# Book Recommendation by Analysing Library User Behaviour Using Ensemble Method

Aleena Ann Shaji
Department of Computer Applications
Amal Jyothi College of Engineering
Kanjirappally, India
aleenaannshaji2024a@mca.ajce.in

Shelly Shiju George
Department of Computer Applications
Amal Jyothi College of Engineering
Kanjirappally, India
shellyshijugeorge@amaljyothi.ac.in

***Abstract*— This study aims to enhance the library user experience by analyzing borrowing history data. By investigating borrowing patterns and book preferences, the research seeks to uncover underlying connections among borrowed books. The primary goal is to utilize this information to provide personalized book recommendations, thereby improving user satisfaction with the library experience. To meet this goal, the research utilizes two machine learning methodologies: Decision Tree and Random Forest algorithms. Initially, the Decision Tree algorithm predicts book preferences based on historical borrowing data. Subsequently, the output from the Decision Tree model is used as input to the Random Forest algorithm for further refinement. This combined approach is designed to create tailored and enjoyable library experiences for users through personalized book suggestions.**

***Keywords*— *library user experience, borrowing history analysis, book preferences, personalized recommendations, Decision Tree, Random Forest.***

## I. INTRODUCTION

In today's digital era, libraries have access to abundant data on how their patrons borrow books. This data serves as a valuable resource, providing deep insights into reader's preferences and interests. By analyzing this information closely, libraries can gain valuable insights into their user's likes and dislikes, ultimately enhancing the library experience.

Behavior analysis emerges as a crucial tool in this context, allowing libraries to identify patterns in borrowing behavior and tailor their services accordingly. This involves studying how people make choices and act when borrowing books, enabling libraries to identify trends and improve their offerings.

Machine learning algorithms like Decision Trees and Random Forests play a key role in understanding this data. While these terms may seem complex, they are actually straightforward concepts. Decision Trees help create a profile of the types of books users prefer based on their borrowing history, resembling a tree with branches representing different user choices.

Random Forests build on this concept by refining predictions from Decision Trees to enhance accuracy. Visualize it as a vast forest of trees, each offering unique perspectives on user reading preferences.

By leveraging these algorithms, libraries can offer tailored recommendations, suggesting books likely to appeal to individual users based on their borrowing patterns. It's akin to having a knowledgeable librarian who understands your preferences and can recommend the perfect book. This not only elevates the library experience but also encourages patrons to explore a diverse range of reading materials.

## II. LITERATURE REVIEW

Zhang, et al. [1] analyze library user behavior using the Apriori optimization algorithm, revealing correlations among borrowed books and advocating for tailored book recommendations. While effective, the Apriori algorithm may face scalability challenges with larger datasets. They also emphasize the importance of a well-organized library collection layout and propose an optimization algorithm merging Apriori and k-means clustering for enhanced analysis and insights into library management.

Khamaj, et al. [2] present an innovative approach employing Reinforcement Learning (RL) and Deep Q Network (DQN) to develop a responsive and tailored user interface, adapting in real-time based on user actions. Traditional interfaces lack personalization, so this study aims to modify interfaces based on individual preferences. By combining RL and DQN, the approach adapts interfaces incrementally, balancing exploration of new interactions with exploiting known high-reward actions. Timestamped insights provide a nuanced understanding of user behavior, enabling timely modifications. While promising, challenges like computing complexity and privacy concerns must be addressed for broader implementation. Future research may focus on optimizing resource allocation and addressing privacy issues.

Liang, et al. [3] explore cross-device behavior in library OPAC usage. They analyze challenges from diverse user data across devices, using a large OPAC transaction log. The study identifies factors for predicting subsequent activities and devices, with initial device activity and time intervals being significant. Operating system features improve prediction accuracy. Despite limitations like small mobile data samples, the study highlights machine learning's potential to improve digital library services.

Ranjan, et al. [4] utilize behavioral analysis and machine learning techniques on logs from the application layer to identify malicious users. By intentionally creating vulnerabilities in an e-commerce web app hosted on AWS, they gather real-time data to identify potential attackers. Their approach enhances infrastructure security by monitoring browsing patterns. Achieving 65-70% accuracy with the Random Forest algorithm, they aim for 90% accuracy, reduced resource usage, and AI automation. They plan to extend analysis to network and transport layer data, but challenges in accuracy and resource utilization remain.

