Steam Store Game Analysis

컴퓨터공학
201714220 정든솔

2021.12.17

CONTENTS

Part 1.

Data Set

1.1 Data

❖ Data

- Steam Store Games (Clean dataset)
  - Games data scraped from Steam Store and SteamSpy APIs
  - Data Collected: May, 2019
  - Data Size: 27,075 Instances
  - Source: https://www.kaggle.com/nikdavis/steam-store-games
  - Data type: object(9), int64(8), float64(1)

	appid	name	release_date	english	developer	publisher	platforms	required_age	categories	genres	steamspy_tags	achievements	positive_ratings	negative_ratings	average_playtime	median_playtime	owners	price
0	10	Counter-Strike	2000-11-01	1	Valve	Valve	windows;mac;linux	0	Multi-player;Online Multi-Player;Local Multi-P...	Action	Action;FPS;Multiplayer	0	124534	3339	17612	317	10000000-20000000	7.19
1	20	Team Fortress Classic	1999-04-01	1	Valve	Valve	windows;mac;linux	0	Multi-player;Online Multi-Player;Local Multi-P...	Action	Action;FPS;Multiplayer	0	3318	633	277	62	5000000-10000000	3.99
2	30	Day of Defeat	2003-05-01	1	Valve	Valve	windows;mac;linux	0	Multi-player;Valve Anti-Cheat enabled	Action	FPS;World War II;Multiplayer	0	3416	398	187	34	5000000-10000000	3.99

Picture 1. Origin Data Set

1.2 Preprocessing

❖ Proprocessing

- release_date
  - pd.to_datetime으로 Timestamp로 변경
- Rating
  - rating = positive_ratings / (positive_ratings + negative_ratings)
  - total_ratings = positive_ratings + negative_ratings
- Owners
  - owner_level ≒ [log(owners)]
  - owners = (owners.MAX + owners.MIN) / 2
- Genres
  - One-hot Encoding + Drop unnecessary genres

	appid	release_date	publisher	required_age	genres	rating	total_ratings	average_playtime	median_playtime	owners	owner_level	price	Action	Adventure	Casual	Early Access	to Play	Gore	Indie	Massively Multiplayer	Nudity	RPG	Racing	Sexual Content	Simulation	Sports	Strategy
0	10	2000-11-01	Valve	0	Action	0.973888	127873	17612	317	15000000	9	7.19	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	20	1999-04-01	Valve	0	Action	0.839787	3951	277	62	7500000	8	3.99	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	30	2003-05-01	Valve	0	Action	0.895648	3814	187	34	7500000	8	3.99	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Picture 2. Preprocessed Data Set

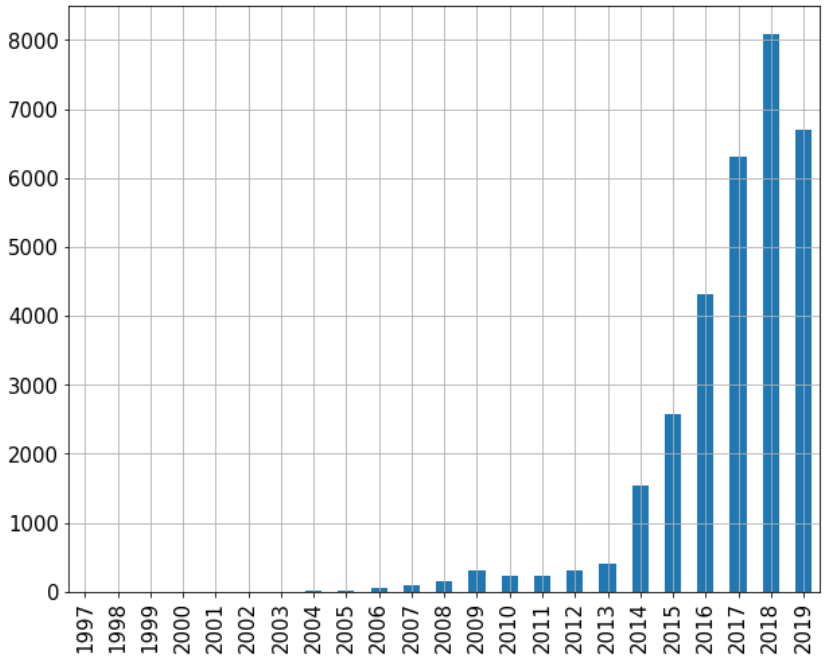
Part 2.

