Dhirubhai Ambani
Institute of Information and Communication Technology

Documentation

for

Dataset - 4

By Group -16

Contributions :

Chinmaya (202218054)	Data Preprocessing, Model Training and testing(ML+ANN)
Riya (202218049)	Apply ML pipeline(Preprocessing+ML models) on Single Author Dataset
Swayista (202218035)	Data Visualization, Model prediction on Best selling rank as target
Asish (202218022)	EDA, ML model on rating average, Regularization
Kashish (202001425)	Documentation,Problem Statement, Dataset Description, Feature Selection

PROJECT PIPELINE

- Dataset description
- Data Preprocessing - checked null values, attribute types, duplicate values etc.
- Data Visualization -
    - KDE plot for numerical columns
    - 'Publication-date' analysis
    - Boxplot of numerical columns over months
- ML pipeline
    - Feature Transformation
    - Feature selection using correlation analysis
    - Splitting of data
    - Scaling of data
    - Model Fitting and Hyperparameter Tuning
        - Logistic Regression
        - Decision Tree Model
        - Random Forest Model
        - KNN classifier model
    - Regularization
    - ANN model
    - Model Evaluation
        - Accuracy
        - F1-score
        - Recall
        - Precision

MAIN FILE : final_CP_02

SecondFile: Final_single_author_CP_04.ipynb

In a toy dataset with a single author, we used machine learning models to predict "Rating Avg" and "Bestsellers Rank." The models included Logistic Regression, Random Forest, Decision Trees, and KNN Classifier. We also applied regularization to some models. Performance metrics were used to evaluate each model's accuracy.

Dataset Description :

Basic description : The dataset is a collection of books that were taken from the bookdepository.com site, as well as additional metadata about the books, such as title, cover image, dimensions, and category information.

- **Authors (list of str):**
  - **Description :** This column contains the name(s) of the author(s) of the book. It is formatted as a list of strings because a book may have one or multiple authors.
  - **Example :**
    - ["Mike Coburn"]
    - ["John Moran", "Carl Williams"]

- **Bestsellers-rank (int):**
  - **Description :** Represents the bestsellers ranking of the book. It's an integer, indicating the book's popularity or rank.
  - **Example :**
    - 49848
    - 11732

- **Categories (list of int):**
  - **Description :** This column indicates the categories or genres that the book belongs to. It is formatted as a list of integers, and you may need to reference an external file (e.g., "authors.csv") for category mappings.

    - **Example :**
        - [214, 220, 237, 2646, 2647, 2659, 2660, 2679]
        - [235, 3386]

- **Description (str):**
    - **Description :** Contains a detailed textual description of the book, offering insights into the book's content.
    - **Example :**
        - "SOLDIER FIVE is an elite soldier's explosive memoir of his time within the Special Air Service (SAS) and, in particular, his experiences during the 1991 Gulf War..."
        - "John Moran and Carl Williams were the two biggest rival drug barons in Australia..."

- **Dimension-x (float in cm) :**
    - **Description :** Indicates the width of the book in centimeters. It is a floating-point number, providing precise measurements.
    - **Example :**
        - 129.0
        - 150.0

- **Dimension-y (float in cm):**
    - **Description :** Represents the height of the book in centimeters. It is also a floating-point number.
    - **Example :**
        - 198.0
        - 203.2

- **Dimension-z (float in mm):**
    - **Description :** Specifies the thickness of the book in millimeters. It's a floating-point number, but note that the unit is millimeters, not centimeters.
    - **Example :**
        - 20.0
        - 25.4

- **edition (str):**

- **Description :** Contains information about the edition of the book, such as edition type or edition number.
- **Example :**
  - "New edition"
  - "Unabridged"

- **edition-statement (str):**
  - **Description :** Includes additional statements or details about the book's edition.
  - **Example :**
    - "Export - Airside ed"
    - "Reprint"

- **for-ages (str):**
  - **Description :** Indicates the recommended age range for the book's readers. It's a string that categorizes the target audience by age.
  - **Example :** "For ages 12 and up"

- **format (int):**
  - **Description :** Represents the format of the book, such as hardcover, paperback, or eBook. You may need to reference an external file (e.g., "formats.csv") for format mappings.
  - **Example :**
    - 1
    - 2

- **id (int):**
  - **Description :** A unique identifier for each book in the dataset. It's an integer, allowing for easy referencing and identification.
  - **Example :**
    - 9781840189070
    - 9781844547371

- **illustrations-note (str):**
  - **Description :** Contains any notes or information regarding illustrations in the book. It's a string, offering details about the book's visual content.
  - **Example :** "Contains color illustrations."