DOI: **10.5281/zenodo.11523782**
ISBN: 978-93-5906-046-0@2024, Dept. of Computer Applications, Amal Jyothi College of Engineering Kanjirappally, Kottayam

Zhao, et al. [5] tackle the challenge of understanding electricity usage patterns with the rise of smart meters, which generate vast amounts of data in power grids. They introduce an ensemble clustering approach to analyze these patterns, considering their complexity and uncertainty. Using Principle Component Analysis (PCA) to simplify the data, they apply single clustering and integrate results to classify users' consumption behavior into distinct modes. Their study, based on real data from 19 Chinese users, sheds light on weekly electricity usage and validates the method's effectiveness. They suggest future improvements, such as exploring different clustering methods, to enhance accuracy.

Cai, et al. [6] introduce a novel unsupervised approach for detecting shilling attacks, which relies on analyzing user rating behavior. Through scrutinizing variations in rating patterns, they pinpoint target items and intentions of malicious users, assembling a group of potentially suspicious users. By evaluating interest and rating preferences, they gauge the level of suspicion to uncover malicious users within this group. Experimental findings validate the efficiency of their method in detecting shilling attacks across diverse datasets. The method reduces computational complexity by focusing on a subset of suspicious users, enhancing efficiency. However, accuracy in identifying target items may impact model performance, warranting further research for improvement. Future work will explore hidden relationships between users and items to enhance attack detection further.

Luo, et al. [7] study how users interact on social networks, particularly focusing on why people repost content. They create a mathematical model to explain this behavior, considering factors like user preferences and how people decide to interact. Their model helps understand why some users repost more than others and how long interactions last. They test their model using real data from Weibo, a social media platform, and find it accurately describes user behavior. However, the model may only be useful for understanding behavior on social networks and might need more testing on other platforms to be fully trusted.

Park, et al [8] explore the utility of Windows Diagnostics, default in Windows 10 and 11, for digital forensics analysis. They scrutinize the recording of user activities such as USB device utilization, web browsing, and network connections within Windows Diagnostics. Subsequently, they devise DiagAnalyzer, a tool designed to automatically assess and represent this data. Their method offers insights into user behavior and its potential application in digital forensics investigations. However, limitations exist, such as variability in log data depending on user settings, highlighting the need for further research in this area.

## III. IMPLEMENTATION

### *A. Machine Learning*

Machine learning, a subset of artificial intelligence (AI), is dedicated to training computers to learn from data rather than relying on explicit programming. Through this approach, algorithms can autonomously improve their performance over time.

The primary aim of machine learning is to develop models capable of recognizing patterns and connections within data. By learning from real-world examples and past experiences, these models enhance their ability to make accurate predictions or decisions, even when encountering new or unfamiliar data.

Machine learning algorithms are versatile and can handle various types of data, including structured data such as tables and spreadsheets, as well as unstructured data like images, text, and audio. They are designed to comprehend the underlying structure of the data they analyze, enabling them to extract meaningful insights and provide valuable predictions or analyses.

### *B. Decision Tree Algorithm*

The Decision Tree algorithm is widely used in supervised learning, where it learns from labeled data to categorize things or predict numerical values. Here's a simplified explanation of how it works: Think of having a bunch of information about various items in a dataset. The Decision Tree begins by analyzing this data to identify which features or attributes are most crucial for making decisions. For instance, if you want to predict if someone will enjoy a particular book, features could be things like the book's genre, author, or length.

Then the algorithm starts splitting the data into smaller groups based on these features. It's like asking a series of questions to separate the data into subsets. For instance, it might start by asking if the book is fiction or non-fiction, then if it's by a specific author, and so forth.

This process forms a tree-shaped structure, where each branch represents a decision based on a feature, and each endpoint, or leaf, indicates the final prediction. So, in our book example, each leaf might indicate whether a person is likely to enjoy the book based on the considered features.

In essence, the Decision Tree algorithm organizes data into a tree structure, making decisions at each step based on different features until it makes a prediction.

### *C. Random Forest Algorithm*

The Random Forest algorithm is a smart technique used to improve prediction accuracy and prevent overfitting, where the model becomes too focused on the training data. It achieves this by combining multiple Decision Trees into a single large "forest."

Here's how it operates: Instead of relying on just one Decision Tree, Random Forest trains many of them. However, each tree is trained on a random subset of the data, not the entire dataset. This randomness ensures that each tree is slightly different. After training, all the trees collaborate to make predictions. Each tree provides its own prediction, and then the Random Forest averages these predictions to generate the final result. This averaging helps to smooth out any errors or peculiarities that individual trees might have.

