Forecast Analysis of Gross Regional Domestic Product based on the Linear Regression Algorithm Technique

Veta Lidya Delimah Pasaribu [1], Fauziah Septiani [1], Suharni Rahayu [1], L Lismiatun [1], Muhammad Arief [1], Angga Juanda [1], M. Yusuf Sunaryo [1], Robbi Rahim [2]

[1]*Universitas Pamulang, Tangerang Selatan, Indonesia*
[2]*Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia*

Abstract **– Statistical data are indispensable for macro-economic planning activities such as the Gross Regional Domestic Product (GRDP) where data can determine the economic development strategies and policies that have been adopted and can be continued in the future. This study draws on quantitative data sources from the Regional Statistical Agency of Jakarta for the period 2017-2019, the subject of the Gross Regional Domestic Product based on current business prices. The aim of this research is to test and predict the level of accuracy of GRDP at current prices based on business fields using the Linear Regression method supported by Rapid Miner software. The results show that the validated Linear Regression algorithm with K-Fold values from 2 to 10 with the sampling type linear sampling and shuffled sampling can be used and implemented with the smallest Root Mean Square Error value of IDR 9,977,431 at k = 10 for the sampling.**

Keywords – Data Mining, Linear Regression, Gross Regional Domestic Product, Root Mean Square Error.

DOI: 10.18421/TEM102-17
https://doi.org/10.18421/TEM102-17

Corresponding author: Robbi Rahim,
Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia.
Email: usurobbi85@zoho.com

Received: 05 March 2021.
Revised: 22 April 2021.
Accepted: 28 April 2021.
Published: 27 May 2021.

© 2021 Robbi Rahim at al; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.

The article is published with Open Access at www.temjournal.com

1. Introduction

Local Gross Domestic Product (LGDP) is a measure of the size of the economy of a region. The information on GRDP can be utilized in order to become public policy. There is a need to develop an extraction approach which can inform more knowledge of the regional economy [1]. Another mechanism that could be employed is data mining [2]. Data mining is an extraction of hidden information from data warehouses [3]. There is the technique of estimation that can do forecasting. The linear regression is used for estimating the estimate value [4],[5],[6]. Linear regression may be used to predict future data based on past data [7].

A significant number of studies have been performed on linear regression [8],[9],[10]. One is [11] analyzes to determine clean water specifications. A linear regression algorithm is proposed in this article. From the results of the study, it is possible when the clean water need is 12,883,649,061 M3 for 2018-2021 and the number of users increases by 21005 in 2021, to apply and enforce the linear regression algorithm. Additional studies [8] are on storm activity predictions. In this paper, a numerical model is built using nonlinear models to estimate the tempest behavior via linear regression. The findings have shown that the model performs well to estimate storm activity with a maximum duration of 1000 iterations and with less than 50 percent of errors. In this case, the forecast concentrates on the Gross Regional domestic product based on the advantages of the linear regression process. This study is focused on market information where new knowledge comes into being, in particular in the Gross Regional Domestic Product, based on business processes from case studies to prevision. In addition, the focus of this study is on evaluating Rapid Mines software for the capacity of linear regression in forecasting.

2. Methodology

The technique used in this analysis is the linear regression method[12],[13],[14] used to predict the Gross Regional Domestic Product based on the data set used over the past 3 years. The study steps can see in Figure 1.

