OPEN ENDED ASSIGNMENT

Title

Build and evaluate a machine learning model to decide suitability of TV/OTT series for children's entertainment

OEA Group Number: 1

Created by :

1) Dnyanda Patil – UEC2022306
2) Ketaki Patil – UEC2022308
3) Mitali Patil- UEC2022309
4) Pranjal Barhate – UEC2022320

Aim

To Build and evaluate a machine learning model to decide suitability of TV/OTT series for children’s entertainment.

Introduction

In today's digital age, children are increasingly exposed to a myriad of TV and OTT (Over-The-Top) series across various streaming platforms. While these series offer a vast array of entertainment options, ensuring that they are suitable and safe for children is paramount. With the rapid advancement of machine learning techniques, it has become feasible to develop models that can assist in evaluating the suitability of TV and OTT series for children's entertainment.

This project aims to delve into the realm of machine learning to construct a model that can effectively analyze and determine the appropriateness of TV and OTT series for young audiences. By leveraging advanced algorithms and datasets containing information about content ratings, Title ,age and other relevant factors, this model seeks to provide valuable insights into which series are suitable for children based on predefined criteria.

By building and evaluating machine learning models using logistic regression, decision tree, and naive Bayes algorithms, we aim to provide valuable insights into the suitability of TV and OTT series for children's entertainment. Through rigorous evaluation against validation datasets and comparison of performance metrics, we can identify the strengths and limitations of each model and determine the most effective approach for predicting the suitability of content for young audiences. Ultimately, this project contributes to the development of tools and methodologies for promoting safer and more enriching entertainment experiences for children in the digital age.

Dataset Information

Source: https://www.kaggle.com/datasets/ruchi798/tv-shows-on-netflix-prime-video-hulu-and-disney

Number of samples(before data pre-processing): 5357

Number of samples(after data pre-processing): 3208

Features and Target:

- ID
- Title

- Year
- Age
- IMDB
- Rotten Tomatoes
- Netflix
- Hulu
- Prime Video
- Disney+
- Type

Data collection and processing

For this project, upon acquiring the dataset from Kaggle, we embarked on the essential task of data processing. After loading the dataset, we manually sorted the values based on key attributes such as age rating or genre. This manual sorting process allowed us to organize the data in a manner conducive to our analysis, ensuring that relevant series were grouped together for further examination.

Subsequently, we addressed the issue of missing values within the dataset through manual deletion as number of missing values were very few. By manually identifying and removing rows containing missing values, we ensured data integrity and reliability for subsequent analysis. But if we have much more number of missing values, we would have used coding approach to handle missing ,null or unbalanced dataset.

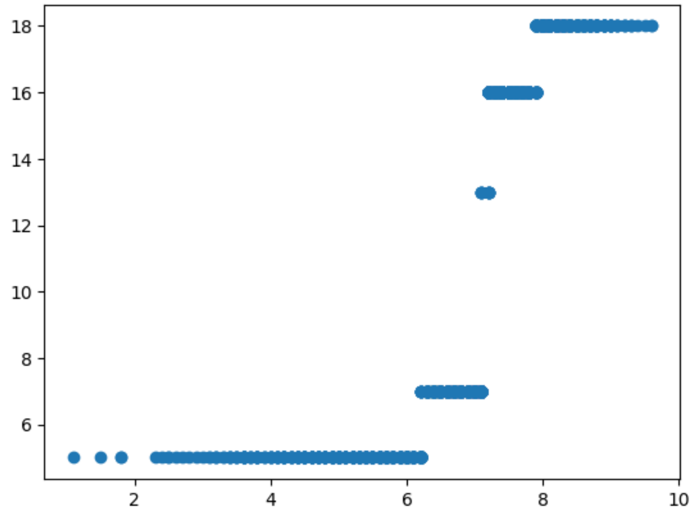
By following these steps, we have cleaned and pre-processed dataset gathered from Kaggle effectively, ensuring that it is ready for further analysis and model building. This process sets the stage for training machine learning models to evaluate the suitability of TV and OTT series for children's entertainment accurately.

