

Exploratory Data Analysis (EDA)

Objective

The objective of this task is to explore the datasets using visualizations, understand data patterns, detect outliers, analyze relationships between features, and identify important features for prediction.

Datasets Used

- Netflix Movies and TV Shows Dataset
 - Iris Dataset
-

Tools Used

- Python
 - Pandas
 - Matplotlib
 - Seaborn
 - Jupyter Notebook / VS Code
-

Dataset Analysis

1. Iris Dataset

- The dataset contains measurements of iris flowers.
- Numerical features include sepal length, sepal width, petal length, and petal width.
- The target variable is species (Setosa, Versicolor, Virginica).

EDA Performed:

- Histograms were plotted to understand feature distributions.
- Count plot was used to check the number of samples per species.
- Box plots were used to identify outliers.
- Correlation heatmap was plotted to understand relationships between features.

Insights:

- Petal length and petal width have strong positive correlation.
- Setosa species is clearly separable from others.
- Very few outliers are present in the dataset.

Important Features:

- Petal length
 - Petal width
-

2. Netflix Movies and TV Shows Dataset.

- The dataset contains information about movies and TV shows available on Netflix.
- Numerical features include release_year and duration.
- Categorical features include type, rating, country, and listed_in (genre).

EDA Performed:

- Histogram was used to analyze the distribution of release years.
- Count plot was used to compare the number of Movies vs TV Shows.
- Box plot was used to check outliers in release_year.
- Correlation heatmap was plotted using numerical features after encoding categorical data.

Insights:

- Movies are more in number compared to TV Shows.
- Most content was released after the year 2010.
- Some older shows appear as outliers.
- Release year shows weak correlation with content type.

Important Features:

- type
 - release_year
 - rating
-

Conclusion

This task helped in understanding how exploratory data analysis is performed using visualizations. EDA makes it easier to understand data behavior, detect outliers, and select important features before applying machine learning models. Both datasets are suitable for supervised machine learning tasks.