

https://dspace.org/

Modelling the effect of production process parameters on the dispersion capabilities of lignosulphonates

Dhanpat, Jennica

2022

https://hdl.handle.net/10539/34896

Downloaded from a DSpace repository. https://dspace.org/



UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Modelling the effect of production process parameters on the dispersion capabilities of lignosulphonates

MSc Research Dissertation

Prepared by

Jennica Dhanpat

(1054378)

Submitted to

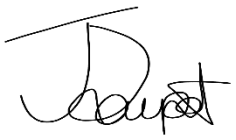
A dissertation submitted to the Faculty of Engineering and Built Environment, University of Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science in Engineering**.**

Supervisors: Dr Kevin Brooks and Mr Antony Higginson

July 2022

Declaration

I, Jennica Dhanpat, declare that this dissertation is my own, unaided work. It is being submitted for the degree of Master of Science in Engineering at the University of Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other university.



Signature of Candidate (Jennica Dhanpat)

.....**29th** .....day of.....**July**...**2022**.....in... **Johannesburg**.....

Part of the research completed in this dissertation has been submitted and accepted by the accredited Control Conference Africa 2021 Program Committee and has been published in a volume of IFAC-PapersOnLine. [Jennica Dhanpat, Antony Higginson, Kevin Brooks (2021), “Estimation of the Effect of Bio-Admixtures on Concrete Workability Using Linear Regression and Support Vector Machines”, IFAC-*PapersOnLine*, 54(21), 133-138].

Modelling the effect of production parameters on the dispersion capabilities of lignosulphonates

Abstract

The Sappi Tugela Mill produces lignosulphonate as a by-product of the Neutral Sulphite Semi-Chemical (NSSC) pulping process. The product is primarily used as a dispersant in the concrete and cement admixture markets, where it is blended with other plasticisers such as polycarboxylate ethers. Attempts to predict the lignosulphonate dispersion characteristics and align them with market demands have been a technical challenge for researchers in this field of industry. Thus, this study aimed to understand and model the effects of process parameters on lignosulphonate dispersion capabilities. Currently, three offline methods are used to assess the dispersion performance of the lignosulphonate produced: the insoluble content, the dispersion index, and the concrete slump test. As a result, predictive models based on process and product knowledge, as well as data analysis using various regression modelling methods, were used to estimate these measurements. RapidMiner and Microsoft Excel were used to develop these models.

The concrete slump data comprised five measurements taken at 15-minute intervals, beginning with an initial value. When the training data were normalised by the initial value, the trends of the data were reasonably linear, implying that all the slump data can be fitted with two parameters, the initial value and time. A general model for predicting concrete slump behaviour was found, in which the slump data was defined by a simple quadratic function and a Neural Net model was developed, using process parameters to predict the initial concrete slump values. Using production data, Random Forest models were developed to predict the insoluble content and dispersion index values. The developed models' results were compared to actual laboratory values, and a simple adaptive approach, of bias updating, was used to improve the models' fit to the data. This work developed a useful system for predicting the lignosulphonate dispersion performance, thus reducing the reliance on laboratory results, and allowing for almost immediate changes to be made to the dispersing agent production process. The modelling approach used in this study proved successful.

The models' predicted values generalized to a reasonable degree to the test sets and captured the trends of the actual values, but the models made significant prediction errors. The reasons for this performance are presented, and this includes the optimal

hyperparameters obtained and the conditions under which these tests are conducted. Improving model performance through bias updating resulted in only minor differences between model predictions and actual laboratory values. Therefore, signifying quality models were developed for a lignosulphonate prediction system or implementation into a suitable control strategy.

Keywords: dispersion, lignosulphonate, plasticizer, soft sensor, pulping, neutral sulphite pulping process, Random Forest, Neural Net, concrete slump, dispersion index, insoluble content.

This dissertation is dedicated to my late father, Gayandeo Dhanpat.

For his endless love, support, encouragement, and wisdom. I hope that this accomplishment brings me one step closer to reaching the dream you had for me.

"A great soul serves everyone all the time. A great soul never dies. It brings us together again and again."

Maya Angelou

Acknowledgements

I would like to express my gratitude:

To my parents and brother, who have seen me work tirelessly and purposefully towards my ambitions. I am grateful for all your encouragement and support in completing this research; without you, I would not have been able to achieve what I have thus far.

To my supervisors, Mr Antony Higginson, and Dr Kevin Brooks, for their insightful tuition, guidance, and supervision on this research.

To Sappi Ltd for providing the project and those who have assisted with their support and guidance on the project.

Contents

Declaration i

Abstract iii

Acknowledgements vi

List of Figures xi

List of Tables xiii

Nomenclature xiv

1 Introduction 1

1.1 Problem statement 2

1.2 Research aims and objectives 4

1.3 Scope of the study 5

1.4 Significance of study 6

1.5 Overview of the dissertation 6

2 Literature review 7

2.1 Lignin 7

2.1.1 Lignin in wood 7

2.1.2 Lignin biosynthesis 7

2.1.3 Physical and chemical structure of lignin 9

2.1.4 Potential uses of lignin 9

2.2 Lignosulphonate 11

2.2.1 Lignosulphonate production 11

2.2.2 Structural properties of lignosulphonate 13

2.2.3 Uses of lignosulphonate 15

2.2.4 Concrete admixtures 16

2.3 Soft sensors in the process industry 19

2.3.1 RapidMiner 20

3 Sappi Tugela plant 23

3.1 Overview of the process 23

3.1.1 Raw material preparation 24

3.1.2 Wood pre-treatment 25
3.1.3 Digestion 25
3.1.4 Spent liquor recovery 26
3.2 Techniques to monitor lignosulphonate dispersion performance 27
3.2.1 Tests 28
3.2.2 Dispersion index 30
3.2.3 Concrete slump 30
3.2.4 Insoluble content 32
3.3 Important properties affecting the lignosulphonate production process and dispersion performance 34
3.3.1 Chemical reactions 34
3.3.2 Wood species 35
3.3.3 Chip size 36
3.3.4 Wood pre-treatment 36
3.3.5 Chip level and movement 37
3.3.6 Temperature 38
3.3.7 Pressure 39
3.3.8 Time 39
3.3.9 Cooking liquor 39
3.3.10 Liquor to wood ratio 40
4 Methodology 42
4.1 A predictive model for concrete slump values 42
4.2 Model development implemented in RapidMiner 45
4.2.1 Data collection 45
4.2.2 Data pre-processing 46
4.2.3 Feature selection and model selection 51
4.2.4 Model training, optimization, and validation 62
4.2.5 Model performance evaluation 63
4.3 Model improvement 64

4.4 Comparative metrics 65
4.4.1 Mean squared error 66
4.4.2 Squared correlation 66
5 Results and discussion 67
5.1 A predictive model for concrete slump values 67
5.1.1 Development of concrete slump function 67
5.1.2 Evaluation of concrete slump function 69
5.1.3 Improved concrete slump function 70
5.2 Model development in RapidMiner 71
5.2.1 Initial concrete slump model 71
5.2.2 Insoluble content 76
5.2.3 Dispersion index 82
6 Conclusion and recommendations 88
7 References 92
Appendix A: Fundamentals of raw materials for pulping 101
A.1 Biological composition of wood 101
A.1.1 Hardwoods and Softwoods 102
A.1.2 Cell wall structure 103
A.2 Chemical composition of wood 105
A.2.1 Cellulose 106
A.2.2 Hemicellulose 106
Appendix B: Selected Models' training and optimization results 108
B.1 Initial Concrete slump 108
B.1.1 Validation results 108
B.1.2 Optimization results 109
B.2 Dispersion Index 110
B.2.1 Validation results 110
B.2.2 Optimization results 111
B.3 Insoluble content 112

B.3.1 Validation results 113
B.3.2 Optimization results 114

List of Figures

Figure 1: The precursors of lignin (Kocurek and Stevens, 1983). 7
Figure 2: In vascular plants, the Phenylpropanoid pathway for lignin biosynthesis (Areskogh, 2011). 8
Figure 3: Basic pulp and papermaking process (International English Language Testing System (IELTS), 2011). 11
Figure 4: Simplified process flow diagram for the manufacture of lignosulphonate from spent cooking liquor of the sulphite pulping process (Aro & Fatehi, 2017). 12
Figure 5: Representation of the structure of a branched lignosulphonate molecule (Areskogh, 2011). 15
Figure 6: Working principle of a Soft Sensor (Curreri et al, 2020). 19
Figure 7: Simple RapidMiner workflow. 21
Figure 8: Key steps of the lignosulphonate production process at Tugela Mill. 24
Figure 9: A continuous digestion process with EMCC (Sappi, 2005) 24
Figure 10: Single effect spray drying process (Ambica Sales Agency, 2016). 27
Figure 11: Lignosulphonate liquor and powder (Sappi, 2016). 27
Figure 12: Example of a Kern Moisture Analyser (John Godrich, 2017). 29
Figure 13: Flow table apparatus (Koehler & Fowler, 2003) 32
Figure 14: Slump measurement (Daily Civil, 2022). 32
Figure 15: Two possible pathways of sulphite delignification (Rydholm, 1965). 35
Figure 16: Concrete slump dataset (created with 34 concrete slump samples). 43
Figure 17: Normalised slump data using approach 1. 43
Figure 18: Normalised slump data using approach 2. 44
Figure 19: Normalised slump data using approach 3. 44
Figure 20: Proposed framework for model development (Rakala et al., 2020; AlBanna, 2016). 45
Figure 21: Insoluble content training and test sets distribution. 48
Figure 22: Dispersion index training and test sets distribution. 49
Figure 23: Initial concrete slump training and test sets distribution. 49
Figure 24: Feature selection workflow implemented in RapidMiner. 54
Figure 25: Representation of the objective of SVM algorithm (Shin, 2021). 55
Figure 26: General decision tree structure (Du & Sun, 2008). 57
Figure 27: Boosting technique representation (Yildirim, 2020). 58
Figure 28: Bootstrap techniques used by Random Forest algorithm (Yildirim, 2020). 60
Figure 29: Neural Network Structure (Geetha & Nasira, 2014). 61
Figure 30: Cross-validation operator in RapidMiner. 63

Figure 31: Normalised concrete slump data with the best fit curve. 68
Figure 32: Actual vs Predicted concrete slump values. 68
Figure 33: Actual vs Predicted concrete slump values from the test set. 69
Figure 34: Actual vs Predicted concrete slump values after bias updating technique. 70
Figure 35: Initial slump prediction on the test set. 74
Figure 36: Actual Vs Predicted values for initial slump using the test set. 74
Figure 37: Initial slump prediction with bias updating technique. 75
Figure 38: Actual vs Predicted values for initial slump with bias updating technique. 76
Figure 39: Insoluble content prediction on the test set. 79
Figure 40: Actual Vs Predicted values for insoluble content using the test set. 80
Figure 41: Insoluble content prediction with bias updating technique 81
Figure 42: Actual vs Predicted values for insoluble content with bias updating technique. .. 81
Figure 43: Dispersion index prediction on the test set. 84
Figure 44: Actual Vs Predicted values for dispersion index using the test set. 85
Figure 45: Dispersion index prediction with bias updating technique. 86
Figure 46: Actual vs Predicted values for dispersion index with bias updating technique. ... 86
Figure 47: Macroscopic view of a transverse section of a tree trunk displaying the fours of the wood: bark (ob &ib), sapwood, heartwood, and pith (P) (Wiedenhoeft, 2010). 102
Figure 48: A simplified structure of a woody cell, displaying the middle lamella (ML) and cell wall layers. (P, S1, S2 and S3) (Côté, 1967) 103
Figure 49: Chemical components distribution in the woody cell wall (Kilian, 1999). 105
Figure 50: Chemical structure of cellulose (Smook,1992). 106
Figure 51: The simplified types of major hemicelluloses in wood (Ingruber, Kocurek and Wong, 1985). 107
Figure 52: Initial slump validation on training set 1 &2. 108
Figure 53: Actual Vs Predicted values for validation performance for initial slump using the training set 109
Figure 54: Initial slump optimization result using training set 1 &2. 109
Figure 55: Actual Vs Predicted values obtained when optimizing initial slump 110
Figure 56: Random Forest prediction of Dispersion index prediction with training set 1. ... 110
Figure 57: Validation results of Actual Vs Predicted values for dispersion index. 111
Figure 58: Optimization of Random Forest model in predicting dispersion index 111
Figure 59: Actual vs Prediction optimization results for selected dispersion index model. . 112
Figure 60: Validation result of the selected insoluble content model. 113
Figure 61: Actual vs Prediction result of the insoluble content model using dataset 2. 113
Figure 62: Optimization result of the insoluble content model. 114
Figure 63: Actual vs Prediction of the optimized insoluble content model 114

List of Tables

Table 1: Hyperparameters for Support Vector Machine learning algorithm. 56
Table 2: Hyperparameters for Decision Tree learning algorithm. 58
Table 3: Hyperparameters for Gradient Boosted Trees learning algorithm. 59
Table 4: Hyperparameters for Random Forest learning algorithm. 60
Table 5: Hyperparameters for Neural Net learning algorithm. 62
Table 6: The concrete slump functions developed. 67
Table 7: Performance results for selected concrete slump function. 69
Table 8: Performance results for tested concrete slump function 70
Table 9: Performance results for bias updated concrete slump function. 71
Table 10: Initial concrete slump performance results for feature and model selection. 71
Table 11: Performance results for the trained initial concrete slump models. 72
Table 12: Neural Net hyperparameters optimized. 72
Table 13: Performance results for optimized Neural Net model. 73
Table 14: Performance results for tested Neural net model. 73
Table 15 : Performance results for corrected Neural net model. 76
Table 16: Insoluble content performance results for feature and model selection 77
Table 17: Performance results for the trained insoluble content models 77
Table 18: Random Forest hyperparameters optimized. 78
Table 19: Performance results for the optimized Random Forest model. 78
Table 20: Performance results for the tested Random Forest models 78
Table 21: Performance results for the corrected Random Forest model 82
Table 22: Dispersion index performance results for feature and model selection. 82
Table 23: Performance results for the trained dispersion index model. 83
Table 24: Random Forest hyperparameters optimized. 83
Table 25: Performance results for optimized Random Forest model 83
Table 26: Performance results for tested Random Forest models 84
Table 27: Performance results for corrected Random Forest model. 87

Nomenclature

Mass (g)	*W*
Solids content (%)	*Solids (%)*
Volume (ml)	*V*
Dispersion Index (dimensionless)	*DI*
Specific gravity(dimensionless)	*SG*
Weight percent	*wt%*
Regularisation parameter	*C*
Correlation	*R*

Subscripts

Average	*avg*
Number of elements	*n*
Model prediction value	*pre*
Corrected predicted value	*cor*
Actual target value	*mea*

Greek

Smoothing parameter	α
Gamma (non-linear SVM) parameter	γ
Intensive loss function	ϵ
Weighting factor	η

Abbreviations

Lignocellulosic feedstock	LCF

Neutral Sulphite Semi-Chemical	NSSC
Strong Red Liquor	SRL
Infrared	IR
European Committee of Standardisation	CEN
European Standard	EN
Gesellschaft mit beschränkter Haftung	GmbH
Yet Another Learning Environment	YALE
Graphical User Interface	GUI
Extended Modified Continuous Cooking	EMCC
High Pressure	HP
Genetic Algorithm	GA
Support Vector Machine	SVM
Mean Squared Error	MSE

1 Introduction

Sustainable economic growth requires the use of safe raw materials for industrial production. Petroleum is currently the most frequently used industrial raw material and is neither sustainable nor environmentally friendly (Kamm & Kamm, 2004). While the energy economy can be based on a variety of alternative raw materials; biomass is, however, more widely available and it is likely to be less expensive and easier to obtain. Thus, biomass provides an alternative source to meet global demands for energy and chemical sources (Areskogh, 2011).

Biomass, particularly plant biomass such as forest and crop residues and woody biomass, is an abundant and carbon-neutral renewable energy resource that has traditionally been used to meet domestic and industrial energy and valuable chemical needs (Kumar & Verma, 2021). To meet the increasing global energy and chemical demands, it has been proposed that a systematic approach for the utilization of biomass as a sustainable alternative resource be developed. This has resulted in the establishment of the biorefinery concept, in which biomass feedstock or by-products of a process, are processed into high-performance products that can often replace non-sustainable oil-based alternatives (Areskogh, 2011).

Biorefining offers numerous advantages, including being environmentally friendly and versatile in terms of product ranges and feed sources. A significant amount of work and resources is currently being invested in biorefinery research, which has resulted in a diverse set of product and process arrangements to consider for future development and implementation. Due to their abundant and favourable feedstock materials and product possibilities, lignocellulosic feedstock (LCF) biorefineries are among the most promising types of biorefineries that can be established (Kamm et. al., 2006: 24).

Their feedstock is plant-based biomass ranging from forestry, pulp, and agricultural waste to papermaking waste such as straw and maize stover; with the majority of lignocellulosic sources used to produce pulp and paper. The LCF has a vast range of feedstock applications since it separates the feedstock into its primary components, cellulose, hemicellulose, and lignin, and allows for specialized reactions to occur for each of the components. As a result, a diverse range of products from traditional petrochemical-based products to innovative bio-based chemicals can be manufactured (Kamm & Kamm, 2004). The perspective role and benefit of biorefineries have prompted several companies in the industry to broaden their business strategy to move to new and adjacent markets, by investing in various biochemical technologies and developing new processes that produce

biomaterials, thereby extracting more value from their feedstock (Matsushita & Yasuda, 2004).

Sappi has adopted the biorefinery concept at their Tugela Mill since 2012 when they began producing lignosulphonate, a by-product of the pulping process. The mill manufactures pulp for internal use and lignosulphonate and containerboard for export. Lignosulphonate is a highly soluble lignin derivative, produced by an advanced lignin extraction process; it was originally considered industrial waste with minimal use. Lignin is a natural wood binding agent that is released during the pulping process (Sappi, 2019). The released material is enhanced by a chemical modification that takes place during the sulphite pulping process, the process of evaporation and spray-drying, to produce lignosulphonate.

Lignosulphonates are used in a variety of industrial and agricultural applications as a dispersing, binding, emulsifying or sequestering agent, as well as a basis for subsequent chemical processing. The lignosulphonate produced at the mill is predominantly used as a dispersant in the concrete and cement admixture market to improve the workability of the cement or concrete mixture. As a result, less water is consumed during the mixing process of the cement or concrete and it reduces the expense of admixture loads.

However, compared to synthetic concrete additives on the market, lignosulphonate has limited quality and performance. Current improvements of lignosulphonates to enhance their dispersant performance, have been focused on filtration and fractionation to obtain lignosulphonates in more specific molecular weight ranges. Quality control systems have also been used to evaluate the consistency and purity of lignosulphonate. While these approaches are sufficient for providing moderate improvements to the lignosulphonate performance, they are not enough to compete as concrete dispersing additives with their synthetic counterparts (Areskogh, 2011). Therefore, it is suggested that more modifications or strategies be established during production to control and produce lignosulphonate with the appropriate competitive properties for the concrete and cement admixture market.

1.1 Problem statement

Unfortunately, there is little research that has been published to understand the effects of the manufacturing process conditions on the final dispersion characteristics of lignosulphonates. Predicting the dispersion capability of lignosulphonate depends in a complex manner on many factors, which have not been properly established yet. Presently, the mill has a quality control management system that is put in place to maintain and align the lignosulphonate

dispersion capabilities to the demands of the concrete and cement admixture market. This process is fully integrated into the operations at the Tugela mill.

The dispersion index and concrete slump test are the two offline methods used to evaluate lignosulphonate dispersion performance. Both tests serve as a benchmark for comparing the lignosulphonate product to competitors. The concrete slump value is significant because it is used to demonstrate that the product's performance does not deteriorate considerably over time.

- The dispersion index is a unitless indicator of the dispersing ability of a cement dispersant. Cement dispersants are a class of ingredients industrially used to reduce the apparent viscosity, by minimizing the use of water while enhancing the rheological properties, of cement slurry (Boughton et al, 1962). A titration method using zinc oxide powder and lignosulphonate samples is used to obtain the lignosulphonate dispersion index (Sappi LQM/BIOSC/M028, 2018).
- The concrete slump test, whereby the lignosulphonate sample is mixed with the standard materials for the concrete mix, is used to determine the workability of the concrete mixture and thus the ease with which concrete flows. The test is carried out by filling the slump cone with the concrete mixture and then slowly and carefully lifting the cone. The diameter of the concrete spread is recorded as the slump. The concrete slump value indicates the degree of workability of the concrete mixture (Sappi W728i023.TUG, 2019). Therefore, this test indicates how well the lignosulphonate performed as a dispersant in the concrete mixture.

The lignosulphonate samples are also analysed to obtain raw data on their solids and ash content, insoluble content, pH, and density. The results of the dispersion index test, the concrete slump test and other lignosulphonate analysis tests, are not immediately available. At the mill, the dispersion index test is performed weekly, while the concrete slump test is performed biweekly. This causes a delay in the process before any corrections can be made to improve the process.

Apart from the time delay caused by relying on the results of the two offline methods used, there are several issues related to the validity and reliability of the results obtained. For the dispersion index test, different lab technicians perform the test; there is subjectivity regarding the endpoint of the titration and the purity of the zinc oxide powder obtained from different suppliers, thereby affecting the results obtained. The concrete slump test is performed using different types of cement and concrete aggregates, and by different lab technicians; thus, affecting the results obtained.

To address the issues associated with the two offline methods in use, the insoluble content of lignosulphonate is currently used as a measure to manipulate the production process thereby, influencing the dispersion performance of the lignosulphonate product. The use of insoluble content to monitor and control the process is based on studies, which show that a high insoluble content in lignosulphonate indicates that less lignin was sulphonated, resulting in a lower lignosulphonate dispersion performance. The lignosulphonate insolubles are the large lignin polymers that were not suitable for sulphonation during the manufacturing process (Klapiszewski et al, 2016).

When a change in the insoluble content is observed, process elements or control parameters are manipulated. The insoluble content of the lignosulphonate product is also a customer concern, and a lower insoluble content is required to ensure customer satisfaction as a purer, cleaner, and better-performing product will be provided. However, a comprehensive set of process parameters that have a direct impact on the insoluble content of lignosulphonate has not been defined.

Efforts to predict the lignosulphonate dispersion characteristics and align them with market demands have thus been a technical challenge for researchers in this field of industry. More work is required to confirm a model to be used and develop process control strategies that will assist in the production of a product of the required quality, while not interfering with the production of Neutral Sulphite Semi-Chemical pulp.

1.2 Research aims and objectives

This study's overall aim is to understand and model the effects of process parameters on the dispersion capabilities of the lignosulphonate, that is produced as a by-product of the Neutral Sulphite Semi-Chemical (NSSC) pulping process at Sappi's Tugela mill. To monitor and control the lignosulphonate dispersion characteristics, soft sensors can be implemented. A soft sensor is a data-driven predictive model that is based on data measured within the plant and accurately provides real-time information required to effectively predict difficult-to-measure variables (Soares & Araújo, 2011), such as the lignosulphonate dispersing capability.

To achieve this overall aim, several objectives are to be fulfilled:

- Identify and isolate the most important process conditions/variables/raw materials that influence the lignosulphonate dispersion capabilities.
- Develop a predictive dispersion capability model based on knowledge of the process and the product, and data analysis using various regression modelling methods and

software. For this study, three predictive models (dispersion index, concrete slump, and insoluble content) will be developed to provide an adequate indication of the dispersing capabilities of the lignosulphonate product.

- Verify and adjust the model, where necessary, to represent the conditions of the process.

The requirements for the developed model, to monitor, control and predict the dispersion capability of the product are:

- It must be accurate enough to allow for effective product quality monitoring.
- It must be simple to be easily integrated into a control strategy while not imposing unrealistic expectations on the control hardware.
- It must be as accurate as possible in reproducing the known behaviours of the process and serving as a useful tool for predicting behaviour.

1.3 Scope of the study

The outcome of this study, the lignosulphonate dispersion soft sensor model, considered all the major processes involved in the production of lignosulphonate at the Sappi Tugela mill. The details of what was taking place in the rest of the mill's pulp and papermaking operations were not given prominence in this study; rather, the focus was on the factors that influenced the lignosulphonate composition, which gives rise to the lignosulphonate dispersion characteristics, and the desired product specifications.

The development of a soft sensor model for the study is a step towards optimizing the lignosulphonate's dispersing capabilities. The insight provided by the developed models can be used to determine what improvements to the Tugela mill operations could be made to improve the lignosulphonate dispersion performance. However, taking note that these improvements may not be the best possible adjustments for the mill's overall operation, as the mill is involved in manufacturing other products, such as corrugating medium and NSSC pulp.

Furthermore, the general concepts and techniques that have been applied for this study, in the selection of variables and the equation form used to develop the models, are generalisable to other industrial operations.

1.4 Significance of study

The development of the soft sensors provides a useful prediction system that can be used to measure and control the dispersion characteristics of the lignosulphonate, thus reducing the reliance on laboratory results and allowing for almost immediate changes to be made to the lignosulphonate production process. Overall, the outcome of this study contributes to the improvement of the product specification and quality, satisfies the customer requirements, and allows for the product to be aligned with the market demands and standards. Furthermore, this study can be used to conduct additional research that can be undertaken to measure and control the properties of other by-products from the Neutral Sulphite Semi-Chemical pulping process.

1.5 Overview of the dissertation

The dissertation describes the steps taken to develop a soft sensor that predicts the lignosulphonate dispersion capabilities based on production process parameters. Several investigative steps had to be completed before the development of the final algorithms. These steps have been organized into chapters in a logical order. Thus, the overview of these chapters is as follows:

Chapter 2 is a review of the related literature required for the study, which includes the product and its applications, production process, and laboratory methods used to test the lignosulphonate dispersion performance. A thorough understanding of the process and methods employed would be necessary for the development of the soft sensor and in the evaluation of the results from the models.

Chapter 3 provides an overview of the Sappi Tugela lignosulphonate production process and the laboratory methods presently used to evaluate the lignosulphonate dispersion capability.

Chapter 4 explains the methodology for soft sensor development. These include historical data collection, data pre-processing, feature selection, learning algorithms selection, model training, validation, and optimization, model performance evaluation, and lastly how were the findings analysed and compared.

Chapter 5 presents the results, along with a discussion of their possible implications.

Chapter 6 concludes the study and provides a summary of the analysis of the results. Recommendations for future work for this study are made, these recommendations include additional modelling work that could potentially improve the models.

2 Literature review

2.1 Lignin

2.1.1 Lignin in wood

Lignin is the second most abundant biopolymer and needs to be removed to turn wood into pulp and paper products. Wood contains between 25 and 35% lignin, although the amount and structure of lignin differ between the types and species of wood. The biological function of lignin in plants is to serve as a bond between the cells and hold them together within the cell wall, providing mechanical support. Lignin is additionally responsible for making the cell walls hydrophobic and protecting the cell wall from any microbiological damage caused by moulds, fungi, and bacteria. It is a very complex, irregular, three-dimensional network polymer that is built upon variously linked phenol and propyl units (Youngs,2011). Wood is not a uniform material; therefore, lignin is not equally distributed throughout the wood. Both the amount and the composition of lignin can differ considerably depending on the species of wood, within the wood species, the growing conditions, and the types of wood cells (Wiedenhoeft, 2010). Additional details on the biological and chemical characteristics of wood are provided in Appendix A.

2.1.2 Lignin biosynthesis

Lignin biosynthesis is made up of three processes which occur within the plant cell: (i) lignin monomer biosynthesis, (ii) transport and (iii) polymerisation. Lignin monomers are the resultant products of the phenylpropanoid pathway beginning with the phenylamine amino acid. After a few steps involving hydroxylation, deamination, methylation and reduction reactions, the lignin monomers are formed in the cytoplasm. These reactions influence which monomers are formed because the monomers differ in how many methoxyl groups are present. The three precursors of lignin are three cinnamyl alcohols: p-coumaryl alcohol, coniferyl alcohol and sinapyl alcohol (Wang, 2011). The precursors of lignin are shown in Figure 1.

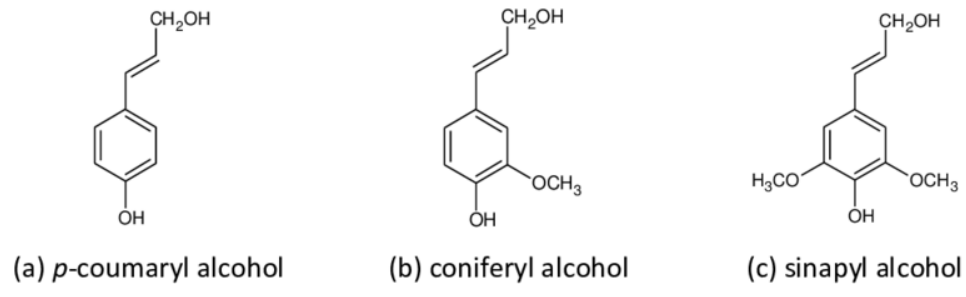


Figure 1: The precursors of lignin (Kocurek and Stevens, 1983).

The lignin monomers are then transferred through the plasma membrane to the apoplast of the plant cell, where they are polymerized. However, little is understood about how these compounds are transported from the cytoplasm to the apoplast. It has been hypothesised that the movement of these small molecules across the cell membrane can occur either by exocytosis or by diffusion and/or transporter-mediated export. Lignin polymerization is initiated by oxidation of the phenylpropane hydroxyl groups, resulting in large and complex lignin macromolecules (Areskogh, 2011).

Lignin polymerisation occurs through simple combination chemistry; lignin polymerisation proceeds mainly due to the 'end-wise' polymerisation process in which lignin monomers are oxidized to phenolic radicals. The oxidised phenolic radicals then undergo cross-coupling reactions with the radicals produced at the free-phenolic ends of the growing lignin polymer. Two distinct oxidative enzymes are present in the plant cell wall, peroxidases, and laccases, and are thought to be responsible for catalysing the phenolic radicals, resulting in the final addition and/or cross-reaction of radical oligomers with the growing lignin polymer(s) (Youngs,2011). The result of the polymerisation is the lignin polymer, shown in Figure 2.

Plant lignins are mainly classified into three classes: annual plant (graminaceous), hardwood (angiosperm) and softwood (gymnosperm) lignin. The lignin varieties are due to the relative abundance of the different monomers in the lignin structure, between species and within species. However, the mechanisms for regulating lignin variation are not yet well established, and neither are the mechanisms for the polymerization of monomers into lignin.

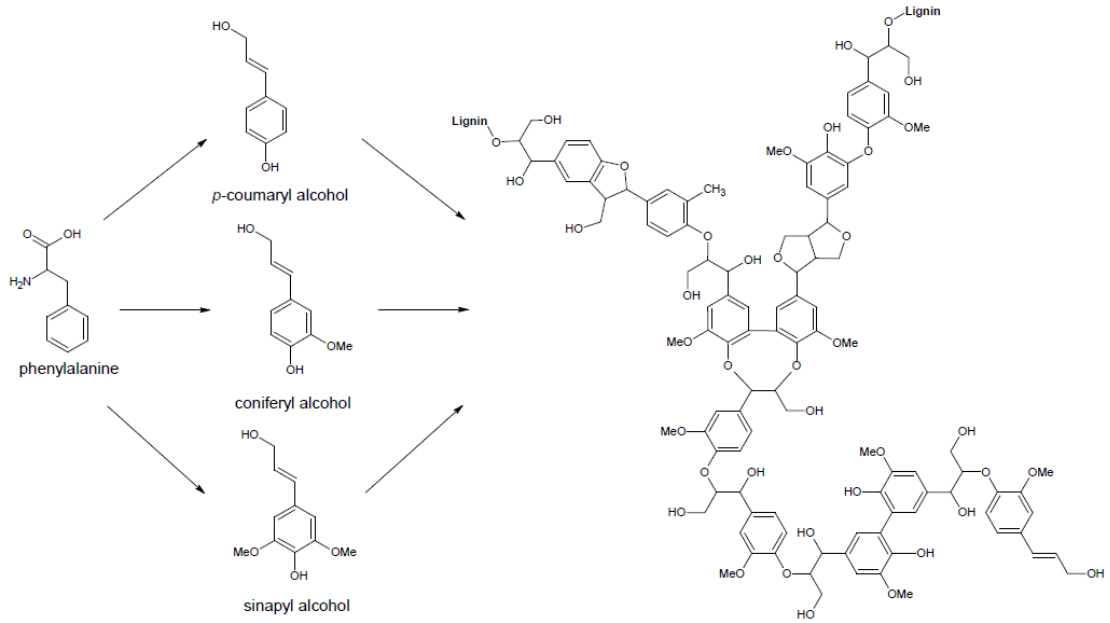


Figure 2: In vascular plants, the Phenylpropanoid pathway for lignin biosynthesis (Areskogh, 2011).

2.1.3 Physical and chemical structure of lignin

Numerous studies have been performed over the years, but the structure of lignin remains unclear. However, what is known about the lignin structure is that, fundamentally, it occurs as an aromatic polymer of an amorphous nature and consists of a heterogeneous three-dimensional network of non-repetitive units and bonding patterns (de Wet-Roos,2016). Thus, lignin has a very high molecular weight (Youngs,2011). As described above, the lignin composition and structure depend on the tree species, the monomers, and their bonds, from which the lignin is polymerised.

The lignin macromolecule consists of several functional groups that influence its reactivity. Lignin mostly consists of phenolic hydroxyl groups, methoxyl groups, and a few terminal aldehyde groups. But its methoxyl group is the most distinctive functional group in lignin. Lignin is responsible for about 90% of the methoxyl content in wood and this group is often used to trace lignin in various connections. The hydroxyl groups are another significant functional group in lignin, accounting for 10% of the total weight of lignin in wood. The complexity of the network structure is further enhanced by the inclusion and abundance of these functional groups in the lignin macromolecule (Wang, 2011).

The bonds between the phenylpropane units and the different functional groups on these units, give lignin its unique and complex structure. Lignin polymer consists primarily of carbon-carbon (C-C) bonds and ether (C-O-C) bonds with β-O-4 bonds. Typically, lignin has one-third carbon-carbon (C-C) bonds and two-thirds ether (C-O-C) bonds with β-O-4 bonds. The dominant bond is the β-O-4 bond. However, the frequency of these different bonds varies depending on the types of wood and are believed to have a significant impact on lignin’s overall reactivity towards the delignification process (Areskogh, 2011).

2.1.4 Potential uses of lignin

The complex lignin structure determines its properties and possible applications, but the properties of the polymer depend on its location within the cell wall (Wang, 2011). Due to its phenolic groups, lignin is insoluble in water but soluble in alkali. It can be made water-soluble by incorporating hydrophilic groups into the lignin structure, such as sulphonated groups. Several applications exist, and they could be identified in the near future. Lignin can be used to generate heat and electricity, used in performance products, as a fuel additive and in syngas. Lignin has the potential to be used in macromolecule-derived products such as carbon fibres, wood binders, polyurethane foams, and aromatics in the future (de Wet-Roos, 2016).

Lignin is also considered to be the only sustainable source for an essential and high-volume class of compounds, aromatics. Lignin can be used to replace the aromatic groups of fossil raw materials in the production of benzene, phenol, and xylene (Holladay et al., 2007). The direct and efficient conversion of lignin into distinct molecules or a low molecular weight class of aromatic molecules is therefore an enticing long-term opportunity (de Wet-Roos, 2016).

2.2 Lignosulphonate

2.2.1 Lignosulphonate production

Wood pulp is the main raw material used in papermaking and is a lignocellulosic fibrous material derived from cellulose fibres in wood, fibre crops and wastepaper. The pulp and paper industry converts wood or recycled fibre into pulp and primary types of paper. The pulping process aims to produce a pulp that removes a significant amount of lignin, while not losing the strength of the fibre, releasing fibres, and removing impurities that will cause discolouration and possible future paper degradation (Liu et al, 2018). Figure 3 depicts the basic pulp and papermaking process, from raw material to finished product.

