

# Machine Learning Implementation for Profit Estimation

Alexander Dharmawan<sup>1</sup>, Tri Purwani<sup>2</sup>, Yani Prihati<sup>3</sup>, Christina Priscilla Putri<sup>4</sup>

<sup>1,2,3,4</sup>Universitas AKI, Semarang, Indonesia

Email: [alexander.dharmawan@unaki.ac.id](mailto:alexander.dharmawan@unaki.ac.id)

## Abstract

The company must have a unique strategy to help grow its business. One of the ways to strengthen the company's business is by estimating the company's profit. This is because with an estimated profit, the company can manage the transactions made. Some companies still use a manual profit determination process that requires several steps. Processes that are carried out manually result in a long time, which can cause delays in completing tasks and results that are less accurate than desired. Multiple linear regression is an algorithm used to determine the relationship between the dependent variable and at least two independent variables. This algorithm is a type of supervised learning algorithm that develops estimation models based on input data. The use of this algorithm is included in part of machine learning. Implementation of machine learning to calculate company profit estimates using the Python programming language. From the estimation results using multiple linear regression and Python programming, the result is that multiple linear regression can be utilized or used to predict company profits. 98% of Profit is influenced by independent factors, namely R&D Spend and Marketing Spend, while the remaining 2% is influenced by variables that are not included in this calculation.

**Keywords:** *Profit Estimation, Machine Learning, Python.*

----- ◆ -----

## A. INTRODUCTION

Profit is the difference between revenue and total costs, which are company expenses during a certain period (Marwansyah & Utami, 2017). Profits obtained by the company can be used for various purposes, including estimating risks in investing, maintaining workplaces or equipment (Elisa, 2018), to be able to pay off existing debts, for future company development, and to improve the welfare of owners and employees, as well as to measure the company's performance in the use of economic resources.

To determine company profit estimates with accurate results, a method or algorithm is needed. The algorithm used to predict profit is a multiple linear regression algorithm. Multiple linear regression is an algorithm used to determine the relationship between the dependent variable and at least two independent variables (Elen Riswana Safila Putri et al., 2021). This algorithm is a type of supervised learning algorithm that develops predictive models based on input data.

The use of this algorithm is included in part of machine learning. Machine learning is a machine that is designed to be able to learn from data (Russell, 2018). The characteristics of machine learning are training, learning, or training (Puteri and Silvanie, 2020). Therefore, machine learning needs to study the data as training data (training set) to be able to get good accuracy results. The model generated during the

training process will be used as a reference in determining the estimated company profit.

Implementation of machine learning to predict company profits using the python programming language. The Python programming language emphasizes code simplicity thereby enabling programmers to develop applications quickly (Chan, 2014). Python programming language can run on various operating systems (Ginting, Kusriani & Luthfi, 2020).

## **B. LITERATURE REVIEW**

### **1. Machine Learning**

Machine learning is a machine that is designed to be able to learn on its own without user guidance (Russell, 2018). If the machine can use the data offered to continuously improve the quality of the machine's output, then the machine can be considered "learning" (Kurniawan, 2022). This is similar to how humans learn, namely by using past experiences to shape the way humans work so that when faced with similar situations in the future, human responses will be better.

Therefore, computer work will be better if it has more data, knowledge, or experience. There are several types of machine learning algorithms, ie:

#### **a. Supervised Learning**

Supervised learning is a subtype of machine learning that requires more human involvement. Supervised learning is used to find patterns in labeled input data, making it possible to produce the correct output data effectively (Russell, 2018). In Supervised Learning, labeled datasets are used (Meng Lee, 2019). Labeled data is data whose validity has been determined previously. Because there is a teacher or supervisor present in the form of labeled data, this learning is referred to as supervised learning.

#### **b. Unsupervised Learning**

Unsupervised learning is used to study the characteristics of unlabeled data sets (Russell, 2018). Unsupervised learning does not require data labeled as input, in contrast to supervised learning. This algorithm is tasked with finding data that has been divided into several categories based on similarities and differences. What Unsupervised Learning does is try to predict patterns in datasets (Meng Lee, 2019). Since there is no supervisor directing the computer to tell what is right and what is not, this is known as unsupervised learning.

