International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 16, No.1 (Aug-24)

http://dx.doi.org/10.12785/ijcds/160142

Decision Tree Analysis Approaches to Classify Sensors Data in a Water Pumping Station

Mostafa Adnan Hadi[1], Alaa Khalaf Hamoud[1], Ahmed Monther Abboud[1], Ahmed Naji Abdullah[1] and Ahmed Khaled Abdullatif[1]

[1]*Department of Computer Information Systems,University of Basrah, Basrah, Iraq*

Received 24 Jun. 2023, Revised 7 Apr. 2024, Accepted 27 Apr. 2024, Published 1 Aug. 2024

Abstract: Water pumping stations play a vital role in the lives of citizens, where a failure in the pumping schedule or the quality of the pumping may affect their lives. The data of the water pumping station may expose the weaknesses in the system of the station, which can be overcome using machine learning approaches. In this paper, six decision tree algorithms are examined to find the optimal one for classifying the data of water pumping stations. The main goal is to determine the fault in the sensors to control the pumping process and overcome future failures. Six algorithms, namely J48, Rep Tree, Random Forest, Decision Stump, Hoeffding Tree, and Random Tree, are examined before and after implementing the feature selection (FS) process. FS is implemented to find the most correlated sensors and remove the less correlated sensors. The FS process affects the accuracies of the algorithms and enhances the resulting accuracies of the algorithms. Random Forest and Random Tree algorithms prove their accuracy in data classification with 100% accuracy after implementing FS and removing the less correlated sensor data. The model can be used as an assistant tool for classifying and predicting the failure of a water pumping station.

Keywords: Decision Tree, Sensors Data, Water Pumping Station, Supervised Machine Learning, Machine Learning Algorithms.

1. INTRODUCTION

Machine learning algorithms are considered the basis of any model used for prediction, analyzing the data, and discovering patterns and anomalies. The field of analyzing data, regardless of its size, whether small or big, faced utilizing supervised and unsupervised machine learning algorithms in the analyzing process, which led to the discovery of valuable patterns that can be used for decision-making. The main characteristics of using machine learning are that it reduces the time consumed to discover the pattern and produces an efficient model. Machine learning algorithms can also utilize different kinds of historical data and find a correlation with the current data.

Water pumping stations have a critical effect on the lifestyle of humans. Different factors and circumstances may affect the pumping station's work and the schedule of water pumping, such as sensor failure, time of pumping, and water availability. Sensor data in a water pumping station may have hidden patterns that can be utilized to optimize the work of the pumping station and diagnose errors. The data can also contain errors, which may cause unreliable decisions related to pumping systems. Different data mining approaches are utilized to analyze sensor data, such as decision trees [1] [2] [3], deep learning [4] [5] [6], regression [7] [8], support vector machine (SVM), and clustering [9] [10]. Decision tree analysis proved its accuracy and performance in different sectors such as education [11] [12] [13], healthcare [13] [14] [15], wireless sensor network [16], stock market, and disaster management [17] [18] [19] [20].

In this paper, machine learning (supervised machine learning) approaches are examined to diagnose the time of failure and predict the sensors that cause the failure in the water pumping station. The objective of the model is to examine machine learning algorithms to find the optimal one for predicting the faults in sensors that are utilized in the water pumping station, which may affect the overall work of the station and then the water pump scheduling. The aim of the study is to ensure high station performance based on learning from the sensor data and using the knowledge gained to predict failure and overcome it.

The proposed approach will help in predicting the potential sensors and the time of failure at the station in order to take proactive action. Four decision tree approaches (decision stump, hoeffding tree, rep tree, and random forest) are utilized and examined to find the optimal algorithm for predicting and classifying sensor data. Real sensor data in the model, where the DT outperforms the other decision tree algorithms in predicting. The DT algorithm proved its accuracy in diagnosing the failure at the station. The paper

E-mail address: mustafa.adnan1998tv@gmail.com,alaa.hamoud@uobasrah.edu.iq, ahmedmontherabboud41@gmail.com,pro1ahmednaji@gmail.com, ahmedkhaledabdullatif41@gmail.com.

is organized as follows: Section 2 presents the related works in the field of implementing machine learning approaches in classifying water pumping station data. Section 3 presents the methodology and framework for implementing and examining decision tree algorithms. Finally, Section 4 explains the concluded points after implementing the model and lists the future works.

2. RELATED WORKS

In [21], Zhaomin Li et al. proposed a strategy based on a deep learning algorithm to overcome the scheduling problem of the water system in the pump station. A deep reinforcement-based suspended sediment concentration is proposed to predict the sediment concentration based on data collected from a withdrawal pumping station in yellow river water. The aim of the strategy is to control wasted energy and reduce energy consumption and annual water usage.

Next in [22], Shiyuan Hu et al. proposed a model based on reinforcement learning algorithms to handle different limitations in real-time pumping stations, such as energy consumption and delay in pumping. The researchers found that deep learning can enhance pump scheduling by transferring the computation offline while enhancing energy consumption. While in [23], Veena Khandelwal et al. presented a model-based machine learning model to exhibit the quality levels of the water. Different physiochemical parameters have been examined in the study, such as sodium, total hardness, nitrate, chloride, magnesium, sulfate, and other physiochemical parameters. The data utilized in this study was collected from 118 points of groundwater stations over a period of about 18 years (2000–2018). Regression random forest and decision tree models were utilized to predict the index of water quality, where the random forest outperformed the decision tree based on the prediction accuracy.

