Text Mining Analysis of Movie Reviews

BDA 622 Marketing Analytics

Jairo Onate | MUID 11062465 | 12/12/2024

PROJECT OVERVIEW

Objective: Utilize text mining and natural language processing (NLP) techniques to analyze movie review data.

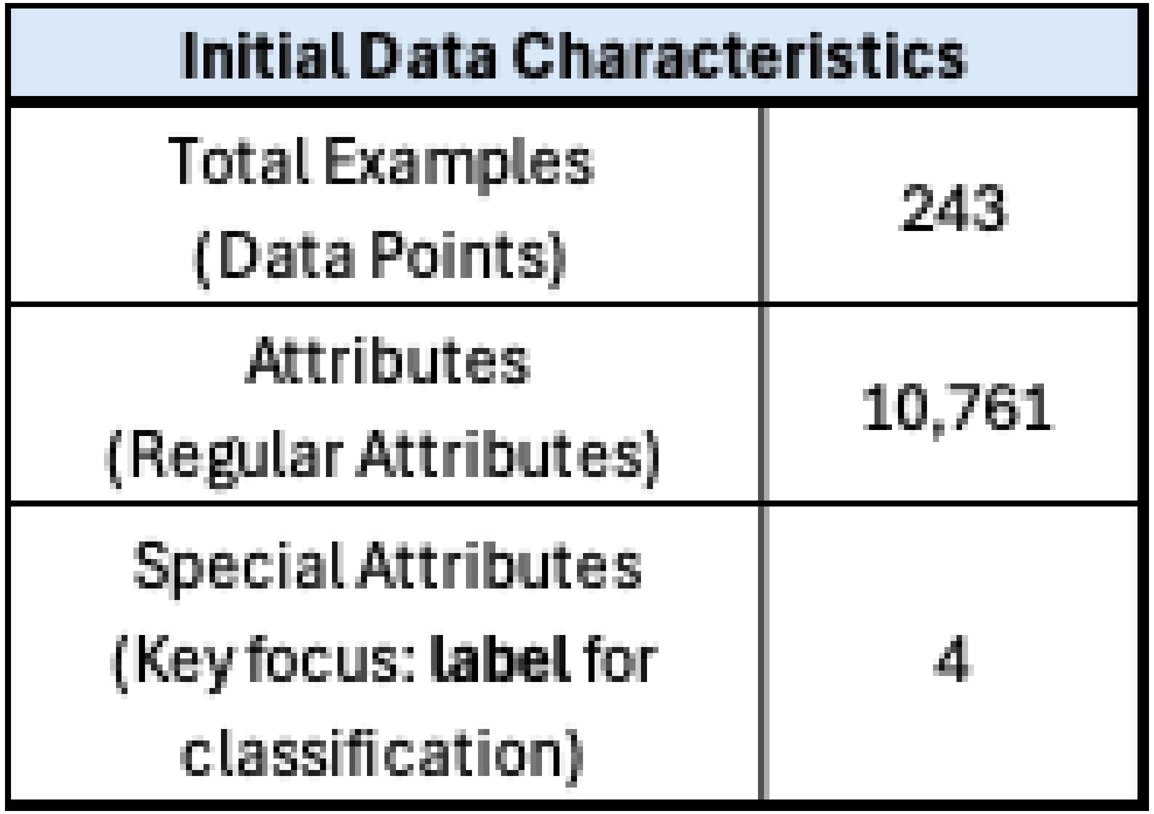
Tool Used: RapidMiner

Key Goal: Automatically classify reviews (positive or negative) and provide insights to inform marketing strategies.

Business Impact: Understand audience sentiment to refine marketing strategies and improve customer satisfaction.

DATASET SUMMARY

- **Data Source:** Text files divided into positive and negative labels.
- **Objective:** Predict sentiment (positive or negative) for incoming reviews.

