ARIKA, Vol. 06, No. 2 Agustus 2012
ISSN: 1978-1105

PREDICTING THE SELLING PRICE OF DRIED *EUCHEUMACOTTONII* IN INDONESIA WITH FOUR CLASSIFIER OF DATA MINING TECHNIQUES

Wilma Latuny
Graduated of Master of Philosophy in Maastricht School of Management/ PhD student in Tilburg University
e-mail: latuny@msm.nl/wlatuny@uvt.nl

ABSTRACT:

Dried Eucheumacottoniiseaweed (DES) is one of the main aquaculture commodities in Indonesia. Local farmers plant, harvest, and dry the seaweed, and subsequently sell it to traders. The selling price of seaweed depends on internal and external factors. To maximize their profit, farmers have to estimate the future price development by means of these factors.

This paper presents a novel data-driven method reports on our attempts to develop methods to aid farmers in making such predictions. More specifically, we apply data mining to the task of predicting the seaweed price eight weeks ahead, at the time of selling. The data mining algorithm, i.e., classifier, requires attributes as its inputs. In our experiments, we selected three internal and three external factors as input attributes. The internal-factorattributes used are: current price, dirty content, and moisture content. The three external-factor attributes all relate to the weather and areminimum temperature, maximum temperature, and precipitation. All attributes are measured at time t. The output of classifier is a binary classification indicating if the seaweed sales price at time t plus eight weeks is larger or smaller than the price at time t. The data is collected from two publicly available sources and consists of 275measurements of six attributes each.We evaluated the performances of four classifiers on our data set using 10-fold cross-validation. The results obtained revealed that we were able to predict the selling price of seaweed with an accuracy of64%. Analysis of the trained classifier revealed that the main attribute used for predicting the future price is the current price.

In conclusion, our data mining experiments suggest that it is feasible to predict the future price development of seaweedselling pricein Indonesia with accuracy above chance level.

Keywords*: seaweed, selling price, data mining.*

INTRODUCTION

The prediction of future sales prices for perishable products is an active topic of research (Li, Xu, & Li, 2010). Making accurate long-term predictions for market prices of agro-products is a considerable challenge because many hard-to-predict factors affect the price development (Li, et al., 2010). In Indonesia, aquaculture has become an important agricultural sector, due to the relative importance of wild and culture fish in rural household consumption (Xiaoshuan, Tao, Revell, & Zetian, 2005). Currently, aquaculture is a major alternative source of income for many farmers in Indonesia. In order to maximize their profit, farmers have to estimate the selling price of seaweed several weeks ahead in time, because of the required production time. Accurate estimates of the selling price, allows farmers to make qualified decisions about investments in the production process. During the rainfall period, farmers had difficulty to determine seaweed selling price due to harvesting failed. In the rainy weather the drying process acquires 7-10 days so that long drying decreases the quality of seaweed neither its moisture content nor its dirty content or impurity. High water content and low purity of seaweed decline selling prices.The development of effective prediction methods to estimate the selling price of seaweed would be of great benefit to Indonesian aquaculture farmers.

Statistics offers a large variety of methods for making predictions (Zhang, Chen, & Wang, 2010). For instance, multivariate linear regression methods have been used for the prediction of agriculture and livestock product price (Zhang, et al., 2010). Since the last decaseaweedseaweed, data mining methods have become popular due to their predictive power and associated evaluation framework, (Li, et al., 2010). An example of a successful application of data mining is Li et al. (2010), who trained a neural network classifier on the task of predicting agriculture prices. It was shown that the one-day, one-week, and one-month prediction accuracies obtained with neural networks outperformed those obtained with time series models. Other example that indicated the successful of data mining technique in agriculture domain is the prediction of rainfall using back propagation neural network model. By using the most widely technique that is artificial neural network the accuracy of prediction obtained 99.79 % accuracy on training of data and 94.28 % accuracy in testing of data. This result is useful for irrigation system of agriculture product to the farmers in India (Enireddy, Varma, Rao, & Satapati, 2010). In addition, the development of decision problem support system for forecasting aquatic product price in China can provide the annual, quarterly, monthly, weekly and 3-daily price forecast for freshwater fishes as well as illuminate the forecast result graphically.

After the model was tested by farmer, now the model can aidto predict of aquatic product,(Xiaoshuan, et al., 2005). Given the success of data mining methods in predicting prices in the agro-culture domain, our goal is to apply data mining to the task of predicting the sales price of seaweed.

