20119035 ①

CSE-320

1)

```
data <- read.csv("demographic data.csv")

bP <- c(-inf, 10000, 25000, inf)

name <- c("Low", "Medium", "High")

data$ Income_label <- cut(student$Income,
                 breaks = bP, labels = name)

data
```

output:-

| Age | State | Gender | Income | Income_label |
|---|---|---|---|---|
| 23 | TN | F | 5000 | Low |
| 13 | AP | M | 1000 | Low |
| 36 | UP | M | 3000 | Low |
| 31 | TN | F | 4000 | Low |
| 58 | PY | M | 10000 | Low |
| 29 | PY | M | 50000 | High |
| 39 | TH | ~~M~~ F | 2000 | Low |
| 23 | UP | F | 20,000 | Medium |

2 Rules

| | |
|---|---|
| 1 | $\{B\} \to \{C\}$ |
| 2 | $\{A\} \to \{D\}$ |
| 3 | $\{B\} \to \{D\}$ |
| 4 | $\{E\} \to \{C\}$ |
| 5 | $\{C\} \to \{A\}$ |

| | |
|---|---|
| 1 | $\{A, B, D, E\}$ |
| 2 | $\{B, C, D\}$ |
| 3 | $\{A, B, D, E\}$ |
| 4 | $\{A, C, D, E\}$ |
| 5 | $\{B, C, D, E\}$ |
| 6 | $\{B, D, E\}$ |
| 7 | $\{C, D\}$ |
| 8 | $\{A, D, E\}$ |
| 9 | $\{A, B, C\}$ |
| 10 | $\{B, D\}$ |

a) $\{B\} \to \{C\}$

| | C | C' | | |
|---|---|---|---|---|
| B | 3 | 4 | 7 | $f_{1+}$ |
| B' | 2 | 1 | 3 | $f_{0+}$ |
| | 5 | 5 | | |
| | $f_{+1}$ | $f_{+0}$ | | |

contingency table

| | Y | $\bar{Y}$ | |
|---|---|---|---|
| X | $f_{11}$ | $f_{10}$ | $f_{1+}$ |
| $\bar{X}$ | $f_{01}$ | $f_{00}$ | $f_{0+}$ |
| | $f_{+1}$ | $f_{-0}$ | |

2. $\{A\} \to \{D\}$

| | D | D' | |
|---|---|---|---|
| A | 4 | 1 | 5 |
| A' | 5 | 0 | 5 |
| | 9 | 1 | |

3. $\{B\} \to \{D\}$

| | D | D' | |
|---|---|---|---|
| B | 6 | 1 | 7 |
| B' | 3 | 0 | 3 |
| | 9 | 1 | |

4. $\{E\} \to \{C\}$

| | C | C' | |
|---|---|---|---|
| E | 2 | 4 | 6 |
| E' | 3 | 1 | 4 |
| | 5 | 5 | |

5. $\{C\} \to \{A\}$

| | A | A' | |
|---|---|---|---|
| C | 2 | 3 | 5 |
| C' | 3 | 2 | 5 |
| | 5 | 5 | |

b) Support

~~Rule~~

B)

Support

| | Support | Rank |
|---|---|---|
| $\{B\} \rightarrow \{C\}$ | $3/10 = 0.3$ | 3 |
| $\{A\} \rightarrow \{D\}$ | 0.4 | 2 |
| $\{B\} \rightarrow \{D\}$ | 0.6 | 1 |
| $\{E\} \rightarrow \{C\}$ | 0.2 | 4 |
| $\{C\} \rightarrow \{A\}$ | 0.2 | 4 |

Confidence

| | confidence | Reank |
|---|---|---|
| $\{B\} \rightarrow \{C\}$ | 0.6 | 3 |
| $\{A\} \rightarrow \{D\}$ | 0.8 | 2 |
| $\{B\} \rightarrow \{D\}$ | 0.857 | 1 |
| $\{E\} \rightarrow \{C\}$ | 0.33 | 5 |
| $\{C\} \rightarrow \{A\}$ | 0.4 | 4 |

Odds Ratio $(X \rightarrow Y)$. $\dfrac{P(x, y)\; P(\bar{x}\bar{y})}{P(x\bar{y})\; P(\bar{X}Y)}$

| | odds Ratio | Rank |
|---|---|---|
| $\{D\} \rightarrow \{E\}$ | 0.325 | 2 |
| $\{A\} \rightarrow \{D\}$ | 0 | 4 |
| $\{B\} \rightarrow \{D\}$ | 0 | 4 |
| $\{E\} \rightarrow \{C\}$ | 0.167 | 3 |
| $\{C\} \rightarrow \{A\}$ | 0.444 | 1 |

(c)

Correlation (confident, Support) = 0.97

correlation (confident, odds Ratio) = - 0.606

$\therefore$

Support is highly correlated

Odds is least correlated

3)

In BOW the created sets of vectors only contain the frequency of the word. In TF-IDF will show the importance of term in document.

TF show frequency of word and
IDF shows the importance of word in the docs

From figure, The Kalam has the most importance and is the word with the highest frequency followed India, Technology and nuclear