MACHINE LEARNING APPLICATION

for

MINERAL PROCESSING

101

Just for the Beginning

Aydin AKTAR

Mineral Processing Engineer

aktaraydin@gmail.com

https://www.linkedin.com/in/aktaraydin/

Table of Contents

1. FOREWORD

Plants controlled by Artificial Intelligence (AI) are not far away. In this report, you will find the Machine learning application and concentrate predictions made with the developed model on data from a Flotation Plant.

After completing Data Science trainings and readings, I started searching for data and finally came across the data shared by Eduardo Magalhães Oliveira on Kaggle.com. I would like to thank him a lot.

Data Address: https://www.kaggle.com/edumagalhaes/quality-prediction-in-a-mining-process

I think all report content will be useful for Mining Professionals who are not familiar with Machine Learning (ML). In the report, when the model development was completed and the predictions started, I asked some example questions of how machine learning might be use for Mineral Processing Plants.

Data scientists who want to learn about the mining industry (especially mineral processing) will be able to find summary information for themselves in the report.

For those who are curious about all coding steps, they can access Jupyter Notebook from link below. Please do not hesitate to contact. We must discuss this issue a lot now. I think there are many topics to discuss. *(Except my English and reporting skills)*

Link for Jupyter Notebook and Presentation:

https://github.com/aktaraydin/ML_for_MineralProcessing

Aydin AKTAR

2. CRISP-DM

Cross-industry standard process for data mining, known as CRISP-DM, is an open standard process model that describes common approaches used by data mining experts. It is the most widely-used analytics model.

In 2015, IBM released a new methodology called *Analytics Solutions Unified Method for Data Mining/Predictive Analytics* (also known as ASUM-DM) which refines and extends CRISP-DM.

(https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining)

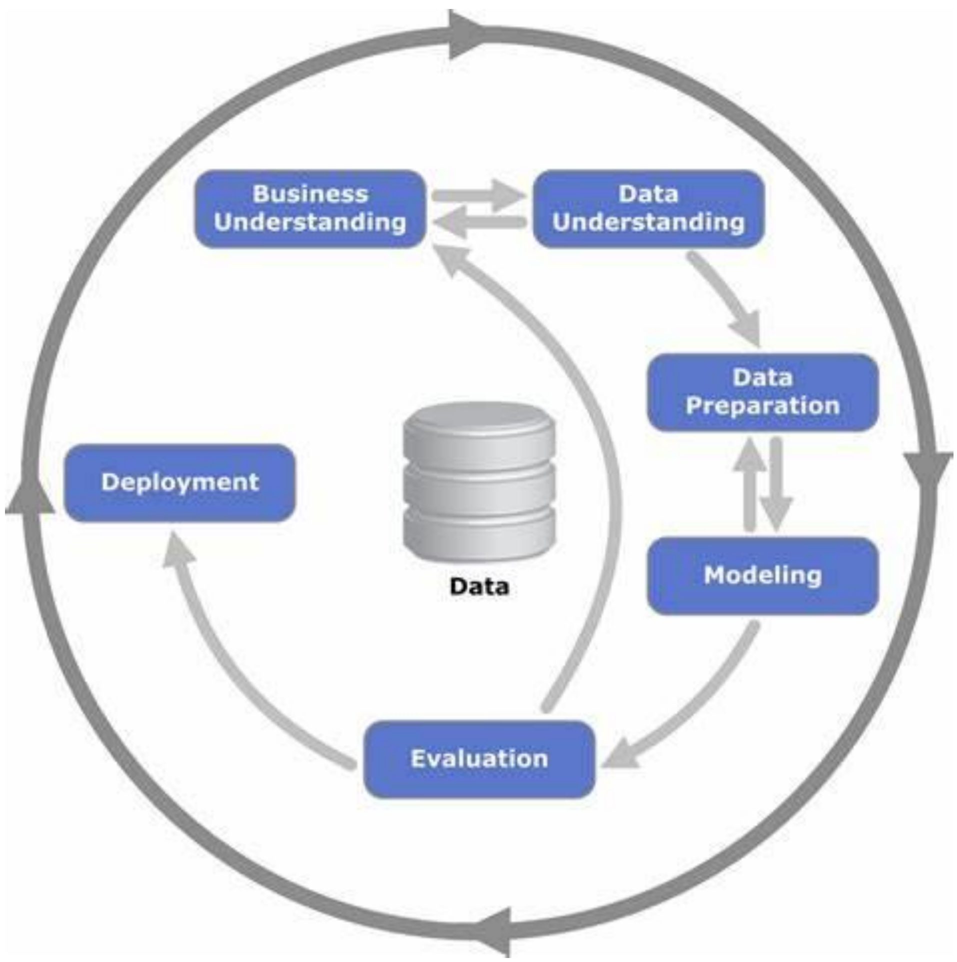


Figure 1 - CRISP-DM

I will prepare the project report over the CRISP-DM cycle.

3. CRISP-DM Step-1: BUSINESS UNDERSTANDING

We cannot use the Mineral or Element in the ore extracted from the earth directly in industry. We must carry out an enrichment operation beforehand. The data we obtained was taken from a Flotation Plant.

3.1. Froth Flotation Definition:

Froth flotation is a process for selectively separating hydrophobic materials from hydrophilic. This is used in mineral processing, paper recycling and waste-water treatment industries. Historically this was first used in the mining industry, where it was one of the great enabling technologies of the 20th century. It has been described as "the single most important operation used for the recovery and upgrading of sulfide ores. The development of froth flotation has improved the recovery of valuable minerals, such as copper- and lead-bearing minerals. Along with mechanized mining, it has allowed the economic recovery of valuable metals from much lower grade ore than previously.(https://en.wikipedia.org/wiki/Froth_flotation)

