

Decision Tree Algorithm for Mining “If Then Else” Rule in Single Slope Basin Solar Still plant

Neha Yadav, Vivek Raich



Abstract: Soft computing dedicatedly works for decision making. In this domain a number of techniques are used for prediction, classification, categorization, optimization, and information extraction. Among rule mining is one of the essential methodologies. “If Then Else” can work as rules, to classify, or predict an event in real world. Basically, that is rule based learning concept, additionally it is frequently used in various data mining applications during decision making and machine learning. There are some supervised learning approaches are available which can be used for rule mining. In this context decision tree is a helpful algorithm. The algorithm works on data splitting strategy using entropy and information gain. The data information is mapped in a tree structure for developing “If Then Else” rules. In this work an application of rule based learning is presented for recycling of water in a distillation unit. By using the designed experimental still plant different attributes are collected with the observed distilled yield and instantaneous efficiency. This observed data is learned with the C4.5 decision tree algorithm and also predict the distilled yield and instantaneous efficiency. Finally to classify and predict the required parameters “If Then Else” rules are prepared. The experimental results demonstrate, the proposed C4.5 algorithm provides higher accuracy as compared to similar state of art techniques. The proposed technique offers up to 5-9% improved outcome in terms of accuracy.

Keywords— If Then Else, rule mining, decision making, Solar Still Plant, fuzzy logic.

I. INTRODUCTION

Solar energy is a healthy source of light and heat. It is harnessed and sometimes good for our health and life. It directly provides us vitamin D. It is an essential source of renewable energy. Solar energy is broadly classified as passive or active, depends upon how the energy is captured and transformed into power. Active energy involves use of photovoltaic systems, to produce harness energy like water heating systems. On the other hand passive techniques include selecting materials with favourable thermal mass or light-dispersing properties. In this work the active energy generation technique is used to distil water. Basically, water is seed of life, but due to climate change and environmental effects the issue of water crisis is rising day by day. Moreover it, in remote areas survival and/or during conditions of disasters, drinking water is least requirement.

Thus, the work is focused on water purification using solar energy. Additionally we want to monitor and predict the possible production ability (productivity) of distillation plant. This paper contributes on following objectives:

1. Designing a single slope basin solar still plant
2. Perform experiments and collection of data with the help of designed water plant
3. Applying predictive techniques, to predict performance and productivity of solar still plant

Therefore, to demonstrate the proposed contributions first a survey is presented. By reviewing the literature we conclude different rule mining techniques that are helpful for predicting distilled yield using the collected attributes. Further, configuration of the developed solar plant is provided for experimentation and observation collection. Finally a data mining algorithm is proposed for accurate prediction of distilled yield. Finally, a comparative performance study is conducted to justify the proposed work.

II. LITERATURE REVIEW

This section offers details about recent contributions and developed methodologies. These methods are helpful to design and develop an accurate prediction system.

Fausto Cavallaro [1] presents his work on Takagi-Sugeno Fuzzy Inference System for Developing a Sustainability Index of Biomass. The Takagi-Sugeno fuzzy inference modelling builds a synthetic index to assess sustainability of production for energy purposes. **Qasem Abdollah Nezhad et al. [2]** Investigate on fuzzy logic controllers based on Takagi -Sugeno & Mamdani model. Author introduced Takagi Sugeno model and comparing it with other controllers. That can hold pendulum in vertical position on cart with more sensitivity and accuracy. **Tomohiro Takagi et al. [3]** offers a Fuzzy Identification Systems and Its Applications to model and Control. They demonstrate it with two industrial processes. First water cleaning process and other is a converter in steel-making process. The fuzzy model is described by IF-THEN rules to represent input-output relations. The main feature of this fuzzy model is to express the local dynamics of fuzzy rules by a linear system. The model is achieved by fuzzy “blending” of the linear systems. **Plamen Angelov et al. [4]** provides On-line Design of Takagi-Sugeno Models. They presents an approach to design of TS models, it evolving structure and can learn recursively with real-time data. The TS model can be decomposed into two sub-problems one is on-line recursive clustering for the rule base learning, and other is estimation of the consequent part of parameters.

Fuzzy Logic techniques were proposed for power demand prediction [5]-[6].

Revised Manuscript Received on January 30, 2020.

