ISSN:2302-8149
e-ISSN:2540-9719

Regression Analysis to Predict the Length of Time to Complete a Thesis based on the Title

[1]Al Aminuddin, [2]Rahmat Hidayat*, [3]Gita Sastria, [4]Astried
[1,4]Program Studi Sistem Informasi, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Riau
[2,3]Program Studi Manajemen Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Riau
[1,2,3,4]Kampus Bina Widya KM 12,5 Simpang Baru Pekanbaru 28291 - Indonesia
*e-mail: rahmat.hidayat@lecturer.unri.ac.id

(***received***: 16 January 2025, ***revised***: 15 February 2025, ***accepted***: 18 February 2025)

Abstract

The selection of thesis titles by students is an important thing to do as part of the graduation requirements in completing undergraduate studies. In general, the difficulty or complexity of the thesis can be reflected through the title of the thesis that is appointed. This can indicate that the more difficult a thesis title, the longer it will take to complete the thesis research. This study utilizes the data mining method using machine learning, namely linear regression, in predicting how long it will take to complete a thesis title. The data used is obtained from the words or text in the thesis title as a feature or independent variable and the completion time in days as the dependent variable to predict the time required for students starting from a thesis proposal seminar to a comprehensive seminar or thesis final session. The regression model produces an evaluation value of the coefficient of determination of 0.999, which is close to the maximum value equal to 1.

Keywords: data mining, linier regression, thesis, TF-IDF.

1 Introduction

At the undergraduate level, students are required to complete a thesis as part of their graduation requirements. A thesis is a scientific research paper conducted under the guidance of a supervisor and presented to an examination panel [1]. This process is designed to develop students' ability to think and work scientifically. Each student is assigned a supervisor based on their chosen thesis title, which is typically selected with guidance from an academic advisor [2]. The selection of a thesis title is a crucial step in the thesis process, as it can influence the duration of completion [3]. In addition to personal interest and the relevance to elective courses, the complexity of the chosen title also plays a role in determining how long a thesis will take to finish. Generally, more complex topics require extended research and analysis, contributing to longer completion times [4].

This study applies Data Mining techniques with Machine Learning to predict the time required to complete a thesis based on its title. Using Linear Regression, a statistical method for analyzing relationships between variables, the research will extract textual features from thesis titles to estimate the time needed from the thesis proposal seminar to the comprehensive seminar or thesis defense [5]. The objective is to evaluate the accuracy of Machine Learning in making such predictions using limited textual features. The findings of this study could contribute to the development of a thesis title management information system that incorporates Machine Learning for predicting thesis completion times. This system could assist students in selecting appropriate titles aligned with their interests while also providing insights into potential time commitments. The research output will include a research paper, dataset, and program code, which can serve as a foundation for future studies in this field.

2 Literature Review

This study is supported by several studies related to the prediction of student thesis completion time using several methods and algorithms. Table 1 is a literature review of similar studies used as references in this study.

