Development of a Software Application for the Prediction of Froth Flotation Reagents Dosage

By

TOLUWANI AJAYI

(Q15764338)

COM726

DISSERTATION

SOLENT UNIVERSITY

Faculty of Business, Law and Digital Technology

09/09/2022

Table of Contents

CHAPTER 1: INTRODUCTION 5

1.1 Background 5

1.2 The Impact of the Metallurgy Industry 5

1.3 Froth Flotation 6

1.4 Reagents 7

1.2 Problem Statement 8

1.4 Aim and Objectives 8

1.5 Project Specifications 8

1.6 Project Plan 9

CHAPTER 2: LITERATURE REVIEW 11

2.1 Reagents in Froth Flotation 11

2.2 Research Related to Reagent Dosage 12

2.3 Artificial Intelligence in Froth Flotation 12

2.4 Reagent Dosage Prediction 13

2.5 Justification for the Proposed Technique 15

2.5.1 *p*-Values for Variable Selection 16

2.5.2 Random Forest Regressor 16

2.5.3 GUI Programming in Python 18

CHAPTER 3: METHODOLOGY 19

3.1 Research Method 19

3.2 Data Collection 20

3.3 Data Pre-processing 21

3.4 Exploratory Data Analysis 21

3.5 Model Development 22

3.6 Performance Metrics 22

3.7 GUI Design 22

3.8 Research Implementation 23

CHAPTER 4: RESULTS 24

4.1 Data Collection 24

4.2 Data Pre-processing 24

4.3 Exploratory Data Analysis 25

4.4 Model Development 32

4.5 Graphical User Interface 33

CHAPTER 5: DISCUSSION 34

5.1 Key Findings 34

5.2 Recommended Implementation of the Product 34

5.3 Benefits of the Software 34

5.4 Limitations 35

CHAPTER 6: CONCLUSION 36

Reflection 36

References 37

Table of Figures

Figure 1: Froth Flotation (Michaud 2021). 7
Figure 2: Top 5 rows of the dataset 24
Figure 3: Output of Pre-processing 25
Figure 4: p-values 26
Figure 5: % Iron Feed 27
Figure 6: % Silica Feed 27
Figure 7: Ore Pulp pH 28
Figure 8: Ore Pulp Density 28
Figure 9: Flotation Column 01 Air Flow 29
Figure 10: Flotation Column 04 Air Flow 29
Figure 11: Flotation Column 05 Air Flow 30
Figure 12: Flotation Column 06 Level 30
Figure 13: Flotation Column 07 Level 31
Figure 14: Correlation Matrix 31
Figure 15: Random Forest Regressor 32
Figure 16: Model Testing 32
Figure 17: GUI 33

CHAPTER 1: INTRODUCTION

1.1 Background

The rise of artificial intelligence and data science applications have led to unprecedented innovations in multiple industries and these innovations have delivered increased productivity, effectiveness, and efficiency (Chen et al 2012). Data science is the retrieval of actionable information directly from gathered data through discovery or hypothesis formulation and testing (Pritzker and May 2015). Over the last decade, there has been a phenomenal increase in the amount of data generated and collated by individuals and organizations for statutory and analytical purposes, this has in turn resulted in massive growth in the applications of the data science discipline (Lamb 2015). Fundamentally, data science aims to help the business, government, healthcare, manufacturing, and other sectors make higher quality decisions and develop more effective strategies by processing data and translating it into intelligence, knowledge, and wisdom (Provost and Fawcett 2013).

Data science and artificial intelligence contribute to strategy design and decision making because they allow predictions to be made from previous data. McCarthy (2004) defined artificial intelligence as the science and engineering of making smart machines and intelligent computer programs. According to IBM (2021), it is a field that adopts a combination of large datasets and computer science to deliver solutions to complex problems. Deductively, as technologies develop and work together, more possibilities will arise to be explored. Hence, artificial intelligence and data science have paved the way for useful technologies such as machine learning and deep learning that are used for predictions.

1.2 The Impact of the Metallurgy Industry

Among the various sectors that benefit from artificial intelligence technologies, one that stands out is the manufacturing and especially the metallurgy industry. This is because the industrial economy of the whole world depends heavily on the effectiveness of producing materials needed for daily life activities. Hence, it is imperative that processes involved in metallurgy are optimised as much as possible to better serve the needs of the world. Metallurgy can be described as a process that involves the retrieval of metals in their natural state from their corresponding ores on a large scale (Quintanilla 2021). According to Ma (2012), despite the exhaustion of the world's reserve of high-grade iron ores, the demand for iron ore continues to rise. Kitchener (1984) further explained that the production of low-grade iron ores play an integral role in the world's economy because high grade ores are processed through very

expensive methods so mass production of metals and inorganic raw materials would be less accessible without a process known as froth flotation. This flotation technique which is one of many various metallurgical processes has been highlighted over many decades as one of the most effective means of separating impurities from iron ores. It presents a low-cost, yet significant method used globally to get metal concentrates from low grade ores by expelling impurities. The technique essentially separates particles within the ore by taking advantages of their surface selectivity. It introduces reagents to isolate hydrophobic particles (particles that are water repellent) from hydrophilic ones (particles that mix with water) and depending on what is to be achieved, the reagents can be adjusted to influence surface selectivity (Mondal et al 2021).

1.3 Froth Flotation

Froth flotation is a process where air bubbles are inputted on a blend of well crushed ores into a system containing water and chemicals. These ingredients facilitate the deposition of the bubbles onto the particles of the required metal and produces its recovery as a froth (Merriam-Webster 2022). This process is the main tonnage separation technique in metallurgy through which desired minerals are retrieved from waste rock deposits (Quintanilla 2021). It is accepted as one of the finest ways of managing fine metals because it takes advantage of the physical and chemical attributes of the materials to distinctly separate them by attaching the required mineral particles on the floated bubbles and augmenting them in a froth state (Cheng et al 2022). The flotation mechanism is generally made up of the three main forms of matter: solid, liquid and air. The ore pulp consisting of solid particles and water is introduced into a flotation device where air and the appropriate reagents can act on it. The composition then causes the surfaces of the water-resistant material to mix with gaseous bubbles so they can be deposited on to the hydrophobic particles. Finally, the resultant mixture gets fed into the froth state at the top of the flotation device. The froth which consequently contains a mixture of bubbles and hydrophobic particles gets separated from the residual pulp consisting of the hydrophilic materials which are then appropriately discarded (Araujo and Peres 1995).

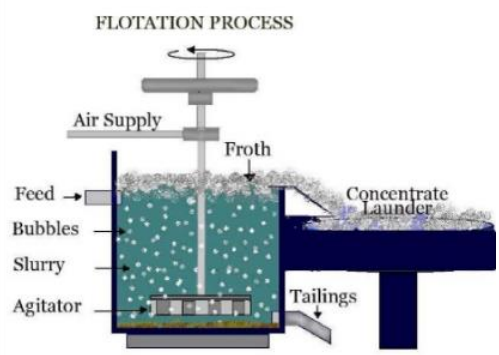

Figure 1: Froth Flotation (Michaud 2021).

1.4 Reagents

This project focuses on the significance of reagents to the flotation process. Froth flotation reagents are generally of two kinds: depressants and collectors. Borregaard (2022) described depressants as chemicals that when introduced into the flotation system, create and manage hydrophilicity of the desired materials within the chamber so they reside within the pulp rather than buoying up with the expected froth. They use dispersion to distinctly depress substances like pyrite, clay and talc that could be present within the pulp by adsorbing onto them and not the desired mineral. They also inadvertently boost the effectiveness of collectors due to the fact that they improve the quality of the recovery of the target material. In contrast, collectors were described as substances that act to increase the target mineral's resistance to water by causing them to mix with the bubbles and float upwards to be collected appropriately.

In the research carried out by Erol et al (2003), it was found that reagents among other elements in a froth flotation system, play a very significant role in determining the purity of the retrieved mineral concentrate. An inordinate amount of reagents can negatively alter the quality of the flotation process in ways such as overly thick grinding fineness, falling slurry level, reflectivity of the foam surface and larger foams in the flotation tank among others so it is imperative for the flotation operators to ascertain the appropriate dosage of the reagents to be used in each iteration (Xinhai, 2022). In addition, Gharai and Venugopal (2018) executed laboratory experiments aimed at exploring the effects of the reagent dosage on the hydrophobic and hydrophilic attributes of ore. It was overcovered that the performance of the flotation system was greatly affected the dosage of the reagents hence it was advised that the measurements need to be as ideal as possible.

1.2 Problem Statement

The froth flotation technique is very essential to the economy of the world. Therefore, research is constantly needed to explore means of optimising it. A flotation process plan that lacks the appropriate dosage of reagents can lead to wastage of resources, compromised quality of the target material and financial loses. Hence, there is a need for means of predicting reagent dosages as accurately as possible.

1.3 Research Questions

- How does the numerical data gathered from previous iron ore flotation operations perform when used to predict the ideal amount of reagents needed for future ore processing?
- Does a random forest regressor algorithm deliver accurate predictions for this project with an R2 score >0.95?
- What effect would the development of the proposed application have on the metallurgy industry?

