# Library Data Analysis Using Machine Learning

Ms. Kirti Jagtap[1], Mrs. Yamini Koli[2], Ms. Kajal Kamble[3], Dr. Seema Chowhan[4]
[1,2]*Department of Computer Science, Baburaoji Gholap College, Pune*
[3,4]*Assistant Professor, Department of Computer Science, Baburaoji Gholap College, Pune*

***Abstract*—This work reveals insights from a library data analysis tool that employs several machine learning algorithms. The data samples that were analyzed in this work consisted of book, user, and rating data extracted from Kaggle, Baburaoji Gholap College Library, and self-generated data for the current research project. Several techniques of data preprocessing were applied, including handling missing data, data cleaning, and data transformation. The technique known as exploratory data analysis revealed some patterns concerning the distribution of books and the behavior of users. The machine learning algorithms used in this work included Decision Tree, ensemble machine learning algorithms like Random Forest, Gradient Boosting, and XGBoost, and K-Means Clustering for classifying data.**

***Index Terms*—Data Analytics, Machine Learning, Library Data Analysis, Clustering, Classification**

## I. INTRODUCTION

Data production in the library is enormous in volume due to a number of features such as books, users, and their ratings. The data is complicated; therefore, the processing of data becomes increasingly hard when its volumes grow bigger. Data analytics is helpful to analyse the data collected. Moreover, machine learning algorithms enable automatic pattern discovery and prediction. In this research, data gathered from various resources will be combined and pre-processed using techniques such as data cleaning, handling missing values, and transformations to enhance the quality of data. In addition to that, exploratory data analysis will be applied to identify the structure of data, including its book-related, user-related, and rating-related attributes.

In addition to these techniques, machine learning algorithms including Decision Tree and ensembles learning models such as Random Forest, Gradient Boosting, and XGBoost will be applied to make a classification of data. To group similar data points, k-means clustering technique will be utilized. Decision Tree algorithm will allow creating interpretable predictions whereas the ensembles technique will enhance performance.

Thus, the goal of the research is to process and analyse data generated by the library effectively.

## II. LITERATURE REVIEW

Various studies have investigated the application of analytical tools and machine learning algorithms in libraries to analyse data efficiently. For instance, Tamba et al. [1] used K-Means algorithm to classify library book data. Also, Mane [2] recognized the significance of big data analytics in improving library services, managing collections, and user satisfaction. It is evident from the research above that analytical tools can help with efficient library management.

Cluster analysis and classification algorithms are some of the machine learning algorithms used to analyze structured data. According to Witten et al. [3], there exist various practical machine learning algorithms that can be applied in data mining. On the other hand, Hastie et al. [4] described statistical algorithms for predicting the results from data. The decision tree algorithm is used for classification purposes due to its ease of implementation and interpretation.

K-Means clustering is one of the most popular algorithms used in unsupervised learning to cluster similar data points together. It works well in discovering hidden patterns and clustering data [5], [6]. Such methods can be used on library databases to cluster books or users based on their common attributes.

Preprocessing data and exploring data are key processes before using machine learning models. According to Raschka and Mirjalili [7], the necessity of data cleaning and transforming was stressed, whereas McKinney [8] focused on the use of Python

in analyzing data. Data preprocessing is vital in ensuring accurate data quality, which results in valid outputs. According to Pedregosa et al. [9], Scikit-learn is an effective library for implementing machine learning algorithms in Python.

Moreover, the literature review also gives some indications about how useful data can be in decision-making processes. Research done on library resources and services [10], along with libraries' function in education [11], suggests that the analysis of data from library services is crucial for enhancing their performance and user experience.

As a result, one can observe that machine learning models such as Decision Tree and K-Means clustering prove to be effective in analysing structured data sets. This paper utilizes these techniques when examining library-related data.