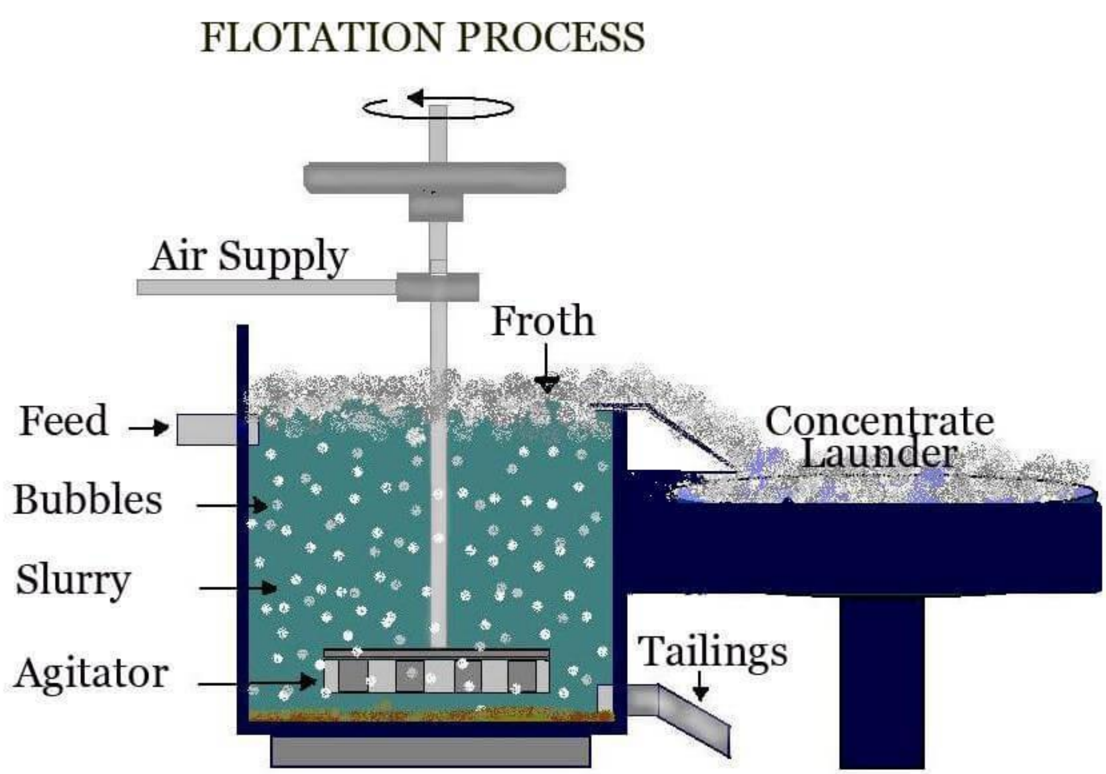


Figure 2 - Forth Flotation Process

(https://www.911metallurgist.com/blog/wp-content/uploads/2013/09/flotation-separators.jpg)

3.2. Reverse Cationic Flotation:

The demand of high-grade iron ore concentrates is a major issue due to the depletion of rich iron-bearing ores and high competitiveness in the iron ore market. Iron ore production is forced out to upgrade flowsheets to decrease the silica content in the palettes. Different types of ore have different mineral composition and texture-structural features which require different mineral processing methods and technologies. (https://iopscience.iop.org/article/10.1088/1742-6596/879/1/012016/pdf)

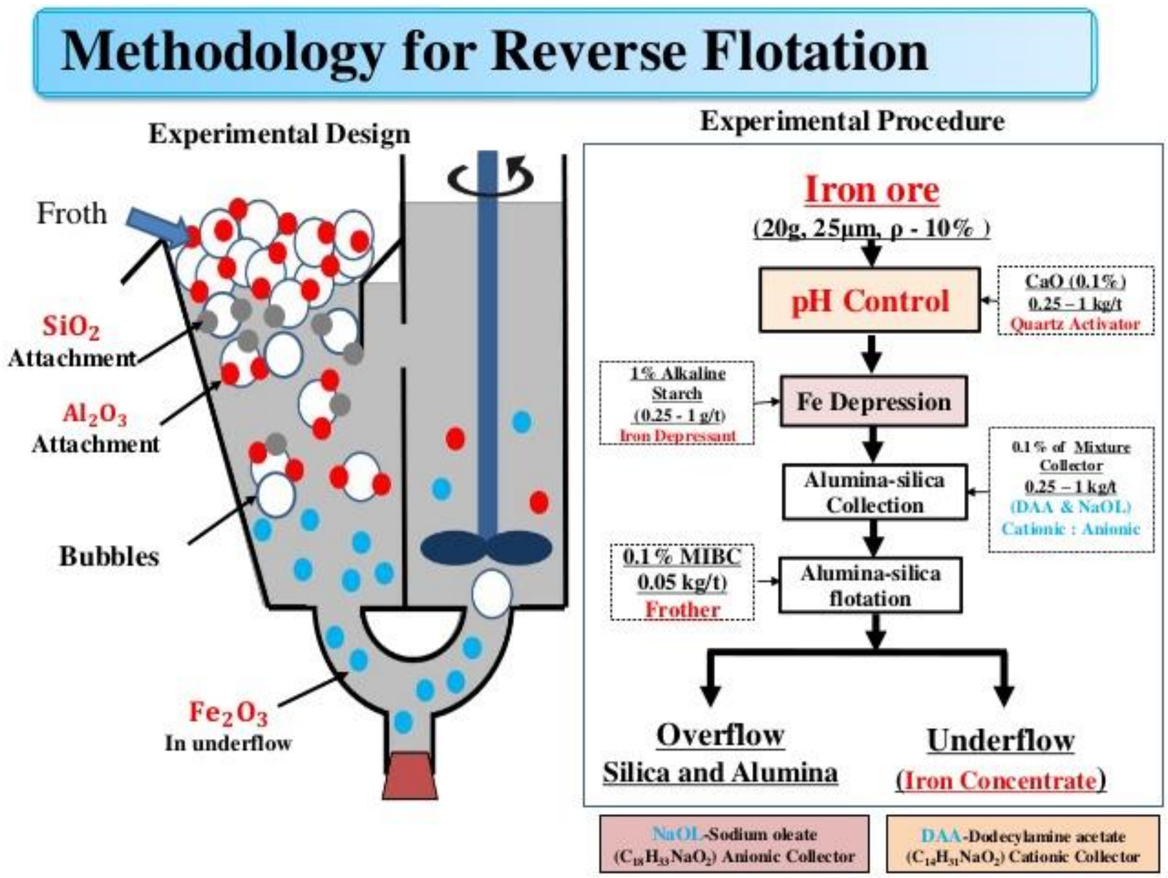


Figure 3 - Reverse Cationic Flotation Process

(https://image.slidesharecdn.com/mmij-conferencetokyo-180522171824/95/development-of-impurities-removal-process-for-lowgrade-iron-ores-using-mineral-processing-technologies-10-638.jpg?cb=1527009664)