1.4 Aim and Objectives

The aim of this project is to develop a software application that predicts the required dosage of depressant and collector reagents needed to process a given volume of iron and silica feed into a desired concentrate within a period of time.

The objectives are:

- To research relevant previous work to determine the feasibility of the proposed solution
- To obtain high quality data, pre-process it and achieve understanding of it
- To research and conclude on the random forest regressor as a machine learning model that will accurately predict the target variables
- To ensure the model performs as required by the pre-stated success and evaluation metrics
- To develop a graphical user interface for the model.

1.5 Project Specifications

The proposed resource for this project is a software application made up of a predictive algorithm. It will be developed using the python language in the Jupyter notebook IDE. The application would require a large dataset made up of multiple variables associated with the froth flotation process such as date/time, percentage feeds, reagent flow, ore pulp flow, pulp

pH, density, flotation columns and the resulting concentrates. The application will be designed to take in a set of inputted variables needed for the prediction so that when prompted, it will deliver the corresponding predictions.

To achieve a software artefact capable of predicting the dosage of reagents that would be needed to process a given volume of ore feed, the random forest regressor would be adopted. The data gathered specifically describes the flotation process of iron ore and silica ore simultaneously and previous work on this dataset was aimed at predicting the resultant percentage concentrate from the system. In this project however, the target variable would be the Starch and Amina flow (reagents) where the p-value technique of feature selection is implemented to realise the most important variables needed to achieve the proposed solution. Also, within this project, exploratory data analysis will be carried out on the data to achieve an understanding of how the variables relate to each other. Finally, the developed model will be linked to a graphical user interface window that prospective users can use to take advantage of the proposed solution within the industry.

The project implements multiple python libraries among which are tkinter, numpy, pandas, sklearn, matplotlib, statsmodels and pickle. These libraries and other tools utilised for the development of this project are summarized in the following chapters.

1.6 Project Plan

The proposed plan for the execution of this project is summarized in the table below.

Tasks	June – July				July - September							
	Wk1	Wk2	Wk3	Wk4	Wk5	Wk6	Wk7	Wk8	Wk9	Wk10	Wk11	Wk12
Background studies												
Formulate research questions												
Search for quality data												
Write up pilot study												
Get ethics approval												
Carry out literature review												
Outline research methodologies												
Pre-process data												
Carry out data understanding												
Implement identified model(s)												
Test and conclude on the results												
Design the GUI												
Results and discussions												
Conclusion												
Submission												
Viva												

CHAPTER 2: LITERATURE REVIEW

This chapter further examines the effects of reagents and their dosages on the performance of a froth flotation system. It features previous research and developments related to improving the flotation process through reagent control. In this portion of the project, past attempts by researchers to adopt artificial intelligence in froth flotation in order to justify the selection of the planned methods are also summarized. Finally, an overview of the proposed prediction technique in literature is provided.

2.1 Reagents in Froth Flotation

In froth flotation, the choice of the reagents used is very vital in determining the recovery and purity of the final product derived from the process. The dosage of the selected reagents, the nature and size of the ore particles to be separated as well as the adsorption of the reagent chemicals on said particle surfaces all contribute immensely to the quality of the concentrate produced. This was concluded from the work carried out by Erol et al (2003) where tests were conducted on the froth flotation performance vis-à-vis the reagents. The results showed that a strong relationship exists between the realised concentrate and the type of the reagents used as well as the quantity. Urbina (2010) further expatiated on the developments in novel mixtures and compositions of both depressant and collector reagents as a mean of improving the performance of froth flotation. These developments have arisen due to the innovative applications of the flotation process which further emphasizes the significance of the reagents in the industry. Furthermore, an in-depth review on the effects of reagents on specific ores like phosphate showed that there were distinct benefits to be accrued from utilising a mixture of surfactant mixtures over single surfactants. The advantages include surface tension, contact angle, adsorption among others (Sis and Chander 2003). In a bid to further illustrate the significance of reagents in the froth flotation technique, experiments were carried out by Vamvuka and Agridiotis (2001) where they compared the results obtained from the system when infused with kerosene with the results gotten when reagents were introduced. It was found that the overall performance of the system was better and produced less product ash levels. These research efforts generally suggest that within the industry, innovations are always explored and introduced to optimise the froth flotation process because it is directly linked to the quality of the metal concentrates.

2.2 Research Related to Reagent Dosage

From the preliminary research conducted, it can be seen that the reagents utilised in the froth flotation process play a significant role. In practice, the type of reagents used is as important as the quantity infused into the system. Xie et al (2017) developed a strategy to be used to determine the appropriate amount of reagents to be used in a flotation process. The experiments conducted revealed that an effective means of determining the right amount of reagent chemicals suitable for an operation is dependent on the ore feed. Hence, an iteration method of arriving at the dosage based on the feed volume was developed and the results of the experiments proved the effectiveness of this strategy. Ozmak and Aktas (2006) tested the results that would be obtained from adding inappropriately high amount of reagent to a flotation process. The experiments showed that the high adsorption of the reagent greatly reduced the required separation between the mineral particles as there was an overflow of hydrophobic particles. The research further concluded that the adsorption of the reagents in a flotation procedure massively affects the froth structure which in turn determines the performance of the system. In addition, Zhang et al (2018) studied the relationship between the froth surface and the dosage of the reagents used in a flotation system. A non-linear model using surface image was developed to approximate the volume of the reagents required to deliver a target froth bubble size. The research concluded that fine tuning the reagent dosage directly impacts the froth surface size which consequently influences the quality of the whole operation. In summary, it can be seen that there have been attempts to predict the optimum reagent dosage within the industry and these methods have yielded promising results. Hence, the aim of this project can be deemed to be feasible.

2.3 Artificial Intelligence in Froth Flotation

As earlier stated, artificial intelligence has grown to be widely applicable in multiple industrial practices and the froth flotation process is not an exception. Wang et al (2018) concluded on a strategy for demonstrating process optimisation by adopting a large number of previous images of froth bubbles clustered accordingly, then analysed and trained them in a convolutional neural network (CNN). The pixel set attributes of each froth bubble image were extracted and the model developed proved to be an effective means of determining the best working conditions for the required froth flotation operation. In other research, a deep neural network architecture was developed by Costa et al (2022) to aid in the multivariate control of the froth flotation system. The benefit of this work was that operators no longer need to spend hours gathering measurements from tedious laboratory analysis to achieve process planning because the results

from the model yielded high accuracy in predictions. Next, in a bid to predict the performance of froth flotation on coal i.e., the combustible recovery and ash content in the coal concentrate, five different machine learning models were implemented on laboratory scale data. It was concluded that when using polymer dosage, pH, polymer conditioning time, sodium metasilicate dosage and impeller speed as the input variables, the Mamdani Fuzzy Logic (MFL) model performed best according to the pre-stated success metrics (Ali et al 2018).

Furthermore, when artificial intelligence models were developed in the research of Jorjani et al (2009) to predict the combustible recovery of coal froth flotation resultant concentrates, the regression model delivered an average correlation coefficient (R-squared) of 0.8 and that of an artificial neural network was 0.95. Moreover, comparative analysis was carried out by Horn et al (2017) on image-based froth data to determine the most important features for performance prediction i.e., feature extraction. The convolutional neural network achieved competitive results and paved the way for future investigation. In addition, Moolman et al (1995) adopted computer vision technology to solve problems in froth flotation. The research discovered that unsupervised neural networks were capable of detecting erratic process control early by monitoring the process behaviour. Although some information was lost in the execution of this technique, smart monitoring and control systems would in practice adopt both supervised and unsupervised learning. Similarly, Aldrich et al (1995) investigated the actual effects of distinct parameters in the interpretation of froth flotation performance using neural networks. The models experimented with varying parameters and combination of parameters in order to achieve the concluded results. It was found that the most vital variables in predicting quality froth were froth stability, mobility and average bubble size. The paper further demonstrated the rapid growth in computer technology and how new generation intelligent automation systems can be developed.

The literature reviewed above indicates that the application of artificial intelligence techniques has widely spread across aspects of froth flotation. This therefore justifies the proposal of this project to adopt an artificial intelligence algorithm to attempt to improve the overall performance in flotation practices.

2.4 Reagent Dosage Prediction

The research contained in this section reveals that there have been attempts made to harness the benefits of controlling and predicting the depressant and collector reagents. For instance, the relationship between the density of iterated froth bubble size and the reagent dosage was

investigated in the cooperative efforts of Zhu et al (2014). In order to recommend a means of determining the optimum reagent dosage by probability density function, the collated images captured from the copper roughing experiment were synthesized into a support vector machine predictive model. The results delivered accuracies within the pre-set success metrics that helped to establish the effectiveness of the proposed method. Next, further research set out to propose suitable methods of discerning appropriate reagent dosage. The idea was to adopt a coordinated optimization technique based on vital attribute variation tendencies and case-based approaches. The sensitivity index of the flotation process is utilised to calculate the main features of reagent dosages on the basis of professional reagent regulation systems in antimony flotation. The conclusion of this research summarized the observable reduction of the tailing indicators and cost of reagent dosage and could present the dawn of the froth flotation optimization (Cao 2018).

