#ITM #Data Mining #Cases

18102081 Soo-oh An 18102072 Sun-ho Kim 20102128 Yu-jeong Hwang

#Table Of Contents

CONTENTS

CONTENTS

CONTENTS

Case 1

Analysis of Popular YouTube Video Content using Data Mining

1

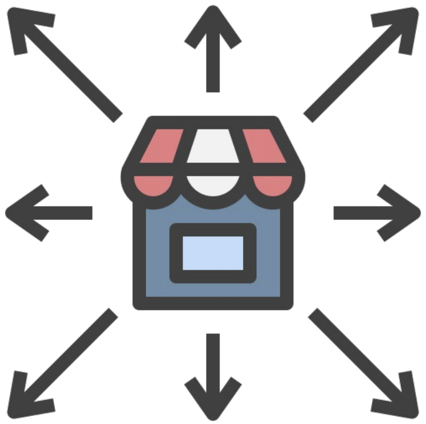
Case 2

Analysis of Utilization and Assessment of Predicting Models for YouTube Science Channel

2

Analysis of Popular YouTube Video Content using Data Mining

1. Motivation



=> Establishes hypothesis that **popular YouTube video content can affect increasing views that are important to YouTube's revenue generation model**

2. Method

Dataset

18.3GB of standardized data for a total of 30 days from January 17, 2020 to February 15, 2020 based on popular YouTube video content

Method of analysis

- Qualitative Analysis: Keyword Analysis with Text Data
- Quantitative Analysis:
    ->Pearson Correlation Analysis
    ->Univariate Multiple Regression
    1) Independent variables: likes, dislikes, comment
    2) Dependent variable: view

3. Qualitative Analysis

Keyword-centered Frequency Analysis through Text Data

Step 1: Run morpheme analysis using Konlpy and Natural Language Toolkit (NLTK) libraries based on collected text datasets.

Step 2: Extract keywords around nouns using the okt.nouns library.

Step 3: Prepare a dictionary of disused words to remove them from the extracted keywords.

