The design of website-based sugarcane forecasting information system

To cite this article: M Effendi et al 2020 IOP Conf. Ser.: Earth Environ. Sci. 524 012024

View the article online for updates and enhancements.
The design of website-based sugarcane forecasting information system

M Effendi, D D Yuliani, D T Setyawan and U Effendi
Department of Agroindustrial Technology, Faculty of Agricultural Technology, Universitas Brawijaya, Malang, Indonesia
E-mail: mas.ud@ub.ac.id

Abstract. Sugarcane (Saccharum officinarum) is widely used as raw material for sugar and MSG. Data on sugarcane production has not been used optimally, except for administrative purposes. The production data can be used to predict the yield of sugar cane of which can be utilized by cooperatives and farmers. This research was conducted to design an information system that can be used to forecast sugarcane yields in the working area of KUD Subur Malang, Indonesia. The information system design process is carried out by implementing Machine Learning. The results of sugarcane yield forecasting using machine learning implementation in KUD Subur showed the best results using the gradient boosting algorithm with 68% model accuracy. Web-based yield forecasting information system can be used as a production forecasting tool for KUD Subur to improve its business processes. Sugarcane forecasting information system can be well received by users.

1. Introduction
Indonesia's growing population requires sufficient availability of basic needs [1, 2]. One of the basic needs of Indonesian is sugar. Sugar is one of the agricultural sector commodities that plays an important role in the national economy [2, 3]. The level of direct consumption of sugar by Indonesian households tends to decrease, but the domestic demand for Indonesian sugar continues to increase due to the development of the food and beverage industry. The high sugar demands cannot be covered with domestic production. In 2018, Indonesia's sugar production shortage was 5.03 million tons. This encourages sugar imports in Indonesia, which tends to increase every year [4]. One of the disadvantages of sugar production in Indonesia is the limitation in the production of sugar cane (Saccharum officinarum). Therefore, the supply of sugar cane becomes important in supporting the national sugar [2, 4].

The production of sugar cane as a raw material for sugar in Indonesia is mostly managed by the KUD [5]. KUD Subur is one of the KUD that operates in sugar cane production management. KUD Subur has a superior program namely the Sugar Cane Partnership in collaboration with sugar factories in Malang. This program provides a 70% credit guarantee for farmers and 30% for factories. KUD Subur helps channel the farmers' sugarcane yields to the sugar factory.

TRK activity data has been recorded manually so that the data generated does not provide great benefits for KUD Subur. Further utilization of data on the results of the TRK program can be done by developing information systems that can minimize errors that often occur in the process of manually...
inputting data, facilitate the dissemination of information to those who need it, and the resulting data can be used as a reference in the future [6]. Utilization of the data is needed as sugarcane production forecasting information. Therefore, this research was conducted to design an information system that can be used to forecast sugarcane production in the working area of KUD Subur.

2. Research Methods
This study was conducted in 2019 in Malang Regency, Indonesia. Research variables used in the research include land area, types of seeds, land location, and rainfall rate obtained from KUD Subur and BPS. The information system compares 5 different algorithms, namely Gradient boosting, Linear regression, Support vector machines, Random forest, and MLPregressor. The information system design process is carried out by implementing Machine Learning in Python [7-9].

Designing machine learning model use some stages [7-9]. The stages of this research are as follows:
1. Data collection. Data used for research consist of the area of sugarcane land, types of seeds of sugarcane, rainfall in Malang, and the number of farmers.
2. Data exploration and preparation. Data exploration is carried out aimed at improving data quality such as eliminating unnecessary data. Because the quality of the resulting model is very dependent on the data used.
3. Model development and training. At this stage, the training set is used as input data
4. Evaluate the model. Model evaluation is carried out to determine the performance of the resulting model, the model generated in the previous stage will be tested using a test set as input data.
5. Improvements to the model. Improvements to the model can be done in several ways to get good performance. Improvements can be done including by changing the type of learning or algorithm used or by making improvements to the data used either by adding data, reducing features in the dataset and others.

