Comparison of Model Prediction for Tile Production in Tabanan Regency with Orange Data Mining Tool

Ida Bagus Putu Jayawiguna[1*], Ida Bagus Alit Swamardika[2], Made Sudarma[3]
[1]Department of Electrical and Computer Engineering, Post Graduate Program, Udayana University
[2,3]Department of Electrical and Computer Engineering, Udayana University
Email: jayawiguna6@gmail.com[1], gusalit@unud.ac.id[2], msudarma@unud.ac.id[3]

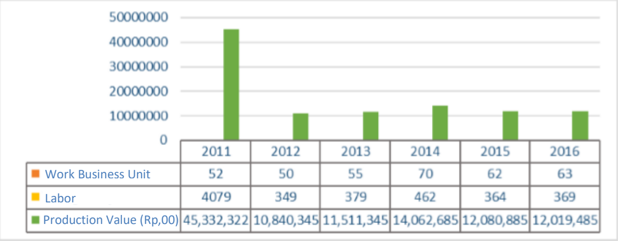
Abstract - Data can not only be saved, but the data can be processed to be a representation of new knowledge. Tile production data in Tabanan Regency is one of them, this data consists of variables that affect the results of tile production in Tabanan Regency. Tile production in Tabanan Regency is relatively declining, therefore a predictive modeling is needed to predict the results of tile production and see which variables most influence tile production. Some data mining techniques such as classification tree, AdaBost, kNN can be used to perform the prediction modeler. In this study predictions using several methods then compare the effectiveness of these methods and look for variables that are more influential on tile production using Linear Regression. Orange data mining tool is used in this study, the use of Orange Data Mining is because this tool is based on GUI so that it can be used easily and can be utilized even by ordinary people. This study produces the best value on the comparison of the methods used and the most influential variables.

***Index of Terms* — *Data Mining, Orange Application, Model Prediction*.**

I. Introduction

Information technology is currently used in various fields that require a lot of data processing. Like, entering data, processing data and then used as a useful information. Large amounts of data in the past can be used as information in the present or future [1]. The data can be processed using methods to dig up information that already exists, for example to find out employee salary categories based on position, age and gender.

Pejaten Village and Nyitdah Village are villages in Kediri Sub-District, Tabanan Regency, a village known as a tile-producing village with a total of 63 industrial business units and capable of absorbing 369 workers [2]. The tile industry in Pejaten Village and Nyitdah Village is a non-formal industry with quite a large number of units and is able to absorb labor. The existence of small and medium industries can overcome the income inequality between communities. Clay tile industry is an industry that is produced using hand media or by using tools or machinery made from clay.



	2011	2012	2013	2014	2015	2016
🟧 Work Business Unit	52	50	55	70	62	63
🟨 Labor	4079	349	379	462	364	369
🟩 Production Value (Rp,00)	45,332,322	10,840,345	11,511,345	14,062,685	12,080,885	12,019,485

Fig1. Number of Business Unit, Labor and Production Value of Tile Industry in Kediri District, Tabanan Regency in 2011-2016

Based on data from the Department of Industry and Trade Tabanan Regency in figure 1 in time series for six years the number of business units, workers and production value of the tile industry experienced fluctuating developments. The number of business units and workers in 2012 decreased when linked from the previous year, as did the production value which decreased by 76.1 percent. However, in 2013 and 2014 the number of business units, workers and the value of tile production increased, but in 2015 business units and labor again declined as well as the value of production decreased by 14.1 percent. The number of business units and clay tile industry workers in 2016 increased again by one business unit and the addition of a workforce of 5 people. However, in 2016,

Data mining is a series of processes to explore the added value of knowledge data sets that are not yet known manually or in other words the process of extracting patterns from data so that it can transform data into information [3]. With the help of data mining techniques, it can be known which variable most influences tile production and can also determine the predicted number of tile production produced based on existing data.

Many tools that can be used in conducting data mining processing techniques, one of which is the Orange Data Mining tool. Orange Data Mining is a data mining tool that is useful for visual programming and exploratory data analysis that can be written in Python. Orange has many components known as widgets. This orange data mining tool supports macOS, Windows and Linux [4].

This study uses the prediction tree widget model, AdaBoost, kNN, and linear Regression to predict tile production to be compared and find the most influential variable on tile production results. The variables used as data are Raw Materials (m3), Energy Fuels (m3) and Labor (Hours).