ISSN:2302-8149
e-ISSN:2540-9719

Table 1. Literature review

Title	Method	Result	Reference
Predict The Thyroid Abnormality Particular Disease Likelihood of The Symptoms' Certainty Factor Value and Its Confidence Level: A Regression Model Analysis	Multiple Linear Regression (MLR) and Multiple Polynomial Regression (MPR)	The MPR model has the best results compared to the MLR model in predicting certain diseases, possibly thyroid disorders, supported by R-squared 94.7%, R-squared adjusted 94.4%, F-value 265.925, dan p-value < 0.05.	[6]
Using Regression Model Analysis for Forecasting the Likelihood of Particular Symptoms of COVID-19.	Multiple Linear Regression (MLR) and Multiple Polynomial Regression (MPR).	The MLR and MPR models are the most accurate regression models for estimating the probability of a disease being associated with COVID-like symptoms.	[7]
The Role of Data Mining in Predicting Adidas Shoe Sales Level Using Simple Linear Regression Algorithm Method	Literature review	The linear regression algorithm has proven to be very effective in predicting Adidas shoe sales, obtaining the regression equation Y = 122.666,5 + 12.008,6X dengan hasil 133.978,6%.	[8]
Implementation of Multiple Linear Regression Method on Palm Oil Tonnage Prediction System at PT. Paluta Inti Sawit	Multiple Linear Regression (MLR)	In general, the level of accuracy of predicting the amount of palm oil tonnage for a month is 99.99%, the lowest level of prediction accuracy on November 4, 2022 has a value of 88%, while the highest prediction accuracy on November 5, 2022 and November 7, 2022 is 100%.	[9]
Gold Value Prediction Using Linear Regression Algorithm	Linear Regression	The linear regression model can be used to predict future gold prices with the MAE evaluation method of 4341.140 which is more accurate than the RMSE which has a value of 4893.132.	[10]
Price Prediction on USDCHF Forex Trading Pair Using Linear Regression	Linear Regression	The best linear regression algorithm was obtained on the Open variable, with a linear regression equation of y=0.0145+0.9849x, the best MSE value was 0.0000328509 and the best RMSE had a value of 0.0057315705.	[11]
Performance Comparison of Decision Tree Regression and Multiple Linear Regression for BMI Prediction on Asthma Dataset	Regression Decision Tree dan Multiple Linear Regression (MLR)	The Multiple Linear Regression (MLR) method has better results than the Decision Tree regression method.	[12]

Prediction of 2023 Rice Harvest Results Using Linear Regression Method in Indramayu Regency	Linear Regression	The algorithm results have MAE, MSE, RMSE, R2-Score values, and the system displays MAE, MSE, RMSE, and R2-Score values with predictions that in 2023 there will be a decline from the previous year.	[13]
Prediction of Mobile Phone Sales in Store X Using Linear Regression Algorithm	Linear Regression	The prediction of sales in the next 3 months has an evaluation value that can be concluded as sufficient or the results of the linear regression equation can be used.	[14]
Data Mining in Predicting The Number of Patients With Linear Regression and Exponential Smoothing	Linear Regression & Exponential Smoothing	Linear Regression predicted better by having 23.90% MAPE, while Exponential Smoothing had 27.62%	[15]
Financial Performance Prediction of PT Astra International Tbk Using Linear Regression and Exponential Smoothing	Linear Regression & Exponential Smoothing	The combination of linear regression and exponential smoothing methods produces predicted values that are closer to the actual values than with each method alone.	[16]

3 Research Methodology

This study has five main processes that are carried out sequentially, namely data collection, data pre-processing, variable weighting, data prediction, and analysis & evaluation results. The process in this study can be observed in Figure 1, each stage has a very meaningful output to be carried out in the next process. The research flow begins with data collection and ends with analysis and evaluation results. A detailed explanation related to each process will be discussed in this chapter.

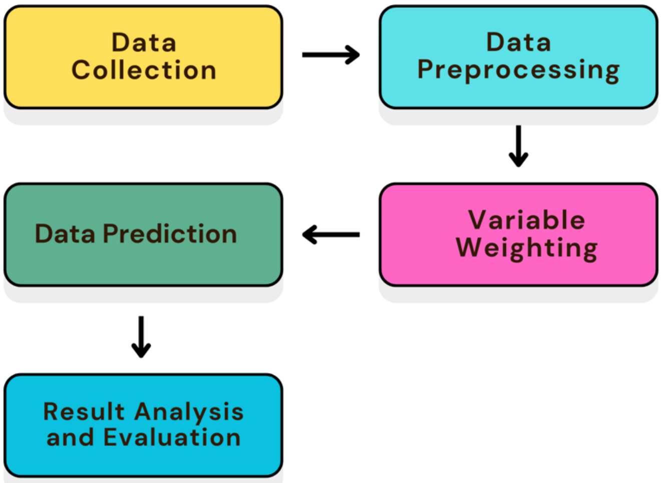

Figure 1. Research stages

3.1 Data Collection

The data used in this study comes from the Informatics Study Program at University X, containing data related to the process of working on student theses. The dataset contains 4 columns of data attributes, namely student name, student number, thesis title and duration of thesis work. The data in this study uses the CSV extension which can be processed directly using the Linear Regression algorithm.

