Sebastian Martin

Master thesis

Implementing machine learning into industrial plant energy supervision

School of technology and innovations
Master thesis
Automation technology

Vaasa 2021

VAASAN YLIOPISTO
School of technology and innovations
Tekijä: Sebastian Martin
Tutkielman nimi: Koneoppimisen soveltaminen teollisuuslaitoksen energia tehokkuuden seurantaan
Tutkinto: Diplomi-insinööri
Oppiaine: Automaatio ja tietotekniikka
Työn ohjaaja: Prof. Timo Mantere
Valmistumisvuosi: 2021 **Sivumäärä:** 50

TIIVISTELMÄ:

Koneoppiminen on konsepti, jossa kone pystyy oppimaan aikaisemmista tapahtumista. Koneoppimisien kiinnostavuus on kasvussa ja tämä on yksi syy, miksi juuri tämä työ valittiin.

Tämän tutkielman tarkoituksena on analysoida, miten hyvin koneoppiminen pystyy analysoimaan energian käyttöä sinkin tuotannossa. Tämä työ on jaettu kahteen osaan, koska elektrolyysin energian-käyttö ei suoraan vaikuta pasuttamon ja rikkihappotehtaan energia kulutukseen. Pasutus ja rikkihappotehdas ovat prosessikokonaisuus, joten energia muutokset pasutuksessa vaikkutavat happotehtaaseen. Prosessissa on tuhansia signaaleja ja jotta saisimme analysoinnista mahdollisimman hyvän, oli olennaista valita vain tärkeimmät signaalit. Elektrolyysi osassa käytettiin Random Forest menetelmää ja pasuttamoon ja happotehtaan analysointiin käytettiin neuroverkko menetelmää. Tämän työn aikana käytettiin useita ohjelmia kuten SQL:ää käytettiin datan poimintaan, Microsoft Exceliä käytettiin data muokkaamiseen ja Orange data mining ohjelmaa käytettiin analysointiin ja visualisointiin.

Tulokset osoittavat, että testit onnistuivat ja näyttävät myös, miten koneoppi pystyy luokittelemaan eri tyyppistä dataa. Myös, siten miten koneoppiminen priorisoi signaaleja, tämä riippuu todella paljon siitä, miten me katsomme signaalien painoarvoa. Mutta kuin simuloimme vääränlaisia luokkia niin tulokset eivät olleet niin hyviä kuten olimme toivoneet, ja tämä riippuu todennäköisesti siitä että simuloitavat arvot eivät ole saman tyyppisiä kuten miltä arvot olisivat todellisessa vikatilanteessa.

AVAINSANAT: Energian analysointi sinkki tuotannossa, Koneoppiminen, Elektrolyysi, Pasutus, Happi liuotus

UNIVERSITY OF VAASA
School of technology and innovations
Author: Sebastian Martin
Topic of thesis: Master thesis : Implementing machine learning into industrial plant energy supervision
Degree: Master of Automation technology
Major of Subject: Automation technology
Supervisor: Prof. Timo Mantere
Year of completing thesis: 2021 **Pages:** 50

ABSTRACT:

Machine learning is a concept in which a machine can learn from its past experiences. The machine learning trend is rising in popularity, and this is the reason this method has been chosen for this thesis.

The purpose of this thesis is to analyze how well machine learning analyzes energy usage in a zinc plant. This thesis work has been split into two pieces as the electrowinning works separately and the roaster and sulphuric acid plant work together. The roaster and sulphuric acid plant are connected and therefore the power changes affect the whole process. There are thousands of signals from the plant so to make the analyses as good as possible only the essential ones were chosen. The electrowinning data was analyzed with Random Forest and Roaster and sulphury acid data was analyzed with Neural Network. During this thesis multiple programs have been exploited such as SQL was used to extract the data, Microsoft excel was for editing data and the analyses were executed with Orange data mining program as it is a good method to test and visualize data.

The results show that the tests are successful and how the different categories can be found with the help of machine learning. And that when watching what the machine learning methods priorities as important factors, it is very different from what we consider as vital information. However, when simulating classes that were not successful there was not as much success as when the simulations were done with the imaginary situations and the values did not match what a real event would look like.

Keywords: Energy analysis in zinc production, Machine learning, Electrowinning, Roaster and sulphuric acid plant

Table of contents

Abbreviations

NN	**Neural Network**
RF	**Random Forest**
SQL	**Structured Query Language**
AI	**Artificial Intelligence**
ML	**Machine Learning**
VPN	**Virtual Private Network**
CA	**Classification accuracy**
SL	**Supervised Learning**
UL	**Unsupervised Learning**
SSL	**Semi-supervised Learning**
RL	**Reinforcement Learning**
SVM	**Support-Vector Machine**
kNN	**k-Nearest neighbor**
NB	**Naive bayes**

1 Introduction

There has been an increase in interest in machine learning during the last two decades. This is thanks to computers being more efficient and the available data online (M.Sugiyama, M.Kawanabe). During this thesis, it was hard to find published papers in the same field since the curiosity of exploiting machine learning is still quite new. (D. Narciso, F. Martins, 2020) Machine learning is a way to engineer a system to mathematically understand the data that is fed into it. The system will learn and train on earlier data and then use that in solving problems. In machine learning, it is most common to use data that does not change over time. Boliden uses a lot of energy and therefore reducing energy usage is a very important matter. With today's technology there are many possible ways to reduce energy usage and a common way to do this is with the help of machine learning. (D. Darlis, M. Latip, N. Zaini, H. Norhazman, 2020)

In this thesis, we use data that is collected at Boliden Kokkola's zinc plant over time. The data that is used in this thesis has not been specifically collected for this purpose. In this thesis, parts that are being analyzed of which two are connected and analyzed together and one is analyzed separately. The data that is analyzed is done so with Orange data mining program. This is a data analysis and visualization program (Orange, 2013).

1.1 Objective of thesis

In this thesis, the main objective is to firstly find the correct data signals for the parts that are being analyzed which are electrowinning, and then roaster and sulphuric acid plant. The roaster and sulphuric acid plant are connected and therefore they are being analyzed together.

During this thesis, it is to test machine learning in zinc production and see how good results it is possible to get by analyzing energy behavior. And then testing that on new parts of data that have not yet been processed by the machine learning models. This will

help in the future with finding errors, classifying, and even predicting certain types of errors with the help of earlier data.

1.2 Structure of thesis

This thesis consists of seven chapters. The first chapter introduces the subject and the objective of the thesis. The second chapter contains relevant theory and technology that has been exploited during this thesis. The third chapter goes through what some of the challenges were during this thesis and how they were solved. The fourth chapter explains what type of data was used during this thesis. The fifth chapter explains how the theories from chapter two have been implemented. The sixth chapter shows the results from all of the parts of data and in the last chapter we conclude and openly discuss how the thesis was and some personal opinions.

2 Foundations

In this chapter, the foundations of the programs and methods used in this thesis are explained. First off machine learning will be explained and how it can be utilized and what different methods there are and when to use which ones, and what has been used in this thesis. Secondly, remote use has been an important tool during this thesis because of the ongoing pandemic. After that the Boliden's process will be briefly explained, mostly focused on the smelters as that is the part of the process that has been analyzed during this thesis. And lastly the data preparation methods and programs that have been used during this thesis.

2.1 Machine learning

Defining machine learning is fuzzy but the general definition of it is, using artificial intelligence (AI) to learn patterns to analyze new data. (A. Mechelli, S. Vieira 2019)

For us, humans learning from our experiences is something very simple. We take it for granted. But that is not the case for machines, and we must teach them how to do it. In this thesis, we teach a system how to classify data, and this way we can automate the system to predict future data.

There are multiple ways to exploit machine learning techniques. In this thesis, I will introduce four common methods. They have different strengths and weaknesses and in a project, it is important to choose the correct one to not waste time and resources.

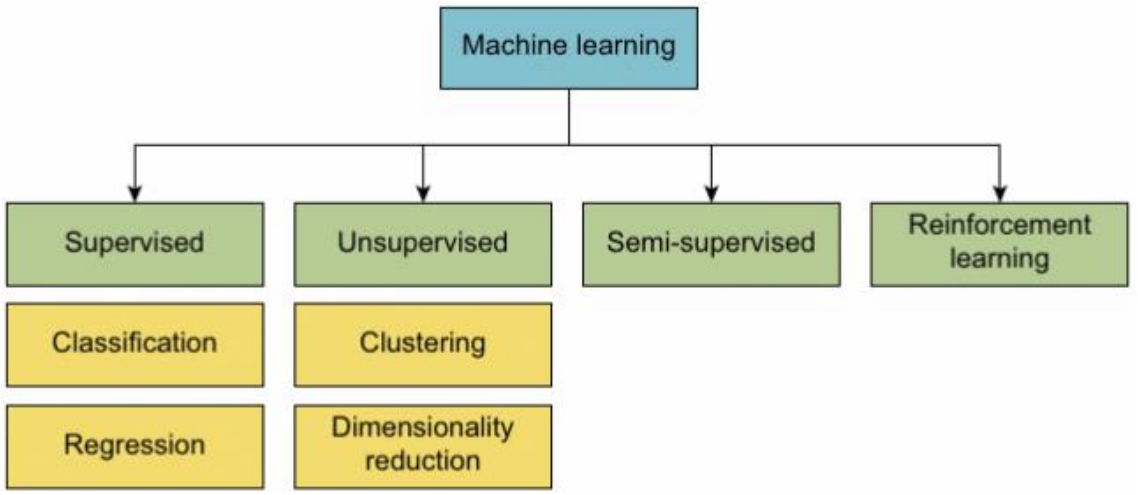

Figure 1. (A. Mechelli, S. Vieria, 2019) Machine learning and different types of analyzing methods.

2.1.1 Supervised learning

In this method, we have a data set where we teach the system what the correct results are. The system will then analyze the parameters, adjust their weight accordingly to come to the same results as us. Some values have positive and some have a negative effect on the result. We focus on creating the best algorithm that understands the relationship between the parameters and the results. (A. Mechelli, S.Vieira 2019: 9-11)

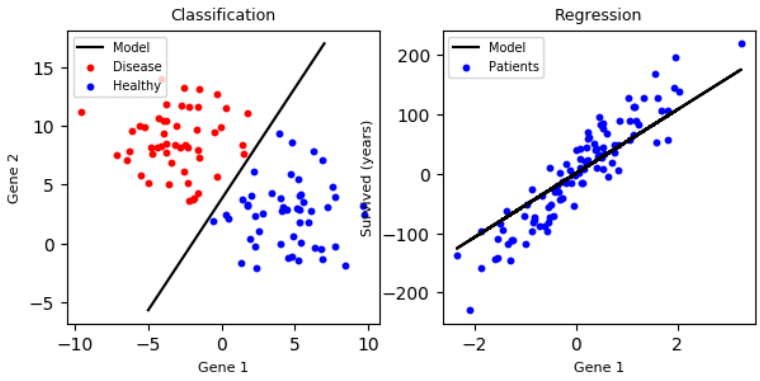

Figure 2. Classification creates classes depending on their values whereas regression sets a numeric value to them.

Multiple techniques have been tested during this thesis to see which was best fitted into this situation.

