STATISTICAL ANALYSIS OF PROCESS IN A HYPOTHETICAL CHEMICAL PLANT

İrem Tanrıverdi
Middle East Technical University
Ankara, Turkey
irem.tanriverdi@metu.edu.tr

Abstract— In this study, the production process in a hypothetical chemical plant is observed. In this chemical plant, various liquid products are produced from one raw material (liquid). The goal is to keep the pH value and the iron value of the liquid in the desired values. Thus, in this paper, pH values and iron levels are predicted in terms of sensor readings and caustic/cinh injection ratios using different machine learning algorithms like logistic regression, decision tree, random forest, neural network and XGBOOST.

***Keywords*— Machine learning, modelling, hypothetical chemical plant, row materials**

I. INTRODUCTION

The raw materials in the hypothetical plant are obtained from five different sources. The process involves heating the input liquid to a certain temperature and applying a series of chemical processes to obtain the final products and wastes from the input product. The input usually contains one or more substances that cause corrosion of the equipment. Caustic soda and corrosion inhibitors are used to minimize corrosion in the equipment. Caustic soda is used to regulate the pH of the liquid. The corrosion inhibitor is used to prevent corrosion of the equipment. A sample of the fluid is taken every 2 hours at a sampling point near the end of the process to measure the pH and iron content. The processing in the laboratory takes 30 minutes, and the ratio of caustic and Cinh injection is adjusted according to the measured pH and iron content. The goal is to keep the pH between 5.5 and 6.5 and the iron value below 1 to minimize corrosion.

II. METHODOLOGY

A. *Dataset*

There are 41 data sets, which include information about production process in a hypothetical chemical plant. In this chemical plant, various fluid products are derived from a fluid which consists of a set of compound substances. These compound substances have distinct densities and specific heat values. Source of the row materials, input temperature, pressure, speed and flow values, caustic ratio and cinh ratios, pH and iron values are provided in these data sets. These data sets contain the production process values collected hourly from 2021-02-01 08:00:00 to 2021-05-04 19:06:00. When the data needed come from multiple sources, it is essential to know how to aggregate them so that we lose as little information as possible and make pairings that make sense given the structure of the data. When datasets represent the same set of observations, datasets are combined horizontally. In such cases, it should be checked if the order of the observations is the same. The primary key is the column or set of columns that uniquely identifies each observation in the data sets. In this study **"date"** column is the primary key of the data sets. Iron and pH datasets only include dates with 2-hour time interval. If inner join is presented, all the other datasets taken as dates with 2-hour time interval. This causes a lot of information loss. Thus, left join must be provided. In this manner, data sets are combined according to their date without any loss of information. The iron and pH datasets contain only dates with a 2-hour timeframe. NA value will be assigned to the time intervals in between.

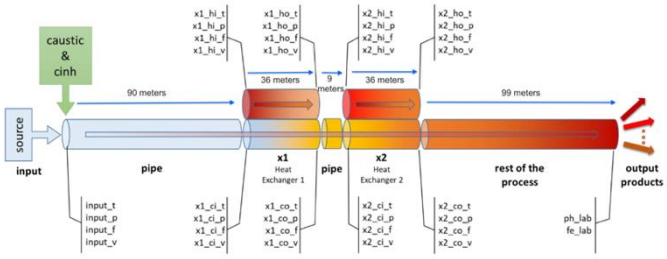

Figure 1: *A part of the process and equipment used*

Based on the flow rate and distances, the data should be synchronized [1], that is, time alignment should be done. When liquid enters, it takes 10 minutes to reach X_1 sensor, 4 minutes to exit X_1 sensor, 1 minute to enter X_2 sensor and 4 minutes to exit X_2 sensor. That is, the liquid comes out of the X_2 sensor in the 19th minute. The first fluid cycle is completed in the 30th minute. Time alignment was made according to these minutes.

B. *Descriptive Statistics*

Table 1*:* Summary of some important variables

	Min.	1st quantile	Median	Mean	3rd quantile	Max.
Caustic	0.005	0.007	0.011	0.012	0.017	0.019
Cinh	0.0006	0.00079	0.00084	0.00084	0.00088	0.0011
pH	3.20	5.76	6.08	6.03	6.25	9.84
Iron	0.43	0.85	0.88	0.88	0.92	1.49
Input flow	0	149294	149994	149273.6	150659	153508
Input pressure	0	$9.51x10^6$	$9.77x10^6$	$9.59x10^6$	$9.84x10^6$	$9.98x10^6$
Input temperature	-40.0	10.4	13.6	12.51	16.9	22.4
Input velocity	14.7	14.9	15.00	29.35	15.07	999.00

In the table above, 5-number summaries and quartile values of numeric variables are observed.

