STAT 223

Laboratory Activity #9 by Akbota Serikkyzy

List of Transactions

(a)

Support({e}) = $\frac{8}{10} = 0.8$
Support({b, d}) = $\frac{2}{10} = 0.2$
Support({b, d, e}) = $\frac{2}{10} = 0.2$

(b)

Support({b, d} $\rightarrow$ {e}) = Support({b, d, e}) = 0.2
Support({e} $\rightarrow$ {b, d}) = Support({b, d, e}) = 0.2
Support is symmetric since Support({b, d} $\rightarrow$ {e}) equal to Support({e} $\rightarrow$ {b, d}).

(c)

Confidence({b, d} $\rightarrow$ {e}) = $\frac{0.2}{0.2} = 1.0$
Confidence({e} $\rightarrow$ {b, d}) = $\frac{0.2}{0.8} = 0.25$
Confidence is not symmetric since they are not equal.

(d)

Lift({b, d} $\rightarrow$ {e}) = $\frac{1.0}{0.8} = 1.25$
Lift({e} $\rightarrow$ {b, d}) = $\frac{0.25}{0.2} = 1.25$
Lift is symmetric since they are equal.

Gini, Entropy, Information Gain

(a)

$$G = 1 - \left(\frac{17}{45}\right)^2 - \left(\frac{15}{45}\right)^2 - \left(\frac{13}{45}\right)^2 \approx 0.663$$

(b)

$$E = -\left(\frac{17}{45}\log_2\frac{17}{45} + \frac{15}{45}\log_2\frac{15}{45} + \frac{13}{45}\log_2\frac{13}{45}\right) \approx 1.576$$

(c)

$$G1 = 1 - \left(\frac{16}{30}\right)^2 - \left(\frac{10}{30}\right)^2 - \left(\frac{4}{30}\right)^2 \approx 0.587$$

$$G2 = 1 - \left(\frac{1}{15}\right)^2 - \left(\frac{5}{15}\right)^2 - \left(\frac{9}{15}\right)^2 \approx 0.524$$

(d)

$$E1 = -\left(\frac{16}{30}\log_2\frac{16}{30} + \frac{10}{30}\log_2\frac{10}{30} + \frac{4}{30}\log_2\frac{4}{30}\right) \approx 0.4$$

$$E2 = -\left(\frac{1}{15}\log_2\frac{1}{15} + \frac{5}{15}\log_2\frac{5}{15} + \frac{9}{15}\log_2\frac{9}{15}\right) \approx 1.231$$

(e)

$$G_s = \frac{30}{45}\cdot 0.587 + \frac{15}{45}\cdot 0.524 \approx 0.566$$

(f)

$$IG = 1.576 - \left(\frac{30}{45}\cdot 0.4 + \frac{15}{45}\cdot 1.231\right) \approx 0.899$$

(g)

Yes because the split decreases the Gini index and provides a positive gain in information. Therefore, it improves purity and better to be kept.

Market Basket Transactions

Rule	Antecedent	Consequent	Support	Confidence	Lift
66	(Eggs, Bread)	(Milk)	0.3981	0.7186	1.0148
22	(Eggs)	(Milk)	0.5619	0.7186	1.0221
67	(Cookies, Eggs)	(Milk)	0.3530	0.7186	1.0234
30	(Hamburger, Pizza)	(Beer)	0.3895	0.6645	1.0598
10	(Pizza)	(Beer)	0.6467	0.6645	1.0662

Table 1: Association Rules

As can be seen in figure 1, Pop is the most popular item to buy, followed by milk and then bread. These items appeared the most often across all transactions.

When it comes to the table 1, it shows just 5 of the total of 63 association rules, those with the highest lift values. These rules are the most interesting because high lift means that the items

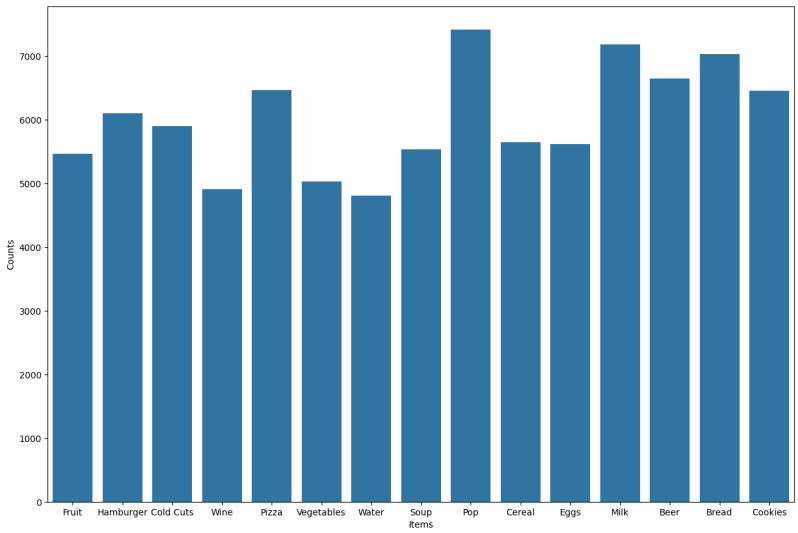

Figure 1: Item Frequency Bar Plot in Market Basket Transactions

appear together more often than would be expected by chance. For example, the rule Pizza and Beer has the highest lift among the five, suggesting that people who buy pizza are more likely than usual to also buy beer.

MLBData.csv

Model	**Accuracy**
Decision Tree (no depth limit)	80.1
Decision Tree (max depth = 3)	82.7

Table 2: Accuracy of Decision

As can be seen in table 2, I trained the decision tree without any depth limit and obtained an accuracy of 80.1%. Then, I tried limiting the tree to a maximum depth of 3 to avoid overfitting. So, the accuracy actually improved to 82.7%, showing that sometimes a simpler tree can perform better by generalizing more.