Logo

Description automatically generated with low confidence

**Lappeenranta University of Technology**

**School of Business and Management**

BM20A6100 Advanced Data Analysis and Machine Learning

**Project work: Data Pretreatment**

**Mining Process – Level A**

Otto Pietikäinen 0518917

Juho Kauppala 0506701

Josh Reeder 0594496

Table of Contents

[Removal of downtime 1](#_Toc83668979)

[Removal of Iron Outlet 3](#_Toc83668980)

[Outlier Removal 3](#_Toc83668981)

[Other Potential for Data Reduction 4](#_Toc83668982)

[Appendix 5](#_Toc83668983)

# Removal of downtime

Looking at a time plot of the data reveals a large period where no data is collected, as shown in Figure 1 below. The other variable time plots are shown in the Appendix by Figures A1-A4.

**Timeline

Description automatically generated with medium confidence**

Figure - Time plots of last 3 variables in dataset

The process was likely shut down for this period. This may potentially affect the model, as the downtime may or may not have change the operation of the system in a way not captured by the process variables. Since the dataset is very large and there is only a small amount of data before the downtime, this portion of the data has been removed, as shown in Figure 2.

Timeline

Description automatically generated

Figure - Time plots of last 3 variables after downtime removal

# Removal of Iron Outlet

Since the iron outlet concentration is determined at the same time as silica outlet concentration, it is assumed to be unavailable in any real-world prediction scenario. Including it would overestimate the model’s actual prediction accuracy. Therefore, this variable can be removed from the dataset.

# Outlier Removal

Outlier removal is required before the data is manipulated for data reduction. Since the multivariate system in question is assumed to have complex dependencies between the variables, the outlier detection system should take multiple variables into account simultaneously. To do this, PCA was used and T2 values were obtained for each data point. The T2 value represents how far the data point is from the (0, 0) of the model, and because data was normalized before PCA, the data points furthest away from the center point get the greatest T2 values and are detected as outliers. After studying the T2 values of the data, was set as the condition for a data point to be considered an outlier. This threshold seems to eliminate the biggest outliers, however, it is subject to change in the future if necessary.

Histograms showing the tails of the T2 values before and after outlier removal are shown in Figure 3.



Figure - Histograms of T2 values before (top) and after (bottom) outlier removal

# Other Potential for Data Reduction

It should be noted that many of the variables update every 20s, while other, including the silica outlet concentration, are only updated every 1hr. The 20s update variables then have 180 data points per 1hr update variable data point. One technique to reduce the data may be to combine 20s update variable samples into a longer time frame, by averaging the individual samples across the new time frame. Also, while not data reduction, since the 1hr update variables are technically changing, the data doesn’t reflect this and keeps it at a constant value until another update in an hour. The performance of the model may potentially be improved by interpolating the 1hr update variables between the exact times when they are found with lab measurements.

# Appendix

Chart

Description automatically generated with low confidence

Figure A1 – Time plots of variables

A picture containing timeline

Description automatically generated

Figure A2 - Time plots of Variables

A picture containing graphical user interface

Description automatically generated

Figure A3 - Time plots of variables

A picture containing chart

Description automatically generated

Figure A – Time plots of variables