Classification of Salt Quality based on Salt-Forming Composition using Random Forest

Classification of Salt Quality based on Salt-Forming Composition using Random Forest

E M S Rochman[1]*, A Rachmad[2], D A Fatah[3], W Setiawan[4], Y Kustiyahningsih[5]

[1,2,3]Departemen of Informatics, Faculty of Engineering, University of Trunojoyo, Madura, Bangkalan, Indonesia

Corresponding author's e-mail: ekamalasari3@gmail.com

Abstract. Salt is part of the chemical that can be used and needed by humans in the field of consumption or industry. The formation of salt can be done in several ways, namely with seawater or lake water that is evaporated to produce salt crystals or through the process of mining rock salt. The results of the salt obtained will have a different composition depending on the process of formation, the difference in composition can affect the quality of the salt produced, so not all salt results are suitable for consumption. Generally, the salt quality classification process is still done manually, but this method takes a long time and is less effective. So, to overcome this problem, this research utilizes data mining science in classifying salt quality automatically using the Machine Learning algorithm, namely Random Forest. The data used in this study is a salt dataset with 7 attributes and 4 target classes totaling 349 data where the data is divided into training data and test data using k-fold cross validation with different k-fold values, namely 5-fold, 10-fold, and 20-fold. folds. The test results obtained indicate that the value of k = 10 has the best performance by achieving an AUC value of 96.1%, then for the classification accuracy is 87.7%, f1 score is 87.6%, precision is 87.7% and recall is 87.7%.

1. Introduction

Salt is part of the chemical that can be used and needed by humans in the field of consumption or industry. Sodium chloride (NaCl) is the main constituent in salt, besides that there are other constituents such as magnesium, calcium sulfate, iodine, and so on [1] [2]. The salt formation can be done in several ways, one of which is by evaporating seawater to produce salt crystals [3]. Therefore, Indonesia, which is a country with a very long coastline, can meet and produce national salt needs independently [4]. In addition to evaporating seawater, salt can also be obtained from lake water or the process of mining rock salt. The resulting salt will have a different composition based on the location of the salt formation or the type of water used.

The difference in composition can affect the quality of the salt produced so not all salt products are suitable for consumption. In categorizing the quality of the salt, it is still done manually by reviewing the composition of the salt, such as the content of NaCl, calcium, magnesium, sulfate, solubility levels, and so on [5]. Classification of salt quality manually makes the process take longer and does not escape classification errors due to several factors, one of which is human error.

To overcome these problems, this research utilizes data mining science to classify salt quality classes automatically using the system. Data mining itself is a process of extracting important information on data through the process of collecting and processing data by applying mathematical techniques, artificial intelligence, statistics, and machine learning [6] [7] [8]. There are many machine learning methods used to classify and analyze diseases, one of which is Random Forest. Random forest is one of the supervised learning algorithms used for data classification based on samples and training data attributes with the ensemble method [9] [10]. So based on this explanation, this study applies the Random Forest method to classify the quality of salt.

2. Methodology

Data mining is an extraction of important information on data through the process of collecting and processing data by applying mathematical techniques, artificial intelligence, statistics, and machine learning [6] [7] [8]. Data mining is often called Knowledge Discovery in Database (KDD) [11] [12], here are some stages in the data mining process shown in Figure.

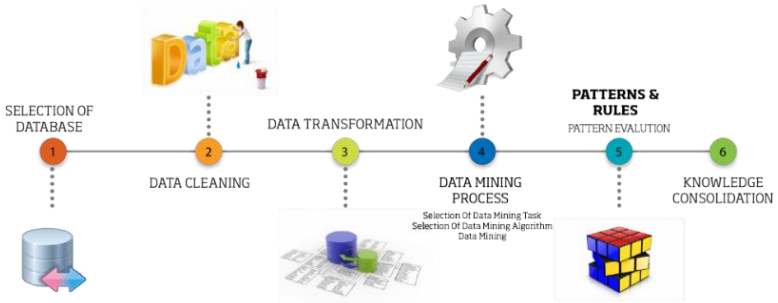

Figure 1. Data Mining Stages

Based on Figure 1, the stages of the data mining process include data selection, data cleaning, data transformation, data mining, pattern evaluation, and knowledge presentation. The following is a description of each of these stages [12] [13].

