# Classification Models Based on Association Rules for Estimation of Key Process Variables in Nuclear Power Plant

Narasimhan S[1]✉ and Rajendran Velayudham[1]

[1]Bharatiya Nabhikiya Vidyut Nigam Limited(BHAVINI) ,Kalpakkam,INDIA & VELS Institute of Science,Technology and Advanced Studies(VISTAS),Chennai, India
[2]VELS Institute of Science, Technology and Advanced Studies (VISTAS) Chennai, India
✉snsimhan@igcar.gov.in

**Abstract**

Nuclear power plant process systems have developed greatly over the years. As a large amount of data is generated from Distributed Control Systems(DCS) with fast computational speed and large storage facilities, smart systems have taken over analysis of the process. These systems are built using data mining concepts to understand the various stable operating regimes of the processes, identify key performance factors, makes estimates and suggest operators to optimize the process. Association rule mining is a frequently used data-mining concept in e-commerce for suggesting closely related and frequently bought products to customers. It also has a very wide application in industries such as bioinformatics, nuclear sciences, trading and marketing. This paper deals with application of these techniques for identification and estimation of key performance variables of a lubrication system designed for a 2.7 MW centrifugal pump used for reactor cooling in a typical 500MWe nuclear power plant . This paper dwells in detail on predictive model building using three models based on association rules for steady state estimation of key performance indicators (KPIs) of the process. The paper also dwells on evaluation of prediction models with various metrics and selection of best model.

**Keywords:** Association Rule Mining; Classification Based on Association; Classification Based on Multiple Association Rules; Classification based on Predictive Association Rules; Data Mining; Data modelling; Data Transformation; Inter Correlation; k-Means Clustering; Principal Component Analysis

## 1 Introduction

Nuclear power plants are built at high capital cost and have a long lead-in period. Due to environmental and societal considerations, their operation is also highly regulated both domestically and internationally.

They operate as base load stations at constant load factors. This calls for optimized design, engineering, erection, commissioning and operation of the systems to reduce capital costs and to increase fuel and thermodynamic efficiency so as to make power generation competitive to other fossil and conventional power generating stations. The systems are designed, engineered and operated to cover all modes of operation within the technical specifications to ensure high availability without compromising on quality, nuclear safety or environment protection. Considering these requirements, plant managers provide extensive training to plant operators, which is moreover demanded by the regulatory authorities for the purposes of granting operating licenses etc. With training, skill and experience, plant operators monitor the process conditions, interpret their observations and take appropriate actions.

Data science has played a major role in implementing concepts like Industry 4.0 and Internet of Things, as highlighted by [1, 2]. Developments in the semiconductor industry have deliveed higher speed, producing server-like computation power within small chipsets. Data science has improved the efficiency, productivity and reliability of plant without much change in the basic principles of mechanics and process systems. Data analytics concepts are now widely accepted in areas like bioinformatics, astrophysics, e-commerce, marketing, traffic management, healthcare, economics etc.

The decision making process of the operator can be made more effective by using computerized tools and models using data science concepts on the large amount of data being created and stored in the computers and servers of the Distributed Control Systems (DCS). These models will enhance the operator's understanding of the process and its dynamics; assist him to predict values so that s/he can take timely action to improve performance.

Association rule mining is a frequently used data-mining concept in e-commerce for suggesting closely related and frequently bought products to customers. It also has a very wide application spread among industries such as bioinformatics, nuclear sciences, trading and marketing. Association rule mining identifies interesting relations between variables in a big database. The algorithm identifies frequent patterns in the data. Based on the frequent patterns it identifies strong rules using some measures of interestingness in the form that "if A happens, then B is likely to happen".

Application of this data mining tool is demonstrated in this study of a representative nuclear power plant process which takes full advantage of readily available plant data. The study helps to identify key performance indicators (KPI) and estimate their optimum values using three classification models based on associations: Classification Based on Association (CBA), Classification Based on Multiple Association Rules (CMAR) and Classification based on Predictive Association Rules (CPAR).

The objectives of this study are to deploy these concepts in the process system to gain a better understanding of the operating modes and to estimate the values of key parameters to prompt operation personnel to adhere to the optimum region of operation. In addition, this study will also aid predictive maintenance and hence result in improved reliability.

Of the many critical process systems of a nuclear reactor, a lube oil system was chosen for the case study to demonstrate the implementation of a new computation concept in nuclear systems that can be adopted for safety systems after due verification and clearance by regulators. The lubrication system is a non-nuclear system, which is important for equipment safety and at the same time facilitates all the possibilities of process system analysis and implementations to an extended level.

The study involves performance optimization analysis in steady state operating conditions. It primarily focuses on identification of various operating regimes, identification of key performance indicators and estimating their values while tweaking the process for enhancement of efficiency. This paper will dwell in detail on predictive model building using association rule mining models for steady state estimation of key performance indicators (KPIs). The paper also dwells on evaluation of these prediction models with various metrics and selection of the model through Resource Operating Characteristics (ROC).

## 2 Literature review

The field of process optimization and development of soft sensing technologies has been a subject of research for many years and there are numerous papers in this area specifically using the data science approach. Various publications deal with monitoring the performance of the process using sensor data analysis. Various models are proposed using several statistical, mathematical and data modelling techniques.

[3] dwells on the two fundamental requirements in a process optimization methodology: performance monitoring and performance improvement. The paper demonstrates the models based on system identification in two chemical engineering applications including a bio diesel process. The author dwelt on performance monitoring of industrial control loops with a focus on the stiction phenomenon, the most common control valve fault. The paper proposes two models for stiction diagnosis based on the numerical optimization and transformation of the industrial dataset. The author demonstrated use of proper data filtering in the presence of noise for correct stiction detection. For process performance improvement the author proposed two models based on PID controller tuning and development of soft sensors using data modelling.

Soft sensors generate new information that is not readily available from on-line instrumentation or laboratory measurements, predict the quality attribute of interest to minimize measurement delays and enable quick control actions. The authors also compared the prediction capability of several modelling techniques using training and validation datasets.

[4] reiterates that since data-driven soft sensors are based on the data measured within the process, they can describe the true conditions of the process better than model-driven soft sensors. The most popular modelling techniques applied to data-driven soft sensors are: Principal Component Analysis combined with a regression model, Partial Least Squares, Artificial Neural Networks, Neuro-Fuzzy Systems and Sup-

port Vector Machines.

[5] focused on newer standards of power plant monitoring and control using Distributed Control Systems (DCS) and reported on their effectiveness in operations. However, during upset or transient conditions, the DCS is flooded with many alarm signals making it complicated and difficult to access critical parameters for monitoring plant conditions. Data mining techniques are best suited in a situation where most of the process data are highly correlated and exhibit multi collinearity. Data mining is used to develop models from data during process steady state conditions. These models establish relationships among variables under normal conditions and these discovered relationships are used to identify unusual conditions from the trained behavior. The paper demonstrates the use of principal component analysis and partial least squares techniques on the data available in the historian server of DCS, and proposes a process-monitoring algorithm.