#### **c. Reinforced Learning**

Reinforcement learning is a type of machine learning algorithm that enables agents to learn through trial and error in an interactive environment using feedback from their own actions and experiences (Russell, 2018). In an uncertain and complex environment, software agents learn to achieve a goal. The purpose of reinforcement learning is to find the best way, method, behavior or approach to be taken in a given situation. Unlike supervised learning which has predefined solutions, reinforcement learning operates differently. With this approach, the computer will be programmed to carry out

commands based on the current state. Because there is no correct response, the computer will learn from past mistakes to prevent future mistakes (Baruque, 2014).

## 2. Multiple Linear Regression

Simple linear regression and multiple linear regression are two types of linear regression. A regression model called simple linear regression is used to explain the relationship between the independent and dependent variables (Ningsih & Dukalang, 2019). The general form of a simple linear regression equation is as follows:  $Y = \alpha + \beta X$

The development of a simple linear regression model is a multiple linear regression model. To determine the relationship between the dependent variable and at least two independent variables, multiple linear regression is one of the methods (Ningsih and Dukalang, 2019; Elen Riswana Safila Putri et al., 2021). Regression with more than one independent variable and one dependent variable is referred to as multiple regression. The general form of the multiple linear regression equation changes with increasing number of independent variables, as follows:  $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$

with,

$Y$  = Dependent Variable

$\alpha$  = constant (Y value if  $X_1, X_2, \dots, X_n = 0$ )

$\beta_1, \beta_2, \dots, \beta_n$  = regression coefficient

$X_1, X_2, \dots, X_n$  = independent variable

## 3. Estimation

Estimation is the process of making predictions about the future based on an analysis of the past (Adiguno, Syahra & Yetri, 2022). Estimates don't have to give a definitive answer as to what will happen, instead, they look for answers that are as close to the actual results as possible. The purpose of estimation is to study what will happen in the future with the greatest probability of happening (Habibi & Suryansyah, 2020). Along with its development, estimation is used as a tool and even estimation has also been used as a consideration in making decisions. Estimates are used in a variety of contexts, including predicting the weather, determining food prices, and many more.

## 4. Profit

Profit is income minus expenses, including taxes, which represent business expenses for a certain period of time (Marwansyah & Utami, 2017). Company profits can be used for several things, such as calculating investment risk, maintaining workplaces or facilities (Elisa, 2018), improving the welfare of owners and employees, and measuring how effectively the company uses its financial resources.

## 5. Python

A high-level programming language that emphasizes code simplicity, allowing programmers to create applications quickly, is the Python programming language (Chan, 2014). Programs developed in the Python programming language require fewer lines of code than programs written in other languages such as C. This results in fewer programming errors and requires less development time. There is such a thing as a library in python. This Python library will keep track of the number of active functions. The Python library is a group of modules, each of which contains code that is included in various types of programs. The existence of libraries makes Python programming more efficient and easier for programmers because there is no need to keep creating the same code for many projects.

## **6. Google Colaboratory**

An open source software called Google Colaboratory or Google Colab can be used on any computer equipped with a web browser (Saiful, 2021). Anyone can develop and run python code using a browser thanks to Google Colaboratory, also known as Google Colab (Geadalfa & Saidah, 2021). Google Colab offers several benefits that can be used for free, including Graphical Processing Unit (GPU), Collaborate features, easy integration, and flexibility. Since Google Colab runs on a web browser, it must have a good internet connection.

## **C. METHOD**

### **1. Data Collection**

Finding and collecting the right data to use in research is the initial stage. Data from this research source includes data on research and development expenditures, management expenditures, marketing expenditures, the amount of profit earned, and location. In this study, data is presented in .csv (Comma Separated Value) format provided by Orbit Future Academy.

### **2. Data Preprocessing**

At this stage, the data is checked and unnecessary data such as outliers and missing values are removed. The purpose of this data processing is to convert data that has not been processed into data that is ready for analysis (Rianto and Yunis, 2021).