4. CRISP-DM STEP-2: DATA UNDERSTANDING

4.1. Examine Data as Data Scientist

Data consists of 24 columns and 737453 rows, covers the dates between March 2017 and September 2017 and some of the data were recorded at 10-second intervals and some at 1-hour intervals.

```
In [23]: mainData.head(10)
Out[23]:
```

	date	% Iron Feed	% Silica Feed	Starch Flow	Amina Flow	Ore Pulp Flow	Ore Pulp pH	Ore Pulp Density	Flotation Column 01 Air Flow	Flotation Column 02 Air Flow	...	Flotation Column 07 Air Flow	Flotation Column 01 Level	Flotation Column 02 Level	Flotation Column 03 Level	Flotation Column 04 Level	Flotation Column 05 Level	Fl (0
0	2017-03-10 01:00:00	55.2	16.98	3019.53	557.434	395.713	10.0664	1.74	249.214	253.235	...	250.884	457.396	432.962	424.954	443.558	502.255	4
1	2017-03-10 01:00:00	55.2	16.98	3024.41	563.965	397.383	10.0672	1.74	249.719	250.532	...	248.994	451.891	429.560	432.939	448.086	496.363	4
2	2017-03-10 01:00:00	55.2	16.98	3043.46	568.054	399.668	10.0680	1.74	249.741	247.874	...	248.071	451.240	468.927	434.610	449.688	484.411	4
3	2017-03-10 01:00:00	55.2	16.98	3047.36	568.665	397.939	10.0689	1.74	249.917	254.487	...	251.147	452.441	458.165	442.865	446.210	471.411	4
4	2017-03-10 01:00:00	55.2	16.98	3033.69	558.167	400.254	10.0697	1.74	250.203	252.136	...	248.928	452.441	452.900	450.523	453.670	462.598	4
5	2017-03-10 01:00:00	55.2	16.98	3079.10	564.697	396.533	10.0705	1.74	250.730	248.906	...	251.873	444.384	443.269	460.449	439.920	451.588	4
6	2017-03-10 01:00:00	55.2	16.98	3127.79	566.467	392.900	10.0713	1.74	250.313	252.202	...	253.477	446.185	444.571	452.306	431.328	443.548	4
7	2017-03-10 01:00:00	55.2	16.98	3152.93	558.777	397.002	10.0722	1.74	249.895	253.630	...	253.345	445.985	461.341	461.640	442.067	441.730	4
8	2017-03-10 01:00:00	55.2	16.98	3147.27	556.030	394.307	10.0730	1.74	250.137	251.104	...	250.884	446.686	478.385	459.103	455.074	439.798	4
9	2017-03-10 01:00:00	55.2	16.98	3142.58	565.857	393.105	10.0738	1.74	249.653	252.202	...	248.137	445.685	478.779	460.665	457.225	453.236	4

10 rows × 24 columns

Figure 4 -Data Segment

```
In [4]: mainData.columns

Out[4]: Index(['date', '% Iron Feed', '% Silica Feed', 'Starch Flow', 'Amina Flow',
           'Ore Pulp Flow', 'Ore Pulp pH', 'Ore Pulp Density',
           'Flotation Column 01 Air Flow', 'Flotation Column 02 Air Flow',
           'Flotation Column 03 Air Flow', 'Flotation Column 04 Air Flow',
           'Flotation Column 05 Air Flow', 'Flotation Column 06 Air Flow',
           'Flotation Column 07 Air Flow', 'Flotation Column 01 Level',
           'Flotation Column 02 Level', 'Flotation Column 03 Level',
           'Flotation Column 04 Level', 'Flotation Column 05 Level',
           'Flotation Column 06 Level', 'Flotation Column 07 Level',
           '% Iron Concentrate', '% Silica Concentrate'],
          dtype='object')
```

Figure 5 - Column Names

```
print('Shape of Main Data = ', mainData.shape)
mainData = mainData.dropna()
print('Shape of Main Data after drop null values = ', mainData.shape)
```

```
Shape of Main Data =  (737453, 24)
Shape of Main Data after drop null values =  (737453, 24)
```

Figure 6 - Is there any null value?

There are no “null” values in data. Marvelous!!!

```
In [5]: mainData.dtypes

Out[5]: date                              object
        % Iron Feed                      float64
        % Silica Feed                    float64
        Starch Flow                      float64
        Amina Flow                       float64
        Ore Pulp Flow                    float64
        Ore Pulp pH                      float64
        Ore Pulp Density                 float64
        Flotation Column 01 Air Flow     float64
        Flotation Column 02 Air Flow     float64
        Flotation Column 03 Air Flow     float64
        Flotation Column 04 Air Flow     float64
        Flotation Column 05 Air Flow     float64
        Flotation Column 06 Air Flow     float64
        Flotation Column 07 Air Flow     float64
        Flotation Column 01 Level        float64
        Flotation Column 02 Level        float64
        Flotation Column 03 Level        float64
        Flotation Column 04 Level        float64
        Flotation Column 05 Level        float64
        Flotation Column 06 Level        float64
        Flotation Column 07 Level        float64
        % Iron Concentrate               float64
        % Silica Concentrate             float64
        dtype: object
```

Figure 7 - Data Columns Data Types

To perform mathematical calculations on data, the data must be in "int" or "float" format. The data type of columns containing numbers is "float64", sufficient for calculation. Date column is in "object" data type, we will need to change the unit before calculations.

4.2. Examine Data as Mineral Processing Engineer:

Data consist of 24 columns Detail description of columns:

Columns:

- **Date:** Date collection date and time.
- **% Iron Feed:** Feed grade of iron-containing ore.
- **% Silica Feed:** Feed grade of silica-containing ore.
- **Starch Flow:** Depressant chemical for Iron (Fe) containing ore.
- **Amina Flow:** Collector chemical for Silica containing ore.
- **Ore Pulp Flow:** The amount of pulp flow fed to the Flotation Columns as the product of the previous process step.
- **Ore Pulp pH:** pH.
- **Ore Pulp Density:** The solid percent of ore fed to Flotation Columns.
- **Flotation Column 01,02,03,04,05,06,07 Air Flow:** The amount of air fed to the Flotations Columns to frothing.
- **Flotation Column 01,02,03,04,05,06,07 Level:** Showing float thickness of Flotation Columns.
- **% Iron Concentrate:** Concentrate grade of iron-containing ore.
- **% Silica Concentrate:** Concentrate grade of silica-containing ore.

