Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 5 Iss. 3 July 2025, pp: 990-999
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Harnessing Machine Learning to Decode YouTube Subscriber Dynamics: Regression Predictive Models and Correlations

Sri Mulyati[1*], Samidi[2]

[1]Master of Science in Management Program, Faculty of Economics and Business, Padjajaran University, Bandung, Indonesia
[2]Master of Computer Science, Budiluhur University, Jakarta, Indonesia

Email: [1]sri23020@mail.unpad.ac.id, [2]samidi@budiluhur.ac.id

Received May 04th 2025; Revised Jun 16th 2025; Accepted Jul 21th 2025; Available Online Jul 31th 2025, Published Jul 31th 2025
Corresponding Author: Sri Mulyati

Abstract

YouTube has grown and become a digital media giant. Content creators continue to struggle with predicting subscriber growth. Due to viewers' changing interests and the vast amount of information, it is challenging to determine which factors most influence subscription behavior. Optimizing content strategy and ensuring channel growth need an understanding of these traits. This study uses linear regression models (LR), neural networks (NN), and Gaussian processes (GP) to predict YouTube subscribers and examine category correlations using video data from various topics. The study of correlation matrix analysis was performed with an absolute root mean square error (RMSE) of 26256351, and the NN prediction model outperformed the LR and GP models. The correlation matrix indicates a slight positive correlation of 0.067 among the YouTube categories. Specifically, the correlation coefficients for population, unemployment rate, and urban population are 0.080, -0.012, and 0.082, respectively. These findings suggest future research to create more intentional content and search for significant factors that increase viewership and marketing audience growth.

Keywords: Machine Learning, Networks, Regression, Subscribers, YouTube

1. INTRODUCTION

YouTube has grown since 2005 and began to dominate the digital world. YouTube has become one of the most popular social media networks, with over 2 billion monthly users by 2023 estimated. Demographically, 2 million people as users are aged 25-34, 7 million are 35-44 years old, and 377 million are between 18-24 years of age [1]. The massive number of users affected the world economy through contributions from content development, digital content, distribution, and business digital advertising turnover, which is predicted to achieve $518.4 billion in the year 2023 [2]. In addition, YouTube has encouraged many people to engage, from bloggers, broadcasters, artists, celebrities, musicians, and event beginners from remote areas, intentionally to create content and find their fortune by monetizing their talents through digital content. McKinsey stated that YouTube videos are estimated to generate income and be sustainable if they grab their customers' engagement [3].

Even though having numerous subscribers has many benefits for creators, many still fail to predict their subscriber growth. Due to the volume of content and changing viewer preferences, identifying the main factors affecting subscriber behavior is tricky. Understanding and perceiving these factors and optimizing content strategy is essential for growth. The digital content ecosystem is such a competitive area, with artists competing for audience attention [4]. The effect of genre on subscribers is significant to research because some genres attract viewers from different demographics. In contrast, others attract subscribers because of their broad appeal to specific interest groups. Deloitte found that content classifications affect audience engagement and loyalty, emphasizing the significance of strategic content planning for channel effectiveness [5].

The primary objective of this study is to develop and compare predictive models for estimating YouTube subscriber growth using three distinct machine learning algorithms: Linear Regression (LR), Neural Networks (NN), and Gaussian Processes (GP). In addition to model development, the research investigates the relationships between categorical YouTube content genres and key demographic indicators, including population size, urban population density, and unemployment rate, about subscriber counts.

Furthermore, the study aims to evaluate the predictive performance of each model through Root Mean Square Error (RMSE) metrics and correlation coefficients, thereby offering insights into their implications for strategic content planning and audience segmentation within digital media platforms.

The related works of Rui et al. [19] previously used Ordinary Least Squares (OLS) and Online Gradient Descent (OGD) models [6]. Additionally, Prachi et al. [7]) used general linear models and LR to analyze YouTube videos [7]. Unlike previous research that focused solely on video-level metrics like duration, likes, or comments [6], this study included additional socio-demographic variables (population, urbanization, and unemployment rate) at the regional to video level in the predictive context. Also, the study examined and compared the performance of three different machine learning models LR, NN, and GP on the same dataset - and thus offered a contextual discussion of their performance and interpretability. The use of socio-demographic, quantitative (video), and qualitative (topic) data in different formats allows for a more nuanced representation of subscribers and changes of subscribers across categories of content as research novelty (see Tabel 1).