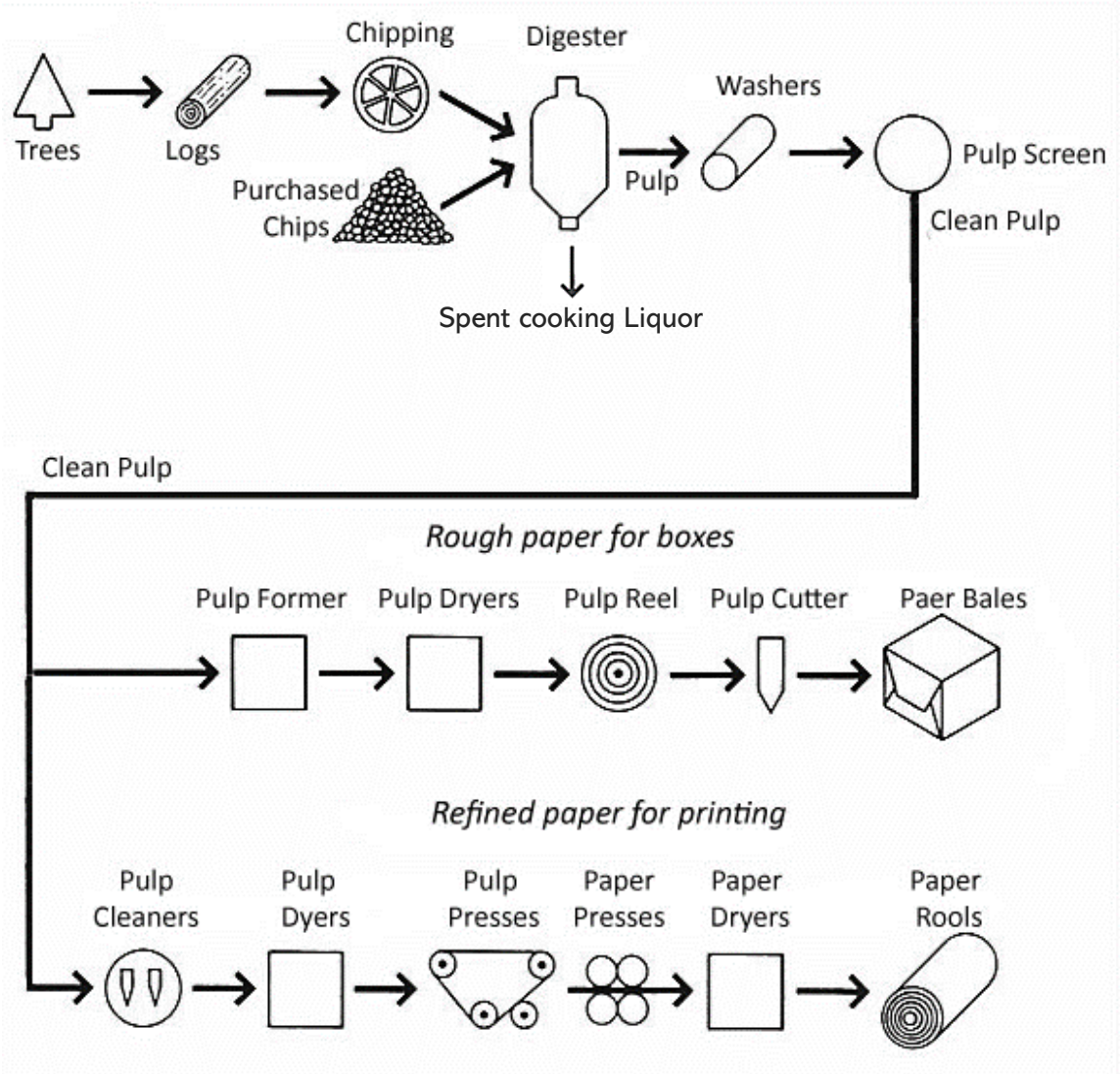


Figure 3: Basic pulp and papermaking process (International English Language Testing System (IELTS), 2011).

Lignosulphonates are sulphonated lignin fragments dissolved in the cooking liquor that can be obtained during sulphite pulping operations or by post-sulphonation of Kraft lignin.

However, most lignosulphonates are produced by the decomposition of the non-cellulose portion of the wood and are then extracted from the spent pulping liquor of sulphite pulping processes. They are by-products of the pulping process and are widely used as biofuel for energy for pulping processes (Hanhikoski, 2014).

Lignosulphonates produced during the sulphite pulping processes, are typically defined by the pH of the process and the bases used. This pulping method can be carried out under different conditions and for different purposes. The conditions of the pulping process give different properties to the lignosulphonate. Lignosulphonates are made using sulphuric acid salts (sulphites or bisulphites) containing magnesium, calcium, sodium, or ammonium at varying pH levels in the cooking process which means that the structure of lignosulphonates differs considerably. Wood chips are digested, under pressure, in a sulphite or bisulphite solution for a cooking period of 4 to 14 hours in a batch or continuous cooking process at 130 to 160 ^{o}C, depending on the pulping chemical used. After the pulping process is complete, lignosulphonates are separated from the pulp by filtration, and the pulp is used to manufacture paper (Sixta, 2006).

The brownish filtrate (spent cooking liquor) contains lignosulphonates, residual pulping chemicals, and hemicelluloses. Lignosulphonates account for 50 to 80wt% of the total solids in the mixture, hemicelluloses account for up to 30wt% and inorganic substances account for approximately 10wt%. Further processing of the filtrate shall require evaporation accompanied by separation (Aro & Fatehi, 2017). After purifying the lignosulphonate, it can be evaporated to a dry matter content sufficient for spray drying and packaging, as shown in Figure 4.

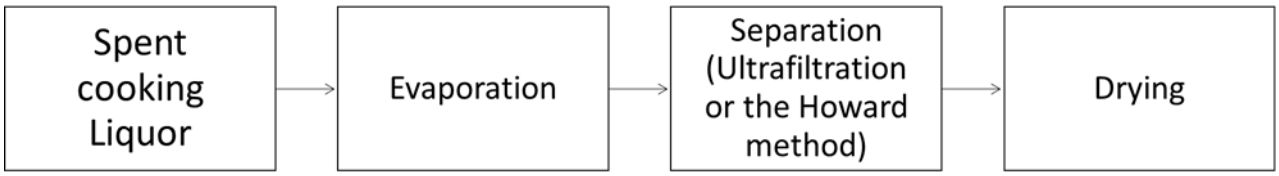


Figure 4: Simplified process flow diagram for the manufacture of lignosulphonate from spent cooking liquor of the sulphite pulping process (Aro & Fatehi, 2017).

To maximise its commercial value, due to the wide range of conditions under which the sulphite pulping process can occur, lignosulphonates should be isolated from the spent liquor. Filtration, evaporation, ultrafiltration, or the Howard method is intended to reduce the content of lignosulphonate impurities to approximately 5% on a dry basis, due to the array of conditions under which sulphite pulping can occur. (Aro & Fatehi, 2017).

Evaporation allows for the water and sulphite content of the spent liquor to be decreased. Subsequently, the filtrate undergoes a purification or separation process to extract lignosulphonates from spent sulphite liquor. As lignosulphonates are water-soluble products, they may not be precipitated by acidification of the spent liquor. The key commercial methods used to isolate lignosulphonates are ultrafiltration and the Howard method. Lignosulphonates have higher molecular weights than other spent liquor components, and this difference allows for good separation (Aro & Fatehi, 2017).

Ultrafiltration is used as a liquid/liquid separation process by which the spent liquor is filtered through a semi-permeable membrane with a certain molecular weight cut-off, such as 20 000 g mol^{-1}. As a result, 40 to 65% of lignosulphonates are derived from sulphite spent liquors by ultrafiltration. However, this approach is not the most economically viable choice for separating lignosulphonates but is the best commercial process currently available. The Howard method is an alternative commercial technique used to recover lignosulphonates. This method is employed when calcium is used as the base during the sulphite pulping process. Calcium oxide is initially added to the spent liquor to precipitate calcium sulphite at a pH of 8.5, which can be filtered and removed. Filtered calcium sulphite can then undergo a pH change and be further filtered to regenerate the pulping chemicals. In the next step, the addition of calcium oxide to the system leads to the creation of calcium lignosulphonate. This approach allows for the recovery of lignosulphonates to be as high as 90 to 95% (Sixta, 2006).

Lignosulphonates are sold in either liquid or powder form. Therefore, once lignosulphonates have been retrieved, the resulting solution may be stored as is or further evaporated to an acceptable solid content at a temperature of 95 to 105^{o}C. Thereafter, the product is spray-dried to a moisture content for the processing of lignosulphonate powder, in accordance with the drying loss specification.

2.2.2 Structural properties of lignosulphonate

Lignosulphonates are sulphonated lignin fragments that have been neutralised and dissolved in the cooking liquor of the sulphite pulping process. However, the structure, properties and uses of lignosulphonate rely on the raw materials, cooking reactants, pulping reactions, and isolation processes. Due to these factors, the structural characteristics of lignosulphonate such as the number of functional groups, degree of sulphonation, molecular weight and available counter ions vary. It is, therefore, necessary to understand how lignosulphonate is produced to ensure that the composition of the lignosulphonate does not differ greatly so that its structural properties and uses are not affected.

Natural lignin is a polymer that has the most varied length, and structure and is hydrophobic, but forms lignosulphonates with unique amphiphilic properties when subjected to the conditions and reactions of the sulphite cooking process. The reactions that take place make it possible for the hydrophilic sulphite and hydroxyl groups, the hydrophobic aromatic structures, and aliphatic chains to form the structural basis for the lignin polymer. Therefore, the phenylpropane structure of lignin is then transformed into a strong anionic ion exchange complex. The distribution of these polar and non-polar groups, including hydroxyl and sulphonic acid groups produced during lignin degradation, are said to provide the properties for lignosulphonate (Antonides, 2000).

The overall lignosulphonate structure is unknown and has been under investigation for several years. However, it can be said that the lignosulphonate structure would be very complex and may have a high degree of crosslinking, due to the complex lignin structure, from which lignosulphonate is a derivative. Several models have been suggested over the years, indicating that lignosulphonates function in aqueous solution as a flexible, coiled or expanded polyelectrolyte (Gupta & Goring, 1958). Lignosulphonates have also been suggested to be made up of a long continuous chain that serves as a backbone with short sidechains.

Side chains can be more branched and joined to the backbone using closed loops. The ether bonds are broken during the cooking stage of the pulping process, due to acidic hydrolysis and the subsequent integration of sulphonate groups into the lignin structure is thought to occur at sites that produce the short sidechains. As a result, the longer backbone is assumed to be more hydrophobic whilst the sidechains are hydrophilic due to the presence of covalently bound sulphonate groups. This arbitrarily branched macromolecule, illustrated in Figure 5, is coiled in the presence of a suitable electrolyte but is loosened and elongated by extracting the salt (Myrvold 2008).

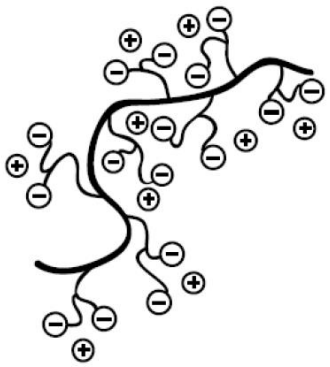


Figure 5: Representation of the structure of a branched lignosulphonate molecule (Areskogh, 2011).

2.2.3 Uses of lignosulphonate

Environmental concern has focused more on the use of plant-derived products or additives such as water-soluble lignosulphonate. Lignosulphonate is an economical material that is easy to use and can quickly become efficient. The use of lignosulphonate extends over a wide range of applications. Lignosulphonate is a multi-functional polymer with antimicrobial, preservative, dispersing, binding, complexing, and emulsifying properties. It has several applications within the food, construction, metallurgical, textile and petroleum industries.

In the food packaging industry, it is used as a de-foaming agent to produce paper and adhesives for products which come into contact with food. It is used as a dust suppressant where it is sprayed on dusty, dirt roads in forestry, rural, farming, construction, and mining areas, to bind the dust particles and create a hard-durable surface (Huang et al, 2019: 13). The dispersing properties of lignosulphonate are suitable for use in the recovery process in oil drilling industries, where they can provide good flow properties as well as be sufficiently stable to allow drilling at high temperatures. They are also used as dispersants for agrochemicals, ceramics, clays, and pigments (Moodley, 2001).

Lignosulphonates' largest use worldwide is as concrete additives and was one of the first dispersants that were used in the cement and concrete admixture industry. They have been used as plasticizers or water-reducing agents since the 1930s and are mixed with concrete or cement to retard the settling time and lessen the amount of water needed. It is also used as a grinding aid for cement, where it is employed to prevent agglomeration and maintain the equipment surfaces clean and clear of debris. Lignosulphonates are also used as binders, in ceramics, urea, pellets for animal feed products, and coal and charcoal briquettes. As lignosulphonates are by-products of the pulping process, it is inexpensive and has a wide range of applications in a variety of industries (Roussel, 2012: 144-208).

2.2.4 Concrete admixtures

Concrete and cement are important components in the modern architectural industry; throughout the history of construction, chemical admixtures have been used in the preparation of concrete. Concrete admixtures are certain products that are added to the primary components (i.e., cement, aggregates, and water) to strengthen or change the concrete properties, compensate for the deficiency of a primary component, or to minimize costs (Wang, 1965). Several concrete admixture types can be used to improve the rheological properties of the concrete mixture.

Concrete admixtures are known to improve the concrete quality, acceleration, manageability, retard the settling time or reduce water consumption for concrete preparation, among other properties that could be altered to get specific results. Examples of chemical admixtures include calcium chloride, calcium sulphate, lignosulphonates, and sodium hexametaphosphate. The properties of concrete are characterised by its flow behaviour and mechanical strength. In most cases, the flow behaviour is governed by the dispersion of cement particles, while the mechanical strength is determined by the ratio of water to cement during concrete preparation (Rodriguez, 2019).

Lignosulphonates are mostly used as dispersants and concrete additives (plasticisers) due to their appropriate molecular weight and anionic charge density, stemming from the presence of certain functional groups (Aro & Fatehi, 2017). Plasticizers or dispersants are organic, or a mixture of organic and inorganic substances used to enhance the plasticity or fluidity of a given material; making it durable, flexible, and easier to handle. Plasticity in polymeric materials such as PVC is enhanced when plasticizers are added, whereas plasticizers in concrete mixtures, minimize the water content for a certain degree of workability. A lower water content results in stronger concrete which is more impervious to water penetration (Ouyang, Oui & Chen, 2006).

In the concrete mixtures, cement particles are dispersed due to electrostatic repulsion force, caused by the formation of a charged layer on the cement grain surface, which has adsorbed the plasticizer molecules, thereby improving the plasticity and workability of the overall concrete mixture (Nagrockiene, Pundiene & Kicaite, 2013). Lignosulphonate has been continually used in the cement and concrete admixture industry as a chemical additive, here they provide several advantages improves the strength of concrete by reducing the water to cement ratio, delaying the settling time, reducing water consumption during the mixing process, improves the workability of the concrete mixture by dispersing the concrete particles and ultimately reduces the expense of admixture loads (Areskogh, 2011).

Despite the fact that cement and concrete have been widely used since the industrial revolution, the mechanisms involved in cement setting and concrete preparation when combined with admixtures are still only partially understood. It is known that the mechanisms for the functioning of lignosulphonates as dispersants in concrete and cement mixtures are through electrostatic repulsion and steric hindrance between the individual cement particles. As lignosulphonates are anionic surface-active polymers, these mechanisms are exerted (Ogbonna, 2009).

Adsorption of lignosulphonate molecules onto the surfaces of cement particle surfaces results in steric hindrance and electrostatic repulsion of the lignosulphonate molecules. This prevents the flocculation of cement particles by promoting their homogeneous dispersion in freshly prepared concrete. Steric repulsion inhibits particle flocculation by keeping particles apart, whereas electrostatic repulsion is caused by the presence of charged groups in the lignosulphonate structure. This mode of action by dispersants was initially described by Uchikawa, Hanehara and Sawaki (Uchikawa et al, 1997) and is broadly accepted. It is understood that the molecular weight and sulphonic group content are the main factors which affect the properties of lignosulphonate and its ability to be used as a dispersant in concrete. Previous studies linked these key factors to the mechanisms involved in the dispersion of lignosulphonates in concrete and cement admixtures. The lignosulphonate molecular weight was shown to influence steric hindrance, while lignosulphonate sulphur content affected electrostatic repulsion forces between the lignosulphonate structure and cement particles (Aro & Fatehi, 2017).

Lignosulphonates are often blended with other plasticisers such as polycarboxylate ethers (PCEs) to enhance the dispersing capacity of lignosulphonates in concrete and cement admixtures and reduce environmental harm and expense. Superplasticisers, such as PCEs, are mainly based on two classes of non-renewable petrochemicals and are produced according to certain requirements which will result in stronger dispersing capabilities than lignosulphonates. Superplasticizers have a high number of charge groups and a substantial molecular weight, which increases repulsive forces by assuring a continuous surplus of charged groups present on the molecule regardless of the number of active sites on the cement particles, hence they can be better dispersants (Matsushita & Yasuda, 2005).

The potential and continuous use of lignosulphonates to function as plasticizers, dispersants or water-reducing agents in the concrete and cement admixture industry is related to its low cost, the need for greater workability of concrete, its composition, which may vary due to reaction conditions, type and age of wood, and its contribution to the improvement of the

concrete properties. However, improvement of molecular weight, purity and charge group content within the macromolecule will further enhance the plasticizing effect on concrete.

2.3 Soft sensors in the process industry

On the topic of process control, there is an area of study centred on the construction of plant models that try to characterize or explain the plant's operational behaviour. While the instrumentation area is a natural complement to process control, there is one component of plant control that demands more than basic instrumentation to achieve its goals. This process control field is concerned with the development of soft sensors that use software and current plant instrumentation to offer either a new process variable or a degree of higher-level information on the plant's state (Harker, 2013).

Several variables are monitored by online sensors while dealing with plant operations, however, some of these variables are difficult to measure or can only be measured infrequently due to high cost or a shortage of sensors. In other instances, the measured variables have long delays due to slow hardware sensors or laboratory analysis, making real-time monitoring of the process impossible (Curreri et al, 2020). Inferential models, such as soft sensors, can be constructed to estimate critical quality factors based on measured online data, eliminating the need for expensive and time-consuming measurement tools or laboratory analysis of the product (Zhu et al, 2020). The second output of a soft sensor is one of plant state. In the process control industry, the concept of a state is well understood. Alarming of plant variables is a critical activity that converts a process variable into an alert state, allowing for monitoring of the plant process and making modifications or enhancements to the process before costly and harmful occurrences occur (Harker, 2013). This usually occurs in an operator monitoring interface system. Figure 6 depicts the basic operation of a soft sensor.



Figure 6: Working principle of a Soft Sensor (Curreri et al, 2020).

A soft sensor has a fast response time since the required response variable prediction is accessible as soon as the input values are gathered. It supports real-time estimates, allowing for the application of strict control policies, and serves as a low-cost alternative to

expensive hardware devices. If the appropriate input variables can be measured online using current instruments, no additional instruments are required (Curreri et al, 2020). When new data, such as a laboratory value, becomes available, the soft sensor model is frequently updated. To ensure that the model's high performance and resilient stability are maintained, several strategies for updating the soft sensor can be implemented. Several adaptive methods for soft sensor updating have been proposed, such as adding a bias (Zhang et al, 2019), a Kalman filter, or using a means and variance update strategy (Wang & Chiang,2018).

These adaptive approaches are used for soft sensor design and are used to periodically correct soft sensor estimates, account for model mismatch and improve prediction accuracy in the presence of noise, disturbances, and variations in operating circumstances (Quelhas & Pinto, 2009). These strategies may help ensure that the model's accuracy is preserved when the soft sensor degrades after a period of online operation due to a change in process dynamics. The use of soft sensors to forecast difficult-to-measure variables such as the lignosulphonate dispersion characteristics, bridge observation gaps, and prevent extra delays until laboratory results are available. As a result, this is the initial step in attempting to successfully monitor and manage the process to optimize a product’s performance, and it serves as the foundation for any subsequent control strategy.

2.3.1 RapidMiner

Data can be analysed using RapidMiner software, which can also be used to create and validate soft sensor models. RapidMiner is an open-source software platform that provides a straightforward integrated framework for data mining, machine learning, text mining, predictive analytics, and business analytics. It is used in business and industry, as well as in research, education, training, prototyping, and application development (Mierswa, 2006). The application, formerly known as YALE (Yet Another Learning Environment), was developed at Germany’s University of Dortmund. RapidMiner Studio, in particular, is software with a graphical user interface (GUI) that allows predictive analytics and data mining workflows to be designed and deployed (Milovic & Milovic, 2012).

Soft sensors could be developed using data mining techniques. Data mining is the extracting of hidden knowledge from databases. It is a versatile technology with a wide range of applications for analysing and forecasting critical information from databases. As a result, data mining techniques have a level of trust in the projected answer in terms of prediction consistency and frequency of correct forecasts. Trends and patterns that are difficult to gain an understanding of using traditional data analysis methods can be uncovered using data mining techniques and machine learning (Ananthapadmanaban & Parthiban, 2014).

Finding trends and patterns aids in forecasting and decision-making. Due to the huge amount of information involved in dealing with plant processes, machine learning is advantageous, as traditional data analysis methods are time-consuming and difficult for people to forecast effectively (Milovic & Milovic, 2012). Data mining has the potential to be both informative and predictive. Predictive data mining is supervised, with a special label or goal variable. It exhibits predictive trends and patterns, such as categorization and regression.

RapidMiner simplifies the entire data mining process, from raw data pre-processing and visualization to validation, optimization, evaluation, and deployment (Miner et al, 2012). All data mining steps in RapidMiner are structured as operator trees, and a typical process workflow consists of numerous operators. An operator is a piece of code that is contained and performs a specific task. This data mining task can be any of the following: importing a dataset, pre-processing it by cleaning and removing misleading samples, lowering the number of attributes by using feature selection techniques, training, and optimizing prediction models, or scoring new datasets using models established previously (Kotu & Deshpande, 2015). The data flow is specified by the connection of ports; for example, in Figure 7, the ‘Retrieve' operator retrieves a dataset from a repository and provides it to the following ‘Select Attributes' operator. The data flow is consistent and adheres to the depth-first search principle. This simplifies the design of the data mining process.

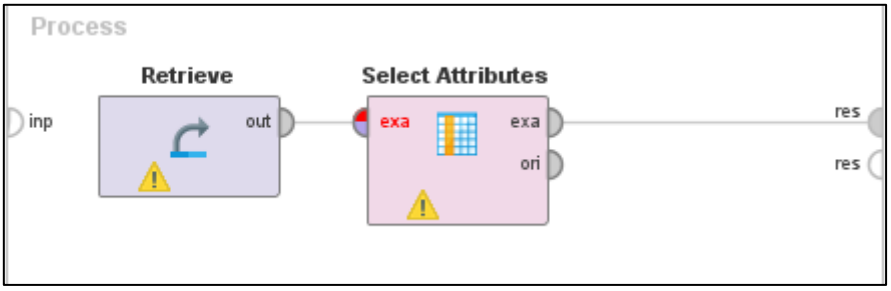


Figure 7: Simple RapidMiner workflow.

RapidMiner allows you to create classification and regression predictive models from a dataset. There are several machine learning techniques available, and the expected output of the trained machine learning model is based on the value to be predicted, which can be categorical/nominal or numerical. Several statistical performance criteria for the modelling work, such as correlation, relative error, and root mean squared error, are offered for the modelling task, allowing performance operators to evaluate model performance precisely and appropriately during model development. Thus, reaching the intended performance can

be a continuous activity by monitoring the model's performance as modifications to the process workflow used to generate the model are implemented (Mierswa, 2006).

3 Sappi Tugela plant

Sappi Tugela Mill is in the town of Mandeni, near the Tugela River in Kwa-Zulu Natal. The location of the mill has made it easier access to the global markets by means of Richards Bay and Durban ports. The mill is currently the oldest purpose-built mill in the Sappi group; it has been in operation since 1945. The Sappi Tugela Mill uses the Neural Sulphite Semi-Chemical (NSSC) pulping process to manufacture pulp for its consumption, containerboard (corrugating medium) and lignosulphonate for export (Sappi,2016). This type of sulphite pulping works well for the variety of products produced at the mill, from the wood species used at the plant.

Yearly, the mill produces 150 000 tonnes of Neutral Sulphite Semi-Chemical pulp for its own use and 200 000 tonnes of corrugated medium, made from virgin and recycled fibres. After joining Sappi Biotech, the Tugela Mill became a source of lignosulphonate in 2012. Annually, the mill produces 25 000 tonnes of lignosulphonate powder and 35 000 tonnes of liquid product. Also, approximately 60 000 tonnes of Refibre are produced annually for the mill's own use (Sappi, 2016).

The mill holds a key role in leading packaging innovation in South Africa and is the only high-performance containerboard packaging mill in the country. Their lignosulphonate product is known to be used in several markets such as the manufacture of concrete and cement admixtures, the manufacture of clay bricks/ceramic tiles and the suppression of road dust (Sappi, 2016). The addition of a spray dryer to the lignosulphonate production process in 2015, has allowed the mill to broaden its product footprint as liquid lignosulphonate can be transformed into a powder format.

3.1 Overview of the process

The NSSC pulping process takes place in two stages of chemical treatment followed by mechanical treatment of the pulp for the extraction of cellulose fibres and delignification of the wood. The key processes involved in the production of lignosulphonate at the Tugela Mill are summarized in Figure 8.

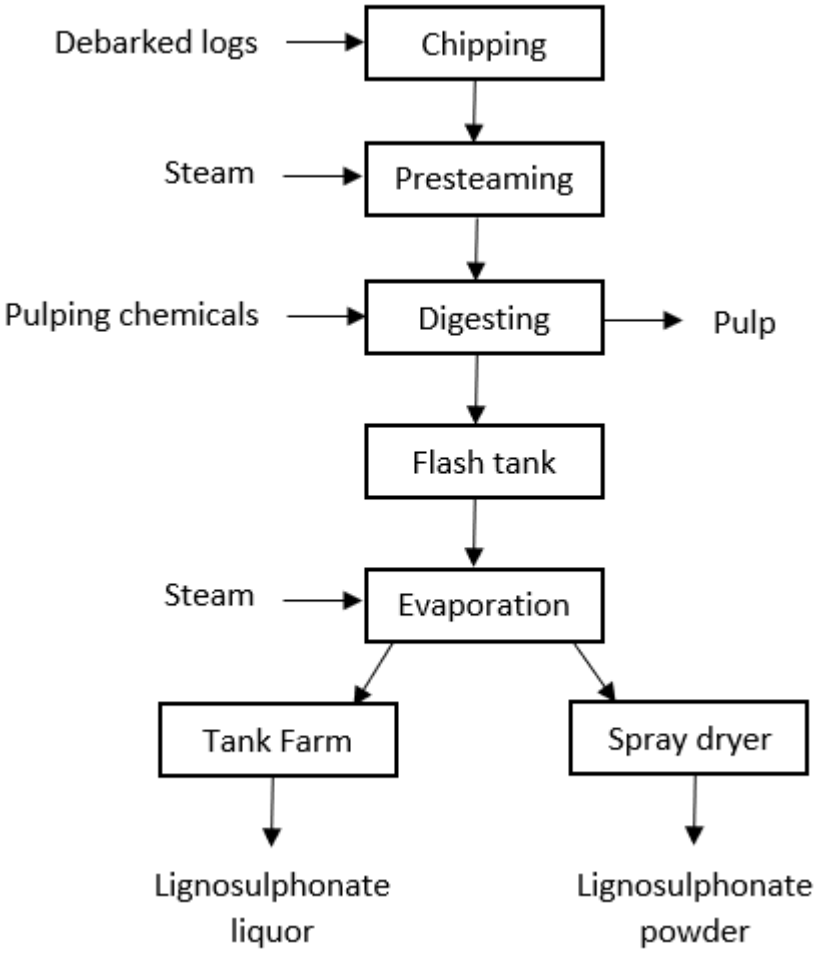


Figure 8: Key steps of the lignosulphonate production process at Tugela Mill.

3.1.1 Raw material preparation

Debarked hardwood logs obtained from nearby plantations in the region, which have been sawed to the appropriate size, are used. A mixture of Eucalyptus wood is used at the mill, namely Eucalyptus Grandis, Nitens and Dunnii wood. Debarking the logs is necessary because the bark absorbs chemicals and remains as dark specs in the pulp, which can be difficult to remove. (Watson & Potter, 2004).

Lightly burnt wood is often used because the wood is exposed to heat and has undergone chemical changes due to the degradation of hemicellulose and lignin. As a result, the wood has more dimensional stability and greater resistance to fungal decay leading to faster cook and less washing of the pulp. During the debarking process of the burnt wood, all charcoal must be removed, and the wood must be used as soon as possible to maintain the moisture content of the wood (Rust, 2015).

The logs are then fed through a woodchipper to reduce the size of the wood. The chipper produces chips with a thickness of 2- 8mm, to ensure a steady flow through refiners and uniform cooking in the digester. The thickness of the chips is the most significant parameter, as it defines the speed and thoroughness of the impregnation of the cooking chemicals in the wood chips. Chips generated from the chipper are stored temporarily in chip piles for approximately 2.5 to 3 hours before their use. Before the chips are fed into the digester, the chips from the chip piles are screened for size and cleaned. Vibrating screens are used to remove undersized and oversized chips.

3.1.2 Wood pre-treatment

The wood chips used for pulping at the mill are not dried in a dryer before it is fed into the digester. The wood chips have approximately 35% moisture, and this moisture is part of the system that feeds into the digester. Air and moisture in the chips are displaced with cooking liquor. The air in the chips is first removed so that the initial penetration of cooking liquors will proceed more rapidly. To remove air before immersing the wood chips in the cooking liquor, they are first pre-steamed in an impregnation vessel, also known as the steaming vessel, to soften them and drive out any trapped air (Killian,1999). A surfactant is also added to ensure effective liquor impregnation, which is essential for satisfactory delignification. The distribution of the wood species fed to the digester at the Tugela Mill is estimated.

3.1.3 Digestion

Wood chips are digested under high temperatures and pressures, in a neutral cooking solution of white liquor and sodium carbonate, in a continuous-type cooking process. Cooking liquor, also known as white liquor, contains two active chemicals, sodium hydroxide and sodium sulphite. The digester at the Tugela Mill is a large, vertical pressured tube, divided into three zones: the impregnation zone (I), the cooking zone (II & III) and the counter-current washing zone (IV), as shown in Figure 9. The digester is a Kamyr continuous digester that employs the extended modified continuous cooking (EMCC) technique, which divides the cooking zone into current and counter regions and distributes the white liquor to several sites. Using such a cooking process has many benefits for pulp and papermaking operations at the mill (Mollereau, 2005).

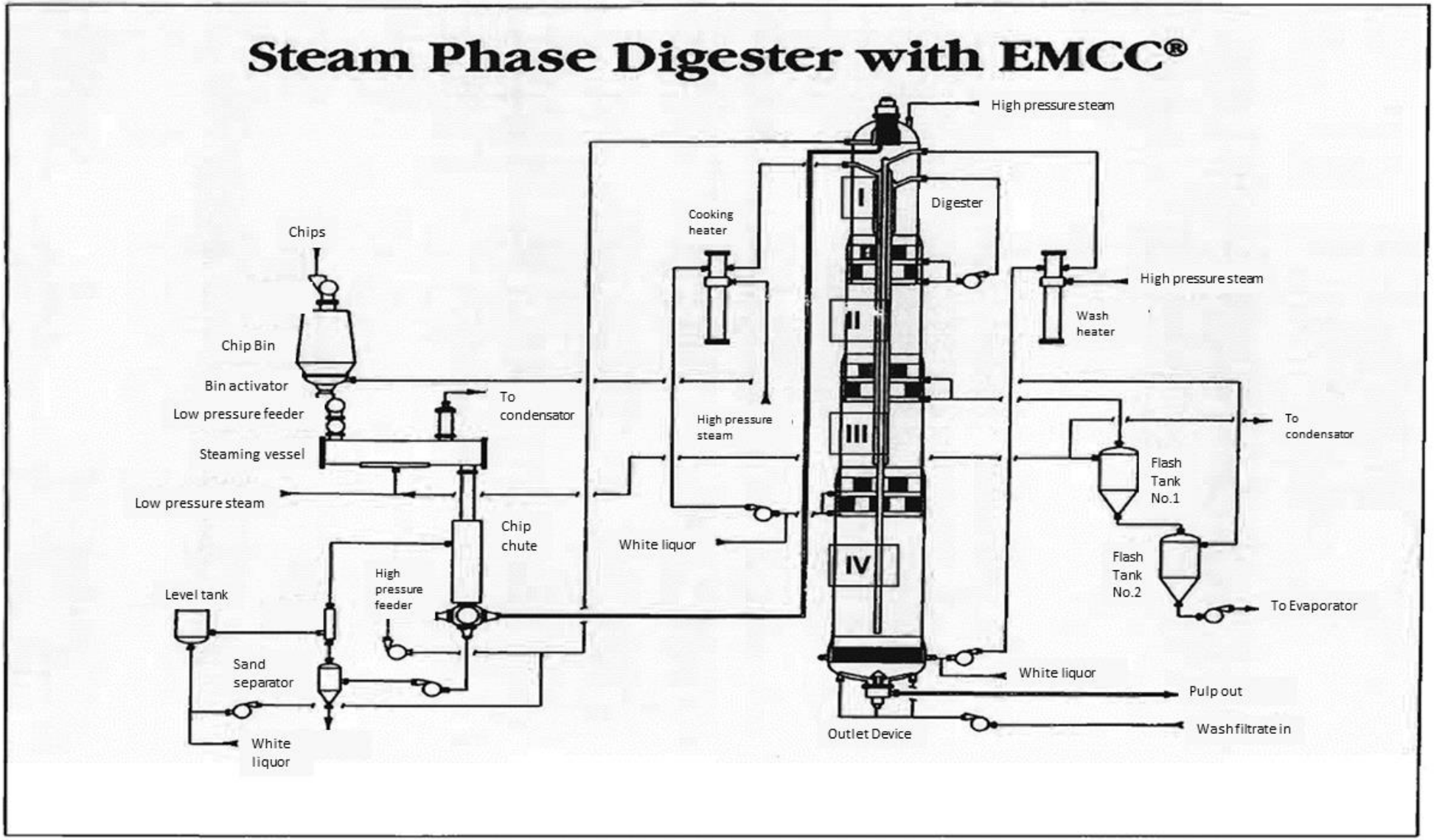


Figure 9: A continuous digestion process with EMCC (Sappi, 2005)

Chips, white liquor, and sodium carbonate are fed into the top of the digester (Impregnation zone, Figure 9). The pressure and temperature at the top of the digester are maintained by a high-pressure steam stream. The chip supply flow must be considered when adjusting the pressure. Thus, ensuring the effective penetration of liquor into the chips is accomplished before rapid delignification begins. In this zone, the chips are heated to 105 - 130 ˚C. It takes about 20 to 25 minutes for the chips to pass through this zone and into the cooking zone.

The liquor and chips then enter the first cooking or heating zone (Zone II, Figure 9), where the chips are rapidly increased to the cooking temperature of 165 to 175 ˚C. This is accomplished by liquor circulation systems, whereby liquor is extracted from this section through screens and circulated through heat exchangers, and then the heated liquor is pumped back into the digester. This is a phase of the digester, this is done twice, with high-pressure steam as the heating medium. Liquor can be extracted through three sets of screens that are along the length of the digester.

The chips and liquor move into the second cooking zone (Zone III, Figure 9). This is where the majority of the delignification occurs and where a maximum cooking temperature of 165 to 175 ^{o}C is maintained (Area *et al*, 2001). The delignification reactions rely on sulphite content, temperature, and time. The sulphite ions in the cooking liquor react with the lignin, and the sodium carbonate serves as a buffer to maintain a neutral cooking liquor and slow down the delignification process.

During the process, lignin is broken down by sulfonation and hydrolysis reactions, which degrades the randomly distributed ether bonds throughout the lignin structure. Sulphonation softens the lignin and makes it more hydrophilic, while hydrolysis breaks the lignin bonds, resulting in the formation of new and smaller soluble lignin fragments. Sulfonation reactions are limited to β-O-4 ethers with free phenolic hydroxyls under neutral sulphite pulping conditions. The net effect of these reactions is therefore light sulphonation of hydroxyl groups on the alpha-carbons in the structural units of the lignin polymers and some cleavage of the lignin β-O-4 ether bonds (Antonides, 2000). When half of the lignin is dissolved, the delignification process slows down, approximately after cooking the wood chips in for 60 to 70 minutes, in the digester cooking zone.

To achieve an up-flow of counter-current diffusion washing, the spent cooking liquor (weak red liquor) is extracted at high pressure from the middle set of extraction screens and sent to flash tanks. This acts to lower the temperature and thereby terminate the pulping reaction. The cooked chips enter the washing zone, where they are partially washed and pumped to a blow unit. The spent cooking liquor contains leftover pulping chemicals, and wood organics

such as hemicelluloses, lignosulphonate, and water (Sixta, 2006). This liquor has a solid content of approximately 13-15%.

3.1.4 Spent liquor recovery

The spent cooking liquor is concentrated to about 48% solids. The concentrated liquor is the lignosulphonate product and is recovered by evaporation. Evaporation is carried out using 3 effects in series, and this method is the most effective way to produce higher liquor solids. However, the limit is set by the rise in viscosity and boiling point that occurs as the solid content increases. The concentrated red liquor (strong red liquor) produced consists of 80% organics, most of which is high molecular weighted lignosulphonate, and 20% inorganics (Aro & Fatehi, 2017). Lignosulphonates are sold in liquid or powder form.