Time-Series & Regression

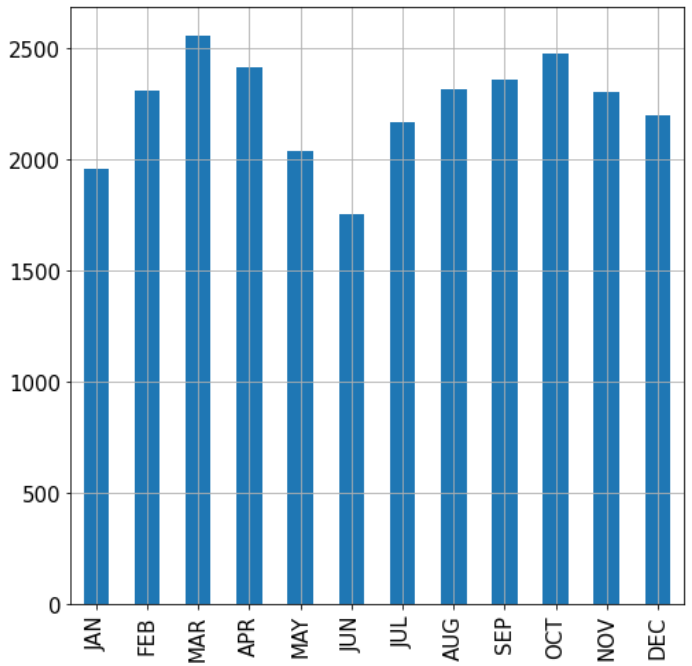
2.1 Time-Series & Correlation

- Time-Series
  - Trend: Incresed (Decreased in 2019)
  - Seasonality: Non-vacation season

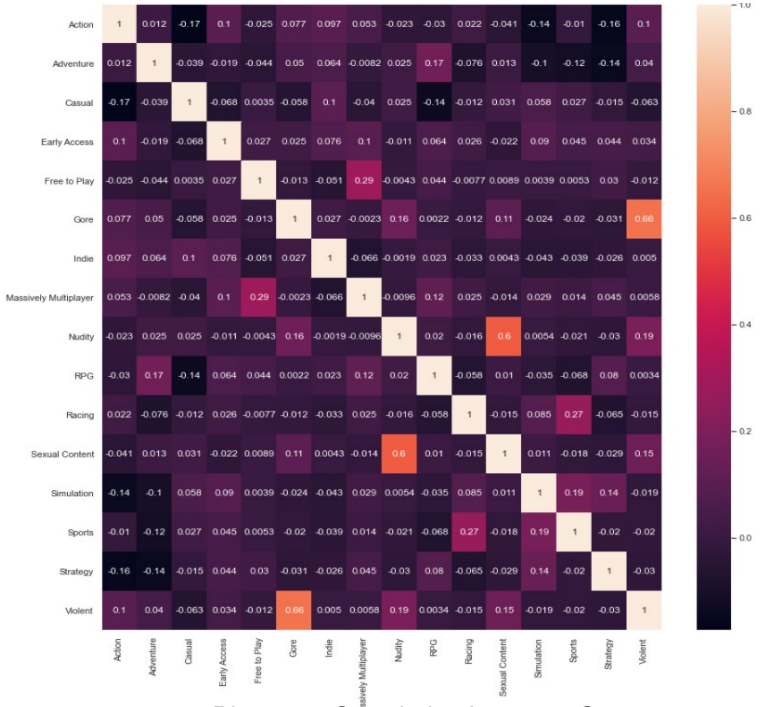
- Correlation
  - Low correlation between Genres



Picture 3. Trend of Data



Picture 4. Seasonality of Data



Picture 5. Correlation between Genres

2.2 Linear Regression

❖ Linear Regression

- Features
  - average_playtime / price / owner_level
- Target
  - rating
- Evaluation
  - Coefficient: 4.30933148 e-07 / 2.37918180 e-03 / 1.04271649 e-02
  - Constant: 0.69223608022612 61MAE: 0.182503
- Evaluation
  - MAE: 0.182503
  - MSE: 0.053695
  - RMSE: 0.231722
  - R2: 0.011840

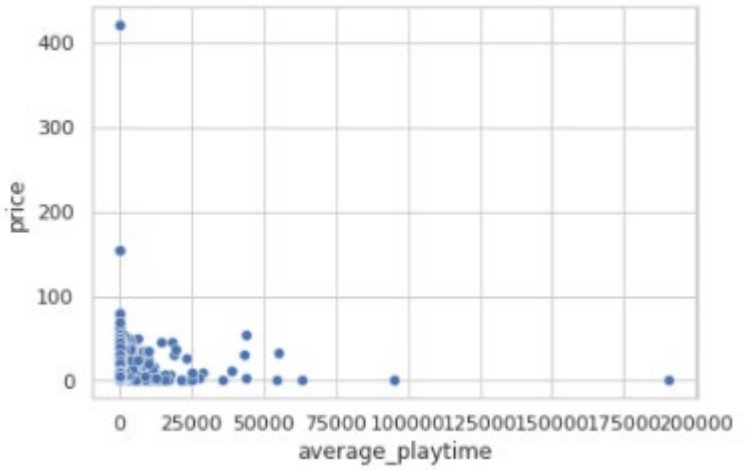
Part 3.