In order to predict the future selling price of seaweed the factors affecting the selling price of seaweed need to be identified. There are two types of factors that influence the price development of agro-products in general and of seaweed in particular: *internal factors* and *external factors*. Examples of internal factors are: the amount of per capita consumption on agro-products, food price for agro-products, and recent agro-product sales prices(Zhang, et al., 2010). Examples of external factors are: transport costs and weather circumstances. For seaweed, the internal and external factors that affect the sales price have been identified. According to the Food Agriculture Organization (FAO) the following two internal factors affect the price of seaweed: (1) the moisture of the seaweed contents ("moisture contents") and (2) the impurity of the seaweed contents ("dirty contents")(Trono, 1990). Moisture contents are an important factor as they dictate the market price of seaweeds. The recommended moisture content is about 35% upon storage (Pelinggon & Tito, 2009). Dirty contents are also of importance because the intensity of cleaning and drying affect the quality of seaweed. Cleaning is done by removing sediments and other unwanted particles by rinsing the seaweed several times while still in the sea. Further removal of epiphytes, other animals (shells, soft corals, sponges) and tying material ('tie-tie') is done on land prior to drying.(3) The recent selling price is the third internal factor that influence seaweed selling price. Although the price has remained the same for some years prediction for example the price from 1987 to 1991, the current selling price will be the starting point to predict the future price.(McHugh, 1990)Further, for external factors, there are three factors influence seaweed price following: (1) precipitation where the large amount of precipitation has a devastating effect on the seaweed cultivation and harvesting stages. For example, the provinces Sulawesi Selatan and Bali have a high precipitation in the period from July to September. These provinces are major seaweed production centers and the precipitation influence the production cycle of cottonii. Extensive rain reduces the level of salinity and increase the sea level temperature. Both reduced salinity and increase water temperature hamper the growth of seaweeds (Bank Indonesia, 2006) and may cause a disease known as 'ice-ice', that kills seaweed. (2)Maximum temperature and (3) minimum temperature affect both the cultivation and harvesting of seaweed. The cottoniiseaweed is planted and harvested preferably at a temperature of 27^0C for minimum temperature and 32^0C for maximum temperature. Any deviation from this temperature range may affect the seaweed negatively. Fortunately, the coastal regions of Indonesia have temperatures that mostly fall into the required range (Bank Indonesia, 2006).

Our study addresses the following three research questions.

1. To what extent can the seaweedselling pricebe predicted 8 weeks ahead?
2. Which factors determine thefuture seaweed selling price?
3. Which data mining methodsare suitable for the prediction of seaweed selling price?

To answer these research questions, we perform data mining experiments using publicly available data on relevant factors and on seaweed price.The rest of the paper is organized as follows; Section 2 presents our data mining method of seaweed selling price prediction.Section 3 describesthe experimental set-up.Section 4reports on the results obtained and discusses. Finally, section 5conclusionand discusses future direction of the research.

DATA MINING METHOD

The data mining method used topredict theselling price ofdried*EucheumaCottonii*seaweed (SEAWEED) is illustrated in Figure 1. The method consists of three components, represented by three bounding boxes. The box on the left represents the input attributes to the data mining algorithm. The two types of attributes are represented by separate boxes labeled "internal factors" and "external factors". The box in the middle represents the data mining algorithm. It takes the attributes as input and generated a prediction as output. The box on the right represents the prediction and the evaluation (i.e., comparison with the actual price).