## III. DESCRIPTION OF DATASETS

There have been various datasets employed in this research for analysing data from the library through machine learning. The datasets involved information related to books, users, ratings, and library transactions. The datasets were gathered from sources including Kaggle, Baburaoji Gholap College Library, and a custom generated dataset.

### A. ChatGPT Dataset

ChatGPT Dataset is known to be a self-made dataset that was prepared for the purpose of analyzing library data. This dataset has 600 entries and includes 24 fields like Book ID, Title, Author, Publication Year, Publisher, ISBN, Category, Number of Pages, Availability, Cost in INR, Number of Issues, Format, and Rating. Most of the books in the dataset are under the category of Computer Science.

Table 1. Summary of library_dataset_600 Dataset

| Feature | Description |
|---|---|
| Total Records | 600 |
| Total Attributes | 24 |
| Main Category | Computer Science |
| Publication Year Range | 1980 – 2025 |
| Average Rating | 3.72 |
| Average Pages | 835 |
| Book Formats | Hardcover, Paperback, eBook |
| Availability Status | Available, Reserved, Issued, Maintenance |

### B. College Data Dataset

Data set for Research Data can be sourced from the library at Baburaoji Gholap College. This data set consists of transaction-related information such as issuing and returning of books. The variables in this data set include ID, Name, Type of Registration, Accession Number, Title of Book, Date of Issue, and Date of Return.

Table 2. Summary of Research Data Dataset

| Feature | Description |
|---|---|
| Total Records | 83 |
| Total Attributes | 7 |
| Valid Records | 70 |
| Missing Records | 13 |
| Unique Users | 42 |
| Unique Books | 64 |
| Register Types | 6 |
| Main Purpose | Transaction analysis |

### C. Books Dataset

Books dataset

This dataset is obtained from Kaggle. The dataset contains variables like ISBN, Book Title, Book Author, Year of Publication, and Publisher. The primary purpose of the dataset is to conduct exploratory data analysis.

Table 3. Summary of Books Dataset

| Feature | Description |
|---|---|
| Total Records | 3725 |
| Total Attributes | 8 |
| Unique Titles | 3616 |
| Unique Authors | 2474 |
| Unique Publishers | 905 |
| Most Frequent Author | Nora Roberts |
| Most Frequent Publisher | Ballantine Books |
| Purpose | Data analysis |

### D. Ratings Dataset

The Ratings data set is acquired from Kaggle and contains information on user ratings of books. Some of the attributes contained in this data set include User ID, ISBN, and Book-Rating.

Table 4. Summary of Ratings Dataset

| Feature | Description |
|---|---|
| Total Records | 1,149,766 |
| Total Attributes | 3 |
| Unique Users | 105,274 |
| Unique Books | 340,545 |
| Rating Range | 0 – 10 |
| Average Rating | 2.87 |
| Most Frequent ISBN | 0971880107 |
| Purpose | Data analysis |

E. Users Dataset

User's data is downloaded from Kaggle, which consists of information like User-ID, Location, and Age of users. It helps us to analyze users' distribution and demographics, among other things.

Table 5. Summary of Users Dataset

| Feature | Description |
|---|---|
| Total Records | 278,177 |
| Total Attributes | 3 |
| Unique Users | 278,177 |
| Unique Locations | 56,882 |
| Average Age | 34.75 |
| Missing Age Values | 110,508 |
| Most Frequent Location | London, UK |
| Purpose | User analysis |

## IV. METHODOLOGY

The chosen methodological approach in the current study will be based on application of machine learning methods to library databases. Several sets of data pertaining to books, customers, user ratings, and transactions were extracted from the Internet resource Kaggle, from Baburaoji Gholap College Library, and were generated manually. The collected data was combined into one database for analysis purposes.

