The purpose of this research was to apply prediction models to predict the trend of watching. I worked with data given by Netflix to get promising results for this multiclass challenge by combining different types of modeling and categorical feature engineering with a random forest, KNN Algorithms and Logistic Regression.

The first step is to explore the Data looking for any noise variable and its structure this step is to explore the data looking for missing value, correcting date format, etc. all those steps are necessary to have a clean data and ready for analysis. After that, we check the variables that contains missing values, after we find those variables. We clean and correct the missing values.

The exploratory data analysis (EDA) and Visualization considered being important to have a clear vision and discovering the data and the question we are trying to answer this EDA on NETFLIX MOVIES AND TV SHOWS, EDA done using python, numpy, pandas, matplotlib, seaborn and plotly. You will find many usefull visualisations and Tables in the project. We have tried to analyze most of the features of the dataset to derive insights.

Before we move to medialization part, the data cleaning and preprocessing with Pandas for medialization. Since using text variable is not effective for modeling this Cleaning and Preprocessing of the data before modelizaton, using different libraries and different method. As converting the text to numerical variables..

Finally, we finish by using three predictive machine-learning models, KNN, Random Forest and Logistic Regression. With analyzing each algorithms and comparing the accuracy of each model in this step of Analysis using different types of Machine Learning Algorithms for classifying and predicting. Using three types of Algorithms; KNN, Random Forest and Logistic Regression.

Logistic regression, k-Nearest Neighbors, and Random Forest classifiers was used before settling on random forest as the model with strongest cross-validation performance. Random Forest feature importance ranking was adopted directly to guide the choice and order of variables to be included as the model underwent refinement.

The full 7787-record training dataset split 80/20 train vs. holdout, and all scores shown below calculated using 5-fold cross validation on the training phase exclusively. Because predictions on the 20% holdout were limited to the very end, this split only utilized once, and the scores were seen only once.