[1] Minute for liquid = flow speed * length of the pipe

- Minimum caustic value to regulate the pH value of the liquid flowing in the equipment is 0.0053 while maximum caustic value to regulate the pH value of the liquid flowing in the equipment is 0.02. Besides, minimum corrosion inhibitor value to prevent corrosion of the equipment is 0.0008 while maximum corrosion inhibitor value to prevent corrosion of the equipment is 0.00107. Minimum pH value of the liquid flowing in the equipment is 3.19 while maximum pH value of the liquid flowing in the equipment 9.84. Moreover, minimum iron value is 0.43, while maximum iron value is 1.49.
- When we look at the range of the range of the input temperature, pressure, flow, and velocity, they are same with the cold in and cold out values.
- On average, pH value of the liquid flowing in the equipment is 6.03 which is between the range (5.6-6.4). Mean pH value in the data is to the desired extent.
- On average, iron value is 0.88 which is smaller than 1. Mean iron value in the data is to the desired extent.

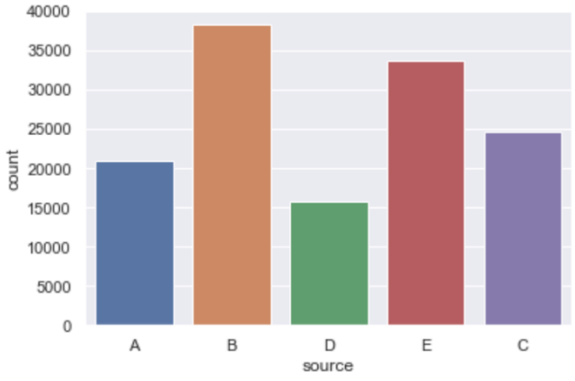

Figure 2: Frequency of the sources

In above bar plot, frequencies of source of the raw materials. As seen, most frequent row material is 'B', while less frequent row material is 'D'. Some variables have missing values in the dataset. We handle missing values in the next parts.

C. Exploratory Data Analysis

We have hourly data, and it can be complicated to visualize. Therefore, the data was made daily and average values for the production process were taken for each day.

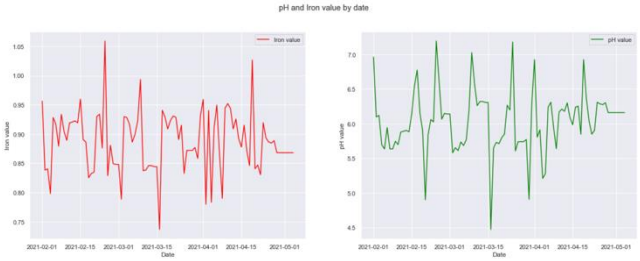

Figure 3: pH and iron values by date

In above line graph, pH and iron values are shown daily. It was observed that daily pH and iron values showed the same pattern. Both pH and iron values show dramatic changes day by day. The changes are so rapid and sharp. For example, in 2021-03-07 average pH value is approximately 7.1, then it drops to 4.5 in 2021-03-15.

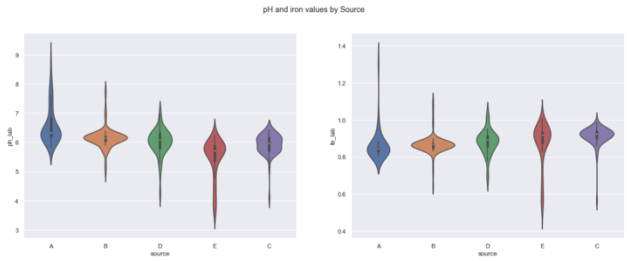

Figure 4: pH and iron values by Source

In above plot, pH and iron values grouped by source of the materials have shown. Range of the pH value of the source A (5.5-10.5) is highest while source E is smallest (3.1-7.1). Mean pH value of source A is highest, while source E is smallest. The shape of the distribution (extremely skinny on each end and wide in the middle) indicates the pH of source B is highly concentrated around its median. Range of the iron value of the source A (0.7-1.6) is highest while source E is smallest (0.2-1.1). Mean iron value of source E is highest, while source A is smallest. The shape of the distribution (extremely skinny on each end and wide in the middle) indicates the iron of source B and C are highly concentrated around its median.

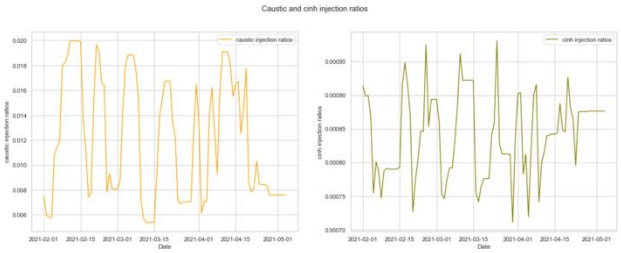

Figure 5: Caustic and cinh ratios by date

In the line plot above, it has been observed how the caustic and Caustic soda (referred to as "caustic") and corrosion inhibitor (referred to as "cinh") values change over time. Continuous increases and decreases are observed in both graphs. These constant changes are not very fast. Caustic and cinh values increase and decrease at opposite times.