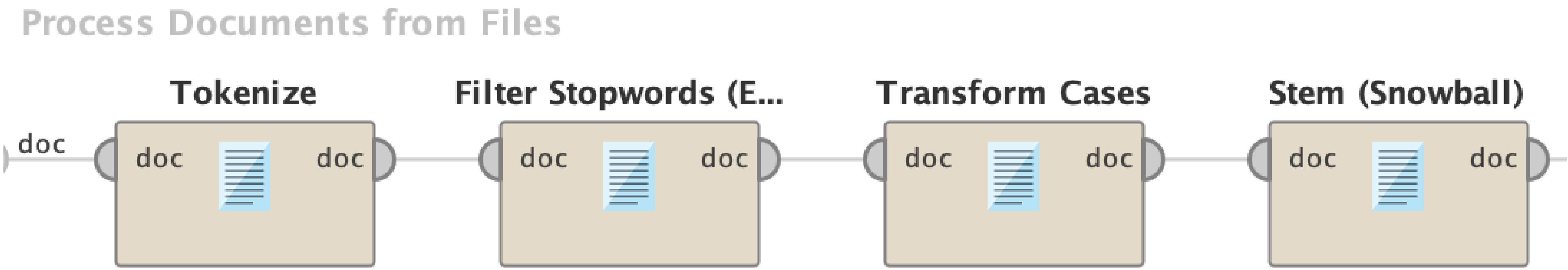


Initial Data Characteristics	
Total Examples (Data Points)	243
Attributes (Regular Attributes)	10,761
Special Attributes (Key focus: **label** for classification)	4

TEXT PROCESSING WORKFLOW

- **Steps Performed:**

1. **Tokenization:** Converted text documents into data points.
2. **TF-IDF Weighting:** Assigned importance to words in the text.
3. **Preprocessing:** Lowercased and stemmed words to focus on root terms.
4. **Pruning:** Applied thresholds (3% minimum, 30% maximum) to reduce dimensionality, keeping 1,864 attributes.

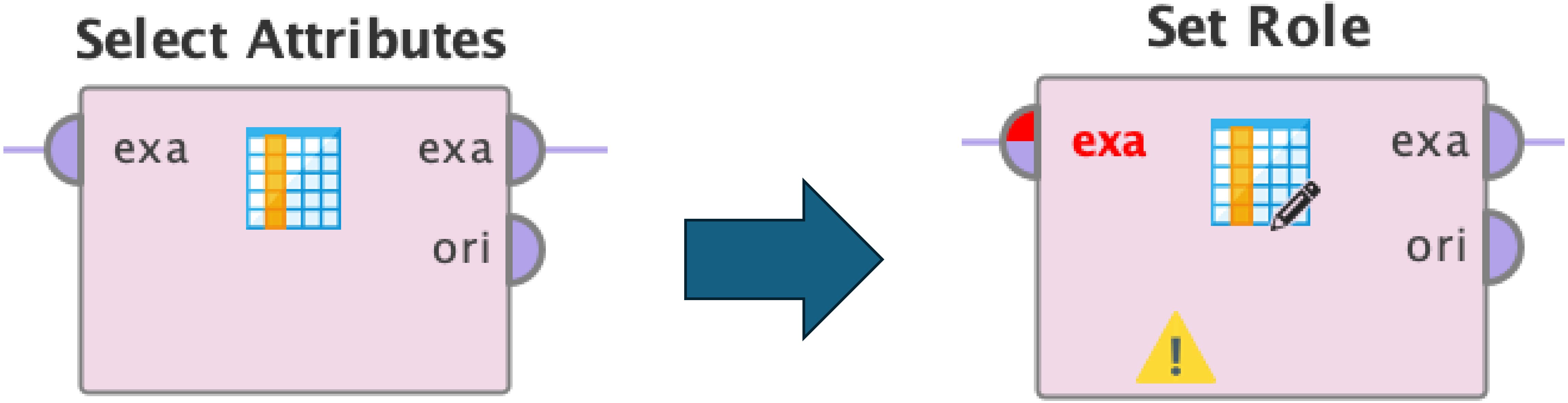


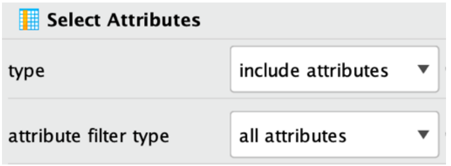
DATA PARTITIONING

- **Process Steps:**

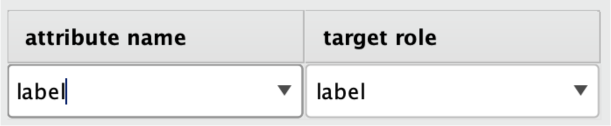
1. **Select Attributes:** Ensured algorithm understands attribute types.
2. **Set Role Operator:** Defined the label attribute as the target for classification.

Goal: Structure data to prepare for model building.