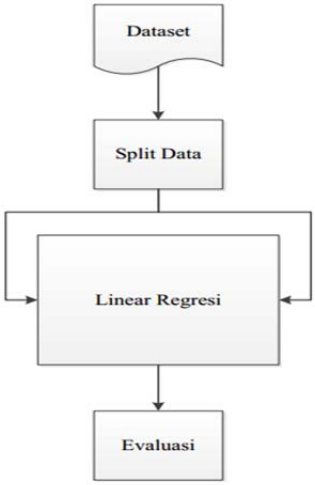


Figure 1. Research steps

The dataset used is the PDRB data based on current business field prices (million rupiah) for the year 2017-2019 for the Jakarta area, which is derived from https://jakarta.bps.go.id, consisting of 17 records with a PDRB attribute based on the current business field price (million rupiah) annually based on the business field. The types of business fields used are agriculture, forestry and fisheries, mining and excavation, processing, electricity and gas procurement, water supply, waste management, waste and recycling, construction, wholesale and retail trade, car and motorcycle storage, transport and storage, housing and food supply, information and communications. Subsequently, we split the data for automatic separation of training data and testing data. In accordance with the predictions in this study, i.e. the use of linear regression is in order when we produce a predictive pattern. The dataset then evaluates the method used. The result is an MRSE (Root Mean Square Root) as in Table 1.

Table 1. PDRB at Current Price by Business Field (Million Rupiah)

Business field	PDRB at Current Price by Business Field (Million Rupiah)		
	2017	2018	2019
Agriculture, Forestry and Fisheries	IDR 2,051,397	IDR 2,125,461	IDR 2,191,304
Mining and excavation	IDR 5,746,496	IDR 6,591,828	IDR 6,187,973
Processing industry	IDR 317,512,261	IDR 341,678,445	IDR 346,988,495
Procurement of electricity and gas	IDR 7,449,251	IDR 9,119,705	IDR 10,336,353
Water supply, waste management, waste and recycling	IDR 838,066	IDR 908,556	IDR 941,031
Construction	IDR 291,528,362	IDR 314,639,322	IDR 329,925,735
Wholesale and retail trade; Car and motorcycle reservation	IDR 394,563,507	IDR 440,082,349	IDR 486,995,838
Transportation and warehousing	IDR 85,669,283	IDR 94,162,613	IDR 104,713,311
Provision of accommodation and food and drink	IDR 116,568,312	IDR 124,334,941	IDR 136,535,154
Information and communication	IDR 180,743,711	IDR 204,640,319	IDR 231,570,995
Financial and insurance services	IDR 251,597,313	IDR 266,918,988	IDR 296,821,467
Real Estate	IDR 145,709,477	IDR 157,341,391	IDR 170,336,213
Company services	IDR 187,465,870	IDR 214,597,420	IDR 254,784,635
Government administration, defense and social security are mandatory	IDR 122,371,391	IDR 143,324,062	IDR 158,939,424
Educational services	IDR 124,037,304	IDR 133,017,522	IDR 141,178,561
Health services and social activities	IDR 39,956,267	IDR 44,191,833	IDR 48,370,935
Other services	IDR 91,545,587	IDR 101,651,451	IDR 114,010,440

Source: Url: https://jakarta.bps.go.id/indicator/52/55/1/pdrb-atas-dasar-harga-berlaku-menurut-lapangan-usaha-.html

3. Results and Discussion

The research discussion is based on the framework and method used namely, linear regression with the use of Rapidminer software. The following is the implementation of a linear dataset regression using Rapidminer software as in Figure 2.