Table 1. Research gaps and hypothesis

References	Variable	Type of Data	Regression Models	Result
Rui et al (2019)	Video duration, upload date, and number of likes and comments.	YouTube video statistics	OLS method, OGD method	1. The OLS method outperforms the OGD method in predicting YouTube views 2. Video duration and number of likes are significant predictors of views [6]
Prachi et al (2024)	Video duration, upload date, and number of likes and comments.	Dataset YouTube Statistic	General linear models, LR models	1. Generalized linear models outperform LR models 2. Video duration and number of likes are significant predictors of views. [7]
Our research	Categories, Subscribers, Population, Urban-population, Unemployment	Dataset YouTube Statistic	LR models, NN, and GP Correlation Matrix	Hypothesis: 1. NN outperformed LR and GP 2. Positive Correlation between YouTube Categories, Population, Urban Population, and Unemployment Rate to YouTube Subscribers

The research questions are:

RQ1 : How does the predictive accuracy of NN compare to LR and GP in modeling YouTube subscriber growth?

RQ2 : What is the strength and significance of the correlation between YouTube categories, population, urban population, and unemployment rate to YouTube subscribers?

2. METHODOLOGY

2.1. Method

This study employs machine learning techniques to predict the number of subscribers on YouTube. It investigates links between video categories using a regression model to enhance the accuracy and interpretability of view prediction as to research objectives. The tools utilized three algorithm models: LR, NN, and GP for prediction, as well as correlation matrix models, to effectively analyze YouTube subscriber growth. The selection of LR, NN, and GP models was guided by their complementary strengths in predictive analytics because LR offers interpretability and baseline performance for linear relationships. A NN captures complex, nonlinear patterns and interactions among variables. GP provide probabilistic predictions with uncertainty quantification, which is suitable for modeling noisy and high-dimensional data. These models were chosen to assess both predictive accuracy and explanatory power. Their performance was benchmarked using RMSE, allowing for a robust comparison across modeling paradigms.

The cross-industry data mining technique is a popular data science method. According to surveys, this method is used 49% of the time in data science initiatives, followed by Scrum and Kanban [8]. Furthermore, most of the CRISP-DM research in data science centers around data mining, artificial intelligence, machine learning, and deep learning. The fields of big data, data analysis, and data analytics exhibit distinct differences. Modeling is needed for data mining, AI, machine learning, and deep learning [9]. Thus, data science projects can use CRISP-DM as a research method and process model. The CRISP-DM methodology is applicable even in highly specialized domains such as deep learning and made possible by utilizing deep learning in artificial intelligence and machine learning. The process model begins with business knowledge, then data understanding, preparation, modeling, evaluation, and deployment [10]. Start with data preparation. Data analysis requires data cleaning and preparation, including standardization and reduction. Maintaining data accuracy, completeness, and consistency (see Figure 1).

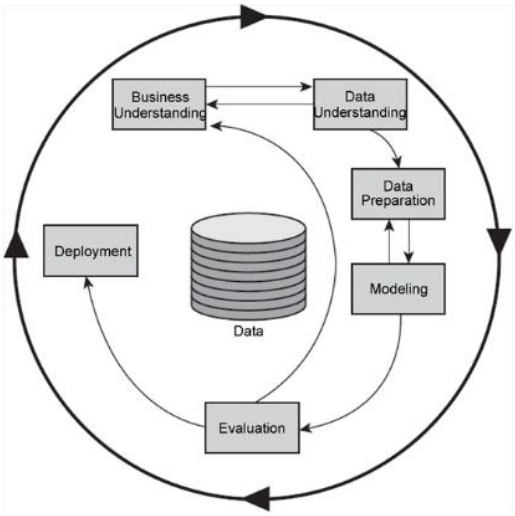

Figure 1. Research Methodology

The initial step, business understanding, was to establish the core research question, which is the difficulty for content creators to forecast subscriber increase due to the complex interrelationship between content categories and socio-demographic factors. This step established the study objective: developing forecasting models to inform strategic content planning and audience targeting at the data understanding phase. The dataset contained variables for YouTube content category, subscriber population, total population, population living in urban areas, and percentage of unemployment. Data preparation involved transforming categorical data into numerical values (18 categories), scaling continuous data, and splitting the dataset into training (90%) and testing (10%) sets. These actions prepared data for machine learning and minimized the chances of bias during model assessment. Three machine-learning methods were executed at the modeling level: LR, NN, and GP. The evaluation process employed RMSE as the primary metric of predictive validity. The deployment phase was conceptualized in actual application terms.

2.2. Datasets and Tools

This research used descriptive quantitative data and sourced data from Kaggle's machine-learning datasets. Kaggle's YouTube video report 2023 is the dataset used for many machine learning projects (https://www.kaggle.com/datasets/nelgiriyewithana/global-youtube-statistics-2023). The form datasets in CSV file datasets include categories, subscribers, unemployment rate, country population, and urban population. There is a total of 836 distinct data points. Of the datasets, 90% (752 data points) were used for training and 10% (84 data points) for testing. The tool used is RapidMiner 10.3, a machine-learning application that can improve the results of data analysis and machine learning and is reliable for various performance analyses of AI projects [10]. The program offers advanced data processing, analysis, and integration operators. Numerous methods for managing missing values and normalizing data ensure that data is error-free and ready for modeling. The variables independent factors are the YouTube category (X1), consisting of 18 genre categories, population (X2), unemployment (X3), and urban population (X4), whereas the dependent variable is the number of subscribers (Y) (see Table 2).

