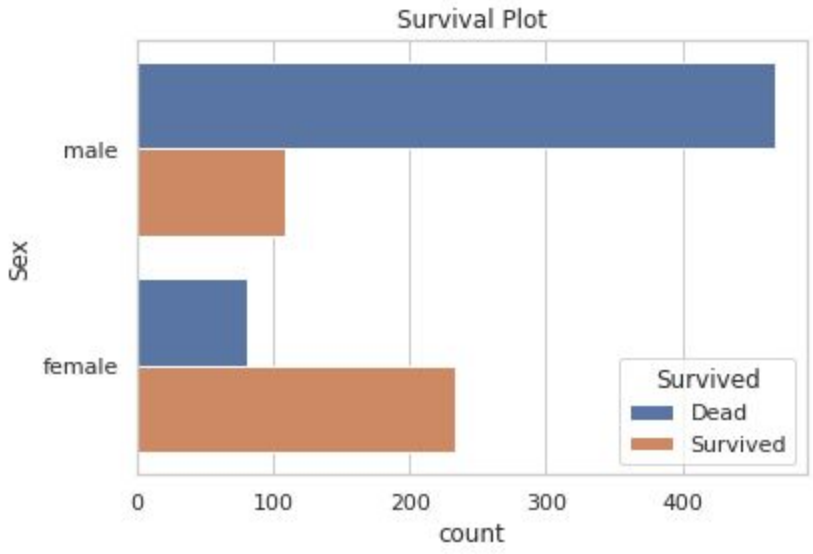
Harisaipravin
17PW13

# Titanic Survival Analysis

Data Mining Lab - Assignment 2

# Association Rule Mining




## Association Rule Mining:-

Association Rule Mining is one of the ways to find patterns in data. It finds:

- features (dimensions) which occur “together”
- features (dimensions) which are “correlated”

These patterns define interesting relationships and interactions between different items.

## Objectives:-

The problem tries to find the association between gender, age and the survival rate. The sinking of the Titanic is one of the most infamous shipwrecks in history. At the time of evacuation, people rushed to save their precious life at all cost. In spite of the support and herculean effort from the crew members and volunteers, thousands of people lost their lives and till this day, it remains as black memory.

To add to this, people were loaded into the safety boat based on their class, gender and age. Thus leading to a lot of ruckus and tension within the crowd. Also there were lots of biasing and few even opted to give their life to save others. This study aims at finding any correlation between the population that survived and died.

This aims to answer the question :-

> "what sorts of people were more likely to survive and die?"

## Data set description:-

Dataset consists of a detailed list of every passenger on board during the lone journey of titanic. The dataset is downloaded from kaggle and has a total of 891 rows indicating the passengers in the fateful journey.

Further the dataset consists of :-

1. PassengerID
2. Survived
3. Pclass
4. Name
5. Sex
6. Age
7. SibSp
8. Parch
9. Ticket
10. Fare
11. Cabin
12. Embarked

## Pre-processing:-

1. Unfilled age indexes are filled with 0.
2. Survived is converted from boolean to Ordinal (Dead or Survived).
3. Age is converted from nominal form to Ordinal form( Child..Adult). This project aims at finding the association between gender and age with survival rate.

```
titanic = pd.read_csv('train.csv')

titanic['Age'].fillna(0, inplace=True)

rep = {0: "Dead", 1: "Survived"}
titanic.replace({'Survived' : rep}, inplace=True)
```

```
def age_r(col, points, labels=None):
  minval = col.min()
  maxval = col.max()
  bp = [minval] + points + [maxval]
  if not labels:
    labels = range(len(points)+1)
  buffer = pd.cut(col,bins=bp,include_lowest=True)
  return buffer

pts = [1, 12, 21, 55 ]
labels = ["Unknown", "Child", "Teen", "Adult", "Old"]
titanic['Age'] = age_r(titanic['Age'], pts, labels)
print(titanic)
```

```
     PassengerId  Survived  Pclass  ...     Fare Cabin Embarked
0              1      Dead       3  ...   7.2500   NaN        S
1              2  Survived       1  ...  71.2833   C85        C
2              3  Survived       3  ...   7.9250   NaN        S
3              4  Survived       1  ...  53.1000  C123        S
4              5      Dead       3  ...   8.0500   NaN        S
..           ...       ...     ...  ...      ...   ...      ...
886          887      Dead       2  ...  13.0000   NaN        S
887          888  Survived       1  ...  30.0000   B42        S
888          889      Dead       3  ...  23.4500   NaN        S
889          890  Survived       1  ...  30.0000  C148        C
890          891      Dead       3  ...   7.7500   NaN        Q

[891 rows x 12 columns]
```

## Rule mining process:

```
in_titanic = titanic[['Pclass','Age', 'Survived', 'Sex']]
in_titanic.head()
```

| | Pclass | Age | Survived | Sex |
|---|---|---|---|---|
| 0 | 3 | Adult | Dead | male |
| 1 | 1 | Adult | Survived | female |
| 2 | 3 | Adult | Survived | female |
| 3 | 1 | Adult | Survived | female |
| 4 | 3 | Adult | Dead | male |

Our data shortlists the set to Pclass, Age, Sex and survival rate. The above diagram shows the shortened data for further processing.

**The parameters considered for this rule mining process are:-**

1. Support
2. Confidence
3. Lift

**Algorithm:-**

Apriori is an algorithm for frequent itemset mining and association rule learning over relational databases.

**Time Required:-**

Time required for apriori may be slow, but given the small dataset. It can be assured a quick runtime. Apriori is often considered a best and efficient approach for small datasets.

## Resulting Rules:-

```
output = apriori(df, min_support=0.2, use_colnames=oht.columns_)

print ('Support of 0.2')
print(association_rules(output, 'support' , min_threshold=0.5))
print ('Confidence of 0.8')
print(association_rules(output, 'confidence' , min_threshold=0.8))
```

```
Support of 0.5
  antecedents consequents  antecedent support  ...      lift  leverage  conviction
0      (male)      (Dead)            0.647191  ...  1.316752  0.126224    2.030636
1      (Dead)      (male)            0.615730  ...  1.316752  0.126224    2.386905

[2 rows x 9 columns]
Confidence of 0.8
     antecedents consequents  ...  leverage  conviction
0         (male)      (Dead)  ...  0.126224    2.030636
1         (Dead)      (male)  ...  0.126224    2.386905
2      (3, male)      (Dead)  ...  0.096581    2.828879
3      (3, Dead)      (male)  ...  0.066171    1.817946
4  (Adult, male)      (Dead)  ...  0.065656    2.007478
5  (Adult, Dead)      (male)  ...  0.068762    2.569173

[6 rows x 9 columns]
```

With tons of testing and repeated runs, It is found that at min_support of 0.2, male and dead have high support with a confidence of more than 0.8.

Hence men are more likely to die than women during shipwrecks or in general during crises. Although it is widely believed that men are more stronger beings in terms of physical strength, this fact seems to contradict that. One possible explanation could be that men tried to help women at first, since they can't handle the extreme situations. But ended up without any support quickly.

**Other factors found :-**

1. Females of PcClass 1 have 96.85 survival confidence.
2. Adults of PcClass 2 have 90.2 survival confidence.