Supervised Learning / Multi Class Classification
1-Problem Statement / movies-classification.

The project is working in the multimedia field of films. We are interested in classifying a Film vote on the IMDb Internet Movie Database platform and other datasets. A possible classification can be carried out on the basis of the features inherent to a generic Film. This classification is useful for understanding which combinations of feature values contribute to bringing a film to success and consequently having a high / low rating on IMDb.

2-Data Description

Data columns (total 37 columns), rows = 3750.

Column	Description
Title	Title and subtitle of film (Dytpe object).
Year	Release year of the movie (Dytpe int64).
lifetime_gross	Gross sales for national film sales in the United States, not taking inflation into account (Dytpe int64).
ratingInteger	Vote on IMDb from 2 to 9 (Dytpe int64).
ratingCount	Number of voters on the IMDb site (Dytpe int64).
duration	Duration of the movie in seconds (Dytpe int64).
nrOfWins	Number of prizes won by the movie (Dytpe int64).
nrOfNominations	Number of nominations of which the film has not won any prize (Dytpe int64).
nrOfPhotos	Number of photos in the IMDb gallery for the movie (Dytpe int64).
nrOfNewsArticles	number of articles written and documented on the film (Dytpe int64).
nrOfUserReviews	number of users who have commented, offered a review or opinion on the movie on IMDb (Dytpe int64).
nrOfGenre	Number of genres of a film with a maximum number of 3 genres (Dytpe int64).
Action	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
Adult	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
Adventure	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).

Animation	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
Biography	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
Comedy	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
Crime	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
Documentary	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
Drama	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
Family	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
Fantasy	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
Horror	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
Music	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
Musical	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
Mystery	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
News	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
RealityTV	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
Romance	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
SciFi	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
Short	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
Sport	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
TalkShow	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
Thriller	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).

War	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).
Western	Boolean domain D = {0,1} that categorize the genre of a film (Dytpe int64).

3-Data Visualization

Data visualization helps us to understand the data by see how the data looks like and what kind of correlation is held by the attributes of data and then determine the features correspond to the output.

pairs plot is the most effective tools (also called a scatterplot matrix). A pairs plot allows us to see both distribution of single variables and relationships between two variables.

From figure

most features contain outliers.

Number of voters on the IMDb site (ratingCount), number of users who have commented, offered a review or opinion on the movie on IMDb (nrOfUserReviews), Two features that are related to each other.

Year, ratingcount. Number of voters on the IMDb site (ratingCount) began to increase after the year 2000.

Vote on IMDb from 2 to 9 (ratingInteger column). Most of data have vote 7(1493from 3750).

Number of prizes won by the movie (nrOfWins), Number of voters on the IMDb site (ratingCount) Two features that are related to each other (direct relationship).

Duration of the movie in seconds, for most movies is between 5,000 and 10,000.

4-Data Preprocessing

For doing data preprocessing in Healthcare Analytics problem. I used feature transformations**.**

Columns- used []

Columns_drop ['title']

Feature Transformations

- ✓ Data Cleaning

- ✓ Work with Missing Data
- ✓ Work with Categorical Data
- ✓ Detect and Handle Outliers
- ✓ Deal with Imbalanced Classes
- ✓ Feature Scaling

5-Train models

The machine learns patterns from data in such a way that the learned representation successfully maps the original dimension to the suggested class without any interference from a human.

Machine learning Algorithms for Classification Problem.

classification tasks have discrete categories, unlike regressions tasks.

- ✓ Logistic Regression
- ✓ K Nearest Neighbors (KNN)
- ✓ Support Vector machine (SVC)
- ✓ Decision Trees
- ✓ Random Forest
- ✓ XGBoost

6-Evaluating Model Performance

All results in file "result.txt"

7-Link of Web Application deployed on Heroku

https://movies-sys.herokuapp.com/