The plant flowsheet is probably (untold) like:

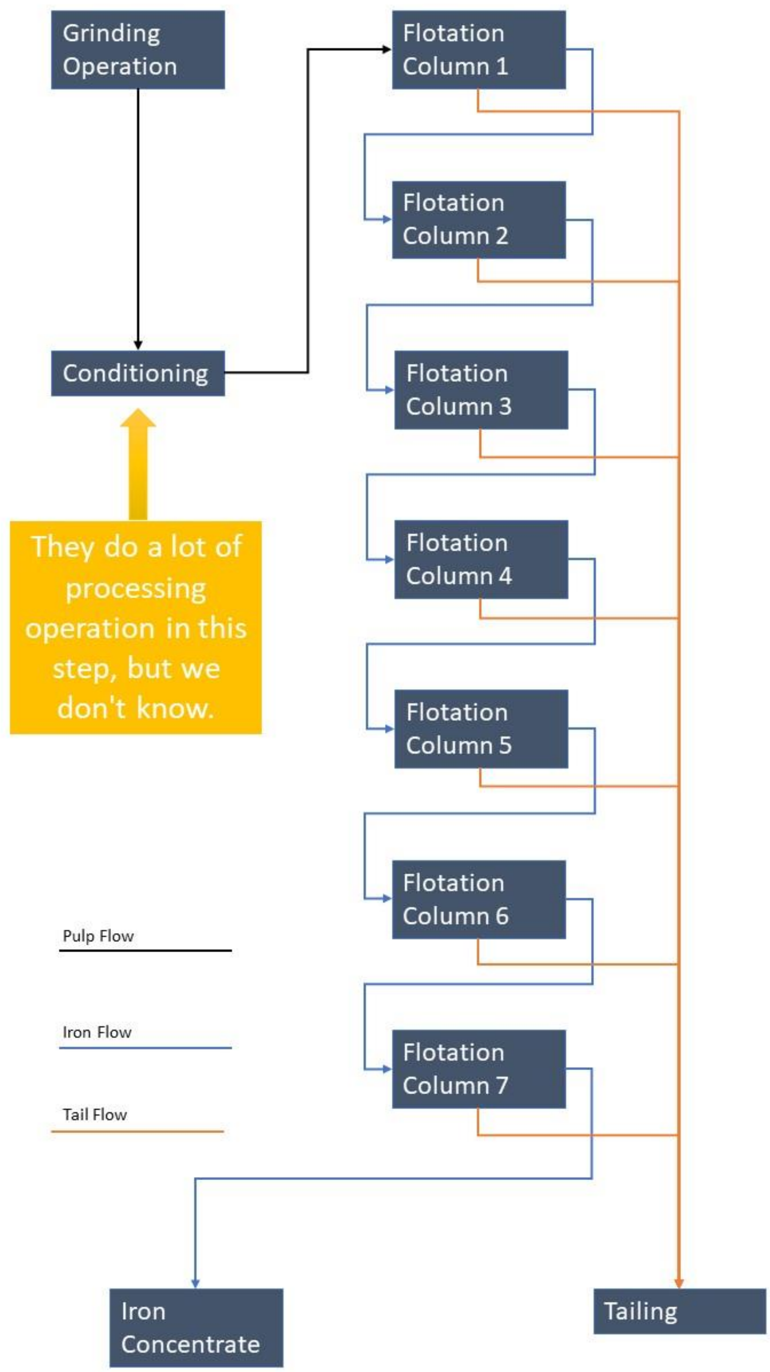


Figure 8 - Plant Flowsheet

Going deep inside to Data:

- Data can be record at 10 second frequency.
- Flows can be measured
- pH can be measured.
- Feed and Concentrate grade can be analyzed instantly.
- Columns has level sensor.

To obtain this data, plant must be operating with established automation system. It is not possible collect this data with this frequency by human hand.

To fully analyze a Flotation Plant, we need to obtain the minimum data below, which is available in a plant operated with Scada.

1. **Recovery:** A concentrate was obtained in plant. X tones with Y grade. These values may be happy to boss but what is the recovery in other words: What percentage of the feed ore could you enrich, what percentage did you send to the waste?

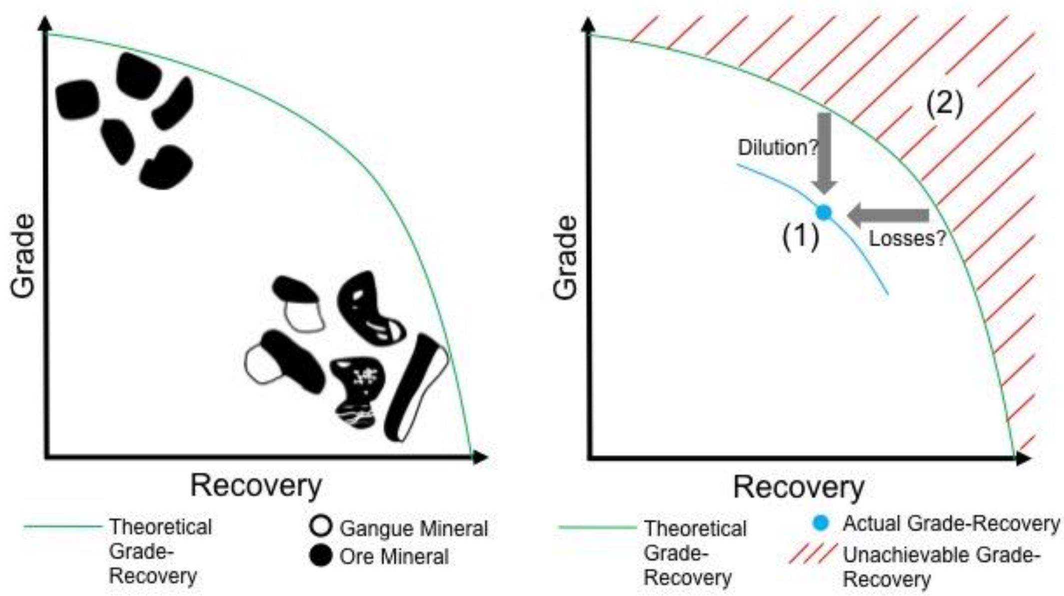


Figure 9 - Grade - Recovery Curve

(https://minassist.com.au/wp-content/uploads/2013/06/Grade-Recovery-explained.jpg)

Short description of the chart: If entire ore is to be recovered by enrichment, concentrate grade will be low. If Concentra grade is to be high, recovery will be low. This is BALANCE...

2. **Liberation Degree, Grinding Performance:** For the ore to react chemically, the ore must be liberated, and this is provided by the grinding and classification units. Huge data is missing. Mill performances, Hydrocyclone values, solid percent's... etc.