Clustering

3.1 Handling Outlier & Scaling(Normalization)

❖ Handling Outlier

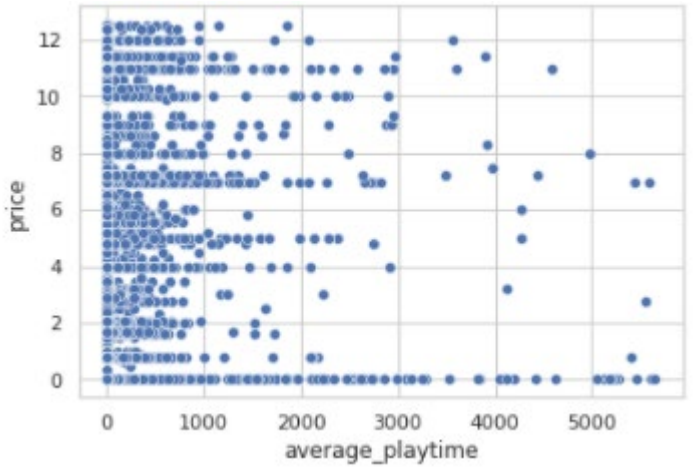
- Price
  - IQR: ± 1.5 * (df.price.quantile(0.75) - df.price.quantile(0.25))
- Average Playtime
  - Z-score: ± 3 * df.average_playtime.std()
  - Average_playtime.quantile(0.75) == 0.0이므로 IQR 불가
  - Normal Distribution이 전제되지 않으므로 Error 존재



Picture 6. Data with Outlier

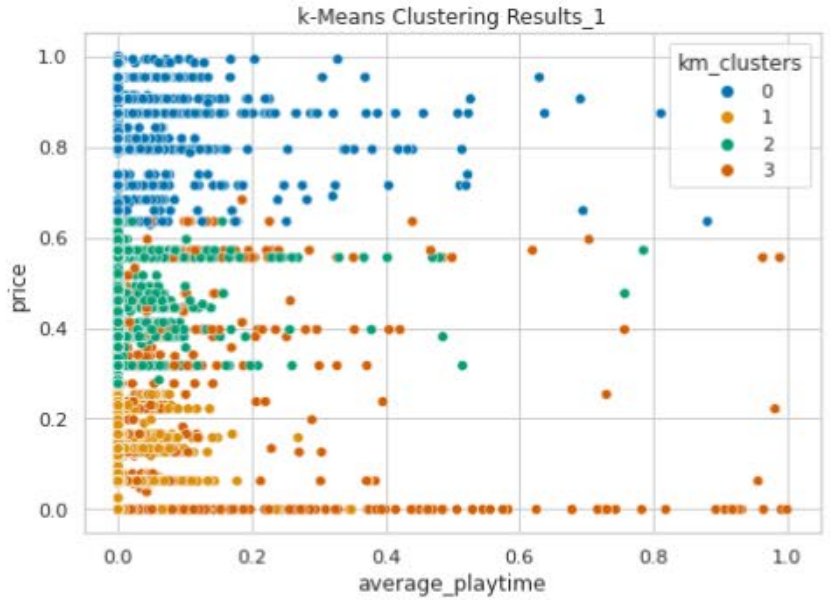
❖ Scaling

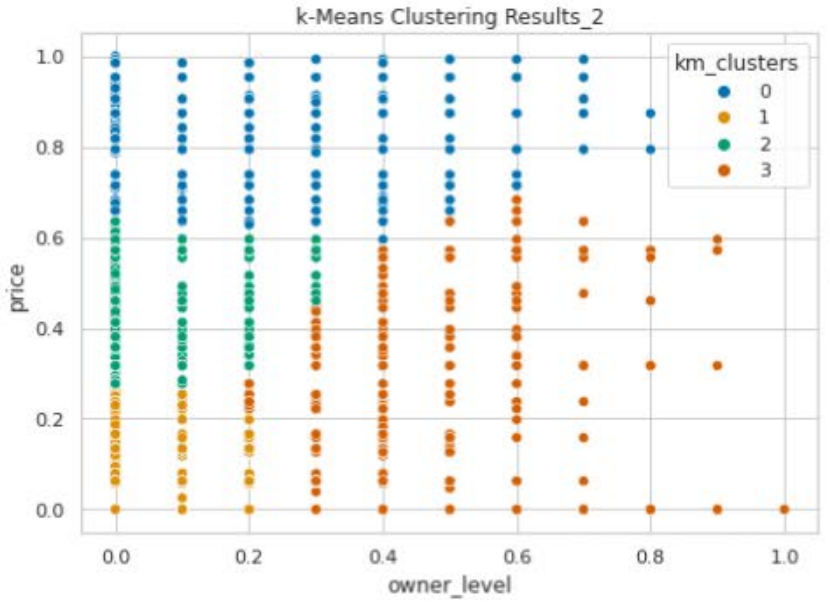
- Normalization (Clusterings use Distance)
  - X[col] = (X[col] - X[col].min()) / (X[col].max() - X[col].min()) RMSE: 0.231722

