Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 7, No. 1, March 2024, pp. 181 – 192
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Modeling The Prediction of Hard Drive Capacity Usage on Server Computers Based on Linear Regression

[1]Wahyuni, [2]Pitrasacha Adytia, [3]Siti Namira Rizqi Astin, [4]Kelik Sussolaikah, [5]Fadly Kasim
[1,3]Department of Informatics Engineering, STMIK Widya Cipta Dharma, Indonesia
[2]Department of Information System, STMIK Widya Cipta Dharma, Indonesia
[4]Department of Informatics Engineering, Universitas PGRI Madiun, Indonesia
[5]Department of Informatics, Universitas Muslim Indonesia, Indonesia
Email: [1]wahyuni@wicida.ac.id, [2]pitra@wicida.ac.id, [3]sitinamira@gmail.com,
[4]kelik@unipma.ac.id, [5]fadly.kasim@gmail.com

Article Info
Article history:
Received Jan 12th, 2024
Revised Feb 27th, 2024
Accepted Mar 17th, 2024

Keyword:
CRISP-DM
Hard Drive Capacity
Linear Regression
Machine Learning
Prediction

ABSTRACT
Bank of XYZ has a server computer that is used to run several information technology application services such as ATMs and others. Because the server computer uses a hard drive, the full hard drive can cause problems with the service not operating properly. Full hard drives occur without being noticed. So that this makes the computer server problematic, resulting in customer dissatisfaction and decreased customer loyalty to Bank XYZ. To solve the problem at XYZ Bank, one of the machine learning algorithms can be used to predict hard drive capacity. The method used to predict hard drive storage or usage. The machine learning algorithm used is Multiple Linear Regression. The results of this study show that the linear regression model successfully predicts the use of hard drive capacity on server computers with a sufficient level of accuracy.But it is still not optimal because only a few servers can be predicted. For further research, may consider using the LSTM (Long Short-Term Memory) algorithm. LSTM is an algorithm that is well-suited for sequence prediction problems, including time series forecasting.

Corresponding Author:
Wahyuni,
Departement of Informatics Engineering,
STMIK Widya Cipta Dharma,
Jl. M.Yamin, No. 25 Samarinda, Kalimantan Timur, Indonesia
Email: wahyuni@wicida.ac.id
DOI: http://dx.doi.org/10.24014/ijaidm.v7i1.28851

1. INTRODUCTION (10 PT)

Prediction is the process of systematically estimating events that are likely to occur in the future, based on past and current information available, to minimize error (the difference between actual events and predicted results). Prediction does not have to provide a clear answer to what will happen, but tries to find an answer that is as close as possible to what will happen [1]

A bank is a business entity that collects funds from the public in the form of deposits and distributes them to the public in the form of credit or other forms in order to improve the lives of the people [2]. Bank XYZ has a server computer that is used to run several information technology application services such as ATMs and others. Because the server computer uses a hard drive, the full hard drive can cause problems with the service not operating properly. Full hard drives occur without being noticed. So that this makes the computer server problematic, resulting in customer dissatisfaction and decreased customer loyalty to Bank XYZ.

Hard drives are high-capacity data storage media that include Mega Bytes (MB), Giga Bytes (GB), and Tera Bytes (TB) [3]. Hard disk is one of the data storage media on a computer which consists of a collection of hard and rotating magnetic disks, as well as other electronic components [4].

To solve the problem at XYZ Bank, one of the machine learning algorithms can be used to predict hard drive capacity. The method used to predict hard drive storage or usage at a future time is to use the

linear regression method and use cloud-based and free tools, namely Google Collaboratory and the Python programming language.

Machine Learning is a field of computer science that involves the study and construction of techniques that enable computers to self-learn based on input data to solve specific problems [5]. Regression is a technique used for two theories. First, regression analyzes are usually used for forecasting and prediction, in which their application has major overlaps with the area of machine learning. Second, regression analysis can be used in some cases to determine causal relations between the independent and dependent variables. Importantly, regressions alone show only relations between a dependent variable and a fixed dataset collection of different variables [6].

The multiple linear regression model is the most commonly applied statistical technique for relating a set of two or more variables [7]. The multiple linear regression (MLR) method is used to predict the yield, because its variability is determined by many independent variables [8] Multiple linear regression analysis predicts changes in the value of certain variables when other variables change. It can be said as various regression because the independent variable is used as a predictor of more than one [9].

Python is an interpretive, object-oriented and semantically dynamic programming language. Python has high-level data structures, dynamic typing and dynamic binding, has a simple syntax and is easy to learn[10]. Python modules have become increasingly popularin astronomy for data analysis. Google CoLaboratory (Google CoLab) is a powerful collaborative tool for coding in Python [11].

Based on these problems, it is necessary to conduct a prediction analysis to predict the use of hard drive storage and hard drive conditions using the Multiple Linear Regression algorithm. So that it is expected to explain how to predict the use of hard drive capacity on the XYZ bank server computer which can minimize problems in server computers which ultimately increase customer satisfaction and customer loyalty at Bank XYZ.

Research that discusses the prediction of hard drives has also been conducted. [12], [13]. However, these studies only predict errors or problems that occur in hard drives. Meanwhile, predicting hard drive capacity in the future has never been done. Where in this research we can find out the prediction of hard drive capacity in the next month. Through this, it will be possible to minimize the problems that will occur and have certain actions prepared if needed.