- **image-checksum (str):**
  - **Description :** A checksum value associated with the cover image of the book. It's a string, likely used for data integrity and verification purposes.
  - **Example :** "97c8e71f2ec114b34f243074d2091077"

- **image-path (str):**
  - **Description :** Provides the file path to the cover image of the book. It's a string indicating the location of the image file.
  - **Example :** "full/c/5/2/c529152ea1246c0cb17d6574d302eae6d2e7fb0a.jpg"

- **image-url (str):**
  - **Description :** Contains the URL to the cover image of the book, allowing online access to the image. It's a string serving as a link to the book's cover.
  - **Example :**
    - "https://d1w7fb2mkkr3kw.cloudfront.net/assets/images/book/lrg/9781/8401/9781840189070.jpg"

- **imprint (str):**
  - **Description :** Includes information about the publisher or imprint of the book. It's a string indicating the publishing entity responsible for the book.
  - **Example :** "Mainstream Publishing"

- **index-date (date):**
  - **Description :** Represents the date when the data was crawled or indexed. It's a date value, typically in a standardized format (e.g., YYYY-MM-DD).
  - **Example :** "2023-10-12"

- **isbn10 (str):**
  - **Description :** Contains the ISBN-10 (International Standard Book Number) for the book. It's a string serving as a unique identifier for the book in a specific format.
  - **Example :** "184018907X"

- **isbn13 (str):**
    - **Description :** Represents the ISBN-13 (International Standard Book Number) for the book. It's another string-based unique identifier for the book, but in a different format from ISBN-10.
    - **Example :** "9781840189070"

- **lang (list of str):**
    - **Description :** This column contains a list of the language(s) in which the book is available. It's a list of strings, indicating language options for the book.
    - **Example :**
        - ["en"]
        - ["en", "es"]

- **publication-date (date):**
    - **Description :** Specifies the date when the book was published. It's a date value in a standardized format.
    - **Example :** "2004-10-14"

- **publication-place (int):**
    - **Description :** An identifier for the place of publication. It's an integer, likely referencing a list of places associated with publication.
    - **Example :** 12345

- **rating-avg (float):**
    - **Description :** Represents the average rating of the book on a scale from 0 to 5. It's a floating-point number, allowing for precise average ratings.
    - **Example :** 4.03

- **rating-count (int):**
    - **Description :** Indicates the number of ratings that the book has received. It's an integer, representing the count of user ratings.

- **title (str):**
    - **Description :** Contains the title of the book. It's a string, providing the book's name.
    - **Example :**

        - "Soldier Five: The Real Truth About The Bravo Two Zero Mission"
        - "Underbelly: The Gangland War"

- **url (str):**
    - **Description :** A relative URL that, when combined with "https://bookdepository.com," forms the complete URL to access the book online. It's a string that facilitates online access to the book.
    - **Example :**
        - "/Soldier-Five-Mike-Coburn/9781840189070"
        - "/Underbelly-Andrew-Rule/9781844547371"

- **weight (float):**
    - **Description :** Specifies the weight of the book in kilograms. It's a floating-point number, providing the book's weight in a standardized unit.
    - **Example :**
        - 224.0
        - 285.76

Preprocessing:

1. Count of Values with 'X' in isbn1:

- The values that appeared in the 'isbn1' column with the value 'X' were retrieved and counted. 'X' appeared 100,102 times in the 'isbn1' column.

2. Cleaning Data:

- Eliminating Unnecessary Columns :
    - In order to simplify the data, we eliminated redundant columns from the dataset.
- Managing Missing Values:
    - Index-date, publication-place, for-ages, and edition columns with more than 80% missing values were dropped.
    - Rows with less than 1% of null values were dropped.

    - For additional processing, selected columns with missing values more than 30% but less than 95%.

3. **Data Imputation :**

- Divided the dataset into columns that were categorized and numerical.
- For numerical columns, missing values were imputed using the corresponding column means.
- Using the mode, or most frequent value, of the corresponding columns, imputed missing values for categorical columns.

