energies

Article

Generation of Hydro Energy by Using Data Mining Algorithm for Cascaded Hydropower Plant

Iram Parvez [1], Jianjian Shen [1,2,*], Ishitaq Hassan [3] and Nannan Zhang [4]

[1] Institute of Hydropower and Hydro informatics, Dalian University of Technology, Dalian 116024, China; irum.pervaiz@mail.dlut.edu.cn
[2] Key Laboratory of Ocean Energy Utilization and Energy Conservation of Ministry of Education, Dalian 116024, China
[3] Department of Civil Engineering, Capital University of Science and Technology, Islamabad 44000, Pakistan; eishtiaq@cust.edu.pk
[4] Basin Projects Operation Management Department China three Gorges Limited Corporation, Yichang 443133, China; zhang_nannan@ctg.com.cn
* Correspondence: shenjj@dlut.edu.cn

Abstract: The thirst of the Earth for energy is lurching towards catastrophe in an era of increasing water shortage where most of the power plants are hydroelectric. The hydro-based power systems are facing challenges in determining day-ahead generation schedules of cascaded hydropower plants. The objective of the current study is to find a speedy and practical method for predicting and classifying the future schedules of hydropower plants in order to increase the overall efficiency of energy by utilizing the water of cascaded hydropower plants. This study is significant for water resource planners in the planning and management of reservoirs for generating energy. The proposed method consists of data mining techniques and approaches. The energy production relationship is first determined for upstream and downstream hydropower plants by using multiple linear regression. Then, a cluster analysis is used to find typical generation curves with the help of historical data. The decision tree algorithm C4.5, Iterative Dichotomiser 3-IV, improved C4.5 and Chi-Squared Automatic Interaction Detection are adopted to quickly predict generation schedules, and detailed comparison among different algorithms are made. The decision tree algorithms are solved using SIPINA software. Results show that the C4.5 algorithm is more feasible for rapidly generating the schedules of cascaded hydropower plants. This decision tree algorithm is helpful for the researchers to make fast decisions in order to enhance the energy production of cascaded hydropower plants. The major elements of this paper are challenges and solution of head sensitive hydropower plants, using the decision-making algorithms for producing the generation schedules, and comparing the generation from the proposed method with actual energy production.

Keywords: short-term scheduling; cascaded hydropower plants; data mining techniques; energy production; generation schedules

check for updates

Citation: Parvez, I.; Shen, J.; Hassan, I.; Zhang, N. Generation of Hydro Energy by Using Data Mining Algorithm for Cascaded Hydropower Plant. *Energies* **2021**, *14*, 298. https://doi.org/10.3390/en14020298

Received: 9 December 2020
Accepted: 29 December 2020
Published: 7 January 2021

1. Introduction

Among other renewable energy resources, such as wind, solar, etc., hydropower is the vital source of producing electricity around the globe. It reduces the emission of greenhouse gases which is one of the main aspects of global warming. In order to minimize the effect of greenhouse gases, the electrical power industry is playing its role by exploiting renewable and clean energy as compared to fossil fuel. Hydro-based power systems are the main sources that contributed towards sustainable energy [1–4]. Precipitation, melting snow and streams are the main supply of inflow for hydropower plants. Further, for short-term hydro scheduling, inflow forecasting is used to identify predictions while stream flow forecasting is helpful for calculating the long-term hydrological cycle, such as evaporation, temperature and precipitation. Flow of the river is considered as the primary source for making the decision related to an ecosystem that is helpful for sustainable hydropower [5,6]. Scheduling of hydropower plants is a difficult task because of variation in the peak demand as human activities follow regular seasonal and yearly periods [7]. The power generation from a hydropower reservoir depends on different parameters, such as yearly discharge, water head, load rate, etc. Good planning and slight computational improvement can give more energy with the same quantity of water [8]. The aim of hydro scheduling is to generate the maximum energy by using the available water. During the past few decades, many studies have been performed on short-term hydro scheduling [9]. Short-term hydro scheduling mainly refers to the determination of the hourly schedules for the hydro plants to satisfy the forecasted demand as well as the technical constraints [10]. It is mathematically classified as non-convex, non-linear and large-scale discrete problem. The scheduling problem is referred to as the unit commitment problem [11]. Moreover, in short-term hydro scheduling, the hours ahead to day ahead time horizon are followed. In power system operations, one of the key issues is unit commitment. There are several problems faced by the electrical power industry which include environmental problems, unit commitment and variation of consumer demand [12]. As hydropower is the viable source of renewable energy, therefore, feasible generation of hydro units has an important role in electricity market. In the electrical power industry, the most significant and critical issue is unit commitment [9]. The main purpose of unit commitment is to realize which generating units are needed to be switched on or off over a given time period subject to given spinning reserve and load forecast constraints for maximum efficiency. Unit commitment is a complex, mixed-integer, non-linear programming problem [13–16]. The objective behind solving unit commitment problem is to balance production with demand while optimizing costs and resources. Thus, the problem of unit commitment needs to be solved in a way that it satisfies the production demand with minimum cost and water resources. The main research questions are as follows: (1) How more electricity can be produced in the water scarcity era? (2) What is the accuracy of decision tree algorithms in forecasting hydro power generation for shorter periods? (3) What is the practicability of data mining algorithms as compared to other mathematical models?

Electric power generation is beneficial for welfare, sustainable development and human life. The prosperity of areas depends on the access of its people to electricity. In today's world, around 20% of the population is still living in the dark; they still have no access to computers, refrigeration, lighting and running water. The objective is not only increasing the production of energy but is also focused on increasing renewable energy. The activities of humans are responsible for greenhouse emissions, which in turn changes the climatic configurations of the earth. In greenhouse gas emissions, carbon dioxide is the key issue that contributes around 76% of these gases. The largest sources of producing carbon dioxide are oil, coal and natural gases, which are used for producing electricity. In order to eliminate the effect of greenhouse gases, the world should focus on renewable energy sources [17,18]. The power generation of a hydropower reservoir depends on different parameters such as inflow, annual discharge and water head, load rate, etc. The inflow parameter is determined by the weather forecast for hydropower generation [19]. Moreover, hydropower is a potential source of energy for Himalayan region countries but Pakistan and Nepal struggle to fulfill the demand of electricity. The demand of energy is increasing 11–13% each year in Pakistan [20]. Hydropower is one of the important pillars of energy in China [21]. In the last two years, 4.1 GW hydropower energy was added in Europe [22]. A systematic life approach was conducted to quantify the hazardous emissions in alpine and non-alpine regions of Europe; it also promotes the use of hydropower as a clean energy resource [23]. Depending on the characteristic of data availability, power system and computational assets, different methods and techniques are used for generating optimal energy. In order to deal with the deregulation environment of the power industry, non-linear programming is used [24,25]. The key point is to maximize the water storage level and revenue. The problem is named as quadratic programming as it consists of quadratic function and the problem is normalized by linear and non-linear constraints. The cascaded reservoir with head sensitive is considered. Moreover, the Portuguese cascaded hydro-based systems are taken as a case study. In the dynamic programming, a recursive relationship is followed by Bellman's Principle. The work of dynamic programming is to find the optimal policy of each state. Moreover, the feasible solution is found by using the backward procedure; this process is carried out stage to stage till it reaches an initial stage [26–28]. In real world problems, complex mathematical measures are taken in mixed integer linear programming (MILP) mechanism. The MILP model is only applied to some unit turbines due to its limited implementation [29,30]. The commitment schedules are found by using the Lagrange relaxation solution technique which is subject to all capacity and reserve constraints in order to fulfill the purpose; the Lagrange multipliers are found, which are useful in finding the optimal solution. The technique uses the economic dispatch measurements in order to satisfy the condition of demand with reserve constraints of a single unit [31]. Furthermore, unit commitment is divided into two groups—one is stochastic unit commitment and the other one is deterministic unit commitment. A large number of studies are carried out on the deterministic unit commitment problem, but numerous research studies also emphasize stochastic unit commitment which explains the formulation and methodology of the stochastic unit commitment problem. In stochastic unit commitment, the main problem is uncertainty. In the order to solve the unit commitment problem, several studies in the literature discussed scenario-based stochastic programming with different methods such as progressive hedging [32,33], dual decomposition [34], benders decomposition [35], spatial decomposition [36], cutting plane [37], dynamic formulation [38] and heuristic methods [39–42]. Moreover, data mining is used to obtain the useful data from large data sets gathered from various sites [43]. The objective of this method is to determine the information that is unknown in the historical data set [44]. Different methods such as Bayesian network [45], artificial neural network [46], clustering [47], classification, regression [48], genetic algorithm [48] and decision trees are used to extract the information from large data sets [43]. Clustering and decision trees are the simplest and most successful methods among all aforementioned methods [49,50]. Clustering is used to recognize a similar group of objects. Samples that belong to the same cluster have similar properties while those having dissimilar behavior fall into another cluster group. Furthermore, different types of clustering methods are used for different data sets such as partitioning methods, model-based methods, grid-based methods, density-based methods, etc. [43]. The decision tree is one of the most successful data mining techniques used for forecasting and classification of the data set [51]. It follows the greedy approach, as it is a supervised learning algorithm. The decision tree represents the data in a way that is easily understandable. The decision tree is feasible in a way that it can handle a large set of data and tackles both numerical and categorical variables [52–56]. The decision tree consists of a variety of algorithms such as C4.5, Iterative Dichotomiser 3 (ID3), Chi-Squared Automatic Interaction Detection (CHAID), Classification and Regression Tree (CART), etc., and each algorithm is based on a specific mechanism. Therefore, it is necessary to evaluate the performance of different decision tree algorithms for hydro scheduling in order to get the efficient and optimized results of energy generation.

The conducted study validates that data mining techniques are best for solving the hydro scheduling problem in the short-term horizon specific for Tianshengqiao cascaded hydropower plant. The current work also intends to be the pioneer study in introducing new research endeavors that comprise the use of data driven modeling in Tianshengqiao cascaded hydropower systems. It also reveals that the decision tree algorithm C4.5 is good in making quick decisions for schedules. The obtained result shows that the decision tree algorithm C4.5 has a minimum percentage of error. Thus, this algorithm can be securely used by engineers in the future for the hydrothermal scheduling problem of cascaded hydropower plants.

The main objectives of the current work are to: (1) find the feasible generation schedule for short-term hydro scheduling of cascaded hydropower plant, (2) examine the electricity demand in the summer and winter season and make optimal generation schedules, (3) make full use of storage and regulation of the main reservoir in order to increase overall efficiency of water utilization of cascaded hydropower plants, and as well the proposed methods can immediately determine the generation schedules of the cascaded hydropower plants, and (4) train different attributes, such as reservoir daily discharge, energy production, water level of reservoir and class, such as generation schedules, collectively to obtain a decision-making library for generating scheduling of cascaded hydropower plants. Several studies have been carried out in the past for scheduling of cascaded hydropower plants with one reservoir [57]. Moreover, it is hard to find the optimal solution for the real-world problem because of different constraints and conditions of optimization. Therefore, it is important to find a quick generation schedule of cascaded hydropower plants for effective utilization of water resources, which is the key objective of this study. The layout of this research work comprises four sections. Section 2 comprises the study area and methodology which include the study area, methodology and data set of the data mining algorithms for both the winter and summer season. Section 3 outlines the results and their analysis. Section 4 is the discussion part, which depicts the comparison of data mining algorithms with the previous optimization techniques. Section 5 explains the conclusions.

2. Study Area and Methodology

2.1. Study Area

In China, during the 1980s, the annual average increase in power generation was by 7.5% and in the next five years the annual average rate improved to 8.5% [58]. In April 1991, the "10-year plan for National Economic and Social Development" and "Eighth 5-Year Plan" were set, which increased the power generation by 5.6% on an annual average from 1991 to 1995; then, 6.3% from 1996 to 2000 [58]. The demand of power was increased in the Guangdong province because this province consists of two coastal cities as well as three economic zones, namely, Zhuhai, Shanton and Shenzhen. In the 1980s, the power generation demand was increased by 12% on a yearly basis and an additional increase was predicted with the commercial growth. Conversely, the supply of power was limited, and there was alarm that in the dry season, the power shortages might occur due to the shortage of water. The sources of generating power in the province were either by coal thermal power generation or by hydropower generation. The coal thermal power generation contributed 83% in power generation, while hydropower generation was only 16% at that time. However, the coal resources needed to be reserved because 73% of the coal was delivered to the other provinces. In order to replace the power generated by coal with hydroelectric power generation and make full use of the abundant water of the southern region, the Chinese government developed the "Hongshui River Comprehensive Use Plan" for the growth of the Hongshui river. In this plan, ten power stations were constructed with a total capacity of 11,120 MW. The objective of this project was to maintain power generation at the other hydropower stations that were at the lower side of the Hongshui river. Thus, Tianshengqiao plant 2 was constructed at the uppermost side of the river to maintain power generation and to secure a power supply source outside of the Guangdong province. Tianshengqiao hydropower plant 1 consists of a concrete-faced rock fill dam (1140 m long crest and 178 m in height), an emptying tunnel, a large chute spillway, a power generation and powerhouse. On the left bank, the powerhouse is located and has four tunnels and four turbines, and the total installed capacity is 1200 MW. Water released from the dam's reservoir also fed Tianshengqiao 2 dam, which is located on downstream. The dam then diverts water to the actual power station and produces 1320 MW [59]. The satellite view of Tianshengqiao hydropower plant is shown in Figure 1.