II. Literature Study

A. *Data Mining*

Data Mining is referring to the process to explore added value in the form of information that has not been known manually from the database. generated into information that will be obtained by extracting and recognizing important or interesting patterns from existing data in the database. Data Mining is mainly used to search for existing knowledge in large databases so it is often called Knowledge Discovery in Databases (KDD) *[5]*. KDD is an activity that includes the collection, use of data, historically to find regularities, patterns or relationships in large data sets *[6]*. Four things are needed in order to effectively data mining such as: right of data, high quality data, examples of which are adequate, and the correct device *[7]*.

Stages of Data Mining, first is data cleaning. Data cleaning needed to eliminating noise and inconsistent data. Next, the integration of data that is merging data from various databases into a new database. After that, data must go to the data selection. Data selection is useful for selecting only the data needed so that the data mining process becomes faster. Furthermore, do data transformation. Then doing the mining process is a major process when the method is applied to find valuable and hidden knowledge of the data. Some methods can be used based on Data Mining grouping. Finally, the presentation of knowledge is done for the visualization and presentation of knowledge of the methods used to acquire the knowledge obtained by the user [8].

B. Classification Tree / Decision Tree

Decision Tree is a tree structure like a flowchart, where rectangular boxes are called nodes. Each node represents a set of records from the original data set. Internal nodes are nodes that have child nodes and leaves (terminal) are nodes that do not have children. The root node is the topmost node. Decision trees are used to find the best way to distinguish classes from other classes. There are five algorithms that are generally used for decision trees: - ID3, CART, CHAID, C4.5 and J48 algorithms [9].

C. KNN (K-Nearest Neighbor)

K-Nearest Neighbor (K-NN) is the method to classify of object based on the training with short distance toward the object [10]. This algorithm tries to classify new data that is not yet known by its class by selecting the number of k data that is located closest to the new data. The most classes of data closest to k are chosen as the predicted class for new data. k is generally determined in an odd number to avoid the appearance of the same amount of distance in the classification process [11]. kNN is often used in classification and also used in predictions and estimates [12].

D. AdaBoost

AdaBoost (Adaptive Boosting) is a machine learning algorithm formulated by Yoav Freund and Robert Schapire. AdaBoost algorithm is an algorithm that builds strong classifiers by combining a number of simple (weak) classifiers [13]. AdaBoost has a solid theoretical foundation, highly accurate prediction, wide, and simplicity [14].

E. Linear Regression

Regression analysis is a statistical method that is widely used in research. The term regression was first introduced by Sir Francis Galton in 1986. In general, regression analysis is a study of the relationship of a variable called a variable that is explained by one or two variables that explain it. The variables that are explained hereinafter are referred to as response variables, while the variables that are explained are commonly called independent variables [15].

F. Orange

Orange Data Mining is a data mining tool that is useful for visual programming and exploratory data analysis that can be written in Python. Orange has many components known as widgets. This orange data mining tool supports macOS, Windows and Linux [4].

III. Research Methodology

In this study proposes a comparison of the Decision tree, kNN, Ada Boost and Linear Regression prediction methods in the Orange Data Mining tool for Prediction of Tile Production in Tabanan Regency. The methodology flow used in this study is illustrated in the following figure.

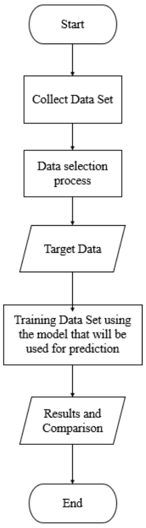


Fig2. Research Methodology Flow

The methodology of this research first is to collect data then from the data obtained, the data selection is done to produce the target data. The target data then becomes material for conducting training data sets with prediction models that are used to produce prediction results and comparisons.