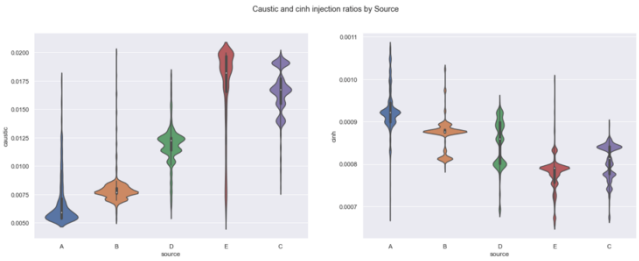

Figure 6: Caustic and cinh ratios by Source

In above plot, caustic and corrosion inhibitor injection ratios grouped by source of the materials have shown. Range of the caustic ratio of the source E (0.004-0.025) is highest while source A is smallest (0.004-0.018). Mean caustic ratio of source E is highest, while source A is smallest. The shape of the distribution (extremely skinny on each end and wide in the middle) indicates the caustic ratio of source B is highly concentrated around its median. Range of the cinh ratio of the source A is highest while source E is smallest. Mean cinh ratio of source A is highest, while source E is smallest. The shape of the distribution (extremely skinny on each end and wide in the middle) indicates the cinh ratio of source B is highly concentrated around its median.

***Figure 7**:* Input values by date

Input temperature, pressure, velocity, and flow values are shown in the graph above. Dramatic increases and decreases in input flow and velocity values have been observed day by day. In other words, input flow and velocity show rapid changes. Input pressure also shows increases and decreases, but these changes occur slowly, unlike flow and velocity. The input temperature is constantly increasing. The input temperature has seasonality also.

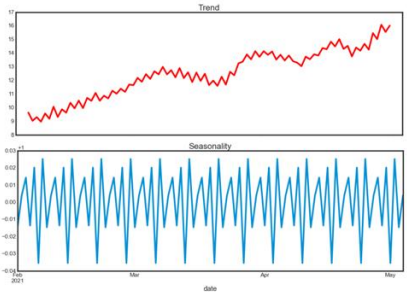

***Figure 8**:* Daily Seasonality for temperature

Seasonal decompose is applied, and seasonality plot is obtained for input temperature. As seen from plot, there is daily seasonality and increasing trend in temperature.

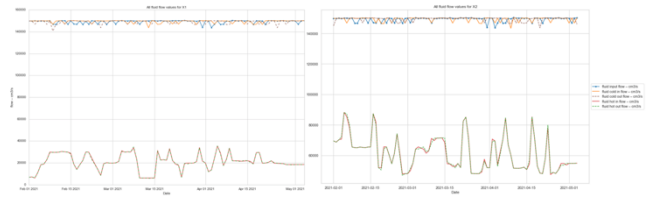

***Figure 9**:* Fluid flow values for X_1 and X_2 sensors

It is seen that cold in flow and cold out flow shows small increases and decreases. In flow values for X_1 and X_2 sensors, both of them present exactly same pattern and same values (around 150000), in input, cold in and cold out flow. Fluid hot in and hot out flow shows sharp increases and decreases.Fluid hot in and hot out flow is between 60000-90000 in X_2 sensor while fluid hot in and hot out flow is between 10000-30000 in X_1 sensor. Fluid hot in and hot out flow shows different pattern in X_1 and X_2 sensors.

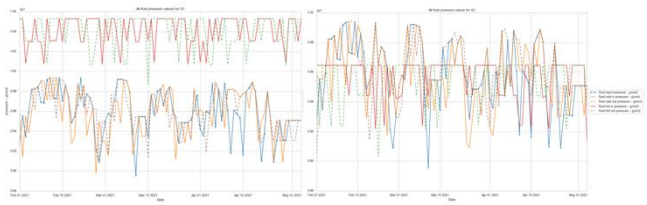

***Figure 10**:* Fluid pressure values for X_1 and X_2 sensors

It is seen in pressure values for X_1 and X_2 sensors, both present same patterns, but their values are different. Fluid hot in pressure and hot out pressure for X_1 sensor is around 1-1.5, while they are around 0.9-0.97 in X_2 sensor. Fluid input, cold in and cold out pressure for X_1 sensor is around 0.88-0.99, and they are exactly same with X_2 sensor. Cold fluid values are equal for X_1 and X_2 sensor, but hot fluid values are different.

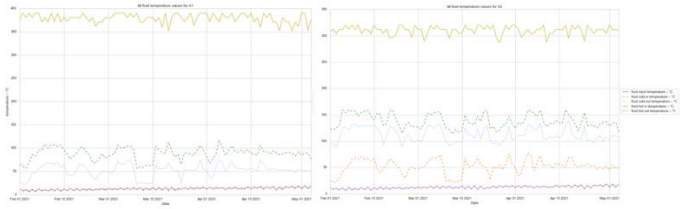

***Figure 11**:* Fluid temperature values for X_1 and X_2 sensors

It is seen in temperature values for X_1 and X_2 sensors, fluid hot out, cold out and hot out temperature values show same pattern, but different values. Fluid cold in shows different pattern in both sensors. Fluid cold out and hot out temperature values are around 30-120 in X_1 sensor while fluid cold out and hot out temperature values are around 98-160 in X_2 sensor. Fluid cold in temperature is around 25 in X_1 while it is between 25-80 in X_2 sensor and in both sensors, it shows different pattern. Cold in, cold out and hot out shows increases and decreases slowly.