1. Data selection is the stage of selecting data to be used as basic analysis.
2. Data cleaning is the stage of removing data that is considered irrelevant and consistent.
3. Data transformation is the stage of converting data into a certain format before the data mining process.
4. Data mining is the stage of finding information using data mining methods.
5. Pattern evaluation is the stage of identifying the pattern of data results obtained from certain methods and evaluating the suitability of the results with the hypothesis.
6. Knowledge presentation is the stage of displaying or visualizing the final result according to the chosen form.

2.1. Data Pre-processing

The application of this process is the first step before the model learning stage is carried out whose purpose is to make the data structured to increase accuracy and simplify the classification process [14] [15]. At this stage, the data to be classified first goes through a data transformation process to change the data to suit the needs.

A. Data Transformation

Data transformation is a technique for changing the scale of data into other forms which aim to equalize the distribution of data according to training needs [16]. Data with variables that have different ranges of values will be normalized to reduce the imbalance of influence between data that have a larger range. Data normalization is a technique of data transformation to change several variables so that they have the same value range, which is between 0 to 1, no data is too large or too small so it will be easier to analyze [17] [18]. The method used to normalize the salt data is the Min-Max Normalization method. The equation below is a formula for the min-max normalization method.

$$z = \frac{x - \min(x)}{[\max(x) - \min(x)]} \quad (1)$$

Information:

z: normalization result,
x: value x (original),
min(x): the minimum value for the variable x,
max(x): the maximum value for the variable x.

2.2. Data Mining Process

In this study, before entering the classification stage, the distribution of training data and test data was first. The process of data sharing is to divide the dataset into k as many as n partitions. This data sharing is known as K-Fold Cross Validation which is a popular statistical method for partitioning data, where the data is divided into two subsets, namely training data for the learning process and test data

for validation or evaluation, which is used to evaluate the performance of a model or method or algorithm [19] [20]. K-Fold Cross Validation can be selected based on the size of the data set [21].

Usually, K-Fold is used to reduce the computation time and also to maintain the accuracy of the estimate. The size of the data divided depends on the specified K value, in this study a k-fold value of 10 is used. In each iteration Cross Validation randomly partitions the original data set given into training sets used to train machine learning algorithms and tests are set to evaluate its performance [22]. The following is an illustration of the process of sharing training data and test data using 10-fold.

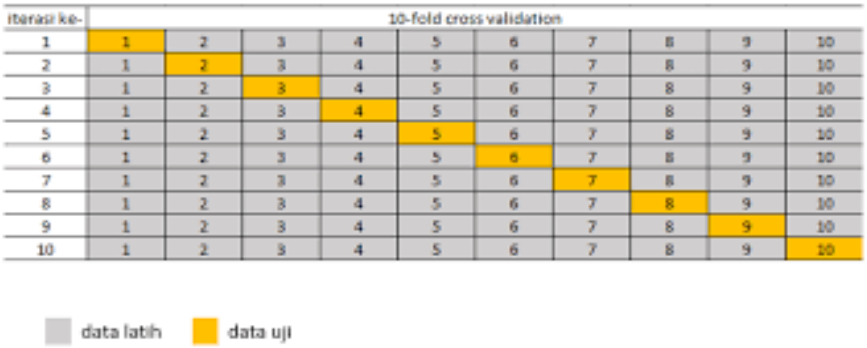

Figure 2. Illustration of 10-Fold

Figure 2 above shows that the dataset will be divided into 10 partitions, wherein in the first iteration the 1st partition is used as test data and the 2nd to 10th partitions as training data. So, the workings of the k-fold cross validation method can be described as follows:

1. The data set will be divided into several partitions, where the number of partitions is represented by the specified value of k.
2. The distribution of data in each iteration will make 1 partition as test data and the rest of the partition as training data.
3. The data sharing process will continue to iterate until it reaches k-fold.

In the data mining process, the method used to build a classification model to be applied to the salt quality detection system is Random Forest

A. Random Forest

Random Forest is a supervised learning algorithm for data classification based on samples and attributes of training data. Where this random forest is one of the algorithms that use the ensemble technique (applying the bagging method and random feature selection) [9]. Ensemble learning is used to improve the unstable classification problem by combining some basic learning to reduce prediction errors.