Step 4: Use the collections library to print out the frequency of keywords

3. Qualitative Analysis

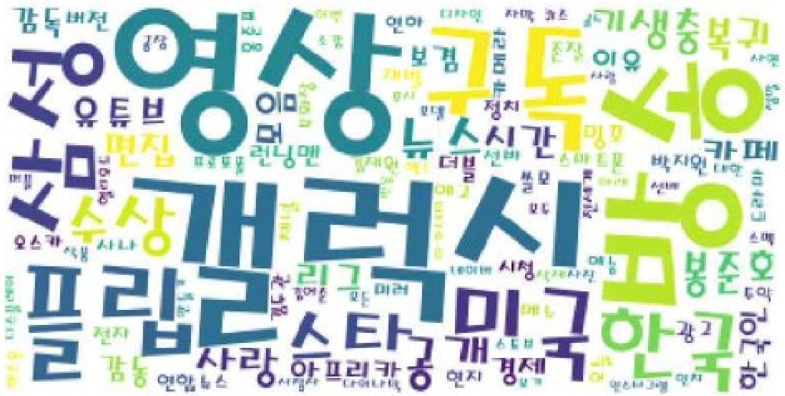


Fig. 7. Wordcloud visualization of the dataset collected on February 15, 2020

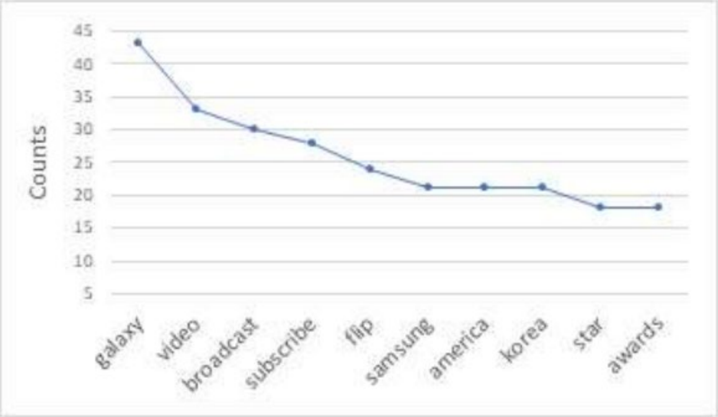


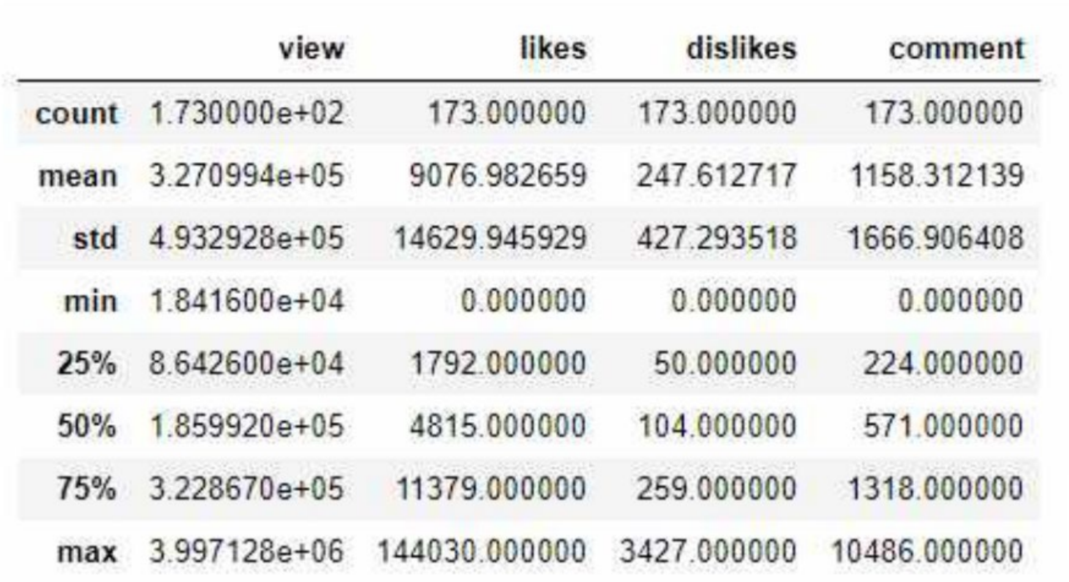
Fig. 8. Graph of keyword frequency

1. On February 9, 2020, when the Academy Awards were held
2. On February 13, 2020 Announces Plan to Sell Package Galaxy Z Flip Thom Browne Edition

=> It could be seen that the popular element of YouTube's popular video content is highly related to the main news of the period.

4. Quantitative Analysis

By using Pandas, matplotlib, and seaborn libraries, the correlation is analyzed between numeric data items based on four variables corresponding to the number of view, likes, dislikes and comments



	view	likes	dislikes	comment
count	1.730000e+02	173.000000	173.000000	173.000000
mean	3.270994e+05	9076.982659	247.612717	1158.312139
std	4.932928e+05	14629.945929	427.293518	1666.906408
min	1.841600e+04	0.000000	0.000000	0.000000
25%	8.642600e+04	1792.000000	50.000000	224.000000
50%	1.859920e+05	4815.000000	104.000000	571.000000
75%	3.228670e+05	11379.000000	259.000000	1318.000000
max	3.997128e+06	144030.000000	3427.000000	10486.000000

Fig. 10. Summary of Statistics Processing with the describe Library

4. Quantitative Analysis

Pearson Correlation Analysis

$$Cov(X, Y) = \frac{\sum_i^n (X_i - \overline{X})(Y_i - \overline{Y})}{n-1} \tag{1}$$

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{\frac{\sum_i^n (X_i - \overline{X})^2}{n-1}}\sqrt{\frac{\sum_i^n (X_i - \overline{X})^2}{n-1}}} \tag{2}$$

=> In the case of covariance, the flow of correlation can be identified, but if the measurement unit sizes of the two variables are different, it is not appropriate to grasp the correlation. Therefore, the Pearson correlation coefficient is utilized