[6] presents the theoretical background of different modelling approaches in condition monitoring applications. The paper evaluates commercial condition monitoring software based on recursive neural networks. Several models were developed and evaluated with different training parameters during healthy and fault detection phases of operation. The study demonstrated that we could train a reliable model capable of automatic fault detection using principal component analysis and correlation analysis.

[7] proposes a method of optimizing process operation at a 300MW power plant based on data mining techniques using real-time data acquired by the Plant Data Acquisition System (DAS). The model thus developed uses fuzzy association rule mining to find optimum values of the process from quantitative data at the power plant. The optimal values guide operators online toward improved performance at the power plant.

[8] describes the research carried out to optimize the performance of steam generation systems in thermal power plants by applying data mining techniques. The aim of the research is to develop models for performance monitoring of the plant for the full range of operating conditions. The model was developed using association rule mining in the historical data of a power plant to ascertain the behavior of plant components and to determine relationships between them for the plant as a whole. It was ascertained that these models are much more accurate than empirical ones determined from the design of the component. Further, they can be used in conjunction with a suitable expert system that will determine deviations from normal plant operations and suggest suitable correction strategies to operators.

[9] highlights the big role played by data mining and analytics in the process industry over several decades. The paper evaluates the existing data mining and analytics applications in the process industry from a machine learning perspective through unsupervised and supervised learning algorithms.

[10] deals with performance monitoring of redundant process sensors for process optimization in a nuclear power plant by developing a combined online early warning signal for the operator on the requirement for calibration checks of the redundant sensors. This paper dealt with the development of a drift monitor through data mining techniques, using mahalanobis distance metrics for a representative process in a nuclear power plant.

All the above studies are oriented toward optimizing performance of processes using various data modelling techniques in power plants in general. In contrast, this paper deals with the study involved in performance optimization analysis in the steady state operating conditions specific to a nuclear power plant. It primarily focuses on the identification of various operating regimes and the identification of associated and correlated variables and their values in each of the operating regimes. In this paper, the data mining techniques are dealt with in detail for performance improvement of the process using a large amount of data collected in the DCS. Dimensionality reduction and data transformation exercises are carried out using multi-collineartity analysis and principal component analysis, which assisted in more reliable model building. The operating regimes are identified using k-means clustering. The time series data is converted to a transaction data type and the associations are identified and pruned with thresholds of significance to identify the key parameters of the process. No soft sensor is modelled in this study. Nevertheless, optimization techniques are developed to identify key performance indicators of the process using association rule mining.

This paper continues the study carried out by [11] where the details of identification and optimized process values of key performance indicators for the chosen process of a nuclear power plant are determined and presented in detail. This paper demonstrates in detail the process of data model building for prediction of values of chosen KPI under changes in other variables using classification models based on association rules. Different stages of the data modelling like data preprocessing, exploratory data analysis, cluster analysis and association rule mining are discussed so that the study is complete and fully presented in the paper, leading to meaningful comparisons that can be made with respect to process optimization by subsequent research. The paper also presents various evaluation metrics for model evaluation and selection.

## 3 The process system

The process system taken for the case study is the oil lubrication and cooling system of a 2.7 MW centrifugal pump commissioned for circulating reactor coolant, as shown in Fig. 1. The pump is used to circulate molten sodium in a loop for extracting heat from a reactor core and feeding the heat to steam generators.

The pump is housed in a fixed shell called a pump tank, having a free liquid level. The space above the liquid is filled with inert cover gas. The pump shaft is guided by a hydrostatic bearing at the bottom and a thrust bearing at the top. Three mechanical seals are provided to prevent leakage of cover gas into the atmosphere. Oil circulates through the bearings and mechanical seals, to provide cooling, lubrication and sealing for the pump shaft assembly. The leakage oil from the bearings and mechanical seals is collected in three catch pots in tandem. The oil collected in the catch pots is drained through individual separate circuits to the leakage collection tank.

Two independent oil circuits are provided to circulate the oil; one for the bottom mechanical seal, the other for the top bearing and mechanical seal assemblies.

Each circuit has two reciprocating pumps for circulating the oil, two filters for cleaning and two blowers for cooling the oil. The pumps, filters and blowers are operated in a redundant configuration so one of the two will always be operating and the other one will be on standby.

The blower cools the oil after it extracts the heat from the seals. The blower is controlled by an electronic controller for the purpose of maintaining the temperature of the mechanical seals of the main centrifugal pump at around 55°C. Alarms and trip are provided in case the oil temperature exceeds the set points. The main centrifugal pump is designed to trip on high seal oil temperature and low oil flow/low oil pressure, which may subsequently lead to reactor trip.

Of the two circuits, the bottom mechanical seal oil circuit is considered for this analysis.