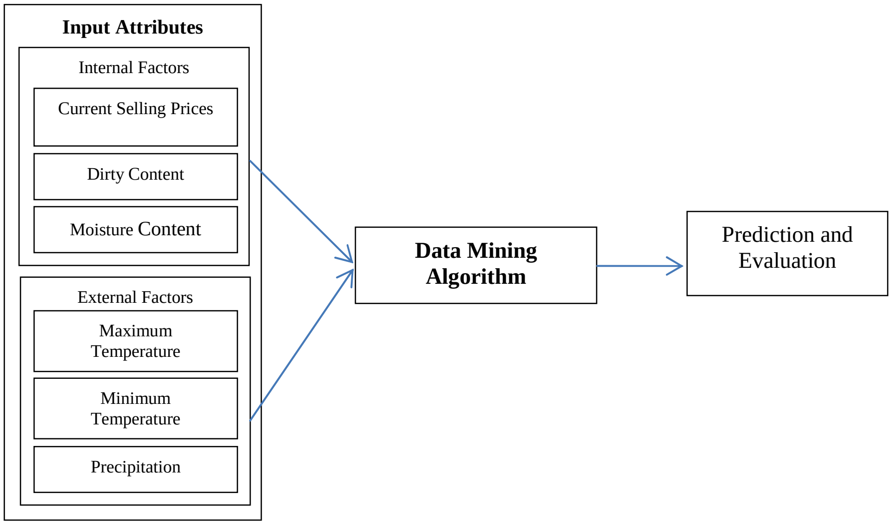


Illustration of the data mining method for the prediction of the selling price of Dried *EucheumaCottonii*Seaweed

The following subsections discuss the three components in more detail.

Attributes

The attributes selected as inputs to the data mining algorithm were identified from literature study. (Tiroba, 2006)

1. Seaweed current selling price (CP) is defined as the price of seaweed sold by farmers to the traders at that time
2. Seaweed dirty content (DC) is defined as level of foreign matter; sands, dirt, raffia, fish, shells, stones, coral pieces or excessive salt were exist on seaweed which can damage processing equipment. These could affect the price paid for seaweed. The impurities is < 3 %(Tiroba, 2006)
3. Seaweed moisture content (MC) is defined as levels of seaweed were dried from water and the ability of seaweed to be compressed into bales. Moisture content is very critical because it dictates the market of seaweed. The recommended moisture content is about 35 % upon storage(Pelinggon & Tito, 2009)
4. Maximum temperature (MaxT) is defined as the maximum temperature on the global climate which can give the best fit to photosynthesis of seaweed where is cultivated and the maximum temperature to dry the seaweed after harvesting. Maximum temperature is 32^0C. (Bank Indonesia, 2006; Chen, Rogoff, & Rossi, 2010)
5. Minimum temperature (MinT) is defined as the minimum temperature on the global climate which can give the best fit to photosynthesis of seaweed where is cultivated and the minimum temperature to dry the seaweed after harvesting. Minimum temperature is 27^0C. (Bank Indonesia, 2006)
6. Precipitation (P) is defined as the density of rainy which decreases salinity of seawater and 'ice-ice' diseases affecting on harvesting and drying of Seaweed.(Bank Indonesia, 2006)

Data Mining Algorithm

The data mining algorithm takes the attributes as input and transforms their values into a prediction value. The determination of the most appropriate data mining algorithm should be largely determined empirically. According to the famous "no free lunch theorem" ((Wolpert & Macready, 1997), algorithm performance depends on the application domain. Therefore, we examine four representative classifiers that often yield different performances: the decision tree classifier, the naïve Bayes classifier, the *k*-nearest neighbor classifier, and the support vector machine classifier.

1. The decision treeclassifier

The key rule of tree models is to divider (in recursive manner) the space covered by the input variables to maximize a score of class purity-meaning (approximately, depending on the particular score selected) that the majority of points in each cell of the partition belong to one class, (Hand, Mannila, & Smyth, 2001). Indeed, to predict the class of value of such a new case and the input of variables are determined, we work down the tree and at each node selecting the proper branch by comparing the new case with threshold value of the variable for that node. In addition, in building tree models, we merely recursively divide the cell of the space of input variables. To divide a given cell, we can explore each possible threshold for each variable to find the threshold split that lead to the ultimate improvement in a specified score function. The score function is evaluated on the basis of training data set elements. Basically, this splitting technique can be continued until each leaf node has a single training data point. On this experiment, we use weka classifier in order to classify the dataset through J.48 decision tree.(Witten, Frank, & Hall, 2011).

2. The *k*-Nearest Neighbor (*k*-NN) classifier

The *k*-Nearest Neighbor classifier is the method to classify a new object, with input vector by examining the *k closest* training data set points to the input and based the assignment of the class that has predominant of points among these *k*,(Witten, et al., 2011).

In principal, there are three important key of this technique; a set of label object, e.g., a set of stored records, a distance or dimensional input space to compute between objects and the value of the *k* as a number of nearest neighbors. In order to classify unlabeled objects, the distance of this object to the labeled object is processed, then *k*-nearest neighbors are identified, and we obtain the class label of these nearest neighbors which used then to determine the class label of the object(Wu, et al., 2007). On weka term, *k*-NN is represented by IB1 and IBk classifier.