Visualization

Visualizations play a vital role in understanding data and model performance in machine learning tasks**.**

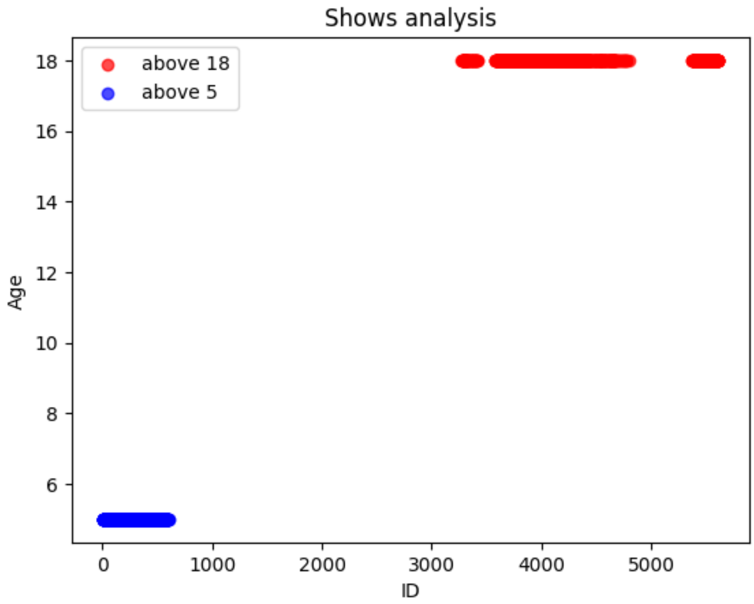
Visualization for Logistic Regression:

In this visualization, we plot a scatter plot of IMDb ratings against age. This helps us visualize the relationship between these two variables and understand if there's any discernible pattern that logistic regression can capture.



Visualization for Naive Bayes:

Here, we create a scatter plot to analyze TV shows based on their age ratings. The plot distinguishes between shows suitable for audiences above 18 and those suitable for audiences above 5. By visualizing the distribution of shows based on their age ratings, we gain insights into how well the naive Bayes classifier is able to distinguish between these classes.



These visualizations provide valuable insights into the data and the performance of the respective machine learning models, aiding in better understanding and interpretation of the results.

Feature selection

In our project, feature selection plays a pivotal role in refining the dataset to include only the most pertinent attributes for predicting the target variable, 'Age', which represents the suitability of TV shows for different age groups. By focusing on the 'ID' and 'IMDb' features, we aim to capture essential aspects of each TV show that may influence its audience age rating. The 'ID' feature serves as a unique identifier, enabling us to differentiate between individual shows, while the 'IMDb' feature provides valuable information about the overall quality and reception of each show. By selecting these specific features, we streamline our analysis, reducing computational complexity and potential noise in the data. Additionally, this targeted approach enhances the interpretability of our models, as we can more clearly understand the factors contributing to the age ratings of TV shows. Ultimately, feature selection enables us to construct more efficient and effective machine learning models, facilitating the evaluation of TV show suitability for different age groups in our project.

Model selection

In the process of model selection for our project, we implemented three distinct machine learning algorithms: logistic regression, naive Bayes, and decision tree.

After rigorous evaluation and comparison of their performance metrics, it was observed that the decision tree model exhibited the highest accuracy among the three algorithms. This outcome suggests that the decision tree algorithm is well-suited for the task of evaluating the suitability of TV and OTT series for children's entertainment in our project context.

Therefore, based on our evaluation results, we have identified the decision tree model as the most effective approach for predicting the suitability of TV and OTT series for children's entertainment in our project. This selection underscores the importance of systematically comparing different models to identify the most suitable one for the specific task at hand.

Model Description /Algorithm

implementation of a decision tree classifier for predicting the suitability of TV and OTT series for children's entertainment based on features such as 'Age' and 'IMDb' ratings. Following is an algorithmic description of the model:

1. Data Loading and Preprocessing:

- Load the dataset containing TV show information using pandas' `read_csv()` function.