Select Attributes

type	include attributes
attribute filter type	all attributes

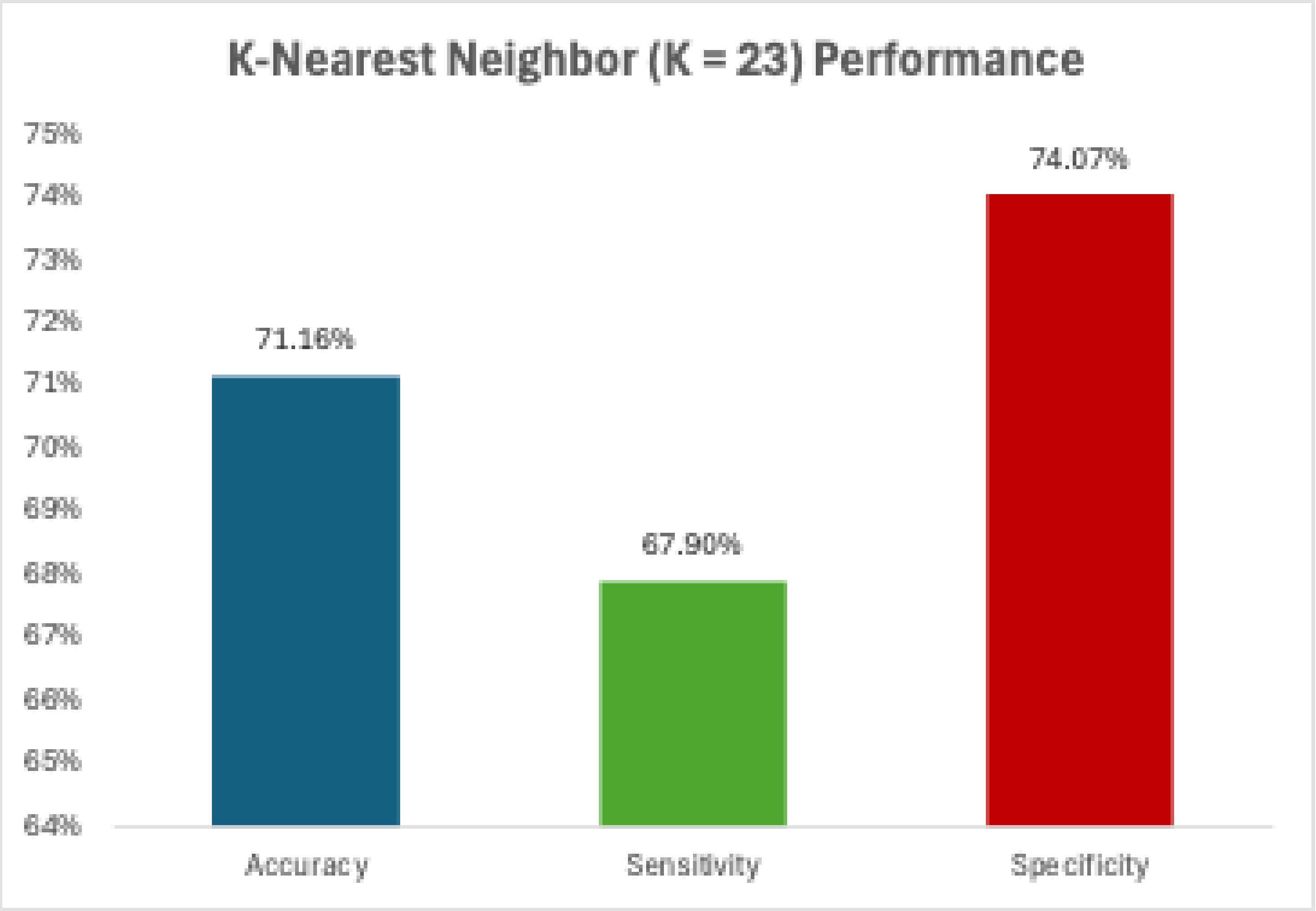


attribute name	target role
label	label

MODEL CONSTRUCTION

- **Key Parameters:**
- **Cross-Validation:** Used 10 folds with stratified sampling.
- **Models Tested:**
  - K-Nearest Neighbors (KNN)
  - Random Forest
  - Decision Tree