3. The naïve Bayes classifier

The Naïve Bayes is the method on the class conditional distribution in which variables are all categorical are straightforward. We simply estimate the probabilities that an object from each lass will fall in each cell of the discrete variables (each possible discrete value of the vector variable X), and the use Bayes theorem to produce a classification. In practice, however, this is often very difficult to implement because of the sheer number of probabilities that must be estimated- estimated-O (k^p) for *p k-valued* variables. For example in this case we have p = 259 and binary variables ($k = 2$) we would need to estimate on the order of 2^{259} probabilities. Assuming (as a rule of thumb) that we should have at least 10 data points for every parameter we estimate (where here the parameters in our model are the probabilities specifying the joint distribution), we would need on the order of 10^{10} data points to accurately estimate the required joint distribution. For *m* classes (m>2) we would need *m* times this number. As *p* grows the situation clearly becomes impractical.

4. The Support Vector Machine classifier

Support vector machine (SVM) is one of the classification methods to distinguish between members of the two classes in the training data. The metric for the concept could be recognized geometrically,(Witten, et al., 2011). For datasets in which split linearly, a linear classification function link to a separating hyperplane that cross the middle of the two classes, separating the two. When this function is determined, the new instance can be classified by merely testing the sign of new function. The new instance determined belongs to the positive class if the new function > 0. SVM ensure to find the best function of linear hyperplanes by maximizing the margin between two classes. Naturally, the margin is defined as the amount of space, or partition between two classes as defined by the hyperplane. Geometrically, the margin corresponds to the shortest distance between the nearest data points to a point on the hyperplane. By this geometric definition can be extended to explore the way to maximize the margin, therefore, albeit there are a large number of hyperplanes, SVM can extract a few qualify as the solution. The reason behind SVM requires to seek the maximum hyperplanes is that SVM proviseaweed the best generalization ability. It allows not only the best classification performance on the training data, but also leaves much room for the correct classification of the future data. (Witten, et al., 2011). On weka classifier, SVM is represented by Sequential Minimal Optimization (SMO) using kernel functions such as polynomial or Gaussian kernels.

Prediction and evaluation

The third component of our data mining method concerns the prediction and the evaluation component. The prediction is defined as the SEAWEED selling price eight weeks ahead of time. This period is about the time needed before seaweed is ready for selling. The reason for this calculation because we count the activity of the farmers from beginning cultivating cottoni until harvesting required 45 cycle days, we add 3-5 days for drying and 2 days activity for shipping. In some cases if the weather was not good especially on rainy season, then the farmer needed 5-10 days for drying and 3-5 days for shipping and transporting. Thus the total days between planting-harvesting-shipping-selling are 50-60 days. The evaluation of the prediction is performed by comparing the predicted price at time t with the actual price at time t plus eight weeks. To estimate how our classifier performs on unseen data, we will employ a standard validation procedure (Witten, et al., 2011).

EXPERIMENTAL SET-UP

We performed the experimental evaluation of our data mining method using the public domain WEKA[1] data mining software. This section describes the set-up of our data mining experiments. We outline the data sets used (3.1), the classifier settings (3.2), and the evaluation and performance criteria.

Datasets

The datasetswere obtained from two public sources, one for the internal factors and the other for the external factors. The source for the internal factors was www.jasuda.net, an online publicly source that provides the selling price of Dried Seaweed *Eucheuma Cottonii*(seaweed) levelof farmersinIndonesiafrom 2005 until beginning of 2012. The amount of datadownloadedis 2134 objects which are bi-weeklyreports originating from 20 provincesof Indonesia from 2005 toFebruary 2012. The dataprovidedconsists ofseaweed selling price on farmer, dirty content, and moisture content that are internal factors influenced seaweed selling price.

The source for the external factors was www.freemeteor.gr, an online publicsource providinghistorical weather data worldwide.We downloaded daily weather reportsfrom the period 1January 2005 – 1 February 2012yielding40.880reportsfromeightweathermeteorology stationsinIndonesia. Data records encompass the daily weather reports, more specifically the maximum temperature, the minimum temperature and the precipitation.

Table 1 lists the six attributes so obtained. The second column specifies the attribute names and their abbreviations and the third column lists the units of measurent. The first attribute will also be used for defining the target (output) value (see below).