### **3. Data Exploration**

Data exploration is the process of analyzing data sets to understand their contents, components, and characteristics (Nafi'ah, 2021). John Tukey introduced the term Exploratory Data Analysis (EDA) to encourage statisticians to explore data and formulate hypotheses. The method that the writer uses for data exploration is visualization. Data visualization is the process of presenting data in a graphical format (Dhingra, Dutt & Banerjee, 2021). Utilizing visual elements will make it easier for readers to understand patterns and content.

### **4. Modelling**

The modeling process is used to find patterns in the data that form the basis of system knowledge to draw conclusions or predictions (Nafi'ah, 2021). The algorithm used for modeling is a multiple linear regression algorithm. The linear regression algorithm has five assumptions that need to be met. The classic assumptions of multiple linear regression are as follows: a) The linear relationship between each feature and the label; b) Normality Test; c) Heteroscedasticity does not occur; d) Multicollinearity does not occur; and e) Autocorrelation does not occur.

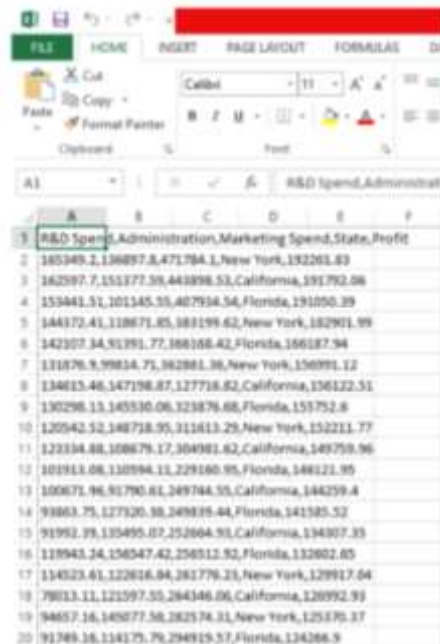
## 5. Evaluation

Once created, the model must go through a series of tests. This helps determine whether the model meets the project requirements for which it has been designed. The model is applied and used only after it is determined that the model is the best model (Nafi'ah, 2021). Testing in this model uses the coefficient of determination. The coefficient of determination indicates that several features can be used to describe the value of the label. A good regression model is one whose coefficient of determination is closest to 1 (Puteri & Silvanie, 2020).

## D. RESULT AND DISCUSSION

### 1. Data Collection

The data is presented in .csv (Comma Separated Value) format and is called "profit\_companies.csv". Sales figures, expenses, and profits are all included in this data. This information will be used to build a model, and this model will help calculate profit based on how much the company is spending. CSV data is stored in Microsoft Excel as shown in Figure 1, containing five variables, namely profit\_companies data, including R&D Spend, Administration, Marketing Spend, State, Profit.



	A	B	C	D	E
1	R&D Spend	Administration	Marketing Spend	State	Profit
2	185349.2	136897.8	471784.1	New York	132261.83
3	162597.7	151377.59	443898.53	California	191792.06
4	153441.57	101145.55	407934.54	Florida	191050.39
5	144172.41	118671.85	381199.62	New York	182901.99
6	142307.34	91391.77	388166.42	Florida	186187.94
7	131876.9	99834.75	382881.38	New York	156091.12
8	134815.48	147198.87	127716.82	California	136122.51
9	130298.13	145530.06	123876.88	Florida	153752.6
10	120542.52	148718.95	111813.29	New York	152211.77
11	12334.88	108679.17	304981.62	California	149759.96
12	101911.08	120594.11	229180.95	Florida	148121.95
13	100871.96	81790.61	249744.55	California	144259.4
14	93863.75	127320.58	249835.44	Florida	141585.52
15	91992.39	135495.87	252864.93	California	134307.35
16	118943.34	158547.42	258512.52	Florida	132882.85
17	114523.61	122836.84	281776.23	New York	129917.64
18	79013.11	121597.55	284346.06	California	128992.93
19	94657.16	148877.58	282574.31	New York	125176.17
20	91389.16	114175.79	294915.57	Florida	134286.9

