**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS**

**Pós-Graduação em Inteligência Artificial e Aprendizado de Máquina**

**ESTATÍSTCA GERAL – TEORIA E APLICAÇÕES**

*André Augusto Laudares Ansani*

*Lucas de Miranda Rodrigues*

*Luiz Fernando Moreira Rodrigues*

**ESTUDO DA BASE DE DADOS: FLOTATION PLANT**

**1. INTRODUÇÃO**

O Dataset contém variáveis de uma planta de flotação, que é um processo de separação de impurezas do minério.

Com essas informações pretende-se prever a quantidade de sílica no minério produzido para que se possa tomar ações antecipadas para melhorar a qualidade do produto.

* 1. **Objetivos**

* Prever a concentração de sílica (% Silica Concentrate) do minério na saída do processo de flotação;
* Prever a concentração de sílica sem usar a concentração de ferro (% Iron Concentrate).

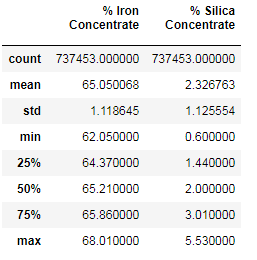
**2. DESCRIÇÃO DA BASE**

* Não há valores nulos em nenhuma das colunas do Dataset.
* Tamanho da população: 737453 registros.
* Colunas: 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 01 Column Level, Flotation 02 Column Level, Flotation 03 Column Level, Flotation 04 Column Level, Flotation 05 Column Level, Flotation 06 Column Level, Flotation 07 Column Level, Iron Concentrate, Silica Concentrate.

**2.2. Análise Descritiva**

Na Tabela 1 encontram-se as principais medidas descritivas sobre as saídas, que é a concentração de ferro e de sílica, do processo de flotação em uma mineradora.

Tabela 1 - Medidas Descritivas

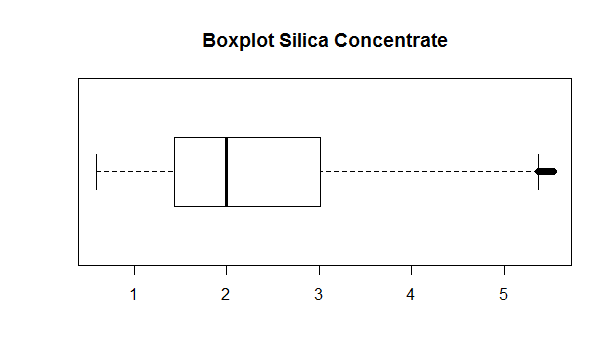
****

**2.3. Análise de Dados Discrepantes e Dispersão dos Dados**

Como objeto para esta análise, é utilizado o gráfico Boxplot, no qual, permite analisar a dispersão dos dados, como valores máximos, mínimos, 1º, 2º e 3º quartil, bem como analisar *Outliers*.

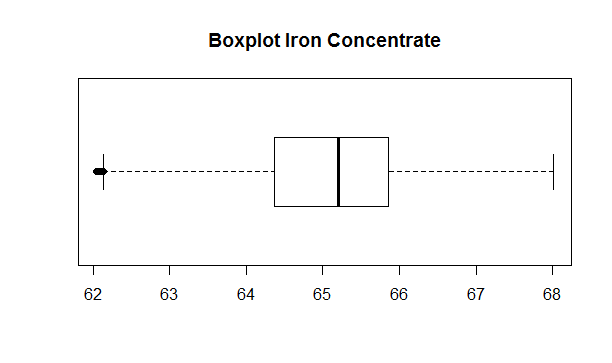
Assim, é possível entender como os dados estão distribuídos e realizar conclusões mais assertivas para diagnóstico de problemas.

Figura 1 - Gráfico Boxplot da Concentração de Sílica em %



Observa-se a presença de Outliers superiores e o conjunto de dados assimétricos para a esquerda.

Figura 2 - Gráfico Boxplot da Concentração de Ferro em %



Percebe-se a presença de Outliers inferiores e o conjunto são levemente assimétricos para a direita.

No Apêndice I encontra-se o Gráfico A1 que mostra o gráfico boxplot para todas as 23 variáveis envolvidas no processo de flotação.

**2.4. Análise de Frequência**

A análise de frequência permite conhecer quais classes ocorrem com maior frequências em um conjunto de dados. E o diagrama de Pareto informa quais classes que se reptem mais e as suas respectivas frequências acumulativas.

Com o Gráfico 3, nota-se que os intervalos de maiores ocorrência são de [65.03 , 65.328[ e de [65.328 , 65.626[. E é no primeiro intervalo mencionado se encontra o valor da média.

Com o Gráfico 4, percebe-se o intervalo que mais ocorre é [1.0930, 1.3395[. E o fenômeno que aconteceu no Gráfico 3 sobre a média, não acontece aqui. Pois o Gráfico 4 está mais deslocado, ou mais assimétrico que outro gráfico.

Figura 3 - Gráfico de Pareto para a Concentração de Ferro

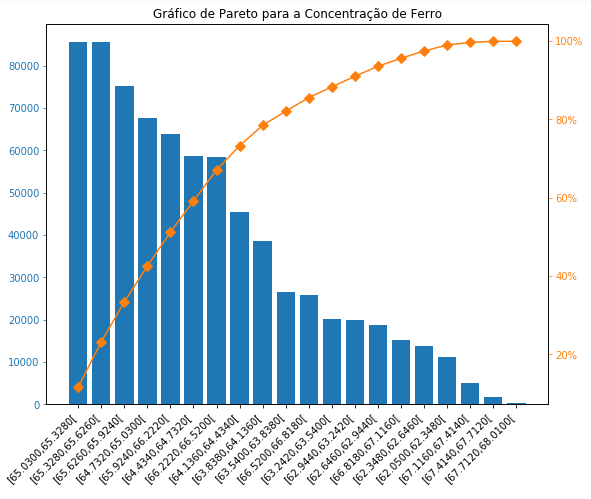
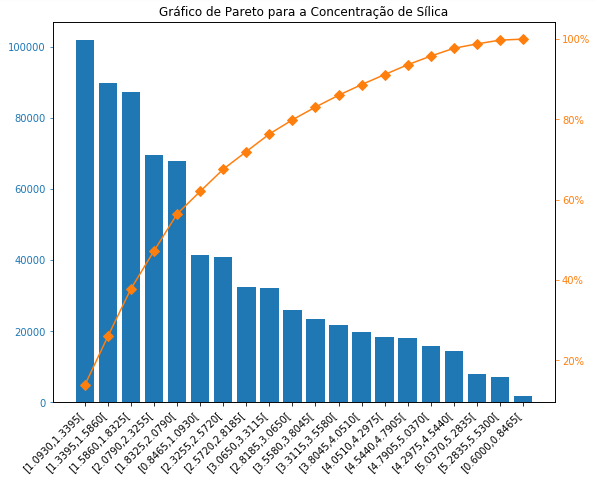