4. **The last data check**

- Verified there were no null values in the output dataset and looked for any remaining missing entries.

Data Visualization, summarizing insights about the dataset through EDA.

- **Selecting Numerical Characteristics:**

    - Identified the dataset's numerical columns while concentrating on continuous data points.

- **Creating Subplots using Numerical Data:**

    - Using box plots for quartile analysis and KDE plots for smooth data density presentation, separate subplots were made for each numerical column.

Publication date analysis :

- 'Publication date' column was converted to the correct date format for conventional analysis.
- From the formatted dates, the year, month, and day components were extracted and separated into distinct columns for further analysis.
- Counted the number of distinct years that are included in the dataset.
- For a more focused analysis, the dataset was filtered to only include novels released between 1960 and 2030.
- Calculated the number of books released in every year for the given period and produced a **bar plot** for display.
- Sorted the months in ascending order, counted the number of books published in each month for the filtered years, and produced a matching bar plot.
- Calculated the number of books released per day for the given years, arranged the days in ascending order, and used a bar plot to display the data.

Bestseller Ranks Analysis:

- The dataset was filtered in order to identify the 100 lowest bestseller ranks and the highest 100 bestseller ranks, which were then categorized using the 'lang_encoded' column.
- To graphically depict the distribution of languages for the lowest bestseller positions, a bar plot was created.

Feature Transformation :

- Analyzed the 'rating-avg' value distribution using a **histogram** to determine the data's range and distribution.
- Considered that in multiclass classification, discrete values are required as classes; therefore, 'rating-avg' must be transformed into separate categories.

- Created a clear depiction of the frequency of various "rating-avg" categories by using a bar plot to illustrate the data.
- Created a lexicon to translate 'rating-avg' data into discrete classes—a critical stage in multiclass classification tasks requiring discrete numerical classes (beginning at 0).
- Modified the attribute data types (such as "rating-avg" and "bestsellers-rank") to provide consistent data formats for modeling and analysis.
- Used the apply() method to apply sentence case transformation to particular columns, improving the dataset's readability and consistency.
- Apply() was used to convert some columns to lowercase in order to maintain consistency in the way text data was presented.
- Used encoding methods to translate category variables into numerical representations so that machine learning models could use them. 'imprint' column was encoded, converting textual imprint data into numerical values for analysis.
- Used encoding techniques to translate the 'lang' column into numerical representations, allowing language data to be handled effectively in modeling and data analysis procedures.

Selecting the Most Interesting Problem:

Rating classification seems to be an interesting classification as compared to others. Identifying books with higher rating can have a significant business impact. Knowing which books are trending or are likely to become bestsellers can help us allocate resources more efficiently and focusing marketing efforts. By accurate prediction of rating of books, one can

1. Enhance user experience for an online book store or library. one can recommend high rated books to users and increase customer engagement and satisfaction.
2. Identifying books that may get higher rating in near future may give one a competitive edge in the market. It allows one to get ahead and promote these books before its competitors do.
3. Rating classification also helps one to make marketing strategies. One can target promotions and discounts on books that are not yet rated but have the potential to get a higher rating, thus boosting sales.
4. Analyzing the performance of books classified as high rated or low rated over time can help refine the popularity classification model and improve its accuracy.

Reasons for choosing this problem:

1. **Market trend analysis:** By continually monitoring the types of books that are getting high rating than others, you can track the trends in the book industry. This knowledge is

vital for staying ahead of market shifts and making informed decisions about which books to promote.

2. **Content curation:** Bookstores and Libraries can curate collections of books with high ratings for specific genres, themes or target audiences making it easier for users to discover trending content.

Conducting Feature Selection through Correlation Analysis:

- **Feature Selection through Correlation Analysis:**
  - Correlation analysis was used to identify important characteristics based on how they related to each other or the target variable.
  - Created a correlation matrix to show the connections between various parameters and give a thorough picture of how interdependent they are.

- **Heatmap Visualization:**
  - To visually depict the correlation matrix to simplify the identification of patterns and correlations between features.

- **Managing Mirror Copies of Correlations:**
  - Found that there was a perfect link between "id" and "isbn13," which made it necessary to eliminate one of these attributes in order to prevent duplication.