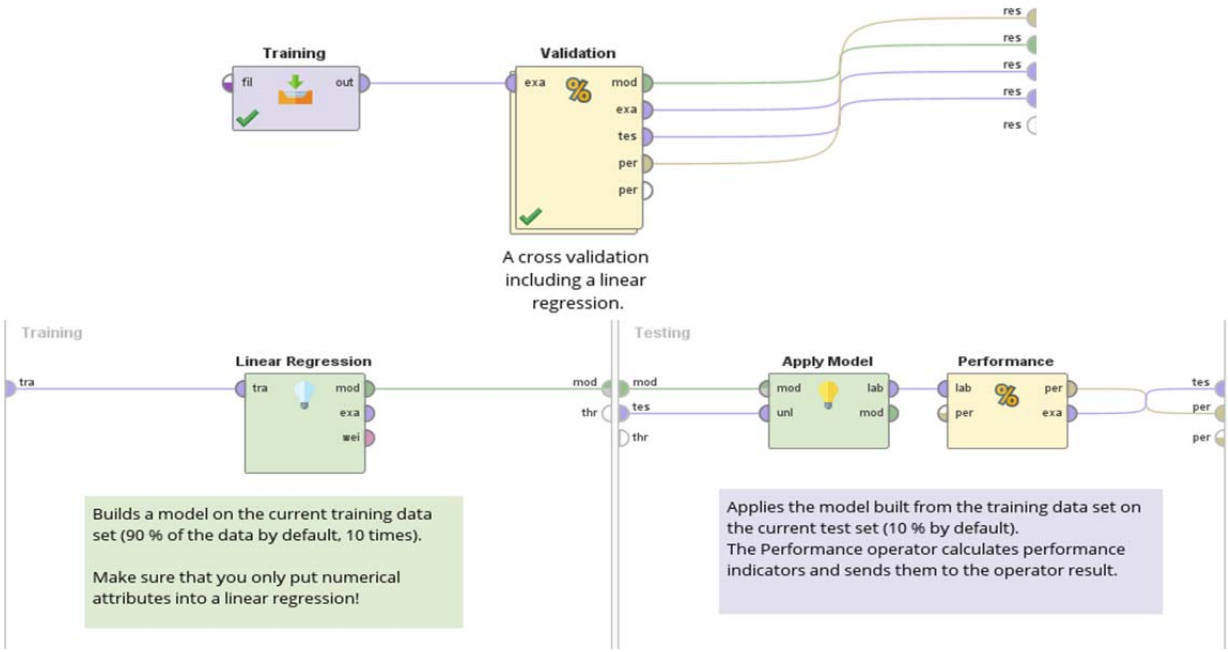


Figure 2. Linear Regression Modelling in Rapidminer

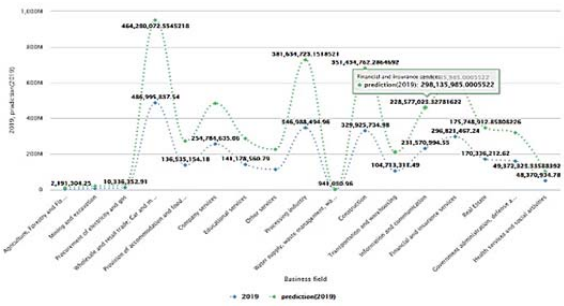
The input method is based on the Read Excel operator from Figure 2 and the data set results can be processed using the Linear Regression Algorithm. The results are given. Apply Model linear regression algorithm, which is useful for data testing, and then connected to results, will be validated by the following X-Validation, which shows how to provide data accurate in root mean square error form (RMSE).

In each experiment, the number of validations is progressively changed by changing the value of the K-Fold = 2 to 10 for the full RMSE (Root Mean Squared Error) effects. The Root Mean Squared Error results for the sampling type = shuffled_sampling are shown in the Table 2 and Figure 3.

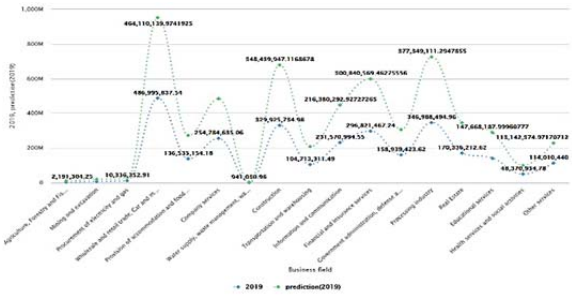
Table 2. RMSE Value for Experiment sampling type = shuffled_sampling

k	RMSE
k=2	IDR 13,220,213
k=3	IDR 13,246,730
k=4	IDR 15,536,813
k=5	IDR 10,631,133
k=6	IDR 11,289,353
k=7	IDR 14,027,466
k=8	IDR 13,460,901
k=9	IDR 14,098,398
k=10	IDR 12,291,283