Furthermore, in an effort to recommend an average appropriate measurement of reagents in coal flotation, Naik et al (2005) critically examined the various effects of sodium meta silicate, kerosene(collector) and frothed (MIBC) reagents on the procedure. The methodology of the project involved adopting a regression model to predict the quality of the recovered material based on varying reagent settings. During the study, it was discovered that increasing the amount of the sodium meta silicate boosted the recovery percentage without really affecting the concentrate grade and an addition in the MIBC reduced the surface tension between the liquid and vapor segment. Therefore, the resultant production consisted of finer bubble size distribution indicating an improved flotation rate. In conclusion, 0.1g/kg of sodium meta silicate, 0.4g/kg collector and 0.075g/kg MIBC delivered a 95.58% recovery and a 91.11% combustible material grade. Similarly, Ai et al (2018) designed a prediction software for reagent dosages due to the need for reduction in the incessant fluctuation of reagent dosages in typical flotation procedures. Images of froth bubble size were synthesized and the correlation between the sizes of the bubbles and dosages of reagents were explored and used to develop a multi-output least square SVR (Support Vector Regressor) model to illustrate their dynamic relationship. Finally, the health of the reagents was assessed and rated by identifying the class of future froths to be predicted. When adopted in a Chinese gold antimony flotation plant, the efficiency of the froth flotation was improved, and the false reagent iteration was reduced.

Cao et al (2021) recognized that typically, the reagent dosage for froth flotation techniques was determined by artificial identification of froth characteristics. However, this method proved low in accuracy and possessed some bias towards the artificial recognition which led to issues

such as reagent wastage and inconclusive concentrate quality. Hence, a new froth image classification technique derived from the maximal-relevance-minimal-redundancy-semi-supervised Gaussian mixture hybrid model for identification of reagent dosage conditions in coal froth flotation was proposed. This innovation consisted of screening for the important features using feature selection based on class information, implementing clustering analysis to determine the groups of the remaining features and training the hybrid model to predict the reagent dosage based on classification of the image features. The results of this proposed technique provided more reliable guidance for the adjustment of the reagent dosage with high accuracy and time control which can deliver reduction in reagent consumption and production accidents within the flotation process. Also, the limitations of previous methods of controlling reagent dosages served as the motivation for the development of a data driven based model for the predictive control of the iterative addition of reagents in tungsten flotation (Wan et al 2021). The proposed technique adopts a Gate Recurrent Unit (GRU) network to process the flotation variables available to predict the concentrate material grade then a cost function is used as an optimizer to track the performance of the reagent dosage control based on the principles of maximizing and minimizing the constraints of the dosages. The results achieved support the proposal as the model delivered satisfactory performance.

The various research papers that have been reviewed indicate that efforts have been made to reap the benefits of predicting and controlling the dosage of reagents used in froth flotation. These benefits include but are not limited to reduction in reagent wastage, optimum performance of the flotation systems, ability to effectively plan for flotation procedures and improved quality of the resultant material grades. Past attempts mostly adopted image-based data for their predictive models, but it limited the solutions to mostly classification-based results that could not be exactly quantified and translated into direct financial planning. Hence, the innovation that this body of work presents is the utilisation of numerical data gathered from a real-life experiment initially aimed at predicting the percentage of the concentrates.

2.5 Justification for the Proposed Technique

This project proposes using the random forest regressor to predict the required starch and amina flow for a pre-determined amount of raw iron and silica ore feed. To prove the validity of this proposal, literature would be review below to justify the feasibility of using random forest regressor to predict time series data effectively as well as the p-value method of feature selection.

2.5.1 *p*-Values for Variable Selection

The dataset to be utilised in the development of this project is made up of float type data. This means that typical random forest feature selection methods are incompatible with the dataset so other methods of determining the most important features related to the target variables must be explored.

p-Values simply refer to the statistical probability that a variable within a dataset observed is due to chance or is directly relevant to the hypothesis in question (Dahiru 2008). It is a statistical measurement utilised in the validation of a hypothesis against already observed data by quantifying the probability of obtaining those results assuming the null hypothesis is true. Essentially, the lower the p-value, the higher the statistical importance of the observed difference i.e., the vital features to be selected for the prediction of the target variables will have the overall lowest *p*-values (Beers 2022).

Zuo et al (2021) adopted a modification of the p-values method of variable selection for regression models and the results presented a case for its implementation as they were highly competitive with other high dimensional methods. Next, the p-values technique was implemented to select the most informative variables from a metabolomic dataset needed to answer a specific research question. The high dimensional nature of the dataset meant that there was a need to select only the important variables that would thereafter be used to solve problems. The p-values method proved to be very effective in the required selection which ultimately led to accurate satisfactory results from the research (Reneen et al 2016). Motivated by the fact that traditional random forest and support vector feature importance methods tended to work better with categorical data, Altmann et al (2010) proposed the p-value technique as a means of selecting the important variables from a dataset. The results of the research indicated that model interpretability is increased with p-values computation and are hence very helpful for determining significance of variables within a dataset. It was also discovered that an integration of random forest feature importance with p-values computation delivered a more accurate prediction that other existing models.

In summary, it can be concluded that adopting the p-values method of variable selection would be an effective means of trimming down the high dimensionality of the dataset to be processed.

2.5.2 Random Forest Regressor

A random forest algorithm functions as a meta estimator that fits multiple classifying decision trees on a number of sub-samples of the given dataset and utilises the averaging technique to

increase the accuracy of the prediction and manage overfitting (scikit-learn 2022). The random forest method is an aggregation classification method introduced by Breiman in 2001, it adopts the bagging technique to draw various unique sample sets which build a decision tree with attributes that are randomly selected. The set of decision trees $\{h(x,\theta k),k=1\ldots\}$, where $h(x,\theta k)$ is a meta-classifier (an unpruned decision tree designed but classification and regression trees (CART) algorithm), x is the input vector while θk is an independent and similarly distributed random vector. Both of these parameters determine the growth of each decision tree within the process (Liu 2014).

The advantages of this model compared to multiple other models such as neural networks, support vector regression, fuzzy logic, k nearest neighbour among others were summarized in the comparative research carried out by Ao et al (2019). After comparing the performance of the various models, it was found that the random forest regressor possesses a strong learning ability, robustness and attributes that support feasibility of hypothesis. Furthermore, random forests generally are classification and regression algorithms that are easy to parametrize, they work well with outliers, are not sensitive to over-fitting and are also useful in deriving essential information such as feature importance and classification error (Horning 2010). In addition, random forests are very effective in dealing with complex dataset and can provide predictions regardless of the sample size of the data. Random forest regressors are also very useful when typical statistical distributional assumptions are not entirely satisfied (Buskirk 2018). The major challenge faced by the random forest algorithm however is the forest scale not being clearly defined. An oversized scale could lead to redundancy and reduced accuracy of the classification, but this can be resolved by programming the number of estimators within the tree (Liu 2014).

Deductively, the random forest according to past researchers is a very good model to use for large datasets. Owing to the nature of the dataset to be utilised in this project being data gathered from an operation, the ability of the random forest to work well with outliers and not be sensitive to overfitting comes very useful. During any operation, miscellaneous events take place that could result in slight derailments of the data recorded i.e., outliers tend to be present when recorded process data so the characteristics of the random forest justify its selection amidst various other predictive models.

2.5.3 GUI Programming in Python

After the development of a predictive algorithm to be used to aid business practices, it is imperative that the designed system be adapted to a graphical user interface. This allows other stakeholders and intended users of the artefact that do not possess technical knowledge of the system to easily make use of the proposed solution. Tkinter is a python-based library that facilitates an easy-to-program user interface. It contains the toolkit needed for the Tk GUI programming. It was initially developed by John Ousterhout in 1988 at the University of California, Berkeley and it is supported by the company he founded, Sun Microsystems. It functions by requiring the importation of the Tkinter module then defining the widgets that the window would consist of namely, labels, buttons, entries etc (Grayson 2000).

The components of a GUI window are as follows; a window that is the rectangular area containing the application that pops up on the display screen, widgets which refers to the building blocks of the application in the GUI such as buttons, frames, text labels etc and frame, the basic unit of organization for the window layout (Shipman 2013).

Conclusively, this chapter has covered multiple credible sources of information that justify the problem statement as well as the proposed solutions detailed in the first section of this report. The problem of ascertaining the appropriate reagent dosage cannot be overemphasized, hence effective techniques have been explored in order to provide useful solutions that can deliver practical business benefits.

CHAPTER 3: METHODOLOGY

In this chapter, the methods used to achieve the aim and objectives of this project would be elaborated on in detail. The project aims to develop a software resource that can take in a set of input data needed for a flotation operation, process it and deliver accurate predictions for the suitable reagent dosage. The methods employed in the development of the project consisting of the data collection, pre-processing, EDA, modelling and GUI design are summarized in the sections that follow.

