Sriwijaya Journal of Informatic and Applications
Vol. 01, No. 01, August 2020, pp. 14-21
ISSN xxxx-xxxx, DOI:xxxxx

Urea Fertilizer Quality Testing with Chi-Squared Automatic Interaction Detection (CHAID) Algorithm

Ahmad Nauvan Zikri Al Ghifran [a,1], Yunita [b,2], Desty Rodiah [c,3]

[a,b,c]*Sriwijaya University, Palembang, Indonesia*
[1] 09021381520075@students.ilkom.unsri.ac.id; [2] yunita@ilkom.unsri.ac.id; [3] destyrodiah@ilkom.unsri.ac.id

ARTICLE INFO

Article history
Received
Revised

Keywords
Expert System
Urea Fertilizer
Quality Testing
Chi-Squared Automatic Interaction Detection (CHAID)

ABSTRACT

PT. XYZ has a Laboratory section in each of its factories that performs its duties manually to determine the quality of the fertilizers to be produced. This manual method is most likely at risk of human error and causes errors in the results of determining the quality of urea fertilizer. An expert system was built using the Chi-Squared Automatic Interaction Detection (CHAID) algorithm which can test the quality of urea fertilizer. The CHAID algorithm applies the decision tree technique where the technique will always branch off two or more as a basis in establishing rules. The system takes the values of the urea fertilizer test parameters as attributes. These attributes are processed to produce the most significant values that will be branches in the decision tree. The parameters used include Nitrogen, Biuret, Moisture, Free Ammonia, Iron, Oil Content, Crushing Strength, and Size Distribution. CHAID algorithm is suitable to be used to test the quality of urea fertilizer because in this study produced 4 different decision trees with an accuracy value of 99% using as much as 100 test data. This number influenced by the amount of training data used to build the rules.

1. Introduction

Expert system is a system that contains information in the form of knowledge from an expert in his field and has the ability to interact with circumstances outside the system, usually in the form of instructions as outlined in a particular programming language [1]. In the industrial sector, an expert system is needed by a company to support the production process of the company itself.

PT. XYZ has a factories for production and laboratory for determine the quality of the fertilizers. Laboratory Section does its job manually by providing quality that matches the parameters used to determine the quality of the urea fertilizer supplied. The method used by the laboratory is to correct the human error, where the laboratory can make mistakes in testing the quality of fertilizers so that it can increase greatly in determining the quality of the urea fertilizer produced. With the opportunity to reduce the problems that occur, the authors want to issue an expert system that can improve the quality of urea fertilizer through parameters that improve the quality of urea fertilizer using Chi-Squared Automatic Interaction Detection (CHAID) algorithm.

CHAID algorithm works by studying the relationship between the dependent variable with several independent variables, then categorizing the sample based on this relationship. The results of the CHAID algorithm are categorized in a tree diagram [2]. CHAID algorithm is an iterative procedure that allows the data to be defined in the appropriate category and will give a sequence of variables as expected.

Based on the description above, to avoid the human error, expert system will be built to test the sample data of urea fertilizer test results. Besides that, the system will display the accuracy of CHAID algorithm.

2. Literature study / Hypotheses development

a. Expert System

Expert system is a system that works inside a computer by relying on human knowledge so that it can be solved as is usually done by experts. expert systems have characteristics that can overcome or solve problems, and provide solutions to user problems related to problems and requirements, and are able to process data that is not possible. Expert systems that must provide or provide accurate and reliable information, are capable of heuristics in order to obtain solutions, and easily facilitate the need to improve with the environment and can be used on various types of computers [3].

b. Chi-Squared Automatic Interaction Detection (CHAID)

CHAID algorithm participation between the dependent variable with the independent variables that support the partnership. Thus, the CHAID algorithm divides certain independent variables and optimal interactions with the dependent variable. The basic principle of the CHAID algorithm is to divide data with the aim of dividing data into subgroups based on the dependent variable. In other words, the CHAID algorithm divides data between the relationship of the dependent variable and the independent variable into several subgroups. The results of the CHAID algorithm are successful in the tree diagram [4].

c. Chi-Square Test

Chi-Square Test is used to find out whether there is a relationship between two specific variables or not at each level. In finding the value of Chi-Square, there are steps to test the hypothesis [5]:

1) Make an initial hypothesis
   a) H0 : Pij = Pi• P•j (variabel i dan variabel j are not related)
   b) H1 : Pij ≠ Pi· P·j (variabel i dan variabel j are related)
2) Determine the level of significance (α), generally at 5% or 0.05.
3) Determine the rejection area:

$$W > x^2_{Table}\,(\alpha;(b-1)(k-1)) \tag{1}$$

4) Find the value from:

$$W = \sum_{i=1}^{b}\sum_{j=1}^{k}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \tag{2}$$

5) Make a decision from the results obtained:
   a) If the value of W enters the rejection area, then H0 is rejected,
   b) If not, H0 is accepted.

d. The Steps of CHAID Algorithm

The CHAID algorithm is divided into 3 stages, namely merging, splitting, and stopping. The tree diagram starts from the root node through these three stages of the node that is formed and is repeated [5].