Figure 1. CSV Dataset

### 2. Data Preprocessing

Data that has not yet undergone any processing is converted into data that can be received and studied by the model being developed during the data preprocessing stage.

```
import numpy as np
import pandas as pd
from scipy import stats
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split #split data test & training
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import variance_inflation_factor
```

**Figure 2. Import Library**

The preparation of the library is the first step in carrying out multiple regression analysis in the Python programming language. Libraries used are NumPy, Pandas, Scipy, Seaborn, Matplotlib, Sklearn, and Statsmodels.

```
data = pd.read_csv('/content/profit_companies.csv')
data.head()
```

	R&D Spend	Administration	Marketing Spend	State	Profit
0	165349.20	136897.80	471784.10	New York	192281.83
1	162587.70	151377.58	443896.53	California	191792.06
2	153441.51	101145.55	407934.54	Florida	191050.39
3	144372.41	118671.85	383199.62	New York	182901.99
4	142107.34	91391.77	366166.42	Florida	166187.94

**Figure 3. Dataset Profit Companies**

Python uses the `data.head()` command when calling data. The first five records are displayed with this command. Figure 3 displays the result of the command. After the data is displayed, the next step is to carry out the cleaning process. The following are some of the data criteria that need to be cleaned: 1) Data that is incomplete (missing value) or has a NaN (null) value; and 2) Duplicate data.

```
[ ] data.isnull().sum()
```

```
R&D Spend      0
Administration 0
Marketing Spend 0
State          0
Profit         0
dtype: int64
```

```
[ ] data.isna().sum()
```

```
R&D Spend      0
Administration 0
Marketing Spend 0
State          0
Profit         0
dtype: int64
```

**Figure 4. Data contains NaN or Null**



```
data.duplicated().value_counts()
```

```
False    200
dtype: int64
```

**Figure 5. Duplicated Data**

### 3. Data Exploration

Data exploration is carried out to obtain data information with the techniques used.

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
#   Column                Non-Null Count  Dtype
---  -
0   R&D Spend              200 non-null   float64
1   Administration         200 non-null   float64
2   Marketing Spend        200 non-null   float64
3   State                  200 non-null   object
4   Profit                 200 non-null   float64
dtypes: float64(4), object(1)
memory usage: 7.9+ KB
```

**Figure 6. Dataset Information**

Data from the profit\_companies.csv dataset shown in Figure 6 Summary DataFrame can be printed using the data info function. It can be concluded that profit\_companies.csv data has 200 entries with 5 columns. Figure 6 shows the data type for each column and the fact that there are no blank data.

### 4. Modelling

The modelling phase is used to identify patterns in the data that serve as a knowledge base for the system to make conclusions or predictions. Multiple linear regression algorithm is an algorithm used in modelling. There are five assumption tests that must be fulfilled for the linear regression algorithm, namely the linearity test, normality test, heteroscedasticity test, multicollinearity test and autocorrelation test.

### 5. Evaluation

The machine learning model must be evaluated to calculate its accuracy because the evaluation results in accurate calculations. The coefficient of determination ( $R^2$ ) can be used to assess the accuracy of the prediction results of the profit\_companies model.

```
print(f'R^2 score: {lin_reg.score(X, Y)}')
```

R^2 score: 0.9812342341599597

**Figure 7. Results Coefficient of Determination**

Figure 7 shows that the research on the profit\_companies prediction accuracy rate is 0.9812342341599597 or 98% using a linear regression algorithm with 13% training data and 87% testing data. This figure shows that R&D Spend and Marketing Spend can accurately predict 98% profit. The remaining 2% is influenced by other factors not included in the model. This model can be used to predict new data because it has a high level of accuracy and passes the classical assumption test.