3.1 Research Method

The research followed a systematic approach guided by the pre-determined objectives. Firstly, in order to justify the relevance and importance of the research work, a review of past literature that support the need for the software was carried out. This represented the first step of the research method and within the literature review, the impact of the target variables (collector and depressant reagents) were expanded upon. The review chapter also expatiated on key components of the project such as p-values, random forest regressor and the graphical user interface. In summary, it was concluded that the proposed technique is suitable for solving the problem of exploring means of accurately predicting reagent flow. Next, the data was obtained from a credible source within the data science community and has been utilised for other projects with different aims and objectives. The extensive variables within the dataset that required further pre-processing and understanding verified the quality of the data. This is because the behaviours of the major features of a froth flotation process as illustrated in literature were represented therein.

The selection of the random forest algorithm as the preferred model to be used for the project was supported by the cited literature that detailed the benefits of the random forest technique. These benefits were critically analysed and deemed to be best suited for the data collected as well as the desired outcomes and success metrics. Consequently, the predictive model was developed and tested before implementing a graphical user interface design. The results from the model testing were satisfactory so the GUI was developed to be straightforward and easy to understand and use. In summary, the approach adopted in the execution of this project including the chosen techniques data collected and the project plan resulted in the successful completion of the project.

3.2 Data Collection

The proposed system employs the sequential data gathered from a continuous flotation process that spanned over seven months where each data entry was recorded every 20 seconds. The data was sourced from Kaggle, a credible data science platform and the variables within the dataset are detailed in the table below.

S/N	Variable	Definition
1	Date	Date and time data was recorded
2	% Iron Feed	Iron composition of ore feed
3	% Silica Feed	Silica composition of ore feed
4	Starch Flow	Depressant reagent dosage
5	Amina Flow	Collector reagent dosage
6	Ore Pulp Flow	The amount of ore being feed into the system
7	Ore Pulp pH	pH of the ore
8	Ore Pulp Density	Density of the ore
9	Flotation Column 01, 02, 03, 04, 05, 06, 07 Air Flow	These variables represent the amount of air being fed into the system through multiple air flow columns
10	Flotation Column 01, 02, 03, 04, 05, 06, 07 Level	These variables indicate the float thickness of the flotation columns
11	% Iron Concentrate	Resultant concentrate grade of the iron ore
12	% Silica Concentrate	Resultant concentrate grade of the silica ore

The variables can be further split into groups that support the manipulation of the features.

- Main feed inputs: Input variables that cannot be controlled such as % Iron feed, % Silica feed and Ore density.
- Pre-feed inputs: Input variables that are controlled and pre-determined based on the process to be executed namely, date, Starch flow and Amina flow.
- Process variables: These are system variables that are controlled and monitored for the process like ore pulp flow, flotation columns air flow and flotation columns level
- Outputs: % Iron and % Silica concentrates which are the original target variables of the dataset.

The dataset is comprised of the float data type and given the breakdown of the data variables, it is evident that the reagents dosages can be controlled and the ability to pre-determine an optimum dosage can be achieved through data science techniques.

3.3 Data Pre-processing

Data pre-processing is an essential step in the execution of any data related project. It is the means by which data is made ready for processing which involves checking for missing values, scaling, normalizing, replacement of missing values among others. Lawton (2022) described data pre-processing as the preparation of data before further processing techniques are executed on it by transforming it into an easily processable format. This project however only required checking for missing values and possible replacement of missing values. The decision to not carry out further pre-processing was motivated by the uniformity in the data type. The insights gotten from the exploration of the data also supported the conclusion of this action.

3.4 Exploratory Data Analysis

The next phase of the project featured implementing techniques that allowed the nature and shape of the dataset variables to be seen and understood. Exploratory data analysis was defined by Patil (2018) as the critical technique used to carry out initial examinations on datasets to uncover patterns, detect outliers, test hypothesis and verify assumptions with the aid of graphs and statistical representations. A count plot algorithm was deployed in order to display the distribution of the values of each variable so the pattern of the values as well as the outliers can be revealed. A count plot is similar to a histogram or bar chart, it is used to display the counts of observations in distinct categories using bars (Waskom 2022). To further take a deep dive into the nature of the data, the correlation matrix function was introduced so that the relationships that exists between the variables can be uncovered. Correlation matrix as defined by CFI (2022) is a table that illustrates the correlation coefficients between variables of a dataset. It is a useful tool that summarizes a large dataset to highlight and display patterns within the data. The function interprets the results of the action on a scale where correlation values from 0.5 and above represent strong positive correlation while values from -0.5 and below indicates a strong negative correlation.

Part of achieving data understanding includes finding the most important variables needed to predict the reagent flows required. This action could also be grouped as data pre-processing because deriving the most important variable would mean that less significant features would be dropped in order to achieve a much easier to work with dataset. Performing this task

employed the technique of finding the p-values of the variables because typical random forest feature importance methods are not compatible with the float data type. The successful execution of this technique delivered a graph that plots the variables of the dataset against their corresponding p-value. The p-value of a variable refers to the probability that the presence of that variable being essential in the prediction of the target variable is due to chance. Hence, the lower the p-value of a variable, the more important it is so variables with p-values of 0 were selected as important for the prediction of the reagents flows.

Finally, the dataset variables that were calculated to have less significance in the prediction of the target variables were dropped.

3.5 Model Development

The random forest regressor was deployed at this stage and appropriate parameters were set in order to achieve satisfactory results. The test size was set to 0.15, random state was set to 0 and the number of estimators was set to 100. The model was trained and the r2 scores were derived in order to justify the selection of the algorithm. Before proceeding with the GUI development, the model was tested with values gotten from the data so the effectiveness of the model can be judged in real time.

3.6 Performance Metrics

The success metrics set adopted for the model is the r2 score. It is the accuracy measure compatible with the random forest algorithm, and it is such that an r2 score of >0.95 represents a very high accuracy. R2 score, pronounced as R squared is a vital performance metric in machine learning sometimes known as coefficient of determination. It is simply the amount of variance that exists in the predictions derived from the model (Kharwal 2021).

3.7 GUI Design

This aspect of the project was developed with the aid of the tkinter library in python. This library enabled the graphical user interface required for this project to be built easily and understandably. The window was designed within a rectangular frame with two columns and 11 rows. The first 9 rows consisted of the labels and entries of the values required from the user to achieve the prediction of the target variables. The last two rows displayed the resultant predictions of the starch and amina reagents flows.

3.8 Research Implementation

The findings from literature related to the problem statement and proposed solutions culminated in the development of a software capable of predicting the flow of reagents that would be effective in the froth flotation of iron and silica ore. The created software solution was tested and verified to be suitable for real world applications as the results gotten from the execution of the in-built model are consistent with the existing data. The figures below capture the these claims as well as possible.

In summary, this chapter described the methods followed in undertaking the project's objectives from the initial research to the research implementation. The following chapters will discuss the results gotten and the conclusions that can be made from them.

CHAPTER 4: RESULTS

From the efforts made to execute this project, this chapter will present and detail the results obtained. From the data collection up till the implementation of the software, the observations made from the outcomes of the research would be summarized.

4.1 Data Collection

As earlier stated, the data was sourced from Kaggle (2021), and it was made up of 700,000 rows and 23 columns. This volume of data however was quite challenging to work with as it took a very long time for the dataset to be loaded into data frame, it also requires a lot of machine power from the computer being utilised for the project so means of reducing the data needed to be sought after. Manku et al (1999) concluded from their joint research that adopting a systematically random approach to sampling a large dataset delivers a quantifiably better performance when working with large datasets. It could be the top 1000, middle 3000 or the bottom 5000 as suggested by Brownlee (2017) in his contribution on how to handle large data files for machine learning. He explained that sampling data from large datasets is a good practice in machine learning because it gives quick spot checks of models and reduces the result turnaround time. For this project, the top 3000 entries were selected as the samples to be utilised in the machine learning of this project. The figure below shows the top 5 rows of the selected sample data.