2.1.1.1 Classification

This method focuses on setting observations into groups, also commonly used as labels or classes. The reason why this is a popular approach is that a lot of problems can be simplified into categories. These types of algorithms can be used in different types of fields like analyzing brain disorders, car dealership salesmen, and industrial plant energy analysis. (A. Mechelli, S.Vieria 2019: 10)

2.1.1.2 Regression

With this method, we teach the algorithm what kind of result we want but we give it a numeric value. The algorithm will then map the inputs so that it can result in a similar output as it has been taught. Regression is like classification, but it is continuous. Regression is often used in economic fields such as risk management or the stock market. (K. Murphy 2012: 9-10)

2.1.2 Unsupervised learning

In unsupervised learning the target value or group is unknown, and the goal is to understand the structure of the data as there is only input data and no output data. The result wanted out of these analyses is to find regularities in the input. There is a structure in the inputs that have certain patterns that occur, some more than others, and to understand which ones do happen and which ones do not. This is called density estimation in statistics. (Alpaydin 2010: 11-12)

2.1.2.1 Clustering

Clustering can be done to help to find specific groups from a dataset. This can be done in e.g., customer relations where the company wants to find a specific customer group and see how their advertisement can be improved. (Alpaydin 2010: 11-13)

2.1.2.2 Dimensionality reduction

This type of method is useful in a scenario where the amount observations is a lot less than the number of features. In these cases, overfitting is a normal issue that must be dealt with by removing features that are not giving vital information that affects the result. (A. Mechelli, S.Vieria 2019: 12)

2.1.3 Semi-supervised learning

This method is used when parts of the data are missing, from either the labels or the target data. These kinds of problems is what are solved by integrating unlabeled data. In some cases, this method performs even better than supervised learning. (A. Mechelli, S.Vieria 2019: 12) An assumption can be made to see when semi-supervised learning and that is that when two events x_1 and x_2 are close to each other, then also the output y_1 and y_2 should be close to each other. Two common methods are the clustering assumption method which rules if certain points are in the same cluster then they are likely to belong in the same class. And Manifold assumption where if there are a high amount of dimensions it is reduced into a smaller amount of dimensions to reduce the curse of dimensionality. (O. S. Chapelle, O. Chapelle, B. Schölkopf & A. Zien (2006).)

2.1.4 Reinforcement learning

Tasks that are optimal for this method are when the learning is a so-called decision-making agent. This agent is set to do a task that requires multiple steps, and the steps it takes are rewarded or punished depending on if they are decisions that get us closer to the end goal or further away from it. This can be viewed as a teacher, although it does

not teach in advance, and only gives critic after each completed step. After a set of steps, the machine will start to understand what type of steps the best is to reach the end goal, this type of method can be exploited in teaching a chess bot. A chess bot will learn when it has done something wrong by losing points or current game score, and the opposite when it has done something correctly and then it wins games or gets a better score. (J. Mueller, L. Massaron 2016, p.169)

2.2 Remote access

With the current ongoing global pandemic, the use of distance work and studies has increased. In this thesis, I have used two methods to get remote access, first VPN (Virtual private network) and then remote desktop.

VPN creates a connection between a network and an outsider. This connection is created to secure sensitive data to be transmitted. VPN is often used in corporations by employees that want to work from home and still be connected to the company intranet. (CISCO)

Remote desktop is a program that can reach a computer that is not located close to you. This means that even if the person is sitting at his home desk with his own computer, can access a computer located in the factory and use the programs and see the files that are located locally on the work computer. (Remote Desktop clients)

2.3 Boliden process

Boliden is a large company in the metal field. They focus mainly on creating zinc, lead ingots, copper cathodes, gold bars, and silver granules. Their process starts from exploration, in existing mines and new locations. During these explorations, the goal is the earlier mentioned resources but also nickel, gold, palladium, platinum, and silver.

After a target has been met and the decision to start mining the next is to mine in that underground mine. The minerals found will be extracted and separated, after this, the minerals will be shipped to smelters.

The last part of the process is the smelters, which this thesis focuses on. In the smelters, the minerals are refined to pure metals. In this thesis, the energy analysis has been done at Boliden Kokkola factory, which is Europe's second-largest zinc factory.

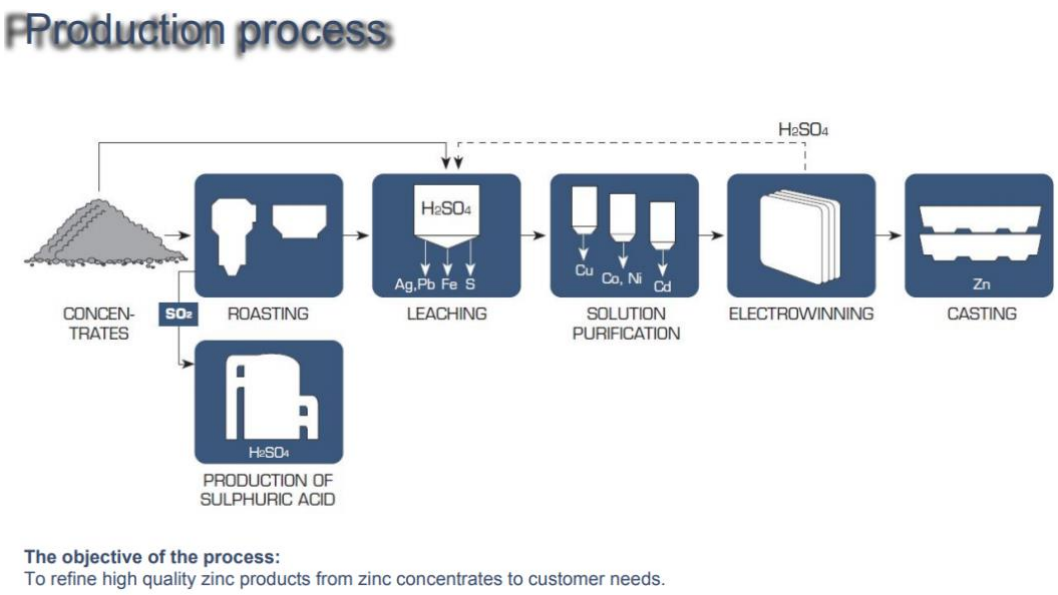

Figure 3. Overall view of the zinc production process (Boliden, 2018)

The process starts when the zinc concentrate has been shipped to Kokkola. Then approximately 2/3 of the concentrate will be fed into an exothermic roaster process where the zinc is processed into an easily dissolvable form (ZnO). The Sulphur from the concentrate is also burned and formed into sulfur dioxide (SO2) gas, which will go to the sulphuric acid production. The remaining concentrate is directly fed into the leaching process.

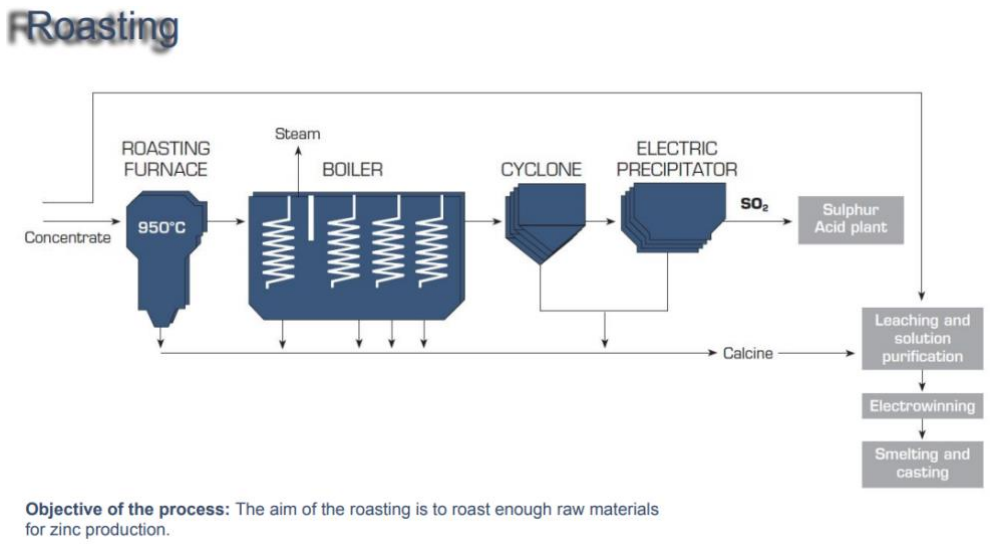

Figure 4. A picture of the roaster. (Boliden, 2018)

The extracted Sulphur is made into sulphuric acid in the sulphuric acid plant which works parallel with the roasting as seen in Figure 3. The overflow steam is generated into electricity and district heating that is operated by another company. The zinc calcine is then grinded in ball mills and then conveyed into silos, once done in the silos they go to leaching and purification where the concentrate is dissolved by sulphuric acid solution. From the zinc concentrate, we get a raw solution which is set through three phases which remove the copper (Cu), then copper, cobalt (Co), and nickel (Ni), and in the last phase cadmium (Cd) is removed, after this we have clean zinc sulfate solution.

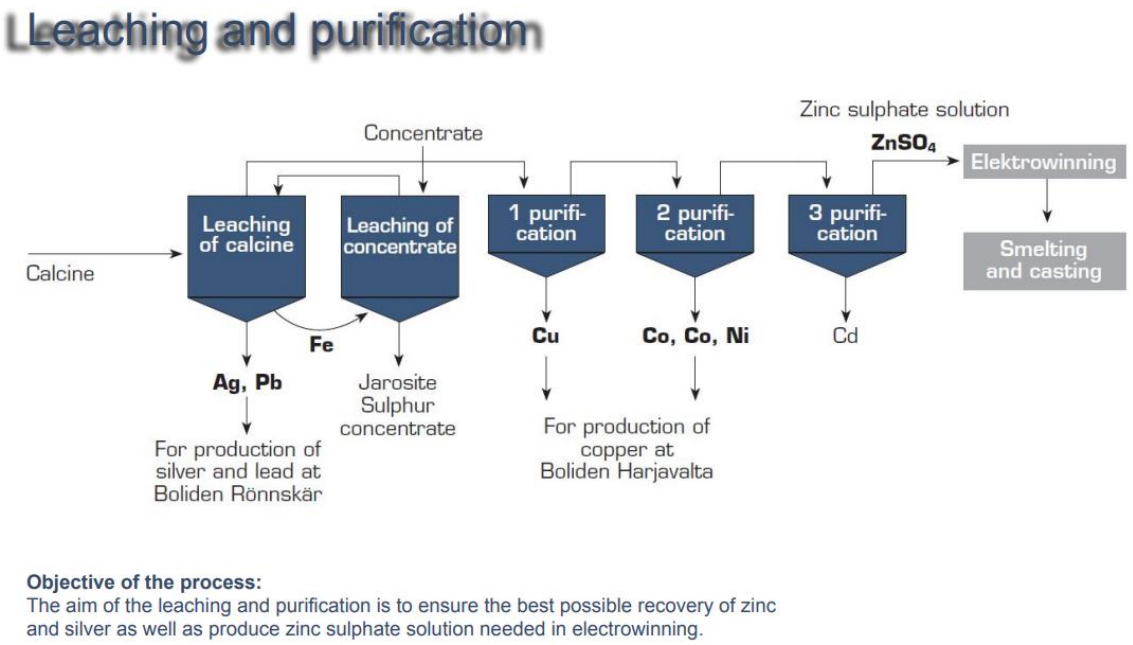

Figure 5. Picture of the leaching and purification process. (Boliden, 2018)

The following step is the electrolysis step where aluminum cathodes and anodes are used to collect the zinc with the help of electric current. This gives zinc cathodes with the purity of 99.995%, the electrolysis interval time is approximately 36 hours.

