# Pattern Mining - Report

## Preprocessing

The given dataset has been preprocessed using feature selection and conversion of all the categorical values of the selected attributes to meaningful labels, which makes them easier to be identified as items in the transactions list. Data smoothing techniques have not been employed since the dataset is relatively small. However, irrelevant columns like StudentID have been removed.

Selected Attributes:  4, 5, 6, 7, 10, 11, 12, 13, 15, 16, 17, 18, 23, 29

## Association Rules

### Using Confidence

```
# Support & Confidence # 1

frequent_itemsets = apriori(data_itemsets, min_support=0.3, use_colnames=True)
rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.4)
top_rules = rules.sort_values(by='confidence', ascending=False).head(10)
top_rules
```

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| 290 | (Less than 5 hours, housewife) | (married) | 0.317241 | 0.875862 | 0.310345 | 0.978261 | 1.116912 |
| 515 | (No Work, Inactive, housewife) | (married) | 0.337931 | 0.875862 | 0.324138 | 0.959184 | 1.095131 |
| 580 | (No Work, state, housewife) | (married) | 0.331034 | 0.875862 | 0.317241 | 0.958333 | 1.094160 |
| 284 | (Present, Less than 5 hours) | (married) | 0.379310 | 0.875862 | 0.358621 | 0.945455 | 1.079456 |
| 279 | (No Work, Less than 5 hours) | (married) | 0.365517 | 0.875862 | 0.344828 | 0.943396 | 1.077106 |
| 434 | (No Work, Bus, housewife) | (married) | 0.365517 | 0.875862 | 0.344828 | 0.943396 | 1.077106 |
| 344 | (No Work, housewife) | (married) | 0.455172 | 0.875862 | 0.427586 | 0.939394 | 1.072536 |
| 231 | (No Work, Inactive) | (married) | 0.448276 | 0.875862 | 0.420690 | 0.938462 | 1.071472 |
| 114 | (Bus, Less than 5 hours) | (married) | 0.324138 | 0.875862 | 0.303448 | 0.936170 | 1.068856 |
| 398 | (m_edu_primary, married) | (housewife) | 0.324138 | 0.710345 | 0.303448 | 0.936170 | 1.317910 |

```
# Support & Confidence # 2

frequent_itemsets = apriori(data_itemsets, min_support=0.4, use_colnames=True)
rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.5)
top_rules = rules.sort_values(by='confidence', ascending=False).head(10)
top_rules
```

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| **101** | (No Work, housewife) | (married) | 0.455172 | 0.875862 | 0.427586 | 0.939394 | 1.072536 |
| **85** | (No Work, Inactive) | (married) | 0.448276 | 0.875862 | 0.420690 | 0.938462 | 1.071472 |
| **24** | (Less than 5 hours) | (married) | 0.510345 | 0.875862 | 0.475862 | 0.932432 | 1.064588 |
| **106** | (No Work, state) | (married) | 0.462069 | 0.875862 | 0.427586 | 0.925373 | 1.056528 |
| **51** | (No Work, Bus) | (married) | 0.448276 | 0.875862 | 0.413793 | 0.923077 | 1.053907 |
| **95** | (No Work, Present) | (married) | 0.496552 | 0.875862 | 0.455172 | 0.916667 | 1.046588 |
| **32** | (No Work) | (married) | 0.662069 | 0.875862 | 0.606897 | 0.916667 | 1.046588 |
| **7** | (Rental) | (Bus) | 0.468966 | 0.675862 | 0.427586 | 0.911765 | 1.349040 |
| **90** | (Inactive, housewife) | (married) | 0.475862 | 0.875862 | 0.427586 | 0.898551 | 1.025904 |
| **129** | (state, housewife) | (married) | 0.524138 | 0.875862 | 0.468966 | 0.894737 | 1.021550 |

[31]
```
# Support & Confidence # 3

frequent_itemsets = apriori(data_itemsets, min_support=0.5, use_colnames=True)
rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.7)
top_rules = rules.sort_values(by='confidence', ascending=False).head(10)
top_rules
```