In [10] Arsene et al. investigated the use of machine learning (ML) techniques for energy management in water pumping systems on small islands. The study aimed to develop a peak shaving strategy for a water pumping system, which would reduce peak demand and improve energy efficiency. The approach used a combination of ML algorithms, including k-means clusters and artificial neural networks, to optimize the system's power consumption. The study evaluated the performance of the peak-shaving strategy using real-world data from a small island in Indonesia. Results and studies showed that the ML-based approach was effective in reducing peak demand, improving energy efficiency in the water pumping system, and reducing carbon emissions in small communities. The study used a range of performance measures, including peak demand reduction, energy consumption, and power factor improvement. The sample size was a water pumping system on a small island in Indonesia. The study was characterized by a high level of methodological accuracy, with the use of appropriate data collection and analysis procedures. The accuracy of the results was high, as the statistical analysis showed a significant improvement in the performance of the water pumping system. The authors note that the results can be applied to other water pumping systems in similar conditions. The study contributes to the development of sustainable energy solutions for small islands and remote communities. The researchers suggested that future research could investigate the scalability and generalizability of the ML-based approach to other types of energy systems.

In [24], Predescu et al. aimed to improve water distribution systems to address environmental and public health outcomes. The study proposes a multiple-model-based control approach that combines different models to achieve better control performance. The control supervisor is developed based on machine learning algorithms that learn from historical data to improve system performance. The research results indicate that the multi-model control supervisor based on advanced learning is more efficient and effective than traditional control methods. The proposed approach is adaptable to changing system conditions and can improve system performance in real-time. However, the study identified some limitations, including the need for a more comprehensive evaluation of the proposed approach using different datasets and system conditions. Additionally, the approach may require further improvement for practical implementation. In summary, the research study developed an advanced learning-based multi-model control supervisor for pumping stations in a smart water distribution system. The proposed approach shows promising results in improving efficiency and effectiveness. However, more research is necessary to validate the approach and improve its implementation.

In [25], Pałczyński et al. attempted to design a model to predict water consumption based on iterative design to solve different problems. Different machine learning approaches have been examined, such as NN, Random Forest, XG-Boost, Decision Tree, and Support Vector Machine, to train and test the model. The results show that the model can estimate the water prediction at different points. The mean absolute error is utilized to examine the algorithm, and the best one is chosen based on the computational power While in [26] Oppedijk et al. implemented short- and long-term methods to predict the waste water pumps. The method is implemented to improve and clean up the data, which is then utilized later in different models. The first model is implemented to predict how much time the station can be turned off without affecting its application, while the second one predicts the load on the pump 24 hours a day.

3. METHODOLOGY

The methodology of the proposed model is implemented based on the framework listed in Figure 1. The steps are: read data, data preprocessing, feature selection, machine learning, and model evaluation. The first step includes collecting the water pumping station data [27], visualizing it, and determining the errors in the data. In the next step, as part of data preprocessing, the description of the water pumping station data is collected from all available sensors, all of which are raw values.

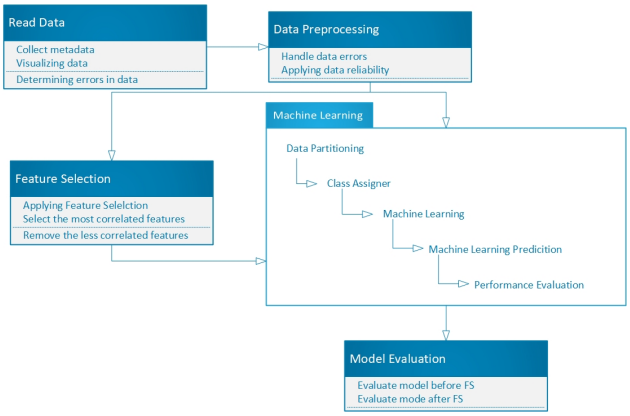

Figure 1. Methodology Framework

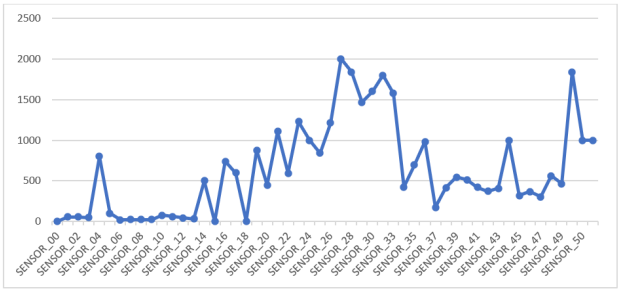

Figure 2. Reading Values of Sensors

The total number of sensors is 52, in addition to the result column (MS), hours (TimeRangeH), and minutes (TimeRangeM), where TimeRangem is the final class. The data contains a rate that contains a number of key areas in this area of final-stage stereotypes. Despite the efforts of a small team responsible for water pump maintenance in a small area outside a big town, there have been seven system failures in the past year, resulting in significant inconvenience and hardship for the local community. The team has not been able to identify any patterns in the data to pinpoint the root cause of these failures, and as a result, they are unsure where and when to focus their attention to prevent future system failures. The dataset consists of raw sensor values, timestamps, and machine status from 52 different units. The dataset can be used for research in various areas, such as industrial automation, predictive maintenance, and machine learning. The information can be used in industrial automation to track machine performance in real-time and identify any problems that might occur while it is in use. This might reduce expensive outages and boost productivity.