After evaporation, the strong red liquor is screened by using decanters, to ensure that the product produced is purer and cleaner. Thereafter, the screened strong red liquor may be stored as is or undergo further processing to produce the lignosulphonate powder. The lignosulphonate powder is produced by heating the strong red liquor to a certain temperature, and the solids content of the liquor is monitored in such a way that the solids content of the liquor is greater than 48%.

This step is essential to ensure successful atomization, and to obtain uniform particle size and free-flowing lignosulphonate powder. Thus, this step affects the viscosity which, in turn, affects the particle size and drying rate of particles. The liquor is then sent to a single effect spray dryer, shown in Figure 10. Evaporation of the liquor is achieved with the drying chamber. The liquor is sprayed through a nozzle which forms small, atomized droplets, into hot air forming dry, free-flowing lignosulphonate powder with a consistent particle size distribution (Aro & Fatehi, 2017). The solids from the dryer are collected in a cyclone, and the resulting lignosulphonate powder has a solids content of 95%.

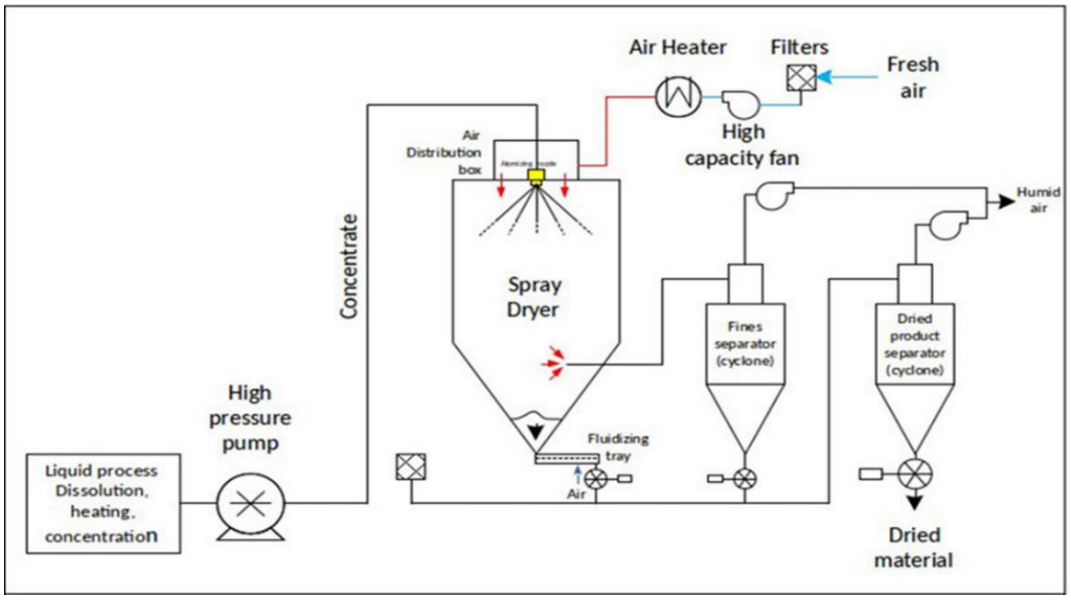


Figure 10: Single effect spray drying process (Ambica Sales Agency, 2016).

The reddish-brown, viscous high lignin content lignosulphonate liquor, obtained after the evaporation process and the brown, fine lignosulphonate powder are the products of the lignosulphonate production process at the Sappi Tugela Mill, shown in Figure 11.



Figure 11: Lignosulphonate liquor and powder (Sappi, 2016).

3.2 Techniques to monitor lignosulphonate dispersion performance

Predicting the dispersion performance of lignosulphonate from process parameters depends on many complex factors, which have not been properly established yet. Currently, the dispersion capability of the lignosulphonate produced at the Sappi Tugela mill is controlled and monitored by offline analysis. This process is fully integrated into the Tugela mill

operations. Two laboratory methods are used to test the lignosulphonate dispersion performance: the dispersion index and the concrete slump tests.

The concrete slump and dispersion index tests are important to showcase the dispersion capabilities of lignosulphonate to the target market. The dispersion index value is an indicator of the dispersing ability of the lignosulphonate. The concrete slump value is important as it is a way to compare product performance amongst competitors and show the customer that the performance of the product does not degrade considerably over time. In industry, the reduction of the concrete slump from the time of original batching to the point when concrete is discharged from the truck mixer or delivery vehicle should be low so that the workability of the fresh concrete or consistency of the concrete is maintained while the concrete mix is delivered to the construction site (Sappi W728i023.TUG,2019).

The insoluble content of the strong red liquor is currently used to try and manipulate the NSSC pulping process to influence the dispersion performance of the lignosulphonate product. At the mill, some elements or control parameters of the digester are manipulated when a change in the spent liquor insoluble content is observed. The theory behind the use of the insoluble content to monitor and control the process is based on studies which have indicated that, when lignin is sulphonated during the cooking process, lignin condensation reactions can occur rendering lignin insoluble in the pulping process because bigger lignin molecular complexes form that is not easily dissolvable, thereby retarding the delignification process (Moodley,2001).

Therefore, high insoluble content in the lignosulphonate product indicates that less lignin was sulphonated, leading to lower lignosulphonate dispersion performance, indicated by the results of the concrete slump and dispersion index tests. The insoluble content of the lignosulphonate product is a customer issue; to ensure customer satisfaction, lower insoluble content is required which leads to a purer, cleaner, better performing product.

3.2.1 Tests

At the mill, the dispersion index is performed weekly and approximately 30 minutes is required to perform the test. The concrete slump test is performed every second week and approximately 60 minutes is required to conduct the test. The results are captured on the Sappi Laboratory Information Management System, where the results of the tests reflect the time, the sample was taken. The insoluble content of the liquor samples is determined every two hours, and the test is completed in approximately 20 minutes. The results of this test are entered into the laboratory information system within the hour of sampling. The laboratory

methods described below are used to evaluate the dispersion performance of lignosulphonate powder samples and the insoluble content of strong red liquor samples.

3.2.1.1 Solids content determination

Before any of the above-mentioned tests may be conducted, the solid content of the lignosulphonate powder or liquid samples is evaluated. This procedure takes 15 minutes to complete and is considered sample preparation because the results are required to complete the methods used to assess the lignosulphonate dispersion capabilities.

Initially, extra pure sea sand is dried in a desiccator after being dried in an oven at 110°C. The IR instrument must be levelled before use, thereafter, an aluminium plate in the three-legged stand is placed. The dry sand is added to the plate, and subsequently, press tare until the mass reading is 0.0. Approximately 1.0g ±0.5g of lignosulphonate liquor is placed onto the dried sand using a 10ml syringe. A spatula is used to dispense roughly 1g of powder onto the dried sand, for powder samples.

Once the samples have been dispensed onto the dried sand, the IR instrument is ready to use. Depending on the KERN Moisture balance IR instrument, such as Figure 12, used to complete the analysis, ensure that it is running on the programmed mode and that the temp/time setting is set to 110C/AFREE. After the measurement, the instrument will come to a halt on its own. Then record the percent solids measurement and remove the aluminium plate containing the sample (Sappi W728i005.TUG, 2017).



Figure 12: Example of a Kern Moisture Analyser (John Godrich, 2017).

3.2.2 Dispersion index

The dispersion index test is performed using a 3.0% solution prepared with the lignosulphonate powder. Dilution is achieved by using the %solids content from the method above, and a chosen 40.0 g of the diluted solution at 3.0% solids. Equation (1) is used to determine the amount of lignosulphonate powder to be added.

$$W_1 = \frac{120}{Solids(\%)} \tag{1}$$

where W_1, is the amount of lignosulphonate powder to be added in grams. Once the amount of lignosulphonate powder is determined, the powder is accurately weighed into a 50ml centrifuge tube. Distilled water for dilution is then added to the tube so that the overall mass within the tube is 40g; ensure that the mixture is adequately mixed to fully solubilise the powder.

The dispersion index test is performed using zinc oxide; 6g of zinc oxide powder is accurately weighed into cups of a baking tray. A burette is filled with 10ml of the prepared solution. Deionised water (2ml) is then carefully added to the ZnO powder. The mixture is left for a minute, to allow the water to soak before mixing the powder thoroughly with a spatula. The lignosulphonate solution is then added dropwise from the burette and blended with the powder after each drop. This is continued until the ZnO slurry/solution drops freely from the spatula, indicating the endpoint. The change in volume of the burette is recorded. The measurement is done in three replications for each sample. The dispersion index is calculated as follows:

$$DI = \frac{144}{V_{avg}} \tag{2}$$

where DI is the dispersion index of the sample and is dimensionless, V_{avg} is the average of the change in volumes recorded for the three runs and 144 is an empirically derived value from the setup of this method (Sappi LQM/BIOSC/MO28,2018).

3.2.3 Concrete slump

The concrete slump test is carried out in four main steps: preparation of the dry mixture, preparation of a 40% lignosulphonate solution, preparation of the concrete mixture and the procedure to perform the slump test on the vibrating table.

Firstly, the dry mixture is prepared by weighing out some reagents into a mixing bowl/agitator. Approximately 1350g of CEN Standard Sand EN 196-1, 728g of cement and 485g of calcium carbonate are weighed out, added into a mixing bowl, and mixed with a

spatula. CEN Standard Sand EN 196-1 is a product of Normensand GmbH and has been used worldwide in the cement industry to test the strength of cement according to cement quality standard EN 196-1 (Normensand GmbH, 2022). The mixing bowl is then attached to a mixer, and the sample is mixed in such a way that a homogeneous mixture is formed. This test requires lignosulphonate samples at a 40% solution therefore, a dilute solution from lignosulphonate powder is made. Dilution is achieved by using the solids content, and a chosen 100ml of the diluted solution at 40% solids. Equation (3) is used to calculate the amount of lignosulphonate powder to be added.

$$Mass\ of\ lignosulphonate\ powder\ (g) = \frac{4000}{Solids(\%)} \qquad (3)$$

Once the mass of lignosulphonate powder to be used is determined, the powder is accurately weighed into a beaker. Approximately 50ml of water is used to dissolve the powder. The dissolved solution is transferred to a 100ml volumetric flask and makeup to volume with distilled water.

The concrete mixture to be tested is prepared by weighing out 218g of water and 1281.5g of the dry reagent mixture into an agitator vessel. The agitator vessel is then fixed to a mixer and mixed for 1 minute on low speed, thereafter, mixed for 1 minute on high speed. Gritting material, 377g, is added to the mixture and mixed on high speed for 1 minute. Approximately, 2.3ml of the prepared 40% lignosulphonate solution is added to the mixture. This mixture is then mixed at high speed for 1 minute.

The slump test can be performed after the concrete mixture has been mixed. A cone with a feeding hopper and a flow table is required, as shown in Figure 13. The concrete cone is approximately half-filled and compacted. After filling the cone to the top, the excess concrete mixture is removed. The vibrating flow table's level/hand wheel is turned 30 times, and measurements are taken. The diameter of the concrete spread (slump) is measured, as shown in Figure 14, and recorded in mm (initial slump recorded at 0 minutes). To calculate the slump loss, the slump is measured at 15-minute intervals on the concrete mixture over one hour. As a result, the concrete mixture is stored in a closed vessel, and the test procedures are repeated for 15, 30, 45, and 60 minutes using the same concrete mixture. Before testing, ensure that the mixture is homogeneous (Sappi W728i023.TUG,2019)

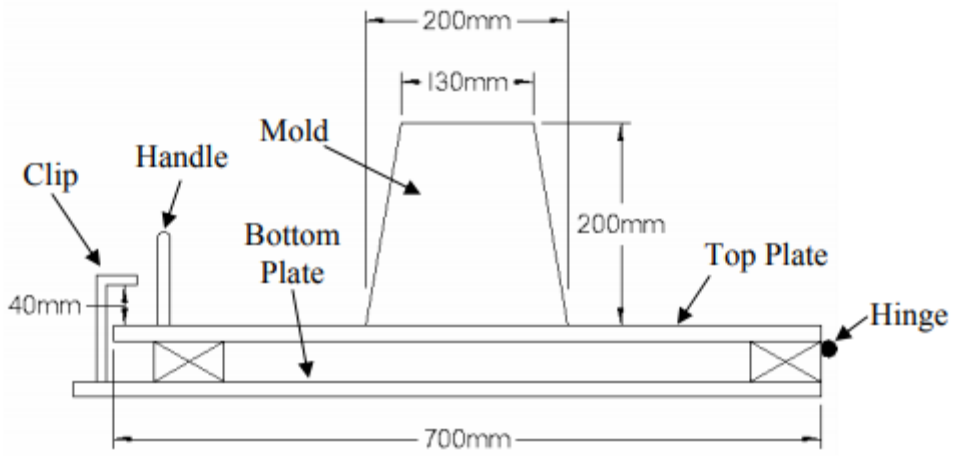


Figure 13: Flow table apparatus (Koehler & Fowler, 2003)

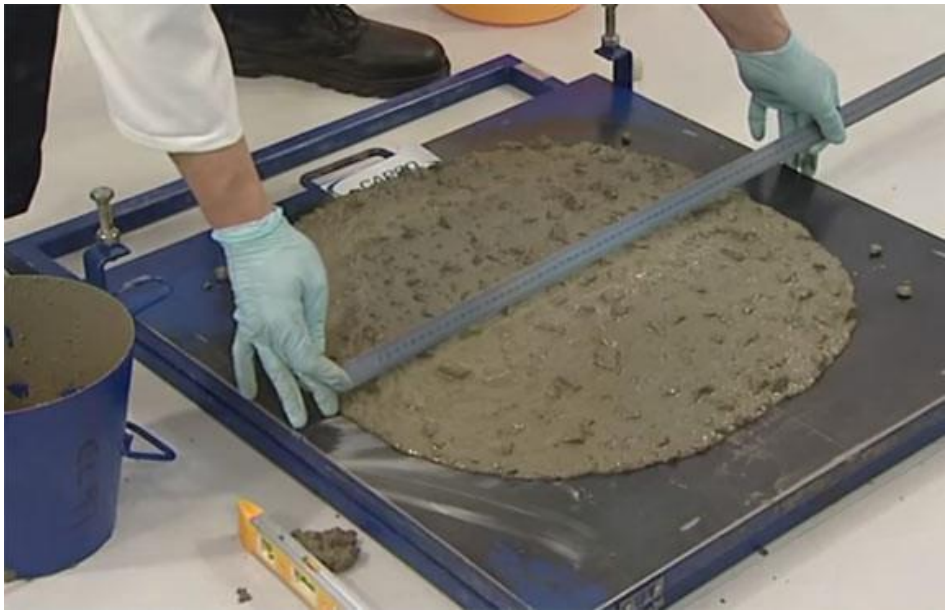


Figure 14: Slump measurement (Daily Civil, 2022).

3.2.4 Insoluble content

The insoluble material of the spent liquor is analysed by separation from its soluble material using centrifugation. Insoluble content determination is performed using a 10% solution of spent liquor samples, therefore, only dilute spent liquor to 10% solids in a 250ml volumetric flask, i.e., if the sample contains 50% solids, use 50ml of spent liquor and dilute up to the mark with distilled water. Equation (4) is used to calculate the amount of liquor to be added.

$$V_1 = \frac{10 \times 250}{Solids(\%)} \quad (4)$$

where V_1 is the volume of spent liquor to be added into the volumetric flask. Weigh an empty centrifuge tube (50ml) without its cap and record the weight accurately ($Mass\ A$). Then, add the 40ml of the dilute 10% liquor sample into the tube. Place cap on the tube and close. Then place the tubes containing the samples into the centrifuge rotor and ensure that the centrifuge rotor is balanced. If one sample is to be analysed, either repeat the steps above to prepare another tube with the liquor sample or fill a tube with water and ensure that the tube with the sample and tube with water have the same weights. To balance the rotor, the opposite sides must be loaded before starting centrifugation.

Centrifuge the samples at 25 ˚C for 10 minutes at 3000rpm. Discard the supernatant from the centrifuged samples and wipe off any excess liquid in the tube. Record the weight of the pellet and the tube ($Mass\ B$). The test is conducted in three replications for each sample (Sappi W728i004.TUG,2014). The insoluble content is calculated as follows:

$$Insoluble\ material\ (\%) = \left(\frac{Mass\ B - Mass\ A}{S.G\ of\ sample\ \times 40_{(volume\ of\ sample}}\right) \times 100 \qquad (5)$$

3.3 Important properties affecting the lignosulphonate production process and dispersion performance

3.3.1 Chemical reactions

The removal of the lignin from the chips is a chemical reaction which depends on cooking temperature, time, and the concentration of the cooking liquor. Many authors have conducted a range of experimental research to determine the specific reactions that occur during the delignification of wood, resulting in the formation of lignosulphonate during sulphite pulping. However, it is understood that three major reactions can occur: sulphonation, hydrolysis, and condensation. It is widely accepted that the delignification process consists of two steps: the sulphonation of lignin followed by sulphitolysis or hydrolysis, which makes lignin soluble in the form of lignosulphonate (Watson, 1992).

The alpha-carbon position in the phenyl propane units that make up lignin is the reactive site for the synthesis of lignosulphonate and is activated by the formation of benzyl ions: benzyl alcohol, benzyl alkyl ether and benzyl aryl ether structures. These structures are sulphonated by the hydrated bisulphite and sulphite ions present in the cooking liquor, resulting in the formation of lignosulphonic acid. However, lignosulphonic acid is insoluble under normal circumstances because its hydrolysable groups, which establish bonds between lignin and carbohydrates, are still intact. Sulphitolysis or hydrolysis are used to make the lignosulphonic acid soluble (Hanhikoski, 2014).

Hydrolysis reactions include the addition of hydrogen to the polymer chain, which causes a substituent group to be replaced by hydrogen ions. In the lignin chains, hydrolysis or sulphitolysis reactions are responsible for the breakdown of lignin-carbohydrate bonds, namely the benzyl alkyl ether bonds, to form soluble lignosulphonate. The two proposed pathways shown in Figure 15 are considered potential routes for the delignification process of sulphite pulping.

The other major reaction that occurs is the condensation reaction, which occurs in both solid and liquid phases, and competes with sulphonation for lignin reactive sites, thus preventing or delaying the delignification process. Condensation reactions that occur during the sulphite pulping process, occur between the lignin phenylpropane monomers under particularly low pH conditions, such as when the lignosulphonate acids are adversely affected, resulting in local high acidity. Excess lignin condensation makes lignin insoluble in the sulphite pulping process and results in greater molar mass compounds with reduced aqueous liquor solubility, even though sulphonation suppresses condensation (Moodley,2001). Therefore,

the condensation reaction mechanism is similar to sulphonation, and the reaction rate depends on the acidity of the solid-liquid interface.

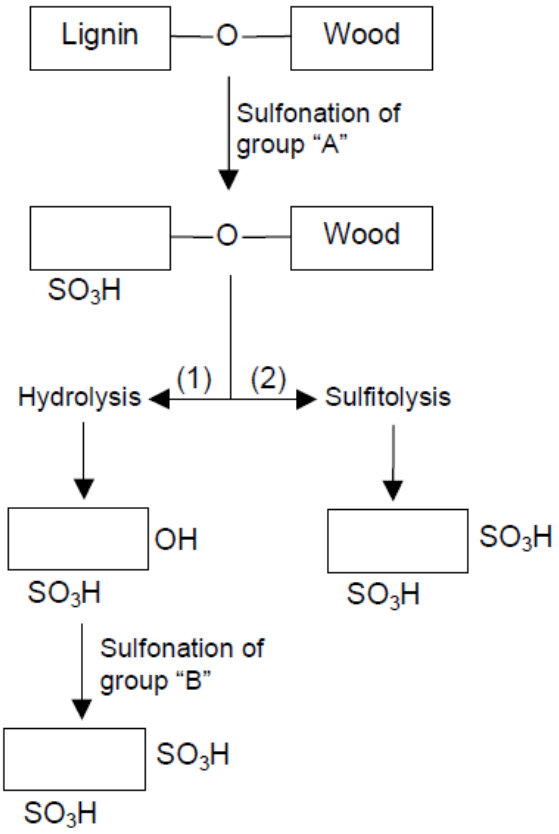


Figure 15: Two possible pathways of sulphite delignification (Rydholm, 1965).

Conditions that favour excessive condensation includes a combination of low sulphite or bisulphite ions, high acidity, and high temperature. The consequence of excessive condensation is a very dark pulp, known as a burnt cook which has adverse effects on the properties of lignosulphonate (Killan,1999).

3.3.2 Wood species

The wood species that is being pulped defines the value of variables used to control the rate and extent of the delignification process, i.e., temperature and time, amongst others, which affects the properties and quality of the products formed from this raw material. Hence, the wood species is an important variable because, for different species, the morphological characteristics and the chemical composition of the wood are different (Keskin-Schneider, 1991). The mill of interest employs neutral semi-chemical pulping, which has mostly been limited to hardwoods (Antonides, 2000). The benefits of hardwood neutral semi-chemical pulping included faster hardwood reaction and lower chemical charge for a given degree of delignification.

However, neutral sulphite liquor sulphonates hardwood lignin to a lower degree than softwood lignin. Eucalyptus wood, also known as Eucalyptus hardwood, is the wood species used at the mill and are quick-growing trees, that provide high-value timber with characteristics that serve the pulp, paper, and solid wood markets, which are found only along South Africa's southern and eastern coasts.

3.3.3 Chip size

The chip size and quality have a major influence on the uniformity of the cook and the quality of the products produced during sulphite pulping. Liquor penetration and diffusion of the cooking liquor take place in the wood structure at varying rates and different directions. The chip thickness and length are therefore both significant because, in chips of a smaller size, more rapid penetration can occur (Killan, 1999). Chips of varying sizes cook at different rates, and in a cook containing chips of varying sizes, the smaller chips absorb more of the cooking liquor and expend it in undesired side reactions. As a result, less liquor is available for the larger chips, which delignify at a slower rate.

Chip thickness more than chip length has been recognized as the most important parameter, in determining the speed and the thoroughness of impregnation of cooking chemicals into the wood chip, thus influencing the quality of products produced. Reduced chip thickness allows for faster-pulping rates and lowers the amount of screen rejects. Smaller wood fragments such as pin chips, fines and sawdust provides lower yields, weaker pulps and use more chemicals (Sappi, 2020). A large percentage of fine material in the chip supply will result in poor liquor circulation, therefore appropriate monitoring of the chip quality entering the process is necessary. There must always be a balance between liquor impregnation and fibre damage (which leads to poorer pulp quality). Smaller chips provide greater impregnation, while larger chips promote less fibre damage (Sappi, 2020)

3.3.4 Wood pre-treatment

The impregnation of the chips with cooking chemicals is an important step in the cooking process and is affected by the wood chip size, moisture content levels of the feedstock, the cooking liquor, and the process conditions. Therefore, these are important pre-treatment considerations that ultimately affect the pulping process and product quality. The degree of cooking is reduced by insufficient impregnation and the level of screen rejects increases (Sappi,2020). Several techniques have been used to improve penetration; the two methods used are to remove the air present in the wood capillaries and by adding a penetrating agent.

As the air in the chips can obstruct the penetration, it has been discovered that pre-steaming the chips in continuous systems to expel the air is beneficial. Small bubbles of entrapped air considerably enhance the barrier to liquid flow through wood capillaries. Pre-steaming the wood chips causes them to expand and release some air; but, as the water's vapour pressure rises, more air is pushed out. This can be accomplished by heating the chips before filling the digester as part of the steam-packing process or by adding steam separately. Steaming at atmospheric pressure must be maintained until a temperature of 100˚C is reached, indicating that all air has been displaced from both the chips and the surrounding vapour (Killan, 1999).

When a chip has a moisture content of 30% or greater, there is enough water in the chip to vaporize and displace all the air in the chip. At or above this moisture level, chip steaming is more effective. Water vapour will condense, and a vacuum will be formed, drawing the cooking liquor into the chip if the digester is filled with liquor at a temperature slightly below the pre-steaming temperature (Killan, 1999). Penetrating agents or surfactants enhance penetration through wetting and emulsifying effects on hydrophobic extractives. They reduce the angle of contact and increase the wettability of the wood chip surfaces, thus increasing liquor penetration into the wood capillaries. A higher concentration of liquor will enter more lignin sites early in the cook when penetrating agents are used thereby, improving the overall chip penetration and the rate of delignification.

3.3.5 Chip level and movement

During the entire cooking process, it is necessary to ensure that the chip flow remains uniform. This is managed by ensuring that the chip feed is consistent. Measuring the quantity of chip feed is the first and the most significant step in chip movement control. Chip properties and pre-steaming effects on the chips affect the chip flow. Supplying chips to the digester in any situation, it is important to monitor the level of the chip bin. However, the chip bin level is difficult to control due to the short residence time of 15 – 45 min, long feeding conveyor and high level in the bin (Pukkila, 2014). Also, if there is any variance in chip quality and size, this can affect chip-level measurements and control.

Pre-steaming of chips must be done under the correct conditions in a steaming vessel. Too low chip level causes steam to pass through the chip layer, leading to incomplete pre-steaming. Inadequate pre-steaming results in problems with the chip sinking in the impregnation zone. As a result, the level control is disturbed. In the chip chute, the first contact with the cooking liquor occurs. The chute's liquid level is kept constant. Chute circulation goes through the HP feeder, where chips are transferred to the feeding circulation. To avoid chip plugs or chips bypassing the HP feeder, the flow in chute

circulation must be controlled. Mass flow in the feeding circulation to the impregnation zone is maximized (Pukkila, 2014).

The setpoint in the impregnation zone for chip level is maximized but limited to not allowing the level to reach the top screw or the screw jams. By manipulating the chip screw speed and the bottom sluice flow to the transfer circulation, the level is regulated. This is further achieved by controlling the bottom scraper's speed and the amount of liquid. Normally, the chip level in the digester is managed by chip feed control or blow flow control (Pukkila, 2014). Blow flow is usually steadied to prevent changes in the delays of the digester and thus stabilize the residence time. For control of the chip digester levels, the sluice flow and bottom scraper may also be used.

3.3.6 Temperature

Temperature is a significant characteristic of the cooking process because it affects the cooking time, rate of diffusion and penetration of cooking liquor into the wood chips, delignification rate, and pulping process production rate. Sulphite pulping is done at temperatures ranging from 130°C to 185°C (Ingruper et al. 1985, p. 4). To obtain the same level of delignification as acid cooks, neutral pH cooks must use higher cooking temperatures and/or longer cooking durations (Virkola, Pusa et al. 1981). In alkaline and neutral sulphite cooks, relatively high temperatures (>175 ºC) are necessary to shorten the cooking time (Wong 1988).

Using higher temperatures than a long cooking time is more practical, so the neutral sulphite cooks are normally carried out between 160˚C and 190˚C (Virkola, Pusa et al. 1981). With an increase in temperature, the relative quantities of lignin and polysaccharides extracted from the wood decrease, yet cellulose loss is not excessive. Selectivity is thus more temperature-dependent for neutral sulphite pulping than Kraft or other alkaline pulping processes, in which selectivity remains relatively constant as the temperature rises (Antonides, 2000). For the NSSC pulping process, the challenges faced when using higher temperatures lie in ensuring sufficiently rapid liquor penetration and chemical distribution, and in increasing pressure.

The temperature at the bottom of the digester is significant. The temperature at the bottom of the digester affects the removal of dissolved lignin, and the dissolved lignin can precipitate back onto the fibre more easily if the temperature is too low.

3.3.7 Pressure

The main aspect of operating pressure in the sulphite cooking process is managing the amount of free sulphur dioxide in the system. This regulates SO_2 partial pressure, which in turn regulates free SO_2 concentration throughout the cooking process, as well as the pH, content of the liquor and the cooking pace. High pressure ensures quick cooking by maintaining a high sulphur dioxide concentration. The digester pressure constraints usually dictate the maximum pressure that can be employed in a mill. As the temperature rises, high free sulphur dioxide concentrations in sodium-based systems prevent sodium sulphite precipitation. Therefore, sustaining the high concentration of free sulphur dioxide concentration throughout the cook is accomplished by maintaining the pressure well above the liquor steam pressure, to prevent relieving excessive amounts of sulphur dioxide (Killan, 1999).

3.3.8 Time

The time for a continuous digester is a function of the production rate and is dependent on the mode of digestion. The cooking schedule/cooking cycle is used in pulping and is defined as a time-temperature schedule, such as an increase with time from the initial temperature to the maximum temperature followed by a period at maximum temperature. Cooking time is affected by the composition of the cooking liquor and cooking temperatures; hence, cooking at a higher temperature and a higher liquor concentration is faster (Keskin-Schneider, 1991). For neutral sulphite cooks, this is true because to achieve the same degree of delignification as Kraft and other sulphite pulping processes, a higher temperature increases the reaction rate for lignin removal.

3.3.9 Cooking liquor

Sulphur dioxide, hydrogen sulphite ions, and sulphite ions are the active components in the sulphite process. Their amounts in the cooking liquor are affected by the pH of the cooking liquor and have a major impact on delignification (Stenius et al 2000, p. 78). The base and SO_2 gas dissolved in water make up sulphite cooking liquor. The four soluble bases commonly utilized in sulphite processes are calcium, magnesium, sodium, and ammonium (Orblin & Fardim, 2011). The selected base is constrained by the process conditions and pulp. Calcium is only employed in the acid sulphite process since it is soluble below pH 2.3. (Ingruper et al. 1985, p. 7). The soluble form of magnesium bisulphite solution is up to pH 5.6 (Ingruper et al. 1985, p. 11).

Sodium and ammonium are soluble throughout the pH range, making them potential bases for the alkaline sulphite reaction (Orblin & Fardim, 2011). However, sodium is more extensively employed in industry than ammonium because ammonium is less thermally

stable. The base is usually required to neutralize the acids created during the cooking process (Sixta,1998). Sodium carbonate and sodium hydroxide are the most utilized alkalising or buffering agents in neutral sulphite pulping. This is to compensate for a sharp reduction in pH at the start of a cook, which could be caused by the neutralization of acetyl groups in wood chips (Antonides, 2000).

Cooking with Na_2CO_3 or NaOH promotes lignin removal but decreases carbohydrate retention (McDonough et al. 1985). The proportion of chemicals charged to the process in the cooking liquor is also an important parameter for the process. Chemical proportions for pulping are determined by the wood species, cooking conditions, the desired degree of delignification, product yield, and quality. However, to finish the cook in an acceptable amount of time and avoid lignin condensation reactions, an excess of chemicals is required (Keskin-Schneider, 1991).

3.3.9.1 Liquor pH

The amount of sulphite, hydrogen sulphite and hydroxyl ions in the cooking liquor determines the pH of the sulphite liquor. However, due to the formation of acids, which are influenced by time, temperature, and pressure, the pH of the cooking liquor decreases. Shifts in pH, which are caused by the concentration of ions in the cooking solution and occur when the temperature or pressure varies, indicate a dependence on these parameters (Hanhikoski, 2014). To guarantee that the delignification process can take place properly, neutrality is necessary and maintained for neutral cooking liquors by a buffer solution.

At a neutral pH level, lignin reactions are such that selective sulphonation of phenolic lignin is active, resulting in lignin fragmentation in part via sulphitolytic breakage of B-aryl ether bonds. Phenyl propane units containing carbonyl groups, whether phenolic or not, are also sulphonated under these conditions. Such structures are largely degraded by cleavage of the B-aryl ether bond if they are part of a B-aryl ether system (Antonides, 2000).

3.3.10 Liquor to wood ratio

The liquor-to-wood ratio is an important parameter to monitor because it enables uniform packing, liquor impregnation, and chip cooking. As a result, a significant amount of liquor is necessary. The ratio is managed by regulating the liquid extractions on different circulations, depending on the impregnation and cooking process utilized. For effective impregnation, the ratios typically range from 3:1 to 5:1, and in the digester, they typically range from 1.5:1 to 4:1. (Killan, 1999). If a low liquor ratio is utilized, the chips are impregnated with an excess of liquor under high pressure at temperatures as high as the maximum pulping temperature.

After chip impregnation, the ratio can be reduced by side relief so that there is less liquor to heat to the cooking temperature, lowering the steam needed. A high concentration of liquor offers a high driving force for diffusion during the impregnation process, resulting in excellent chemical penetration and lower screen rejection (McGovern, 1979).

When the digester is heated by cycling the liquor through an external heat exchanger, a higher liquor to wood ratio is required. As less steam is required to heat the digester charge and evaporate the spent liquor, a low liquor to wood ratio reduces steam consumption. However, enough liquor must remain in the digester to maintain good circulation. The chemical charge on dry wood chips, wood density, chip packing, and chemical concentration all influence the liquor-to-wood ratio (Antonides, 2000).

4 Methodology

This chapter describes the methodology used to develop the lignosulphonate dispersion performance prediction system for the Sappi Tugela lignosulphonate production process. The study was divided into two sections and numerical prediction models were developed using Microsoft Excel and RapidMiner software. Part A involved creating a function to predict concrete slump values using regression modelling techniques, and Part B involved developing models to predict insoluble content, dispersion index, and initial concrete slump values based on the lignosulphonate production process parameters. An in-depth look at the proposed approach is provided, as well as theoretical justification from literature.

4.1 A predictive model for concrete slump values

Concrete slump data collected over 12 months in 2020 and 2021, was extracted from the provided spreadsheets to create a dataset of 48 samples. The concrete slump data consisted of five measurements taken at 15-minute intervals, beginning with an initial concrete slump value. Stratified sampling was used to divide the dataset into two subsets; the model was developed on one subset and tested on the other. In general, datasets are divided into subsets with a 70:30% ratio (Moore et al, 2019); thus, the datasets for this study were divided into subsets with a 70/30 train/test split. The function was created with a subset of 34 concrete slump samples, and its performance on unseen data was evaluated with 14 concrete slump samples.

Stratified sampling resulted in a similar distribution of data among the subsets. The stratified sampling method generates random subsets while ensuring that the class distribution in the subsets matches the distribution of the entire dataset (May et al., 2010). A similar distribution for the subsets was required to determine whether the developed model generalized well to previously unseen data, an indication of how well the model would perform in real-world applications (Batista et al, 2004). A plot of the training subset is shown in Figure 16 and it was discovered that the data points for each measurement taken, lie on a straight line and follow the same trend. This meant that all the data could be fitted with a simple polynomial function based on two variables: time and the initial concrete slump value. This approach, rather than developing multiple prediction concrete slump models, allows for ease of use and implementation.

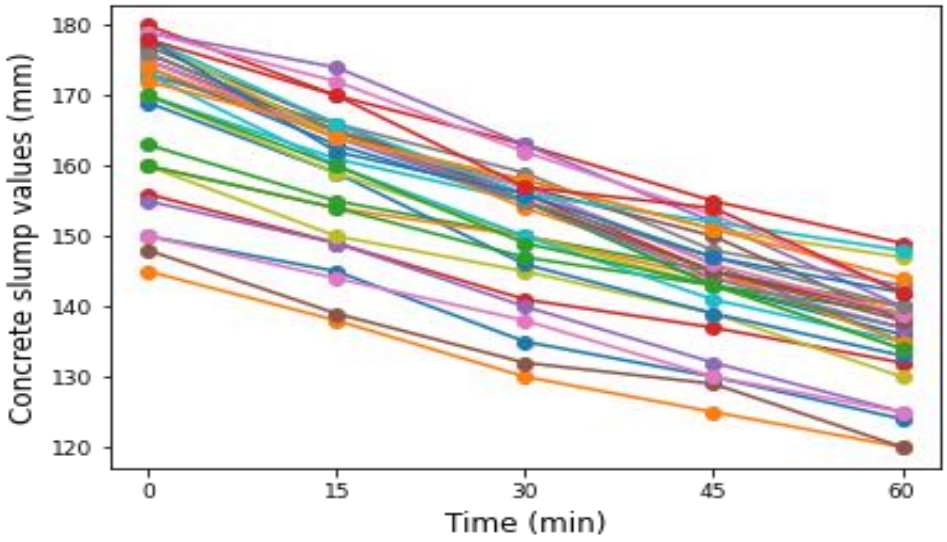


Figure 16: Concrete slump dataset (created with 34 concrete slump samples).