The main contributions of the study can be summarized as follows: (1) One of the machine learning algorithms can be used to predict hard drive capacity; (2) The method used to predict hard drive storage or usage at a future time is to use the linear regression method; (3) Predicting hard drive capacity in the future has never been done; (4) In this research we can find out the prediction of hard drive capacity in the next month.

2. RESEARCH METHOD

The research methodology used is Cross Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM is an industry-independent process model for data mining. It consists of six iterative phases from business understanding to deployment [14]. CRISP-DM, which is the most popular framework teams use to execute data science projects, provides an easy to understand description of the data science project workflow (i.e., the data science life cycle) [15]. A widely applied method in data mining is CRISP-DM. CRISP-DM (Cross-Industry Standard Process Model for Data Mining) describes the data mining process in six stages, namely: Business Understanding; Data Understanding; Data Preparation; Modeling; Evaluation; Deployment (figure 1) [16].

Here are the stages of the CRISP- DM process as shown in Figure 1:

2.1. Business Understanding

Is the process of understanding the characteristics of the business to be analyzed in depth [17]. Business Understanding is the process of determining business objectives, understanding the situation and conditions at the time of the research and setting a goal of the research conducted into problems that are solved by data mining [18]. This research is directly related to the data provided to see the condition of the problems that occur and have an impact on Bank XYZ.

2.2. Data Understanding

Data understanding includes activities to prepare, assess data needs, and includes data collection [19]. At the data understanding stage, there are several things that are done, including collecting initial data, describing data, exploring data, and verifying data quality. In this study using a dataset in the form (.csv). The attributes in the dataset are Unnamed, Id, HostId, ServerName, Address, Alias, Date, CPU, RamFree, RamFreeGB, RamFreePercentage, RamUsed, RamUsedGB, RamUsedPercentage, RamTotal, RamTotal GB,

RamStatus, DriveFree, DriveFree GB, DriveFree Percentage, DriveUsed, DriveUsedGB, DriveTotal, DriveTotalGB, DriveTotalPercentage, DriveStatus, Ping.

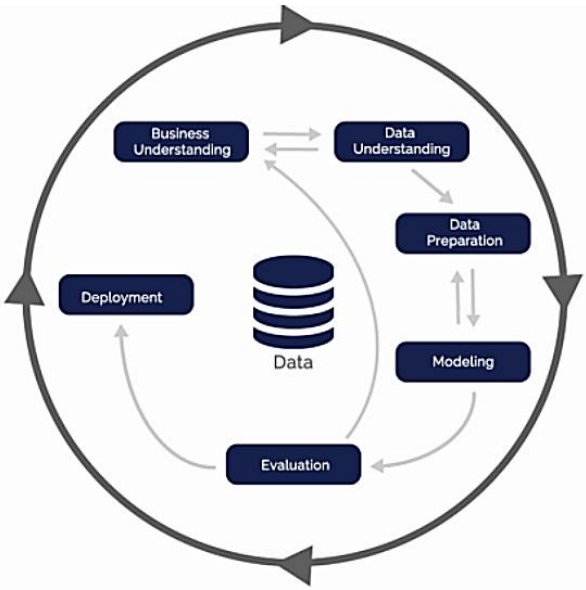

Figure 1. Crisp-DM Methodology

2.3. Data Preparation

Data preparation may be one of the most difficult steps in any machine learning project. The reason is that each dataset is different and highly specific to the project [20]. The data preparation phase generally includes three steps: Data cleaning, i.e., handling missing data and outliers, data reduction, i.e., reducing the data size by aggregation, elimination redundant feature, etc. and data normalization [21]. In this stage, building the final data set from raw data. There are several things that will be done including data cleaning, data selection, records and attributes, and data transformation to be used as input in the modeling stage.

2.4. Modelling

The model building stage involves selecting appropriate modeling techniques and applying them to the prepared data, to meet specific business objectives [22]. Selection of data mining techniques, algorithms and determining parameters with optimal values. The prediction patterns generated by this data mining technique are used to predict the time of full use of hard drive capacity. This stage displays and provides information on the performance of linear regression. By using a linear regression model which is one of the methods in statistics used to describe the linear relationship between one or more independent variables (known as "predictor variables") and a dependent variable (known as "response variable"). The main objective of linear regression is to construct a straight-line equation that best represents the relationship between these variables.

2.5. Evalution

This stage is to evaluate the performance of the data model and check all the processes that have been carried out, to ensure that no data and stages have been missed[23]. Evaluation is carried out in depth with the aim that the results at the modeling stage are in accordance with the goals to be achieved in the business understanding stage. At the evaluation stage, there are several things that are done, among others: Evaluate the results, Review the process, Determine the next step. This stage is carried out after all models are made, and the program can run, where all software, additional programs, and all programs involved in building the system are tested to ensure the system can run according to the design or not. To calculate the amount of error predicting the use of hard disk capacity on server computers. The resulting output is the accuracy of predicting the use of hard drive capacity. The following is the evaluation formula used:

1. MeanSquare Error (MSE)
   MSE is an alternative method for evaluating individual forecasting techniques. The MSE method is a useful indicator and provides absolute values as opposed to relative information in the MAPE method[24].

2. Root Mean Squared Error (RMSE)
   Root Mean Square Error (RMSE) is a measure that is often used to find the difference between the predicted values in the model[25]. In simple terms, RMSE is a method to calculate the bias in a forecasting model. The accuracy of the estimation measurement is indicated by the RMSE result having a small value (close to zero)[26].