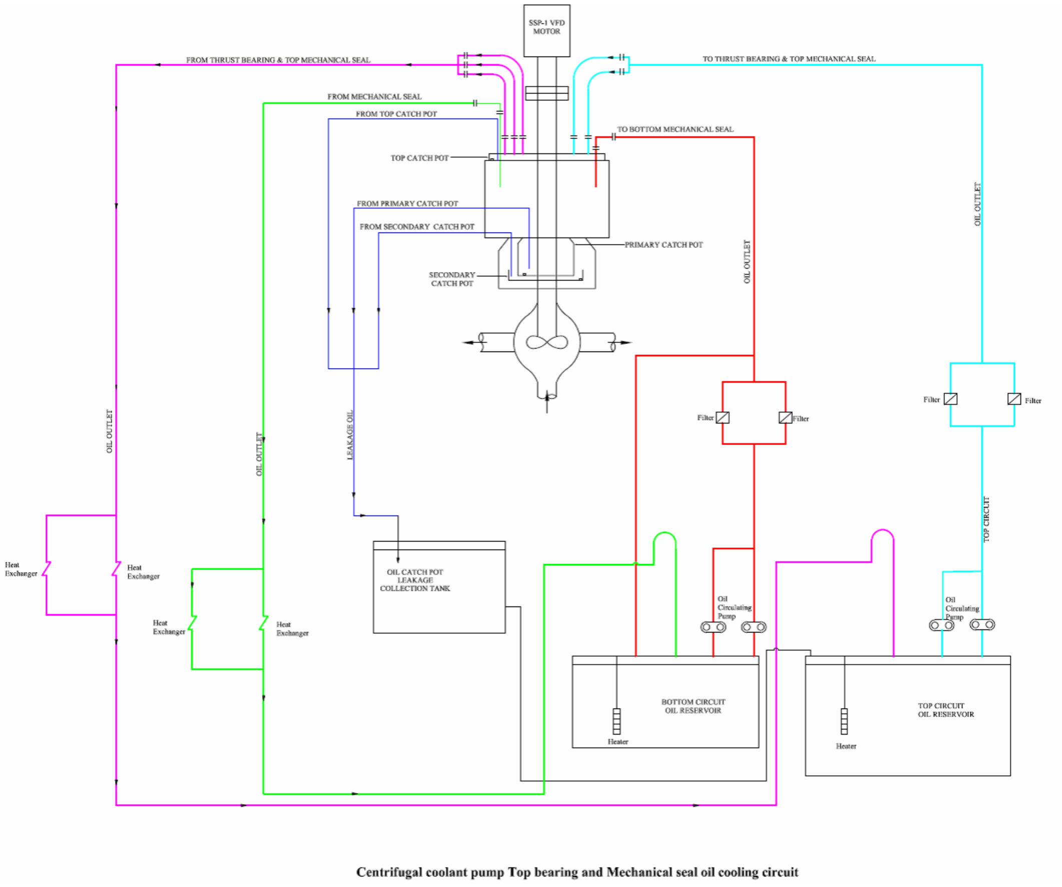

Figure 1: Oil Cooling Circuit for Bottom Mechanical Seal

## 4 Data preprocessing

The instrumentation of the system consists of both analog and digital sensors to monitor various parameters of the bottom circuit oil cooling system.

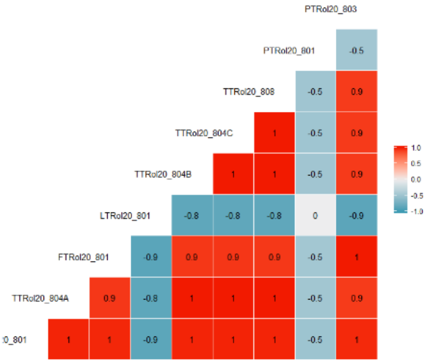

Figure 2: Correlation among variables

The DCS scans the sensors, collect the data and stores in the server for presentation to the operator in mimic form. The data is generated every five seconds and stored in the redundant process computer.

Raw data available in the process computer is converted to a spreadsheet file and then imported to a data analysis tool 'R'. Around 400,000 observations were extracted from the server for the period of one month and analyzed. There were some anomalies in the data like missing values, out of range values due to sensor unavailability, sensor unable to connect to the distributed control system, issues in database server & historians, and mismatch in connections etc., which were excluded from data processing. In addition to this, there are some digital signals which represent the healthiness of the system showing high all the time, and this will not give us any information for analysis purposes. These signals were also excluded. As a result, we were left with approximately 400,000 samples having nine feature sets that correspond to the analog measurements of temperatures, pressure, level and flow at various locations of the process system for this particular analysis, as listed in Table 1.

Table 1: Feature sets for Analysis

| Parameter ID | Parameter Description | Range of Measurement |
|---|---|---|
| TTRol20_801 | Oil inlet temperature | 0-600°C |
| TTRol20_804A | Oil outlet temperature | 0-600°C |
| TTRol20_804B | Oil outlet temperature | 0-600°C |
| TTRol20_804C | Oil outlet temperature | 0-600°C |
| TTRol20_808 | Return oil temperature after cooling by blowers | 0-600°C |
| PTRol20_801 | Oil Filter Differential Pressure | 0-7 kg/cm2 |
| PTRol20_803 | Oil inlet pressure | 0-7 kg/cm2 |
| LTRol20_801 | Bottom Circuit Oil tank Level | 0-500 mm |
| FTRol20_801 | Oil outlet flow | 0-50 m3/hr. |

## 5 Exploratory data analysis

A key tool in data mining is data exploration. Visual inspection of the data itself can provide some observations. Further, detailed quantitative measures are analyzed by deploying various exploratory tools. Such tools will produce important analytical inferences with minimum knowledge about the complex process system, so that one can make meaningful and realistic inferences just by exploring the data.

As can be seen from Fig. 2, the inter correlation study of the variables indicate that the temperature, pressure and flow of oil on both inlet and outlet of the seals are interdependent and hence exhibit multicollinearity. Further, the range of measurements for each type of process measurements are different, motivating the use of principal component analysis to eliminate the effect of multicollinearity by dimensionality reduction and orthogonal transformation.

To find out the right number of principal components, we calculate the eigenvalues for each of the principal components. It is observed from Tables 2 and 3 that 93.84% of the variances can be represented by just three principal components.

With this, the amount of data was reduced to 33% without affecting the individual variances, removing the multicollinearity completely. To optimize the components, we use orthogonal rotation methods.

Each component represents a set of variables according to their weight on the component. PC1 explains variances of all temperature measurements. PC2 explains the variances of pressure and flow in the process and PC3 explains the variances of all pressure sensors. We use these three components for clustering and plot in a 3D space for understanding the distribution of data as shown in Fig. 3.

Table 2: Major Principal Components with Eigen Values

| Principal Components | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Eigen Values | 6.517 | 1.156 | 0.773 | 0.439 | 0.087 | 0.017 |

Table 3: Principal Components and Variances Explained

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 |
|---|---|---|---|---|---|---|
| ss loading | 6.52 | 1.16 | 0.77 | 0.44 | 0.09 | 0.02 |
| Proportion variance | 0.72 | 0.13 | 0.09 | 0.05 | 0.01 | 0 |
| Cumulative variance | 0.72 | 0.85 | 0.94 | 0.99 | 1 | 1 |