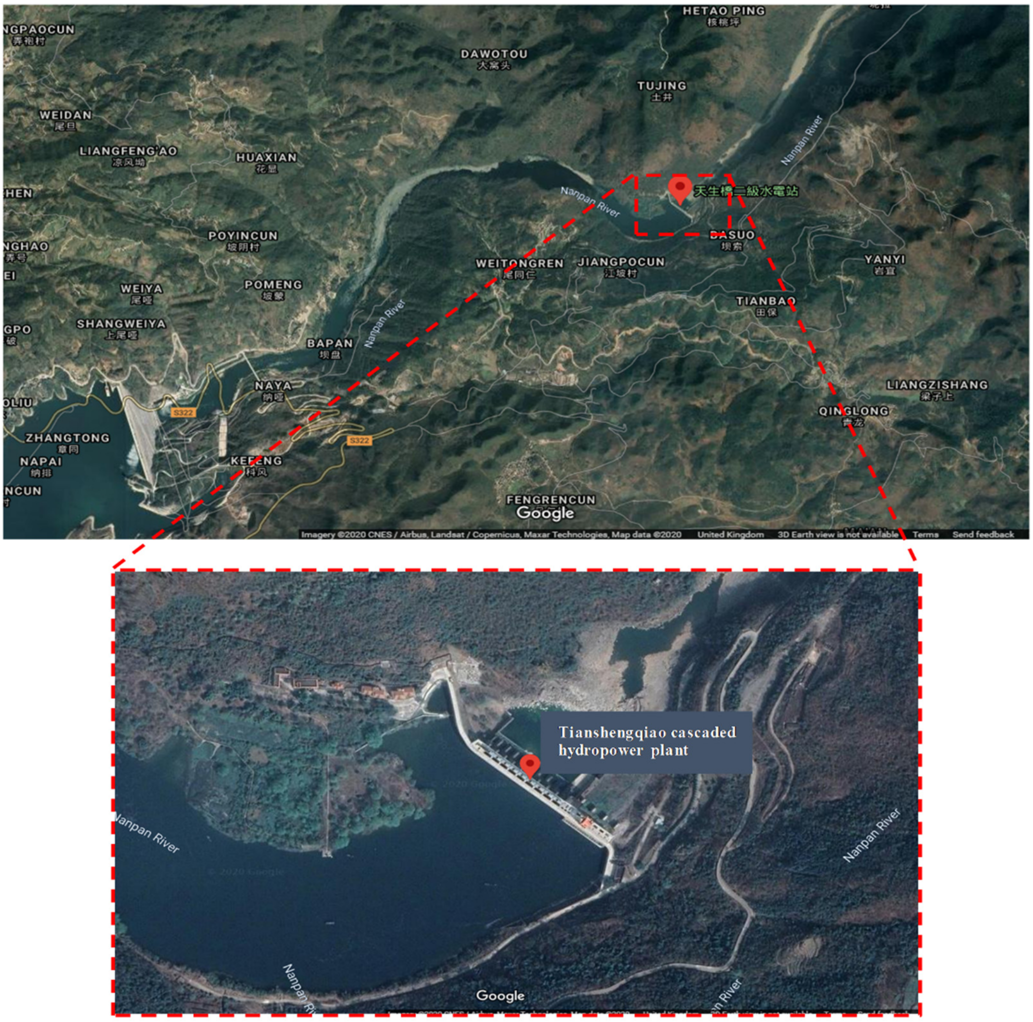


Figure 1. Satellite image of the Tianshengqiao cascaded hydropower plants (by Google maps).

2.2. Methodology

The operation of a hydropower system is a typical engineering challenge. For cascaded hydropower plants, the workability of the feasible solutions is more essential than optimality in mathematics; therefore, for making the viable generation schedules, practicability of solution is very important. Furthermore, mathematical modelling is not able to satisfy the complex conditions and constraints, which results in lowering the feasibility of the model. Therefore, it is important to find quick and feasible generation schedules by using the data mining decision tree algorithms. Decision tree algorithms require less effort for data preparation, and they are good for predicting the continuous values. It gives quick generation values as scaling of data is not required in decision tree algorithms. The limitation of this work is that small fluctuation in data may alter the feasible decision tree structure which is not suitable for long-term hydro scheduling because of the large number of variables. It also does not satisfy the security and environmental constraints. The multiple regression analysis is used to find the energy production. The generation curves are determined by K-mean cluster analysis. For short-term generation scheduling of cascade hydropower plants, the decision trees, such as C4.5, improved C4.5, ID3-IV and CHAID algorithms, are used. The algorithms are solved in the data mining tool, SIPINA. The comparison between different algorithms is also discussed in detail. The overall framework of the methodology is shown in Figure 2.

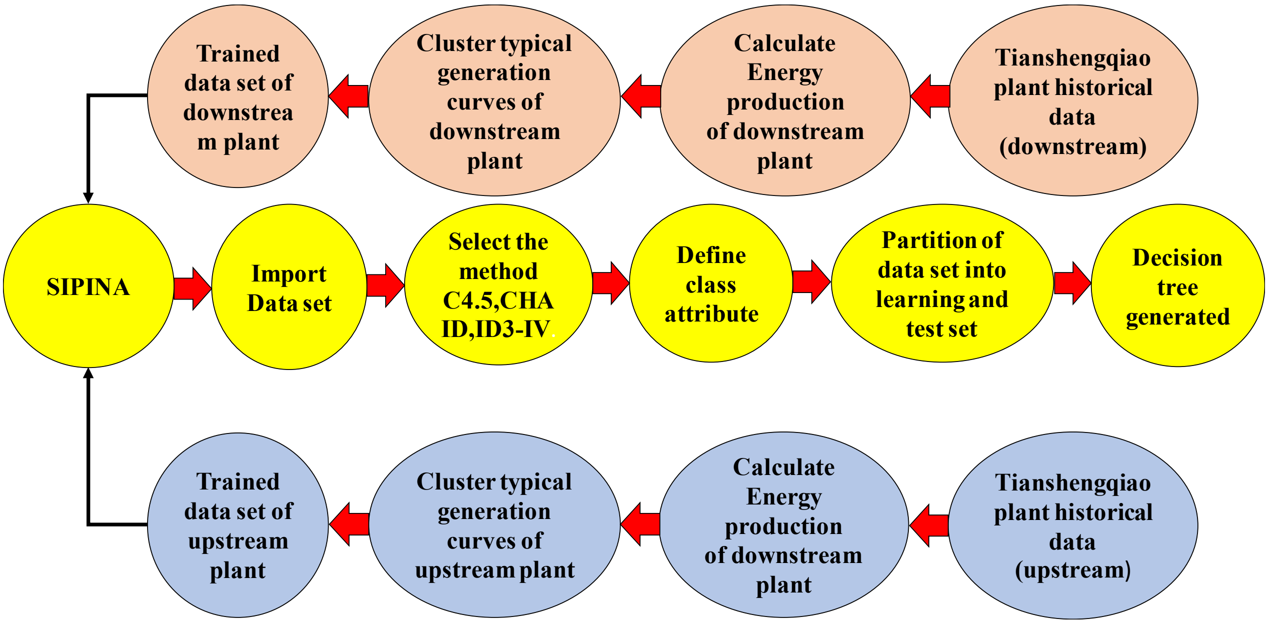

Trained data set of downstream plant
Cluster typical generation curves of downstream plant
Calculate Energy production of downstream plant
Tianshengqiao plant historical data (downstream)
SIPINA
Import Data set
Select the method C4.5,CHAID,ID3-IV
Define class attribute
Partition of data set into learning and test set
Decision tree generated
Trained data set of upstream plant
Cluster typical generation curves of upstream plant
Calculate Energy production of downstream plant
Tianshengqiao plant historical data (upstream)

Figure 2. Overall framework.

2.2.1. Proposed Equations for Energy Production

The following steps are followed for multiple regression analysis and K-mean cluster analysis:

Energy production of upstream and downstream hydropower plant is determined. Water level, discharge and power generation in each hour are used to find the energy production of the upstream and downstream hydropower plants, and multiple regression analysis is used which results in Equation (1).

$$y \;=\; G_0{+}G_1{*}Z_{\text{up}} \;+G_3*\text{N} \tag{1}$$

where y is the water discharge, Zup is the water level and N is the power generation of each hour.

2.2.2. K-Mean Cluster Analysis

The cluster analysis method is used to find the typical generation curves. The K-mean clustering algorithm is used to cluster the generation curve. Firstly, the daily historical data of 24 h for two years are taken and then the generation curves are obtained by using Equations (2) and (3).

$$X_t \;=\; P_t/P_{max} \tag{2}$$

$$\text{X} \;=\; (X_{1,}X_2\,,X_3\,\ldots X_n) \tag{3}$$

where X_t is the eigenvector vector of the generation curve in time t; P_t is the power value of the generation curve in time t; P_{max} is the maximum generation of the hydropower plant within one day; X gives a set of observations spanning up to "n" observations.

In the next step, the "k" number of generation curves is selected randomly. Each sample shows the initial cluster center and remaining samples of the generation curve are assigned to the most similar cluster based on its distance to each cluster center. Therefore, for each cluster, a new center is required to recalculate. For this, an iteration is required until a square error criterion function shown in Equation (4) converges:

$$W \;= \sum_{i=1}^{k} \sum_{a_{k \epsilon C_i}} (a_k - b_i)^2 \tag{4}$$

where W is the sum of the square errors of all generation curves in the historical data set; a_k is the generation curve of a group of cluster vectors; b_i represents a cluster vector, which is the center of cluster C_i.

2.2.3. C4.5 Algorithm

The first step is to find a set of sample "F". The entire sample set consists of water discharge, initial water level, inflow, demand, and energy production.

Then, take the generation schedule as a category attribute, and count the typical generation schedules that include various types of generation schedules and calculate the information expectation of samples using Equation (5).

$$Info(F) = -\sum_{i=1}^{n} q_i log_2(q_i) \tag{5}$$

Whereas n is the total number of generation schedules; q_i is the rate of samples with a generation schedule number and i is the number of samples.

In the next step, the information gain rate of non-category attributes is calculated. If a non-category attribute B has the values y, correspondingly, the unit generation schedules are divided into "y" categories. The information expectation and gain of attribute B for each non-category attribute are calculated by using Equations (6) and (7), respectively. Equations (8) and (9) are used for the information gain rate of attribute B.

Sample information expectation of attribute B:

$$Info_B(F) = \sum_{k=1}^{y} \frac{|F_k|}{|F|} \times Info(F) \tag{6}$$

Information gain of attribute B:

$$Gain\ (B) \ = \ Info(F) \ - \ Info_B(F) \tag{7}$$

Information gain rate of attribute B:

$$Split\ Info(B) = -\sum_{k=1}^{y} \frac{|F_k|}{|F|} \times log_2\left(\frac{|F_k|}{|F|}\right) \tag{8}$$

$$Gain\ Ratio(B) \ = \ \frac{Gain\ (B)}{SplitInfo(B)} \tag{9}$$

Whereas, F is the sample set; F_k is the number of samples of the kth element included in attribute B.

2.2.4. Chi-Squared Automatic Interaction Detection (CHAID) Algorithm

CHAID can be used for prediction as well as classification, and for detection of interaction between variables [60]. CHAID can only work with categorical data. This algorithm uses the Chi-Squared test between response or target variable (dependent) and input variables (independent). Based on the result of the Chi–Square test, it chooses the best splitting variable from the input variables. There are three steps of this algorithm, i.e., merging, splitting and stopping.

- Merging
  1. The merging finds the best split predictor. The non-significant categories are merged for every predictor variable "S". If "S" is used to split the node, every final category of "S" will give one child node as outcome. The adjusted value of p is calculated in the merging step that is further used in splitting.
  2. If X has only one category, then the process should be stopped, and the p-value needs to be adjusted as 1.

3. If X has two categories, then go to step vii; otherwise, determine the pair of S that is the most similar. The most similar pair is the pair that gives the largest value with respect to dependent variable W.
4. If the largest *p*-value of the pair is greater than alpha-level merge, which specifies the user, then this pair is substituted in a single compound l and the new set of S category is formed. If its *p*-value is not greater than the alpha-level merge, then merge any category with a smaller number of observations with the most similar category that is measured based on the largest *p*-value and follow step vi.
5. If the compound category consists of three or more original categories based on the smallest *p*-value within the compound category, then find the best binary split.
6. If the best binary split is not obtained, then again go to step ii.
7. Merge the category with the smaller number of observations with the most similar category, which is measured on the basis of the largest *p*-value.
8. Finally, use Bonferroni adjustments to compute the adjusted *p*-value.

- Splitting

  The merging step contains the best split for each predictor. The predictor is chosen in the splitting step for selecting the best split node. From the merging step, the *p*-value is obtained, and this value is then compared with each predictor. Further, the predictor with the smaller *p*-value is selected. If the adjusted *p*-value is less than or equal to the user-specified alpha-level, then split the node; if not, then do not split the node and this node is considered as a terminal node.

- Stopping

  1. In the stopping step, there are some rules that check if the tree growing process should be stopped or not according to the following rules:
  2. In the node, if all the values of the dependent variables are the same, then it will not split.
  3. If the values of each predictor are the same, then the node will not split.
  4. If the depth of the current tree is more than the depth of the user-specified tree, then the tree should stop growing.
  5. If the size of the node is less than the user-specified minimum node size, then the node will not split.
  6. If the number of the child node is 1, then also the node will not split.

2.2.5. ID3-IV Algorithm

The splitting criterion of ID3-IV depends on the information gain process [61]. The root node is the topmost decision node, which is also called the predictor. The split attribute is the one that has the highest information gain. From training instances, the tree is created by using information gain. The growing of the tree stops when the information gain is zero.

1. First, select the calculation of entropy attribute and target attribute.
2. Then, the attribute with the highest information gain is measured.
3. The node is created by using that attribute. The above steps are applied iteratively to new branches of the tree and, after satisfying the stopping criterion, the growth of the tree needs to be stopped.

2.2.6. Improved C4.5

The C4.5 algorithm is improved by using the principal component analysis technique. The key idea behind principal component analysis is to choose the variables in the data set for generating the C4.5 decision tree and, for each node, select several attributes. The stepwise procedure of improved C4.5 is given below:

1. First, the initialization of the data sample set is required.
2. Then, the simplification of the data set's principal component analysis is performed.
3. For each principal component, calculate the information gain rate.

4. Splitting node is selected based on the largest information gain rate of the principal component and generates the subset of the data.
5. Repeat step iii and step iv till all the components of the decision tree are utilized.
6. Pruning is carried out to generate the decision tree.