Three different approaches were used to develop the functions, made evident by how the dataset was normalized. Linear, quadratic, cubic, and fourth-degree polynomial functions were used to fit the data. The dataset was normalized as follows:

1. By dividing each concrete slump value by the initial concrete slump values, Figure 17 was obtained.

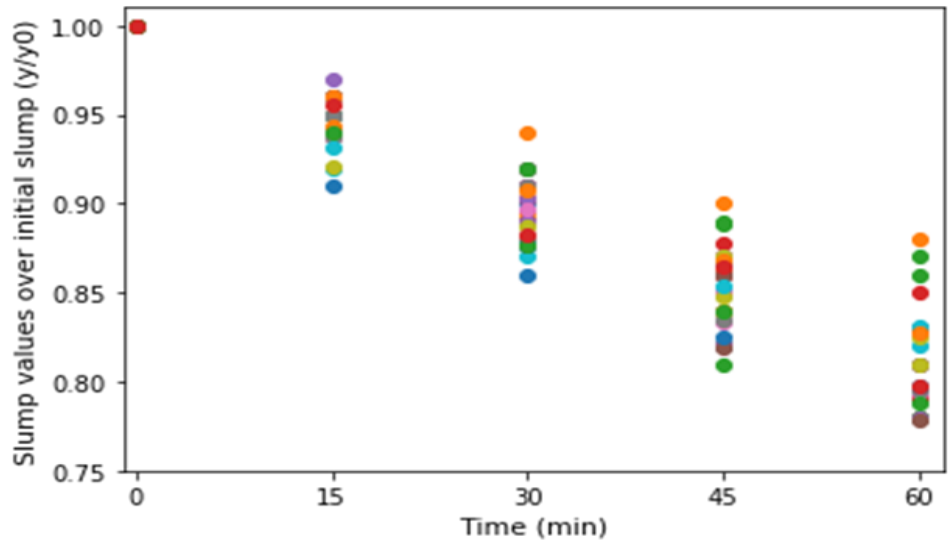


Figure 17: Normalised slump data using approach 1.

2. The initial concrete slump value was subtracted from each concrete slump value, Figure 18 was obtained. The difference in values is an indication of the improvement in the workability of the concrete mixture at each instance.

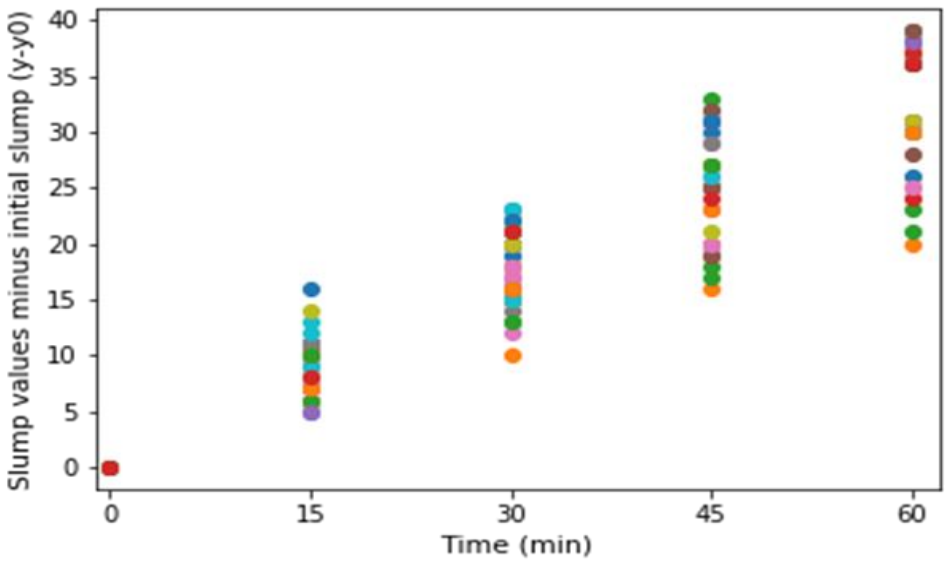


Figure 18: Normalised slump data using approach 2.

3. After subtracting each concrete slump value from the initial concrete slump values, the resulting values were divided by the initial concrete slump values and Figure 19 was obtained.

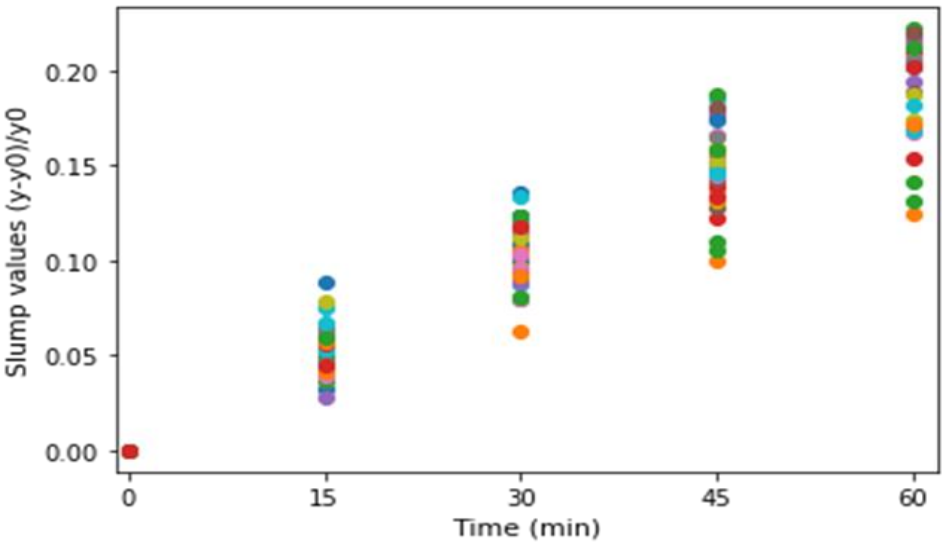


Figure 19: Normalised slump data using approach 3.

The proposed function could predict all four concrete slump measurements based on production process parameters, provided that the initial concrete slump value is affected by process variables. The chosen function was evaluated on the unseen dataset to determine how well it would perform in a real-world application.

4.2 Model development implemented in RapidMiner

The proposed architecture for developing the models is depicted in Figure 20. The workflow was divided into six major sections: data collection, data pre-processing, feature selection, model selection, model training using various learning algorithms, and model performance evaluation. RapidMiner software was used to analyse the data, as well as to create and validate the soft sensor models, and Microsoft Excel was used for model testing and comparison.

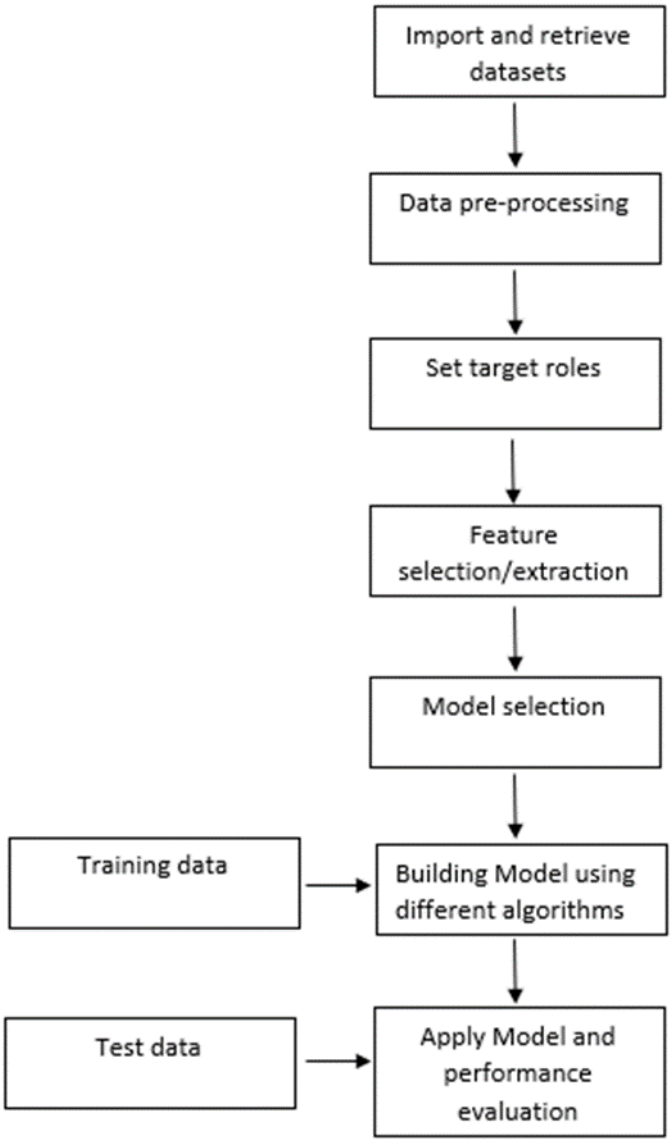


Figure 20: Proposed framework for model development (Rakala et al., 2020; AlBanna, 2016).

4.2.1 Data collection

To reflect the ongoing trends of the lignosulphonate production process and the desired product specifications, the process data collected for the study was limited to several months

in 2019, 2020, and 2021. The data collected was assumed to be reliable and an accurate representation of the processes and measurements taken during the lignosulphonate production process, from chip feeding to spray dryer operations.

The Tugela Mill's data historian system provided data on a minute-by-minute basis. The data was already synchronized based on time lags between the various lignosulphonate process equipment to ensure ease of use and accuracy. The laboratory results for samples of final and intermediate lignosulphonate products collected at various stages of the process were entered into the information system so that the results reflect when the sample was collected rather than when the results were obtained. As a result, the models developed do not ignore the obvious distributed nature of the process.

As mentioned, beginning in 2020, the concrete slump test and dispersion index test were conducted biweekly and weekly per month. These tests' results were provided on Excel spreadsheets and were not recorded in the mill's information system. As a result, the values were extracted from the spreadsheets, and the initial concrete slump and dispersion index raw datasets were created using the available process data at these instances.

A dataset of 48 initial concrete slump values collected over 12 months in 2020 and 2021 was used. A dataset of dispersion index data was collected over seven months in 2019, and a dataset of 42 dispersion index values was collected over six months in 2020. As a result, the raw dataset of dispersion index values used was a combination of data collected from Excel spreadsheets and the mill's historian system over 13 months in 2019 and 2020. Lastly, in 2019 and 2020, a dataset of insoluble content values was collected over 11 months from the data historian system. All raw datasets had approximately 500 attributes of various types such as numerical, categorical, and date-time attributes.

4.2.2 Data pre-processing

Raw dataset pre-processing

The raw datasets were examined to acquire a clear understanding of the process, identify any obvious problems, and gain an overview of the overall structure of the data to be handled at this early stage of model development. RapidMiner was used to perform several data pre-processing steps to acquire datasets that were more efficient and applicable for predicting the target variables, as low-quality data can result in inaccurate or low prediction results. The pre-processing steps used to transform the raw data addressed a wide range of disturbance-related issues, such as missing values, data normalization, noise reduction, outlier identification and replacement, and attribute selection, among others.

Upon examining the datasets for the insoluble content and the 2019 dispersion index, obtained from the mill's data historian system, revealed that previous laboratory results were repeated until a new sample was taken, and the result entered. Consequently, a constant value was generated over a specific timeframe. To remove these constant values, which do not accurately represent the process, the datasets were filtered on a 40-minute basis, which reflected the time delay between performing the test and entering the laboratory results of the samples into the system. The datasets for the 2019 and 2020 dispersion index values were then combined.

The datasets were pre-processed before being divided into test and training sets, which included the following steps: attribute selection, data cleansing, and filtering. Numerical models were developed to predict the target variables; thus, the datasets were filtered based on this attribute value type. All numeric attributes were chosen, including those of the real and integer types. Following that, data cleansing techniques were applied to the numeric datasets to better prepare them for machine learning, such as the removal of low-quality data and missing values. The columns of data that are of very low quality for machine learning were removed using three quality measures (Llyas & Chu, 2019):

- Missing values were calculated by dividing the total number of missing values in the column by the total number of rows.
- Stability was calculated by dividing the number of rows by the count of the most frequent non-missing values in the column.
- Validity was the proportion of column values that are not marked as missing, infinity, id, or stable.

Following that, the cleansed datasets were filtered based on the attributes that contained missing values, resulting in numeric datasets with no missing values. Finally, the insoluble content and dispersion index datasets were filtered to fall within the range of the lignosulphonate product specifications. The dispersion index dataset was filtered on a range of dispersion index values of 80 to 100, and the insoluble content dataset was filtered on a range of insoluble content values of 3 to 6.5% of liquor. There were approximately 430 numerical attributes in the resulting datasets.

Splitting dataset into training and test subsets

Splitting the entire dataset into test and training subsets is critical for model development because the selected model will be trained on the training dataset (used to fit the model) and evaluated on the test dataset (Zhu et al, 2020). The datasets were divided into training and test subsets with a 70:30% ratio. A machine learning model was designed to make accurate

predictions on new, previously unseen data; thus, the goal was to train a model to generalize well on new data. Generalizable models are preferred over fine-tuned models, which may perform well on the training dataset but poorly on unseen data (Batista et al, 2004).

This was only possible if the test dataset met the following requirements:

- It was large enough to produce statistically significant results; and
- It was representative of the training dataset. As a result, data distribution and all pre-processing steps were considered. A model trained on a dataset with vastly different pre-processing steps and data distributions than the test set will underperform during evaluation.

The datasets underwent stratified sampling, to ensure that the distributions of the test and training datasets were almost equivalent. Figure 21, Figure 22 and Figure 23 were the distributions of the training and test sets obtained after using stratified sampling.

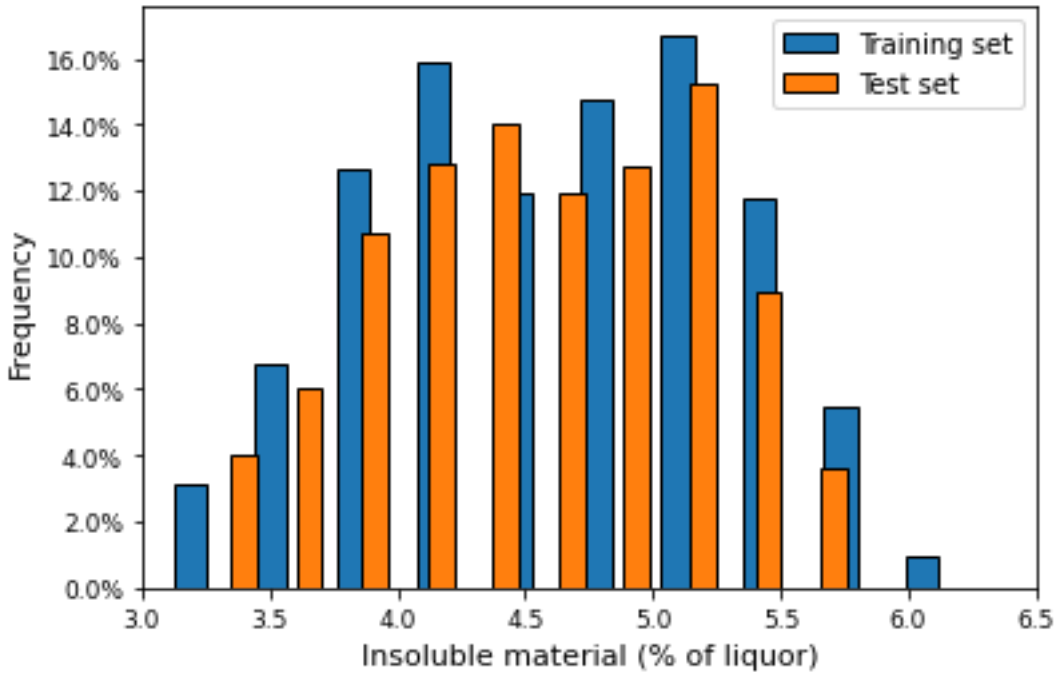


Figure 21: Insoluble content training and test sets distribution.

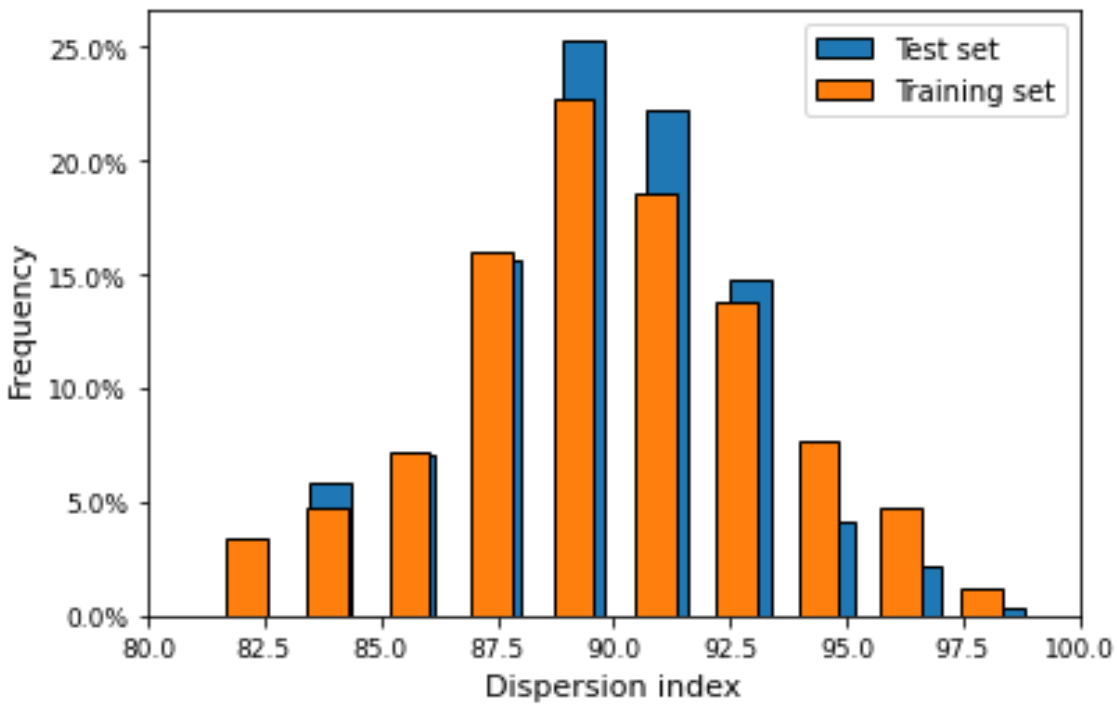


Figure 22: Dispersion index training and test sets distribution.

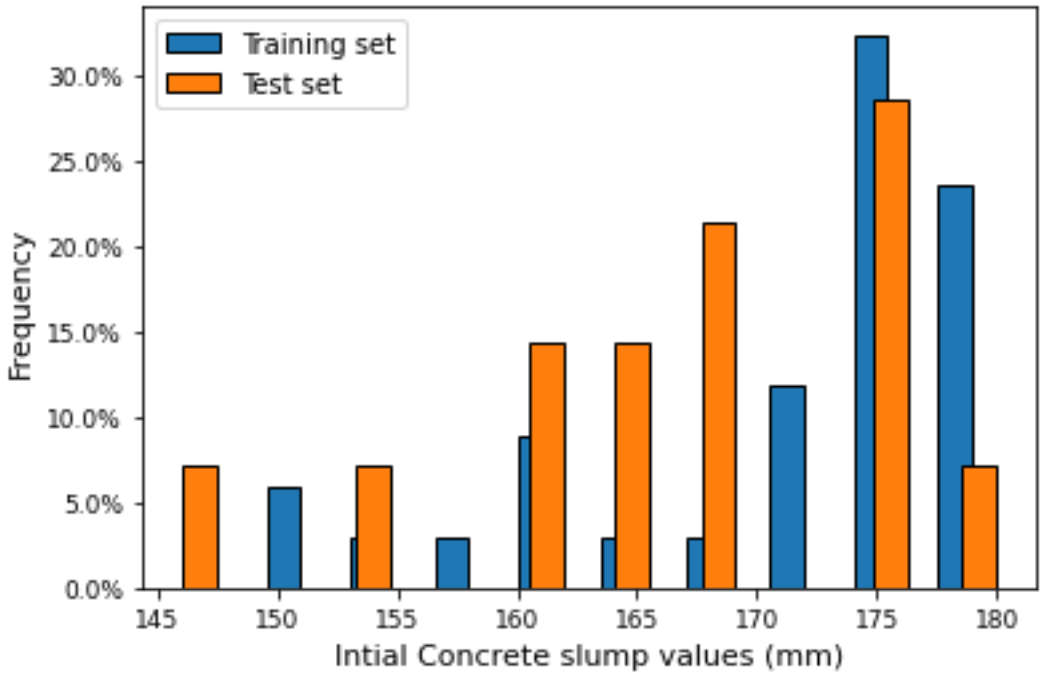


Figure 23: Initial concrete slump training and test sets distribution.

Training set pre-processing

Following the splitting of the datasets, the training datasets underwent two additional data pre-processing steps: normalization and noise reduction. Normalisation and noise reduction are effective approaches for potentially improving the quality of developed models.

Normalisation and noise reduction were applied to the entire dataset to transform and prepare it for model development. As a result, to avoid data leakage, these techniques were completed after the data was divided into training and test subsets.

When information not included in the training set is used to develop the model, this is known as data leakage. The additional information would allow the model to learn or know something it otherwise would not have known, invalidating the model's estimated performance. When the model is used on real-world data, undetected data leakage can lead to exaggerated and poor model performance (Tingle, 2019). Statistics from the data are commonly used for normalization, and noise reduction filters are dependent on the current and previous filtered values; hence, if these steps were completed on the entire dataset rather than separately on the training set, information about the test set would have leaked into the training set.

Normalization

Normalization ensures that all the numeric columns in the dataset are approximately on the same scale. Each column is rescaled so that the average of the resulting column is 0.0 and the standard deviation of all columns is 1.0. As a result, the different scales have no effect on machine learning algorithms. RapidMiner supports four different normalization methods; for this study, the training datasets were transformed using z-transformation, also known as statistical normalization. This method of normalizing subtracts the mean of the data from all values before dividing them by the standard deviation. The resulting data distribution has a zero mean value and a variance of one. It is a widely used and effective normalization method. It also retains the original distribution of the dataset, which was required for the trained model to perform well on the test set (Jo, 2019).

Noise reduction

Subsequently, to gain a better overview and understanding of the dataset trends and patterns, a simple filter was used to detect and smooth the noisy insoluble content and dispersion index training sets. The presence of noisy data in a dataset can significantly influence the prediction of any useful information. Many empirical studies have shown that noise in a dataset reduces accuracy and results in poor prediction results. Noise in process data can be attributed to measuring devices, electrical equipment, random errors introduced during data collection or capture, or the process itself (Seborg et al, 2011).

An exponential filter was chosen for the study; this is a low-pass filter that is commonly used to smooth noisy experimental data. It can be used to dampen high-frequency fluctuations. Its operation is described for a value at point t.

$$x_F(t) = \alpha x(t) + (1-\alpha)x_F(t-1) \quad (6)$$

According to Equation (6), the filtered measurement is a weighted sum of the current measurement $x_F(t)$ and the filtered value at the previous time step $x_F(t-1)$. Smoothing becomes weaker as the value of α increases, and when $\alpha = 1$, the measurement is not smoothened. Whereas $\alpha \to 0$, has a stronger smoothening effect; a pattern of the data begins to emerge as data peaks begin to diminish, but the actual measurement is gradually ignored (Seborg et al, 2011). To smooth the insoluble content and dispersion index data, the alpha parameter was adjusted to 0.4 and 0.3, respectively. These alpha values resulted in curves that reflected the trend or profile of the original training datasets.

4.2.3 Feature selection and model selection

To improve model performance, a suitable attribute or feature selection method can be used. Therefore, determining the most relevant features for the problem at hand was one of the primary data-mining tasks. Statistics, data crunching applications, pattern recognition, and machine learning have all identified feature selection as a vigorous and active research area. The primary goal of feature selection for model development is to reduce the complexity of the learning algorithm by creating a subset of available input variables from a set of available features. In both practice and theory, it has been shown to be effective in reducing the complexity of learned results, increasing prediction accuracy, and improving learning efficiency for a specific algorithm (Arunadevi & Nithya, 2016).

As a result of the significant increase in data velocity, veracity, volume, and variety, the feature selection process was required. The feature selection process was used collectively in the study to determine which of the RapidMiner regression predictive learning algorithms would be improved further to develop useful, accurate models. Deciding on a final learning algorithm from a set of candidate algorithms for model training was a critical process. Since the learning algorithm is the heart of the soft sensor, selecting the best type was vital for its performance (Slikovc et al, 2011).

The feature and model selection process consisted of three steps:

1. Use research and data analysis strategies to select a set of features.
2. Use feature selection algorithm/s to determine a good set of selected features. This was essentially a search method, with a performance measurement required to indicate how well a feature subset performed.

3. Use the default parameters of different learning algorithms selected, to compare the best feature subset. The accuracy and performance of various types of learning algorithms were compared, which aided in the selection of the best algorithm.

Reduction of pre-processed training sets

Based on the review of the process, raw materials, lignosulphonate product, laboratory methods used to evaluate the final and intermediate lignosulphonate products, the process flow diagram, and relevant P&IDs, it was determined that the lignosulphonate production process was essentially divided into four major parts: chip feeding, digesting, evaporation, and spray drying operations. As a result, the laboratory and process parameters of these sub-processes were key for developing useful models for predicting the target variables. The attributes of the pre-processed training datasets were reduced from the available process data to contain the significant attributes of the main sub-processes of the lignosulphonate production process.

As stated, the insoluble content of the strong red liquor (SRL) was used to adjust the process. Since the SRL was sampled at the end of the evaporation process, the insoluble content prediction model was developed using variables from the chip feeding, digester, and multiple-effect evaporator sub-processes. The dispersion index and initial concrete slump prediction models were based on variables from the chip feeding to the spray dryer sub-process, as these values are determined using samples of the final lignosulphonate powder product.

The final training datasets included numerical attributes that were all readily available and can be manipulated. The final training datasets for predicting insoluble content, dispersion index, and initial concrete slump values contained 80, 128, and 187 attributes, respectively. The final training dataset sizes were 3662, 2003 and 34, respectively which were used for training the models to predict the insoluble content, dispersion index, and initial concrete slump values.

The study's objective was to develop dispersion capability models using theory and data analysis. As a result, the final training datasets were also further reduced to include only the most important attributes of the main sub-processes based on process and product knowledge, as well as engineering judgment. A total of 37 numerical attributes were chosen, these were considered the parameters that most affected the process. This included parameters such as temperatures, flowrates, pressures, density, and pH among others, that affected reactions and the operation of each piece of equipment, resulting in optimal conditions not being achieved during the production process. The results of laboratory

analysis of the lignosulphonate product and intermediate process products were also considered. The variables chosen were based on the availability and the reliability of the data obtained, to ensure that improvements to the process can be made without significant delay.

All the models developed were trained and evaluated using three cases to achieve the stated goal and demonstrate the significance and influence of various variables chosen as input to the models. First, input variables based on data analysis were used (referred to as training dataset 1), followed by input variables based on theory (referred to as training dataset 2), and finally, variables based on both data analysis and theory were used (referred to as training dataset 1 and 2).

Feature selection algorithm

The genetic algorithm (GA) optimisation method and weighted correlation were used to select the best attributes to predict the target variables from the training datasets. The approach was used to identify useful correlations or relationships that are not direct indicators of the target attribute and automatically discard some features that do not provide relevant information for the problem. As a result, a set of input variables was established that are not highly correlated but have a significant impact on the target variable.

GA has produced competitive results, and it has been used in several studies to generate meaningful solutions to optimization and search problems (Tharwat & Gabel, 2019). It is a biologically motivated optimization technique that is inspired by Darwin's survival of the fittest principle and natural evolution processes such as heredity, mutation, selection, and crossover (Oliveri & Massa, 2011). Three key principles are at work in feature selection: ‘mutation' refers to switching features on and off, and 'crossover' refers to the trading of used features.

The method was dependent on the first step, which is the creation of a population of features; there are no fixed rules for the size of the starting population. However, if we want to do a crossover, we need at least two features in the population. Generally, the population size should be between 5 and 30% of the total number of attributes (Mierswa, 2018). Therefore, all the parameters for optimized selection by GA were left at their default values, and the population size was adjusted to values ranging from 5 to 30% of the total number of attributes in a dataset. The GA for feature selection in RapidMiner is a nested method, which means it is a sub-process and dependent on the learning algorithm used. Therefore, it iteratively runs the sub-process to determine the model's optimal performance given a set of input variable values when the population size parameter was varied.

Weighted correlation is concerned with the use of subject weights in the calculation of the correlation coefficient between the target attribute and each attribute in the dataset; the higher the weight assigned to the attributes, the more relevant it is considered. This weighting technique is based on correlation, and the attribute weight is the absolute or squared value of correlation (Costa, 2011). The feature selection workflow used in this study, and was implemented in RapidMiner, is shown in Figure 24

The maximum number of input variables used to develop the dispersion index and insoluble content models were kept at 20, while the maximum number of input variables used to develop the initial concrete slump model was kept at 10. The maximum number of input variables was chosen upon consideration of the dataset sizes and model complexity; the fewer features used, the more efficient the data analysis will be from a computational standpoint. It will also make learning predictability easier for the various algorithms that were applied to the datasets.

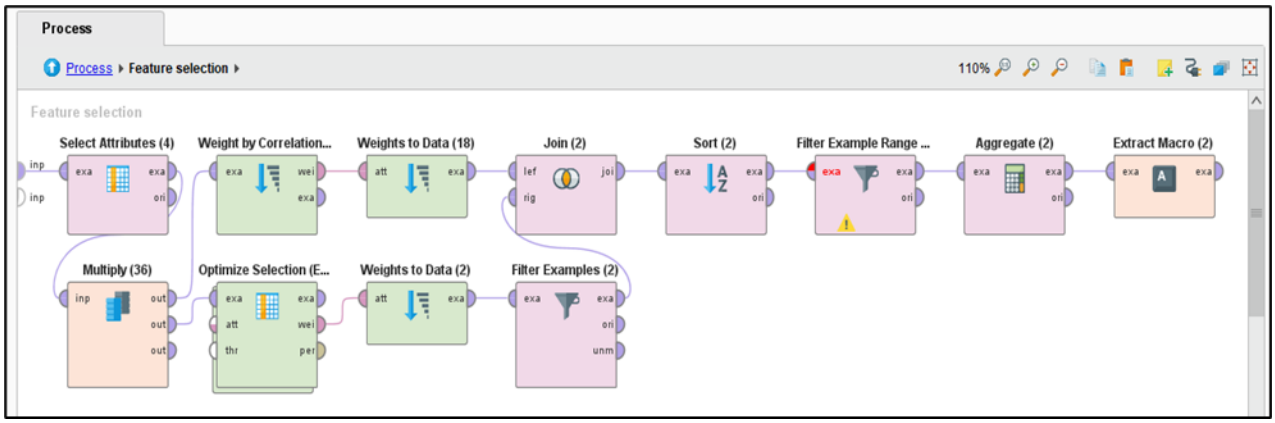


Figure 24: Feature selection workflow implemented in RapidMiner.

Model selection

The feature selection approach was used to eliminate learning algorithms that performed poorly given the set of input variables chosen.

Learning methods

RapidMiner has 62 predictive learning algorithms for developing models, but only 17 of these 62 learning algorithms can handle numerical target attributes. The learning algorithms used to evaluate the datasets for the study are discussed. The set of learning algorithms selected ranges in complexity, thus providing a good baseline for evaluating the various dataset sizes and the complicated relationships that exist between the number of attributes and the target attributes.

Support Vector Machine

Support Vector Machines (SVM) are one of the most widely used learning algorithms derived from statistical learning theory, proposed by Vladimir Vapnik (Vapnik, 1995). SVM is a machine learning approach used for classification and regression analysis that is based on structural risk minimization and statistical learning theory. It has been recognized as a statistical learning method that performs exceptionally well, and it is widely used in a variety of fields due to its effective learning and generalization capabilities (Wang & Huang, 2011). Based on the data provided, the SVM algorithm constructs a hyperplane or set of hyperplanes in a high or infinite-dimensional space. The hyperplane divides the data into different classes, and the algorithm finds the points from each class that is closest to the line; these points are known as support vectors, as observed in Figure 25

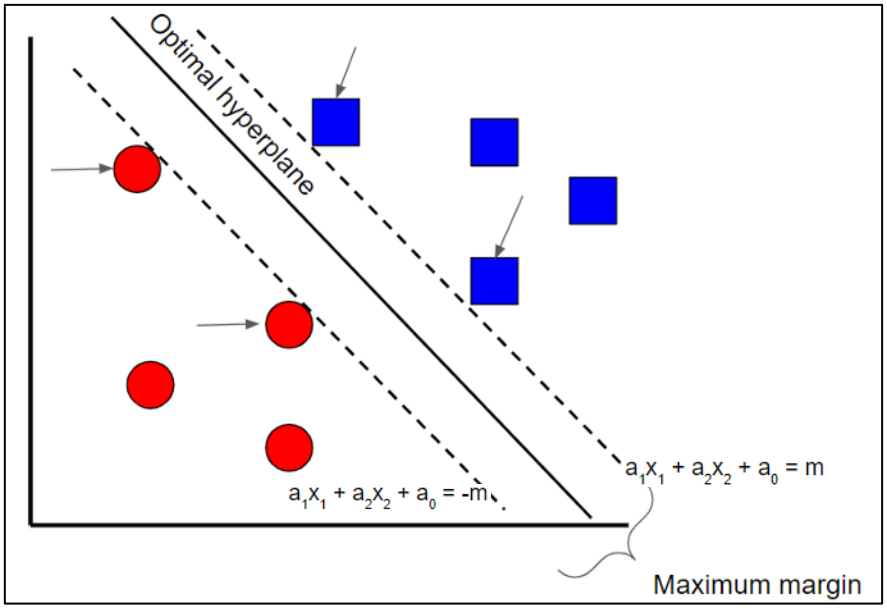


Figure 25: Representation of the objective of SVM algorithm (Shin, 2021).

The distance between the line and the support vectors is used to calculate the margin. The margin is estimated by using a Lagrangian formulation to solve the constrained optimization problem. As a result, the goal of the algorithm is to maximize the margin (Matheny et al, 2007). The mathematical expression of the goal of the SVM algorithm is provided by Equation (7.

$$MINIMIZE_{a_0,\ldots,a_m}: \sum_{j=1}^{n} MAX\left\{0,1 - (\sum_{i=1}^{m} a_i x_{ij} + a_0) y_j\right\} + \lambda \sum_{i=1}^{m} (a_i)^2 \quad (7)$$

Where n is the number of data points, m is the number of attributes and, x_{ij} is the i[th] attribute of the j[th] data point. The first part of the equation is focused on minimizing the error of the

number of falsely classified points SVM makes, while the second part of the equation, highlighted in grey, focuses on maximising the margin (Shin, 2021).

Although SVMs are widely used for both linear and nonlinear problems, the presence of nonlinearly separable data has a significant impact on SVM performance. As a result, SVMs employ kernel functions to map the dataset to a high-dimensional space where the data can be separable linearly. The use of kernel functions helps to develop a more accurate classifier or predictor. Various types of kernel functions can achieve this goal, including linear, polynomial, and radial basis functions, amongst others (Wang & Huang, 2011).

The hyperparameters to be optimized, to improve the performance of the SVM algorithm can be determined based on the kernel functions used to aid the SVM algorithm; given in Table 1, are the specific hyperparameters for radial and linear SVMs (Eitrich & Lang, 2005; Gaspar et al, 2012; Chapelle et al, 2002, Smola & Schölkopf, 1998).

Table 1: Hyperparameters for Support Vector Machine learning algorithm.

Hyperparameter	Definition	Range
C (Regularization parameter)	Controls the trade-off between maximization of margins and minimization of errors.	0.001 – 100
ϵ (Intensive loss function)	Errors that are within a certain distance of the true value are ignored. It affects the smoothness of the SVM's response and the number of generated support vectors.	0 – 0.001
γ (Gamma)	Used for non-linear SVM. Controls the distance of influence of a single training point.	0.0001 – 10

Linear Regression

Linear Regression is a fundamental machine learning algorithm that conducts regression analysis. It forecasts a target value based on a set of independent variables and it is widely used to determine the relationship between variables and prediction. During training, the model attempts to fit a linear equation between the input variables and the output variable. The linear regression model is shown by Equation (8).

$$y_i = a_1 x_{i,1} + a_2 x_{i,2} + \cdots + a_n x_{i,n} + b \qquad (8)$$

The motive of the linear regression algorithm is to find the best values for $a_n$an unknown regression coefficient using the input variable x_i and the target variable $y_i$and b the intercept. The complexity of a model like linear regression refers to the number of coefficients used in the model. Linear regression provides the best predictive accuracy for a linear relationship, is less sensitive to outliers, and frequently requires no further tuning (Neter et al.,2005). Feature selection in RapidMiner can be redone to optimize the linear regression algorithm, as the algorithm relates attribute weight to the target variable and decides on the best features to use to develop the linear regression model.

