# Business Bankruptcy Prediction Using the KDD Methodology

## Abstract

This paper explores the use of the Knowledge Discovery in Databases (KDD) methodology for predicting business bankruptcy based on financial data. We systematically applied the KDD approach through data selection, preprocessing, transformation, and modeling, using Random Forest and SMOTE for class balancing. The results show the potential of data-driven bankruptcy prediction for informed financial decision-making.

## Introduction

Predicting business bankruptcy is crucial in today's economy, helping stakeholders assess financial stability and mitigate risks. This study employs the KDD methodology, a structured approach to data mining, to assess the bankruptcy risk of companies based on financial data. The main goal is to predict bankruptcy likelihood by analyzing financial metrics and using machine learning models.

## Methodology

The project follows the KDD process, which includes:

- Data Selection: Merging financial data with bankruptcy status labels.
- Data Preprocessing: Handling missing values and transforming non-numeric data.
- Data Transformation: Scaling features and addressing class imbalance with SMOTE.
- Data Mining: Training a Random Forest model for bankruptcy prediction.
- Interpretation and Evaluation: Evaluating model accuracy and predictive performance.

### Data Selection and Preprocessing

The data consists of financial metrics and a label indicating bankruptcy status. The initial steps include merging datasets and replacing missing or incorrect values (e.g., '?' with NaN).

Sample Code:

```
data.replace('?', np.nan, inplace=True)
data = data.apply(pd.to_numeric, errors='coerce').dropna()
```

### Data Transformation

Scaling is applied to the features for model consistency. Additionally, SMOTE is used to handle class imbalance by oversampling the minority class (bankrupt companies).

Sample Code:

```
from imblearn.over_sampling import SMOTE
smote = SMOTE(random_state=42)
X_resampled, y_resampled = smote.fit_resample(X, y)
```

### Data Mining - Model Training

The model chosen is a Random Forest classifier due to its robustness with structured data. The dataset is split for training and validation, and the model's performance is evaluated on the validation set.

Sample Code:

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
```

```
X_train, X_val, y_train, y_val = train_test_split(X_resampled, y_resampled, test_size=0.2, random_state=42)
model = RandomForestClassifier(random_state=42)
model.fit(X_train, y_train)
y_pred = model.predict(X_val)
print(classification_report(y_val, y_pred))
```

### Evaluation

The model achieved high accuracy, effectively distinguishing between bankrupt and non-bankrupt companies. The precision, recall, and F1-score were evaluated to provide a complete picture of the model’s effectiveness. Further tuning or experimenting with additional models could enhance performance.

## Results and Analysis

The Random Forest model's predictions were evaluated on the validation set, yielding a classification report with precision, recall, and F1-score metrics. The model's performance indicates promising results in predicting business bankruptcy, with balanced class representation due to SMOTE. The class distribution after applying SMOTE, as well as the confusion matrix from validation predictions, highlight the effectiveness of the model in predicting bankruptcy risk.

## Conclusion

This study demonstrates the effectiveness of the KDD methodology in predicting bankruptcy, providing a structured framework for data preprocessing, transformation, and modeling. The predictive insights generated from financial data are valuable for businesses, investors, and policymakers, enabling informed financial decisions. Future work could explore additional machine learning models or tuning techniques to further enhance prediction accuracy.

## References

1. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37.
2. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.
3. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.