* Correspondence Author

Neha Yadav*, Mathematics Department S.D.Bansal College of Technology, Umariya, Indore (M.P.) India

Vivek Raich, Research Center of Mathematics, Govt. Holkar Science College, Indore (M.P.) India

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

Fuzzy Logic has been applied successfully to a large number of applications. This work presents model of FLS, comprising the control rules and term sets of variables with their relates fuzzy sets, with the help of extended set theory to handle partial memberships issue, and enabling to express human concepts [7]. Jigeesh has made a preliminary work to simulate a solar water desalination using fuzzy logic rule based reasoning system [8].

Shanmugan, [9] discussed Fuzzy logic modeling of floating cum tilted – wick solar still. In this Fuzzy modeling and simulation a floating cum tilted – wick solar still has been developed it makes transparent to qualitative interpretation and analysis. A set of fuzzy rules have been developed based on general analogy between changes in solar radiation intensity, yield of distillate output change in the weak. **Panchal [10]** study Effect of different parameter on double slope solar still productivity. Author observe single basin solar still for converting waste water into potable water. The aim of study is to find effect of various parameters on performance of solar still using water depth inside the still, sprinkler and various dies. They proved that black die will increase the output of solar still, sprinkler increase the condensation rate with lower depth of water. Black die is a good parameter to increase the distilled output with low water depth.

Shanmugam et al. [11] discussed on Fuzzy logic modelling of single slope single basin solar still. Authors developed the thermal analysis of single slope single basin solar still with fuzzy logic. Experimental observation has been carried out on 9 May 2012 in Dhanalakshmi college of Engineering Chennai Tamilnadu. Qualitative interpretation and fuzzy rule have been developed between change solar radiation intensity, yield of distilled output Mamdani model has been used to predict a distillate output. **Hrushikesh Kulkarni et al. [12]** performed experimental evaluation of still using phase change. The different designs of solar still with PCM are analysed. Phase change and thermal energy storage materials also play an important role to enhance internal energy of system. Top cover cooling is also one of the methods to induce faster condensation inside the solar still. **J.I. Orisaleye et al. [13]** developed and evaluate solar water still with characterization of water quality before and after distillation.

III. EXPERIMENTAL SOLAR PLANT DESIGN

This section introduces the experimental solar water distillation plant. The figure 3.1 demonstrate the proposed solar still, the dimension of our experimental plant is 119*80*30 cm as a box. That box is made with GI sheet of 3 MM thickness. Additionally to cover this box a 5 MM thick glass is used, with a fixed slop of 15°. A J-shape drainage is developed to collect distillate yield as output in measuring jar. The basin of this plant is made of G.I. sheet and thin copper sheet. To enable it for absorbing solar radiation black paint is coated over it. Finally, basin temperature, water temperature, condensing cover temperature has been measured. Additionally solar radiation intensity and ambient temperature has been measured using digital thermometer.



Figure 3.1 experimental solar still

The collected experimental records are used with a data model or method to predict the performance of water still plant. It is used as a data sample over the supervised learning model. The example of prepared dataset is given in table 3.1.

Table 3.1 example dataset

S.No.	basin Temperature (°C)	Water Temp (°C)	Glass Temperature (°C)	Solar Radiation (W/m²)	Distilled water yield	
					Experimental	Instantaneous Efficiency (%)
1	43	48.9	37	800	0.010	12.47
2	45	54.4	41	850	0.040	11.33

The table contains observations for different purification and temperature relevant attributes. Finally the experiments are conducted with developed plant, for one month between Date 23 Nov 2018 to 22 Dec 2018. The total 30 days samples are collected additionally in a single day 18 observations are collected. Initially entire dataset contains total 30*18= 540 instance among we found variations in each sample. Therefore to reduce the error possibility in collected samples, thus those samples are removed which are not complete or for cloudy weather days. So the samples of 5 days are removed thus 90 samples are removed and only 450 samples are used for experimentation.

IV. PREDICTIVE DATA MODELLING

This section provides understanding about proposed model for predicting distillation performance of solar still plant. Additionally summarized process step of predictive data model is also described as algorithm steps.

A. Predictive data modeling

The technique for predicting distilled yield of implemented water distillation plant is demonstrated in figure 4.1. This is a data mining model which generate “IF THEN ELSE” rules for predicting performance of water purification.