```
Input R&D Spend      = 142634
Input Marketing Spend = 9344.343991

Prediksi Profit yang akan didapat adalah 147931.7487658281
```

**Figure 8. Estimation Results on New Data**

Figure 8 shows the predictions made using the new data with an accuracy rate of 98%. The profit prediction result is 147931.7487658281 if the R&D Spend value is 142634 and the Marketing Spend value is 9344.343991.

## E. CONCLUSION

Multiple linear regression algorithms can be utilized or used to predict company profits. 98% of Profit is influenced by independent factors, namely R&D Spend and Marketing Spend, while the remaining 2% is influenced by variables that are not included in the calculation. The use of Google Colab is very helpful in the process of calculating linear regression. Google Colab is used online and can be used on various devices. Users can enter R&D Spend and Marketing Spend data, which will be processed by the system using a machine learning model which will then display Profit results to the user.

## REFERENCES

1. Adiguno, S., Syahra, Y., & Yetri, M. (2022). Prediksi Peningkatan Omset Penjualan Menggunakan Metode Regresi Linier Berganda. *Jurnal Sistem Informasi Triguna Dharma (JURSI TGD)*, 1(4), 275-281.
2. Baroque, B., & Corchado, E. (2011). *Fusion methods for unsupervised learning ensembles* (Vol. 322). Berlin, Germany: Springer.
3. Chan, J. (2015). *Learn Python in one Day and Learn it Well: Python for Beginners with Hands-On Project: The Only Book You Need to Start Coding in Python Immediately*. CreateSpace Independent Publishing.
4. Tiwari, S., Dogan, O., Jabbar, M. A., Shandilya, S. K., Ortiz-Rodriguez, F., Bajpai, S., & Banerjee, S. (2022). Applications of Machine Learning Approaches to Combat COVID-19: A Survey. *Lessons from COVID-19*, 263-287.



5. Putri, E. R. S., Novianti, F., Yasmin, Y. R. A., & Novitasari, D. C. R. (2021). Prediksi Kasus Aktif Kumulatif covid-19 di Indonesia Menggunakan Model Regresi Linier Berganda. *Transformasi: Jurnal Pendidikan Matematika dan Matematika*, 5(2), 567-577.
6. Elisa, E. (2018). Prediksi Profit pada Perusahaan dengan Klasifikasi Algoritma C4. 5. *Kumpulan Jurnal Ilmu Komputer (Klik)*, 5(02), 179-189.
7. Giyanda, G., & Saidah, S. (2021). Auto Machine Learning dengan Menggunakan H2O AutoML untuk Prediksi Harga Bitcoin. *Jurnal Ilmiah KOMPUTASI*, 20(2), 189-198.
8. Ginting, V. S., Kusri, K., & Luthfi, E. T. (2020). Penerapan Algoritma C4. 5 dalam Memprediksi Keterlambatan Pembayaran Uang Sekolah Menggunakan Python. (*JurTI*) *Jurnal Teknologi Informasi*, 4(1), 1-6.
9. Habibi, R., & Suryansah, A. (2020). *Aplikasi Prediksi Jumlah Kebutuhan Perusahaan* (Vol. 1). Kreatif.
10. Kurniawan, D. (2022). *Pengenalan Machine Learning dengan Python*. Elex Media Komputindo.
11. Marwansyah, S., & Utami, A. N. (2017). Analisis Hasil Investasi, Pendapatan Premi, dan Beban Klaim terhadap Laba Perusahaan Perasuransian di Indonesia. *Jurnal Akuntansi, Ekonomi dan Manajemen Bisnis*, 5(2), 213-221.
12. Lee, W. M. (2019). *Python machine learning*. John Wiley & Sons.
13. Nafi'ah, H. (2021) *AI Project Cycle*, *Medium.com*. Available at: <https://medium.com/@hannnfh/ai-project-cycle-ccd67c3dd21d>.
14. Ningsih, S., & Dukalang, H. H. (2019). Application of Interval Successive Method in Multiple Linear Regression Analysis. *Jambura Journal of Mathematics*, 1(1), 43-53.
15. Puteri, K., & Silvanie, A. (2020). Machine Learning untuk Model Prediksi Harga Sembako dengan Metode Regresi Linear Berganda. *Jurnal Nasional Informatika (JUNIF)*, 1(2), 82-94.
16. Russell, R. (2020). *Machine Learning: Step-by-Step Guide to Implement Machine Learning Algorithms with Python*. (Knxb).
17. Saiful, A. (2021). Prediksi Harga Rumah Menggunakan Web Scrapping dan Machine Learning dengan Algoritma Linear Regression. *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, 8(1), 41-50.