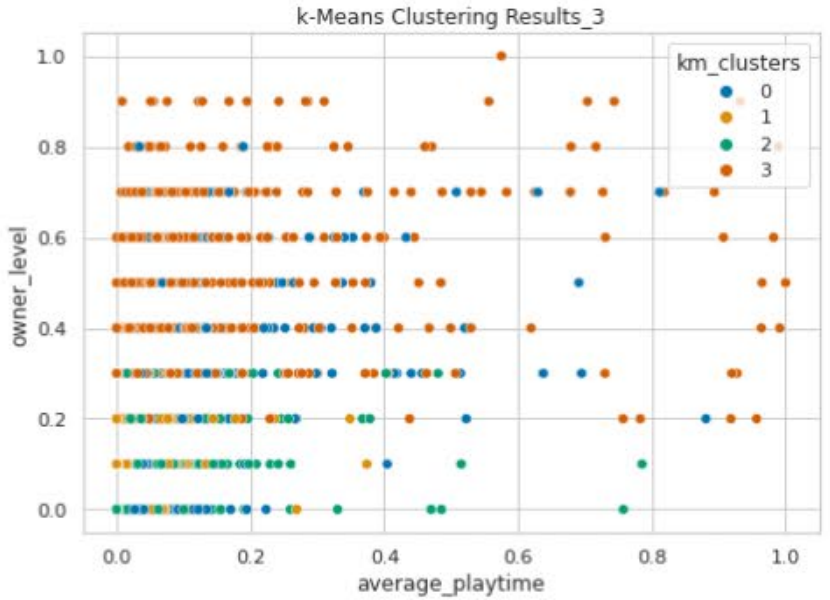


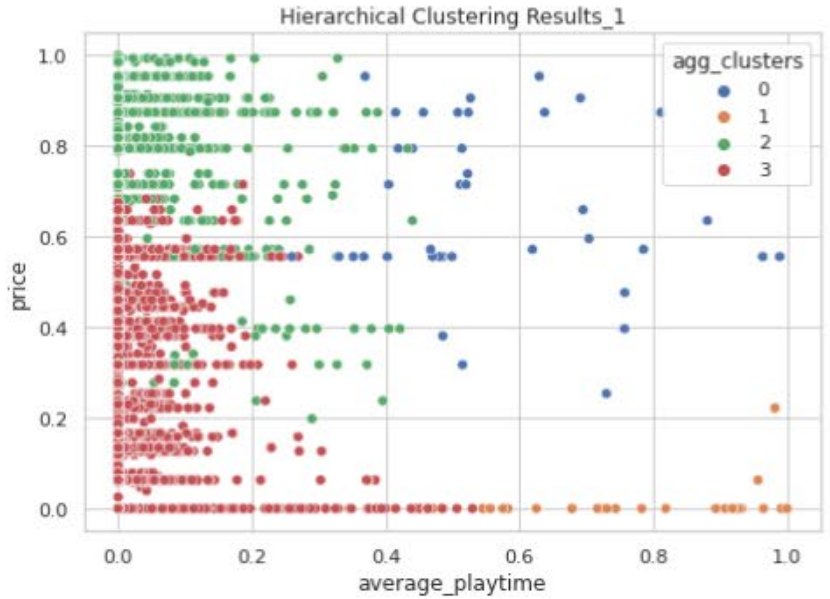
Picture 7. Data without Outlier

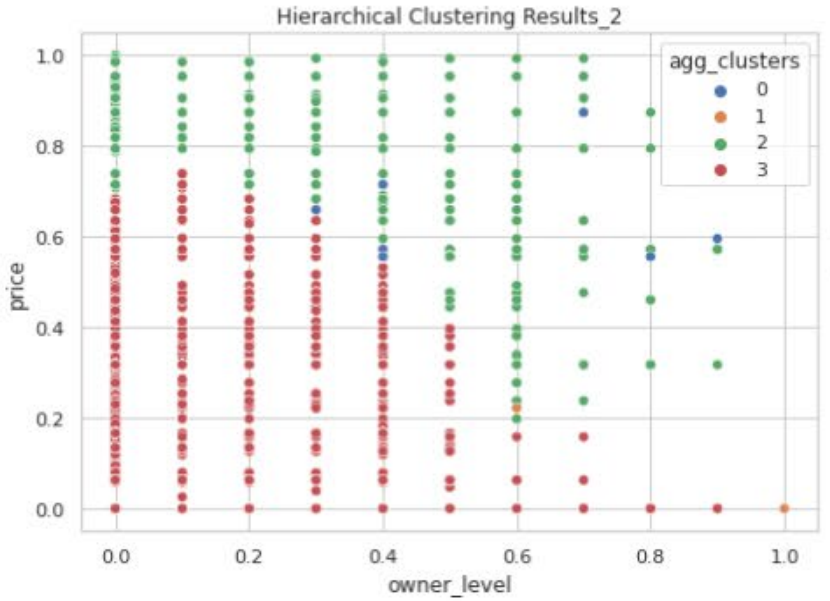
3.2 Clustering

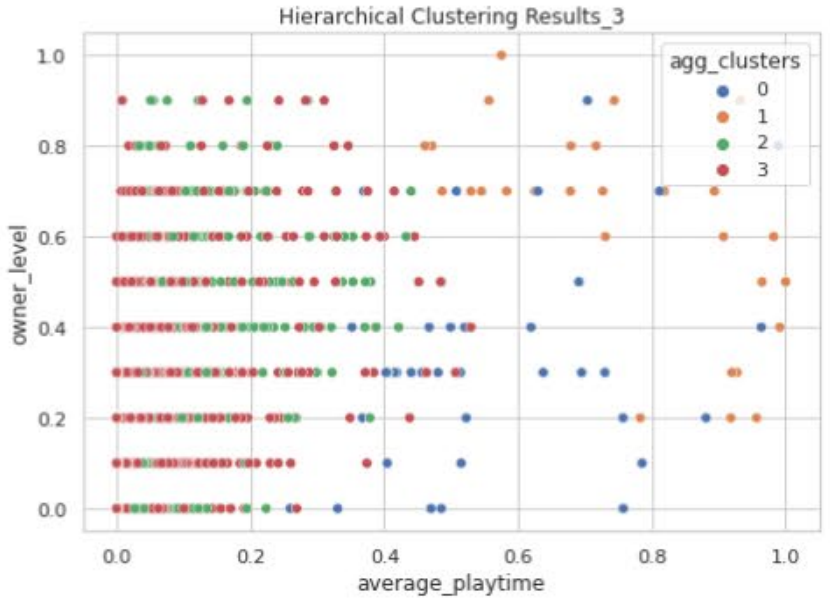












Part 4.