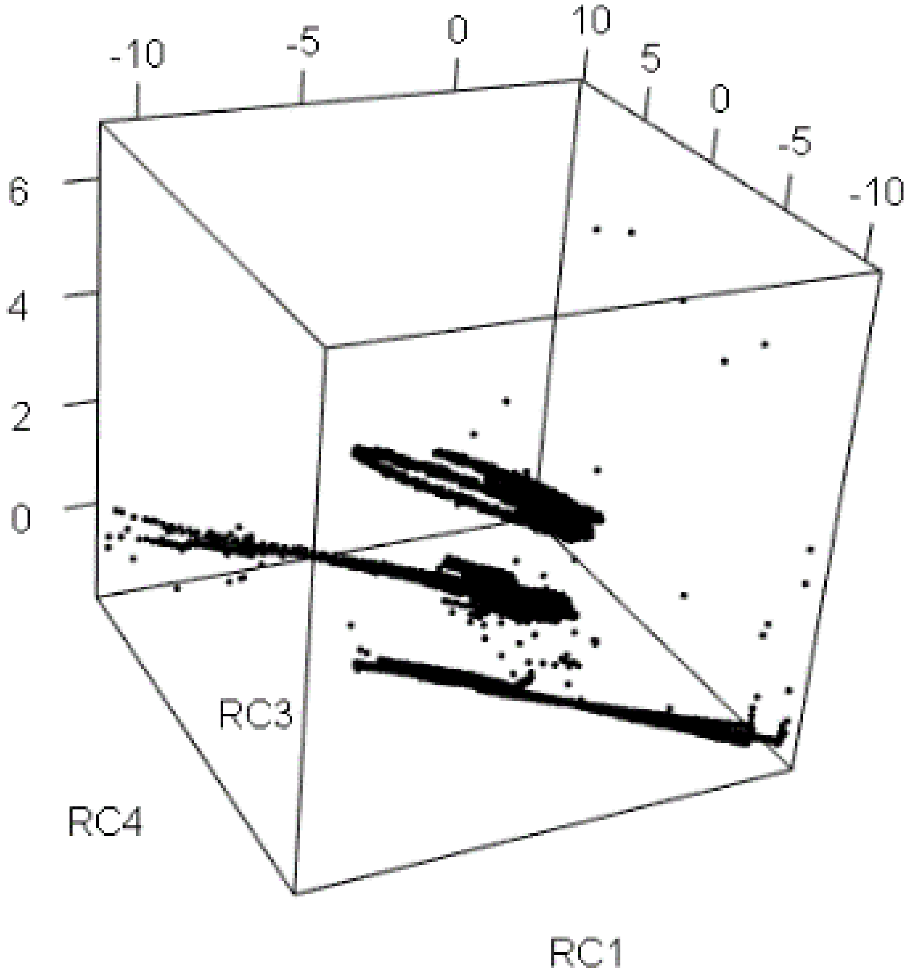

Figure 3: Plot of Principal Components

## 6 Clustering

Clustering is an important technique employed in data mining to identify data points that are similar in nature. For a process system, similar data points mean the operating region of the instances is similar in nature. This will help to distinguish the modes of operation as well. We normalize the data before deploying the clustering algorithm to make sure that all variables are treated on the same scale. The distances between each data point are calculated and based on the distance the data points are grouped, so that the inter-cluster distance is maximum and the intracluster distance is minimum.

We use K means clustering in this study as the computational power required and complexity is minimum. The normalized data points are distributed into an 'N' dimensional space and random centroids are generated in it. During the first iteration, the distance from each point to each centroid was calculated. Minimum distance to centroid will acquire the data point for the corresponding cluster. Then the centroid will be recalculated by taking the average of all its data points. After several iterations with the centroids remaining unchanged, we obtain the final cluster distribution of the data. Generally, the number of clusters will be decided by eigenvalues. From the 3D plot of the principal components in Fig. 3 it was evident that the number of clusters shall be three. K- means clustering was done by assigning three random centroids and after 50 iterations, the best fit was taken by considering the minimum intra-cluster and maximum inter-cluster distances. After clustering, it was found that there were 2 major clusters and 1 minor cluster with 197900,205855 and 10975 data points respectively. Three operating regions of the process system were identified, as in Table 4.

Table 4: Clusters for Principal Components

| Parameter | | Cluster-1 | Cluster-2 | Cluster-3 |
|---|---|---|---|---|
| Cluster Centres | PC1 | 0.0853 | -3.942 | 0.128 |
| | PC2 | -0.037 | 4.485 | 0.274 |
| | PC3 | 1.032 | -0.807 | -0.949 |
| Cluster Size by number of data sets | | 197900 | 10975 | 205855 |
| Custer size in % | | 47.718 | 2.646 | 49.636 |

## 7 Association rule mining

Association rule mining is a data science algorithm used to identify interesting relations between variables in a big database. The algorithm identifies frequent patterns in data. Based on the frequent patterns, it identifies strong rules using some measures of interestingness in the form that "if A happens, then B is likely to happen".

This is a very frequently used data mining concept in e-commerce for suggesting closely related and frequently bought products to customers. Nevertheless, it has a very wide application spread among industries such as bioinformatics, nuclear sciences, trading and marketing. As presented by [12],

Let $\mathbf{I} = \{\mathbf{i_1,i_2,i_3,.............i_n}\}$ be a set of **n** binary attributes called items. Let $\mathbf{D} = \{\mathbf{t_1,t_2,t_3,.........t_m}\}$ be a set of **m** transactions called database.

Each transaction in **D** has a unique transaction ID and contains a subset of the items in **I** . The association rule is built as

$A \to B\ where\ A \neq 0, B \neq 0, A \cap B = \phi$

$A$ called antecedent or left-hand-side (LHS) and $B$ consequent or right-hand-side (RHS). Interesting rules from the set of all possible rules are selected based on minimum thresholds on support, confidence and Lift. Rules are considered strong and eligible if their support and confidence values are higher than the thresholds,

The support value $s$ for the rule is given by the percentage of transactions in **D** that contain $A \cup B$(i.e., the union of sets $A$ and $B$ say, or, both $A$ and $B$). This is taken to be the probability $P(A \cup B)$.

The confidence $c$ for the rule $A \to B$ is given by the percentage of transactions in **D** containing $A$ that also contain $B$. This is taken to be the conditional probability, $P\left(\frac{B}{A}\right)$.

Lift is the measure of correlation between item sets $A$ and $B$. The occurrence of item set $A$ is independent of the occurrence of item set $B$ only if $P(A \cup B) = P(A)P(B)$; otherwise item sets $A$ and $B$ are dependent and correlated as events. The lift between the occurrence of $A$ and $B$ can be measured by computing the ratio between $P(A \cup B)$ and $P(A)P(B)$.

If the lift value is less than one, then the occurrence of $A$ is likely to lead to the absence of $B$. If the value is greater than one, then $A$ and $B$ are positively correlated. If the value is equal to one, then $A$ and $B$ are independent and there is no correlation between them.

To apply frequent pattern mining in the process under study, the process system time series data is converted into a transactional item like structure by converting the continuous data to categorical factors and slicing it to 25 bins and naming it as cards. The database is now converted to a transactional type with various item sets to facilitate association rule mining.

Each cluster is considered as a store, each variable is considered as different sections and each bin is considered as a card/item from the section. Association rule mining was done for each cluster separately by considering each row as a transaction data with item sets from each variable/section.

Association rule mining is done and rules with maximum lift were considered for each cluster individually. In addition, the rules were plotted in Fig. 4,5,6 after pruning the confidence and support to a minimum value of 0.4 as there will be so many rules that are insignificant.