2.3. Data Set

Sample sets for decision tree algorithms of Tianshengqiao plant 1 for the winter season are shown in Table 1. The table consists of different parameters such as water level, discharge, energy production, inflow and generation schedule. The percentage (%) of energy production is found by energy production (in MW) divided by load (in MW). The percentage (%) of energy production is found by energy production (in MW) divided by load (in MW). In Table 1, it is observed that the percentage (%) of energy production varies from 78% to 208%, being the lowest at T15 and highest at T17. The average value is 125.88% with a standard deviation of 26.03. The time frame of scheduling is 24 h (T1 to T24). Five schedules are taken as class in the data set. In the winter season, the level of water is almost constant, having a value of around 771 m. The discharge of water is low in the winter season because a large part of the winter rainfall accumulates as snow and, thus, influences the water level of the river in the summer season. During eleven (11) different hours in a day, the discharge goes beyond 500 m^3/s and does not move below 256 m^3/s. The generation schedules are classified because of energy production and discharge. Target schedules are divided into 15 steps. It is shown in Table 1 that more energy production means mostly more discharge. In Table 2, the sample set of Tianshengqiao plant 1 for the summer season is presented. In Table 2, it is observed that the percentage (%) of energy production varies from 86% to 127%, being the lowest at T10 and highest at T12. The average value is 101% with a standard deviation of 7.54. Various category attributes such as discharge, water level, energy production, inflow, load, and class generation schedule are presented in Table 2. The data are collected for 24-hours' time span. Six schedules are proposed for hydropower generation scheduling. The discharge in summer is greater, having highest discharge of nearly 1049m^3/s. Similar to discharge, the load is also high during the summer season because the gap between the outside and inside temperature is larger and air conditioner is used to keep the temperature low for cooling purposes. Besides this, the inflow in summer is greater as compared to that in the winter season. The generation schedules are made since percentage of energy production is executed in twenty steps. The percentage of energy is obtained by taking the load and energy production variables.

The sample set of Tianshengqiao plant 2 for the winter season is demonstrated in Table 3. The table consists of different attributes, such as water level, discharge, energy production, inflow, load and class, such as generation schedule. The time frame of scheduling is 24 h; the total number of schedules is seven. In the winter season, the level of water is almost constant, having a value of around 642 m. The inflow level and discharge of water are low in the winter season because a large part of the winter rainfall gathers as snow and the water influences the river in the summer season when released. The maximum discharge is around 500 m^3/s during different hours of the day and the minimum discharge ranges around 300 m^3/s. The ratio of energy production and maximum discharge is the main parameter of the decision tree for making the generation schedule. The schedules are predicted because of all the attributes in the winter season. It is shown in Table 3 that more energy production resulted in more value of discharge. Table 4 illustrates the sample set of Tianshengqiao plant 2 for the summer season with various category attributes. The number of schedules is seven and the data are collected for a period of 24 hours. The discharge in summer is greater, which ultimately increases the level of energy production. The load is more in the summer season because everyone is using air conditioner for cooling purposes, which ultimately increases the load on the power stations. The ratio of energy production and maximum discharge plays an important role in generation scheduling.

Table 1. Data set of Tianshengqiao plant 1 for the winter season.

Time (h)	Inflow (m^3/s)	Discharge (m^3/s)	Load (MW)	Water Level (m)	Energy Production (MW)	Percentage Energy Production	Target (Schedule)
T1	377	509	589	771.51	707	120	Schedule 6
T2	312	508	587	771.49	708	121	Schedule 6
T3	415	315	404	771.50	435	108	Schedule 4
T4	295	259	301	771.50	352	117	Schedule 3
T5	112	258	302	771.48	351	116	Schedule 3
T6	140	256	299	771.48	354	118	Schedule 3
T7	178	317	302	771.48	440	146	Schedule 4
T8	248	508	564	771.48	714	127	Schedule 6
T9	140	507	591	771.47	713	121	Schedule 6
T10	119	554	587	771.46	783	133	Schedule 7
T11	308	551	590	771.46	780	132	Schedule 7
T12	351	512	403	771.46	726	180	Schedule 6
T13	280	314	298	771.46	443	149	Schedule 4
T14	168	257	301	771.46	364	121	Schedule 3
T15	338	258	470	771.45	368	78	Schedule 3
T16	317	432	600	771.46	615	103	Schedule 5
T17	377	487	331	771.46	689	208	Schedule 6
T18	334	256	302	771.46	359	119	Schedule 3
T19	407	332	532	771.46	469	88	Schedule 4
T20	459	566	592	771.46	802	135	Schedule 7
T21	431	513	596	771.45	730	122	Schedule 6
T22	515	512	594	771.46	726	122	Schedule 6
T23	378	512	596	771.46	725	122	Schedule 6
T24	416	484	594	771.46	685	115	Schedule 6

Table 2. Data set of Tianshengqiao plant 1 for the summer season.

Time (h)	Inflow (m^3/s)	Discharge (m^3/s)	Load (MW)	Water Level (m)	Energy Production (MW)	Percentage Energy Production	Target (Schedule)
T1	750	777	733	746.92	777	106	Schedule 10
T2	833	748	727	746.91	744	102	Schedule 9
T3	743	528	539	746.91	524	97	Schedule 7
T4	698	530	497	746.90	522	105	Schedule 7
T5	835	530	499	746.90	522	105	Schedule 7
T6	561	530	500	746.92	530	106	Schedule 8
T7	475	530	505	746.91	526	104	Schedule 8
T8	439	531	578	746.89	518	90	Schedule 7
T9	509	679	730	746.89	667	91	Schedule 8
T10	559	828	951	746.89	816	86	Schedule 10
T11	417	1049	973	746.89	1037	107	Schedule 13
T12	488	957	736	746.86	932	127	Schedule 12
T13	576	784	743	746.85	755	102	Schedule 10
T14	573	783	740	746.84	750	101	Schedule 10
T15	544	782	740	746.83	745	101	Schedule 9
T16	445	783	741	746.82	741	100	Schedule 9
T17	538	783	741	746.80	733	99	Schedule 9
T18	737	783	742	746.80	733	99	Schedule 9
T19	737	783	746	746.84	750	101	Schedule 10
T20	647	790	749	746.83	753	101	Schedule 10
T21	740	790	749	746.80	741	99	Schedule 9
T22	878	790	749	746.81	745	99	Schedule 9
T23	879	791	749	746.82	749	100	Schedule 10
T24	647	760	749	746.82	718	96	Schedule 9

Table 3. Data set of Tianshengqiao plant 2 for the winter season.

Time (h)	Inflow (m^3/s)	Discharge (m^3/s)	Load (MW)	Water Level (m)	Energy Production (MW)	Percentage Energy Production	Target (Schedule)
T1	252	151	220	642.42	251	114	Schedule 3
T2	174	162	221	642.82	255	115	Schedule 3
T3	67	73	92	642.86	95	103	Schedule 1
T4	67	58	64	642.84	69	108	Schedule 1
T5	67	55	64	642.87	62	97	Schedule 1
T6	67	64	64	642.91	77	120	Schedule 1
T7	67	126	115	642.92	187	163	Schedule 2
T8	222	305	466	642.72	514	110	Schedule 6
T9	497	378	650	642.44	655	101	Schedule 7
T10	508	416	652	642.84	708	109	Schedule 8
T11	509	402	650	643.14	672	103	Schedule 8
T12	512	405	651	643.48	665	102	Schedule 7
T13	303	447	647	643.82	727	112	Schedule 8
T14	260	396	648	643.36	653	101	Schedule 7
T15	260	390	657	642.92	659	100	Schedule 7
T16	257	385	659	642.48	666	101	Schedule 7
T17	491	387	659	642.05	685	104	Schedule 8
T18	510	411	661	642.40	715	108	Schedule 8
T19	514	412	662	642.73	705	106	Schedule 8
T20	414	441	660	643.07	744	113	Schedule 8
T21	253	395	658	642.98	665	101	Schedule 7
T22	252	401	657	642.50	694	106	Schedule 8
T23	252	312	581	642.00	553	95	Schedule 6
T24	252	200	251	641.79	361	144	Schedule 4

Table 4. Data set of Tianshengqiao plant 2 for the summer season.

Time (h)	Inflow (m^3/s)	Discharge (m^3/s)	Load (MW)	Water Level (m)	Energy Production (MW)	Percentage Energy Production	Target (Schedule)
T1	573	479	865	639.91	812	94	Schedule 8
T2	323	296	488	640.26	507	104	Schedule 5
T3	293	263	439	640.36	453	103	Schedule 4
T4	293	258	439	640.47	444	101	Schedule 4
T5	293	260	439	640.60	448	102	Schedule 4
T6	292	262	439	640.72	452	103	Schedule 4
T7	292	259	440	640.83	447	110	Schedule 4
T8	292	281	440	640.95	484	120	Schedule 5
T9	526	474	673	640.99	806	100	Schedule 8
T10	773	626	1060	641.17	1059	104	Schedule 10
T11	797	677	1103	641.68	1146	109	Schedule 11
T12	776	713	1104	642.09	1206	98	Schedule 11
T13	512	500	872	642.30	852	107	Schedule 8
T14	287	414	662	642.34	708	98	Schedule 7
T15	289	381	663	641.91	653	94	Schedule 6
T16	431	365	663	641.59	625	102	Schedule 6
T17	449	394	664	641.82	674	101	Schedule 6
T18	449	392	665	642.01	671	102	Schedule 6
T19	450	391	666	642.20	669	102	Schedule 6
T20	461	393	667	642.40	673	102	Schedule 6
T21	469	394	665	642.63	676	102	Schedule 6
T22	466	392	661	642.88	673	102	Schedule 6
T23	466	391	658	643.12	670	101	Schedule 6
T24	469	385	660	643.36	661	100	Schedule 6

3. Results and Analysis

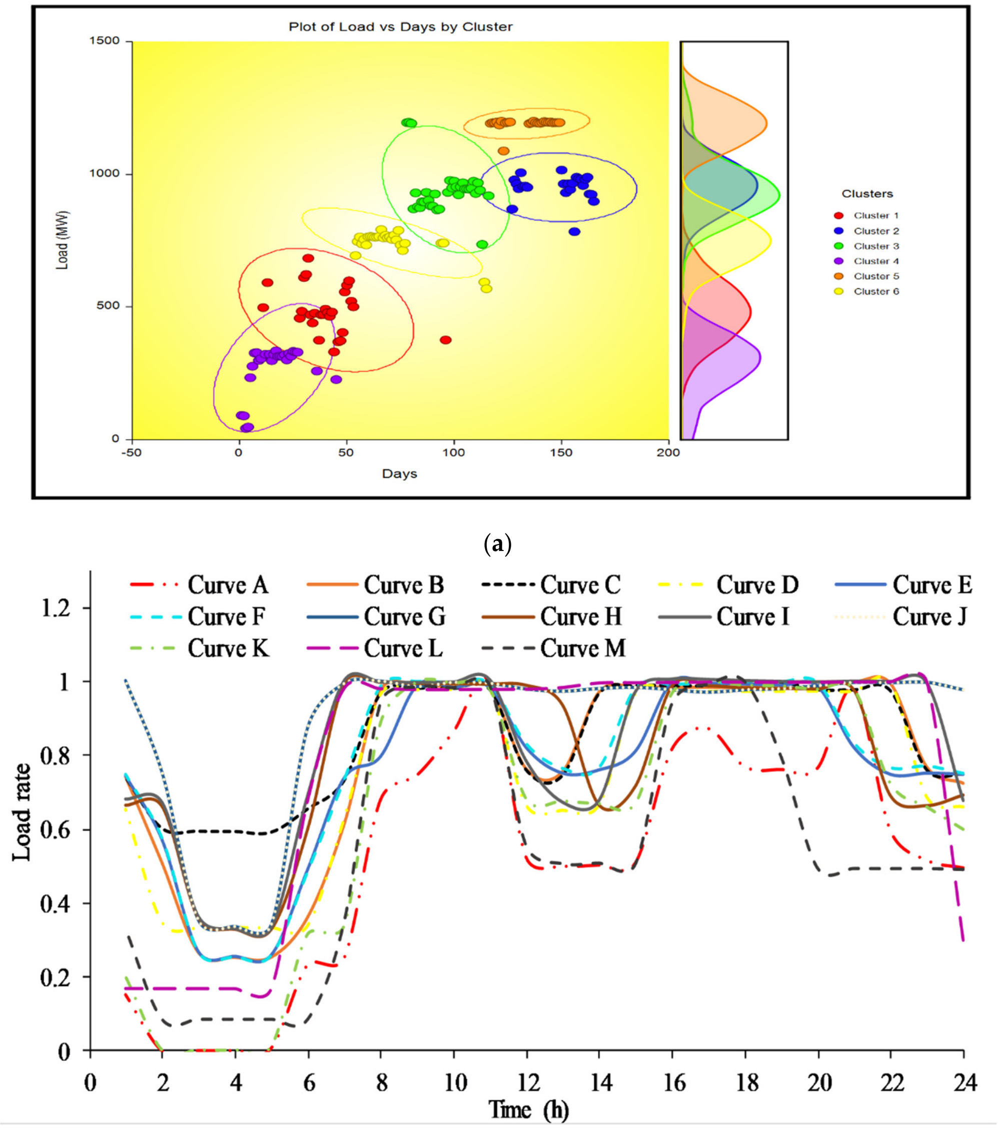
3.1. K-Mean Cluster Analysis and Typical Generation Profiles

The cluster analysis of Tianshengqiao plant 1 for the winter season and the summer season is shown in Figure 3a. The results show that the demand in the summer season was higher than that of the winter season as is evident from Figure 3a. In addition, the demand, with respect to time, varies continuously after every two to three hours in the winter season while the demand with respect to time was almost constant except for peak hours in the summer season. The low electricity demand was observed due to less use of electricity in the winter season. On the other hand, the excessive use of electricity in the summer season was the basis for higher electricity demand. Moreover, during the winter season, the maximum demand was 600 MW and the minimum was 300 MW. In contrast, for the summer season, the demand of electricity was more because more electricity was used for cooling purposes in the summer season, and, during the peak hours, it reached almost to 1000 MW. Typical generation curves are shown in Figure 3b. The cluster analysis curve of Tianshengqiao plant 2 for the winter and summer seasons is shown in Figure 4a and typical generation curves are shown in Figure 4b, respectively. The outcome of the analysis showed that the electricity demand was greater in the summer season as compared to that in the winter season. The demand of electricity almost remains continual for both the summer and winter seasons. Additionally, the load of electricity in winter is nearly 700 MW as compared to that in the summer season, which is almost 1300 MW. In the summer season, temperature and severe heat events increase, so the trend of consuming energy changes from fuel oil, natural gas to the use of hydropower and become more dominant for production of electricity because of more availability of water.