A. Tile Production Data Set

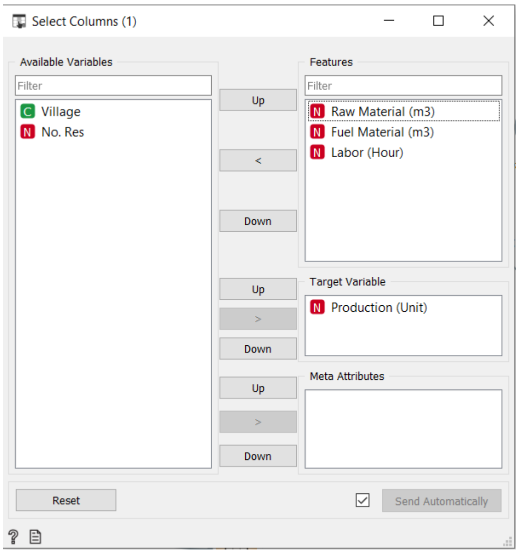
The tile production data set is in the form of primary data obtained through direct interviews with several tile industry owners in Tabanan Regency. The initial attributes of this data set are 5 with 1 destination attribute and 52 instance data. Descriptions of the datasets are explained in the following table.

TABLEI
ATTRIBUTE ON TILE PRODUCTION DATA SETS

No	Attribute	Type	Description
1	Village	Categorical	Village where tile industry
2	Production	Numeric	In unit
3	Labor	Numeric	In hour
4	Raw material	Numeric	In M3
5	Fuel Energy	Numeric	In M3

B. Data Selection Process / Prepocessing

In preprocessing this tile production dataset has no missing value in the data so only the data selection is done. The missing value in the instance data will interfere with the prediction process. Some models in data mining for prediction cannot process due to missing value.



Select Columns (1)

Available Variables

Filter

C Village
N No. Res

Features

Filter

N Raw Material (m3)
N Fuel Material (m3)
N Labor (Hour)

Target Variable

N Production (Unit)

Meta Attributes

Fig3. Selection Data using Select Columns Widget

Figure 3 is the process of selecting data using the widget select columns where the attributes used are raw materials, labor, raw materials and production target attributes.

C. Data Mining Process (Prediction Model)

In analyzing the performance of several prediction models on the orange tool, a comparison of methods in data mining is performed to choose the best method with high accuracy in predicting the Tile Production dataset.

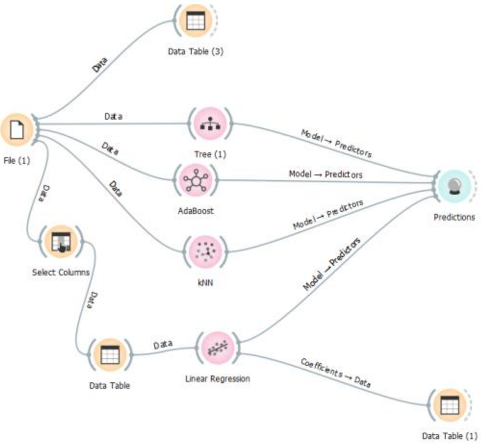


Fig4. Widget Design Process Modeler Prediction Data Set Tile Production

In Figure 4 is a widget design using prediction models that exist in Orange software in data mining in the form of trees, adaboost, kNN and linear regression inputted data sets that have been processed before. Then all the data is brought in prediction mode.

D. The Process of Testing Prediction Models

In the process of testing the prediction model that has been made before, it takes a test data set to find out the predicted results.

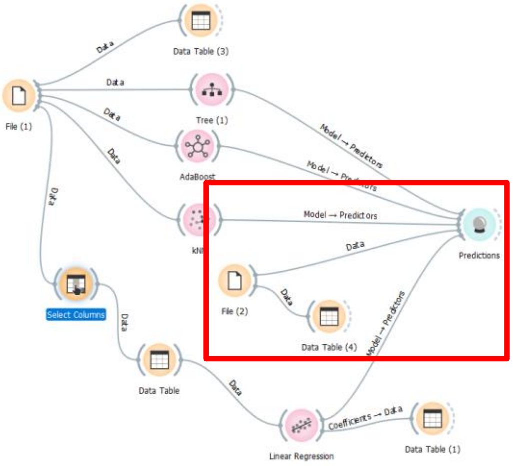


Fig5. Design Prediction Model Trial Widget

In Figure 5 is a widget design that has been added to the prediction trial process for the prediction model. In the picture in the red box is a trial data set that is entered into the prediction process to find out the predicted results of tile production.

E. The Process of Comparison of Prediction Model Results

In the process of comparing the results of the predicted models used the Test and Score widget is needed to calculate the success rate of predictions between each prediction model.