The dataset can be used in predictive maintenance to build machine learning models that can anticipate when a machine will fail based on sensor readings and machine statistics. This can help cut down on upkeep expenses and prevent unplanned downtime. The dataset can also be used for machine learning studies, particularly in the fields of anomaly detection and time-series forecasting. Machine learning models can be taught to spot anomalies and forecast future values by looking at trends in sensor readings and machine statistics over time. The data types of the columns are (sensors are DECIMAL, the timestamp is date, and the machine status is varchar), while the highest value of each sensor is depicted in figure 2 . The figure shows that the (sensor 00) has the smallest value (2.54), while (sensor 27) with highest value (2000).
Data processing step also involves many steps, such as uploading data from a CSV file to a MySQL server, then editing the data using Pentaho Data Integration 9.3. Next, the empty sensor values are filled with zero values and the date sections (month, day, hour, and minute). The data preprocessing also involves converting the values of the machine status field from text values to numeric values as follows: Normal to 1, Recovery to 2, Broken to 3. The SQL statement command is utilized to divide the time into labels such as the hours of the day into four sections (A, B, C, D) based on the 24-hour system:

- From 0 to 5 = A
- From 6 to 11 = B
- From 12 to 17 = C
- From 18 to 23 = D
- Each hour into four sections (Q1, Q2, Q3, Q4), each section 15 minutes long.

After that, the average was found based on the readings of the ranked attributes and the data that was less than the average value, which was (0.0013147). The step of implementing decision tree algorithms involves examining six algorithms, namely Random Tree, J48, Random Forest, Decision Stump, Rep Tree, and Hoeffding Tree. Random Forest and J48 are well-known algorithms in machine learning and are commonly used in many tasks involving regression and classification. Algorithms work widely across various fields of industries, including health, marketing, and finance. These algorithms are usually used when dealing with large amounts of data to classify them into decision tree forms, and according to the results, the proposed methodology is evaluated [28] [17].
The Random Forest algorithm is characterized by the creation and collection of several decision trees to reduce overfitting, improve quality and accuracy, and form studied forests through the use of randomly selected inputs or from a set of inputs in each node to develop them. Each tree and extraction accuracy of the forests studied compares favorably with Adaboost; this class has many advantages (strong boost, fast, useful results for error, strength, and correlation). It trains the tree on a set of random sub-data and sub-properties. When predicting, vote each tree into a more accurate and predictable category given that. The decision is made by the class with the most votes to predict the final class [28] [29].
Random forests are formed by pre-determined random features from random selection, and in each of them, one can select the most useful variables for the problem

at hand, leading to reduced data dimensionality and improved model performance [3]. The performance of the model in random forests is good with a large number of features and a small number of samples because random feature selection through the algorithm reduces the risk of providing high-dimensional data [31]. The node produces small-sized groups of variables to split. The tree is grown using the CART methodology. The main steps in the Random Forest algorithm [30]:

1. Random data is taken from the sub-data in the full data source.
2. To take advantage of the subset, several features are selected according to the decision feature.
3. Adopt the construction of the tree from the previously approved sub-data.
4. Iterate the method more than once to generate many decision trees and predict the outcome from the predictions of all decision trees.

The Random Forest algorithm has many advantages, such as being swift compared to other algorithms, having less burden, and having no objection to dealing with the lost data. It does not have the problem of dealing with noisy and missing data because each decision tree is trained on a different subset, so the missing data in a particular tree does not affect other trees, and in addition to the voting system, it reduces noise in the data [31]. The J48 is a decision tree algorithm used for data classification and mining tasks. It is based on the C4.5 algorithm [32][33], and is part of the Weka data mining software [34].

The algorithm divides the data repeatedly, depending on the attribute, into small groups with different features. In this case, the feature that provides the best is determined and divided according to the size of information acquisition for each subgroup, and this process is repeated more than once until all groups from the subgroup belong to the same group. Category Otherwise, the additional division does not provide any type of optimization [33]. The algorithm contains several parameters that can be set manually or automatically through cross-validation techniques, allowing us to determine the minimum conditions required for the leaf, knot splitting, and the required factor for pruning the tree [32].

It can handle lost and noisy data. Easily, and one of its advantages is that it also deals with multi-class problems and is sensitive to values, which allows for increasing the size of the data if the tree is large or complex [35]. The J48 algorithm can be used in a wide range of various fields, such as finance, health care, etc. For example, it is used in cardiology, stock identification, and data classification [36]. The J48 algorithm has many advantages over other algorithms. It is easy to understand, can handle missing data, and can be used for binary and multiclass classification tasks.

A low-error pruning tree, also known as a REP tree, is a decision tree algorithm widely used in machine learning for classification tasks. The algorithm works by iteratively dividing the training data into subsets based on feature values and building a tree structure that captures the decision-making process of assigning class labels to new states. The REP tree defines decision nodes by maximizing the information acquisition ratio, which is a metric that measures the effectiveness of a feature in separating instances of different classes. This approach ensures that the tree is optimally constructed for accurate mapping. To further enhance classification accuracy, the REP tree is pruned using a technique called low-error pruning. This method eliminates branches that have a negligible impact on the classification accuracy, thereby reducing the complexity of the tree and improving its overall performance. The REP Tree is a powerful decision tree algorithm that can effectively handle classification tasks in machine learning, and its use of information gain ratio and reduced error pruning make it a popular choice for many practitioners [37][38][39][40].