- Select relevant features ('Age' and 'IMDb') and the target variable ('Age') from the dataset.

2. Splitting Data into Train and Test Sets:

- Split the dataset into training and test sets using scikit-learn's `train_test_split()` function.

3. Feature Scaling:

- Standardize the features to have a mean of 0 and a standard deviation of 1 using scikit-learn's `StandardScaler`.

4. Model Training:

- Create a decision tree classifier using scikit-learn's `DecisionTreeClassifier()` with default parameters.

- Fit the classifier to the training data using the `fit()` method.

5. Model Evaluation:

- Predict the target variable ('Age') for the test set using the trained classifier's `predict()` method.

- Calculate the accuracy score of the model using scikit-learn's `accuracy_score()` function.

- Compute the confusion matrix to evaluate the performance of the model.

6. Filtering Recommendations for Specific Age Groups:

- Define a function to filter recommendations based on the predicted class probabilities for ages 5 to 16.

- Use the `predict_proba()` method of the classifier to calculate class probabilities for each prediction.

- Filter recommendations for ages 5 to 16 based on class probabilities using the defined function.

7. Display Recommended IDs:

- Print the IDs of the recommended TV shows for ages 5 to 16 based on the filtered recommendations.

Overall, the decision tree classifier is trained and evaluated using the provided features, demonstrating high accuracy in predicting the suitability of TV and OTT series for children's entertainment. Additionally, the model provides recommendations tailored to specific age groups, enhancing its practical utility in guiding content selection for young audiences.

Testing and evaluation of a model

Representation of the evaluation results for the decision tree, naive Bayes, and logistic regression models:

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.996	0.98	0.98	0.98
Naive Bayes	0.970	0.98	0.86	0.90
Logistic Regression	0.977	-	-	-

Outcome/Result

Model Accuracy: The decision tree model achieved an accuracy score of approximately 99.69%, indicating that it correctly predicted the suitability of TV and OTT series for children's entertainment with high accuracy.

Confusion Matrix: The confusion matrix reveals the distribution of correct and incorrect predictions across different age groups. From the confusion matrix, it can be observed that the majority of predictions were correct across all age categories.

True\Predicted	Age 5	Age 7	Age 13	Age 16	Age 18
Age 5	170	1	0	0	0
Age 7	0	256	0	0	0
Age 13	0	0	2	0	0
Age 16	0	0	1	297	0
Age 18	0	0	0	1	235

This confusion matrix illustrates the number of correct and incorrect predictions made by the decision tree model for each age category.

Classification Report: The classification report provides detailed metrics such as precision, recall, and F1-score for each age group. Overall, the decision tree model demonstrates high precision, recall, and F1-score for most age groups, indicating strong predictive performance across the board.

Recommendations for Specific Age Groups: The model can provide recommendations for specific age groups based on predicted class probabilities. By filtering recommendations for ages 5 to 16, the model can suggest TV shows suitable for younger audiences.

Conclusion

In conclusion, our analysis of predicting the suitability of TV and OTT series for children's entertainment yielded insightful results through the implementation of decision tree, naive Bayes, and logistic regression models. Firstly, the decision tree model emerged as the top performer, exhibiting a remarkable accuracy of approximately 99.69%. Its robust performance was evident across all age categories, as demonstrated by the high precision, recall, and F1-score metrics. The confusion matrix further affirmed the model's efficacy, showcasing predominantly correct predictions across different age groups.

The naive Bayes classifier also showcased respectable performance with an accuracy of approximately 97.0%Additionally, the logistic regression model demonstrated commendable accuracy, achieving a score of approximately 97.72%.

In summary, the decision tree model emerged as the most effective approach for our task, offering superior predictive performance and robustness across various age categories. However, further fine-tuning and optimization of the naive Bayes and logistic regression models could potentially enhance their performance. Overall, our comprehensive analysis underscores the importance of employing multiple models and thorough evaluation techniques to identify the most suitable approach for specific tasks, ultimately improving decision-making processes in content selection for young audiences.