http://infor.seaninstitute.org/index.php/infokum/index
JURNAL INFOKUM, Volume 9, No. 2, Juni 2021 ISSN : 2302-9706

LINEAR REGRESSION ANALYSIS TO PREDICT THE LENGTH OF THESIS COMPLETION

Fristi Riandari [1], Hengki Tamando Sihotang [2], Aura Nissa Galuh Suanda [3], Samuel Lumban Tobing [4]

[1, 2, 3, 4] STMIK Pelita Nusantara
fristy.rianda@ymail.com, hengki_tamando@yahoo.com

Article Info
Received 31 May 2021
Revised 28 June 2021
Accepted 30 June 2021

Abstract

Students who carry out knowledge in the undergraduate program will certainly be faced with the preparation of a thesis at the end of their study period. However, every year students still find it takes longer than the time specified in completing their thesis. This is caused by several things, such as students who are working, working hours that do not support the implementation of thesis preparation, students who already have families and other factors. This of course makes universities have to prepare special strategies in order to reduce the number of students who cannot complete their thesis on time in the future, one of which is with a decision support. This can be done by utilizing university big data. Prediction of the length of time for completion of college student thesis can be done by utilizing data mining and a simple linear regression approach. Using 1 independent variable, namely the average inhibiting factor (Working Status, Working Hours, Work Sip, Guidance Media, Status) (X1) and the number of days of thesis completion being the dependent variable (Y). After looking for the regression value of b and constant a, then the simple linear regression equation model is: Y = 280.450 + 1.650 X.

Keywords: big data, prediction, simple linear regression.

1. Introduction

The popularity of the term big data since the last few decades shows that big data can be used to produce information in the future, this is evidenced by the many research topics related to big data. In the era of digital transformation, big data has taken on an important role, such as being able to generate valuable information for predict [1]. One of the areas of science that utilizes big data in the settlement of the problems that faced that field of science data mining.

Data mining is a technology computer which classify, categorize and predict the data in a number of large [2]. Data mining has been realized by a variety of methods including analysis of statistical and empirical, analysis of the network of social, engineering ML, and techniques of NLP [3]. One of the benefits with their data mining is to generate information that can be assets to increase power competitiveness of an institution. Data mining is associated with learning machine is to extract information that is useful from the data size of a large [4]. The importance of data mining is based on several reasons. First, the availability of sources of power academically are abundant. Second, because the reservoir data easily accessible to this. Third, the availability of data in a number of large on the collaboration of scientists, share documents, and publications allows the evaluation of the impact of science from various entities including papers, authors and journals. Measurement of the impact of science is considered to be important for the government and the field of business for the process of making decisions such as the allocation of funds, the identification

of gaps research, determination of ranking universities, the period of office, and the decision of recruitment [3]. Citing the reason for the four linked importance of data mining that can help in making a decision, then the research is the use of data mining is done to perform the prediction. In essentially the prediction is alleged or estimates on the occurrence of an incident or event in time that will come. And that became the purpose of the research it is generating an information that makes it easy to draw up policies the days to come.

Analysis of regression is one of the techniques of data mining is often used to determine how the level of correlation between variables dependent or result can be predicted through variables independent or cause, as an individual. Regression analysis is closely related to correlation, where every regression must have a correlation, but each correlation is not necessarily able to proceed to the regression process. The approach that is used in the research is that the approach of regression linear. Regression linear is metodestatistik which to determine the effect of variable dependent (descriptors) of the variable independent (free). Correlation and regression both have a very close relationship. Each regression always have a correlation, but each correlation is not necessarily proceed towards regression. Correlations were not followed by regression, is the correlation between the two variables that do not have a relationship casual/cause as a result of, or relationship functional. To set the two variables have a relationship causal or not, should be based on theories or concepts of the two variables are. Linear regression is based on a functional or causal relationship between one independent variable (Y) and one dependent variable (X). The results will allow us to see the accuracy of the simple linear regression approach in predicting the factors that affect the length of the thesis completion time [5].

Research to approach the analysis of regression linear multiple has been used in various fields , a study that predicts the cost of HR BIM in the year 2020 by [6], explained that in most cases , the cost of personnel working BIM only be calculated through proporsiluas floor dirty in accordance with the project practical , which always carry a risk which is not unexpected untukmanajer project because the method of regression linear simple it contains risks tinggikesalahan estimate of the phase of construction . By because it, research it will develop hybrid barumetodologi which combines Forest Random and Regression Linear Simple untukmenghilangkan fault prediction cost labor work BIM on the stage of construction.