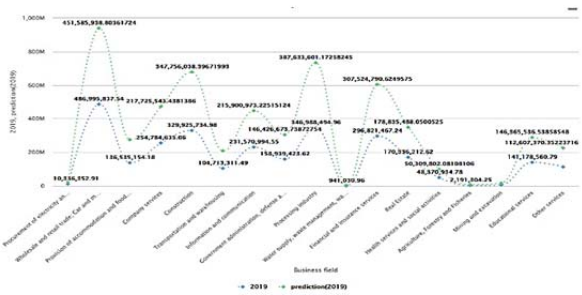
The following is a graphic image of each K-Fold value at sampling type = shuffled_sampling



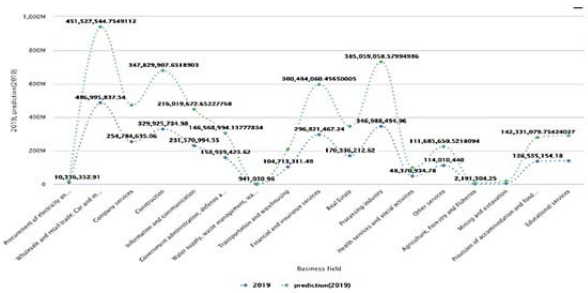
K=2



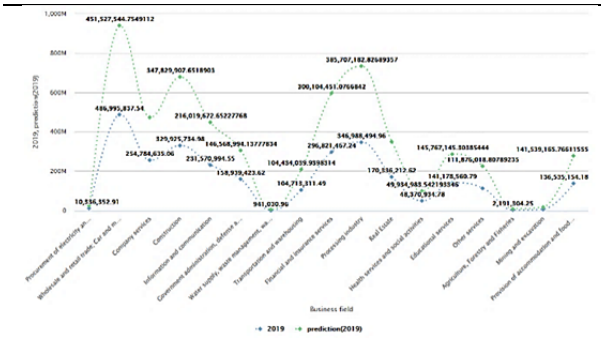
K=3



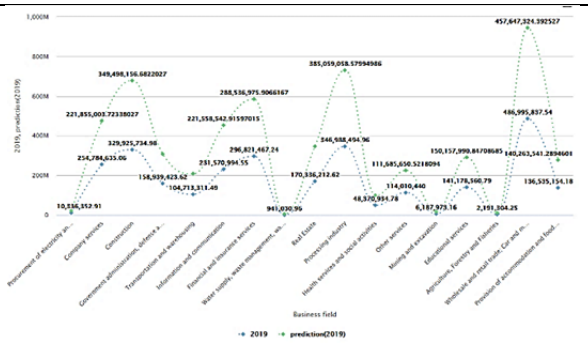
K=4



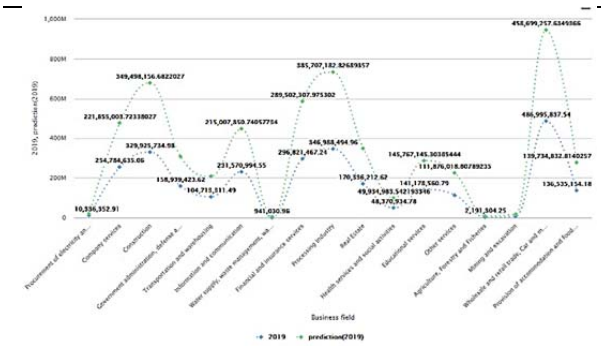
K=5



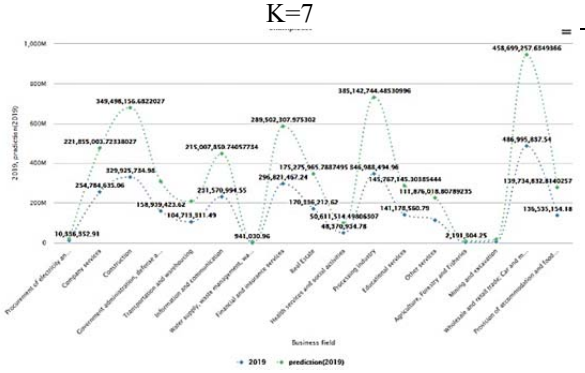
K=6



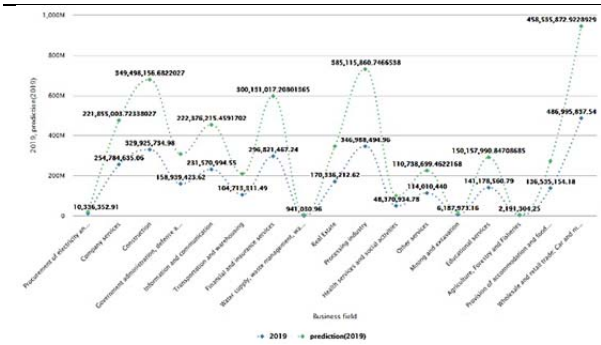
K=7



K=8



K=9



K=10

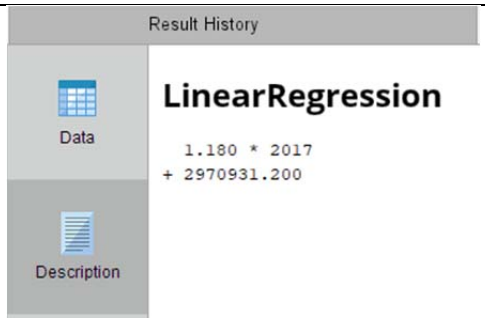


Figure 3. Linear regression statistical equation at sampling type = shuffled_sampling

Following, the number of validations is progressively changed by changing the value of the K-Fold = 2 to 10 for the full RMSE (Root Mean Squared Error) effects. The Root Mean Squared Error results for the sampling type = linear_sampling are shown in the Table 3.

Table 3. RMSE Value for Experiment sampling type = linear_sampling

k	RMSE
k=2	IDR 17,213,321
k=3	IDR 16,491,677
k=4	IDR 13,602,989
k=5	IDR 13,588,801
k=6	IDR 12,451,125
k=7	IDR 10,006,758
k=8	IDR 11,400,609
k=9	IDR 10,578,159
k=10	IDR 9,977,431

Figure 4 is a graphic image of each K-Fold value at sampling type = linear_sampling