3.2. C4.5 Outcomes

Figure 5 shows a decision tree generated by using the algorithm C4.5. The tree has seven nodes, four leaves and a maximum depth of 3. The node of each tree represents the analysis of each class. Moreover, it also shows the number of collected samples and percentage of confidence. The features and threshold values are also shown in the decision tree. Threshold values are used for the separation between two classes, and features are responsible for determining the best values. The generation schedule is illustrated in the tree. The rule of the tree describes that if the discharge is <532 m^3/s, the most suitable schedule 6 is selected having the confidence index of 100%. When the discharge is $\geq$532 m^3/s, then the decision tree with confidence index of 100% suggests schedule 7. The quality of the rule is influenced by the percentage of its confidence index. Moreover, schedule 4 and schedule 5 are suitable when discharge is $\geq$286. Schedule 7 is suggested when the value of discharge is $\geq$532m^3/s. Figure S1 shows a decision tree developed by using the algorithm C4.5 for the summer season of Tianshengqiao plant 1. The node of each tree shows the class. Moreover, it also presents the number of collected samples and percentage of confidence. The features and threshold values are also displayed in the decision tree. The objective of the threshold values is used for the separation between two classes, and the features are responsible for determining the best values. The generation schedule is illustrated in the tree. If the value of discharge is <768 m^3/s, water level is <746 m and inflow is <551 m^3/s, then schedule 9 is selected, having a 100% confidence index. Moreover, other schedules are ignored for this rule because of the null value of their confidence indices. Furthermore, the best rule for schedule 7 is when level of inflow is $\geq$629m^3/s; if the discharge is $\geq$892 m^3/s, then schedule 12 and schedule 13 are recommended. Figure S2 shows a decision tree generated using C4.5 algorithms for Tianshengqiao plant 2 during the winter season. The generation schedule for a 24 h time period is illustrated. The tree has nine nodes, five leaves and a maximum depth of 4. Moreover, it also indicates the number of collected samples and percentage of confidence. The features and threshold values are also revealed in the decision tree. Further, threshold values are used for the separation between two classes, and features are responsible for determining the best values. The best rule for schedule 3 is proposed when discharge

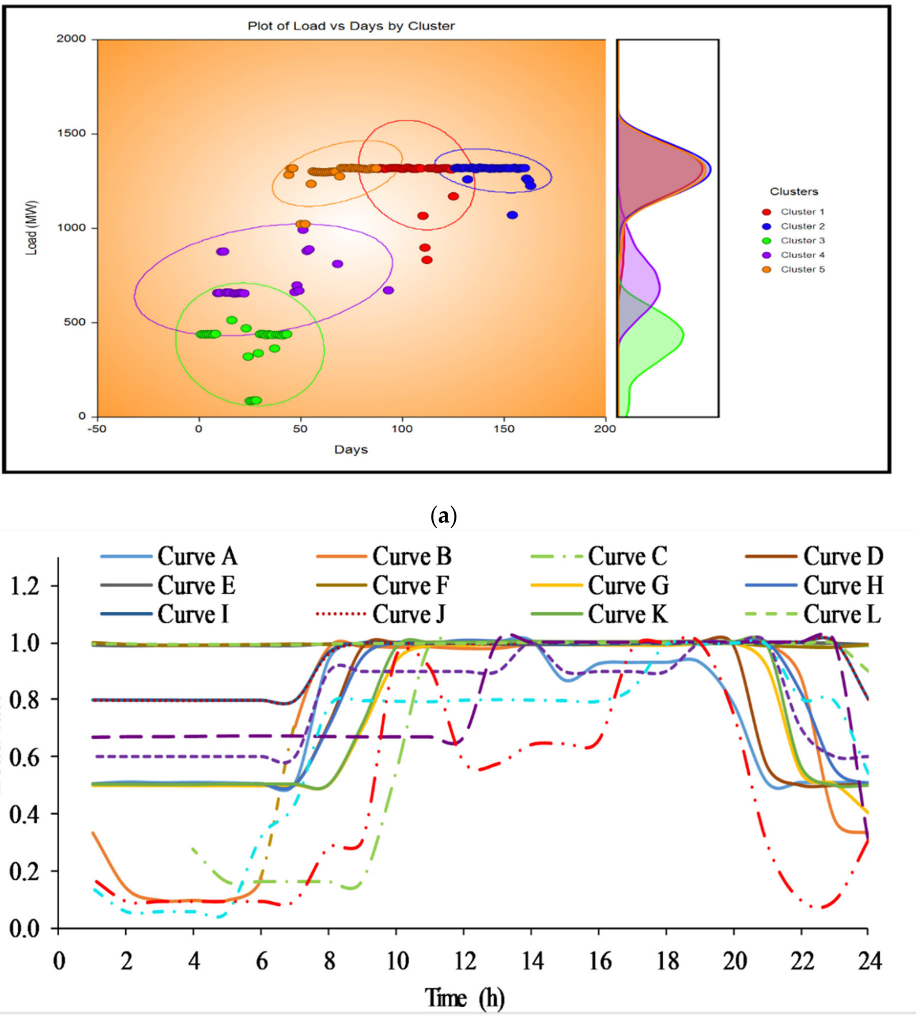
<181 m^3/s; schedule 1 and schedule 2 are suggested when inflow is <120.5 m^3/s based on their confidence indices. When the level of discharge is <398 m^3/s, then schedule 7 is recommended, and schedule 6 is suitable when the inflow is ≥120.5 m^3/s. The decision tree obtained by the algorithm C4.5 for Tianshengqiao 2 of summer season is presented in Figure 6. The 24-h generation schedule is illustrated in the decision tree. The tree has nine nodes, five leaves and a maximum depth of 4. In the decision tree, the node of each tree represents the analysis of each class. Furthermore, it also shows the number of collected samples and percentage of confidence, features and threshold. Threshold values are used for the separation between two classes and features are responsible for determining the best values. According to the rule, when the inflow ≥673 m^3/s, then schedule 11 is most suitable with the confidence index of 67%. When the discharge is <404 m^3/s, then the decision tree obtaining confidence index of 100% suggests schedule 6. The quality of the rule is influenced by the percentage of its confidence index. When the discharge <330 m^3/s, then schedule 4 is suggested. When the inflow is <673 m, schedule 8 can be adopted, having a 75% confidence index.



(**a**)

(**b**)

Figure 3. Tianshengqiao plant 1: (**a**) cluster analysis; (**b**) typical generation curves.



(**a**)

(**b**)

Figure 4. Tianshengqiao plant 2: (**a**) cluster analysis; (**b**) typical generation curves.

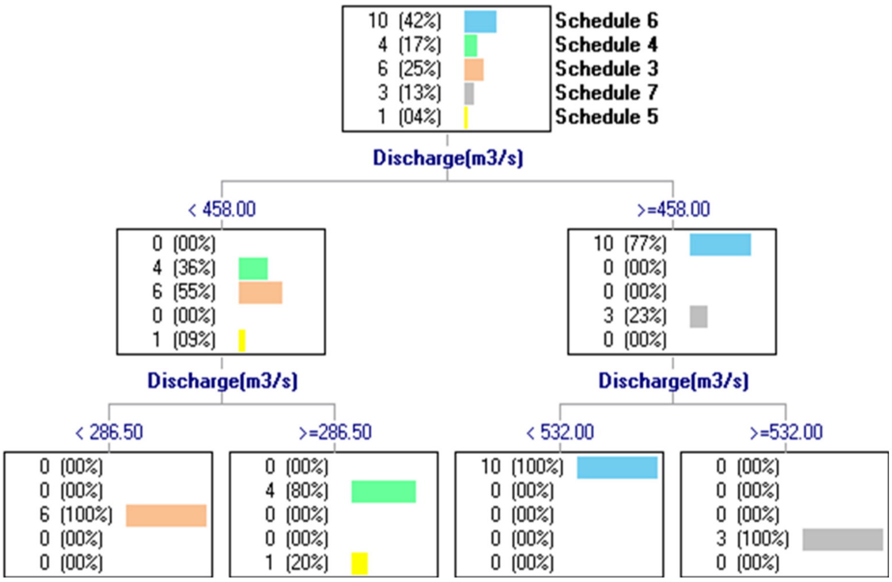


Figure 5. Decision tree of Tianshengqiao plant 1 generated by C4.5 for the winter season.

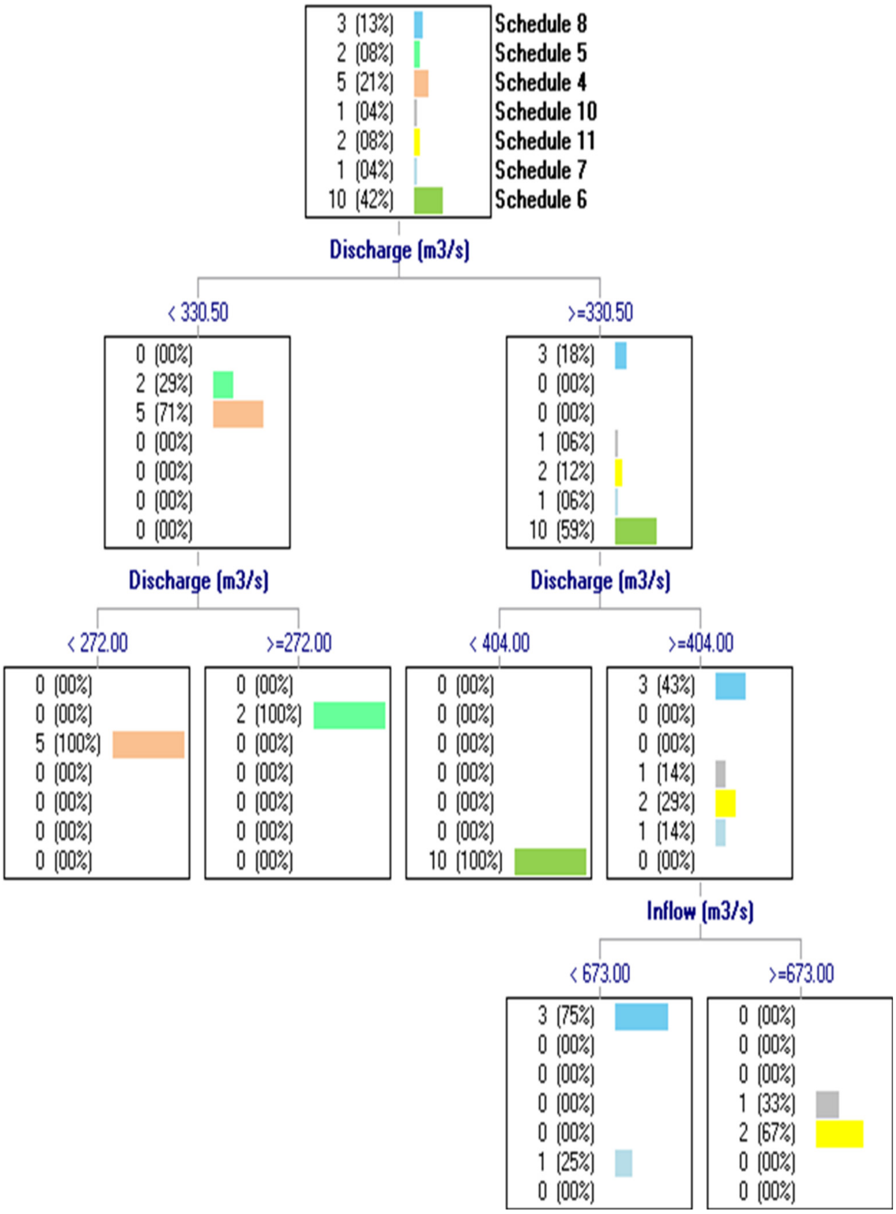


Figure 6. Decision tree of Tianshengqiao plant 2 generated by C4.5 for the summer season.

3.3. CHAID Findings

The decision tree established by the CHAID algorithm for Tianshengqiao 1 of the winter season is demonstrated in Figure S3. Five schedules are determined for a period of 24 h. From 2 a.m. to 8 a.m., and in night hours, schedule 6 is desirable to consider with a 100% confidence index. In addition, schedule 4 is the best suited from 7 a.m. to 11 a.m. and some hours of the afternoon, having a confidence index of 50%. Schedule 3 is suitable in the early morning hours, schedule 5 is ignored because of its low confidence index. In the decision tree, attribute time presents the threshold values. The decision tree generated by the CHAID algorithm for Tianshengqiao 1 of the summer season is shown (refer Figure 7). The tree presents different timings suitable for six schedules. In the morning until 9 a.m., schedule 10 is considered as a suitable schedule and the same schedule is appropriate for the 19th and 20th hours of the night. For 2 a.m., in the afternoon hours and two night hours, schedule 9, having a 100% confidence index, needs to be followed. Furthermore, schedule 7 is the best-suited schedule from the 3rd hour to the 11th hour with a confidence index of 44%. Schedule 12 and schedule 13 have low confidence indices so they can be ignored during the peak load hours. Figure 8 exhibits the decision tree developed by the

CHAID algorithm for Tianshengqiao 2 for the winter season. Four schedules are shaped based on the data set.

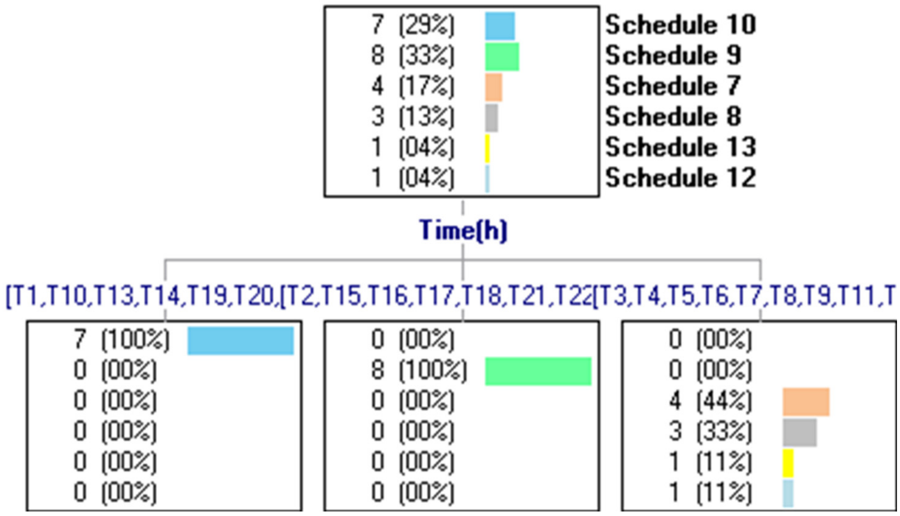


Figure 7. Decision tree generated by Chi-Squared Automatic Interaction Detection (CHAID) Tianshengqiao plant 1 for the summer season.