Random Forest is a method that makes modeling using several decision trees or in other words, a collection of several decision trees [9] and [23]. Where in each tree there is an estimated classification that we can consider (called a vote) to combine each possibility from each tree, then choose the classification with the greatest number of classifications to produce optimal and stable predictions [24]. In the random forest algorithm, there are m random subsets of p attributes (variables/features), where to find out how many optimal trees we can use, we can use the formula = to find out the value of m.

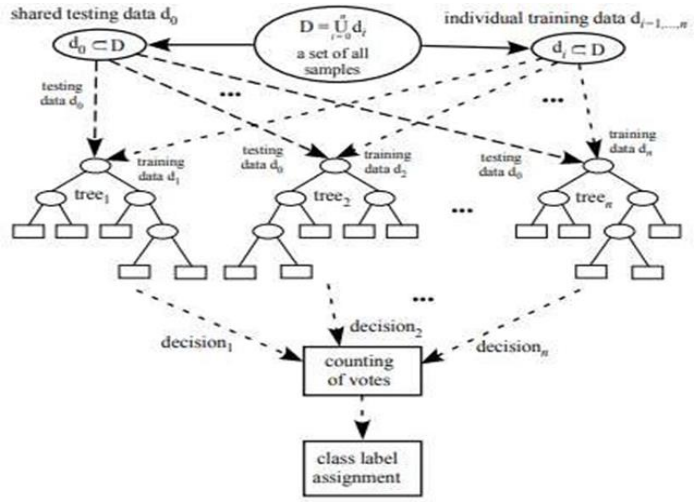

Figure 3. Random Forest Classification

In the random forest method, there are several processes, as follows:

1. Bootstrapping Stage
   This stage is the stage to create a subset by taking a random sample with a return size of n in the data cluster [24].
2. Random Feature Selection Stage
   Stages of building a tree until it reaches its maximum size. Where the selection of the disaggregation on the predictor variable m is chosen randomly, with m << p, after that the best sorter will be selected based on the predictor m [24]. For the response to a prediction observation, it is done by combining the results of predictions of as many as k trees based on the majority vote (majority vote).
   For the process of determining the majority vote on a tree, the first step is to calculate the entropy value using equations (2) and (3) in determining the level of impurity of a feature and information gain using equation (4) to choose the best feature/feature selection. Here's a formula for calculating the entropy value for an attribute:

$$entropy(S) = \sum_{i=1}^{c} p_i log_2 p_i \quad (2)$$

Information:
S: a set of datasets
c: the number of classes
pi: the probability of class i frequency in the dataset

Here's a formula for calculating the entropy value using two attributes:

$$entropy(T, X) = \sum P(c)E(c) \quad (3)$$

Information:
(T, X): features T and X
P(c): the probability of feature class
E(c): entropy result of a feature class

For the feature selection process using the formula of gain:

$$gain(A) = entropy(S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} x\ entropy(S_i) \quad (4)$$

Information:
S: a set of dataset
A: features
|Si|: number of samples values i
S: sum of all data

entropy(Si) : sample entropy value i

2.3. Evaluation

Model evaluation in research is needed to find out how well the performance of a classification model is confusion matrix is a method that can be applied in measuring the accuracy or performance of the classification model. The confusion matrix displays the performance results of the classification model in the form of a matrix [25]. Four possible outputs are resulting from the comparison between the predicted classification and the actual classification, namely false positive (FP) and false negative (FN), true positive (TP), and true negative (TN). The output can be seen in the following table.

Based on the table above, it can be concluded that if the results of the confusion matrix are in the true positive (TP) column, the results are correct and identified as positive. If the result is in the false positive (FP) column, then the result is false and identified as positive. If the result is in the false negative column (FN) then the result is wrong and identified as negative and if the result is in the false negative column (TN) then the result is correct and identified as negative.

From the results of the confusion matrix, accuracy, recall, precision, and f-measure values can be generated. Accuracy is the degree of closeness between the actual results and the predicted results. The recall is the success rate of the system in retrieving information. Precision is the level of conformity of the answers generated from the system with the information desired by the user. F-measure is the average result of the combination of recall and precision calculations [26]. In this study, the value of AUC (area under the curve) will also be calculated or it can be called probability which is a method for calculating under the ROC curve, where the higher the AUC value, the classification method used can be applied properly in a study [27].

1. Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

2. Precision

$$Precision = \frac{TP}{TP + FP} x\ 100\% \quad (6)$$

3. Recall

$$Recall \quad = \frac{TN}{TN + FN} x\ 100\% \quad (7)$$

4. F-measure

$$F - measure \quad = 2\ x\ \frac{Recall\ x\ Precision}{Recall + Precision} x\ 100\% \quad (8)$$

5. AUC

$$AUC \quad = \frac{1}{2} x\ (\frac{TP}{TP + FN} + \frac{TN}{TN + FP}) \quad (9)$$

Information:

TP: the number of data correctly predicted positive,

TN: the number of data with the original class is positive but the prediction result is negative,

FN: the number of data correctly predicted negative,

FP: the number of data with the original class is negative but the prediction result is positive.

2.4. Play Results

A. Data Collection

The data used to complete this research is the table salt quality dataset which in this dataset has 7 attributes (water content, not dissolved, calcium, magnesium, sulfate, NaCl(wb), and NaCl(db)) and the target class of this research is 4, namely the quality of K1, K2, K3, and K4 salts. The dataset used in this study can be seen in the Table.

Table 1. Salt Dataset

Data	*Water Content*	*Not Dissolved*	*Calcium*	*Magnesium*	*Sulfate*	*NaCl(wb)*	*NaCl(db)*	*Grade*
1	7,82779	0,0423309	0,24028	0,80734	1,34723	83,9156	91,0422	K1
2	8,10811	0,590494	0,216902	0,954371	1,18318	85,012	92,513	K1
3	8,78378	0,0293832	0,20077	0,950314	1,11296	84,3214	92,4413	K1
4	5,87121	0,211922	0,541328	0,903783	1,11528	88,182	93,6823	K2
5	4,66418	0,175555	0,418727	0,530387	1,47868	89,2901	93,6585	K2
6	8,49057	0,181191	0,242135	0,813574	1,27676	85,0348	92,9246	K2
7	8,74525	0,202144	0,311962	0,481313	1,22794	84,4195	92,5097	K2
8	11,3674	0,198742	0,340359	0,884933	1,02472	80,2015	90,4875	K3
9	7,69231	0,654811	0,431922	0,777459	1,4162	84,2949	91,3195	K3
10	8,03059	0,296823	0,221097	0,411241	0,933233	88,264	95,9711	K4

B. Analysis

In this section, it will be explained how the research was carried out using the proposed method, as a solution to the problems raised. The stages of the process will be visualized in the Input-Process-Output diagram below.

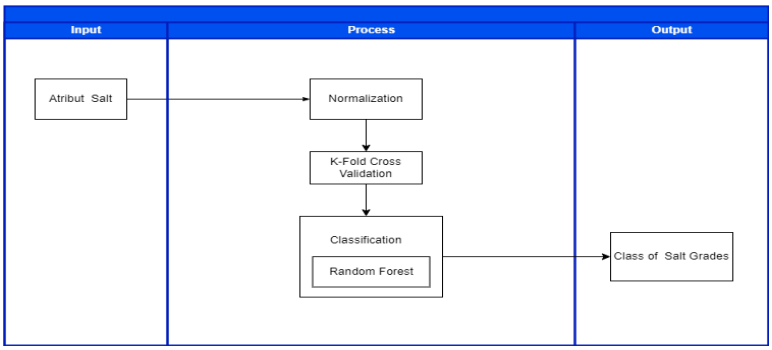

Figure 4. IPO Diagram

Based on Figure 4 above, the process of making a classification system is divided into several parts, namely, the input process, the core process, and the output process. The description of the three parts of the stages is as follows.