Figura 4 - Gráfico de Pareto sobre a Concentração de Sílica

****

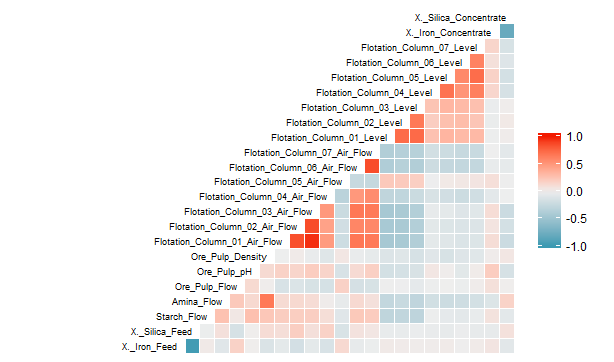
No Apêndice I tem o Gráfico A2 mostrando os histogramas das 23 variáveis.

**2.5. Análise de Correlações**

A Figura 5 mostra uma matriz de correlação. E ela evidencia as correlações entre cada uma das variáveis do dataset.

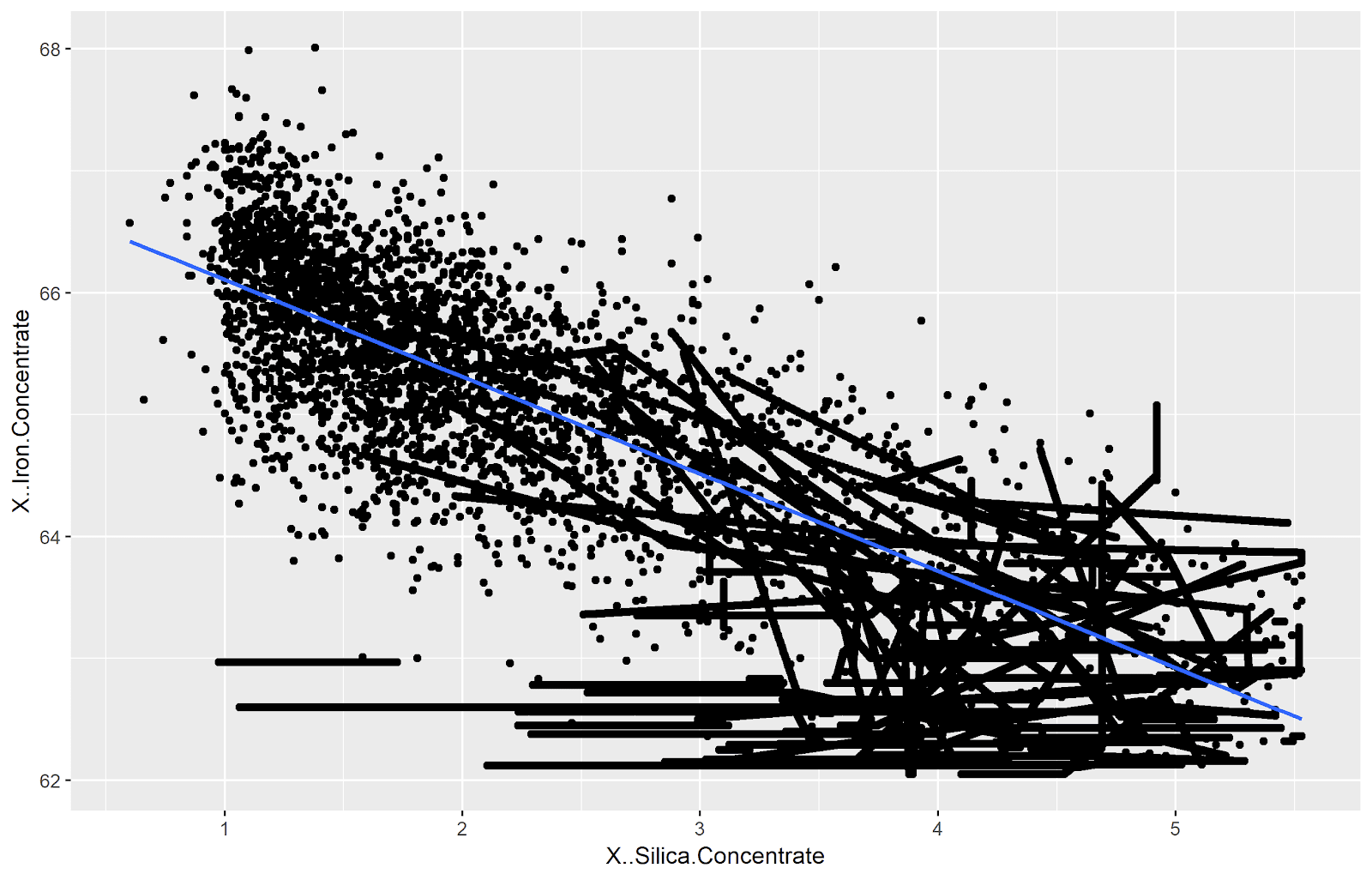
.

Figura 5 - Matriz de Correlação



Observando a matriz de correlação, percebe-se que a porcentagem de Sílica e a porcentagem de Ferro na saída do processo são muito correlacionadas, o que pode influenciar no modelo gerado. Abaixo é mostrado um gráfico de dispersão que mostra essa correlação.

Figura 6- Gráfico de Dispersão



Executando a função cor.test() com as variáveis de ferro e sílica no R, obtemos:

t = -1147.2, df = 737450, p-value < 2.2e-16  
alternative hypothesis: true correlation is not equal to 0  
95 percent confidence interval:  
 -0.8013781 -0.7997389  
sample estimates:  
     cor   
-0.80056

Esse resultado mostra que a correlação está entre -0.801 e -0.799 com 95% de confiança, indicando que as variáveis realmente são muito correlacionadas.