6) Merging
   a) Form two-way contingency tables for independent variables on the dependent variable.
   b) Calculate Chi-Squared statistics for each category that can be combined into one, to test its freedom in a 2 x J contingency sub table that has been formed by the pair of categories with the dependent variable that has J categories.
   c) For paired Chi-Squared values, calculate the p-value together. Among the insignificant pairs, combine one pair of the most similar categories (pairs that have the smallest paired Chi-Squared value and the largest p-value) into a single category, then move on to the fourth step.
   d) Re-check the level of significance for the new category after combining it with other categories in the independent variable. if there are still significant pairs, repeat the previous steps. If everything is significant, continue to the next step.

e) Calculate the p-value that has been corrected Bonferroni based on the tables that have been merged. Bonferroni correction is done if the independent variable has more categories than the existing category on the dependent variable.

7) Splitting
   a) Choose the independent variable that has the smallest (most significant) p-value to use as a split node.
   b) If the p-value is less than or equal to the alpha specification level (α), a split node will be performed on this independent variable. If there is no significant p-value, a split is not done and the node is transformed into a terminal node.

8) Stopping
   a) If the node has a significant value to the dependent variable and or independent variable, the node will not be split.
   b) If the tree has now reached the tree's maximum limit of specifications, growth will stop.
   c) If the node size is less than the minimum node size specification or contains a small number of observations, the node will not be split.
   d) If the result of a split node results in a child node with a value less than the minimum child node size specification or contains a small number of observations, the node will not be split.

e. Urea Fertilizer

Urea is a chemical fertilizer that contains high levels of Nitrogen (N). Nitrogen is a nutrient that is needed by plants. This urea fertilizer is in the form of white crystal grains. Urea is a fertilizer that dissolves easily in water and has properties that are very easy to suck water (hygroscopic). In general, urea fertilizer has a nitrogen content of 46%. To determine the nitrogen level, laboratory tests are usually carried out. Laboratory tests are also useful for determining the quality of the urea fertilizer produced.

3. Methodology

f. Data

The data used as the object of this study are secondary data obtained from certificates of laboratory urea fertilizer test results of PT. XYZ The data has been collected as many as 250 data during the last 4 years, from 2016 to 2019. From the certificate of the test results, it can be seen that the parameters in the urea fertilizer test include Nitrogen, Biuret, Moisture, Free Amonia, Iron, Oil Content, Crushing Strenght, and Size Distribution content. In addition to these parameters, there are standard data on the size of the tests performed and each parameter has a different standard. Table 1 shows the standard for each parameters [6]:

Table 1. Parameters of Urea Fertilizer Quality Testing

Test Parameters	Details	Measurement Standards	
		Worthy	***Not Worthy***
Nitrogen	Independent Variable	≥ 46% wt	< 46% wt
Biuret	Independent Variable	≤ 0,5% wt	> 0,5% wt
Moisture	Independent Variable	≤ 0,5% wt	> 0,5% wt
Free Ammonia	Independent Variable	≤ 150 ppm	> 150 ppm
Iron	Independent Variable	≤ 0,7 ppm	> 0,7 ppm
Oil Content	Independent Variable	≤ 50 ppm	> 50 ppm
Crushing Strenght	Independent Variable	≥ 14,20 kg/cm^2	< 14,20 kg/cm^2
Size Distribution	Independent Variable	≥ 98% wt	< 98% wt
Fertilizer Quality	Dependent Variable	1	0

In addition to the data on the certificate, there are data obtained from the results of interviews, namely the cause and solution data that will be carried out if there is a problem in the production of urea fertilizer. So in this study, we will use three attributes, consisting of test parameters, causes, and solutions.

g. Research Testing Criteria

In the testing phase of this research, the test data will go through several stages in detail, there are:

1) Training Data
   Prepare the data that will be used in this research

2) Knowledge-base
   In this study, using RBR because knowledge is represented in the form of IF-THEN rules. There are 2 parts in knowledge-base :

   a) *Chi-Squared Automatic Interaction Detection*
      After the training data has been processed, the application of the CHAID algorithm is started. The CHAID algorithm uses the decision tree concept, where nodes always divide by two or more. The independent variable will be the basis for determining the quality of urea fertilizer. While the dependent variable is to determine whether the quality of the fertilizer is feasible or not to be distributed. In applying the decision tree concept, the dependent variable will be the root node, and the independent variable will be the branch.

   b) *Save the Decision Tree*
      The results are stored in advance so that they can be processed to the next stage. The results of the evaluation of the CHAID algorithm on the training data in the form of rules that have been formed.