I make it my duty to say it repeatedly: WE NEED TO THINK THE FLOTATION PLANT AS A MONOLITHIC

3. **Pumps Amperes, Pressures, Flows:** After the ore is ground, it is mixed with water and turned into pulp. It needs to be transported for processing the ore. Pumps do the transport. It gives us valuable information mass metal balance, plant flow performance…etc.

4. **Feed Tons:** It is needed for calculate Mass Balance, Metal Balance, Recovery… etc.

5. **Feed and Tail Grades for Every Columns:** Each cell enriches the concentrate of the previous one. If measurement of the cell performance is done, which this facility can do, then the power of our machine learning will be increased.

6. **Mineralogy:** In flotation plants, it is generally enriched with minerals, not elements.

   Element: https://en.wikipedia.org/wiki/Chemical_element

   Mineral: https://tr.wikipedia.org/wiki/Mineral

   Having ore mineralogy data means finding a way to process that mineral with machine learning. I even think that by combining laboratory data, plant data and mineralogy data, even the future performance of the plant can be predicted, and even future mineralogy and plant parameters can be found. Wouldn't it be nice?

Because of the missing data (I mentioned above) In my opinion this data useless for real life. We will continue the analysis as we have no other choice.

5. CRISP-DM STEP-3: DATA PREPARATION

There is no need for unit conversions in data. I just changed the "date" column from the "object" data type to the "datatime64".

5.1. Grouping Rows with Hourly Frequency

I indicated about the need to consider the plant data as a totality. So how do we do this on data? By doing loop-based analysis. My solution is to consider each row of data on a loop. We feed 100 tons of ore to the plant, enrich it, and complete the cycle. In this data I set each cycle as 1 hour. That is, we take a photograph of the plant every hour and analyze it on this photograph. If we had more regular data, we could do these cycles even for 1 minute, then we would have more cycles for machine learning.

```
In [7]: mainData['date'] = pd.to_datetime(mainData['date'])
        #grouping the data according to the hours and get their average values.
        cycle_data = mainData.groupby(pd.Grouper(key='date',freq='H')).mean()
        # cycle_data.insert(0,'Date',cycle_data.index)
        cycle_data.reset_index(inplace = True)

        #some rows have 'null' values because of timing. We need to drop them
        print('Shape of Cycle Data = ', cycle_data.shape)
        cycle_data = cycle_data.dropna()
        print('Shape of Cycle Data after drop null values = ', cycle_data.shape)

        Shape of Cycle Data =  (4415, 24)
        Shape of Cycle Data after drop null values =  (4097, 24)
```

Figure 10 - Data Grouping according with Hourly Frequency

The grouping process has a few advantages and disadvantages.

Advantages:

- Each analysis will be able to do each data cycle on an hourly frequency.
- Date column can be dropped. The number of columns fell to 23.
- The number of rows fell to 4097 from 737453. Every row means calculations on computer.

Disadvantage:

- The number of rows fell to 4097 from 737453. The more rows we have for machine learning, the better results we get. This large data loss will adversely affect our estimation results.

5.2. Divide Data

Flotation plant product (at least for this missing data) Iron and Silica concentrates namely the last two columns. We will develop separate models for silica and iron in machine learning, so we will separate the columns from the main data as two data types named "iron_concentrate" and "silica_concentrate".

The remaining columns are essentially Flotation conditions required to enrich the concentrate. We also separate them as "flotation_conditions".

If we need to pay attention, the concentrate contains both silica and iron. As we said while describing the recovery, we can never make 100% enrichment, that is, the flotation conditions must be met in the silica too.

```
In [35]: #seperate data as flotation_conditions and concentrates
         flotation_conditions = data.iloc[:,1:22]
         concentrates = data.iloc[:,22:]
         silica_concentrate = concentrates.iloc[:,1].values
         iron_concentrate = concentrates.iloc[:,0].values

         print('Shape of flotation_conditions = ', flotation_conditions.shape)

         Shape of flotation_conditions =  (4097, 21)
```

Figure 11 - Divide data as "flotation conditions" - "iron_concentrate" and “silica_concentrate”

6. CRISP-DM STEP-4: MODELING

6.1. P-Value:

"The concepts of *p-value* and *level of significance* are important aspects of hypothesis testing and statistical methods like regression. However, they can be a little tricky to understand, especially for beginners, and a good understanding of these concepts can go a long way in understanding machine learning.**" – (https://medium.com/@ODSC/the-importance-of-p-values-in-data-science-6cb7c7380881)**

