A CLASSIFICATION MODEL FOR WATER QUALITY ANALYSIS USING DECISION TREE

Consolata Gakii and Jeniffer Jepkoech

Department of Mathematics, Computer science and Information Technology (MCIT).

University of Embu P.O Box 06 Embu

ABSTRACT: *A classification algorithm is used to assign predefined classes to test instances for evaluation) or future instances to an application). This study presents a Classification model using decision tree for the purpose of analyzing water quality data from different counties in Kenya. The water quality is very important in ensuring citizens get to drink clean water. Application of decision tree as a data mining method to predict clean water based on the water quality parameters can ease the work of the laboratory technologist by predicting which water samples should proceed to the next step of analysis. The secondary data from Kenya Water institute was used for creation of this model. The data model was implemented in WEKA software. Classification using decision tree was applied to classify /predict the clean and not clean water. The analysis of water Alkalinity,pH level and conductivity can play a major role in assessing water quality. Five decision tree classifiers which are J48, LMT, Random forest, Hoeffding tree and Decision Stump were used to build the model and the accuracy compared. J48 decision tree had the highest accuracy of 94% with Decision Stump having the lowest accuracy of 83%.*

KEYWORDS: Data Mining, classification model, Decision tree, Weka Tool, water quality

INTRODUCTION

Supervised learning is a machine learning algorithm which receives feature vector and the target pattern as an input to build a model. The model can be used to recognise new patterns and assign a target to them. Applications of supervised learning include classification (e.g. classifying players according to their behaviour during a game) and regression (e.g. Predicting household prices according to features) Unsupervised learning is a machine learning algorithm which only receives the feature vector as an input, and its task is to find similar groups of items with comparable features. The essential application of unsupervised learning is clustering, such as determining the distribution of data items within a multidimensional space (Biggio & Roli, 2018).

Classification is an instance of supervised learning that includes a training phase to create a model (Classifier). Its task is to predict the class of items in a data set using a certain model of a classifier. The model is constructed using already-labelled items of similar data sets. This step allows classification techniques to be considered as a supervised machine learning method. Data Mining is the process of finding patterns in a large scale of data which are interesting, new, useful and meaningful (Zaki et al., 2014). Data mining can be considered as an interdisciplinary field of study consisting of areas such as databases, statistics, machine learning and artificial intelligence Classifier has been widely applied in machine learning, such as pattern recognition, medical diagnosis, credit scoring, banking and weather prediction. Different classifier models are created by using different classification algorithms, which can be divided into four main categories: Decision Tree Classifier, Probabilistic Classification, Support Vector Machines and Linear Discriminant Analysis (Zaki et al., 2014). These classifiers are discussed in the following subsections, with consideration of Decision Tree Classifiers which are used for experiment in this research.

LITERATURE REVIEW

Classification is an important problem in machine learning and data mining. It has been widely applied in many real-world applications. To build a classifier, a user first needs to collect a set of training examples/instances that are labelled with predefined classes. A classification algorithm is then applied to the training data to build a classifier that is subsequently employed to assign the predefined classes to test instances (for evaluation) or future instances (for application).

Muharemi et al., 2018 proposed the Nearest Neighbor Algorithm (KNN) and the Neural Network of Classification based on Logistic Regression to obtain an adequate solution to address the problem of changes in the quality of drinking water. Haghiabi et al., (2018) investigated the performance of artificial intelligence techniques that include the artificial neural network (ANN), the group data management method (GMDH) and the support vector machine (SVM) to predict the components of the water quality of the Tireh River located in southwestern Iran. During the development process of ANN and SVM, it was found that tansig and RBF as transfer and core functions have the best performance among the tested functions (Haghiabi et al., 2018) Chou et al., (2018) conducted a study to determine the water quality in the reservoir using data collected over ten years in Taiwan. Four well-known artificial intelligence techniques, artificial neural networks (ANN), support vector machines, classification and regression trees, and linear regression were used to analyze reference scenarios and sets. Then, an easy-to-use interface was developed that integrates a metaheuristic regression model to evaluate predictive performance and compare it with those of the two constituent scenarios. The ANN model was more accurate than the other unique models, sets and meta heuristic regression hybrids (Chou et al., 2018) Zhang et al., (2017) proposed a new anomaly detection algorithm for water quality data using double-movement windows over time, which can identify historical pattern anomaly data in real time. The algorithm is based on statistical models, autoregressive linear combination model. The algorithm has been tested using water quality PH data at 3 months from a real water quality monitoring station on a river system. The experimental results show that their algorithms can significantly decrease the false positive rate and have a better anomaly detection performance than the AD and ADAM algorithms.