REP Tree algorithm takes into account the correlations between traits when constructing the decision tree, which improves its accuracy in data classification. The algorithm's use of low-error pruning allows noisy data to be addressed more effectively by removing branches that do not contribute significantly to classification accuracy [9] [41]. The study compares the performance of the REP tree across several datasets, including the Iris and breast cancer datasets. The authors reported that on all datasets, the REP Tree algorithm achieves higher accuracy and lower error rates. They also observed that the performance of both algorithms improves with increasing sample size and number of traits.

Random trees are a popular machine learning algorithm for classification and regression problems because they can handle high-dimensional data, nonlinear relationships between variables, and missing data. The algorithm works by building multiple decision trees, each of which is trained on a subset of the data and a subset of the features. The hierarchy for classifying network traffic consists of two levels. The first level uses a random tree that is trained on a subset of the most crucial features, which are selected by utilizing a genetic algorithm. The primary purpose of this tree is to categorize traffic into two distinct groups: normal and abnormal. Moving on to the second level, multiple random trees are trained on diverse feature subsets using distinct decision criteria. The second level is responsible for identifying the type of attack the abnormal traffic belongs to, such as denial of service, probe, and user-to-root [42] [43].

Random trees are versatile tools that have found applications in various fields, such as data mining, image processing, and computational biology. In data mining, random trees are frequently used for classification and regression tasks. The leaf nodes of the tree correspond to class labels or predicted values, and the path from the root to the leaf nodes defines a set of conditions on the input features that lead to the assigned label or value. In image processing, random trees can be used for image segmentation and object recognition. By training the tree on a set of images, it is possible to represent the color or

texture distribution of different regions of the image. In computational biology, random trees are often employed for phylogenetic analysis and protein structure prediction. The tree can be utilized to represent the evolutionary history of a set of sequences. Overall, the flexibility and broad applicability of random trees make them a valuable tool in various scientific fields. Random trees can be extended in various ways, such as by using different splitting criteria at different levels of the tree, by incorporating additional randomness in the splitting process, or by using ensemble methods to combine multiple trees [44] [45] [46].

The Hoeffding Tree is a decision tree learning algorithm that is designed to work in online learning scenarios where data is received in a continuous stream. It was introduced by Pedro Domingos and Geoff Hulten in 2000. The Hoeffding Tree algorithm is named after Wassily Hoeffding, who developed the concept of concentration inequalities. The algorithm uses Hoeffding's inequality to determine the sample size needed to make a decision with a high degree of confidence. The algorithm of the Hoeffding Tree has been extended and modified in several ways, including incorporating SVM and k-nearest neighbors (KNN) classifiers, using adaptive windowing, and applying kernel functions for non-linear data [47] [48] [49] [48]–[50].

The Hoeffding Tree algorithm has been applied to a variety of domains, including web and mobile applications, e-intrusion detection, e-commerce purchases, network sensors, weather prediction, social networks, bioinformatics, and many more. One potential limitation of the Hoeffding Tree algorithm is that it assumes that the data distribution is stationary over time. If the data distribution changes significantly, the algorithm may not be able to adapt quickly enough. Despite its limitations, the Hoeffding Tree algorithm remains a popular choice for online learning scenarios where data is received in a continuous stream and the model must be updated incrementally. To address this limitation, several researchers have proposed extensions to the Hoeffding Tree algorithm [50] [51] [52].

The performance of the proposed model is measured based on four main values of evaluation namely (True Positive (TP) rate, False Positive (FP) rate, Precision, and Recall). TP rate measures the proportion of positive instances that are correctly identified by the classifier. It is calculated as follow:

$$TPrate = TP/(TP + FN)$$

Where TP is the number of true positives, FN is the number of false negatives.

FP rate measures the proportion of negative instances that are incorrectly classified as positive. It is calculated as follow:

$$FPrate = FP/(FP + TN)$$

Where FP is the number of false positives and TN is the number of true negatives [53].

Figure 3. TP rate, and FP rate before Applying FS

Precision measures the proportion of true positives among the instances that are classified as positive. It is calculated as follow:

$$Precision = TP/(TP + FP)$$

Recall, also known as sensitivity or TP rate, measures the proportion of positive instances that are correctly identified by the classifier. It is calculated as follow:

$$Recall = TP/(TP + FN)$$

F-Measure, also known as F1 score, is a harmonic mean of precision and recall. It is calculated as follow [54] [55][56]:

$$F1 = 2 * (Recall * Precision)/(Recall + Precision)$$

A. *Results and Discussion*

In this section, the results of implementing a model based on machine learning algorithms will be examined before and after the FS step. This process will show the effectiveness of FS on the accuracy of the model. The first step after implementing FS is applying decision tree algorithms to compare the algorithm criteria. Figure 3 shows the performance criteria (TP rate and FP rate) after applying all six algorithms. The figure shows that Random Tree outperforms the other algorithms, followed by Rep Tree algorithm with (0.8). Hoeffding tree and Decision Stump scored (0.245) while Random Tree and J48 scored (0). Regarding FP rate, Random Tree, Random Forest, and J48 scored 0, while Rep Tree, Decision Stump, and Hoeffding Tree scored (0.067, 0.25, and 0.245 respectively).