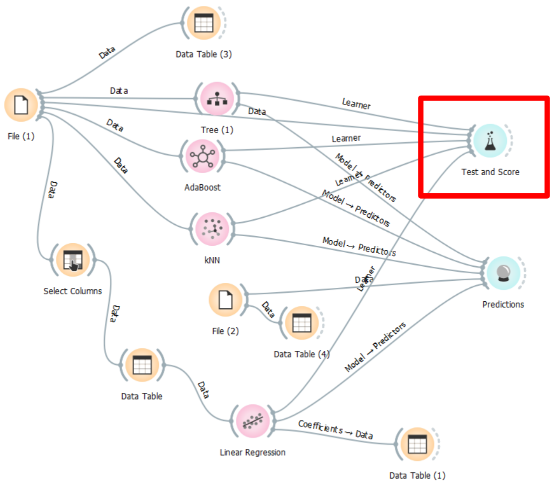


Fig6. Widget Design Calculates the Success of Prediction Models

In Figure 6 is a widget design that has been added to the process of calculating the success rate of the prediction model using the Test and Score widget which is correlated with the prediction model used previously.

F. *The Process of Calculating the Most Influential Variables*

In the process of calculating the most influential variables using the Linear Regression Widget with the variable labor, raw materials, fuel and the target variable is the tile production produced.

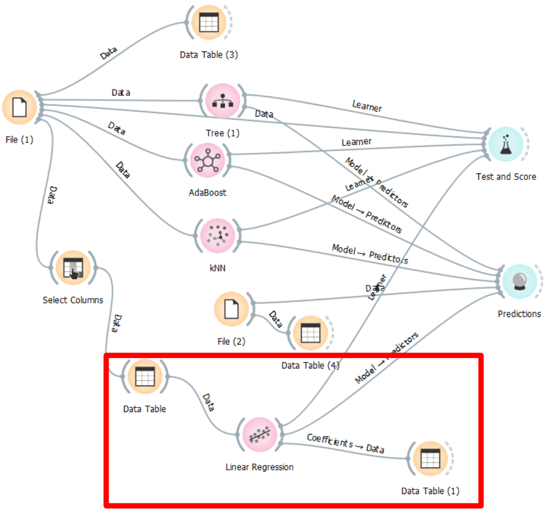


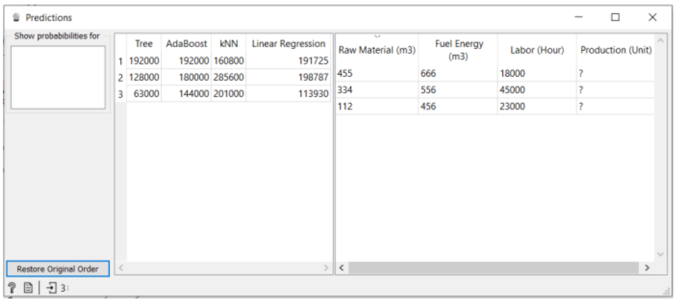
Fig7. Widget Design Counts the Most Sounding Variables

In Figure 7 is the widget design that has been added to the process of calculating the influential variables using linear regression widget.

IV. Result and Discussion

A. *Prediction Model Trial Results*

The results of the prediction model trials are performed using a data set of trials with 3 data instances and 3 data attributes namely raw materials, fuel, and labor. Here are the results.



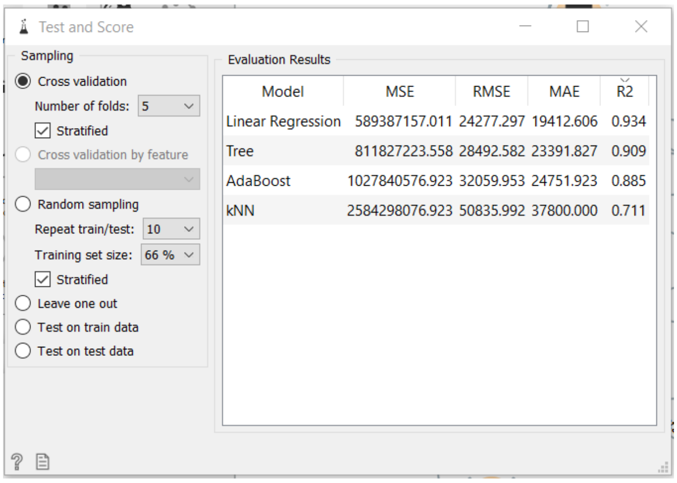
	Tree	AdaBoost	kNN	Linear Regression	Raw Material (m3)	Fuel Energy (m3)	Labor (Hour)	Production (Unit)
1	192000	192000	160800	191725	455	666	18000	?
2	128000	180000	285600	198787	334	556	45000	?
3	63000	144000	201000	113930	112	456	23000	?