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| **5** | (No Work) | (married) | 0.662069 | 0.875862 | 0.606897 | 0.916667 | 1.046588 |
| **18** | (Bus, housewife) | (married) | 0.572414 | 0.875862 | 0.510345 | 0.891566 | 1.017930 |
| **4** | (Inactive) | (married) | 0.600000 | 0.875862 | 0.531034 | 0.885057 | 1.010499 |
| **16** | (state) | (married) | 0.710345 | 0.875862 | 0.627586 | 0.883495 | 1.008715 |
| **12** | (housewife) | (married) | 0.710345 | 0.875862 | 0.627586 | 0.883495 | 1.008715 |
| **8** | (Present) | (married) | 0.758621 | 0.875862 | 0.668966 | 0.881818 | 1.006800 |
| **3** | (Bus) | (married) | 0.675862 | 0.875862 | 0.593103 | 0.877551 | 1.001928 |
| **19** | (Bus, married) | (housewife) | 0.593103 | 0.710345 | 0.510345 | 0.860465 | 1.211334 |
| **1** | (Bus) | (housewife) | 0.675862 | 0.710345 | 0.572414 | 0.846939 | 1.192292 |
| **20** | (housewife, married) | (Bus) | 0.627586 | 0.675862 | 0.510345 | 0.813187 | 1.203185 |

**Insights:**

- If the students do not do any additional work, they usually have married parents
- If the mothers are housewives, they are usually married
- Students with married parents have good attendance in classes and study fewer hours
- Students with bus as their primary transport have married parents and their mothers are usually housewives

These rules are very interesting as they give a peek into the lives of students with married parents. Such students usually do not work, study less but maintain good attendance in classes. Additionally, the mothers of such parents are mostly housewives.

## Using Interest

```
# Support & Interest # 1

frequent_itemsets = fpgrowth(data_itemsets, min_support=0.3, use_colnames=True)
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1.2)
top_rules = rules.sort_values(by="lift", ascending=False).head(10)
top_rules
```

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| **79** | (Rental) | (No Work, Bus) | 0.468966 | 0.448276 | 0.303448 | 0.647059 | 1.443439 |
| **76** | (No Work, Bus) | (Rental) | 0.448276 | 0.468966 | 0.303448 | 0.676923 | 1.443439 |
| **59** | (Rental, married) | (Bus, housewife) | 0.413793 | 0.572414 | 0.337931 | 0.816667 | 1.426707 |
| **56** | (Bus, housewife) | (Rental, married) | 0.572414 | 0.413793 | 0.337931 | 0.590361 | 1.426707 |
| **58** | (Rental, housewife) | (Bus, married) | 0.400000 | 0.593103 | 0.337931 | 0.844828 | 1.424419 |
| **57** | (Bus, married) | (Rental, housewife) | 0.593103 | 0.400000 | 0.337931 | 0.569767 | 1.424419 |
| **62** | (Rental) | (Bus, housewife, married) | 0.468966 | 0.510345 | 0.337931 | 0.720588 | 1.411963 |
| **53** | (Bus, housewife, married) | (Rental) | 0.510345 | 0.468966 | 0.337931 | 0.662162 | 1.411963 |
| **41** | (Rental) | (Bus, married) | 0.468966 | 0.593103 | 0.386207 | 0.823529 | 1.388509 |
| **38** | (Bus, married) | (Rental) | 0.593103 | 0.468966 | 0.386207 | 0.651163 | 1.388509 |

[35]
```
# Support & Interest # 2

frequent_itemsets = fpgrowth(data_itemsets, min_support=0.4, use_colnames=True)
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1.1)
top_rules = rules.sort_values(by="lift", ascending=False).head(10)
top_rules
```