Next, Figure 4 shows the performance criteria (precision and recall) of six algorithms before applying FS. Regarding precision, the highest precision goes to Random Tree with (1), followed by Rep Tree with (0.8). Hoeffding Tree scored (0.256) respectively, while Decision stump, Random Forest, and J48 scored 0. Regarding recall, the highest precision goes to Random Tree with (1), followed by Rep Tree with (0.8). Decision stump and Hoeffding tree scored (0.251, and 0.256) respectively, while Random Forest and J48 scored 0.

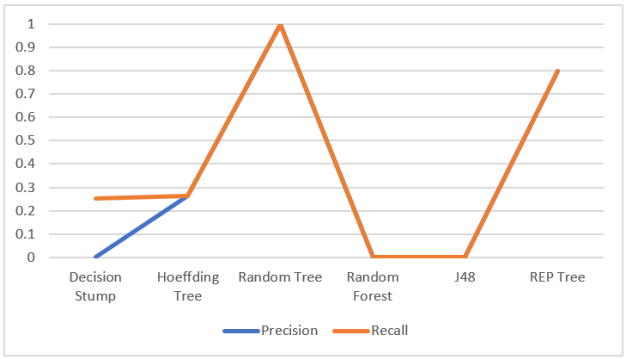

Figure 4. Precision and Recall before FS

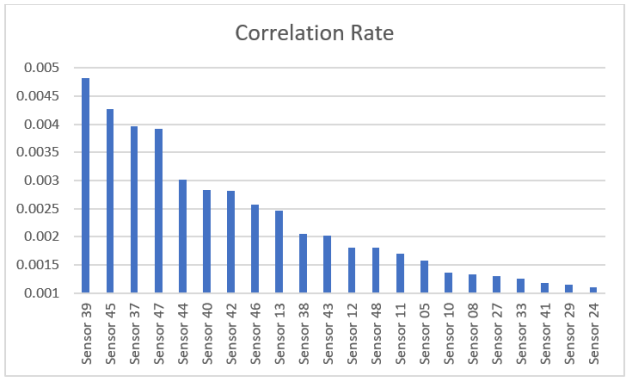

Figure 5. Features Correlation with Final Class

After that, the model's performance will be examined after applying the FS step. The FS approach (CorrelationAttributeEval) is utilized with ranker to determine the correlation between the attributes or features of a given dataset and the target variable or output. Following the operation, the average correlation coefficient value was calculated for all the features in the dataset. Based on this average value, the attributes with correlation coefficients below the average were removed from the dataset. This process aimed to eliminate features that were weakly correlated with the output variable and may not contribute much to the accuracy of the predictive model. Figure 5 shows the correlation rate of the sensor data with the final class of the pump station dataset. The correlation rate is filtered so that the less correlated features (below 0.001) are removed from the fissure, while the most correlated features (22 features) are shown in the figure. The figure shows the most correlated feature (sensor 39) that has the highest correlation (0.00458) with the final class, followed by sensors (45, 37, 47, 44, 40, 42, 46, 13, 38, 43, 12, 48, 11, 05, 08, 27, 33, 41, 29, and 24). The sensor (24) scored the lowest correlation (0.001108) with the final class.

Figure 6 shows the performance criteria (TP rate and FP rate) after applying all six algorithms. Regarding the TP rate, the figure shows that Random Tree and Random Forest outperform the other algorithms with (1), followed by J48

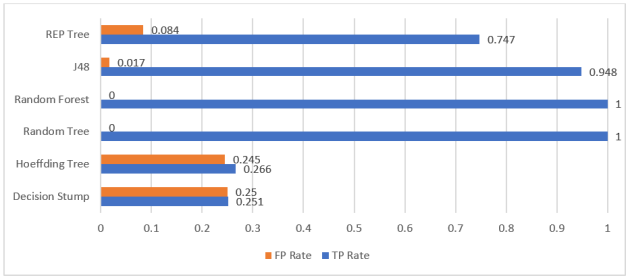

Figure 6. TP rate, and FP rate after FS

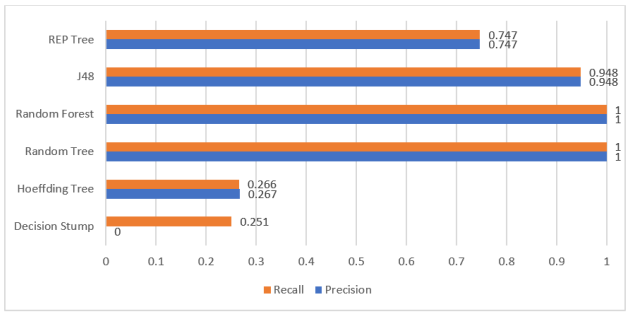

Figure 7. precision and recall after FS

with (0.948). Rep Tree scored (0.747), while Hoeffding Tree and Decision Stump scored (0.266 and 0.251), respectively. Regarding the FP rate, Random Tree and Random Forest scored the lowest values with (0), while J48 scored (0.017). Rep Tree, Hoeffding Tree, and Decision Stump scored (0.084, 0.245, and 0.25) respectively.