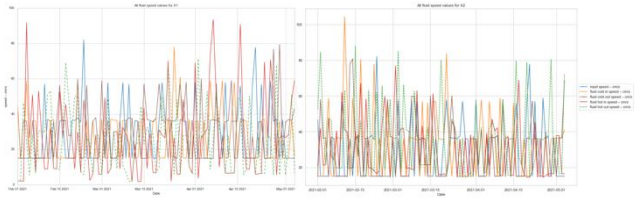

***Figure 12**:* Fluid speed values for X_1 and X_2 sensors

It is seen from the speed values for X_1 and X_2 sensors, both present similar patterns, but their values are different. Fluid hot in speed in X_1 sensor is higher than X_2 sensor. X_1 hot in speed value is around 0-90 while it is around 0-60 in X_2 sensor. Fluid hot out speed in X_1 sensor is smaller than X_2 sensor. X_1 hot out speed is around 0-70 while it is around 10-90 in X_2 sensor. Fluid cold in speed in X_1 sensor is smaller than X_2 sensor. X_1 cold in speed is around 10-70 while it is around 10-120 in X_2 sensor.

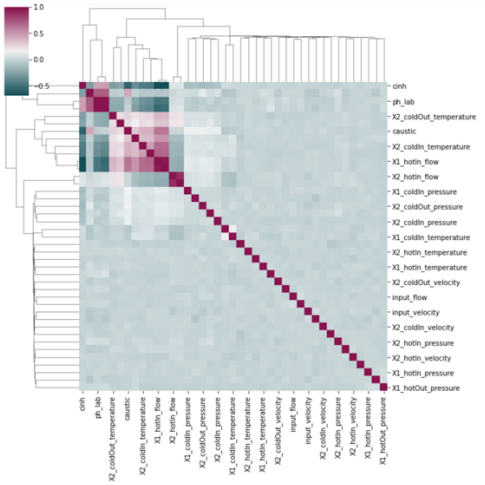

***Figure 13**:* Correlation between features

Red color represents positive correlation and green color represent negative correlation between features.

D. Outliers and Missingness

Firstly, outliers are detected using boxplots and detecting all features interquartile ranges. Then, outliers are trimmed. That means, outlier values are excluded from analysis. By applying this technique data becomes thin when there are more outliers present in the dataset. Its main advantage is its fastest nature. Let's look at plot of some variables with outlier and without outlier values.

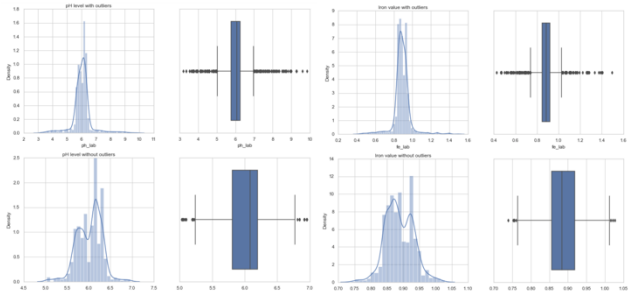

***Figure 14**:* Variables with and without outliers

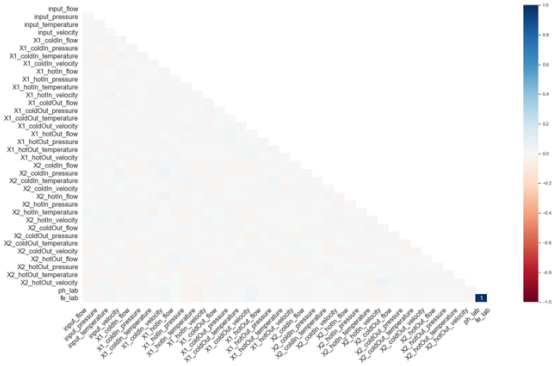

***Figure 15**:* Correlation of missingness

Heatmap shows the correlation of missingness between every 2 columns. In the above heatmap, we observe that correlations are very close to 0. That means there is no dependence between the occurrence of missing values of two variables. It was found that the probability of missing responses depends on the set of observed responses, but not on the exact missing values that are predicted. The result is that we have random missing individuals. We can use **mice** because we have MAR. Multivariate Imputation over Chained Equations (MICE) is an acronym for Multivariate Imputation over Chained Equations. Unlike a single imputation (such as the mean), multiple imputations are created to account for uncertainty in missing variables. It specifies one imputation model per variable to impute data on a variable-by-variable basis.