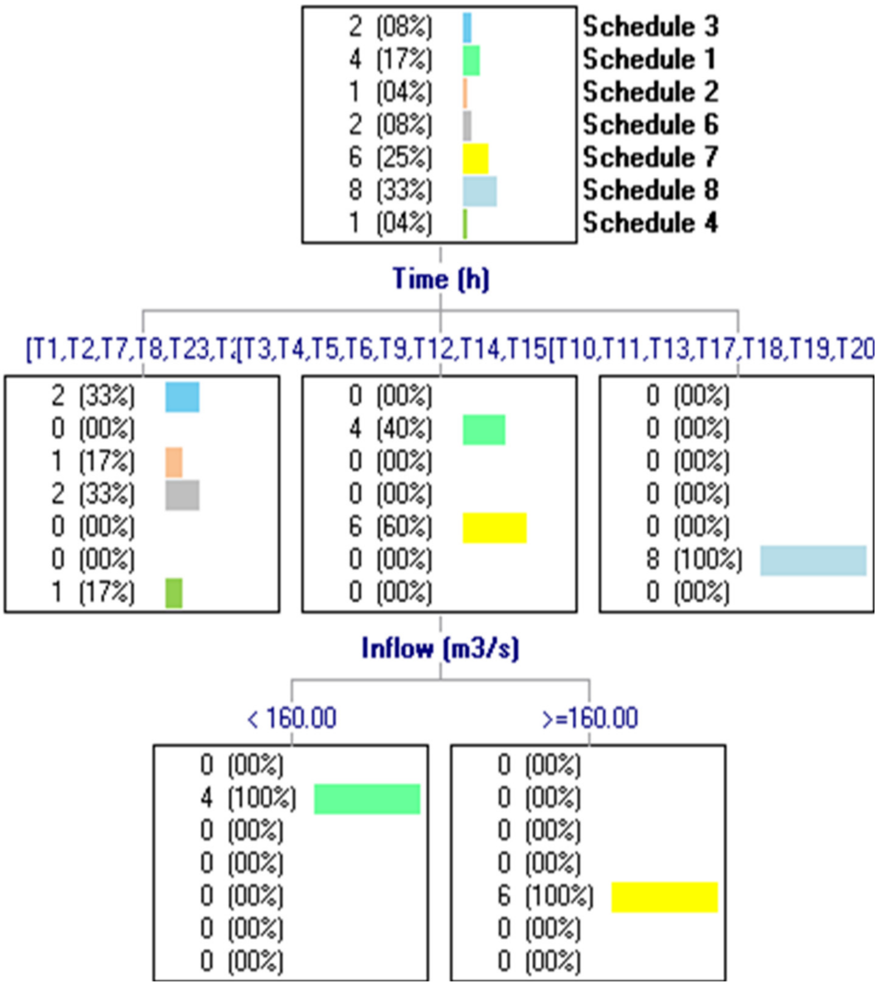


Figure 8. Decision tree of Tianshengqiao 2 plant generated by CHAID for the winter season.

The period covers 24 h. When the flow is <160 m^3/s, schedule 1 with a 100% confidence index is appropriate; schedule 7, having a confidence index of 100%, can be considered as optimal when the inflow is ≥160 m^3/s. Furthermore, schedule 8 is feasible in the evening time. The decision tree generated by the CHAID algorithm for the Tianshengqiao 2 plant in the summer season is presented (see Figure S4). The schedules are determined for 24 h and seven schedules can be followed in the morning, afternoon, evening and night. Schedule 6 is suitable for the evening and in the nighttime; schedule 4 is recommended for the morning. Furthermore, schedule 8, schedule 5, schedule 11, and schedule 7 are ignored due to low confidence indices.

3.4. ID3-IV Results

Figure 9 shows that during the winter season schedule 6 is best suited to follow with a confidence index of 100% when the value of discharge is <532m^3/s; schedule 7 is suggested when the discharge is ≥ 532 m^3/s. When the discharge value is <458, schedule 4 is suitable, while the rest of the schedules can be ignored when the load is higher due to low percentage of confidence index. The decision tree of the Tianshengqiao 1 plant for the summer season generated by ID3 is presented in Figure S5. The figure clearly shows the daily schedules for power generation. When the discharge is ≥892 m^3/s, schedule 13 and schedule 12 are appropriate to select. The rest of the schedules should not be considered according to the rule of discharge. When the value of discharge is <713.5 m^3/s, schedule 7 is selected with a 57% confidence index. Schedule 10 and schedule 9 are suitable when discharge is ≥713 m^3/s. S6 illustrates the decision tree of Tianshengqiao 2 plant for the winter season developed by ID3. The daily schedule of units for power generation is indicated. When the discharge is <398 m^3/s, then schedule 2, schedule 4, schedule 6, and schedule 3 are carefully chosen as per demand. The rest of the schedules for the previous rule can be overlooked because of their nil value. When the value of discharge is <398m^3/s, then schedule 7 is followed. Schedule 8 is suitable when discharge is ≥398m^3/s.

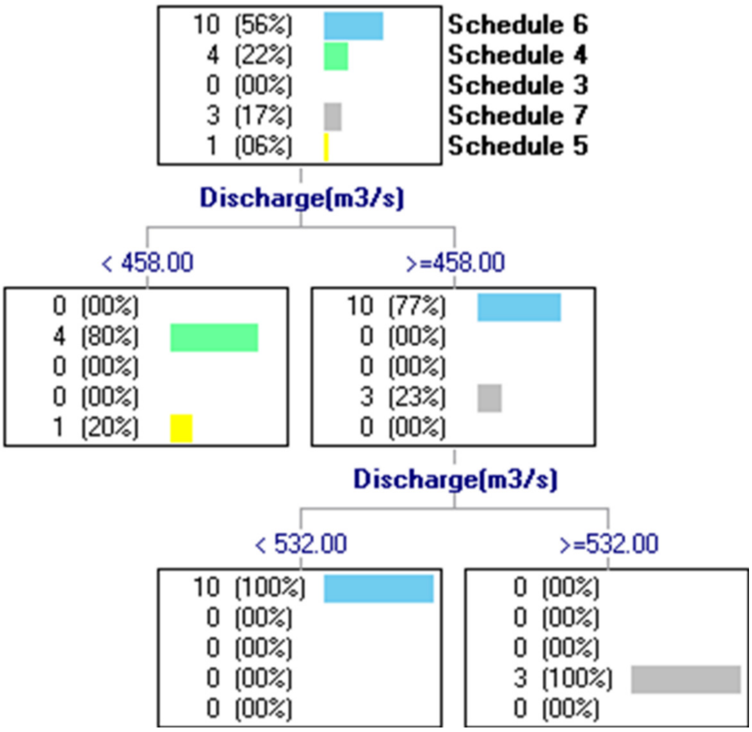


Figure 9. Decision tree by using the Iterative Dichotomiser 3 (ID3) algorithm Tianshengqiao plant 1 for the winter season.

The decision tree of Tianshengqiao plant 2 for the summer season gives information about the generation schedule of 24 h, as shown in Figure 10. The best rule for schedule 4 is suitable when discharge is <272 m^3/s; schedule 5 is appropriate when the discharge is ≥272 m^3/s. Schedule 6 is well fitted when the discharge is ≥404 m^3/s, and schedule 5 is appropriate when the discharge is ≥272 m^3/s. Schedule 6 is well fitted when the discharge is ≥404 m^3/s. Schedule 7, schedule 11 and schedule 10 are ignored, having low confidence. Different rules are suitable for different schedules according to the desirable requirement. It is not necessary that for one rule (discharge <272 m^3/s) all the schedules are recommended. Therefore, according to the required condition, the different schedules are considered, as schedule 4 with a 77% confidence index can be used and schedule 5, having a 29% confidence index, can be ignored at the same time as per the load requirement.

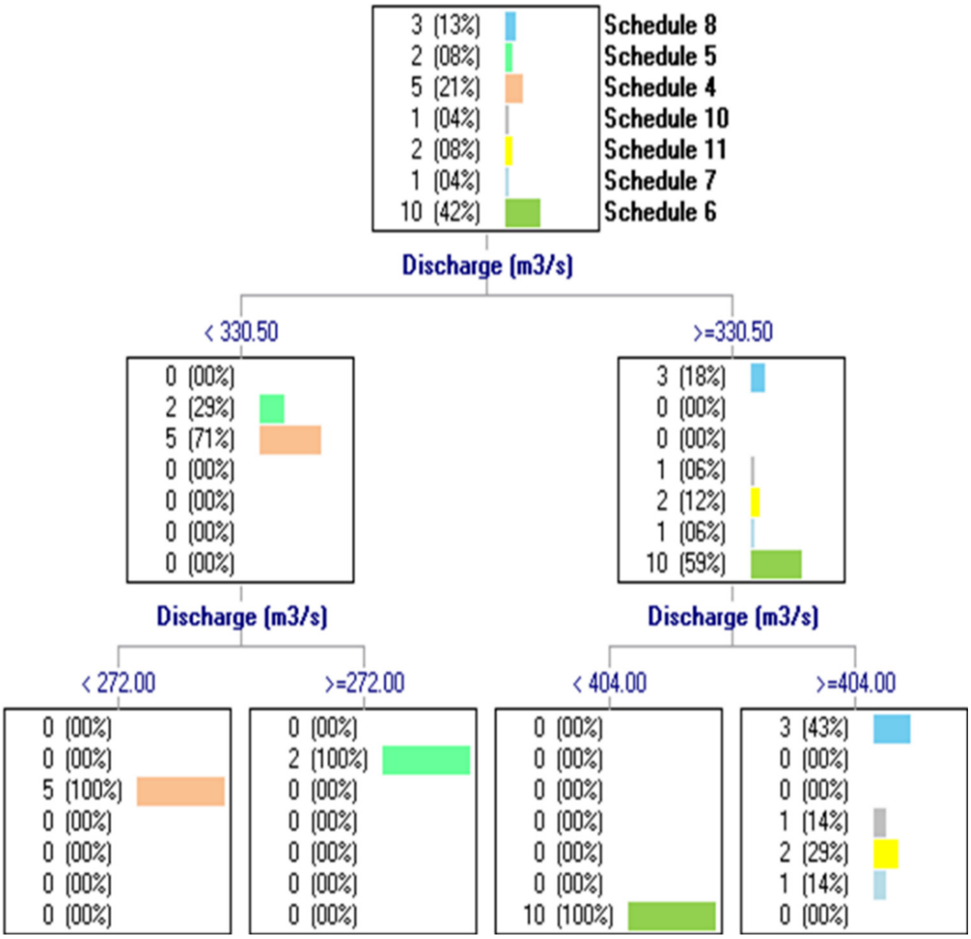


Figure 10. Decision tree by using the ID3 algorithm Tianshengqiao plant 2 for the summer season.

3.5. Improved C4.5

Figure S7 represents the decision tree generated by improved C4.5 algorithms for Tianshengqiao plant 1 of the winter season. It indicates five schedules that are suitable for each hour in 24 h. When the value of discharge is ≥287 m^3/s, then schedule 6 with confidence index 33%, schedule 4 with confidence index 33% and schedule 7 are the most applicable. When the discharge is <287.5, schedule 3 is more likely to be followed; those of the other schedules can be ignored because of zero percentage of confidence index. The decision tree established by improved C4.5 algorithms for Tianshengqiao plant 1 of the summer season is demonstrated in Figure 11. It is noted that the six schedules are suitable for each hour in the 24 h period. When the inflow is ≥629.m^3/s, then schedule 7, with a confidence index of 100%, is appropriate, and schedule 8, with a confidence index of 75%, is appropriate when the value of inflow is <629.5 m^3/s. When the discharge is <713.5 m^3/s, schedule 7 and schedule 8 are more likely to be considered. The best rule for schedule 10 is recommended when water level is ≥746 m^3/s. The schedules that have zero percent of confidence index should be ignored.

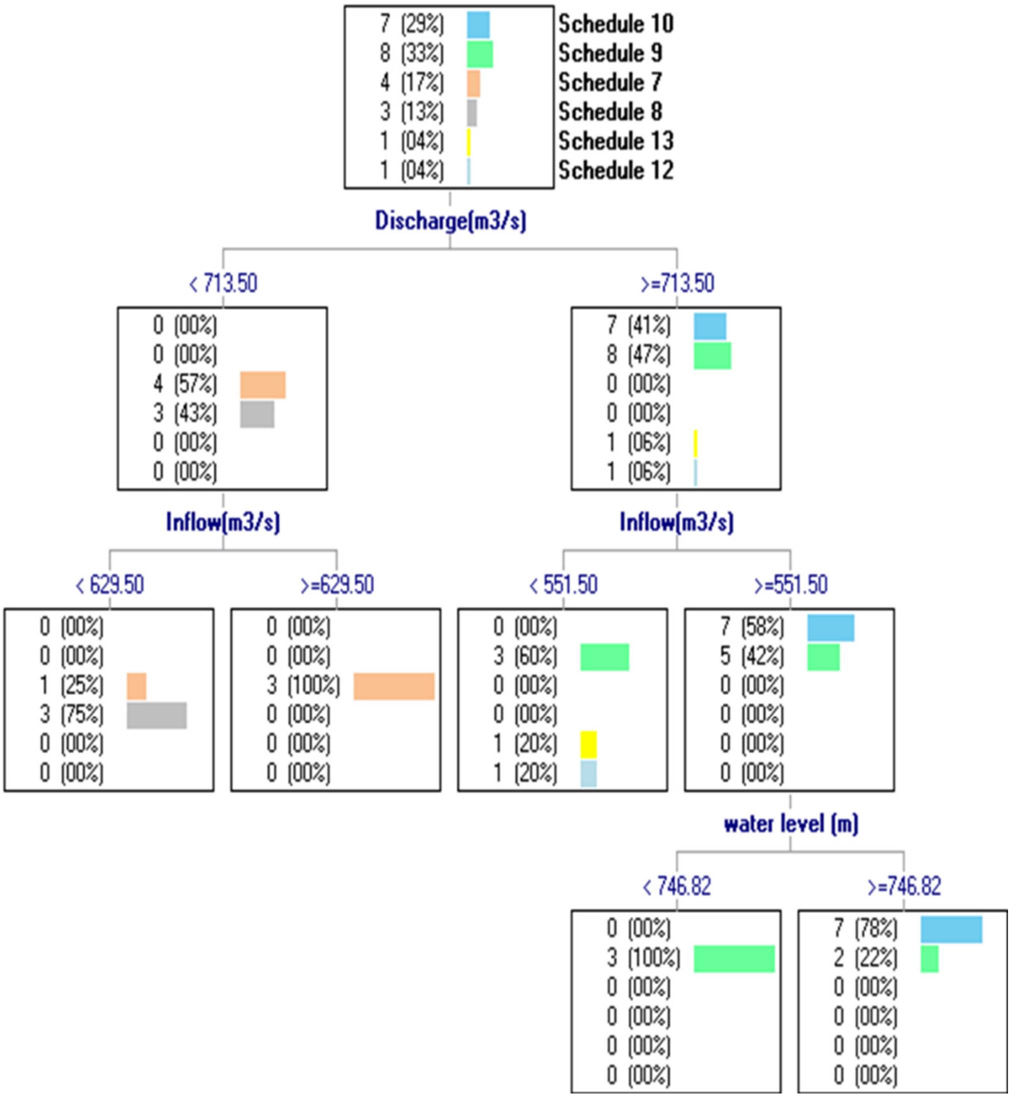


Figure 11. Decision tree by using improved C4.5 Tianshengqiao plant 1 for the summer season.