Another study in 2020 concerning the Prediction of Soil Compression Index [7], in this study, the soil investigation report used for the city of Al-Nasiriya was carried out by the national construction laboratory. The result of the nature of physical showed that soil the city of Al-Nasiriya has a plateau low to high compressibility and are classified as clay anorganikplastisitas moderate to high. Regression linear simple-analysis sion applied to estimate kompresiindeks are not directly through the use of several properties tanahindeks as limits Atterberg. The correlation coefficient of the linear regression model indicated a reasonable relationship between the compression index and the parameters proposed in this paper. The results obtained show that the model terbaikmampu predict the number kompresiindeks and the accuracy of which is high based on likuiditasbatas in the analysis of regression. There are also reasons a deal that can be measured between the measured and the diprediksinilai index compression in addition to the presence of aperbedaan little between the value that is calculated using crunches statistics for limits Atterberg. Empirical equations were derived from regression linear simple models can be unreliable used in design engineering to the area of study with confidence themselves are high.

Research in the year 2019 [8] who do experiment field in summer hot and humid (Transplantation Aus) to realize the loss hasilpadi varieties Purbachi vulnerable were inoculated blight bacteria (BB). Treatment consists of BBinokulasi at various stages of the growth of plants such as tillers maximum (MT), the initiation of panicle (PI), boots (Bt), stages of flowering and heading differ include controls (no inoculation BB). Disease severity index (DSI) was measured at 14 days after inoculation (DAI) and harvest. Data on weight of 1000 grains and results adalahdicatat at the time of harvest. Variations significantly on DSI was

observed in the growth of the plants were inoculated BB berbedatahapan. MT, PI and Boot stadia inoculation showed similar (DSI 7.1-8.0) but DSI is higher than flowering danpos stages of inoculation (3.2-5.3) even controls (0.00) at 14 days after planting. However, all treatments showed the same DSI 9.0 at harvest. Disease blight bacteria can affect the weight of the grain to limit certain though not signifikanantara treatment (0.1 to 4.5%). DSI showed a negative correlation with the weight of 1000 grains (r=-0.77*) and was the same as the result (r=-0.97**). Yields ranged from 2.4 to 3.4 t/ha between treatments. Losing hasildiamati 5.8 to 30.4% in the treatment of the inoculated BB. The MT, PI and Boot stage inoculations affected the yield a lot resulting in a yield loss of 21-30.4%. Can be concluded that the varieties are prone to be affected by nyatakehilangan results up to 30.4% with an outbreak BB are severe. The equation regression simple = 4,09-0,211X (= Results, X = Score severity BB) is recommended for the prediction of loss results in varieties prone in the summer heat.

And that became the purpose of the research it is generating an information that makes it easy to draw up policies the days to come with the analysis of regression linear simple which uses one variable independent (X) and a variable dependent (Y).

2. Method

2.1 Research Method

1. Data Collection Method
   a. Literature Review
      Learn the theories and concepts of Data Mining, Prediction, Simple Linear Regression and various studies related to this research topic.
   b. Expert Interview
      Interviewing parties who have the authority to monitor the implementation of student thesis preparation at universities that are research locations with the aim of obtaining data that will be used as test data.
2. Data Analysis: Application of Simple Linear Regression
   At this stage the data that has been obtained and has been determined as a fixed variable and independent variable will then be processed using simple linear regression in accordance with the steps in the approach.

2.2 Data Mining

Data mining, often also called knowledge discovery in database (KDD), is an activity that includes collecting, using historical and data to find regularities, patterns or relationships in large data sets [12]. KDD is a process of finding useful information from databases. Where in the process includes understanding the application field, making target data determined from the raw data contained in the database, as well as data preprocessing and data cleaning [13]. The main purpose of KDD is to extract high-level knowledge from low-level information, or in other words, to automatically process from large amounts of raw data, identify the most significant and meaningful patterns, present it as appropriate knowledge to achieve user goals. [14].

Data mining is one of the implementation models that is applied to find a pattern that is able to make predictions based on previous data in the period and is used to explore knowledge from large amounts of data. [15]. Data Mining is a process of extracting data or filtering data by utilizing a large enough data set through a series of processes to obtain valuable information from the data. Data Mining can be applied to various fields that can contain a number of data. According to Daryl Pregibon that “Data mining is a mix of statistics, artificial intelligence, and database research” which is still developing [16]. Data Mining is a process of extracting data or filtering data by utilizing a large enough data set through a series of processes

to obtain valuable information from the data. Data Mining can be applied to various fields that can contain a number of data. According to Daryl Pregibon that "Data mining is a mix of statistics, artificial intelligence, and database research" which is still developing [17]. Various techniques are available in data mining for knowledge extraction including classification, prediction, estimation, association and grouping [18].