1. Data Input Process
   The input section is a process for entering data to be classified in this study, where the input data used is a salt dataset with 7 features and 4 classes or targets.
2. Data Pre-processing
   At this stage, a data transformation technique is used to normalize the data with variables that have a range of different values to reduce the imbalance of influence between data that has a larger range. The method used to normalize the salt data is the Min-Max Normalization method so that all data have values in the range 0 to 1.
3. Data Sharing Process
   At this stage, the training data and test data are divided using the k-fold cross-validation method using k-fold values of 5, 10, and 20. The training data is used to build a model that is formed with a certain amount of data and for test data taken from the remaining data that is not used in the training process that is used to test the performance of the model that has been trained.
4. Classification Process

At the classification stage, a learning process is carried out to build a system model using the Random Forest algorithm method.

5. Output
   The output produced after the entire process is run is in the form of class predictions of salt attributes based on modeling with the methods applied in this study.

3. Result and Analysis

In this study, 349 salt data were divided into training data and test data using k-fold cross validation with different k-fold values, namely 5-fold, 10-fold, and 20-fold until the results of the evaluation of the classification model were obtained using the Random Forest method with a number value of the tree is 10. The results of the method performance evaluation can be seen in Table 2 below.

Table 2. Evaluation Results Using the Random Forest Method

Fold	AUC	Classification Accuracy	F1 Score	Precision	Recall
5	95.6 %	86.0 %	86.0 %	86.0 %	86.0 %
10	96.1 %	87.7 %	87.6 %	87.7 %	87.7 %
20	95.6 %	86.0 %	85.9 %	86.1 %	86.0 %

Table 2 shows the test results from each of the different k-fold values that have differences in accuracy values that are not too significant. Of the three k-fold values, it can be seen that the value of k = 10 has the best performance by achieving an AUC value of 96.1%, then for classification accuracy is 87.7%, f1 score is 87.6%, precision is 87.7% and recall is 87.7%.

4. Conclusion

The conclusion that can be drawn based on the research conducted is the classification using the Random Forest method provides a good accuracy value in classifying the salt dataset, which amounts to 349 data records with 7 attributes into 4 target classes, namely the best accuracy is obtained when using the fold value = 10 with an AUC value of 96.1 %, classification accuracy is 87.7%, f1 score is 87.6%, precision is 87.7% and recall is 87.7%. Thus, the application of the Random Forest method is very capable of classifying data well.

5. Acknowledgment

Thank you to the Institute for Research and Community Service and the Faculty of Engineering, the University of Trunojoyo Madura for allowing researchers to complete this research. This research article is supported by UTM DIPA Fund No. 163/UN46.4.1/ PT.01.03/2022 for Penelitian Mandiri - UTM.

References

[1] Y. Kustiyahningsih, E. Rahmanita, A. Rachmad, & J. Purnama, "Integration interval type-2 FAHP-FTOPSIS group decision-making problems for salt farmer recommendation." *Commun. Math. Biol. Neurosci.* 2021 (2021).

[2] A. Khozaimi, Y. D. Pramudita, E. M. S. Rochman, and A. Rachmad, "Decision Support System for Determining the Quality of Salt in Sumenep Madura-Indonesia." *Journal of Physics: Conference Series*. Vol. 1477. No. 5. IOP Publishing, 2020.

[3] A. Khozaimi, Y. D. Pramudita, E. M. S. Rochman, and A. Rachmad, "Decision Support System for Determining the Quality of Salt in Sumenep Madura-Indonesia." *Journal of Physics: Conference Series*. Vol. 1477. No. 5. IOP Publishing, 2020.

[4] Rusdi, "The Effect Factors Of Supply Salt In Indonesia," Scientific Journal of Reflection, vol. 1, pp. 141-150, 2018.

[5] E. Tarmizi, K. Sunandar and A. D. K. Wibowo, "Thermodynamic evidence of giant salt deposit formation by serpentinization: an alternative mechanism to solar evaporation," Nature Research, 2019.

[6] R. R. Asaad and R. M. Abdulhakim, "The Concept of Data Mining and Knowledge Extraction Techniques," Qubahan Academic Journal, pp. 17-20, 2021.

[7] J. F. P. d. Costa and M. Cabral, "Statistical Methods with Applications in Data Mining: AReview of the Most Recent Works," Mathematics, vol. 10, no. 6, pp. 1-22, 2022.