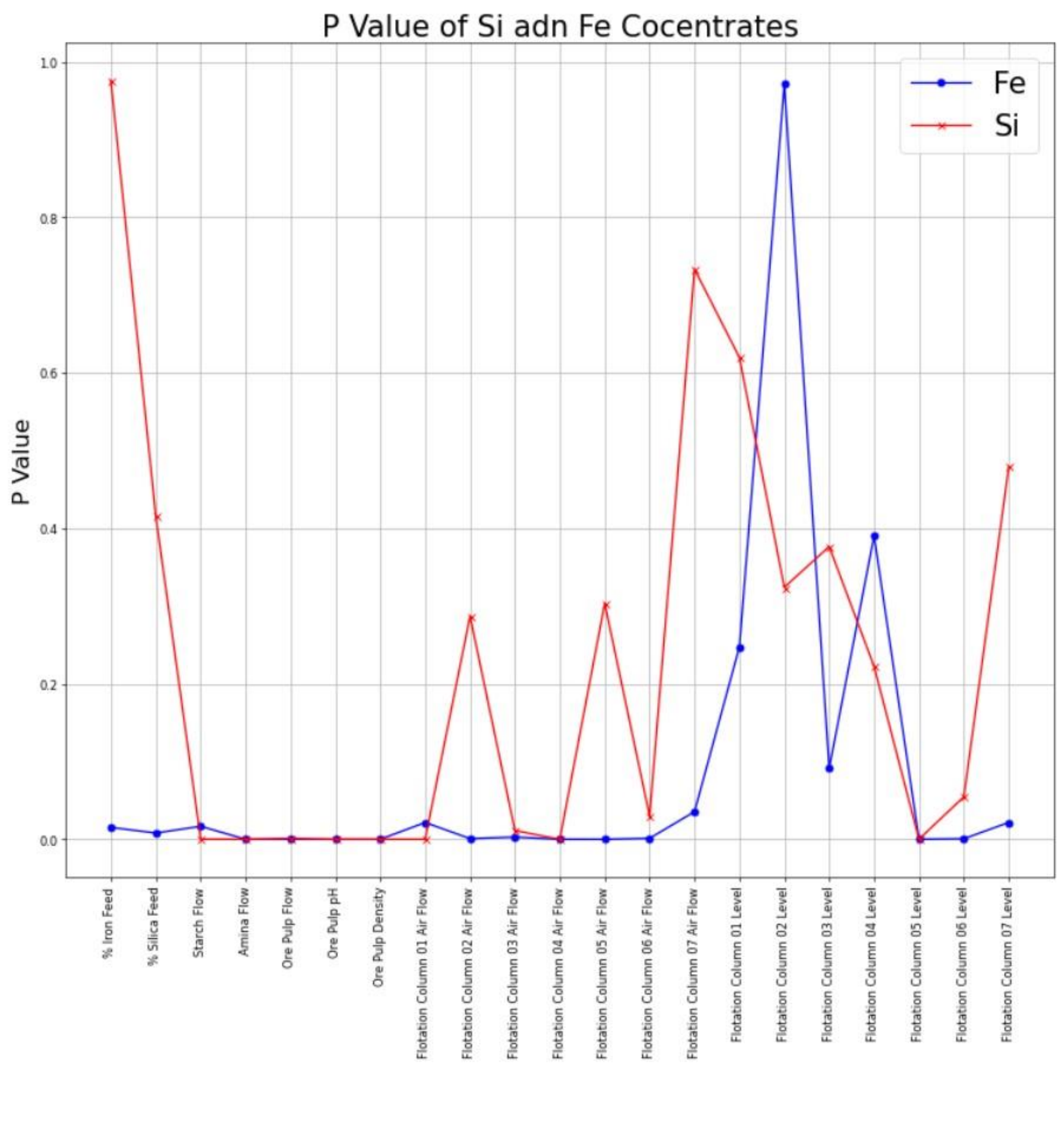


Figure 12 - P value of Si and Fe Concentrates

Flotation plant is a dynamic facility. If we change any parameter anywhere in the plant, all plant parameters will be recalculated and adapt immediately to the new situation. While we are processing ore, we feed and process 1 ton of ore and then do not feed another 1 ton of ore. No!! Continuity and balance are the main dynamics. There is a continuous supply of ore, there is continuous ore preparation and there is continuous concentrate production. For this reason, we cannot consider the data independently.

When we examine the P values, we see that some values do not affect the concentrated grades. But now we know that this is due to our lack of data. What would you think if you saw that the cocoa had no effect on the chocolate produced in the chocolate factory?

6.2. Confidence Interval:

The degree of confidence at which we are sure the interval will span the true parameter is *Confidence level*

e.g. 95% confidence interval contains the estimated parameter with probability 0.95 - i.e. in 1 case out of 20 it will miss the real parameter. (http://mlwiki.org/index.php/Confidence_Intervals)

```
#Confidence Intervals

m_fe = iron_concentrate.mean()
se_fe = iron_concentrate.std()/math.sqrt(len(iron_concentrate))
ci_fe = [m_fe - se_fe*1.96, m_fe + se_fe*1.96]


m_si = silica_concentrate.mean()
se_si = silica_concentrate.std()/math.sqrt(len(silica_concentrate))
ci_si = [m_si - se_si*1.96, m_si + se_si*1.96]

print  ('Confidence interval of Fe Concentrate:' ,ci_fe)
print  ('Confidence interval of Silica Concentrate:' ,ci_si)
```

```
Confidence interval of Fe Concentrate: [65.01585076976251, 65.0843176461516]
Confidence interval of Silica Concentrate: [2.292315650512625, 2.361191793958939]
```

Figure 13 - Confidence Interval Values

Flotation facilities can be considered as intermediate enrichment facilities. We separate a certain amount of iron in ore from waste material and sell it in concentrated form. The product we obtain is not pure iron, but iron separated from the waste material according to the material in the feed.

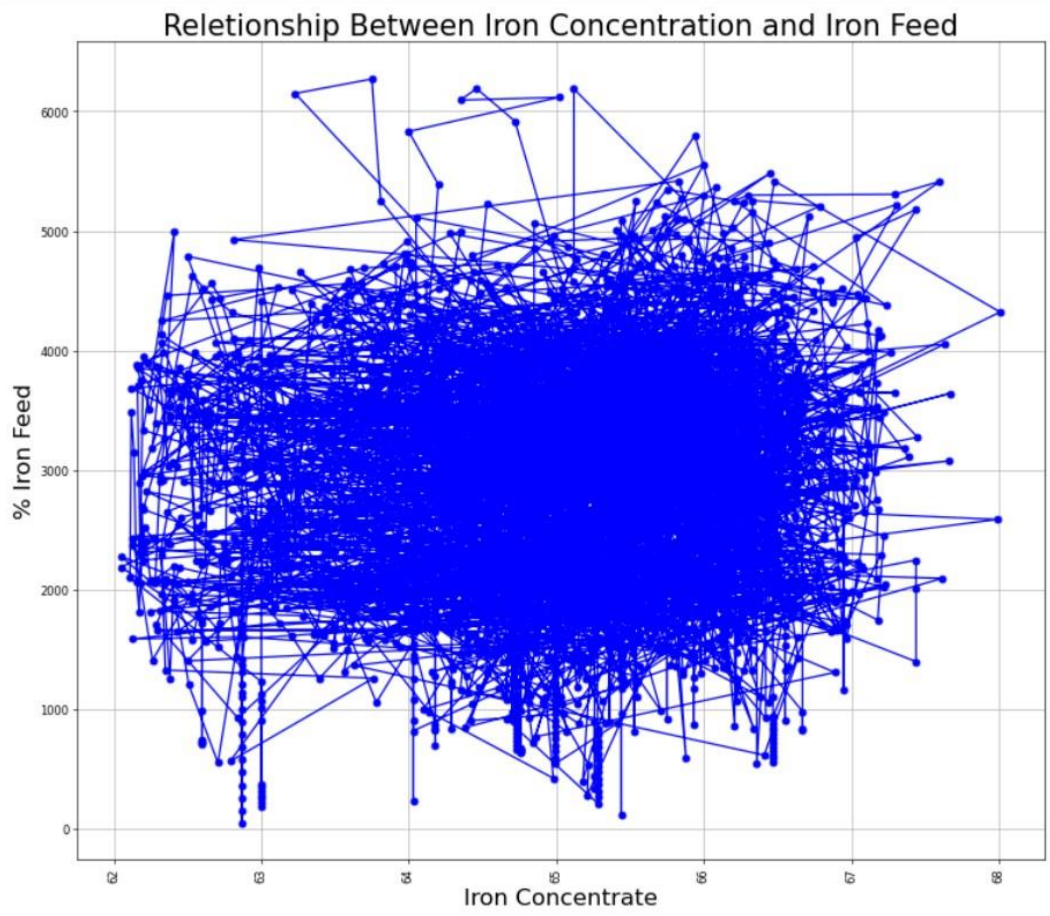
Metallurgical engineers apply the pouring process to the concentrate we sell and obtain pure iron. There is a grade range they demand from us for this operation:

%67 – 68 Fe Grade

%1.6 – 1.7 Si Grade

In the model we developed according to the confidence interval, the silica values are more than demanded. In the first stage, we need to warn the engineers in the field. (Assuming our data is reliable.)