Mohammad pour et al. (2015) investigated the problem with water quality, using three different algorithms, SVM and two methods of artificial neural networks. The performance is compared using R2, RMSE, MAE. On the results they achieved, the SVM algorithm is competitive with neural networks. This work led us to remodel SVM and ANN of Muharemi et al. (2018a).As it promises to give better results. The best result was achieved using the artificial neural network with non-linear autoregressive. In 2009, Xiang & Jiang applied the least squares support vector machine (LS-SVM) with particle swarm optimization methods to predict water quality and overcome the weaknesses of the usual back propagation algorithms by being moderate to meet and simple to reach the extreme minimum value. They discovered that through the simulation tests, the model shows a great capacity to estimate the water quality of the Liuxi River Xiang and Jiang (2009). The recurrent neural network as a dynamic system, whose next state and output depend on the current status and input of the network, is recently applied to large-scale vision speech problems (Gregor et al., 2015). The RNNs and the LSTMs are quite good at extracting patterns in the input feature space, where the input data covers sequences that are too long. They can model problems almost perfectly with multiple input variables, which provides a great benefit in forecasting time series, where classical linear methods can be difficult to adapt to multiple or multiple input forecast problems (Che et al. al., 2018).Mohammad et al. (2015) investigated the problem with water quality, using three different algorithms, SVM and two methods of artificial neural networks. The performance is compared using R2, RMSE, MAE. On the results they achieved, the SVM algorithm is competitive with neural networks.

METHODOLOGY

Data analysis tool used in this study

The data analysis tool used in this study is WEKA which is a data mining software developed by the University of Waikato in New Zealand that apparatus data mining algorithms using the JAVA language. Weka is a milestone in the history of the data mining and machine learning research communities, because it is the only toolkit that has gained such widespread adoption. Weka is a bird name of New Zealand. WEKA is a modern feature for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The WEKA project aims to provide a comprehensive collection of machine learning algorithms and data pre-processing tools to researchers. WEKA implements algorithms for data pre-processing, classification, regression, clustering and association rules; It also includes visualization tools. WEKA would not only afford a toolbox of learning algorithms, but also a framework inside which researchers could implement new algorithms without having to be concerned with supporting infrastructure for data manipulation and scheme evaluation. WEKA is open source software issued under General Public License. The data file normally used by Weka is in ARFF file format, which consists of special tags to indicate different things in the data file foremost: attribute names, attribute types, and attribute values and the data.Various decision tree algorithms are used in classification like ID3, AD Tree, REP, J48, FT Tree, LAD Tree, decision stamp, LMT, random forest, random tree etc. In this work the decision trees which were considered are ***J48, LMT, Random Forest and Random tree*** for comparison purposes and all those decision tree algorithms are found in WEKA.

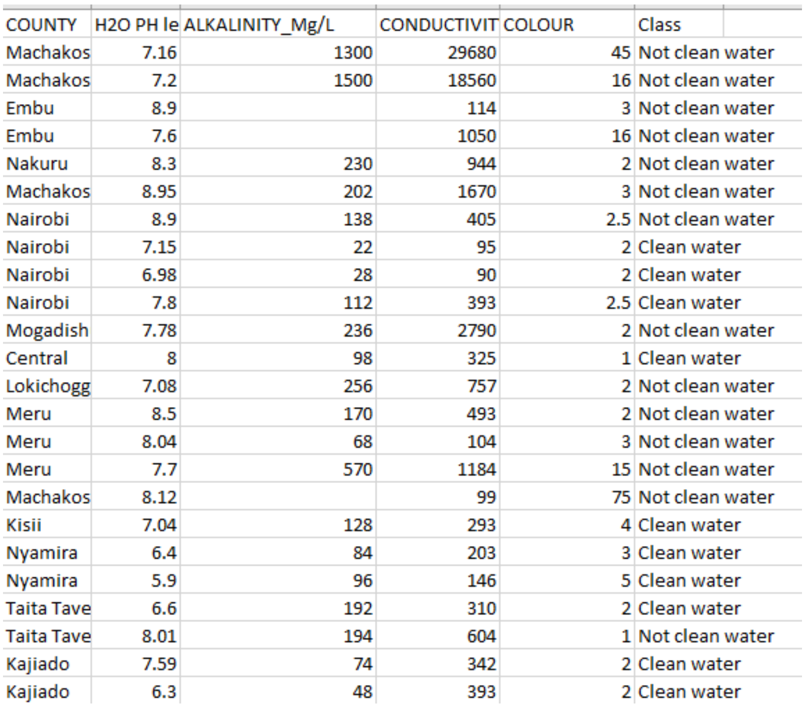
Datasets

In this study secondary data was used to build the decision tree and this data was downloaded from http://www.opendata.go.ke/datasets/kewi-water-test-and-results. The data is from the 47 counties in Kenya with water samples analysed.