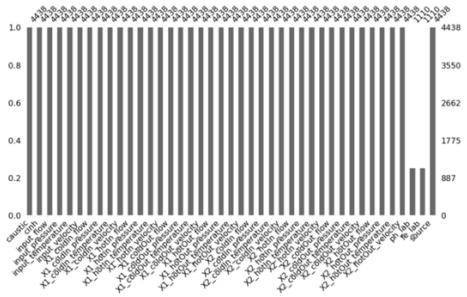

***Figure 16**:* Missingness plot after imputation

We see that after imputation, there is no missing value in the data anymore. The reason why the missing value appears in pH and iron is that the two values are measured every 2 hours.

E. Modeling

Aim in this section is to predict the pH level and iron level of the liquid flowing in the equipment. It is wondered how pH level of the liquid affected by the caustic soda (referred to as "caustic"), corrosion inhibitor (referred to as "cinh"), input temperature, pressure, flow, velocity, and sensor readings. Firstly, multiple linear regression had conducted to predict pH level and then decision tree regression was conducted.

Data sets usually contain numerical features that have been measured in different units. The values for all features must be transformed to the same scale. That is why before modeling, variables are scaled. An important aspect of modeling is the division of data into training and test sets. Data should be divided as train, validation, test set because different techniques will be tried. The validation set is used for hyperparameter tuning and the test set is used to evaluate the last selected model. Also, data should not be split randomly. Since the data changes very slowly, validation and testing are very similar to those who are in the training set. Instead, data can be split up to a certain time for training set and the rest can be divided into the similarly validation and test.

In this part two different models have conducted for each pH level and iron level. **It has been tried to conduct models with scaled features and polynomial features.**

1. Prediction of pH level

pH level is predicted with two kinds of covariates which are scaled covariates and polynomial covariates. Firstly, decision tree regression is conducted. After that support vector machine is deployed.

a) Decision Tree Regression for prediction of pH

Decision tree includes some parameters like *"max depth", "ccp alpha", "min weight fraction leaf", "min samples leaf",* etc. Hyperparameter tunning should be done using validation set to increase model efficiency and to prevent overfitting problem. These parameter values were chosen according to the points at which the mean square error is the smallest. After hyperparameter tunning, K-fold cross validation applied which ideal parameters of decision tree and root mean square error is obtained at the end to ensures that every observation from the original dataset has the chance of appearing in training and test set.

- It is observed that in the model conducted with **scaled covariates,** changing *"ccp alpha"* and *"min weight fraction leaf"* increases the MSE, so these two taking as 0. When *"max depth"* is 7, validation error and training error is minimum, so "max depth" is set as 7. After that K-fold cross validation is applied with 20 splits, and root mean square error is obtained using test set.
- After that decision tree model with **polynomial covariates** is deployed. Changing *"ccp alpha"* and *"min weight fraction leaf"* increases the MSE, so these two taking as 0. When *"max depth"* is 6, validation error and training error is minimum,

so “max depth” is set as 7. After that K-fold cross validation is applied with 20 splits, and root mean square error is obtained using test set.

In above figure, result of decision tree regression for pH level is observed.

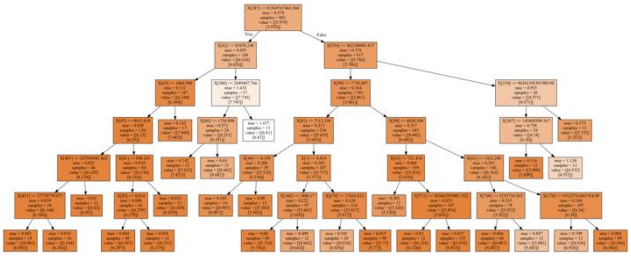

Figure 17: Decision Tree for pH level

b) Support Vector Machine (SVM)

Support Vector regression is a type of Support vector machine that supports linear and non-linear regression. It should be chosen a kernel and parameter and regularization if needed. (Gaussian Kernel and noise regularization are an instance for both steps). The most important SVR parameter is Kernel type. It can be linear, polynomial or gaussian SVR. We have a non-linear condition, so we can select polynomial or gaussian but here we select RBF (a gaussian type) kernel. Epsilon in the epsilon-SVR model is also tunned. It specifies the epsilon-tube within which no penalty is associated in the training loss function with points predicted within a distance epsilon from the actual value. After hyperparameter tunning, K-fold cross validation applied which ideal parameters of SVM, and root mean square error is obtained at the end to ensures that every observation from the original dataset has the chance of appearing in training and test set.