3.2 *Preprocessing* Data

The collected data is cleaned and formatted in the preprocessing stage according to the data format needed as input in the prediction process. Data preprocessing or data preprocessing is a preparatory stage in processing data that aims to facilitate the data processing process [17]. Data preprocessing usually includes case folding, tokenization, POS tagging, cleaning data from elements that are not needed in the analysis (stopwords), stemming, and lemmatization [18]. In this study, the data preprocessing process performs stopword removal and attribute deletion in the dataset. So that a clean dataset will have 2 attribute columns in the form of thesis title and duration of work.

3.3 Variable Weighting

At this weighting stage, the weight is determined for each word/sentence in a text. In the process of weighting this variable, the Term Frequency Inverse Document Frequency (TF-IDF) algorithm is used. TF-IDF is a method for weighting words in text based on the frequency of occurrence of the word and the inverse weight of the frequency of occurrence of the words in all sentences [19]. In the TF-IDF equation [20] it is written as in formula (1). In the formula (1) $n_{i,j}$ shows the number of occurrences of a word in a document d$_j$. $\sum_k n_{k,j}$is the total number of occurrences of a word in a document. Furthermore, |D| is the total number of documents, while $|d_j \in D: t_j \in d_j|$ is the number of documents that the keyword t_j showed.

$$TFIDF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \; x \; log \frac{|D|}{|d_j \in D: t_j \in d_j|} \tag{1}$$

3.4 Data Prediction

At this data processing stage, data training is carried out on data that already has features and completion time to then carry out the testing process on new data in order to find out the prediction time according to the thesis title. The data prediction process at this stage uses the Linear Regression algorithm, which is a data analysis technique in predicting unknown data values by using other related and known data values [21]. The purpose of Linear Regression is to understand and predict the value of the dependent variable based on the values of the independent variables that have been given [22]. In the equation, Linear Regression can be written as in formula (2). In formula (2) it is known that a is a constant, X is an independent variable, and b is a regression coefficient of variable X [23].

$$Y = axbX \tag{2}$$

3.5 Result Analysis and Evaluation

At the analysis stage, an evaluation will be carried out on the prediction process of the data mining or machine learning method that has been applied. Accuracy and errors will be documented to be concluded in order to gain new insights. The evaluation uses the determination coefficient R-Square (R2), which is a measurement to see how much the exogenous variable explains the endogenous variable [24]. The value is zero to one, the closer to one, the independent variable provides all the information needed to predict the variation of the endogenous variable.

4 Result and Discussion

Before applying the linear regression method to predict the completion time of students' theses based on the title of the thesis, it is necessary to first determine two main variables to be used, namely the independent variable and the dependent variable. In this study, the independent variable or commonly called the free variable (X) is the title of the student's thesis, while the dependent variable or bound variable (Y) is the length of time it takes students to complete the thesis in days.

4.1 Data Collection

The data used are research title data and thesis completion time totaling 563 data records. Table 2 shows samples or examples of data used in this study.

Table 2. Data used

No.	Title	Days
1	Algorithm for changing non-standard sentences in tweets into standard sentences	29
2	Application of genetic algorithms to production scheduling optimization for makespan minimization	73
3	Application of Nguyen Widrow's weight initialization algorithm to diagnose diabetes mellitus using the backpropagation neural network method.	67
4	Searching for book borrowing patterns at the UIN Suska Riau library using the prefixspan algorithm	72

5	Implementation of the term frequency inverse document frequency algorithm to determine the level of similarity of internship reports	74
6	Classification of alay accounts on Twitter using the naive bayes classifier method	100
7	The system for determining the place of residence uses the modified k nearest neighbor method mknn	93
8	Hijaiyah letter character recognition using the learning vector quantization 3 lvq 3 method and the modified direction feature mdf method	99
9	Classroom chair monitoring system using convolutional neural network cnn method	102
10	Android-based monitoring application for prospective BCA council member activities, case study of Riau PKS DPW	100
…	**…**	**…**
561	Kutai language text stemming algorithm based on morphological rules	254
562	Implementation of the frequent pattern growth fp growth algorithm to find parameters that influence low programming course grades.	257
563	Kuantan Malay language text stemming algorithm based on morphology	749

