Forecasting determination of housing development schedule using machine learning clustering method

Allwin M. S, Andi Tanoto, Forbes, Willy, Ashari

1 Faculty of Technology and Computer Science, Universitas Prima Indonesia, Indonesia
*allwinsimarmata@unprimdn.ac.id

Abstract. Machine learning is a part of AI (Artificial Intelligence) which is a focus of development of system that can learn by “itself” without being reprogram many times by human. Machine Learning Application need a Data for a training subject before it produce an output. An Application like Machine Learning is usually in a specific domain or it couldn’t be interpreted generally for all problems. Machine Learning does not mean without data. Which is means all Machine Learning Application needs Data for training material and to be analysed so it could produce an output. Before Machine Learning Application could work, it needs Data for training, the result of the training will be test with the same Data or the opposite data. The output of Machine learning usually come out as a prediction data with the trustworthy label. Clustering is grouping a data without based on some class data, moreover Clustering could be used for labelling a data in some class that haven’t been known the type. Therefore Clustering is usually be categorized as unsupervised learning method.

1. Introduction
In todays technology development it’s really need a concept of determining a proper decision, directed and corresponding so it could give the best result in a activity. So do as for building a housing in particular location. In determining the begining starting point of building to the end of the building project is needed a solid plan. In addition to avoiding a mistake in building process its required a very detail calculation so it could affect the income that will be reach.

The utilization of Machine Learning is a improvement of artificial intelligence, by developing every connection of data and combined with a flexible algorithm so it could provide a value and give more efficient solution for solving the occurring problem. Clustering is a technique that used for grouping data in Data Mining. The principle of Clustering is maximize similarity of member in a same class and minimize the similarity of the class or cluster. Clustering can be used for data that have a few attribute that will be mapping as a multi-dimensional space. In building a housing in some location its really require percision for a marketing expert and builder to scheduling every business process. Strating from mining the data and the fact at the location, location mapping and soil texture, at the location, market opportunity determination, the interest rate determination, building schedule determination, lauching schedule and the other factors that will determine the building process later. That’s why the problem is how are we supposed to design a forecasting method that could provide a solution and a maximum conclusion and moreover what is the effect of using Machine Learning to process all of the data for the scheduling determination.

In this paper, writer is more accentuate in determining a proper scheduling based on every factor from before. In the data design basically is already complex and structured and Machine Learning will have a important role
for increasing the quality of the schedule determination so it could give a major effect in increasing the house selling rate later.

2. Research Methodology

2.1 Machine Learning

In this technology era everyone is not strange with the term AI (Artificial Intelligence) almost every things use a main concept of Artificial Intelligence. Machine Learning is one of the branch knowledge of Artificial Intelligence which is a machine is designed can learn. Machine Learning basically referring to the changes in system that running certain task that related to Artificial Intelligence. The Task as introduction, diagnose, planning, robotic controlling, forecasting, etc [4].

Changes in the system of Machine Learning probably could be increase of the system that already running or a new combination in the system. The important thing in Machine Learning is how we can make a clear fundamental that can protect the program from a useless or senseless result [5].

Machine Learning is like a “tool”, like a mathematical formula. How to use is depends on what domain it used. By that you must really understand that only understanding the Machine Learning techniques is not enough. You must know the application domain because the usage of the Machine Learning technique is can be different at different domain [3].

The important thing in Machine Learning development is choosing sample or training data. If a training data can’t represent data population, then the model result from the training is not good. Therefore usually there is a test data as a balance. Machine is trained by using the training data, then the result is tested using the test data[7].

The methods of Machine Learning consists of:

1. Supervised Learning

Supervised Learning is mean that in this learning, there is a teacher that teach and the student that will be teach.



Figure 1. Supervised Learning

Like the figure 2.1 show a teacher writing a number 8, 6, 2 as an example then the teacher give the correct way to read for every number. As example the number symbolize the input, then the correct way to read it is the desired-output. This the set of input and desired-output is called training data. The purpose of supervised learning generally is for classified data like classify news text to become categorized like sports, politic, national, etc. if there is only two option then is called binary classification. While more than two category is called, multi-label classification.