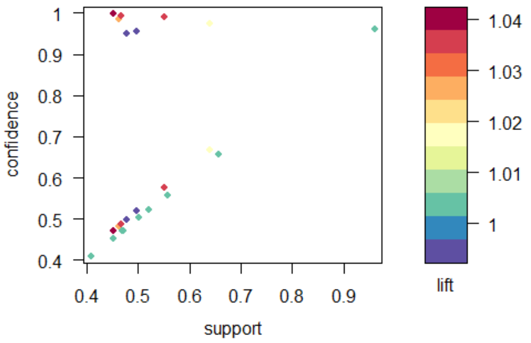

Figure 4: Rules Distribution for Cluster-1

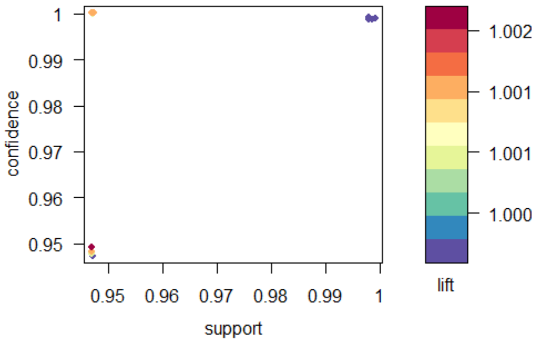

Figure 5: Rules Distribution for Cluster-2

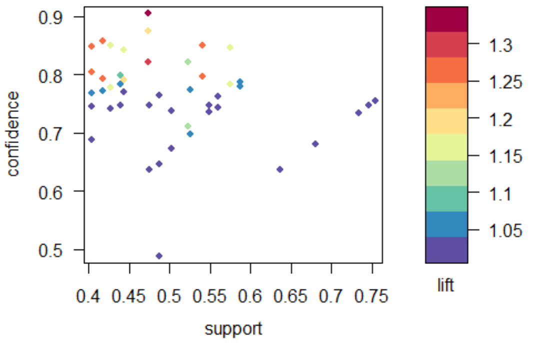

Figure 6: Rules Distribution for Cluster-3

# 8 Key performance indicators

Association mining has resulted in a set of rules based on frequent patterns. From the very strong rules with high support, confidence, lift, and variables, which are part of these rules, the key performance indicators were identified as follows:

PTRoI20_803 Oil inlet pressure to bottom mechanical seal

TTRoI20_801 Oil inlet temperature to bottom mechanical seal

FTRoI20_801 Oil outlet flow from mechanical seal

TTRoI20_808 Return oil temperature after cooling by blowers

These variables play a major role in the performance of the system. Any change in value of one of the variable affects the others. Hence, these variables can be considered key parameters for the process. By controlling these parameters, the process can be tuned to provide optimized performance.

# 9 Model Building

The objectives of this work are to know the operating regions and to understand the behavioral characteristics among the variables. There are various machine-learning models available for building predictive models for such systems.

In the present study, since the continuous time series data are transformed to a transactional type with categorical attributes and association rules are identified to understand the key parameters, these rules are extended further to develop a model for prediction. As presented by [12], the association rule classification method follows the procedure below:

1. Identify frequent item sets by association rule mining.
2. Generate association rules based on the frequent item sets that satisfies the threshold criteria for support and confidence.
3. Develop a rules-based classifier by organizing the association rules.

Performance of the associative classification models is based on the method of mining the frequent item sets and the method of analysis of the mined rules that are used for classification. The various models applied in the study are as follows:

## 9.1 Classification Based On Association

Classification Based on Association (CBA) follows an iterative approach by making multiple passes on the data, to identify frequent item sets similar to the Apriori algorithm. The length of the rules mined determines the passes. It is equal to the length of the longest rule mined. The rules satisfying minimum confidence and support thresholds are then ordered on decreasing precedence on the values of confidence and support. If more than one rule have the same antecedent, then we select the rule with the highest confidence to represent the set. A new data tuple is classified based on the first rule satisfying the tuple. In case a new tuple does not satisfy any of the rules, the classifier assigns a default class for the tuple based on a default rule with the lowest precedence. Thus, we form a decision list with the set of rules.

## 9.2 Classification Based on Multiple Association rules

As elaborated by [13], Classification based on Multiple Association Rules (CMAR) carries out frequent item sets mining using the FP-Growth algorithm using a tree structure called FP-tree. Hence, only two passes are required on the data to identify the frequent item sets and generate rules that satisfy the threshold conditions of support and confidence. CMAR also develops a tree to prune the rules, store and retrieve rules efficiently. The rules are pruned based on confidence, correlation and coverage. We prune special rules with low confidence if general rules with high confidence are available. Further, we also prune the rules for which the antecedent and the class are not positively correlated.

The identified rules are grouped based on the class labels. All rules within a group share the same class label and each group has a distinct class label.

For a new data tuple, unlike CBA, which classifies based on the most confident rule that satisfy the tuple; CMAR calculates a weighted $\chi^2$ value for each group. It then assigns the class label of the group that satisfies the tuple with the highest $\chi^2$ value.

CMAR is more accurate than CBA. It is also more efficient in terms of runtime, scalability and memory usage.

### 9.3 Classification Based on Predictive Association rules

Classification based on Predictive Association Rules (CPAR) as deliberated by [14] differs from CBA and CMAR on the rule generation algorithm. It generates rules to distinguish preferred class tuples from all others. For every rule generation, we remove the data tuples of the preferred class that satisfies the rule. This iteration is continued until all the tuples of the preferred class in the data set are covered and all the classes are covered. Hence, fewer rules are generated compared to CBA and CMAR. Classifier rule sets are formed with the generated rules based on the class labels. The rules are ordered according to their Laplace accuracy.

For a new data tuple, CPAR will use the best k rules of each group to predict the class label for the expected accuracy. Thus we eliminate the influence of lower rank rules in decision-making. CPAR is comparable to CMAR in terms of accuracy. However, CPAR is faster and efficient for a large data set, since the rule sets are smaller than CMAR.

### 9.4 Model Deployment

The models were built using R code version 3.6.3 for the total data collected for this study. However, developing, training and testing the model to estimate the accuracy using the same data set will result in misleading values. The developed model is trained with the given data and if the same data is again used for testing, the estimates obtained will be optimistic due to over specialization of the model to the data.

Table 5: Model Output parameters

| ModelParameters | | TTRol20_808 | TTRol20_801 | FTRol20_801 | PTRol20_803 |
|---|---|---|---|---|---|
| | Absolute minimum support count | 14409 | 14409 | 14409 | 14409 |
| CBA | Rules before pruning | 17 | 16 | 17 | 17 |
| | Rules after Pruning | 5 | 2 | 9 | 6 |
| CMAR | No.of classes | 14 | 13 | 13 | 14 |
| | No.of rules | 190 | 249 | 240 | 261 |
| CPAR | No.of classes | 14 | 13 | 13 | 14 |
| | No.of rules | 131 | 125 | 81 | 91 |

**Holdout** method [12] is adopted for building and testing the model by randomly dividing the data set into a *training set* and a *test set*. The model is built using the training set and the model's accuracy is tested with the test set. This estimate will pessimistic, since the model is built only with a portion of the data.

Hence, a complete dataset of 205855 data tuples that were converted to a transactional database for association rule mining was split into training data set and testing data in the ratio of 70:30. Then the model was built on the training set for each of the parameters: oil outlet flow from mechanical seal (FTRol20_801), oil outlet temperature from the coolers (TTRol20_808), oil inlet pressure to the mechanical seal (PTRol20_803) and oil inlet temperature to the mechanical seal (TTRol20_801).

The CMAR and CPAR models were built using LUCS_KDD implementation tool as [15]

The CBA model was implemented with the R package 'arulesCBA' as per [16].

Both CBA and CMAR models are built with the thresholds of 0.1 for support and 0.5 for confidence. CPAR model is set to use the best of five rules for classification. The parameters of the model output are as per Table 5.

The predictions were made on the test data set to evaluate model performance for each variable.