	Starch Flow	Amina Flow	% Iron Feed	% Silica Feed	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 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	Flotati Colur 07 Lev
0	3019.53	557.434	55.2	16.98	395.713	10.0664	1.74	249.214	253.235	250.576	...	250.884	457.396	432.962	424.954	443.558	502.255	446.370	523.3
1	3024.41	563.965	55.2	16.98	397.383	10.0672	1.74	249.719	250.532	250.862	...	248.994	451.891	429.560	432.939	448.086	496.363	445.922	498.0
2	3043.46	568.054	55.2	16.98	399.668	10.0680	1.74	249.741	247.874	250.313	...	248.071	451.240	468.927	434.610	449.688	484.411	447.826	458.5
3	3047.36	568.665	55.2	16.98	397.939	10.0689	1.74	249.917	254.487	250.049	...	251.147	452.441	458.165	442.865	446.210	471.411	437.690	427.6
4	3033.69	558.167	55.2	16.98	400.254	10.0697	1.74	250.203	252.136	249.895	...	248.928	452.441	452.900	450.523	453.670	462.598	443.682	425.6

5 rows × 23 columns

Figure 2: Top 5 rows of the dataset

4.2 Data Pre-processing

In order to ensure the quality of the data to be utilised in the development of the predictive software, the first pre-processing technique implemented was checking for missing values as they can gravely derail the results obtained. The figure below shows the result of scrutinising the data for missing values. It was found that there were no missing values in the data, so no further action was required.

```
These are the missing data
 Empty DataFrame
Columns: [Starch Flow, Amina Flow, % Iron Feed, % Silica Feed, Ore Pulp Flow, Ore Pulp pH, Ore Pulp Density, Flotation Co
lumn 01 Air Flow, Flotation Column 02 Air Flow, Flotation Column 03 Air Flow, Flotation Column 04 Air Flow, Flotation Col
umn 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, Flo
tation Column 07 Level, % Iron Concentrate, % Silica Concentrate]
Index: []

[0 rows x 23 columns]
```

Figure 3: Output of Pre-processing

4.3 Exploratory Data Analysis

The methods of achieving data understanding that were adopted at this stage of the project included countplot, correlation matrix and p-values. The p-value function delivered a graph that plotted the variables against their p-values indicated the probability of their relevance to predicting the starch and amina flows respectively. The variables that had a p-value of 0 were deemed to be the important features needed to deliver accurate predictions of the target variables and they were % Iron Feed, % Silica Feed, Ore Pulp pH, Ore Pulp Density, Flotation Column 01 Air Flow, Flotation Column 04 Air Flow, Flotation Column 05 Air Flow, Flotation Column 06 Level and Flotation Column 07 Level as illustrated in the following graph. The % silica concentrate, although indicated by the p-values graph as being essential in the prediction of the reagents cannot be included in the design of the model. This is because the software is designed to be applied in real world practice therefore, the factory operators would be unable to fill in the % silica concentrate before the start of the operation. The selected variables fall under features that can be controlled and determined before the start of the operation.

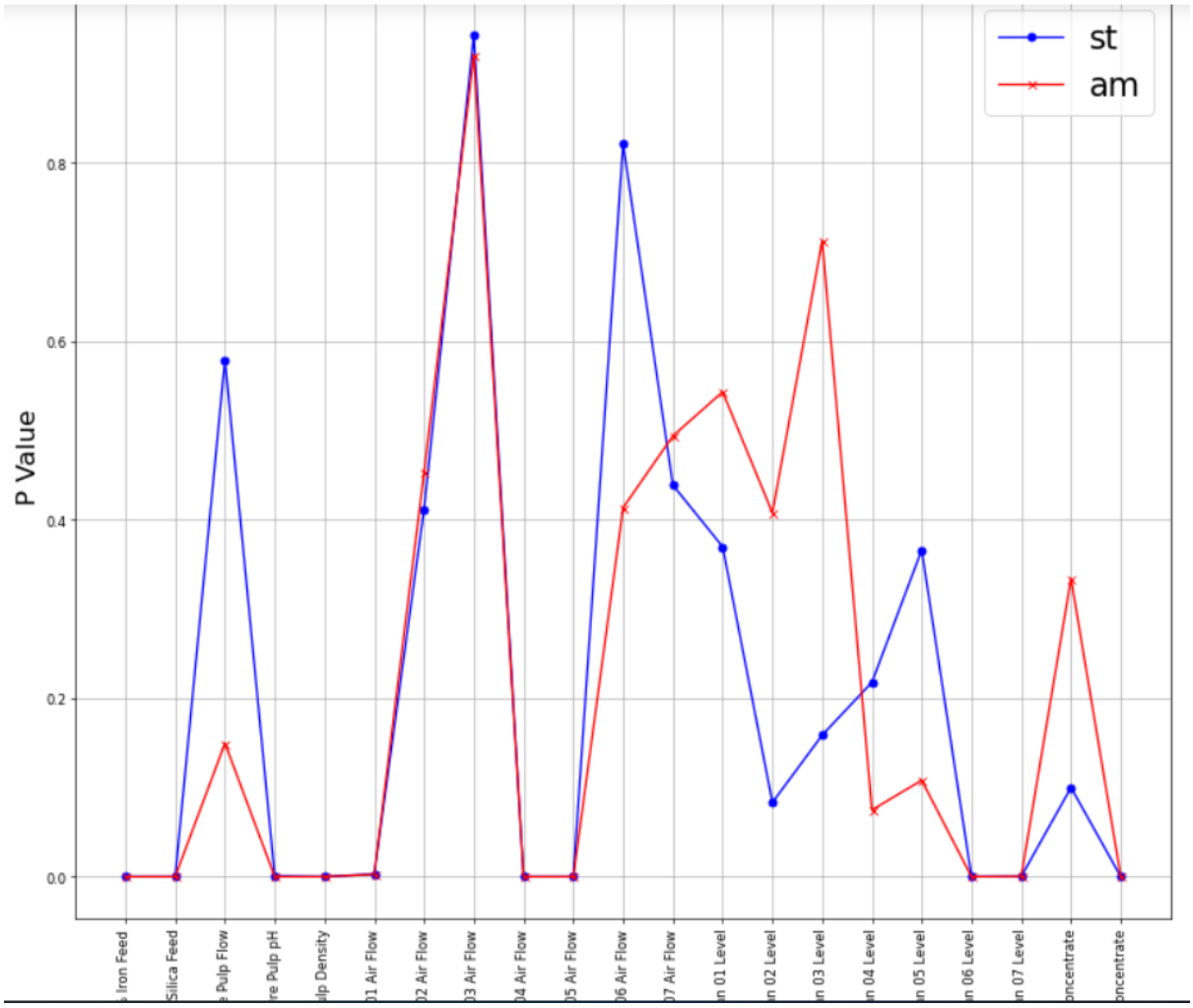

Figure 4: p-values

Having deduced the vital variables needed for the development of the predictive model in order to achieve a more straightforward design, the results obtained from the countplot can hence be elaborated upon.

- % Iron Feed: the countplot graph for this feature shows that the percentage composition of the ore being fed into the system remained relatively the same as the exact values were 54.95, 55.2 and 55.99 which average out to 55.3%.

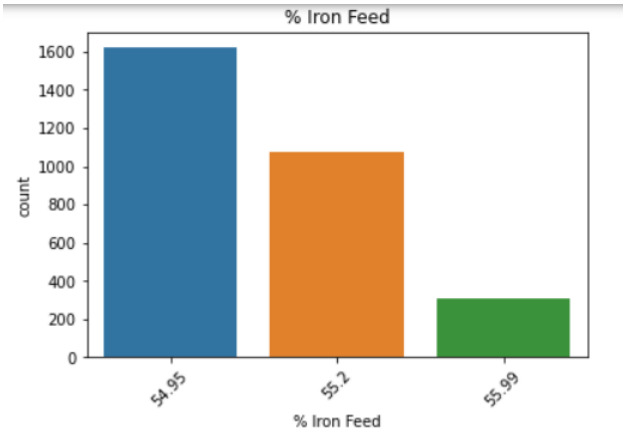

Figure 5: % Iron Feed

- % Silica Feed: Similarly, the proportion of the ore feed contained an average of 17.18% as shown in the graph below. The ratio of the metals in within the ore stayed approximately constant throughout the process.

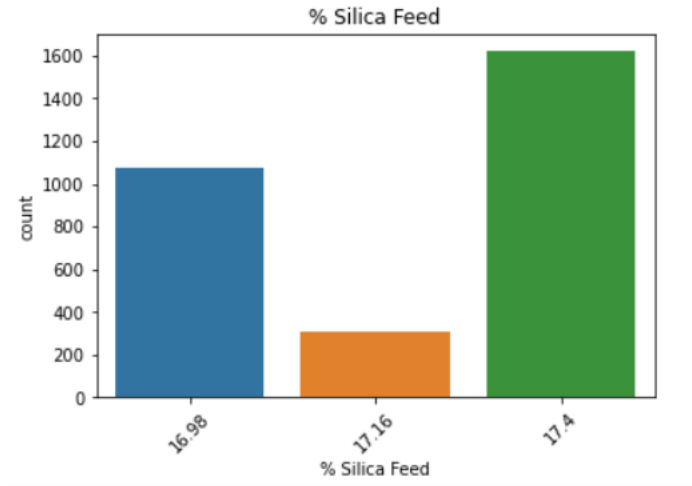

Figure 6: % Silica Feed

- Ore Pulp pH: As result below illustrates, the pH of the ore pulp increased approximately uniformly throughout. The pH of a substance can be altered by the addition of water; hence it can be inferred that the rises in the plot may have been caused by the air flow. For this project, it is assumed that the outliers are negligible.