Tianshengqiao plant 2 of the winter season. Seven schedules are made for satisfying the load during the 24 h period as shown in Figure 12. When the level of discharge is ≥98.5 m^3/s with an 88% confidence index, then schedule 8 is suggested. Schedule 1 is suitable inflow <120.5 m^3/s with an 80% confidence index. The best rule for schedule 7 is recommended when the value of discharge is <398.5 m^3/s. The decision tree generated by the improved C4.5 algorithms of Tianshengqiao plant 2 for the summer season is shown (refer Figure S8). Seven schedules are established for satisfying the demand during the time period of 24 h. The tree consists of nine nodes, which represent the decision rules (value of discharge and inflow), confidence index and threshold values. When the inflow is <673 m^3/s, then schedule 8 is suitable, having a 100% confidence index. The quality of the rule is determined by the percentage of the confidence index. Furthermore, the best time to follow schedule 4 is acquired when the threshold value of the load is <439.5 MW.

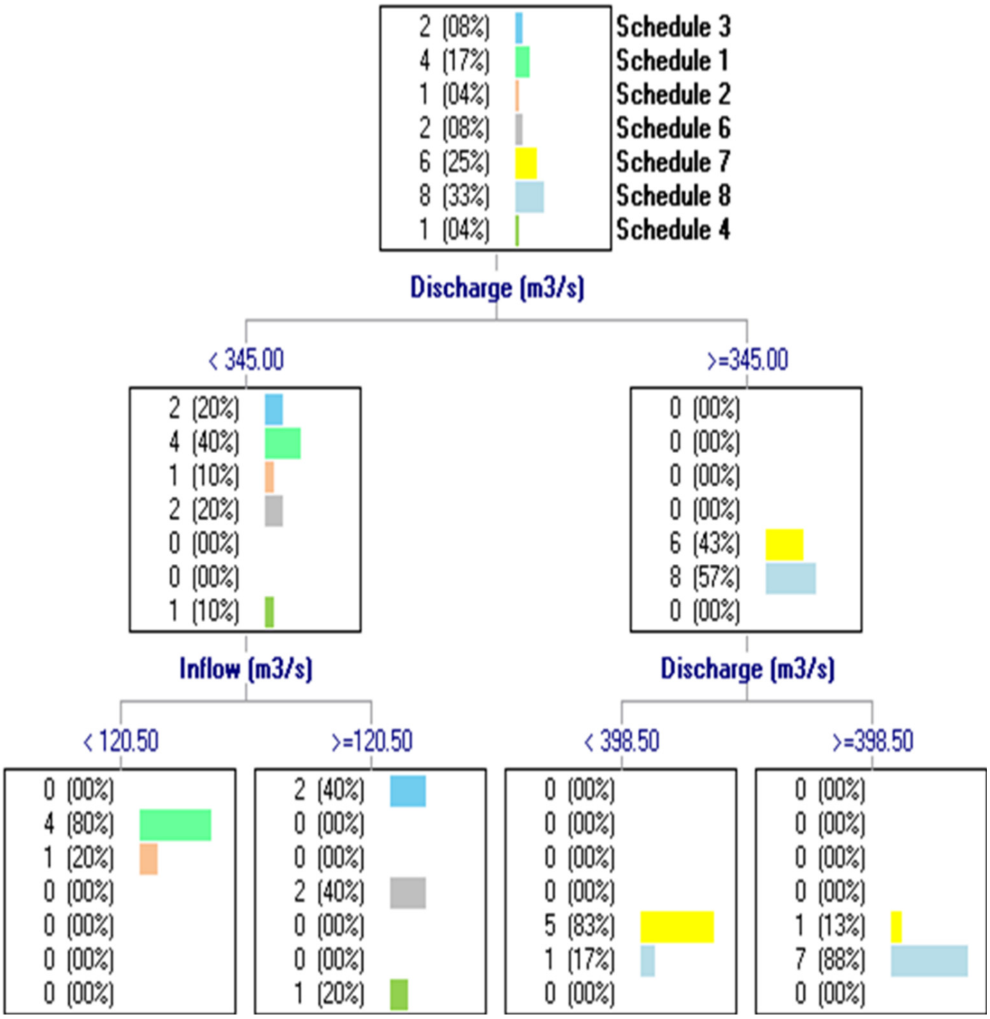
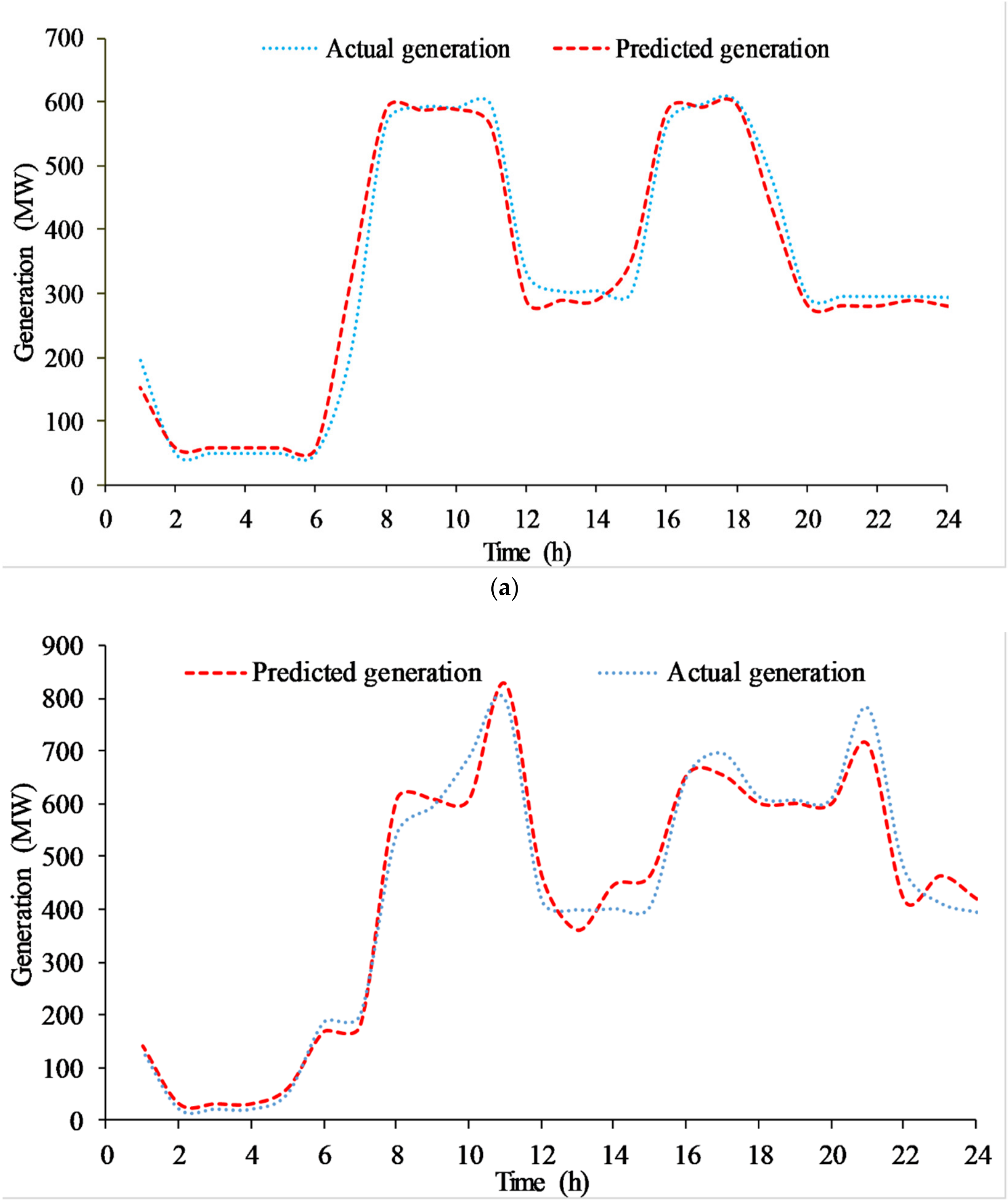


Figure 12. Decision tree by using improved C4.5 Tianshengqiao plant 2 for the winter season.

3.6. Generation Profile of Tianshengqiao Cascaded Hydropower Plants

The winter season and summer season generation profiles for Tianshengqiao plant 1 are illustrated in Figure 13a,b. The obtained result indicates that the predicted and actual generations of electricity in the summer season are higher than those of the winter season, as is evident from these figures. The proposed method provides more power generation in summer because the availability of water is greater in summer; it also showed the difference between the actual production and the predicted generation by using the proposed method. In the summer season, the snow and ice that are formed during winter melt off and produce more water, that in return is responsible for more power generation. Additionally, the power generated during peak hours in the winter season is 700 MW, which was less than the power generated during the summer season, i.e., 800 MW. The Tianshengqiao plant 2 generation profile is shown in Figure 14 for both the winter and summer season. The generation profile exhibits that the proposed method provides more generation than demand, as shown in Figure 14a,b. The power generated from the proposed method was more in the summer than winters. It also depicts the difference between the actual production and the predicted generation from the proposed method. During the winter, the precipitation is in the form of snow, which melts during the summer season. Additionally, the power generated during peak hours in the summer season is greater (1100 MW) as compared to that of power generated in the winter season (750 MW).



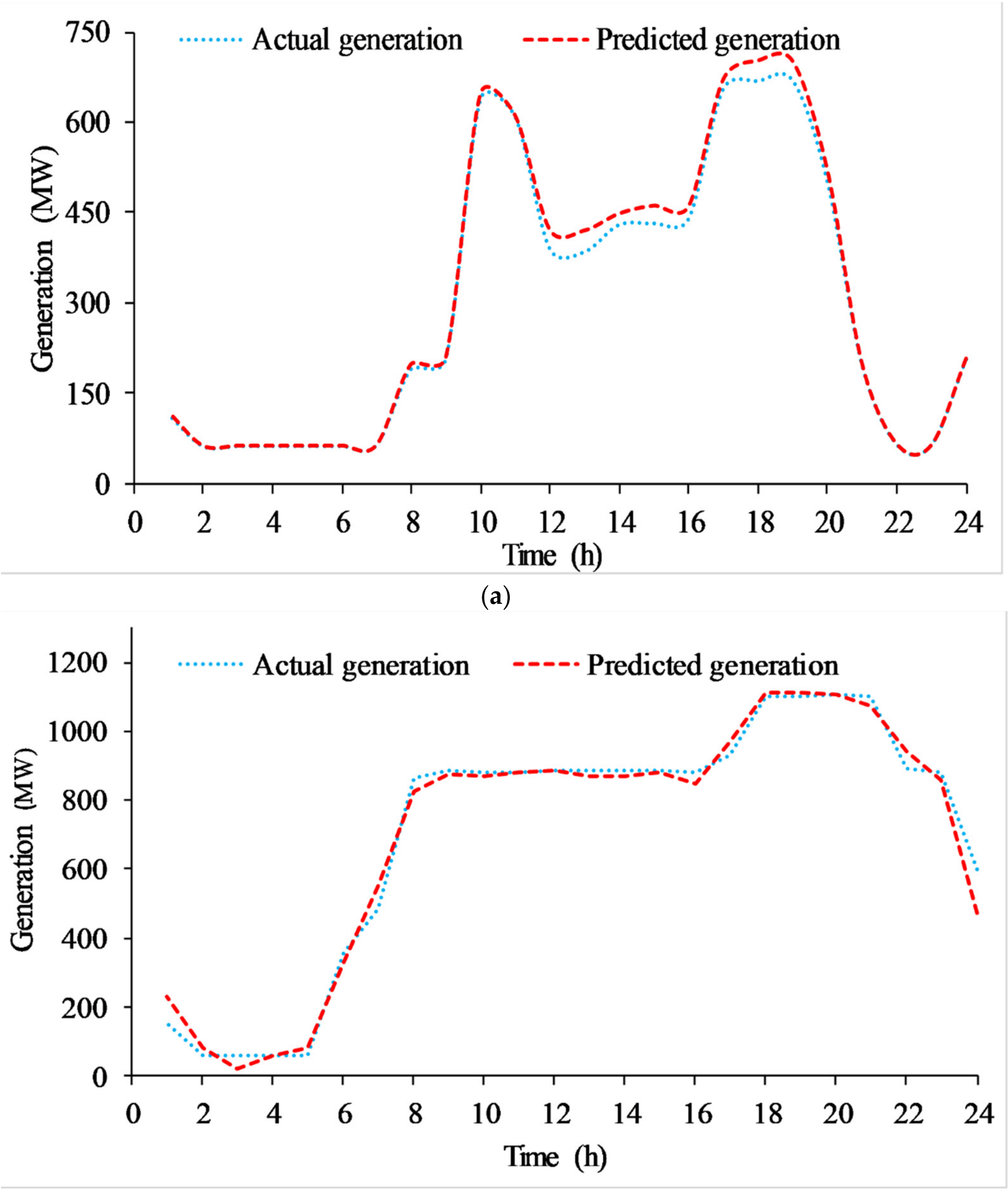
(**a**)

(**b**)

Figure 13. Generation profile of the Tianshengqiao plant 1: (**a**) winter season; (**b**) summer season.