Next, Figure 7 shows the performance criteria (precision and recall) of six algorithms before applying FS. Regarding precision, the highest precision goes to Random Tree and Random Forest with (1), followed by J48 with (0.948). Rep Tree scored (0.747), while Hoeffding Tree and Decision Stump scored 0.267 and 0, respectively. Regarding recall, the highest recall goes to Random Tree and Random Forest with (1), followed by J48e with (0.948). Rep Tree, Hoeffding Tree, and Decision Stump scored (0.747, 0.266, and 0.251), respectively.

B. Comparison to Most Recent Related Works

The recent works and results in the field of applying machine learning algorithms with water pumping station data encourage us to give our attention to this field due to its importance and effects on different fields such as energy management, water distribution and consumption control, and water waste and consumption prediction. Different approaches with different methodologies are applied to enhance the water pump station systems, which affect the lives of citizens directly and indirectly. However, different algorithms are examined where the step of FS is not implemented. FS can be considered an important step because it can discover the most important factors that affect the work of the water pumping station, in addition to improving the accuracy of the results and the work of the station. The

results of this step can highly affect the accuracy of the prediction systems where the machine learning algorithms will be trained on data with a high correlation to the final class.

4. CONCLUSION AND FUTURE WORKS

The dataset almost holds information that can be discovered using machine learning approaches. Decision tree approaches have proven their accuracy in classifying different kinds of datasets. The accuracies are proven in the field of water pumping stations that have been conducted in this paper. Decision tree algorithms (Random Forest and Random Tree) have proven the highest accuracy in predicting and classifying sensor data from water stations. The effect of FS on the prediction accuracy shows that the accuracy is improved after applying FS. FS can be implemented to show the most corelated features (sensors) with the target sensor. This approach can be effectively used in determining the features that impact the pumping process in order to improve the work flow of the station. The results show that the model based on decision tree approaches can be utilized as a tool for classifying and predicting failures in water pumping stations. Many machine learning approaches can be implemented to find the group of sensors that affect station failure and predict the time of failure.

REFERENCES

[1] P. Nancy, S. Muthurajkumar, S. Ganapathy, S. Santhosh Kumar, M. Selvi, and K. Arputharaj, "Intrusion detection using dynamic feature selection and fuzzy temporal decision tree classification for wireless sensor networks," *IET Communications*, vol. 14, no. 5, pp. 888–895, 2020.

[2] U. Saeed, S. U. Jan, Y.-D. Lee, and I. Koo, "Fault diagnosis based on extremely randomized trees in wireless sensor networks," *Reliability engineering & system safety*, vol. 205, p. 107284, 2021.

[3] I. G. A. Poornima and B. Paramasivan, "Anomaly detection in wireless sensor network using machine learning algorithm," *Computer communications*, vol. 151, pp. 331–337, 2020.

[4] J. Xu, H. Wang, J. Rao, and J. Wang, "Zone scheduling optimization of pumps in water distribution networks with deep reinforcement learning and knowledge-assisted learning," *Soft Computing*, vol. 25, pp. 14 757–14 767, 2021.

[5] A. M. Moreno-Rodenas, A. Duinmeijer, and F. H. Clemens, "Deep-learning based monitoring of fog layer dynamics in wastewater pumping stations," *Water Research*, vol. 202, p. 117482, 2021.

[6] G. Seo, S. Yoon, M. Kim, C. Mun, and E. Hwang, "Deep reinforcement learning-based smart joint control scheme for on/off pumping systems in wastewater treatment plants," *IEEE Access*, vol. 9, pp. 95 360–95 371, 2021.

[7] J. Filipe, R. J. Bessa, M. Reis, R. Alves, and P. Póvoa, "Data-driven predictive energy optimization in a wastewater pumping station," *Applied Energy*, vol. 252, p. 113423, 2019.

[8] K. Banerjee, V. Bali, N. Nawaz, S. Bali, S. Mathur, R. K. Mishra, and S. Rani, "A machine-learning approach for prediction of water contamination using latitude, longitude, and elevation," *Water*, vol. 14, no. 5, p. 728, 2022.

[9] X. Liu, Y. Zhang, and Q. Zhang, "Comparison of eemd-arima, eemd-bp and eemd-svm algorithms for predicting the hourly urban water consumption," *Journal of Hydroinformatics*, vol. 24, no. 3, pp. 535–558, 2022.

[10] A. Predescu, C. Negru, M. Mocanu, C. Lupu, and A. Candelieri, "A multiple-layer clustering method for real-time decision support in a water distribution system," in *Business Information Systems Workshops: BIS 2018 International Workshops, Berlin, Germany, July 18–20, 2018, Revised Papers 21*. Springer, 2019, pp. 485–497.

[11] S. Alija, A. Hamoud, and F. Morina, "Predicting textbook media selection using decision tree algorithms," *Balkan Journal of Applied Mathematics and Informatics*, vol. 5, no. 2, pp. 27–34, 2022.

[12] A. Hamoud, A. S. Hashim, and W. A. Awadh, "Predicting student performance in higher education institutions using decision tree analysis," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, pp. 26–31, 2018.