- It is observed that in the model conducted with **scaled covariates,** setting kernel type as gamma gives smallest MSE. Regularization parameter (c) is set as 100. When *“epsilon”* is 0.05, validation error and training error is minimum, so “epsilon” is set as 0.05 and “kernel” is set as “rbf”. After that K-fold cross validation is applied with 20 splits, and root mean square error is obtained using test set.
- After that SVM model with **polynomial covariates** is deployed. Kernel type is set as gamma since it gives smallest MSE. When *“epsilon”* is 0.1, validation error and training error is minimum, so “epsilon” is set as 0.1 and “kernel” is set as “rbf”. After that K-fold cross validation is applied with 10 splits, and root mean square error is obtained using test set.

c) Model Performance Comparison for prediction of pH level

Table 2: Model performance comparison

Models	MSE	RMSE
Decision Tree with Scaled Covariates	0.12	0.35
Decision Tree with Polynomial Covariates	0.10	0.32
SVM with Scaled Covariates	0.092	0.30
SVM with Polynomial Covariates	0.25	0.50

When we look at the test set performance, mean square and root mean square error of SVM with Scaled Covariates are smallest. Besides, root mean square of the Decision Tree with Polynomial Covariates is smallest. Error values of Decision Tree with Polynomial Covariates is close to SVM with Scaled Covariates. It is observed that polynomial features did a good job in decision tree, while this transformation did not work well in support vector machine model Thus, SVM with Scaled Covariates outperforms the other models.

2. Prediction of iron value

Iron value is predicted with two kinds of covariates which are scaled covariates and polynomial covariates. Firstly, decision tree regression is conducted. After that XGBOOST is deployed.

a) Decision Tree Regression for prediction of Iron

Hyperparameter tunning for decision tree again applied using validation set to increase model efficiency and to prevent overfitting problem. After hyperparameter tunning, K-fold cross validation applied which ideal parameters of decision tree and root mean square error is obtained at the end to ensures that every observation from the original dataset has the chance of appearing in training and test set.

- It is observed that in the model conducted with **scaled covariates,** changing *“ccp alpha”* and *“min weight fraction leaf”* increases the MSE, so these two taking as 0. When *“max depth”* is 10, validation error and training error is minimum, so “max depth” is set as 10. After that K-fold cross validation is applied with 30 splits, and root mean square error is obtained using test set.
- After that, decision tree model with **polynomial covariates** is deployed. Changing *“ccp alpha”* and *“min weight fraction leaf”* increases the MSE, so these two taking as 0. When *“max depth”* is 8, validation error and training error is minimum, so “max depth” is set as 7. After that K-fold cross validation is applied with 10 splits, and root mean square error is obtained using test set.

In above figure, result of decision tree regression for iron value is observed.

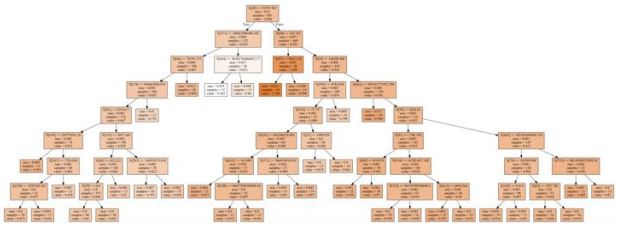

Figure 18: Decision Tree for Iron value

b) XGBOOST

Gradient boosting refers to a class of ensemble machine learning algorithms. Ensembles are constructed from decision tree models. Trees are added one at a time to the ensemble and fit to correct the prediction errors made by prior models. This is a type of ensemble machine learning model referred to as boosting. Models are fit using any arbitrary differentiable loss function and gradient descent optimization algorithm. XGBOOST includes some parameters like *“reg alpha”* which is L1 regularization term on weights, *“reg lambda”* which is L2 regularization term on weights, *“Learning rate”* which is Boosting learning rate (“eta”). Hyperparameter tunning is done using validation set to increase model efficiency

and to prevent overfitting problem. These parameter values were chosen according to the points at which the mean square error is the smallest. Before modeling, dataset needs to be converted into DMatrix. It is an optimized data structure that the creators of XGBoost made.

- It is observed that in the model conducted with **scaled covariates,** when “eta:0.6”, "reg lambda: 0.9”, "reg alpha:0.6” and “num boost round: 100”, MSE is the smallest.
- After that, XGBOOST model with **polynomial covariates** is deployed. When “eta:0.6”, "reg lambda: 0.911”, "reg alpha:0.6” and “num boost round: 100”, MSE is the smallest.
- In above figure, result of XGBOOST model output for iron value is observed using shap values [2].