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| 27 | (Rental) | (Bus) | 0.468966 | 0.675862 | 0.427586 | 0.911765 | 1.349040 |
| 26 | (Bus) | (Rental) | 0.675862 | 0.468966 | 0.427586 | 0.632653 | 1.349040 |
| 10 | (Bus, state) | (housewife) | 0.482759 | 0.710345 | 0.420690 | 0.871429 | 1.226768 |
| 13 | (housewife) | (Bus, state) | 0.710345 | 0.482759 | 0.420690 | 0.592233 | 1.226768 |
| 2 | (Bus, married) | (housewife) | 0.593103 | 0.710345 | 0.510345 | 0.860465 | 1.211334 |
| 5 | (housewife) | (Bus, married) | 0.710345 | 0.593103 | 0.510345 | 0.718447 | 1.211334 |
| 4 | (Bus) | (housewife, married) | 0.675862 | 0.627586 | 0.510345 | 0.755102 | 1.203185 |
| 3 | (housewife, married) | (Bus) | 0.627586 | 0.675862 | 0.510345 | 0.813187 | 1.203185 |
| 29 | (housewife) | (Rental) | 0.710345 | 0.468966 | 0.400000 | 0.563107 | 1.200742 |
| 28 | (Rental) | (housewife) | 0.468966 | 0.710345 | 0.400000 | 0.852941 | 1.200742 |

[36]
```
# Support & Interest # 3

frequent_itemsets = fpgrowth(data_itemsets, min_support=0.5, use_colnames=True)
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1)
top_rules = rules.sort_values(by="lift", ascending=False).head(10)
top_rules
```

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| 13 | (Bus, married) | (housewife) | 0.593103 | 0.710345 | 0.510345 | 0.860465 | 1.211334 |
| 16 | (housewife) | (Bus, married) | 0.710345 | 0.593103 | 0.510345 | 0.718447 | 1.211334 |
| 15 | (Bus) | (housewife, married) | 0.675862 | 0.627586 | 0.510345 | 0.755102 | 1.203185 |
| 14 | (housewife, married) | (Bus) | 0.627586 | 0.675862 | 0.510345 | 0.813187 | 1.203185 |
| 10 | (Bus) | (housewife) | 0.675862 | 0.710345 | 0.572414 | 0.846939 | 1.192292 |
| 11 | (housewife) | (Bus) | 0.710345 | 0.675862 | 0.572414 | 0.805825 | 1.192292 |
| 24 | (No Work) | (married) | 0.662069 | 0.875862 | 0.606897 | 0.916667 | 1.046588 |
| 25 | (married) | (No Work) | 0.875862 | 0.662069 | 0.606897 | 0.692913 | 1.046588 |
| 7 | (housewife) | (state) | 0.710345 | 0.710345 | 0.524138 | 0.737864 | 1.038741 |
| 6 | (state) | (housewife) | 0.710345 | 0.710345 | 0.524138 | 0.737864 | 1.038741 |

**Insights:**

- Students living on rent usually commute by bus
- Students commuting by bus and not working are usually living on rent
- If the mothers are housewives, their kids mostly attended state schools
- Married households with mothers being housewives and students commuting by bus are usually living on rent

These rules generated using lift focus more on the socio-economic condition of the students showcasing that the students commuting by bus, graduates of state schools are usually living on rent. Furthermore, these students mostly have married parents.

## Real Life Problems

1) **How do students from different accommodation types, with varying transportation methods, and working status manage their attendance and GPA?**

2) **How does the high school graduation type paired with the parent's occupation affect the transportation to the university and participation in sports & activities?**

3) **Do students who come from divorced or single-parent families, receive higher scholarships, and have part-time jobs face challenges in weekly study hours and GPA?**

4) **How do parents' education levels, scholarship type, and high school type together influence a student's GPA?**

**Generated Rules:**

1. Students commuting by bus and not doing any additional work have good attendance in classes
2. Students living on rent and not working commute by bus
3. Students who graduated from state school and do not take part in activities usually have mothers who are housewives
4. Students who do not take part in activities and have housewife mothers, commute by bus
5. Students with 50% scholarship and married parents do not do any additional work
6. Students not working and studying fewer hours have married parents
7. Students with married parents do not have additional work
8. Students receiving 50% to 75% scholarship are graduates of state school
9. Students whose fathers graduated from high school went to state school
10. Students whose mothers graduated from primary school went to state school