other.