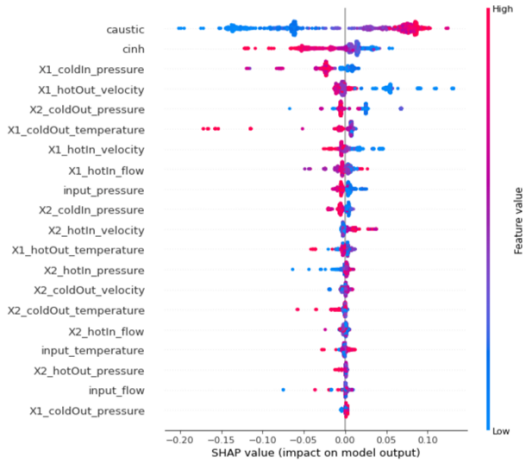

***Figure 19**:* Impact of the covariates on model output

The summary plot shows the global importance of features. It shows the distribution of feature contributions to model output using the SHAP values of each feature for each observation. Each point is an observation. pink color represents high impact, blue color represents low impact. We can say that caustic, cinh and X1 cold in pressure have high impact on iron value.

c) Model Performance Comparison

***Table 3**:* Model performance comparison

Models	MSE	RMSE
Decision Tree with Scaled Covariates	0.002	0.048
Decision Tree with Polynomial Covariates	0.0008	0.028
XGBOOST with Scaled Covariates	0.0015	0.039
XGBOOST with Polynomial Covariates	0.0007	0.027

When we look at the test set performance, mean square and root mean square error of XGBOOST with polynomial Covariates are smallest. It is observed that polynomial features did a good job in both decision tree and XGBOOST models compared with scaled features (MSE of model with polynomial features is smaller than model with scaled features).

F. Clustering

The k-means clustering method is an unsupervised machine learning technique used to identify clusters of data objects in a dataset. The first step is to randomly select k centroids, where k is equal to the number of clusters you choose. Centroids are data points representing the center of a cluster. Data sets usually contain numerical features that have been measured in different units. The values for all features must be transformed to the same scale. Since the dataset 41 features the visualization in 2D is impossible. Therefore, observations are represented by points in the plot, using principal components. Then, the data are ready to be clustered.

a) Choosing the Appropriate Number of Clusters

The elbow method and silhouette coefficient are often used as evaluation techniques. To perform the Elbow method, run several k-means, increment k with each iteration, and record the SSE. “k” value is chosen at a point **smallest WSS value** is obtained. The Silhouette Coefficient is a measure of cluster cohesion and separation. It quantifies how well a data point fits into its assigned cluster. Larger numbers indicate that samples are closer to their clusters than they are to other clusters. K-means models were conducted for each k between the values 2 and 10. Then, obtain silhouette score for each k. Then, choose optimal value of k which **has the highest average silhouette score**.

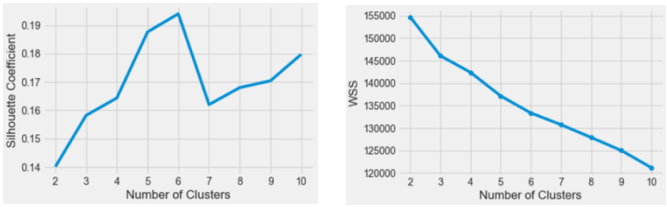

***Figure 20**:* WSS and silhouette scores

As seen from the plot, when k=6, we reach the highest average silhouette score and smallest WHH value. That is why number of clusters is taking as 6.

b) Fitting Cluster

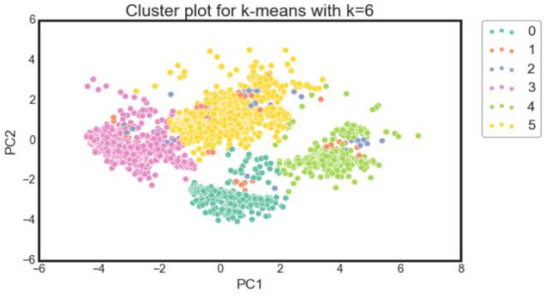

***Figure 21**:* Cluster plot for k-means

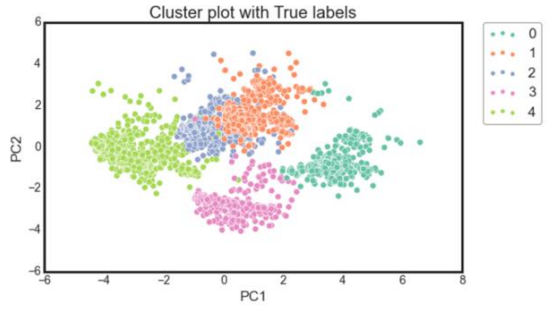

***Figure 22**:* Cluster plot for true labels

[2] “Shapley Additive explanations” values are the most advanced method for interpreting results from tree-based models. It represents feature importance based on the marginal contribution to the model outcome.

Between these 2 cluster plots in above there is not so many differences. The number of clusters in the plots is different, respectively 6 for k-means and 5 for true labels. When we look at the predicted labels in the k-means model, we see that there are 6 labels which are ["0", "1", "2", "3", “4”, “5”]. True label values are ["0", "1", "2", "3", "4”] (original data values for source). The K-means model overestimate 1 more label based on true label values. If we consider that underestimating is not good, we can say that the k-means model is good in terms of including all true label values.