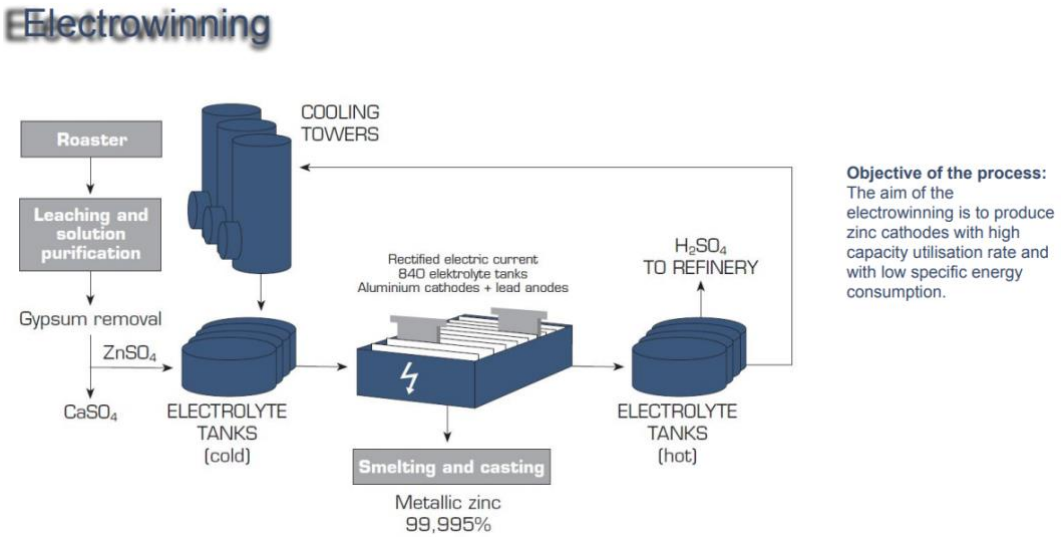

Figure 6. This figure represents how the Electrowinning process looks like. (Boliden, 2018)

And lastly, the process moves on to the foundry where the electric furnaces melt the zinc cathodes. The product then varies on the customer's demand. (Boliden, 2018)

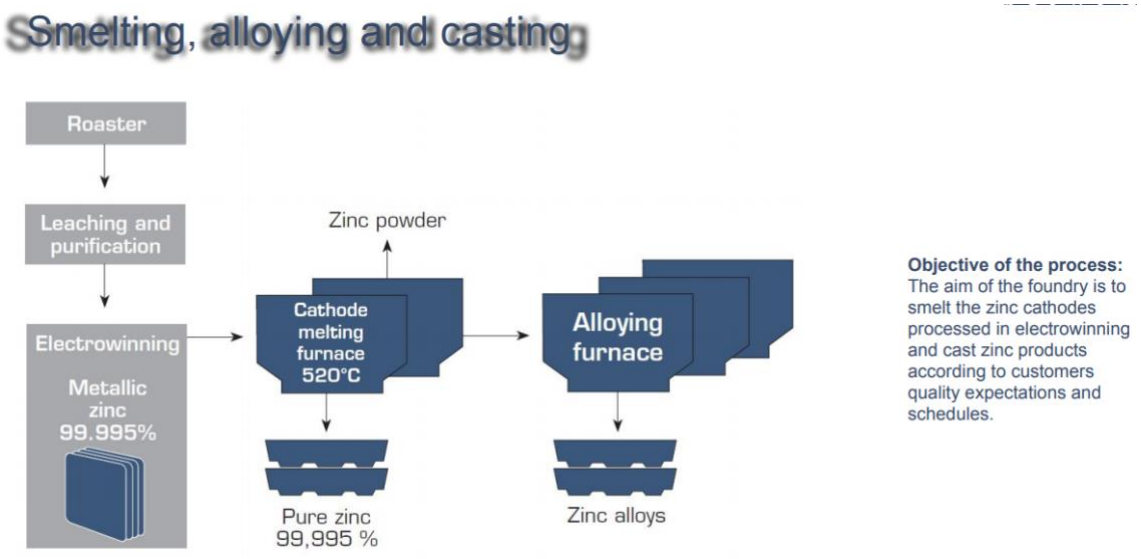

Figure 7. This figure shows how the foundry works. (Boliden, 2018)

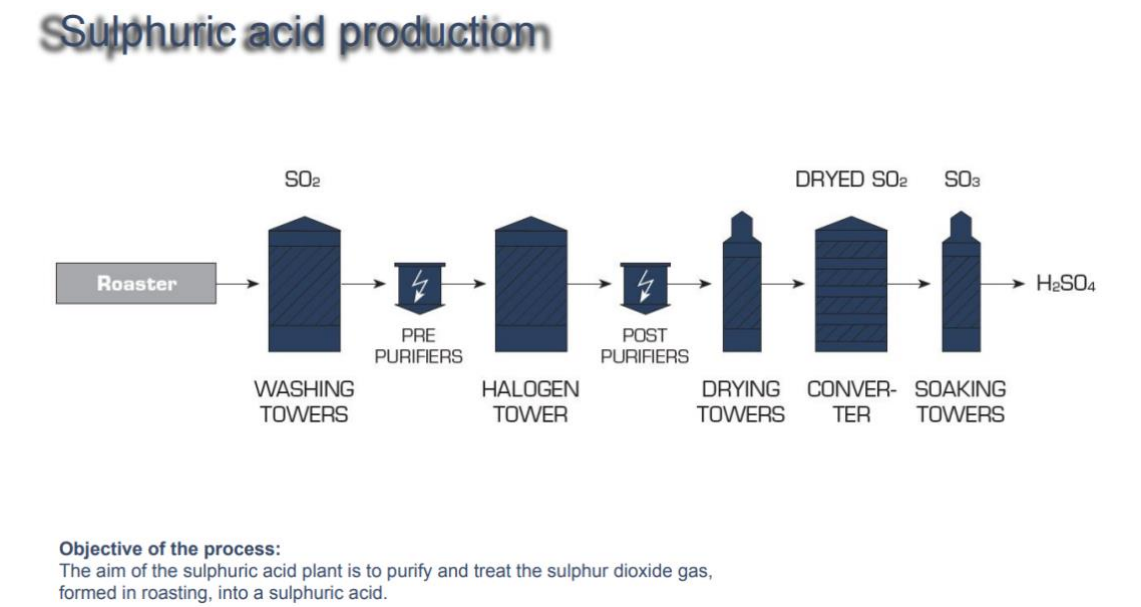

Figure 8. Picture of the sulphuric acid plant which works with the roaster. (Boliden, 2018)

2.4 Source of data

The electrowinning was analyzed separately and then the roaster and sulphuric acid plant were analyzed together as they are connected with two main air blowers, 2 SO2 blowers which are located at the roaster, and one main gas blower which is located at the sulphuric acid plant.

The data that was gathered for the electrowinning is represented in the following picture. 840 electrowinning cells are connected in groups of 30 cells. The groups are called row pairs, 'rivit' in the picture below. One row pairs energy consumption is the smallest unit that can be monitored during this thesis. The current that goes through the cells and cell groups are fed by 4 rectifiers, (TS 1 – 4). Each rectifier feeds 210 cells in 7 cell groups in its own separate current loop, of which each capacity has about 30 MW power. The total effect of the electrowinning process shown in the picture is approximately 120MW. Each part of the picture has its own signals. For example, "RIVIT" meaning rows have their own signals, as well as the four current loops which contribute to the total effect of the electrowinning.

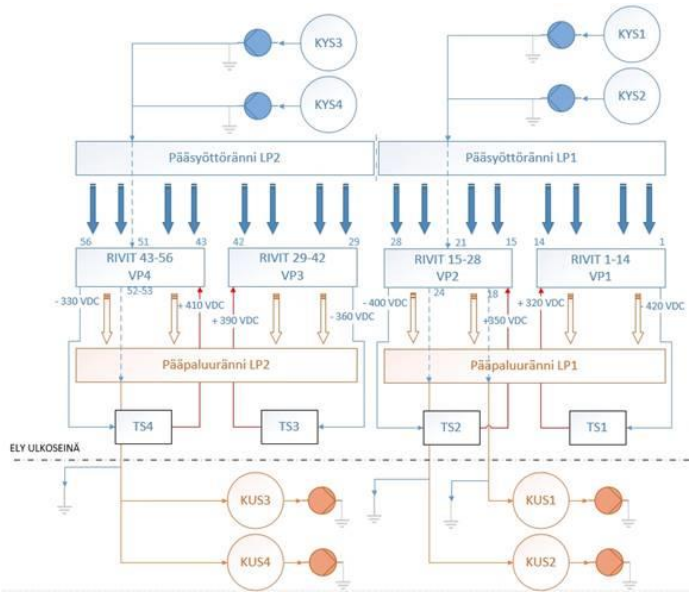

Figure 9. This picture represents electrowinning. (K. Palola, 2020)

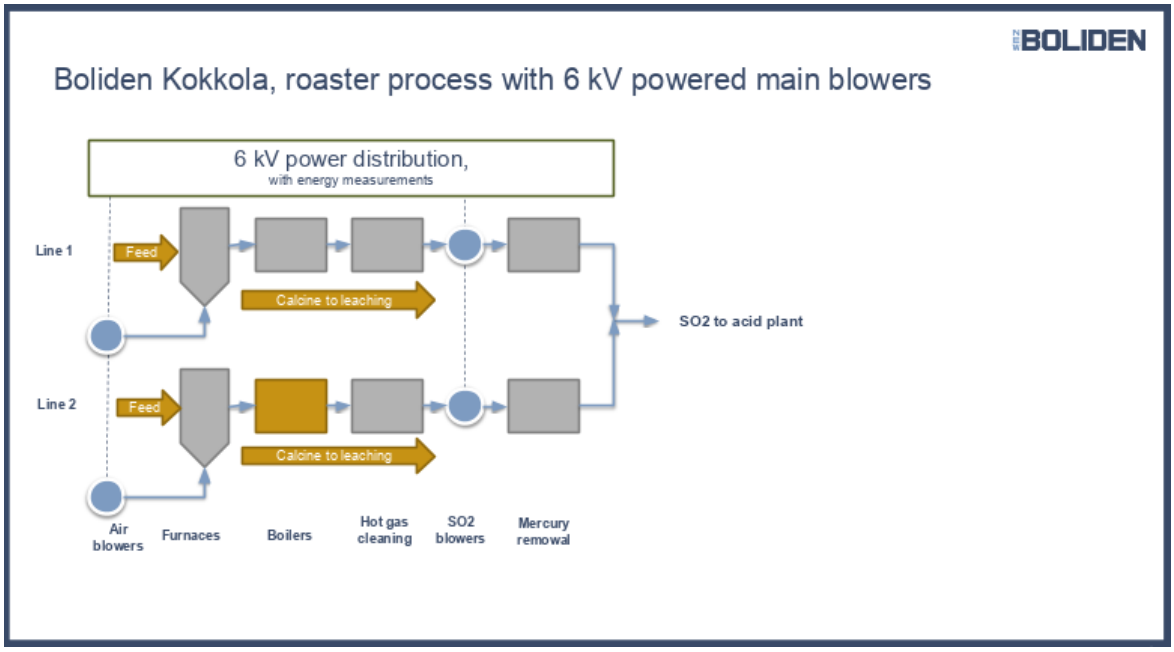

Figure 10. This picture represents what the process looks like in the roaster.