[8] V. Marriboyina and L. C. Reddy, "A Review on Data mining from Past to the Future," International Journal of Computer Applications, vol. 15, pp. 19-22, 2011.
[9] V. Y. Kulkarni and D. P. K. Sinha, "Random Forest Classifier: A Survey and Future Research Directions," International Journal of Advanced Computing, vol. 36, pp. 1144-1153, 2013.
[10] K. Fawagreh, M. M. Gaber and E. Elyan, "Random forests: from early developments to recent advancement," Systems Science & Control Engineering, vol. 2, pp. 602-609, 2014.
[11] P. A. Widya and M. Sudarma, "Implementation of EM Algorithm in Data Mining for Clustering Female Cooperative," International Journal of Engineering and Emerging Technology, vol. III, pp. 75-79, 2018.
[12] Shivali, J. Birla and Gurpreet, "Knowledge Discovery in Data-Mining," International Journal of Engineering Research & Technology (IJERT), vol. 3, no. 10, pp. 1-5, 2015.
[13] M. A. Jassim and S. N. Abdulwahid, "Data Mining preparation: Process, Techniques and Major Issues in Data Analysis," in IOP Publishing, 2020.
[14] V. Agarwal, "Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis," International Journal of Computer Applications, vol. 131, pp. 30-36, 2015.
[15] K. Maharana, S. Mondal and B. Nemade, "A review: Data pre-processing and data augmentation techniques," Global Transitions Proceedings, vol. 3, pp. 91-99, 2022.
[16] D. K. Lee, "Data transformation: a focus on the interpretation," Korean Journal of Anesthesiology, vol. 73, no. 6, pp. 503-508, 2020.
[17] G. Aksu, C. O. Guzeller and M. T. Eser, "The Effect of the Normalization Method Used in Different Sample Sizes on the Success of Artificial Neural Network Model," International Journal of Assessment Tools in Education, vol. 6, pp. 170-192, 2019.
[18] Hendri, T. Wahyuningsih and E. Rahmawanto, "Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor(kNN) Algorithm to Test the Accuracy of Types of Breast Cancer," International Journal of Informatics and Information System, vol. 4, pp. 13-20, 2021.
[19] S. Hulu, P. Sihombing and Sutarman, "Analysis of Performance Cross Validation Method and K-Nearest Neighbor in Classification Data," International Journal of Research and Review, vol. 7, no. 4, pp. 69-73, 2020.
[20] G. Buron and U. Stanczyk, "Standard vs. non-standard cross-validation: evaluation of performance in a space with structured distribution of datapoints," Procesia Computer Science, pp. 1245-1254, 2021.
[21] D. Normawati and D. P. Ismi, "K-Fold Cross Validation for Selection of Cardiovascular Disease Diagnosis Features by Applying Rule-Based Data mining," Signa and Image Pocessing Letters, vol. 1, pp. 22-32, 2019.
[22] N. Darapureddy, N. Karatapu and T. K. Battula, "Research of Machine Learning Algorithms using K-Fold Cross Validation," International Journal of Engineering and Advanced Technology (IJEAT), vol. 6, no. 6S, pp. 215-218, 2019.
[23] J. Ali, R. Khan, N. Ahmad and I. Masqood, "Implementing Random Forest Algorithm in GEE: Separation and Transferability on Built-Up Area in Central Java, Indonesia," International Journal On Informatics Visualization, vol. 6, pp. 74-82, 2022.
[24] G. S. Saragih, Z. Rustam, D. Aldila, R. Hidayat, R. E. Yunus and J. Pandelaki, "Ischemic Stroke Classification using Random Forests Based on Feature Extraction of Convolutional Neural Networks," International Journal on Advanced Science, Engineering and Information Technology, vol. 10, pp. 2177-2182, 2020.
[25] R.-C. Chen, C. Dewi, S.-W. Huang and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," Journal of Big Data, pp. 1-26, 2020.
[26] A. Rachmad, N. Chamidah, and R Rulaningtyas. "Mycobacterium tuberculosis images classification based on combining of convolutional neural network and support vector machine." Commun. Math. Biol. Neurosci. 2020 (2020)
[27] A. Rachmad, N. Chamidah, and R Rulaningtyas. "Classification of mycobacterium tuberculosis based on color feature extraction using adaptive boosting method." *AIP Conference Proceedings*. Vol. 2329. No. 1. AIP Publishing LLC, 2021.