Multi-label Classification

4.1 Classification without Balancing Data

❖ Classification

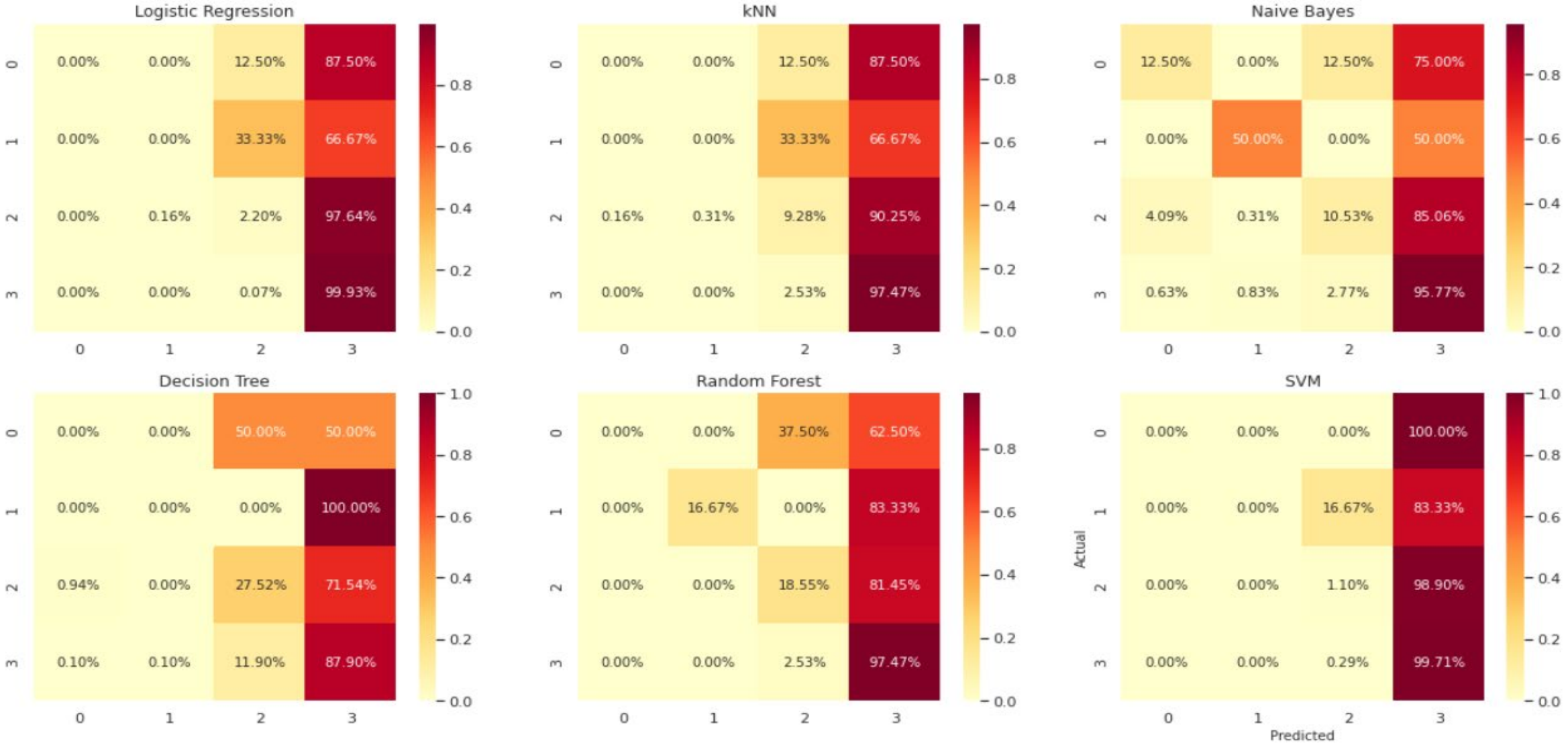
- Logistic Regression
- K-Nearest Neighbors
- Gaussian Naïve Bayes
- Decision Tree Classification
  - Feature Importances: TOP 20