Table 2. YouTube Data

No	Youtube Category (X1)	Population (X2)	Unemployment (X3)	Urban Population (X4)	Subscriber (Y)
1	Autos & Vehicles	212559417	12.08	183241641	21600000
2	Comedy	212559417	12.08	183241641	44200000
3	Education	270203917	4.69	151509724	12900000
4	Entertainment	270203917	4.69	151509724	25000000
5	Film & Animation	212559417	12.08	183241641	14200000
6	Gaming	270203917	4.69	151509724	20200000
7	How to & Style	212559417	12.08	183241641	14500000
8	Movies	1366417754	5.36	471031528	28400000
9	Music	212559417	12.08	183241641	66500000
10	News & Politics	270203917	4.69	151509724	15000000
11	Nonprofits & Activism	328239523	14.7	270663028	38600000
12	People & Blogs	1366417754	5.36	471031528	14400000
13	Pets & Animals	328239523	14.7	270663028	23700000
14	Sports	212559417	12.08	183241641	12300000

No	Youtube Category (X1)	Population (X2)	Unemployment (X3)	Urban Population (X4)	Subscriber (Y)
15	Science & Technology	83132799	3.04	64324835	19800000
16	Shows	1366417754	5.36	471031528	70500000
17	Trailers	1366417754	5.36	471031528	36600000
18	Travel & Events	126014024	3.42	102626859	12500000
…	…..	….	….	…..	…..
836	People & Blogs	126014024	3.42	102626859	12400000

3. RESULTS AND DISCUSSION

3.1. Data Processing

The analysis of YouTube data using LR, NN, and GP as the predictive model and correlation model (see Figure 2 and Figure 3).

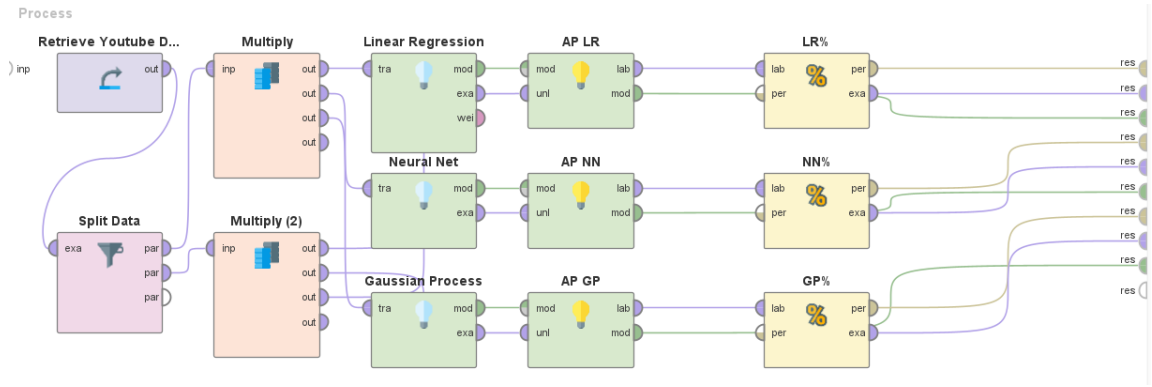

Figure 2. Predictive Model

Figure 2 above presents a comprehensive workflow diagram constructed in RapidMiner which illustrates the procedural architecture of a predictive modeling framework applied to YouTube analytics. The model integrated multiple machine learning techniques such as LR, NN, dan GP to evaluate and compare predictive performance across different algorithmic paradigms.

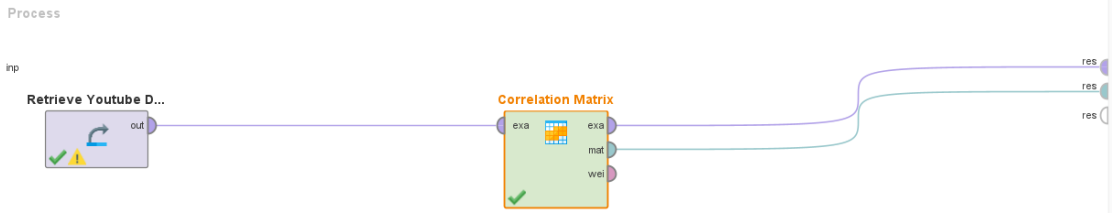

Figure 3. Correlation Model

Figure 3 illustrates a data processing workflow. It is designed to analyze YouTube related data through correlation analysis. It accessed a predefined dataset containing YouTube metrics, such as YouTube category, population, urban population, and unemployment rate.

