K.I.E.T. Group of Institutions

Ghaziabad

Name: Arnav Bhardwaj

Branch: CSE(AI) - A

Roll No: 62

Date: 22/04/2025

Project Report On -. Classify Book Genres

Introduction

In the modern literary landscape, the vast diversity of books makes genre classification an essential yet complex task. With the rise of digital libraries and recommendation systems, accurately categorizing books into genres helps improve discoverability and enhances user experience. Traditionally, genre classification has relied on manual curation or direct analysis of book content. However, metadata—such as author popularity, book length, and keyword frequency—offers a scalable and efficient alternative for automated genre prediction.

This project explores the use of book metadata to classify books into their respective genres using machine learning techniques. By transforming numerical and categorical metadata into textual features, and applying vectorization and classification algorithms, the system aims to predict a book's genre with minimal human intervention. The model is trained and evaluated using a real-world dataset that includes fields like author popularity, book length, number of keywords, and predefined genres.

This approach demonstrates how machine learning can effectively leverage structured metadata for genre classification, offering a lightweight yet powerful solution for publishers, retailers, and readers alike.

Methodology

Data Collection

The dataset includes book metadata: author popularity, book length, number of keywords, and genre labels.

Data Processing

Missing values were removed. Metadata fields were combined into a single text feature and vectorized using TF-IDF. Genres were label-encoded for model training.

Modeling

A Random Forest Classifier was trained to predict genres based on the processed features. Model performance was evaluated using classification metrics and a confusion matrix.

Visualization

Basic visualizations were created, including histograms, bar graphs, and scatter plots to understand data distribution and relationships.

Tools Used

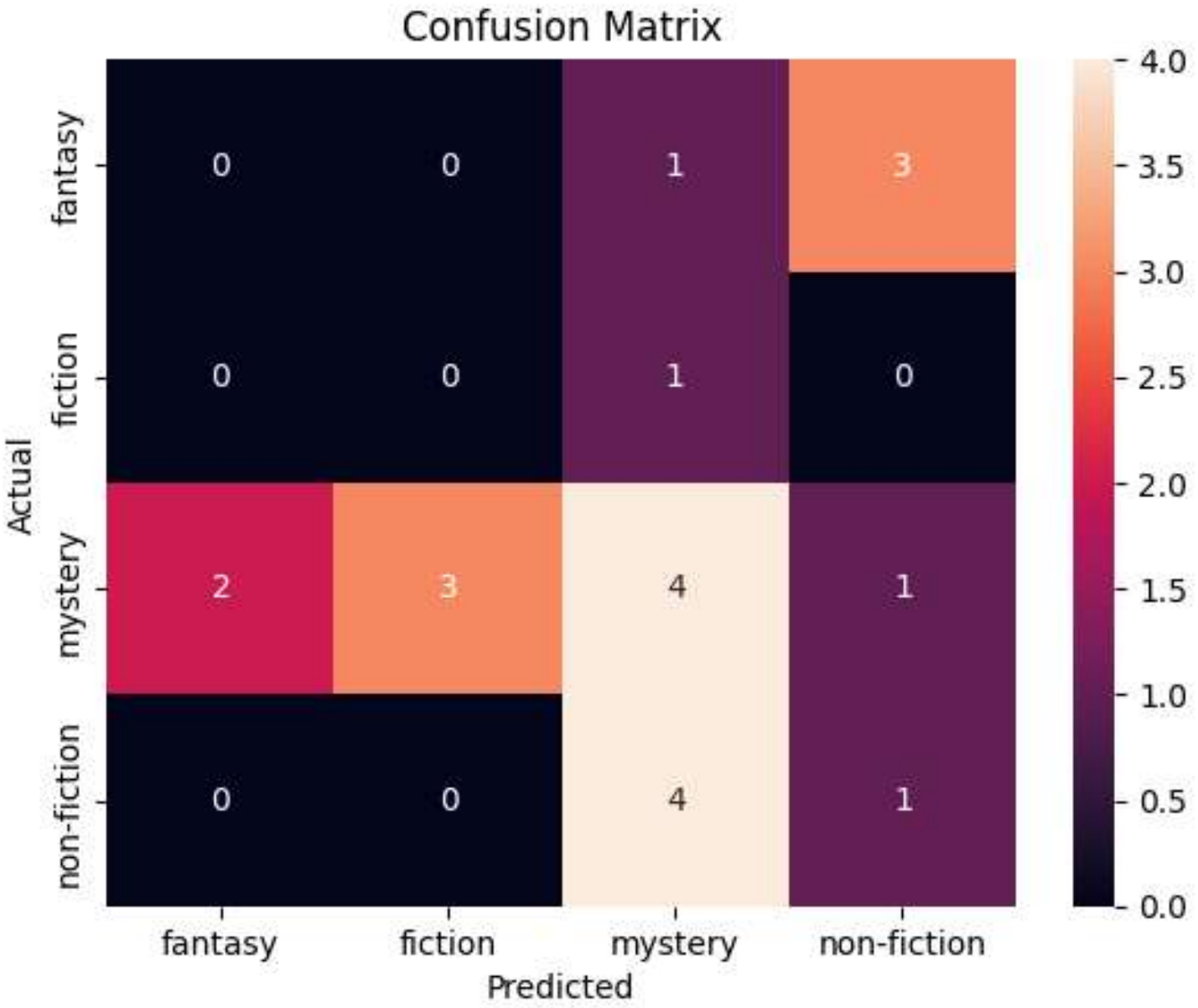
Python, Pandas, Scikit-learn, Matplotlib, and Seaborn.

Code

```
!pip install pandas scikit-learn seaborn matplotlib

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import
TfidfVectorizer
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report,
confusion_matrix

from google.colab import files

df = pd.read_csv("/content/book_genres.csv")
df.head()

df.info()
df.isnull().sum()

df.dropna(inplace=True)
df.rename(columns={"AuthorName": "author",
"PageCount": "length", "Tags": "keywords", "Genre":
"genre"}, inplace=True)

df.head()

print(df.columns)

le = LabelEncoder()
df['genre_encoded'] = le.fit_transform(df['genre'])
df[['genre', 'genre_encoded']].drop_duplicates()

df['text_data'] = df['author_popularity'].astype(str) + ' ' +
df['book_length'].astype(str) + ' ' +
df['num_keywords'].astype(str)
df[['text_data']].head()
```

```
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(df['text_data'])

y = df['genre_encoded']

X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2,
random_state=42)

model = RandomForestClassifier()
model.fit(X_train, y_train)
```

```
y_pred = model.predict(X_test)

print("Classification Report:\n",
classification_report(y_test, y_pred,
target_names=le.classes_))

# Confusion matrix heatmap
conf_matrix = confusion_matrix(y_test, y_pred)
sns.heatmap(conf_matrix, annot=True, fmt='d',
xticklabels=le.classes_, yticklabels=le.classes_)
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix")
plt.show()
```



Output/Result

Observation from the analysis is given Above:

. References/Credits:

- Dataset: [source: Kaggle]
- Libraries: Pandas, Matplotlib, Seaborn