The interface of the forecasting information system is written in Bahasa Indonesia for the convenient of the users.

3. Results and Discussion
3.1. Dataset preparation
The dataset is prepared through the data acquisition, exploration and pre-processing stages. Data acquisition is done by collecting and preparing data from KUD Subur. The data used are the last 10 years with 5 attributes and 1 class with a total of 140 data rows. The attributes used include land area, year, land area, rainfall rate, number of farmers, and location. The class used is sugarcane yield. Data exploration was carried out to obtain information on the characteristics of the analysed data. Data exploration is done to get a dataset that can be read by Python in the form of a .csv file. Null value analysis is also carried out during data exploration which gets missing value ratio of all attributes of 0, except for the number of farmers which is worth 0.02857. It means 2.8% of data is lost. Domain knowledge analysis is also carried out during data exploration to obtain feature descriptions from datasets as in Table 1.

<table>
<thead>
<tr>
<th>No</th>
<th>Feature</th>
<th>Function</th>
<th>Data type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Location</td>
<td>Independent</td>
<td>Object</td>
<td>Location of sugarcane cultivation</td>
</tr>
<tr>
<td>2</td>
<td>Year</td>
<td>Independent</td>
<td>Integer</td>
<td>Year of cultivation</td>
</tr>
<tr>
<td>3</td>
<td>Land area</td>
<td>Independent</td>
<td>Float</td>
<td>The total land area of each location</td>
</tr>
<tr>
<td>4</td>
<td>Rainfall rate</td>
<td>Independent</td>
<td>Float</td>
<td>The level of rainfall at each location</td>
</tr>
<tr>
<td>5</td>
<td>Number of farmers</td>
<td>Independent</td>
<td>Integer</td>
<td>Total farmers in each location</td>
</tr>
<tr>
<td>6</td>
<td>Yield</td>
<td>Dependent</td>
<td>Float</td>
<td>The yield of sugar cane production in each location</td>
</tr>
</tbody>
</table>
Data Pre-processing is done by missing value handling, label encoder, join all data, and feature correlation. This stage is to get dataset that is ready to process at the machine learning model selection stage. Missing value handling produces a dataset with 136 rows through the process of data amputation with null values [6]. The dataset obtained needs to be label encoder on location attributes that were initially categorical into numerical data type. The label encoder results are combined in the same dataset by joining all data. Feature correlation visualizes the effect of location, year, number of farmers, rainfall, and land area attributes on yield. The results of feature correlation can be seen in Figure 1. Feature correlation shows the attribute with the highest correlation is the land area with a relationship value of 0.08. The attribute that has the least effect on production results is the location with a correlation value of 0.23. The feature correlation results indicate that the higher the area, the higher the production yields.

![Figure 1. Feature correlation results](image)

### 3.2. Designing machine learning models

Machine learning model design is done by comparing linear regression algorithm, Gradient Boosting regressor, Decision three, Survey vector regressor, and MLPregressor and determining the attributes associated with prediction of sugarcane yield. The design of the model is done by comparing the accuracy of the model and the prediction results of several algorithms used. The best predicted algorithm model will be used as a machine learning model.

The design of machine learning models is done through the stages of training and evaluation, optimization, and model comparison. Model training and evaluation measure the accuracy of the training data and test data for each algorithm. Evaluation results can be seen in Table 2. The gradient boosting algorithm provides the highest accuracy value compared to other algorithms. The results of the evaluation are optimized in order to get a better model by looking for each algorithm's hyperparameter. SVM was excluded due to its low accuracy. The best hyperparameter is used to compare models between algorithms as in Table 3. The results of the model comparison show the gradient boosting algorithm remains the best algorithm and is used as the basis for developing information systems.
Table 2. Model evaluation results