Supervised Learning Equation :

\[
p(y|x, w)
\]

The understanding of supervised learning is remembering the equation 1.1. There is three important things in supervised learning which is input, desired output, and learning parameters. Based on the model, we can classify it. As conceptual, the classification is defined like equation 1.2 which is chosen the most optimal label that will give to a certain instant of data.
Classification Equation:

\[
y_i = \operatorname{arg\,max}_{y_t \in \mathcal{C}} p(y_t|x_i; \mathbf{w})
\]

(2)

Other than supervised learning, there still other studying method which is unsupervised learning, and semi-supervised learning.

2. Semi-supervised Learning

Semi-supervised learning is similar to supervised learning, the difference is the process of labelling the data. In supervised learning, there is a “teacher” that will provide an input-output “answer”. While in semi-supervised learning there is no explicit “answer” that is made by teacher. The answer is obtain from other source (example is from Clustering result).

3. Unsupervised Learning

If supervised learning there is a teacher that teaching, than in unsupervised learning there is no teacher. The example of unsupervised learning is clustering. For example if you open a department store, for the customer to easier to shop, you grouping the things, but the label for the things is not defined. What you do is make a groups based on the characteristics of the things. The example of unsupervised learning algorithm is K-nearest-neighbor. The different of supervised learning and unsupervised learning is that supervised learning have desired output, while unsupervised learning there is no desired output.

\[
p(x, \mathbf{w})
\]

Figure 2. Unsupervised Learning

The real population have distribution \(q(x)\), we want to estimate \(q(x)\) by taking a few sample, then learn it. Learning is done by optimizing \(p(x, \mathbf{w})\) by optimizing \(\mathbf{w}\) parameter. The difference between the estimation and the real function is called generalization loss.

Unsupervised learning equation:

\[
p(x|\mathbf{w})
\]

(3)

The key to understand unsupervised learning is by remembering the equation 1.3 there is two important things that is input and learning parameter.
2.2 Clustering

Clustering is a grouping data with similar properties. Data for clustering do not have label (class). Generally, clustering algorithm can be categorized into two types based on the desired result that is partitional and hierarchical. The example or partitional algorithm is K-means, while example of hierarchical algorithm is agglomerative clustering[9].

K-means, K-means algorithm is grouping data into the number of K groups according to what we define. This algorithm is also called flat clustering, which is means one group has equal position with the other groups[6]. The stages of K-means algorithm are as follows:
1. Determine the number of K group that we want.
2. Initiation of centroid for each groups.
3. Calculate the proximity of data to centroid, then input the data into whose the centroid has the closest properties to it.
4. Reselect the centroid for each group, from the group member.
5. Repeat the previous steps until there is no member changes for all groups.

There is 2 important things in K-means algorithm that is choosing centroid and calculation the similarity of the data. One of the simplest method to choose the centroid is by choosing random. In reality the initiation of the centroid for every group can be done by random, but in the next step, generally centroid is chosen by picking the average of the value. Given a set of data \( \{d_1, d_2, \ldots, d_n\} \) in a cluster in a feature vector form, so the centroid or \( c \) is calculated using

Equation to find Centroid:

\[
\hat{c} = \frac{1}{n} \sum_{i=1}^{N} \sum_{e=1}^{p} d_{i} [e]
\]

That is the average value of every feature vector for all member of the cluster, where the \( n \) is the amount of cluster and \( d_i \) is the document to-i. With this, centroid generally can be not the element of the cluster member.

3. Analysis and design

3.1 Analysis

In this analysis it’s expected that machine could think independently in determining the solution based on the data mining that has been clustered. Before it reach at machine learner level, first we gather the data as detail as possible related to scheduling the house building from the term of starting point of building, determination of scheduling, determining the building process and etc. Next will be sort every raw data first by clustering method that will provide a maximum mined data to give a better solution. Then the data that have been sorted will be connected to Machine Learning algorithm that will be arranged so the flow of every process could make machine work automatically.

3.2 Design

3.2.1 Design of Basic knowledge

In the design of this machine learning system is consist of few steps. The steps is pictured as follows.

![Figure 4. Steps of system design](image)
First of all is analyze the requirement of the data that will be used as calculation or learning subject. Data mining that related to build scheduling process start from the geography aspect to the economy aspect. Then every data is determined in 3 level or stages which is low, medium, and high. As example like the soil texture is start from wavy at low level, to medium, and flat for high level. After every data level is determined then the next step is grouping the data so the data will be easier to be determined based on each their own parameter.

