Assessment Report

on

"Classify Book Genre"

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

CSE-AI

By

PRATEEK RAI

202401100300179

(CSE-AI (C) (32))

Under the supervision of

Sir Mayank Lakhotia

KIET Group of Institutions, Ghaziabad

22 April, 2025

Classify Book Genre

Introduction :

In the digital era, the volume of books being published and accessed through online platforms has grown exponentially. With this explosion of content, manually categorizing and recommending books has become increasingly impractical. Genre classification is vital for digital libraries, e-commerce platforms, and recommendation systems, as it significantly enhances user experience by enabling efficient browsing, searching, and personalized recommendations.

Traditionally, book genres are assigned manually by editors or librarians, which can be subjective and inconsistent. Automating this process using machine learning allows for a scalable, data-driven, and more consistent approach. This project leverages book metadata—specifically author popularity, book length, and the number of keywords—as features to build a model that predicts the genre of a book.

The implementation of such a classifier could help publishers in categorizing large catalogs, aid booksellers in organizing inventories, and enhance the recommendation engines for readers. It can also be applied in educational contexts for tagging academic resources. The versatility and simplicity of using metadata without the need for full-text analysis make this approach both computationally efficient and practical.

This project utilizes a Decision Tree classifier to train on a labeled dataset of books and evaluate its performance using multiple classification metrics, offering insights into the feasibility of genre classification through metadata alone.

Problem Statement :

To build a machine learning model that classifies books into predefined genres (such as mystery, fantasy, fiction, and non-fiction) using metadata. This can streamline cataloging and enhance user experience in book recommendation systems.

Objectives

- Preprocess the dataset for training a machine learning model.
- Train a Decision Tree classifier to predict book genres.
- Evaluate model performance using standard classification metrics.
- Visualize the confusion matrix using a heatmap.

Methodology

Data Collection: A CSV dataset containing metadata for books including author popularity, book length, and number of keywords is used.

Data Preprocessing:

- One-hot encoding of categorical variables (if any).
- Numerical features retained as-is or normalized if needed.

Model Building:

- The dataset is split into 70% training and 30% testing.
- A Decision Tree classifier from scikit-learn is trained on the dataset.

Model Evaluation:

- Metrics used: Accuracy, Precision, Recall, and F1-score.
- Confusion matrix generated and visualized using Seaborn heatmap.

Data Preprocessing

- The input CSV is loaded using pandas.
- Features (X) include author popularity, book length, and number of keywords.
- Target (y) is the genre.

- One-hot encoding is applied to any categorical inputs.
- Dataset is split into training and testing sets (70:30 split).

Model Implementation

The Decision Tree Classifier is chosen for its interpretability and simplicity. The model is trained using scikit-learn and evaluated on the test set.

Evaluation Metrics

- **Accuracy:** Measures overall correctness of the model.
- **Precision:** Proportion of true positive predictions among all positive predictions.
- **Recall:** Proportion of true positives correctly identified.
- **F1 Score:** Harmonic mean of precision and recall.
- **Confusion Matrix:** Visualized using a heatmap

Results and Analysis

- Sample test data and predictions were evaluated.
- The confusion matrix showed mixed classification performance across genres.
- Example Classification Report:

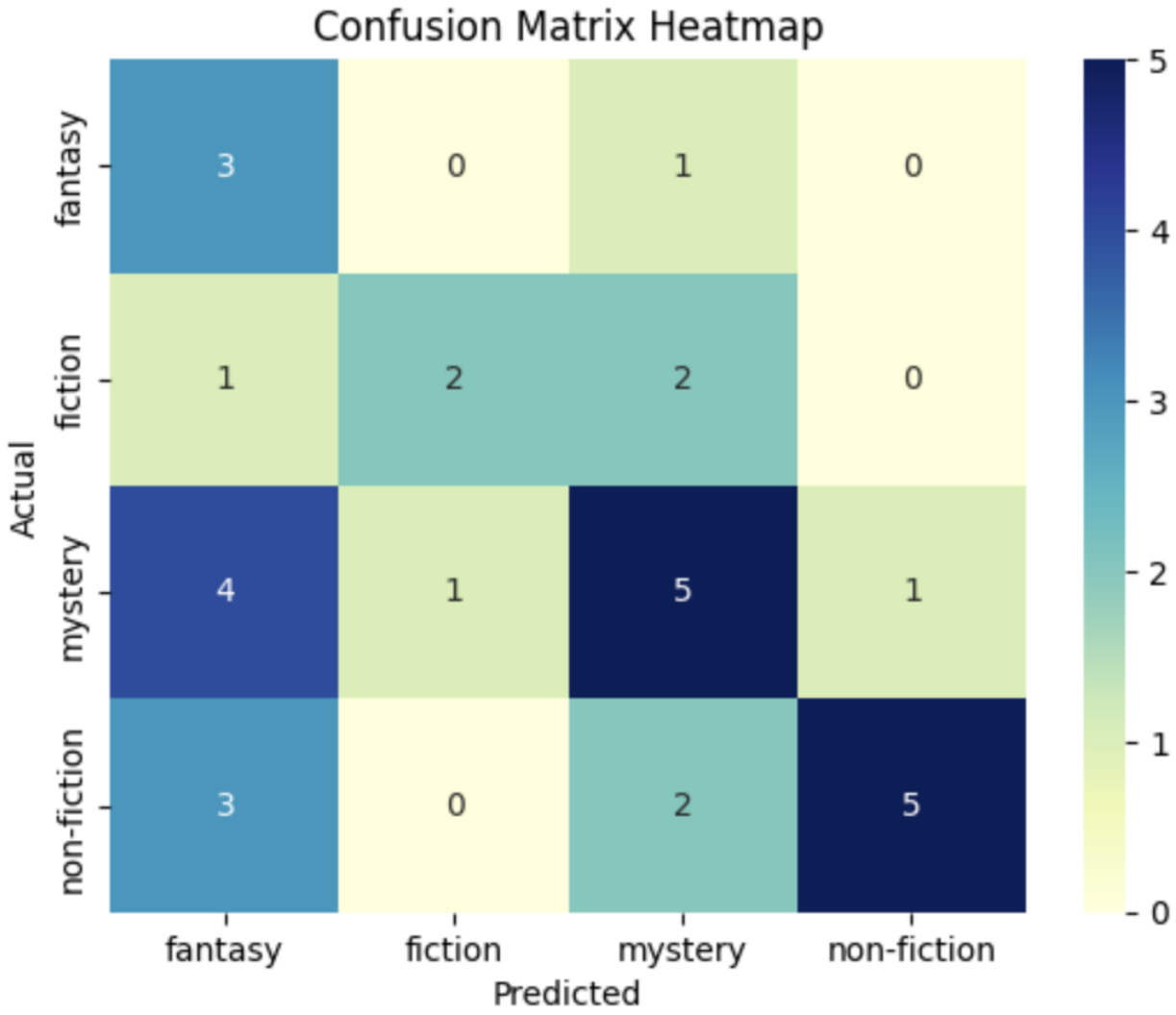
```
Sample data:
   author_popularity  book_length  num_keywords    genre
cell output actions
0          1.052297          776             5  mystery
1         48.950098          674             5  mystery
2          2.323401          633            19  fantasy
3         41.564184          169            12  mystery
4         65.129649          992            18  fantasy

Classification Report:
              precision    recall  f1-score   support

     fantasy       0.27      0.75      0.40         4
     fiction       0.67      0.40      0.50         5
     mystery       0.50      0.45      0.48        11
 non-fiction       0.83      0.50      0.62        10

    accuracy                           0.50        30
   macro avg       0.57      0.53      0.50        30
weighted avg       0.61      0.50      0.52        30
```

HEATMAP


Confusion Matrix Heatmap
Actual
fantasy
fiction
mystery
non-fiction
3
0
1
0
1
2
2
0
4
1
5
1
3
0
2
5
fantasy
fiction
mystery
non-fiction
Predicted
5
4
3
2
1
0

Conclusion

This project demonstrates a machine learning approach to classify book genres based on simple metadata features. While the Decision Tree model provided reasonable accuracy, performance can be improved using ensemble methods, better feature engineering, or deep learning approaches. Nevertheless, the system offers a foundation for automated book classification systems.

References

- scikit-learn documentation
- pandas documentation
- Seaborn visualization library
- Research articles on book classification and content-based recommendation systems