[13] A. Hamoud, "Applying association rules and decision tree algorithms with tumor diagnosis data," *International Research Journal of Engineering and Technology*, vol. 3, no. 8, pp. 27–31, 2017.

[14] A. Kilic, "Artificial intelligence and machine learning in cardiovascular health care," *The Annals of thoracic surgery*, vol. 109, no. 5, pp. 1323–1329, 2020.

[15] R. Manikandan, R. Patan, A. H. Gandomi, P. Sivanesan, and H. Kalyanaraman, "Hash polynomial two factor decision tree using iot for smart health care scheduling," *Expert Systems with Applications*, vol. 141, p. 112924, 2020.

[16] I. A. Najm, A. K. Hamoud, J. Lloret, and I. Bosch, "Machine learning prediction approach to enhance congestion control in 5g iot environment," *Electronics*, vol. 8, no. 6, p. 607, 2019.

[17] M. Nabipour, P. Nayyeri, H. Jabani, A. Mosavi, and E. Salwana, "Deep learning for stock market prediction," *Entropy*, vol. 22, no. 8, p. 840, 2020.

[18] M. Nabipour, P. Nayyeri, H. Jabani, S. Shahab, and A. Mosavi, "Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis," *IEEE Access*, vol. 8, pp. 150 199–150 212, 2020.

[19] J. Karnon, "A simple decision analysis of a mandatory lockdown response to the covid-19 pandemic," *Applied Health Economics and Health Policy*, vol. 18, no. 3, pp. 329–331, 2020.

[20] P. Yariyan, S. Janizadeh, T. Van Phong, H. D. Nguyen, R. Costache, H. Van Le, B. T. Pham, B. Pradhan, and J. P. Tiefenbacher, "Improvement of best first decision trees using bagging and dagging ensembles for flood probability mapping," *Water Resources Management*, vol. 34, pp. 3037–3053, 2020.

[21] Z. Li, L. Bai, W. Tian, H. Yan, W. Hu, K. Xin, and T. Tao, "Online control of the raw water system of a high-sediment river based on deep reinforcement learning," *Water*, vol. 15, no. 6, p. 1131, 2023.

[22] S. Hu, J. Gao, D. Zhong, R. Wu, and L. Liu, "Real-time scheduling of pumps in water distribution systems based on exploration-enhanced deep reinforcement learning," *Systems*, vol. 11, no. 2, p. 56, 2023.

[23] V. Khandelwal and S. Khandelwal, "Ground water quality index

prediction using random forest model," in *Proceedings of International Conference on Recent Trends in Computing: ICRTC 2022*. Springer, 2023, pp. 469–477.

[24] A. Predescu, C.-O. Truică, E.-S. Apostol, M. Mocanu, and C. Lupu, "An advanced learning-based multiple model control supervisor for pumping stations in a smart water distribution system," *Mathematics*, vol. 8, no. 6, p. 887, 2020.

[25] K. Pałczyński, T. Andrysiak, M. Głowacki, M. Kierul, and T. Kierul, "Analysis of long-range forecast strategies for iot on urban water consumption prediction task," in *Computational Intelligence in Security for Information Systems Conference*. Springer, 2022, pp. 3–11.

[26] W. Oppedijk, N. Tiben, D. Gebbran, and T. Dragičević, "Flexibility prediction in wastewater-energy nexus using machine learning," in *IECON 2022–48th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2022, pp. 1–6.

[27] "Water pump station," Kaggle.com. [Online]. Available: https://www.kaggle.com/datasets/nphantawee/pump-sensor-data

[28] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *The Stata Journal*, vol. 20, no. 1, pp. 3–29, 2020.

[29] C. L. Devasena, "Comparative analysis of random forest, rep tree and j48 classifiers for credit risk prediction," *International Journal of Computer Applications*, vol. 975, no. 8887, pp. 30–36, 2014.

[30] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[32] M. Abdullah-Al-Kafi, I. J. Tasnova, M. Wadud Islam, and S. K. Banshal, "Performances of different approaches for fake news classification: An analytical study," in *International Conference on Advanced Network Technologies and Intelligent Computing*. Springer, 2021, pp. 700–714.

[33] S. L. Salzberg, "C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993," 1994.

[34] E. Frank, M. A. Hall, and I. H. Witten, *The WEKA workbench*. Morgan Kaufmann, 2016.

[35] D. T. Larose, "An introduction to data mining," *Traduction et adaptation de Thierry Vallaud*, 2005.

[36] B. Egüz, F. E. Çorbacı, and T. Kaya, "Stock price prediction of turkish banks using machine learning methods," in *Intelligent and Fuzzy Techniques for Emerging Conditions and Digital Transformation: Proceedings of the INFUS 2021 Conference, held August 24-26, 2021. Volume 2*. Springer, 2022, pp. 222–229.

[37] S. A. Shubho, M. R. H. Razib, N. K. Rudro, A. K. Saha, M. S. U. Khan, and S. Ahmed, "Performance analysis of nb tree, rep tree and random tree classifiers for credit card fraud data," in *2019 22nd International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2019, pp. 1–6.

[38] T. O. Olaleye, S. M. Akintunde, C. Akparanta, T. A. Avovome, O. Oluyen, and A. Akparanta, "Opinion mining analytics for spotting omicron fear-stimuli using reptree classifier and natural language processing," *International Journal for Research in Applied Science & Engineering Technology*, pp. 995–1005.