4.2 *Data Preprocessing*

The collected data is then preprocessed so that it can be processed into the linear regression method. The data preprocessing stage begins by filtering data on the independent variable which is the student's Thesis Title, data filtering is done to remove words that are not influential and are considered unimportant, such as removing conjunctions, conjunctions, and so on. The data filtering process utilizes Indonesian stopword removal with Python Sastrawi. The results of the filtering process are then processed further at the cleaning stage, in this case to remove symbols, numbers, punctuation, and formats other than letters in the words contained in the Thesis title. The following are the results of the data filtering and cleaning process in Table 3.

Table 3. Data filtering and cleaning results

No.	Title
1	standard sentence conversion algorithm tweet standard sentence
2	application of genetic algorithm optimization of production scheduling makespan minimization
3	application of Nguyen Widrow's weight initialization algorithm to diagnose diabetes mellitus using the backpropagation neural network method
4	search for book borrowing patterns in the UIN Suska Riau library using the prefixspan algorithm
5	Implementation of the term frequency inverse document frequency algorithm determines the level of similarity of the internship report
6	classification of alay twitter accounts using the naive bayes classifier method
7	modified k nearest neighbor mknn residence determination system
8	hijaiyah letter character recognition learning vector quantization lvq method modified direction feature mdf method
9	classroom chair monitoring system convolutional neural network method cnn
10	android-based monitoring application for activities of prospective members of the BCA council, study of the Riau PKS Dpw
…	**…**
561	Kutai language text stemming algorithm based on morphological rules
562	Implementation of the frequent pattern growth fp growth algorithm finds parameters that influence low programming course grades.
563	morphology based kuantan malay language text stemming algorithm

4.3 Variable Weighting

Then the process of changing the format of the independent variable in this case the Thesis Title into a numeric format is carried out so that the weight of each word of the student's thesis title is obtained using the TF-IDF method. The TF-IDF method is a method for calculating the weight of each word in a collection of sentences that is most commonly used in information retrieval. TF-IDF is efficient, easy to use and is known to have accurate results [9]. TF-IDF works by calculating the Term Frequency (TF) and Inverse Document Frequency (IDF) values for each word in a document file. Table 4 is the result of weighting with TF-IDF.

Table 4. Weighting results with TF-IDF

No	Variable X												Variable Y
	abc	absensi	access	achmad	aco	acute	adaboost	adaptif	adaptive	addictive	...	zero	Number_day
1	0	0	0	0	0	0	0	0	0	0	…	0	29
2	0	0	0	0	0	0	0	0	0	0	…	0	73
3	0	0	0	0	0	0	0	0	0	0	…	0	67
4	0	0	0	0	0	0	0	0	0	0	…	0	72
5	0	0	0	0	0	0	0	0	0	0	…	0	74
6	0	0	0	0	0	0	0	0	0	0	…	0	100
7	0	0	0	0	0	0	0	0	0	0	…	0	93
8	0	0	0	0	0	0	0	0	0	0	…	0	99
9	0	0	0	0	0	0	0	0	0	0	…	0	102
10	0	0	0	0	0	0	0	0	0	0	…	0	100
…	…	…	…	…	…	…	…	…	…	…	…	…	…
562	0	0	0	0	0	0	0	0	0	0	…	0	257
563	0	0	0	0	0	0	0	0	0	0	…	0	749

The results of the TF-IDF process obtained features that were also used as independent variables as many as 1665 independent variable features.

4.4 Data Prediction with Linear Regression

The prediction analysis process with the linear regression algorithm produces various coefficient values for each independent variable feature. The regression coefficients produced for each feature can be seen in Table 5 below.