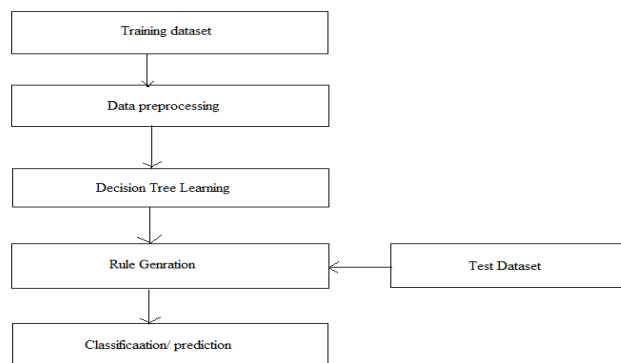


Figure 4.1 proposed data model

Training dataset: The collected experimental outcomes using the solar water still plant is basically organized as the dataset.

The data set is used to learn the data samples with a soft computing methodology and provide the predictive outcomes. Therefore complete set of information is produced to the system for learning.

Data pre-processing: Data pre-processing is an essential step of data mining and machine learning. That process is used for improving quality of data. Therefore it helps to refine attributes and noisy contents from data to improving performance of algorithm learning. In this work, data is refined to find and remove missing samples and instances which containing noise. After cleaning up data, it is used for data modelling.

Decision tree learning: The system needs to generate decision rules therefore C4.5 or J48 decision tree algorithm is used. This decision tree is an extension of a popular decision tree namely ID3. That is use the concept of entropy and information gain to create the data partitions. Additionally attribute with highest information gain is selected to create decision tree. The C4.5 algorithm recourses partitioned sub-lists to create a complete decision tree. The algorithm considers the following constraints.

1. If samples in dataset contain same class then it simply creates a leaf node as decision tree.
2. If information gain computation is not feasible then algorithm creates a node higher up then tree using the expected value of class.
3. If previously-unseen class encountered, the algorithm creates a decision node using the target value.

Before providing the steps of decision tree development it is required to understand information gain. Therefore it is required to discuss entropy first. Let's assume that resultant decision tree classifies data into two classes, i.e. P (positive) and N (negative). Therefore the entropy S based on this binary classification is:

$$E(S) = -P(Pos) \log_2 P(Pos) - P(neg) \log_2 P(neg)$$

P (pos): ratio of positive samples, P (neg): ratio of negative samples

For cutting down depth of a decision tree, while traversing the same, selection of the best possible characteristic is required to split tree branches, it is clearly shown attribute with minimum entropy will be superlative pick. The information gain can be termed as required drop in entropy in relation with individual attribute during splitting. The information gain, Gain (E, A) of an attribute A can be computed using,

$$Gain(E, A) = Entropy(s) - \sum_{n=1}^v \frac{E_v}{E} \times Entropy(E_v)$$

The concept of gain can be utilized to decide positions of attributes to construct decision tree. Every node is positioned the attribute with maximum gain among the attributes that is not considered in path of root yet. The intention of this is:

1. To generate small sized tree to identify patterns by using decision tree splitting.
2. To attain desired level of unfussiness of decisional approaches.

C4.5 decision tree is developed by Quinlan, as an algorithm. This algorithm returns decision tree as learning outcome [14]. The following steps can be used for generating decision tree using input dataset:

INPUT: A set of data (D) with the means of discrete variables.

OUTPUT: A decision tree T which is constructed by passing data set.

- 1) A node (X) is created;
- 2) If the instance falls in same class.
- 3) Make node (X) as leaf node and assign a class label C;
- 4) If the attribute list is empty,
- 5) Make node(X) a leaf node and assign a class label of most frequent class;
- 6) Choose an attribute which has highest information gain, and then marked as test-attribute;
- 7) If X in role of test-attribute; (To recognize the value for every test-attribute for dividing samples)
- 8) Generate a new branch of tree that is suitable for test-attribute from node X; (Let Bi is a group of test-attribute in samples)
- 9) If Bi is NULL,
- 10) Add a new leaf node, with class label of most common class;
- 11) ELSE
- 12) Add a leaf node and returned by Generate-decision-tree.