## 10 Model Evaluation

Evaluation of a model is an important step in ascertaining its suitability for the given problem and efficiency on performance. Various metrics are derived for evaluating model performance for both balanced and imbalanced classes of data sets.

In a multiclass environment, evaluation is carried out for each of the classes. Positive tuples (P) indicate the tuples with main class of interest and all other tuples are Negative tuples (N).

Considering that the model is evaluated on a labelled test data set, P indicates the quantity of positive tuples and N indicates the quantity of negative tuples. For each tuple, the prediction of class label by the model is compared with the actual class label of that tuple. Further to this, four more measures are to be calculated for computing the evaluation metrics.

**True positives (TP)** : The term refers to the quantity of positive tuples that were correctly classified by the model.

**True negatives (TN)** : The term refers to the quantity of negative tuples that were correctly classified by the model.

**False positives (FP)** : The term refers to the quantity of negative tuples that were incorrectly classified as positive.

**False negatives (FN)** : The term refers to the quantity of positive tuples that were incorrectly classified as negative.

While true metrics such as TP and TN indicate how the model is predicting correctly, the false metrics FP and FN indicate wrong predictions. All these measures are indicated in the Confusion Matrix drawn up for the purpose of evaluating the performance of the model, as in Table 6. A **confusion matrix** is a table of representation of actual and predicted classes. An entry $C_{i,j}$ in any cell of the confusion matrix tells the quantity of data tuples with actual class$i$ which was estimated by the model as class $j$ . The accuracy level of the model is indicated by the quantities along the diagonal cells (TP&TN): the higher the diagonal quantity, the higher the accuracy.

For an ideal case, the accuracy is 100% when TP+TN is equal to P+N and FP=FN=zero.

The confusion matrix in Fig.6 helps us evaluate the capability of a model in recognizing tuples of various classes.

Table 6: Confusion Matrix

| | | | Predicted Class | |
|---|---|---|---|---|
| | | Positive | Negative | Total |
| Actual Class | Positive | TP | FN | P |
| | Negative | FP | TN | N |
| | Total | P' | N' | P'+N' or P+N |

The general performance metrics for any prediction model is the accuracy and error rate. As per [12], the **accuracy** of a model for a given test data is defined as the percentage of test data tuples that are correctly classified by the model. That is,

$$\text{Accuracy (in \%)} = \frac{(TP+TN)}{(P+N)} \times 100$$

This metric can also be called the **recognition rate** of the model.

The **error rate** or **misclassification rate** of model is defined as the ratio (in percentage terms) between test data tuples that are incorrectly classified and the total test data tuples.

$$\text{Error Rate (in \%)} = \frac{(FP+FN)}{(P+N)} \times 100$$

Normally, the accuracy and error rates are the most effective metrics when the distribution of the classes are relatively balanced.

With the accuracy values calculated, a **kappa** value indicates how well the model is predicting compared to random predictions. The metric indicates the accuracy of the system compared to random accuracy.

$$Kappa = \frac{(Total\ Accuracy - Random\ Accuracy)}{(1-Random\ Accuracy)} \quad \text{where}$$

$$Random\ Accuracy = \frac{(NN' + PP')}{(P+N)(P' + N')}$$

The **no information rate** or **the prevalence** of model is defined as the largest class percentage in the data. The idea is that a useful model should do better than you could do by always predicting the most common class.

$$\text{No Information Rate} = \frac{P}{(P+N)}$$

Other than these global metrics, various other metrics are derived for each of the class labels as per Table 7.

The suitability of a given metrics depends on the type of problem, class prevalence and application of the model.

Table 7: Model Metrics

| Model Metrics | Description | Derivation |
|---|---|---|
| Sensitivity/Recall/True Positive Rate(TPR) | True Positive (recognition) Rate/Measure of Completeness | TP/P |
| Specificity/Selectivity/True Negative Rate(TNR) | True Negative (recognition) Rate,Measure of Completeness | TP/N |
| Precision/Positive Prediction value(PPV) | When the prediction is positive,how often is it correct(Exactness) | TP/(TP+FP) |
| Negative Predicted Value(NPV) | when the prediction is negative,how often is it correct | TN/(TN+FN) |
| Miss Rate/False Negative Rate(FNR) | probability of false positive prediction | 1-Sensitivity |
| Fallout/False Positive Rate(FPR) | probability of missing a genuine class | 1-Specificity |
| Prevalence | Class percentage in the data | P/(P+N) |
| Balanced Accuracy | Average of sensitivity and selectivity | (Sensitivity+Specificity)/2 |
| F1-Score | Harmonic average of precision and recall | (2*Precision * Recall)/(Precision+Recall) |

### 10.1 Model Evaluation under This Study

The various metrics of the models in this study are calculated for each of the classes using the R package 'caret' as per [17]. The results of the summary metrics are tabulated in Table 8.

The results on summary metrics indicate that:

1. The global accuracy of all the models is in the range 71% to 80% except in the case of CPAR.
2. CPAR model exhibits the lowest accuracy of 21.51% for the parameter TTRol20_801.
3. CMAR model exhibits comparatively higher accuracy levels for all parameters compared to the other two models.
4. Kappa values for all models for parameters TTRol20_808 and PTRol20_803 indicate moderate agreement and performance.
5. All three models exhibit a very low kappa value for the parameters TTRol20_801 and FTRol20_801 indicating that these parameter values have higher variance compared to the other parameters.

The other quality metrics for the most prevalent class for each parameter are tabulated in Table 9.

The results of these metrics indicate that:

1. The performance of all three models are similar for the parameter TTRol20_808.
2. Both CBA and CMAR models exhibited highest sensitivity for the parameters TTRol20_801 and FTRol20_801.The sensitivity of CPAR model is lower than the other models. Specifically, CPAR model exhibits very low sensitivity for the parameter TTRol20_801 for the most prevalent class.
3. Both CBA and CMAR models exhibit poor selectivity for the parameters TTRol20_801 and FTRol20_801. CPAR is a good model for all parameters in terms of selectivity.
4. CPAR is more precise in classification compared to the other two models. The model has the highest precision value for PTRol20_803.
5. CMAR outperforms the other two models in the precision of predicting negative classes, with its NPV value higher than other models for all parameters.
6. CBA did not predict any negative class for any of the tuples in predicting parameter TTRol20_801. The model is highly selective for that parameter. CPAR is poorer than the other two models in this aspect.
7. Both CBA and CMAR models show high type-1 error (false positive rate) for the parameters TTRol20_801 and FTRol20_801. CPAR shows high type-2 error (false negative rate) for the same parameters.
8. The balanced accuracy for the prevalent class is higher for CPAR compared to the other two models. The accuracy values are also consistent for all parameters in the range 0.7 to 0.8. However, both CBA and CMAR models show lower accuracy values for the parameters TTRol20_801 and FTRol20_801 in the most prevalent class. Their accuracy values range from 0.5 to 0.78.
9. The F1 score for the CBA model is in the range 0.82 to 0.85 for all parameters. This shows the CBA model is performing consistently in the estimation of all parameters. However, both CMAR