- **Data Preparation for Analysis:**
  - The data was ready for further analysis, modeling, and interpretation by integrating the chosen features.

Model Training and Evaluation

Multi class Logistic Regression Using Multinomial

- **Prediction Process:**
    - Started the prediction task by predicting results based on the provided data using a machine learning model.

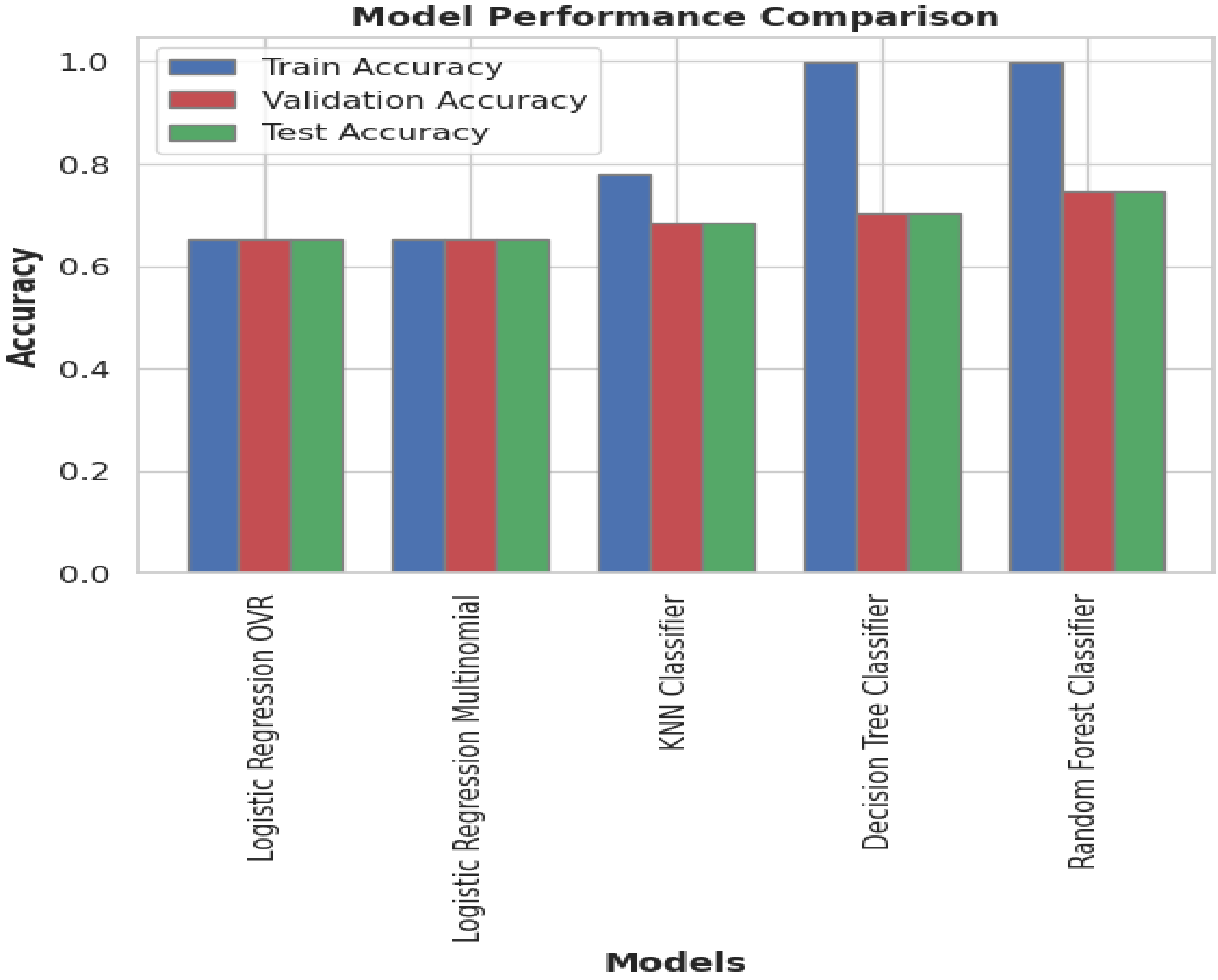
- **Creation of a Decision Tree Model:**
    - Created a Decision Tree model that was customized to meet the precise prediction criteria. Decision trees are a prevalent machine learning approach for both regression and classification tasks.