Decision Tree

Decision trees are a type of predictive learning algorithm that is widely used in data mining and machine learning, because of their ability to capture underlying relationships and their ease of implementation, setup, and interpretation (AlBanna, 2016). The decision tree is built as a hierarchical tree structure with nodes, as shown in Figure 26, and the algorithm's goal is to create a model that predicts the value of the target attribute based on several input variables. Each node represents a test on an attribute, each branch represents the test's outcome, and each leaf node contains the prediction made about the target variable (Parthiban, 2014).

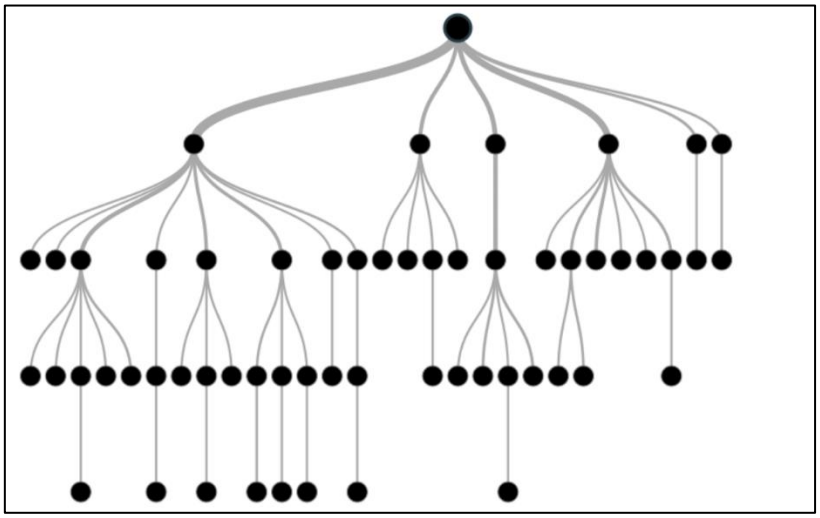


Figure 26: General decision tree structure (Du & Sun, 2008).

From each node of the decision tree, the Sum of Product (SOP) form is used. It is a collection of product representations. Every branch from the tree's root to a leaf node with the same class is a conjunction (product) of values; various branches ending in that class form a dis-junction (sum) (Sweta, 2020).

Decision trees are a rapid and useful solution for predicting instances in large datasets with many variables and have a high level of accuracy, but the performance is usually dependent on the dataset's characteristics. To avoid data over-fitting, pre- and post-pruning can be used to reduce model structure and complexity (AlBanna, 2016). To optimize the model for regression, tune the maximal depth hyperparameter (Boehmke & Greenwell, 2020; Prettenhofer & Louppe, 2014).

Table 2: Hyperparameters for Decision Tree learning algorithm.

Hyperparameter	Meaning	Range
Maximal depth	Controls the depth of individual trees, allowing for the capture of unique feature interactions.	5 – 20

Gradient Boosted tree

The gradient boosted tree model could be a regression or classification decision tree model. The algorithm generates predictive results by gradually improving estimations. The gradient boosted tree algorithm constructs one decision tree at a time, with each new tree assisting in the correction of errors made by the previously trained tree; this method is known as boosting (Ganjisaffar et al., 2011). Boosting is a meta-algorithm and a nonlinear regression procedure that aids in improving model accuracy by reducing variance and bias. Gradient boosted tree algorithms, build a series of decision trees by applying algorithms to gradually changed data in a sequential manner, as shown in Figure 27, resulting in a strong ensemble of prediction models (Panthong & Srivihok, 2015).

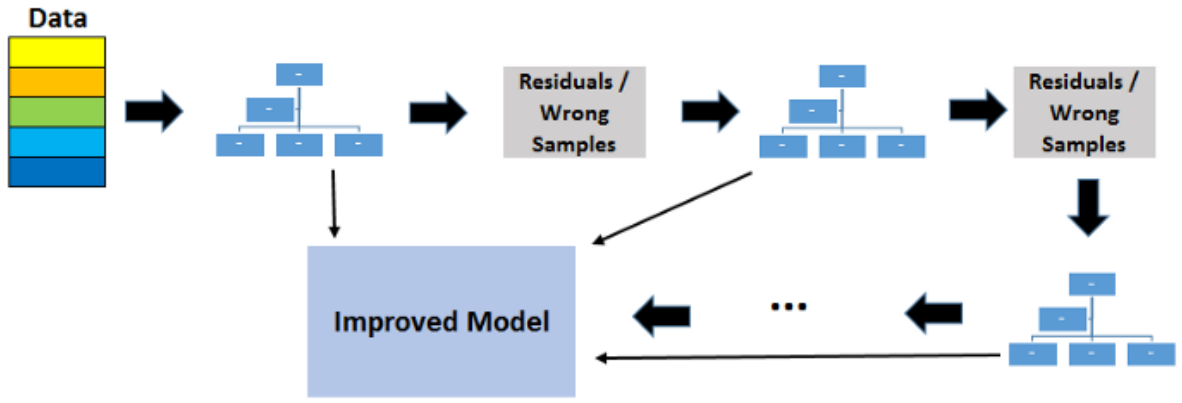


Figure 27: Boosting technique representation (Yildirim, 2020).

At each iteration, the observations with the worst prediction from the previous iteration are given more weight. The goal is to improve on the previous model's results with each iteration, focusing on the observations that were the furthest from the truth. While boosting trees improves their accuracy, it can also reduce the model's speed and human interpretability. However, this algorithm is continually used as it is adaptable and tree boosting is used to reduce issues that may arise, thus resulting in highly accurate models (Ye et al., 2009). Averaging techniques are used to combine the results from each tree along the way. The final model is a weighted average of all models created. The following hyperparameters, given in Table 3, are optimized to improve the gradient boosted tree model (Boehmke & Greenwell, 2020; Prettenhofer & Louppe, 2014).

Table 3: Hyperparameters for Gradient Boosted Trees learning algorithm.

Hyperparameter	Meaning	Range
Number of trees	The total number of trees that will be trained in the sequence. Each tree is grown to correct the mistakes of the previous tree.	100 – 1000
Maximal depth	Controls the depth of individual trees, allowing for the capture of unique feature interactions.	5 – 20
Minimum number of rows	The minimum number of rows to assign to terminal nodes; determines the tree's complexity.	10 -50

Random forest

The decision tree learning algorithm is the foundation for the random forest algorithm, but it possesses properties that allow collective decision-making to overcome the shortcomings of a single decision tree (Blachnik & Kordos, 2020). The algorithm generates a set number of random decision trees independently from random data samples, yielding a single, aggregated result. These trees are created using bootstrapped subsets of the training data. Thus, the model attempts to create an uncorrelated forest of trees, whereby the prediction of the overall forest is more accurate (Gray et al., 2016). The low correlation between models is important because uncorrelated models can provide more accurate predictions than any individual prediction.

In comparison to a single decision tree, the model is thus considered more robust and less likely to overfit training data. The random forest's final predictions are made by averaging the

predictions of each tree (AlBanna, 2016). Bootstrap aggregation is a machine learning meta-algorithm that generates random samples from the entire training dataset using a random selection of rows and columns with replacement, as shown in Figure 28. The random forest algorithm employs this approach to improve the model's stability and accuracy. It reduces variance and aids in the prevention of overfitting (Panthong & Srivihok, 2015).

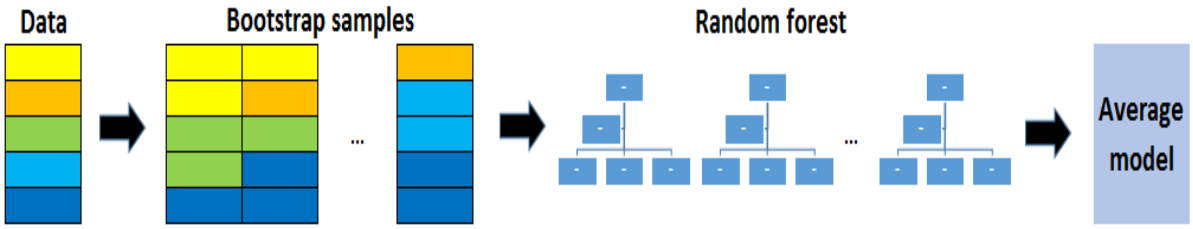


Figure 28: Bootstrap techniques used by Random Forest algorithm (Yildirim, 2020).

The following hyperparameters, shown in Table 4, are optimized to improve the random forest model (Boehmke & Greenwell, 2020; Prettenhofer & Louppe, 2014; Oshiro et al, 2012).

Table 4: Hyperparameters for Random Forest learning algorithm.

Hyperparameter	Meaning	Range
Number of trees	The total number of trees that will be trained in the sequence.	64 – 128
Maximal depth	Controls the depth of individual trees, allowing for the capture of unique feature interactions.	5 – 20

Neural Net

A neural network, often known as an artificial neural network, is a computer or mathematical model based on biological brain networks' structure and function. As shown in Figure 29, neural networks are arranged into layers, each of which has linked nodes with an activation function. The structure of the model is made up of a network with an input layer that communicates with one or more hidden layers. The neural net is made up of a network of artificial neurons that work together to process input using a connectionist approach to computation.

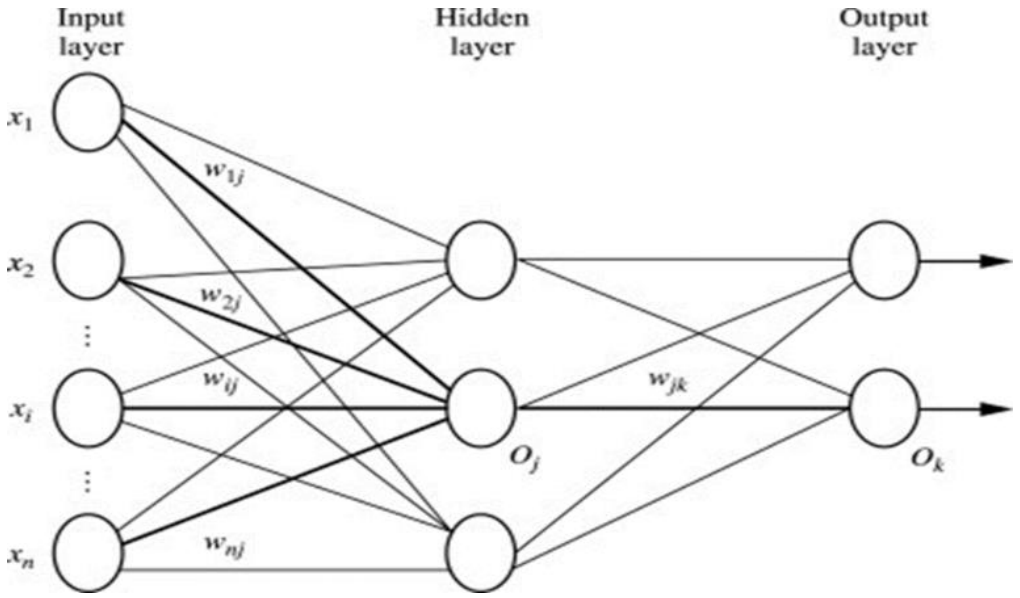


Figure 29: Neural Network Structure (Geetha & Nasira, 2014).

During the learning phase, most neural networks are adaptive systems that change their structure based on external or internal information that passes through the network (Caudill, 1987). Modern neural networks have been used to simulate complicated interactions between inputs and outputs, as well as to detect patterns in data, making this model incredibly powerful but potentially resulting in an unreadable structure. RapidMiner's Neural Net learning algorithm is a feed-forward neural network that only transports data in one direction, from input nodes to hidden nodes and then to output nodes.

A feed-forward artificial neural network with numerous node layers is known as a multilayer perceptron. A multilayer perceptron has at least three layers of nodes: an input layer, one or more hidden layers, and an output layer, each of which is fully connected to the one before it. Multilayer perceptrons use backpropagation to train their networks (Werbos, 1974; Werbos, 1994). Propagation and weight update are the two phases of the backpropagation method. The two phases are repeated until the performance of the network is sufficient.

By comparing the output values to the correct answer, backpropagation algorithms calculate the error. The error is then passed back into the network, and the algorithm utilizes it to alter the weights of each link to reduce the error's value by a small amount (Morariu et al., 2009). This approach is continued for several training cycles until the network converges to a state with a modest computation error. The neural network can be enhanced by adjusting the parameters shown in Table 5 (Boehmke & Greenwell, 2020; Smith, 2018).

Table 5: Hyperparameters for Neural Net learning algorithm.

Hyperparameter	Definition	Range
Momentum	Adds a fraction of the prior weight update to the current one; prevents local maxima and smoothens optimization directions.	0 – 1
Training cycles	Specifies the number of neural network training cycles to be used.	200 – 1000
Hidden layers	The number of hidden layers in the neural network topology is described by this value. Between the input and output layers are the node layers.	1 - 3
Neurons per layer/hidden layer sizes	The actual number of neurons/nodes in each layer; describes the size of the hidden layer.	2 – 20
Learning rate	Determines by how much the weights at each step are changed.	0.001 – 0.01

4.2.4 Model training, optimization, and validation

The chosen learning algorithms were trained and optimized using k-fold cross-validation to assess their effectiveness and improve model accuracy further by determining the optimal set of hyperparameters.

Cross-validation

A common cross-validation approach is K-fold cross-validation. The full dataset is utilized to train and evaluate the model after partitioning the training dataset, giving a good idea of how well the model will generalize to new data. The original sample is randomly partitioned into k equal-sized subsets in k-fold cross-validation. A single subset of the k subsets is used to validate the trained model, while the remaining k-1 subsets are utilized to train the model. The cross-validation procedure is then performed k times, with each of the k subsets only serving as validation data once. The accuracy of the model is the average of the accuracy of each fold (Arunadevi and Nithya, 2016).

As it allows the use of all the data for training and validation, k-fold cross-validation is a very useful approach for a small dataset. To predict the dispersion index and insoluble content, the learning algorithms were trained using 10-fold cross-validation, so the training datasets were divided into 10 equal-sized subsets. When the dataset size is small, less than 40, it is recommended that k be reduced to 5. As a result, to predict the initial concrete slump values, the learning algorithms were trained using 5-fold cross-validation, and the training dataset was divided into 5 equal-sized subsets. The use of 10-fold and 5-fold cross-validation to estimate model accuracy provided an acceptable compromise for the bias-variance trade-off in terms of computational efficiency (de Rooij & Weeda, 2019).

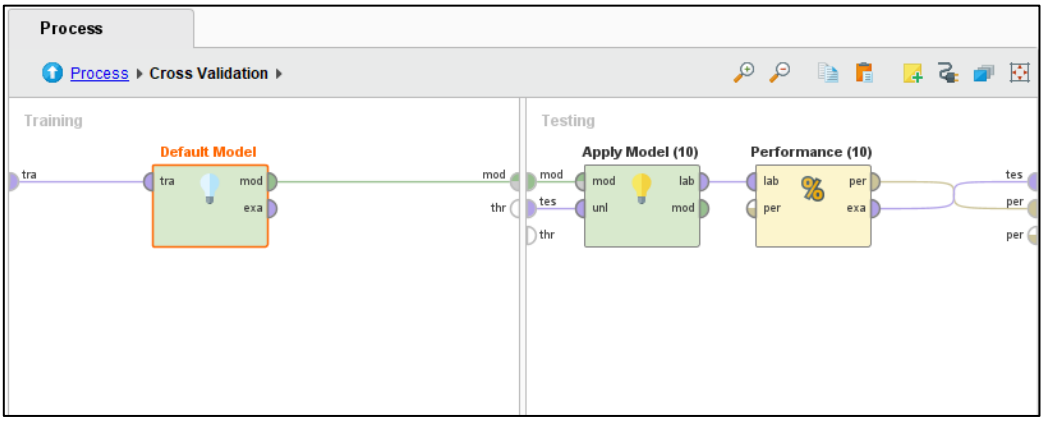


Figure 30: Cross-validation operator in RapidMiner.

Hyperparameter selection

To select and determine the optimal set of hyperparameters for the learning algorithms that provided the highest accuracy, the Genetic Algorithm optimization method was used. In most circumstances, this strategy is more appropriate than grid search or greedy search and yields better results. In RapidMiner, this is a nested method, which means that it has a sub-process that it executes several times to find the model's hyperparameter values in a given range where the optimal values of these parameters are found. The population size was determined based on the number of input variables used; thus, a population size of 20 was used when determining the optimal set of hyperparameters for the dispersion index and insoluble content models, and population size of 10 was used when determining the optimal set of hyperparameters for the initial concrete slump models.

4.2.5 Model performance evaluation

Following the selection of the appropriate soft sensor model structures, the models were tested on new and unseen test datasets. A good model makes good predictions on new data while avoiding overfitting and underfitting while making the most accurate predictions on the training dataset. As a result, good test set performance is a useful indicator of the model's

generalizability, indicating that the model was able to predict the target attributes based on the knowledge learned during training. Therefore, the model that performed the best on the unseen dataset was chosen as the final model to predict the target variable, because a model chosen for its accuracy on training data rather than accuracy on unseen data is likely to perform poorly on the unseen test dataset. The reason for this is that the model is not as generalized, and the training dataset had a specialized structure (Batista et al, 2004).

The test sets were pre-processed in the same way that the training datasets were. The test set must be nearly identical to the training datasets in every way. This was necessary to determine whether the trained models generalized well to an unknown dataset. Thus, pre-processing steps like normalization and noise reduction were performed on the test sets, and the distribution of data for the test and training sets was similar, which was achieved when the datasets were split by stratified sampling.

4.3 Model improvement

The primary goal of industry is to operate as close to or within range to the point where profit maximization is possible. This means that no product should deviate from the required specifications, and the product should have the least amount of product quality give-away possible (Quelhas & Pinto, 2016). Models developed to predict these specifications should be improved to achieve this goal through process monitoring and control.

As a result, an adaptive procedure, such as bias updating, was used to improve the fit of the evaluation results. This technique is typically used online, where the difference between a target variable value provided by the lab and the most recent value predicted by the soft sensor can be used to update the soft sensor by a bias (Mercangöz & Doyle, 2008; Singh, 1997). This simple technique could be used to improve the fit of the developed models and is widely used in industry and literature for optimizing purposes or for soft sensor inferences (Sharmin et al. 2006; Mu et al., 2006; Tran et al., 2005).

This study used a method similar to the one mentioned by Zhang (Zhang et al., 2019). The method concluded that the models' prediction results could be updated by introducing an offset smoother bias, as described in Equation (9).

$$y_{cor}(m) = y_{pre}(m) + bias_n(m) \quad (9)$$

where $y_{pre}(m)$ and $y_{cor}(m)$ denote the model prediction and the corrected predicted value for mth instance respectively, and $bias_n(m)$ is the updated bias written as:

$$bias_n(m) = \eta bias_n(m-1) + (1-\eta)bias(m) \quad (10)$$

$$bias(m) = y_{mea}(m) - y_{pre}(m) \quad (11)$$

where η is the weighting factor satisfying 0<η>1, $bias(m)$ is the current bias with the initial value $bias(0) = 0$, and $y_{mea}(m)$ is the actual target value.

Microsoft Excel was used to investigate the bias updating technique. The method was devised by combining values obtained from the developed soft sensors and laboratory results of the target variables. Initially, Equation (11) was used to calculate the error or current bias,$bias(m)$, between these values. Following that, the updated bias, $bias_n(m)$, was calculated using Equation (10) and the calculated values from the previous step. Equation (9) was used to calculate the updated prediction values. The updated bias values were added to the values predicted by the soft sensor to provide a new corrected estimate of the target variable. This method allows the model to learn from previous values and provide useful predictions. The performance metrics used for the study were determined using the acquired updated prediction values and laboratory results.

The value of η was determined for the study, using the built-in tool Solver, which allows for iterative calculations without the need for programming. When the optimum value toward the mean square error value was obtained, the weight factor was discovered. Using this simple approach allows for a faster iteration process for solving this parameter. The minimal mean squared error value was chosen as the objective function because it indicates how well the prediction performed. The constraint conditions are obtained by taking into account the weight factor, η ϵ R, and the fact that this parameter should be between 0 and 1. Since these problems are non-linear in nature, the evolutionary solving method was chosen. As a result, it was discovered that when η was 0.3, there was a significant improvement in the bias and performance of the models, as well as a minimum mean square error.

4.4 Comparative metrics

Two performance metrics were used throughout the study to evaluate the performance of the models developed. The accuracy and performance of the learning algorithms and developed models were compared by observing the mean squared error and squared correlation values, which were used to determine how successful the prediction was, selected the best feature subset, assist in deciding which algorithms to train and optimize, and how well the model performed on unseen data.

4.4.1 Mean squared error

The mean squared error (MSE) is a regression error metric that measures the average of the squares of the errors. In this case, the error is the average squared difference between the observed/actual values and the predicted values. It tells us whether the predictions made are accurate or misleading in comparison to the actual values; thus, the larger the MSE value obtained, the greater the error. Equation (12) is used to determine MSE (Diez et al., 2014; Spiegel, et al., 2001).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{12}$$

where n is the number of data points, y_i was the value of the target variables and $\hat{y}_i$ was the predicted target values.

4.4.2 Squared correlation

Correlation (R) is a number ranging from 1 to -1 that describes the strength of a linear relationship between two variables. The correlation is either -1 or 1 only when the relationship is perfectly linear. The correlation will be close to +1 if the relationship is strong and positive. If it is both strong and negative, it will be close to -1. If there is no obvious relationship between the variables, the correlation will be close to zero. Like correlation, squared correlation, or the coefficient of determination (R^2), indicates how closely two variables are related.

R^2 measures the fraction of variation (ranging from 0 to 1) between two values; it is the proportion of data that is closest to the fitted regression line. Variance for regression is a measure of variability determined by the average squared deviations from the mean; the aim is to have a low value. Equation (13) shows how to calculate squared correlation (Diez et al., 2014; Spiegel, et al., 2001).

$$R^2 = \left(\frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \cdot \sqrt{n(\sum y^2) - (\sum y)^2}} \right)^2 \tag{13}$$

where n is the number of data points, x was the value of the target variables and y was the predicted target values.

5 Results and discussion

The findings from the methodology taken to develop the lignosulphonate dispersion performance capability prediction system are presented and discussed. The study evaluated the performance of the proposed models by observing feature selection, validation results of the trained models, and performance evaluation of the selected models. The mean square error and squared correlation values obtained were used to assess the predictions made by the models.

5.1 A predictive model for concrete slump values

5.1.1 Development of concrete slump function

From the three approaches taken to develop functions to predict the concrete slump values, that are dependent on time and the initial concrete slump value, the values that indicated any abnormal behaviour were removed from the dataset. To capture the trends of the approaches used linear, quadratic, cubic, and 4th order polynomial functions were fitted to the datasets. Each approach yielded squared correlation values ranging from 0.9 to 0.92, with varying mean squared error values. Table 6 is a summary of concrete slump functions that performed the best for each approach taken; where y is the value of the slump at time t, y_0 is the value of the slump at time 0, and t is time in minutes.

Table 6: The concrete slump functions developed.

	Function	R^2	MSE
1	$y = y_0(9.2 \times 10^{-6}t^2 - 0.004t + 0.99)$	0.92	11.7
2	$y = y_0 + 0.002t^2 - 0.63t - 0.02$	0.9	17.9
3	$y = y_0 - y_0(-5.1 \times 10^{-10}t^4 + 8.7 \times 10^{-8}t^3 - 1.4 \times 10^{-5}t^2 + 0.004t + 6.1 \times 10^{-18})$	0.92	11.7

The first and third functions obtained have the same performance; however, when the complexity of these functions was considered, the first function represented a simple model that provided strong performance results; thus, this function was the function of choice to be used to determine the concrete slump values. Figure 31 depicts normalised concrete slump data with the best fit curve for the data.

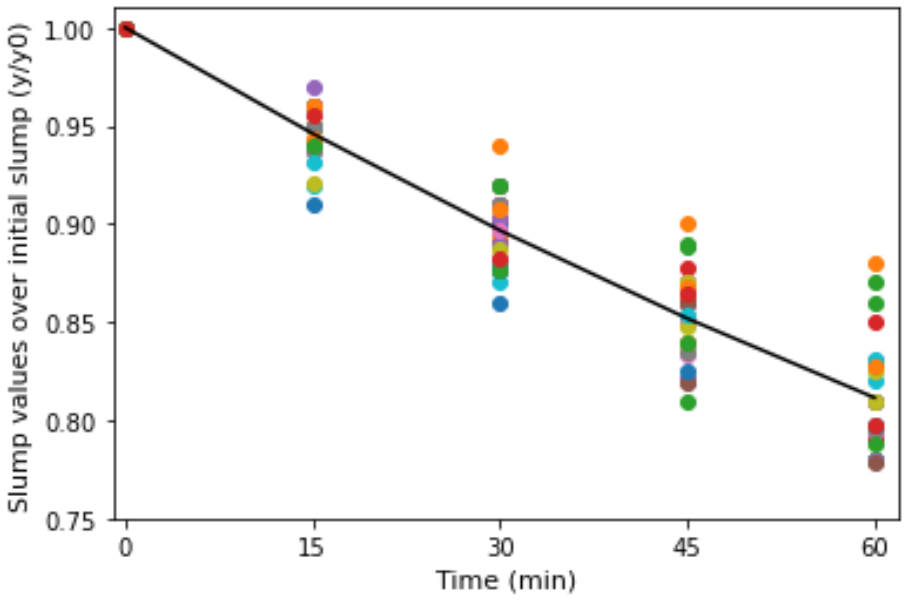


Figure 31: Normalised concrete slump data with the best fit curve.

Figure 32 depicts the function's prediction result. The function successfully established a relationship between the input variables and the target variables, allowing the function to predict slump values at varying time intervals with small prediction errors, as evidenced by the data points falling close to or on the fitted regression line. Further evaluation of the chosen function yielded the performance results shown in Table 7.

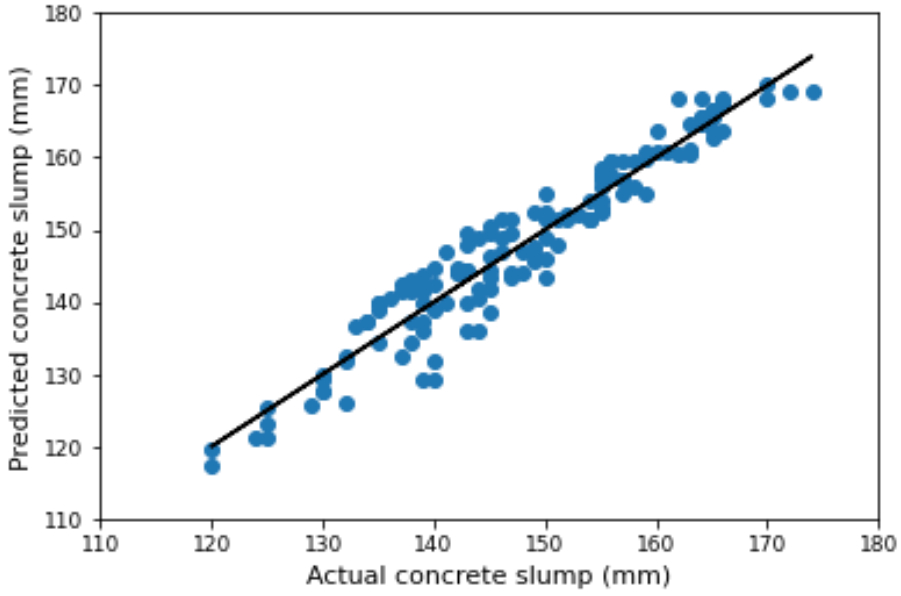


Figure 32: Actual vs Predicted concrete slump values.

Table 7: Performance results for selected concrete slump function.

Time (minutes)	15	30	45	60
R^2	0.94	0.9	0.81	0.71
MSE	5.4	8.1	13.9	19.6

The model performed reasonably well but its prediction ability decreased at times 45 and 60 minutes, as indicated by the reasonable squared correlation and high mean squared error values obtained when compared to the performance results obtained at times 15 and 30 minutes. These results can be attributed to the model struggling to capture the variation in concrete slump values in these instances, and this variation in values could be due to the conditions under which the slump test was conducted.

5.1.2 Evaluation of concrete slump function

The test set was pre-processed to remove concrete slump values that showed abnormal behaviour (outliers). The performance results of the tested function showed an overall squared correlation value of 0.81, a drop of 10.4 %, and a mean squared error value of 13.6, an increase of 13.2 %, compared to the previous performance results. Further evaluation of the selected function yielded the performance results shown in Table 8 for the tested model.

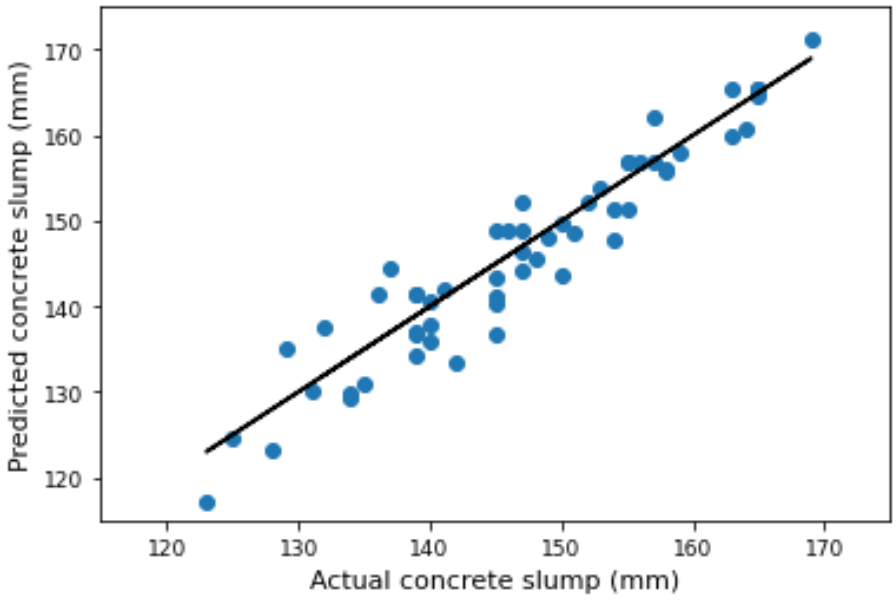


Figure 33: Actual vs Predicted concrete slump values from the test set.

Table 8: Performance results for tested concrete slump function

Time (minutes)	15	30	45	60
R^2	0.96	0.86	0.76	0.61
MSE	3.8	11.0	15.8	23.7

The model's prediction ability deteriorated; however, slump values at 60 minutes provided the largest performance reduction. At this time, the poor performance can be attributed to the wide variation in concrete slump values which the model was not able to capture sufficiently, as illustrated in Figure 33. Overall, the developed function provided a useful prediction system for determining concrete slump values using two dependent variables. At times 15 and 30 minutes, the function achieved a balance between variance and bias, and a strong predicting accuracy for the test set is observed. The function fitted the test set reasonably well, but prediction errors are visible, as evidenced by the mean square error values and predicted data points falling within a certain distance of or on the fitted regression line.

5.1.3 Improved concrete slump function

To account for the shortcomings of the developed function, a simple bias updating technique was used to improve the selected function's findings. Figure 34 shows an improved prediction system in which the predicted slump values fall on the fitted regression line, hence small prediction errors and a strong balance between bias and variance were achieved.

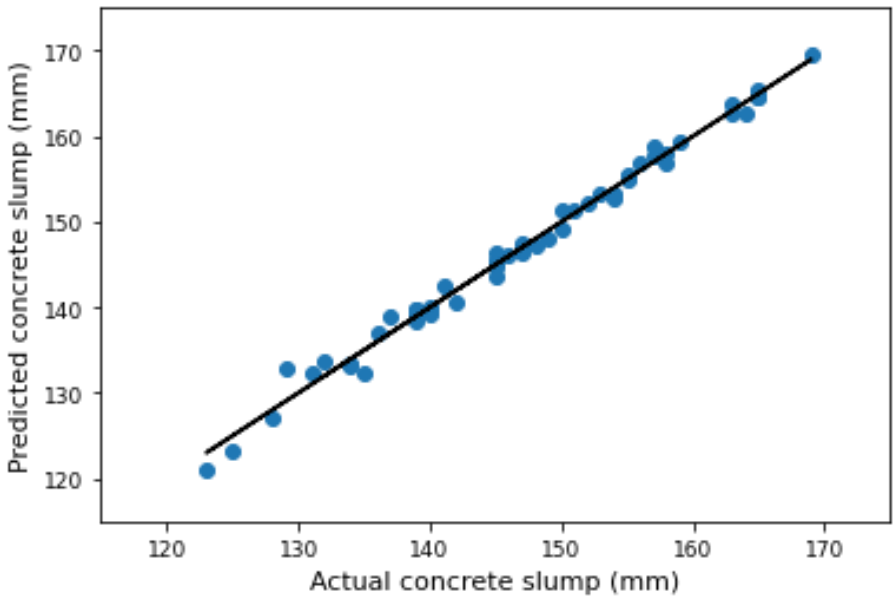


Figure 34: Actual vs Predicted concrete slump values after bias updating technique.

The model's overall performance improved significantly, with an increase of 8.4% to achieve a squared correlation value of 0.98 and a decrease of 90.4 % to achieve a mean squared error value of 1.3. Shown in Table 9, were the performance results obtained when using the simple corrected function.

Table 9: Performance results for bias updated concrete slump function.

Time (minutes)	15	30	45	60
R^2	0.98	0.98	0.97	0.95
MSE	0.41	0.71	1.4	2.5

5.2 Model development in RapidMiner

5.2.1 Initial concrete slump model

The training set obtained after pre-processing steps were taken, contained 186 numerical attributes and 34 examples (referred to as training dataset 1). The dataset was further filtered to include attributes chosen based on process knowledge, resulting in a dataset with 37 attributes and 34 examples (referred to as training dataset 2). The variables in both training datasets were from the lignosulphonate production process chip feeding to spray dryer operations.

Feature selection and model selection

The best features for predicting the initial concrete slump values were obtained when the GA method's population size parameter varied from 2 to 56. Table 10 shows the best performance results obtained for the learning algorithms, as well as the population sizes used to achieve these results.

Table 10: Initial concrete slump performance results for feature and model selection.

	Training dataset 1			Training dataset 2		
Learning algorithm	Population size	Statistic		Population size	Statistic	
		R^2	MSE		R^2	MSE
Support Vector Machine	35	0.55	36.9	10	0.54	69.0
Neural Net	28	0.61	20.3	8	0.59	25.3
Random Forest	20	0.64	18.7	8	0.58	19.6
Decision Tree	50	0.56	38.6	7	0.51	52.6
Linear Regression	30	0.25	74.1	5	0.15	66.9

As shown in Table 10, the Neural Net and Random Forest learning algorithms performed better than the other selected learning algorithms, and thus these algorithms were chosen for training and validation.

Model training, validation, and optimization

The validation results of the trained models are shown in Table 11.

Table 11: Performance results for the trained initial concrete slump models.

Training dataset	Random Forest		Neural Net	
	R^2	MSE	R^2	MSE
1	0.83 +/- 0.14 (average: 0.72)	26.2 +/- 4.5 (average: 28.7)	0.7 +/- 0.22 (average: 0.77)	28.6 +/- 7.9 (average: 26.4)
2	0.81 +/- 0.15 (average: 0.75)	25.1 +/- 8.6 (average: 25.5)	0.72 +/- 0.24 (average: 0.77)	25.1 +/- 14.1 (average: 25.1)
1 and 2	0.81 +/- 0.21 (average: 0.76)	26.3 +/- 12.3 (average: 26.4)	0.61 +/- 0.34 (average: 0.76)	25.5 +/- 12.5 (average: 25.1)

The trained Neural Net models performed better overall based on the input attributes chosen and when the desired performance results, a high squared correlation and low MSE values were obtained. Therefore, the Neural Net model was optimized using the model's hyperparameters. When the following hyperparameters were obtained, the best performance for predicting the initial concrete slump values was achieved, as shown in Table 12 and Table 13.

Table 12: Neural Net hyperparameters optimized.

Training dataset	Parameters	Value
1	Momentum	0.5
	Training cycles	1000
	Hidden layers	1
	Neurons per layer/hidden layer sizes	7
	Learning rate	0.01
2	Momentum	0.7
	Training cycles	800
	Hidden layers	2
	Neurons per layer/hidden layer sizes	5 and 3
	Learning rate	0.009
1 and 2	Momentum	0.8
	Training cycles	900
	Hidden layers	2
	Neurons per layer/hidden layer sizes	10 and 4
	Learning rate	0.003

Table 13: Performance results for optimized Neural Net model.