4. Quantitative Analysis

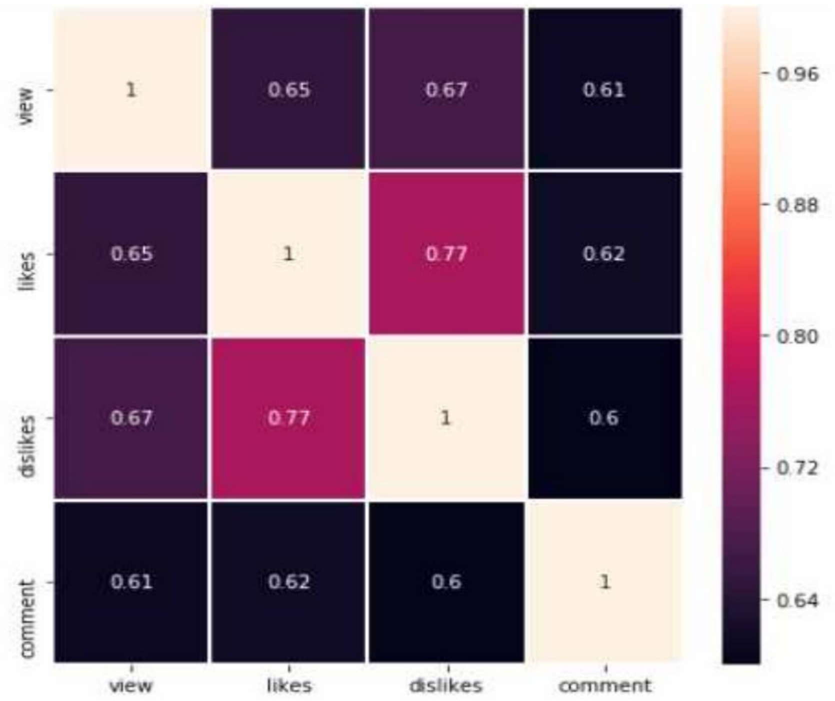


Fig. 11. Pearson Correlation Analysis of Variables

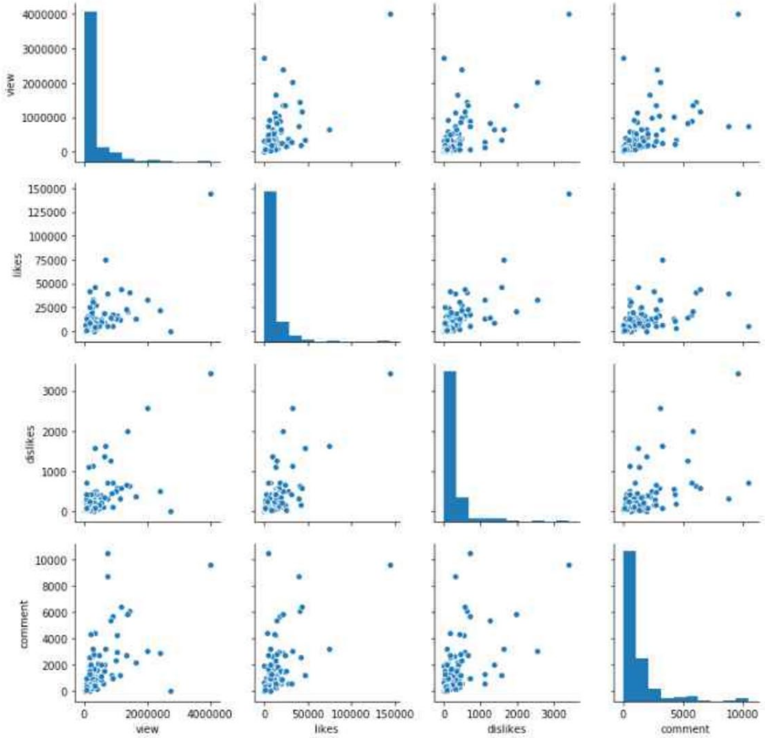
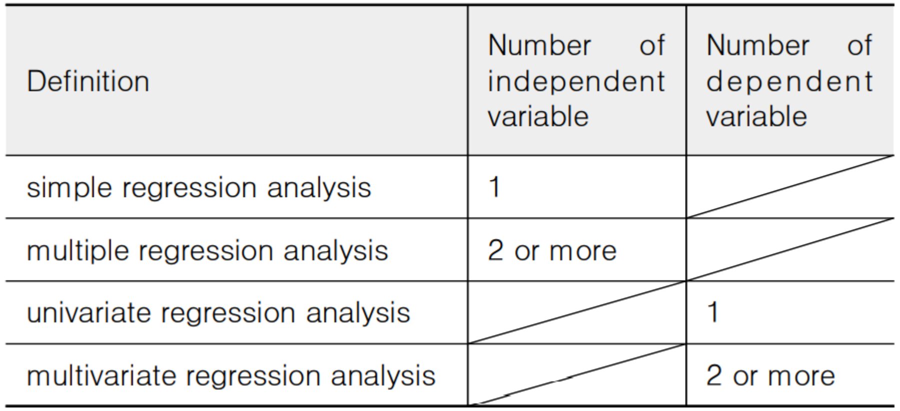


Fig. 12. Visualize the Correlation of Variables

4. Quantitative Analysis

Linear regression Analysis

Table 2. Regression types



Definition	Number of independent variable	Number of dependent variable
simple regression analysis	1	
multiple regression analysis	2 or more	
univariate regression analysis		1
multivariate regression analysis		2 or more