[39] S. Kalmegh, "Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news," *International Journal of Innovative Science, Engineering & Technology*, vol. 2, no. 2, pp. 438–446, 2015.

[40] W. N. H. W. Mohamed, M. N. M. Salleh, and A. H. Omar, "A comparative study of reduced error pruning method in decision tree algorithms," in *2012 IEEE International conference on control system, computing and engineering*. IEEE, 2012, pp. 392–397.

[41] V. N. Uzel, S. S. Turgut, and S. A. Özel, "Prediction of students' academic success using data mining methods," in *2018 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 2018, pp. 1–5.

[42] J.-F. Le Gall, "Random trees and applications," 2005.

[43] A. Ahmim, L. Maglaras, M. A. Ferrag, M. Derdour, and H. Janicke, "A novel hierarchical intrusion detection system based on decision tree and rules-based models," in *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 2019, pp. 228–233.

[44] L. Addario-Berry, A. Brandenberger, J. Hamdan, and C. Kerriou, "Universal height and width bounds for random trees," *Electronic Journal of Probability*, vol. 27, pp. 1–24, 2022.

[45] F. P. Lestari, M. Haekal, R. E. Edison, F. R. Fauzy, S. N. Khotimah, and F. Haryanto, "Epileptic seizure detection in eegs by using random tree forest, naïve bayes and knn classification," in *Journal of Physics: Conference Series*, vol. 1505, no. 1. IOP Publishing, 2020, p. 012055.

[46] M. R. C. Acosta, S. Ahmed, C. E. Garcia, and I. Koo, "Extremely randomized trees-based scheme for stealthy cyber-attack detection in smart grid networks," *IEEE access*, vol. 8, pp. 19 921–19 933, 2020.

[47] R. B. Kirkby, "Improving hoeffding trees," Ph.D. dissertation, The University of Waikato, 2007.

[48] S. Alija, E. Beqiri, A. S. Gaafar, and A. K. Hamoud, "Predicting students performance using supervised machine learning based on imbalanced dataset and wrapper feature selection," *Informatica*, vol. 47, no. 1, 2023.

[49] A. K. Hamoud, A. S. Hashim, and W. A. Awadh, "Areview on internet of things architecture for big data processing," *Iraqi Journal for Computers and Informatics*, vol. 46, no. 1, 2020.

[50] A. Muallem, S. Shetty, J. W. Pan, J. Zhao, and B. Biswal, "Hoeffding tree algorithms for anomaly detection in streaming datasets: A survey," *Journal of Information Security*, vol. 8, no. 4, 2017.

[51] S. Hoeglinger and R. Pears, "Use of hoeffding trees in concept based data stream mining," in *2007 Third International Conference on Information and Automation for Sustainability*. IEEE, 2007, pp. 57–62.

[52] H. M. Gomes, J. Read, and A. Bifet, "Streaming random patches for evolving data stream classification," in *2019 IEEE international conference on data mining (ICDM)*. IEEE, 2019, pp. 240–249.

[53] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.

[54] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 1, pp. 97–107, 2013.

[55] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[56] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.

Mostafa Adnan Hadi is an IT specialist with a Bachelor degree in Information Technology from the University of Basrah, Basrah, Iraq. His journey in the technical field has been marked by a genuine passion for learning and a hands-on approach to programming, particularly in Java and Python. He takes pride in achieving the third position in a university-level software competition, showcasing my dedication to my craft. Beyond academics, he has earned certifications in Project Management (PMP) and Human Resources, underlining my commitment to both technical and managerial aspects of IT.His focus lies in data collection and processing, and I enjoy exploring the latest techniques in this field. In the ever-evolving world of technology, I believe in staying ahead by embracing new trends and applying them in practical ways. He aims to contribute my skills and knowledge to the dynamic IT landscape, and I look forward to new challenges and opportunities for growth.

Alaa Khalaf Hamoud is an Assist. Prof. in Computer Information Systems, University of Basrah, Iraq. He received BSc degree from Computer Science Department, University of Basrh in 2008 with first ranking college student. He also received his MSc degree from the same department with first ranking department student. He participated in (seven months) IT administration course in TU berlin-Germany. His scientific interests are data mining, and data warehousing.

Ahmed Monther Abboud is a proficient programmer skilled in various programming languages and technology domains. Holding a bachelor's degree in computer science and information technology with a specialization in computer information systems from the University of Basrah, Basrah, Iraq. He has honed his abilities in designing and developing data analytics solutions, demonstrating a versatile understanding of the dynamic fields of computer science and information technology. His fields of interest are data mining, DBMS, and data warehousing.

Ahmed Naji Abdullah is an outstandingly skilled programmer in different programming languages. He earned his bachelor's degree in computer information systems from the University of Computer Science and Information Technology, University of Basrah, Basrah, Iraq, as one of the top students in his field in 2021. With his analytical mind, his fields of interest are data mining, data warehousing, and DBMS.

Ahmed Khaled Abdullatif is a graduate of computer science and information technology. He has a passion for technology and innovation. He holds a bachelor's degree in computer science from the University of Basrah, where he has strong experience in various programming languages, networks, and software development. I participated in many programming competitions and have many electronic certificates in various technologies, in addition to my focus on solving software problems and having a keen interest in emerging technologies.