6.3. Linear Relationship:



<Figure size 4320x216 with 0 Axes>

Figure 14 - Linear relationship between data values.

I just added this chart as an example. The relationship of other flotation conditions on the concentrate is not linear. It has become clear that the Multi Linear Regression model cannot be used. I will focus on Random Forest machine learning model development.

7. CRISP-DM STEP-5: EVALUATION

7.1. Random Forest:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance. (https://en.wikipedia.org/wiki/Random_forest)

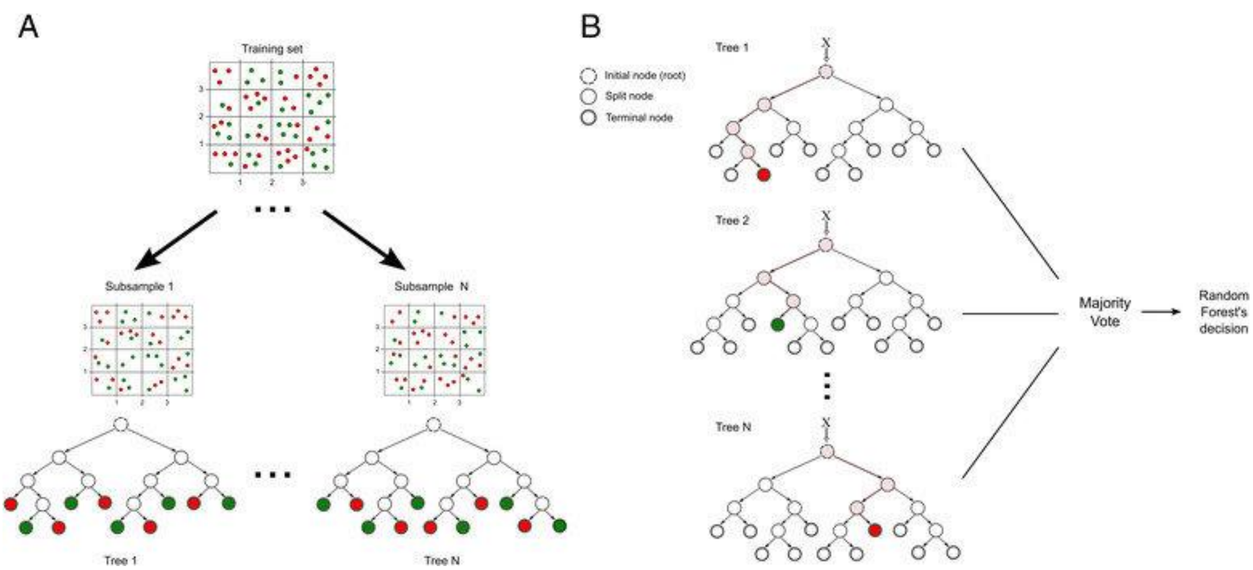


Figure 15 - Random Forest Model

(https://www.researchgate.net/profile/Mariana_Recamonde-Mendoza/publication/280533599/figure/fig5/AS:267770621329410@1440852899493/Random-forest-model-Example-of-training-and-classification-processes-using-random.png)

7.2. R-squared in Regression Analysis

R-squared is a statistical measure that represents the goodness of fit of a regression model. The ideal value for r-square is 1. The closer the value of r-square to 1, the better is the model fitted. (https://www.geeksforgeeks.org/ml-r-squared-in-regression-analysis/)