Rule generation: the developed data structure (i.e. decision tree) is used here as input to the system for extracting "IF THEN ELSE" rules. To extract rules using decision tree the branches of tree is converted as decision rules and leaf nodes are works as decisions. The example of rule extraction using decision tree is given using figure 4.2:

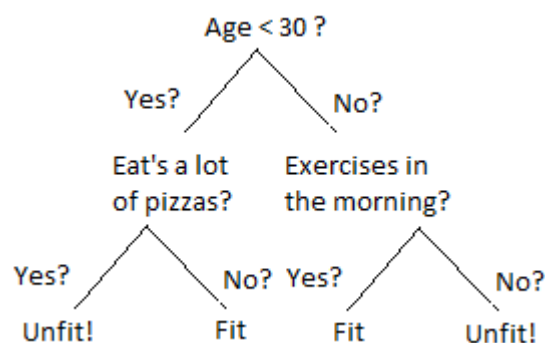


Figure 4.2 example of decision tree

The figure contains a decision tree example. This decision tree is composed with attributes as nodes and attribute values are labelled over branches. Based on corresponding attribute values and name the leaf node is traversed. The leaf node contains the target value or predictive outcome. The above example of decision tree helps to construct four decisional rules for instance.

If "age < 30 = YES" and "Eats' a lot of Pizzas = YES" then "prediction = Unfit".

In above given decision tree we can create three more rules that help us to classify the similar patterns. In the similar manner as given in example decision tree is converted into rules. These rules are further used for predicting target values.

Test dataset: However data mining and machine learning methods needs some examples for learning. Similarly, validation of developed model requires a set of unknown data samples. Additionally, using the input attributes the prediction is performed. This set of data is termed as the test dataset. Here the initial dataset is used for selecting 30% of random data samples as test dataset.

Classification/prediction: Finally test dataset is used with generated rules. Using these rules each input test samples are evaluated to get the decision class label. The estimated class label for a single instance of data is recognized as the predicted data. The predicted values are used for cross validation and performance evaluation of the prepared data model.

B. Proposed algorithm

The functional details about the proposed data model are described in previous section. This section offers the summarized steps of process involved as the algorithm steps. Thus the processing of input samples and produced predicted outcomes are demonstrated. The algorithm steps are notified using table 4.1.

Table 4.1 proposed algorithm

Input: observation table T
Output: predicted class label C
Process: <ol style="list-style-type: none"> 1. $D_n = readInputData(T)$ 2. $for(i = 1; i \leq n; i++)$ <ol style="list-style-type: none"> a. $P_i = preprocess(D_i)$ 3. $end\ for$ 4. $T_{model} = C45.TrainModel(P_n)$ 5. $R_k = ExtractRules(T_{model})$ 6. $Test_m = P_n.Split(30\%, random)$ 7. $for(j = 1; j \leq m; j++)$ <ol style="list-style-type: none"> a. $for(l = 1; l \leq k; l++)$ <ol style="list-style-type: none"> i. $C = R_l.Classify(Test_j)$ b. $end\ for$ 8. $end\ for$ 9. Return C

The table 4.1 shows the data processing using the proposed data model, namely C4.5 decision tree algorithm. According to reported process steps of algorithm, it accepts observations of experiments conducted with the solar still plant. The collected data is denoted in this algorithm as T . This data is transformed into a 2D vector of n instances and defined here as D_n . In further each sample of data D_n is pre-processed and the pre-processed instance of data is stored in a variable of similar size P_n . This variable is used with the C4.5 decision tree algorithm for preparing decision tree that tree data structure is defined in this algorithm as T_{model} .

That decision tree further processed for generation of rules, the k number of rules are extracted from decision tree and stored in a variable R_k . Now we have the k number of decision making rules for prediction of solar still plant performance. Thus m number of test data samples is evaluated with the help of rules and their class labels are estimated using rules. Finally a list of class labels C is generated for all the m number of test samples. That is the final outcome of the system.

C. State of art methods

In literature a number of approaches using soft computing available which are helpful for extracting the decisional rules for classification and prediction. In these methods some of the techniques are developed on the basis of genetic and naturally inspired algorithm. Additionally some of the techniques are designed using fuzzy logic and statistical methods. In this context a noteworthy contribution is identified by *Vivek et al [15], [16]*. In these research article the authors are designed a Single Slope Single Basin Solar Still Using Fuzzy Logic, to predict distillate output and Instantaneous Efficiency with the help of Mamdani Model and Takagi Sugeno Inference respectively. Authors present Takagi Sugeno Model and Mamdani Model for predicting distillate output and Instantaneous Efficiency. These two methods are compared with the proposed model for demonstrating the efficiency and accurateness of the proposed decision tree based data model.