Table 5. Regression coefficient of variable X

No.	Variable X	Regression Coefficient
1	abc	267.3335
2	absensi	-54.8471
3	access	246.4174
4	achmad	68.2539
5	aco	134.0115
6	acute	-246.2476
7	adaboost	-20.3050
8	adaptif	43.3842
9	adaptive	-85.0337
10	addictive	-133.2021
…	...	…
1664	zc	-24.9169
1665	zero	-24.9169

In Table 5, it can be seen that the coefficient value for the abc variable is positive with a value of 267.3335. This value means that assuming the other independent variables are ignored, if the independent variable in this case the abc variable increases by 1%, it can affect the completion time of students' theses by 267.3335. Furthermore, if the independent variable in this case the attendance variable increases by 1%, it can affect the completion time of students' theses by -54.8471 assuming the other independent variables are ignored. The same thing is also interpreted in other independent variables.

4.5 Results Analysis and Evaluation

At this stage, evaluation and analysis of the prediction results carried out in the previous stage are carried out. The results of the evaluation of the linear regression model can be seen in Table 6. Prediction of the completion time of student theses is carried out by dividing the data into training data and test data using the Hold-out technique with a ratio of 90% training data to create a prediction model with linear regression and 10% test data to test the linear regression model. The results of the prediction of the completion time of student theses based on the specified regression model are as follows.

Table 6. Prediction results with test data

No	Title	Actual Day	Prediction Day
1	implementation of chain code and learning vector quantization lvq for Japanese katakana character recognition	101	347
2	design of information security assessment system for hospital information technology infrastructure using ISO 27001 framework and case study of Arifin Achmad Regional Hospital	106	228
3	clustering of Indonesian soccer league player performance using the k means algorithm, case study of the 2016 Indonesian soccer championship	106	288
4	comparison of mean filter algorithms in reducing noise in digital images	117	175
5	A priori algorithm to determine the pattern of library book borrowing, case study of the library and archives service of Pekanbaru city	105	67
6	application of a combination of naive bayes and mknn methods for predicting land case decisions	143	-127
7	application of simple additive weighting saw in the selection of industrial plantation crops	103	119
8	application of MFCC and BPNN methods for recognizing Hijaiyah letters	175	385
9	digital al quran application to assist in memorizing al quran using google speech api	236	357
10	Android based daily prayer application with Google Speech API	233	312
…	…	…	…
55	application of classification and regression trees method for network attack classification case study of nsl kdd dataset	286	277
56	classification of bank health levels using modified k nearest neighbor and particle swarm optimization	251	310
57	classification of types of agarwood based on fruit texture and color using the adaptive neuro fuzzy interference system anfis method	261	304

The results of the prediction of the completion time of students' theses based on the existing regression model as in Table 6 are depicted in the graph in Figure 3.