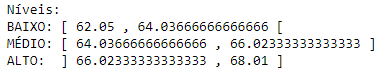
**3. PROBABILIDADES**

As probabilidades permitem uma análise das chances de obtenção de cada resultado de um experimento. Partindo desse conceito, fez-se uma análise para a concentração de ferro e para a concentração de sílica.

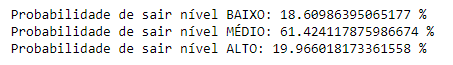
Porém para realizar as probabilidades, separou os valores encontrados, que são do tipo quantitativo contínuo, em 3 (três) níveis iguais denominados em: ALTO, MÉDIO e BAIXO.

**3.1. Probabilidade para a concentração de ferro**

Os níveis encontrados para a concentração de ferro em porcentagem foram:



E a ocorrência desses níveis ou eventos que aconteceram no dataset foi:

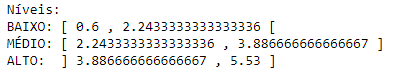


Observa-se que o nível que mais ocorreu durante o experimento foi o MÉDIO com uma probabilidade de 61,42%, aproximadamente. Essa alta ocorrência pode ter influenciado no valor da média encontrada para a concentração de ferro que foi de 65,05. Pois esse valor encontra-se de dentro do intervalo desse nível e muito próximo da média do intervalo MÉDIO que é, aproximadamente, 65,03.

Nota-se também que o nível BAIXO teve pouca ocorrência.

**3.2. Probabilidade para a concentração de sílica**

Os intervalos para a concentração de sílica foram:



E a probabilidade encontrada para esses níveis foram:

https://lh5.googleusercontent.com/1lmUc0L3R-acAyVLLvMONV8edmpcWEjWx1P7b5cjjZ9v0i7pejc_RPQZWvsDyodAobweHuxqBE316le5S6Zj-KGO_mfECXSPP5vM6X6Ofp9kFOV_MCRu2zJ6eVxzSHM305tGjveB

O nível BAIXO foi o que mais ocorreu durante o experimento.   E isso pode ter influenciado no valor da média encontrada, já que ela encontra-se nesse mesmo intervalo. O nível MÉDIO teve uma ocorrência de, aproximadamente, 27,15% que também pode ter influenciado na média fazendo com que ela seja um pouco maior do que a média encontrada no intervalo BAIXO.

O nível que menos ocorreu foi o ALTO.

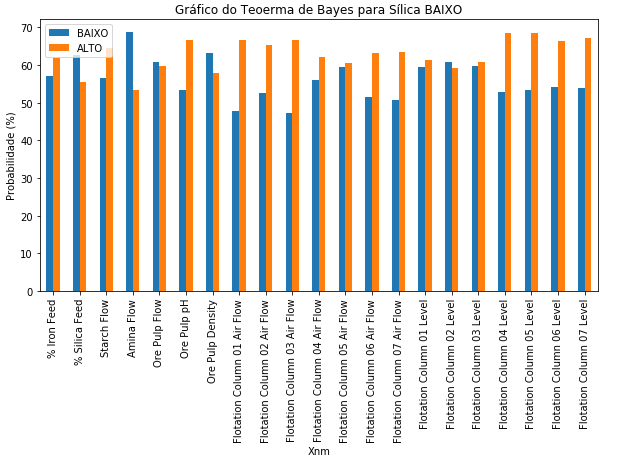
**3.3. Teorema de Bayes**

Com esses valores de probabilidades pode-se calcular:

* P(Concentração de Sílica BAIXO | Concentração de Ferro ALTO) = 97.80%
* P(Concentração de Sílica BAIXO | Concentração de Ferro MÉDIO) = 64,91%
* P(Concentração de Ferro ALTO | Concentração de Sílica BAIXO) = 32,44%
* P(Concentração de Ferro MÉDIO | Concentração de Sílica BAIXO) = 66,23%

E a P(Concentração de Sílica BAIXO | Xnm), sendo Xn as demais variáveis e m o nível de intervalos das X variáveis que é ALTO e BAIXO (metade para cada nível) é dado pela gráfico a seguir.

Figura 7 - Gráfico de barras para o Teorema de Bayes para Sílica BAIXO dado a ocorrência de Xnm

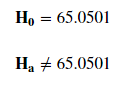


**4. TESTE DE HIPÓTESE E INTERVALO DE CONFIANÇA**

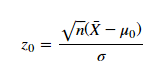
**CASO 1:** Um experimento realizado entre os dias 10/03/2017 e 09/09/2017 afirma que a média de concentração de ferro em porcentagem  é de 65,0501% depois do processo de flotação de uma mineradora. Para recusar ou aceitar essa afirmação com um nível de significância de 5%, colhe-se 100 amostras aleatórias com média igual a 64.9897 e desvio padrão igual a 1.1342.

A distribuição t pode ser aproximada da distribuição normal quando o tamanho da amostra é maior ou igual a 30, mesmo desconhecendo o desvio padrão populacional.

As definições das hipóteses são:



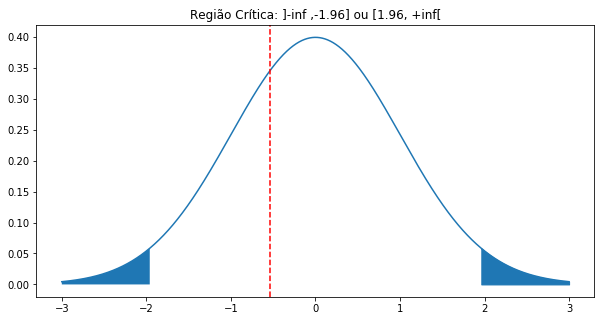
O estimador usado é dado por:



https://lh4.googleusercontent.com/6xXqkuclHAKMn5pxpA1F6DCY4xWcREiFzarMgEGRTq_lWRrpXuklLVbiwFtJI5JbgzfBLYnS3Ap7nZ2YI9K_0fehmWJgjkdo0n9F3NPJbLVSxqVPcth6NKcVlrMFNprd7zgQmk6I

A Região Crítica é dada por:

Figura 8 - Gráfico da Região Crítica



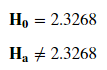
Como o valor de z0 está fora da região crítica, se aceita H0. Ou seja, há evidências que a média da concentração em porcentagem de ferro é igual a 65.05401%.