In the roaster either line one is active or line two, which one is active does not affect the total effect of the roaster, but the specific signal values change a lot since one of the lines is completely turned off.

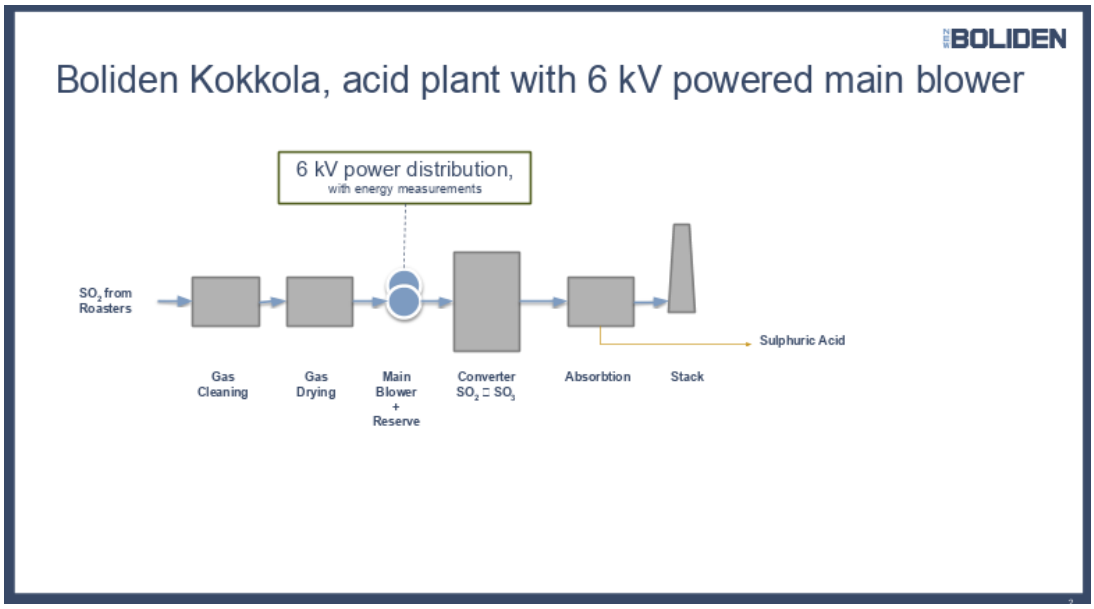

Figure 11. This picture represents the sulphuric acid part.

This picture of the sulphuric acid plant is connected with the earlier picture of the roaster, meaning that if the power in this part lowers it can lower in the roaster, but it can also mean that it increases, in case there is something wrong with the sulphuric acid motor and the roaster has to work harder to get the air to flow.

2.5 Data preparation

Data preparation is an important part of machine learning since unclear data can mislead the algorithms to the wrong results. In this thesis the data preparations have been filling null values where the empty values were filled, classification to set the classes so we could use supervised learning methods, and data modifications for simulations to learn the algorithms what type of situations are wrong and which ones are correct. Programs that have been used are SQL coding in the internal program, Excel, and Orange.

2.5.1 SQL

With computers storing data it is important to have a way to access it easily. SQL is a database-specific programming language. It is a tool used to create maintain and extract certain parts of data from databases. (A. Taylor, 2013: 5).

In this thesis, SQL has been used to extract certain signal values from certain time intervals.

2.5.2 Orange

Orange is a program that can be used to analyze and easily visualize data. It is a way to program via a graphic interface by connecting the different widgets one wants to utilize (Orange, 2013). Being able to utilize a graphic interface rather than coding everything by hand saves a lot of time in situations where it is unclear how to come to the right conclusion. As it is possible to test which method is the best without having to code it by hand.

2.6 Summary

In this chapter, we have learned that machine learning is a method that can be utilized to learn the patterns of a data set and how to utilize them. It is important with quantity and quality when learning a system. The more data and the better we have prepared it, the better the result is.

3 Challenges

In this chapter some of the challenges will be explained and how they have been solved. Some of the challenges have been to work offline, this had to be done for security reasons. And the challenge of working offline is it takes a lot more time to add new programs and data to the dataset. Also, in the program Orange that has been used during this thesis, it is not possible to add widgets while offline. Data preparation had to be done as it was extracted with SQL, so it was quite messy, and in this chapter, it is explained how these issues were solved. During this thesis, multiple machine learning methods were tested. And in this chapter, it is explained how it was done.

3.1 Working offline

During this thesis, there has been an additional challenge as the pandemic has forced the writer of this thesis to work distantly.

This was done with Cisco AnyConnect to connect to Boliden's internal network. After the connection was stable remote access was used to be able to connect to the computer that is stored in Kokkola Boliden industrial plant. On this computer, there is stored a virtual computer that the data was stored on, and the thesis was made. Unfortunately, the virtual computer does not have internet access so all the data that was added onto it, was transferred from the local computer placed in Kokkola. Adding data to the thesis analysis had to be fetched via the local computer and then transferred over to the virtual computer. The same process had to be done when adding new programs such as anaconda and orange, and then run the install file on the virtual computer.

3.2 Data preparation

The data that was extracted from the system was with the help of Historian interactive SQL, this is a program that is used to view and extract data. This program is meant for

users with some SQL knowledge as it does not have wizards in it (Ge digital solutions, 2021).

The values that were extracted with Historian interactive SQL were in metric prefixes. A metric prefix is the unit prefix at the end of a number explaining the multiplier of that value (k for kilo, M for mega, and G for Giga). Although the metric prefixes were missing meaning some of the values that were supposed to be added together had to be changed into the same metric prefix. This was done in Microsoft Excel by multiplying the values, so they matched with the same metric prefix.

By extracting the values, they came with an additional symbol, this caused issues since Microsoft Excel does not recognize this as an integer. This issue was simply fixed with the find & replace gadget that is built into Microsoft excel.

3.3 Testing

With the built-in testing method in Orange and the common usage of using 80% of the data to learn the algorithm and 20% to test it to see which method is the best. With this style both Neural Network (NN) and Random Forest (RF) methods gave 99% accuracy but when testing with completely new data the results gave very different answers. The first set of data was tested from July 2020 to August 2020, where the results gave 99% accuracy. But when applying it to the second set of data which was from the end of December to the middle of January the results were completely off, they classified everything as a normal group. This resulted in having to change the main method of analyzing data.

Z.Fan, Y. Zuo, D.Jiang, X.Cai (2015) explain why Random Forest is a popular method in machine learning:

> Random forest classifier is one of the most successful ensembles learning algorithms which has been proven to a very popular and powerful technique in the pattern recognition and machine learning community for high-dimensional classification problems. (p.691)

It is executed with decision trees which are trees that make decisions that are done in for every step that the tree goes further down starting from the root (base) and based on these decisions the tree comes to different solutions. (S. Marsland, 2014: 249-251)

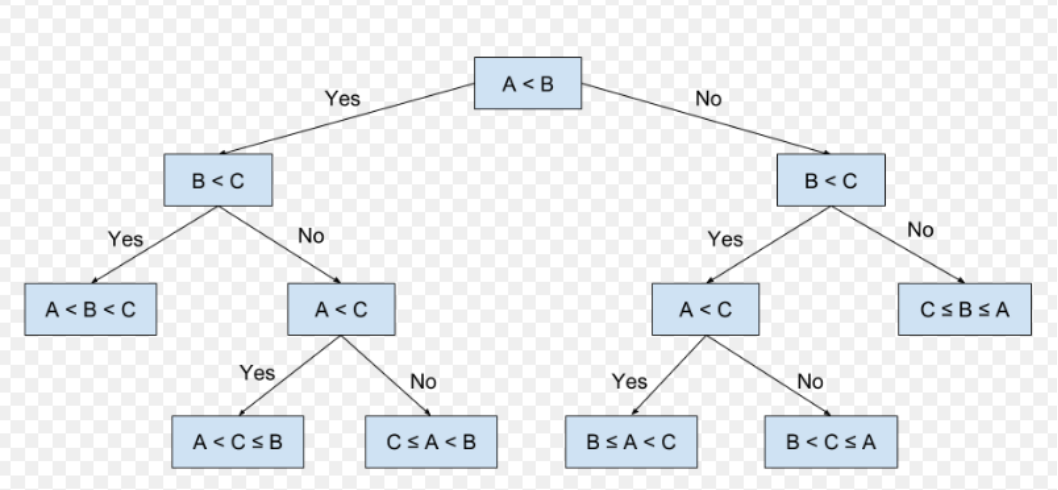

Figure 12. This example shows how a decision tree makes the choices. (O. Niculasescu, 2018)

In Random Forest the idea is simple, using a lot of trees (a forest) should be better than using only one tree. Each tree analyzes features a bit differently, the interesting part is how it chooses to randomize the data. An interesting part is that at each node the there is a random amount of features, and the tree can only pick from those features, meaning each tree and node looks a lot different. (S. Marsland, 2014: 275)

Random forest is a good method that can be used for energy analysis which does not use as much computational power as other methods popular methods. (A. Gloria, J.

Cardoso, P. Sebasliao, 2020) And is one of the methods that were exploited during this thesis, and it was the one that gave the best results with the electrowinning data.

In the roaster and sulphuric acid part, the Neural network method showed the best results. It is the same case as in the electrowinning, so the analyzing methods do not show the same results in new data as in the trained data even though the results are similar.

Faul, A. C., (2019) Explains very simply what a Neural Network is:

> “Neural Network are dynamic system characterized by non-linear, distributed, parallel and local processing”.

A neural network is built on neurons, there are three types of neurons, inputs, hidden, and outputs. All the inputs are input neurons, the hidden neurons mean that they are part of a mathematical function that creates predictions. And output neurons are gathered to produce the result based on the hidden neurons predictions. The mathematical factors are differently weighted, meaning that more important features are weighted higher than the less important ones. (A. C. Faul, 2019:135)

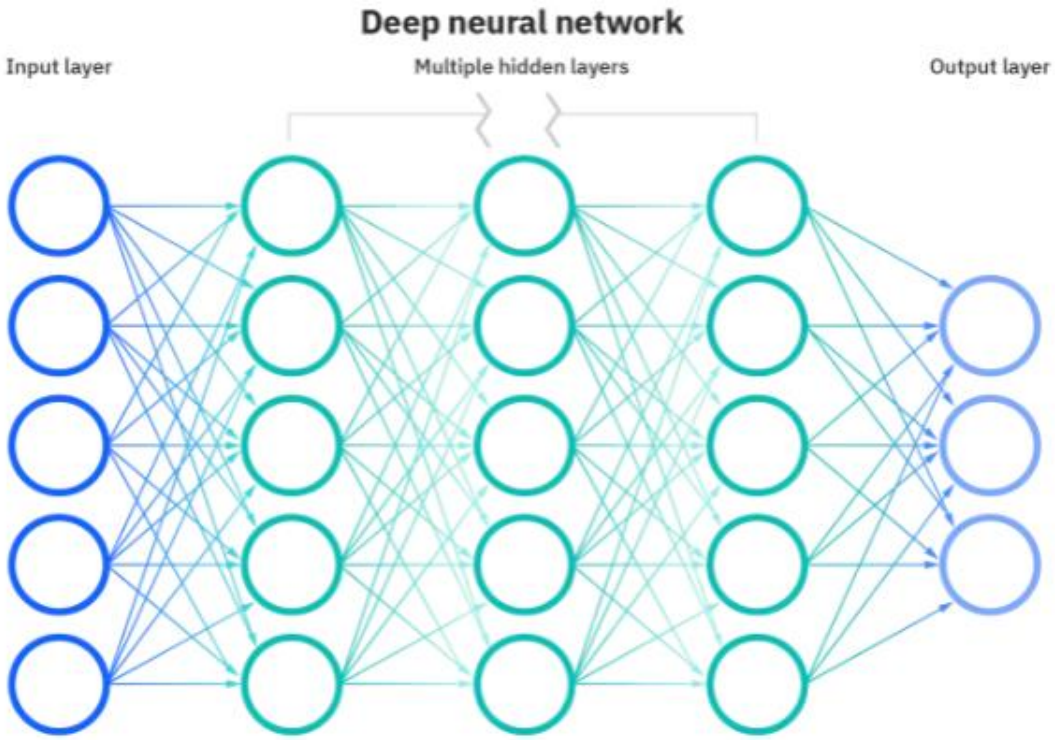

Figure 13. This example shows how a neural network operates. (Neural networks, 2020)

This was the same case in the electrowinning, Neural network has earlier had success in energy analysis and during this thesis, it gave the best result out of all the tested methods and this is the reason it was used. Different methods have to be exploited because the data is different. (O. Olanrewaju, C. Mbohwa, 2017)

3.4 Plotting the data

One variable was added as “test time” which does affect the Random Forest and showed to be an important variable in the results. It was also so that the data could be plotted and the “test time” variable could be used as an X-axis, which is not possible using only the date from the SQL code as it does not register as a date on in Orange.