- Random Forest Classification
  - Feature Importances: TOP 20
- Support Vector Machine

❖ Data Split & Feature

- Train set : Test set == 80 : 20
- Feature: Clustering에 사용된 Feature 제외

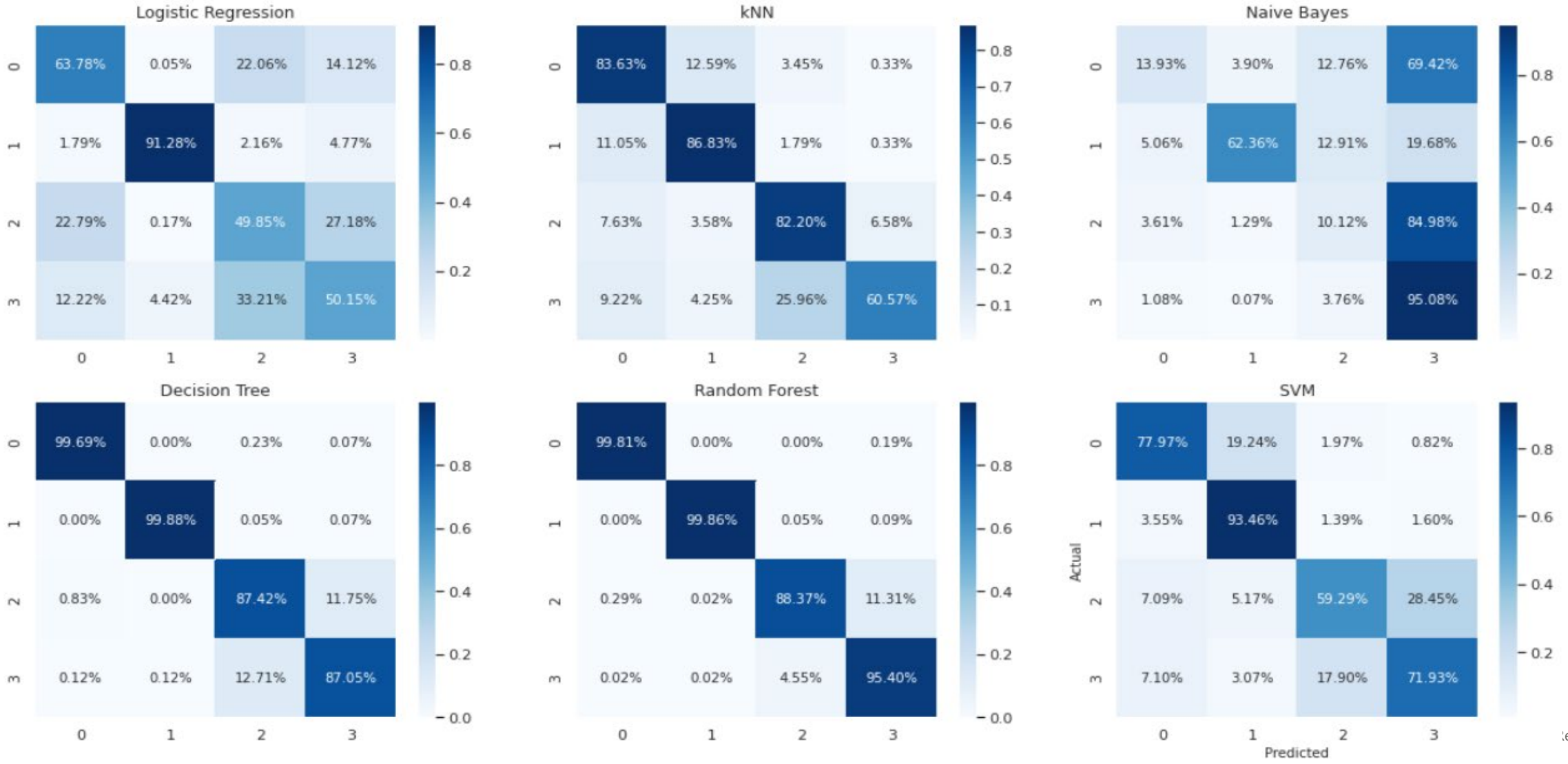
4.1 Classification without Balancing Data