Table 8: Model Summary Metrics Values

| Variable | Model | Accuracy | CI interval | Kappa value |
|---|---|---|---|---|
| TTRoI20_808 | CBA | 0.7621 | (0.7587, 0.7654) | 0.4861 |
| | CMAR | 0.7638 | (0.7604, 0.7671) | 0.491 |
| | CPAR | 0.7634 | (0.7601, 0.7668) | 0.4901 |
| TTRoI20_801 | CBA | 0.7415 | (0.738, 0.7449) | 0 |
| | CMAR | 0.7463 | (0.7428, 0.7497) | 0.0373 |
| | CPAR | 0.2151 | (0.2118, 0.2183) | 0.0331 |
| FTRoI20_801 | CBA | 0.7348 | (0.7313, 0.7383) | 0.0118 |
| | CMAR | 0.7417 | (0.7382, 0.7452) | 0.0576 |
| | CPAR | 0.7192 | (0.7157, 0.7228) | 0.3418 |
| PTRoI20_803 | CBA | 0.7623 | (0.7589, 0.7656) | 0.4541 |
| | CMAR | 0.8023 | (0.7992, 0.8055) | 0.5581 |
| | CPAR | 0.7761 | (0.7728, 0.7793) | 0.5412 |

and CPAR have low scores for the parameter TTRoI20_801. In terms of F1, the score of CPAR model is poorer than the other two.

## 10.2 Model Selection

Given the evaluation metrics of the models, it may be simpler to select the model with the highest accuracy. However, the metrics provided only the estimates with 95% confidence limits. The accuracy of the models is expected to be higher than the no information rate, indicating that prediction by the model is not a chance that it always predicts the highest class. Hence, the overall accuracy rate with a 95 percent confidence interval is calculated and tested for a one-sided hypothesis to ascertain if the accuracy is statistically higher than the "no information rate".

The probability value p for the success event is calculated using the binomial distribution tables and the value is compared with the level of significance $\alpha$. If $p>\alpha$ then we do not reject the null hypothesis that accuracy of the model is not more than the no information rate. If $p<\alpha$ we accept the alternative hypothesis. The p-values obtained in this test are compared with the significance level of 0.05 to select a suitable model. The accuracy, no information rate and the p-values obtained for all the models evaluated in this study are set out in Table 10.

The test indicates that

1. All three models exhibit a statistically significant accuracy level more than the no information rate for the parameters TTRoI20_808 and PTRoI20_803.
2. For the other two parameters TTRoI20_801 and FTRoI20_801, only CMAR model exhibits a statistically significant accuracy level that is higher than the no information rate.
3. Further, the overall accuracy values of CMAR are higher than the other two models for all parameters.

## 10.3 Receiver Operating Characteristics Curves

The Receiver Operating Characteristics (ROC) curve is a popular model comparison tool based on two major evaluation metrics-sensitivity and specificity. The ROC curve helps us to ascertain the relationship between true positive rates and false positive rates. The area under the ROC curve indicates the capability of accurate prediction by the model. Sensitivity gives True Positive Rate and 1-specificity gives False Positive Rate. The plot uses False Positive Rate(FPR) on the X-axis and True Positive Rate(TPR) on the Y-Axis. The diagonal line represents random guessing with equal probability of TPR and FPR. Thus, the accuracy of the model is based on how close the curve is to the diagonal line and the area covered. A model with perfect accuracy would have covered the total area and hence will be 1.0. The ROC plot and the area under the curve (AUC) are calculated with the R package 'pROC' as per [18]. The various AUC values for the models in this study are as in Table 11.

The table shows that the overall AUC values of the CMAR model for all parameters is higher compared to the other two models. Since the CBA model predicted all the data as uniclass, we cannot calculate the AUC values for the parameter TTRoI20_801. Based on the results we can conclude that CMAR has better predictive ability than CBA and CPAR.

Table 9: Model Metrics Values

| Metrics | Model | TTRol20_808 | TTRol20_801 | FTRol20_801 | PTrol20_803 |
|---|---|---|---|---|---|
| Prevalence | | 0.6366 | 0.7415 | 0.7345 | 0.6809 |
| Sensitivity/ | CBA | 0.8518 | 1 | 0.9981 | 0.8378 |
| Recall/ | CMAR | 0.8518 | 1 | 0.9998 | 0.8454 |
| True Positive Rate(TPR) | CPAR | 0.8518 | 0.2306 | 0.7845 | 0.7458 |
| Specificity/ | CBA | 0.6184 | 0 | 0.01683 | 0.6288 |
| Selectivity/ | CMAR | 0.6187 | 0.02618 | 0.04404 | 0.7128 |
| True Negative Rate(TNR) | CPAR | 0.6184 | 0.822 | 0.6057 | 0.8453 |
| Positive Predicted | CBA | 0.7964 | 0.7415 | 0.73748 | 0.8281 |
| Value(PPV)/ | CMAR | 0.7965 | 0.74651 | 0.7432 | 0.8627 |
| Precision | CPAR | 0.7963 | 0.788 | 0.8463 | 0.9114 |
| Negative | CBA | 0.7043 | NA | 0.76243 | 0.645 |
| Predicted | CMAR | 0.7044 | 1 | 0.98769 | 0.6836 |
| Value(NPV) | CPAR | 0.7043 | 0.2714 | 0.5039 | 0.6091 |
| Miss Rate/ | CBA | 0.1482 | 0 | 0.0018 | 0.1621 |
| False Negative | CMAR | 0.1482 | 0 | 0.0001 | 0.1545 |
| Rate(FNR) | CPAR | 0.1482 | 0.7694 | 0.2155 | 0.2541 |
| Fall Out/ | CBA | 0.3816 | 1 | 0.98317 | 0.3712 |
| False Positive | CMAR | 0.3812 | 0.97382 | 0.95596 | 0.2871 |
| Rate(FPR) | CPAR | 0.3816 | 0.178 | 0.3943 | 0.1547 |
| | CBA | 0.7351 | 0.5 | 0.50747 | 0.7333 |
| Balanced | CMAR | 0.7352 | 0.51309 | 0.52192 | 0.7791 |
| Accuracy | CPAR | 0.7351 | 0.7674 | 0.6951 | 0.7955 |
| | CBA | 0.8231 | 0.8515 | 0.84822 | 0.8329 |
| F1 Score | CMAR | 0.8232 | 0.51309 | 0.85261 | 0.854 |
| | CPAR | 0.8231 | 0.3568 | 0.8142 | 0.8203 |