Training dataset	Statistics	
	R^2	MSE
1	0.7 +/- 0.23 (average: 0.76)	23.8 +/- 13.2 (average: 24.1)
2	0.72 +/ - 0.31 (average: 0.85)	15.4 +/- 8.2 (average: 15.6)
1 and 2	0.69 +/- 0.23 (average: 0.81)	23.1 +/- 8.1 (average: 19.7)

The performance of the optimized Neural Net models improved significantly. The squared correlation values improved between 5 to 10%, and the mean squared error decreased between 15 to 39%. The models performed relatively well such that a suitable relationship between the input variables and the target variable was made, allowing the model to understand the data and predict the initial concrete slump values.

Model evaluation

The prediction results of the Neural Net models are given in Table 14.

Table 14: Performance results for tested Neural net model.

Training dataset	Statistics	
	R^2	MSE
1	0.64	65.6
2	0.55	72.7
1 and 2	0.59	47.6

The input variables chosen based on training datasets 1 and 2 performed better when considering that the MSE value obtained was much lower than the approaches used to develop the initial concrete slump prediction model. As a result, the Neural Net model developed with training datasets 1 and 2 was chosen as the final model to predict the initial concrete slump values which are used to assess lignosulphonate dispersion capability. Figure 35 and Figure 36 depict the final Neural Net model's prediction results.

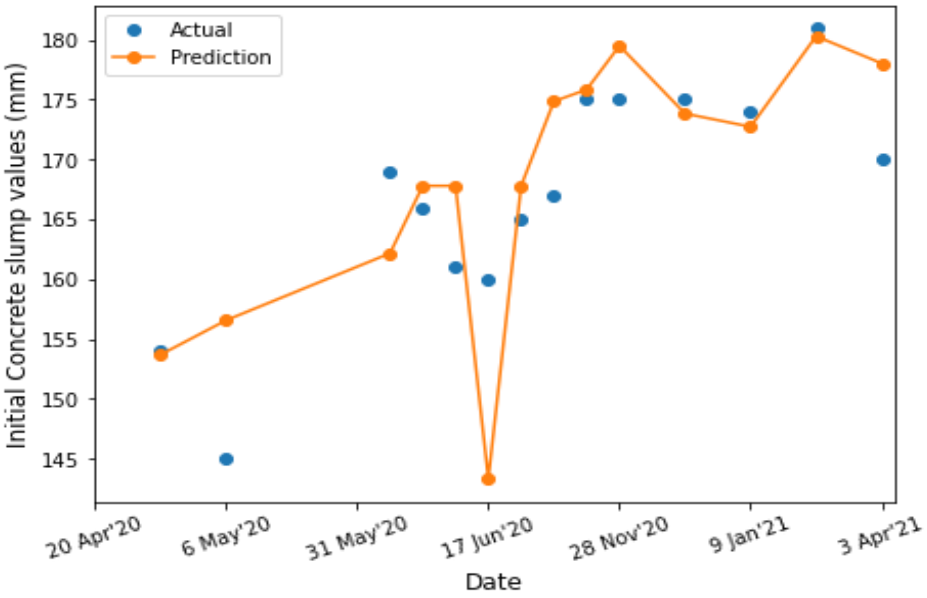


Figure 35: Initial slump prediction on the test set.

A reasonable prediction system was provided by the model. Figure 35 shows that the model's prediction values followed the trend of the actual initial concrete slump values, but some values were underpredicted or overpredicted. Some of the extreme initial concrete slump values can be attributed to how the concrete slump test was conducted and process conditions; however, the model's inability to predict certain values correctly can also be attributed to the optimal hyperparameters obtained. The model hyperparameters obtained through optimization were chosen based on the model achieving the best performance results.

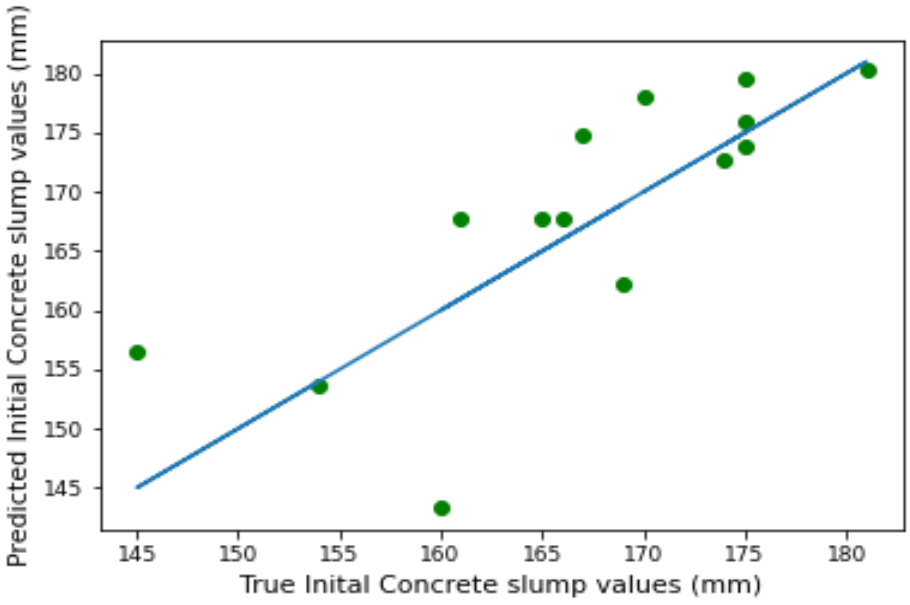


Figure 36: Actual Vs Predicted values for initial slump using the test set.

A low squared correlation value was found, indicating that the model had a fair predicting accuracy for the test set, as shown in Figure 36, such that the model was able to account for 64% variance. However, a considerably higher mean squared error value was obtained, implying that the model made high prediction errors of the actual slump values, as evidenced by data points that are not significantly close to or within an acceptable distance from the fitted regression line.

Improved initial concrete slump model

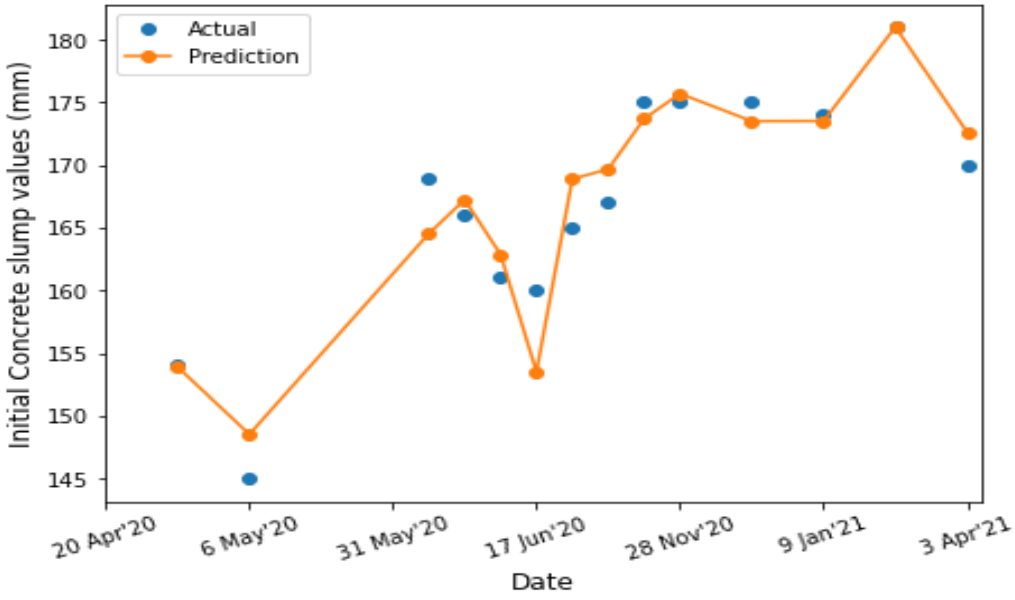


Figure 37: Initial slump prediction with bias updating technique.

Figure 37 shows that by updating the model's bias, the model's shortcomings were reduced to an acceptable level. The updated model provided a reasonable framework for generalization. Figure 38 shows that the model reaches a balance between two extremes, high variance, and high bias, as it does not significantly underfit or overfit the data.

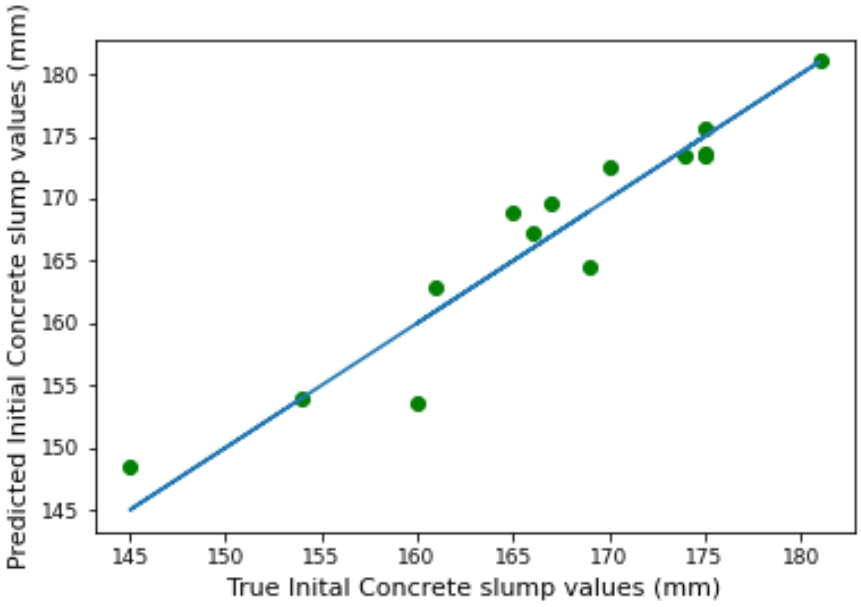


Figure 38: Actual vs Predicted values for initial slump with bias updating technique.

Table 15 : Performance results for corrected Neural net model.

Statistics	
R^2	0.91
MSE	8.1

A high squared correlation value was obtained, thus the model achieved high predictive accuracy for the unseen dataset using the bias updating technique. The model fitted the unknown data to an acceptable degree, as illustrated by the majority of the data points falling close to the fitted regression line, as shown in Figure 38. The mean squared value decreased significantly, indicating that the corrected model predicted the actual slump values with only small errors made.

5.2.2 Insoluble content

The training set (referred to as training dataset 1) had 78 numerical attributes and 3 662 examples after pre-processing and splitting the insoluble content dataset into test and training sets. The dataset was further filtered to include attributes chosen based on knowledge of the process, product, and factors influencing lignosulphonate dispersing capability; this dataset (referred to as training dataset 2) contained 17 attributes and 3 662 examples. The variables included in both training datasets were from chip feeding to evaporator set operations in the lignosulphonate production process.

Feature selection and model selection

The population size of the GA optimization method was varied between 2 and 24 to determine the best features for predicting the insoluble content. Table 16 shows the best performance results obtained for the learning algorithms, as well as the population sizes used to achieve these results.

Table 16: Insoluble content performance results for feature and model selection

	Training dataset 1			Training dataset 2		
Learning algorithm	Population size	Statistic		Population size	Statistic	
		R^2	MSE		R^2	MSE
Support Vector Machine	12	0.51	0.49	2	0.47	0.34
Neural Net	22	0.51	0.84	4	0.48	0.38
Gradient Boosted Tree	3	0.62	0.25	6	0.58	0.3
Random Forest	18	0.64	0.18	4	0.68	0.21
Decision Tree	37	0.42	0.34	6	0.38	0.64
Linear Regression	15	0.36	0.58	2	0.26	1.12

The Gradient Boosted Tree and Random Forest algorithms performed the best among the selected learning algorithms, so these algorithms were trained and validated.

Model training, validation, and optimization

Table 17: Performance results for the trained insoluble content models

Training dataset	Random Forest Model		Gradient Boosted Tree	
	R^2	MSE	R^2	MSE
1	0.84 +/- 0.02 (average: 0.83)	0.09 +/- 0.01 (average: 0.09)	0.72 +/- 0.02 (average: 0.72)	0.19 +/- 0.01 (average: 0.19)
2	0.7 +/- 0.03 (average: 0.7)	0.15 +/- 0.01 (average: 0.15)	0.55 +/- 0.03 (average: 0.58)	0.26 +/- 0.02 (average: 0.26)
1 and 2	0.84 +/- 0.02 (average: 0.84)	0.08 +/- 0.01 (average: 0.08)	0.71 +/- 0.02 (average: 0.71)	0.19 +/- 0.01 (average: 0.26)

The validation results of the trained models are shown in Table 17. The performance results obtained revealed that the trained Random Forest models performed better overall as the learning algorithm obtained high R^2 and low MSE values, which are desirable and a good indication of the models' accuracy. Therefore, the hyperparameters of the Random Forest model were optimized. The optimal set of Random Forest hyperparameters that achieved the best performance for predicting insoluble content is shown in Table 18.

Table 18: Random Forest hyperparameters optimized.

Training dataset	Parameters	Value
1	Number of trees	96
	Maximal depth	20
2	Number of trees	113
	Maximal depth	19
1 and 2	Number of trees	85
	Maximal depth	19

Table 19: Performance results for the optimized Random Forest model.

Dataset	Statistic	
	R^2	MSE
1	0.92 +/- 0.01 (average: 0.92)	0.04 +/- 0.01 (average: 0.03)
2	0.82 +/- 0.03 (average: 0.82)	0.09 +/- 0.02 (average: 0.09)
1 and 2	0.92 +/- 0.01 (average: 0.92)	0.04 +/- 0.01 (average: 0.04)

The performance of the Random Forest models was improved by optimization of the model hyperparameters. The squared correlation values increased by 10%, while the mean squared error decreased by 40 to 60%. Overall, the optimized trained models fitted the training datasets well with the given inputs, establishing a successful relationship between the input variables and the target variable, whereby the model understood the data well and was able to predict the insoluble content values with small prediction errors.

Model evaluation

The performance results of tested Random Forest models are shown in Table 20.

Table 20: Performance results for the tested Random Forest models

Training dataset	Statistic	
	R^2	MSE
1	0.78	0.1
2	0.77	0.09
1 and 2	0.76	0.09

From the performance evaluation results, the 20 input variables chosen based on training dataset 2, aided by the feature selection technique, performed better than the other approaches used to develop the insoluble content prediction model. As a result, the

optimized Random Forest model developed with training dataset 2 was chosen as the final model to predict the insoluble content values used to manipulate the lignosulphonate process to produce lignosulphonate with the desired dispersion capabilities. Figure 39 and Figure 40 show the prediction results of the Random Forest model that was chosen.

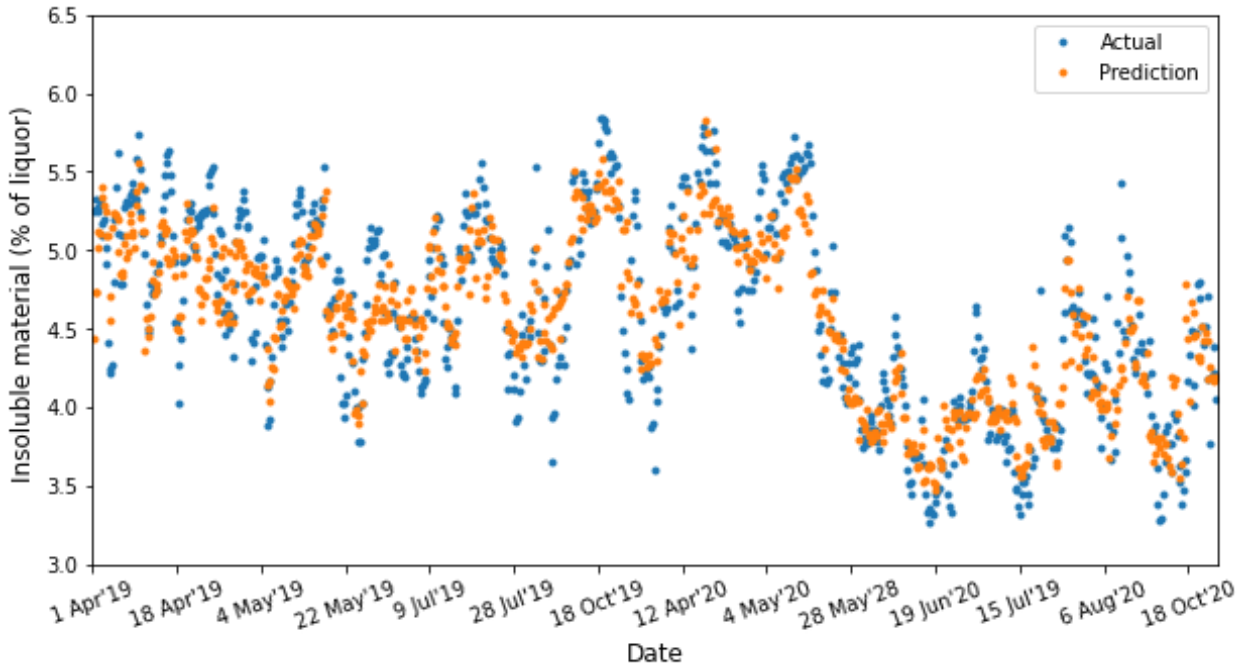


Figure 39: Insoluble content prediction on the test set.

The chosen Random Forest model provided an insoluble content prediction system that was able to capture the trend of the actual insoluble content values; the algorithm learned from the data with the specific input variables chosen and generalised to an appropriate level of knowledge for predicting the insoluble content values. Figure 40 depicts the model's balance between variance and bias, the model does not underfit or overfit the test data.

The model, however, was unable to predict the extreme values, which were values that were less than 4% of liquor and greater than 5.5% of liquor. This can be attributed to the model's optimal hyperparameters. The model's hyperparameters were chosen when the model reached its maximum efficiency; however, the model can be overfitted at high values of maximal depth and number of trees. Therefore, the model predicted the train data relatively well but struggled to adjust to the changes in the new data.



Figure 40: Actual Vs Predicted values for insoluble content using the test set.

A high squared correlation value was obtained shown in Table 20, thus indicating that the model had high predicting accuracy for the test set. The model fits the data well enough to account for 77% of the variance as evidenced by how closely most of the data points have fallen to the regression line in Figure 40. The mean squared error increased by 16% compared to the optimized trained model, thus implying that the model made prediction errors, as demonstrated by some data points not falling significantly close to or on the fitted regression line.

Improved insoluble content model

The corrected Random Forest model resulted in a significantly better prediction system. Figure 41 shows that by updating the model's bias, the overall prediction accuracy of actual insoluble content values improved such that the model fitted the unseen data.

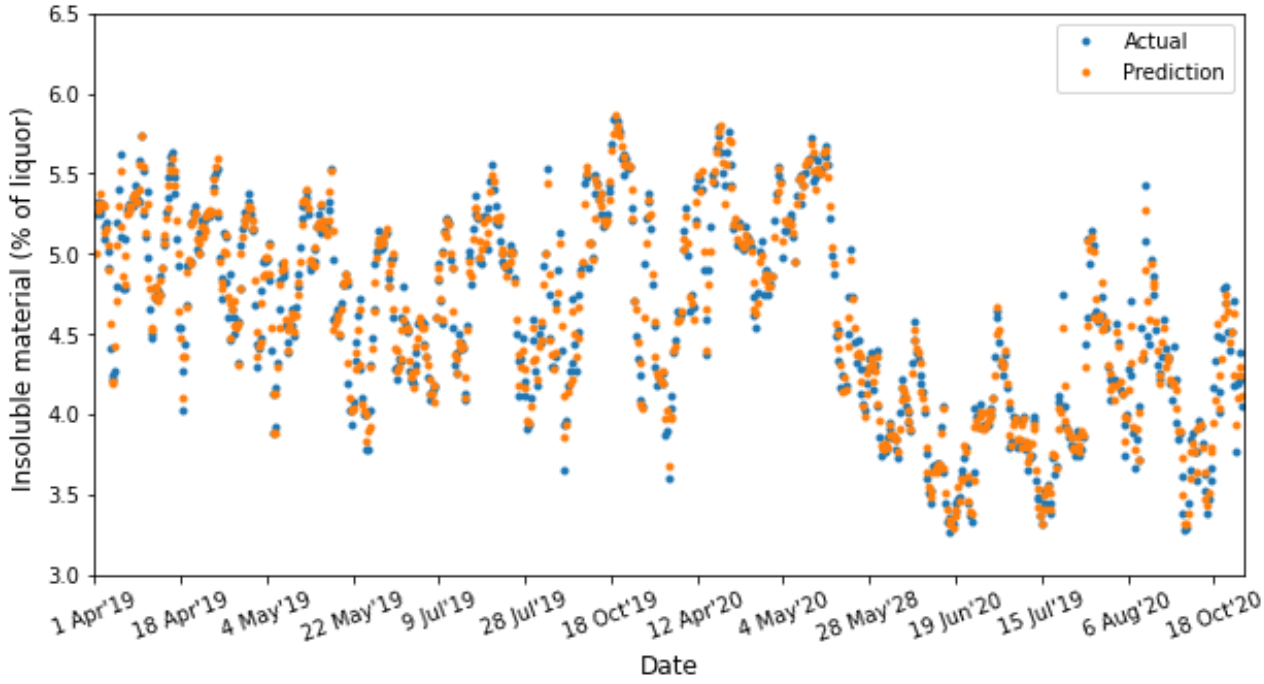


Figure 41: Insoluble content prediction with bias updating technique

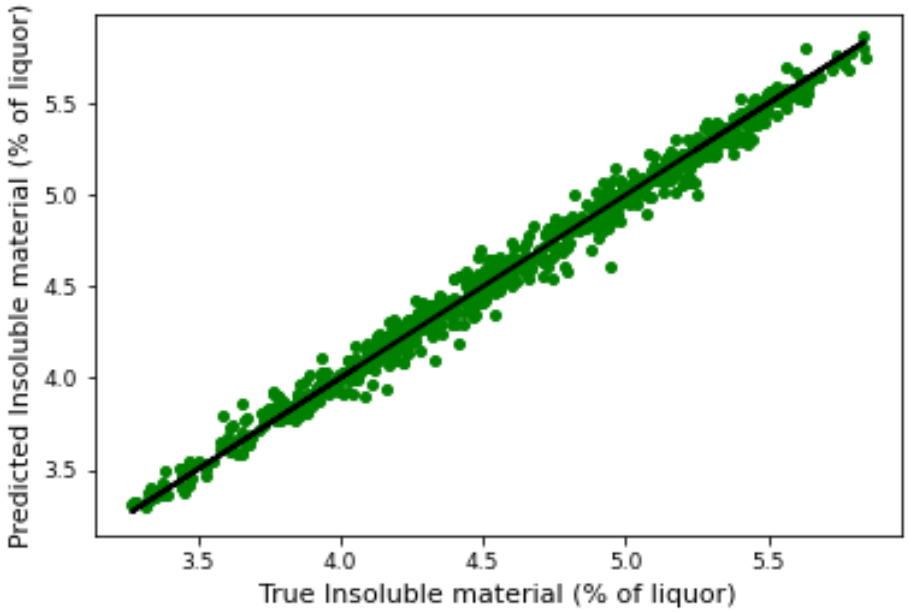


Figure 42: Actual vs Predicted values for insoluble content with bias updating technique.

Table 21: Performance results for the corrected Random Forest model

Statistic	
R^2	0.98
MSE	0.005

As indicated in Figure 42 and Table 21, the performance improved by 20 – 30%, the model provided strong predictive accuracy to the test set, and small prediction errors between the actual insoluble content values were made when the bias updating technique was implemented.

5.2.3 Dispersion index

The training set had 126 numerical attributes and 2 003 examples after pre-processing the dispersion index dataset and splitting it into test and training sets with similar data distribution (referred to as training dataset 1). The dataset was further filtered to include attributes chosen based on process knowledge, resulting in a dataset with 28 attributes and 2 003 examples (referred to as training dataset 2). The variables in both training datasets were from the lignosulphonate production process's chip feeding to spray dryer operations.

Feature selection and model selection

To determine the best features for predicting the dispersion index, the GA method's population size parameter varied from 2 to 38. The table shows the best performance results obtained for the learning algorithms, as well as the population sizes used to achieve these results.

Table 22: Dispersion index performance results for feature and model selection.

	Training dataset 1			Training dataset 2		
Learning algorithm	Population size	Statistic		Population size	Statistic	
		R^2	MSE		R^2	MSE
Support Vector Machine	12	0.41	9.5	8	0.40	9.7
Neural Net	47	0.40	9.7	6	0.41	9.7
Gradient Boosted Tree	23	0.68	4.0	4	0.66	5.2
Random Forest	35	0.71	1.2	8	0.73	1.9
Decision Tree	15	0.47	9.7	4	0.52	9.9
Linear Regression	7	0.39	9.9	2	0.36	10.6

As shown in Table 22, the Gradient Boosted Tree and Random Forest algorithms performed better than the other learning algorithms; thus, these algorithms were chosen for training and validation.

Model training, validation, and optimization

Table 23 shows the validation results of the Random Forest and Gradient Boosted Tree trained models.

Table 23: Performance results for the trained dispersion index model.

Training dataset	Random Forest Model		Gradient Boosted Tree	
	R^2	MSE	R^2	MSE
1	0.88 +/- 0.02 (average: 0.88)	1.53 +/- 0.23 (average: 1.53)	0.76 +/- 0.02 (average: 0.76)	4.26 +/- 0.31 (average: 4.26)
2	0.84 +/- 0.02 (average: 0.84)	2.01 +/- 0.24 (average: 2.01)	0.68 +/- 0.04 (average: 0.68)	5.16 +/- 0.37 (average: 5.16)
1 and 2	0.87 +/- 0.02 (average: 0.87)	1.69 +/- 0.23 (average: 1.69)	0.75 +/- 0.03 (average: 0.75)	4.28 +/- 0.39 (average: 4.27)

As per the validation results, the trained Random Forest models performed better overall based on the input attributes chosen. As a result, the hyperparameters of the Random Forest model were optimized. The best performance for predicting the dispersion index values by implementing the Random Forest learning algorithm was achieved when the following hyperparameters were obtained, shown in Table 24.

Table 24: Random Forest hyperparameters optimized.

Training Dataset	Parameters	Value
1	Number of trees	97
	Maximal depth	18
2	Number of trees	119
	Maximal depth	18
1 and 2	Number of trees	122
	Maximal depth	17

Table 25: Performance results for optimized Random Forest model

Training Dataset	Statistic	
	R^2	MSE
1	0.93 +/- 0.02 (average: 0.93)	0.88 +/- 0.17 (average: 0.88)
2	0.89 +/- 0.01 (average: 0.9)	1.4 +/- 0.16 (average: 1.4)
1 and 2	0.93 +/- 0.01 (average: 0.92)	1.0 +/- 0.18 (average: 1.0)

The performance was significantly improved by optimization. The squared correlation values improved by 5%, and the mean squared error decreased by 28 to 60%. The optimized models established a successful relationship between the input variables and the target variable, allowing the model to understand the data well and predict dispersion index values with low prediction errors made.

Model evaluation

The evaluation results obtained for Random Forest models are shown in Table 26.

Table 26: Performance results for tested Random Forest models

Training Dataset	Statistic	
	R^2	MSE
1	0.73	2.6
2	0.64	3.3
1 and 2	0.66	3.5

According to Table 26, the input variables chosen based on training dataset 1 significantly outperformed the approaches used to develop the dispersion index prediction model. As a result, the optimized Random Forest model developed with training dataset 1 was chosen as the final model to predict the dispersion index values used to assess lignosulphonate dispersion capability. Figure 43 and Figure 44 show the prediction results of the final Random Forest model that was chosen.

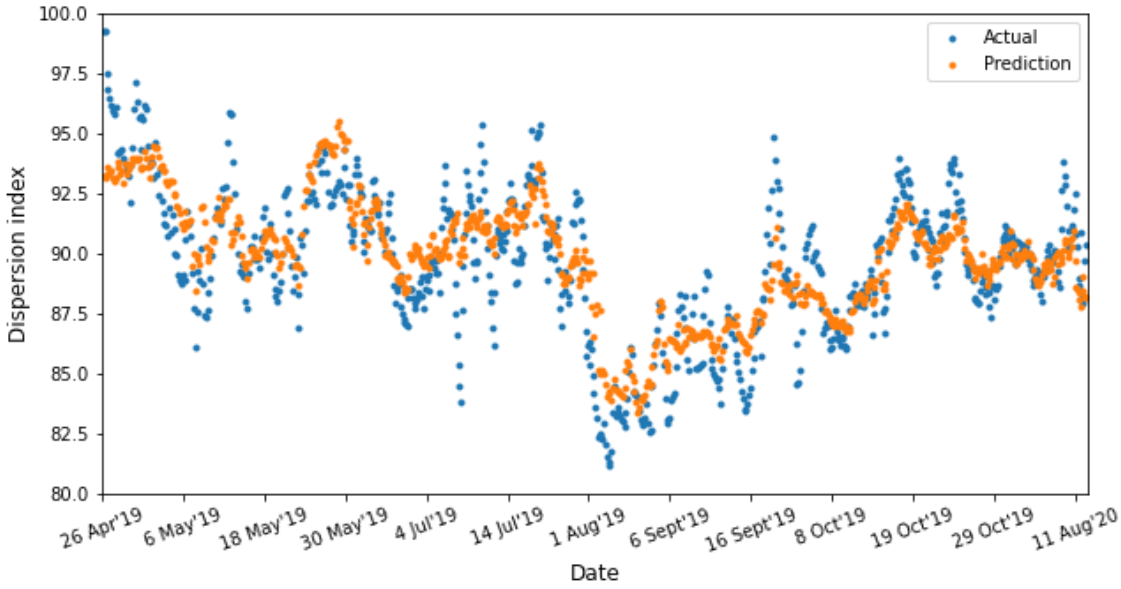


Figure 43: Dispersion index prediction on the test set.

A reasonable prediction system was provided by the model. Figure 43 shows that prediction values obtained followed the trend observed by the actual dispersion index values, but the model was unable to predict the most extreme values. This is partly related to the optimal hyperparameters obtained. The model hyperparameters obtained through optimization were chosen so that the model performed optimally. Therefore, it can be said that the model learned and fit the test set to a reasonable degree, but it was unable to adapt fully to the test set.

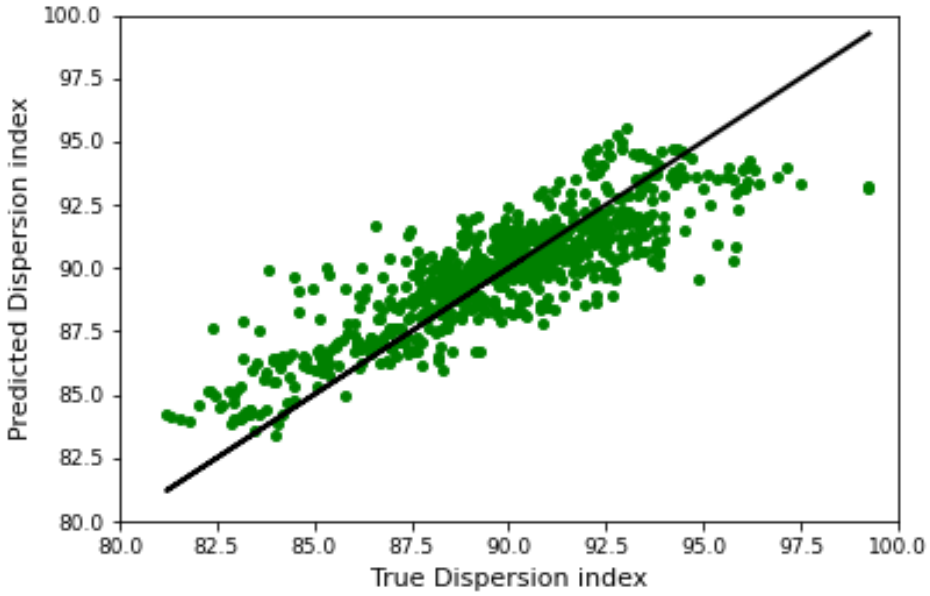


Figure 44: Actual Vs Predicted values for dispersion index using the test set.

A reasonably high squared correlation value was obtained, indicating that the model had good predicting accuracy for the test set and only 27% of the variance was unexplained by the model, as shown in Figure 44, where it was observed that the majority of the data points have fallen close to the fitted regression line. The model, however, produced a relatively high mean squared error value, resulting in high prediction errors made of the actual dispersion index values.

5.2.3.1 Improved dispersion index model

The improved Random Forest model resulted in significantly better prediction performance.

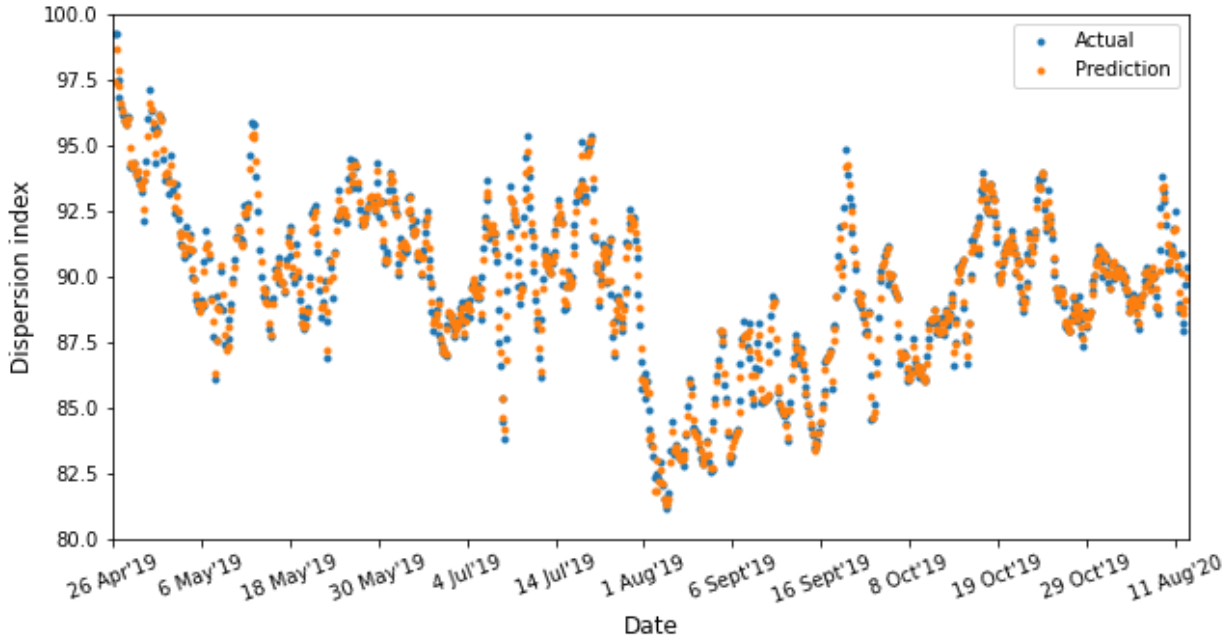


Figure 45: Dispersion index prediction with bias updating technique.

Figure 45 shows that after updating the model's bias, it was able to account for and predict the extreme dispersion index values. The revised model provided a basis for generalization. Figure 46 shows how well the model fitted the unknown data and an almost perfect correlation between the predicted and actual dispersion index values was achieved.

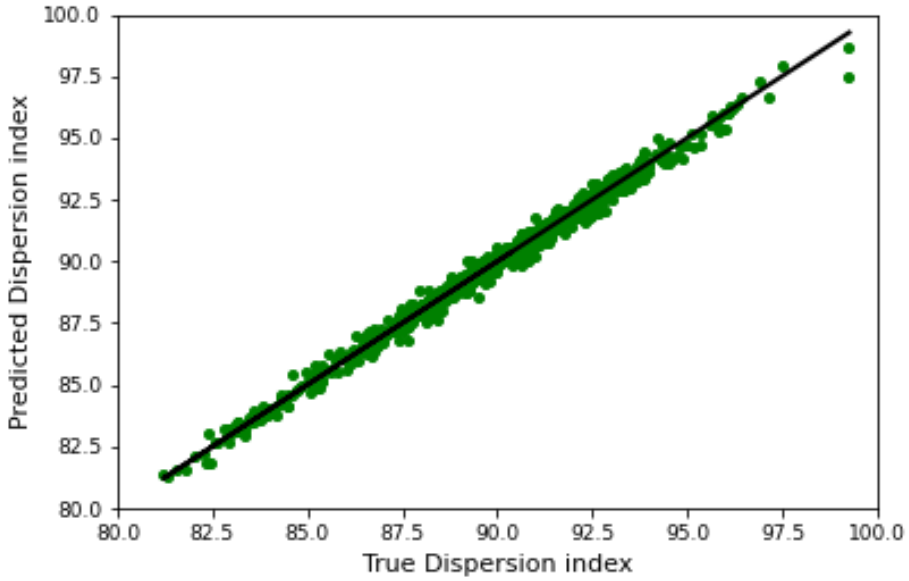


Figure 46: Actual vs Predicted values for dispersion index with bias updating technique.