3) Consulting Environtment
   Used by people who are not experts to consult or as a benchmark for the process that has been run. There are 3 parts in consulting environment:

   a) *Consulting Data*
      Consultation data serves to measure the level of accuracy in the final results. Consultation data has the same function as test data, but in this study test data were created based on what was tested in determining the quality of urea fertilizer. The more data used, the more accurate it is because the errors are higher to be detected. The test data is actually almost similar to the training data, it's just distinguished by the amount of data. Therefore, it is possible for some of the training data to be used in the test data.

   b) *Read the Rules Built*
      Before finding a solution to the cause of the decline in the quality of urea fertilizer, the CHAID algorithm rules that have been stored, will be read again with the code. The concept uses rules obtained in the previous process, to perfect the results in the later stages.

   c) *Finding Solutions*
      The final stage is used for the process of finalizing a case or problem. Through several previous stages, it is expected to produce the right and accurate solution. You do this by looking at and comparing each standard parameter to be tested with the standard parameters that have been stored according to the rules. Wanted which rules meet all the criteria of the new parameter standard.

4. Result and Discussion

The software will be built to test the quality of urea fertilizer using the CHAID algorithm with the following capabilities:

1) Import data as training data;
2) Entering data from .csv (comma separated values) format into SQL (Structured Query Language);
3) Conduct training from training data;
4) Calculating the Chi-Squared value;
5) Calculates the p-value based on the Chi-Squared value obtained previously;
6) Select the most significant parameters (the largest Chi-Squared value and the smallest p-value) to be used as a new branch;
7) Discard parameters that have become branches;
8) Produces calculation results in the form of rules

Tests are carried out using 100 urea fertilizer test data with total of 250 data. The data includes the values of each parameter, along with the causes, solutions, data quality, and rule-based quality. The process of testing will be carried out in accordance with the software architecture. The experiment was carried out 10 times on each data with the consideration that the training data pattern will affect the rules that are built and also the accuracy in identifying the quality of urea fertilizer which can be seen from its accuracy value.

Table 2. Splitting of Training Data and Testing Data

Experiment	Training Data	Testing Data
1	A, B, C[a]	D, E[b]
2	A, B, D	C, E
3	A, B, E	C, D
4	A, C, D	B, E
5	A, C, E	B, D
6	A, D, E	B, C
7	B, C, D	A, E
8	B, C, E	A, D
9	B, D, E	A, C
10	C, D, E	A, B

[a.] A = Data between 1-50, B = Data between 51-100, C = Data between 101-150
[b.] D = Data between 151-200, E = Data between 201-250

Based on all experiments conducted, the result are:

Table 3. Experiments Result

Experiment	Training Data	Testing Data	Total True	Total False	Accuracy (%)
1	A, B, C	D, E	98	2	98%
2	A, B, D	C, E	99	1	99%
3	A, B, E	C, D	99	1	99%
4	A, C, D	B, E	94	6	94%
5	A, C, E	B, D	94	6	94%
6	A, D, E	B, C	86	14	86%
7	B, C, D	A, E	96	4	96%
8	B, C, E	A, D	98	2	98%

Experiment	Training Data	Testing Data	Total True	Total False	Accuracy (%)
9	B, D, E	A, C	99	1	99%
10	C, D, E	A, B	99	1	99%

The results of testing experiments that have been carried out, obtained the level of accuracy or percentage value of the match between the results of the quality data and the results of the quality of the system that is the second, third, ninth, and tenth experiment of 99%. Based on all four test experiment, each has a different decision tree. The decision tree of all four test experiments will be shown below:

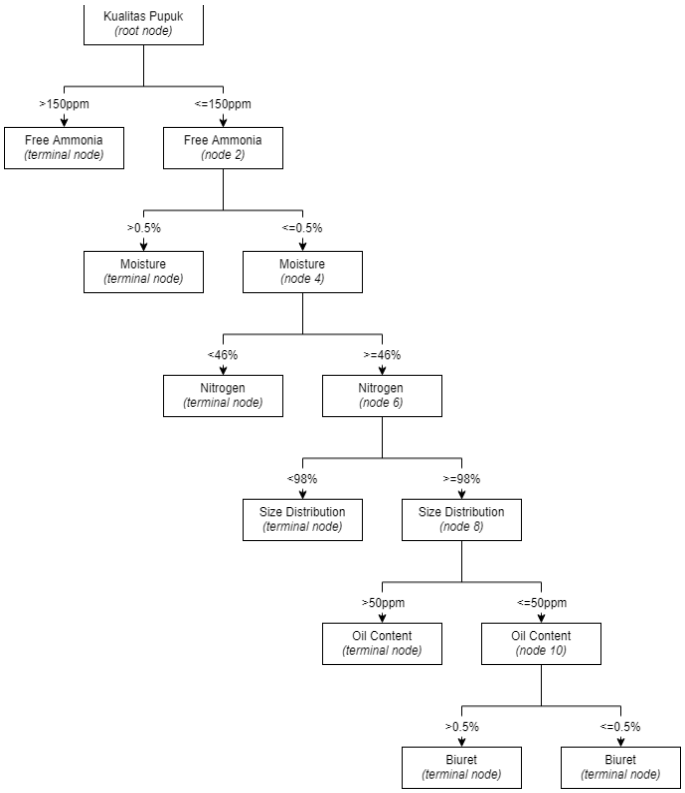

Fig. 1. The Decision Tree of 2nd Experiment.

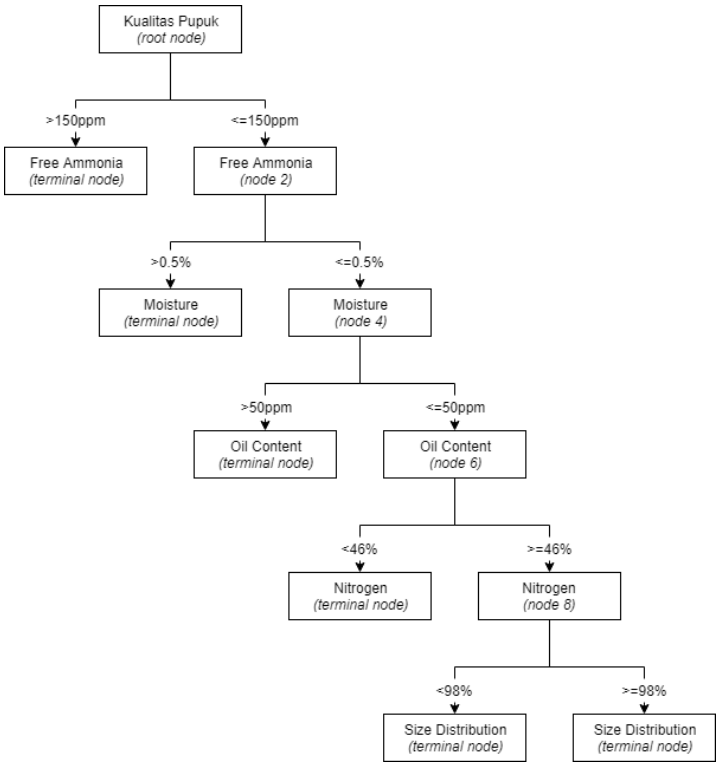

Fig. 2. The Decision Tree of 3rd Experiment.

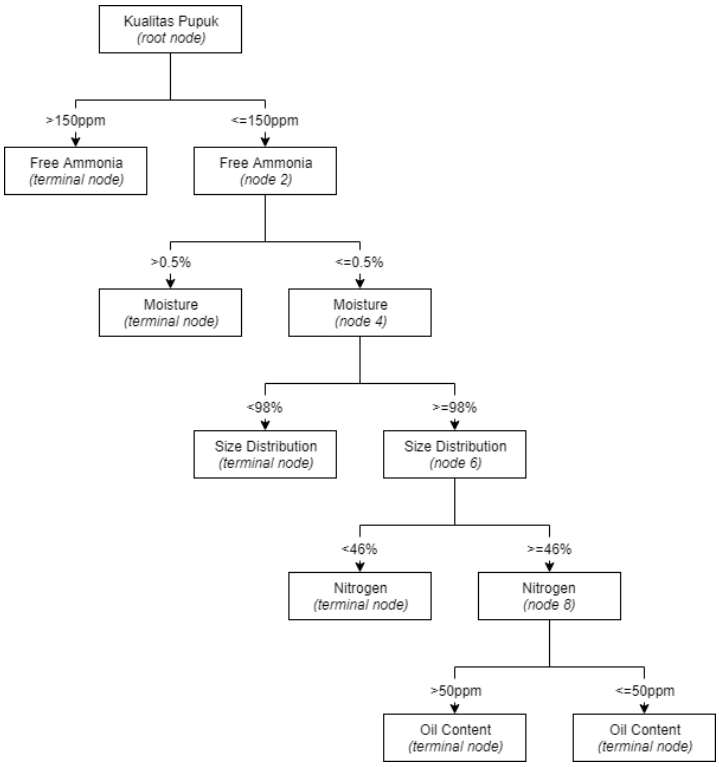

Fig. 3. The Decision Tree of 9th Experiment.