Univariate multiple regression analysis

Independent variables: likes, dislikes, comment

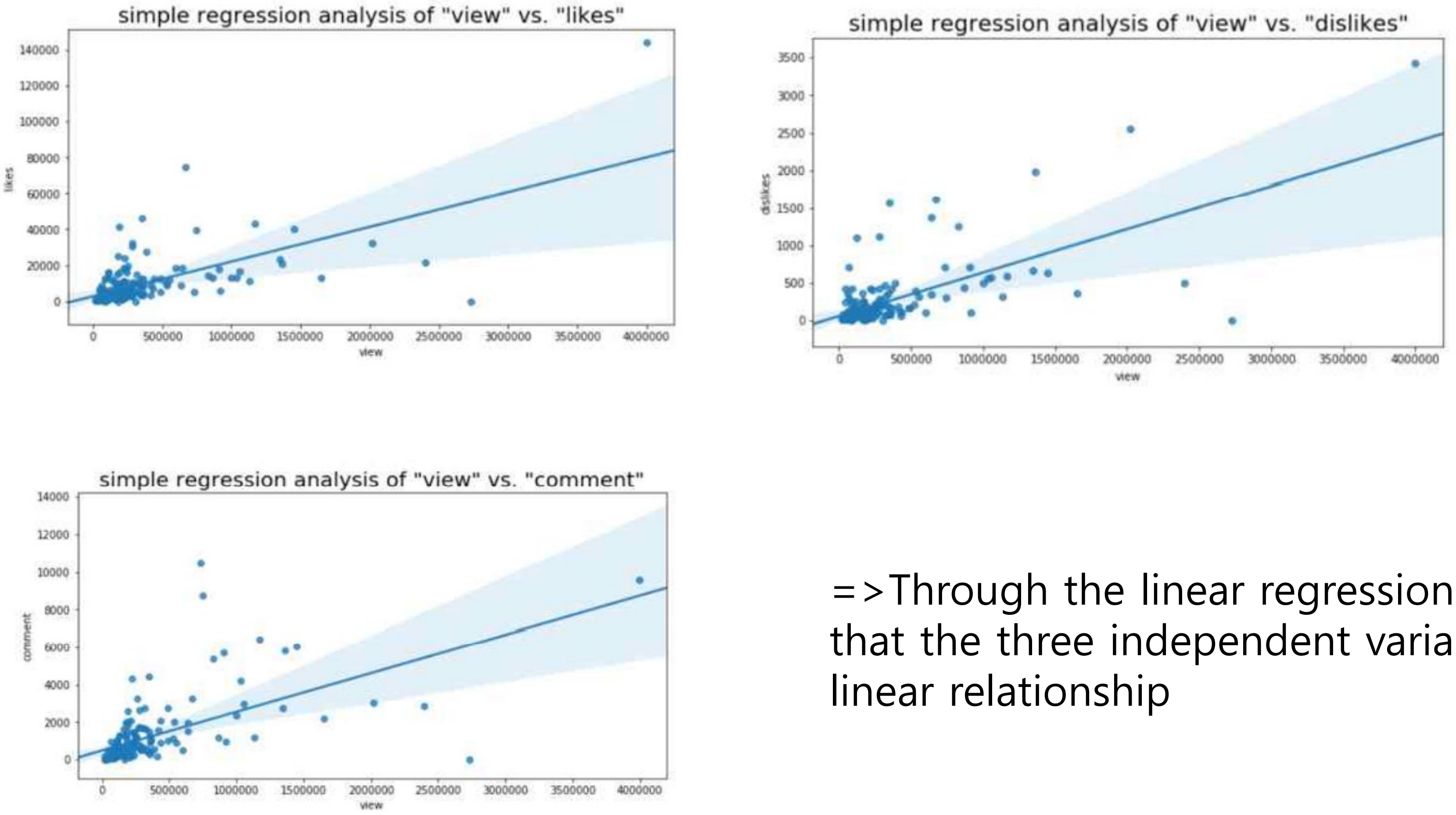
Dependent variable: view

$$Y = a_0 + a_1 X_1 + a_2 X_2 \quad \cdots \quad a_n X_n \tag{4}$$

$$a = \frac{\sum_{i=1}^{n} (x - mean(x))(y - mean(y))}{\sum_{i=1}^{n} (x - mean(x))^2} \tag{5}$$

4. Quantitative Analysis

Fig. 13. Linear Regression Analysis of Variables



simple regression analysis of "view" vs. "likes"

simple regression analysis of "view" vs. "dislikes"

simple regression analysis of "view" vs. "comment"

=>Through the linear regression line, it can be seen that the three independent variables and view are in a linear relationship

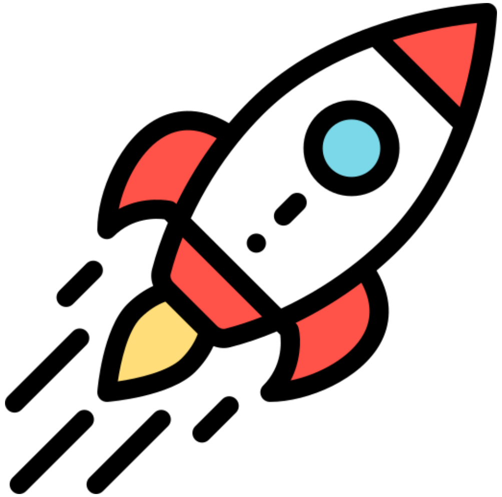
5. Expected Effect

=>Using data mining results based on popular YouTube video content, it is expected that it can be used as training data by connecting to deep learning programming based on supervised learning

=>By implementing artificial intelligence services for YouTube creators, research can be expanded to increase the value of YouTubers' personal brand and help them predict and generate profits

Analysis of Utilization and Assessment of Predicting Models for YouTube Science Channel

1. Problem



In modern science and technology society,
the public needs scientific knowledge,
but such education is insufficient.

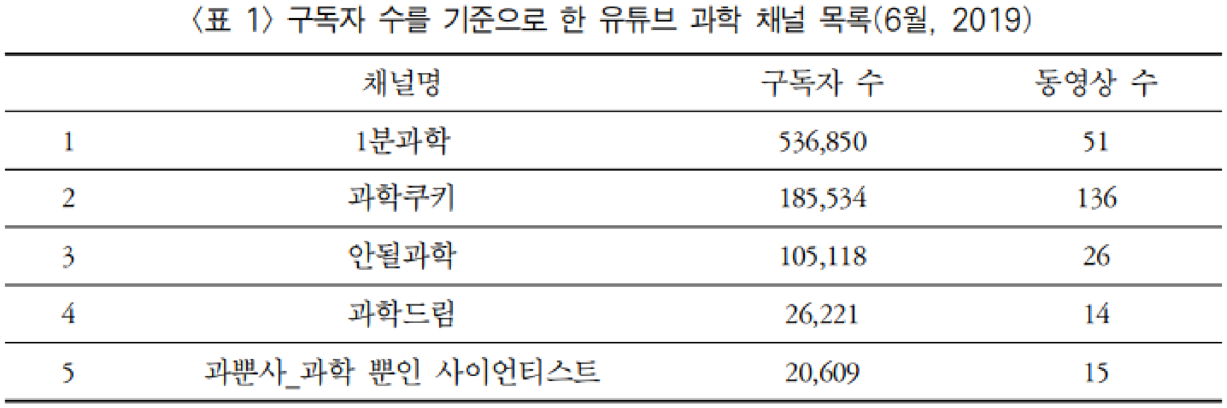
YouTube is drawing attention as a path of
science education outside of school.