V. RESULTS ANALYSIS

This section offers detail about conducted experiments and measured parameters for justifying the proposed technique.

a. Accuracy

The accuracy of the proposed data model is described in this section additionally other state of art data models are also compared. Basically, the accuracy of a data mining system is the ratio of success for predicting accurate patterns based on the generated rules. The following equation can be used for measuring the performance in terms of percentage accuracy.

$$accuracy (\%) = \frac{\text{total correctly predicted}}{\text{total samples for prediction}} \times 100$$

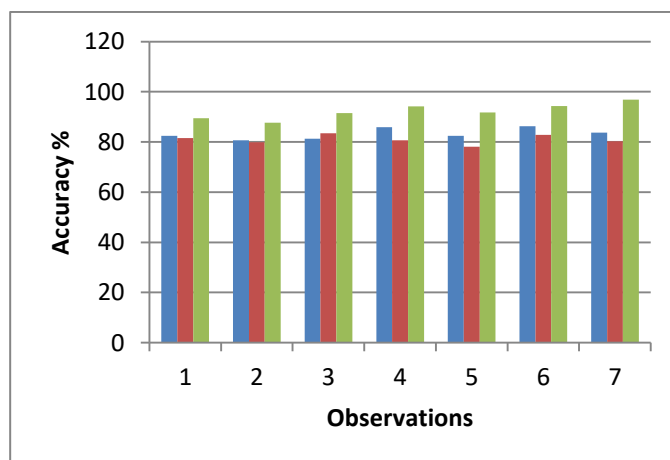


Figure 5.1 accuracy (%)

The prediction accuracy of proposed and other two state of art techniques are reported in figure 5.1. That is a line graph which is constructed using experimental observations. The X axis of this line graph shows the experiments conducted and Y axis shows the corresponding accuracy of the techniques. The measured accuracy is reported here in percentage (%). Based on the experimental analysis classical fuzzy logic based data models are producing similar accuracy. Additionally proposed C4.5 decision tree based rule mining technique provides higher accuracy as compared to both fuzzy based approaches.

b. Time complexity

The time complexity is recognized here as the algorithm run time. In order to produce required rules using input training samples. The time requirements of the proposed system are measured using this function.

$$\text{time required} = \text{algorithm end time} - \text{start time}$$

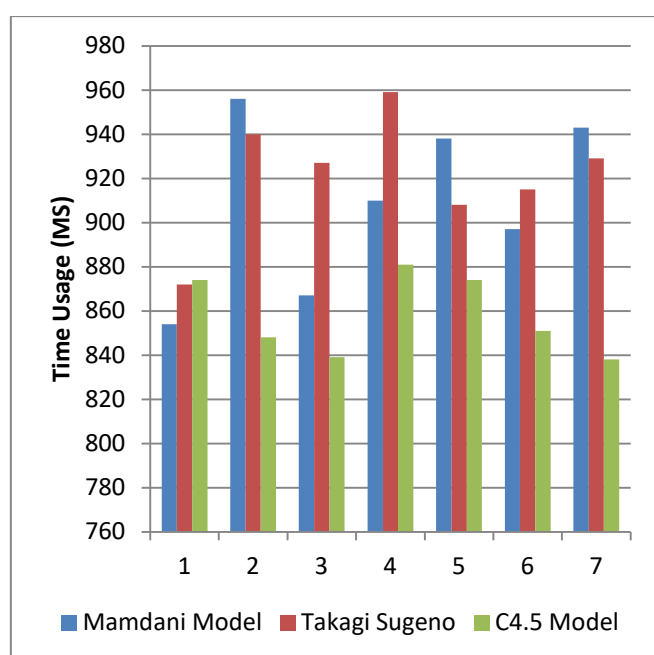


Figure 5.2 time complexity (MS)

The running time of the algorithms is measured here in terms of millisecond (MS). For all three algorithms the similar size of dataset is used for measuring and comparing the performance or time requirements. The figure 5.2 is a line graph which contains method wise experimental observations. The X axis of the diagram shows the experiments and Y axis shows the time consumed for generating “IF THEN ELSE” rules. According to the demonstrated performance of “IF THEN ELSE” rule mining algorithms proposed model works faster than other state of art fuzzy logic based decision rule mining algorithms. Therefore proposed method is efficient and less time consuming.