Table 10: Model Selection Parameters

| Variable | Model | Accuracy | No Info rate | P-Value[Acc>NIR] |
|---|---|---|---|---|
| TTRol20_808 | CBA | 0.7621 | 0.6366 | <2.2e-16 |
| | CMAR | 0.7638 | 0.6366 | <2.2e-16 |
| | CPAR | 0.7634 | 0.6366 | <2.2e-16 |
| TTRol20_801 | CBA | 0.7415 | 0.7415 | 0.5021 |
| | CMAR | 0.7463 | 0.7415 | 0.002996 |
| | CPAR | 0.2151 | 0.7415 | 1 |
| FTRol20_801 | CBA | 0.7348 | 0.7345 | 0.4441 |
| | CMAR | 0.7417 | 0.7345 | 2.64E-05 |
| | CPAR | 0.7192 | 0.7345 | 1 |
| PTRol20_803 | CBA | 0.7623 | 0.6809 | <2.2e-16 |
| | CMAR | 0.8023 | 0.6809 | <2.2e-16 |
| | CPAR | 0.7761 | 0.6809 | <2.2e-16 |

Table 11: Area under the curve values

| Model | TTRol20_808 | TTRol20_801 | FTRol20_801 | PTrol20_803 |
|---|---|---|---|---|
| CBA | 0.7709 | NA* | 0.3974 | 0.7171 |
| CMAR | 0.8756 | 0.881 | 0.9462 | 0.8956 |
| CPAR | 0.8558 | 0.8658 | 0.932 | 0.8669 |

## 11 Conclusion

This study was carried out to gain a better understanding of the process system with data science concepts and to optimize the system in light of the newly-acquired insights. The main objective was to identify the operating regions that the human intellect may not be able to visualize, as the data contains multiple variables and hundreds of thousands of data points. The operating regions identified were analyzed to obtain a viewpoint on performance of the process. It was observed that the three regions identified correspond to: start up, operating in normal region and operating in refined region of operation. From the monitored variables, association rule mining techniques and multicollinearity analysis were used to identify the key parameters that contribute to the stability of the system. The identified key performance indicators significantly contribute towards the stability of the system and make it easier for the operators to choose the optimal operating region. Three predictive models were built based on the association rules for each of the key parameters and performance was evaluated by calculating the various metrics. All the models were evaluated for their capability based on test performance on various metrics. Considering these metrics, we can select CMAR as the best model of the three based on the statistical comparison of Accuracy with the No Information rate and the Area under the ROC curve. We can further improve prediction accuracy by using other high-level models such as Bayesian classification, decision tree induction and support vector machines. Further, we can also apply other ensemble techniques such as boosting, bagging and stacking to improve predictive accuracy for all key parameters. The study was done using historical data recorded in the DCS system, but in the future this can be conducted on live data directly coming out of the system; that will help to monitor the system with fewer parameters and will alarm the operator if the system leaves the normal operating range. In addition to this, the control engineer can tune the parameters for better stability using the key performance indicators and the prediction models, so that the efficiency of the process improves with no additional resources.

### 11.1 Acknowledgments

sincerely thank all my seniors and colleagues who have encouraged and stood by me in every manner for this study. I would especially thank Dr. Kallol Roy, Chairman Managing Director and Shri V. Rajan Babu, Director (Technical), Bharatiya Nabhikiya Vidyut Nigam Limited (BHAVINI) for kind permission given to me to take up the study in the process system and to have access to the data for the purpose of study.

## References

[1] Dragan Vuksanović, Jelena Ugarak, and Davor Korčok. Industry 4.0: the Future Concepts and New Visions of Factory of the Future Development. In *Proceedings of the International Scientific Conference - Sinteza 2016*. Singidunum University, 2016.

[2] Saurabh Vaidya, Prashant Ambad, and Santosh Bhosle. Industry 4.0 – A Glimpse. *Procedia Manufacturing*, 20:233–238, 2018.

[3] Ramos Sofia, Ana Brásio. *Industrial processes monitoring methodologies*. PhD thesis, Department of Chemical Engineering, Faculty of Science and Technology, University of Coimbra, 2015.

[4] Petr Kadlec, Bogdan Gabrys, and Sibylle Strandt. Data-driven Soft Sensors in the process industry. *Computers & Chemical Engineering*, 33(4):795–814, apr 2009.

[5] J. Cregan M. Flynn, D. Ritchie. Data mining techniques applied to power plant performance monitoring. In *IFAC Proceedings Volumes (IFAC-PapersOnline)*, volume Volume-38,Issue-1,pages-369-374, 2005.

[6] Juha Juselius. Advanced condition monitoring methods in thermal power plants. Master's thesis, LAPPEENRANTA UNIVERSITY OF TECHNOLOGY LUT School of Energy Systems, 2018.

[7] Jian qiang Li, Cheng lin Niu, Ji zhen Liu, and Luan ying Zhang. Research and Application of Data Mining in Power Plant Process Control and Optimization. In *Advances in Machine Learning and Cybernetics*, pages 149–158. Springer Berlin Heidelberg, 2006.

[8] T. Ogilvie, E. Swidenbank, and B.W. Hogg. Use of data mining techniques in the performance monitoring and optimisation of a thermal power plant. In *IEE Two-day Colloquium on Knowledge Discovery and Data Mining*. IEE, 1998.

[9] Zhiqiang Ge, Zhihuan Song, Steven X. Ding, and Biao Huang. Data Mining and Analytics in the Process Industry: The Role of Machine Learning. *IEEE Access*, 5:20590–20616, 2017.

[10] S. Narasimhan and Rajendran. Application of Data Mining Techniques for Sensor Drift Analysis to Optimize Nuclear Power Plant Performance. *International Journal of Innovative Technology and Exploring Engineering*, 9(1):3087–3095, nov 2019.

[11] V.Rajendran. S.Narasimhan. Optimization of a Process System in Nuclear Power Plant- A Data Mining Approach. *Grenze International Journal of Engineering and Technology, Special Issue*, Grenze ID:6.2.1, 2020.

[12] Jian Pei. Jiawei Han, Micheline Kamber. *Data mining : concepts and techniques*. Morgan Kaufmann Publishers, 2012.

[13] Wenmin Li, Jiawei Han, and Jian Pei. CMAR: accurate and efficient classification based on multiple class-association rules. In *Proceedings 2001 IEEE International Conference on Data Mining*. IEEE Comput. Soc, 2001.

[14] Xiaoxin Yin and Jiawei Han. CPAR: Classification based on Predictive Association Rules. In *Proceedings of the 2003 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, may 2003.

[15] Coenen, F. LUCS-KDD implementations of CPR (Classification based on Predictive Association Rules).

[16] Hahsler M, Johnson I. Classification Based on Association Rules [R package arulesCBA version 1.2.0].

[17] Max Kuhn. Classification and Regression Training [R package caret version 6.0-86]. *Comprehensive R Archive Network (CRAN)*, 2020 web page = https://cran.r-project.org/package=caret.

[18] Natacha Hainard Alexandre Tiberti Natalia Lisacek Frédérique Sanchez Jean-charles Müller Markus Robin, Xavier Turck. pROC : an open-source package for R and S + to analyze and compare ROC curves. *BMC Bioinformatics*, Issue 12, 2011.