Water quality parameters which is used for creating Model

Many parameters can influence the surface water quality. In this study four parameters are selected for the investigation. The parameters which were used to determine whether the water is clean is PH level, Alkalinity, conductivity and colour. To label the data the Kenya Bureau of standards (KEBS) and World Health Organisation (WHO) guidelines were used which are as follows:

PARAMETER UNIT	WHO and KEBS STANDARDS
PH level	6.5-8.5
Alkalinity	Max 500
Conductivity	Max 2500
Colour	Max 15



COUNTY	H2O PH le	ALKALINITY_Mg/L	CONDUCTIVIT	COLOUR	Class
Machakos	7.16	1300	29680	45	Not clean water
Machakos	7.2	1500	18560	16	Not clean water
Embu	8.9		114	3	Not clean water
Embu	7.6		1050	16	Not clean water
Nakuru	8.3	230	944	2	Not clean water
Machakos	8.95	202	1670	3	Not clean water
Nairobi	8.9	138	405	2.5	Not clean water
Nairobi	7.15	22	95	2	Clean water
Nairobi	6.98	28	90	2	Clean water
Nairobi	7.8	112	393	2.5	Clean water
Mogadish	7.78	236	2790	2	Not clean water
Central	8	98	325	1	Clean water
Lokichogg	7.08	256	757	2	Not clean water
Meru	8.5	170	493	2	Not clean water
Meru	8.04	68	104	3	Not clean water
Meru	7.7	570	1184	15	Not clean water
Machakos	8.12		99	75	Not clean water
Kisii	7.04	128	293	4	Clean water
Nyamira	6.4	84	203	3	Clean water
Nyamira	5.9	96	146	5	Clean water
Taita Tave	6.6	192	310	2	Clean water
Taita Tave	8.01	194	604	1	Not clean water
Kajiado	7.59	74	342	2	Clean water
Kajiado	6.3	48	393	2	Clean water

A Section of model training data

Experimental Setup

In the Experiment 80% of datasets was used as a training set for developing a model and 20% was used as the test data. ***J48, LMT, Random Forest and*** **Hoeffding tree** decision tree classifiers were used in this study. These classifiers are part of Weka data mining tool.

J48- The J48 decision tree is a predictive self-learning model that determines the target value of a new sample based on various attribute values of the available data. The different attributes are designated by the internal nodes of a decision tree. The branches between the nodes indicate the possible values that these attributes may have in the experimental samples, while the endpoints indicate the final value of the dependent variable (Pham et al., 2017).

LMT - A classification model associated with a monitored training algorithm combining logistic prediction and decision tree learning is the Logistic Model Tree (LMT). Logistic model trees use a decision tree with linear regression models on the leaves to provide a linear regression model by section.

Random forest - The random forest is a method of learning sets for classification, regression and other tasks, where a large number of decision trees are created at the time of the decision and the class is output, which is the average classification or prediction mode of individual trees. Random forests rectify the habit of decision trees that go with their training set. Random forests are a method of calculating the mean of several deep decision trees formed in different parts of the same training set, with the aim of reducing the variance. This is to the detriment of a slight increase and a loss of interpretability, but generally increases significantly the presentation of the final model. (Pham et al., 2017).

Hoeffding Tree - A Hoeffding Tree (VFDT) is an incremental decision tree analysis algorithm at any time, capable of learning from massive data streams. It is assumed that the examples do not change over time. Hoeffding trees take advantage of the fact that a small sample is often enough to select an optimal split attribute. This idea is mathematically supported by the reinforcement limit, which quantifies the number of observations needed, to estimate some statistics with a prescribed precision (in our case, the quality of an attribute) (Adhikari et al. 2018).

RESULTS AND DISCUSSION

J48 classifier was tested with confidence factor ranging from 0.1 and the number of minimum instances per node (minNumObj) was held at 2, and cross validation folds for the Testing Set (crossValidationFolds) was held at 10 during confidence factor testing.