In data processing, the prior steps address missing values and convert category variables to numbers of reproduced data. Then, 90% of the data is separated into training and 10% testing sets. First, LR, NN, and GP are employed to assess YouTube category, population, urban population, and unemployment rate. Further, these traits are classified and linked. Next, the regression model is trained with the training data and evaluated using RMSE rate comparative and correlation matrix to find the optimum regression model.

3.2. RMSE Comparatives

Experiments measure RMSE and Coefficient of Determination. RMSE is a square root of the average difference between predicted and forecasted values (Chicco et al., 2021). The formula for RMSE is:

$$\boldsymbol{RMSE} = \left[\sqrt{\frac{1}{m}\sum_{t=1}^{m} = \lfloor yi - Xi \rfloor^2}\right] \quad (1)$$

The RMSE result sees Tabel 3.

Table 3. RMSE Result

Algorithm	RMSE	Value
LR	26340816.061+/-0.000	
NN	26256351.509+/-0.000	NN<LR<GP
GP	27048822.217+/-0.000	

After comparing RMSE result of LR, NN, and GP for effectiveness, NN outperforms LR and GP models in RMSE, NN captures complex variable correlations, improving predictions with an absolute RMSE value of 26256351.509, NN model accuracy varies, with significant underestimates. The weak positive correlation between YouTube categories and subscriber counts suggests that category alone cannot predict subscriber counts. Additional variables and advanced models could improve forecasts and help researchers understand YouTube subscription characteristics.

3.3. Output Analysis

The tables and figures below display the results of model projections of YouTube subscriber numbers prediction using LR, NN, and GP. A careful investigation of these variables can reveal model correctness and factor relationships using a correlation matrix.

3.3.1 Linear Regression (LR)

LR is a statistical technique for making quantitative predictions. Fitting a LR to observed data entails establishing the correlation between a variable and one or more explanatory factors[11]. RapidMiner's operator creates a LR model from an example set. This approach works with numeric data. Hence, before applying this operator, the nominal to numerical operator should be used and prepared. The output port yields the regression model, and finally, this model can be applied to new datasets. The result of LR see Tabel 4 and Figure 4.

Table 4. Example set: 84 examples, 1 special attributes (subscriber), 4 regular attributes (YouTube category; population; unemployment rate; urban population)

Row No	Subscribers	Predictions (Subscribers)	YouTube Category	Population	Unemployment Rate	Urban Population
1	245000000	25216135.018	9	1366417754	5.38	471031528
2	80100000	21678482.652	7	66834405	3.85	55908316
3	64600000	27205292.846	15	1366417754	5.36	471031528
4	38200000	24180292.851	9	328239523	14.7	270663028

Subscribers and predictions, for example sets resulted in the YouTube category (9) being music "T-Series" from India; subscribers' prediction is 25.216.135. YouTube category (7) how to & style, "5-Minute Crafts" from the United Kingdom, subscribers' prediction is 21.678.482. YouTube category (15) shows "Color TV" from India, and subscribers' prediction is 27.205.292. YouTube category (9) music "One-Direction," from The United States, subscribers' prediction is 24.180.292. There are 16 categories, each represented by a unique number.

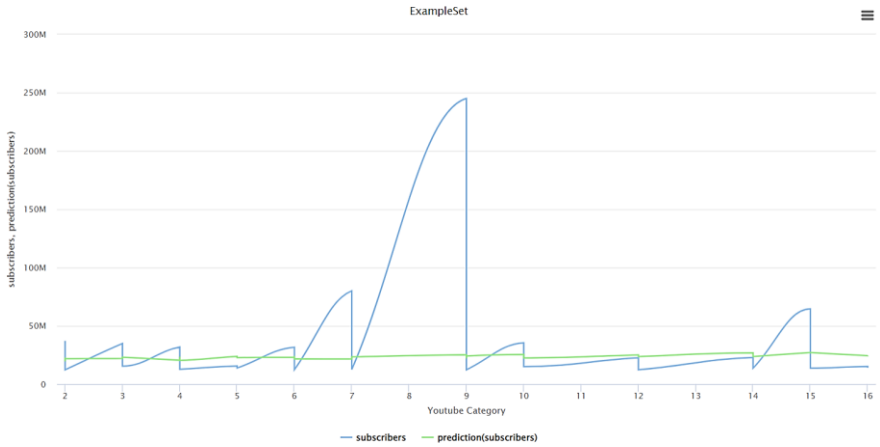

Figure 4. Diagram LR Prediction

3.3.2 Neural Network (NN)

NN are powerful and flexible in computer science and machine learning. They have altered how complicated questions are solved and how big data is consumed [12]. NN in business applications is used to

search for tendencies, predict, clarify, and often forecast. An important component of data science and business analytics is known as NN computing [18**].** The NN result see Figure 5 and Tabel 5.