3.7. Comparison of Different Algorithms

The comparison of various parameters for different algorithms is presented in Table 5. The obtained results show the performance ability of decision tree algorithms. According to the outcomes, different parameters, such as node, leaves, maximum depth, execution time, and error rate, are calculated using SIPINA. Four algorithms of the decision tree are employed on the data sets of Tianshengqiao plant 1 and Tianshengqiao plant 2 for both the winter and summer season. In the winter season, the decision tree of the improved algorithm has a maximum rate of percentage error of 30% with three nodes, two leaves and maximum depth of two that makes it unsuitable. Furthermore, the CHAID algorithm is also not feasible for making schedules for hydro generation because of its 16% error rate. Limitations of the CHAID algorithm are that it can only work with categorical variables and cannot handle pruning methods and cross validation. The CHAID algorithm only accepts the ordinal and nominal categorical predictors. If the predictors are continuous, then they are converted first in ordinal predictors. The C4.5 and ID3-IV show the optimal solution with error rates of 4% and 5%, respectively, for the winter season. The execution time of C4.5 is less as compared to ID3-IV; so C4.5 is well suited for Tianshengqiao plant 1 for the winter season. The results of C4.5, ID3-IV and the results for the Tianshengqiao plant 1 summer season show that the improved C4.5, ID3-IV and CHAID fail to give an optimal solution because of maximum error rate. During the winter season, the data set of Tianshengqiao plant 2 is better classified by CHAID and C4.5 with a 16% error rate. Nevertheless, the results of the improved C4.5 and ID3-IV are not acceptable with 25% and 30% error rates. The algorithm with a poor rate of error is ID3-IV, with a 30% error. For the summer season data set, C4.5 is the most feasible algorithm for making schedules on a daily basis; while ID3-IV also performed well for this data set. The execution times of all four algorithms differ for both plants and seasons, as is also shown (see Table 5). The number of attributes is the same for all the algorithms.



(**a**)

(**b**)

Figure 14. Generation profile of the Tianshengqiao plant 2: (**a**) winter season; (**b**) summer season.

Table 5. Comparison between decision tree algorithms.

	Tianshengqiao Plant 1 (Winter Season)						**Tianshengqiao Plant 1 (Summer Season)**					
Decision Tree Algorithms	**Nodes**	**Leaves**	**Max Depth**	**Attributes**	**Error Rate**	**Execution Time**	**Nodes**	**Leaves**	**Max Depth**	**Attributes**	**Error Rate**	**Execution Time**
C4.5	7	4	3	8	4%	93 ms	13	7	6	8	8.3%	188 ms
Improved C4.5	3	2	2	8	30%	109 ms	9	5	4	8	20%	109 ms
ID3	5	3	3	8	5%	110 ms	5	3	3	8	40%	125 ms
CHAID	4	3	2	8	16%	93 ms	4	3	2	8	20%	125 ms
	Tianshengqiao Plant 2 (Winter Season)						**Tianshengqiao Plant 2 (Summer Season)**					
Decision Tree Algorithms	**Nodes**	**Leaves**	**Max Depth**	**Attributes**	**Error Rate**	**Execution Time**	**Nodes**	**Leaves**	**Max Depth**	**Attributes**	**Error Rate**	**Execution Time**
C4.5	9	5	4	8	16%	94 ms	9	5	4	8	8%	110 ms
Improved C4.5	7	4	3	8	25%	125 ms	9	5	4	8	12%	187 ms
ID3	5	3	3	8	30%	110 ms	7	4	3	8	16%	110 ms
CHAID	4	3	2	7	4.1%	93 ms	4	3	2	8	25%	110 ms

Note: ms denotes milli second.

4. Discussion

In the past, work has been carried out in solving the hydro scheduling problem while considering different constraints. The algorithms and models are different, even if they lie within the same category. Wang, Shahidehpour, Kirschen, Mokhtari and Irisarri [62], Frangioni, Gentile and Lacalandra [63], Orero and Irving [64], Beltran and Heredia [65], Virmani, Adrian and Mukherjee [66] and Gröwe-Kuska, Kiwiel, Nowak, Römisch and Wegner [67] used augmented Lagrange relaxation, sequential Lagrange and MILP, Lagrange relaxation decomposition and genetic algorithm, augmented Lagrange relaxation, decomposition techniques (block coordinate descent and auxiliary problem principle), Lagrange relaxation, stochastic Lagrange relaxation methods, respectively. The purpose of the non-linear method is to examine the small change in head, discharge and water storage. Catalão et al. [68] proposed the non-linear method for resolving the issue of hydrogenation scheduling; the results showed that this method gave an optimal solution in less time. Patra et al. [69] used the dynamic programming method and the results obtained are efficient during peak hours. Esmaeily et al. [70] presented the MILP in order to solve the unit commitment problem; results showed that the speed of execution is fast. In solving the real world's problems, the Augmented Lagrange Relaxation is robust, fast and efficient. The augmented Lagrange relaxation and decomposition techniques (auxiliary problem principle and block coordinate descent) are implemented as a global optimizer. Wong [71] implemented this technique and the results obtained were found satisfactory. Zheng et al. [11] used benders decomposition and found that the results obtained were good but the convergence was not faster. Classification techniques, such as decision trees, showed that without doing the complex mathematical formulations, the decision tree can be used for making quick decisions by taking the useful information from the large data sets. Different decision tree algorithms, such as C4.5, ID3-IV, etc., deal with continuous and discrete variables and are good in solving the poor practicable outcomes of the different operations. From the above discussion, it is found that classification techniques, such as decision tree algorithms, can be used to find quick decisions for short-term hydro scheduling of cascaded hydropower plant. It is also feasible to find speedy unit generation schedules. Decision tree algorithm C4.5 is the remedy for all troubles, allowing one to choose proper decisions for schedules, with the target being maximal hydro-energy production. More often for energy plants, the aim is related to the economic value of electric energy production because the price of electric energy depends on the electric energy demand especially. The cascaded hydropower plant is a water reservoir and increases and decreases the electric energy production in a short time, aligning the maximal and minimal electric energy demand. Therefore, in this study, different data mining algorithms are compared and discussed for short-term hydro-scheduling.

5. Conclusions

In this study, hydropower is considered as a clean source of renewable energy, and, in order to generate more energy, the data mining algorithms and techniques such as clustering, regression and decision tree algorithms are used to explore more valuable information from the available data of the power system. The data set is collected for both winter and summer seasons; the generation obtained from the decision tree algorithms are then determined, resulting in the best algorithms that show fast generation schedules being found. The current work evaluates the procedure of four decision tree algorithms, namely, C4.5, improved C4.5, ID3-IV and CHAID for finding the quick decisions of hydro scheduling. The error rate of all four algorithms is computed. The main conclusions are as follows:

1. The values of energy production are found by using the multiple regression analysis for both winter and summer seasons of the Tianshengqiao cascaded hydropower plant. The energy production increases with increase in discharge.
2. The K-mean cluster analysis technique is used for the generation curves based on historical data, and the cluster analysis identified the most similar generation curves for each season of each hydropower plant.

3. The case study of the Tianshengqiao cascaded hydropower plants is considered, which is on the mainstream of Hongshui River. Data sets are established for both winter and summer seasons for upstream and downstream power stations. The data sets consist of water level, load, discharge, inflow, energy production and the generation schedules of each hour.
4. The results obtained from the ID3-IV algorithm showed the best performance on the data set of Tianshengqiao 1 (winter season) because of the good number of splits, but for the other three types of data sets, the power of prediction is weaker because of its high percentage error.
5. The CHAID algorithm depicts overall reasonable classification except for the winter season of Tianshengqiao plant 2 with 4.1% error rate.
6. The results exhibited by improved C4.5 are not satisfactory for three cases; however, it precisely predicts the outcome of the summer season of Tianshengqiao plant 2 with 12% error.

Based on this research, among all the algorithms, C4.5 is the best suited for making quick and optimal decisions for the hydro generation scheduling problem because it has a low error rate; also, its tree is more compact and well-classified, and its number of splits and execution time are in acceptable range.

It is suggested that number of measured points along with multiple input parameters must be increased for more precise and accurate results in the future. There is also a need to include constraints such as spinning reserve, security constraints, etc. The problem of the cascaded hydropower plant should be solved by using approximate dynamic programming compared with data mining algorithms outcomes. The results of the study can be cross checked and validated for long-term hydro scheduling and hydrothermal scheduling.

Supplementary Materials: The following data are available online at https://www.mdpi.com/1996-1073/14/2/298/s1, Figure S1: Decision tree of Tianshengqiao plant 1 generated for C4.5 summer season, Figure S2: Decision tree of Tianshengqiao plant 2 generated for C4.5 winter season. Figure S3: Decision tree generated by CHAID of Tianshengqiao plant 1 for winter season. Figure S4: Decision tree generated by CHAID of Tianshengqiao plant 2 for summer season. Figure S5: Decision tree by using ID3 algorithm Tianshengqiao plant 1 for winter season. Figure S6: Decision tree by using ID3 algorithm Tianshengqiao plant 2 for summer season. Figure S7: Decision tree by using improved C4.5 Tianshengqiao plant 1 for winter season. Figure S8: Decision tree by using improved C4.5 Tianshengqiao plant 2 for summer season.

Author Contributions: Conceptualization, I.P. and J.S.; methodology, I.P.; software, I.P.; validation, I.P.; formal analysis, I.P.; investigation, I.P.; resources, J.S.; data curation, I.P. and N.Z.; writing—original draft preparation, I.P.; writing—review and editing, J.S., I.H., and N.Z.; visualization, J.S. and I.H.; supervision, J.S.; project administration, J.S.; funding acquisition, J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (No. 52079014), and open research fund of the Hubei Provincial Key Laboratory for Operation, Control of Cascaded Hydropower Station (2019KJX06) and Dalian Youth Science and Technology star (2019RQ004) and the Fundamental Research Funds for the Central Universities (DUT19JC43).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are possibly available on request from the corresponding authors.

Acknowledgments: The authors would like to acknowledge the financial support from the China Scholarship Council (CSC) for higher studies of Iram Parvez at the Dalian University of Technology, China. The careful review and constructive suggestions by the anonymous reviewers are gratefully acknowledged.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ID3	Iterative Dichotomiser 3
CHAID	Chi-Squared Automatic Interaction Detection
CART	Classification and Regression Tree
Zup	Water level
N	Power generation
X_t	Eigenvector vector of generation curve in time t
P_t	Power value of generation curve in time t
P_{max}	Maximum generation of hydropower plant within one day
X	Set of observations spanning up to "n" observations
W	Sum of square errors of all generation curves in historical data set
a_k	Generation curve of a group of cluster vectors
b_i	A cluster vector, which is the center of cluster $\mathrm{C_i}$
q_i	Rate of samples with generation schedule number and i is the number of samples
F	Samples
F_k	Number of samples of kth element included in attribute B

References

1. Zamfir, A.; Colesca, S.E.; Corbos, R.-A. Public policies to support the development of renewable energy in Romania: A review. *Renew. Sustain. Energy Rev.* **2016**, *58*, 87–106. [CrossRef]
2. Koutsoyiannis, D.; Efstratiadis, A.; Karavokiros, G. A Decision support tool for the management of multi-reservoir systems. *JAWRA J. Am. Water Resour. Assoc.* **2002**, *38*, 945–958. [CrossRef]
3. Shen, J.; Zhang, X.; Wang, J.; Cao, R.; Wang, S.; Zhang, J. Optimal operation of interprovincial hydropower system including Xiluodu and local plants in multiple recipient regions. *Energies* **2019**, *12*, 144. [CrossRef]
4. Shen, J.-J.; Shen, Q.-Q.; Wang, S.; Lu, J.-Y.; Meng, Q.-X. Generation scheduling of a hydrothermal system considering multiple provincial peak-shaving demands. *IEEE Access* **2019**, *7*, 46225–46239. [CrossRef]
5. Suwal, N.; Huang, X.; Kuriqi, A.; Chen, Y.; Pandey, K.P.; Bhattarai, K.P. Optimisation of cascade reservoir operation considering environmental flows for different environmental management classes. *Renew. Energy* **2020**, *158*, 453–464. [CrossRef]
6. Kuriqi, A.; Pinheiro, A.N.; Sordo-Ward, A.; Garrote, L. Water-energy-ecosystem nexus: Balancing competing interests at a run-of-river hydropower plant coupling a hydrologic—Ecohydraulic approach. *Energy Convers. Manag.* **2020**, *223*, 113267. [CrossRef]
7. Hammid, A.T.; Sulaiman, M.H.B.; Abdalla, A.N. Prediction of small hydropower plant power production in Himreen Lake dam (HLD) using artificial neural network. *Alex. Eng. J.* **2018**, *57*, 211–221. [CrossRef]
8. Liu, D.; Guo, S.; Shao, Q.; Liu, P.; Xiong, L.; Wang, L.; Hong, X.; Xu, Y.; Wang, Z. Assessing the effects of adaptation measures on optimal water resources allocation under varied water availability conditions. *J. Hydrol.* **2018**, *556*, 759–774. [CrossRef]
9. Séguin, S.; Côté, P.; Audet, C. Self-scheduling short-term unit commitment and loading problem. *IEEE Trans. Power Syst.* **2015**, *31*, 133–142. [CrossRef]
10. Hechme-Doukopoulos, G.; Brignol-Charousset, S.; Malick, J.; Lemaréchal, C. The short-term electricity production management problem at EDF. *Optima Newsl.* **2010**, *84*, 2–6.
11. Zheng, Q.P.; Wang, J.; Pardalos, P.M.; Guan, Y. A decomposition approach to the two-stage stochastic unit commitment problem. *Annals Oper. Res.* **2013**, *210*, 387–410. [CrossRef]
12. Carøe, C.C.; Schultz, R. A Two-Stage Stochastic Program for Unit Commitment under Uncertainty in a Hydro-Thermal Power System. 1998. Available online: https://www.semanticscholar.org/paper/A-Two-Stage-Stochastic-Program-for-Unit-Commitment-Car%C3%B8e-Schultz/2c2620b14608f8fcb52e3c47c0d50ee35cfa9b7c (accessed on 28 December 2020).
13. Carrión, M.; Arroyo, J.M. A computationally efficient mixed-integer linear formulation for the thermal unit commitment problem. *IEEE Trans. Power Syst.* **2006**, *21*, 1371–1378. [CrossRef]
14. Shen, J.J.; Shen, Q.Q.; Cheng, C.T.; Zhang, X.F.; Wang, J. Large-Scale Unit Commitment for Cascaded Hydropower Plants with Hydraulic Coupling and Head-Sensitive Forbidden Zones: Case of the Xiluodu and Xiangjiaba Hydropower System. *J. Water Resour. Plan. Manag.* **2020**, *146*, 05020023. [CrossRef]
15. Razavi, S.-E.; Nezhad, A.-E.; Mavalizadeh, H.; Raeisi, F.; Ahmadi, A. Robust hydrothermal unit commitment: A mixed-integer linear framework. *Energy* **2018**, *165*, 593–602. [CrossRef]
16. Wang, J.; Guo, M.; Liu, Y. Hydropower unit commitment with nonlinearity decoupled from mixed integer nonlinear problem. *Energy* **2018**, *150*, 839–846. [CrossRef]
17. Przychodzen, W.; Przychodzen, J. Determinants of renewable energy production in transition economies: A panel data approach. *Energy* **2020**, *191*, 116583. [CrossRef]
18. Fernández-Muñoz, D.; Pérez-Díaz, J.I. Contribution of non-conventional pumped-storage hydropower plant configurations in an isolated power system with an increasing share of renewable energy. *IET Renew. Power Gener.* **2019**, *14*, 658–670. [CrossRef]
19. Ahmad, S.K.; Hossain, F. Maximizing energy production from hydropower dams using short-term weather forecasts. *Renew. Energy* **2020**, *146*, 1560–1577. [CrossRef]