1. Problem

Purpose

- To find out the public's interest, usage, and characteristics of the YouTube science channel
- Obtain viewing statistics for each channel and analyze the accuracy of the analysis model

2. Method



〈표 1〉 구독자 수를 기준으로 한 유튜브 과학 채널 목록(6월, 2019)

	채널명	구독자 수	동영상 수
1	1분과학	536,850	51
2	과학쿠키	185,534	136
3	안될과학	105,118	26
4	과학드림	26,221	14
5	과뿐사_과학 뿐인 사이언티스트	20,609	15

Dataset

- A channel suitable for the size of data analysis and machine learning was selected.
- Select two channels with more subscribers(over 100,000) and videos(over 50) > "1 Minute Science" and "Science Cookies."

2. Method

Comparative Analysis of Channel Discriminant Prediction Models

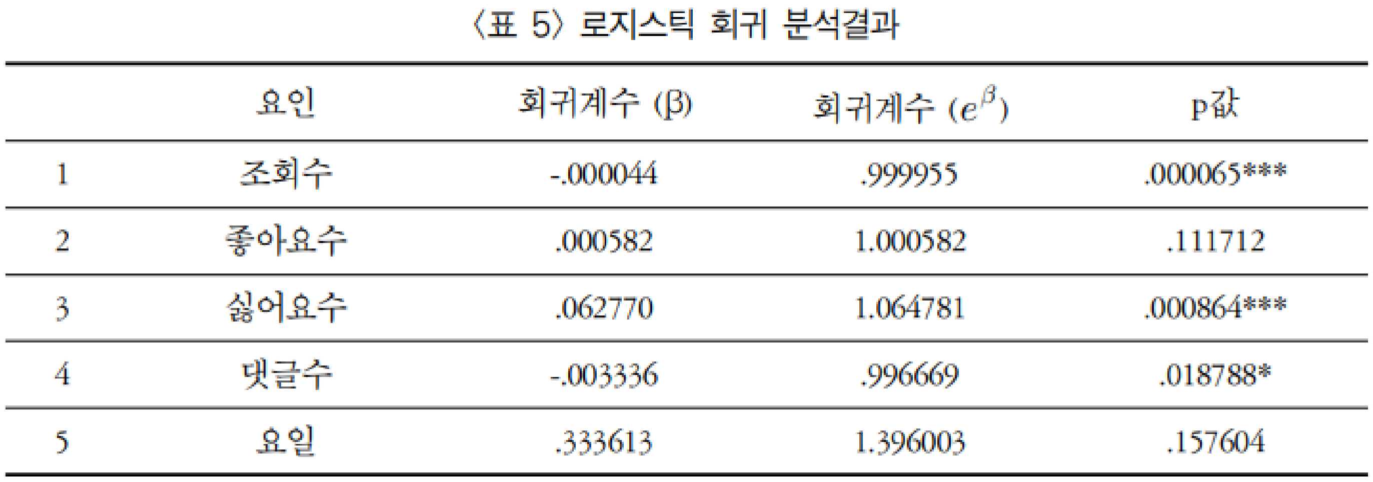
Case1) Logistic Regression/Decision Tree/Random Forests

Case2) knn(k-nearest-neighbors)/SVM(support vector machine)