Table 27: Performance results for corrected Random Forest model.

Statistics	
R^2	0.98
MSE	0.08

The performance results improved significantly compared to the model's previous performance results, the squared correlation increased by 25.0% and approximately 95.0% improvement in the mean squared error. Thus, the model had high predictive accuracy for the test set, with only minor errors in predicting the actual dispersion index values and representing a better-quality model as the MSE value obtained is close to zero.

6 Conclusion and recommendations

The Sappi Tugela mill produces lignosulphonate, which is used as a surfactant and dispersant in the concrete and cement admixture markets. To predict the dispersing capabilities of the product and align them to market demands, soft sensors were developed to model the effects of process parameters on the dispersion capabilities of lignosulphonate. Soft sensors are data-driven predictive models that estimate a target variable in real-time. To determine the dispersion performance of the lignosulphonate product at the mill, three offline methods are used: dispersion index, concrete slump, and insoluble content tests.

The study was divided into two sections; Part A of the study involved developing a function to predict concrete slump values using a regression modelling technique, and Part B involved developing models to predict insoluble content, dispersion index, and initial concrete slump values based on the lignosulphonate production process parameters.

The concrete slump data consisted of process data related to the dispersant's manufacturing process and five concrete slump measurements were taken at 15-minute intervals, beginning with an initial slump value. The slump dataset was subjected to normalization, allowing it to be represented by a simple quadratic function that was dependent on time and the initial slump value. The function could predict the four concrete slump measurements if the initial slump value was affected by process variables.

The selected function achieved strong validation results, with a square correlation value of 0.92 and a mean squared error of 11.7. The function successfully established a relationship between the input variables and the target variables, allowing it to predict slump values at varying time intervals with small prediction errors; however, the model's prediction ability decreased at times 45 and 60 minutes when compared to performance results obtained at times 15 and 30 minutes. The function’s evaluation results showed an overall squared correlation value of 0.91, a drop of 10.4%, and a mean squared error value of 13.6, an increase of 13.2%.

The predictability of the model deteriorated; slump values at 60 minutes provided the greatest performance reduction. At this time, the poor performance can be attributed to a wide variation in concrete slump values that the model was unable to capture adequately, and the variation in values could be due to the conditions under which the slump test was performed. To account for the shortcomings of the developed function, the bias updating technique was used to improve the findings of the selected function. The model's overall

performance improved significantly, with an 8.4% increase to a squared correlation value of 0.98 and a 90.4% decrease to a mean squared error value of 1.3.

RapidMiner was used to create numerical predictive models. To provide a better idea of how well the model would perform in real-life applications, the developed models were evaluated on test sets that went through the same pre-processing steps and had a similar data distribution as the training set. To improve the evaluation results of the selected models, the models were further improved using an adaptive procedure such as bias updating.

The models were trained using three cases to achieve the goal of developing predictive lignosulphonate dispersion capability models based on process and product knowledge, as well as data analysis, in accordance with the objectives of this research study. This approach was used to demonstrate the importance and influence of various variables selected as input to the models. The Genetic Algorithm (GA) optimization method was used to select the most significant variables influencing the target variable and to tune model parameters.

To determine the initial concrete slump values, a Neural Net model with ten input variables was chosen. The model’s input variables came from the lignosulphonate production process's chip feeding to the spray dryer operations. The model performed well in validation, with a squared correlation value of 0.81 and a mean squared error value of 19.7. However, when the model was evaluated on the test set, it yielded a low squared correlation value of 0.64 and a high mean squared error value of 47.6. The low evaluation results could be attributed to how the concrete slump test was performed and process conditions; however, the model's inability to correctly predict certain values could also be attributed to the optimal hyperparameters obtained.

Overall, the initial concrete slump model provided a reasonable prediction system, and further improvement of the model through bias updating reduced the model's shortfalls to an acceptable level. The updated model provided a reasonable framework for generalization; it achieved a balance between the two extremes, high variance, and high bias, by not significantly underfitting or overfitting the data. The improved model predicted the initial concrete slump values with small errors, as evidenced by a high squared correlation value of 0.91 and a mean squared error value of 8.1.

To predict the dispersion index values, a Random Forest model with 20 input variables taken from the chip feeding to spray drying operations of the lignosulphonate production process was used. The model performed well in terms of validation, with a squared correlation value

of 0.93 and a mean squared error of 0.88. The optimized model successfully established a relationship between the input variables and the target variable, allowing the model to grasp the data and anticipate dispersion index values with low prediction errors. However, the model's performance declined, with a squared correlation value of 0.73 acquired, a 21.5% decrease, and a mean squared error of 2.6 achieved, a 66.2% rise. These evaluation results might be attributed to the obtained prediction values, which followed the trend of the actual dispersion index values but were unable to predict the most extreme values. This could be connected in part to the ideal hyperparameters obtained. The model hyperparameters were chosen in such a way that the model performed optimally.

As a result, the dispersion index model learned and fitted the test set to a respectable extent, but it was unable to properly adapt to the test set due to overtraining or over-specified hyperparameters. Bias update enhanced model performance, and this method resulted in the model accounting for and predicting extreme dispersion index values. The updated model served as a foundation for generalization. The performance results improved to the point where the squared correlation climbed by 25%, yielding a value of 0.98, and the mean squared error improved by about 95%, yielding a value of 0.08. Therefore, the model had high predictive accuracy for the test set, with just minimal mistakes in predicting the actual dispersion index values and represented a higher-quality model because the MSE value achieved was close to zero.

Finally, a Random Forest model with 20 input variables based on the process's chip feeding and evaporator set operations was used to predict the insoluble content values of the lignosulphonate liquor. A squared correlation value of 0.82 and a mean squared error of 0.09 were achieved, indicating that the model fitted the training set well and created a satisfactory link between the input variables and the target variable, resulting in minor prediction errors. When evaluated on the test set, the model provided an insoluble content prediction system that was able to capture the trend of the actual insoluble content values; the algorithm learned from the data with the specific input variables chosen and generalized to an appropriate level of knowledge for predicting the insoluble content values. Extreme values, on the other hand, were unlikely anticipated, which may be attributed to the model's optimal hyperparameters.

The insoluble content model was evaluated and yielded a high squared correlation value of 0.77, however, the mean squared error increased by 16% when compared to the optimal trained model, showing that the model made prediction errors. Improving the model's predictability resulted in a considerably better prediction system. When the bias updating technique was used, the performance improved by 20 – 30%, the model offered good

predictive accuracy to the test set, and modest prediction errors between the actual insoluble content values were made. Using the adaptive approach, a squared correlation value of 0.98 and a mean squared error of 0.05 were attained.

The study's proposed approach proved that machine learning and data mining techniques can be utilized to uncover valuable correlations or relationships that are not direct indicators of the target attribute. For the sample size and number of attributes employed, Neural Net and Random Forest have proven to be effective learning algorithms. This study developed a useful prediction strategy for predicting insoluble content, dispersion index, and concrete slump values, decreasing reliance on laboratory results and allowing for almost immediate modifications to the lignosulphonate production process.

The method used demonstrated the significance of model parameter settings in the training process, their impact on prediction performance, the significance and influence of selecting various variables as input to the models, and the utility of the Genetic Algorithm (GA) optimization method for attribute selection and hyperparameter determination to improve model efficiency.

Furthermore, the data can be used to undertake additional research to improve the prediction system's accuracy, model selection, and attribute selection. Future work for this study includes acquiring a larger dataset of concrete slump values to provide a better understanding of the repeatability and variability of the values obtained from the concrete slump test, as well as a better understanding of the observations and conclusions made regarding the value of the regression coefficient.

A larger dataset allows for the implementation of complicated algorithms, which improves efficiency and allows for a higher and larger selection of model parameters and attribute selection. A larger selection of input variables could be used to better capture the complicated relationships or correlations that exist with the target variable to improve the predictability of the insoluble content and dispersion index models. Adaptive procedures, such as learning algorithms, could be examined to improve soft sensor performances. These algorithms can be implemented online and allow the model to adapt to changes in the process, offering useful estimations of the target variables as the process expands or changes.

7 References

AlBanna, GA (2016), “Data Mining Techniques Implementation To Improve Healthcare Among Diabetic Patients”, Master’s thesis, The British University, Dubai, UAE.

Ambica Sales Agency (216), “Spray Drying Process”, https://ambicasalesblog.wordpress.com/2016/06/22/spray-drying-process/ [2021, September 3].

Antonides, F (2000), “Simultaneous Neutral Sulphite Semichemical Pulping of Hardwood and Softwood”, Master’s Thesis, University of Kwa-Zulu-Natal, Durban, South Africa.

Area, MC, Felissa, FE, Venica, A and Valade, JL (2001),” NSSC Process Optimization: Pulping, Pulps and Spent Liquors”, *TAPPI Journal*, Vol 84: No. 4.

Areskogh, D (2011), “Structural Modifications of Lignosulphonates”, PhD Thesis, Royal Institute of Technology, Stockholm, Sweden.

Aro, T and Fatehi, P (2017), “Production and Application of Lignosulphonates and Sulfonated Lignin”, *ChemSusChem*, 10, 1861-1877.

Arunadevi, J and Nithya, MJ (2016) “Comparison of Feature Selection Strategies for Classification using Rapid Miner”, *International Journal of Innovative Research in Computer and Communication Engineering*, 4(7), 556-563.

Bajpai, P (2018), “Biermann’s Handbook of Pulp and Paper: Volume 1: Raw Material and Pulp Making”, Elsevier Publishing, Amsterdam.

Batista G, Prati, RC and Monard, MC (2004) “A study of the behavior of several methods for balancing machine learning training data”, *ACM SIGKDD Explorations Newsletter*, 6(1), 20-29.

Blachnik, M and Kordos, M (2020) “Comparison of Instance Selection and Construction Methods with Various Classifiers”, *Applied Sciences*, 10 (3933), 1-19.

Boehmke, B and Greenwell, B (2020) Hands-On Machine Learning with R, CRC Press, Florida, USA.

Boughton, LD, Pavlich, JP and Wahl, WW (1962), “The Use of Dispersants In Cement Slurries To Improve Placement Techniques”, paper presented at Fall Meeting of the Society of Petroleum Engineers of AIME, 7-8 October 1962, Los Angeles, United States of America.

Caudill, M (1987) “Neural networks primer, part I”, *AI Expert*, 2, 46-52.

Chapelle, V, Vapnik, V, Bousquet, O and Mukherjee, S (2002) “Choosing Multiple Parameters for Support Vector Machines”, *Machine Learning*, 46, 131-159.

Cobb, P, & Gravemeijer, KPE. (2008),” Experimenting to support and understand learning processes” In Kelly, AE and Lesh, RA (Eds.), Handbook of design research methods in education innovations in science, technology, engineering, and mathematics learning and teaching, Routledge, New York: 68-95.

Costa JFP (2011) “Weighted Correlation”, in Lovric M. (eds) International Encyclopaedia of Statistical Science. Springer, Berlin.

Côte, W.A., Jr. (1967), “Wood Ultrastructure”, University of Washington Press, Syracuse, NY, USA.

Curreri, F, Fiumara, G and Xibilia, MG (2020), “Input Selection Methods for Soft Sensor Design: A Survey”, *Future Internet*, 12, 97.

de Rooij, M and Weeda, W (2019) “Cross-validation: A method Every Psychologist Should Know”, *Advances in Methods and Practices in Psychological Science*, 3(2), 248-263.

de Wet-Roos, D (2016), “Polymers and chemicals derived from lignin”, Sappi Technology Centre, Pretoria, South Africa.

Diez, D M, Barr, CD, and Çetinkaya Rundel, M (2014), “Introductory Statistics with Randomization”,

Du, C and Sun, D (2008), “Object Classification Methods*”, Computer Vision Technology for Food Quality Evaluation*, 81-107.

Eitrich, T and Lang, B (2005) “Efficient optimization of support vector machine learning parameters for unbalanced datasets”, *Journal of Computational and Applied Mathematics*, 196, 425-436.

Fibo Intercon (2019), “Concrete Workability and the Slump Test”, https://fibointercon.com/articles/concrete-workability/ [2021, September 3].

Firouzabadi, MD and Hatam, A (2013), “Modelling NSSC Pulping to Predict and Optimize Pulp Yield”, *Wood Research*, 59(5), 739-746.

Flatt, RJ and Schober, I (2012), “Understanding the Rheology of Concrete”, Chapter 7: Superplasticizers and the rheology of Concrete, Woodhead Publishing, Cambridge.

Forests NSW (2008), “About wood, Primefact 541”, NSW Department of Primary Industries, New South Wales, Australia.

Ganjisaffar, Y, Caruana, R and Lopes, CV (2011) “Bagging gradient-boosted trees for high precision, low variance ranking models”, paper presented at the 34th international ACM SIGIR conference on Research and development in Information Retrieval, New York, USA, 85–94.

Gasper, P, Carbonell, J and Oliveira, JL (2012) “On the parameter optimization of Support Vector Machines for binary classification”, *Journal of Integrative Bioinformatics*, 9(3), 201-212.

Geetha, A and Nasira, GM (2014), “Artificial Neural Networks’ Application in Weather Forecasting – Using RapidMiner*”, International Journal of Computational Intelligence and Informatics*, Vol.4: No. 3, 177-182.

Gray, K, van Zyl, T and Celik, T (2016), “Virtual Wind Sensors: Improving Wind Forecasting Using Big Data Analysis”, Master’s thesis, University of the Witwatersrand, Johannesburg, South Africa.

Hanhikoski, S (2014), “High yield nucleophile cooking of wood chips”, Master’s Thesis, VTT Technical Research Centre of Finland, Espoo, Finland.

Harker, WG (2013), “Real-Time Furnace Froth State Detection using Hidden Markov Models”, Master’s dissertation, University of the Witwatersrand, Johannesburg, South Africa.

Henriksson, G (2009), “Pulp and Paper Chemistry and technology-Wood Chemistry and Wood Biotechnology”, Walter de Gruyter, Berlin.

Holm, A and Niklasson, R (2018), “The effect on wood components during soda pulping”, Master’s Thesis, Chalmers University of Technology, Gothenburg, Sweden.

Huang, J, Fu, S and Gan, L (2019), “Lignin Chemistry and Applications”, Elsevier Inc, Amsterdam, Netherlands.

IETLS (2011), “How to describe an image that depicts a process on your IELTS Task 1 question response”, http://ieltsielts.com/how-to-describe-an-image-that-depicts-a-process-on-your-ielts-task-1-question-response/ [2020, November 04].

Ingruber, OV, Kocurek, MI and Wong, A (1985), “Sulphite Science & Technology: Pulp & Paper Manufacture”, Vol. 4, 3rd ed., The Joint Textbook Committee of the Pulp & Paper Industry, McGraw-Hill, New York.

Jo, JM (2019) “Influence of the pre-processing process of normalization of big data on the performance of machine learning”, *The Journal of the Korea Institute of Electronics and Communication Sciences,*14(3), 547–552.

John Godrich (2017), “Kern Moisture Analysers”, https://johngodrich.co.uk/product/kern-moisture-analaysers/ [2021, September 3].

Kamm, B and Kamm, M (2004), “Principles of Biorefineries”, *Applied Microbiology and Biotechnology*, 64(2), 137-145.

Kamm, B, Gruber, P and Kamm, M (2006), “Biorefineries – Industrial Processes and Products: Status Quo and Future Directions”, Wiley – VCH, Weinheim.

Keskin-Schneider, A (1991), “Engineering Reaction Kinetics in Sulfite and Sulfite-Anthraquinone pulping”, PhD Thesis, McGill University, Montreal, Canada.

Kilian, A (1999), “Control of an acid sulphite batch pulp digester based on a fundamental process model”, Master’s Thesis, University of Pretoria, Pretoria, South Africa.

Kocurek MJS and Stevens CFB (1983), “Sulphite Science & Technology: Pulp and Paper manufacture”, Vol. I. The Joint Textbook Committee of the Paper Industry, McGraw-Hill, New York.

Koehler, EP and Fowler, DW (2003), “Summary of Concrete Workability Test Methods”, Research report by International Center of Aggregates Research, University of Texas, Austin, United States of America.

Kotu, V and Deshpande, B (2015), “Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner”, Elsevier Inc, Amsterdam, Netherlands.

Liu, Z, Wang, H and Hui, L (2018), “Pulping and papermaking of non-wood fibers”, IntechOpen, London.

Llyas, IF and Chu, X (2019),” Data Cleaning, Association for Computing Machinery”, New York.

May, RJ, Maier, HR and Dandy, GC (2010) “Data splitting for artificial neural networks using SOM-based stratified sampling*”, Neural Networks*, 23(2), 283-294.

Mercangöz, M and Doyle, FJ (2008) “Real-time optimization of the pulp mill benchmark problem”, *Computers and Chemical Engineering*, 32, 789-804.

Mierswa, I (2018) “Better Machine Learning Models with Multi-Objective Optimization”, https://rapidminer.com/resource/webinar-better-machine-learning-models-multi-objective-optimization/ [2021, February 26].

Mierswa, IW (2006), “YALE: Rapid prototyping for complex data mining tasks”, *Association for Computing Machinery – Knowledge Discovery in Databases*, 935-940.

Miner, G, Elder, J, Fast, A, Hill, T, Nisbet, R and Delen, D (2012), “Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications”, Academic Press, Waltham.

Mishra, G (2016), “Concrete Slump Test for Workability-Procedure and Results”, https://theconstructor.org/concrete/concrete-slump-test/1558/ [2020, August 18].

Mollereau, A (2005), “Real-time observer Model for a Kraft wood digester”, Master’s Thesis, University of Kwa-Zulu Natal, Durban, South Africa.

Moodley, B (2001), “Characterisation of Sappi Saiccor Pulp Mill’s Effluent”, Master’s Thesis, University of Kwa-Zulu Natal, Durban, South Africa.

Moore, MM, Slonimsky, E, Long, AD, Raymond, SW and Ramesh, LS (2019) “Machine learning concepts, concerns and opportunities for a pediatric radiologist”, *Pediatr Radiol*, 49, 509–516.

Morariu, N, Iancu, E and Vlad, S (2009) “A neural network model for time series forecasting”, *Romanian journal of economic forecasting*,12, 213-223.

Mu, S, Zeng, Y, Liu, R, Wu, P, Su, H and Chu, J (2006) “Online dual updating with recursive PLS model and its application in predicting crystal size of purified terephthalic acid (PTA) process”, *Journal of Process Control*, 16, 557-566

Nagrockiene, D, Pundiene, I and Kicaite, A (2013), “The effect of the cement type and plasticizer addition on concrete properties”, *Construction and Building Materials*, 45(2013), 324-331.

Neter, J, Kutner, KH, Nachtsheim, CJ and Li, W (2005), “Applied linear statistical models”, McGraw-Hill, Irwin.

Oliveri, G and Massa, A (2011) “Genetic algorithm (GA)-enhanced almost difference set (ADS)-based approach for array thinning”, *IET Microwaves, Antennas and Propagation*, Vol. 5(3),305–315.

Oshiro, TM, Perez, PS and Baranauskas, JS (2012) “How many Trees in a Random Forest?”, Lecture notes: Computer Science, University of Sao Paulo, 29 June 2012, Sao Paulo, Brazil.

Ouyang, X, Qui, X and Chen, P (2006), “Physiochemical characterisation of calcium lignosulphonate-A potentially useful water reducer”, *Colloids and Surfaces A: Physiochem. Eng. Aspects*, 282-283(2006), 489-497.

Owusu, PA, Asumadu-Sarkodie, S (2016), “A review of renewable energy sources, sustainability issues and climate change migration”, *Cogent Engineering*, 3:1167990.

Panthong, R and Srivihok, A (2015) “Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm”, *Procedia Computer Science*, 72, 162-169

Parthiban, G (2014), “Prediction of Risk of Heart Disease for Diabetic Patients using Data mining”, PhD Thesis, Dr MGR Educational and Research Institute, Chennai, India.

Prettenhofer, P and Louppe, G (2014), “Gradient Boosted Regression trees”, Lecture notes, University of Liege, delivered 24 February 2014.

Quelhas, AD and Pinto, JC (2009) “Soft senor models: Bias updating revisited*”, IFAC Proceedings Volumes*, 42(11), 679-684.

Rakala, N, Subasi, MM and Subasi, E (2020) “Evolutionary Feature Selection for Machine Learning”, *SAS Global Forum*, 5135.

Rodrigues, J (2019), “Seven Must-Use Concrete Admixtures (Additives)”, https://www.thebalancesmb.com/common-used-concrete-admixtures-845036 [2020, 01 November].

Roussel, N (2012), “Understanding the Rheology of concrete”, Woodhead Publishing Limited, Cambridge, England.

Rydholm, S. A. (1965), “Pulping processes”, 1st ed., Interscience Publications, London.

Sappi (2005), “Tugela Technical Manual”, Sappi Tugela Mill, Mandeni, South Africa

Sappi (2016),” Sappi Fact Sheet – Lignex”, Sappi Ltd, Johannesburg, South Africa

Sappi (2019),” Sappi and Borregaard celebrate 20 years of biorefinery investment”, Sappi Ltd, Johannesburg, South Africa

Sappi (2020), “Saiccor Technical Manual”, Sappi Saiccor Mill, Umkomaas, South Africa.

Sappi LQM/BIOSC/M028 (2018), “Dispersion index of lignosulphonate samples using ZnO”, Sappi Technology Centre, Pretoria, South Africa.

Sappi W728i004.TUG (2014), “Determination of insoluble content in lignosulphonate liquor using 10% dilution method”, Sappi Tugela Mill, Mandeni, South Africa.

Sappi W728i005.TUG (2017), “Determination of %Solids using Infrared (IR)”, Sappi Tugela Mill, Mandeni, South Africa.

Sappi W728i023.TUG (2019), “Performance Testing of Lignosulphonate: Micro-Concrete Test”, Sappi Tugela Mill, Mandeni, South Africa.

Seborg, DE, Edgar, TF, Mellichamp, DA and Doyle, FJ (2011), “Process Dynamics and Control”, Third Edition, International Student Version, John Wiley & Sons, Hoboken.

Shapiro, J and Blackman, R (2020), “Four steps for drafting an ethical data practices blueprint”, https://techcrunch.com/2020/07/24/four-steps-for-an-ethical-data-practices-blueprint/ [2020, August 2].

Sharmin, R, Sundararaj, U, Shah, S, Griend, LV and Sun, Y (2006) “Inferential sensors for estimation of polymer quality parameters - Industrial application of a PLS based soft sensor for an LDPE plant”, *Chemical Engineering Science*, 61, 6372-6384.

Shin, T (2021), “A Mathematical Explanation of Support Vector Machines”, https://towardsdatascience.com/a-mathematical-explanation-of-support-vector-machines-e433ffe04362 [2021, October 02].

Singh, A (1997), “Modeling and Model Updating in the Real-Time Optimization of Gasoline Blending”, Master thesis, University of Toronto, Toronto, Canada.

Sixta, H, (1998), “Comparative evaluation of different concepts of sulfite pulping technology”, *Lenzinger Berichte*, 78, 18–27.

Sixta, H (2006), “Handbook of Pulp”, Wiley, New Jersey.

Sjöström, E. (1993), “Wood Chemistry. Fundamentals and Applications”, 2nd edition, Academic Press, San Diego, CA, USA, 293 p.

Sliškovíc, D, Grbíc, R and Hocenski, Z (2011) “Methods for Plant Data-Based Process Modeling in Soft-Sensor Development”, *AUTOMATIKA*, 52(4), 306-318.

Smith, LN (2018), “A Disciplined Approach to Neural Network Hyper-Parameters: Part 1 – Learning Rate, Batch Size, Momentum, and Weight Decay”, US Naval Research Laboratory Technical Report, 5510-026.

Smola, A and Schölkopf, B (1998),” A Tutorial on Support Vector Regression”, NeuroCOLT Technical Report NC-TR-98-030, Royal Holloway College, University of London, United Kingdom.

Smook GA (1992), “Handbook for pulp and paper technologists”, 2nd edition. Angus Wilde Publications, Vancouver, USA.

Soares, S and Araújo, R (2011), “Design and Application of Soft Sensor in the Paper Pulp Industry Using Small Datasets”, Proc. 16th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA 2011), 1-8.

Spiegel, M, Schiller, J and Srinivasan, A (2001), “Schaum’s Easy Outlines of Probability and Statistics”, McGraw-Hill, New York.

Sweta, K (2020), “Introduction to Decision tree”, https://blog.probyto.com/introduction-to-decision-tree/ [2021, October 02].

Tharwat, A and Gabel, T (2019) “Parameters optimization of support vector machines for imbalanced data using social ski driver algorithm”, *Neural Computing and Applications*, 32(69), 25-38.

Tingle, M (2019) “Preventing Data Leakage in Your Machine Learning Model”, https://towardsdatascience.com/preventing-data-leakage-in-your-machine-learning-model-9ae54b3cd1fb [2021, March 31].

Tran, M, Varvarezos, DK and Nasir, M (2005) “The importance of first-principles, model-based steady-state gain calculations in model predictive control - a refinery case study”, *Control Engineering Practice*, 13, 1369- 1382.

Wang, D (2011), “Basis Lignin Chemistry”, National ZTE University, Taiwan.

Wang, LKP (1965), “Effect of calcium lignosulphonate on properties of concrete at early ages”, Master’s Thesis, University of Missouri, Rolla, United States of America.

Wang, Z and Chiang, L (2018), “Hard and Soft Sensors Fusion for Process Monitoring: An Industrial Application”, paper presented at ISA Analysis Division Symposium, 22 April 2018, at Galveston, Texas.

Werbos, P (1974), “Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences”, PhD Thesis, Harvard University, Boston, USA.

Werbos, PJ (1994), “The roots of backpropagation: from ordered derivatives to neural networks and political forecasting”, Wiley-Interscience, Hoboken.

Wiedenhoeft, A (2010), “Wood Handbook-Wood as an Engineering Material”, USDA, Madison, United States of America.

Ye, J, Chow, JH, Chen, J and Zheng, Z (2009) “Stochastic gradient boosted distributed decision trees”, paper presented at the 18th ACM conference on Information and knowledge management (CIKM '09), New York, USA, 2061–2064.

Yildirim, S (2020), “Gradient Boosted Decision Trees- Explained”, https://towardsdatascience.com/gradient-boosted-decision-trees-explained-9259bd8205af [2021, October 02].

Youngs, RI (2009), “History, Nature and Products of wood”, Forests and Forest plants Vol 2, EOLSS Publishers, Oxford, United Kingdom.

Zhang, S, Chu, F, Deng, G and Wang, F (2019) “Soft Sensor Model Development for Cobalt Oxalate Synthesis Process Based on Adaptive Gaussian Mixture Regression”, *IEEE Access*, 7, 749 -763.

Zhu, X, Rehman, KU, Wang, B and Shahzad, M (2020), “Modern Soft-Sensing Modelling Methods for Fermentation Processes”, *Sensors*, 20, 1771.

Appendix A: Fundamentals of raw materials for pulping

Wood is an incredibly valuable natural resource, that has been used and adapted by humans throughout history. It is a complex biological structure that is made up of a variety of chemicals and cell types, that work together to fulfil the requirements of a living plant. Over several years, wood has evolved to fulfil three key functions of a living plant, which are to conduct water from the roots to the leaves, to provide the plant body with mechanical support, and to store and synthesize biochemicals (Wiedienhoeft, 2010).

The various important properties that wood exhibits, make it a useful material that can be used in many different industries such as pulp and paper, textiles, wood composite production and construction. Large areas of land have been set aside and managed solely to sustain global timber production. Forests must be managed sustainably to provide a steady supply of wood to meet current and future needs (Forests NSW, 2008).

The biological structure and composition of wood is the main raw material to produce pulp and lignosulphonate. The chemical and biological components of the wood are essential for the pulping industry and lignosulphonate production; therefore, it will be discussed in detail.

A.1 Biological composition of wood

Wood usually implies the trunk of living, growing trees and performs the role of support and transport (Forests NSW, 2008). Support allows the tree to remain upright despite the heights reached and the transportation of water and nutrients from the bottom to the top of the tree. Wood consists of four layers, shown in Figure 47, and each layer is composed of hollow, elongated and spindle-shaped cells that are arranged in such a way that they lie parallel to each other in the direction of the tree trunk. The wood is, therefore, essentially fibrous in nature and the arrangement and characteristics of these fibrous cells affect wood properties such as strength and stiffness. Therefore, it determines how useful wood can be for a variety of applications.

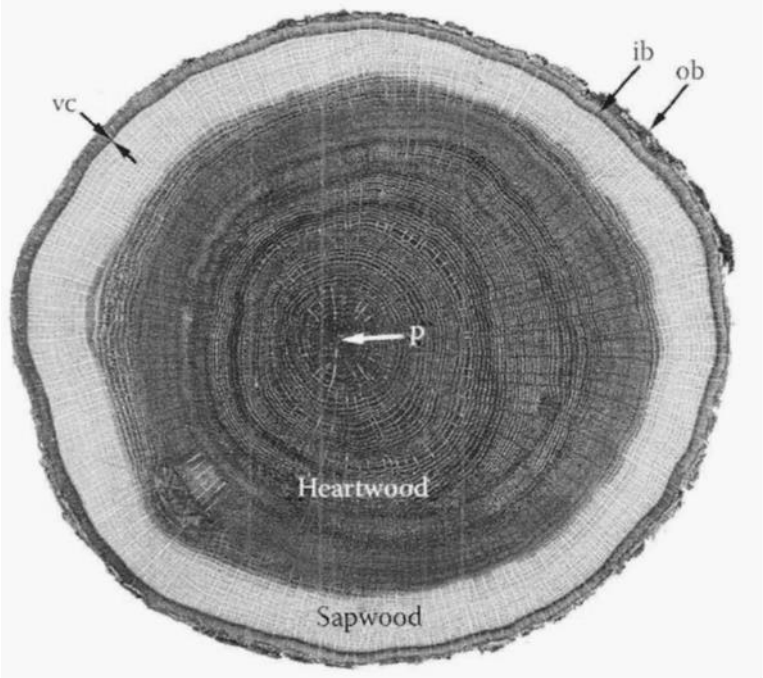


Figure 47: Macroscopic view of a transverse section of a tree trunk displaying the fours of the wood: bark (ob &ib), sapwood, heartwood, and pith (P) (Wiedenhoeft, 2010).

A.1.1 Hardwoods and Softwoods

All species of trees are classified as either hardwood or softwood, these key types of wood differ in their chemical composition as well as function. Thus, because of their different properties, the various types of wood make them ideal for different purposes. Softwood trees are common as gymnosperms, while hardwoods as angiosperms. Gymnosperms are trees with needles rather than leaves; they may be kept for at least a few years and are commonly found in colder regions of the world, such as spruce and pine trees. Angiosperms are trees with broad leaves; they are the most prevalent species of tresses and may be found in most parts of the world. Eucalyptus and red cedar are two examples of angiosperms (Henriksson et al., 2009).

The terms softwood and hardwood refer to the water-conducting cells of a living tree rather than the wood's softness or hardness. Softwoods have a simple structure, and their water-conducting cells are called xylem tracheids, and they have a tapering shape, whereas hardwoods have a more complicated structure, and their water-conducting cells are called xylem vessels, and they are tubular-shaped (Forests NSW, 2008). Softwoods provide long and strong fibres that lend strength to paper and are used for boxes and packaging, whereas hardwoods produce shorter fibres, resulting in a weak paper that is thinner, opaque, and better suited for printing.

A.1.2 Cell wall structure

It is critical to understand the structure of the wood cells to understand where the various chemical components that influence the process are located. This is significant because the reactants in the cooking liquor would have to enter the wood and hence its cells to reach the right chemical components within the cell (Kilian,1999). The chemical composition of the wood will be examined in the following section, but the major components must be identified. The plant cells in wood are composed of cellulose and hemicellulose, which are held together by lignin. Cellulose is a crucial structural component of plant cell walls, while lignin holds these cells together.

The majority of the characteristics of wood are provided by its cell walls. The cell walls have a regular shape and are composed of numerous layers, which all work together to provide mechanical support and allow the circulation of water and biochemicals throughout the plant. As illustrated in Figure 48, the cell wall is divided into three major regions: the central lamella, the primary wall, and the secondary wall. The cell wall comprises many important components for each of these regions: cellulose microfibrils, hemicelluloses, and a matrix or encrusting substance, often pectin in primary walls and lignin in secondary walls. Microfibrils are groups of cellulose molecules that are extremely long, strong, and thin (Wiedenhoeft, 2010).

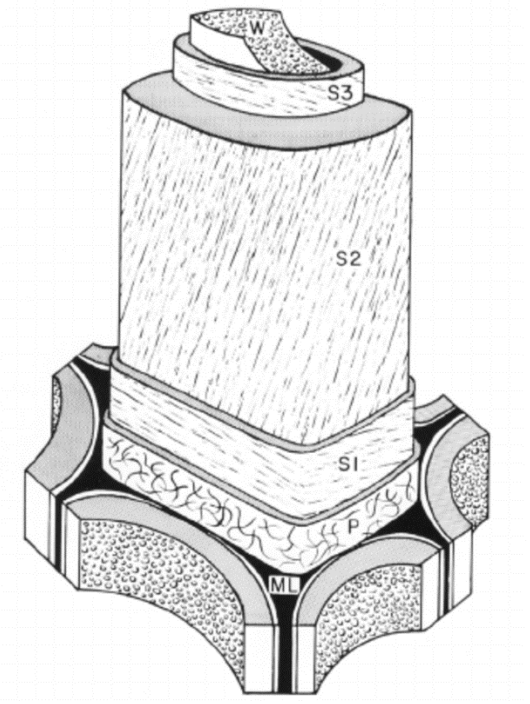


Figure 48: A simplified structure of a woody cell, displaying the middle lamella (ML) and cell wall layers. (P, S1, S2 and S3) (Côté, 1967)

The middle lamella is the outer layer of the cell wall continuum that is placed between the cells and acts to bind the cells together, allowing water and biochemicals to pass between the cells. This layer is rich in pectin in non-woody organs, whereas the intermediate lamella is lignified in wood. The primary wall is created as the next layer, just inside the middle lamella. The vast, random orientation of cellulose microfibrils defines the primary wall. This thin layer is made up of cellulose, hemicellulose, and pectin, and is entirely encased in lignin. Since it is fundamentally comparable to the middle lamella, it is referred to as the middle lamella compound, which includes the middle lamella and the primary wall on both sides (Wiedenhoeft, 2010).

The secondary wall is the last cell wall region and is found in almost all wood cells. The secondary wall is made up of three layers that form after growth. The S1 layer is the secondary layer's outer layer; it is thin, lignin-rich, and closely resembles the primary wall to which it is intimately linked. The primary secondary wall, the S2 layer, is almost securely linked to the S1 layer (Mollereau, 2005). The S2 layer is the thickest secondary wall layer and possibly the most important cell wall layer in determining cell properties and, hence, wood properties on a macroscopic scale. The S3 layer, which is hemicellulose rich, thin, and surrounds the central canal, is the interior of the S2 layer (Mollereau, 2005).

In the pulping industry, it is critical to understand the distribution of the essential chemical components in wood. This is necessary for the pulping reactants to permeate the cell walls and reach the lignin to dissolve it (Kilian,1999). The distribution of wood's chemical components varies from layer to layer, and the usual distribution of these substances in cell walls is depicted in Figure 49.

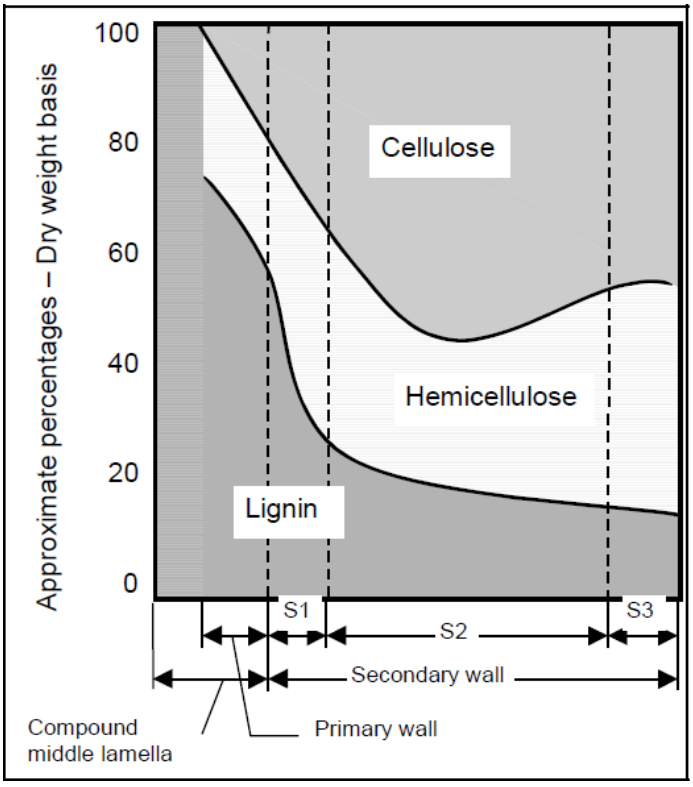


Figure 49: Chemical components distribution in the woody cell wall (Kilian, 1999).