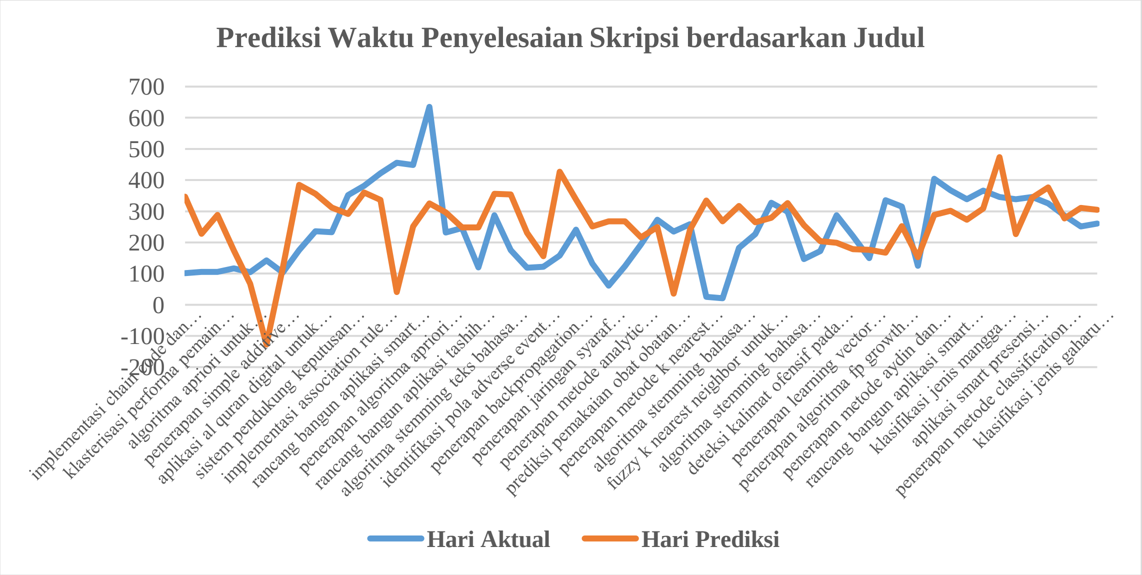

Figure 3. Prediction results with test data

Based on Figure 3, it can be concluded that there are various prediction results from the actual value. There are predictions that are close to the actual value, besides that there are also prediction results that produce negative values, where initially the actual day was positive.

Table 7. Prediction Evaluation Results

Evaluation Components	Evaluation Results
MAE (Mean Absolute Error)	105.17
MSE (Mean Squared Error)	16,331.50
RMSE (Root Mean Squared Error)	127.79
R² (R-Square)	-0.1337

Table 7 shows MAE (Mean Absolute Error): 105.17 indicates the average absolute difference between prediction and actual value is quite large. MSE (Mean Squared Error) and RMSE (Root Mean Squared Error): MSE of 16,331.50 and RMSE of 127.79 indicate that the model error is quite large. R² (R-Square): -0.1337 on the test data indicates that the model fails to explain the variability of the target data. Negative values indicate the model is no better than the simple average (mean) of the target data.

5 Conclusion

The findings of this study indicate that the linear regression model is ineffective in predicting the target variable, number_of_days. The model exhibits very low R² values in both training and testing datasets, with the test data even showing a negative R² value, indicating that the model performs worse than a simple mean-based prediction. Additionally, the predictor variable used, thesis_title, is statistically insignificant, suggesting that it does not meaningfully contribute to predicting thesis completion time. Given these limitations, it is crucial to explore more advanced machine learning techniques that can better capture complex relationships in the data. Random Forest and Gradient Boosting are particularly promising alternatives to linear regression due to their ability to handle non-linearity, feature interactions, and high-dimensional data. Random Forest is an ensemble learning method that constructs multiple decision trees and combines their outputs to enhance prediction accuracy. It is highly robust to overfitting and can capture complex, non-linear relationships that linear regression cannot model effectively. Additionally, Random Forest is less sensitive to irrelevant features, making it useful in scenarios where individual predictors, such as thesis title, may have weak direct effects but interact with other factors. On the other hand, Gradient

Boosting builds trees sequentially, with each tree correcting errors made by previous ones. This approach makes it highly effective for complex datasets, as it optimizes prediction performance by focusing on difficult-to-predict cases. Methods like XGBoost, LightGBM, or CatBoost further enhance efficiency and accuracy, making them ideal for identifying subtle patterns in thesis completion times. To improve predictive accuracy, future research should implement Random Forest and Gradient Boosting while incorporating additional variables that may have a greater influence on thesis completion time. Factors such as thesis difficulty level, student-related attributes (motivation, research experience, and academic background), and availability of resources (access to advisors, research materials, and institutional support) should be considered. These advanced machine learning techniques, combined with a richer feature set, can potentially yield more reliable predictions than linear models, leading to a better understanding of the factors affecting thesis completion time.

6 References

[1] P. A. Sanistasya *et al.*, "*Coaching Clinic Skripsi Hack* bagi Mahasiswa Administrasi Bisnis Universitas Mulawarman," *JMM (Jurnal Masyarakat Mandiri)*, vol. 7, no. 3, pp. 2577–2587, 2023, [Online]. Available: https://journal.ummat.ac.id/index.php/jmm/article/view/14069

[2] A. Lusi, A. P. Nalle, and K. R. Saba, "Hubungan Antara Kecemasan Akademik dengan Self-Efficacy pada Mahasiswa yang sedang menyusun Skripsi di Rumpun Ilmu Pendidikan FKIP Universitas Nusa Cendana," *Jurnal Bimbingan Konseling Flobamora*, vol. 1, no. 2, 2023, [Online]. Available: https://ejurnal.undana.ac.id/index.php/JBKF/article/view/12292