We preprocessed the raw data in two steps. First, we identified the dates at which a seaweedselling price was reported from at least one location. Second, for each date we averaged over the selling prices reported on that date to obtain estimates of the attributes for Indonesia. Since the reports at any location were generated quite irregularly, using average values ensures a more or less regular temporal sampling of the price and the other internal attribute values.

Using the preprocessed data, we selected all reports at time t for which the seaweed selling price was reported at time t plus 50-60 days. This was the case for 275 reports. As a consequence, our final data set consists of 275 instances. and 3-5 days for shipping and transporting. Thus the total days between planting-harvesting-shipping-selling are 50-60 days.

Attributes of SEAWEED selling prices prediction

No.	Attribute	Unit of Measurement
1.	Current Selling Price (CP)	Rupiah (Rp)
2.	Dirty Contents (DC)	Percent (%)
3.	Moisture Content (MC)	Percent (%)
4.	Maximum Temperature (MaxT)	Degree of Celcius (^{0}C)
5.	Minimum Temperature (MinT)	Degree of Celcius (^{0}C)
6.	Precipitation (P)	Millimeter (mm)

[1]

Classifier

Each of the four classifiers evaluated (naïve Bayes, k-nearest neighbor, decision tree, and support vector machine) requires the setting ofparameters. Table 2 lists the parameters and their settings for the four classifiers.

The classifier, parameter and output of anlysis

Classifier	Parameter	Output of analysis
Naïve Bayes	• Normal distribution for numeric attributes assumption (Current selling price, DC, MC, MaxT, MinT) • Mean, standard deviation after classifying • Correctly and incorrectly classification • Confusion Matrix • Frequency count of nominal value (binary class increase =1 and decrease =0)	• The number of classification occur per each attribute • The values of the mean, the standard deviation, the weight sum and precision have been classified • The percentage of correctly classified instances • The number of instances have been classified correctly or incorrectly
k-Nearest Neighbor Classifier	• k (number of neighbors) = 1,5,10 • Geometric distances between k nearest neighbor of vector dimension of instances • Correctly and incorrectly classification • Confusion Matrix • Frequency count of nominal value (binary class increase =1 and decrease =0)	• The percentage of correctly classified instances • The number of instances have been classified correctly or incorrectly
Decision Tree	• Obtain the leaves pruned as result the attributes to predict • Correctly and incorrectly classification • Confusion Matrix • Frequency count of nominal value (binary class increase =1 and decrease =0)	• The significant attributes in which as influence factors to predict after pruning process • The percentage of correctly classified instances • The number of instances have been classified correctly or incorrectly
Support Vector Machine	• Polynomial Kernel • Machine Linear showing normalization attributes • Number of Kernel's evaluation • Correctly and incorrectly classification • Confusion Matrix • Frequency count of nominal value (binary class increase =1 and decrease =0)	• The attributes in which include in support vector • The percentage of correctly classified instances • The number of instances have been classified correctly or incorrectly

Evaluation and Performance Criteria

The evaluation of the classifier requires target values, i.e., the true output values. These were defined as follows. If the selling price at time t plus 50-60 days is larger than the selling price at time t, the target is set to 1. In all other cases it is set to 0.

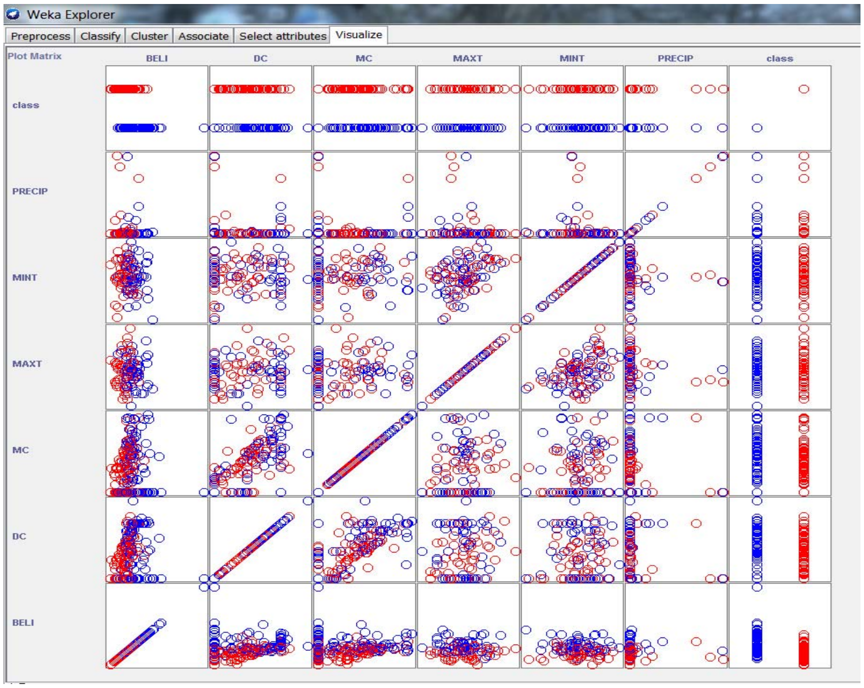
To estimate the prediction error of trained classifiers, we employ 10-fold cross-validation ((Witten, et al., 2011). Our performance criteria are a prediction accuracy that differs significantly from chance level (50%) and a balanced confusion matrix in the sense that the true positives and true negatives are about equal.