3. Mean Absolute Percentage Error (MAPE)
   Mean Absolute Percentage Error (MAPE) is a measure of relative error[27]. The use of the mean absolute percentage error (MAPE) as a measure of forecast accuracy should be avoided because they argue it treats forecast errors above the actual observation differently from those below this value. The smaller the MAPE value, the closer the estimated value is to the true value[28].

2.6. Deployment

The last stage of CRISP-DM is deployment, where the evaluated model is embedded in a user interface for easy use. However, in this research, the stages carried out only up to the modeling and evaluation of the model. Starting with business understanding where observation of the problem is carried out, then preparing tools, importing libraries and preparing data sets for the data understanding stage. at the data understanding stage, data collection, reading data, exploring data and verifying data are carried out. then enter the data preparation stage where at this stage attribute selection, data transformation, and splitting data as much as 80% training data and 20% setting data. Then the model training is carried out and continued with model evaluation. after completion, the prediction model can be used.

3. RESULTS AND ANALYSIS

In accordance with the methodology used, it will be explained in detail the stages carried out in this research. Which starts from Business Understanding to the evaluation stage. The research conducted only reaches the evaluation stage, where at that stage the final results are obtained regarding the model created whether it can predict the hard disk capacity properly or not.

3.1. Business Understanding

At this stage, we identify the problems that occur, what factors influence the prediction of hard drive usage storage and hard drive conditions using existing data. Checking the relationship between each attribute in the data that can be used as a reference to predict hard drive capacity so as to maximize understanding of existing data.

Customers want to get services that are always active in the use of ATMs so that they can access their accounts to carry out various financial transactions, namely cash withdrawal transactions and non-cash transactions, such as checking balances, paying credit card bills, paying electricity bills, buying credit, and so on. The use of ATMs at Bank XYZ offers convenience and speed in conducting banking transactions. The systemized work of the two products is due to previous hard work, resulting in good service and providing convenience to customers and service providers in making transactions.

All banks certainly hope that customers are always satisfied with the services provided and always hope to serve customers without any obstacles and problems when making transactions at ATMs. But what happened was not as expected by Bank XYZ because there was always an error in making transactions at the ATM and could not be handled because the hard drive was full without any prior warning so that the ATM could not make transactions. As a result, customer dissatisfaction with bank services occurs, customer trust in the bank decreases and loyalty to the bank also decreases. So that researchers want to provide solutions to Bank XYZ to deal with these problems by predicting hard drive capacity using existing data. Therefore, the business objective of this research is to design a machine learning model that can be used as a predictor. The results of this prediction are expected to help management in making decisions or policies to improve Bank XYZ's services to customers.

3.2. Data Understanding

At this stage, data collection is first carried out, all the necessary data will later be processed thoroughly. The original dataset contains 26 attributes, but later feature selection will be carried out. In this study, the data collected has 91 server computers, starting from the end of 2018 to 2021.

Next, the library import stage is carried out. Import pandas as a python library for data analysis, NumPy (numerical operations), Matplotlib / Seaborn (plotting or data visualization), Scikit-Learn (machine learning and evaluation models). Can be seen in Figure 2 tools and libraries.

```
import datetime as dt
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import math
sns.set()

%matplotlib inline

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_percentage_error, mean_squared_error
```

Figure 2. Import Library

There are 91 server computers and 6 drive statuses which include OK, WARNING, CRITICAL, REQUEST, TABLE, and check. After obtaining the drive status, it was decided to use only 3 statuses, namely OK, WARNING and CRITICAL. As well as changing the drive status into numeric numbers to make it easier to correlate, such as OK = 0, WARNING = 1, CRITICAL = 2. For column 0 or OK is a status that states that the drive is safe, 1 or WARNING is a status that indicates that there are some potential problems that need to be considered on the ram and drive. 2 or CRITICAL is a status that refers to damage or problems with the drive. The plot() function to see the data comparison between drive status and drive used percentages using barcharts, the results of the plot() function show that hard disk storage shows a warning status if it reaches 80%, and critical >90%. There are 2705 datasets whose hard disk storage usage reaches >80%.

3.3. Data Preparation

Server computer data that will be used are:

1. Apu-conven, The data used from this server is from 30-12-2018 until it reaches warning on 12-04-2019 and the period of time to reach warning takes 102 days. Can be seen in Figure 3 below.

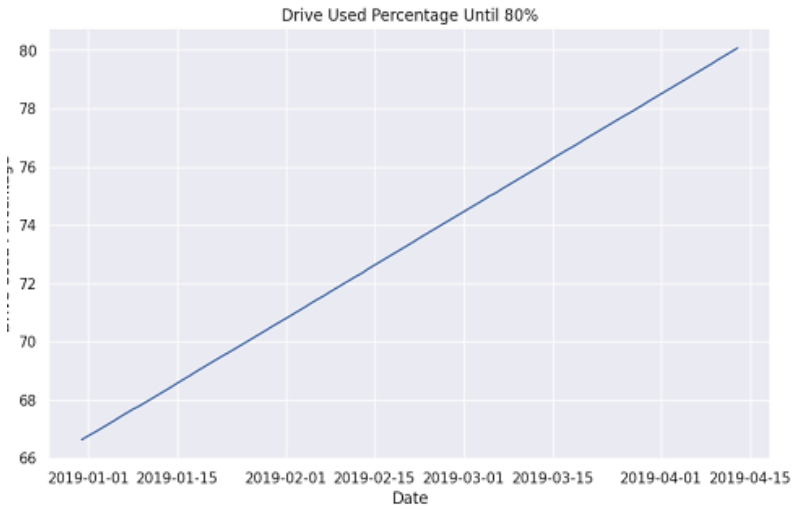

Figure 3. Apu-Convenient Capacity Graph

2. Server-Helpdesk-SLIK-DRC, The data used from this server is from 01-01-2020 until it reaches warning on 17-06-2020 and the period of time to reach critical requires 132 days. Can be seen in Figure 4 below.