O Intervalo de Confiança de 95% é:

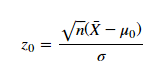
https://lh4.googleusercontent.com/vaDPahCy2fvXGga0Ks5cfq-T7bo4i3IOmywhu7xAAvbElsqmeH4_V5xkTyLWQm1XCQts-gNDOcP_P3txWDOZFcc7D7yi99Ak4vrWt7hCxuB6W7Dybh5Jnt0O4UYaKKaXyPerU-Ah

**CASO 2:** Um experimento realizado entre os dias 10/03/2017 e 09/09/2017 afirma que a média de concentração de sílica em porcentagem  é de 2,3268% depois do processo de flotação de uma mineradora. Para recusar ou aceitar essa afirmação com um nível de significância de 5%, colhe-se 100 amostras aleatórias com média igual a 2.3606 e desvio padrão igual a 1.1475.

As definições das hipóteses são:



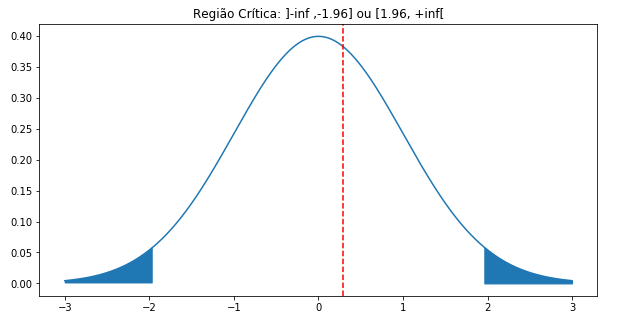
O estimador usado é dado por:



https://lh6.googleusercontent.com/T9FSp3MzPqoZUDvVo9cTl9X5T0Jm0Co5uNckblbt7geV3kIuKml2m99bSjuTsPbcSbrXtpe9ww9T2d9d6Qie8ztgSRJHkx6XzF-jXp0XWMC9W1nvXcTs1g92k8gkmTin7zdDm45K

A Região Crítica é dada por:

Figura 9 - Gráfico da Região Crítica



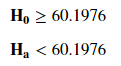
Percebe-se que z0 não faz parte da região crítica, então aceita-se H0. Ou seja, há evidências que a média da concentração em porcentagem de sílica é igual a 2.3268%.

O Intervalo de Confiança de 95% é:

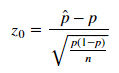
https://lh6.googleusercontent.com/Xvn1-5TYA7Pn93MTDinL43KXSpvpl3cJfk3I-nPdU5uVEVu12NeZaYQGf9Q3kmspe6s9cDnYtWMljQ9FzL5crhjwLVMbI4Cqiujhs4HSMEDU-O_XYuIah5hmRJxSfm3xoGNthRVf

**CASO 3:** Um experimento realizado entre os dias 10/03/2017 e 09/09/2017 afirma que a proporção de sílica do nível BAIXO é menor ou igual a 60,1976%. Para recusar ou aceitar essa afirmação com um nível de significância de 5%, colhe-se 100 amostras aleatórias com uma proporção de 68%.

As definições das hipóteses são:



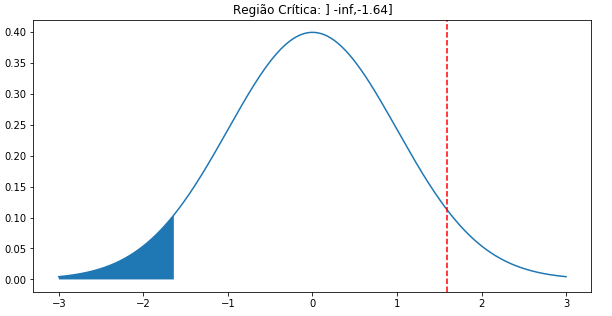
O estimador usado é dado por:

****

**https://lh5.googleusercontent.com/jP9sTWGEmncvun4ak9FGTva2k6svjIoWSHngEAAzEGGzCbAAxCwbRyUYGlFzdU_nUxiCgCU2-VSZu-aRJc3Sua8W-iiF7Is0wWbnP9xbsjTthZE8ie9GDbFlu7NU_wr2CXUeInAD**

A Região Crítica é dada por:

Figura 10 - Gráfico da Região Crítica

****

Percebe-se que z0 não faz parte da região crítica, então aceita-se H0. Ou seja, há evidências que porcentagem do nível BAIXO de sílica é maior ou igual a 60.1976%.

O Intervalo de Confiança de 95% é:

**https://lh3.googleusercontent.com/_y71bX-E9-5QntnJwC6net8o3xF1AGGBA1vg1ozHq0LdqPQYSS3BeHxqN5KLZJalZ3P7kY7IJDlnoCzv2TnXDfZsXFdgx0a40GEY4gesLWDEGsVoyb1CSNZlfVz8SEYIM72WJEol**

**5. SOLUÇÃO PROPOSTA**

A solução proposta para prever a porcentagem de sílica no minério foi utilizar a técnica de regressão linear.