Fig8. Prediction Results for Trial Data Sets

Figure 8 shows the results of tile production using a trial data set that was entered into the prediction model of Tree, AdaBoost, kNN, and Linear Regression. Prediction results from the four prediction models are not much different.

B. *Comparison of Prediction Mode Results*

The results of comparison tests of predictive models are performed using the Test and Score widget on the Orange tool. The results of the comparison are MSE (Mean Squared Error), RMSE, MAE (Mean absolute error) and R2 (oefficient of determination) values.



● Cross validation
Number of folds: 5
☑ Stratified
○ Cross validation by feature
○ Random sampling
Repeat train/test: 10
Training set size: 66 %
☑ Stratified
○ Leave one out
○ Test on train data
○ Test on test data

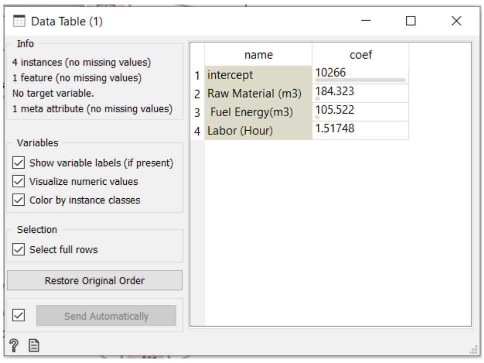
Model	MSE	RMSE	MAE	R2
Linear Regression	589387157.011	24277.297	19412.606	0.934
Tree	811827223.558	28492.582	23391.827	0.909
AdaBoost	1027840576.923	32059.953	24751.923	0.885
kNN	2584298076.923	50835.992	37800.000	0.711

Fig9. Comparison of Prediction Model Results

Figure 9 shows the results of a comparison score between the linear regression model, tree, AdaBoost kNN. From these results it can be seen that the Linear Regression prediction model has the highest score among the four other prediction models. So it can be said the results of predictions with linear regression prediction models in this case study prediction of tile production has a high degree of accuracy.

C. *Results of the Most Influential Variables*

The results of calculating the variables that most influence on tile production are carried out using Linear Regression with input variables for labor, raw materials, and fuel.



Data Table (1)

Info

4 instances (no missing values)
1 feature (no missing values)
No target variable.
1 meta attribute (no missing values)

Variables

☑ Show variable labels (if present)
☑ Visualize numeric values
☑ Color by instance classes

Selection

☑ Select full rows

	name	coef
1	intercept	10266
2	Raw Material (m3)	184.323
3	Fuel Energy(m3)	105.522
4	Labor (Hour)	1.51748

Fig10. Calculation Results of the Most Influential Variables

Figure 9 shows the results of the coefficient scores for the variables entered. The result is that the raw material variable has the highest score, which means that the raw material variable is the most influential variable on tile production in Tabanan Regency.

V. CONCLUSION

It can be concluded that the Orange data mining tool can be used easily to perform prediction modeling with the Tree, Linear Regression, kNN, and AdaBoost models. Prediction results for tile production based on the attributes of raw materials, fuel, and labor get fairly accurate results where for the prediction modeling that gets the highest score is the Linear Regression model. The tile production in Tabanan Regency is from the results of the linear regression calculation which is most influenced by the variable Raw Materials so that if the raw material is getting thinner it can result in the lack of tile production produced. With the results of this study the authors hope that these results can become new knowledge for related parties and industry owners to increase the production of the Tile Industry in Tabanan Regency.

REFERENCE

[1] A. E. Pramadhani och T. Setiadi, ”Penerapan Data Mining Untuk Klasifikasi Prediksi Penyakit Ispa (Infeksi Saluran Pernapasan Akut) Dengan Algoritma Decision Tree (ID3),” *Jurnal Sarjana Teknik Informatika,* vol. 2, nr 1, pp. 831-839, 2014.