Results of the model

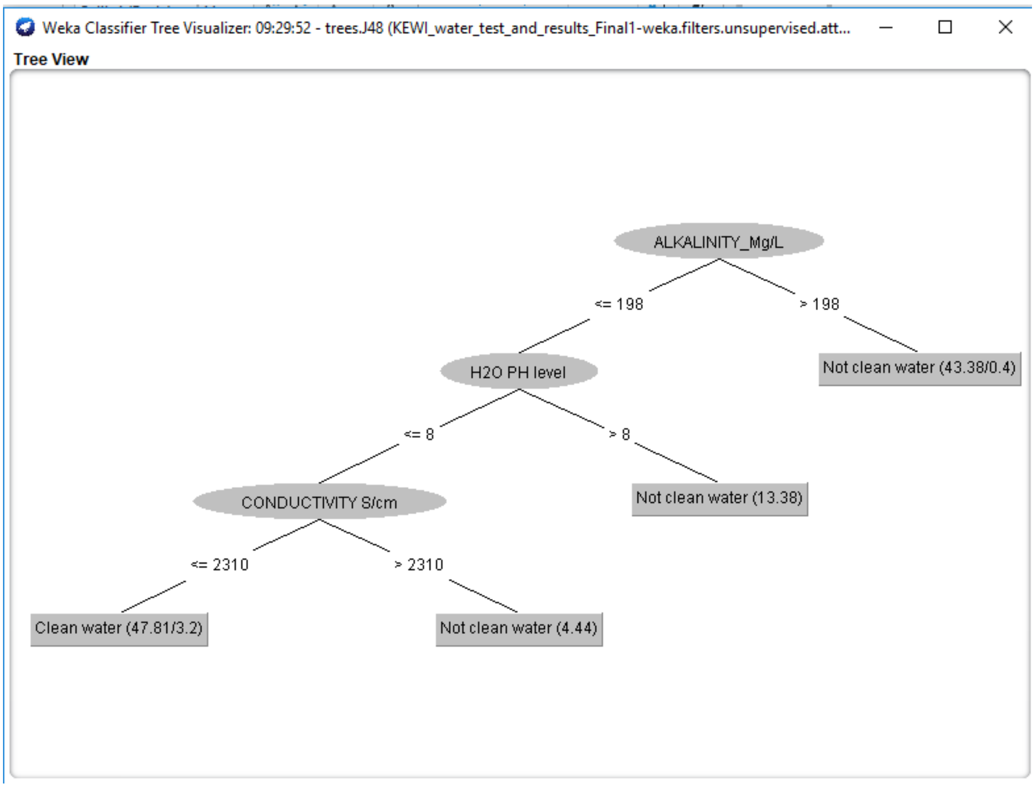


Figure 2.0 J48 decision tree for water quality analysis

Based on the classification model. If...Then rules can be extracted from weka generated tree. The rules of experiment are shown in figure 2 and are as follows:
Rule 1: IF Alkalinity>198 Then class Not clean (43.3%)
Rule2: IF PH_level >8 Then class Not clean (13.3%)
Rule 3: IF conductivity >2310 Then class Not clean (4.40%) else Clean water (47.8%)

Table1. Decision tree classifiers summary

Decision tree	Accuracy	Time taken to build the model
J48	93.6%	0 second
LMT	89.9%	0.06 seconds
Random forest	91.7%	0.05 seconds
Hoeffding tree	80.7%	0.03 seconds
DecisionStump	83.4%	0 second

Table 1 shows accuracy for different decision tree classifiers. The results show that Decision stump classification algorithm takes minimum time to classify data but gives less accuracy of 83%. J48 have quite good accuracy with model having accuracy of 94% and also with minimum time taken to build the model. LMT classifier gave a higher accuracy of 89.9% as compared with Decision stump but time taken to build classification model is much higher than other classifiers with 0.06 seconds.

Table 2: A confusion matrix

	A (Not clean water)	B (Clean water)
A (Not clean water)	TP	FN
B (Clean water)	FP	TN

TP (True Positive): It denotes the number of records classified as true while they were true.
FN (False Negative): It denotes the number of records classified as false while they were true.
FP (False Positive): It denotes the number of records classified as true while they were false.
TN (True Negative): It denotes the number of records classified as false while they were false.
Results obtained from the five classifiers are summarised in table 3 below

Table. 3: Summary of confusion matrix

Decision tree	Mean absolute error	a	b
J48	0.08	60	4
		3	42
LMT	0.13	59	5
		6	39
Random forest	0.1	60	4
		5	40
Hoeffding tree	0.1	45	19
		2	43
DecisionStump	0.23	48	16
		2	43

The matrix that shows the possible prediction results is called a confusion matrix. There are different evaluation criteria that can be obtained from these values. Accuracy is basically the ratio of correct predictions. However, accuracy has limitations in evaluating the prediction performance. Especially, accuracy does not show how the cases of minority class are classified, when the class distribution is imbalanced (Provost & Fawcett, 2013) in this analysis, the J48 decision tree model correctly predicted the positive class for Not clean water 60 times and incorrectly predicted it 4 times. The model correctly predicted the negative class for not clean water 42 times and incorrectly predicted it 3 times. The Hoeffding tree classifier had the highest incorrect predictions of 19 times for Not clean water and correctly predicted the positive class of Not clean water 45 times.