c. Space complexity

The space complexity of an algorithm can be defined as the requirements of main memory for execution of a program. The following function is used for computation of memory usages of implemented algorithms.

$$\text{memory usage} = \text{total assigned memory} - \text{total free space}$$

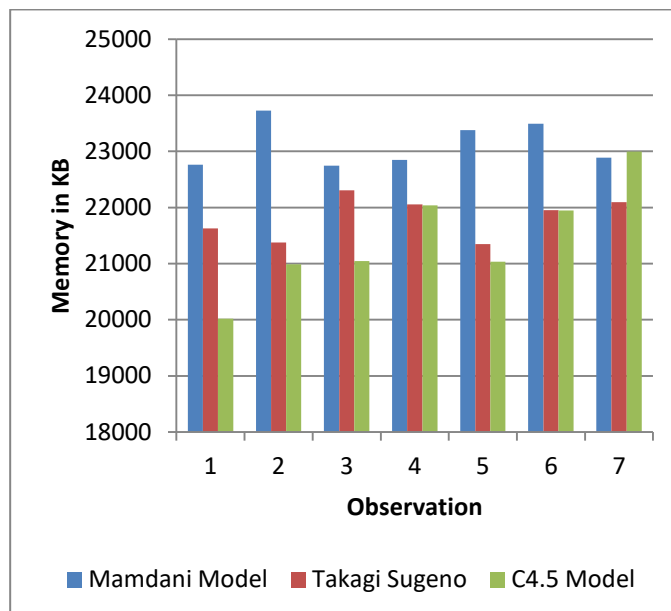


Figure 5.3 memory usages

Basically, when a process is appeared for execution, the system assigns an amount of main memory to that process. And when the program is executed the memory is utilized by the process to hold data and instruction. The balance amount of main memory is measurable during this process. Thus difference between available space and free space is measured as memory usages. The memory usages for “IF THEN ELSE” rule mining is demonstrated in figure 5.3. It is a line graph for memory usages of the proposed and state of art methods. The memory usage is measured here in terms of kilobytes (KB). The proposed decision tree algorithm is efficient as compared to traditional fuzzy based approaches. The classical methods are demonstrating higher memory usage as compared to proposed technique. The memory usages can be an indicator for the amount of rules is developed for classification. The amount of rules are stored over main memory and utilized when required for predicting the performance of solar still plant.

V. CONCLUSIONS

This section provides summary of efforts placed to predict performance of solar water distillation plant. Additionally future extension of the work is also reported.

A. Conclusion

Now in these days’ data mining techniques are frequently used in various kinds of data analysis and prediction system development. Data mining techniques have the potential to estimate or compute the decisions by analysing available attributes. In this work decision tree algorithm is used for mining IF THEN ELSE rules. In literature a number of rule mining algorithms are available among them fuzzy logic based rules and decision tree based techniques are much popular. These rules are used for decision making and prediction in different applications. Thus in this work we utilized the decision tree based rules for predicting performance of the solar still plant.

In this context, first an experimental solar distillation plant is established. Using this plant experiments are conducted to collect observational data. Using this observation table a dataset is designed. That data set is used further for learning about the different temperature levels using supervised learning algorithm C4.5 decision tree. The decision tree algorithm consumes the entire learning data and producing a tree data structure. In order to prepare required decision tree data splitting and other calculation required such as entropy and information gain. Basically ID3 decision tree is revised for optimizing the performance. Thus modified version of ID3 algorithm is known as C4.5 or J48 decision tree. Advantage of C4.5 algorithm is that, it's small size and low run time. Therefore, tree consumes less amount of main memory also. The generated decision tree is used for extracting decision rules. Using rules class labels for test dataset instances is predicted because test dataset is unlabelled. The classified data using the generated rules is measured as accuracy of the proposed system. The implementation of the proposed distilled yield prediction system is accomplished using WEKA data mining tool and with the help of JAVA based NETBEANS IDE. After implementation of the system performance is evaluated and reported in table 6.1.

Table 6.1 performance summary

Parameters	Mamdani Model	Takagi Sugeno	C4.5 Model
Accuracy	Moderate	Moderate	Higher
Time consumption	Fewer higher	Moderate	Low
Memory usages	Fewer higher	Moderate	Low

According to the listed results in table 6.1, the proposed decision tree based rule mining technique is found efficient and less time and effort consuming. Therefore that technique is helpful for predicting the performance of a water treatment plant performance prediction.