### 3.2.2 Clustering Design

Clustering design is a grouping process form the data that have been gathered from data mining process in field. The data is as follows.

#### Table 1. List of data

| NO | PARAMETER                        | Level  |
|----|----------------------------------|--------|
|    |                                  | 1      | 2    |
| 1  | Soil Texture                     | 2      | 6.5  |
| 2  | Land Classification              | 5      | 7.5  |
| 3  | Soil Layer                       | 6.5    | 3    |
| 4  | Land Coordination (From Settlement) | 5     | 0.5  |
| 5  | Land Location (City)             | 2.5    | 9.5  |
| 6  | NJOP Cost (Tax)                  | 5      | 4.5  |
| 7  | Land Location (Competitor)       | 4      | 1    |
| 8  | Site Plan                        | 5      | 6.5  |
| 9  | Consumer                         | 8      | 6    |
| 10 | Competitor                       | 4      | 3    |
| 11 | Operational Cost                 | 8      | 8    |

Can be seen from figure above that the data is consist of few parameter. That Parameter in the form of criteria that have been analyze the necessity as the things that will affecting the building schedule. Level in the data is the value of the parameter that have been research. The purpose of the level is to make the calculation and grouping of the data is easier. By dividing the data into two level so the calculation of the data in the area that relatively massive. Because the value of the land is non numeric but the clustering process could be done by numerical data, therefore the data is analyze and declared the value of every each parameter.

For example in the parameter of soil texture, if the value is higher than that means the soil texture in that level is more flat and specially for land coordination (from settlement), land location (city), and land location (competitor) and the cost. The better value is the higher value that means if the value is high is not farther or higher the cost but otherwise. So the conclusion of data from above is, if the value is higher that means the criteria of the land is better. Next step is calculation of the data. Process of calculation of the data is use a Euclidean distance calculation method with the equation as follows:

\[
D(p, c)_n = \sqrt{\sum_{i=0}^{n}(p_i - c_i)^2}
\]  (1.5)

Where \( p \) is the data, \( i \) is the iteration or looping, \( n \) is the total of data, and \( c \) is the centroid or the middle point of a cluster. The starting value of centroid for the calculation will be taken one data randomly from all of the data where the centroid value for first cluster is 2 and 4,5 and the centroid for second cluster is 4 and 1. By using the Euclidean distance equation then the calculation of the first to the last data at cluster 0 and cluster 1. Data is calculated two times because of the data will be clustered into two groups. From the calculation result is obtained from the first data or the soil texture data that the proximity from the cluster is 2. Then the first data is calculated again to the second cluster by using the second centroid. As well as for the whole data that will be calculated
from the first to the last data which is the eleventh data. From the whole calculation of the data then is obtain the result as follows.

Table 2. Clustering result table

| Parameter | Cluster 0 | Cluster 1 |
|-----------|-----------|-----------|
| D1        | 2         | 5.85      |
| D2        | 4.2       | 6.5       |
| D3        | 4.2       | 3.2       |
| D4        | 2.64      | 0.86      |
| D5        | 5         | 8.63      |
| D6        | 0         | 2.8       |
| D7        | 2.8       | 0         |
| D8        | 3.6       | 5.5       |
| D9        | 6.1       | 6.4       |
| D10       | 1.3       | 2         |
| D11       | 6.9       | 8         |

After the result of the clustering is obtained then the next things to do is grouping the minimum value of the result data. Minimum data is the smallest value by comparing the result of each data of two cluster, for example in data D1 in cluster 0 in valued 2 while in D1 at cluster 1 is valued 5.85 then the smallest value is 2. The minimal result that has been determined will be made into a table list so it could easier to observed and gather.