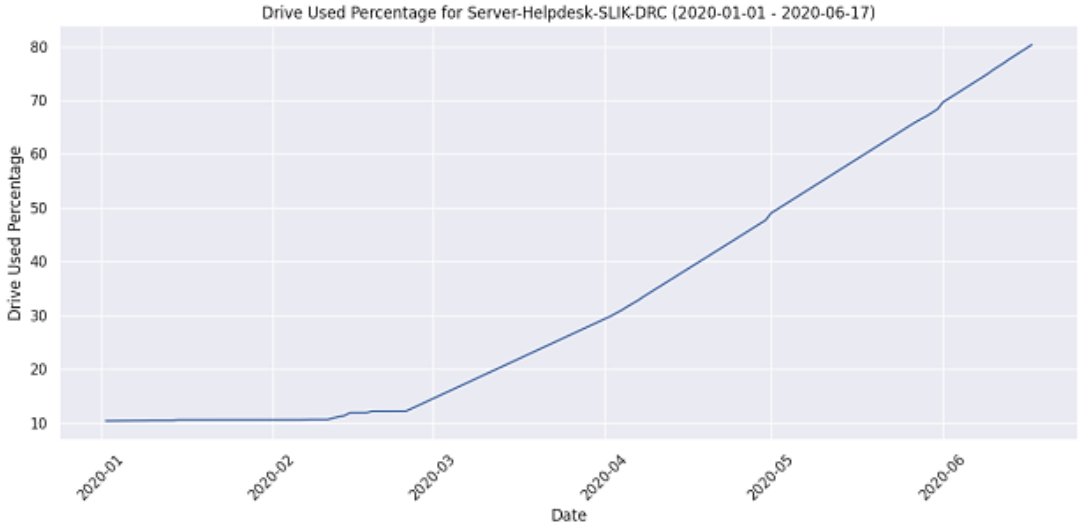

Figure 4. Capacity Chart of Server-Helpdesk-SLIK-DRC

3. Graphon198_BSSB3, The data used from this server is from 06-11-2020 until it reaches critical on 30-05-2021 and the period of time to reach critical takes 718 days. Can be seen in Figure 5 below.

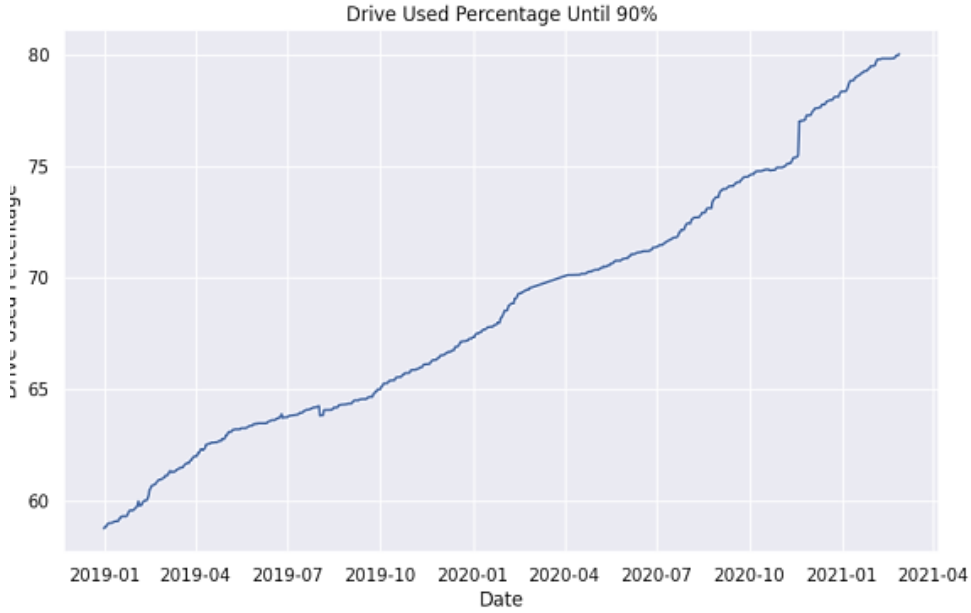

Figure 5. Capacity graph of graphon198_BSSB3

4. Interface-sknsy-dc, The data used from this server is from 06-11-2020 until it reaches critical on 30-05-2021 and the period of time to reach critical takes 205 days. Can be seen in Figure 6.