In simpler terms, Random Forest is akin to having a team of diverse experts (the Decision Trees), each offering their own

DOI: **10.5281/zenodo.11523782**
ISBN: 978-93-5906-046-0@2024, Dept. of Computer Applications, Amal Jyothi College of Engineering Kanjirappally, Kottayam

viewpoint. By combining these insights, we obtain a more dependable and accurate prediction overall. This versatility makes Random Forest an effective tool for making predictions across different scenarios.

*D. Ensemble Method*

The Ensemble Method is similar to teamwork in machine learning, where multiple models join forces to enhance prediction accuracy.

Consider Random Forest as an illustration. Instead of depending solely on a single Decision Tree, Random Forest assembles a team of them. Each Decision Tree offers its own prediction, and by combining these diverse perspectives, the overall prediction becomes more accurate.

Ensemble techniques, such as Random Forest, capitalize on the variety among models to boost overall performance. By blending predictions from different models, we can generate more reliable and resilient predictions. It's comparable to gathering insights from a panel of experts, each contributing their unique viewpoint, to reach a more trustworthy conclusion.

*E. Recommendation*

Recommendation systems act as friendly helpers, utilizing machine learning techniques like Decision Tree and Random Forest to analyze users' borrowing history. From there, they suggest books tailored to each user's preferences and interests based on past borrowing behavior.

The objective is to enhance the library experience by providing personalized recommendations that align with individual reading preferences. By customizing suggestions to suit each user's tastes and past interactions, the aim is to boost satisfaction with the library service. It's akin to having a dedicated book advisor who anticipates what you'll enjoy reading, making your visits to the library more pleasurable and fulfilling.

## IV. METHOD OF IMPLEMENTATION

*A. Import necessary modules*

The necessary libraries are imported, including pandas for data handling, Streamlit for creating the user interface, and scikit-learn for machine learning functionalities.

```
import pandas as pd
import streamlit as st
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
```

Fig. 1

*B. Load and preprocess data*

The code loads three CSV files containing information about books, borrowers, and borrowing history.

```
books_df = pd.read_csv('books.csv')
borrowers_df = pd.read_csv('borrowers.csv')
history_df = pd.read_csv('historys.csv')
```

Fig.2

*C. Merge dataframe appropriately*

It merges the three datasets based on common columns to create a single dataframe containing information about borrowers, books, and their borrowing history.

```
merged_df = pd.merge(history_df, borrowers_df, on='borrower_id')
merged_df = pd.merge(merged_df, books_df, on='book_id')
```

Fig.3

*D. Train/ Test split*

The merged data is split into training and testing sets. Features (X) include borrower and book IDs, and the target variable (y) is the book category.

```
X = merged_df[['borrower_id', 'book_id']]
y = merged_df['category']

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

Fig.4

*E. Training Models*

Two models are trained - a Decision Tree Classifier and a Random Forest Classifier - using the training data.

```
decision_tree_model = DecisionTreeClassifier()
decision_tree_model.fit(X_train, y_train)
random_forest_model = RandomForestClassifier()
random_forest_model.fit(X_train, y_train)
```

Fig. 5

*F. Predict borrower's preference using Decision Tree and use it as input to RandomForest for recommendation*

This function takes a borrower ID as input and recommends books based on the borrower's historical borrowing data. It predicts the borrower's preference using the Decision Tree model and recommends books in the same category.

```
def recommend_books(borrower_id):
    borrower_history = merged_df[merged_df['borrower_id'] == borrower_id]
    borrower_name = borrower_history['fullname'].iloc[0]

predicted_preference = decision_tree_model.predict(borrower_history[['borrower_id', 'book_id']])
recommended_books = books_df[books_df['category'] == predicted_preference[0]]

  recommended_borrowers = borrower_history['borrower_id'].unique()
  return recommended_books, borrower_name, recommended_borrowers
```

Fig. 6

DOI: **10.5281/zenodo.11523782**
ISBN: 978-93-5906-046-0@2024, Dept. of Computer Applications, Amal Jyothi College of Engineering Kanjirappally, Kottayam

*G. Evaluate models accuracy*

The accuracy of both models is calculated using the test data and displayed in the UI.

