# Machine Learning Insights for Netflix Data Analysis

### 1. Model Selection

- **Logistic Regression** and **Random Forest** models were selected to analyze the Netflix dataset. These models were used for classification tasks, where the goal was to predict content type (Movie vs. TV Show) based on features like duration, number of seasons, and release year.

### 2. Logistic Regression Results

- **Accuracy**: 1.00 (100% accuracy in predicting the type of content)
- **Precision**: 1.00
- **Recall**: 1.00
- **Confusion Matrix**:

  [[1856  1] (True Negatives: **1856**, False Positives: 1)
  [ 0 780]] (False Negatives: 0, True Positives: 780)

- Logistic Regression perfectly classified the type of content in the dataset, with almost no misclassifications. This shows that the features provided (e.g., `duration_in_minutes`, `num_seasons`) are highly predictive of the content type.

### 3. Random Forest Results

- **Accuracy**: 1.00
- The Random Forest model also achieved perfect classification accuracy, reinforcing the effectiveness of the selected features in predicting whether content is a Movie or TV Show.

### 4. Cross-Validation

- **Logistic Regression Cross-Validation Score**: 1.00
- **Random Forest Cross-Validation Score**: 1.00
- Both models maintained 100% accuracy during cross-validation, which suggests the models are robust and unlikely to overfit the data. Cross-validation helps confirm the consistency of the model's performance across different subsets of the dataset.

### 5. Feature Importance (Random Forest)

- Feature Importance analysis in the Random Forest model identified the following key features:
  - **`duration_in_minutes`**: 58.96% importance (dominates the prediction)
  - **`num_seasons`**: 38.91% importance
  - **`release_year`**: 1.57% importance
  - **`director_count`**: 0.39% importance

    - **`genre_count`**: 0.16% importance
    - **`country_count`**: 0.00% importance
- The duration in minutes and the number of seasons are the most important features for classifying content type, which aligns with intuitive expectations (movies have a duration, TV shows have seasons). Other features, like director count and genre count, have a marginal impact on predictions.

**6. Hyperparameter Tuning (Random Forest)**

- The model was tuned using GridSearchCV, optimizing the following hyperparameters:
    - **Best Parameters**: {'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}
    - **Best Score**: 1.00
- Hyperparameter tuning improved the performance of the Random Forest model by finding the optimal combination of model parameters, further validating the model's reliability.

## Conclusion from Machine Learning Analysis:

- Both Logistic Regression and Random Forest models achieved perfect accuracy in predicting whether a piece of Netflix content is a movie or a TV show. This suggests that the selected features, particularly the `duration_in_minutes` and `num_seasons`, are highly informative.
- The high performance across multiple validation strategies, such as cross-validation and hyperparameter tuning, indicates that the models are robust, not overfitting, and generalize well to unseen data.
- The insights derived from the models highlight the importance of understanding content properties (e.g., duration and seasons) when analyzing or predicting Netflix content categories.