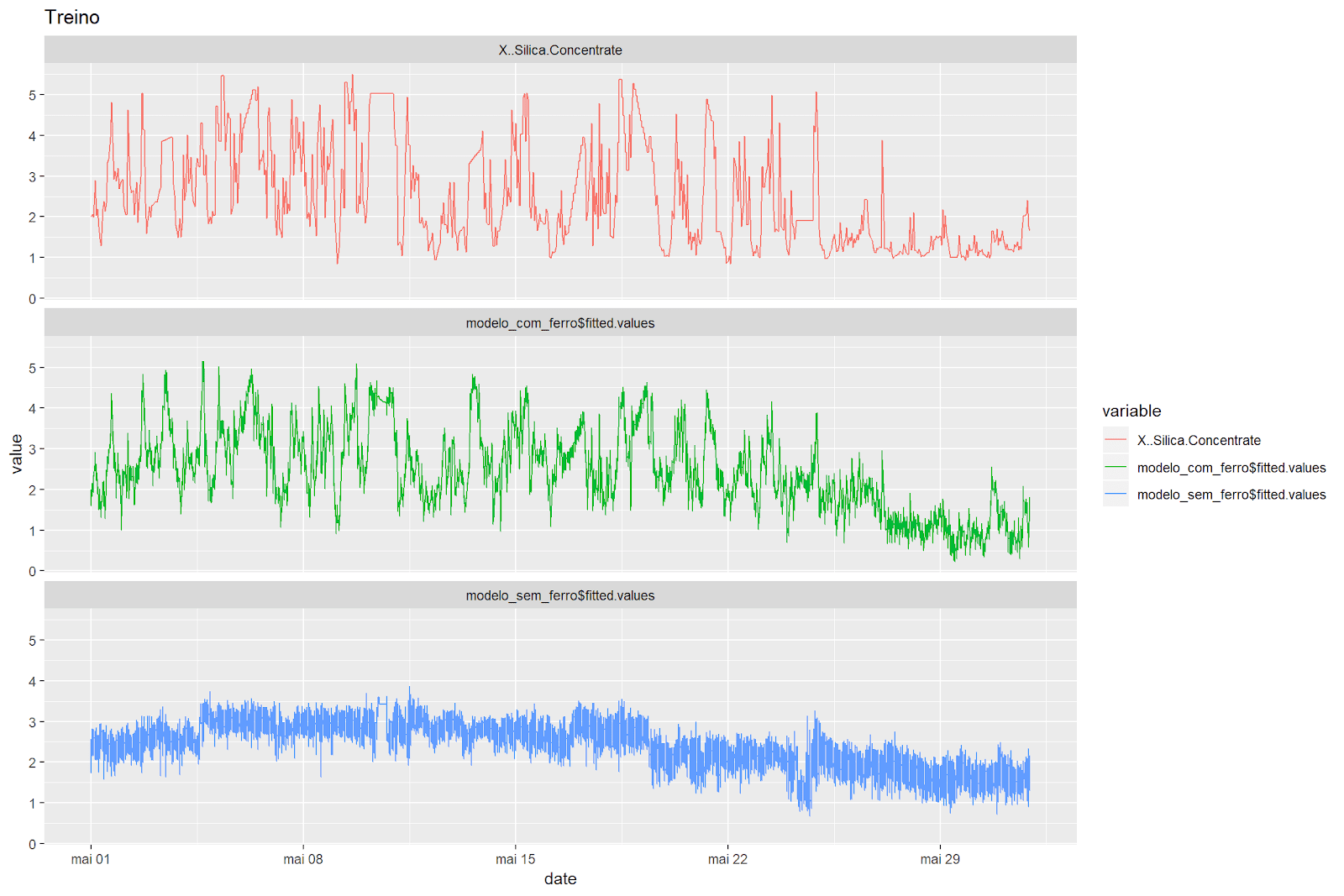
O dataset foi dividido em dois: um para treino e outro para teste. Com o dataset de treino, foram gerados dois modelos de regressão linear: um utilizando todas as variáveis como entrada, e outro que não utiliza a variável de porcentagem de ferro. Em seguida a função step() foi aplicada aos dois modelos para selecionar as variáveis mais significativas.

Os modelos gerados foram então utilizados para prever a sílica no dataset de teste.

**6. VALIDAÇÃO DA SOLUÇÃO**

A seguir é mostrada a série real de sílica e as predições dos modelos para o dataset de treino, e o summary() dos modelos gerados.

Figura 11 - Gráfico de Treino



**Descrição de modelo sem Ferro**

Residuals:

Min   1Q  Median   3Q Max

-3.0598 -0.7337 -0.2029  0.5971 4.3254

Coefficients:

                           Estimate    Std. Error t value Pr(>|t|)

(Intercept)                1.066e+02  2.709e+00 39.350  < 2e-16 \*\*\*

date                      -8.104e-08  1.793e-09 -45.208  < 2e-16 \*\*\*

X..Iron.Feed              -9.551e-02  2.302e-03 -41.486  < 2e-16 \*\*\*

X..Silica.Feed            -6.229e-02  1.741e-03 -35.775  < 2e-16 \*\*\*

Starch.Flow               -2.581e-05  1.829e-06 -14.112  < 2e-16 \*\*\*

Amina.Flow                 2.414e-03  3.086e-05 78.218  < 2e-16 \*\*\*

Ore.Pulp.Flow             -2.176e-03  3.098e-04 -7.025 2.15e-12 \*\*\*

Ore.Pulp.pH               -9.089e-02  6.845e-03 -13.278  < 2e-16 \*\*\*

Ore.Pulp.Density          -1.361e+00  3.511e-02 -38.767  < 2e-16 \*\*\*

Flotation.Column.01.Air.Flow  -1.417e-02 4.987e-04 -28.417 < 2e-16 \*\*\*

Flotation.Column.02.Air.Flow  -6.540e-04 1.226e-04 -5.333 9.69e-08 \*\*\*

Flotation.Column.03.Air.Flow   2.938e-03 5.158e-04 5.697 1.22e-08 \*\*\*

Flotation.Column.04.Air.Flow   1.127e-01 1.015e-03 111.053 < 2e-16 \*\*\*

Flotation.Column.05.Air.Flow  -1.364e-02 6.645e-04 -20.523 < 2e-16 \*\*\*

Flotation.Column.06.Air.Flow   2.216e-03 1.525e-04 14.533 < 2e-16 \*\*\*

Flotation.Column.01.Level -3.764e-04  2.505e-05 -15.026  < 2e-16 \*\*\*

Flotation.Column.02.Level -7.910e-04  2.270e-05 -34.842  < 2e-16 \*\*\*

Flotation.Column.03.Level 4.497e-04  1.806e-05 24.894  < 2e-16 \*\*\*

Flotation.Column.04.Level 3.098e-04  2.790e-05 11.104  < 2e-16 \*\*\*

Flotation.Column.05.Level -1.459e-03  3.682e-05 -39.610  < 2e-16 \*\*\*

Flotation.Column.06.Level 6.739e-04  2.851e-05 23.639  < 2e-16 \*\*\*

Flotation.Column.07.Level -4.574e-04  3.268e-05 -13.997  < 2e-16 \*\*\*

---

Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.038 on 368704 degrees of freedom

Multiple R-squared:  0.1988, Adjusted R-squared:  0.1988

F-statistic:  4358 on 21 and 368704 DF,  p-value: < 2.2e-16

**Descrição do modelo com Ferro**

Residuals:

Min   1Q  Median   3Q Max

-3.2743 -0.4001 -0.0226  0.3833 2.9519

Coefficients:

                           Estimate Std. Error  t value Pr(>|t|)

(Intercept)                1.707e+02  1.650e+00 103.462  < 2e-16 \*\*\*

date                         -9.262e-08 1.090e-09 -84.989  < 2e-16 \*\*\*

X..Iron.Feed               1.068e-02  1.395e-03 7.656 1.92e-14 \*\*\*

X..Silica.Feed             2.581e-02  1.061e-03 24.338  < 2e-16 \*\*\*

Amina.Flow                 4.548e-05  1.896e-05 2.399  0.01645 \*

Ore.Pulp.Flow              2.056e-03  1.869e-04 11.003  < 2e-16 \*\*\*

Ore.Pulp.pH                2.332e-01  4.053e-03 57.523  < 2e-16 \*\*\*

Ore.Pulp.Density           5.136e-01  2.125e-02 24.169  < 2e-16 \*\*\*

Flotation.Column.01.Air.Flow -2.666e-03  3.057e-04 -8.721 < 2e-16 \*\*\*

Flotation.Column.02.Air.Flow -1.903e-03  7.458e-05 -25.511 < 2e-16 \*\*\*

Flotation.Column.03.Air.Flow -8.825e-04  3.170e-04 -2.784 0.00538 \*\*

Flotation.Column.04.Air.Flow  4.560e-02 6.221e-04 73.302 < 2e-16 \*\*\*

Flotation.Column.05.Air.Flow  8.238e-03 4.041e-04 20.385 < 2e-16 \*\*\*

Flotation.Column.06.Air.Flow  6.175e-03 2.198e-04 28.093 < 2e-16 \*\*\*

Flotation.Column.07.Air.Flow -2.710e-03  2.210e-04 -12.263 < 2e-16 \*\*\*

Flotation.Column.01.Level    -1.174e-04 1.511e-05 -7.773 7.70e-15 \*\*\*

Flotation.Column.02.Level    -5.838e-04 1.379e-05 -42.350  < 2e-16 \*\*\*

Flotation.Column.03.Level 4.768e-05  1.090e-05 4.373 1.23e-05 \*\*\*

Flotation.Column.04.Level 2.512e-04  1.453e-05 17.291  < 2e-16 \*\*\*

Flotation.Column.06.Level 6.721e-04  1.575e-05 42.681  < 2e-16 \*\*\*

X..Iron.Concentrate          -7.754e-01 9.693e-04 -799.971  < 2e-16 \*\*\*

---

Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6301 on 368705 degrees of freedom

Multiple R-squared:  0.7049, Adjusted R-squared:  0.7049

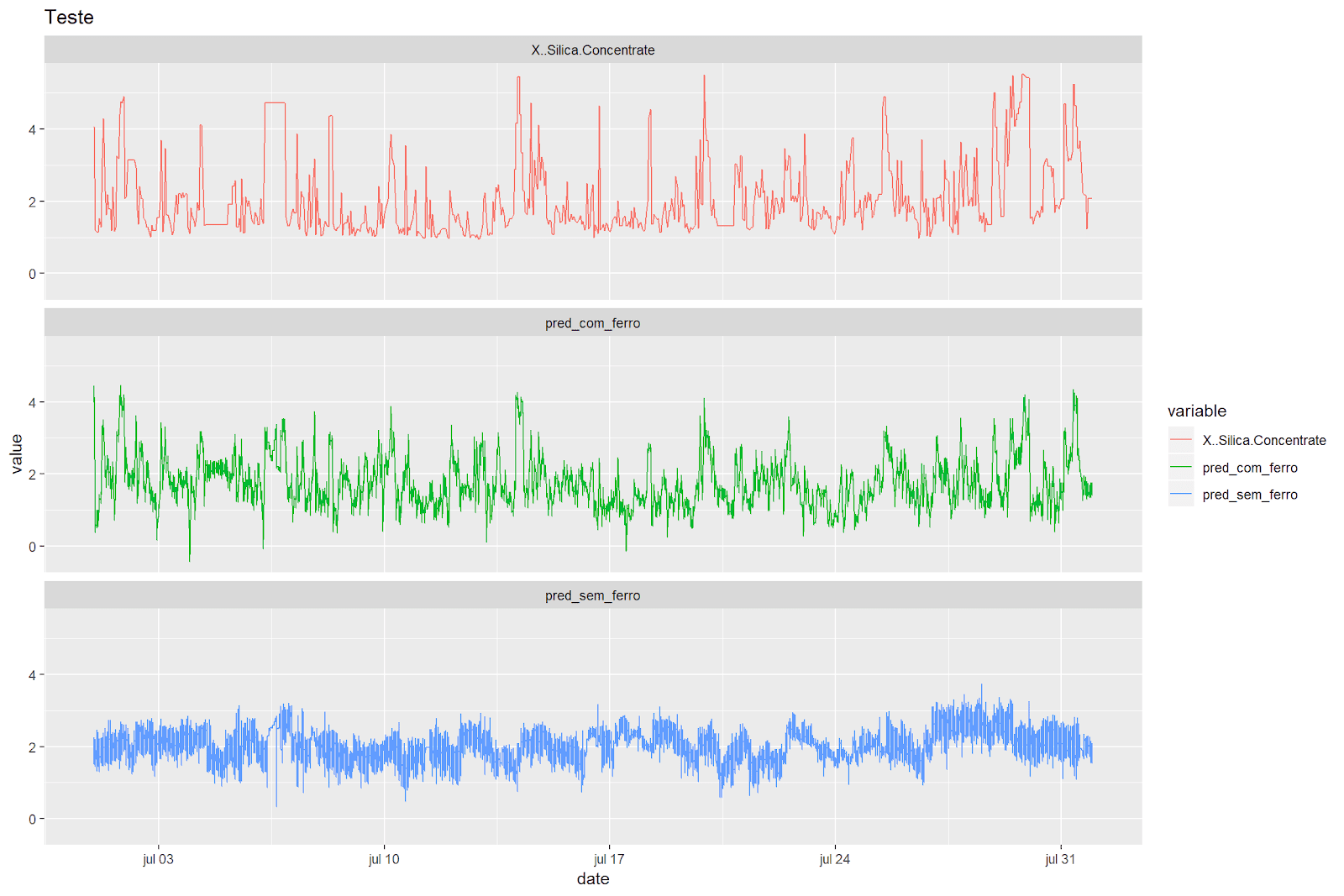
F-statistic: 4.404e+04 on 20 and 368705 DF,  p-value: < 2.2e-16

Observando o R² dos modelos gerados, nota-se que as predições do modelo com ferro foram melhores que as do modelo sem ferro, devido à grande correlação entre sílica e ferro.