3.5 Summary

In this chapter it was explained how working offline has affected this thesis and what comes with it. The data preparation and how SQL has been utilized and what had to be done in excel to make it possible to use the datasets in Orange.

Testing is important in machine learning to see what method is the best and in this chapter is has been explained how that works. The method chosen for the first part of the data is Random Forest and Neural Network for the second part. In this chapter, it has also been explained how they work. Also, how the variables have been analyzed during this thesis has been explained.

4 Data background

In this chapter, it will be explained why the data looks the way it does and what parts are supposed to be analyzed. This chapter will also explain why the analyzing methods were chosen for certain parts. In machine learning, it is one of the important parts to have a lot of background data as that is what the system learns from and the more data you have the more accurate the result will be.

4.1 Zinc electrowinning

The zinc electrowinning process is a large energy consumer, and in this thesis, the goal is to correctly classify the events and try to find irregularities. The power stays stable in normal conditions but when cell cleaning takes place the power has to be dropped down to zero and the disconnection of a row pair happens. When the total power drops it goes all the way down to zero then one of the cells in the cell group stays disconnected for X amount of time and is cleaned, during this time the total power is approximately 17% lower. After this, the power drops back to zero and the cell is connected back, and the power rises back to the original state. Some real-life events can happen such as lightning strikes which will result in the power dropping down and these cannot be foreseen or prevented. Also, other events such as maintenance can affect the total power being lower for a while.

The objective of this thesis is to classify these events and to find irregularities of the cell group disconnection.

In this part Random Forest was used, this method gave the best results when the changes in power are very large. The other methods do not give the same results and could not differ the real-life events such as lightning strikes from regular disconnecting events, while Random Forest was able to do that.

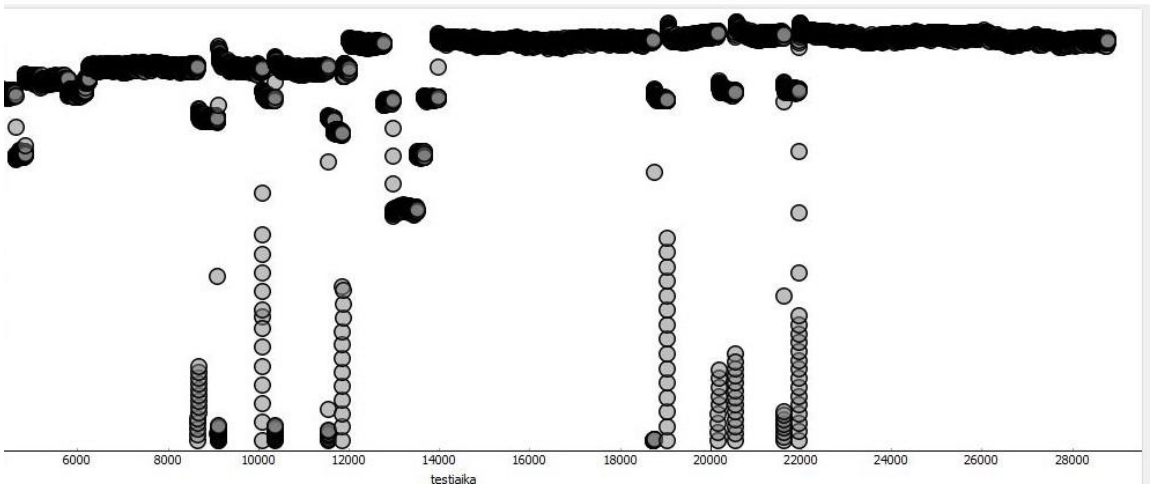

Figure 14. This is picture represents how the data can look and what happens during cell group disconnection.

4.2 Roasting & sulphuric acid plant

The data in the roasting part is a bit different than the zinc electrowinning data, the data is more stable and the changes are in longer periods of time. The data still moves in the roaster and sulphuric acid, there are some dips come in the data and those are when the roaster is being cleaned.

The reason the roasting data and sulphuric acid data are analyzed together is that they are connected through the pipe. The roasting part pushes the exhaust air and the sulphury acid part drags the air out, so if there are issues with the roasting motor the sulphury motor has to compensate and work harder.

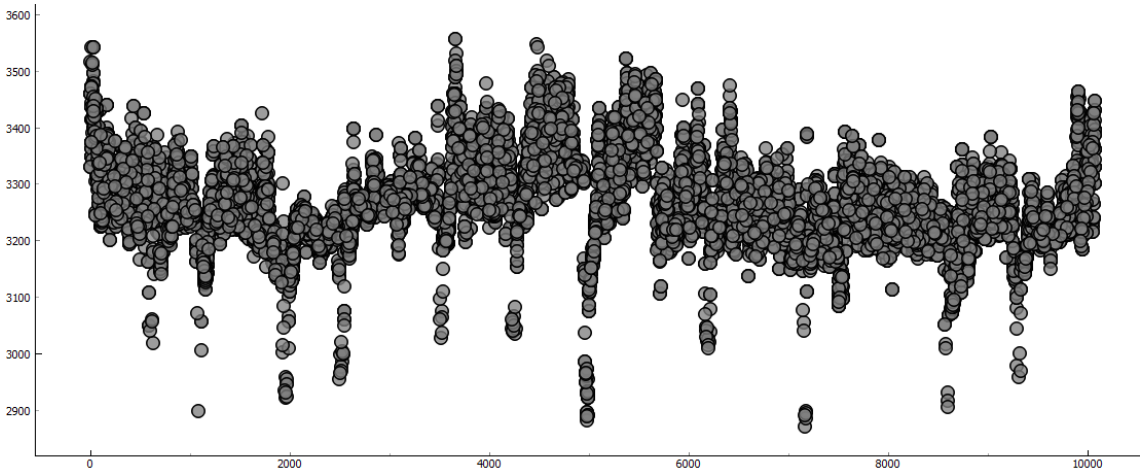

Figure 15. This is picture represents how the total effect acts in the roasting and sulphury acid.

The goal here was to identify the difference between normal behavior and the dips in data. The effect sink in the data is the results of lowering the effect because it has been cleaned during this time.

In this part of the thesis, Neural Network was used. The results were not good at first and the parameters in the neural network had to be changed. This is most likely because the Neural Network was too tightly taught. In the next chapter, it will be explained more in detail.

4.3 Summary

In this chapter, we learned what type of data was used in this thesis and how it behaves. The reasoning should be known so that when analyzing the data, it is known why certain things happen and when to interfere with the learning.

5 Implementing theories

In this chapter, the methods that have been explained earlier will be implemented into practice. What kind of methods have been used and why they have been chosen. Data preparation will explain how the different programs have been used, and what has been done to them. It will also be explained how machine learning has been utilized during this thesis.

5.1 Data preparation

In machine learning data preparation is a key component as the better the data is prepared the better the methods can be taught and utilized in the future. (L. Sanhudo, J. Rodrigues, Ê. Filho, 2021) As is it in this thesis. during this thesis, the data was prepared by using SQL and Microsoft Excel.

The first step in the process was to transfer data to the virtual computer, as the virtual computer does not have direct access to Boliden's servers. The data was fetched on the computer that the virtual computer was stored on, from there it was transferred over to the virtual computer.

Once the data was on the virtual computer SQL coding was used to get the data. Issues that appeared were some values came from every second and a lot of values were missing a value completely since the system took values for every time the value of a signal had changed. This was fixed by taking the exact value for every starting minute.

```
SET RowCount = 0, SamplingMode=calculated, CalculationMode=Average, TimeZone=Server,
DaylightSavingtime=False, starttime='31-aug-2020 06:00:00', endtime='14-sep-2020 06:00:00',
IntervalMilliseconds=1m SELECT timestamp, *Signal tag* FROM ihTrend ORDER BY timestamp
```

Figure 16. This picture shows what the code looks like when fetching data with SQL coding.

In this thesis, the focus is on the power consumption. In the electrowinning it was only one signal for power, this was easier to analyze as it could be done alone and set that as a target. But in the roasting and sulphuric acid, 3 signals that contributed to the total power used, and to be able to analyze them simultaneously they were added together. The signals coming in, were motors for the air flowing through the systems, two of them are parallel and only one of them is always on, and the third signal is for the roaster that drags the air through the sulphuric acid plant.

During the thesis, it was also tested to add a sliding average to recognize the cell cleaning average and to make it easier for the program to understand if it was a correct or an incorrect cell cleaning. The sliding average was done on the power signal, the sliding average did not influence the result.

The date and time that were extracted with SQL into excel could not be read as a time and therefore it was harder to analyze since the X-axis could not be used to show time, to fix this a variable was added as test time. The test time variable was simply to start from 1 and added +1 for every row. This showed to have a large impact according to the ranking widget in Orange. This also allows us to see plots where the X-axis is used as a timeline.

When extracting data with SQL there came additional symbols in the values which were causing issues because excel could not recognize there as integers. To remove them it was simply to copy one of the symbols and remove them via the excel find and replace command.

In the Zinc electrowinning data, the data had to be categorized into the correct categories that were discussed in the last chapter. 2 main categories were focused most on and that was the normal state and the disconnection phase. Although there were more categories as the analysis part was from a long period and during that time there were services that were done and lightning that affected the power in the circuit.

5.2 Machine learning

When the data was ready to be analyzed it was done so with Orange data-mining program. The input data is in a Microsoft Excel file because this is the way that SQL gives the data, Orange can read and it is easy to modify.

5.2.1 Electrowinning

The data were categorized into groups, in the electrowinning, it was set to normal, disconnect/connect, outside event, or power reduction. When the power was at approximately 30 MW it is categorized as normal, then when it drops down to zero and then rises back up to approximately 17% less than its earlier power it classifies it as a disconnect/connect. Power reduction is a class when the power is lower than 30 MW but still functions as normal. During the analyses in this thesis, there were outside events that occurred and that was when the lightning struck, and the power went down to zero.