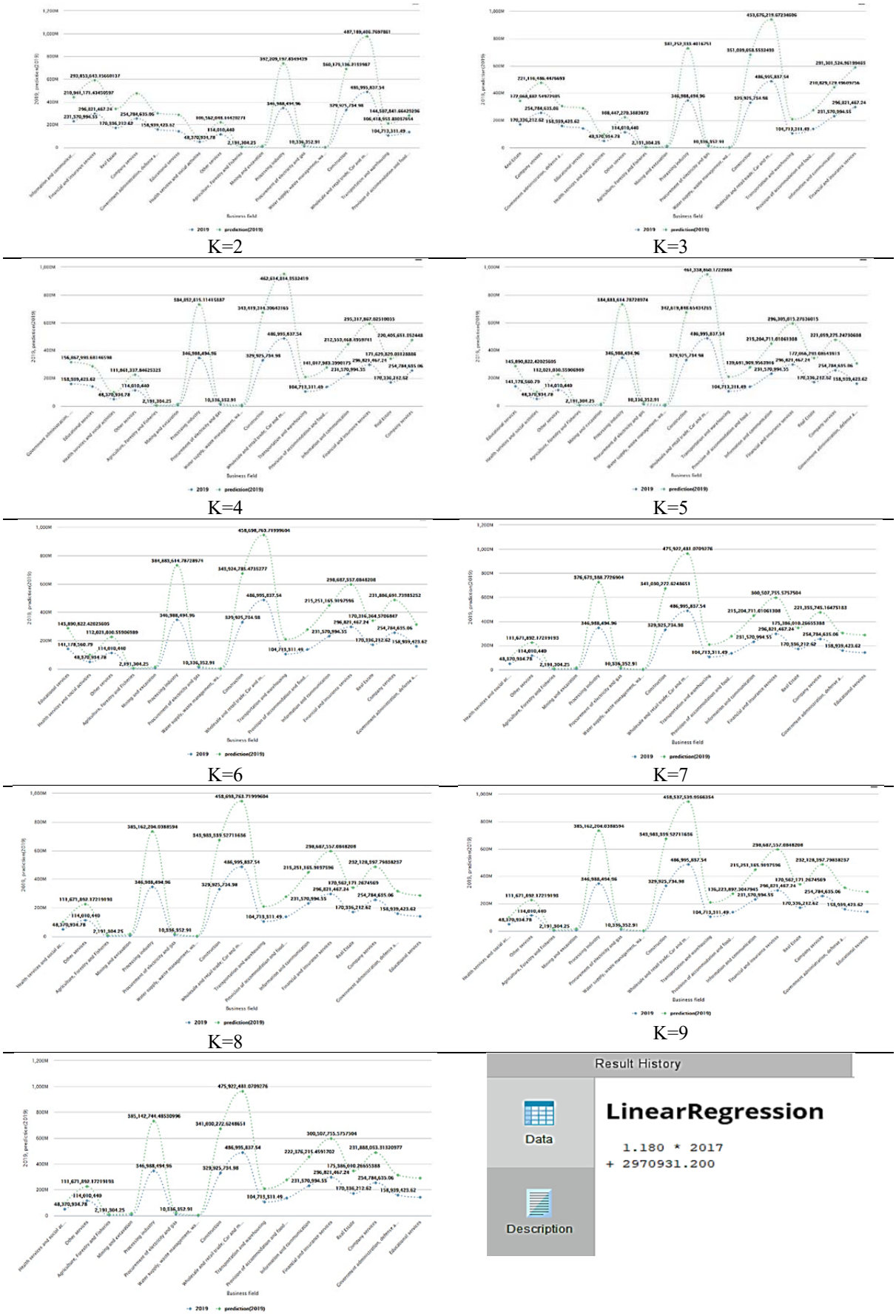


Figure 4. Linear regression statistical equation at sampling type = linear_sampling

Based on the results of the research and tests carried out, the smallest RMSE value was obtained in the sampling type = linear_ sampling with k = 10 with an RMSE value of IDR 9,977,431. Following are the complete results of the prediction using k-fold = 10 at sampling type = linear_sampling as shown in the following Table 4:

Table 4. Prediction Results (k=10; sampling type = linear_sampling)

Business field	2019		prediction(2019)	
Health services and social activities	IDR	48,370,935	IDR	49,593,227
Other services	IDR	114,010,440	IDR	111,671,892
Agriculture, Forestry and Fisheries	IDR	2,191,304	IDR	6,711,880
Mining and excavation	IDR	6,187,973	IDR	11,050,124
Processing industry	IDR	346,988,495	IDR	385,142,744
Procurement of electricity and gas	IDR	10,336,353	IDR	12,404,459
Water supply, waste management, waste and recycling	IDR	941,031	IDR	3,553,890
Construction	IDR	329,925,735	IDR	341,030,273
Wholesale and retail trade; Car and motorcycle reservation	IDR	486,995,838	IDR	475,922,481
Transportation and warehousing	IDR	104,713,311	IDR	104,267,812
Provision of accommodation and food and drink	IDR	136,535,154	IDR	140,708,656
Information and communication	IDR	231,570,995	IDR	222,376,215
Financial and insurance services	IDR	296,821,467	IDR	300,507,756
Real Estate	IDR	170,336,213	IDR	175,386,010
Company services	IDR	254,784,635	IDR	231,888,053
Government administration, defense and social security are mandatory	IDR	158,939,424	IDR	155,244,136
Educational services	IDR	141,178,561	IDR	149,809,652

4. Conclusion

The inference can be drawn from these results:

a) The Linear Regression algorithm can provide accurate results or accurate calculation of gross regional national product forecasts based on the fields of operation and, of course, the necessary steps have to be taken.

b) The lowest root mean square error (RMSE) Linear Regression Algorithm results in K-Fold = 2 samples with the product of the sample type = linear sampling with IDR 9,977,431 results.

