Knowledge Discovery in Databases (KDD): An In-Depth Analysis Using the Titanic Dataset

Pallavi Vangari

Abstract

Knowledge Discovery in Databases (KDD) is a comprehensive process that encompasses various stages to extract valuable information from vast amounts of data. Using the Titanic dataset as a case study, this paper provides a detailed walkthrough of each KDD stage, demonstrating the importance and intricacies involved in converting raw data into actionable insights.

1 Introduction

With the explosive growth of data in recent years, it's essential to have systematic processes to extract valuable information. Knowledge Discovery in Databases (KDD) is such a process, and this paper aims to elucidate its various stages using a hands-on approach with the Titanic dataset.

2 The KDD Process

The KDD process can be broadly divided into the following stages:

1. Data Cleaning
2. Data Integration
3. Data Selection
4. Data Transformation
5. Data Mining
6. Pattern Evaluation and Knowledge Presentation
7. Knowledge Consolidation

3 Case Study: The Titanic Dataset

The Titanic dataset contains passenger information and their survival status when the Titanic ship tragically sank. It offers a mix of categorical and numerical features, making it an ideal candidate for this demonstration.

3.1 Data Cleaning

Objective: Handle missing values, noise, and outliers to prepare a clean dataset.
Implementation: Upon inspection, we identified missing values in the Age, Fare, and Cabin columns. The following strategies were adopted:

- The Cabin column was transformed into a binary indicator, Cabin_Known, showing if a cabin was known or not.
- Median values were used to impute missing entries in the Age and Fare columns.

3.2 Data Integration

Objective: Combine data from multiple sources.
Implementation: Since we only had a single dataset, data integration was not required in this case.

3.3 Data Selection and Transformation

Objective: Choose relevant data for analysis and convert it into a suitable format.
Implementation:

- We dropped columns like PassengerId, Name, and Ticket which were deemed unnecessary for the predictive model.
- Categorical features like Sex and Embarked were transformed into numerical format using one-hot encoding.

3.4 Data Mining

Objective: Apply algorithms to extract patterns.
Implementation: The primary goal was to predict survival based on passenger attributes. We employed a logistic regression model, which is a commonly used algorithm for binary classification tasks.

3.5 Pattern Evaluation and Knowledge Presentation

Objective: Interpret and visualize the discovered patterns.
Implementation: The logistic regression model achieved a surprisingly perfect accuracy of 100% on the test set. While the result might be seen as a success, such high accuracy warrants caution. It suggests potential overfitting, data leakage, or the dataset being overly simplistic.

3.6 Knowledge Consolidation

Objective: Utilize the discovered knowledge for decision-making or further processes.
Implementation: The insights from the dataset suggest that certain features like Pclass, Age, Sex, and Embarked location could play a pivotal role in determining survival. This knowledge can be used for various applications, such as developing survival prediction systems or historical analysis.

4 Conclusion

KDD is a holistic process, encompassing various stages that convert raw data into valuable insights. Using the Titanic dataset, we demonstrated each of these stages in detail. While our model achieved perfect accuracy, it is essential to approach such results with caution. The KDD process, in real-world scenarios, is iterative, often requiring revisiting previous stages based on outcomes.

References

[1] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37.

[2] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.