**6.3. Previsões no dataset de teste**

A seguir é mostrada a série real de sílica e as predições dos modelos para o dataset de teste.

Figura 12 - Gráfico de Teste



A qualidade da previsão dos modelos no dataset de teste caiu, mas nota-se que a previsão do modelo com ferro ainda foi melhor.

**7. CONCLUSÃO**

A análise estatística permite descrever o comportamento dos dados, evidenciando como eles estão distribuídos no banco de dados, o seu nível de dispersão, as classes que ocorrem com maior frequência e as probabilidades com elas associados.

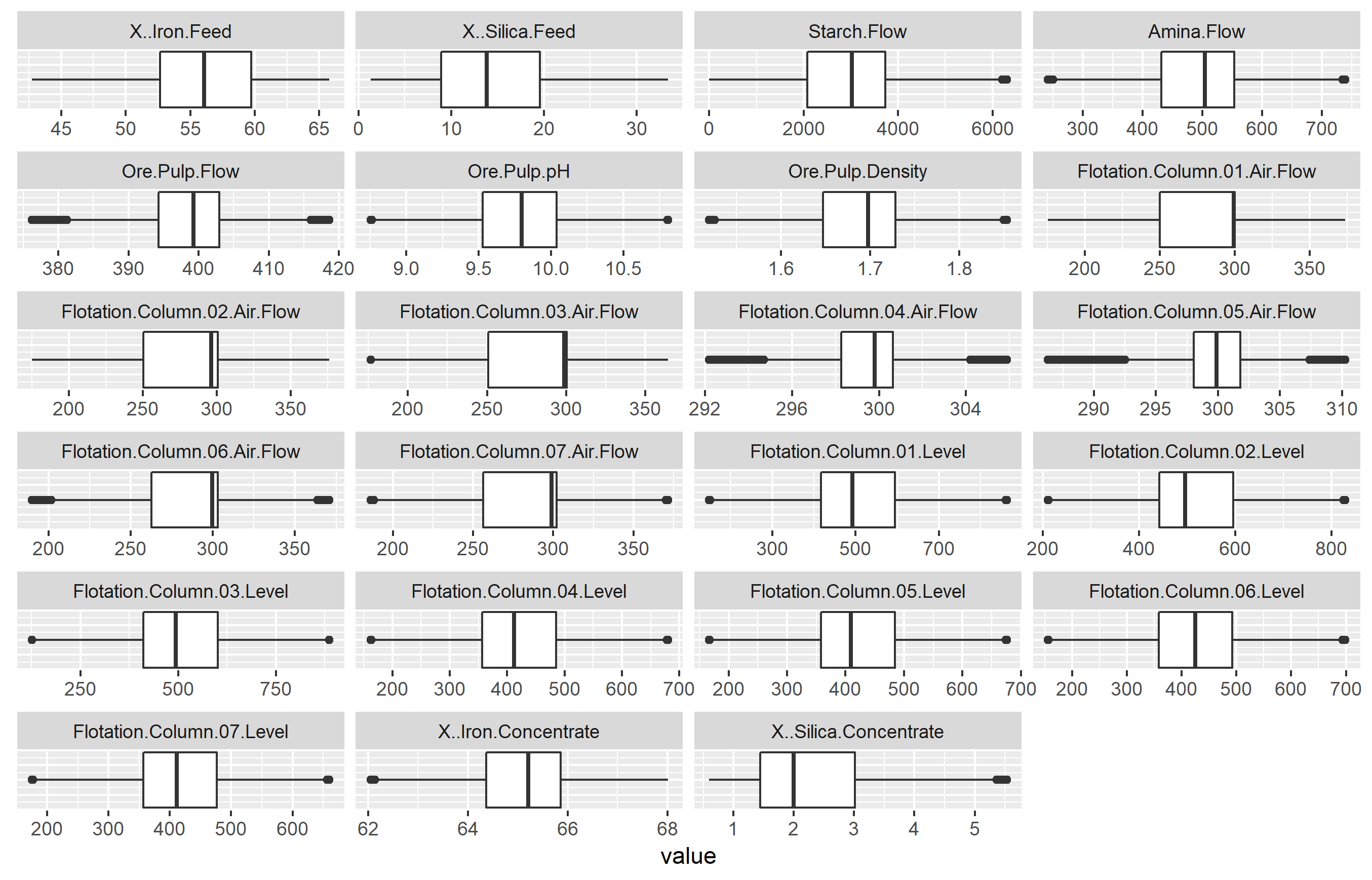
A inferência, através do teste de hipótese, possibilita a aceitação ou rejeição de uma hipótese sobre algum parâmetro populacional, como a média, o desvio padrão ou/e certa ocorrência de um evento dentro dessa população a partir de uma amostra.

Já a estimação permite saber o intervalo que se encontra um parâmetro populacional. E esse intervalo é conhecido como Intervalo de Confiança.

Com tudo isso, pode-se fazer uma comparação com os resultados dos modelos de predição, percebendo que tais valores preditos não estejam muito distantes daqueles valores encontrados por uma análise estatística, inferencial e estimativa.