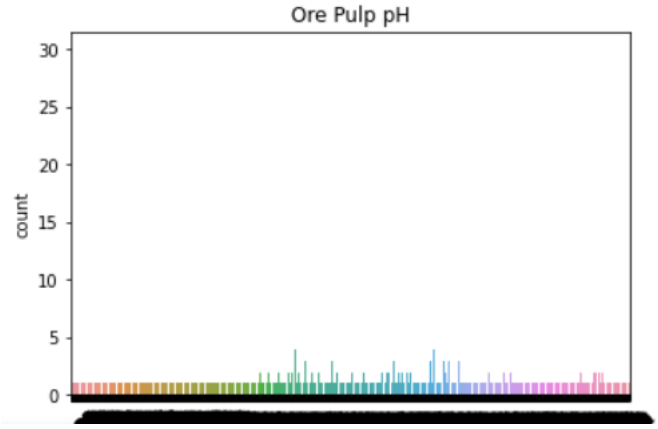

Figure 7: Ore Pulp pH

- Ore Pulp Density: the density of the ore pulp categorically remained the same except for a slight hike which can also be deemed negligible in the process. Density, a function of the mass of a substance divided by its volume is one of the significant variables needed for the prediction of the reagent flow is expected to stay constant for the process.

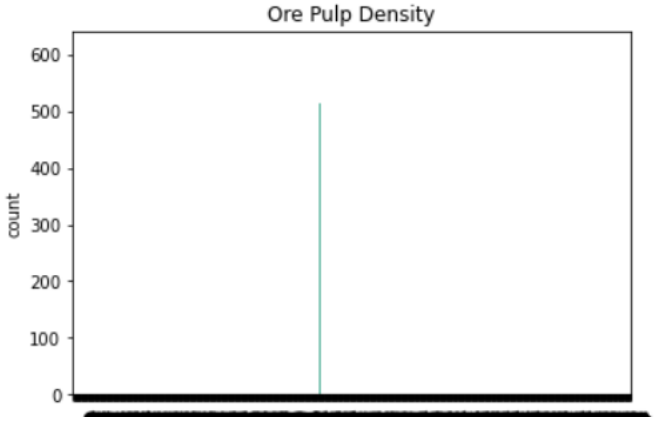

Figure 8: Ore Pulp Density

- Flotation Column 01 Air Flow: this variable as shown in the graph below is expected to be altered intermittently during the process. As earlier detailed in the previous chapter it is the amount of air being fed into the system but overall, it increased within a given range of values.

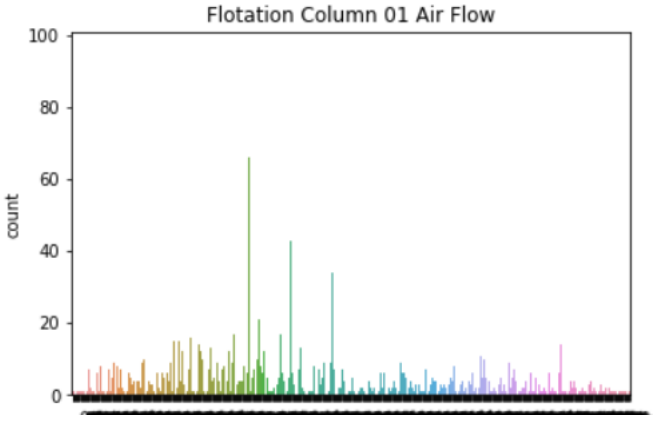

Figure 9: Flotation Column 01 Air Flow

- Flotation Column 04 Air Flow: as suggested in the flotation column 01 air flow, the amount of air being fed into the system per column is generally as uniform as possible, so it does not affect the flotation process. The graph below clearly indicates that the setting of the air flow in the 4th column stayed the same all through the process.

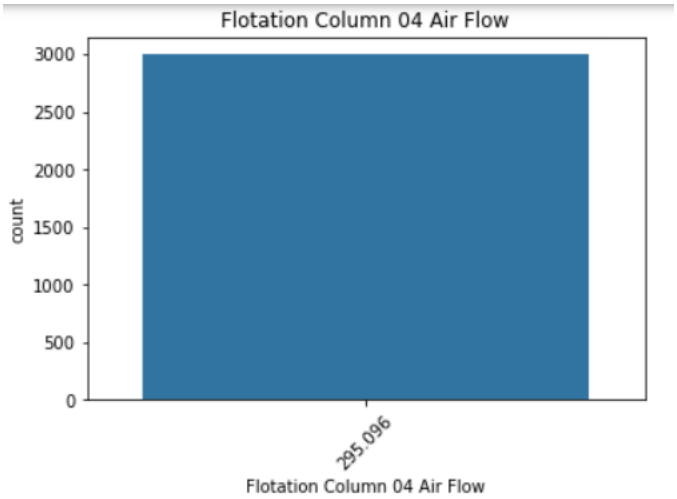

Figure 10: Flotation Column 04 Air Flow

- Flotation Column 05 Air Flow: Similarly, the value of the air flow amount stayed uniform through out the operation in order to maintain the operations behaviour.

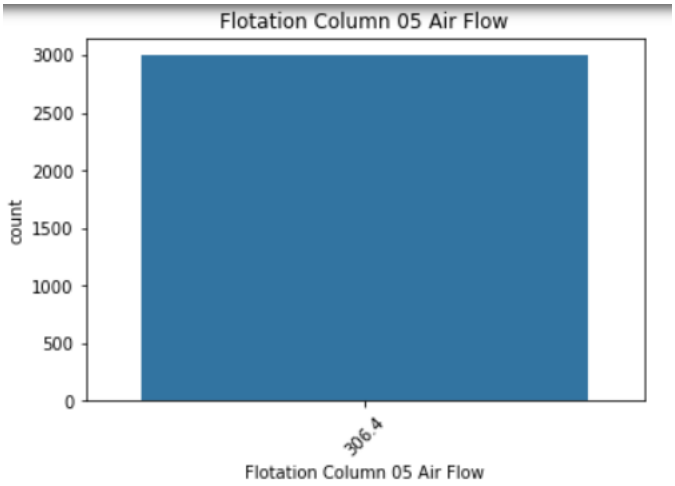

Figure 11: Flotation Column 05 Air Flow

- Flotation Column 06 Level: The p-values graph revealed that the level of the flotation thickness is essential to predicting the amount of starch and amina reagents required for the operation. The graph below illustrates the behaviour of the level throughout the operation, and it can be seen that it is mostly consistent in its increment at this level. The disruptive increments could be down to a number of systemic factors that can be further researched.

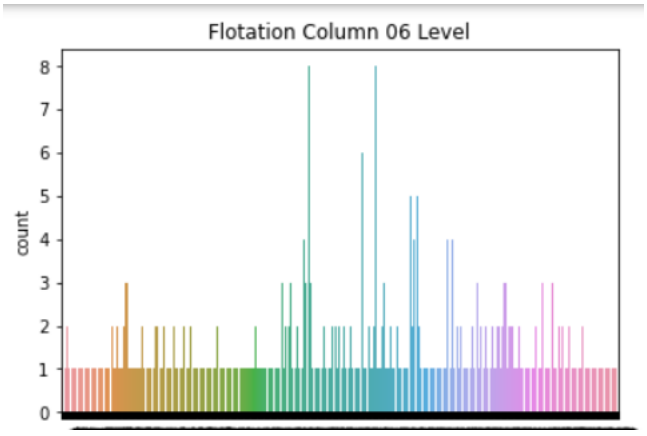

Figure 12: Flotation Column 06 Level

- Flotation Column 07 Level: Similar to the behaviour of the 6th column level, the flotation thickness level here stayed increasing uniformly and the outliers can be researched in further study.

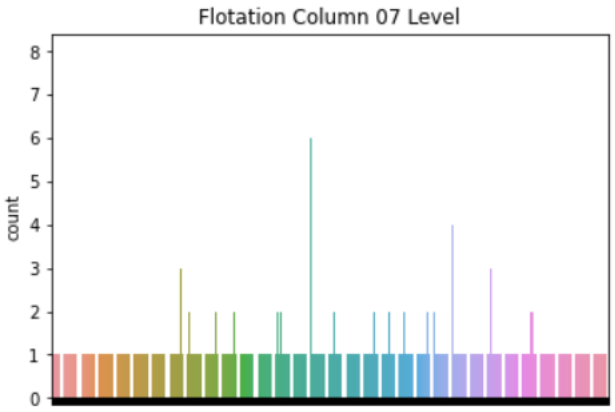

Figure 13: Flotation Column 07 Level

Next, the correlation matrix was plotted to uncover any relationship that exist between the starch and amina flow and other variables. The figure below illustrates the result of the correlation action, and it can be seen that no significant correlation exists between the reagents and other variables. The only observable relationships present are the ones that exist between the flotation column levels, understanding the positive correlation between these variables are inconsequential to the study so they can be ignored till further study.