Table 3. Table of minimal value grouping

| Parameter | Cluster 0 | Cluster 1 |
|-----------|-----------|-----------|
| D1        | 1         | 0         |
| D2        | 1         | 0         |
| D3        | 0         | 1         |
| D4        | 0         | 1         |
| D5        | 1         | 0         |
| D6        | 1         | 0         |
| D7        | 0         | 1         |
| D8        | 1         | 0         |
| D9        | 1         | 0         |
| D10       | 1         | 0         |
| D11       | 1         | 0         |

After the minimal value of the data is obtained then the next step is determining the new centroid from the data that have been calculated. The determination of the centroid is determined based on the total of all the minimum result data that have been obtain by last calculation as example if the minimum is in the D1 then the value of D1 is summed with all other minimum value at the same cluster then divide by how many minimum data is been selected for the simple word is the new centroid is the mean of every minimum data from before. The result from the calculation that will be the new centroid value from the first cluster which is 4.3 and 5.1 that the value will be used to recalculate with the first data. After the first centroid for the first cluster is obtained then we will calculate for obtaining the centroid for the second cluster. After the centroid for the second cluster is obtained then the next step is do the calculation of the data from beginning with the new centroid value. Same as before, all the data must be calculated with the new centroid and the result obtained is as follows.
Table 4. Calculation result with the new centroid

| Parameter | Cluster 0 | Cluster 1 |
|-----------|-----------|-----------|
| D1        | 2.58      | 6.6       |
| D2        | 2.5       | 5.86      |
| D3        | 3         | 1.5       |
| D4        | 4.6       | 1.7       |
| D5        | 4.7       | 8.9       |
| D6        | 2.3       | 5.3       |
| D7        | 4.1       | 2.45      |
| D8        | 1.56      | 5.2       |
| D9        | 3.8       | 4.7       |
| D10       | 2.1       | 2.8       |
| D11       | 4.7       | 6.7       |

Table 5. The minimum value of the data from the calculation using new centroid

| Parameter | Cluster 0 | Cluster 1 |
|-----------|-----------|-----------|
| D1        | 1         | 0         |
| D2        | 1         | 0         |
| D3        | 0         | 1         |
| D4        | 0         | 1         |
| D5        | 1         | 0         |
| D6        | 1         | 0         |
| D7        | 0         | 1         |
| D8        | 1         | 0         |
| D9        | 1         | 0         |
| D10       | 1         | 0         |
| D11       | 1         | 0         |

After the minimum value is determined, compare the result. If the result is not different from the first calculation then the clustering result is the optimal result, but if the result is different. The data will be calculated again by using the new centroid until the result is not different from the last result.
3.2.3 Machine Learning Design

In the Machine learning design we used a semi-supervised learning, because the data needed for machine learning to learn is from the result of data mining in field and the optimization by using K-means Clustering algorithm. As for the work flow of the machine learning system is as follows.

![Machine Learning Flowchart](image)

**Figure 5. Machine learning Flowchart**

At the first step in flowchart figure above is data gathering or data mining obtained from field. The data obtained is the data that match with the requirement of the machine learning calculation like soil texture, land classification soil layer, etc. then after all the data is gathered then the next process is grouping or clustering the data. As specified in the clustering method before. The determination of clustering is helpful for easier calculation in machine learning system. By grouping the data corresponding as the criteria then will easier to obtain a result and the result obtained from machine learning process is better.

Next step is the calculation using machine learning approach. The idea of this process is the calculation from the data that have been clustered before to discover how beneficial a land condition from the characteristics that have been parameterized. For example if a land have a wavy soil texture characteristics, the land classification is not productive, the operational cost is expensive, and the land location is relatively far from nearby settlement then machine learning will calculate whether the building is better to start from this land or the other land. All the criteria that have been determined in the beginning will be the main reference for the machine learning. Then the forecasting result will be stored for consideration material or learning material for machine learning to learn, so for the machine learning will provide a better or the same result for the next forecasting. If the result is same as the last result after being calculated many times, then that result must be the best result.
4. Implementation

In the implementation the idea of the result from machine learning will be a prediction because the machine learning will be work as a learner because the final result whether the building point is determined by human decision or the user. Machine learning will only calculate and provide a result based on the data that have been clustered that this result will probably influence the decision made by human. For example like the data that have been calculated before, is obtained as follows.