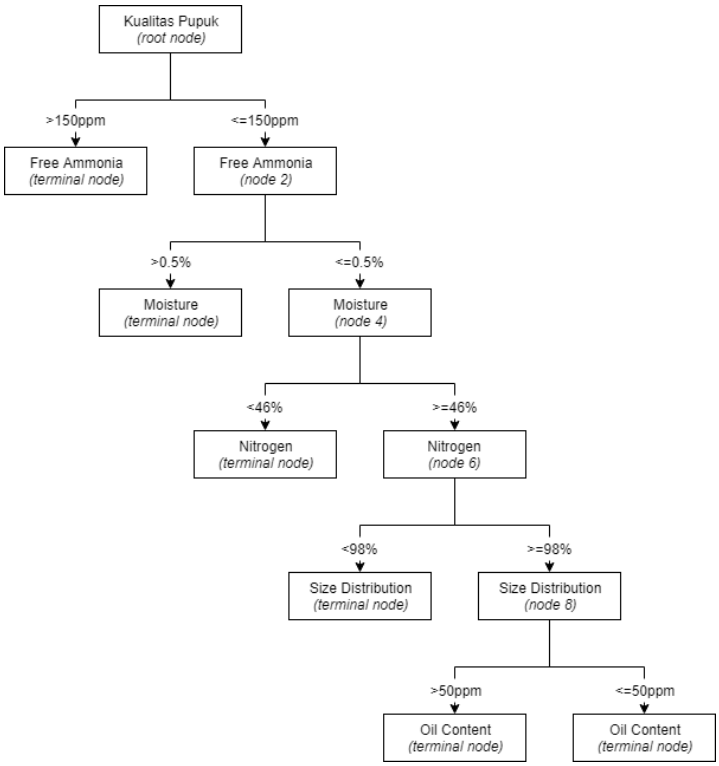

Fig. 4. The Decision Tree of 10th Experiment.

1. Conclusion

The role of training data is very influential in calculating the CHAID algorithm in forming a decision tree, because it also affects the rules that are built up. The used of training data will be processed with the CHAID algorithm to produce a decision tree in a knowledge-base. In this study, the knowledge-base was used to test the quality of urea fertilizer. That quality testing is based on determining the most significant independent variable on the dependent variable, then the variable is separated and carried out repeatedly until there are no more significant independent variables, so a decision tree is formed.

Software development by implementing CHAID algorithm produce 4 decision trees with very high accuracy, which is 99%. The accuracy value is influenced by the training data which will affect the rules established.

For further research, use more data so the result of decision tree can be optimized. Also try a combination of training data and randomized testing data to get a more varied decision tree.

References

[1] Syarafina, "Sistem Cerdas Pendeteksi Penyakit Pernapasan dengan Algoritma Classification and Regressiong Tree," in thesis. Palembang: Sriwijaya University, 2018.
[2] E. A. Hasibuan, A. N. Harahap, "Aplikasi Metode CHAID dalam Menganalisis Kecenderungan Penelitian Skripsi Mahasiswa pada Program Studi Pendidikan Matematika," in Edumatika, vol. 1, no. 2. Padang Sidempuan: Graha Nusantara University, 2018, pp.63-72.
[3] H. Mustafidah, H. Prawijaya, D. Aryanto, "Expert System for Diagnosing Computer Malfunction and Giving Advice to Repair It," in Juita, vol. I, no. 3. Purwokerto: Muhammadiyah University of Purwokerto, May 2011, pp. 71-76.
[4] H. Permana, "Klasifikasi dengan Metode CHAID dan Penerapannya pada Klasifikasi Alumni FMIPA UNY," in thesis. Yogyakarta: UNY, 2011.
[5] R. R. Nazar, "Penerapan Metode CHAID dan CART pada Klasifikasi Preeklampsia," in thesis. Yogyakarta: UII, 2018.
[6] Pusri, "Manual and Explanation Book of Factory Laboratories IV," unpublished