References

[1]. Chusaeni, A. M. (2013). *Analisis Sektor Lapangan Usaha Yang Mempengaruhi Nilai Pdrb (Produk Domestik Regional Bruto) Atas Dasar Harga Berlaku Dan Konstan Tahun 2000 Kabupaten Kendal Periode 1997-2011* (Doctoral dissertation, Universitas Negeri Semarang).

[2]. Elsi, Z. R. S., Pratiwi, H., Efendi, Y., Rusdina, R., Alfah, R., Windarto, A. P., & Wiza, F. (2020, July). Utilization of Data Mining Techniques in National Food Security during the Covid-19 Pandemic in Indonesia. In *Journal of Physics: Conference Series* (Vol. 1594, No. 1, p. 012007). IOP Publishing.

[3]. Li, T. (2019). Mining association rules on enrollment information of higher vocational colleges using the apriori algorithm. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, *23*(4), 775-781.

[4]. Pertiwi, M. W., & Indrajit, R. E. (2017). Metode Regresi Linier Untuk Prediksi Pengadaan Inventaris Barang. *Simnasiptek 2017*, *1*(1), 27-30.

[5]. Pratama, S. (2016). Prediksi Harga Tanah Menggunakan Algoritma Linear Regression. *Technologia: Jurnal Ilmiah*, *7*(2).

[6]. Yanto, R. (2020). Implementasi Data Mining Prediksi Kebutuhan Tenaga Listrik Di Kota Lubuklinggau. *Techno. Com*, *19*(2), 197-206.

[7]. Warsino, W. (2017). Metode Peramalan Permintaan Jasa Penerjemahan Bahasa Asing Dengan Algorithma Linear Regression, Menggunakan Rapidminer. Studi Kasus: Azzam Translator Bekasi. *Santika: Jurnal Ilmiah Sains dan Teknologi*, *7*(2), 621-628.

[8]. Suparta, W., & Putro, W. S. (2017, March). Using multiple linear regression model to estimate thunderstorm activity. In *IOP Conference Series: Materials Science and Engineering* (Vol. 185, No. 1, p. 012023). IOP Publishing.

[9]. El Aissaoui, O., El Madani, Y. E. A., Oughdir, L., Dakkak, A., & El Allioui, Y. (2020). A Multiple Linear Regression-Based Approach to Predict Student Performance. *Journal: Advances in Intelligent Systems and Computing Advanced Intelligent Systems for Sustainable Development (AI2SD'2019)*, 9-23.

[10]. Shen, Y., Hua, J., Jin, C., & Huang, D. (2019, April). TCL: Tensor-CNN-LSTM for Travel Time Prediction with Sparse Trajectory Data. In *International Conference on Database Systems for Advanced Applications* (pp. 329-333). Springer, Cham.

[11]. H. Di Kesuma and R. Yanto.,(2019). "Data mining prediction production of water needs using simple linear regression algorithms," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 1, pp. 3575–3578.

[12]. C. S. Sasikumar and A. Kumaravel.,(2019). "Linear regression and support vector machine for classification of e-learning students' engagement and performance," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 6, pp. 2696–2700.

[13]. Sharma, I., & Kakchapati, S. (2018). Linear Regression Model to Identify the Factors Associated with Carbon Stock in Chure Forest of Nepal. *Scientifica*, *2018*.

[14]. T. Mohd, S. Jamil, S. Masrom, F. Architecture, and U. T. Mara., (2020). "Multiple Linear Regression on Building Price Prediction with Green Building Determinant," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 9, pp. 1137–1148.