20. Hussain, A.; Sarangi, G.P.; Pandit, A.; Ishaq, S.; Mamnun, N.; Ahmad, B.; Jamil, M.K. Hydropower development in the Hindu Kush Himalayan region: Issues, policies and opportunities. *Renew. Sustain. Energy Rev.* **2019**, *107*, 446–461. [CrossRef]
21. Wang, Y.; Yan, W.; Zhuang, S.; Zhang, Q. Competition or complementarity? The hydropower and thermal power nexus in China. *Renew. Energy* **2019**, *138*, 531–541. [CrossRef]
22. Wagner, B.; Hauer, C.; Habersack, H. Current hydropower developments in Europe. *Curr. Opin. Environ. Sustain.* **2019**, *37*, 41–49. [CrossRef]
23. Mahmud, M.P.; Huda, N.; Farjana, S.H.; Lang, C. A strategic impact assessment of hydropower plants in alpine and non-alpine areas of Europe. *Appl. Energy* **2019**, *250*, 198–214. [CrossRef]
24. Ikura, Y.; Gross, G. Efficient large-scale hydro system scheduling with forced spill conditions. *IEEE Trans. Power Appar. Syst.* **1984**, 3502–3520. [CrossRef]
25. Catalão, J.P.D.S.; Mariano, S.J.P.S.; Mendes, V.M.F.; Ferreira, L.A.F.M. Nonlinear optimization method for short-term hydro scheduling considering head-dependency. *Eur. Trans. Electr. Power* **2010**, *20*, 172–183. [CrossRef]
26. Håberg, M. Fundamentals and recent developments in stochastic unit commitment. *Int. J. Electr. Power Energy Syst.* **2019**, *109*, 38–48. [CrossRef]
27. Feng, Z.K.; Niu, W.J.; Cheng, C.T.; Wu, X.Y. Optimization of large-scale hydropower system peak operation with hybrid dynamic programming and domain knowledge. *J. Clean. Prod.* **2018**, *171*, 390–402. [CrossRef]
28. Feng, Z.K.; Niu, W.J.; Cheng, C.T.; Wu, X.Y. Optimization of hydropower system operation by uniform dynamic programming for dimensionality reduction. *Energy* **2017**, *134*, 718–730. [CrossRef]
29. Feng, Z.; Niu, W.; Wang, S.; Cheng, C.; Song, Z. Mixed integer linear programming model for peak operation of gas-fired generating units with disjoint-prohibited operating zones. *Energies* **2019**, *12*, 2179. [CrossRef]
30. Shen, J.; Cheng, C.; Shen, Q.; Lu, J.; Zhang, J. Overview of China's hydropower absorption: Evolution, problems, and suggested solutions. *IET Renew. Power Gener.* **2019**, *13*, 2491–2501. [CrossRef]
31. Ghaddar, B.; Naoum-Sawaya, J.; Kishimoto, A.; Taheri, N.; Eck, B. A Lagrangian decomposition approach for the pump scheduling problem in water networks. *Eur. J. Oper. Res.* **2015**, *241*, 490–501. [CrossRef]
32. Rachunok, B.; Staid, A.; Watson, J.P.; Woodruff, D.L.; Yang, D. Stochastic unit commitment performance considering Monte Carlo wind power scenarios. In Proceedings of the 2018 IEEE international conference on probabilistic methods applied to power systems (PMAPS), Boise, ID, USA, 24–28 June 2018; pp. 1–6.
33. Gade, D.; Hackebeil, G.; Ryan, S.M.; Watson, J.P.; Wets, R.J.B.; Woodruff, D.L. Obtaining lower bounds from the progressive hedging algorithm for stochastic mixed-integer programs. *Math. Program.* **2016**, *157*, 47–67. [CrossRef]
34. Papavasiliou, A.; Oren, S.S.; Rountree, B. Applying high performance computing to transmission-constrained stochastic unit commitment for renewable energy integration. *IEEE Trans. Power Syst.* **2014**, *30*, 1109–1120. [CrossRef]
35. López-Salgado, C.J.; Añó, O.; Ojeda-Esteybar, D.M. Stochastic unit commitment and optimal allocation of reserves: A hybrid decomposition approach. *IEEE Trans. Power Syst.* **2018**, *33*, 5542–5552. [CrossRef]
36. Scuzziato, M.R.; Finardi, E.C.; Frangioni, A. Comparing spatial and scenario decomposition for stochastic hydrothermal unit commitment problems. *IEEE Trans. Sustain. Energy* **2017**, *9*, 1307–1317. [CrossRef]
37. Jiang, R.; Guan, Y.; Watson, Y.-P. Cutting planes for the multistage stochastic unit commitment problem. *Math. Program.* **2016**, *157*, 121–151. [CrossRef]
38. Analui, B.; Scaglione, A. A dynamic multistage stochastic unit commitment formulation for intraday markets. *IEEE Trans. Power Syst.* **2017**, *33*, 3653–3663. [CrossRef]
39. Wang, W.; Li, C.; Liao, X.; Qin, H. Study on unit commitment problem considering pumped storage and renewable energy via a novel binary artificial sheep algorithm. *Appl. Energy* **2017**, *187*, 612–626. [CrossRef]
40. Shahbazitabar, M.; Abdi, H. A novel priority-based stochastic unit commitment considering renewable energy sources and parking lot cooperation. *Energy* **2018**, *161*, 308–324. [CrossRef]
41. Perez-Ortega, J.; Almanza-Ortega, N.N.; Romero, D. Balancing effort and benefit of K-means clustering algorithms in Big Data realms. *PLoS ONE* **2018**, *13*, e0201874. [CrossRef]
42. Jo, K.-H.; Kim, M.-K. Stochastic unit commitment based on multi-scenario tree method considering uncertainty. *Energies* **2018**, *11*, 740. [CrossRef]
43. Liao, S.-H.; Chu, P.-H.; Hsiao, P.-Y. Data mining techniques and applications–A decade review from 2000 to 2011. *Expert Syst. Appl.* **2012**, *39*, 11303–11311. [CrossRef]
44. Bharati, M.; Ramageri, M. Data mining techniques and applications. *Indian J. Comput. Sci. Eng.* **2010**, *1*, 301–305.
45. White, I.R. Bayesian network meta-analysis. *Stata J.* **2019**, *15*, 951–985. [CrossRef]
46. El Jerjawi, N.S.; Abu-Naser, S.S. Diabetes prediction using artificial neural network. *Int. J. Adv. Sci. Technol.* **2018**, *121*, 54–64.
47. Chain, P.; Arunyanart, S. Using cluster analysis for location decision problem. In *IOP Conference Series: Materials Science and Engineering*; IOP Publishing: Bristol, UK, 2019.
48. Schroeder, L.D.; Sjoquist, D.L.; Stephan, P.E. *Understanding Regression Analysis: An Introductory Guide*; Sage Publications: Newcastle, UK, 2016.
49. Brodny, J.; Tutak, M. Analyzing similarities between the European Union countries in terms of the structure and volume of energy production from renewable energy sources. *Energies* **2020**, *13*, 913. [CrossRef]
50. Romesburg, C. *Cluster Analysis for Researchers*; Lily Press: Morrisville, NC, USA, 2004.

51. Zhou, Y.; Lai, C.; Wang, Z.; Chen, X.; Zeng, Z.; Chen, J.; Bai, X. Quantitative evaluation of the impact of climate change and human activity on runoff change in the Dongjiang River Basin, China. *Water* **2018**, *10*, 571. [CrossRef]
52. Liu, J.; Ning, B.; Shi, D. Application of improved decision tree C4. 5 algorithms in the judgment of diabetes diagnostic effectiveness. *J. Phys. Conf. Ser.* **2019**. [CrossRef]
53. Wahbeh, A.H.; Al-Radaideh, Q.A.; Al-Kabi, M.N.; Al-Shawakfa, E.M. A comparison study between data mining tools over some classification methods. *Int. J. Adv. Comput. Sci. Appl.* **2011**, *8*, 18–26.
54. Milanović, M.; Stamenković, M. CHAID decision tree: Methodological frame and application. *Econ. Themes* **2016**, *54*, 563–586. [CrossRef]
55. Ren, N.; Zargham, M.; Rahimi, S. A decision tree-based classification approach to rule extraction for security analysis. *Int. J. Inf. Technol. Decis. Mak.* **2006**, *5*, 227–240. [CrossRef]
56. Dai, Q.-Y.; Zhang, C.-P.; Wu, H. Research of decision tree classification algorithm in data mining. *Int. J. Database Theory Appl.* **2016**, *9*, 1–8. [CrossRef]
57. Shen, J.; Cheng, C.; Zhang, X.; Zhou, B. Coordinated operations of multiple-reservoir cascaded hydropower plants with cooperation benefit allocation. *Energy* **2018**, *153*, 509–518. [CrossRef]
58. Report by Mitsue Mishima (OPMAC), Tianshengqiao First Hydropower Project (1)–(4) September 2004: China main Contractor, South China Hydropower Construction Association. *Japan International Cooperation Agency (JICA)*. Available online: https://www.jica.go.jp/english/our_work/evaluation/oda_loan/post/2005/pdf/2-21_full.pdf (accessed on 28 December 2020).
59. Report on Field survey of Tianshengqiao Hydro Electric Power Generation Project (1)-(6) 2002: China. *Japan International Cooperation Agency (JICA)*. Available online: https://www.jica.go.jp/english/our_work/evaluation/oda_loan/post/2002/pdf/003_full.pdf (accessed on 28 December 2020).
60. Reddan, T.; Corness, J.; Harden, F.; Mengersen, K. Analysis of the predictive value of clinical and sonographic variables in children with suspected acute appendicitis using decision tree algorithms. *Sonography* **2018**, *5*, 157–163. [CrossRef]
61. Ru-Ping, L. Research of Decision Tree Classification Algorithm in Data Mining. *J. East China Inst. Technol. Nat. Sci.* **2010**, *2*. Available online: http://en.cnki.com.cn/Article_en/CJFDTotal-HDDZ201002018.htm (accessed on 28 December 2020).
62. Wang, S.J.; Shahidehpour, S.M.; Kirschen, D.S.; Mokhtari, S.; Irisarri, G.D. Short-term generation scheduling with transmission and environmental constraints using an augmented Lagrangian relaxation. *IEEE Trans. Power Syst.* **1995**, *10*, 1294–1301. [CrossRef]
63. Frangioni, A.; Gentile, C.; Lacalandra, F. Sequential Lagrangian-MILP approaches for unit commitment problems. *Int. J. Electr. Power Energy Syst.* **2011**, *33*, 585–593. [CrossRef]
64. Orero, S.; Irving, M. A combination of the genetic algorithm and Lagrangian relaxation decomposition techniques for the generation unit commitment problem. *Electr. Power Syst. Res.* **1997**, *43*, 149–156. [CrossRef]
65. Beltran, C.; Heredia, F.J. Unit commitment by augmented lagrangian relaxation: Testing two decomposition approaches. *J. Optim. Theory Appl.* **2002**, *112*, 295–314. [CrossRef]
66. Virmani, S.; Adrian, E.C.; Imhof, K.; Mukherjee, S. Implementation of a Lagrangian relaxation based unit commitment problem. *IEEE Trans. Power Syst.* **1989**, *4*, 1373–1380. [CrossRef]
67. Gröwe-Kuska, N.; Kiwiel, K.C.; Nowak, M.P.; Römisch, W.; Wegner, I. Power management under uncertainty by Lagrangian relaxation. In Proceedings of the 16th International Conference Probabilistic Methods Applied to Power Systems PMAPS, Liege, Belgium, 18–21 August 2020; Volume 2.
68. Catalão, J.P.S.; Mariano, S.J.P.S.; Mendes, V.M.F.; Ferreira, L.A.F.M. Scheduling of head-sensitive cascaded hydro systems: A nonlinear approach. *IEEE Trans. Power Syst.* **2008**, *24*, 337–346. [CrossRef]
69. Patra, S.; Goswami, S.; Goswami, B. Fuzzy and simulated annealing based dynamic programming for the unit commitment problem. *Expert Syst. Appl.* **2009**, *36*, 5081–5086. [CrossRef]
70. Esmaeily, A.; Ahmadi, A.; Raeisi, F.; Ahmadi, M.R.; Nezhad, A.E.; Janghorbani, M. Evaluating the effectiveness of mixed-integer linear programming for day-ahead hydro-thermal self-scheduling considering price uncertainty and forced outage rate. *Energy* **2017**, *122*, 182–193. [CrossRef]
71. Wong, S.Y.W. An enhanced simulated annealing approach to unit commitment. *Int. J. Electr. Power Energy Syst.* **1998**, *20*, 359–368. [CrossRef]