**APÊNDICE**

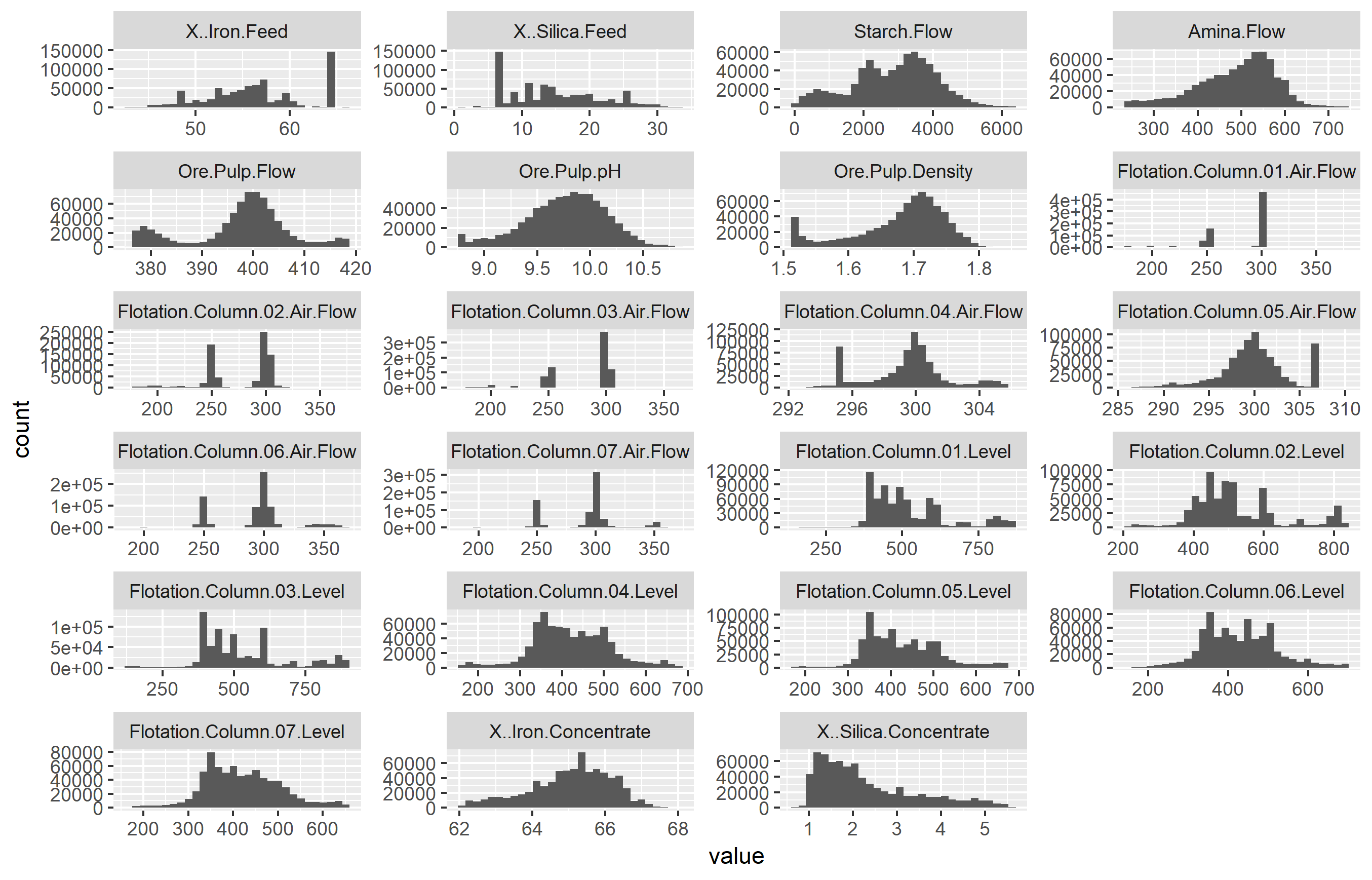
**Figura A1 – Gráfico Boxplot para as 23 variáveis**

****

**Tabela A1 – Descrição dos gráficos boxplot**

|  |  |  |  |
| --- | --- | --- | --- |
| **Iron Feed** | **Silica Feed** | **Starch Flow** | **Amina Flow** |
| Não apresenta outliers e distribuição de dados simétrica. | Não possui outliers e possui uma distribuição levemente assimétrica positiva. | Possui outliers superiores e conjunto de dados distribuídos assimétricamente para o 3Q. | Possui outliers superiores e inferiores e possui uma distribuição levemente assimétrica negativa. |
| **Ore Pulp Flow** | **Ore Pulp pH** | **Ore Pulp Density** | **Flotation Column 01 Air Flow** |
| Possui outliers superiores e inferiores e possui uma distribuição simétrica. | Possui outliers superiores e inferiores e possui uma distribuição simétrica. | Possui outliers superiores e inferiores e distribuição levemente assimétrica negativa | Não possui outliers e o conjunto de dados, é totalmente assimétrico negativamente. Pois a mediana está no 3Q. |
| **Flotation Column 02 Air Flow** | **Flotation Column 03 Air Flow** | **Flotation Column 04 Air Flow** | **Flotation Column 05 Air Flow** |
| Não apresenta outliers e apresenta distribuição assimétrica negativa | Possui outliers inferiores e o conjunto de dados, é totalmente assimétrico negativamente. Pois a mediana está no 3Q. | Possui outliers inferiores e superiores e o conjunto de dados é assimétrico negativamente. | Possui outliers superiores e inferiores e os dados apresentam uma dispersão simétrica. |
| **Flotation Column 06 Air Flow** | **Flotation Column 07 Air Flow** | **Flotation Column 01 Level** | **Flotation Column 02 Level** |
| Possui outliers inferiores e superiores o conjunto de dados, é totalmente assimétrico negativamente. | Possui outliers inferiores e superiores o conjunto de dados, é totalmente assimétrico negativamente. | Apresenta outliers inferiores e superiores e o conjunto de dados são assimétricos positivos. | Apresenta outliers inferiores e superiores e o conjunto de dados são assimétricos positivos. |
| **Flotation Column 03 Level** | **Flotation Column 04 Level** | **Flotation Column 05 Level** | **Flotation Column 06 Level** |
| Apresenta outliers inferiores e superiores e o conjunto de dados são assimétricos positivos. | Apresenta outliers inferiores e superiores e o conjunto de dados são assimétricos positivos. | Apresenta outliers inferiores e superiores e o conjunto de dados são assimétricos positivos. | Apresenta outliers inferiores e superiores e o conjunto de dados são simétricos. |
| **Flotation Column 07 Level** | **Iron Concentrate** | **Silica Concentrate** |  |
| Apresenta outliers inferiores e superiores e o conjunto de dados são assimétricos positivos. | Possui outliers inferiores e o conjunto de dados são simétricos. | Possui conjunto de dados superiores e conjunto de dados assimétricos positivos. |  |

**Figura A2 – Gráfico de Histograma para as 23 variáveis**

****