[2] I. A. P. Sri Handayani och I. B. P. Purbadharmaja, ”ANALISIS ECONOMIC OF SCALE DAN EFISIENSI PENGGUNAAN INPUT TERHADAP OUTPUT PADA INDUSTRI GENTENG DI KECAMATAN KEDIRI KABUPATEN TABANAN,” *E-Jurnal EP Unud,* vol. 8, nr 5, pp. 974-1002, 2019.

[3] P. A. Widya och M. Sudarma, ”Implementation of EM Algorithm in Data Mining for Clustering Female Cooperative,” *International Journal of Engineering and Emerging Technology,* vol. 3, nr 1, pp. 75-79, 2018.

[4] . M. S. Kukasvadiya och D. H. Divecha, ”Analysis of Data Using Data Mining tool,” *International Journal of Engineering Development and Research,* vol. 5, nr 2, pp. 1836-1840, 2017.

[5] Wahyudin, I. P. Ari Wijaya och I. B. Alit Swamardika, ”Data Mining for Clustering Revenue Plan Expense Area (APBD) by using K-Means Algorithm,” *International Journal of Engineering and Emerging Technology,* vol. 2, nr 1, pp. 87-93, 2017.

[6] D. Ardiada, P. Agus Ariawan och M. Sudarma, ”Evaluation of Supporting Work Quality Using K-Means Algorithm,” *International Journal of Engineering and Emerging Technology,* vol. 3, nr 1, pp. 52-55, 2018.

[7] P. Mamang Weking, I. G. N. Wira Partha och A. Ibi Weking, ”Application ofData Mining withSupport Vector Machine (SVM) inSellingPrediction Trend of Spiritual Goods (Case Study:PT. XBali),” *International Journal of Engineering and Emerging Technology,* vol. 4, nr 1, pp. 20-24, 2019.

[8] L. G. Putri Suardani, I. M. Adi Bhaskara och M. Sudarma, ”Optimization of Feature Selection Using Genetic Algorithm with Naïve Bayes Classification for Home Improvement Recipients,” *International Journal of Engineering and Emerging Technology,* vol. 3, nr 1, pp. 66-70, 2018.

[9] P. Saini, S. Rai och A. K. Jain, ”Decision Tree Algorithm Implementation Using Educational Data,” *International Journal of Computer-Aided technologies,* vol. 1, nr 1, pp. 31-41, 2014.

[10] D. A. Putri Wulandari, K. A. BudiPermana och M. Sudarma, ”Prediction of Days in Hospital Dengue Fever Patients using K-Nearest Neighbor,” *International Journal of Engineering and Emerging Technology,* vol. 3, nr 1, pp. 23-25, 2018.

[11] I. G. Harsemadi, I. M. Sudarma och N. Pramaita, ”Implementasi Algoritma K-Nearest Neighbor pada Perangkat Lunak Pengelompokan Musik untuk Menentukan Suasana Hati,” *Teknologi Elektro,* vol. 16, nr 1, pp. 15-20, 2017.

[12] I. W. Agus Surya Darma och M. Sudarma, ”The Identification of Balinese Scripts' Characters based on Semantic Feature and K Nearest Neighbor,” *International Journal of Computer Applications,* vol. 91, nr 1, pp. 14-18, 2014.

[13] S. Mulyati, Yulianti och A. Saifudin, ”PENERAPAN RESAMPLING DAN ADABOOST UNTUK PENANGANAN MASALAH KETIDAKSEIMBANGAN KELAS BERBASIS NAÏVE BAYES PADA PREDIKSI CHURN PELANGGAN,” *JURNAL INFORMATIKA UNIVERSITAS PAMULANG,* vol. 2, nr 4, pp. 190-199, 2017.

[14] I. G. N. Agung Surya Mahendra, I. B. Leo Mahadya Suta och M. Sudarma, ”Classification of Data Mining with Adaboost Method in Determining Credit Providing for Customers,” *International Journal of Engineering and Emerging Technology,* vol. 4, nr 1, pp. 31-36, 2019.

[15] Amrin, ”DATA MINING DENGAN REGRESI LINIER BERGANDA UNTUK PERAMALAN TINGKAT INFLASI,” *Jurnal Techno Nusa Mandiri,* vol. 13, nr 1, pp. 74-79, 2016.