Data mining is a process of analyzing data from different perspectives and concluding it into important information that can be used to increase profits, reduce costs, or even both. Technically, data mining can be referred to as the process of finding correlations or patterns from hundreds or thousands of fields from a large relational database. One of the benefits of data mining is to produce information that can be an asset to increase the competitiveness of an institution [11].

2.2 Regresi Linier

Statistics play an important role in big data because many statistical methods are used for big data analysis. Statistical software provides rich functionality for data analysis and modeling, but can only handle small amounts of data. Regression can be seen in many widely used fields such as business, social and behavioral sciences, biological sciences, climate prediction, and so on. Regression is the study of the relationship between variables and independent variables, and the relationship between independent variables and dependent variables is expressed through the regression equation. Multivariate linear regression model is a correlation between a variable and a number of independent variables [19]. Regression analysis is used in big data statistical analysis because the regression model itself is popular in data analysis [20]. Linear algorithms have several advantages, one of which is their simple structure [21]. There are two types of linear regression analysis based on the number of input variables, namely the first type is called simple linear regression which takes a single input variable, while the other type has more than one input variable and is called multiple linear regression [22].

2.3 Simple Linear Regression

Regression analysis is a statistical method that observes the relationship between the dependent variable Y and a series of independent variables X1,…,Xp. The purpose of this method is to predict the value of Y for a given value of X. Simple linear regression model is the simplest regression model that has only one independent variable X. Regression analysis has several uses, one of which is to predict the dependent variable Y [4]. The equation for the simple linear regression model is as follows [9].

$Y = a + bX$ (1)

Y is the predicted dependent variable, X is the independent variable, a is intercep, i.e. the value of Y when X = 0, and b is the slope, i.e. the average change of Y to the change of one unit of X. Coefficients a and b are regression coefficients where the value a and b can be found using the following equation.

$b = n(\Sigma xy) - (\Sigma x)(\Sigma y)/n(\Sigma x2) – (\Sigma x)2$

$a =\Sigma y - b(\Sigma x)/n$

The value of a is the slope, b is the intercept and n is the number of data used in the calculation.

3. Results and Discussion

3.1 Testing Data Set

In this study, the dataset is sourced from the university database in 2020, the dataset used is related to the dependent variable/responses (Y) and the independent variable/predictor (X) which have been adjusted to the prediction needs.

Dataset processing with a simple linear regression approach is done by following the steps in the model. And to analyze this case, the things to do are:

1. Objectives: do the indicators that are determined as variable X affect the completion time of Thesis?
2. Variable : X (Independent Variable/Predictor) =

Table 3. Independent Variabel

No.	*Delay Factor*
1	Working Status (WST)
2	Working Hours (WH)
3	Working Sip (WS)
4	Guidance Media (GM)
5	Status (S)

Table 2. Preliminary data

No_Id	*WST*	*WH*	*WS*	*GM*	*S*	*Result (Day)*
1	75	50	50	65	60	180
2	25	20	15	35	60	240
3	25	30	35	65	60	180
4	25	20	35	35	40	255
5	75	50	50	65	60	180
6	25	20	35	65	60	205
7	25	20	15	35	60	210
8	25	20	15	35	60	240
9	75	50	50	35	60	225
10	25	30	35	65	60	180

Furthermore, after the data has been successfully transformed, the average value for each student will be calculated based on the weight value obtained from the variables that have been set to get the value of the X variable. And after looking for the average value, the results can be assessed.

Table 3. Average Value Of The Delay Factor

No_Id	*WST*	*WH*	*WS*	*GM*	*S*	*Result (Day)*
1	75	50	50	65	60	180
2	25	20	15	35	60	240
3	25	30	35	65	60	180
4	25	20	35	35	40	255
5	75	50	50	65	60	180
6	25	20	35	65	60	205
7	25	20	15	35	60	210
8	25	20	15	35	60	240
9	75	50	50	35	60	225
10	25	30	35	65	60	180

Table 4. Help Table Made To Make It Easier To Do Calculations

No.	*X*	*X^2*	*Y*	*Y^2*	*XY*
1	60	3600	180	32400	10800
2	31	961	240	57600	7440
3	43	1849	180	32400	7740
4	31	961	255	65025	7905
5	60	3600	180	32400	10800
6	41	1681	205	42025	8405