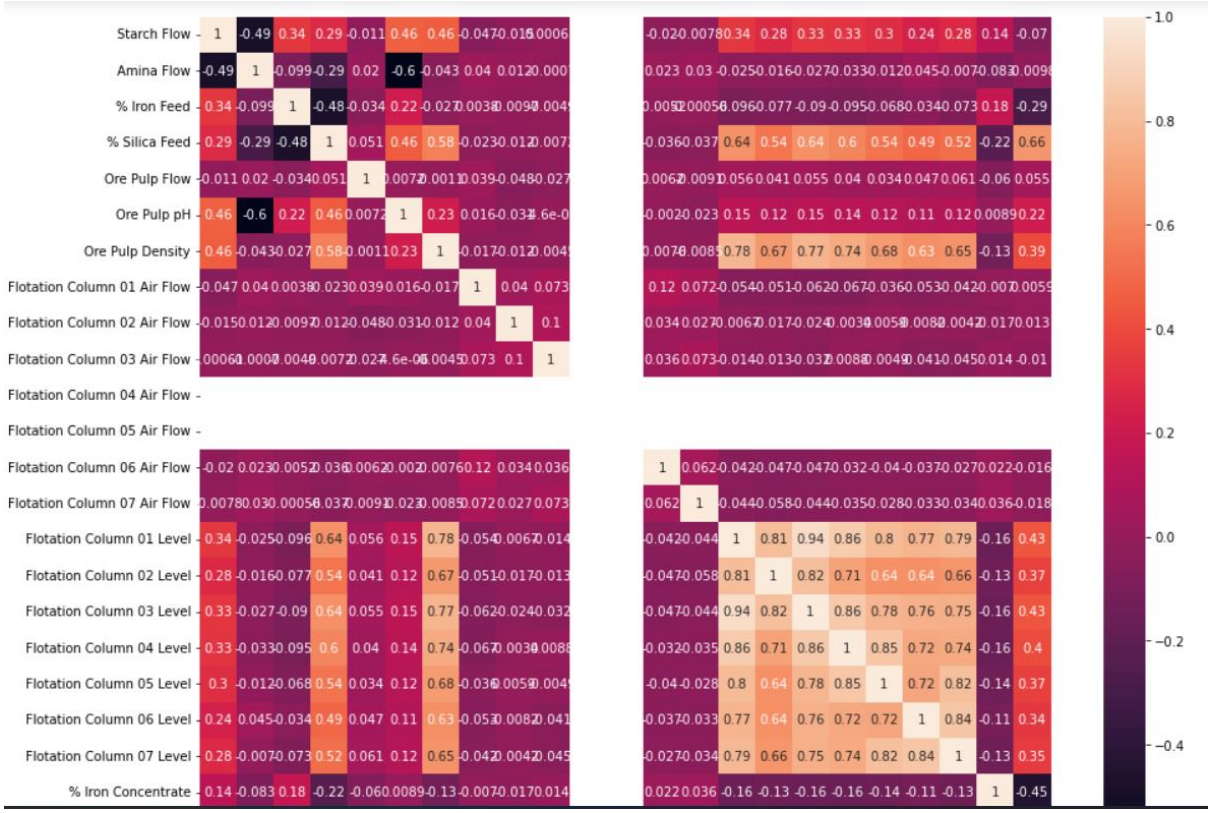

Figure 14: Correlation Matrix

4.4 Model Development

The random forest regressor algorithm was chosen for this execution of this project. Although the starch and amina flow are predicted with the same variables, the system was designed for the reagents to have individual models so their accuracy can be measured appropriately. The figure below demonstrates the development of the individual models for the prediction of each of the reagents and their corresponding R2 score. As pre-determined, the threshold of the R2 score for a satisfactory random forest predictive regressor is >0.95. Hence, the individual models both performed excellently as they both had R2 scores of 0.984 and 0.991 respectively.

```
In [12]:
regressor_st = RandomForestRegressor(random_state = 0, n_estimators = 100)
regressor_st.fit(flotation_parameters,starch_flow)
y_pred_st = regressor_st.predict(flotation_parameters)

print('R2 Score of Random Forest Regression with Only starch',r2_score(starch_flow,y_pred_st))
```

```
R2 Score of Random Forest Regression with Only starch 0.9849948586421621
```

```
In [13]:
regressor_am = RandomForestRegressor(random_state = 0, n_estimators = 100)
regressor_am.fit(flotation_parameters,amina_flow)
y_pred_am = regressor_am.predict(flotation_parameters)

print('R2 Score of Random Forest Regression with Only amina',r2_score(amina_flow,y_pred_am))
```

```
R2 Score of Random Forest Regression with Only amina 0.9912115625545493
```

Figure 15: Random Forest Regressor

```
In [14]:
predictions = {'% Iron Feed':55.2,
               '% Silica Feed':16.98,
               'Ore Pulp pH':10.0697,
               'Ore Pulp Density':1.74,
               'Flotation Column 01 Air Flow':250.203,
               'Flotation Column 04 Air Flow':295.096,
               'Flotation Column 05 Air Flow':306.4,
               'Flotation Column 06 Level':443.682,
               'Flotation Column 07 Level':425.679,
               }
```

```
In [15]:
predict_values = []
for value in predictions.values():
    predict_values.append(value)

predict_st = regressor_st.predict([predict_values])
predict_am = regressor_am.predict([predict_values])

print('Predicted starch flow =',predict_st)
print('Predicted amina flow =',predict_am)
```

```
Predicted starch flow = [3051.7059]
Predicted amina flow = [560.90483]
```

Figure 16: Model Testing

The values that were used to test the model were gotten from the main dataset to visually gauge and compare the results from the models with the actual values. This is illustrated in the following screenshot and the comparison is summarized in the table below.

Predicted values	Actual values
Starch Flow: 3051.7059	Starch Flow: 3033.69
Amina Flow: 560.904	Amina Flow: 558.167

4.5 Graphical User Interface

Given that the model functioned satisfactorily, for it to be deployed in real life practice, it needed to be well packaged. Using the tkinter library in python, a gui for the prediction model was developed to take in the required inputs from the user, process it and deliver the predicted output. The screenshot below demonstrates the deployment of the software application with the results being similar to what was gotten before the gui development.

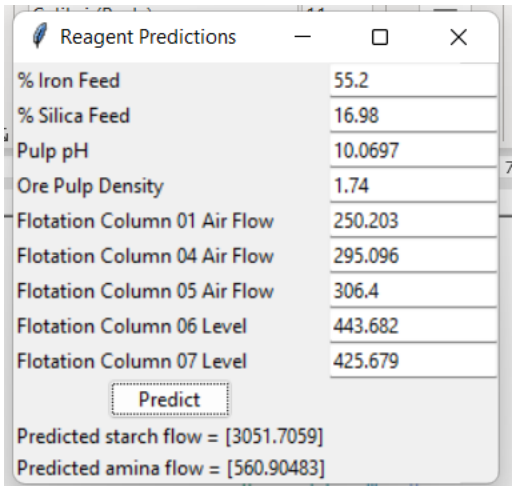

Figure 17: GUI

CHAPTER 5: DISCUSSION

Following the results illustrated in the previous chapter, the findings, recommendations, benefits and limitations of this project will be detailed below. The discussions from this chapter will facilitate the formation of the conclusion section of the report.

5.1 Key Findings

The exploratory data analyses revealed the behaviour of the key variables needed to predict the desired outcome. The uniform increment was illustrated and averaged out to show that the values of the variables tended to stay approximately similar throughout the recorded operation. This uniformity indicates that the output from the model can be used as a guide for operation planning and the benefits of the designed solution can be accrued. Although the correlation matrix technique yielded information that was adjudged to not be useful to the project, it was worth carrying out because the lack of correlation suggests that each variable functions independent of the rest when predicting reagent flow. The random forest regressor was also found to be an effective predictive model because it delivers an excellent R2 score and this fosters confidence in the results obtained from the execution of the model.

5.2 Recommended Implementation of the Product

The output of the prediction software is the starch and amina flow. Flow is the rate at which a fluid moves and is the product of multiplying the volume of the liquid by time. Hence, the output of the system can be utilised in two ways. The first one is that the derived suitable reagent flow informs the user of the rate at which the reagents need to be fed into the froth flotation system during the operation. Secondly, the predicted flow can be used to calculate the total volume of reagent needed for a set period of flotation operation which in turn inform the financial and operations plan for a proposed flotation operation.

5.3 Benefits of the Software

- The predicted values serve as a guide to the average flow of reagents that would be compatible with the inputted composition and density of the ore to be processed
- Project management related to a froth flotation operation can be assisted by the results obtained from utilising the product
- Wastage of depressant and collector reagents will be reduced due to the optimised project planning afforded by the use of this solution

- The quality of the processed ore will be enhanced through the accurate dosing of the reagents as concluded from past research.
- Time taken to prepare for a flotation process is reduced since an easy method of reagent estimation has been introduced as opposed to the previously tedious methods.

5.4 Limitations

- The dataset used centred around iron and silica ores. The system is yet to be tested to be functional with other ores
- Other predictive models were not explored and compared to the performance of the selected random forest regressor
- The larger dataset was not incorporated in the predictive model so the accuracy of the prediction could be improved
- More exploratory data analysis could have been carried out to achieve deeper understanding of the dataset
- The graphical user interface designed was quite basic and therefore can be further customized to fit the desires of the potential users of the software

CHAPTER 6: CONCLUSION

In summary, the aim of this project was to develop a software application that predicts the required dosage of depressant and collector reagents needed to process a given amount of iron and silica composition feed. To achieve this, the objectives that needed to be met were:

- To research related previous work to determine the feasibility of the proposed solution
- To obtain high quality data, pre-process it and achieve understanding of it
- To research and conclude on the random forest regressor as a machine learning model that will accurately predict the target variables
- To ensure the model performs as required by the pre-stated success and evaluation metrics
- To develop a graphical user interface for the model.