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy of training data model</th>
<th>Accuracy of testing data model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient Boosting</td>
<td>0.82</td>
<td>0.76</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>0.65</td>
<td>0.72</td>
</tr>
<tr>
<td>Random Forrest Regressor</td>
<td>0.90</td>
<td>0.73</td>
</tr>
<tr>
<td>MLP Regressor</td>
<td>0.65</td>
<td>0.73</td>
</tr>
<tr>
<td>Survey Vector Machine</td>
<td>-0.13</td>
<td>-0.04</td>
</tr>
</tbody>
</table>

Table 3. Model comparison results

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Prediction</th>
<th>Target</th>
<th>Gap</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient Boosting</td>
<td>22.134</td>
<td>32.719</td>
<td>10.585</td>
<td>68</td>
</tr>
<tr>
<td>Linear regression</td>
<td>16.160</td>
<td>32.719</td>
<td>16.559</td>
<td>49</td>
</tr>
<tr>
<td>Random Forrest Regressor</td>
<td>19.066</td>
<td>32.719</td>
<td>13.653</td>
<td>58</td>
</tr>
<tr>
<td>MLP Regressor</td>
<td>50.760</td>
<td>32.719</td>
<td>18.041</td>
<td>45</td>
</tr>
</tbody>
</table>

Figure 2. The interface implementation

3.3. Information system development

Web-based information system was developed to predict or forecast sugarcane yields from KUD Subur. Website users can be divided into two, namely admin and user. Admin can add, delete, and modify the data contained in the system and make predictions of sugarcane yield. The admins are KUD Subur’s administrative officers. Users can only perform prediction of sugarcane yield in the
system. The user is the sugarcane land owner. Website specifications developed based on the requirement analysis results, including:

1. Website consists of 3 pages, namely the main page, data management page, and prediction page.
2. The main page is a page that contains a dashboard of sugarcane yield data of KUD Subur over the past 10 years. The main page is the page that appears when the user opens the website.
3. Data management page is a page that can only be accessed by the admin. On the data management page, admin can add, delete, and modify the dataset. Data management page has a login menu to access the page.
4. Prediction page is a page for predicting sugarcane yield. Before making a prediction, the user enters some inputs consisting of location, land area, number of farmers, year, and rainfall rate as independent variable to be predicted. The prediction page can be accessed by the land owner or KUD Subur administration officer.

The interface design of the information system website page is done before implementation. This design covers the entire interface. The agreed design is then continued with the implementation of the system interface. The interface implementation can be seen in Figure 2. The main interface page displays a dashboard of sugarcane production data over the past 10 years, total sugar production, number of farmers, and the area of sugarcane land that is sheltered by KUD Subur. The prediction page consists of prediction input forms and prediction results. The attributes inputted included location, land area, number of farmers, year, and rainfall rate.

3.4. System testing
The test is carried out to find out the results of the implementation of the system is running according to design in meeting user needs. Testing is done based on the results of the analysis of user needs. Tests conducted include black box testing and user acceptance testing. Blackbox test results indicate the system has successfully met functional requirements according to user needs analysis. User acceptance testing (UAT) is carried out on systems that have been developed by focusing on 3 criteria, namely performance, usability, and accuracy. UAT calculation results on the performance criteria of 68.75%, usability of 87.5%, and accuracy of 87.5%.

4. Conclusions
The sugarcane yield forecasting using implementation of machine learning in KUD Subur Malang, Indonesia showed that the best result was using the gradient boosting algorithm with 68% model accuracy. Web-based yield forecasting information system can be used as a production forecasting tool for KUD Subur to improve its business processes. Sugarcane forecasting information system can be well received by users. Further performance improvement is increasing the usability and accuracy of information systems.

References
[1] BPS-Statistics Indonesia 2013 Indonesia Population Projection 2010-2035 (Jakarta:BPS-Statistics Indonesia) [In Indonesian]
[4] Pusdatin Kementan 2019 Outlook Tebu (Jakarta: Pusdatin Kementan) [In Indonesian]
Kesejahteraan Petani" 2015 [In Indonesian]