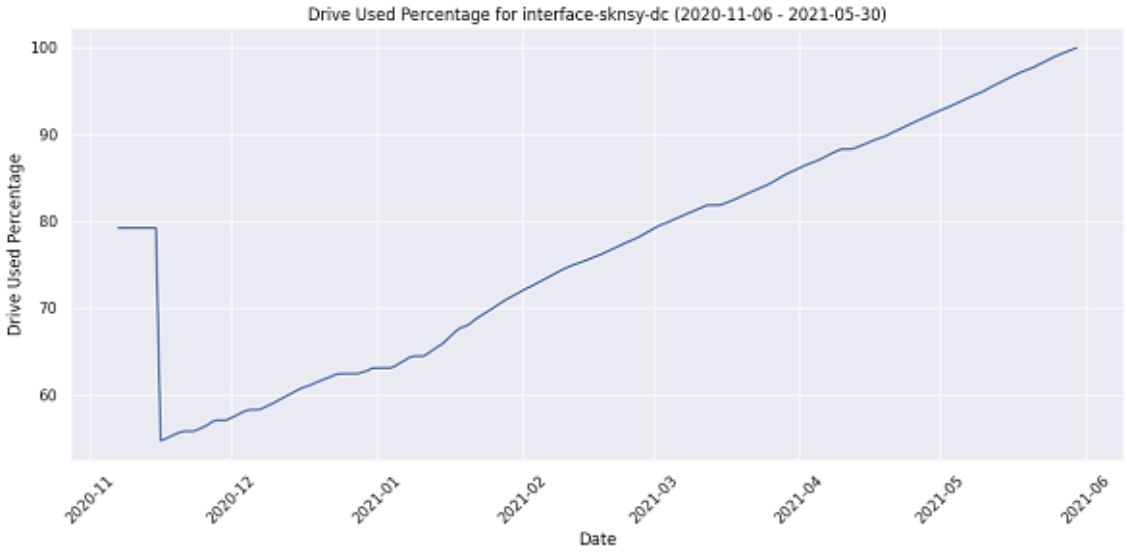

Figure 6. Capacity graph of interface-sknsy-dc

5. Sknsybackup, The data used from this server is from 01-11-2020 until it reaches critical on 09-03-2021 and the period of time to reach critical takes 126 days. Can be seen in Figure 7.

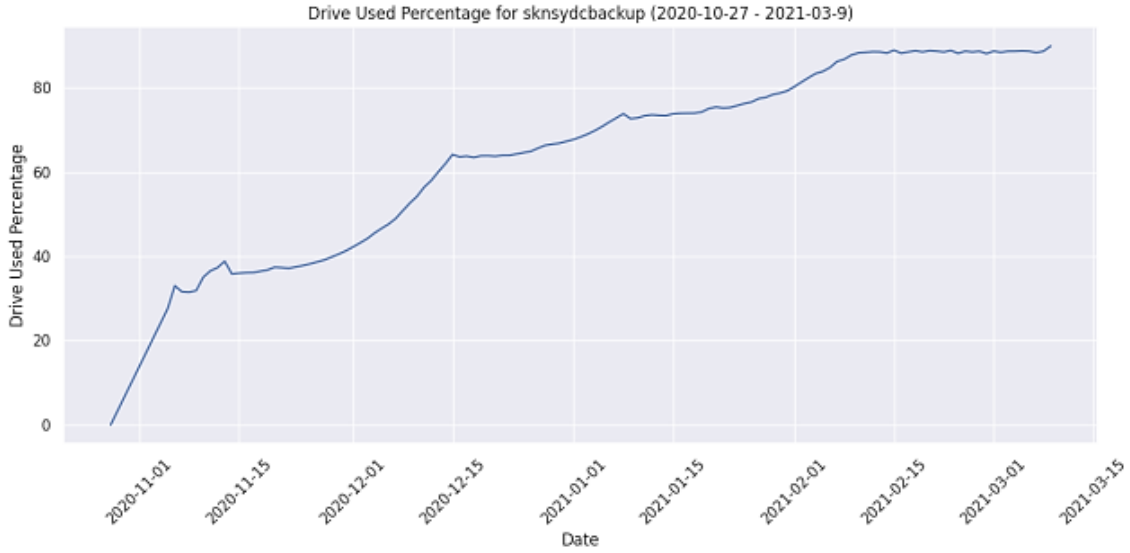

Figure 7. Sknsybackup Capacity Chart

3.4. Modelling

After exploring the data and doing data preparation, then splitting the data, where variable X (independent) is the number of days, and y (dependent) is the actual data from the drive used percentage. Will be divided as much as 80% of the data into training data and 20% of the data into test data. After the data is divided, it will be continued by entering the Linear Regression algorithm. The regressor.score(X_test, y_test) function is used to measure the performance of the linear regression model on the test data (X_test and y_test). Specifically, it calculates the coefficient of determination (R-squared) of the model against the test data. The coefficient of determination (R-squared) is an evaluation metric that provides information on how well the regression model fits the given data. R-squared values range between 0 and 1, and the higher the R-squared value, the better the model can explain the variation in the data. The results of the test data on each server can be seen in table 1 R-squared test data results.

Table 1. R-squared test data results.