- Average Accuracy: 0.847881



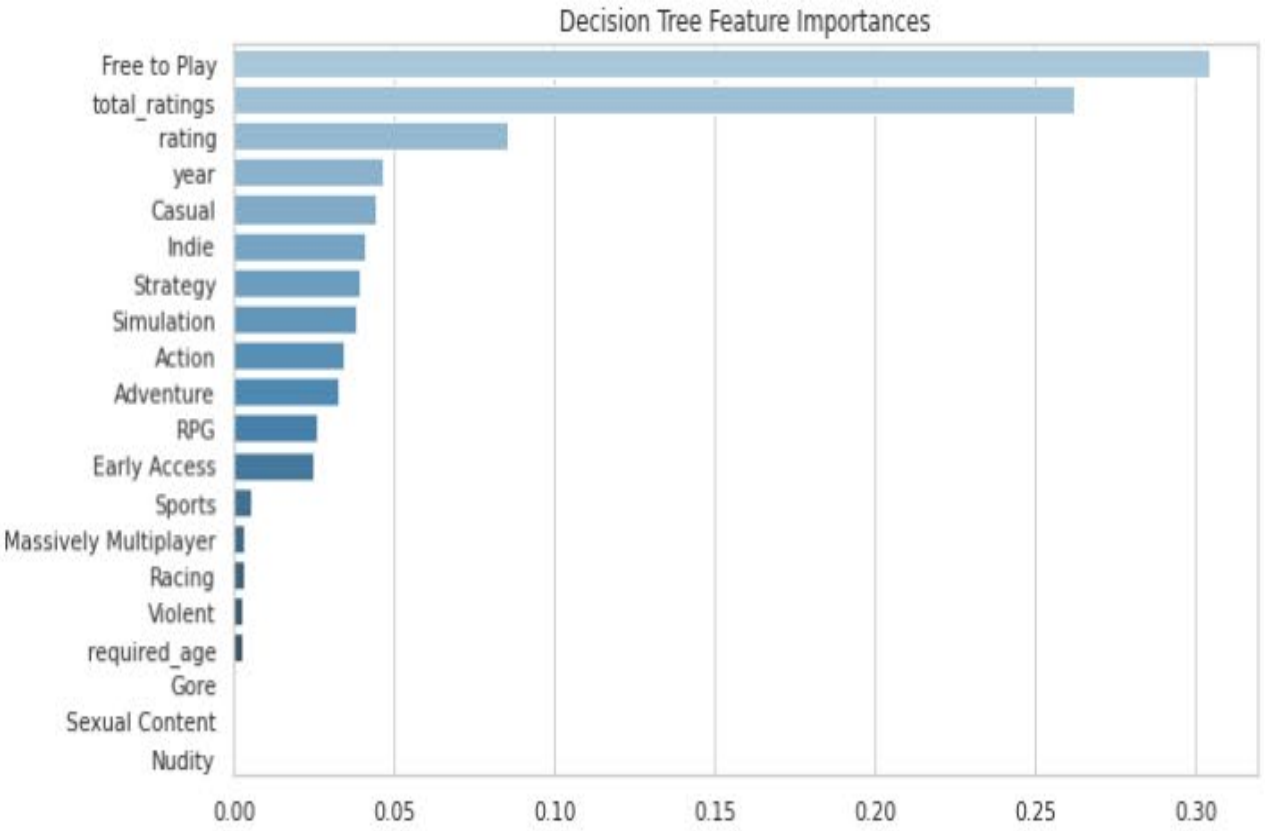
4.2 Classification with Over-sampling Data