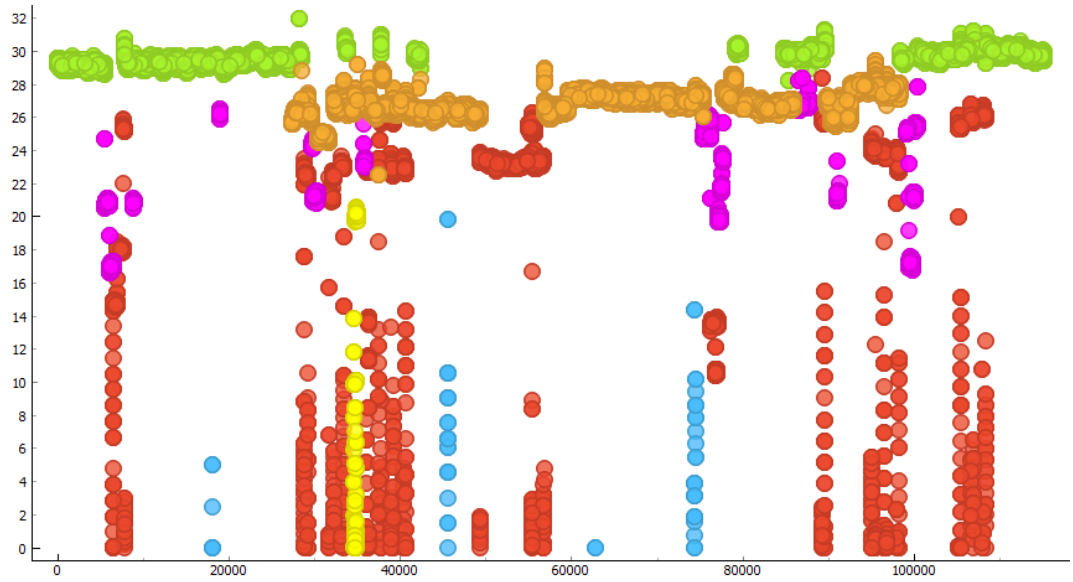

Figure 17. This graph represents how the categorizing works. This is the data that is used to train the set.

All these categories were identified correctly by using Random Forest, other methods gave similar results during the test/train, but when testing it on new data the results were very different. The reason for this is not completely sure but one reason for this might be that the new data was very different for the machine compared to the trained data set.

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Neural Network	1.001	0.999	0.999	0.999	0.999
kNN	1.000	0.999	0.999	0.999	0.999
Random Forest	0.916	0.977	0.973	0.974	0.977
Tree	0.204	0.975	0.971	0.971	0.975
SVM	0.790	0.936	0.929	0.927	0.936
Naive Bayes	0.664	0.875	0.884	0.902	0.875

Figure 18. In this picture, we can see how the results came out in the electrowinning, and even though Random Forest is not highest on the list it gave the best results in new data**.**

The CA is a shortening for classification accuracy. Meaning how accurately the methods can learn and apply the correct class during the train/test phase. During this thesis, the weight was, 80% of the data was used to train the model and 20% was used in testing the model.

During this thesis ranking was exploited, this is a feature in the Orange program. with this feature, it is possible to learn more about how the program understands data. In the electrowinning part, the analyses were as expected. The power was by far the one that explained the classification most. After that, the rest of the signals fell a lot in value.

5.2.2 Roasting and sulphuric acid

The same concept is used analyzing the roasting and sulphuric acid plant as in the electrowinning data. The data that was taken with SQL was put in a Microsoft Excel sheet in which the data was categorized. The focus in this part was different as it does not include as drastic changes in the data as there is in the electrowinning part. The focus was on categorizing normal, cleaning, and error events.

The normal part is categorized as data that was approximately between 3.2 and 3.4, and 3.8 and 4.0 MW. At the first analysis, all the data was somewhere between 3.0 and 3.6 MW, but newer data was between 3.6 and 4.0 MW meaning the data that was taught in the 3.0 to 3.6MW region did not learn how to analyze the data that was valued between 3.6 and 4.0 MW. The solution that was used to add both these data groups together and re-train them. After this, the data did not categorize everything in the 3.6 to 4.0 MW region wrong. The reason for the difference in 3.0 to 3.6 and 3.6 to 4.0 is that the feed into the roaster was less during the first part of the data and then it was higher during the second part of the data. This led to that the new data was harder to read at first but once the NN learned it correctly the results were better.

The error category was data that stood out and was not acting as it should, for instance, a part of the data was significantly higher than it should be.

Cleaning events are parts of the data that fall lower than the normal state. The cleaning means that the pipes run on lower effects and are cleaned during that.

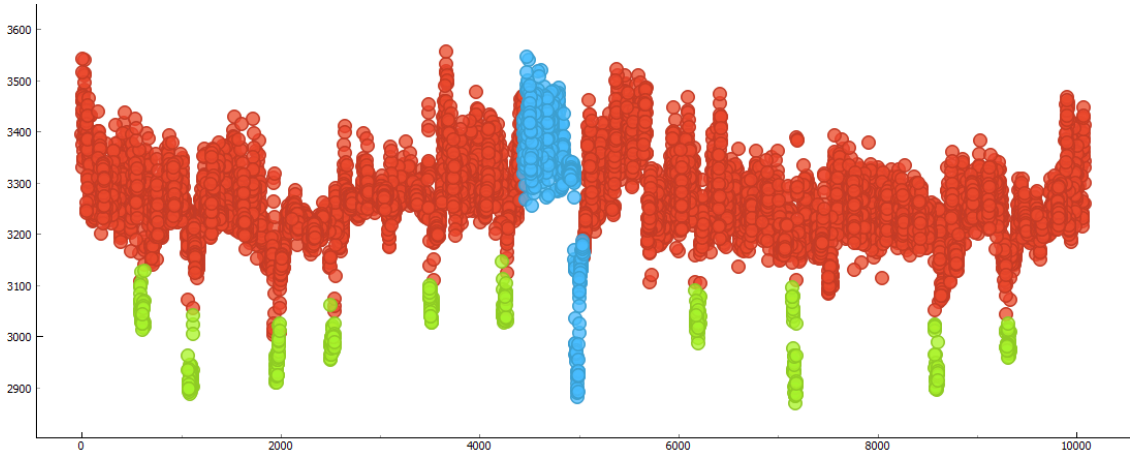

Figure 19. This picture shows the different categories of data.

In the picture above we can see part of the data that was set in to train the data set. The green parts are categorized as cleaning, the blue as error and, the red as the normal state.

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
kNN	0.967	0.978	0.977	0.977	0.978
Neural Network	0.975	0.968	0.964	0.965	0.968
Random Forest	0.953	0.965	0.960	0.965	0.965
Tree	0.500	0.918	0.878	0.842	0.918
SVM	0.554	0.761	0.806	0.879	0.761
Naive Bayes	0.884	0.700	0.772	0.913	0.700

Figure 20. This picture shows how effective the different types were with analyzing roaster and sulphuric acid data.

This was the result that the ranking after changing the amount neurons used in hidden layers, regularizations, and the maximum number of iterations to get the best possible result out of the NN used in this thesis. At first, the NN placed higher than kNN but after giving it more flexibility the test/train results lowered but the final result improved.

Before the changes were made the NN classified everything as normal but after giving it more flexibility it classified the cleaning and the normal state correctly.

In the roasting and sulphuric acid data, the ranking page came in as a bit of a surprise, it seems to be that the program does not value the power from the engines as high as expected. Rather it considers the angel of the airflow vent more important. This shows how the computer assumes things differently than the human since a human would value the motor speed higher than the angle of the airflow vent.

5.3 Summary

In this chapter, we see what happens to the data throughout the process. How we extract the data with SQL. Then use it in a Microsoft Excel file, and what can be done to make it readable with Orange and how to make visualize it better for us humans. It is also explaining how the data behaves and how it should be approached.

6 Results

This chapter it is evaluated how accurately the machine learning methods are and how the data preparation came out. The goal of these results was that in the electrowinning part to learn to identify the difference between the different classes, and to simulate failed connection/disconnections, and make the Random Forest recognize the difference between failed and successful connections/disconnections. The goal for the roaster and sulphuric acid data was to train and test the difference between normal data, cleaning events, and some error events.

6.1 Data

To understand the data, it is important to understand what type of data was used in the analyses during this thesis. The data that was used during this thesis varied on which part of the process was currently analyzed.

6.1.1 Extracted data.

The data that was extracted was set into a Microsoft Excel sheet. The SQL could not extract the full amount of data, so it had to been done multiple times to get enough data to teach the Random Forest and Neural Network correctly. In this part, there are 39 signals that the Random Forest analyses.

12.06.2020 06:04:00	Norm	29.39999962	39.99375153	735.6	106	104	106	102.9285714	103	104	103
12.06.2020 06:05:00	Norm	29.35000038	39.91957474	735.8	106	104	105.75	103	103	104	103
12.06.2020 06:06:00	Norm	29.30000114	40.00911331	736	106	104	105.5	102	103	104	103
12.06.2020 06:07:00	Norm	29.39999962	39.98814774	735	106	104	105.25	102.1	103	104	103

Figure 21. This picture shows what type of data was used in the electrowinning.

In the electrowinning data set, there were thirteen factors that contributed to the training and testing of the data set. This also includes the moving average which was set to be.

$$X = \frac{i - 100\,; i + 100}{200}$$

Here X is the value of the moving average and i is set to be the row. This is set to take the average value of the last 100 and coming the 100 values. This was set to improve the analyzing and according to the ranking page this had a major impact but when testing the difference with new data the results were very similar.

The data in the Roaster and sulphuric acid looks quite different as the data does not have as big changes in it as it does in the electrowinning. In this part there were

31.08.2020 06:03:00	Norm	3331	22370.01	10.71946335	25.6145401	41291.51172	53.35698318	99.98664856	100.3013649	2.680541992	0	0.291193008
31.08.2020 06:04:00	Norm	3516	24385.74	10.75761032	25.39536095	40930.30469	52.14389229	100.0038147	100.2441406	5.4046731	0	0.278477162
31.08.2020 06:05:00	Norm	3436	24886.66	10.87205315	25.25584412	40159.28125	50.93080139	100.2326965	100.2822952	1.250743866	0	0.316624701
31.08.2020 06:06:00	Norm	3460	23933.08	11.02464294	25.6893177	40193.52344	51.23598099	99.93133545	100.2441444	2.414399385	0	0.278477162
31.08.2020 06:07:00	Norm	3375	22110.33	10.94834805	25.58368111	40245.26563	51.35042191	100.0801086	100.2059937	5.587149143	0	0.288014047
31.08.2020 06:08:00	Norm	3543	23654.61	10.96742201	25.5005722	41257.98438	53.73846054	100.0038147	100.2822952	8.08864E-05	0	0.297550932
31.08.2020 06:09:00	Norm	3444	22438.63	10.98649597	25.32143211	41206.08203	53.62401581	100.0801086	100.2441406	0.127611995	0	0.307087816
31.08.2020 06:10:00	Norm	3473	24296.43	10.94834805	25.45516205	40317.94922	51.23598099	100.067393	100.2568588	13.79796505	0	0.316624701

Figure 22. This picture shows some of the roaster & sulphuric acid data.

6.2 Machine learning predictions

The data that was put into the training set can be seen in Figure 17. In the following picture, we can see the result of the Random Forest method. Boliden's requirement was the success rate would be above 90%. This is however only the results for the test and train data and the 90% result had to be done on new data to see if the model works. In the train set the Random Forest correctly classified approximately 250 000 and misclassified approximately 5000. The reason for misclassifications is unknown.