No	Server Name	*R-squared*	Description
1.	Apu-Konven	0.9999945895879	Regression models have a very high ability to explain variations in data.
2.	Server-Help desk-SLIK-DRC	0.4957753047320179	the regression model is able to explain about 49.58% of the variation in the data.
3.	Graphon198_BSSB3	0.9814623488360749	the regression model is able to explain about 98.15% of the variation in the data
4.	Interface-sknsy-dc	0.9401976563092542	the regression model is able to explain about 94.02% of the variation in the data.
5.	sknsydcbackup	-0.16672348358719424	Invalid

After looking at the results from using the R-squared score function. The majority of the server computers had a good fit with the linear regression model, indicated by the close R-squared scores. This indicates that the linear regression model is able to well explain the variation in drive usage data and drive status on these server computers. However, the sknsydcbackup server computer had a poorer fit with the linear regression model, as shown by the lower R-squared score. This suggests that the linear regression model was not able to effectively explain the variation in the data on these server computers. This could be due to various factors, such as the presence of a non-linear relationship between features and targets. Therefore, we did not proceed to model the sknsydcbackup server. The above results show the importance of evaluating the linear regression model for each server computer individually. By evaluating the R-squared score, we can identify the server computers that fit the model and those that do not, and know the extent to which the model can explain the variation in the data on each server computer.

y_pred = regressor.predict(X_test) is used to predict the target (dependent) value based on the pre-trained linear regression model. The resulting prediction of the amount of hard drive capacity in the next 30 days for each server has a good R-Squared score. Here are the results for each server:

Table 2. Apu-Konven Server Prediction Results

No	Date	Result Prediction
0	2021-06-17	79.936423
1	2021-06-18	80.066901
2	2021-06-19	80.197379
3	2021-06-20	80.327857
4	2021-06-21	80.458335
5	2021-06-22	80.588812
6	2021-06-23	80.719290
7	2021-06-24	80.849768
8	2021-06-25	80.980246
9	2021-06-26	81.110724
10	2021-06-27	81.241201
11	2021-06-28	81.371679
12	2021-06-29	81.502157
13	2021-06-30	81.632635
14	2021-07-01	81.763113
15	2021-07-02	81.893591
16	2021-07-03	82.024068
17	2021-07-04	82.154546
18	2021-07-05	82.285024
19	2021-07-06	82.415502
20	2021-07-07	82.545980
21	2021-07-08	82.676457
22	2021-07-09	82.806935
23	2021-07-10	82.937413
24	2021-07-11	83.067891
25	2021-07-12	83.198369
26	2021-07-13	83.328847
27	2021-07-14	83.459324
28	2021-07-15	83.589802
29	2021-07-16	83.720280

Table 2 shows the Apu-Konven Server Prediction Results for the next 30 days. Where the result is 83.7% which means it is in WARNING status.

Table 3. Prediction Results Server-Helpdesk-SLIK-DRC

No	Date	Result Prediction
0	2021-06-17	56.537438
1	2021-06-18	56.976333
2	2021-06-19	57.415228
3	2021-06-20	57.854123
4	2021-06-21	58.293017
5	2021-06-22	58.731912
6	2021-06-23	59.170807
7	2021-06-24	59.609701
8	2021-06-25	60.048596
9	2021-06-26	60.487491
10	2021-06-27	60.926386
11	2021-06-28	61.365280
12	2021-06-29	61.804175
13	2021-06-30	62.243070
14	2021-07-01	62.681964
15	2021-07-02	63.120859
16	2021-07-03	63.559754
17	2021-07-04	63.998649
18	2021-07-05	64.437543
19	2021-07-06	64.876438
20	2021-07-07	65.315333
21	2021-07-08	65.754227
22	2021-07-09	66.193122
23	2021-07-10	66.632017
24	2021-07-11	67.070912
25	2021-07-12	67.509806
26	2021-07-13	67.948701
27	2021-07-14	68.387596
28	2021-07-15	68.826490
29	2021-07-16	69.265385

Table 3 shows the Server-Helpdesk-SLIK-DRC Prediction Results for the next 30 days. Where the result is 69.3% which means it is in OK status.

Table 4. Servergraphon198_BSSB3 Prediction Results

No	Date	Result Prediction
0	2021-06-17	76.661737
1	2021-06-18	76.686470
2	2021-06-19	76.711203
3	2021-06-20	76.735936
4	2021-06-21	76.760669
5	2021-06-22	76.785402
6	2021-06-23	76.810135
7	2021-06-24	76.834868
8	2021-06-25	76.859600
9	2021-06-26	76.884333
10	2021-06-27	76.909066
11	2021-06-28	76.933799
12	2021-06-29	76.958532
13	2021-06-30	76.983265
14	2021-07-01	77.007998
15	2021-07-02	77.032731
16	2021-07-03	77.057463
17	2021-07-04	77.082196
18	2021-07-05	77.106929
19	2021-07-06	77.131662
20	2021-07-07	77.156395
21	2021-07-08	77.181128
22	2021-07-09	77.205861
23	2021-07-10	77.230594
24	2021-07-11	77.255326
25	2021-07-12	77.280059
26	2021-07-13	77.304792
27	2021-07-14	77.329525
28	2021-07-15	77.354258
29	2021-07-16	77.378991

Table 4 shows the Servergraphon198_BSSB3 Prediction Results for the next 30 days. Where the result is 77.4% which means it is in OK status.

3.5. Evaluation

Next, test the results of the error value that researchers have done on the drive used percentage.

Table 5. Apu-Conven Error Value Testing Results

Evaluation	Evaluation Result
Mean Squared Error (MSE)	9.014501807998079e-05
Root Mean Squared Error (RMSE)	0.00949447302802956
Mean Absolute Percentage Error (MAPE)	0.00010750744288971825

The resulting MSE value is about 9.014501807998079e-05. It shows that the model has a very small difference between the prediction and the actual value, which indicates good prediction quality. The RMSE value is about 0.00949447302802956, the lower the RMSE value, the better the model is at predicting the data. The MAPE value is about 0.000107, indicating that the average prediction error in percentage terms is very small, almost close to zero.