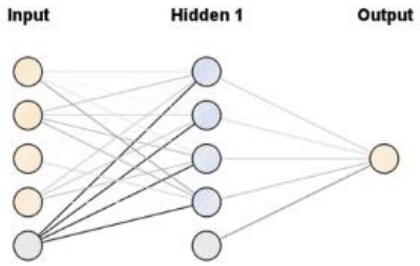

Figure 5. Improved NN

Table 5. Example set: 84 examples, 1 special attribute, 4 regular attributes

Regression (Linear)	Weight
Node 1 (YouTube Category)	-0.317
Node 2 (Population)	-0.250
Node 3 (Unemployment)	-0.376
Node 4 (Urban Population)	0.443
Threshold	-0.799

These weights Tabel 5 represent the linear influence of each input feature on the output prediction. The negative weight (Node 1-3) suggests that increases in the YouTube category index, population, and unemployment are associated with decreased predicted subscriber count. The positive weight (Node 4) of the urban population indicated that a higher urban population contributes positively to subscriber growth. The threshold acts as a bias term, shifting the activation function to calibrate the output. Statistically, the value of the NN model is that node 1 has a negative weight, affecting the forecast. If node 1 (YouTube category) outputs more, the final prediction will reduce by 0.317 units while keeping the other nodes the same. Node 2 (population) has a negative weight and destroys the prediction. Every increase in Node 2's (population) output decreases the final forecast by 0.250 units while maintaining the other nodes the same. The forecast is most negatively affected by Node 3 (unemployment), which has the highest negative weight. A negative connection between node 3's (unemployment) output and the final forecast affected a 0.376 decline. This effect persists while other nodes remain constant. Node 4 has a positive weight, affecting the prediction. Each 0.443 unit increase in node 4's (urban population) output improves the final forecast. Even with unchanged node values, this impact persists.

The threshold forecast decreases by 0.799 units at -0.799. Equation Y is the NN regression model result (see Tabel 6):

$$Y=(-0.317 \times Node1)+(-0.250 X Node2)+ (-0.376x Node3)+(0.443xNode4) -0.799$$

Table 6. NN performance

Row No	Subscribers (Y)	Predictions (Subscribers)	YouTube Category (X1)	Population (X2)	Unemployment Rate (X3)	Urban Population (X4)
1	245000000	24522840.315	9	1366417754	5.36	471031528
2	80100000	29123985.795	7	66834405	3.85	55908316
3	64600000	28790879.972	15	1366417754	5.36	471031528
4	38200000	22207090.839	9	328239523	14.7	270663028

Based on model NN (Figure 6), subscribers' predictions for YouTube category (9) are music "T-Series" from India, and subscribers' prediction is 24.522.840. YouTube category (7) how to & style, "5-Minute Crafts" from the United Kingdom, subscribers' prediction is 29.123.985. YouTube category (15) shows "Color TV" from India; subscribers' prediction is 28,790.879. YouTube category (9) music "One-Direction," from the United States, subscribers' prediction is 22.207.090.

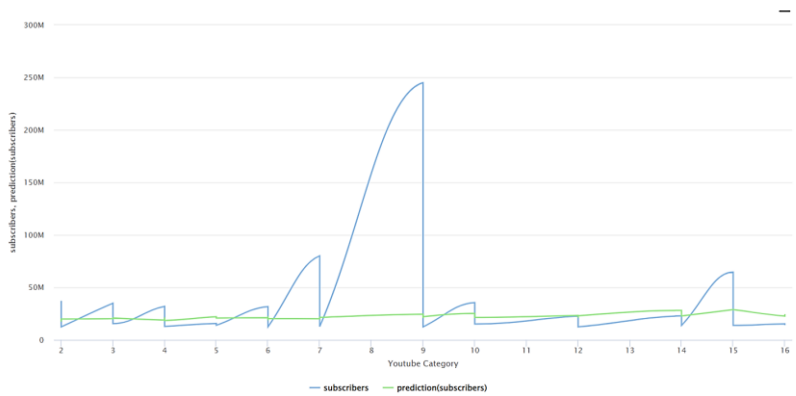

Figure 6. Diagram NN Prediction

3.3.3 Gaussian Process (GP)

The multivariate Gaussian distribution has different properties because it gives the probability density function for any finite number of variables in a GP. Utilizing the GP operator in RapidMiner allows for capturing highly complex input-output dependencies [13]. This piece of software is used to solve regression and classification problems and estimate the level of certainty with the prediction. Output the GP model predictions on new data using the output port. This model provides standard errors to check the reliability of the forecasts. Defining a GP distribution for the variable f:

$$p(f)=GP(f;\mu,K) \tag{2}$$

Based on the GP model and subscribers' predictions result of this study, the YouTube category (9) is music "T-Series" from India, subscribers' prediction is 30.240.631. YouTube category (7) how to & style, "5-Minute Crafts" from the United Kingdom, subscribers' prediction is 23.471.123. YouTube category (15) shows "Color TV" from India; subscribers' prediction is 72,856,686. YouTube category (9) music "One-Direction," from the United States, subscribers' prediction is 23.122.320 (see Table 7):