Data preprocessing is needed for achieving data quality. Missing, duplicate, and inconsistent data should be addressed at this stage of work. Further, exploratory data analysis (EDA) is conducted to reveal any patterns and connections in various data attributes with the use of statistic measures and graphical representation. Application of machine learning algorithms enables researchers to build models that are used in further analysis and prediction. Such algorithms as Decision Tree and ensemble learning methods including Random Forest, Gradient Boosting, and XGBoost can be employed in classification tasks. K-Means is useful for grouping and revealing some hidden patterns within the dataset. In summary, the methodological approach in question includes data collecting, preprocessing, EDA, model building, and analysis.

A. Data Preprocessing

The data was pre-processed to ensure high-quality data before applying any machine learning algorithm. The missing data, duplicate data, and inconsistent data were managed through data preprocessing.

The categorical data, like Availability Status, were encoded as numerical data using Label Encoding. The numerical data, such as Pages, Price INR, Times Issued, and Rating, were used for machine learning.

Table 6. Data Preprocessing Summary

| Task | Description |
|---|---|
| Missing Value Handling | Median imputation |
| Noise Removal | Invalid age values removed |
| Duplicate Removal | Duplicates removed |
| Data Integration | Merged using ISBN & User-ID |
| Encoding | Label Encoding |
| Normalization | Min-Max Scaling |
| Final Dataset | 138,027 × 12 |

B. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to uncover any underlying trends in the data. Several visualizations were created to examine the availability status, book types, rating distribution, user age distribution, and trends over time regarding the year of publication. The analysis revealed some key patterns regarding the user behavior and borrowing activity.

Figures Included in EDA

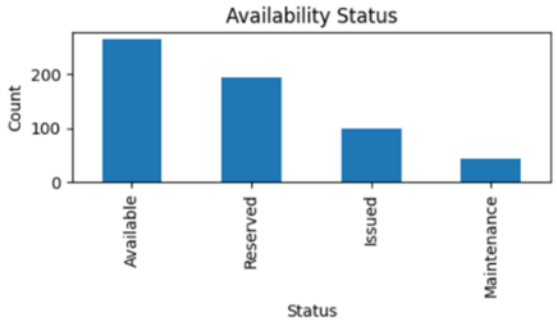


Fig. 1. Distribution of Availability Status

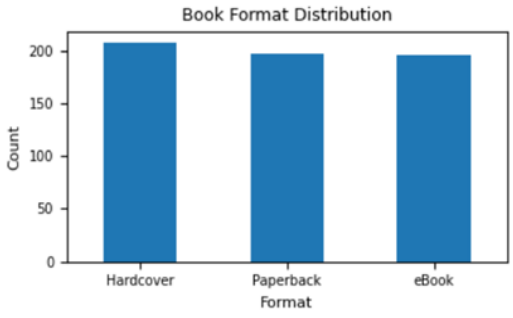


Fig. 2. Book Format Distribution


Fig. 3. Distribution of Book Ratings


Fig. 4. Top 10 Most Rated Books

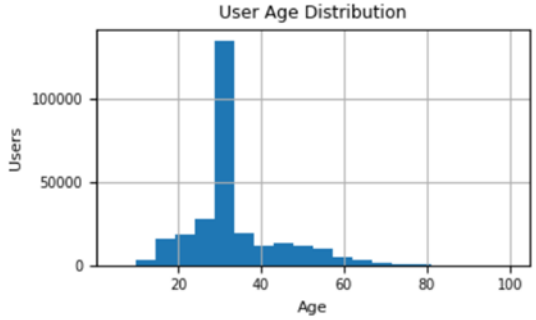


Fig. 5. Age Distribution of Users

The above charts offer a visual representation of user patterns, book distribution, and trends in the system, as observed in exploratory data analysis (EDA).

## V. IMPLEMENTATION AND ALGORITHMS

The system was developed using Python within a Jupyter notebook, using packages like Pandas, NumPy, Matplotlib, and Scikit-Learn. The data used was obtained from sources like Kaggle, college libraries, and other manually created datasets, including Books, Users, Ratings, and Transactions.