Table 6. Results of Server-Helpdesk-SLIK-DRC Error Value Testing

Evaluation	Evaluation Result
Mean Squared Error (MSE)	35.08701352448341
Root Mean Squared Error (RMSE)	5.923429203129165
Mean Absolute Percentage Error (MAPE)	0.3029661778564563

The MSE value is about 35.0870. If the MSE is high, it means that the prediction and real data are quite different. The RMSE value is about 5.9234. The MAPE value is about 0.3029, which means that the average prediction error in percentage terms is about 30.29%. This shows that the average prediction error in percentage terms is relatively large.

Table 7. Error Value Testing Results graphon198_BSSB3.

Evaluation	Evaluation Result
Mean Squared Error (MSE)	0.6537686037579203
Root Mean Squared Error (RMSE)	0.808559585780739
Mean Absolute Percentage Error (MAPE)	0.010071482053369055

The MSE value is about 0.6538, the RMSE value is about 0.8086, and the MAPE value is about 0.0101. The results of these evaluation metrics indicate that the predictions have a low error rate. The low MSE, RMSE, and MAPE values indicate that the prediction is quite close to the true value and has a good level of accuracy.

4. CONCLUSION

Of the 91 server computers, only a few can be predicted, for example in the case of the server computer "sknsydcbackup" produces an invalid R-squared, so it cannot be continued to the modeling stage. The model can predict hard drive capacity for the next 30 days, but after evaluation, the Apu-Konven and graphon198_BSSB3 servers have the best accuracy and have a small error rate. As for Server-Helpdesk-SLIK-DRC, it needs to be reviewed. Some servers may have more predictable characteristics, while others may be affected by other factors not covered in the model. This confirms the importance of understanding the unique characteristics of each server in developing a more accurate and reliable prediction model. And in other words, hard drive capacity prediction modeling using machine learning using Linear Regression algorithm is successful and provides results that vary from server to server. Although the Linear Regression algorithm successfully predicts hard disk capacity, it is still not optimal because only a few servers can be predicted. Therefore, further research is needed on this topic. For further research, you may consider using the LSTM (Long Short-Term Memory) algorithm. LSTM is an algorithm that is well-suited for sequence prediction problems, including time series forecasting.

REFERENCES

[1] F. D. Putra, Psikojurnalistik Analisis Wacana Kompas.Com Tentang Perilaku Komunikasi Pembuat Kebijakan Kabinet I. Guepedia, 2021.
[2] P. P. Indonesia, : Undang-Undang Tentang Perubahan Atas Undang-Undang Nomor 7 Tahun 1992 Tentang Perbankan. Indonesia: LN. 1998/ No. 182, TLN NO. 3790, LL SETNEG : 32 HLM, 1998.
[3] Widada, Kitab Teknisi Komputer, Laptop, Printer, dan Monitor untuk Pemula. Yogyakarta: MediaKom, 2014.
[4] S. Farizy and E. S. Harianja, "Pengembangan Media Penyimpanan dalam Sistem Berkas," Jurnal Ilmu Komputer, vol. 3, no. 2, pp. 5–9, Apr. 2020.