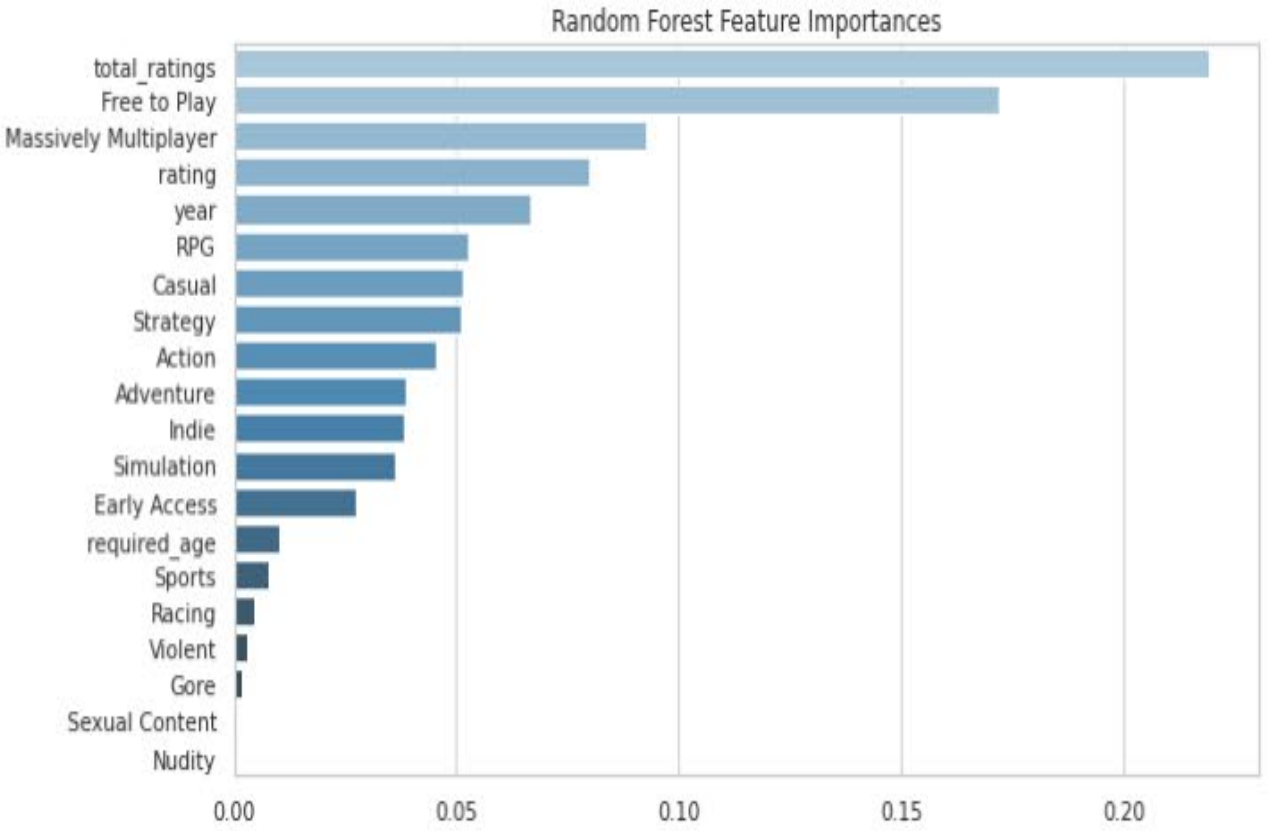
❖ Average Accuracy: 0.755116



4.2 Classification with Over-sampling Data

❖ Feature Importances





Part 5.

Difficulty & Future Task

5.1 Difficulty & Future Task

❖ Difficulty

- Data set
  - Multi-valued Attributes: Genres, steamspy_tags
  - Overly Categorical Data: Owners
  - Worthless Data: average_playtime, median_playtime

- Feature & Target
  - Genres: Multi-label
  - Game Type: Multi-class

❖ Future Task

- Need more valuable data and preprocessing to make worth analysis

Thank You

201714220 정든솔