[3] A. Winyo, T. Trisno, and T. Kurra, "Analisis Algoritma Asosiasi untuk memilih Judul Mahasiswa Skripsi Stimkom Stella Maris Sumba," *Multidisciplinary Indonesian Center Journal (MICJO)*, vol. 1, no. 1, pp. 404–411, 2024, [Online]. Available: https://e-jurnal.jurnalcenter.com/index.php/micjo/article/view/46

[4] F. Marsela, A. Bakar, and R. Shopia, "Analisis Faktor Penyebab Keterlambatan Penyelesaian Studi pada Mahasiswa Prodi Bimbingan dan Konseling," *Syifaul Qulub: Jurnal Bimbingan dan Konseling Islam*, vol. 4, no. 1, pp. 46–53, Jul. 2023, doi: 10.32505/syifaulqulub.v4i1.6169.

[5] G. Maulana and R. D. Dana, "Prediksi Hasil Produksi Jagung di Jawa Barat dengan Metode Algoritma Regresi Linear menggunakan *Google Collab*," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 1, pp. 827–837, 2024, [Online]. Available: https://ejournal.itn.ac.id/index.php/jati/article/view/8816

[6] R. Ridlo Al-Hakim *et al.*, "*Predict the Thyroid Abnormality Particular Disease Likelihood of The Symptoms' Certainty Factor Value and its Confidence Level*: *A Regression Model Analysis*," 2023. [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[7] A. Pangestu *et al.*, "*Using Regression Model Analysis for Forecasting the Likelihood of Particular Symptoms of COVID-19*," *Sistemasi: Jurnal Sistem Informasi*, vol. 13, no. 1, pp. 167–176, 2024.

[8] S. Supardi *et al.*, "Peran Data Mining dalam memprediksi Tingkat Penjualan Sepatu Adidas menggunakan Metode Algoritma Regresi Linear Sederhana," *Jurnal Ekonomi Manajemen Sistem Informasi*, vol. 4, no. 5, pp. 883–890, 2023, [Online]. Available: https://dinastirev.org/JEMSI/article/view/1556

[9] R. Andrianto and F. Irawan, "Implementasi Metode Regresi Linear Berganda pada Sistem Prediksi Jumlah Tonase Kelapa Sawit di PT. Paluta Inti Sawit," *Jurnal Pendidikan Tambusai*, vol. 7, no. 1, pp. 2926–2936, 2023, [Online]. Available: https://jptam.org/index.php/jptam/article/download/5658/4751

[10] W. Andriani, G. Gunawan, and A. E. Prayoga, "Prediksi Nilai Emas menggunakan Algoritma Regresi Linear," *Jurnal Ilmiah Informatika Komputer*, vol. 28, no. 1, pp. 27–35, 2023, [Online]. Available: https://ejournal.gunadarma.ac.id/index.php/infokom/article/view/8096

[11] M. Edi, E. Utami, and A. Yaqin, "Prediksi Harga pada Trading Forex Pair USDCHF menggunakan Regresi Linear," *Jurnal Manajemen Informatika (JAMIKA)*, vol. 13, no. 2, pp. 109–119, 2023, [Online]. Available: https://ojs.unikom.ac.id/index.php/jamika/article/view/9826

[12] A. T. Nurani, A. Setiawan, and B. Susanto, “Perbandingan Kinerja Regresi *Decision Tree* dan Regresi Linear Berganda untuk Prediksi BMI pada Dataset Asthma,” *Jurnal Sains dan Edukasi Sains*, vol. 6, no. 1, pp. 34–43, 2023, [Online]. Available: https://ejournal.uksw.edu/juses/article/view/8438

[13] D. Yanti, Martanto, and A. Bahtiar, “Prediksi Hasil Panen Padi Tahun 2023 menggunakan Metode Regresi Linier di Kabupaten Indramayu,” *Jurnal Informatika Terpadu*, vol. 9, no. 1, pp. 18–23, Mar. 2023, doi: 10.54914/jit.v9i1.657.