RESULTS AND DISCUSSION

We start by visualizing the data set in terms of the attributes (4.1). Subsequently, we present the classification results for the four classifiers (4.2). Finally, we discuss the results by analyzing the decision tree and by examining the relation of the most predictive attribute with the target value (4.3).

Visualization of datasets

Using WEKA's visualization module, we plotted the pair-wise attribute scatter plots shown in Figure 2. The 7 columns and rows of the figure show the joint distribution of the six attributes values (first six columns and rows) and of the class or target value (seventh column and row). Within each plot, each circle represents two attribute (or class) values. The colors of the circles represent the class. Red represent 0 (decreased price) and blue 1 (increased price). To alleviate the dense clustering of data points, we added a slight amount of jitter to the positions of the data points.



visualization of the relationships between 6 attributes

Classification results

The classification results obtained for the four classifiers are presented in Table 3. For each classifier, the percentages correctly and incorrectly classified instances are listed. In addition, confusion matrices are given. Result acquire that the Naïve Bayes and SVMtechnique provide the best performance of classifier compare than other technique. There are 180 correctly classified instance in or 64.45% for both Naïve Bayes and SVM. More result can be seen in table

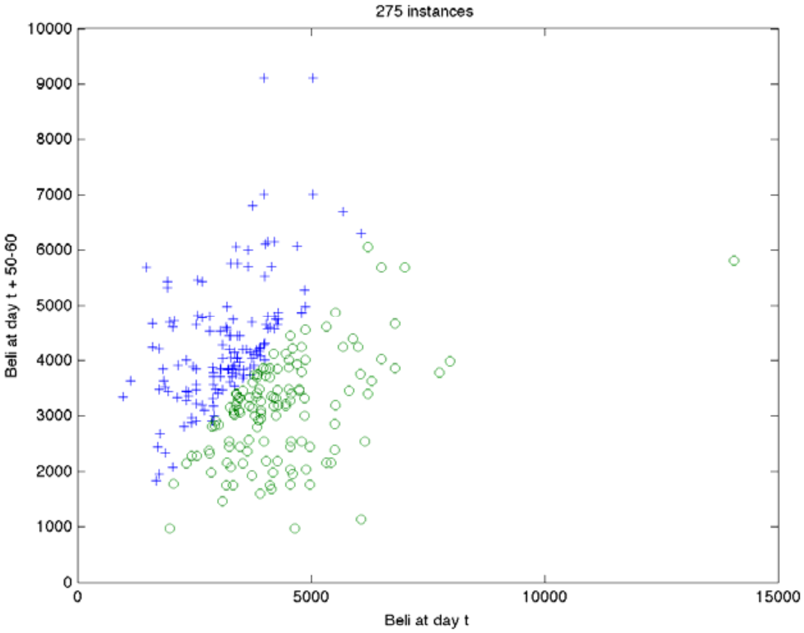
. Overview of the classification results

No.	Classifier Technique	Percentage of instances	Confusion Matrix		
			Binary		
		Correctly Classified		a=0	b=1
1.	Naïve Bayes	64%	a=0	75	56
			b=1	39	105
2.	IB1	60%	a=0	75	56
			b=1	39	105
3.	IBk k =10	63%	a=0	80	51
			b=1	58	56
4.	J48	64%	a=0	76	55
			b=1	43	101
5.	SVM	64%	a=0	67	64
			b=1	31	113

The results contribute to the farmers' decision to predict the SEAWEEDseaweed. If the farmer'sseaweed desire to know whether increase or decrease of the selling prices, they would revert to the number of Current Selling price to forecastthe next price happened.