Case3) Artificial Neural Network

3. Result

Case1) Logistic Regression/Decision Tree/Random Forests

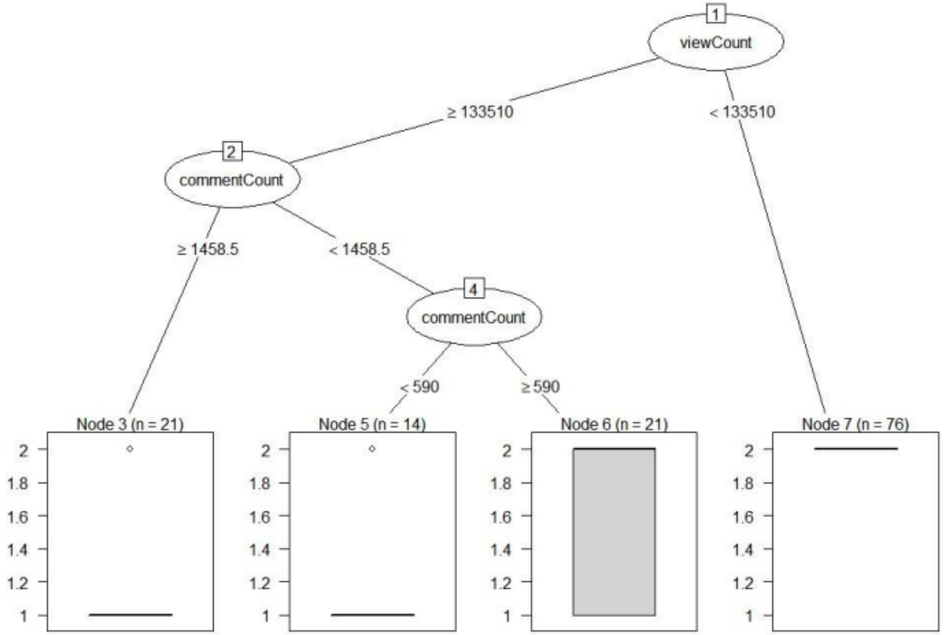


〈표 5〉 로지스틱 회귀 분석결과

	요인	회귀계수 (β)	회귀계수 (e^{β})	p값
1	조회수	-.000044	.999955	.000065***
2	좋아요수	.000582	1.000582	.111712
3	싫어요수	.062770	1.064781	.000864***
4	댓글수	-.003336	.996669	.018788*
5	요일	.333613	1.396003	.157604

3. Result

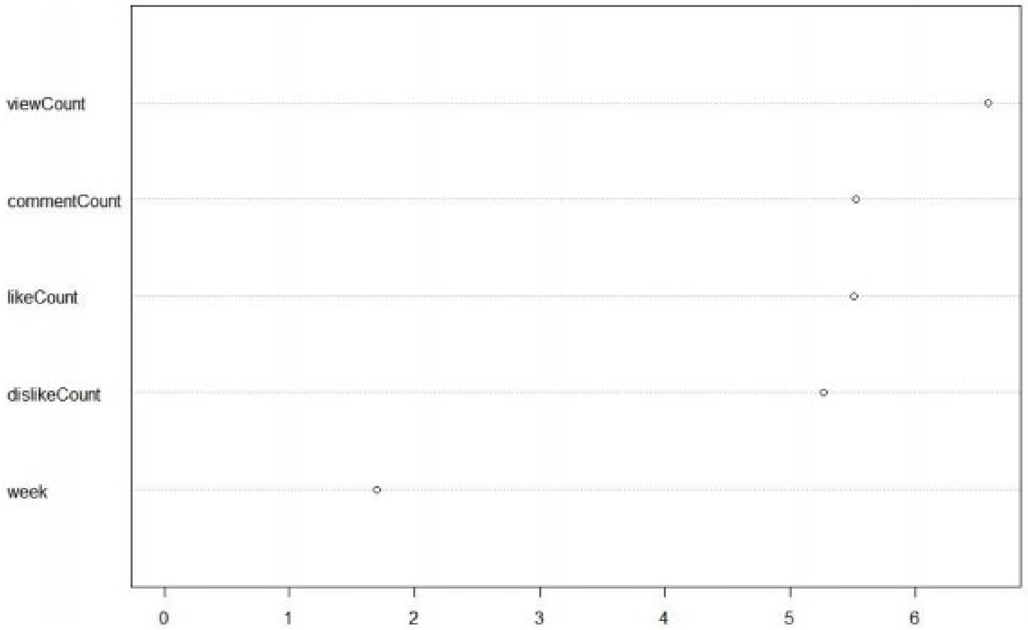
Case1) Logistic Regression/Decision Tree/Random Forests



〔그림 5〕 채널 판별 예측에 대한 의사결정트리 분석

3. Result

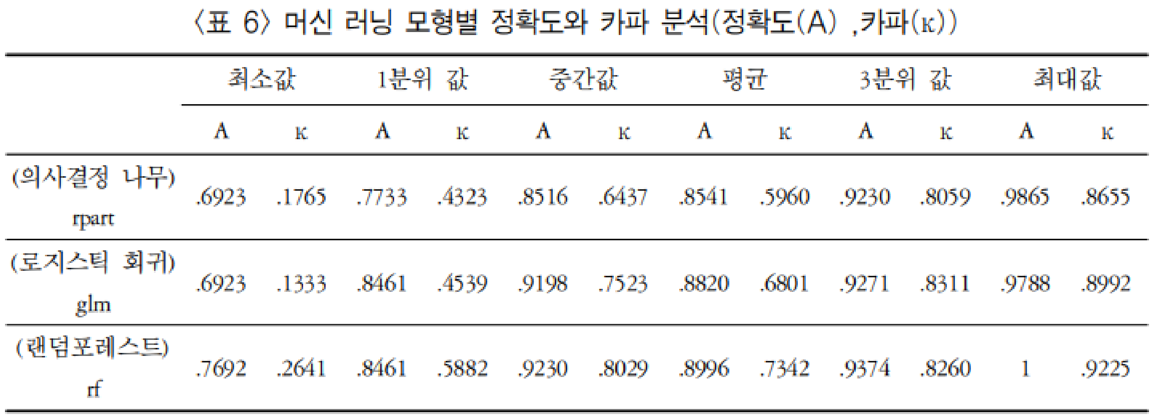
Case1) Logistic Regression/Decision Tree/Random Forests