```
decision_tree_accuracy = decision_tree_model.score(X_test, y_test)
random_forest_accuracy = random_forest_model.score(X_test, y_test)
st.write(f"Decision Tree Model Accuracy: {decision_tree_accuracy}")
st.write(f"Random Forest Model Accuracy: {random_forest_accuracy}")
```

Fig. 7

## V. RESULT

The result here shows the recommendation of the particular user based on their borrowed history.

The evaluation of models accuracy indicates a Decision Tree Model Accuracy of 0.5 and a Random Forest Model Accuracy of 0.166.

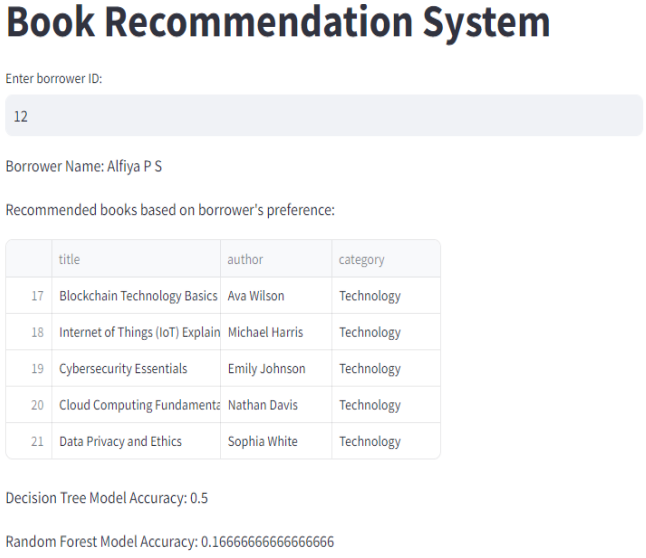


Fig. 8

## VI. CONCLUSION

This research utilizes machine learning techniques, specifically Decision Tree and Random Forest algorithms, to improve the library user experience through the examination of borrowing history data. By investigating borrowing patterns and book preferences, the research aims to provide personalized book recommendations, ultimately improving user satisfaction with the library experience. The Decision Tree algorithm predicts book preferences based on historical borrowing data, which is then refined using the Random Forest algorithm. The evaluation of the model's accuracy indicates a Decision Tree Model Accuracy of 0.5 and a Random Forest Model Accuracy of 0.166, suggesting room for improvement. Despite the modest accuracy, the combined approach demonstrates the potential to tailor library services to individual user preferences, paving the way for further research in optimizing recommendation systems for libraries.

## REFERENCES

[1] Zhang X., Zhang J., (2023), "Analysis and research on library user behavior based on apriori algorithm", Measurement: Sensors, Vol. 27, pp. 1-6.

[2] Khamaj A., Ali A. M., (2024), "Adapting user experience with reinforcement learning: Personalizing interfaces based on user behavior analysis in real-time", Alexandria Engineering Journal, Vol. 95, pp. 164-173

[3] Liang S., Wu D., (2019), "Predicting Academic Digital Library OPAC Users' Cross-device Transitions", Data and Information Management, Vol. 3, pp. 40-49

[4] Ranjan R., Kumar S. S., (2022), "User behaviour analysis using data analytics and machine learning to predict malicious user versus legitimate user", High-Confidence Computing, Vol.2, pp. 1-10

[5] Zhao Q., Li H., Wang X., Pu T., Wang J., (2019), "Analysis of user's electricity consumption behavior based on ensemble clustering", Global Energy Interconnection, Vol. 2, pp. 479-488

[6] Cai H., Zhang F., (2019), "Detecting shilling attacks in recommender systems based on analysis of user rating behavior", Knowledge-Based Systems, Vol. 177, pp. 22-43

[7] Luo G., Zhang Z., Diao S., (2022), "Empirical analysis and modelling social network user interaction behavior and time characteristics based on selection preference", Information Sciences, Vol. 608, pp. 1202-1220

[8] Park S., Lee S., (2022), "DiagAnalyzer: User behavior analysis and visualization using Windows Diagnostics logs", Forensic Science International: Digital Investigation, Vol. 43, pp. 1-7

[9] Hasan J., Horvat M., (2024), "An application of the Random Forest algorithm for the prediction of Solar Envelope 'Floor Space Index' based on spatiotemporal parameters", Journal of Building Engineering, Vol. 86, pp. 1-26

[10] Cai C., Yang C., Lu S., Gao G., Na J., (2023), "Human motion pattern recognition based on the fused random forest algorithm", Measurement, Vol. 222, pp. 1-12

DOI: **10.5281/zenodo.11523782**
ISBN: 978-93-5906-046-0@2024, Dept. of Computer Applications, Amal Jyothi College of Engineering Kanjirappally, Kottayam