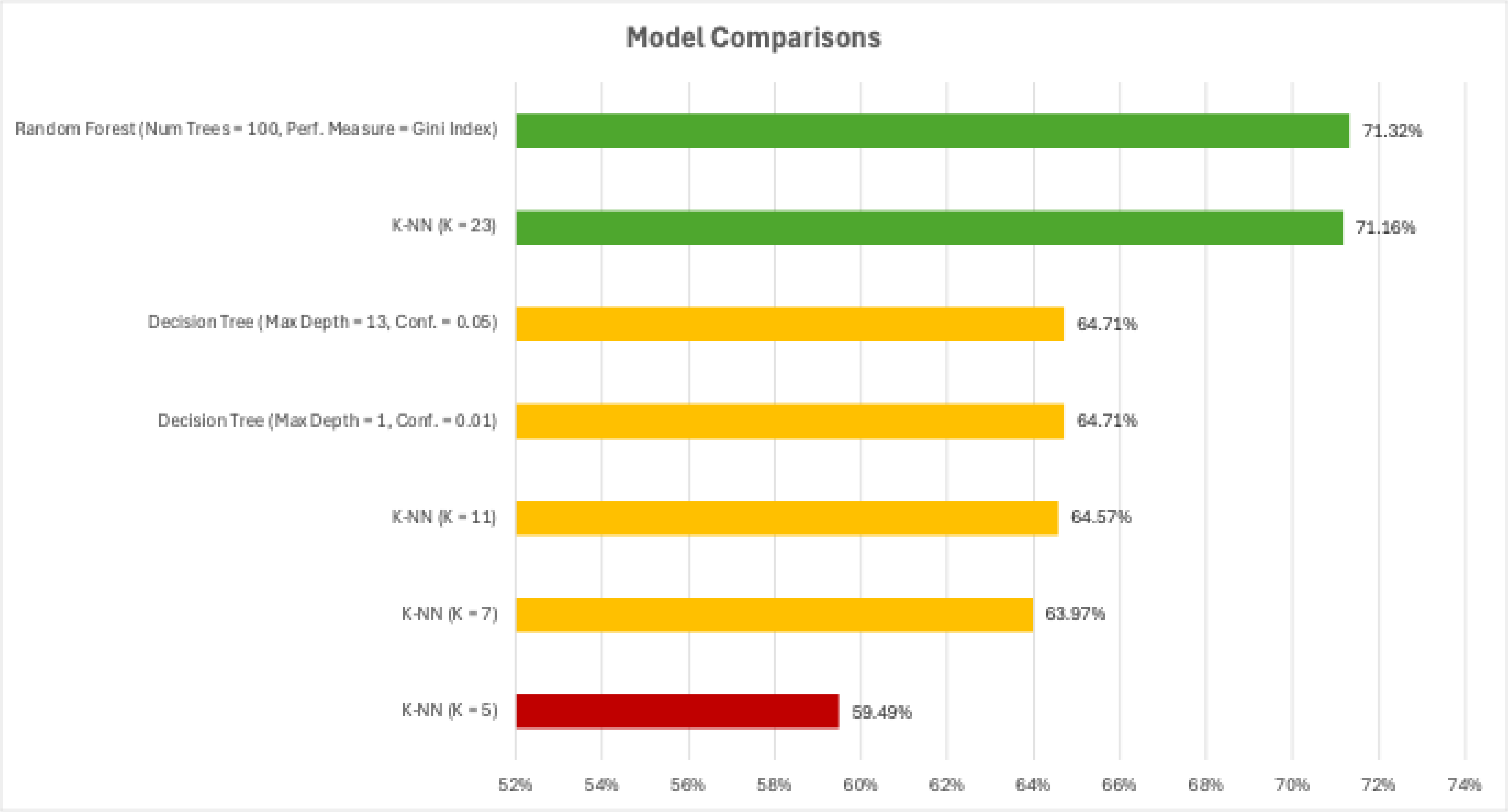
  Best Model: KNN with K=23.
- **Distance Metric:** Cosine similarity
- **Performance Metric Across Models:** Accuracy
- **Key Insight:** Effective classification of audience sentiment for informed decision-making.



BUSINESS IMPLICATIONS

- **Key Benefits for Studios:**
- **Positive Sentiment Analysis:** Highlight popular elements (e.g., performances, storylines) in promotions.
- **Negative Sentiment Insights:** Address criticisms to improve future productions.
- **Time Savings:** Automate review analysis to focus on strategic marketing.
- **Outcome:** Develop data-driven strategies to enhance audience engagement and satisfaction.

MODEL PERFORMANCE VALIDATIONS



CONCLUSION

- **Project Summary:** Demonstrated effective use of text mining for sentiment classification.
- **Best Model:** KNN (K=23) with accuracy of 71.16%.
- **Marketing Impact:** Provides actionable insights to improve customer engagement and refine marketing strategies.
- **Future Applications:** Expand analysis to other industries or additional datasets to enhance generalizability.

THANK YOU