[14] Y. Aqsho Ramadhan, A. Faqih, and G. Dwilestari, “Prediksi Penjualan Handphone di Toko X menggunakan Algoritma Regresi Linear,” *Jurnal Informatika Terpadu*, vol. 9, no. 1, pp. 40–44, Mar. 2023, doi: 10.54914/jit.v9i1.692.

[15] A. Wibowo, D. Iskandar, and W. A. S. Wibowo, “Data Mining dalam Prediksi Jumlah Pasien dengan Regresi Linear dan *Exponential Smoothing*,” *Jurnal Sistem Informasi dan Sains Teknologi*, vol. 5, no. 1, 2023, [Online]. Available: https://dirdosen.budiluhur.ac.id/0007097901/2022-1/B_Data_Mining_Dalam_Prediksi.pdf

[16] P. Herwanto, N. Marliani, and R. Rosida, “Prediksi Kinerja Keuangan PT Astra International Tbk dengan Regresi Linier dan *Exponential Smoothing*,” *Infotronik : Jurnal Teknologi Informasi dan Elektronika*, vol. 8, no. 1, p. 12, Jun. 2023, doi: 10.32897/infotronik.2023.8.1.2734.

[17] F. M. Sarimole and K. Kudrat, “Analisis Sentimen terhadap Aplikasi Satu Sehat pada Twitter menggunakan *Algoritma Naive Bayes* dan *Support Vector Machine*,” *Jurnal Sains dan Teknologi*, vol. 5, no. 3, pp. 783–790, 2024, [Online]. Available: http://ejournal.sisfokomtek.org/index.php/saintek/article/view/2702

[18] E. Miranda, V. Gabriella, S. A. Wahyudi, and J. Chai, “*Text Classification for Analysing Indonesian People’s Opinion Sentiment for Covid-19 Vaccination*,” *SISTEMASI*, vol. 12, no. 2, p. 438, May 2023, doi: 10.32520/stmsi.v12i2.2759.

[19] S. N. Cahyani and G. W. Saraswati, “*Implementation of Support Vector Machine Method in Classifying School Library Books with Combination of TF-IDF and Word2Vec*,” *Jurnal Teknik Informatika (Jutif)*, vol. 4, no. 6, pp. 1555–1566, Dec. 2023, doi: 10.52436/1.jutif.2023.4.6.1536.

[20] S.-W. Kim and J.-M. Gil, “*Research Paper Classification Systems based On Tf-Idf and Lda Schemes*,” *Human-Centric Computing and Information Sciences*, vol. 9, no. 1, p. 30, Dec. 2019, doi: 10.1186/s13673-019-0192-7.

[21] M. B. Gultom, P. A. Simbolon, and N. S. Nainggolan, “Prediksi Tingkat Pengangguran berdasarkan Pendidikan menggunakan Regresi Linear (Studi Kasus : Kota Medan),” 2024. [Online]. Available: https://www.kaggle.com/

[22] E. C. Sitohang, F. E. Ginting, and Y. M. B. Sembiring, “Prediksi Jumlah Perokok dan Dampaknya terhadap Kesehatan Masyarakat menggunakan Regresi Linear,” in *Seminar Nasional Inovasi Sains Teknologi Informasi Komputer*, 2024, pp. 512–516. [Online]. Available: https://ejournal.ust.ac.id/index.php/SNISTIK/article/view/3683

[23] V. R. Prasetyo, H. Lazuardi, A. A. Mulyono, and C. Lauw, “Penerapan Aplikasi *RapidMiner* untuk Prediksi Nilai Tukar Rupiah terhadap US Dollar dengan Metode Linear Regression,” *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 7, no. 1, pp. 8–17, May 2021, doi: 10.25077/TEKNOSI.v7i1.2021.8-17.

[24] D. H. Perkasa and M. Magito, “Determinan Faktor *Blue Economy* dalam Aplikasi Praktis SDM Perhotelan di Pulau Tidung Kepulauan Seribu,” *Jesya*, vol. 7, no. 1, pp. 840–852, Jan. 2024, doi: 10.36778/jesya.v7i1.1495.