After following the methodology and analysing the results of the project, it can be concluded that the proposed solution answered the research questions by achieving the aim and objectives of the project.

The software can hence be launched and utilised in real world scenarios to reap the inherent benefits of the solution.

Reflection

Undertaking this project has helped to strengthen my confidence and understanding of the data science discipline and I look forward to working on more exciting and impactful projects in the future.

References

Aktas, M. O. a. Z., 2006. Coal Froth Flotation: Effects of Reagent Adsorption on the Froth Structure. *ACS Publications.*

Alexsander C.A.A. Costa, F. V. C. L. R. A. L. C. T. A. P. B., 2022. Deep architecture for silica forecasting of a real industrial froth flotation process. *Elsevier.*

Andre Altmann, L. T. O. S. T. L., 2010. Permutation importance: a corrected feature importance measure. *Oxford Academic.*

Armando Correa De Araujo, A. E. C. P., 1995. Froth Flotation: Relevant Facts and the Brazilian Case. *Brazil.*

Beers, B., 2022. p-value: what is it, how to calculate it and why it matters. *Investopedia.*

Bin-fang Cao, Y.-f. X. W.-h. g. C.-h. Y. a. J.-q. L., 2018. Coordinated optimization setting of reagent dosages in roughing-scavenging process of antimony flotation. *SpingerLink.*

Borregaard, 2022. *Borregaard.* [Online] Available at: https://www.borregaard.com/markets/mineral-processing/all-you-need-to-know-about-mineral-processing/

Brownlee, J., 2017. *Machine Learning Mastery.* [Online] Available at: https://machinelearningmastery.com/large-data-files-machine-learning/ [Accessed 7 September 2022].

Buskirk, T. D., 2018. *Surveying the forests and sampling the trees: an overview of classification and regression trees and random forests with applications in survey research,* Boston: Center for Survey Research and Department of Management Science and Information Systems.

C. Aldrich, D. W. J. V. D. D. B., 1995. The interpretation of flotation froth surfaces using digital image analysis and neural networks. *Elsevier.*

CFI, 2022. *Corporate Finance Institute.* [Online] Available at: https://corporatefinanceinstitute.com/resources/excel/study/correlation-matrix/#:~:text=A%20correlation%20matrix%20is%20simply,patterns%20in%20the%20given%20data.

D. Vamvuka, V. A., 2001. The effect of chemical reagents on lignite flotation. *Elsevier.*

D. W. Moolman, C. A. J. V. D., 1995. The monitoring of froth surfaces on industrial flotation plants using connectionist image processing techniques. *Elsevier.*

Dahiru, T., 2008. p-Value, a true test of statistical significance? a cautionary note. *National Library of Medicine.*

Danish Ali, M. B. H. L. A. O. K. M., 2018. An evaluation of machine learning and aritificial intelligence models for predicting the flotation behaviour of fine high-ash coal. *Elsevier.*

E. Jorjani, H. A. P. A. S. S. C. C. S. M. M. S., 2009. Prediction of coal response to froth flotation based on coal analysis using regression and artificial neural network. *Elsevier.*

Fawcett, F. P., 2013. Relationship to big data and data driven decision making. *Mary Ann Libert Inc.*

Grayson, J. E., 2000. *Python and Tkinter Programming.* Greenwich: Manning Publications.

Gurmeet Singh Manku, S. R. B. G. L., 1999. Random sampling techniques for space efficient online computation of order statistics of large datasets. *ACM Digital Library.*

H. Sis, S. C., 2003. Reagents used in the flotation of phosphate ores: a critical review. *Elsevier.*

Horning, N., 2010. *Random Forests: An Algorithm for image classification and generation of continuous fields data sets,* New York: International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences.

Hsinchun Chen, R. H. L. C. a. V. C. S., 2012. Business Intelligence and Analytics. *JSTOR.*

IBM, 2021. *IBM.* [Online] Available at: www.ibm.com/cloud/learn/what-is-artificial-intelligence [Accessed 2022].

Jianyoung Zhu, W. G. C. Y. H. X. X. W., 2014. Probability density function of bubble size based reagent dosage predicitive control for copper roughing flotation. *Elsevier.*

Jin Zhang, Z. T. M. A. W. G., 2018. Nonlinear modelling of the relationship between reagent dosage and flotation froth surface image by Hammerstein-Weiner Model. *Elsevier.*

Kharwal, A., 2021. *The clever programmer.* [Online] Available at: https://thecleverprogrammer.com/2021/06/22/r2-score-in-machine-learning/#:~:text=The%20R2%20score%20is%20a,predictions%20explained%20by%20the%20dataset.

Kitchener, J. A., 1984. The Froth Flotation Process: Past, Present and Future- In Brief. *SpringerLink.*

Kun Wan, D. X. Y. C., 2021. Data driven based model predictive control of reagents addition for tungsten flotation. *IEEE Xplore.*

Lawton, G., 2022. *SearchDataManagement.* [Online] Available at: https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing

Liu, Y., 2014. *Random Forest Algorithm in Big Data Environment,* Beijing, China: School of Economics and Management, Beihang University.

Ma, M., 2012. Froth Flotation of Iron Ore. *International Journal of Mining Engineering and Mineral Processing.*

Mari van Reenen, C. J. R. J. A. W. J. H. V., 2016. Variable selection for binary classification using error rate p-values applied to metabolomic data. *SpringerLink.*

McCarthy, J., 2004. What is artificial intelligence. *Stanford university.*

Merriam-Webster, 2022. *Merriam-Webster.* [Online] Available at: https://www.merriam-webster.com/dictionary/froth%20flotation

Michaud, D., 2021. *911metallurgist.* [Online] Available at: https://www.911metallurgist.com/blog/froth-flotation-process

Mingxi Ai, Y. X. D. X. W. G. C. Y., 2018. Data-driven flotation reagent changing evaluation via union distribution analysis of bubble size and shape. *Wiley Online Library.*

Mousumi Gharai, R. V., 2018. Effect of reagent dosages on the entrainment factor of flotation of copper ore at laboratory scale. *Research Gate.*

Murat Erol, C. C. Z. A., 2003. The effect of reagents and reagent mixtures on froth flotation of coal fines. *ResearchGate.*

Patil, P., 2018. *Towards Data Science.* [Online] Available at: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

Pradyumna K. Naik, P. S. R. R. V. N. M., 2005. Interpretation of interaction effects and optimization of reagent dosages for fine coal flotation. *Elsevier.*

Quintanilla, P., 2021. Modelling for Froth Flotation Control: A review. *Science Direct.*

Sangita Mondal, A. A. U. M. B. S., 2021. Froth Flotation Process and its Application. *ResearchGate.*

scikit-learn, 2022. *scikit-learn.* [Online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

Shipman, J. W., 2013. *Tkinter 8.5 reference: a GUI for Python*, s.l.: New Mexico Tech Computer Center.

S, L. N., 2015. Using data science and big analytics to make healthcare green. *IEEE Xplore.*

Urbina, R. H., 2003. Recent developments and advances in formulations and applications of chemical reagents used in froth flotation. *Taylor Francis Online.*

Waskom, M., 2022. *Seaborn.* [Online] Available at: https://seaborn.pydata.org/generated/seaborn.countplot.html

Wenyan Cao, R. W. M. F. X. F. H. W. Y. W., 2021. A new froth image classification method based on the MRMR-SSGMM hybrid model for recognition of reagent dosage condition in the coal flotation process. *SpringerLink.*

Xiaoli Wang, C. S. C. Y. Y. X., 2018. Process working condition recognition based on the fusion of morphological and pixel set features of froth for froth flotation. *Elsevier.*

Xinhai, 2022. *Xinhai Mining Processing.* [Online] Available at: www.xinhaimining.com/newo/895.html

Yali Cheng, F. M. H. L. J. C. X. F. .., 2022. Effect of reagent interaction on froth stability of coal flotation.. *Elsevier.*

Yi Zuo, T. G. S. a. J. D. B., 2021. Variable selection in GLM and Cox models with second generation p-values. *Cornell University.*

Yile Ao, H. L. L. Z. S. A. Z. Y., 2019. The linear random forest algorithm and its advantages in machine learning assisted logging regression modelling. *Elsevier.*

Yonfang Xie, J. W. D. X. C. Y. W. G., 2017. Reagent Addition Control for Stibium Rougher Flotation Based on Sensitive Froth Image Features. *IEEE Xplore.*

Z.C. Horn, L. A. J. M. C. A. B. H., 2017. Performance of convolutional neural networks for feature extraction in froth flotation sensing. *Elsevier*.