CONCLUSION

In this study, water quality model was implemented using decision tree technique. The analysis of water Alkalinity,pH level and conductivity can play a major role in assessing water quality. Five decision tree classifiers which are J48,LMT,Random forest, Hoeffding tree and DecisionStump were used to build the model and the accuracy compared. J48 decision tree had the highest accuracy of 94% with DecisionStump having the lowest accuracy of 83%.The decision tree classifier provided with quality water parameters set by WHO and KEBS can be used to predict whether drinking water is clean or not.

References

Ashraf, N., Ahmad, W., & Ashraf, R. (2018). A Comparative Study of Data Mining Algorithms for High Detection Rate in Intrusion Detection System. *Annals of Emerging Technologies in Computing (AETiC)*, *2*(1).

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, *84*, 317-331.

Elkano, M., Galar, M., Sanz, J., & Bustince, H. (2018). CHI-BD: a fuzzy rule-based classification system for big data classification problems. *Fuzzy Sets and Systems*, *348*, 75-101.

Li, P., Li, J., Huang, Z., Gao, C. Z., Chen, W. B., & Chen, K. (2018). Privacy-preserving outsourced classification in cloud computing. *Cluster Computing*, *21*(1), 277-286.

Manavalan, B., Shin, T. H., & Lee, G. (2018). PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine. *Frontiers in microbiology*, *9*, 476.

Nikam, S. S. (2015). A comparative study of classification techniques in data mining algorithms. *Oriental journal of computer science & technology*, *8*(1), 13-19.

Sahani, R., Rout, C., Badajena, J. C., Jena, A. K., & Das, H. (2018). Classification of intrusion detection using data mining techniques. In *Progress in Computing, Analytics and Networking* (pp. 753-764). Springer, Singapore.

Zaki, M. J., Meira Jr, W., & Meira, W. (2014). *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.

Chou, J. S., Ho, C. C., & Hoang, H. S. (2018). Determining quality of water in reservoir using machine learning. Ecological informatics, 44, 57-75.

Chandrasekaran, S., Freise, M., Stork, J., Rebolledo, M., & Bartz-Beielstein, T. (2017). Gecco 2017 industrial challenge: Monitoring of drinking water quality.

Che, Z., Purushotham, S., Cho, K., Sontag, D., & Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. Scientific Reports, 8(1), 6085.

García, S., Luengo, J., & Herrera, F. (2015). Data preprocessing in data mining. Springer.

Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., & Wierstra, D. (2015). Draw: A recurrent neural network for image generation. arXiv preprint arXiv:1502.04623.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Vol. 112. Springer.

Mohammadpour, R., Shaharuddin, S., Chang, C. K., Zakaria, N. A., Ab Ghani, A., & Chan, N. W. (2015). Prediction of water quality index in constructed wetlands using support vector machine. Environmental Science and Pollution Research, 22(8), 6208–6219.

Muharemi, F., Logofatu, D., Andersson, C., & Leon, F. (2018a). Approaches to building a detection model for water quality: A case study. Modern Approaches for Intelligent Information and Database Systems (p. 173–183). Springer.

Muharemi, F., Logofătu, D., & Leon, F. (2018b). Review on general techniques and packages for data imputation in r on a real-world dataset. Springer. Nguyen, M., & Logofătu, D. (2018). Applying tree ensemble to detect anomalies in real-world water composition dataset. International Conference on Intelligent Data Engineering and Automated Learning (pp. 429–438). Springer.

Haghiabi, A. H., Nasrolahi, A. H., & Parsaie, A. (2018). Water quality prediction using machine learning methods. Water Quality Research Journal, 53(1), 3-13.

Muharemi, F., Logofătu, D., Andersson, C., & Leon, F. (2018). Approaches to building a detection model for water quality: a case study. In Modern Approaches for Intelligent Information and Database Systems (pp. 173-183). Springer, Cham.