7	31	961	210	44100	6510
8	31	961	240	57600	7440
9	54	2916	225	50625	12150
10	43	1849	180	32400	7740
∑	425	19339	2095	32400	86930

Then look for the value of the regression coefficient b using the formula:
$b = n(\Sigma xy) - (\Sigma x)(\Sigma y)/n(\Sigma x2) - (\Sigma x)^2$
$b = 10(86930)-(425)(2095)/10(19339)-(425)^2$
$b = 21075/12765 = 1{,}650$

Next look for the value of the constant a by using the formula:
$a = \Sigma y - b(\Sigma x)/n$
$a = (2095)(19339)-(425)(86930)/10(19339)-(425)^2$
$a = 3579955/12765 = 280{,}450$

So the simple linear regression equation model is:
$Y = 280{,}450 + 1{,}650\ X$

4. Conclusions

Simple linear regression analysis can have a simple structure using 1 independent variable, namely the average inhibiting factor (Working Status, Working Hours, Work Sip, Guidance Media, Status) (X1) and the number of days of thesis completion being the dependent variable (Y). that can be accepted and used as a tool in predicting the completion time of student thesis work in universities. Future research is expected to be able to compare methods that can be used in predicting to see the level of accuracy in each method.

Reference

[1] L. Ardito, R. Cerchione, P. Del Vecchio, and E. Raguseo, "Big data in smart tourism: challenges, issues and opportunities," *Curr. Issues Tour.*, vol. 22, no. 15, pp. 1805–1809, 2019, doi: 10.1080/13683500.2019.1612860.

[2] A. Yang, Y. Han, C.-S. Liu, J.-H. Wu, and D.-B. Hua, "D-TSVR Recurrence Prediction Driven by Medical Big Data in Cancer," *IEEE Trans. Ind. Informatics*, vol. 3203, no. c, pp. 1–1, 2020, doi: 10.1109/tii.2020.3011675.

[3] A. Dridi, M. M. Gaber, R. M. A. Azad, and J. Bhogal, "Scholarly data mining: A systematic review of its applications," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, no. October, pp. 1–23, 2020, doi: 10.1002/widm.1395.

[4] T. M. Song and J. Song, "Prediction of risk factors of cyberbullying-related words in Korea: Application of data mining using social big data," *Telemat. Informatics*, vol. 58, p. 101524, 2021, doi: 10.1016/j.tele.2020.101524.

[5] D. Muriyatmoko, "Analisa Volume Terhadap Sitasi Menggunakan Regresi Linier Pada Jurnal Bereputasi di Indonesia," *J. Ilm. Simantec*, vol. 6, no. 3, pp. 129–134, 2018.

[6] J. Z. Wang and J. Zhang, "Predicting in BIM Labour Cost with a hybrid approach Simple Linear Regression and Random Forest," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 565, no. 1, 2020, doi: 10.1088/1755-1315/565/1/012108.

[7] E. A. Mandhour, "Prediction of Compression Index of the Soil of Al-Nasiriya City Using Simple Linear Regression Model," *Geotech. Geol. Eng.*, vol. 38, no. 5, pp. 4969–4980, 2020, doi: 10.1007/s10706-020-01339-w.

[8] T. H. Ansari, M. Ahmed, S. Akter, M. S. Mian, M. A. Latif, and M. Tomita, "Estimation of Rice Yield Loss Using a Simple Linear Regression Model for Bacterial Blight Disease," *Bangladesh Rice J.*, vol. 23, no. 1, pp. 73–79, 2020, doi: 10.3329/brj.v23i1.46083.

[9] A. Hijriani, K. Muludi, and E. A. Andini, "Implementasi Metode Regresi Linier Sederhana Pada Penyajian Hasil Prediksi Pemakaian Air Bersih Pdam Way Rilau Kota Bandar Lampung Dengan Sistem Informasi Geofrafis," *Inform. Mulawarman J. Ilm. Ilmu Komput.*, vol. 11, no. 2, p. 37, 2016, doi: 10.30872/jim.v11i2.212.

[10] A. Bengnga and R. Ishak, "Prediksi Jumlah Mahasiswa Registrasi Per Semester Menggunakan Linier Regresi Pada Universitas Ichsan Gorontalo," *Ilk. J. Ilm.*, vol. 10, no. 2, pp. 136–143, 2018, doi: 10.33096/ilkom.v10i2.274.136-143.