B. Future work

The main aim of the proposed work is to enhancing IF THEN ELSE rule mining technique which is used traditionally. Thus decision tree based rules mining is proposed and implemented to predict distillate yield for solar water plant. In near future the following work is proposed.

1. Exploring more techniques that are helpful in “IF THEN ELSE” rule mining such as ACO (ant colony optimization), and other genetically inspired methods.
2. Work will be enhanced with some other kinds of opaque data modeling techniques for enhanced prediction ability of plant with continuous data streams.
3. Working with some rule optimization techniques that minimize the comparison cycles during classification process.

REFERENCES

1. Fausto Cavallaro “A Takagi – Sugeno Fuzzy Inference System for Developing a Sustainability Index of Biomass” Sustainability,7 ,12359 -12371 (2015).
2. Qasem Abdollah Nezhad, Javad Palizvan Zand and Samira Shah Hoseini “An Investigation on fuzzy logic controllers (Takagi-Sugeno & Mamdani) In Inverse Pendulum System (IJFLS) Vol.3, No3, July (2013).
3. Tomohiro Takagi and Michio Sugeno “Fuzzy Identification of Systems and Its Applications to Modeling and Control” IEEE, Vol. SMC-15, NO. 1, Jan/ Feb (1985).
4. Plamen Angelov, Dimitar Filev “On-line Design of Takagi-Sugeno Models”: IFSA 2003, LNAI 2715, pp. 576-584, (2003).
5. Hossain, A., Ataur, R., Rahman, M., Hasan, S.K. and Jakir, H., 2009, “Prediction of Power Generation of Small Scale Vertical Axis Wind Turbine Using Fuzzy Logic,” Journal of Urban and Environmental Engineering (JUEE), 3(2), pp. 43-51.
6. Al-Anbuky, A., Bataineh, S. and Al-Aqtash, S., 1995, “Power demand prediction using fuzzy logic,” Control Engineering Practice, 3(9), pp. 1291-1298.
7. Carman, K., 2008, “Prediction of soil compaction under pneumatic tires a using fuzzy logic approach,” Journal of Terramechanics, 45, pp. 103-108.
8. Jigesh, N., 2003, “Fuzzy Rule Based solar Desalination Expert system, proc.” International Symposium on Information Technology, Kuala Lumpur Malaysia, (IT Sim 2003), pp. 433-438.
9. S.Shanmugan, Fuzzy logic modeling of floating cum titled –wick solar still. International journal of recent scientific research vol 4, issue, 5 pp 579- 582, May 2013
10. N. Hitesh Panchal. Effect of different parameter on double slope solar still productivity. International Journal of advances in Engg Science vol 1. Issue 2 April 2011.
11. S. Shanmugan and G.Krishnamoorthi Fuzzy Logic Modeling of Single Slope Single Basin Solar Still .international Journal of Fuzzy Mathematics and systems. Volume 3. Number 2(2013) pp. 125-134.
12. Hrushikesh Kulkarni, Chinmay Kute, Chirag Patel, Akshay Tavse, Prof. Lokesh R. Dhumne Experimental Investigation and Performance Evaluation of Solar Still Using Phase change material (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 05 | May-2018
13. J. I. ORISALEYE, S. O. ISMAIL, M. OGBONNAYA, A. A. OGUNDARE. Development and performance evaluation of solar water still. Tome XI [2018] | Fascicule 1 [January – March]
14. Kundan Kumar Mishra, Rahul Kaul, “Audit Trail Based on Process Mining and Log”, International Journal of Recent Development in Engineering and Technology, Volume 1, Issue 1, Oct 2013
15. Vivek raich, Neha Yadav. Single Slope Single Basin Solar Still Using Fuzzy Logic, IOSR journal of mathematics vol 14 issue 3may June 18.
16. Vivek raich , Neha Yadav. Takagi Sugeno Inference for Single Slope Single Basin Solar Still Journal of Computer and Mathematical Sciences, Vol.10(4),670-679 April 2019c

AUTHORS PROFILE



Neha Yadav, Department of Mathematics, S.D.Bansal college of Tecnology umariya mhow,B.sc., M.Sc. research area fuzzy set theory



Dr. Vivek Raich ,B.sc., M.sc, Ph.D. Research Center of Mathematics, Govt.Holkar Science College, Indore (M.P.) India Research area fuzzy set , fixed point theory