Table 7. GP performance

Row No	Subscribers (Y)	Predictions (Subscribers)	YouTube Category (X1)	Population (X2)	Unemployment Rate (X3)	Urban Population (X4)
1	245000000	30240631.348	9	1366417754	5.36	471031528
2	80100000	23647123.076	7	66834405	3.85	55908316
3	64600000	72856686.274	15	1366417754	5.36	471031528
4	38200000	23122320.723	9	328239523	14.7	270663028

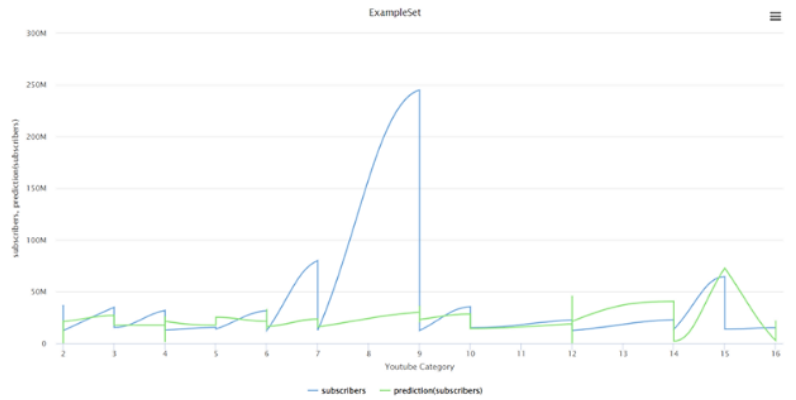

Figure 7. Diagram GP Prediction

Overall, LR, NN, and GP model results predicted subscriber counts differently. Predictions for each sample are calculated to assess model accuracy. LR and NN models accurately represent the "T-Series" music category on YouTube by overestimating its subscribers. On the other hand, they drastically underestimate "how to & style" represented by "5-Minute Craft". YouTube categories (9) music "T-Series and (15) series "Color TV had overstated subscribers in the GP model. However, the subscriber counts for

"How to & Style," notably "5-Minute Crafts," and YouTube category (9) music "One-Direction," was severely underestimated.

3.3.4 Correlation Result

The coefficient of determination, known as the correlation coefficient, measures how well a model fits on a data set. Pearson correlation coefficient formula:

$$r = \frac{\sum(Xi-X)(Yi-Y)}{\sqrt{\sum(Xi-X)^2 \sum(Yi-Y)^2}} \quad (3)$$

Xi and Yi are variables, while X and Y are the average of those variables. A correlation matrix GP measures -1 to 1. 1 represents a perfect positive linear relationship, -1 represents a perfect negative relationship, and 0 depicts no relationship [14]. When it comes to correlation, RapidMiner's user-friendly interface simplifies correlation analysis. It generates and displays correlation matrices quickly. The correlation results please see Tabel 8.

Table 8. Correlation Result

Attribute	Coefficient	St. Error	St. Coefficient	Tolerance	t-Stat	p-Value	Code
YouTube Category	331526.305	17862.846	0.071	0.999	1.940	0.053	*
Population	-0.001	0.003	-0.024	0.644	-0.273	0.785	
Urban Population	0.010	0.010	0.087	0.706	0.986	0.324	
(Intercept)	18878061.367	1754355.810	?	?	10.761	0	****

Multiple LR statistical indicators are shown in Table 8. YouTube subscribers are the dependent variable, while YouTube category, population, and urban population are independent. Important metrics include coefficient, standard error, standardized coefficient, tolerance, t-statistic, and p-value. Each YouTube Category unit increases subscribers by 331,526.305 while maintaining all other factors the same. The YouTube Category may significantly affect subscriber counts because the p-value (0.053) is somewhat higher than 0.05. However, it is a very mild correlation. The negative coefficient implies a slight reduction in subscriber counts as the population grows; it is not statistically significant (p-value = 0.785). Population size has little impact on subscriber numbers. The coefficient shows a little increase in subscribers with urbanization. Nevertheless, the p-value (0.324) implies no statistically meaningful relationship. The intercept is the starting number of subscribers when all independent variables are 0. The significant p-value of 0 indicates that the intercept is far from zero.

The regression study shows that the size of the population and the urban population does not impact the number of subscribers. However, the YouTube category might influence it. The significant intercept indicates a consistent initial subscriber level that remains unaffected by all circumstances. Introducing new variables or employing nonlinear modeling techniques may enhance the accuracy of the model's predictions. The result of correlation coefficient (see Tabel 9).