G. Proposing New Corrosion Control Method

Let first calculate the cost of cinh and caustic ratio given method in figure 21 by assuming the unit cost of cinh is equal to twice the unit cost of caustic. Then, new method should be proposed to minimize the total cost of chemicals injected for corrosion control while ensuring the pH value is between 5.5 and 6.5 and the iron value is less than 1.

if $5.6 \leq ph\_lab \leq 6.4$ then $r_{caustic\_next} = r_{caustic\_current}$
if $ph\_lab < 5.6$ then $r_{caustic\_next} = 1.1 * r_{caustic\_current}$
if $ph\_lab > 6.4$ then $r_{caustic\_next} = 0.9 * r_{caustic\_current}$
if $0.80 \leq fe\_lab \leq 0.95$ then $r_{cinh\_next} = r_{cinh\_current}$
if $fe\_lab < 0.80$ then $r_{cinh\_next} = 0.95 * r_{cinh\_current}$
if $fe\_lab > 0.95$ then $r_{cinh\_next} = 1.05 * r_{cinh\_current}$

Figure 23*:* Old corrosion control method

pH and iron level were predicted in previous parts. After finding new formulas for caustic, new pH values can be predicted by taking the test set of the updated data and putting the updated test set into the model trained with the old data. After the pH values are predicted, it can be observed between which values the pH values are. These new pH predictions can be compared with the old pH prediction values and observed if the new caustic method works or not. Similarly, after new method for cinh is presents, updated data can be obtained. After that test set can be obtained from updated data, and new prediction values can be obtained using updated test set. Then, it can be compared updated prediction values for iron value with old prediction values of iron level. If predicted iron values with new method smaller than 1 more compared to old prediction method, new formula for cinh works. Besides, aim is to minimize cost of cinh and caustic. Hence, the amount used of cinh, and caustic should be reduced. To reduce amount of cinh and caustic, coefficients in the formula should be reduced. Cinh and caustic values are included in the data with time periods 8:30, 10:30, 12:30, 14:40... (Half an hour after the measurement of pH and iron values). Thus, take the sum of cinh and caustic values to calculate cost of them. Let assume unit cost of cinh is 100$, and unit cost of caustic is 200$. Total cost of cinh with old corrosion control is 93.38$, while cost of the caustic is 2688.18$.

If $0.80 \leq fe\_lab \leq 0.95$ then, $r_{cinh\_next} = r_{cinh\_current}$
If $fe\_lab < 0.80$ then, $r_{cinh\_next} = 0.1 * r_{cinh\_current}$
If $fe\_lab > 0.95$ then, $r_{cinh\_next} = 1.01 * r_{cinh\_current}$

Figure 24*:* New cinh method

When this new cinh method is applied, cost of cinh be 10.73$, and prediction of iron level with new method is less than 1. Prediction of iron level with old method is also less than 1. However, since cost of this new method is smaller than old method, updated method of cinh is better.

If $5.6 \leq ph\_lab \leq 6.4$ then, $r_{cinh\_next} = r_{cinh\_current}$
If $ph\_lab < 5.6$ then, $r_{cinh\_next} = 1.05 * r_{cinh\_current}$
If $ph\_lab > 6.4$ then, $r_{cinh\_next} = 0.9 * r_{cinh\_current}$

Figure 25*:* New caustic method

When this new caustic method is applied, cost of caustic be 197.6$, and prediction of pH level with new method is between the desired range. Prediction of pH level with old method is also between the desired range. However, since cost of this new method is smaller than old method, updated method of cinh is better.

III. CONCLUSION

In this study, to understand better the data structure, exploratory data analysis techniques such as graphical tools and descriptive statistics are used. Then, the mechanism of missing values was investigated and imputed to improve the data quality. Outlier values are observed. Subsequently, the effect of sensor readings and cinh and caustic injection ratios on pH level and iron value had examined. Finally, pH level and iron values of sample from the fluid in raw materials had tried to predict by using several machine learning algorithms like XGBOOST, Support Vector Machine, and Decision Tree. The results had shown in the previous chapter. According to the data provided, it is observed that caustic, cinh, X1 cold in pressure and X1 hot out velocity are the most efficient common factors in each model for pH value and iron value of the fluid in the study. Besides, decision tree with scaled features seems to be good at estimating test data for pH value and the XGBOOST with polynomial features is good at estimating the test data for iron value successfully. After modelling part number of source from which row materials is received are tried to be discovered using k-mean clustering method. Optimal number of clusters is obtained as 6, result of the k-means clustering. Ground truth is 5 group. It can be said that k-means model is good in terms of including all true label values.

IV. REFERENCES

[1]. Hayes, S. (2021, June 8). *Finding seasonal trends in time series data with python*. Medium. Retrieved February 3, 2022, from https://towardsdatascience.com/finding-seasonal-trends-in-time-series-data-with-python-ce10c37aa861

[2]. *Basic shap interaction value example in xgboost*. Basic SHAP Interaction Value Example in XGBoost - SHAP latest documentation. (n.d.). Retrieved February 3, 2022, from https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/tree_based_models/Basic%20SHAP%20Interaction%20Value%20Example%20in%20XGBoost.html