Discussion

We analyzed the trained decision tree to determine the most informative attribute. The structure of the tree consisted of a single level defined by the attribute current selling price. Figure 4 is a scatter plot of the SEAWEED price at day t versus the SEAWEED price at day t +50-60 days. The decreases and increases in price are represented by – and + signs, respectively. As can be seen from the plot, the value of the current price is weakly related to the future value of the price. Roughly speaking, a lower value of the price (Beli) at day t, predicts an increase of the price at day t+50-60. This weak relation is exploited by our data mining algorithms.



Current price at day t versus at day t+50-60days

CONCLUSIONS AND FUTURE WORK

We state our conclusions by answering the three research questions.

1. To what extent can the SEAWEED selling price be predicted 8 weeks ahead?
2. Which factors determine the future SEAWEED selling price?
3. Which data mining methodsare suitable for the prediction of SEAWEED selling price?

First, using data mining it is possible to predict the SEAWEED selling price in Indonesia with an accuracy of 64 %. Second, the mainpredicting factor of the SEAWEED selling price is the current SEAWEED selling price. Third, the best predictions are obtained with the Naïve Bayes classifier and the Support Vector Machine classifier.

Of course, our conclusions are based on a restricted sample of SEAWEED selling price data and a limited number of internal and external factors. In our future research we will explore larger data sets and additional factors. In particular, we aim at extracting guidelines from the trained classifiers in order to discover guidelines for farmer prediction to maximize their profit.

REFERENCES

Bank Indonesia. (2006). Pola pembiayaan usaha kecil (PPUK). Jakarta: Direktorat kredit, BPR dan UMKM Bank Indonesia.

Chen, Y.-c., Rogoff, K. S., & Rossi, B. (2010). Predicting agri-commodity prices: an asset pricing approach (D. o. Economics, Trans.) (Vol. I, pp. 45). Washington: University of Washington.

Enireddy, V., Varma, K. V. S. R. P., Rao, S. P., & Satapati, R. (2010). Prediction of rainfall using backpropagation neural network model. *International Journal on Computer Science and Engineering (IJCSE), 02* 1119-1121.

Hand, D., Mannila, H., &Smyth, P. (2001). *Principles of data mining*: Massachusetts Institute of Technology.

Li, G.-q., Xu, S.-w., & Li, Z.-m. (2010). Short-Term Price Forecasting For Agro-products Using Artificial Neural Networks. *Agriculture and Agricultural Science Procedia, 1*(0), 278-287.

McHugh, D. J. (1990). *Prospects for eucheuma marketing in the world and the future of seaweed farming in the pacific.* Paper presented at the The regional workshops on seaweed culture and marketing Fiji.

Pelinggon, R. E., & Tito, O. D. (2009). Enhancing the demands of AFNR graduates through curricular intervention using modular approach with high S & T *Seaweed production*

Tiroba, G. (2006). *A practical guide for seaweed farmers, Seaweed extension officers, Buying agents, Fisheries officers and Exporters.*: CoSPSI.

Trono, G. C. J. (1990). *Lesson from the history of seaweed culture in the Philippines and the trend of seaweed farming in Southeast Asia.* Paper presented at the The regional workshops on seaweed culture and marketing, Fiji.

Witten, I. A., Frank, E., & Hall, M. A. (2011). Data Mining practical machine learning tools and technique

Wolpert, D. H., & Macready, W. G. (1997). No Free Lunch Theorems for Optimization. *IEEE Transaction on Evolutionary, I*, 67-82.

Wu, X., Kuma, V., Quinlan, R. J., Ghosh, J., Yang, Q., Motoda, H., et al. (2007). The top 10 on Data Mining Algorithm. Retrieved from http://www.cs.uvm.edu/~icdm/algorithms/10Algorithms-08.pdf doi:DOI 10.1007/s10115-007-0114-2

Xiaoshuan, Z., Tao, H., Revell, B., & Zetian, F. (2005). A forecasting support system for aquatic products price in China. *Expert Systems with Applications, 28*(1), 119-126.

Zhang, W., Chen, H., & Wang, M. (2010). *A forecast model of agricultural and livestock product price.* Paper presented at the International federation for information processing (IFIP), Beijing.