According to Figure 49, the intermediate lamella and primary wall of the wood cell have a larger percentage of lignin. The middle lamella, on the other hand, is more lignin-rich than the secondary wall (approximately 55% of the ML material), but because the secondary wall eventually covers a considerable percentage of the wood fibre cell wall, it will have the highest total amount of lignin (Kilian,1999). The cooking duration can be adjusted to enable enough time for the reactants to diffuse through the cell walls, depending on the grade of pulp desired (Kilian,1999).

A.2 Chemical composition of wood

Wood is made up of three elements: carbon, hydrogen, and oxygen. Carbon and oxygen are the most abundant elements, accounting for around 49 and 44% of the total weight. The remaining 7% is primarily hydrogen, with minor amounts of nitrogen and metallic ions (Youngs, 2009). Wood's organic components can be divided into two categories: macromolecular components and low molecular mass components. Macromolecular components include polysaccharides such as cellulose, hemicellulose, and lignin, whilst low molecular components include organic and inorganic components known as extractable and ash, respectively (Keskin-Schneider, 1991).

However, as previously said, wood is not a homogenous substance and is composed of a number of chemical components that vary in quantity with species, within species, and within the tree trunk and cell walls (Kilian, 1999). The primary interest in papermaking is cellulose fibre, which accounts for 40 to 50% of the total dry wood mass, while hemicellulose accounts for 20 to 30% of the total dry wood mass (Holm & Niklasson, 2018). Lignin binds the fibres tightly together within the wood, creating strength.

A.2.1 Cellulose

Cellulose is the most abundant organic biopolymer on the planet. The amount of cellulose in wood differs between softwoods and hardwoods, as well as across different portions of the tree. It is the major component of tree cell walls and an important component due to the impact it has on wood characteristics. Polymerization of glucose into long linear chained polymers results in the formation of cellulose. Cellulose has the chemical formula $(C_6 H_{10} O_5)_n$, where n is the degree of polymerization (DP) (Rydholm,1965). Depending on the type of wood, the number of glucose units in the cellulose chain can range from 10 000 to 20 000, although the normal DP of wood cellulose is about 8000 units. Figure 50 depicts the chemical structure of a cellulose molecule.

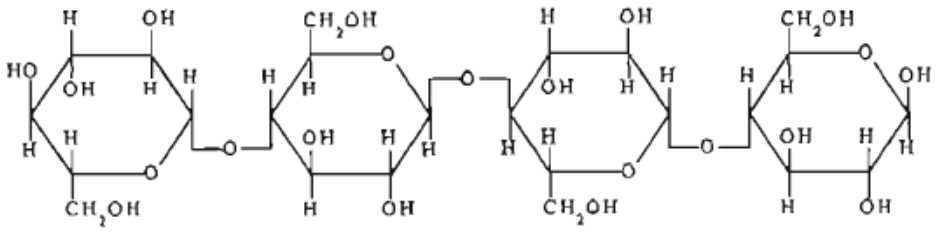


Figure 50: Chemical structure of cellulose (Smook,1992).

The molecular structure of cellulose gives it unique features like degradability, chirality, and hydrophilicity. Due to its high tensile strength and the use of dilute acids, cellulose decomposes easily to generate simple sugars despite being insoluble in water and a few organic solvents. The cellulose structure's polymeric links are such that the chains form in an extended fashion, allowing the molecule to fit tightly together, permitting the formation of tremendous associative forces responsible for the immense strength of cellulosic materials (Mollereau, 2005).

A.2.2 Hemicellulose

Wood is made up of various hemicelluloses, which work together to form one of the most important types of wood components. Hemicelluloses are non-cellulosic polysaccharides

found in wood that provide structural support for the plant's cell wall. These polymers have a low molecular weight, are branched, and are made up of a variety of pentose and hexose sugar monomers. Hemicellulose has a much lower degree of polymerization, ranging from 50 to 300 and has an entirely amorphous structure (Youngs, 2009). Figure 51 depicts the composition of hemicellulose, which includes xylan, glucuronoxylan, arabinoxylan, glucomannan, and xyloglucan. Since the polymer is easily soluble in alkali and hydrolyzed by acids, hemicelluloses are less resistant to deterioration than cellulose during chemical pulping. The relative number of hemicelluloses in wood depends on the wood type and whether the plant is stressed, i.e., compressed or tensed (Teleman, 2009).

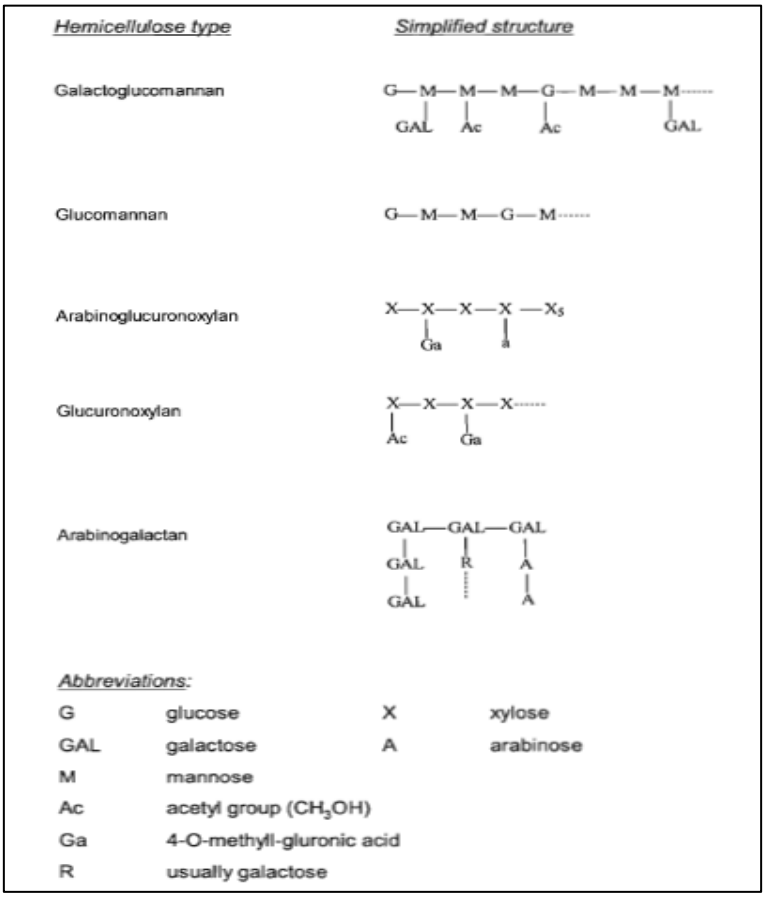


Figure 51: The simplified types of major hemicelluloses in wood (Ingruber, Kocurek and Wong, 1985).

Appendix B: Selected Models' training and optimization results

The performance results of the selected models in Chapter 5 are documented. This is shown in Figure 52 to Figure 63. These plots simply provide a visualization of how the selected models performed before and after optimization.

B.1 Initial Concrete slump

To predict the initial concrete slump value, a Neural net with 10 input variables from datasets 1 and 2 was selected.

B.1.1 Validation results

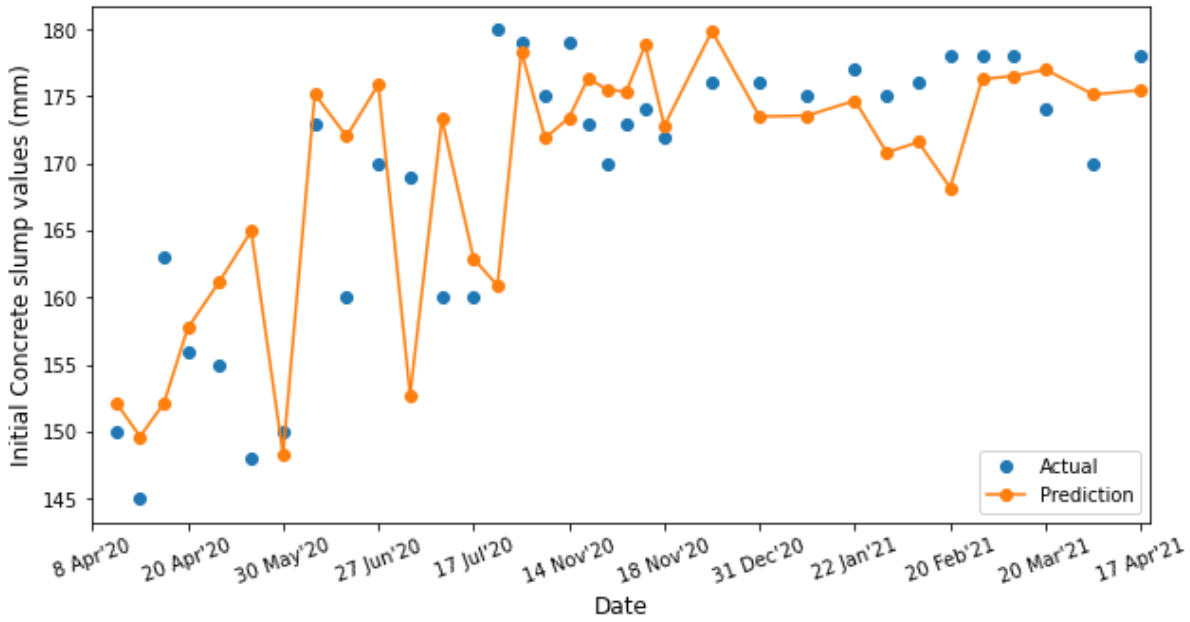


Figure 52: Initial slump validation on training set 1 &2.

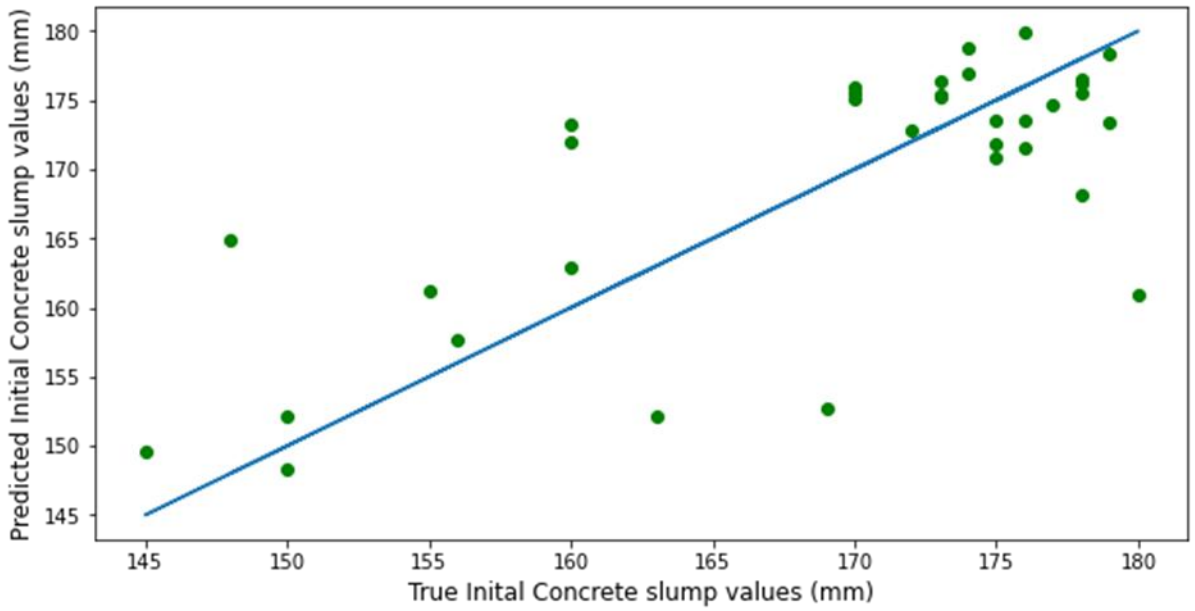


Figure 53: Actual Vs Predicted values for validation performance for initial slump using the training set

B.1.2 Optimization results

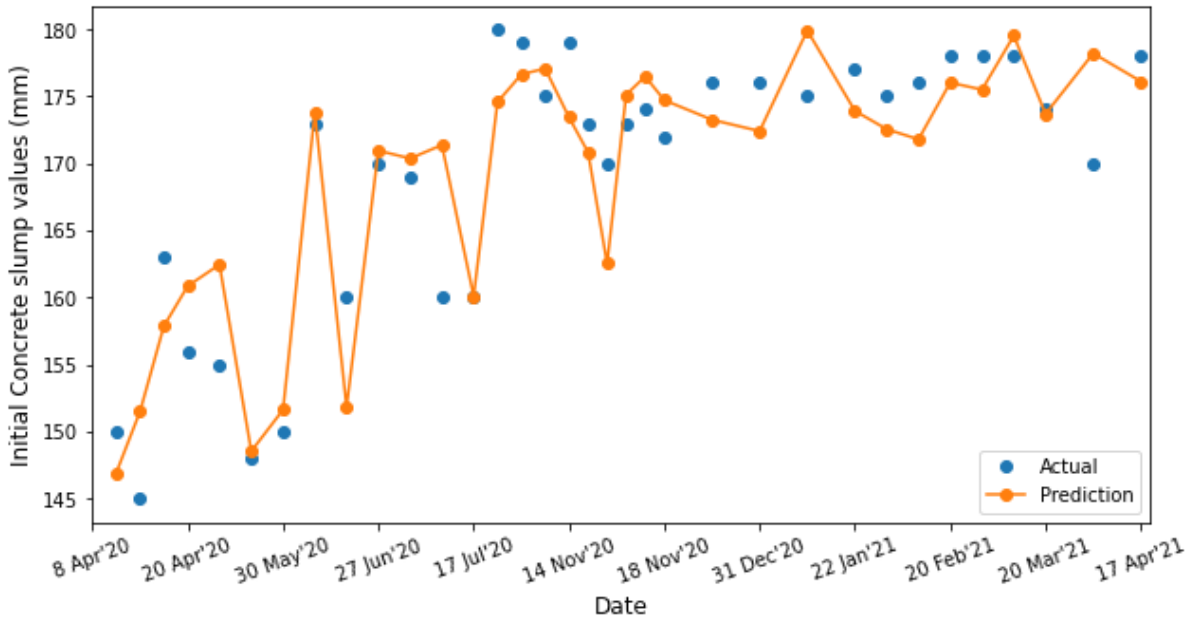


Figure 54: Initial slump optimization result using training set 1 &2.

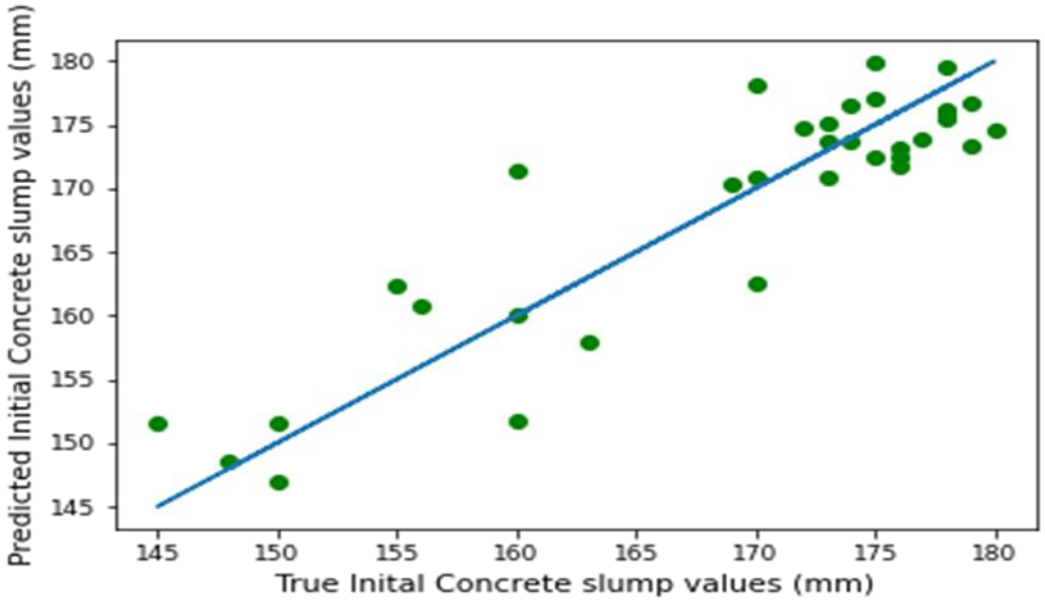


Figure 55: Actual Vs Predicted values obtained when optimizing initial slump

B.2 Dispersion Index

To estimate the dispersion index values, a Random Forest model with 20 input variables using dataset 1 was selected based on evaluation performance.

B.2.1 Validation results

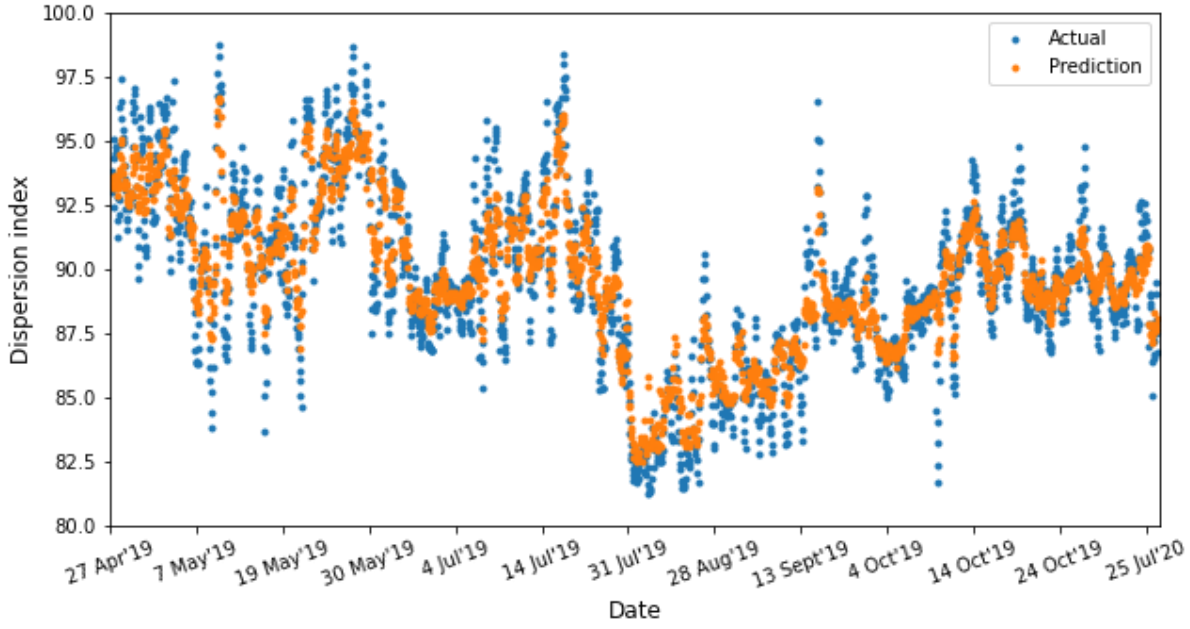


Figure 56: Random Forest prediction of Dispersion index prediction with training set 1.

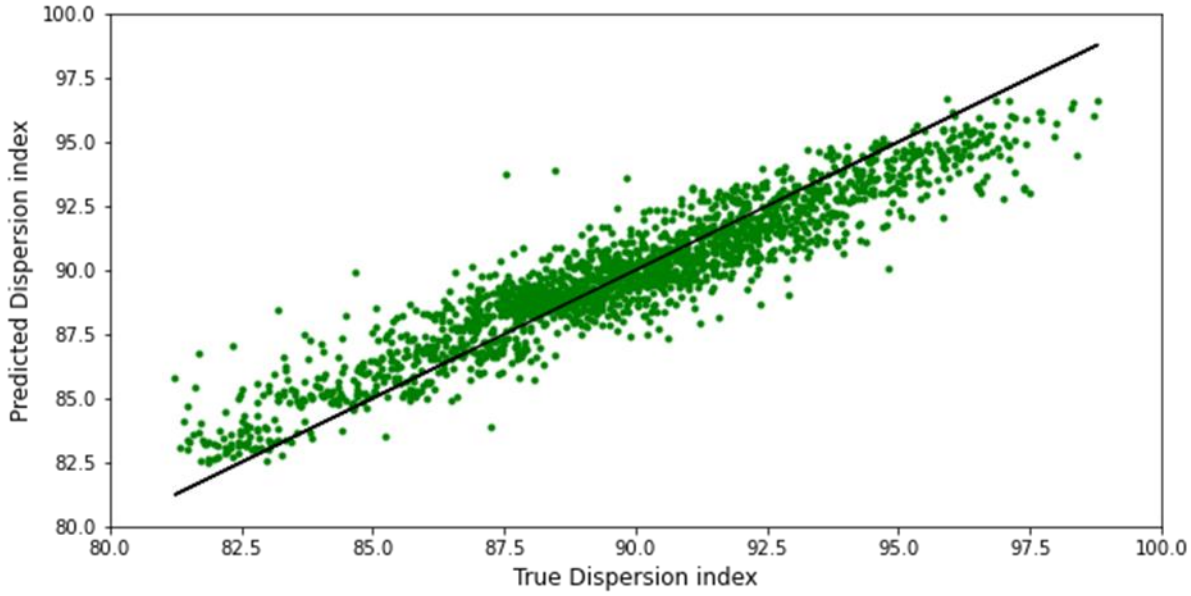


Figure 57: Validation results of Actual Vs Predicted values for dispersion index.

B.2.2 Optimization results

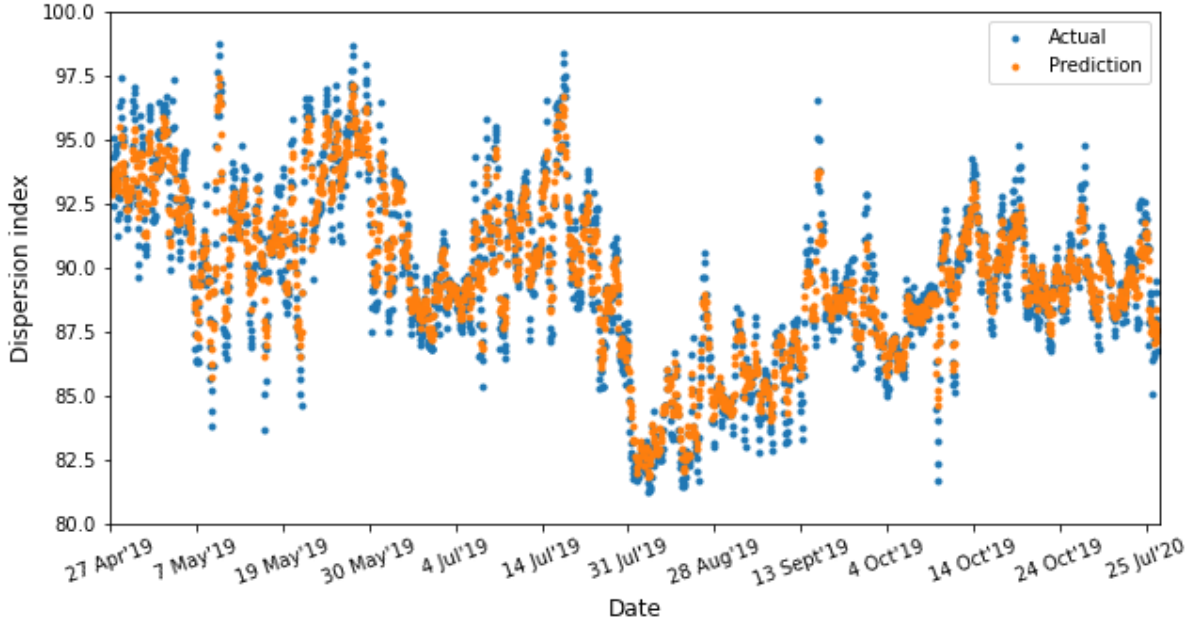


Figure 58: Optimization of Random Forest model in predicting dispersion index

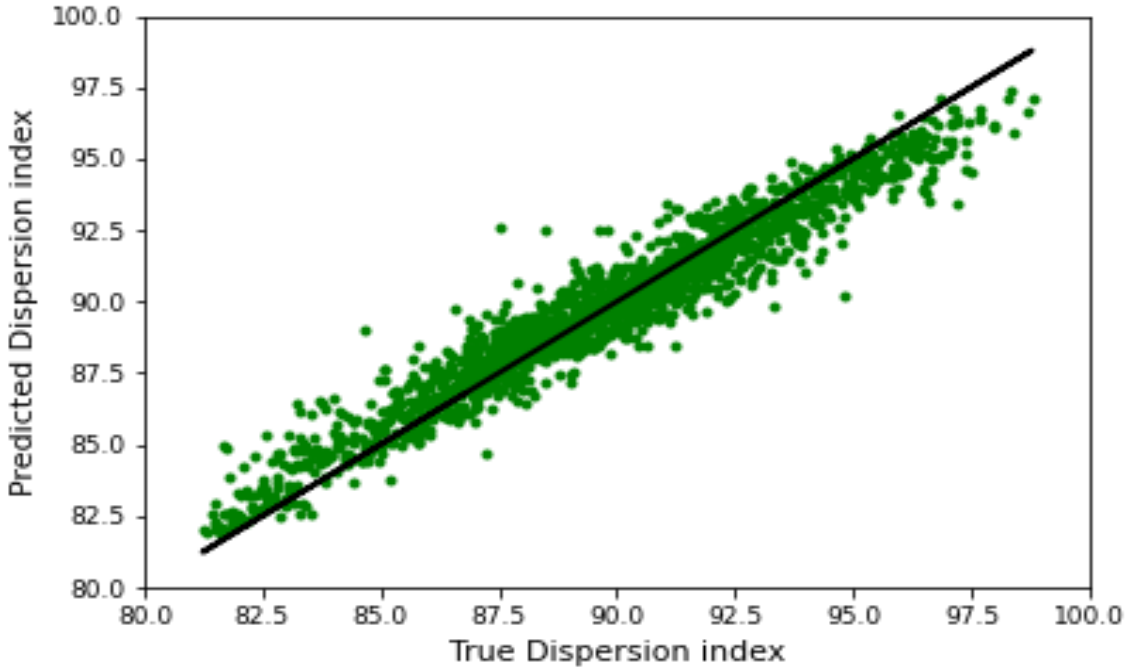


Figure 59: Actual vs Prediction optimization results for selected dispersion index model.

B.3 Insoluble content

A Random Forest model, using 20 input variables from dataset 2, was selected to predict insoluble content values of the lignosulphonate product.

B.3.1 Validation results

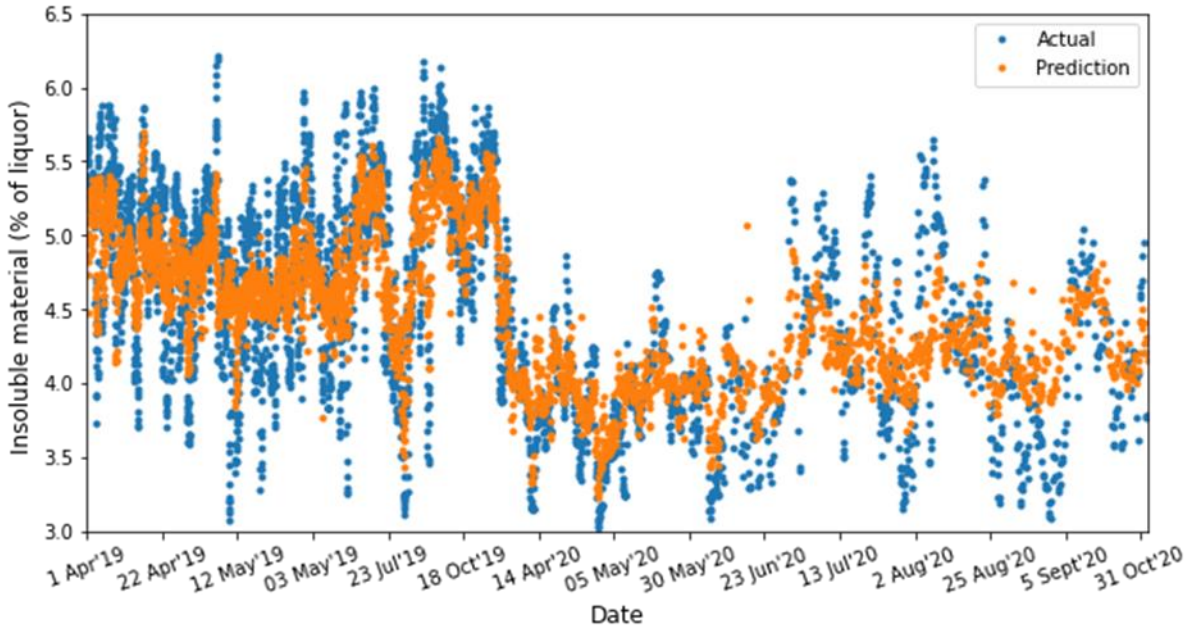


Figure 60: Validation result of the selected insoluble content model

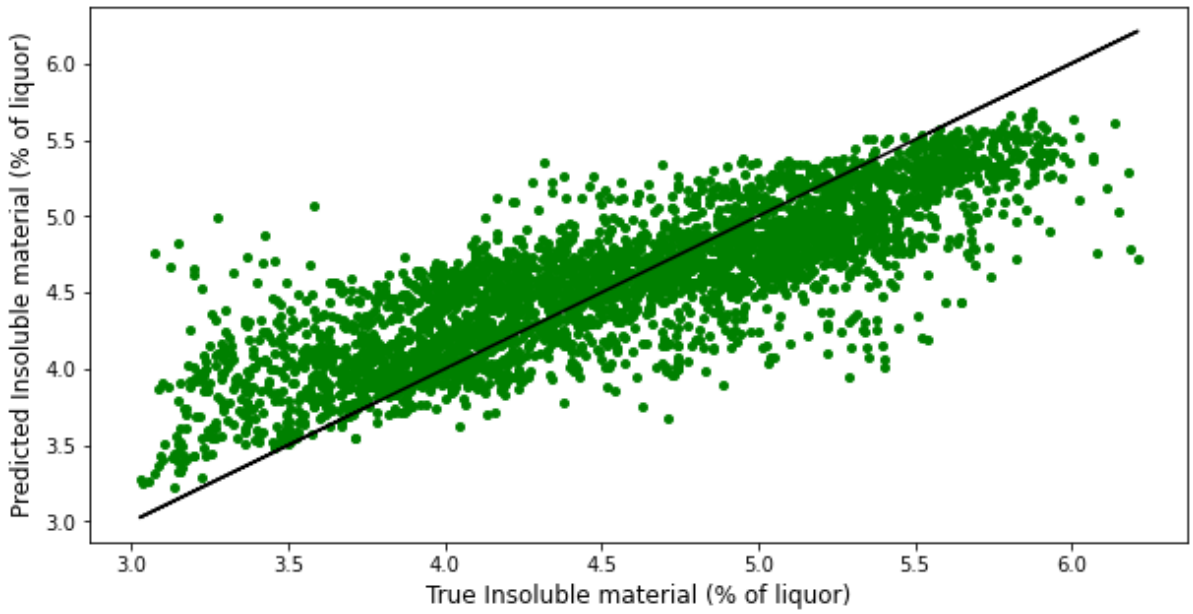


Figure 61: Actual vs Prediction result of the insoluble content model using dataset 2.

B.3.2 Optimization results

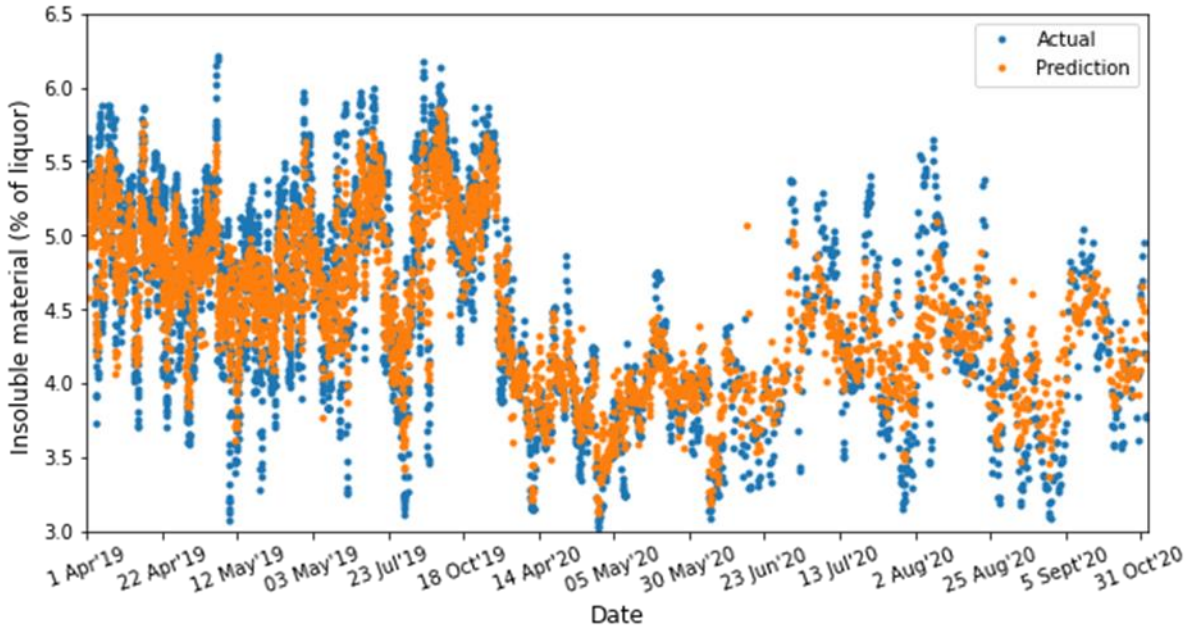


Figure 62: Optimization result of the insoluble content model.

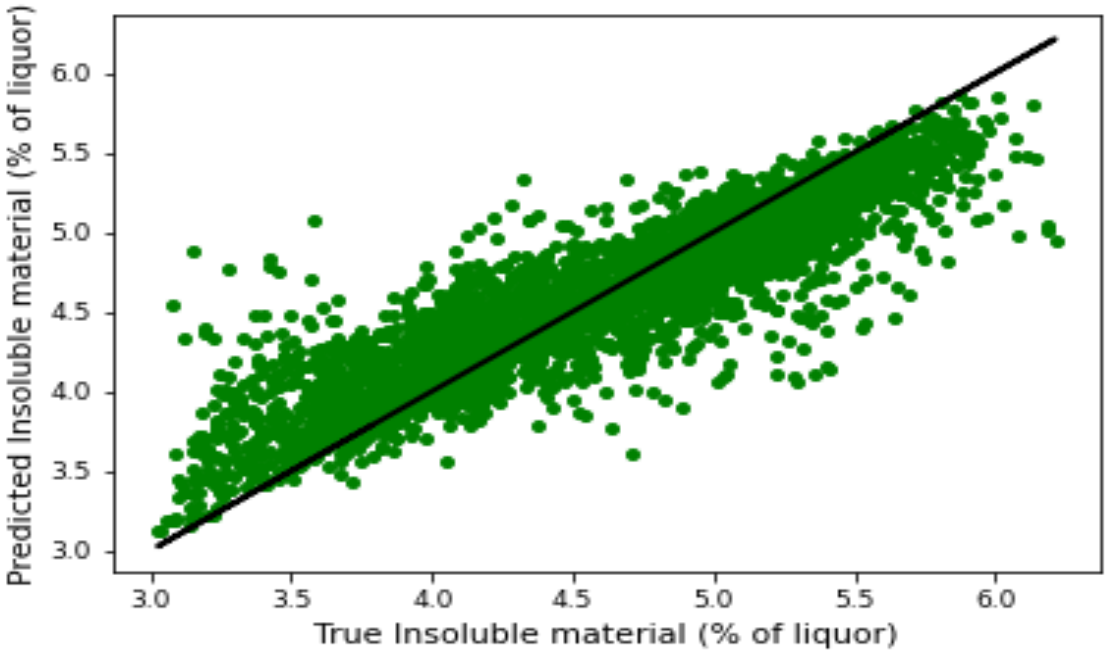


Figure 63: Actual vs Prediction of the optimized insoluble content model