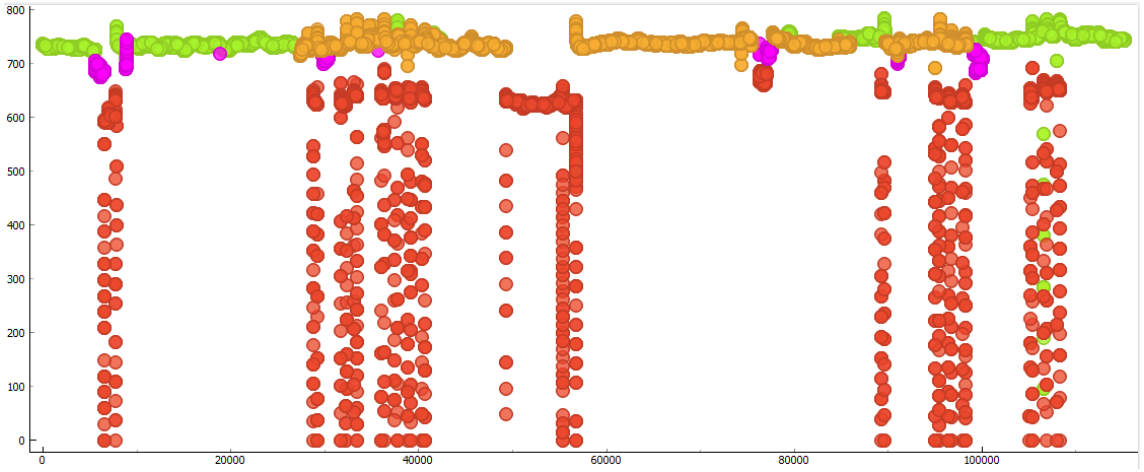

Figure 23. This picture shows the correctly classified events in electrowinning with Random Forest.

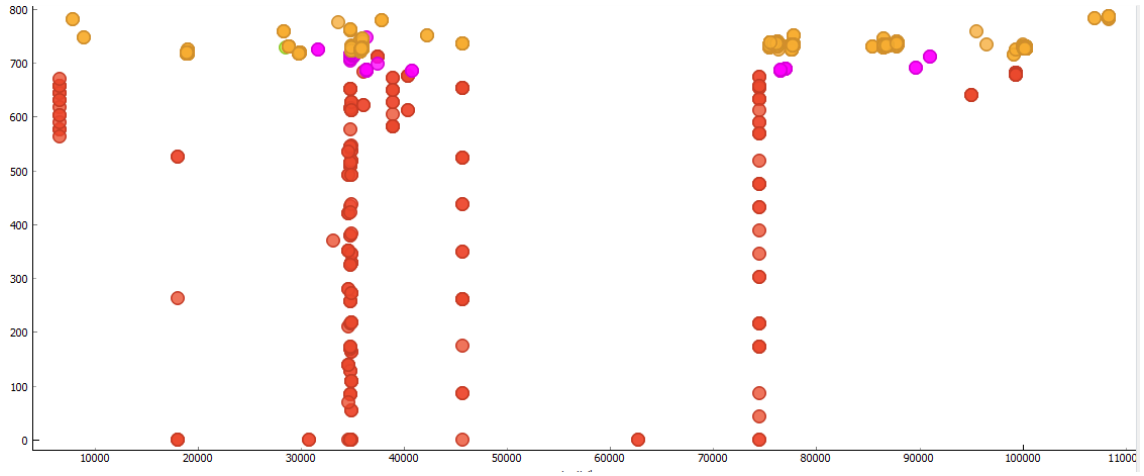

Figure 24. This picture shows the misclassified events in electrowinning with Random Forest.

This was the result of Random forest when it gave the best result on new data. This could be even higher but then the method would be too tightly trained in this data and would not be able to handle another set of data as well. During this thesis, it was also tested on separating the connection/disconnection class into two different classes but that caused more issues than it solved.

Similar procedures were done with the roaster and sulphuric acid data. In this data, 70 signals were contributing to the data set. Although three of them were combined for the

total power used. In training and testing the roaster and sulphuric acid it predicted approximately 41 400 correctly and 400 wrong. This was when the Neural Network values had been changed to give the best possible outcome for new data.

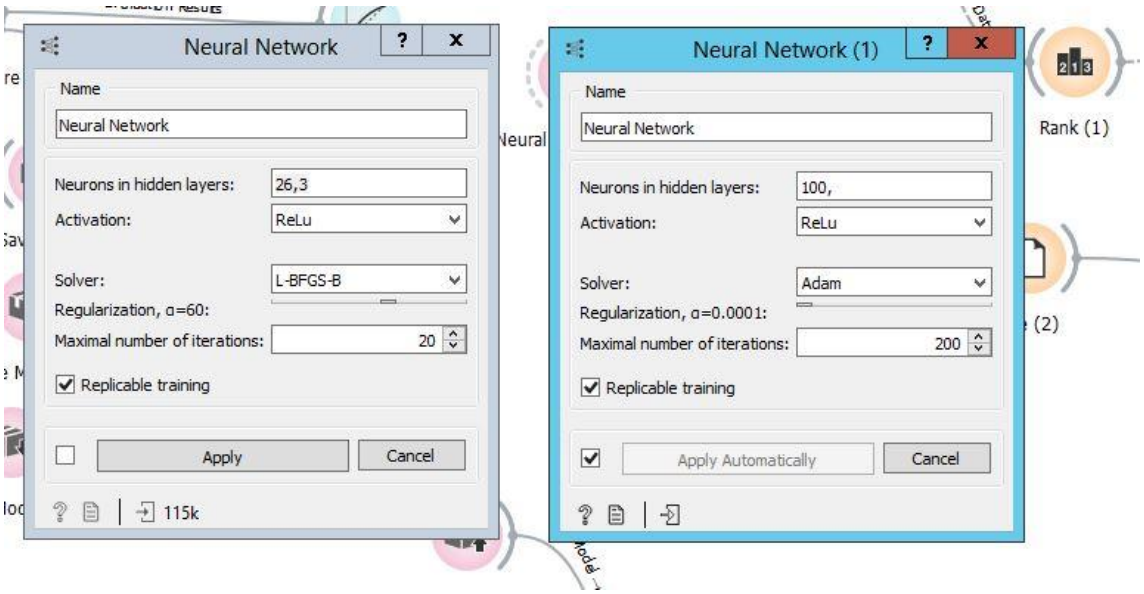

Figure 25. This shows how the values in the neural network were changed. The one on the left side is changed while the one on the right is the original values that orange sets as a default.

When these methods were trained and tested, they were saved as models into Orange. Then these models were loaded and ran through the new data.

6.3 Loading methods to new data

With the new models loaded into the Orange system, they were tested on the new data. this was not done in this order since the models were modified many times before they were predicted the data in the best possible way.

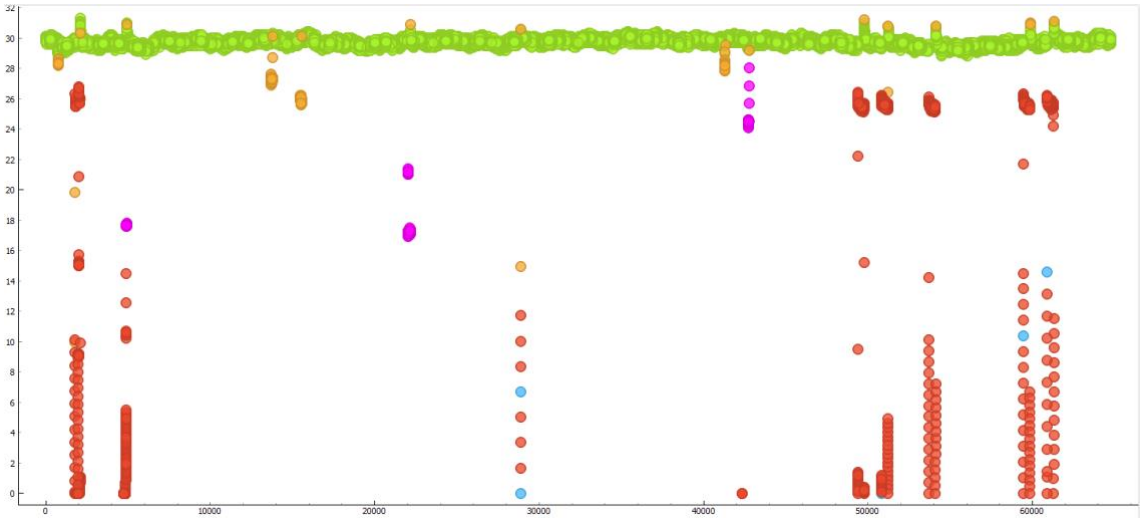

Figure 26. This is the graph that is the result with the trained Random Forest method and new data from the electrowinning.

In this picture, we can see that the model is acting very much as a human would classify these events. The classification of the data electrowinning data was approximately 99.8% accurate. 0.2% is still misclassified and that is unclear why, this does not matter too much as it quite easy to see where it has gone wrong. Some events are classified as real-life events which should not be there, but they fall under the requirement that the result should be above 90%. This model was very successful. Other models were also tested during this thesis to get the best possible results. The methods that were used were Neural Network, Random Forest, Support-Vector Machine, Naïve, Bayes, and K-Nearest Neighbor. As they have shown good results in energy analysis. (D. Liu, Q. Yang, F. Yang, 2019) (M. Goyal, M. Pandey, R. Thakur, 2020)

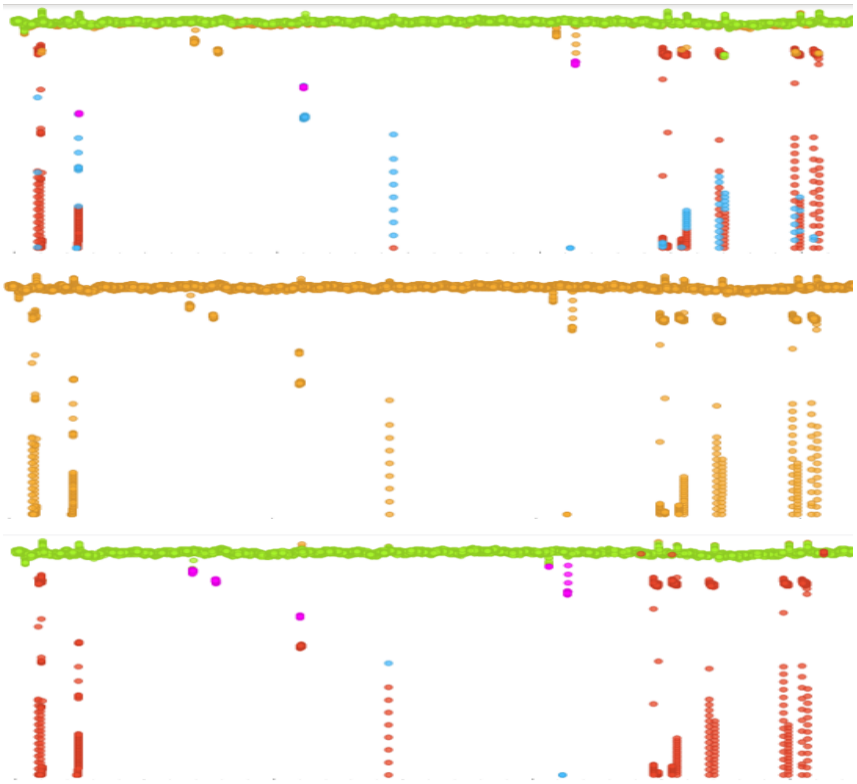

Figure 27. This picture shows the results of some of the other methods. The first one is Neural Network, second the k-Nearest neighbor, and lastly Support-Vector Machine.