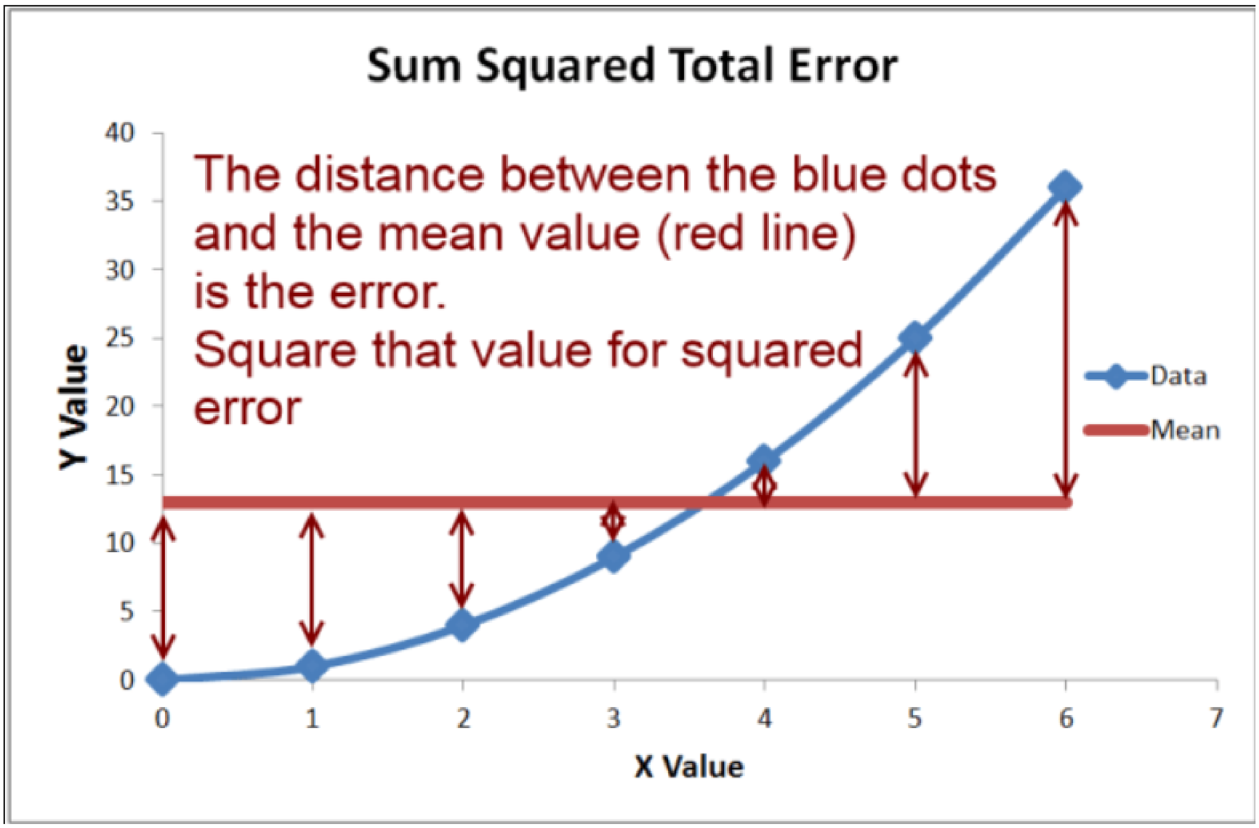


Figure 16 - R-Squared graph

7.3. Random Forest Regressor Application

The first time I divided the data as train test split at 0.33 degree and applied Random Forest machine learning, but I got very low r-Square values. Finally, I used all the data for training and got the following results. Despite the lack of data, the result is satisfactory.

```
'''Iron Random Forest Model'''
regressor_Fe = RandomForestRegressor(random_state = 0, n_estimators = 100)
regressor_Fe.fit(flotation_conditions,iron_concentrate)
y_pred_Fe = regressor_Fe.predict(flotation_conditions)

print('R2 Score of Random Forest Regression with Only Iron',r2_score(iron_concentrate,y_pred_Fe))
```

```
R2 Score of Random Forest Regression with Only Iron 0.9086194504298444
```

```
'''Silica Random Forest Model'''
regressor_Si = RandomForestRegressor(random_state = 0, n_estimators = 100)
regressor_Si.fit(flotation_conditions,silica_concentrate)
y_pred_Si = regressor_Si.predict(flotation_conditions)

print('R2 Score of Random Forest Regression with Only Silica',r2_score(silica_concentrate,y_pred_Si))
```

```
R2 Score of Random Forest Regression with Only Silica 0.9090494050116655
```

Figure 17 - Random Forest ML results

8. CRISP-DM STEP-6: DEPLOYMENT

Machine learning training is done. Now it is time to prediction.

First Predictions: I will feed to ML random values as conditions and demand to predict concentrate values.

Second Predictions: I will feed to ML mean values of the main data and demand to predict concentrate values.

```
predictions = {'% Iron Feed':50.5,
               '% Silica Feed':13.3,
               'Starch Flow':3500.0,
               'Amina Flow':580.0,
               'Ore Pulp Flow':400.0,
               'Ore Pulp pH':10.11,
               'Ore Pulp Density':1.69,
               'Flotation Column 01 Air Flow':250.0,
               'Flotation Column 02 Air Flow':150.0,
               'Flotation Column 03 Air Flow':270.0,
               'Flotation Column 04 Air Flow':190.0,
               'Flotation Column 05 Air Flow':230.0,
               'Flotation Column 06 Air Flow':200.0,
               'Flotation Column 07 Air Flow':240.0,
               'Flotation Column 01 Level':480,
               'Flotation Column 02 Level':210.0,
               'Flotation Column 03 Level':550.0,
               'Flotation Column 04 Level':620.0,
               'Flotation Column 05 Level':610.0,
               'Flotation Column 06 Level':615.0,
               'Flotation Column 07 Level':616.0,
               }
```

Figure 18 - Random Values for Prediction

```
predict_values = []
for value in predictions.values():
    predict_values.append(value)

predict_Fe = regressor_Fe.predict([predict_values])
predict_Si = regressor_Si.predict([predict_values])

print('Predicted Fe Concentrate =',predict_Fe,'%')
print('Predicted Silica Concentrate =',predict_Si,'%')
```

```
Predicted Fe Concentrate = [64.96165] %
Predicted Silica Concentrate = [2.46170555] %
```

Figure 19 – First Predicted Concentrate Grades

```
predict_from_means = flotation_conditions.describe().mean()
predict_Fe = regressor_Fe.predict([predict_from_means])
predict_Si = regressor_Si.predict([predict_from_means])

print('Predicted Fe Concentrate from Conditions mean  =',predict_Fe,'%')
print('Predicted Silica Concentrate from Conditions mean =',predict_Si,'%')
```

```
Predicted Fe Concentrate from Conditions mean  = [64.83364983] %
Predicted Silica Concentrate from Conditions mean = [2.96153289] %
```

Figure 20 - Second Predicted Concentrate Grades

We obtained the values we determined in the Confidence Intervals. The results are also very close to the average tenor values.

After all, we have a working ML application.

9. CONCLUSION

9.1. For Data Scientists:

I tried to tell you about an enrichment plant in the report. Since I have not seen a report prepared for this style before, I explained it in detail.

- I mentioned that the plant should be considered as a total and the data to be obtained should be on minimum basis.
- Cycle-based analysis will provide us the basis for dealing with engineers in the field.
- I talked about the importance of waste as well as concentrate.
- I talked about the differences between production and enrichment.

I hope you now feel closer to the mining world.

9.2. For Mining Professionals

Since there are only concentrate grades expected here, I developed a method to estimate them.

Do not restrict the estimates that can be made to concentrated grades only. Questions set the entire limit in machine learning. All the fun starts in this section.

For example:

- Chemical consumption is also a problem. What is the optimum amount of chemicals I should feed by experimenting with plant conditions?
- What can I do to optimize water consumption?
- How should I optimize the plant for the target concentration and/or tail grades?
- Ore mineralogy has changed. I have laboratory data. What are my parameters for optimum processing? Do you think the tale of actual facility data will differ from laboratory results?
- Can I optimize the facility machinery-equipment maintenance tracking according to the ore?
- Can I predict plant machinery-equipment failures and optimize warehouse stocks and workforce? (Predictive Maintenance - https://towardsdatascience.com/how-to-implement-machine-learning-for-predictive-maintenance-4633cdbe4860)

Whole answers are YES you can.

With the analysis made over the same data set, we can write thousands more questions and get results.

Our touchstone data on the road to IoT (internet of things) and how we store them. If we take care of our data and enrich them, the transition of AI will be easier. The more data you get, the more accurate results are inevitable.

Artificial intelligence, which analyzes the ore coming from the mine and gives us the facility parameters as ready and calculate equipment maintenance possibilities, is now very close.

See you on the next project.

Aydin AKTAR

Good Luck!