In the process of preprocessing, missing values, duplication, encoding of categorical data, and normalization were addressed. EDA was done using graphs and statistics techniques to determine patterns about the distribution of books, behavior of users, and their ratings. Classification algorithms like Decision Trees and clustering algorithms like K-Means have been implemented for the dataset. Evaluation of models was done using the split ratio of 80:20.

From the implementation above, it has been demonstrated that Python can be used to analyze large sets of data efficiently.

## VI. RESULTS AND DISCUSSION

This section discusses the use of machine learning models for classification and prediction of the proposed solution. The research uses Decision Tree, Random Forest, Gradient Boosting, and XGBoost classifiers to make predictions on whether a particular book is available or not, depending on attributes like Pages, Price INR, Times Issued, and Rating. The models are developed using an 80:20 split of training data and testing.

1) Decision Tree Classifier

Classification: The Decision Tree Classifier uses the classification technique by labeling the target variable, which is Availability_Status, into pre-defined categories. The classifier is built on the final preprocessed data set (138,027 × 12), which comes from the combination of different datasets, namely Books, Users, Ratings, and Transactions. Since Availability_Status is a categorical variable, the Label Encoding method is used to encode the variable into numeric values. The classifier categorizes the book according to its availability status, for example, Available, Reserved, Issued, and Maintenance.

Prediction: The prediction is done by predicting the value of Availability_Status using the same data as the model input. The data set used for this experiment is split into 80% training and 20% testing in the ratio of 80:20. The model analyzes the training data to predict the availability status of books during the testing process. The accuracy of the Decision Tree model is 0.50.

2) Random Forest Classifier

Classification: The Random Forest Classifier does classification by dividing the target variable, Availability Status, into various categories. This

model uses the last preprocessed data set that contains 138,027 rows and 12 columns that is formed using four different data sets such as Books, Users, Ratings, and Transactions. Because Availability Status is categorical, this variable is changed to numeric through the method of Label Encoding. This model categorizes the books based on various features such as pages, Price INR, Times Issued, and Rating.

Prediction: The Random Forest algorithm predicts through estimating the Availability_Status value using different decision trees. The data set is divided between the training and test datasets in the ratio of 80:20. Accuracy can be improved through the combination of different tree results. The accuracy rate obtained was 0.70, indicating superior prediction ability than the Decision Tree algorithm.

3) Gradient Boosting Classifier

Classification: The Gradient Boosting Classifier makes predictions about the target Availability_Status by placing it into one of the several predefined classes. This classifier is fitted to the processed final dataset (138,027 × 12) obtained from the processing of Books, Users, Ratings, and Transactions datasets. As Availability_Status is categorical, it is encoded using label encoding. The classifier predicts the category of the book (e.g., Available, Reserved, Issued, and Maintenance) based on the features such as Pages, Price INR, Times Issued, and Rating.

Prediction: The Gradient Boosting algorithm makes predictions by predicting the availability status based on the iteration method which decreases the errors made by previous algorithms. The data set is split in an 80/20 split whereby 80 percent of the data set is used for training while 20 percent is for testing. This algorithm predicts with an accuracy of 0.70.

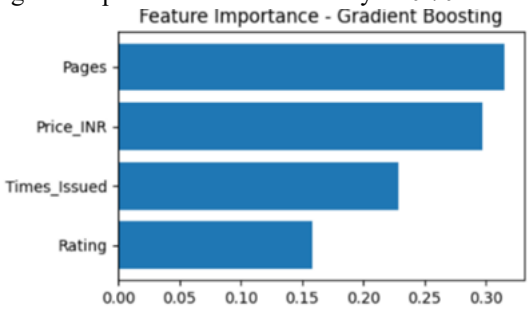


Fig. 6. Feature Importance of Gradient Boosting Classifier

4) XGBoost Classifier

Classification: XGBoost Classifier categorizes the Availability_Status into various classes for prediction purpose. The model is built on the final dataset, which includes 138,027 × 12, created through the combination of Books, Users, Ratings, and Transactions Datasets. As Availability_Status is a categorical data type, it is transformed to numerical format by applying the Label Encoding technique. This classifier categorizes the book as available, reserved, issued, or maintenance, depending upon the input variables such as Pages, Price INR, Times Issued, and Rating.