In the picture above we can see that the Support-Vector Machine method also performed well but when comparing the correctly classified and misclassified classes Random Forest gave a better result.

During this thesis, the failed connect/disconnect events were simulated. These events were such that the ones the power rises back up it does so but not completely. This would mean that the power reduction would be 25% rather than 17%. However, these simulations were not successful. This is because during this thesis only the total power was changed and not the other signals. The reason the other signals were not changed was that there had been no such events and it would be very complex to calculate what the other values would be, during such a failed connect/disconnect event.

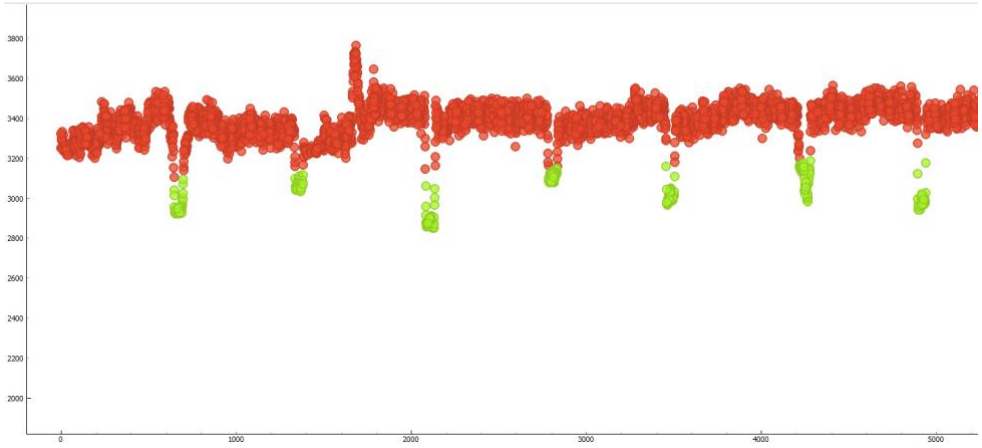

Figure 28. This is the graph that is the result of the trained Neural network method and new data from the roaster and sulphuric acid plant.

This graph shows How the neural network classified the roaster and sulphuric acid plant data. The classification was approximately 99,4% accurate here. The remaining 0.6% is unclear how to improve, during this thesis there was a lot of testing changing these parameters, and trying different methods to improve the result but as this result was the best and it is better than the 90% requirement this was accepted. Other methods that were tested during this thesis were. The methods that were tested were Neural network, Random Forest, Support-vector machine, Naive Bayes, K-nearest neighbor, and Normal Tree analysis.

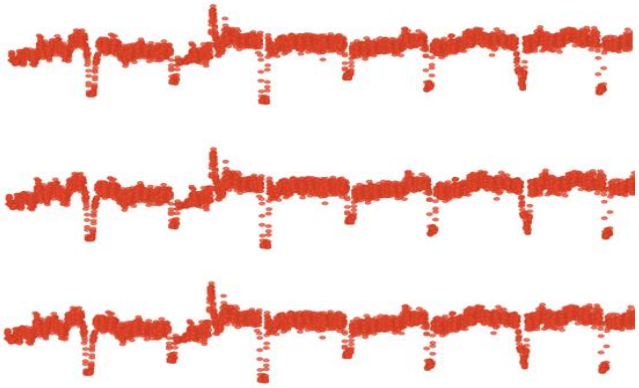

Figure 29. This picture shows the results of some of the other methods. The First one is Random Forest, the second k-Nearest neighbor, and lastly Support-Vector Machine.

As we can see the Neural Network gave by far the best result.

It is simpler than the electrowinning data. In this data part, it is easy to see where the cleaning has happened and what should be classified where. In this part of the data there were no simulations, and this result is a success.

6.4 Summary

In this chapter, the results of each part of this thesis have been explained. Overall, this was a success except that the simulations were not successful as they did not contain sufficient amounts of data. When starting to extract the data there were some issues at first with learning the signal tags and getting the correct SQL code. The data itself required a lot of work before it was able to be processed by Orange. The process was slowed down by having to work remotely and all the questions had to be asked via email or meetings. Orange was a huge advantage in this thesis as it allowed to test quicker than having to code everything manually to test different methods. The electrowinning data could have been optimized a little bit better but as there was more to this thesis than only that part of data this result was good, to train and test the earlier data was quite slow. The roaster and sulphuric acid data were very successful but also easier to analyze. There was no error class in the newer data part, but according to the test/train, it would have been able to detect it. And simulating error events is not something that would have been very successful as all the 70 signals would have had to been taken into consideration.

7 Conclusion and discussion

In this chapter, we will discuss what advantages there are to using machine learning in energy analyses, and what the results can be used in. The results will be reviewed and how they could be improved.

In this thesis, the energy usage at Boliden Kokkola was analyzed with the help of machine learning methods. The data that was used in this thesis was not prepared in advance so it was not the best data to be analyzed but this also means the data is from everyday use so with that in mind it is good to have normal data instead of specifically chosen data to teach from because that could cause inconsistencies when testing it with newer data. The process to teach and to test was quite slow as it was done in Orange and could be further optimized by coding it by hand in Python. The process could also be speeded up if some signals were removed but that could result in worse results.

The results that were initially set for this thesis were to find error events and make the machine learning agent notice them and warn about them in advance. This was however unfortunately not achieved as there were not enough past events to teach from. Then we simulated those events but as we do not know how the events would act in a real situation, they were unsuccessful. However, the normal events and even lightning strikes were successfully detected and taught. So, the results of this thesis can be applied when searching for certain errors or classifying future events.

For future topics, this could be improved by having more and better background data to teach the machine learning agent from. The extraction of data could also be improved as it takes a lot of time extracting it piece by piece since SQL is limited how much data it can extract at a time and this limit gets filled quite quickly when 70 signals were being extracted.

References

A. Faul (2019) A Concise Introduction to Machine Learning. Fetched from https://www-taylorfrancis-com.proxy.uwasa.fi/books/mono/10.1201/9781351204750/concise-introduction-machine-learning-faul

A. Gloria, J. Cardoso, P. Sebasliao (2020) Improve energy efficiency of irrigation systems using smartgrid and random forest. Fetched from: https://ieeexplore-ieee-org.proxy.uwasa.fi/document/9221776

A. Mechelli and S. Vieira (2019) Machine learning: Methods and applications to brain disorders. Fetched from https://ebookcentral-proquest-com.proxy.uwasa.fi/lib/tritonia-ebooks/detail.action?docID=5979314

A. Taylor (2013) SQL for Dummies. Fetched from https://ebookcentral-proquest-com.proxy.uwasa.fi/lib/tritonia-ebooks/detail.action?docID=1335264

Boliden (2018) https://www.boliden.com/operations/smelters/boliden-kokkola

CISCO How does a virtual private network (VPN) (2020) work?
https://www.cisco.com/c/en_uk/products/security/vpn-endpoint-security-clients/what-is-vpn.html

D. Darlis, M. Latip, N. Zaini, H. Norhazman, (2020) Random Forest approach for energy consumption behavior analysis. Fetched from https://ieeexplore-ieee-org.proxy.uwasa.fi/document/9188072

D. Liu, Q. Yang, F. Yang, (2019) Forecasting Energy Consumption of Office Building by Time Series Analysis Methods based on Machine Learning Algorithm. Fetched from https://ieeexplore-ieee-org.proxy.uwasa.fi/document/9107816

D. Narciso, F. Martins (2020) Application of machine learning tools for energy efficiency in industry: A review. Fetched from https://www.sciencedirect.com/science/article/pii/S2352484719308686

G. Akhil, I. Satyakumar, K. Sarath, M.Rahul, P. Warrier (2017) Controlling the motor direction by tilting of head and using KNN classifier to identify the pattern. Fetched from https://ieeexplore-ieee-org.proxy.uwasa.fi/document/8089169

Ge digital solutions, 2021 https://www.ge.com/digital/documentation/historian/version72/c_historian_interactive_sql_app.html

J. Mueller, L. Massaron (2016) Machine learning for dummies. Fetched from https://ebookcentral-proquest-com.proxy.uwasa.fi/lib/tritonia-ebooks/detail.action?docID=4526803

K. Murphy (2012) Machine learning: A probabilistic perspective. Fetched from https://ebookcentral-proquest-com.proxy.uwasa.fi/lib/tritonia-ebooks/detail.action?docID=3339490

K. Palola (2020) Teollisen sinkkielektrolyysinprosessin Sähköinen Vastinkytkentä

L. Sanhudo, J. Rodrigues, Ê. Filho, (2021) Multivariate time series clustering and forecasting for building energy analysis: Application to weather data quality control. Fetched from: https://www.sciencedirect.com/science/article/abs/pii/S2352710220336287

M. Goyal, M. Pandey, R. Thakur (2020) Exploratory Analysis of Machine Learning Techniques of predict Energy Efficiency in Buildings. Fetched from https://ieeexplore-ieee-org.proxy.uwasa.fi/document/9197976

Microsoft Remote desktop clients (2020) https://docs.microsoft.com/en-us/windows-server/remote/remote-desktop-services/clients/remote-desktop-clients

M. Sugiyama and M. Kawanabe (2012) Introduction to machine learning. Fetched from https://ebookcentral-proquest-com.proxy.uwasa.fi/lib/tritonia-ebooks/detail.action?docID=4107705

Neural Networks (2020) IBM Cloud education. Fetched from https://www.ibm.com/cloud/learn/neural-networks

O. Niculaescu (2018) Fetched from https://elf11.github.io/2018/07/01/python-decision-trees-acm.html

O. Olanrewaju, C. Mbohwa (2017) Assessing the possible potential in the global energy consumption: Integrated artificial neural network and data envelopment analysis. Fetched from: https://ieeexplore-ieee-org.proxy.uwasa.fi/document/8290152

O. S. Chapelle, O. Chapelle, B. Schölkopf & A. Zien (2006). Semi-Supervised Learning. Fetched from: https://books.google.fi/books?id=A3ISEAAAQBAJ&printsec=frontcover&dq=Semi-Supervised+Learning.+chapelle&hl=sv&sa=X&redir_esc=y#v=onepage&q=Semi-Supervised%20Learning.%20chapelle&f=false

Orange - Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: Data Mining Toolbox in Python, Journal of Machine Learning Research 14(Aug): 2349−2353.

S. Marsland (2014) Machine learning: An Algorithmic Perspective, Second Edition Fetched from: https://books.google.fi/books?hl=sv&lr=&id=y_oYCwAAQBAJ&oi=fnd&pg=PP1&dq=Machine+learning:+An+Algorithmic+Perspective,+Second+Edition&ots=-yfYPlAl5P&sig=5WuL2BRTI-5YUhkOuoKLEQXSI5c&redir_esc=y#v=onepage&q=Machine%20learning%3A%20An%20Algorithmic%20Perspective%2C%20Second%20Edition&f=false

Z. Fan, Y. Zuo, D. Jiang, X. Cai (2015) Prediction of acute hypotensive episodes using random forest based on genetic programming, 2015 IEEE Congress on Evolutionary Computation (CEC)