〔그림 6〕 랜덤포레스트 분석의 중요요인 분석

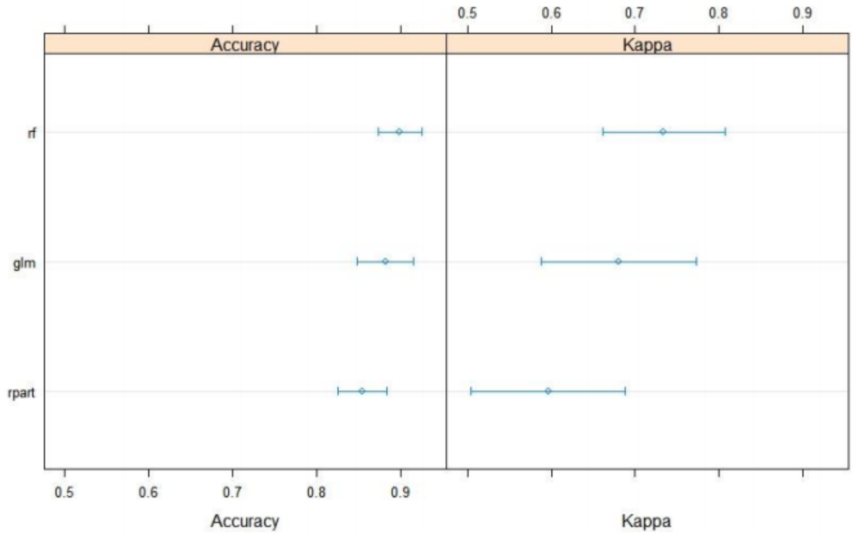
3. Result

Comparison of the previous three models



〈표 6〉 머신 러닝 모형별 정확도와 카파 분석(정확도(A) ,카파(κ))

	최소값		1분위 값		중간값		평균		3분위 값		최대값	
	A	κ	A	κ	A	κ	A	κ	A	κ	A	κ
(의사결정 나무) rpart	.6923	.1765	.7733	.4323	.8516	.6437	.8541	.5960	.9230	.8059	.9865	.8655
(로지스틱 회귀) glm	.6923	.1333	.8461	.4539	.9198	.7523	.8820	.6801	.9271	.8311	.9788	.8992
(랜덤포레스트) rf	.7692	.2641	.8461	.5882	.9230	.8029	.8996	.7342	.9374	.8260	1	.9225

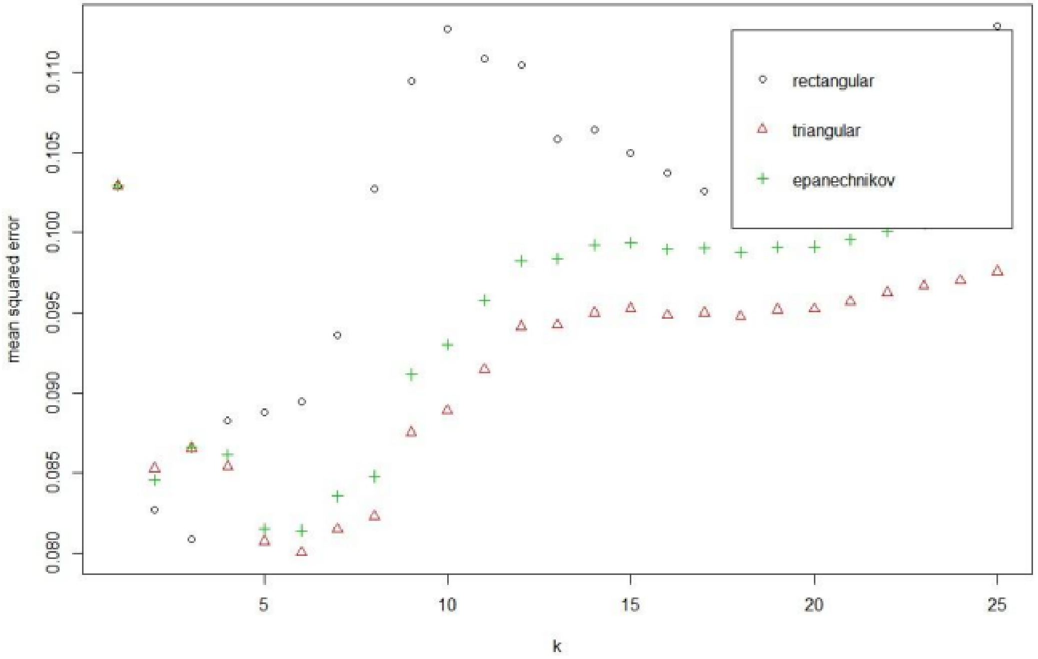


〔그림 7〕 머신 러닝 모형별 비교

Based on the accuracy value and the kappa value, the YouTube channel discrimination prediction was better in the order of random forest(rf), logistic regression(glm), and decision tree model(rpart).