[11] W. M. Baihaqi, M. Dianingrum, and K. A. N. Ramadhan, "Regresi Linier Sederhana Untuk Memprediksi Kunjungan Pasien di Rumah Sakit Berdasarkan Jenis Layanan dan Umur Pasien," *J. SIMETRIS*, vol. 10, no. 2, pp. 671–680, 2019.

[12] H. Santoso, I. P. Hariyadi, and Prayitno, "Data Mining Analisa Pola Pembelian Produk Dengan Menggunakan Metode Algoritma Apriori," *Tek. Inform. ISSN 2302-3805*, no. 1, pp. 19–24, 2016, [Online]. Available: http://ojs.amikom.ac.id/index.php/semnasteknomedia/article/download/1267/1200.

[13] I. Virgo, S. Defit, and Y. Yunus, "Klasterisasi Tingkat Kehadiran Dosen Menggunakan Algoritma K-Means Clustering (Studi Kasus Institut Agama Islam Batusangkar)," *J. Sistim Inf. dan Teknol.*, vol. 2, no. 1, pp. 24–29, 2020, doi: 10.37034/jsisfotek.v2i1.22.

[14] R. A. Putra and S. Defit, "Data Mining Menggunakan Rough Set dalam Menganalisa Modal Upah Produksi pada Industri Seragam Sekolah," *J. Sistim Inf. dan Teknol.*, vol. 1, no. 4, pp. 72–78, 2019, doi: 10.35134/jsisfotek.v1i4.18.

[15] M. Guntur, J. Santony, and Y. Yuhandri, "Prediksi Harga Emas dengan Menggunakan Metode Naïve Bayes dalam Investasi untuk Meminimalisasi Resiko," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 2, no. 1, pp. 354–360, 2018, doi: 10.29207/resti.v2i1.276.

[16] A. Hidayad, S. Defit, and S. Sumijan, "Penerapan Algoritma K-Means Clustering untuk Melihat Hubungan Kegiatan Tahfiz dengan Hasil Belajar (Studi Kasus Madrasah Aliyah Negeri 1 Bukiktinggi)," *J. Sistim Inf. dan Teknol.*, vol. 2, no. 2, pp. 41–47, 2020, doi: 10.37034/jsisfotek.v2i2.34.

[17] H. Sulastri and A. I. Gufroni, "Penerapan Data Mining Dalam Pengelompokan Penderita Thalassaemia," *J. Nas. Teknol. dan Sist. Inf.*, vol. 3, no. 2, pp. 299–305, 2017, doi: 10.25077/teknosi.v3i2.2017.299-305.

[18] P. A. Ariawan, "Optimasi Pengelompokan Data Pada Metode K-means dengan Analisis Outlier," *J. Nas. Teknol. dan Sist. Inf.*, vol. 5, no. 2, pp. 88–95, 2019, doi: 10.25077/teknosi.v5i2.2019.88-95.

[19] X. Xu, Z. Sun, L. Wang, J. Fu, and C. Wang, "A Comparative Study of Customer Complaint Prediction Model of Time Series, Multiple Linear Regression and BP Neural Network," *J. Phys. Conf. Ser.*, vol. 1187, no. 5, 2019, doi: 10.1088/1742-6596/1187/5/052036.

[20] B. Amil *et al.*, "No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title," *J. Chem. Inf. Model.*, vol. 21, no. 1, pp. 1–9, 2020, [Online]. Available: https://doi.org/10.1016/j.tmaid.2020.101607%0Ahttps://doi.org/10.1016/j.ijsu.2020.02.034%0Ahttps://onlinelibrary.wiley.com/doi/abs/10.1111/cjag.12228%0Ahttps://doi.org/10.1016/j.ssci.2020.104773%0Ahttps://doi.org/10.1016/j.jinf.2020.04.011%0Ahttps://doi.o.

[21] F. Wang, Z. Shi, A. Biswas, S. Yang, and J. Ding, "Multi-algorithm comparison for predicting soil salinity," *Geoderma*, vol. 365, no. February 2019, p. 114211, 2020, doi: 10.1016/j.geoderma.2020.114211.

[22] H. Rawashdeh *et al.*, "Intelligent system based on data mining techniques for prediction of preterm birth for women with cervical cerclage," *Comput. Biol. Chem.*, vol. 85, no. February, p. 107233, 2020, doi: 10.1016/j.compbiolchem.2020.107233.