Table 6. The final calculation

| Parameter                           | Cluster 0 | Cluster 1 |
|-------------------------------------|-----------|-----------|
| Soil Texture                        | 2.58      | 6.6       |
| Land Classification                 | 2.5       | 5.86      |
| Soil Layer                          | 3         | 1.5       |
| Land Coordination (From Settlement) | 4.6       | 1.7       |
| Land Location (City)                | 4.7       | 8.9       |
| NJOP Cost (Tax)                     | 2.3       | 5.3       |
| Land Location (Competitor)          | 4.1       | 2.45      |
| Site Plan                           | 1.56      | 5.2       |
| Consumer                            | 3.8       | 4.7       |
| Competitor                          | 2.1       | 2.8       |
| Operational Cost                    | 4.7       | 6.7       |

If from the clustering calculation before we use a \( D_1, D_2, \ldots, D_{11} \) variable then in this result table is not. Because the variable is only used to make the mathematical calculation process easier. By obtaining the final result from the clustering it will make machine learning easier to determine the decision of the user. If from the last clustering we determine the minimum value from the result, in the implementation for machine learning is also needed. The purpose of determine the minimum value is as stated from before, that is for easier in the process of determining the decision. The determination of the minimum value is as follows.

Table 7. The minimum determination from the final result

| Parameter                           | Cluster 0 | Cluster 1 |
|-------------------------------------|-----------|-----------|
| Soil Texture                        | 1         | 0         |
| Land Classification                 | 1         | 0         |
| Soil Layer                          | 0         | 1         |
| Land Coordination (From Settlement) | 0         | 1         |
| Land Location (City)                | 1         | 0         |
| NJOP Cost (Tax)                     | 1         | 0         |
| Land Location (Competitor)          | 0         | 1         |
| Site Plan                           | 1         | 0         |
| Consumer                            | 1         | 0         |
| Competitor                          | 1         | 0         |
| Operational Cost                    | 1         | 0         |
As we can see in the cluster, where is the minimum value is can be found. By using the criteria from the beginning we said that if the value of the data is higher then the result is better. With the condition as stated before if we apply the result to machine learning then machine learning most likely will predicted cluster 0 or the first cluster will be the optimal result to become the starting point of the building process because many of the parameter value is higher in the first cluster.

5. Conclusion

The usage of Machine Learning using Clustering method could help providing a solution and also could be a learning material so the machine could provide a best solution.

By using Clustering method, the process of data mining will be easier because data is grouped specifically and also could make machine work independently for providing solution.

References

[1] Tom M. Mitchell, 1997. Machine learning McGraw-Hill Science Engineering Math.
[2] Tan, P. N., Steinbach, M. & Kumar, 2006. V, Introduction to Data Mining, Boston: Pearson Addison Wesley.
[3] Ian Witten and Eibe Frank, 2015. Data Mining – Practical Machine Learning. Tools and Technique. San Francisco: Morgan Kaufman Publishers 12 5.
[4] Jonathan Gratch and Stacell Marsella, 2014. Computationally modelling human emotion. Communications of the ACM. 57 12.
[5] Jan Wira Gotama Putra, 2018. Introduction to Learning Concept of Machine and Deep Learning, Edisi 1.1 July 11 05.
[6] Nils J. Nilsson, 1998. Introduction To Machine learning. Robotics Laboratory Department of Computer Science Stanford University 1 10.
[7] Takao Terano and Tsuyoshi Murata, 2017. Spring lecture on machine learning. Lecture of Tokyo Institute of Technology.
[8] Yoshua Bengio Jerome Louradour, Ronan Collobert, and Jason Weston, 2009. Curriculum learning. In proceedings of the 26th Annual International conference on Machine Learning, ICML ’09. 41 48.
[9] Yan-Kai Lin, Cun-Chao, Tu Yu Zhao, Zhi-Yuan, Liu Ayana, Shi-Qi shen and Mao-Song Sun, 2017. Recent advance on neural headline generation. Journal of computer science and technology. 32(4) 76 78.
[10] Jan Wira Gotama Putra, 2017. Rhetorical sentences classification for automatic title generation in scientific article. TELKOMNIKA. 15(2). 65 66.
[11] Hans-Martin, Ramsl Daraksha Parceen and Michael Strube, 2015. Topical coherence for graph-based extractive summarization. In conference on empirical methods in natural language processing. 194 954.
[12] Shai Shalev-Shwartz and Shai Ben-David, 2014. Understanding Machine learning: From Theory to Algorithms Published, Cambridge University Press 2 24.