[5] X. D. Hoang and N. T. Nguyen, "Detecting Website Defacements Based on Machine Learning Techniques and Attack Signatures," Computers, vol. 8, no. 2, p. 35, May 2019, doi: 10.3390/computers8020035.
[6] 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. 4, pp. 140–147, Dec. 2020, doi: 10.38094/jastt1457.
[7] J. D. Jobson, "Multiple Linear Regression," in Applied Multivariate Data Analysis: Regression and Experimental Design, J. D. Jobson, Ed., New York, NY: Springer New York, 1991, pp. 219–398. doi: 10.1007/978-1-4612-0955-3_4.
[8] M. Piekutowska et al., "The Application of Multiple Linear Regression and Artificial Neural Network Models for Yield Prediction of Very Early Potato Cultivars before Harvest," Agronomy, vol. 11, p. 885, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:235567813
[9] D. Alita, A. D. Putra, and D. Darwis, "Analysis of classic assumption test and multiple linear regression coefficient test for employee structural office recommendation," IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:238781028
[10] P. Purwanto and S. Sumardi, "Perancangan Klasifikasi Tanaman Herbal Menggunakan Transfer Learning Pada Algoritma Convolutional Neural Network (CNN)," Jurnal Ilmiah Infokam, vol. 18, no. 2, pp. 105–118, Dec. 2022, doi: 10.53845/infokam.v18i2.328.
[11] K. Tock, "Google CoLaboratory as a Platform for Python Coding with Students," in RTSRE Proceedings, Our Solar Siblings, Dec. 2019. doi: 10.32374/rtsre.2019.013.
[12] J. Li, R. J. Stones, G. Wang, X. Liu, Z. Li, and M. Xu, "Hard drive failure prediction using Decision Trees," Reliab Eng Syst Saf, vol. 164, pp. 55–65, Aug. 2017, doi: 10.1016/j.ress.2017.03.004.
[13] J. Li et al., "Hard Drive Failure Prediction Using Classification and Regression Trees," in 2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, IEEE, Jun. 2014, pp. 383–394. doi: 10.1109/DSN.2014.44.
[14] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," Procedia Comput Sci, vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.
[15] J. S. Saltz, "CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps," in 2021 IEEE International Conference on Big Data (Big Data), IEEE, Dec. 2021, pp. 2337–2344. doi: 10.1109/BigData52589.2021.9671634.
[16] Y. Suhanda, I. Kurniati, and S. Norma, "Penerapan Metode Crisp-DM Dengan Algoritma K-Means Clustering Untuk Segmentasi Mahasiswa Berdasarkan Kualitas Akademik," Jurnal Teknologi Informatika dan Komputer, vol. 6, no. 2, pp. 12–20, Sep. 2020, doi: 10.37012/jtik.v6i2.299.
[17] Y. A. Singgalen, "Analisis Sentimen dan Sistem Pendukung Keputusan Menginap di Hotel Menggunakan Metode CRISP-DM dan SAW," Journal of Information System Research (JOSH), vol. 4, no. 4, pp. 1343–1353, Jul. 2023, doi: 10.47065/josh.v4i4.3917.
[18] F. N. Dhewayani, D. Amelia, D. N. Alifah, B. N. Sari, and M. Jajuli, "Implementasi K-Means Clustering untuk Pengelompokkan Daerah Rawan Bencana Kebakaran Menggunakan Model CRISP-DM," Jurnal Teknologi dan Informasi, vol. 12, no. 1, pp. 64–77, Mar. 2022, doi: 10.34010/jati.v12i1.6674.
[19] A. Pambudi, "PENERAPAN CRISP-DM MENGGUNAKAN MLR K-FOLD PADA DATA SAHAM PT. TELKOM INDONESIA (PERSERO) TBK (TLKM) (STUDI KASUS: BURSA EFEK INDONESIA TAHUN 2015-2022)," Jurnal Data Mining dan Sistem Informasi, vol. 4, no. 1, p. 1, Mar. 2023, doi: 10.33365/jdmsi.v4i1.2462.
[20] J. Brownlee, Data Preparation For Machine Learning, V1.2. 2020.
[21] O. Masmoudi, M. Jaoua, A. Jaoua, and S. Yacout, "Data Preparation in Machine Learning for Condition-based Maintenance," Journal of Computer Science, vol. 17, no. 6, pp. 525–538, Jun. 2021, doi: 10.3844/jcssp.2021.525.538.
[22] R. Pratama, M. I. Herdiansyah, D. Syamsuar, and A. Syazili, "Prediksi Customer Retention Perusahaan Asuransi Menggunakan Machine Learning," Jurnal Sisfokom (Sistem Informasi dan Komputer), vol. 12, no. 1, pp. 96–104, Mar. 2023, doi: 10.32736/sisfokom.v12i1.1507.
[23] Y. Christian and K. O. Y. R. Qi, "Penerapan K-Means pada Segmentasi Pasar untuk Riset Pemasaran pada Startup Early Stage dengan Menggunakan CRISP-DM," JURIKOM (Jurnal Riset Komputer), vol. 9, no. 4, p. 966, Aug. 2022, doi: 10.30865/jurikom.v9i4.4486.
[24] S. Fachrurrazi, "PERAMALAN PENJUALAN OBAT MENGGUNAKAN METODE SINGLE EXPONENTIAL SMOOTHING PADA TOKO OBAT BINTANG GEURUGOK," TECHSI Teknik Informatika, vol. 6, no. 1, pp. 19–30, Apr. 2015.
[25] P. Goodwin and R. Lawton, "On the asymmetry of the symmetric MAPE," Int J Forecast, vol. 15, no. 4, pp. 405–408, Oct. 1999, doi: 10.1016/S0169-2070(99)00007-2.
[26] I. P. Sutawinayam, I. N. G. A. Astawa, and N. K. D. Hariyanti, "PERBANDINGAN METODE JARINGAN SARAF TIRUAN PADA PERAMALAN CURAH HUJAN," Logic : Jurnal Rancang Bangun dan Teknologi, vol. 17, no. 2, pp. 92–97, Oct. 2017.
[27] F. Ahmad, "PENENTUAN METODE PERAMALAN PADA PRODUKSI PART NEW GRANADA BOWL ST Di PT.X," JISI: Jurnal Integrasi Sistem Industri, vol. 7, no. 1, p. 31, May 2020, doi: 10.24853/jisi.7.1.31-39.
[28] D. Purwanti and J. Purwadi, "Metode Brown's Double Exponential Smoothing dalam Peramalan Laju Inflasi di Indonesia," Jurnal Ilmiah Matematika, vol. 6, no. 2, p. 54, Oct. 2019, doi: 10.26555/konvergensi.v6i2.19548.

BIBLIOGRAPHY OF AUTHORS

Wahyuni, S.Kom., M.Kom. is a lecturer at the Informatics Engineering study program at STMIK Widya Cipta Dharma. the author is interested in machine learning and data science.

Pitrasacha Adytia, S.T., M.T. is a lecturer in the Information Systems study program at STMIK Widya Cipta Dhrama. The author is very interested in researching and writing in the fields of Data Science, Intelligent Systems and Machine Learning.

Siti Namira Rizqi Astin is a final year student of the informatics engineering study program at stmik widya cipta dharma. the author is interested in researching machine learning.

Kelik Sussolaikah, S.Kom., M.Kom. is a lecturer in the Informatics Engineering study program at Universitas PGRI Madiun. The author is very interested in researching and writing in the fields of Data Science.

Fadly Kasim, S.T., M.Kom. is a lecturer in the Informatics study program at Universitas Muslim Indonesia. The author is veri interested in researching and writing in the fields of Internet of Things.