3. Result

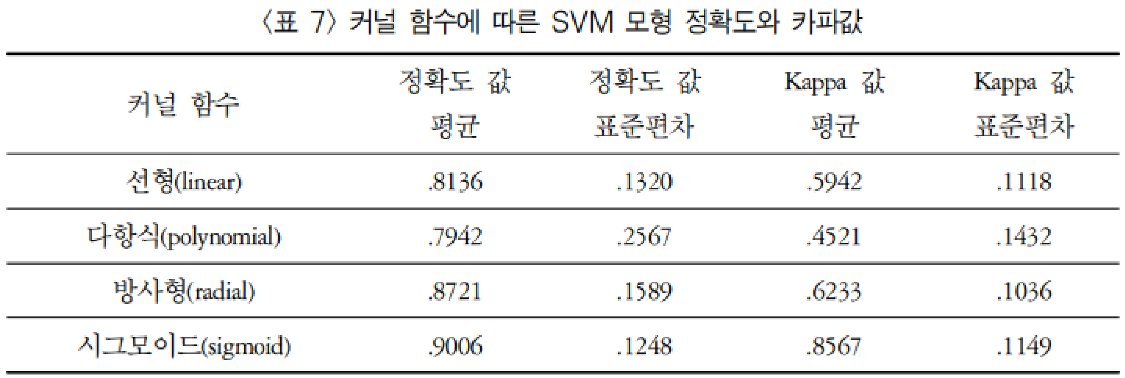
Case2) knn(k-nearest neighbors)/SVM(support vector machine)



〔그림 8〕 가중치 조건 및 k값에 따른 오류율

3. Result

Case2) knn(k-nearest neighbors)/SVM(support vector machine)



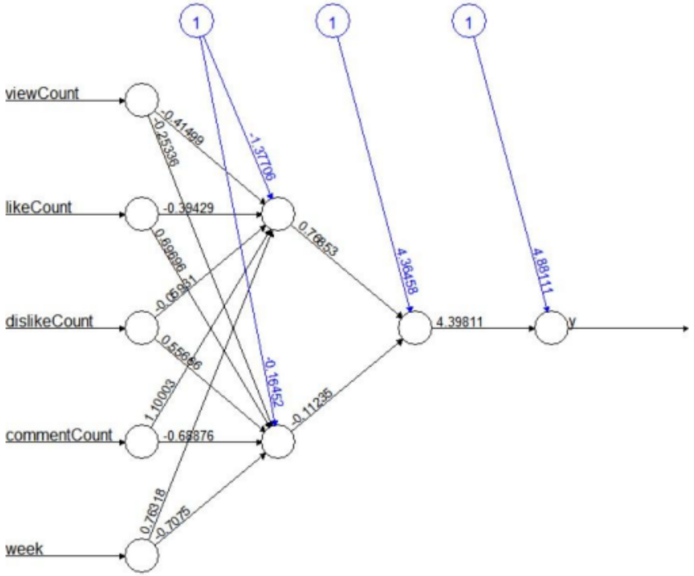
〈표 7〉 커널 함수에 따른 SVM 모형 정확도와 카파값

커널 함수	정확도 값 평균	정확도 값 표준편차	Kappa 값 평균	Kappa 값 표준편차
선형(linear)	.8136	.1320	.5942	.1118
다항식(polynomial)	.7942	.2567	.4521	.1432
방사형(radial)	.8721	.1589	.6233	.1036
시그모이드(sigmoid)	.9006	.1248	.8567	.1149

3. Result

Case3) Artificial Neural Network

유튜브 과학 채널에 대한 이용실태 분석 및 채널 판별 예측 모형 평가-소셜 빅데이터 분석 및 머신 러닝 활용을 중심으



〔그림 9〕 인공신경망 분석 결과(로지스틱 활성 함수 사용)

3. Result

Total Prediction Model Analysis Results

- SVM's sigmoid kernel function is 90.06% accurate, making it the most powerful model
- The random forest model and logistic regression analysis also have high accuracy of 89.96% and 88.20%, respectively
- The accuracy of decision tree model and knn analysis is lower compared to previous models
- The artificial neural network model did not improve the performance

4. Research Evaluation

Significance

- Big data analysis and machine learning should be used to quickly identify public interests and educate them in areas where interest is increasing
- Limited data alone produced more than .80 accurate and reliable results

Limitation

- Since a specific channel was studied, it is limited to say that it reflected all levels of indicators occurring on YouTube
- Possible variables other than machine learning could not be considered

Reference

- Hye-Suk Kim (2020), Analysis of Popular YouTube Video Content using Data Mining, Journal of Digital Contents Society Vol. 21, No. 4, pp. 673-681, Apr. 2020

- Hyunguk Kim (2020), Analysis of Utilization and Assessment of Predicting Models for YouTube Science Channel - Focusing on using Social big data analysis and Machine learning, Journal of Educational Technology 2020, Vol 36, No 2, pp.383-412

- Kim, Jongho, Kim, Kihan.(2021).Determinants of Consumption for YouTube Sport Contents: A Big Data Analysis using Multi-level Regression.Korean Journal of Sport Management,26(3),28-51.