Prediction: In the case of XGBoost, the model makes the prediction by forecasting the Availability_Status value using an efficient boosted algorithm. In the process, the entire dataset is divided in an 80%:20% ratio to test the performance. This model detects the intricate patterns in the data and generates reliable predictions. It has attained an outstanding accuracy of 0.75, which makes it the most accurate model amongst all those tested.

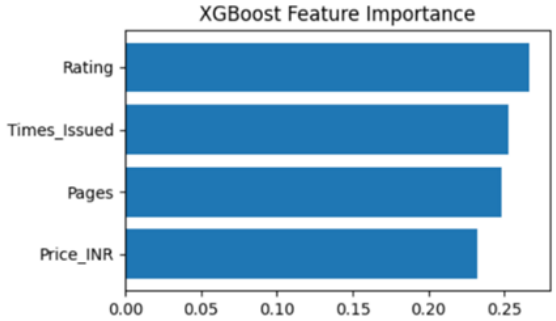


Fig. 7. Feature Importance of XGBoost Classifier

5)Model Comparison

Classification: The classification operation for all models entails putting the target variable “Availability Status” into classes like Available, Reserved, Issued, and Maintenance. The classification operation is performed on the processed data set (138,027 x 12), with features like Number of Pages, Price in INR, Times Issued, and Rating.

Prediction: Prediction accuracy for each model is based on the prediction of Availability_Status value using the same data set and predictors.

| Model | Accuracy |
|---|---|
| Decision Tree | 0.50 |
| Random Forest | 0.70 |
| Gradient Boosting | 0.70 |
| XGBoost | 0.75 |

The highest prediction accuracy score is achieved using the XGBoost model, proving that ensemble learning models are more accurate than the single Decision Tree model.

## VII. CONCLUSION

An intelligent library analytics system using machine learning approaches has been designed in this research based on books, users, rating and transaction data sets. Data pre-processing and data exploration has been performed in order to enhance data quality and detect useful patterns.

Various machine learning models such as Decision Tree, Random Forest, Gradient Boosting, and XGBoost have been applied in order to classify and predict the availability status of books in addition to K-Means to perform clustering. Out of all machine learning models, XGBoost yielded the best result (0.75).

It is clearly evident that machine learning algorithms have proven to be useful in the management of libraries. Possible future enhancements in this regard may include the application of deep learning models and real-time system, among others.

## REFERENCES

[1] S. P. Tamba et al., "Book data grouping in libraries using the K-means clustering method," 2019. [Online]. Available: https://www.researchgate.net/publication/335660447_Book_data_grouping_in_libraries_using_the_k-means_clustering_method

[2] V. S. Mane, "Big Data Analytics in Libraries: Enhancing Collection Management and User Experience," 2023. [Online]. Available: https://www.researchgate.net/publication/390626843_Big_Data_Analytics_in_Libraries_Enhancing_Collection_Management_and_User_Experience

[3] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques.

[4] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning.

[5] C. M. Bishop, Pattern Recognition and Machine Learning.

[6] C. C. Aggarwal, Data Mining: The Textbook.

[7] S. Raschka and V. Mirjalili, Python Machine Learning.

[8] W. McKinney, Python for Data Analysis.

[9] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research. [Online]. Available: https://arxiv.org/abs/1201.0490

[10] "Use of library resources and services: A study of review of literature." [Online]. Available: https://www.researchgate.net/publication/371778711_Use_of_library_resources_and_services_A_study_of_review_of_literature

[11] "Role of Libraries in the Context of Research and Educational Institutions." [Online]. Available: https://www.ijtsrd.com/papers/ijtsrd67201.pdf