Table 9. Correlation Coefficient

Attributes	Subscribers (Y)	YouTube Category X1)	Population (X2)	Unemployment Rate (X3)	Urban Population (X4)
Subscribers (Y)	1	0.067	0.080	-0.012	0.082
YouTube Category (X1)	0.067	1	-0.044	-0.013	-0.045
Population (X2)	0.080	-0.044	1	-0.233	0.912
Unemployment Rate (X3)	-0.012	-0.013	-0.233	1	0.131
Urban Population (X4)	0.082	-0.045	0.912	0.131	1

The relationship between subscribers (Y) *and* YouTube category (X1): 0.067 indicates a mild positive connection. These statistics suggest that the YouTube category does not affect subscribers. Population (X2): The correlation is 0.080, indicating a slight positive connection. Population growth may affect subscriber growth. The unemployment rate (X3) has a -0.012-correlation coefficient, indicating a weak negative relationship. Unemployment has little effect on subscribers, urban population (X4): A weak positive association (0.082). Increased urban populations may boost subscriptions slightly. The correlation matrix shows a 0.067 positive correlation between YouTube categories and subscriber counts. This correlation coefficient suggests a weak positive association despite its low value. Despite the weak correlation, the prediction results imply that YouTube subscriber counts may be accurately forecasted for each category.

Ultimately, the hypothesis that NN performance outperforms LR and GP is proved. However, the low correlations between subscribers and other variables show how difficult YouTube subscriber forecasting is, suggesting that adding variables or using more advanced modeling methods may improve forecast accuracy. The strong correlation between the general and urban populations shows urbanization patterns. In contrast, other relationships, such as unemployment rates and population sizes, are intriguing but weak and may require further study.

3.4. Discussion

This report provides a comparative evaluation of predictive models. As demonstrated with an RMSE value of 26256351.509, (NN) was able to predict better YouTube subscriber growth than (LR) and (GP). The NN model outperformed LR and GP models based on the lowest RMSE value, showing that it is better at both nonlinearities and detecting latent behaviors, which are increasingly important when modeling digital engagement. The weak correlations of 0.067 categories, 0.080 population, and 0.082 urban population demonstrate that demographic and social indicators had little explanatory power as moderator variables, as discussed in [15]. Their study demonstrated that emotional tone, linguistic style, and interactivity can better predict levels of digital consumer engagement than static demographic indicators.

Compared with previous findings of Rui et al. [6] and Prachi et al.[7] which emphasized linear and generalized linear models for predicting YouTube engagement based on observable metrics (e.g., video duration, likes). This study demonstrates the superior predictive capability of (NN) in modeling subscriber growth. Still, the low correlation coefficients suggest that demographic variables contribute minimally, reinforcing the primacy of content-driven and behavioral features in digital engagement [15]. The weak significance of the YouTube category ($p = 0.053$) suggests that genre-specific targeting must be matched with assessed content creation strategies. Zhao [16], using entropy-based model techniques and principal component analysis, demonstrates that these types of modeling are superior at portraying the behavior of trending videos than simple categorical models. Pellegrino [17] supports the need for emotive and socially embedded models in discussing the relevance of social comparison and digital identity as psycho-social aspects of consumer choice. In an increasingly digital world, consumers display omnichannel behaviors [18], increasingly complicating the notion of predictive modeling.

In the digital engagement and marketing world, this study's findings highlight that macro-level segmentation is only a partial solution. Behavioral and psychological data must be utilized in addition to it. Ramaswamy and Ozcan [19] argue that value co-creation through personalized, participatory content is central to sustaining viewer loyalty. Furthermore, Pan et al. [20] stated that influencer credibility and perceived authenticity mediate the relationship between content type and engagement outcomes, inferring that genre cannot be solely relied upon to predict subscription behavior.

Future research should include longitudinal perspectives and advanced ensemble methods like XGBoost and Light GBM to align with the volatility and complexity of temporal and behavioral aspects. Additionally, integrating watch time metrics, comment sentiment, and viewer retention can be further developed to facilitate the planning of a predictive model while also supporting a more strategic and ethically orientated approach for practitioners in the era of immersive marketing in the digital landscape.

4. CONCLUSION

Understanding the dynamics of YouTube subscriber growth is critical for content creators, marketers, and platform strategists. By leveraging machine learning models, this research provides actionable insights into which content and demographic variables most influence subscriber behavior. The result of predictive accuracy comparativeness RMSE of RapidMiner shows that NN is better than LR and GP models. The RMSE of the NN model is significantly lesser at 26256351.509, which is much higher than the previous one and shows higher predicted accuracy and model performance. The correlation matrix reveals the result of 0. 067. There is a slightly positive correlation between the YouTube categories and the number of subscribers. The other variables, such as population, are 0.080 and 0.082, and the urban population has a mild positive relationship. The unemployment rate has almost a negative relationship of -0.012 to YouTube subscriber numbers. The YouTube Category may not affect subscriber counts because the p-value (0.053) is somewhat higher than 0.05. This figure points to a relatively weak positive association regarding the correlation coefficient. The dominance of the positive weight for urban populations suggests that urbanization may be a stronger driver of YouTube engagement than other factors. Meanwhile, the negative weights imply that broader population size and unemployment may not directly translate into higher subscriber count, possibly due to limited digital access and engagement in those segments. These findings can inform targeted content development, enhance digital marketing strategies, and contribute to the broader discourse on algorithmic personalization and audience analytics in social media platforms.