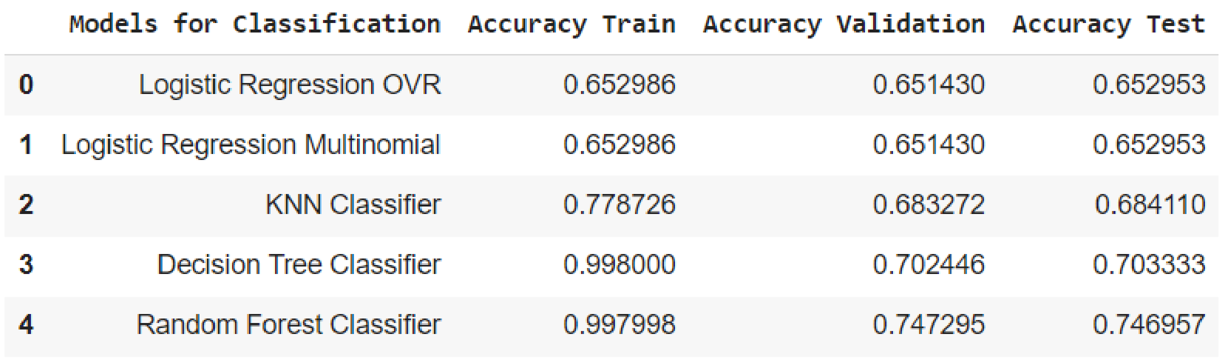
- **Printing Prediction Outcomes:**
    - Printed and presented the Decision Tree model's findings, illustrating the predicted conclusions or classifications for the supplied input data.

Multi class Logistic Regression Using OVR-- One v/s Rest

- Same as above model

Model's Performance and Comparison





	Models for Classification	Accuracy Train	Accuracy Validation	Accuracy Test
0	Logistic Regression OVR	0.652986	0.651430	0.652953
1	Logistic Regression Multinomial	0.652986	0.651430	0.652953
2	KNN Classifier	0.778726	0.683272	0.684110
3	Decision Tree Classifier	0.998000	0.702446	0.703333
4	Random Forest Classifier	0.997998	0.747295	0.746957

Regularization :

For Logistic Regression -

- Using L2 regularization in the multiclass and multinomial classification situations did not result in significant changes in the model's performance.

ANN Model :

Initialization of the Model:

The neural network model's initialization, laying the groundwork for further configuration and training.

Adding a Layer to the Model:

Added layers to the model architecture that allowed information to flow through the network and defined its structure.

Model Gathering:

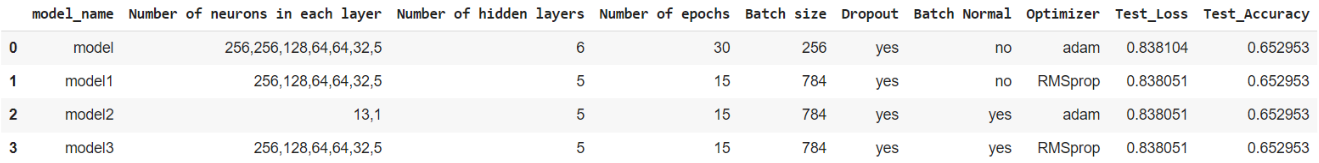
Built the model, ready for training, by defining the evaluation metrics, loss function, and optimizer.

Model Instruction:

Utilizing the training dataset, the model was trained to identify patterns and relationships in the data.

Assessment of the Test Set:

Evaluated the model's performance using the test dataset to determine its efficacy and accuracy with unknown data.



	model_name	Number of neurons in each layer	Number of hidden layers	Number of epochs	Batch size	Dropout	Batch Normal	Optimizer	Test_Loss	Test_Accuracy
0	model	256,256,128,64,64,32,5	6	30	256	yes	no	adam	0.838104	0.652953
1	model1	256,128,64,64,32,5	5	15	784	yes	no	RMSprop	0.838051	0.652953
2	model2	13,1	5	15	784	yes	yes	adam	0.838051	0.652953
3	model3	256,128,64,64,32,5	5	15	784	yes	yes	RMSprop	0.838051	0.652953