REFERENCES

[1] Metamorworks, "Competition Issues concerning News Media and Digital Platforms," 2021. [Online]. Available: https://www.oecd.org/daf/competition/competition-issues-in-news-

[2] B. Auxier, A. Bucaille, K. Westcott, and D. Ortiz, "As seen in your feed: Shopping goes social, trending past US$1 trillion in annual sales," 2023.

[3] R. Peres, M. Schreier, D. A. Schweidel, and A. Sorescu, "The Creator Economy: An Introduction and a Call for Scholarly Research," *International Journal of Research in Marketing.*, 2024, [Online]. Available: https://ssrn.com/abstract=4663506

[4] OECD, "Directorate For Financial and Enterprise Affairs Competition Committee," 2021. [Online]. Available: https://www.oecd.org/daf/competition/competition-issues-in-news-media-and-digital-platforms.htm

[5] Deloitte, "Social Commerce," 2023.

[6] L. T. Rui, Z. A. Afif, R. D. R. Saedudin, A. Mustapha, and N. Razali, "A regression approach for prediction of Youtube views," *Bulletin of Electrical Engineering and Informatics*, vol. 8, no. 4, pp. 1502–1506, Dec. 2019, doi: 10.11591/eei.v8i4.1630.

[7] Prachi, Siddhi, Deepa, and Manju, "Exploring Regression Models for Youtube Views Prediction with Interpretable Insights," in *2024 IEEE 9th International Conference for Convergence in Technology, I2CT 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/I2CT61223.2024.10544146.

[8] J. Abasova, P. Tanuska, and S. Rydzi, "Big data—knowledge discovery in production industry data storages—implementation of best practices," *Applied Sciences (Switzerland)*, vol. 11, no. 16, Aug. 2021, doi: 10.3390/app11167648.

[9] A. Rianti *et al.*, "CRISP-DM: Metodologi Proyek Data Science," in *Seminar Nasional Teknologi Informasi dan Bisnis (SENATIB)*, Universitas Duta Bangsa Surakarta, Jul. 2023.

[10] Ramesh. Sharda, *Analytics, data science, & artificial intelligence*. Pearson, 2020.

[11] D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 1, no. 2, pp. 140–147, Dec. 2020, doi: 10.38094/jastt1457.

[12] V. Trinh, "A Comprehensive Review: Applicability of Deep Neural Networks in Business Decision Making and Market Prediction Investment," Jan. 2025, [Online]. Available: http://arxiv.org/abs/2502.00151

[13] A. Zeng, H. Ho, and Y. Yu, "Prediction of building electricity usage using Gaussian Process Regression," *Journal of Building Engineering*, vol. 28, Mar. 2020, doi: 10.1016/j.jobe.2019.101054.

[14] V. Kotu and B. Deshpande, *Data science: concepts and practice*. Morgan Kaufmann., 2019.

[15] A. C. Munaro, R. Hübner Barcelos, E. C. Francisco Maffezzolli, J. P. Santos Rodrigues, and E. Cabrera Paraiso, "To engage or not engage? The features of video content on YouTube affecting digital consumer engagement," *Journal of Consumer Behaviour*, vol. 20, no. 5, pp. 1336–1352, Sep. 2021, doi: 10.1002/cb.1939.

[16] Y. Zhao, "Visualization and Scoring Models: Trending Videos Discovery and Recommendation based on Information Entropy Method and Principal Component Analysis," in *2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering, ICBAIE 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 451–465. doi: 10.1109/ICBAIE56435.2022.9985885.

[17] A. Pellegrino, *Decoding Digital Consumer Behavior Bridging Theory and Practice*. Springer Nature Singapore Pte Ltd., 2024. doi: https://doi.org/10.1007/978-981-97-3454-2.

[18] P. Rodríguez-Torrico, R. San José Cabezudo, and S. San-Martín, "Building consumer–brand relationships in the channel-mix era. The role of self–brand connection and product involvement," *Journal of Product and Brand Management*, vol. 33, no. 1, pp. 76–90, Jan. 2024, doi: 10.1108/JPBM-10-2022-4181.

[19] V. Ramaswamy and K. Ozcan, "What is co-creation? An interactional creation framework and its implications for value creation," *J Bus Res*, vol. 84, pp. 196–205, Mar. 2018, doi: 10.1016/j.jbusres.2017.11.027.

[20] A. Tatar, P. Antoniadis, M. D. de Amorim, and S. Fdida, "From popularity prediction to ranking online news," *Soc Netw Anal Min*, vol. 4, no. 1, pp. 1–12, Jan. 2014, doi: 10.1007/s13278-014-0174-8.